



Data Mining

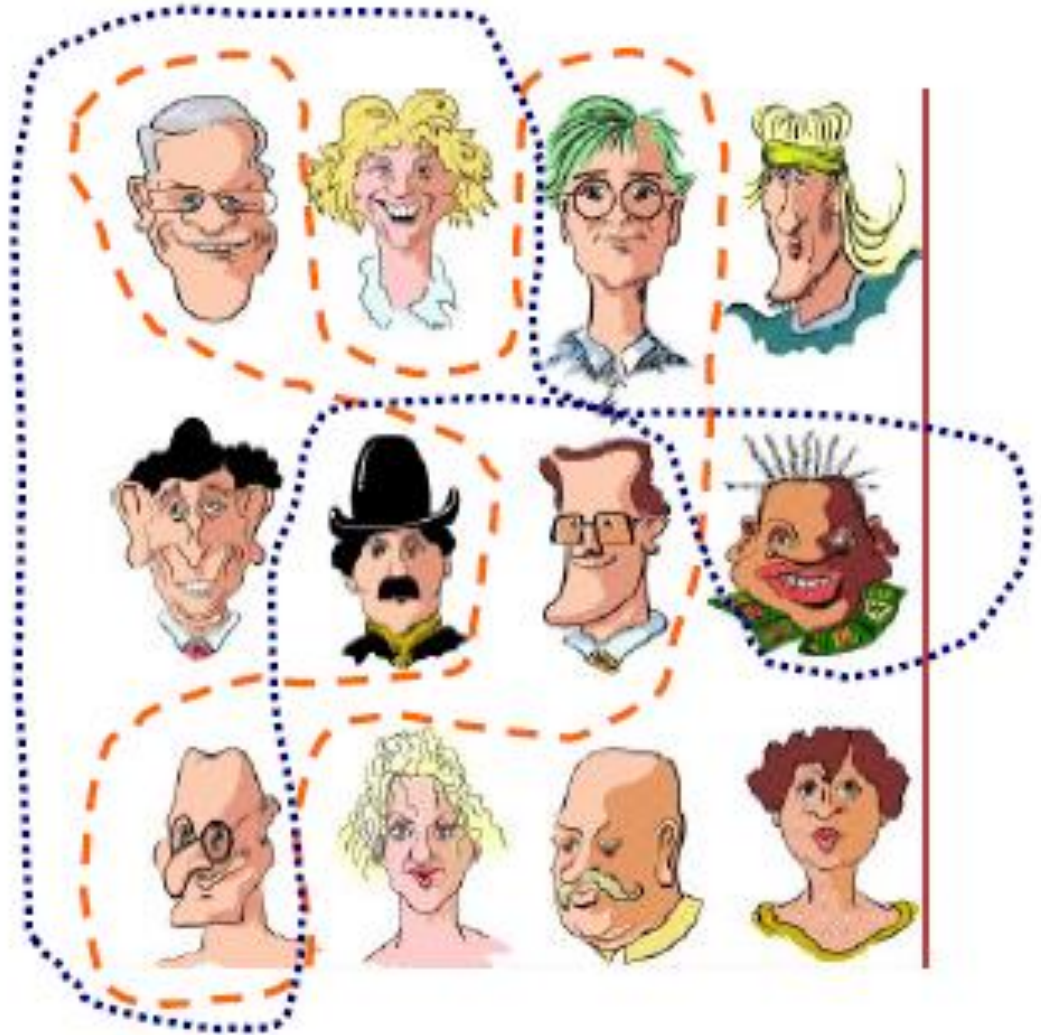
Mohammed Fethi KHALFI

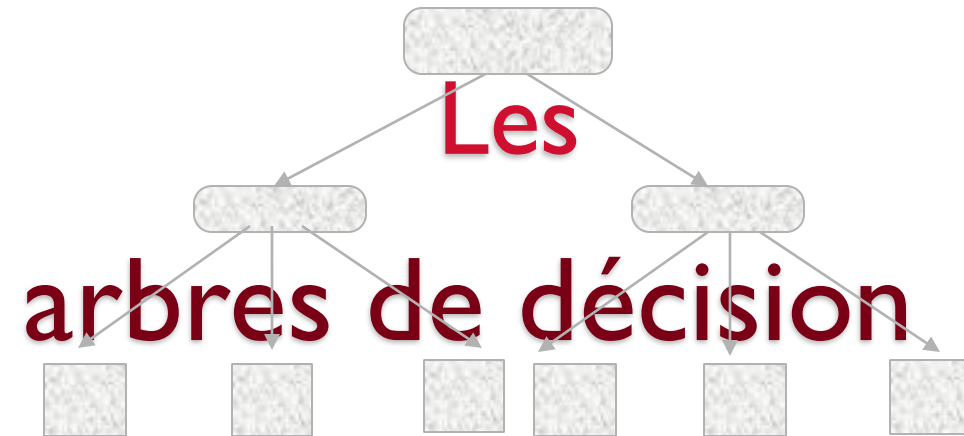
Fethi.Khalfi@yahoo.fr

arbres de décision II

Exemple: Regroupement de personnes

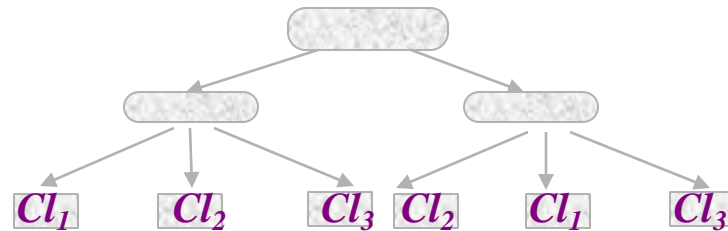
*Sexe,
lunettes,
sourire,
chapeau*



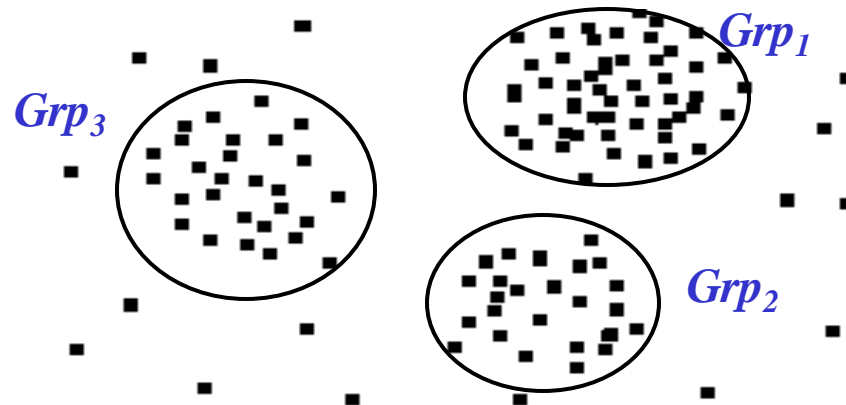


La classification

- **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

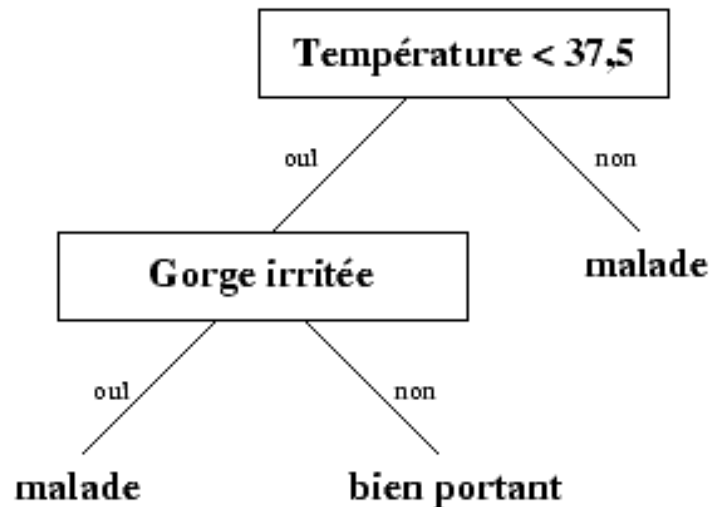
- **Approche:**

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

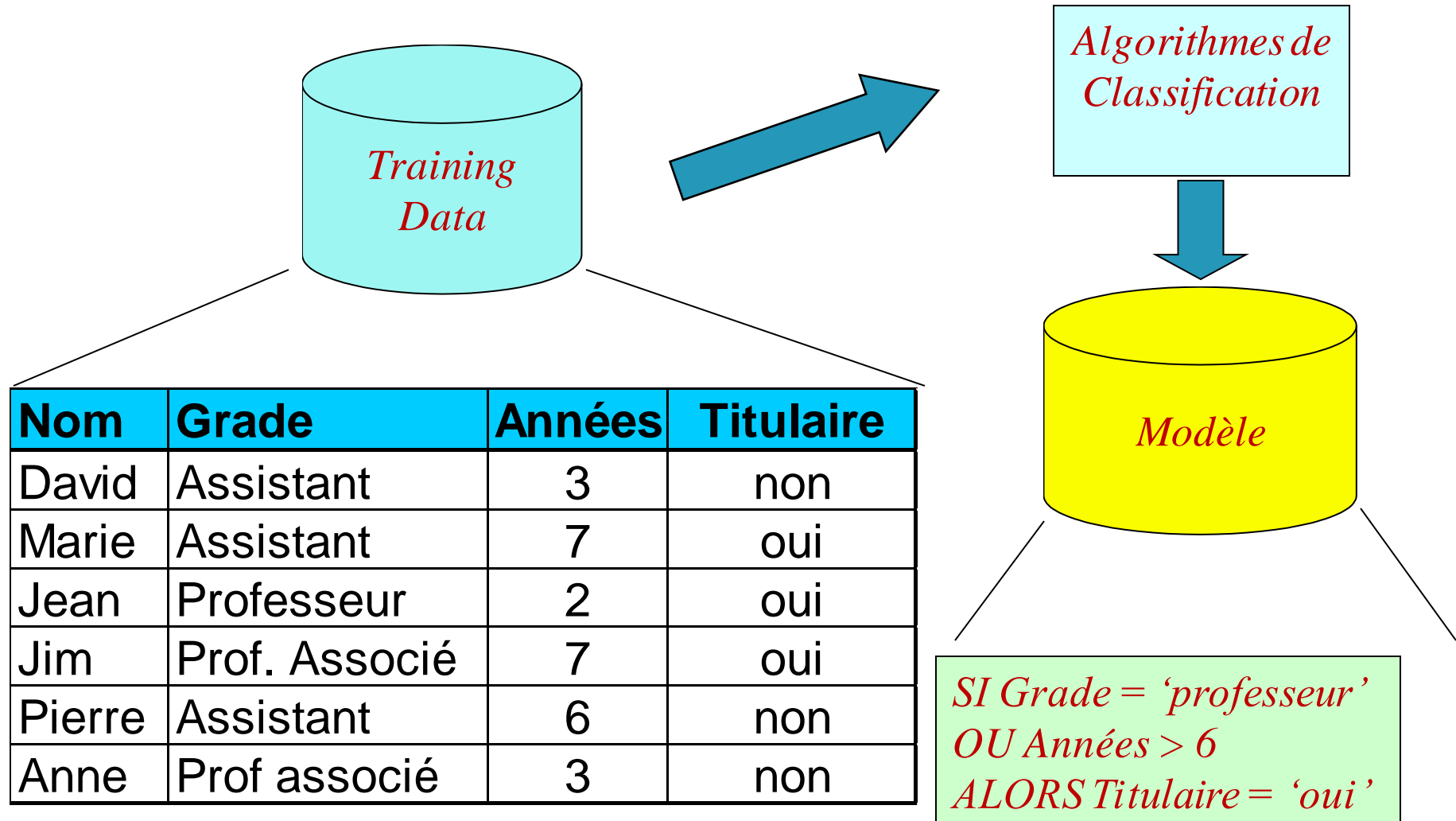


Un patient ayant une température de 39 et ayant la gorge non irritée sera classé comme

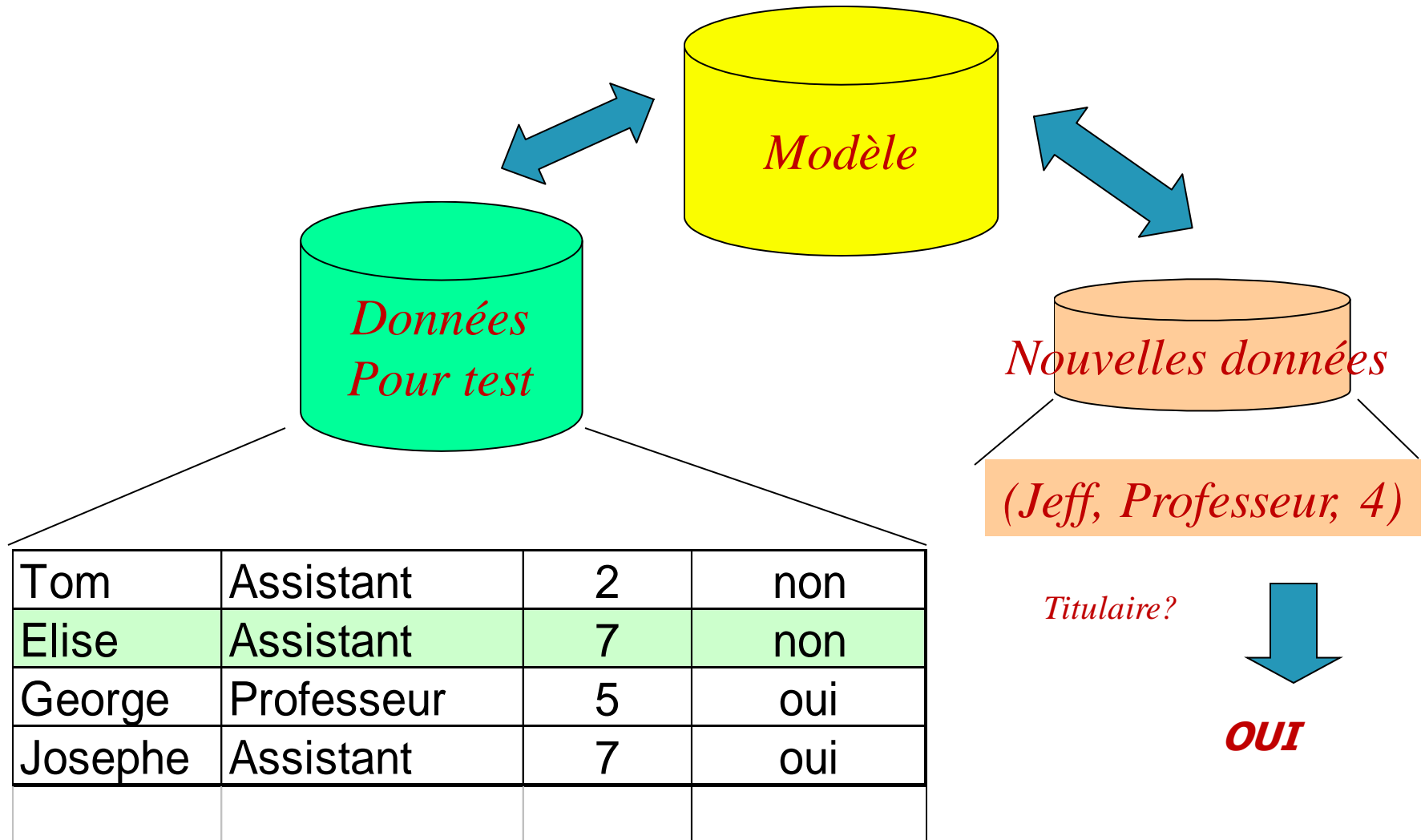
Arbres de décision

- 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

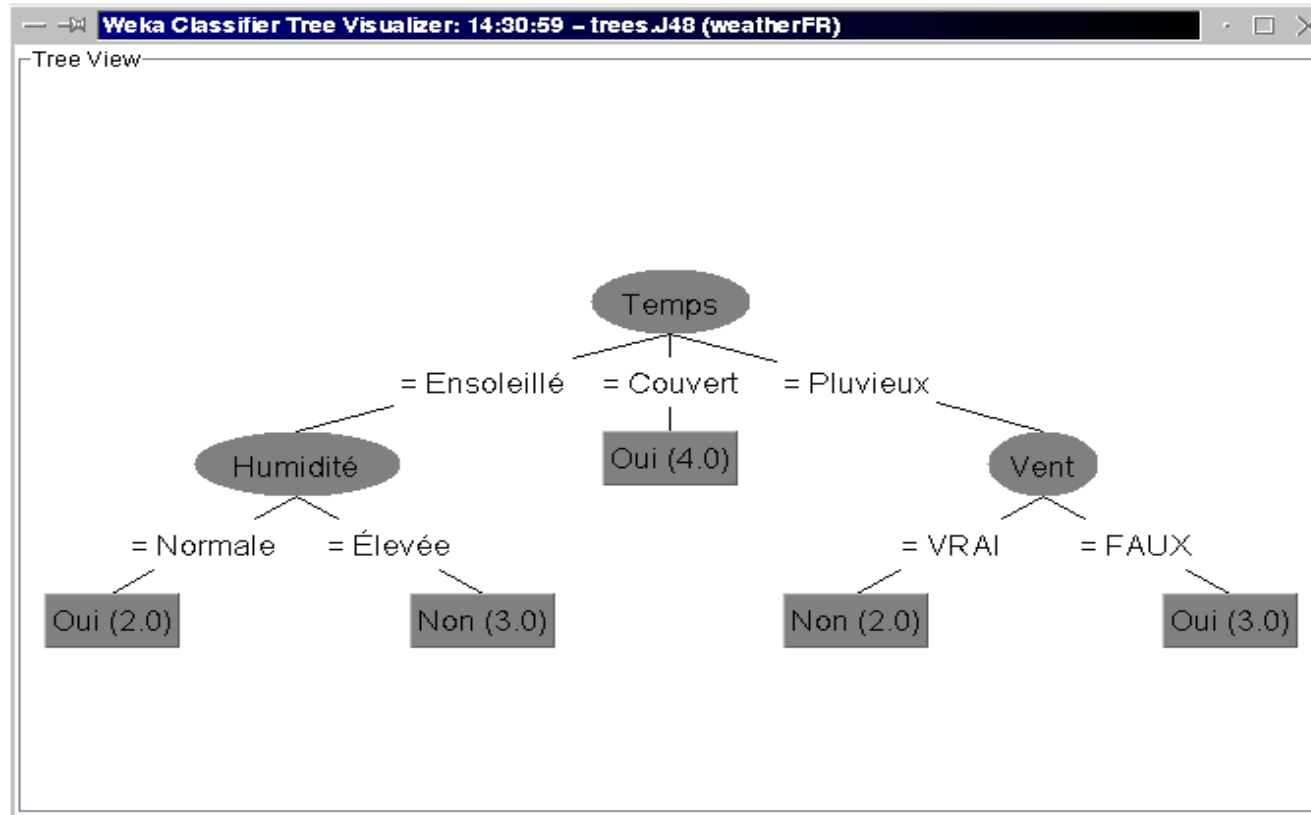
14
Exemples

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

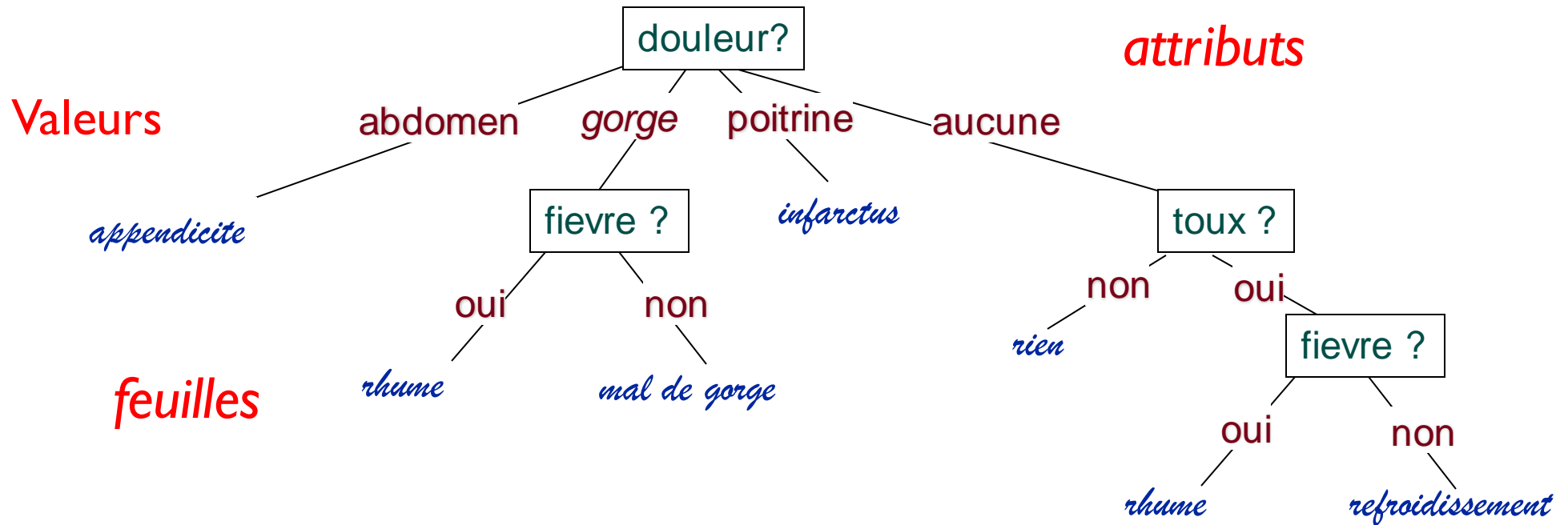
Attributs

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**

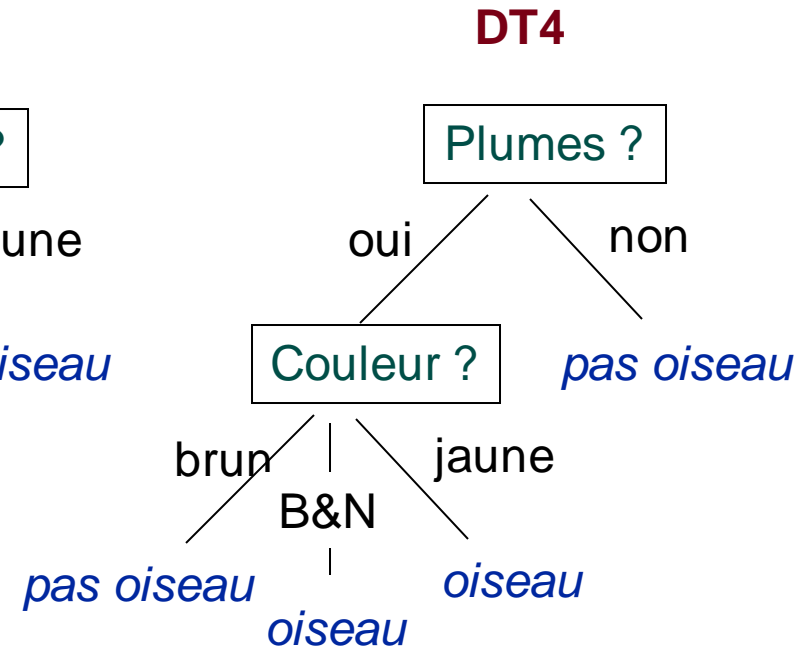
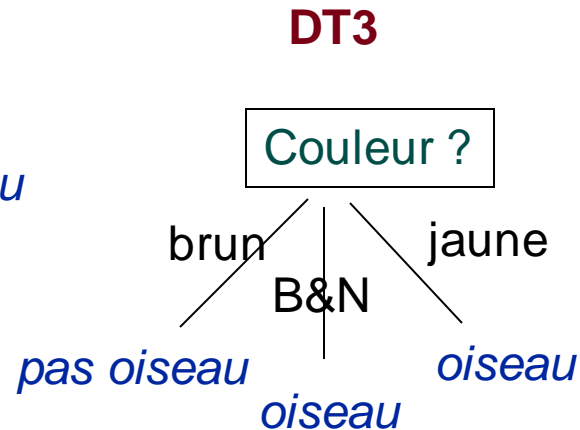
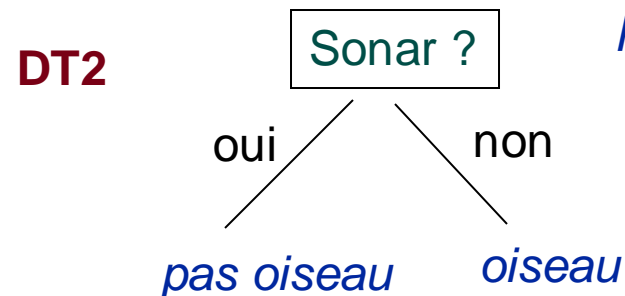
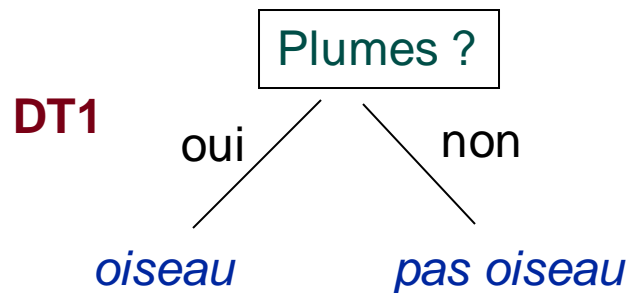
E.g. Un arbre de décision pour le diagnostic

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 CART
- l'algorithme CART (Classification And Regression Tree) basé sur l'indice de Gini et sur l'élaboration de noeuds binaires.

3- d'arbres de décision : CART

- CART choisit donc l'attribut et le seuil qui maximisent la décroissance de l'impureté du nœud par rapport à la cible.
- En classification, la mesure de l'impureté utilisée est l'index (ou impureté) de Gini plus il est bas plus il est pur.
- L'impureté (ou l'index de Gini IG(S)) ou Mesure du désordre : GINI pour un nœud S est calculée comme suit :

$$\underline{Gini = 1 - \sum_{i=1}^C (p_i)^2} \quad \text{for } i=1 \text{ to number of classes}$$

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

Exemple

For each Attribute: (let say **pif**)
Calculate Gini Index 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	nbr
soleil	2	3	5
pluie	3	2	5
couvert	4	0	4

Exemple

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

pif	p	n	nbr
soleil	2	3	5
pluie	3	2	5
couvert	4	0	4

$$Gini = 1 - \sum_{i=1}^c (p_i)^2$$

- Gini (pif = soleil)= $1 - [(2/5)^2 + (3/5)^2] = 0,48$
- Gini (pif = couvert)= $1 - [(4/4)^2 + (0/4)^2] = 0$
- Gini (pif = *pluie*)= $1 - [(3/5)^2 + (2/5)^2] = 0,48$

Calculate sum of gini index:

$$\bullet \text{Gini (pif)} = (5/14) 0,48 + (4/14) 0 + (5/14) 0,48 = 0,342$$

Exemple

For each Attribute: (let say **humid**)
Calculate Gini Index 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	nbr
haute	3	4	7
normale	6	1	7

Exemple

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

Humid	p	n	nbr
haute	3	4	7
normale	6	1	7

$$Gini = 1 - \sum_{i=1}^c (p_i)^2$$

- Gini (Humid= *haute*) = $1 - [(3/7)^2 + (4/7)^2] = 0.489$
- Gini (Humid= *normale*) = $1 - [(6/7)^2 + (1/7)^2] = 0.244$

Calculate **sum of gini index**:

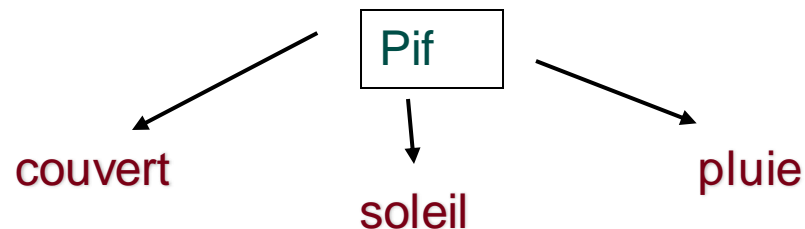
$$\bullet \text{Gini (Humid)} = (7/14) \cdot 0.489 + (7/14) \cdot 0.244 \dots = 0.367$$

Exemple

Pick the **highest gain** attribute.

Attribut	Gain
pif	0,342
temp	0,439
humid	0,367
vent	0,428

Root Node:



Exemple

Pick the **highest gain** attribute.

Root Node:

soleil

Pif

pluie

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	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

couvert

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

Exemple

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

pif= « soleil »

Pif	Temp	Humid	Vent	Golf	Temp	p	n	nbr
soleil	chaud	haute	faux	NePasJouer	chaud	0	3	3
soleil	chaud	haute	vrai	NePasJouer	frais	1	0	1
soleil	chaud	haute	faux	NePasJouer	bon	1	0	1
soleil	frais	normale	faux	Jouer				
soleil	bon	normale	vrai	Jouer				

- Gini (temp= *chaud*)= $1 - [(0/3)^2 + (3/3)^2] = 0,2$
- Gini (temp = *frais*)= $1 - [(1/1)^2 + (0/1)^2] = 0$
- Gini (temp = *bon*)= $1 - [(1/5)^2 + (0/5)^2] = 0$

Calculate sum of gini index:

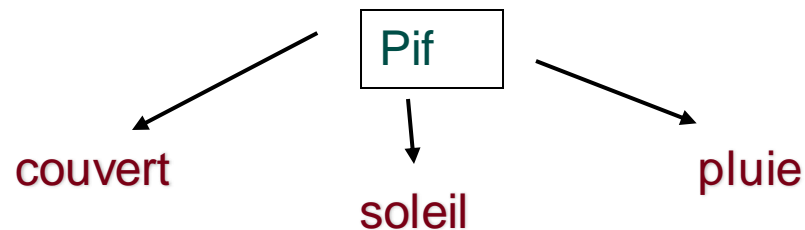
- Gini (temp)= $(3/5) \dots\dots + (1/5) \dots\dots + (1/5) \dots\dots = 0,2$

Exemple

Pick the **highest gain** attribute.

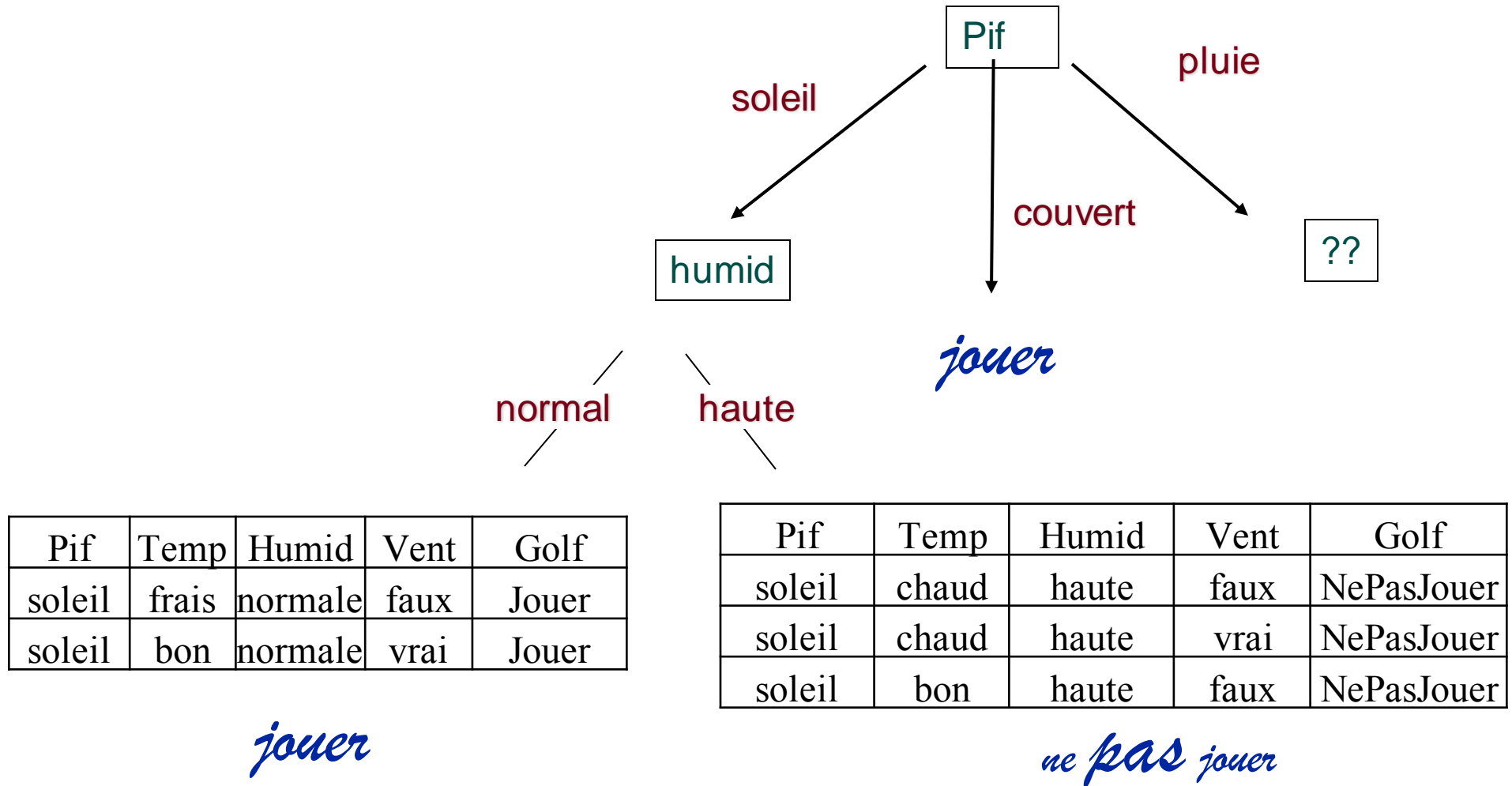
Attribut	Gain
temp	0,2
humid	0
vent	0,466

Root Node:



Exemple

Pick the **highest gain** attribute.



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

Temp	p	n	nbr
frais	2	1	3
bon	1	1	2

- Gini (temp= *frais*)= $1 - [(2/3)^2 + (1/3)^2] = \dots\dots$
- Gini (temp = *bon*)= $1 - [(1/2)^2 + (0/2)^2] = \dots\dots$
-

Calculate sum of gini index:

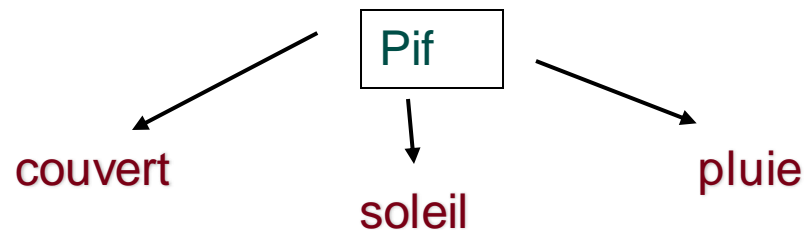
- Gini (temp)= $(3/5) \dots\dots + (2/5) \dots\dots = \dots\dots$

Exemple

Pick the **highest gain** attribute.

Attribut	Gain
temp	0.466
humid	0.466
vent	0

Root Node:



Exemple

Pick the **highest gain** attribute.

