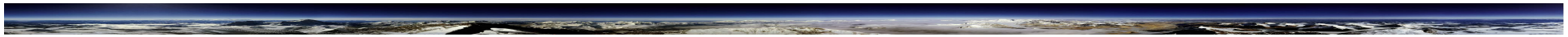
Une introduction à la **fouille de données**

Illustrations en *agronomie*,

alimentation, sciences de la vie …

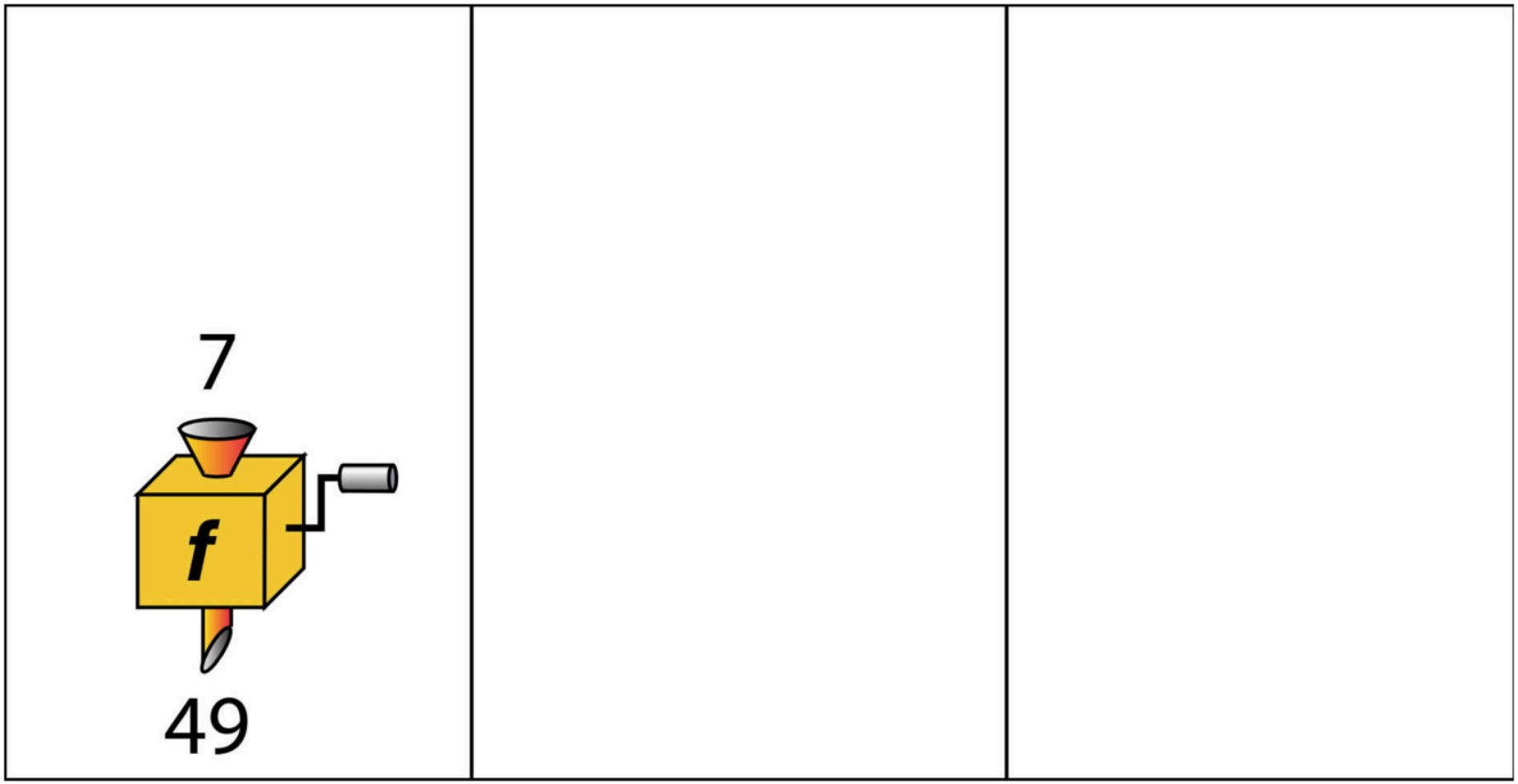


Antoine Cornuéjols

AgroParisTech – INRA MIA 518

antoine.cornuejols@agroparistech.fr

Algorithmes … et … apprentissage



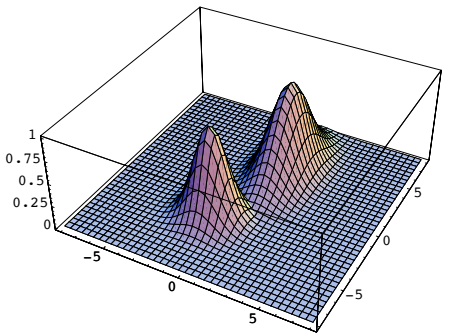
Plan

1. Grands types d'apprentissage
2. Méthodes d'apprentissage
3. En pratique
4. Ce que l'on sait faire et les défis à relever
5. L'IA : une révolution ?

Apprentissage descriptif

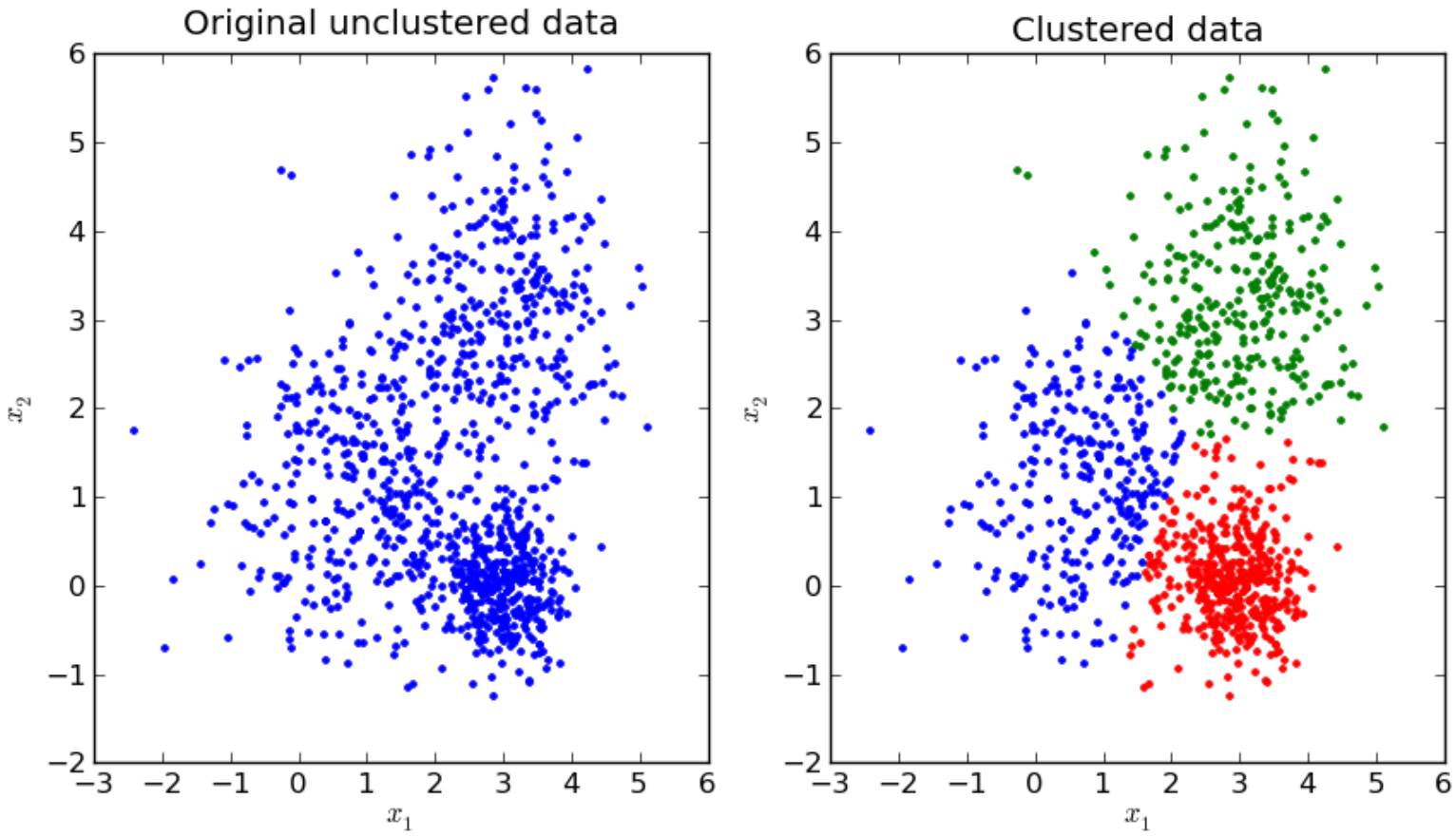
non supervisé

- Catégorisation de consommateurs
  - Base de données sur les répondants de la base Nutrinet
    - ~ 280 000
    - Données sur *âge*, *nb de personnes dans la famille*, *catégorie socio-professionnelle*, …
    - Données sur consommations alimentaires sur une certaine durée
  - Y a-t-il émergence de **groupes** distincts ?



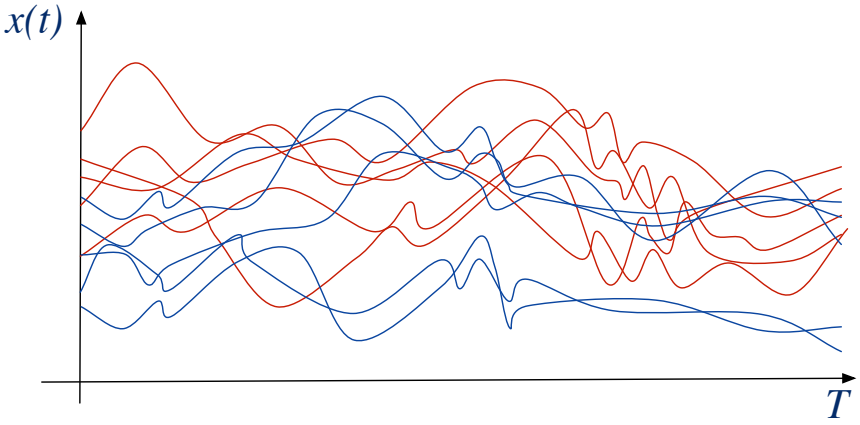
Les **grands types** d'apprentissage

- Apprentissage « **descriptif** » (non supervisé)



- Catégorisation des dynamiques de populations d'insectes ravageurs
  - **Bases de données** sur des courbes d'évolution (éventuellement multi-variées)

Y a-t-il émergence de types d'insectes ravageurs ?



Clustering / Catégorisation

- Extraire des **régularités**
  - Base de données sur les **consommations alimentaires**
  - Peut-on identifier des « patterns » de consommation ?
  - Des « motifs » fréquents
    - (pizza, coca-cola, glace)
    - (steak, frites, vin)
    - (poisson, haricots verts, eau minérale)

Recherche de motifs fréquents

Frequent Item Sets

Recherche de règles d'association

Apprentissage descriptif

À propos d'un *échantillon d'apprentissage* $s = \{(x_i)\}_{1,m}$

identifier des régularités rendant compte de *S*

- E.g. sous la forme de **clusters** (e.g. *mélange de Gaussiennes*)
  - CLUSTERING
- E.g. sous la forme de **motifs fréquents** (fouille de données)

pour résumer, suggérer des régularités, comprendre …

Apprentissage prédictif supervisé

Apprentissage prédictif (*supervisé*)

- Un *échantillon d'apprentissage*

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$$

f

h

Prédiction pour de **nouveaux** exemples x –*h*-> y ?

- Reconnaissance d'insectes ravageurs
  - Base d'images d'insectes dans des cuvettes
    - *Reconnaissance du type d'insectes*
    - *Comptage*

- Reconnaissance d'animaux malades ou en chaleur

  - Mesures en continu sur leur comportement

    - Vidéos
    - Capteurs « embarqués »

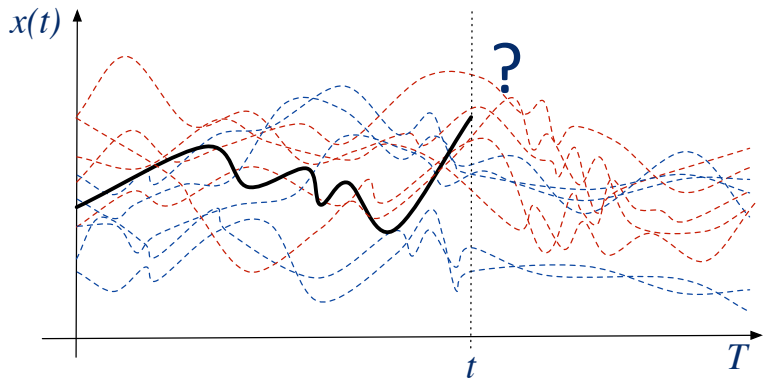
    - Mobilité (nb de pas / minute ; distance parcourue à l'heure)
    - Lieux visités
    - …

    ➢ *Reconnaissance de **comportements types***

- Prédiction de consommation d'aliments protéinés

  - Historique de consommations

    - *Prédictions à 3 mois, 1 an, 3 ans, …*



Apprentissage prescriptif

pour « intervenir »

Apprentissage prescriptif

- Apprentissage « **prescriptif** » (recherche de ***causalités***)

1. J'observe que les gens qui mangent des glaces
   sont souvent en maillot de bain
2. Je voudrais vendre davantage de glaces

→ Je demande aux gens de se mettre en maillot de bain

- Quelles **recommandations** faire à un consommateur pour qu'il baisse sa consommation d'aliments carnés ?

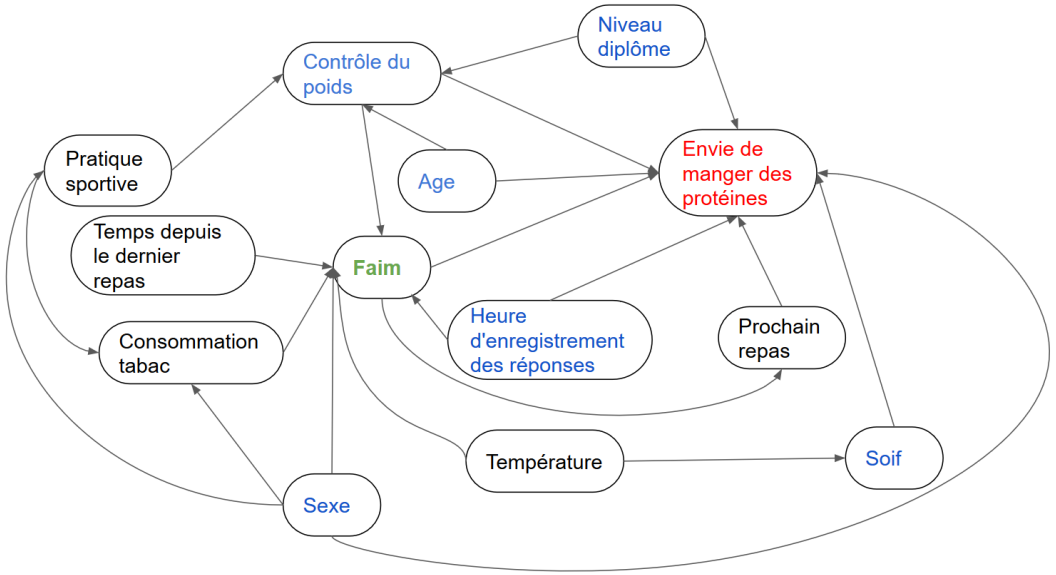
- Quel impact **si on double le prix** de … ?

- Quel rendement aurais-je eu l'année dernière **si j'avais** planté du ... au lieu de ...

La recherche de relations causales

Qu'est-ce qui **cause** l'appétence pour des plats protéinés ?

- La **faim** ?
- **L'heure** dans la journée ?
- Le **genre** ?
- L'aspect **visuel** ?
- L'aspect **olfactif** ?
- La richesse en **protéines** des **repas précédents** ?
- …

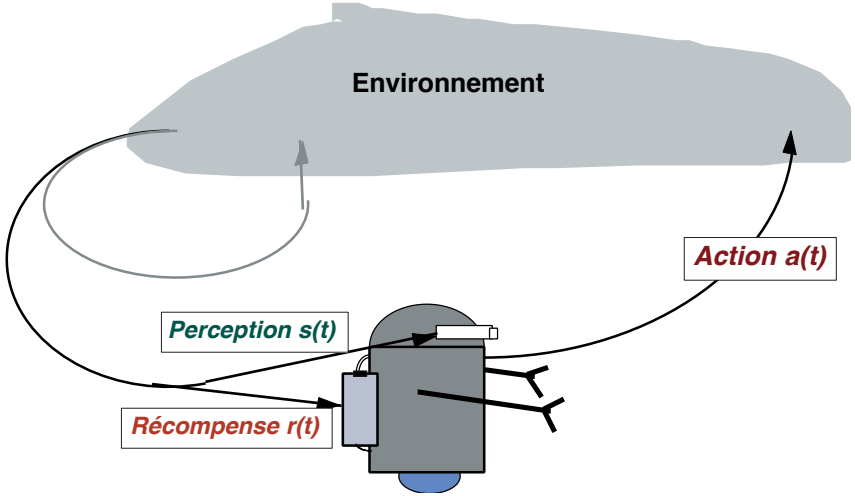


Apprentissage de contrôle

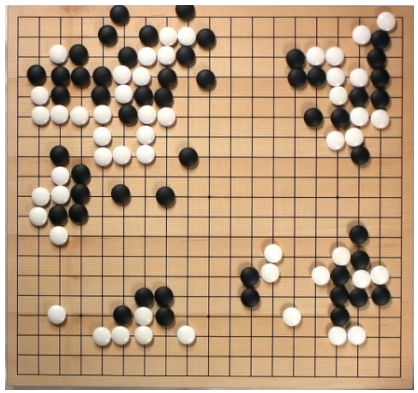
par renforcement

Apprentissage « par **renforcement** »

- comment (ré)agir



1. Contrôler une **ferme**
2. Reinforcement learning for **developing country agriculture** (Colombie, thèse de doctorat)



- L'agriculture numérique

Plan

1. Grands types d'apprentissage
2. Méthodes d'apprentissage
3. En pratique
4. Ce que l'on sait faire et les défis à relever
5. L'IA : une révolution ?

Méthodes pour l'apprentissage descriptif

Ce que l'on cherche

1. Découvrir des régularités
2. Comprimer l'information / ré-exprimer les données

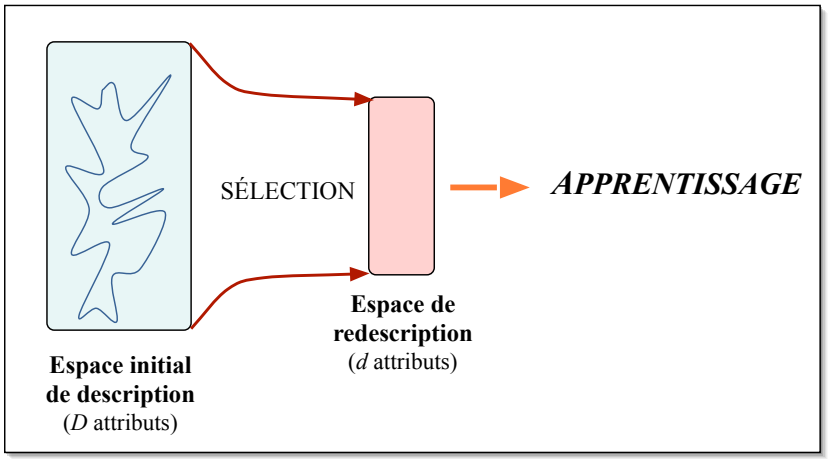
Ce que l'on cherche

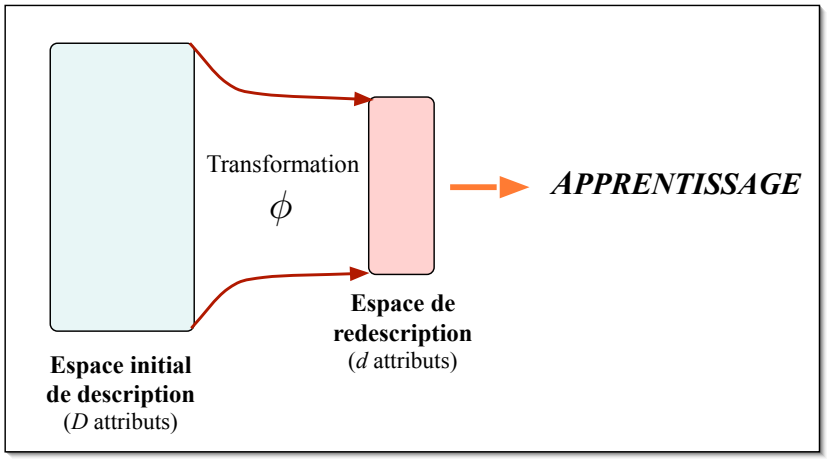
1. Découvrir des régularités
2. Comprimer l'information / ré-exprimer les données

Des **méthodes**

1. Changer l'espace de **représentation**
   - Diverses analyses en composantes : ACP, ACI, NMF, … (LVQ)
   - Sélection d'attributs

Changement d'espace de description

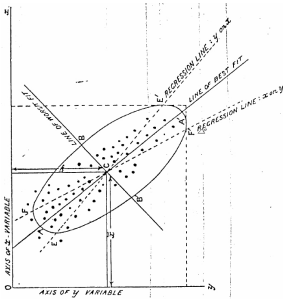




Sélection de descripteurs

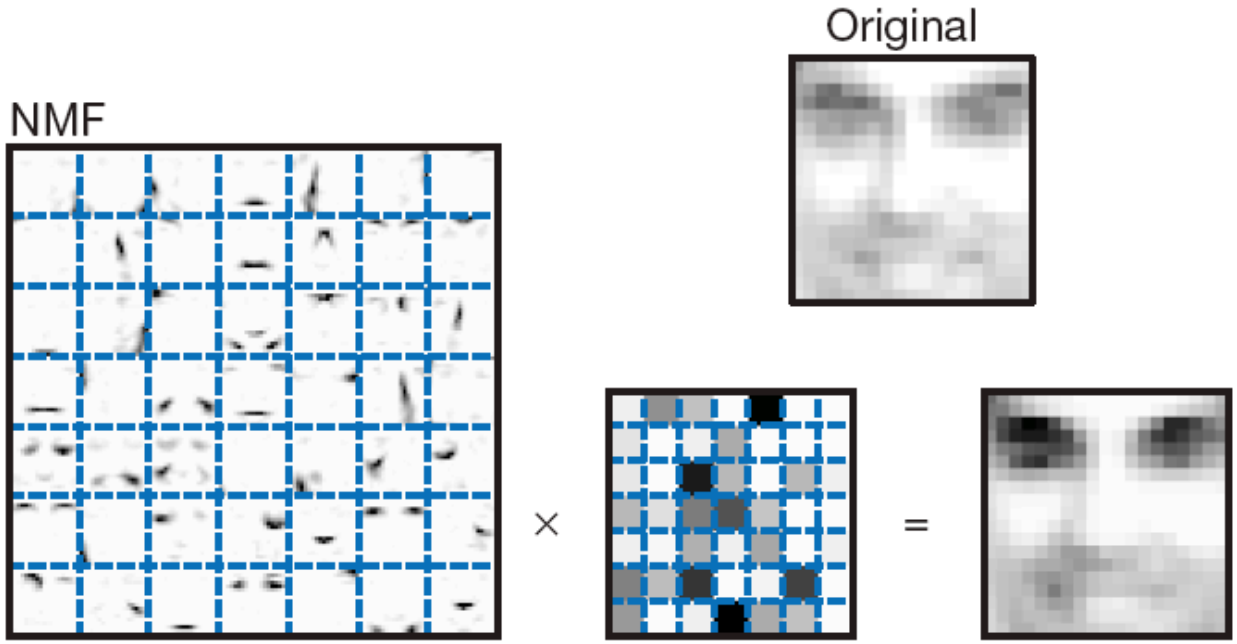
Changement de descripteurs

- ACP
- ICA
- NMF



Changement d'**espace de représentation**

La Non Negative Matrix Factorization (NMF)



Ce que l'on cherche

1. Découvrir des régularités
2. Comprimer l'information / ré-exprimer les données

Des **méthodes**

1. Changer l'espace de **représentation**
   - Diverses analyses en composantes : ACP, ACI, NMF, … (LVQ ?)
   - Sélection d'attributs
2. Mettre en évidence des « **patterns** »
   - Clustering / Mélanges de Gaussiennes
   - SOM (Self-Organizing Maps) / GTM (Graphical Topographic Mapping)
   - Motifs fréquents / règles d'association

Exemple

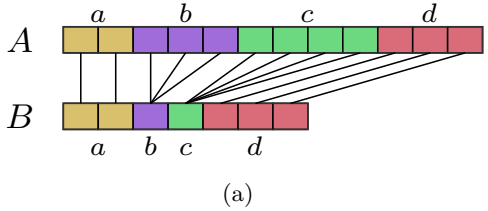
Déterminer des **classes de consommateurs** à partir de leurs **séquences** de *consommation d'aliments*

S_{238} = <(*thé, croissant*)$_1$, (*brocoli, pâtes, vin*)$_2$, …, (*pizza, soda*)$_t$>

– Il faut **une mesure de "distance"**

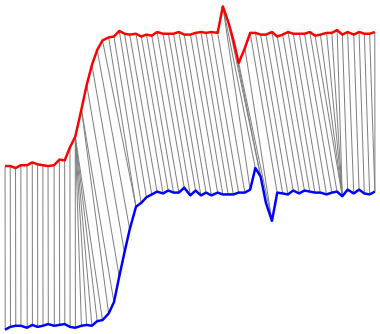
Exemple

Alignement de **séquences**



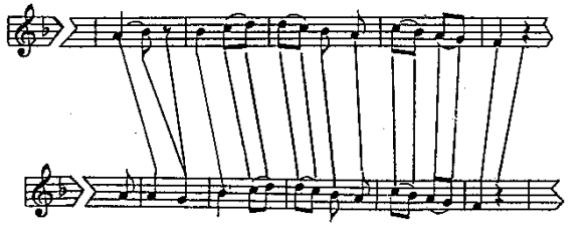
(a)

Exemple d'alignement de deux séquences **symboliques**



(b)

Exemple d'alignement de deux **séries temporelles**



(b) Exemple d'alignement de deux partitions représentant deux versions de l'Alleluia de Mozart (Mongeau et Sankoff, 1990).

Tiré de [Germain Forestier HDR (2017), p.38, 39]

Exemple

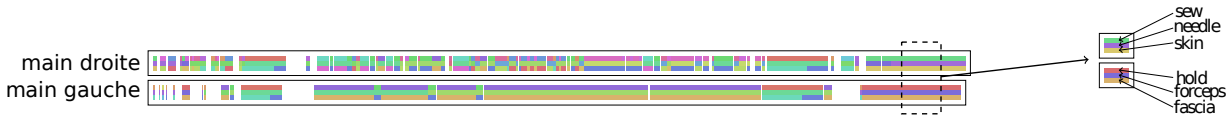
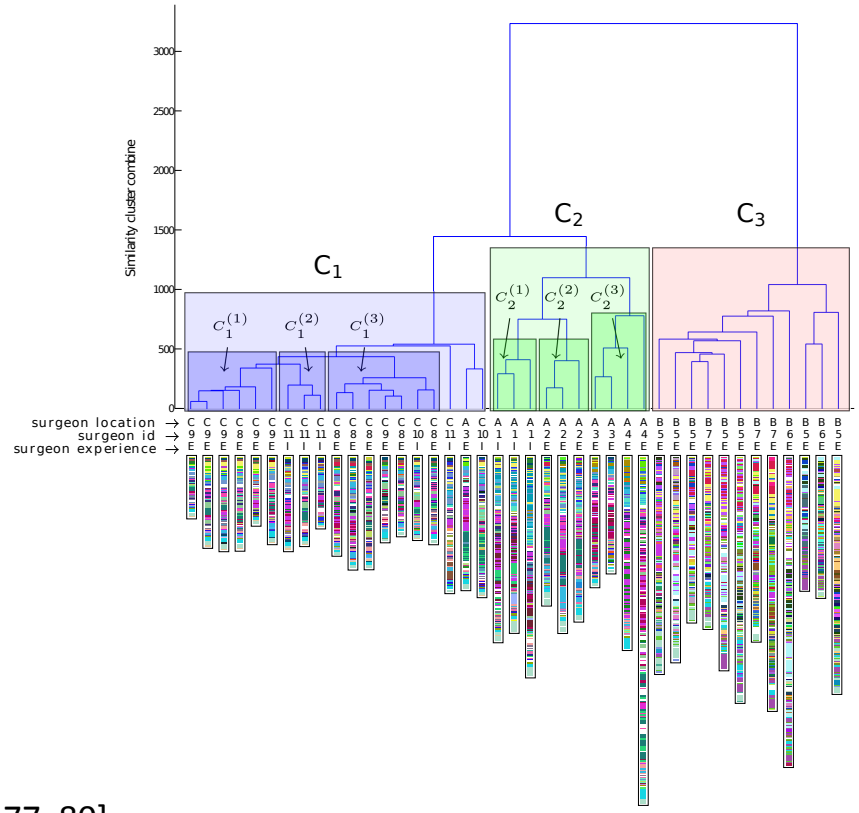


FIGURE 3.2 – Exemple de séquences d'activités chirurgicales avec les mains droite et gauche du chirurgien.



Tiré de [Germain Forestier HDR (2017), p.77, 80]

Exemple

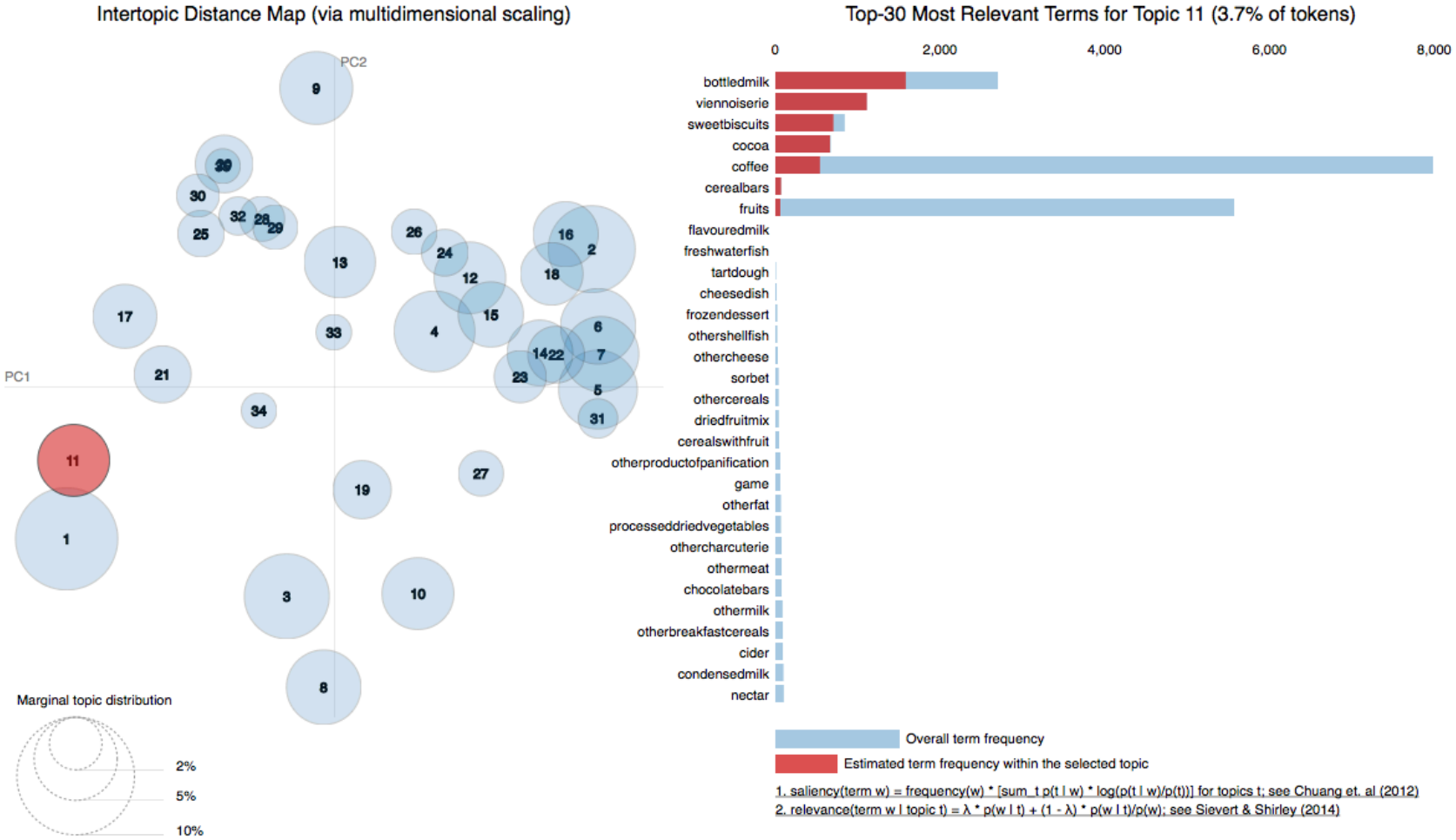
Déterminer des **classes de consommateurs** à partir de leurs **séquences** de *consommation d'aliments*

S_{238} = <(*thé, croissant*)$_1$, (*brocoli, pâtes, vin*)$_2$, …, (*pizza, soda*)$_t$>

- Il faut une mesure de "distance"
  - **Difficile** car distance entre séquences de multi-sets de taille variable
  - Une solution :
    changer d'espace de représentation pour arriver dans un **espace vectoriel**
  - Espace de **variables latentes**
    - Un repas est un mélange de topics
      (comme un document est un mélange de thèmes)

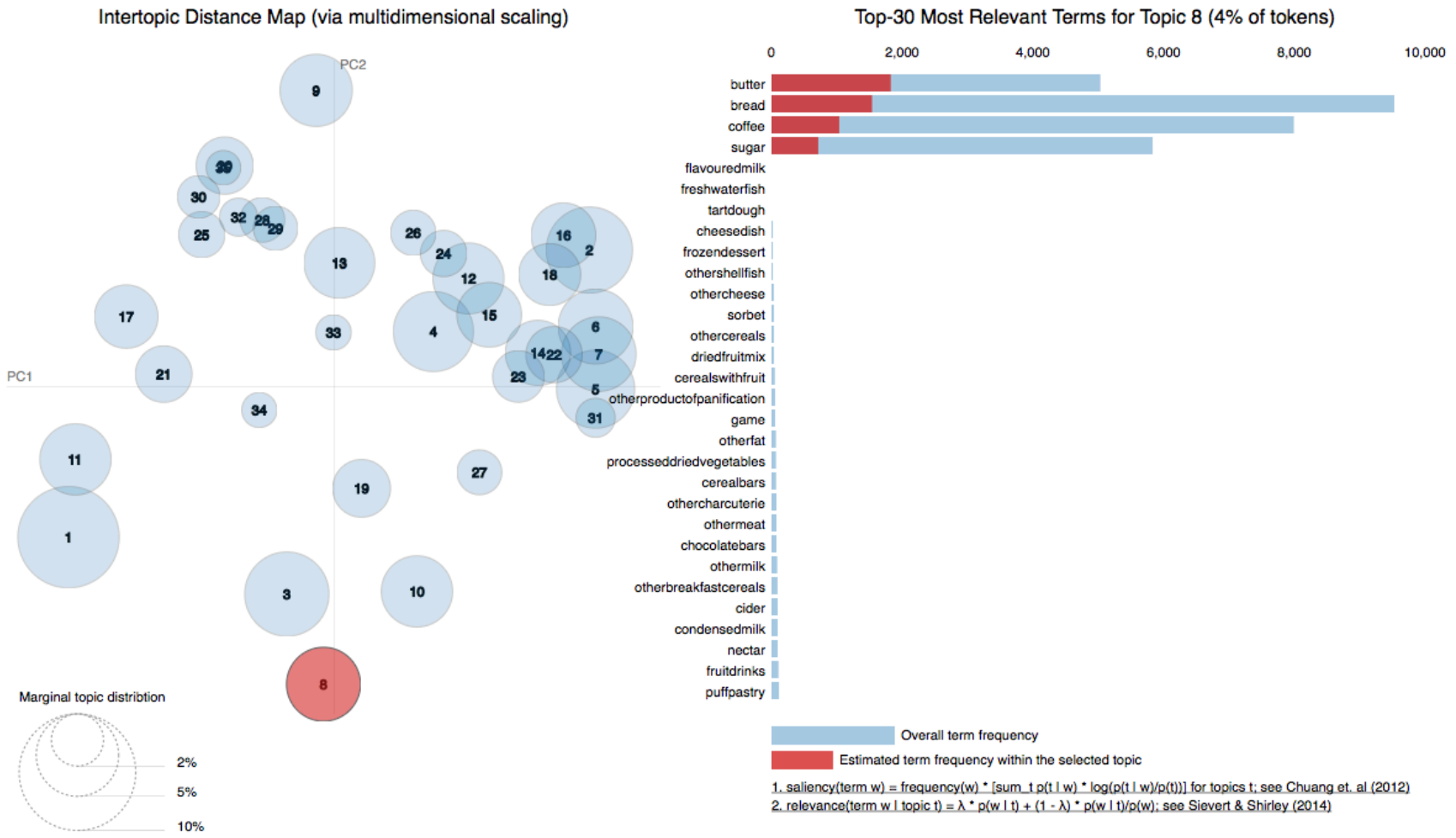
Exemple (suite)

LDA



Exemple (suite)

LDA



Exemple (suite)

En utilisant l'analyse latente de Dirichlet (LDA)

LDA avec K = 45 topics

(pain, margarine, sucre, café) → 0,48 % T4
0,51 % T43

(pain, petits pois, bœuf, dessert, eau minérale) → 0,33 T31
0,21 T24
0,18 T35
0,18 T36

Topic 4 :
0,46 margarine
0,18 pain
0,11 café

Topic 31:
0,44 petit pois
0,12 huile
0,12 eau du robinet
0,10 beurre

Topic 43 :
0,56 café
0,41 sucre

Topic 24:
0,28 PDT
0,26 desserts
0,13 crêpes etc

Topic 35 :
0,89 eau minérale
0,10 fruits

Topic 36
0,75 bœuf
0,11 garniture mixte

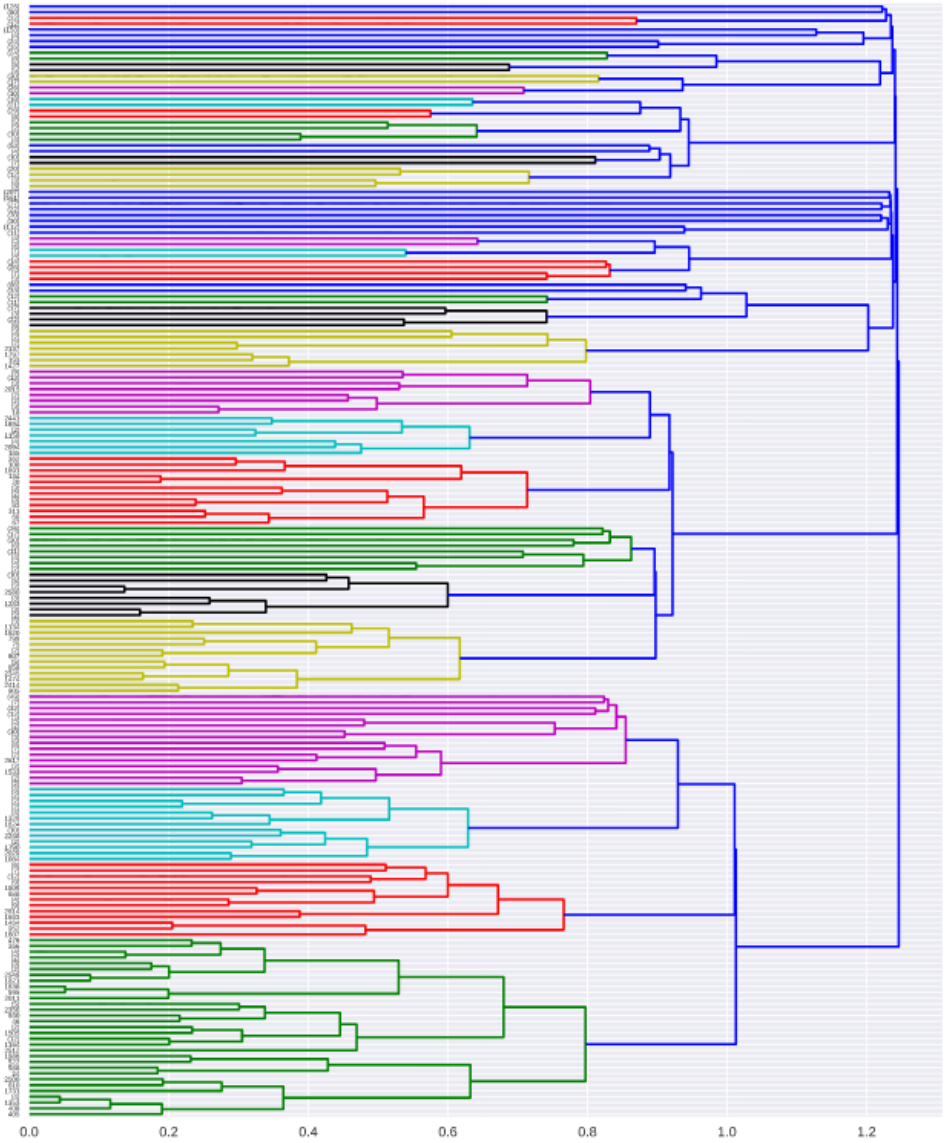
Exemple (suite)

- Recodage des séquences
  - Supposons les topics : A, B, C et D
  - S_{749} = ‹ $(0.6\ A + 0.4\ B)_1$, $(0.02\ B + 0.98\ C)_2$, …, $(0.3\ C + 0{,}7\ D)_t$ ›

- **Distance euclidienne** dans le nouvel espace de représentation
  - Somme de distances entre vecteurs

Exemple (suite)

Classification **A**scendante **H**iérarchique (AHC)

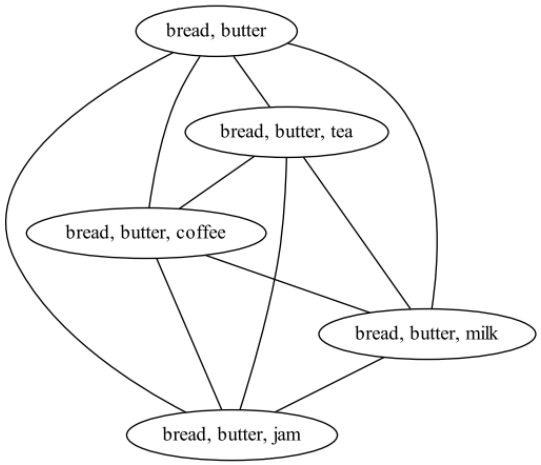


Recommandation personnalisée

Modification des habitudes de consommation

Par recherche de trajectoires de « plus faible coût d'acceptabilité »

Motifs fréquents



[Sema Akkoyunlu et al. (2017),
« *Investigating substitutability of food items in consumption data* ».
Workshop Health in RecSys-2017]]

	Breakfast and lunch		Breakfast		Lunch	
Food Item	Substitute item (ordered by score)	Score	Substitute item (ordered by score)	Score	Substitute item (ordered by score)	Score
Bread	Rusk	0.2234	Rusk	0.3716	Fruits	0.0497
	Viennoiserie	0.1359	Viennoiserie	0.2010	Yogurt	0.0490
	Cakes	0.0745	Cakes	0.1243	Potatoes	0.0468
Coffee	Tea	0.2799	Tea	0.4219	Sodas	0.065
	Cocoa	0.1729	Chicory	0.2550	Yogurt	0.0642
	Chicory	0.1486	Cocoa	0.2255	Fruits	0.0633
Tea	Coffee	0.2799	Coffee	0.4219	Cakes	0.0536
	Cocoa	0.1721	Chicory	0.1965	Viennoiserie	0.0417
	Chicory	0.1289	Cocoa	0.1462	Coffee	0.0412
Cocoa	Chicory	0.2171	Chicory	0.2211	Cereal bars	0.25
	Coffee	0.1729	Coffee	0.2077	Preprocessed vegetables	0.0526
	Tea	0.1289	Tea	0.1965	Hamburgers	0.0256
Butter	Margarine	0.2413	Margarine	0.4030	Margarine	0.0602
	Honey/jam	0.0924	Chocolate spread	0.1240	Fruits	0.0431
	Chocolate spread	0.0786	Honey/jam	0.1175	Sauces	0.0431
Milk	Juice	0.1409	Yogurt	0.1815	Doughnut	0.0869
	Yogurt	0.1264	Juice	0.1504	Other milk	0.0666
	Sugar	0.1089	Tap water	0.1361	Milk in powder	0.0625
Wine	Sodas	0.0814			Sodas	0.0860
	Beer	0.0704	/	/	Tap water	0.0755
	Tap water	0.0412			Beer	0.0746
Pizza	Sandwich baguette	0.2429			Sandwiches baguette	0.2810
	Other sandwiches	0.1729	/	/	Other sandwiches	0.2177
	Meals with pasta or potatoes	0.1513			Meal with pasta or potatoes	0.1658
Potatoes	Pasta	0.1111			Pasta	0.1142
	Green beans	0.0922	/	/	Green beans	0.0941
	Rice	0.0602			Rice	0.0616

Table 2: Top 3 substitutable items for several items for breakfast and lunch

Méthodes pour l'apprentissage prédictif

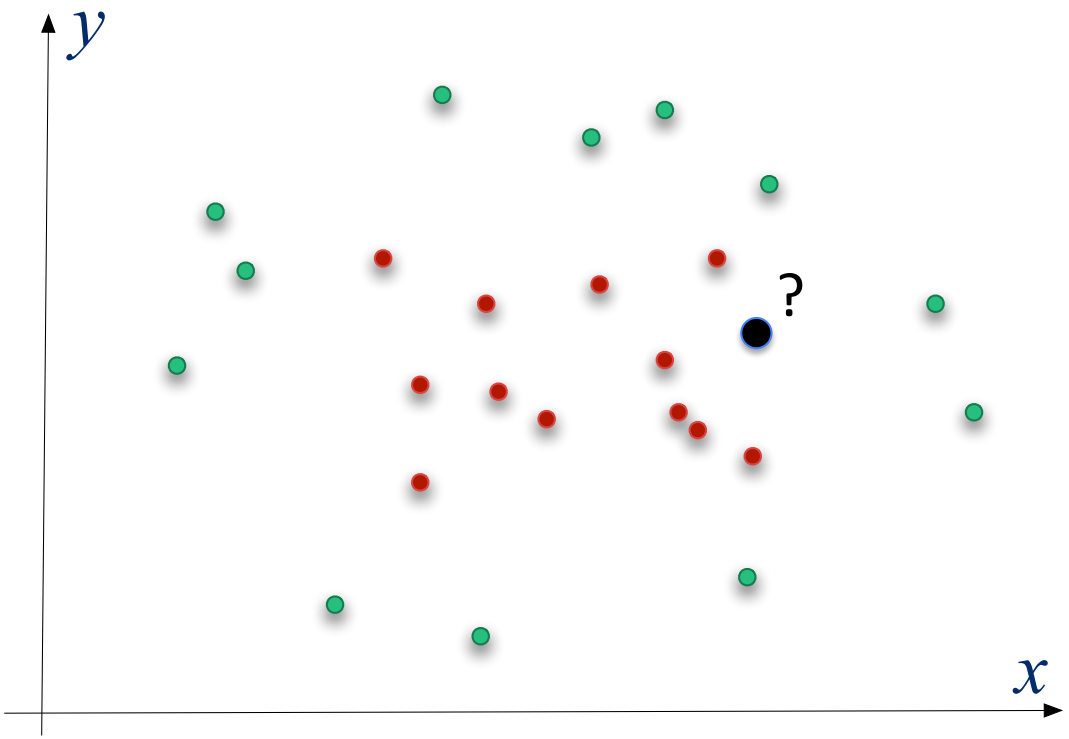
Ce que l'on cherche

1. Des « règles » pour prédire / décider *x* -> *y*
2. Éventuellement **la règle** pour justifier / comprendre

Des **méthodes**

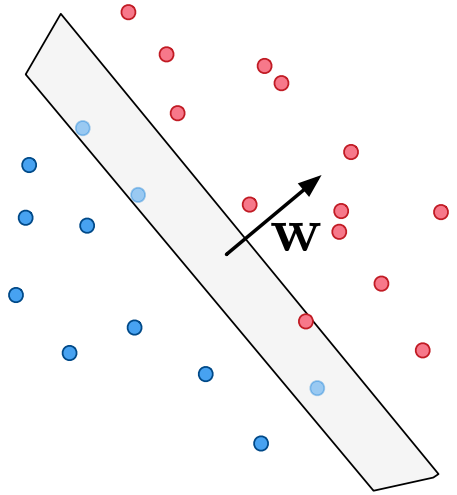
1. De nature **géométrique**
2. De nature modélisation **statistique**
3. De nature **symbolique**

K-ppv (*k* plus proches voisins)



Règles de nature géométrique

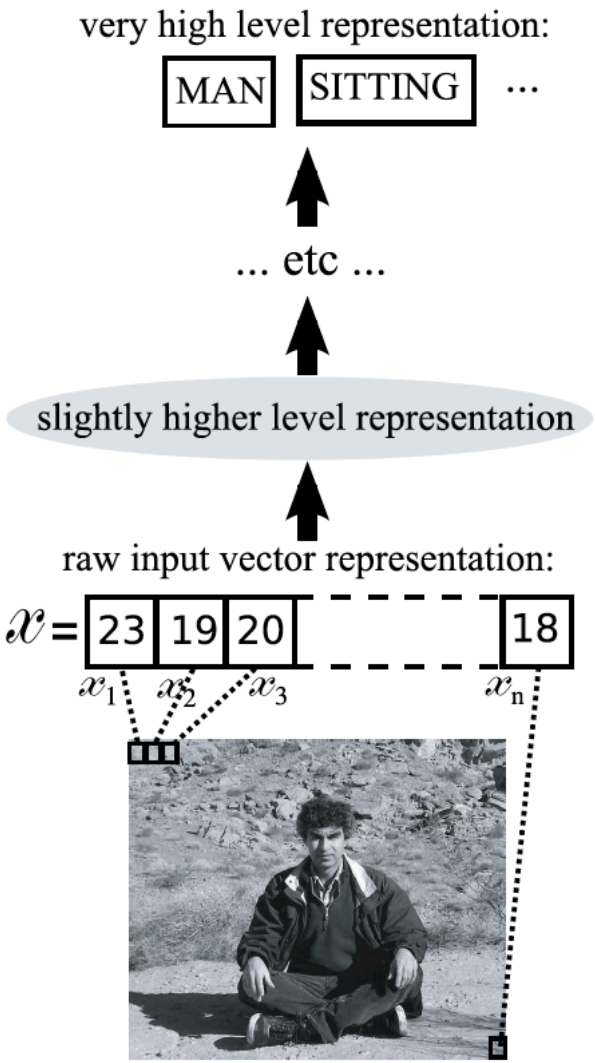
1. Séparateurs **linéaires**
   - Régression logistique / perceptron
2. Séparateurs **non** linéaires
   - Réseaux de neurones
   - SVM (Séparateurs à Vastes Marges)



Inférence par minimisation d'un
risque empirique régularisé

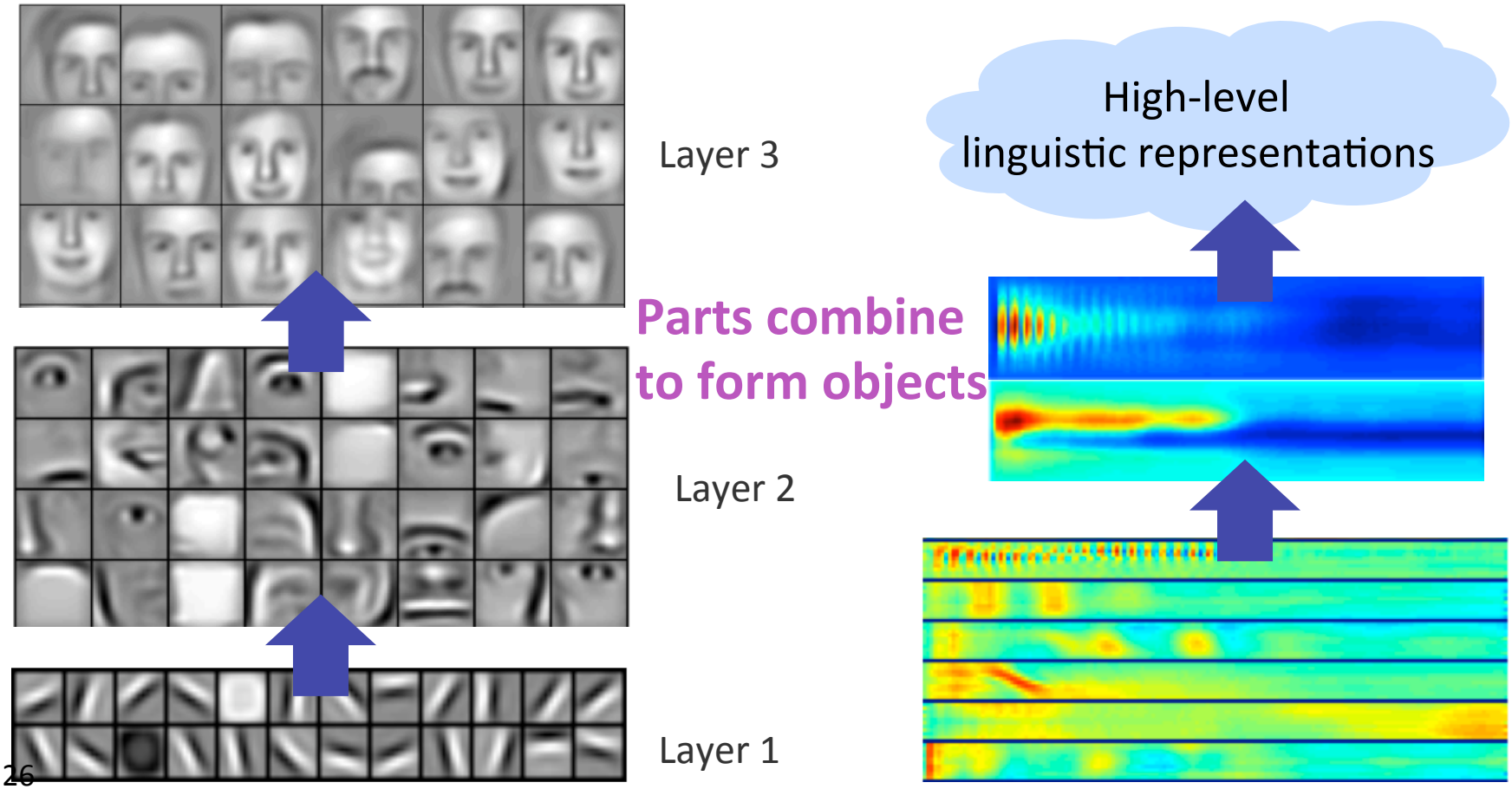
Les réseaux de neurones **profonds**

« Deep belief networks »



Apprentissage de représentations hiérarchiques

- Apprentissage de représentations hiérarchiques



26

Les « réseaux de neurones profonds »

- Des réseaux de neurones artificiels

  1. à grand nombre de couches
  2. et **très grand nombre de paramètres**
  3. qui apprennent des **représentations hiérarchiques**
  4. et **décomposent les calculs**

Illustration : ImageNet

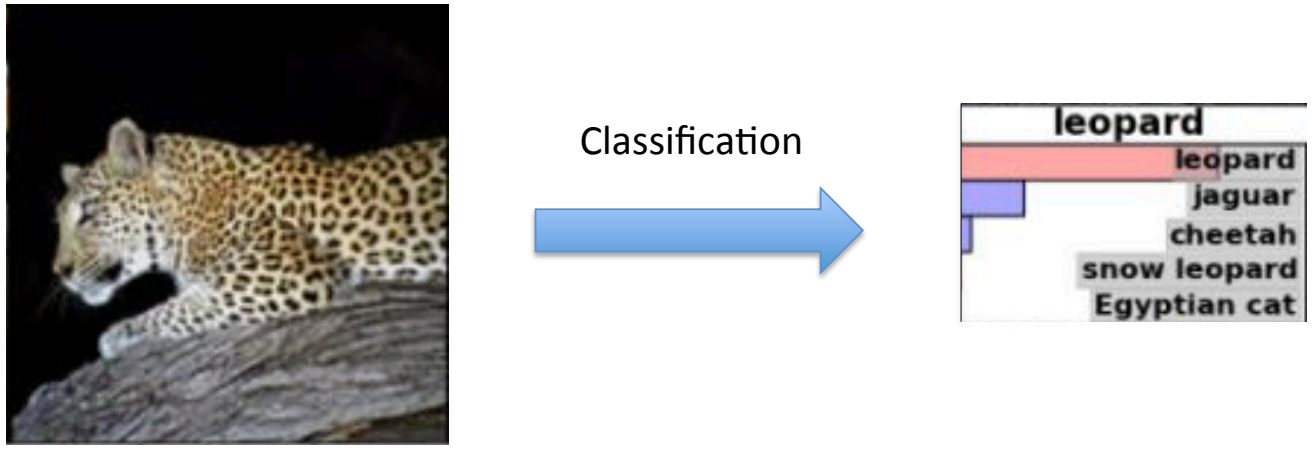
La compétition ImageNet

- Plus de **15M d'images** haute résolution étiquetées
- Environ **22K catégories**
- Récoltées sur le Web et étiquetées par Amazon Mechanical Turk

Illustration : ImageNet

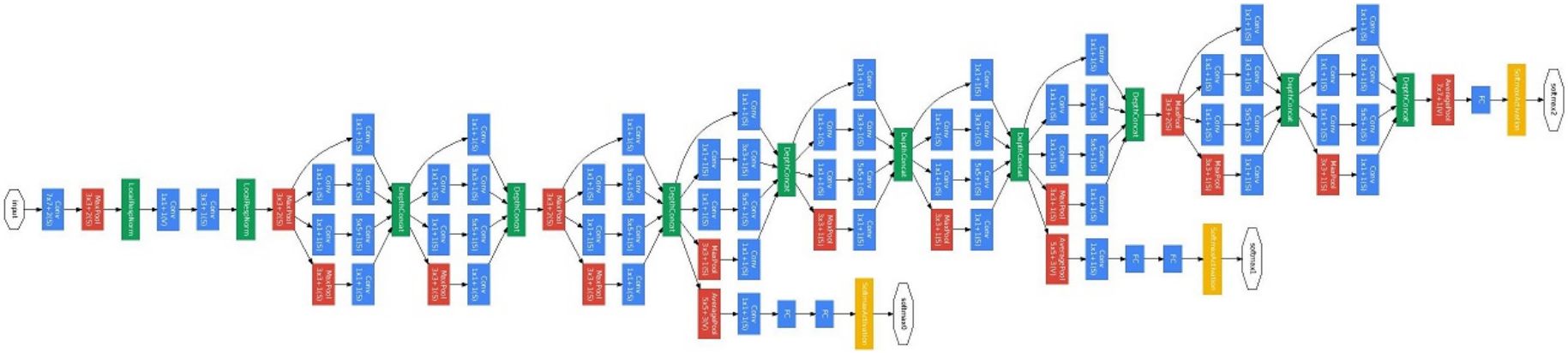
La compétition ImageNet

- Plus de **15M d'images** haute résolution étiquetées
- Environ **22K catégories**
- Récoltées sur le Web et étiquetées par Amazon Mechanical Turk



GoogleNet

- Un **mécano** de réseaux de neurones



Illustration

Système développé par Google et U. de Stanford

- Reconnaissance de visages
  - Sous conditions de lumière diverses
  - Sous tout angle
- Apprentissage non supervisé
  - 9 couches ; 10^9 connexions
  - 10 millions d'images
  - 3 jours de calcul sur 16 000 processeurs
- Amélioration des performances de 70% / état de l'art

Un « bolide » délicat à piloter

Requiert

1. beaucoup de **données** (en général)
   - Des millions d'images
   - Des dizaines de milliers de documents
2. du **savoir-faire** (des data scientists)
   - Nombreuses « **astuces** » d'ingénierie
   - Utilisation de réseaux déjà appris (**transfert**)
   - L'état de l'art **progresse très vite**
3. des machines adaptées
   - Puissance **calcul** : clusters et/ou cartes graphiques
   - **Mémoire** centrale importante (≥ 128 Go)

Enseigné dans certaines écoles et universités

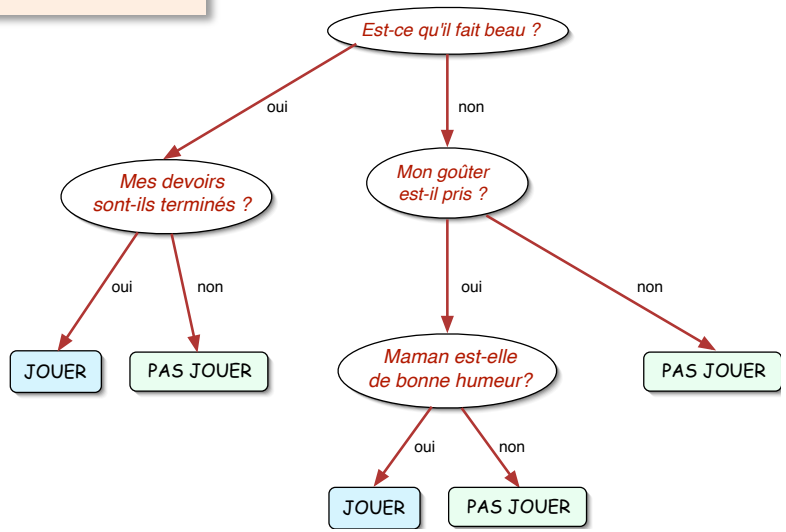
Règles de nature modélisation statistique

1. Mélange de distributions (e.g. gaussiennes)
2. Modèles graphiques
3. Chaînes de Markov / HMM

Inférence par maximum de vraisemblance

Règles de nature symbolique

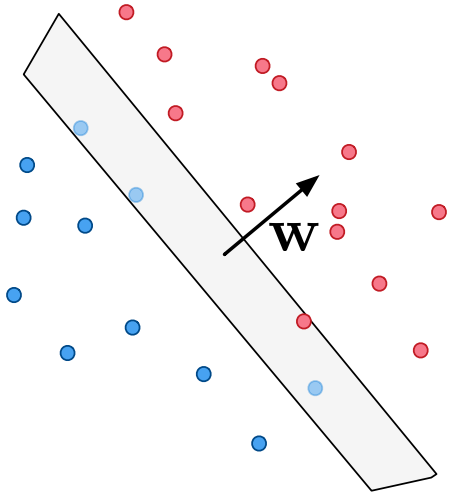
1. Arbres de décision
2. Inférence de grammaires
3. Inférence de systèmes de règles



Inférence par méthode itérative heuristique

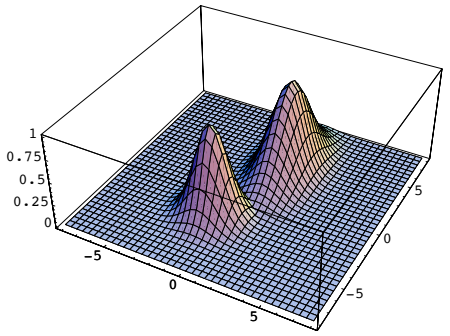
Différence entre …

- **Apprentissage** automatique
  - On cherche une règle / frontière de décision
  - Approche « discriminative »

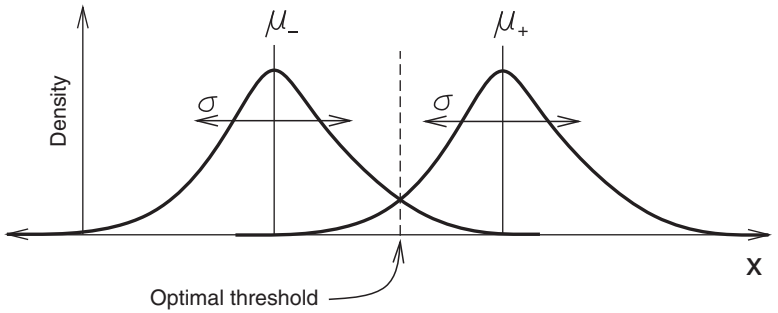


Différence entre …

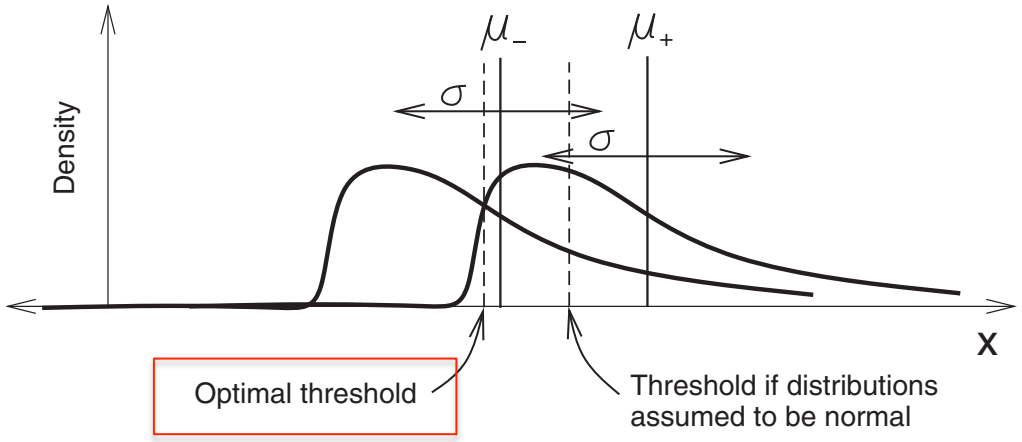
- **Apprentissage** automatique
  - On cherche une règle / frontière de décision
  - Approche « discriminative »



- Approche **statistique**
  - On passe par la définition / estimation d'une distribution de probabilité
  - Et on applique la règle de Bayes
  - Approche « générative »

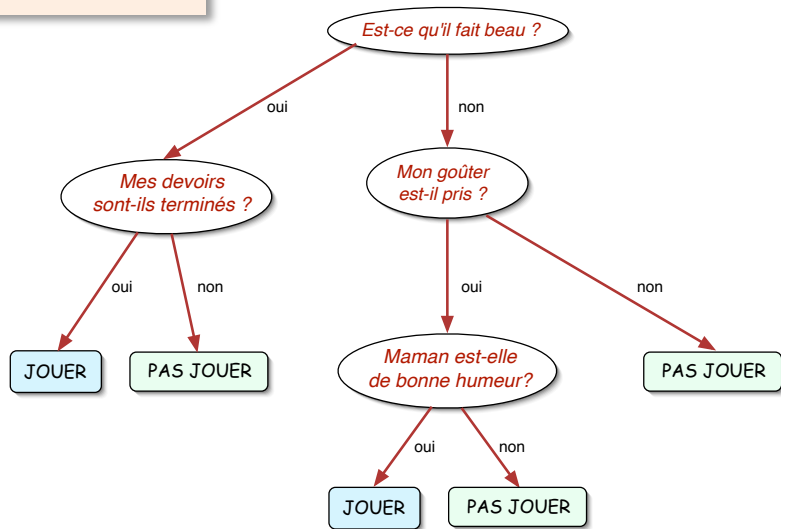


Une différence … significative



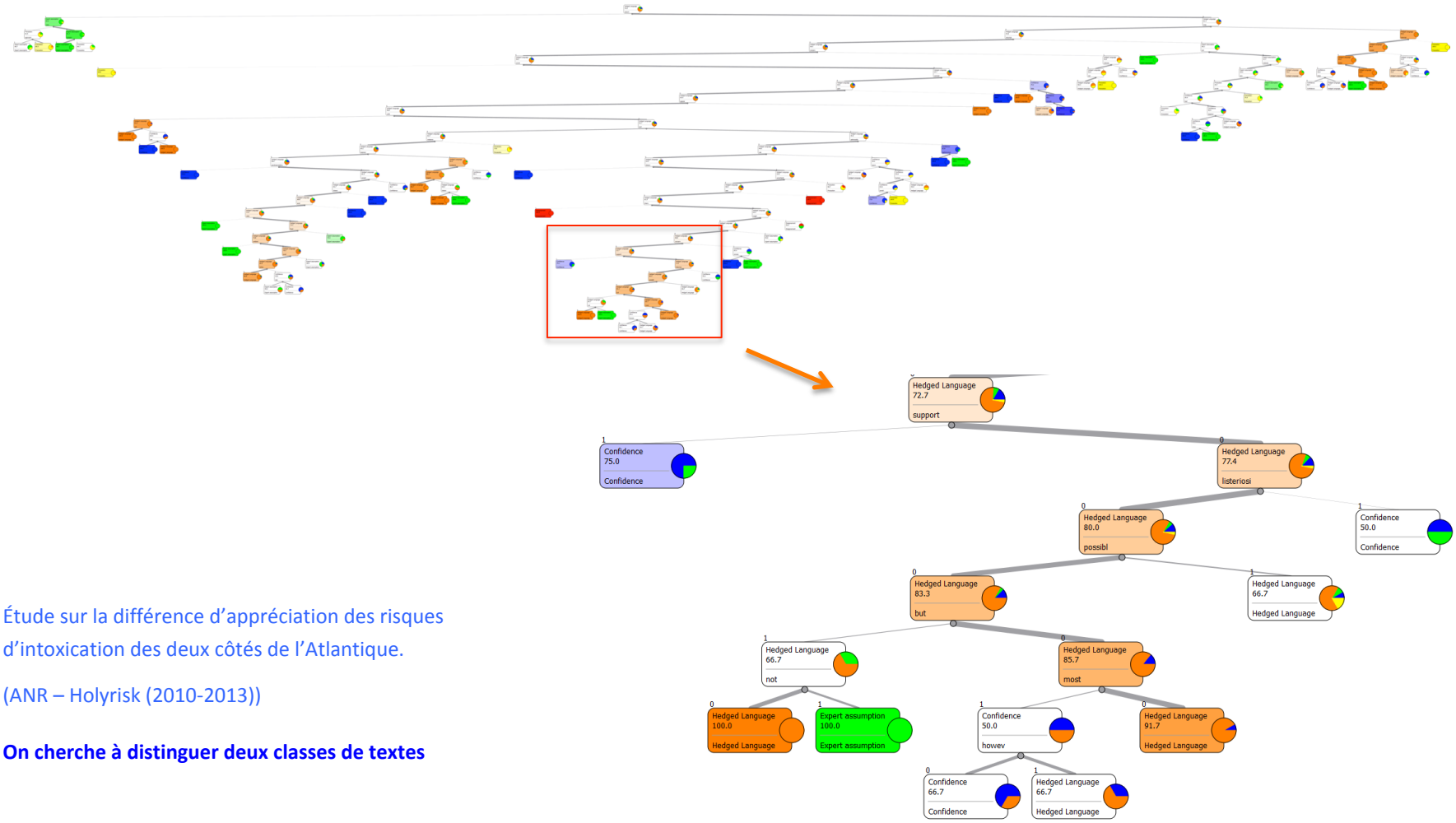
Règles de nature symbolique

1. Arbres de décision
2. Inférence de grammaires
3. Inférence de systèmes de règles



Inférence par méthode itérative heuristique

Exemple : arbre de décision



Étude sur la différence d'appréciation des risques d'intoxication des deux côtés de l'Atlantique.

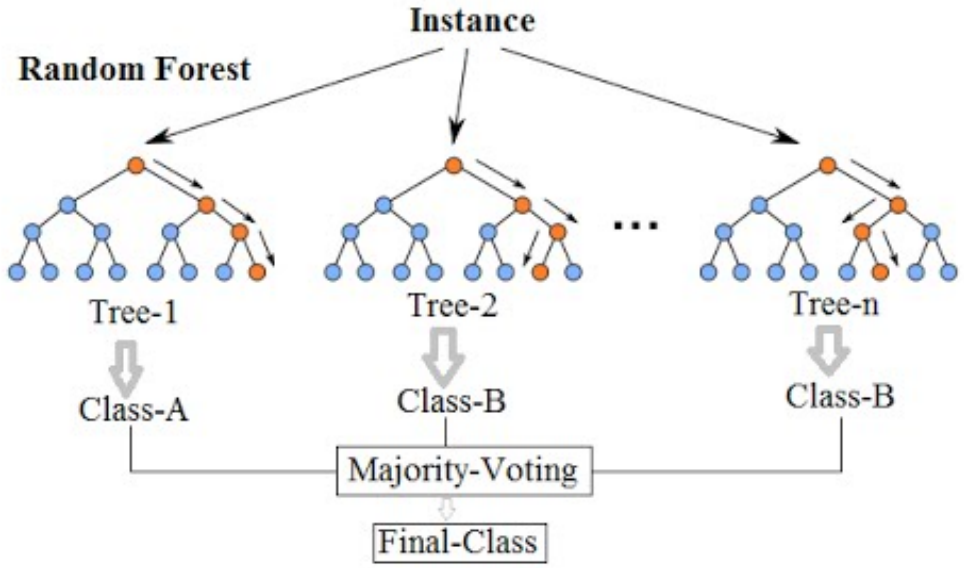
(ANR – Holyrisk (2010-2013))

On cherche à distinguer deux classes de textes

Méta-méthodes

– Par combinaison de méthodes (ensemble methods)

Bagging, boosting, random forests, …

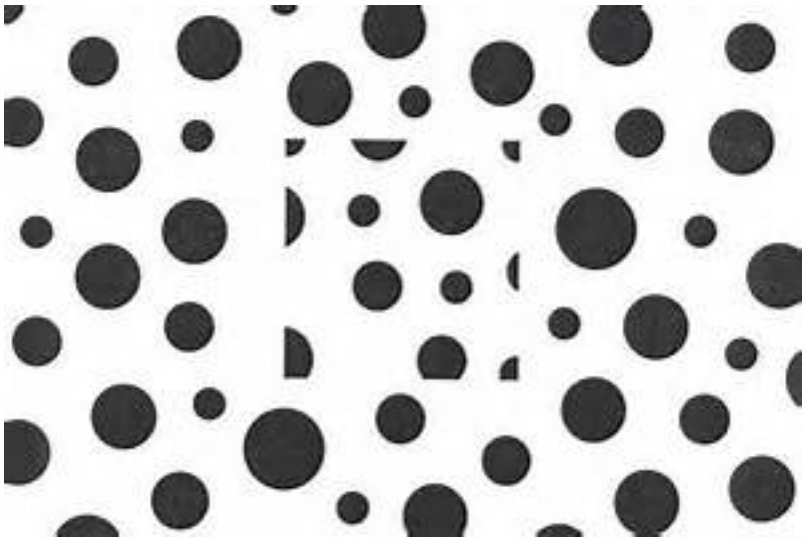


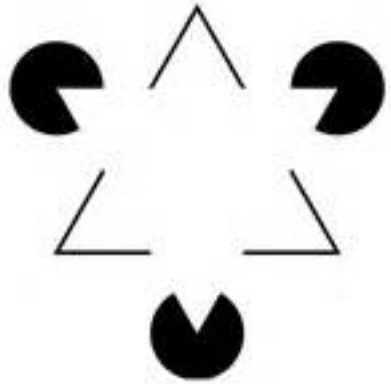
Une question cruciale : l'évaluation

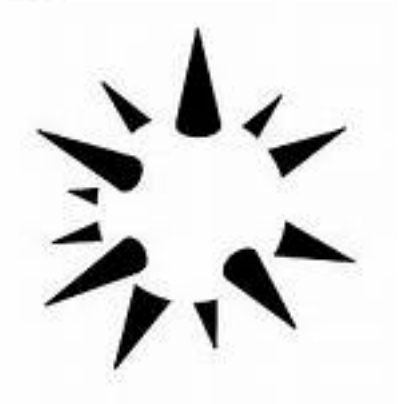
- Comment **sélectionner** les régularités trouvées ?

- Comment les **évaluer** ?

L'apprentissage – une extrapolation

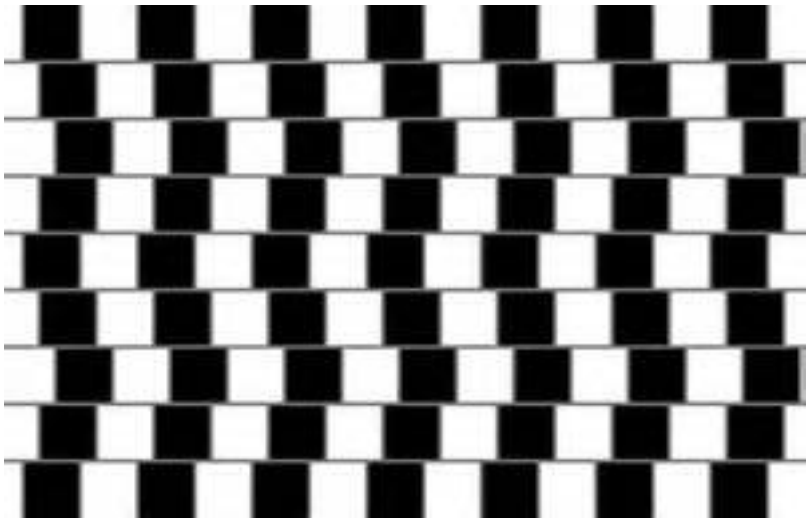




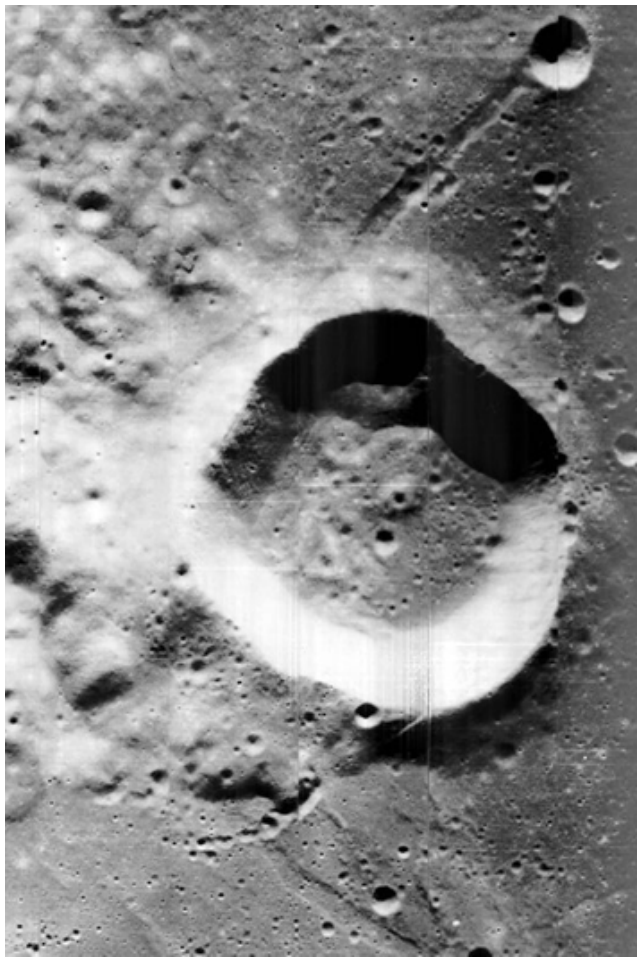


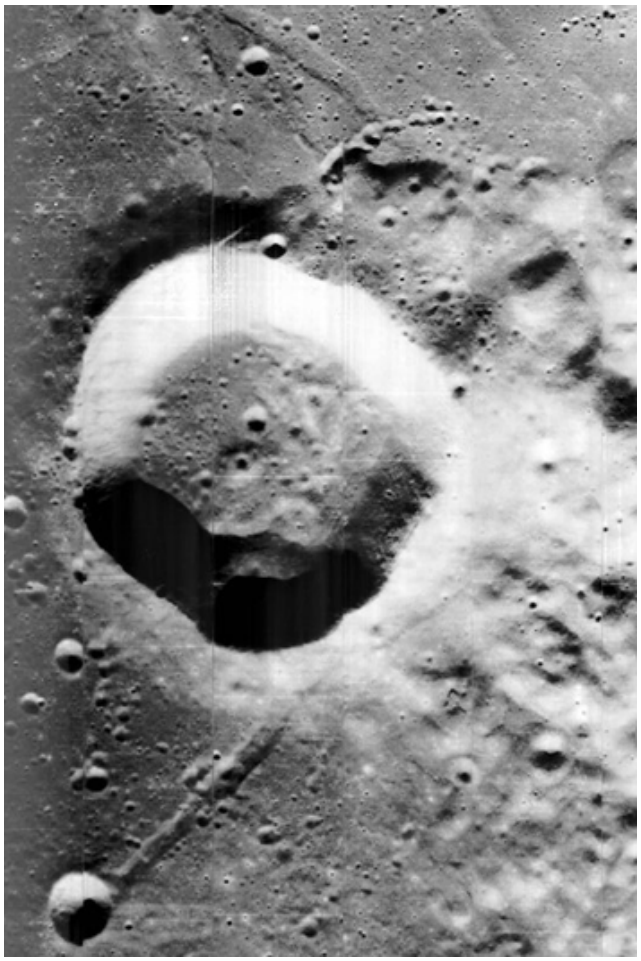
Une extrapolation – soumise à des **choix**

Des **biais** pouvant conduire à des illusions



Induction et illusions

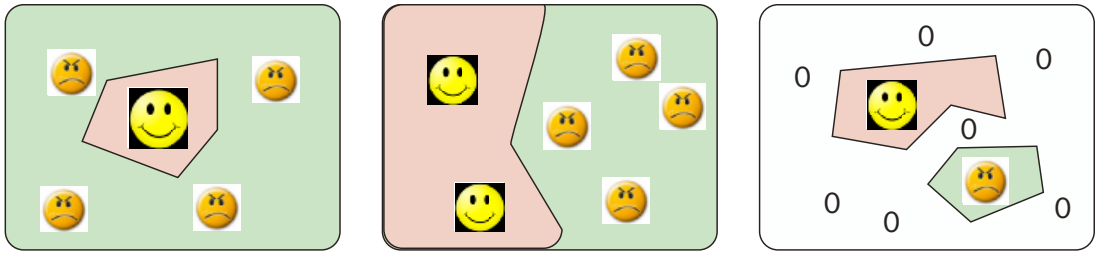




Cratère ou colline ?

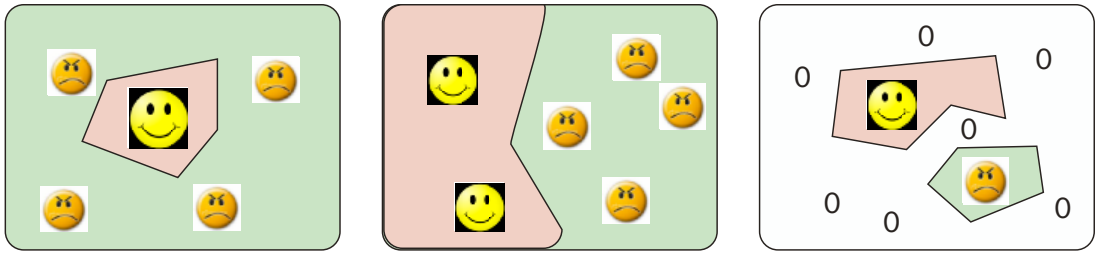
Le no-free-lunch theorem

Possible

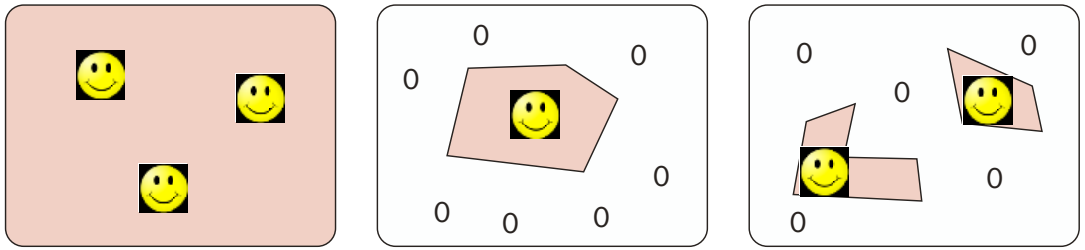


Le no-free-lunch theorem

Possible



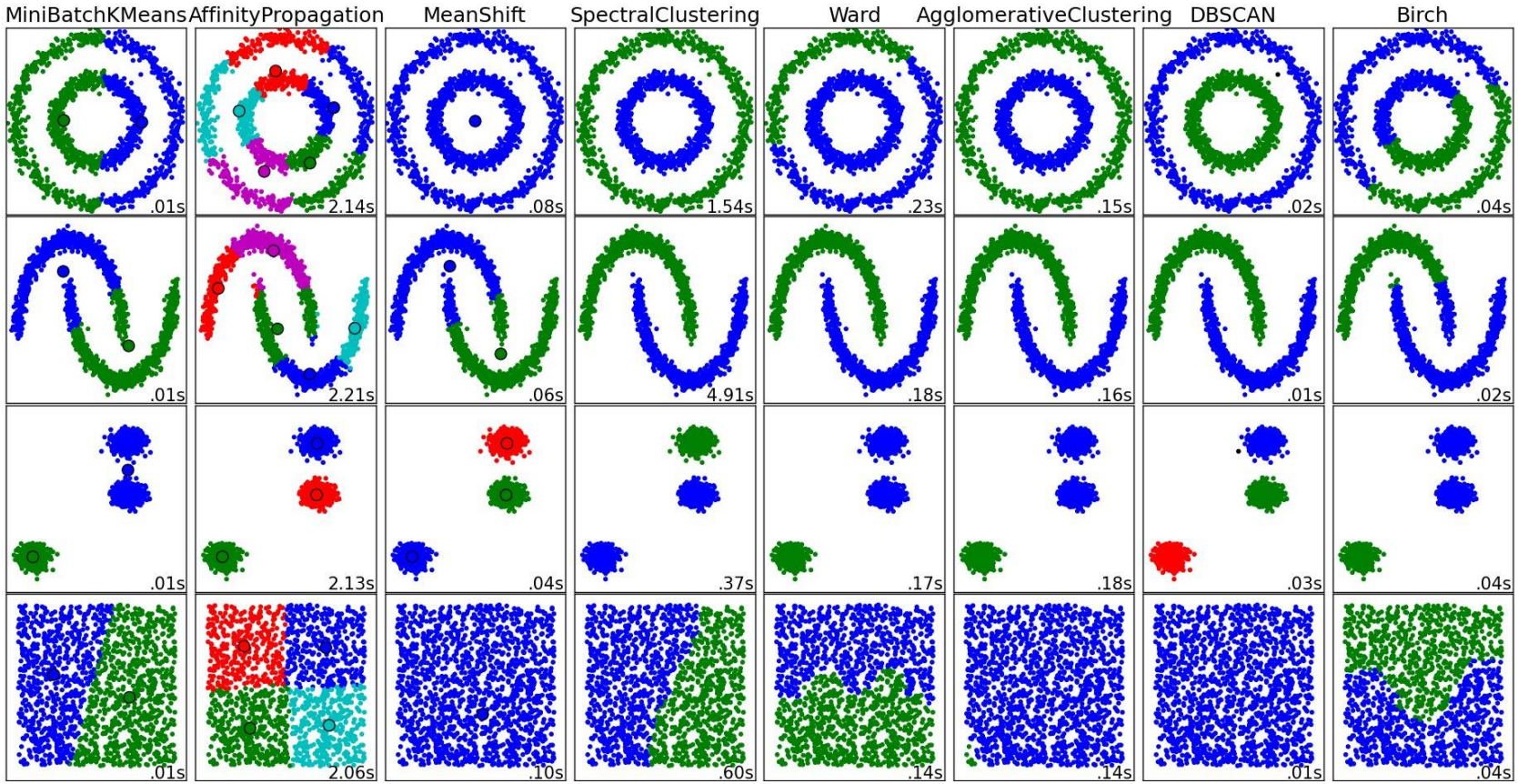
Impossible



Il faut **choisir** le **bon algorithme** pour la **classe de problèmes** étudiée

Illustration : le **clustering**

Les biais a priori sont **nécessaires** et … il faut bien les **choisir**



Propriétés des méthodes

Interprétabilité des hypothèses produites

- ✦ Méthodes **linéaires**
- ✦ **Arbres** de décision
- ■ K-plus proches voisins
- – **SVM** (Séparateurs à Vastes Marges)
- – **Réseaux de neurones**

Propriétés des méthodes

Passage **à l'échelle**

Nombre d'exemples

- ✦ Méthodes **linéaires**
- ✦ **Arbres** de décision
- ✦ **Réseaux de neurones**
- ✦ Méthodes **statistiques** par estimation de densité
- ▪ **SVM** (Séparateurs à Vastes Marges)
- – K-plus proches **voisins**

Nombre de dimensions

- ✦ Méthodes **linéaires**
- ✦ **SVM** (Séparateurs à Vastes Marges)
- ▪ **Réseaux de neurones**
- ▪ **Arbres** de décision
- – K-plus proches **voisins**
- – Méthodes **statistiques** par estimation de densité

Plan

1. Grands types d'apprentissage
2. Méthodes d'apprentissage
3. En pratique
4. Ce que l'on sait faire et les défis à relever
5. L'IA : une révolution ?

En pratique

1. Obtenir les **données**
2. Bien penser le **recueil des données**
3. Importance des **prétraitements**
4. Importance de la disponibilité des **experts métier**
5. Les questions **juridiques**

Obtenir les données

Souvent **difficile** !!!

- Les données ne sont **pas encore disponibles**
- Le donneur d'ordre n'est **pas détenteur des données**
  - Pas le même service / département
- Les données sont **protégées par des droits**
- Une partie des données **reste à recueillir**

Bien penser le recueil des données

Essentiel !!!

- Exemple : **I**nternet **d**es **O**bjets (IoT)
  - Objets **faciles** et **agréables** à utiliser
  - **Mais**
    - Ne recueille pas les données nécessaires
    - Développement « agile »
      - ✓ Changements de formats
      - ✓ Changements des mesures recueillies

2 ans de perdus

Tout reprendre à zéro

Les prétraitements

- **90%** du temps d'un projet
- Mise dans un **format adéquat**
- **Nettoyage**
  - **Bruit** dans les données
  - Données **manquantes**
  - Données **aberrantes**
  - **Doublons**
  - **Normalisation** des mesures
  - **Discrétisation** de valeurs continues
  - **Rendre continues** des valeurs discrètes
- Élimination des **attributs redondants** / calcul de **nouveaux attributs**
- **Précision / incertitude**
- Intégration de plusieurs **sources de données (hétérogènes)**
- **…**

Choix d'un bon critère de performance

Disponibilité des experts métier

Essentiel !!!

- **Comprendre** le problème
- Établir un **vocabulaire commun**
- **Évaluer** les résultats
- Orienter / **ré-orienter**
- **S'approprier** les résultats / assurer la suite

Les questions juridiques

Essentiel !!!

- Données **personnelles**
- **Obtenir l'autorisation**
  - CNIL
  - RGPD
    - À partir du **25 mai 2018**, le Règlement Général Européen sur la Protection des Données (RGPD) affecte toutes les organisations traitant les données personnelles identifiables (DPI) de résidents européens.

Plan

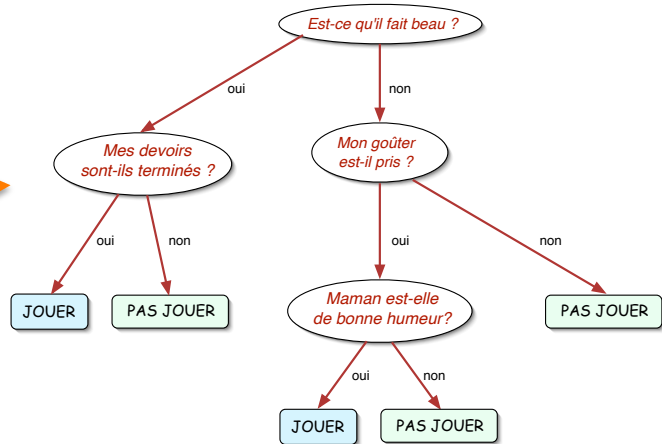
1. Grands types d'apprentissage
2. Méthodes d'apprentissage
3. En pratique
4. Ce que l'on sait faire et les défis à relever
5. L'IA : une révolution ?

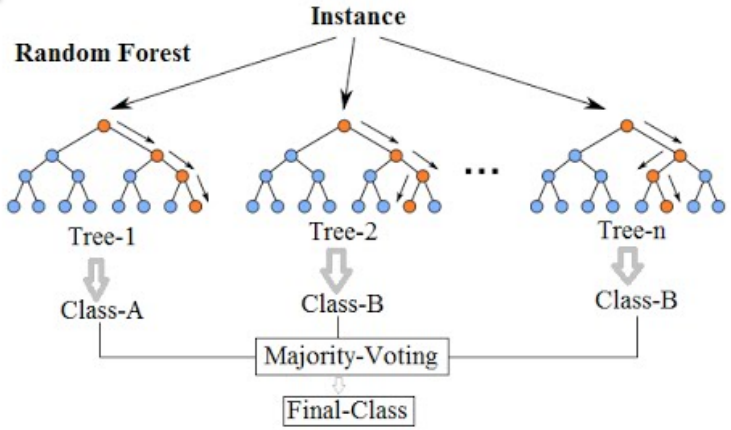
Ce que l'**on sait faire**

- Apprentissage prédictif
  - En environnement **stationnaire**
  - À partir de (très) **nombreux exemples**
  - Classification / régression
- Apprentissage descriptif
  - Problème de la **validation**
- Apprentissage de recommandation
- Apprentissage de contrôle / commande (app. par renforcement)

Nombreuses méthodes d'apprentissage

- Réseaux de neurones
- Arbres de décision
- Méthodes d'ensemble
- Apprentissage bayésien
- Chaînes de Markov et HMM
- Outils de fouille de données
- …





Les méthodes et algorithmes

- Librairies / méthodes / algorithmes
  - Sont dans le **domaine public** !!!
    - Publications scientifiques
    - Forums
    - Conférences
    - Librairies (e.g. ScikitLearn)

- Des « **recettes** » privées
  - Réseaux de neurones profonds
  - Traitement d'images / télédétection
  - Connaissances métiers (e.g. alimentation)

Les moyens calcul

Les moyens calcul

- Important
- Mais **pas forcément très honéreux**
  - **Station de travail** avec 8 cartes graphiques et 128 Go de mémoire centrale
  - **Cluster** de machines
  - **Utilisation de cloud computing**
- Problème… **évolue vite**

  et dépend de ce que l'on veut faire

Les « data scientists »

- **Compétences** attendues
  1. Apprentissage artificiel / Statistiques
     - Bonne compréhension des questions et des hypothèses sur lesquelles reposent les méthodes
  2. Compétences en informatique
     - Algorithmique
     - Bases de données
     - Réseaux
  3. Capacités relationnelles

En très forte demande

100 000 en France
à l'horizon 2022 !!

- **Formations**
  - Quelques dizaines d'heures
  - **Master** ou équivalent
  - **Doctorat**

Grand risque de déconvenue
si pas les bons recrutements

Les défis à relever

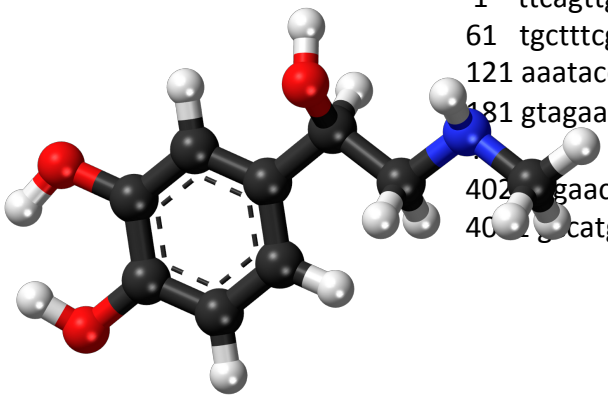
Défis : Apprendre …

1. … à partir de peu d'exemples
2. … à partir de **sources** de données et de connaissances multiples et hétérogènes
3. …
   - des **hypothèses interprétables** (la **sortie**)
   - de manière **explicable** (l'**algorithme**)
4. … à extraire des relations causales
5. … en environnement non stationnaire
   - **Dérive** de concept
   - **Transfert**
   - Apprentissage « **au long cours** »

Intégration de multiple sources de données

- Annotation de protéines

Protéine « sp|P00004|CYC_HORSE » is activated by …

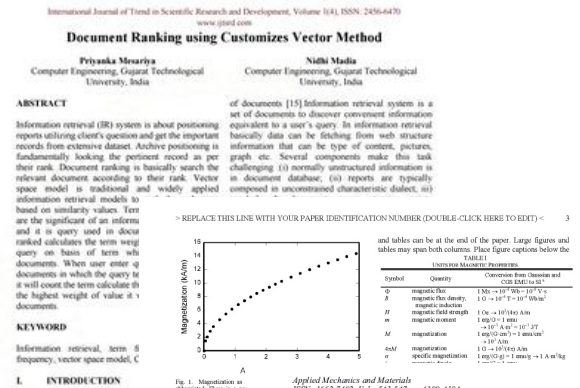


1 ttcagttgtg aatgaatgga cgtgccaaat agacgtgccg ccgccgctcg attcgcactt
61 tgctttcggt tttgccgtcg tttcacgcgt ttagttccgt tcggttcatt cccagttctt
121 aaataccgga cgtaaaaata cactctaacg gtcccgcgaa gaaaaagata aagacatctc
181 gtagaaatat taaaataaat tcctaaagtc gttggtttct cgttcacttt cgctgcctgc

402 gaacacgcc gaggctccat tcatagcacc acttcgtcgt cttaatcccc tccctcatcc
40 catggcgg tgcaaaaaat aaaaagaact c

Intégration de multiple sources de données

- GIEC
  - Documents scientifiques multiples
  - Tableaux
  - mesures



Document Ranking using Customizes Vector Method

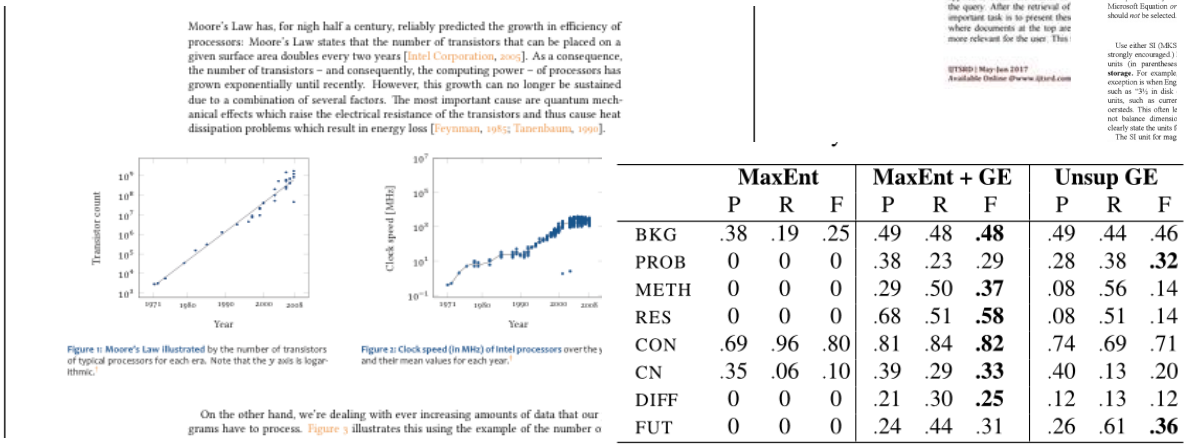
Priyanka Mesariya, Computer Engineering, Gujarat Technological University, India

Nidhi Madia, Computer Engineering, Gujarat Technological University, India

ABSTRACT

KEYWORD

I. INTRODUCTION



Moore's Law has, for nigh half a century, reliably predicted the growth in efficiency of processors: Moore's Law states that the number of transistors that can be placed on a given surface area doubles every two years [Intel Corporation, 2005]. As a consequence, the number of transistors – and consequently, the computing power – of processors has grown exponentially until recently. However, this growth can no longer be sustained due to a combination of several factors. The most important cause are quantum mechanical effects which raise the electrical resistance of the transistors and thus cause heat dissipation problems which result in energy loss [Feynman, 1985; Tanenbaum, 1990].

Transistor count

Year

Figure 1: Moore's Law Illustrated by the number of transistors of typical processors for each era. Note that the y axis is logarithmic.

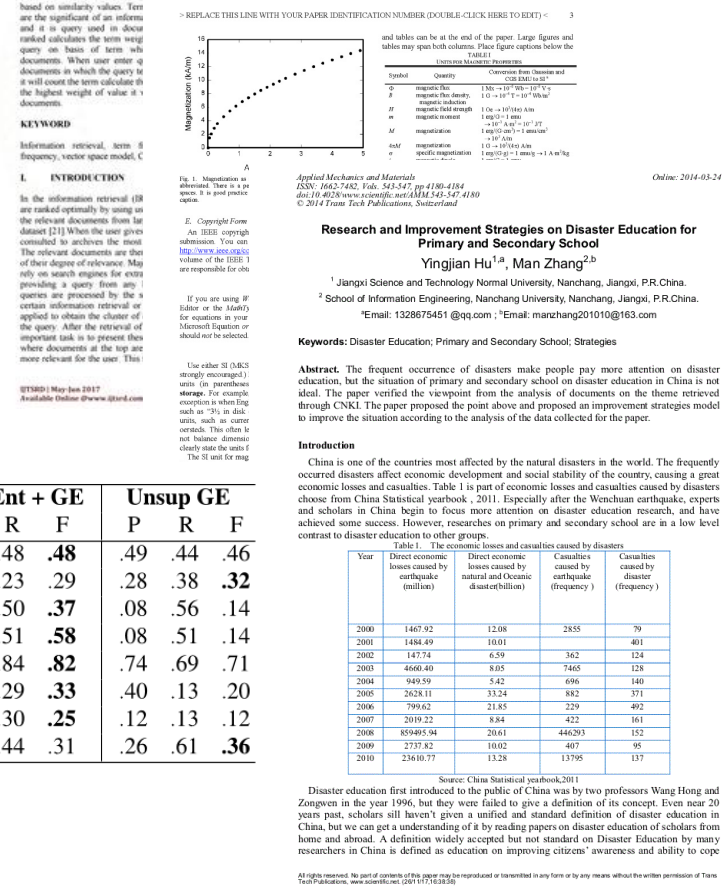
Clock speed [MHz]

Year

Figure 2: Clock speed (in MHz) of Intel processors over the y and their mean values for each year.

On the other hand, we're dealing with ever increasing amounts of data that our grams have to process. Figure 3 illustrates this using the example of the number o

	MaxEnt			MaxEnt + GE			Unsup GE		
	P	R	F	P	R	F	P	R	F
BKG	.38	.19	.25	.49	.48	**.48**	.49	.44	.46
PROB	0	0	0	.38	.23	.29	.28	.38	**.32**
METH	0	0	0	.29	.50	**.37**	.08	.56	.14
RES	0	0	0	.68	.51	**.58**	.08	.51	.14
CON	.69	.96	.80	.81	.84	**.82**	.74	.69	.71
CN	.35	.06	.10	.39	.29	**.33**	.40	.13	.20
DIFF	0	0	0	.21	.30	**.25**	.12	.13	.12
FUT	0	0	0	.24	.44	.31	.26	.61	**.36**



Research and Improvement Strategies on Disaster Education for Primary and Secondary School

Yingjian Hu[1,a], Man Zhang[2,b]

[1] Jiangxi Science and Technology Normal University, Nanchang, Jiangxi, P.R.China.

[2] School of Information Engineering, Nanchang University, Nanchang, Jiangxi, P.R.China.

[a]Email: 1328675451 @qq.com ; [b]Email: manzhang201010@163.com

Keywords: Disaster Education; Primary and Secondary School; Strategies

Abstract. The frequent occurrence of disasters make people pay more attention on disaster education, but the situation of primary and secondary school on disaster education in China is not ideal. The paper verified the viewpoint from the analysis of documents on the theme retrieved through CNKI. The paper proposed the point above and proposed an improvement strategies model to improve the situation according to the analysis of the data collected for the paper.

Introduction

China is one of the countries most affected by the natural disasters in the world. The frequently occurred disasters affect economic development and social stability of the country, causing a great economic losses and casualties. Table 1 is part of economic losses and casualties caused by disasters choose from China Statistical yearbook , 2011. Especially after the Wenchuan earthquake, experts and scholars in China begin to focus more attention on disaster education research, and have achieved some success. However, researches on primary and secondary school are in a low level contrast to disaster education to other groups.

Table 1. The economic losses and casualties caused by disasters

Year	Direct economic losses caused by earthquake (million)	Direct economic losses caused by natural and Oceanic disaster(billion)	Casualties caused by earthquake (frequency)	Casualties caused by disaster (frequency)
2000	1467.92	12.08	2855	79
2001	1484.49	10.01		401
2002	147.74	6.59	362	124
2003	4660.40	8.05	7465	128
2004	949.59	5.42	696	140
2005	2628.11	33.24	882	371
2006	799.62	21.85	229	492
2007	2019.22	8.84	422	161
2008	859495.94	20.61	446293	152
2009	2737.82	10.02	407	95
2010	23610.77	13.28	13795	137

Source: China Statistical yearbook,2011

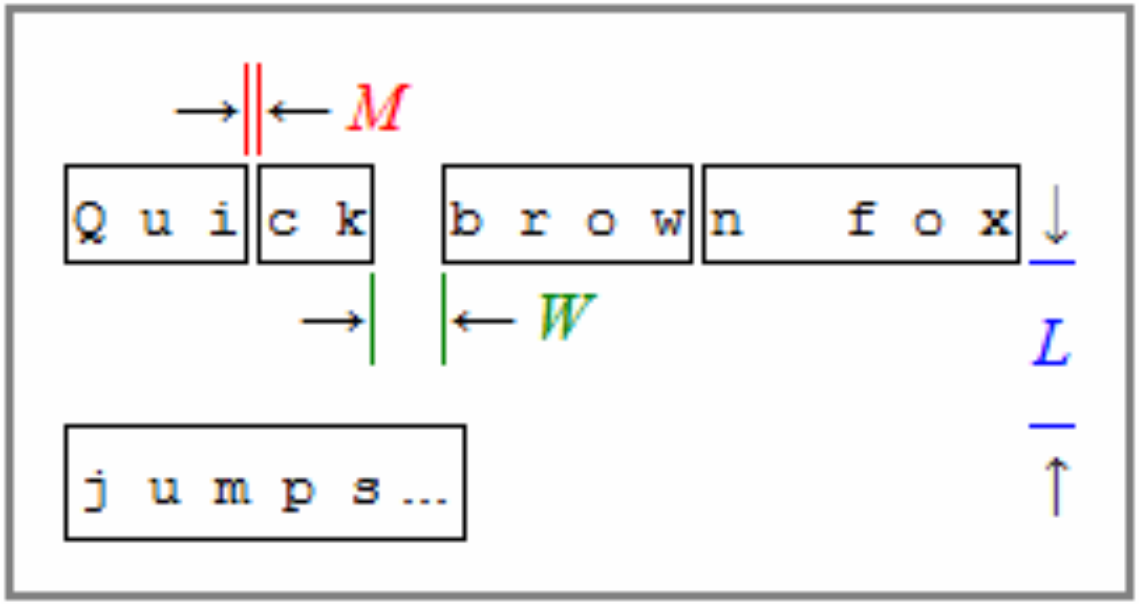
Disaster education first introduced to the public of China was by two professors Wang Hong and Zongwen in the year 1996, but they were failed to give a definition of its concept. Even near 20 years past, scholars sill haven't given a unified and standard definition of disaster education in China, but we can get a understanding of it by reading papers on disaster education of scholars from home and abroad. A definition widely accepted but not standard on Disaster Education by many researchers in China is defined as education on improving citizens' awareness and ability to cope

Le traitement des documents en .pdf

- Structure des pages … en .pdf

Le traitement des documents en .pdf

Segmentation en mots, en paragraphes, notes de bas de pages, …



- **Un point**, est-ce : une **fin de phrase**, indication d'une **initiale**, un **point décimal** dans un nombre, … ?

Plan

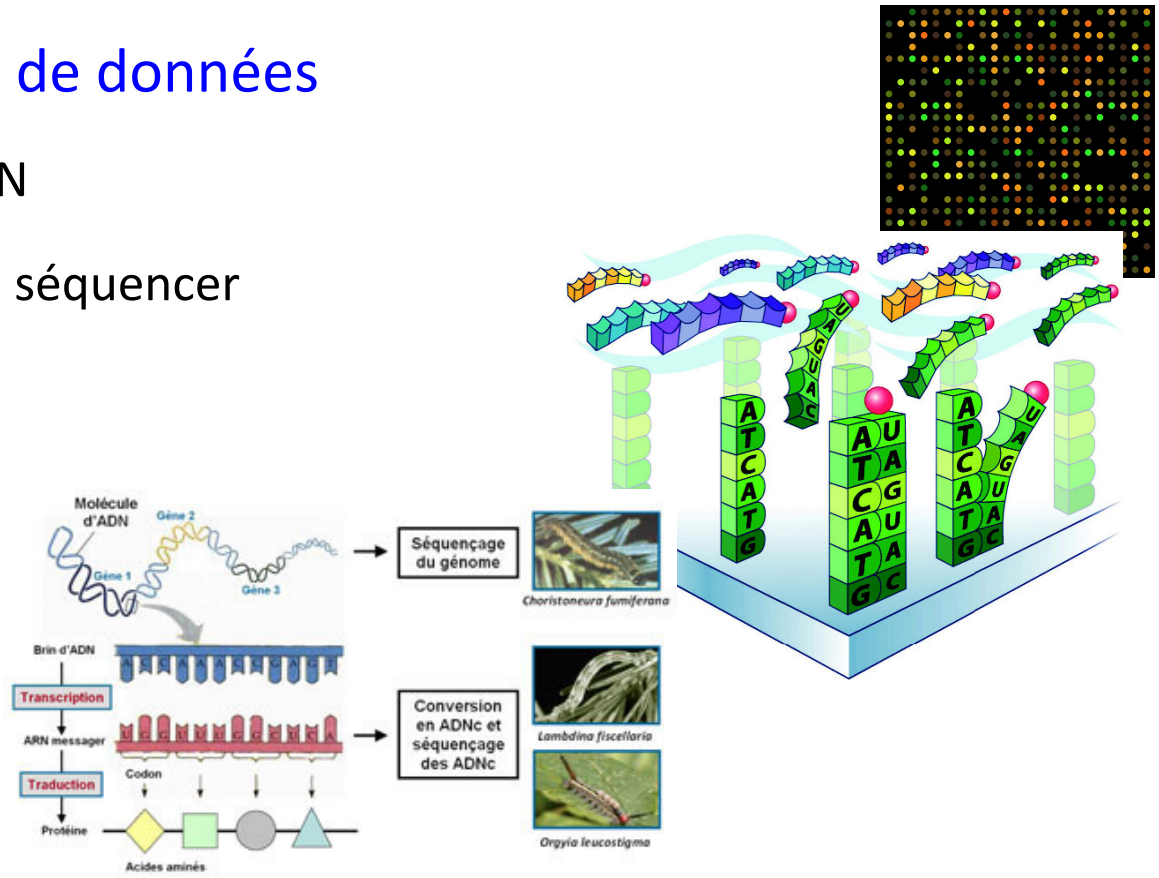
1. Grands types d'apprentissage
2. Méthodes d'apprentissage
3. En pratique
4. Ce que l'on sait faire et les défis à relever
5. L'IA : une révolution ?

Exemples de domaines renouvelés

- La bio-informatique

  - Gros volume de données

    - Puces à ADN
    - Machines à séquencer



Exemples de domaines renouvelés

- La sociologie

  - Gros volume de données
    - Réseaux sociaux
    - Smarphones
    - Websites consultations

Exemples de domaines renouvelés

- La e-medecine (le me-data)

  - Gros volume de données
    - Smarphones
    - Objets connectés
    - Forums
    - WATSON
    - Google Flu

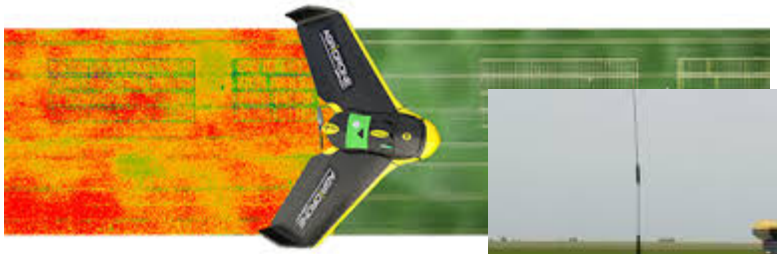
"I have no idea what's wrong with you. I just collect information. My computer makes the decisions."

Exemples de domaines renouvelés

- L'agriculture numérique

  - Gros volume de données

    - Capteurs
    - Drones
    - Réseaux sociaux et pro



Exemples de domaines renouvelés

- Le domaine juridique

  - Gros volume de données

    - Archives numérisées
    - Réseaux sociaux et professionnels

Exemples de domaines renouvelés

- Le « surgical data science »
  - Gros volume de données
    - Capteurs dans les salles d'opération

Des **contre**-exemples

- L'**alimentation**
  - Enquête **Nutrinet**
    - ~ 277 000 internautes théoriquement sur des années
    - **Mais**
      - à 80% des femmes
      - Milieux socio-professionnels élevés
      - Abandonnent après quelques jours

Manque de données représentatives

- L'**éducation**
  - Peu de données sur ce qui se passe en classe ou devant un écran

- Pour aller plus loin

http://www2.agroparistech.fr/ufr-info/membres/cornuejols/Research/Tr-Sup-Agro-Montpellier-03-12-2018-v3x4.pdf

Conclusions

Motivation

Concepts difficiles à coder à la main

- Un robot qui marche dans des zones dévastées
- Sélection de personnes à recruter
- Prédispositions pour certains types de cancer

→ Apprentissage **à partir d'exemples**

Les passages à l'échelle … petite

Savoir traiter de (très) **petits** **volumes de données**

Compenser le manque d'information dans les données

- Par de la **connaissance experte**
- **Enrichissement** des données
  - Ontologies
  - Web sémantique
  - Wikipedia and Co
- Question de la **validation des résultats**
  - Les experts

Grands types d'apprentissages …

… illustrés

En pratique

Un peu de recul :

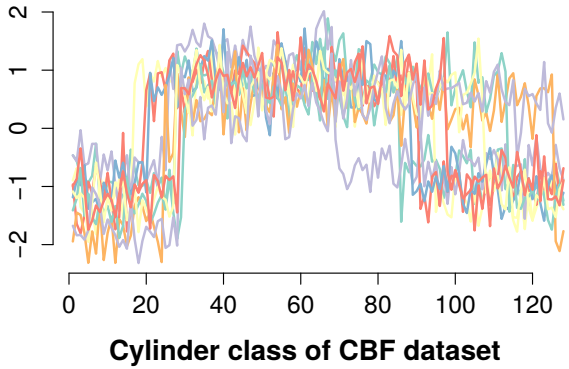
Que sait-on faire

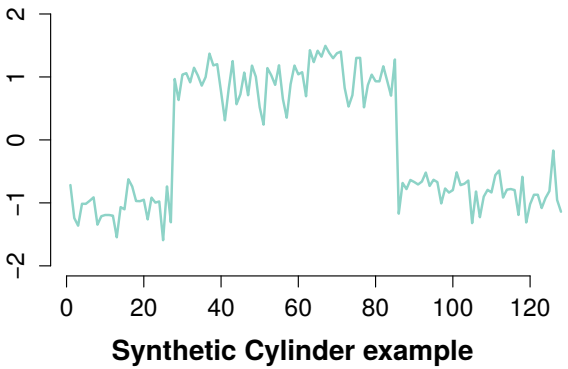
et **où sont** les **limites** ?

L'IA : Une révolution ?

Exemple (suite)

Calcul de prototype





Classe Cylinder du jeu de données CBF

Séquence prototypique calculée en moyennant les séquences de cette classe

Tiré de [Germain Forestier HDR (2017), p.60]

Apprentissage prédictif

- Si f est une *fonction* ***continue***
  - Régression

- Si f est une *fonction* ***discrète***
  - Classification

- Si f est une *fonction* ***binaire*** (Booléenne)
  - Apprentissage de concept