



Data Mining

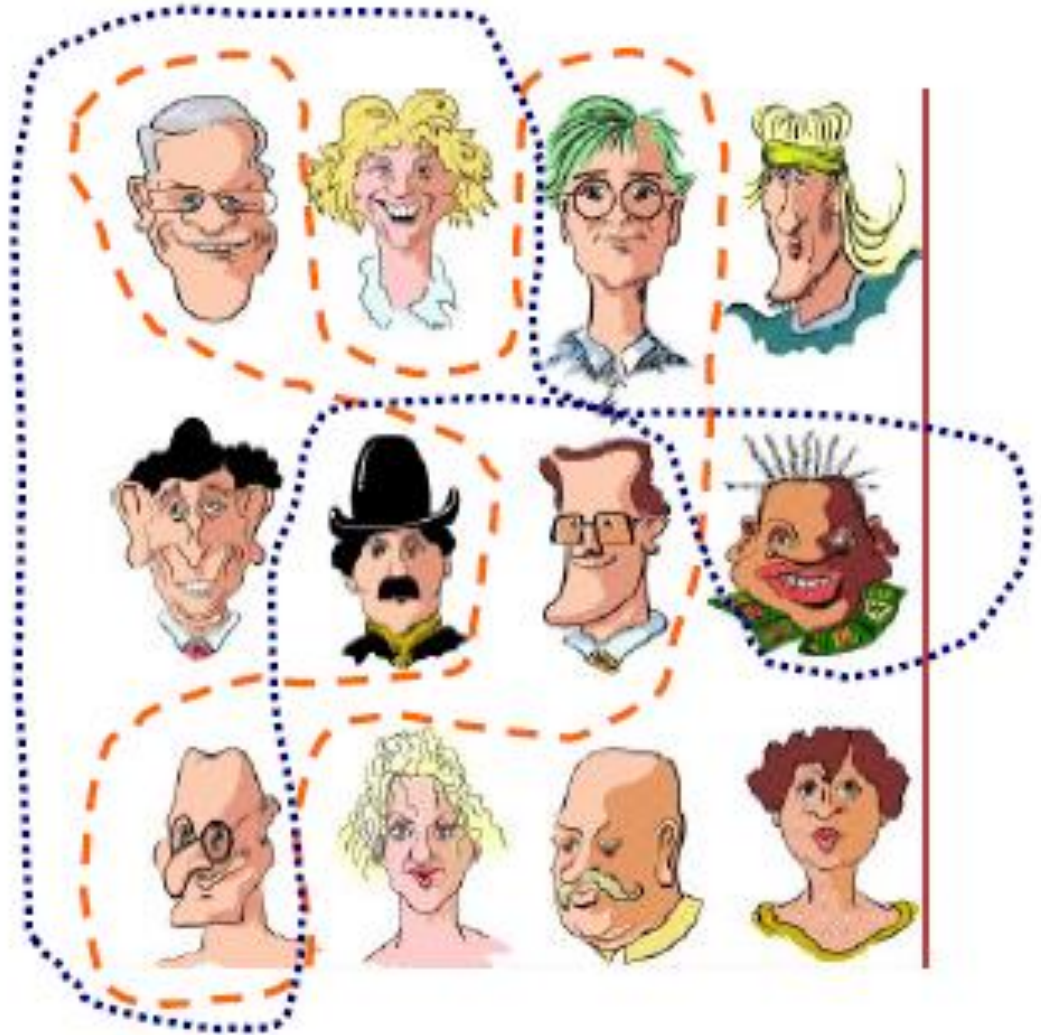
Mohammed Fethi KHALFI

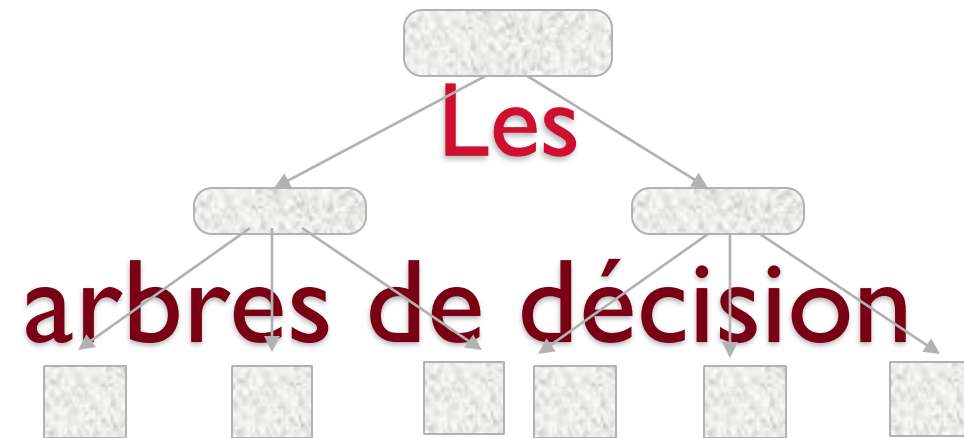
Fethi.Khalfi@yahoo.fr

Arbres de décision

Exemple: Regroupement de personnes

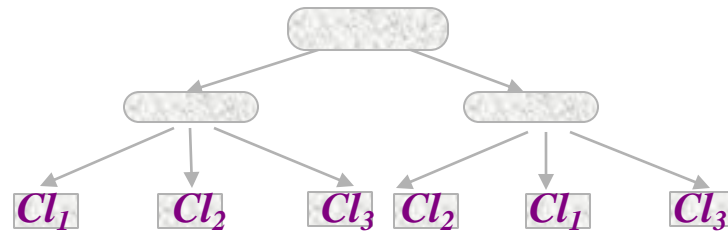
*Sexe,
lunettes,
sourire,
chapeau*



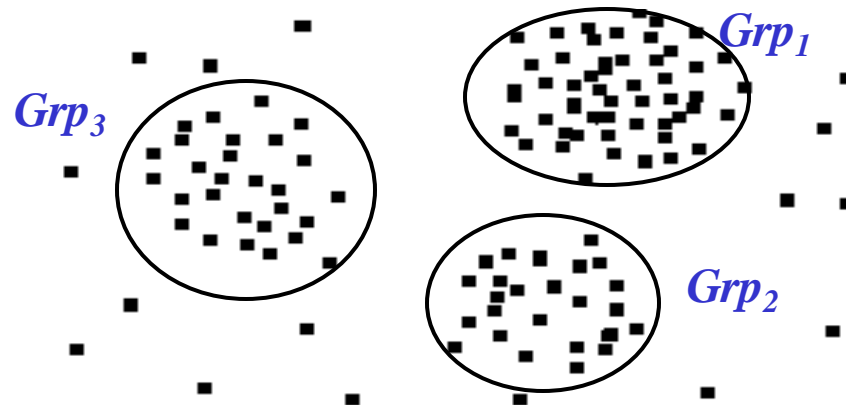


La classification

- **Supervisée : on connaît les classes**



- ***Non supervisée : on ne connaît pas les classes***



La classification

- Supervisée : **on connaît les classes**
 - ♦ Bayésienne
 - ♦ Réseaux neuronaux
 - ♦ Arbres de décision (**Apprentissage**)
 - ♦ ...
- Non supervisée : **on ne connaît pas les classes**
 - ♦ K-moyennes, nuées dynamiques, CLARANS,...
 - ♦ Classification Ascendante Hiérarchique (**Analyse des données**)

Problèmes difficiles

- Pour certains domaines d'application, il est essentiel de produire des procédures de classification compréhensibles par l'utilisateur.
- Comment interpréter les symptômes de mon patient ?
- Ma voiture ne démarre pas, comment dois-je procéder ?
- À quelle heure dois-je me lever pour être en cours à 9h30 ?
- Comment caser ces bagages dans le coffre de ma voiture ?
- Puis-je encore optimiser mon emploi du temps ?
- Est-ce que cet étudiant peut faire un bon Master ?
- Puis-je écrire un résumé de 100 lignes de cet article ?
- La traduction de ce poème est-elle bonne ?

Induction d'arbres de décision

- Les arbres de décision répondent à cette contrainte car ils représentent graphiquement un ensemble de règles et sont aisément interprétables.
- Les algorithmes d'apprentissage par arbres de décision sont efficaces, disponibles dans la plupart des environnements de fouille de données. Ils constituent l'objet de ce cours.

Classification: arbre de décision

- **Entrée:**

BD = Exemples classés décrits par des attributs

- **Sortie:**

Arbre classifiant les exemples en classes

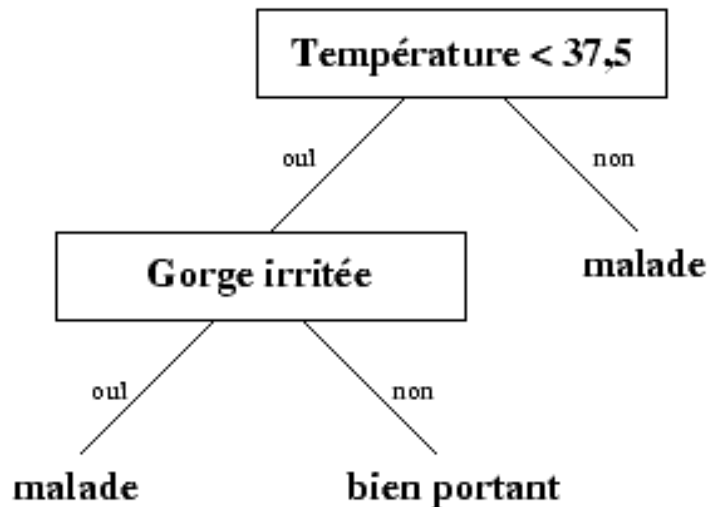
- **Approche:**

Organiser les exemples en arbre, les feuilles sont les classes

- **Méthodes:** Cart, C4.5 ...

Induction d'arbres de décision

- Exemple : La population est constituée d'un ensemble de patients. Il y a **deux classes** : **malade** et **bien portant**. Les descriptions sont faites avec les deux attributs : Température qui est un attribut à valeurs décimales et gorge irritée qui est un attribut logique. On considère l'arbre de décision de la figure

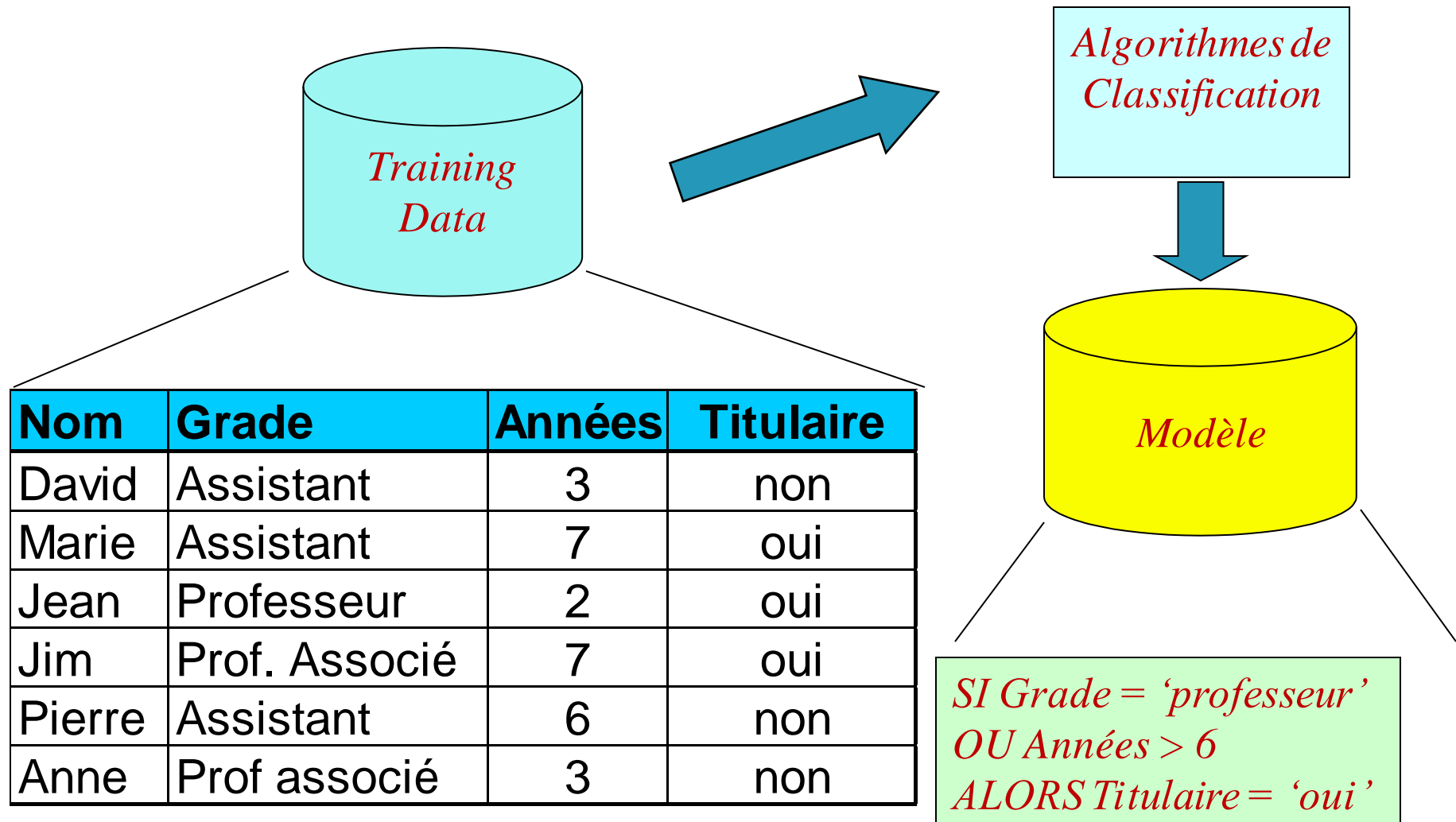


Un patient ayant une température de 39 et ayant la gorge non irritée sera classé comme

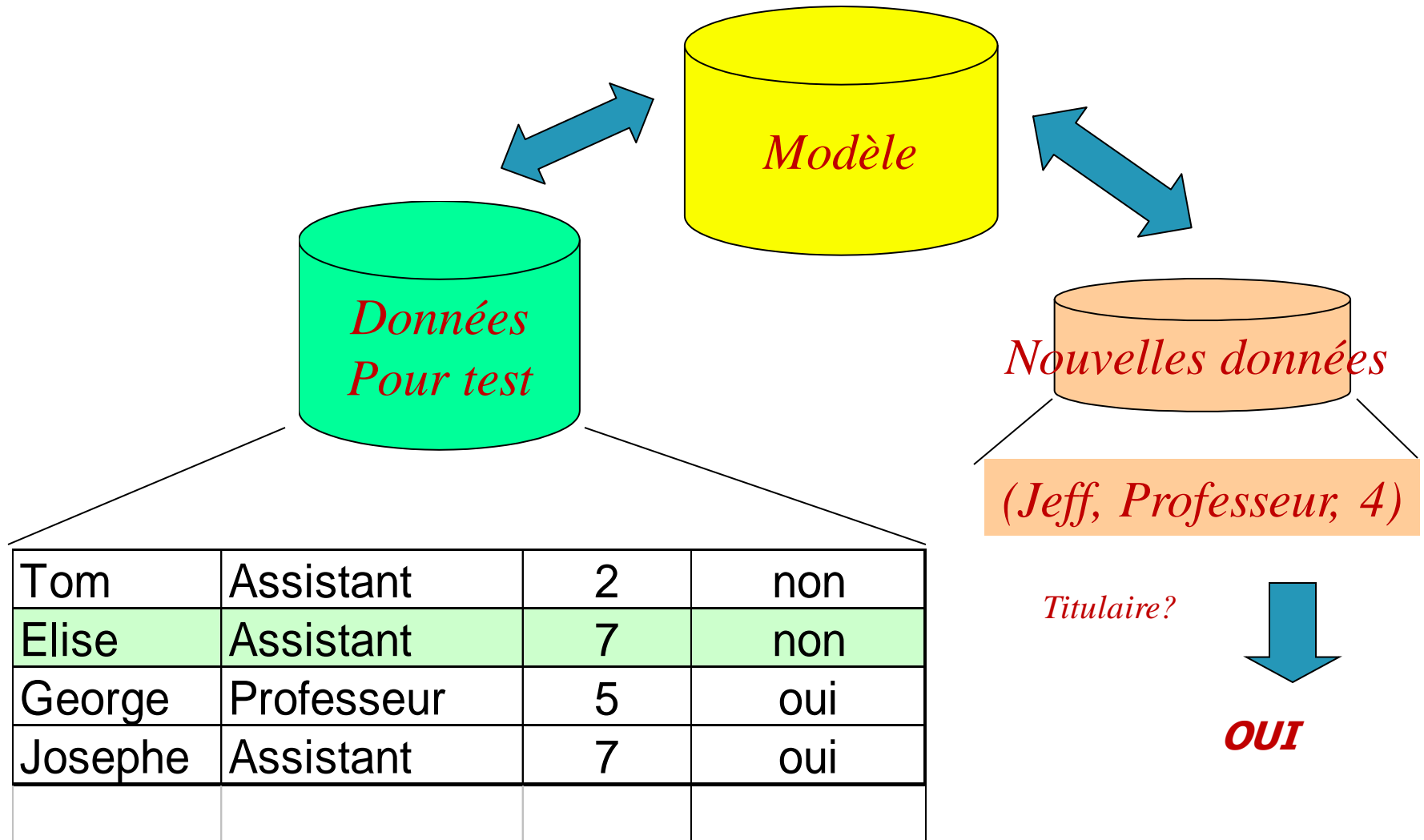
Arbres de décision

- Basée sur la théorie de l'information
- Fonctionnant pour des variables continues ou discrètes
- Recherche itérative de variables discriminantes
- Produisant des modèles faciles à interpréter (sous forme de règles SI ... ALORS ... SINON)

Processus de Classification (I): Construction du modèle



Processus de Classification (2): Prédiction



Induction d'arbres de décision :

Exemple de données météorologiques

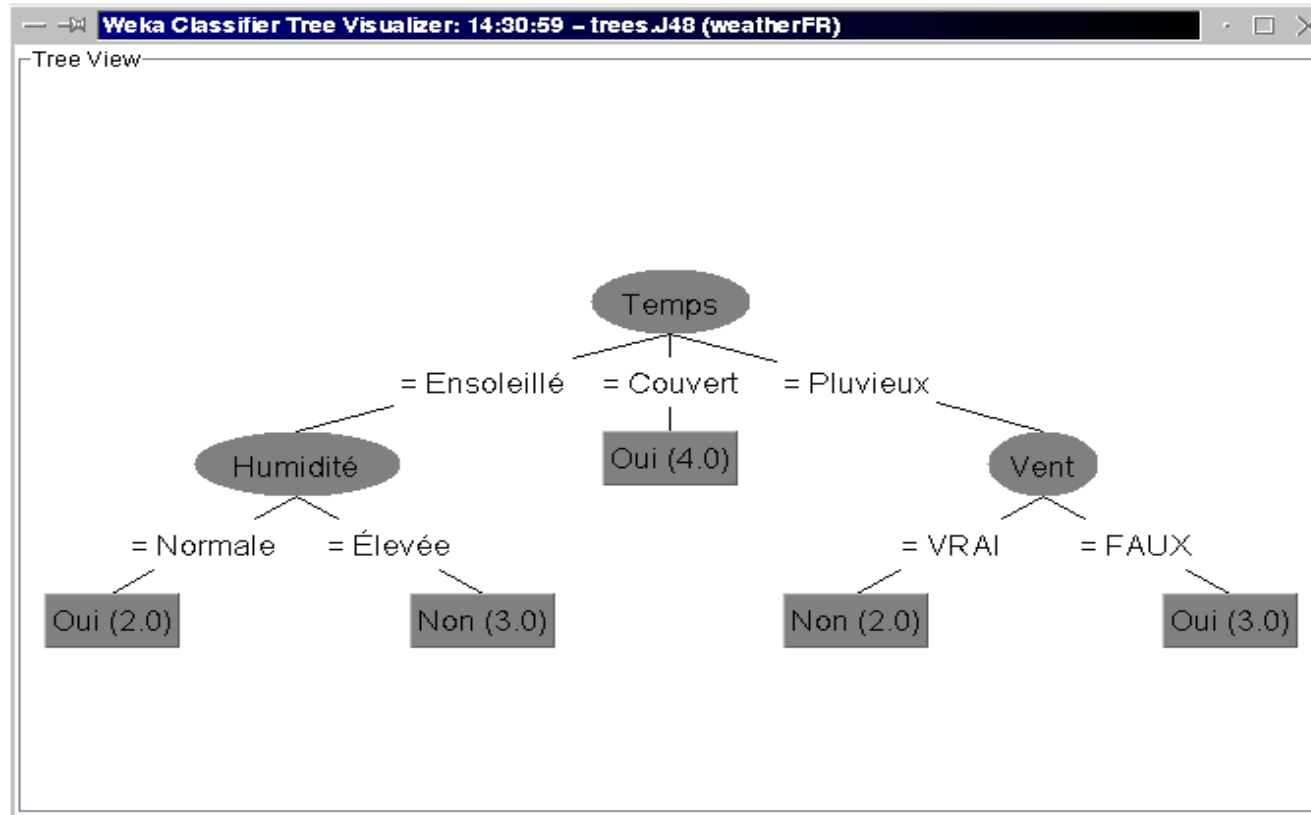
Attributs prédictifs				Attribut de classes
Temps	Température	Humidité	Vent	Tennis ?
Ensoleillé	Chaude	Élevée	FAUX	Non
Ensoleillé	Chaude	Élevée	VRAI	Non
Couvert	Chaude	Élevée	FAUX	Oui
Pluvieux	Modérée	Élevée	FAUX	Oui
Pluvieux	Fraîche	Normale	FAUX	Oui
Pluvieux	Fraîche	Normale	VRAI	Non
Couvert	Fraîche	Normale	VRAI	Oui
Ensoleillé	Modérée	Élevée	FAUX	Non
Ensoleillé	Fraîche	Normale	FAUX	Oui
Pluvieux	Modérée	Normale	FAUX	Oui
Ensoleillé	Modérée	Normale	VRAI	Oui

Exemple : Est-ce que les conditions sont favorables pour jouer au tennis?

Classifier l'instance suivante:

<Ciel = Ensoleillé, Température = chaud, Humidité = élevé, Vent = fort>

Induction d'arbres de décision :



*Nouvelle
journée*

Temps	Température	Humidité	Vent	Tennis ?
Ensoleillé	Frais	Élevée	VRAI	?

Induction d'arbres de décision :

Attributs	Pif	Temp	Humid	Vent
Valeurs possibles	soleil,couvert,pluie	chaud,bon,frais	normale,haute	vrai,faux

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

la classe

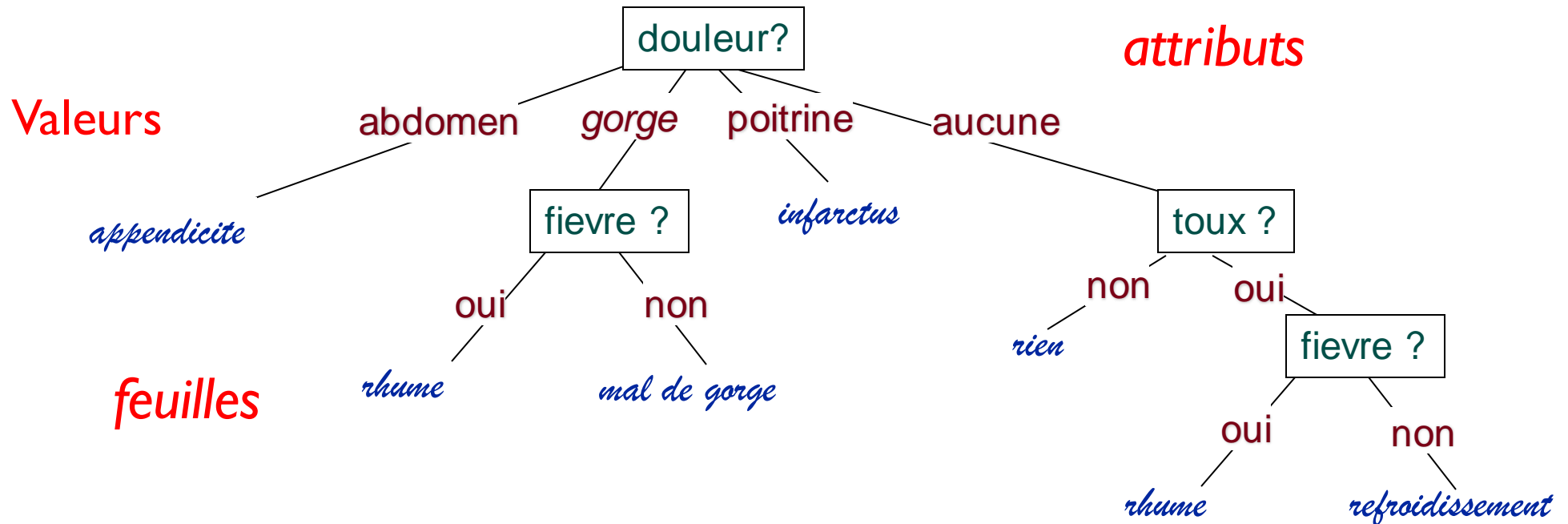
Attributs

instance

Objectif: prévoir si l'on va jouer au golf (ou pas)

I - Les arbres de décision : exemple

- Les arbres de décision sont des classifieurs pour des instances représentées dans un formalisme attribut/valeur
 - Les **nœuds** de l'arbre testent les attributs
 - Il y a une **branche** pour chaque valeur de l'attribut testé
 - Les **feuilles** spécifient les catégories (deux ou plus)




I - Les arbres de décision : le problème

- Chaque **instance** est décrite par un vecteur d'attributs/valeurs

	<u>Toux</u>	<u>Fièvre</u>	<u>Poids</u>	<u>Douleur</u>
Marie	non	oui	normal	gorge
Fred	non	oui	normal	abdomen
Julie	oui	oui	maigre	aucune
Elvis	oui	non	obese	poitrine

- **En entrée** : un ensemble d'instances et leur classe (correctement associées par un “professeur” ou “expert”)

	<u>Toux</u>	<u>Fièvre</u>	<u>Poids</u>	<u>Douleur</u>	
Marie	non	oui	normal	gorge	 Diagnostic rhume appendicite
Fred	non	oui	normal	abdomen	
.....					

- L'algorithme d'apprentissage doit construire un **arbre de décision**

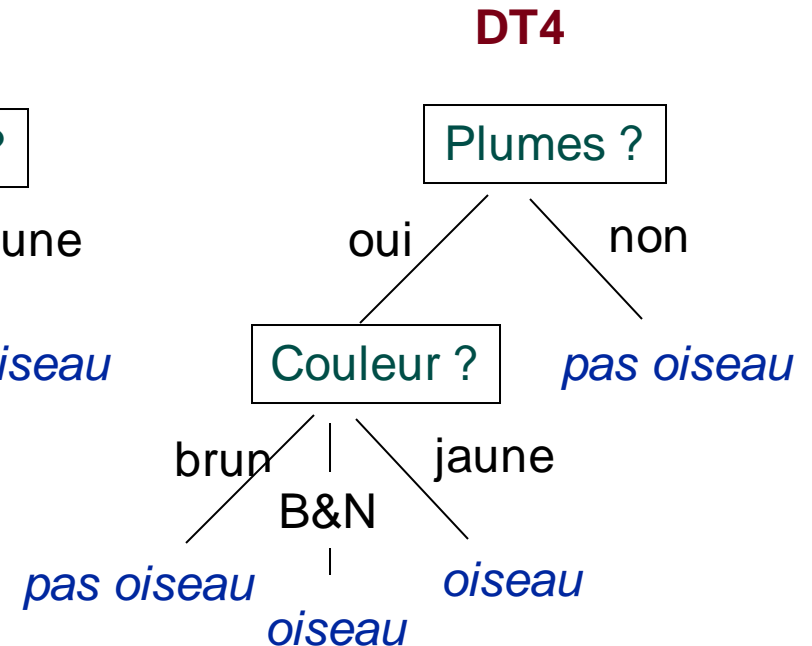
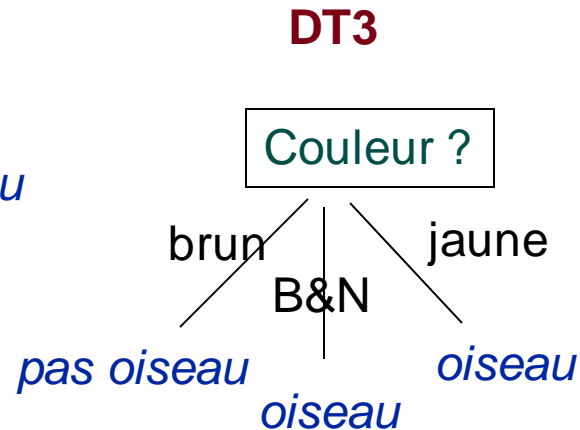
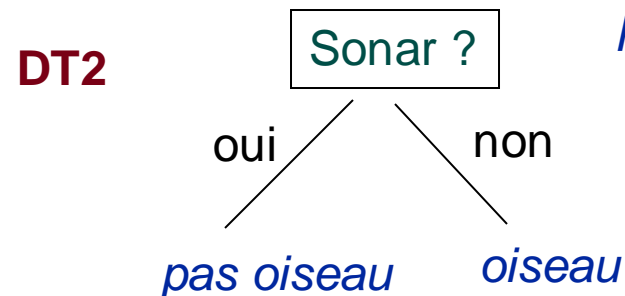
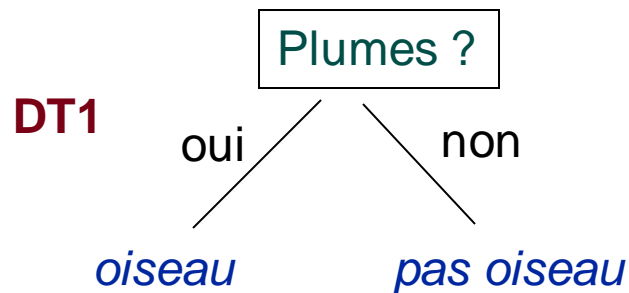
E.g. Un arbre de décision pour le diagnostic

Une des principales applications de l'apprentissage !

2- Les arbres de décision : le choix d'un arbre

	Couleur	Ailes	Plumes	Sonar	<u>Concept</u>
Faucon	jaune	oui	oui	non	<i>oiseau</i>
Pigeon	B&N	oui	oui	non	<i>oiseau</i>
chauve-souris	brun	oui	non	oui	<i>pas oiseau</i>

Quatre arbres de décision cohérents avec les données:



Algorithmes

- Pour construire un tel arbre, plusieurs algorithmes existent : ID3, CART, C4.5,...etc. On commence généralement par le choix d'un attribut puis le choix d'un nombre de critères pour son nœud. On crée pour chaque critère un nœud concernant les données vérifiant ce critère.
- L'algorithme continue d'une façon récursive jusqu'à obtenir des nœuds concernant les données de chaque même classe.

Algorithmes

- **Algorithme ID3**
- ID3 construit l'arbre de décision récursivement. A chaque étape de la récursion, il calcule parmi les attributs restant pour la branche en cours, celui qui maximisera le gain d'information. C'est-à-dire l'attribut qui permettra le plus facilement de classer les exemples à ce niveau de cette branche de l'arbre. Le calcul se fait à base de l'entropie de Shannon déjà présentée. L'algorithme suppose que tous les attributs sont catégoriels ; si des attributs sont numériques, ils doivent être décritisés pour pouvoir l'appliquer. ID3 utilise l'algorithme Construire_arbre précédent.

3- Induction d'arbres de décision : Exemple [Quinlan,86]

Attributs Valeurs possibles		Pif soleil,couvert,pluie	Temp chaud,bon,frais	Humid normale,haute	Vent vrai,faux
--------------------------------	--	-----------------------------	-------------------------	------------------------	-------------------

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

← la classe

3- Le critère entropique

(1/3)

- **L'entropie d'information est:**

$$Entropy = -\frac{p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

- **Calculate Average Information:**

$$I(Attribut) = \sum \frac{p_i + n_i}{p+n} Entropy(A)$$

- **Calculate Information Gain:** (Difference in Entropy before and after splitting dataset on attribute A)

$$Gain = Entropy(S) - I(Attribut)$$

Exemple

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

P = 9

N = 5

Total = 14

- Calculate **Entropy(S)**:

$$Entropy = \frac{-p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy(S) = \frac{-9}{9+5} \log_2 \left(\frac{9}{9+5} \right) - \frac{5}{9+5} \log_2 \left(\frac{5}{9+5} \right)$$

$$Entropy(S) = \frac{-9}{14} \log_2 \left(\frac{9}{14} \right) - \frac{5}{14} \log_2 \left(\frac{5}{14} \right) = 0.940$$

Exemple

For each Attribute: (let say **pif**)
Calculate Entropy for each Values, i.e for soleil', pluie',couvert

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

For each Attribute: (let say **pif**) *Exemple*
 Calculate Entropy for each Values, i.e for soleil', pluie',couvert

Pif	Golf
soleil	NePasJouer
soleil	NePasJouer
soleil	NePasJouer
soleil	Jouer
soleil	Jouer

Pif	Golf
pluie	Jouer
pluie	Jouer
pluie	NePasJouer
pluie	Jouer
pluie	NePasJouer

Pif	Golf
couvert	Jouer
couvert	Jouer
couvert	Jouer
couvert	Jouer

pif	p	n	Entropy
soleil	2	3	0,971
pluie	3	2	0,971
couvert	4	0	0

Example

For each Attribute: (let say **pif**)
Calculate Entropy for each Values, i.e for soleil ', pluie',couvert

pif	p	n	Entropy
soleil	2	3	0,971
pluie	3	2	0,971
couvert	4	0	0

$$Entropy = \frac{-p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy (pif= soleil) = -\frac{2}{5} \log \left(\frac{2}{5} \right) - \frac{3}{5} \log \left(\frac{3}{5} \right) = 0.971$$

$$Entropy (pif= *pluie*) = -1 \log(1) - 0 \log(0) = 0$$

$$Entropy (pif= *couvert*) = -\frac{3}{5} \log \left(\frac{3}{5} \right) - \frac{2}{5} \log \left(\frac{2}{5} \right) = 0.971$$

Exemple

Calculate **Average Information Entropy**:

$$I(\text{Attribute}) = \sum \frac{p_i + n_i}{p + n} \text{Entropy}(A)$$

$$I(\text{pif}) = \frac{p_{\text{soleil}} + n_{\text{soleil}}}{p + n} \text{Entropy}(\text{pif} = \text{soleil}) +$$

$$\frac{p_{\text{pluie}} + n_{\text{pluie}}}{p + n} \text{Entropy}(\text{pif} = \text{pluie}) +$$

$$\frac{p_{\text{couvert}} + n_{\text{couvert}}}{p + n} \text{Entropy}(\text{pif} = \text{couvert})$$

pif	p	n	Entropy
soleil	2	3	0,971
pluie	3	2	0,971
couvert	4	0	0

$$I(\text{pif}) = \frac{3 + 2}{9 + 5} * 0.971 + \frac{2 + 3}{9 + 5} * 0.971 + \frac{4 + 0}{9 + 5} * 0 = 0.693$$

Exemple

Calculate Gain: attribute is pif:

$$Gain = Entropy(S) - I(Attribute)$$

pif	p	n	Entropy
soleil	2	3	0,971
pluie	3	2	0,971
couvert	4	0	0

$$Entropy(S) = 0.940$$

$$Gain(Outlook) = 0.940 - 0.693 = 0.247$$

pif

Exemple

For each Attribute: (let say **humid**)
Calculate Entropy for each Values, i.e for haute', normale

Humid	Golf
haute	NePasJouer
haute	NePasJouer
haute	Jouer
haute	Jouer
haute	NePasJouer
haute	Jouer
haute	NePasJouer

Humid	Golf
normale	Jouer
normale	NePasJouer
normale	Jouer
normale	Jouer
normale	Jouer
normale	Jouer
normale	Jouer

Humid	p	n	Entropy
haute	3	4	0,985
normale	6	1	0,591

Exemple

Calculate **Average Information Entropy**:

$$I(\text{Attribute}) = \sum \frac{p_i + n_i}{p + n} \text{Entropy}(A)$$

Humid	p	n	Entropy
haute	3	4	0,985
normale	6	1	0,591

$$I(\text{Humid}) = \frac{p_{\text{haute}} + n_{\text{haute}}}{p + n} \text{Entropy}(\text{Humid} = \text{haute}) +$$

$$\frac{p_{\text{Normal}} + n_{\text{Normal}}}{p + n} \text{Entropy}(\text{Humid} = \text{Normal})$$

$$I(\text{Humid}) = \frac{3 + 4}{9 + 5} * 0.985 + \frac{6 + 1}{9 + 5} * 0.591 => 0.788$$

Exemple

Calculate Gain: attribute is pif:

$$Gain = Entropy(S) - I(Attribute)$$

Humid	p	n	Entropy
haute	3	4	0,985
normale	6	1	0,591

$$Entropy(S) = 0.940$$

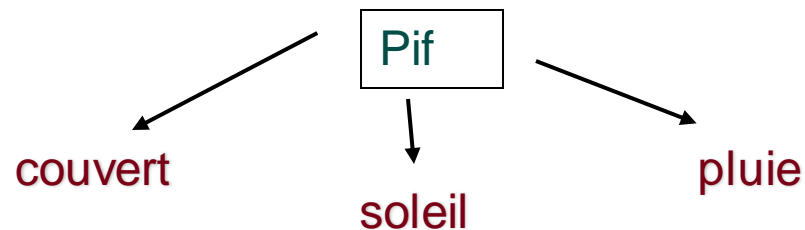
$$Gain(Humid) = 0.940 - 0.788 = 0.152$$

Exemple

Pick the **highest gain** attribute.

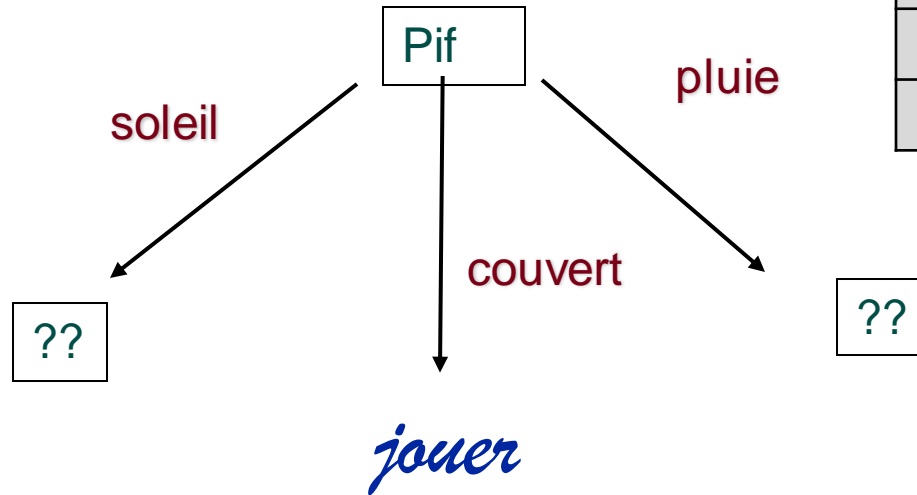
Attribut	Gain
pif	0,247
temp	0,029
humid	0,152
vent	0,048

Root Node:



Exemple

Pick the **highest gain** attribute.



Pif	Golf
couvert	Jouer
couvert	Jouer
couvert	Jouer
couvert	Jouer

Exemple

For each Attribute: (let say **Temp**)
Calculate Entropy for each Values, i.e for soleil', pluie',couvert

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

Exemple

Repeat the same thing for sub-trees till we get the tree.

pif= « soleil »

Pif	Temp	Humid	Vent	Golf
soleil	chaud	haute	faux	NePasJouer
soleil	chaud	haute	vrai	NePasJouer
soleil	bon	haute	faux	NePasJouer
soleil	frais	normale	faux	Jouer
soleil	bon	normale	vrai	Jouer

pif= « pluie »

Pif	Temp	Humid	Vent	Golf
pluie	bon	haute	faux	Jouer
pluie	frais	normale	faux	Jouer
pluie	frais	normale	vrai	NePasJouer
pluie	bon	normale	faux	Jouer
pluie	bon	haute	vrai	NePasJouer

	Doublant	Série	Mention	Classe
1	Non	Maths	ABien	Admis
2	Non	Techniques	ABien	Admis
3	Oui	Sciences	ABien	Non Admis

Exemple

Repeat the same thing for sub-trees till we get the tree.

pif= « soleil »

Pif	Temp	Humid	Vent	Golf
soleil	chaud	haute	faux	NePasJouer
soleil	chaud	haute	vrai	NePasJouer
soleil	bon	haute	faux	NePasJouer
soleil	frais	normale	faux	Jouer
soleil	bon	normale	vrai	Jouer

P=2
N=3
Total=5

- ENTROPY:

$$Entropy = \frac{-p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

$$Entropy(S_{oleil}) = \frac{-2}{2+3} \log_2 \left(\frac{2}{2+3} \right) - \frac{3}{2+3} \log_2 \left(\frac{3}{2+3} \right)$$

$$\Rightarrow 0.971$$

Exemple

For each Attribute: (let say **humid**)
Calculate Entropy for each Values, i.e for haute,normale

Pif	Humid	Golf
soleil	haute	NePasJouer
soleil	haute	NePasJouer
soleil	haute	NePasJouer
soleil	normale	Jouer
soleil	normale	Jouer

Humid	p	n	Entropy
haute	0	3	0
normale	2	0	0

- Calculate **Average Information Entropy** $I(\text{Humid} = 0)$
- Calculate **Gain**: $\text{Gain} = 0.971$

Exemple

For each Attribute: (let say **humid**)
Calculate Entropy for each Values, i.e for haute,normale

Pif	Temp	Golf
soleil	chaud	NePasJouer
soleil	chaud	NePasJouer
soleil	bon	NePasJouer
soleil	frais	Jouer
soleil	bon	Jouer

temp	p	n	Entropy
chaud	0	2	0
bon	1	1	0
frai	1	0	1

- Calculate **Average Information Entropy**:

$$I(\text{Temp}) = 0.4$$

- Calculate **Gain**:

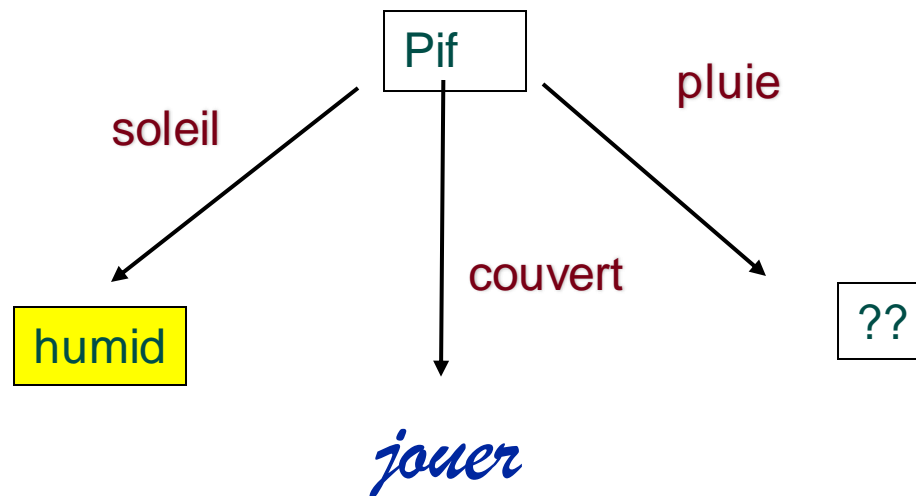
$$\text{Gain} = 0.571$$

Active Windows

Exemple

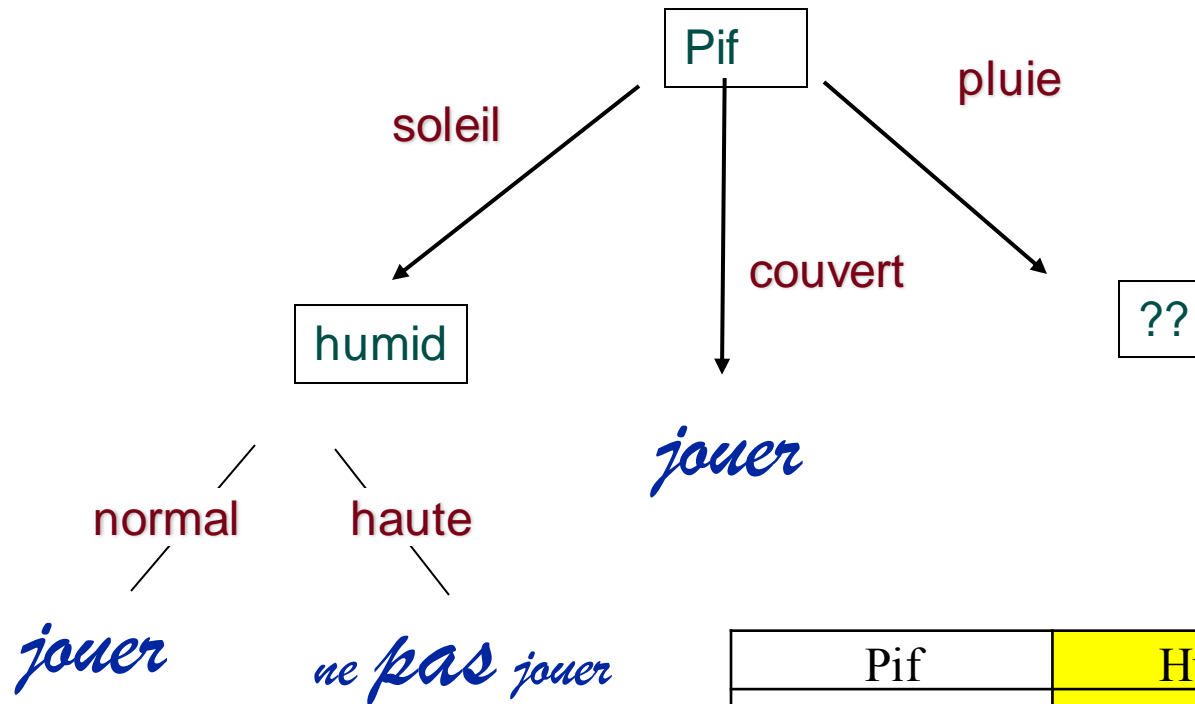
Pick the **highest gain** attribute.

Attribut	Gain
temp	0,571
humid	0,971
vent	0,02



Exemple

Pick the **highest gain** attribute.



Pif	Humid	Golf
soleil	haute	NePasJouer
soleil	haute	NePasJouer
soleil	haute	NePasJouer
soleil	normale	Jouer
soleil	normale	Jouer

Exemple

Repeat the same thing for sub-trees till we get the tree.

pif= « pluie »

Pif	Temp	Humid	Vent	Golf
pluie	bon	haute	faux	Jouer
pluie	frais	normale	faux	Jouer
pluie	frais	normale	vrai	NePasJouer
pluie	bon	normale	faux	Jouer
pluie	bon	haute	vrai	NePasJouer

P=3

N=2

Total = 5

- ENTROPY:

$$Entropy = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$$Entropy(\text{'Pluie'}) = \frac{-3}{3+2} \log_2\left(\frac{3}{3+2}\right) - \frac{2}{3+2} \log_2\left(\frac{2}{2+3}\right)$$

=>0.971

Exemple

For each Attribute: (let say **humid**)
Calculate Entropy for each Values, i.e for haute,normale

pif= « pluie »

Pif	Humid	Golf
pluie	haute	Jouer
pluie	normale	Jouer
pluie	normale	NePasJouer
pluie	normale	Jouer
pluie	haute	NePasJouer

Humid	p	n	Entropy
haute	1	1	0
normale	2	1	0

- Calculate **Average Information Entropy**: $I(\text{Humidity}) = 0.951$
- Calculate **Gain**: $\text{Gain} = 0.020$

Exemple

For each Attribute: (let say **temp**)
Calculate Entropy for each Values, i.e for haute,normale

pif= « pluie »

Pif	Temp	Golf
pluie	bon	Jouer
pluie	frais	Jouer
pluie	frais	NePasJouer
pluie	bon	Jouer
pluie	bon	NePasJouer

temp	p	n	Entropy
bon	2	1	0
frai	1	1	0

- Calculate **Average Information Entropy:** $I(\text{Temp}) = 0.951$
- Calculate **Gain:** $\text{Gain} = 0.020$

Exemple

For each Attribute: (let say **vent**)
Calculate Entropy for each Values, i.e for vrai, faux

pif= « pluie »

Pif	Vent	Golf
pluie	faux	Jouer
pluie	faux	Jouer
pluie	vrai	NePasJouer
pluie	faux	Jouer
pluie	vrai	NePasJouer

temp	p	n	Entropy
vrai	0	2	0
faux	3	0	0

- Calculate **Average Information Entropy**:

$$I(\text{vent}) = 0$$

- Calculate **Gain**:

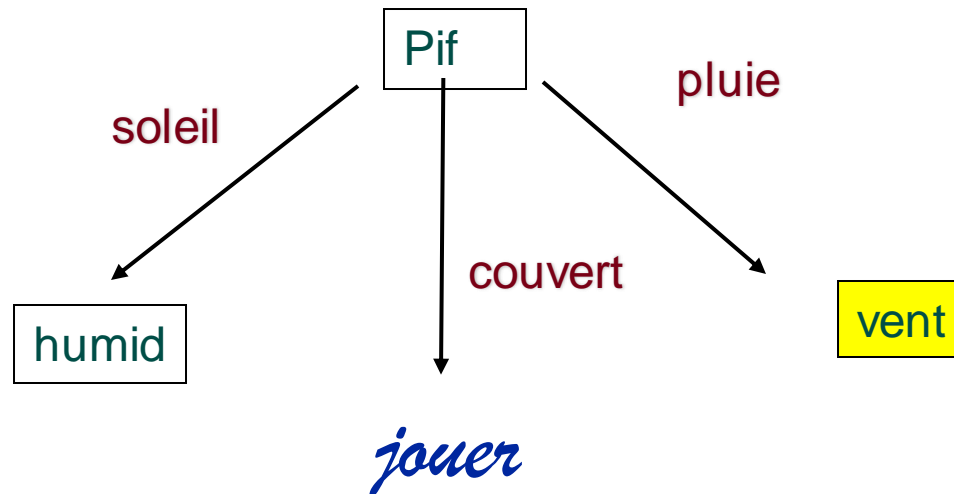
$$\text{Gain} = 0.971$$

Active Windows

Exemple

Pick the **highest gain** attribute.

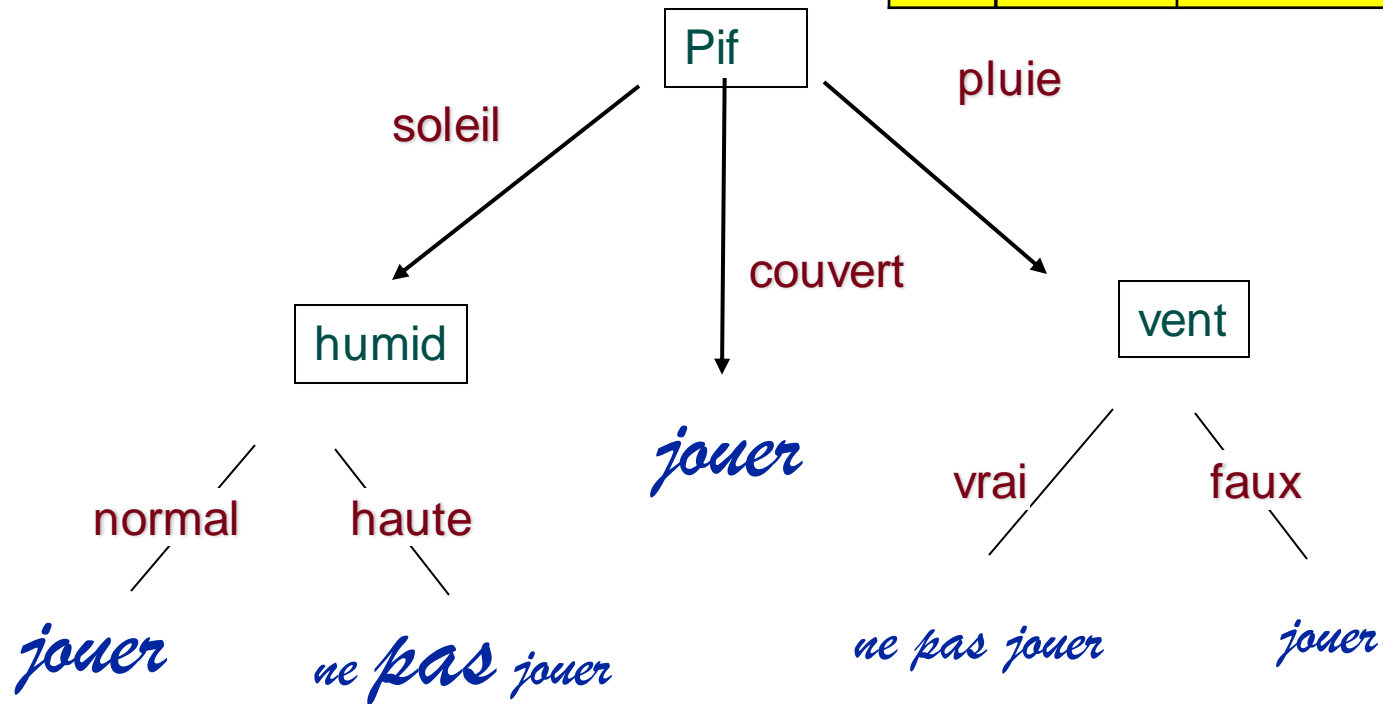
Attribut	Gain
temp	0,02
humid	0,02
vent	0,971



Exemple

Pick the **highest gain** attribute.

Pif	Vent	Golf
pluie	faux	Jouer
pluie	faux	Jouer
pluie	vrai	NePasJouer
pluie	faux	Jouer
pluie	vrai	NePasJouer



Exemple

For each Attribute: (let say **pif**)
Calculate Entropy for each Values, i.e for soleil', pluie',couvert

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer