Découverte de règles, dans le contexte de la science des données

Université de Caen-Normandie

Bruno Crémilleux

Apprentissage de règles

En apprentissage automatique symbolique, on cherche à obtenir des procédures de classification compréhensibles par l'utilisateur humain.

Beaucoup de travaux en apprentissage de règles :

- **apprentissage de règles par couverture** (algorithmes AQ, CN2,...). Principe :
  - chercher une règle qui couvre une partie des exemples positifs
  - enlever les exemples positifs couverts de la base d'apprentissage initiale
  - recommencer récursivement le processus jusqu'à ce qu'il n'y ait plus d'exemples positifs à couvrir

  A la fin du processus, chaque exemple de la base d'apprentissage est couvert par au moins une règle
- **arbres de décision**
- ...

Arbre de décision

Un exemple : play / do not play (Quinlan, 1986)

Contexte supervisé (i.e., chaque exemple a une étiquette).

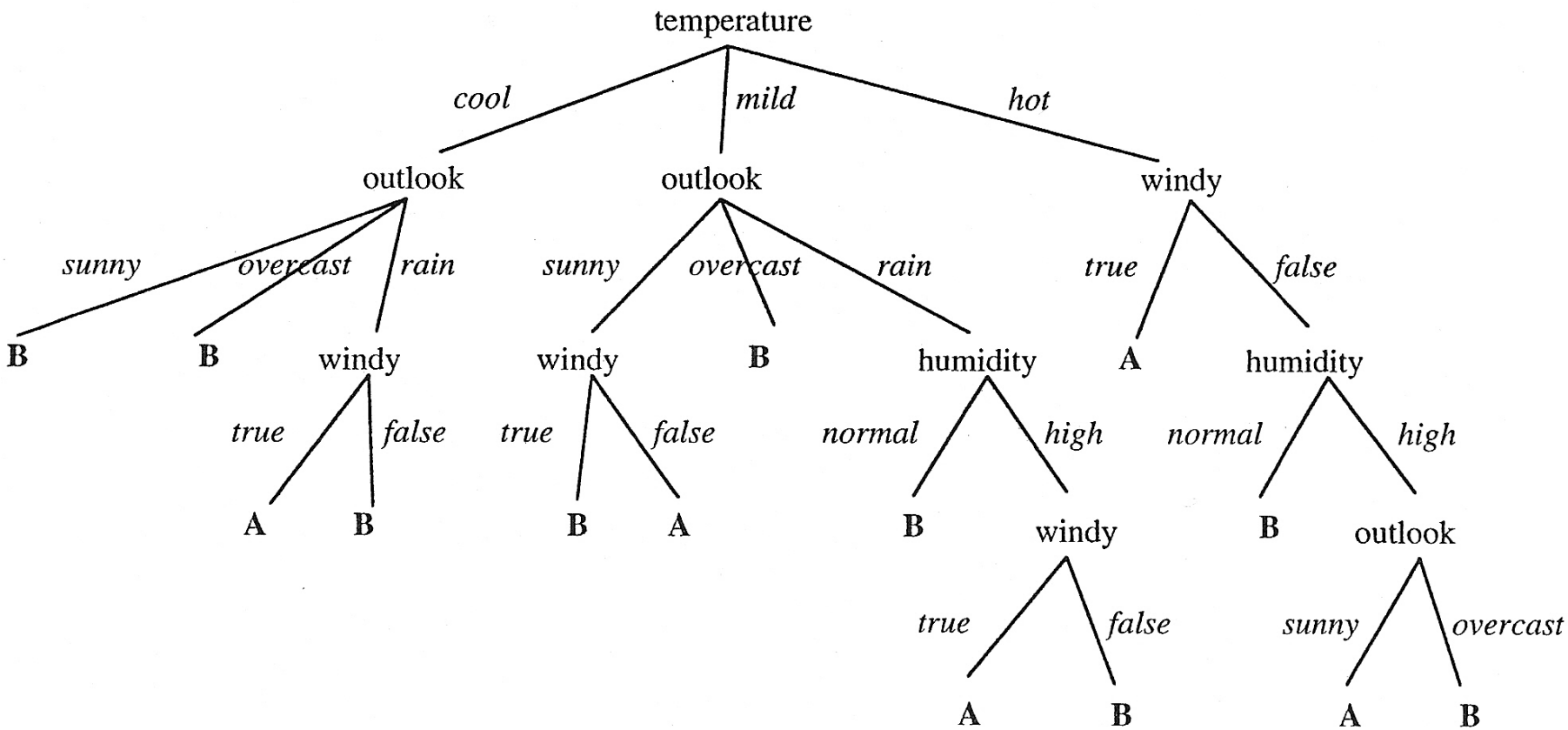
No	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	false	A
2	sunny	hot	high	true	A
3	overcast	hot	high	false	B
4	rain	mild	high	false	B
5	rain	cool	normal	false	B
6	rain	cool	normal	true	A
7	overcast	cool	normal	true	B
8	sunny	mild	high	false	A
9	sunny	cool	normal	false	B
10	rain	mild	normal	false	B
11	sunny	mild	normal	true	B
12	overcast	mild	high	true	B
13	overcast	hot	normal	false	B
14	rain	mild	high	true	A

Principe de construction d'un arbre

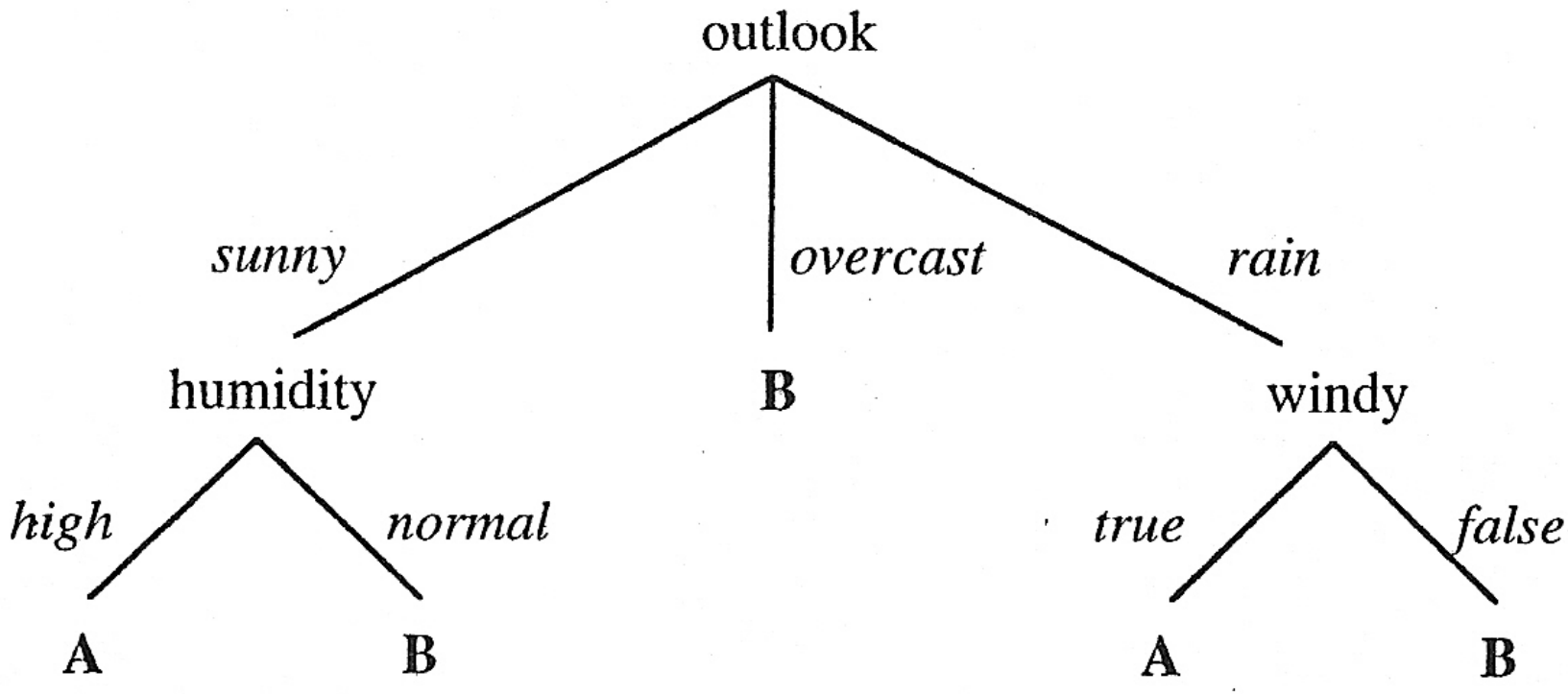
- partir de l'arbre vide
- si le nœud courant est une feuille alors
  attribuer une classe au nœud courant
  sinon choisir un attribut, partitionner le nœud courant, appliquer récursivement la construction de l'arbre sur les sous-arbres
- élaguer l'arbre de décision obtenu

pas de retour-arrière lors de la construction

Un arbre possible sur l'exemple "play / do not play"

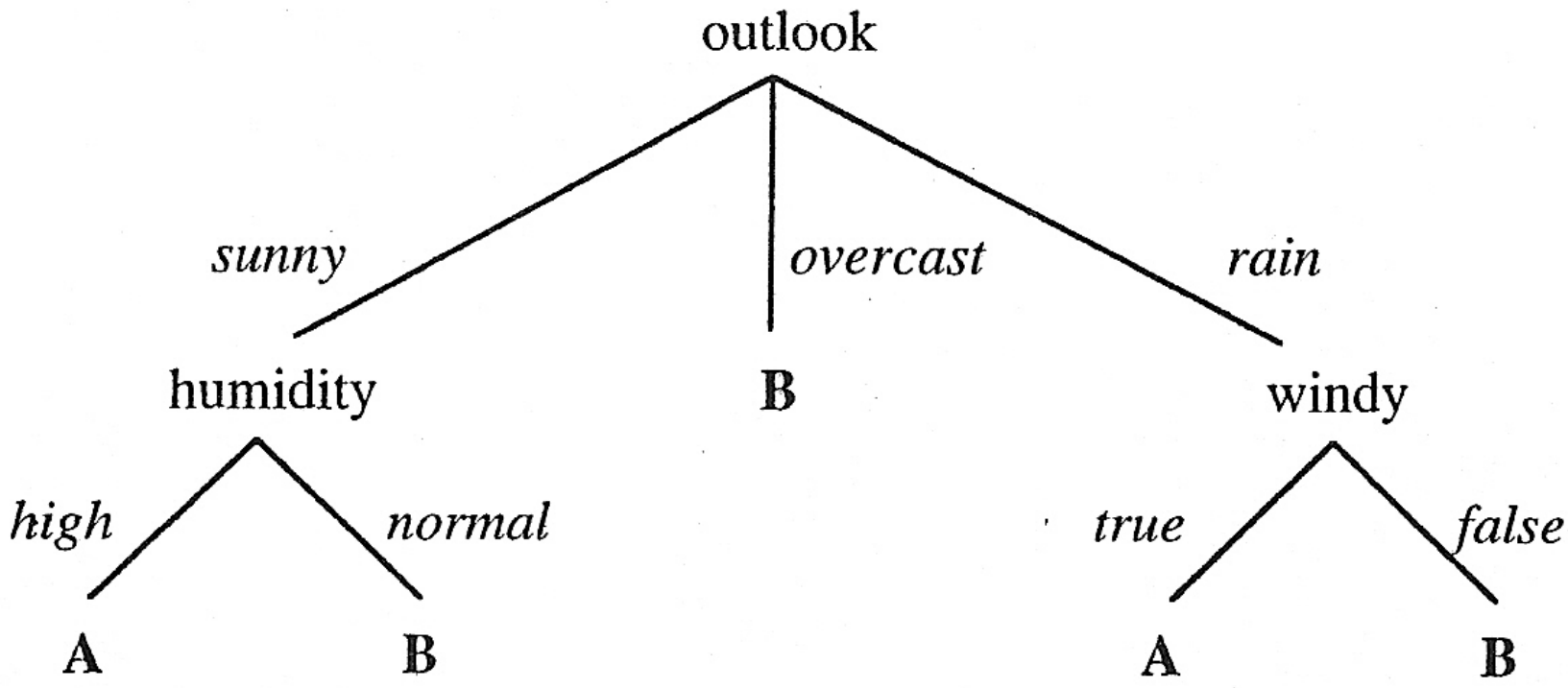


Un autre arbre possible sur l'exemple "play / do not play"



quel arbre préfére-t-on choisir ?

Un autre arbre possible sur l'exemple "play / do not play"



quel arbre préfére-t-on choisir ?

le deuxième est plus compréhensible et plus efficace pour classer de nouveaux exemples

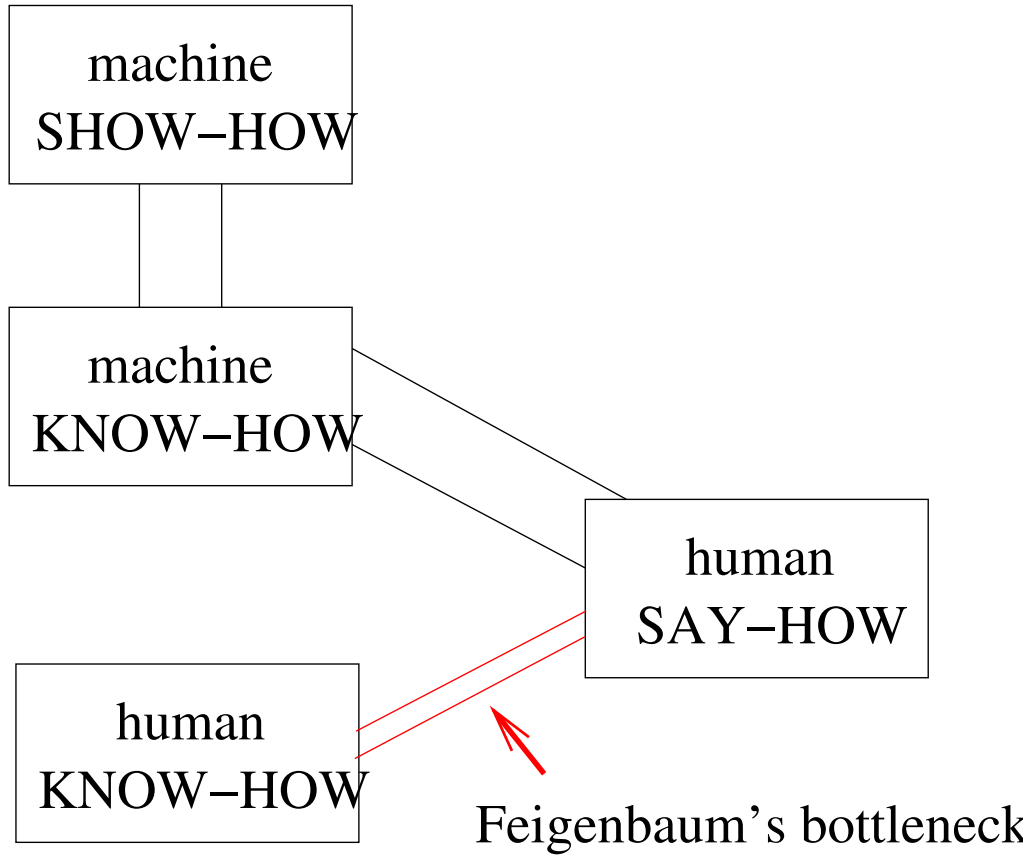
Règles issues de l'arbre sur l'exemple "play / do not play"

- if outlook = overcast then B
- if outlook = sunny and humidity = high then A
- if outlook = sunny and humidity = normal then B
- if outlook = rain and windy = true then A
- if outlook = rain and windy = false then B

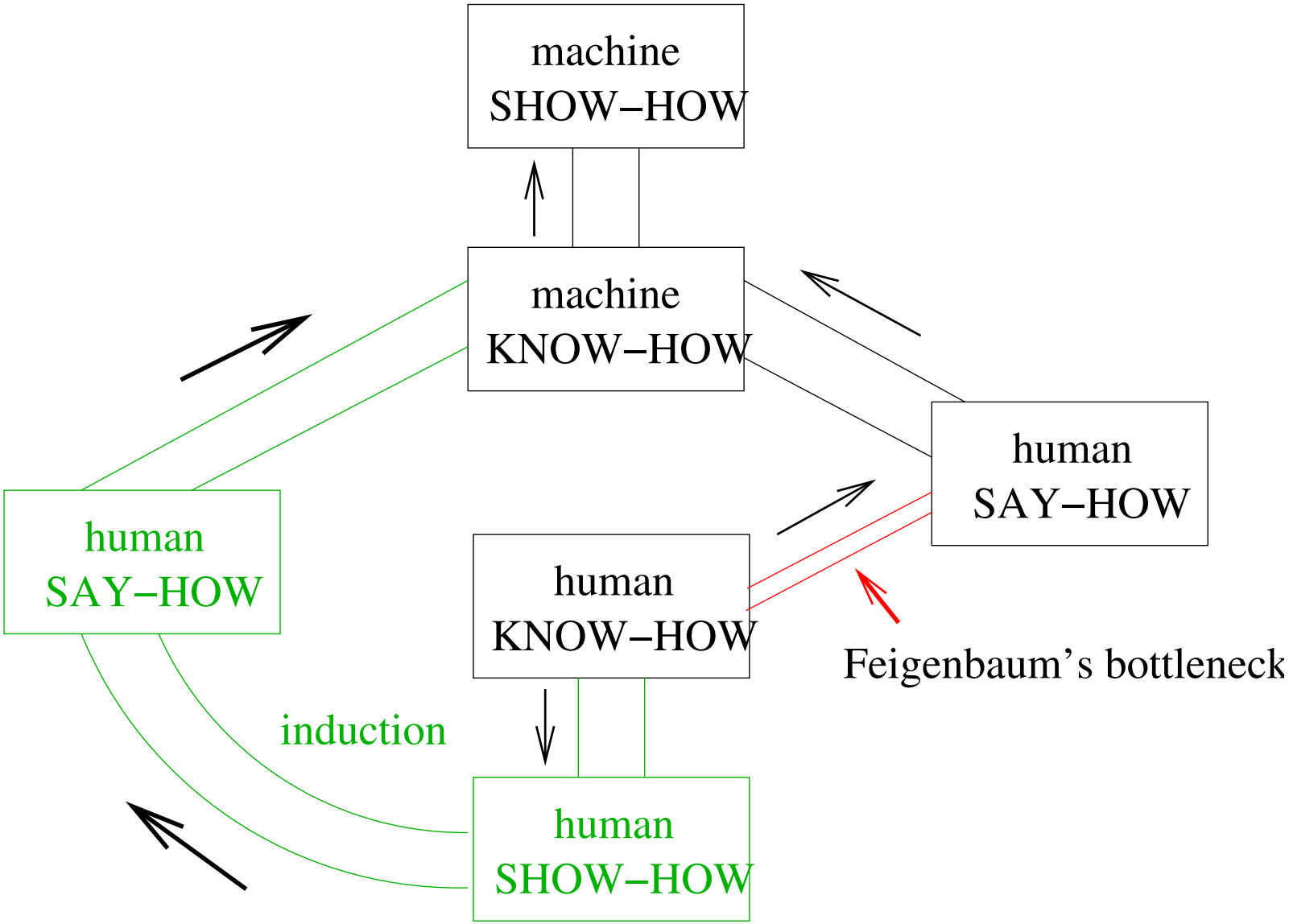
Intérêt d'apprendre des règles

- pour construire des classifieurs
- pour l'explication de phénomènes (caractère explicatif des règles)
- pour l'élaboration de la base de connaissance d'un système expert

Learning from examples (Michie, 1986)



Learning from examples (Michie, 1986)



Et le big data est arrivé...

"Léa poste une vidéo sur un réseau social, cela génère un like de Bart, qui génère lui-même un commentaire d'Eliana"

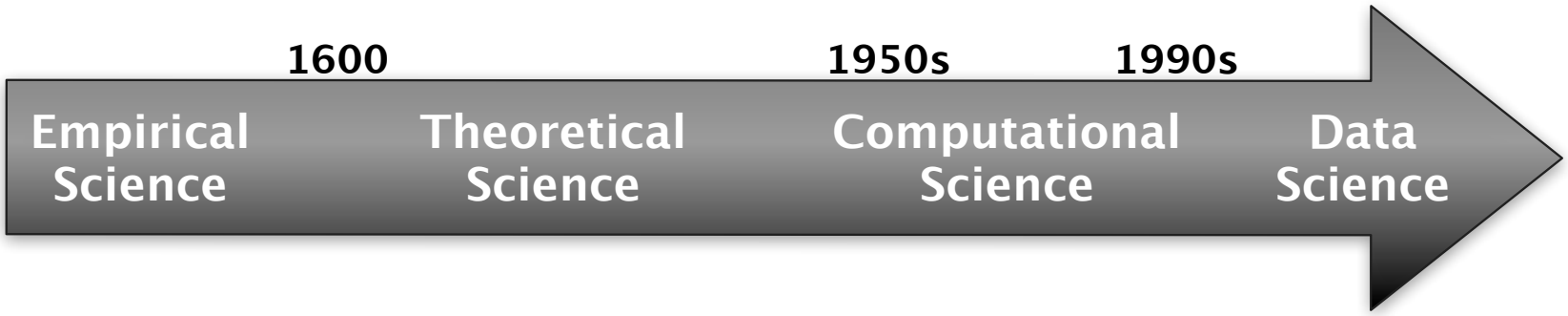
Masses des données qui transitent sur le web :
photos, vidéos, commentaires/forum, horaires de train, likes sur réseaux sociaux, historiques de données météo, etc.

Des données qui se caractérisent par :

- Variété
- Volumétrie
- Vélocité
- Véracité

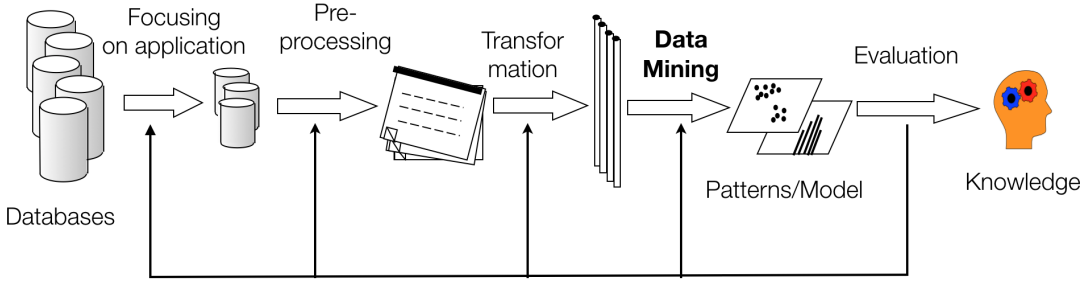
des gisements d'information et de connaissances nécessitant informaticiens, statisticiens, thématiciens pour les exploiter
➥ science des données

Évolution des sciences
sciences des données : 4ème pilier



- **science empirique** : observations de phénomènes naturels, extraction de lois générales par raisonnement inductif
- **science théorique** : modèles (mathématiques) pour comprendre un certain univers
- **science computationnelle** : simulation de phénomènes complexes pour comprendre ou valider des théories
- **science des données** : collecte massive de données, les masses de données guident la découverte de connaissances

Découverte de connaissances
. . . et data mining



Iterative and Interactive Process

data mining
(i.e. “fouille de données”) :
cœur du processus de découverte de connaissance

Comment “faire parler les données ?”

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth:
From Data Mining to Knowledge Discovery: An Overview.
Advances in Knowledge Discovery and Data Mining, pp. 1-34, 1996

Fouille de données versus extraction d'information

- **extraction d'information :** repérer les passages d'une collection de textes qui parlent de la maladie d'Alexander

- **fouille de données :**
  - découvrir que les gènes KHDRBS1 NONO TOP2B FMR1 semblent former un groupe de synexpression
  - découvrir une relation entre 2 gènes et la qualifier (e.g., active/inhibe, contexte biologique)

Data mining : caractéristiques (1/2)

- pas d'hypothèse initialement formulée sur les données
- capacités à faire ressortir des régularités locales (motifs)

Exemple : *analyse de données du LCR chez un enfant atteint de méningite*

si % de polynucléaires $> 80\%$
et nombre d'éléments par $ml > 900$
alors en faveur d'une méningite bactérienne

Autre exemple de règle :

si $0.34 \times$ protéinorachie
$- 0.33 \times$ gluchorachie
$+ 0.53 \times$ % de polynucléaires
$+ 0.15 \times$ nombre d'éléments par mm^3
> 0.7
alors en faveur d'une méningite bactérienne

Data mining : caractéristiques (2/2)

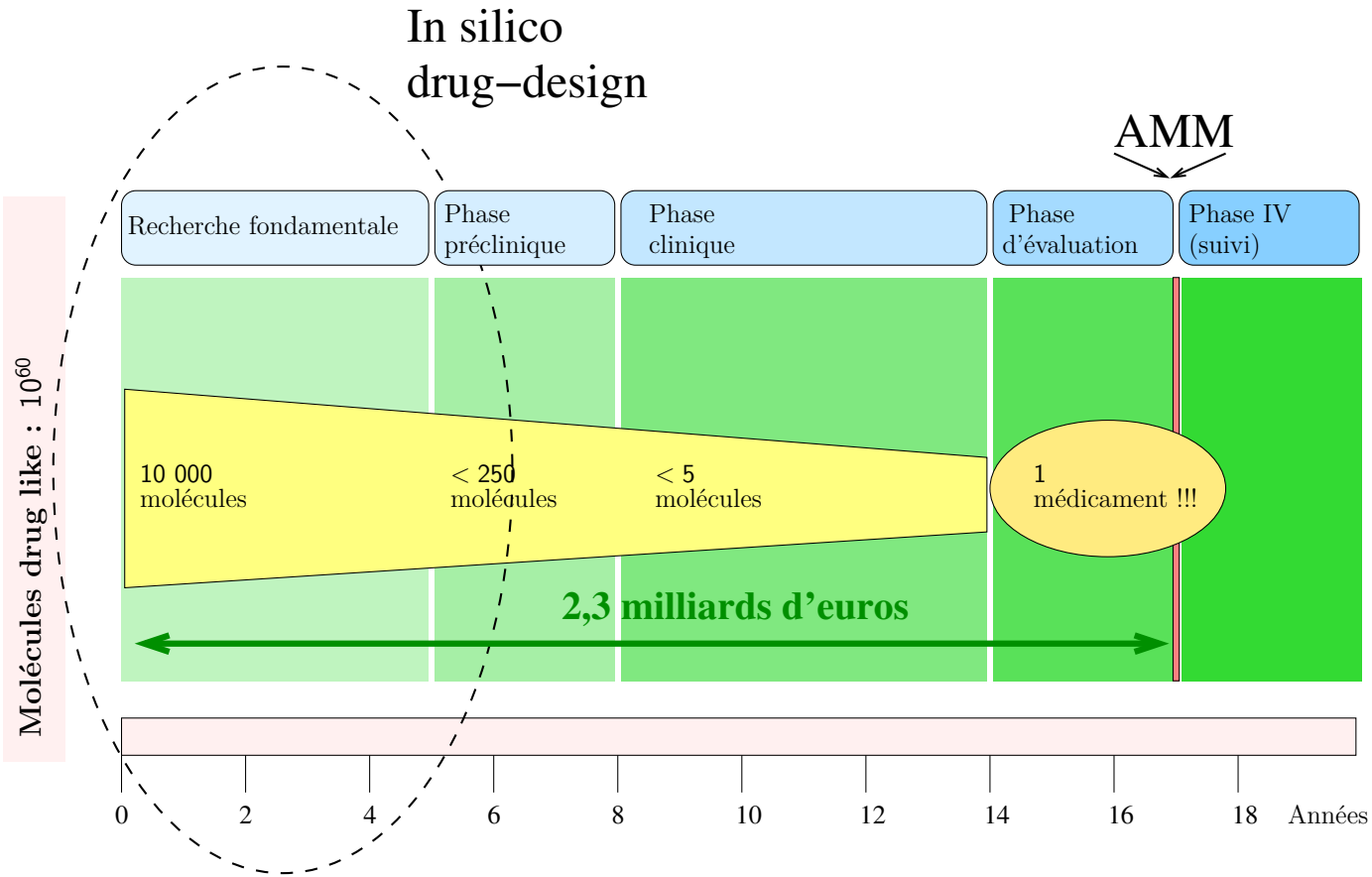
- une grande variété de :
  - méthodes : description versus prédiction, classification, clustering, ensembles de motifs, recherche d'exceptions, d'informations rares, de tendances, de ruptures,...
  - motifs : règles, clusters, contrastes, ensembles de règles,...
- au début de la discipline, "dogme" de fournir toutes les solutions à une requête (question/problème)

Data mining : caractéristiques (2/2)

- une grande variété de :
  - méthodes : description versus prédiction, classification, clustering, ensembles de motifs, recherche d'exceptions, d'informations rares, de tendances, de ruptures,...
  - motifs : règles, clusters, contrastes, ensembles de règles,...
- au début de la discipline, "dogme" de fournir toutes les solutions à une requête (question/problème)

  c'est le cas de l'algorithme "brute force" qui vous est suggéré dans l'exemple "fil rouge"

Un exemple : fouille de données pour la conception de médicaments



Kirkpatrick and Ellis, Chemical space, Nature 2004, vol. 432, pp 823– 823

Phrma profile 2015, http://www.phrma.org/sites/default/files/pdf/2015_phrma_profile.pdf

Quelle méthode ?
Motifs de contraste (1/2)

	d_1	d_2	d_3	d_4	d_5
mol_1	X				X
mol_2	X	X	X		X
mol_3				X	
mol_4	X		X		
mol_5	X		X	X	
mol_6	X		X		X
mol_7					X
mol_8		X			
mol_9	X	X			X
mol_{10}	X	X			

2 classes :

T : toxique

NT : non toxique

X : motif

exemple : {d1,d2}

{d1,d2} est présent dans les molécules [2,9,10]

Fréquence :

$F(\{d1, d2\}) = 3$

Motifs de contraste (2/2)

	d_1	d_2	d_3	d_4	d_5
mol_1	X				X
mol_2	X	X	X		X
mol_3				X	
mol_4	X		X		
mol_5	X		X	X	
mol_6	X		X		X
mol_7					X
mol_8		X			
mol_9	X	X			X
mol_{10}	X	X			

GR ("growth rate") pour mesurer le contraste :

$$GR_T(X) = \frac{|NT| \times F(X,T)}{|T| \times F(X,NT)}$$

{d1,d3} est présent dans :

- les molécules toxiques [2,4,5]
- les molécules non-toxiques [6]

$$GR_T(\{d1,d3\}) = \frac{5 \times 3}{5 \times 1} = 3$$

Motif émergent : $GR_{classe}(X) \geq mingr$

but : étant donné *mingr*, extraire tous les motifs émergents.

Quelle est la difficulté ?

Considérons une description très simple des molécules :

➥ n descripteurs binaires (présence/absence de fragments moléculaires)

Quelle est la taille de l'espace de recherche ?

Quelle est la difficulté ?

Considérons une description très simple des molécules :

➥ n descripteurs binaires (présence/absence de fragments moléculaires)

Quelle est la taille de l'espace de recherche ? 2^n (c'est vite grand...)

Exemple de temps de calcul :

(1 micro-seconde est nécessaire pour traiter une donnée)

Taille (n)	$log_2 n$	n	$nlog_2 n$	n^2	2^n
10	3×10^{-6}	10×10^{-6}	30×10^{-6}	100×10^{-6}	10^{-3}
100	7×10^{-6}	100×10^{-6}	700×10^{-6}	0.01	
1000	10×10^{-6}	10^{-3}	0.01	1	
10 000	13×10^{-6}	0.01	0.13	1.7 minute	
100 000	17×10^{-6}	0.1	1.7	2.8 heures	

Quelle est la difficulté ?

Considérons une description très simple des molécules :

➥ n descripteurs binaires (présence/absence de fragments moléculaires)

Quelle est la taille de l'espace de recherche ? 2^n (c'est vite grand...)

Exemple de temps de calcul :

(1 micro-seconde est nécessaire pour traiter une donnée)

Taille (n)	$log_2 n$	n	$nlog_2 n$	n^2	2^n
10	3×10^{-6}	10×10^{-6}	30×10^{-6}	100×10^{-6}	10^{-3}
100	7×10^{-6}	100×10^{-6}	700×10^{-6}	0.01	10^{14} siècles
1000	10×10^{-6}	10^{-3}	0.01	1	
10 000	13×10^{-6}	0.01	0.13	1.7 minute	
100 000	17×10^{-6}	0.1	1.7	2.8 heures	

Quelle est la difficulté ?

Considérons une description très simple des molécules :

➥ n descripteurs binaires (présence/absence de fragments moléculaires)

Quelle est la taille de l'espace de recherche ? 2^n (c'est vite grand...)

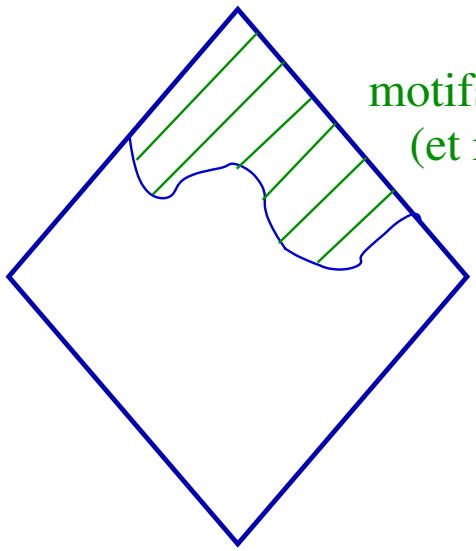
Exemple de temps de calcul :

(1 micro-seconde est nécessaire pour traiter une donnée)

Taille (n)	$log_2 n$	n	$nlog_2 n$	n^2	2^n
10	3×10^{-6}	10×10^{-6}	30×10^{-6}	100×10^{-6}	10^{-3}
100	7×10^{-6}	100×10^{-6}	700×10^{-6}	0.01	10^{14} siècles
1000	10×10^{-6}	10^{-3}	0.01	1	astronomique
10 000	13×10^{-6}	0.01	0.13	1.7 minute	astronomique
100 000	17×10^{-6}	0.1	1.7	2.8 heures	astronomique

Heikki Mannila : "data mining is the art of counting"

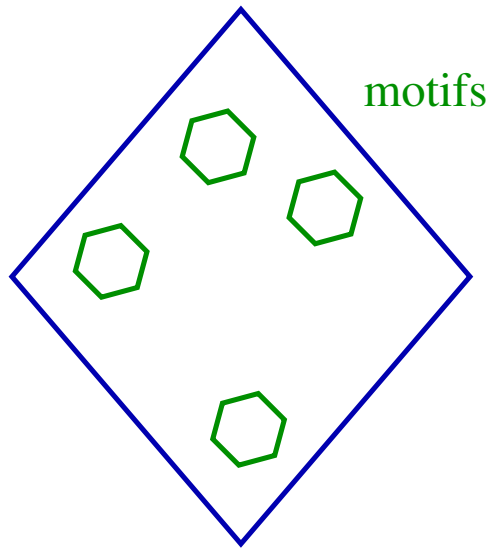
Mais faut-il parcourir *tout* l'espace de recherche ?



motifs fréquents
(et règles d'association)

➥ non

(mais cela peut rester coûteux)



motifs de contraste

➥ a priori oui

Pourquoi non-monotonicité pour les motifs de contraste ?

quand on spécialise un motif, le numérateur et le dénominateur diminuent, mais cela peut être aussi bien le numérateur que le dénominateur qui diminue le plus vite

"solution" : élagage aux bornes (approche "branch and bound")