Arbres de décision

Laurent.Bougrain@loria.fr

Bibliographie

Livres

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. : *Classification And Regression Trees*, Wadsworth and Books, Pacific Grove, CA, 1984

QUINLAN J.R. : *C4.5 : programs for machine learning*, Morgan Kaufmann, San Mateo, CA, 1993

Chapitres de livres

CORNUEJOLS A., MICLET L. : *Apprentissage artificiel : concept et algorithmes*, chapitre 11, Ed. Eyrolles, 2010

LEBART L., MORINEAU A., PIRON M. : *Statistique exploratoire multidimensionnelle*, section 3.5, ed. Dunod, 1995

Articles

QUINLAN J.R. : *Induction of decision Trees*, Machine Learning, 1 (1), p 81-106, 1986

Cours en ligne

MICLET L. : *http://people.irisa.fr/Laurent.Miclet/documents/transplivre/arbresdecision2.pdf*

Laurent.Bougrain@loria.fr

Aperçu

Introduction

Représentation par arbre de décision

Construction d'un arbre de décision

Gain d'information

Elagage

Laurent.Bougrain@loria.fr

Comment prédire le type de contraception utilisé par des femmes mariées à partir de données démographiques et socio-économiques les concernant ?

En utilisant un ensemble de situations connues qui serviront à établir un modèle permettant de prédire le type de contraception utilisé pour de nouvelles femmes.

Ce modèle peut être construit à l'aide d'un arbre de décision

Laurent.Bougrain@loria.fr

Un exemple de banque de données

Considérons la banque de données issue du National Indonesia Contraceptive Prevalence Survey, 1987.

Elle comporte : 1473 exemples, 10 variables (dont la classe)

3 classes ont été distinguées : no-use, long_term, short_term

Variable	Type	Valeurs
Wife's Age	Continu	
Wife's Education	Ordinal	1=low, 2, 3, 4=high
Husband's Education	Ordinal	1=low, 2, 3, 4=high
Number of children	Continu	
Wife's religion	Binaire	Non-Islam, Islam
Wife's now working ?	Binaire	yes, no
Husband's occupation	Nominal	1, 2, 3, 4
Standard-of-living index	Ordinal	1=low, 2, 3, 4=high
Media exposure	Binaire	Good, Not good
Contraceptive method used	Nominal	No-use, Long-term, Short-term

Laurent.Bougrain@loria.fr

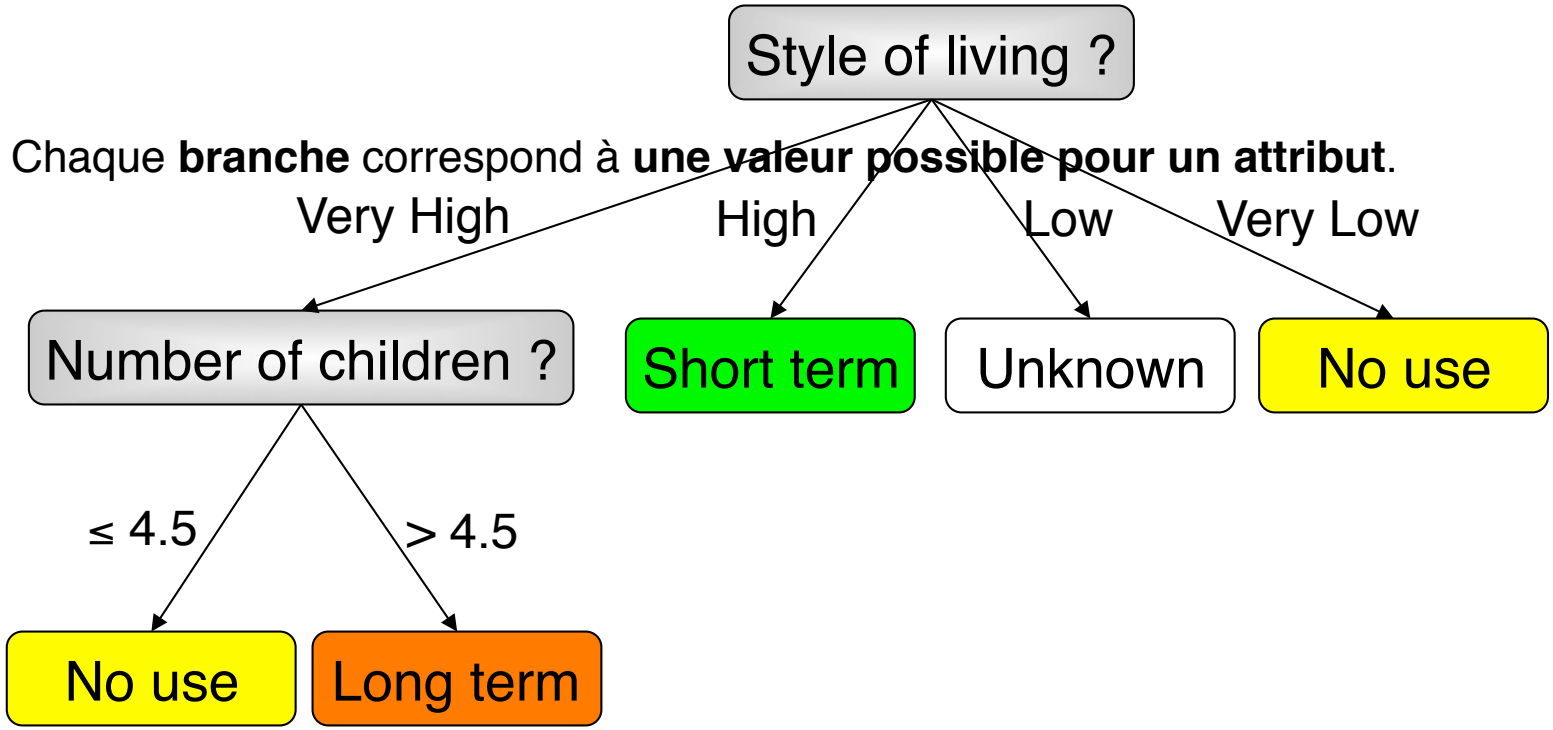
Un exemple de situations enregistrées

Age	Education	HusbandEducation	Children	Religion	Working	HusbandWork	Living	Media	Contraception
numerical	symbolic	symbolic	numerical	symbolic	symbolic	symbolic	symbolic	symbolic	symbolic
26	very_high	very_high	0	Islam	no	3	very_high	good	**No-use**
24	low	low	4	Islam	no	3	very_high	good	**No-use**
25	low	very_high	6	Islam	no	3	very_high	good	**Long_Term**
30	very_high	very_high	4	Islam	no	2	very_high	good	**No-use**
41	very_high	very_high	5	non-Islam	no	2	very_high	good	**Long_Term**
22	high	very_high	3	Islam	no	2	high	good	**Short_term**
35	low	high	6	Islam	no	2	high	good	**Short_term**
38	high	high	7	Islam	no	3	very_low	good	**No-use**
33	very_high	very_high	5	Islam	no	3	high	good	**Short_term**
22	high	high	2	Islam	no	2	very_high	good	**No-use**

Laurent.Bougrain@loria.fr

Un exemple d'arbre de décision obtenu

Chaque **nœud non-terminal** représente un **test sur un attribut**.



Chaque **branche** correspond à **une valeur possible pour un attribut**.

Chaque **nœud terminal** correspond à une **classe**.

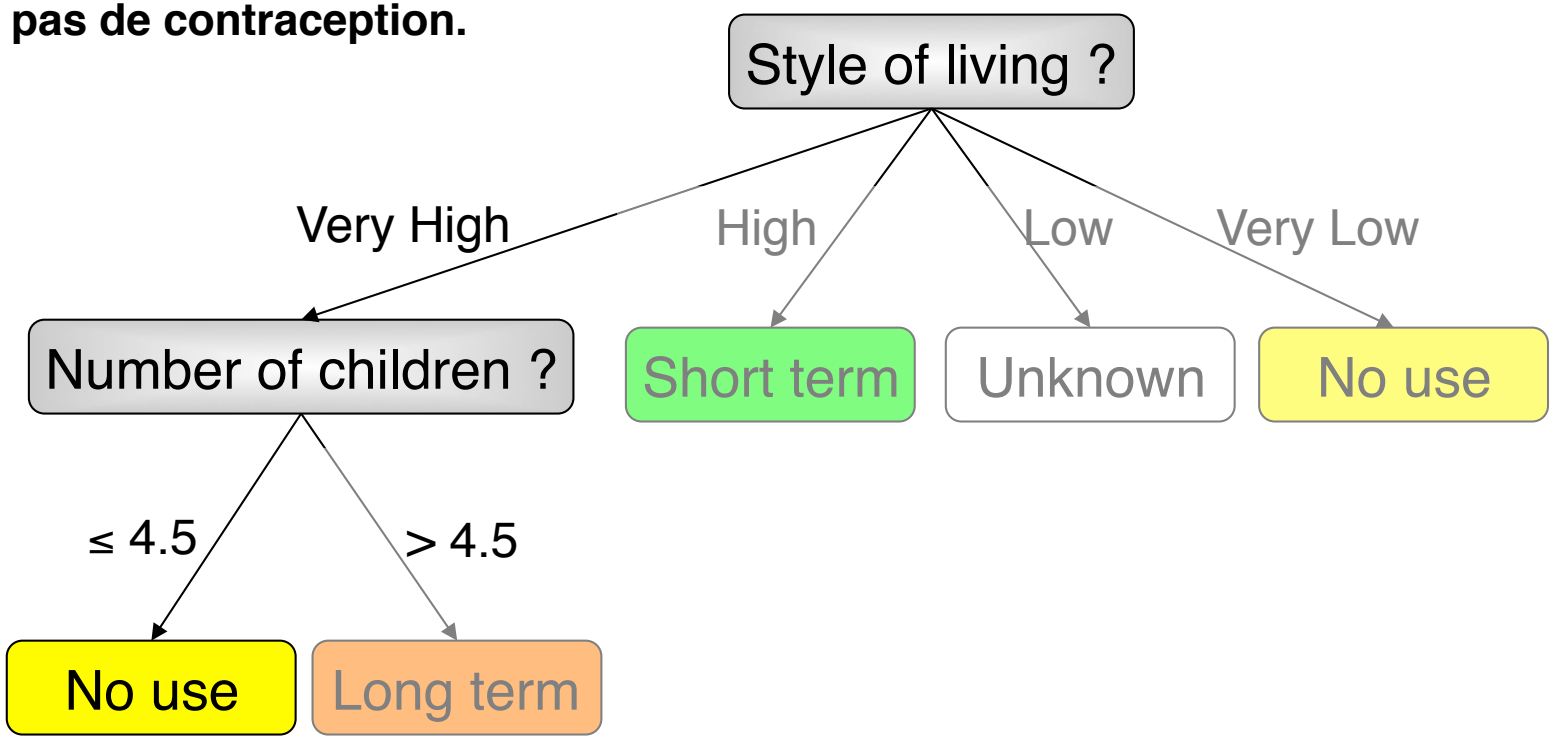
Laurent.Bougrain@loria.fr

Un exemple de prédiction

On relève les données démographiques et socio-économiques d'une nouvelle femme.

Age	Education	HusbandEducation	Children	Religion	Working	HusbandWork	Living	Media
20	low	high	3	Islam	no	3	very_high	good

L'arbre de décision classifie cette femme dans le groupe des femmes qui n'utilise pas de contraception.



Laurent.Bougrain@loria.fr

Construction (principe)

Construire tous les arbres de décision possibles afin d'identifier le meilleur n'est pas une solution envisageable.
On cherche donc à construire intelligemment l'arbre de décision par une **induction descendante** (top-down induction of decision tree):

Procédure : **ConstruireArbre(X)** *(X est l'ensemble des exemples)*
Si tous les exemples de X appartiennent à la même classe
Alors créer une feuille portant le nom de cette classe
Sinon
  Si un critère d'arrêt est vérifié (taux minimum de bonne classification, seuil minimum de mélange, nombre minimum d'exemples…) ou s'il n'y a plus de tests de discrimination disponibles
  Alors créer une feuille portant le nom de la classe dominante dans X
  Sinon
    Déterminer le test de séparation est le plus discriminant en fonction des valeurs de chaque attribut
    Créer un nœud étiqueté par l'attribut utilisé par le test
    Pour chaque séparation X_i de X **faire** ConstruireArbre(X_i)

Laurent.Bougrain@loria.fr

Exemple de sélection du test le plus discriminant

La liste des attributs et de leur type est la suivante :

Age	Education	HusbandEducation	Children	Religion	Working	HusbandWork	Living	Media
continu	ordinal	ordinal	continu	binaire	binaire	nominal	ordinal	binaire

Attribut numérique

Un attribut numérique fournira un test qui séparera les exemples en deux sous-ensembles (c'est-à-dire deux branches).

Un attribut continu a un nombre fini de valeurs puisque l'ensemble d'apprentissage est fini. On trie les valeurs de l'attribut. Chaque valeur (ou médiane entre deux valeurs successives) constitue un point de séparation en deux sous-ensembles. On cherche parmi ces valeurs la valeur, i.e. le seuil du test, qui sépare au mieux les individus de chaque classe.

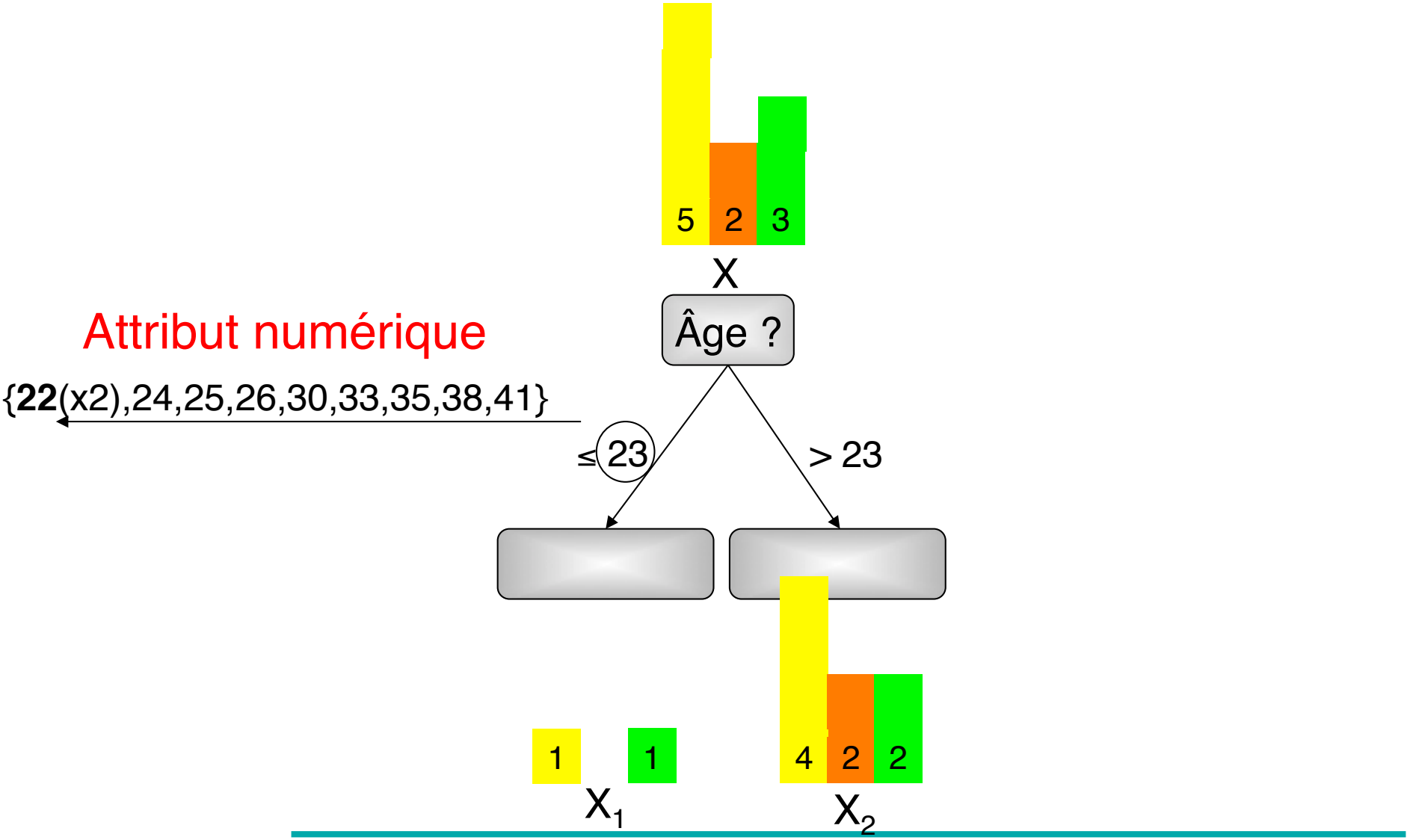
Attribut symbolique

Un attribut symbolique à k valeurs fournira un test spécial qui donnera k sous-ensembles (c'est-à-dire k branches).

Certains arbres de décision créent un test binaire à partir d'un attribut symbolique à k valeurs (exemple : couleur (rouge, bleu, vert) fournira un test "couleur=rouge?" oui ou non)

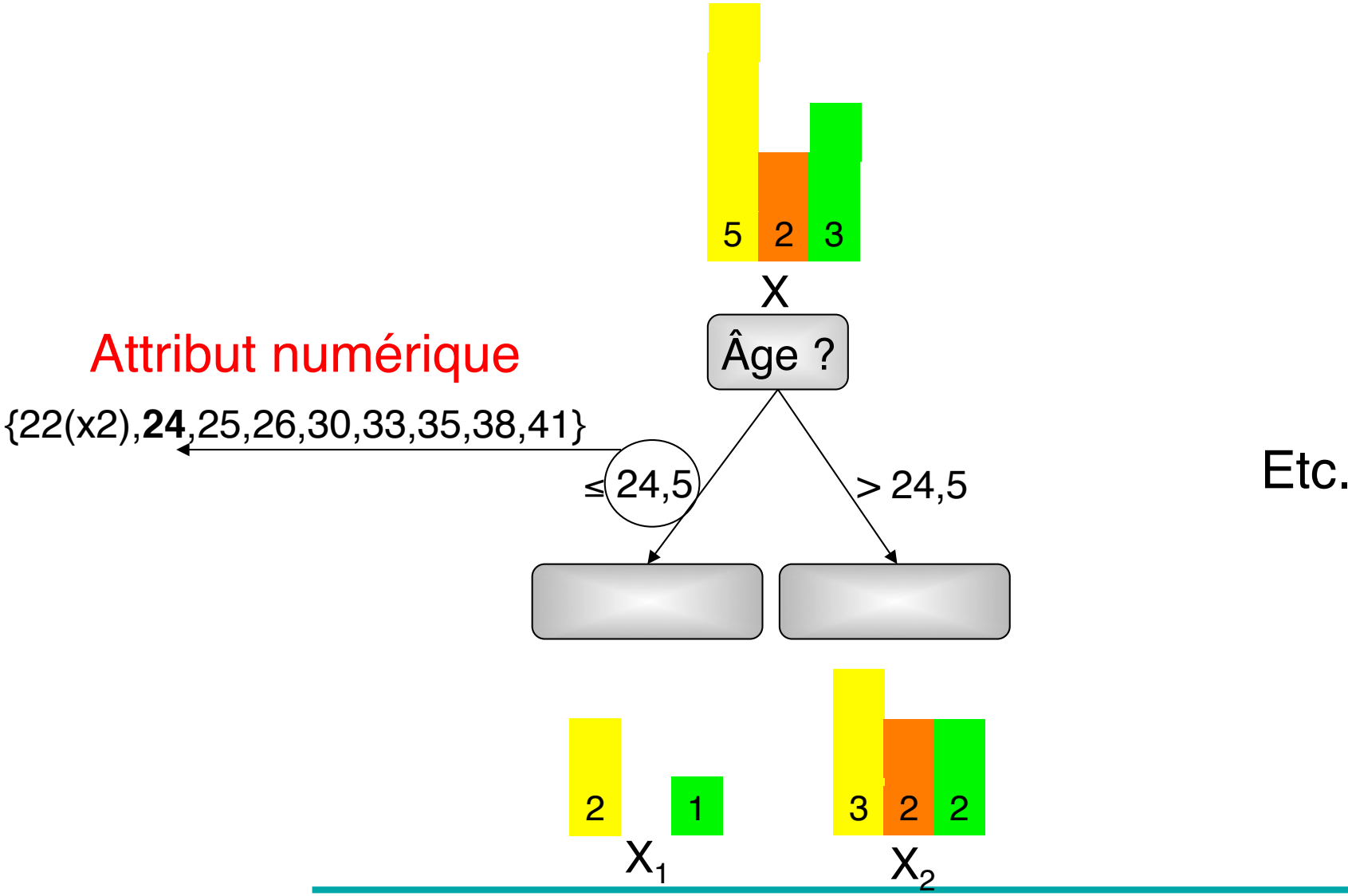
Laurent.Bougrain@loria.fr

Discrimination par un test sur un attribut numérique



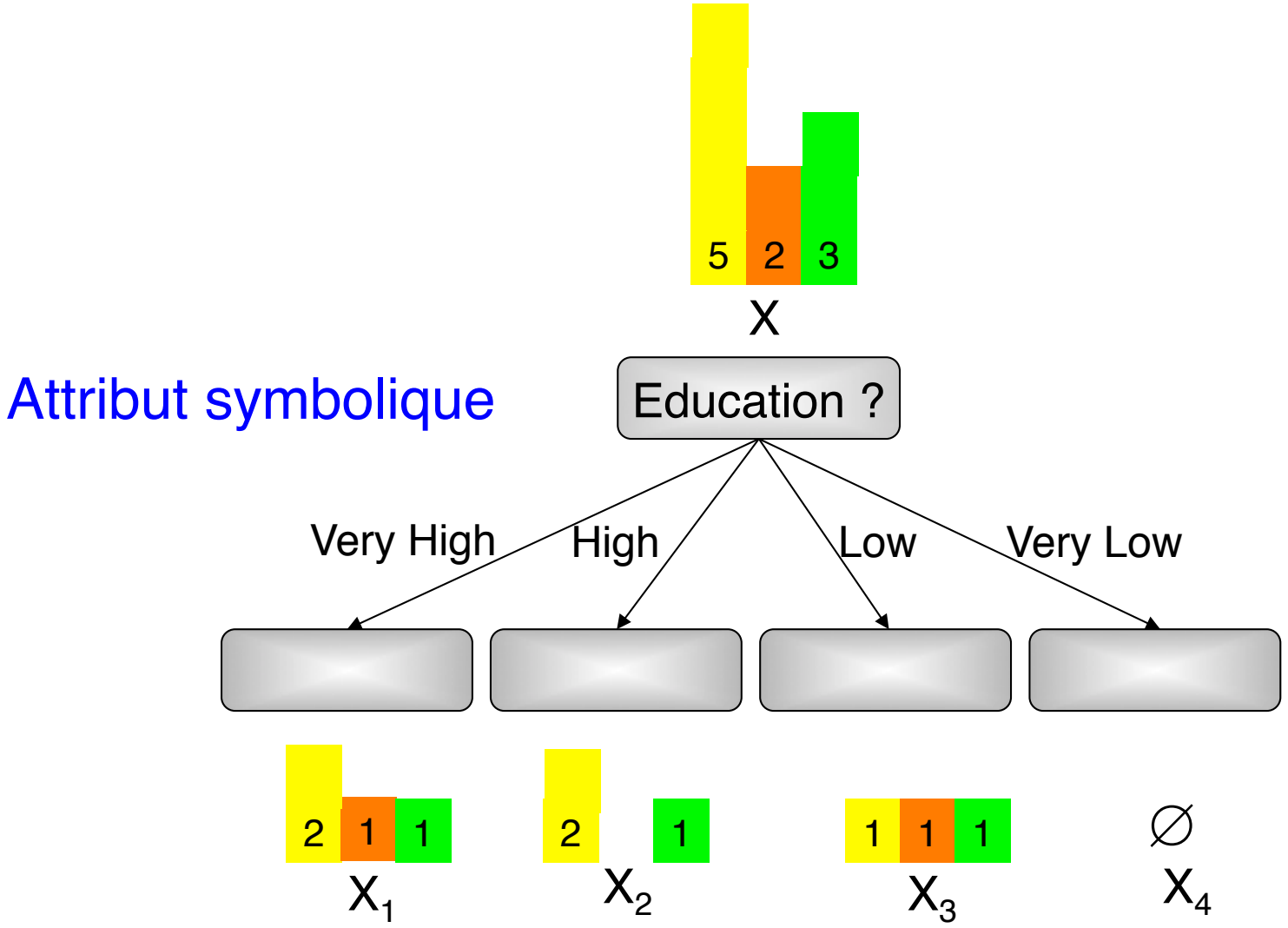
Laurent.Bougrain@loria.fr

Discrimination par un test sur un attribut numérique



Laurent.Bougrain@loria.fr

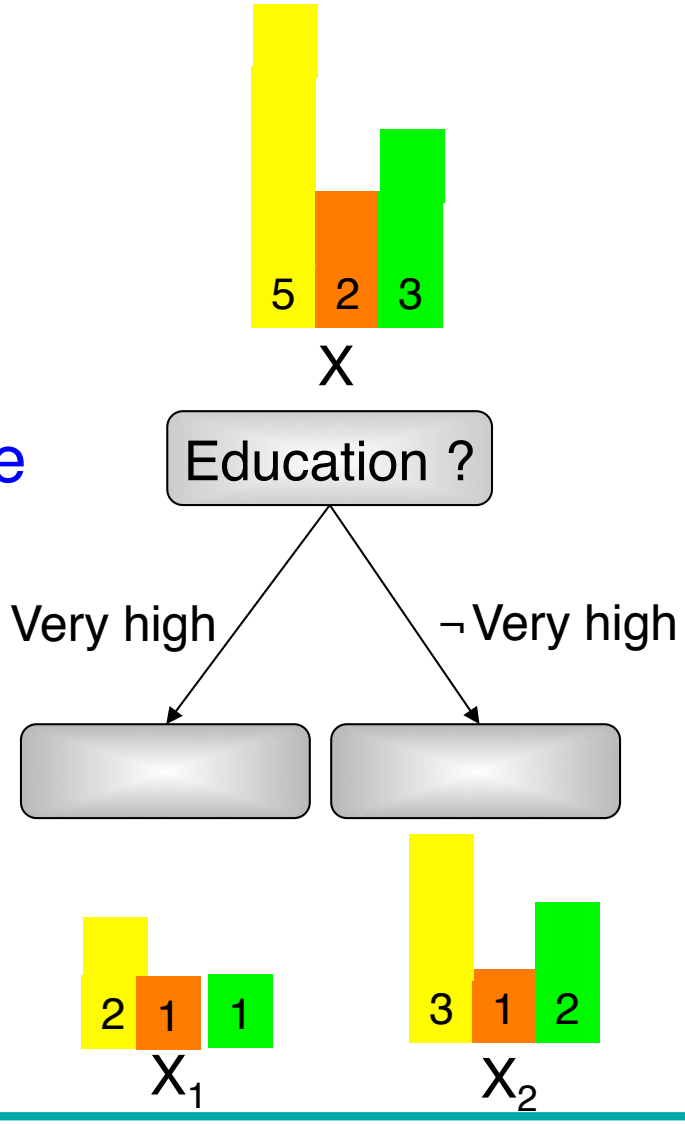
Discrimination par un test sur un attribut symbolique (cas n-aire)



Laurent.Bougrain@loria.fr

Discrimination par un test sur un attribut symbolique (cas binaire)

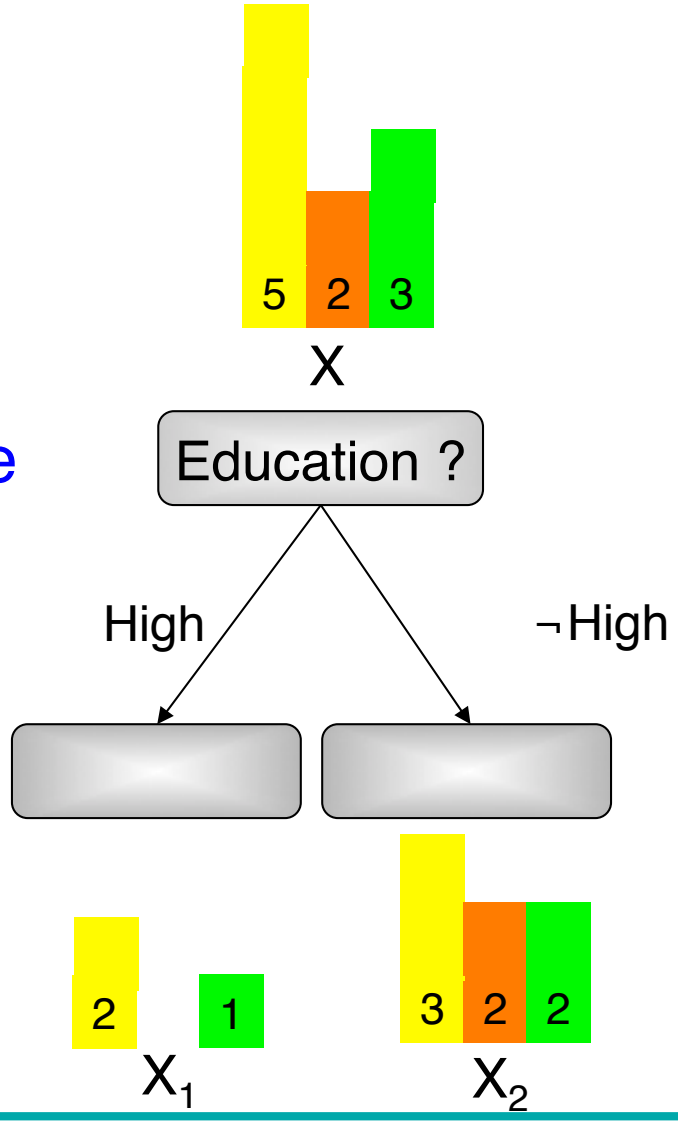
Attribut symbolique



Laurent.Bougrain@loria.fr

Discrimination par un test sur un attribut symbolique (cas binaire)

Attribut symbolique



Laurent.Bougrain@loria.fr

Interprétation probabiliste : exemple

X

$P(C = c_1)=5/10$

$P(C = c_2)=2/10$

$P(C = c_3)=3/10$

A ?

a_1 a_2

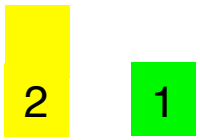
$P(A = a_1)=3/10$

$P(A = a_2)=7/10$

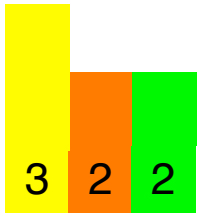
$P(A = a_1, C = c_1)=2/10$

$P(A = a_1, C = c_2)=0/10$

$P(A = a_1, C = c_3)=1/10$



X_1



X_2

$P(A = a_2, C = c_1)=3/10$

$P(A = a_2, C = c_2)=2/10$

$P(A = a_2, C = c_3)=2/10$

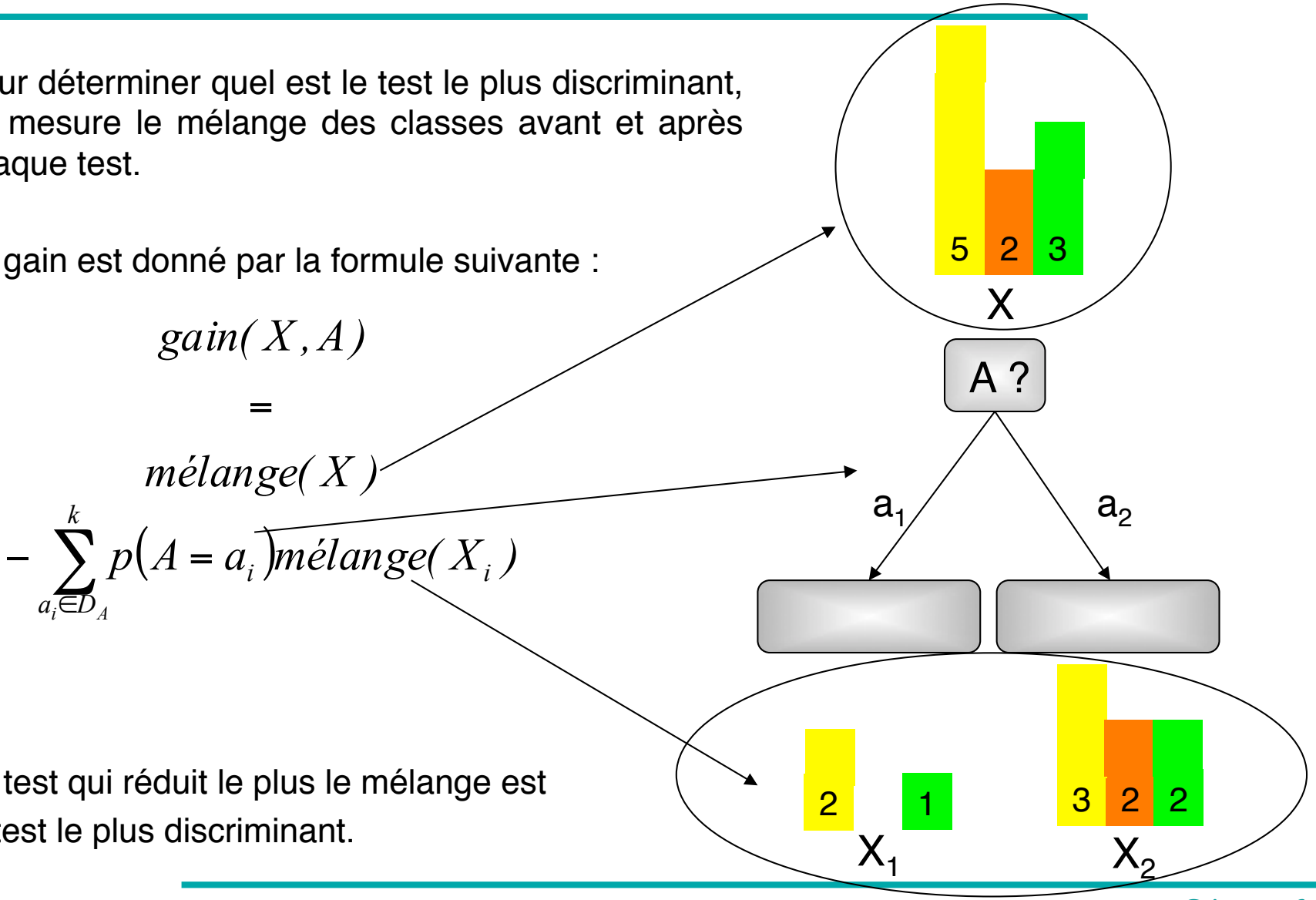
Laurent.Bougrain@loria.fr

Choix du test le plus discriminant : gain en information

Pour déterminer quel est le test le plus discriminant, on mesure le mélange des classes avant et après chaque test.

Le gain est donné par la formule suivante :

$$gain(X, A) = mélange(X) - \sum_{a_i \in D_A}^{k} p(A = a_i)\, mélange(X_i)$$



Le test qui réduit le plus le mélange est le test le plus discriminant.

Laurent.Bougrain@loria.fr

Choix du test le plus discriminant : mesure du mélange

Plusieurs indicateurs peuvent être utilisés pour mesurer le mélange :

1. L'entropie

Elle mesure l'incertitude à déterminer la classe C d'un individu du noeud.

Elle a pour formule :

mélange(X) = $H(C) = -\sum_{c_j \in D_c} P(C=c_j) \log_2 P(C=c_j)$

Remarque : le gain correspondra à l'information mutuelle (cf. théorie de l'information)

2. L'indice d'impureté de Gini

Il mesure la probabilité que 2 éléments choisis aléatoirement (avec remise) dans une population composée de k classes distinctes soient différents.

mélange(X) = $i(C) = \sum_{c_j \in D_c} \sum_{c_j \neq D_c} P(C=c_j) P(C=c_{j'}) = 1 - \sum_{c_j \in D_c} P(C=c_j)^2$

Dans les deux cas :

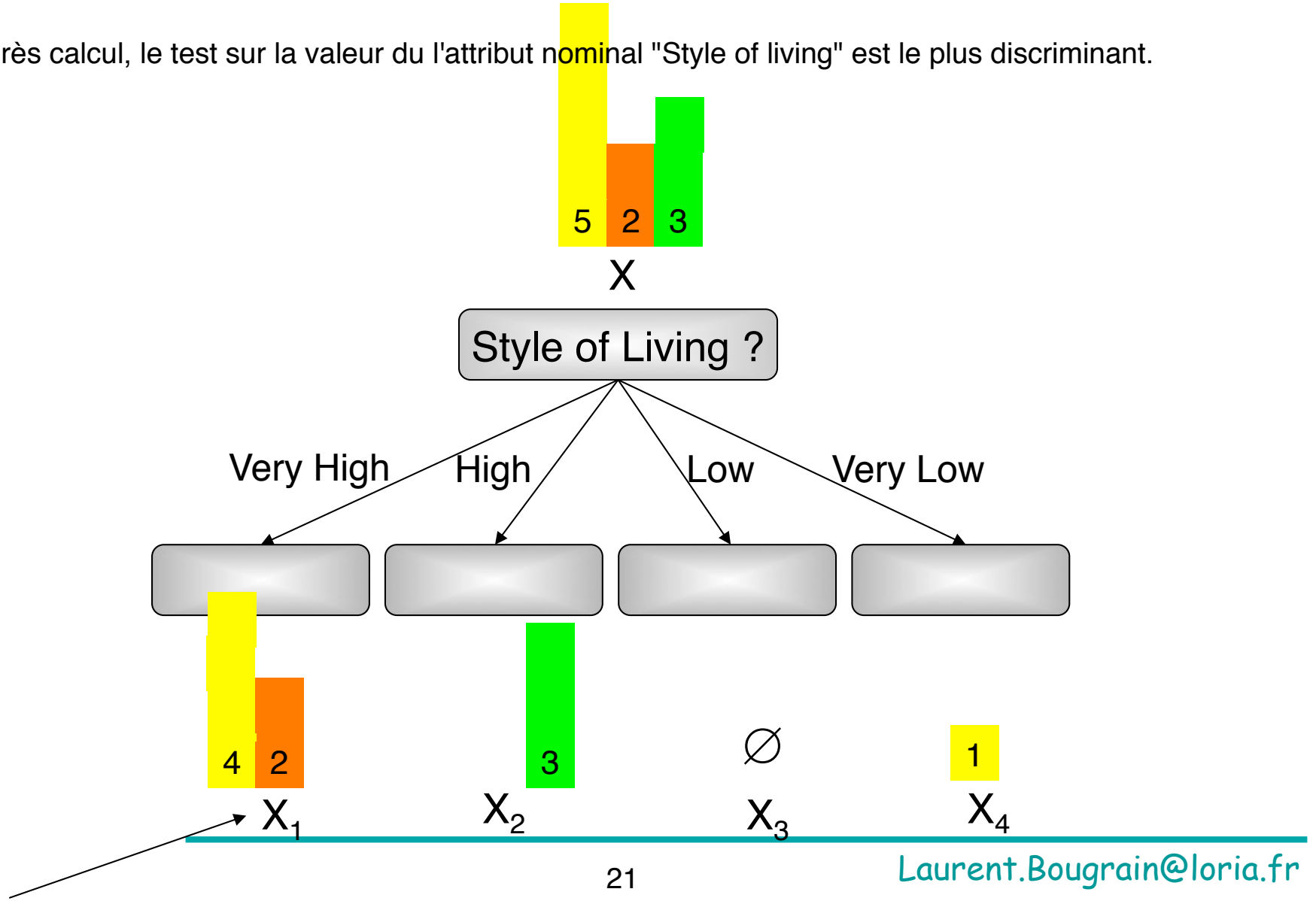
la valeur est minimale quand tous les individus sont de la même classe.

la valeur est maximale quand les exemples sont équirépartis dans les classes.

Laurent.Bougrain@loria.fr

Exemple de sélection du test le plus discriminant

Après calcul, le test sur la valeur du l'attribut nominal "Style of living" est le plus discriminant.



Laurent.Bougrain@loria.fr

Ici, il y a encore un mélange. On recherche pour cet ensemble le test le plus discriminant.

Choix du test le plus discriminant : mesure du mélange (suite)

Plusieurs indicateurs peuvent être utilisés pour mesurer le mélange :

3. Le gain normalisé (aussi appelé le ratio du gain)

La mesure de mélange favorise les attributes possédant de nombreuses valeurs.

Pour réduire ce biais, Quinlan propose de normaliser le gain en fonction de la répartition des valeurs de l'attribut

$$ratio(C,A) = \frac{gain(C,A)}{H(A)} = \frac{H(C) - H(C \mid A)}{H(A)}$$

Laurent.Bougrain@loria.fr

Elagage

Pourquoi élaguer ?

L'arbre sera généralement développé jusqu'à obtenir des feuilles pures quitte à générer un grand nombre de tests. Mais cet arbre est basé sur un échantillon particulier : le corpus d'apprentissage dont les éléments peuvent être bruités.

Il est souvent préférable d'avoir un modèle moins complexe qui s'interprétera plus facilement et qui généralisera un peu les exemples vus (pour éviter un apprentissage par cœur).

Ici, le nombre de nœuds constitue un critère de complexité simple et efficace.

Deux méthodes permettent d'effectuer cette opération

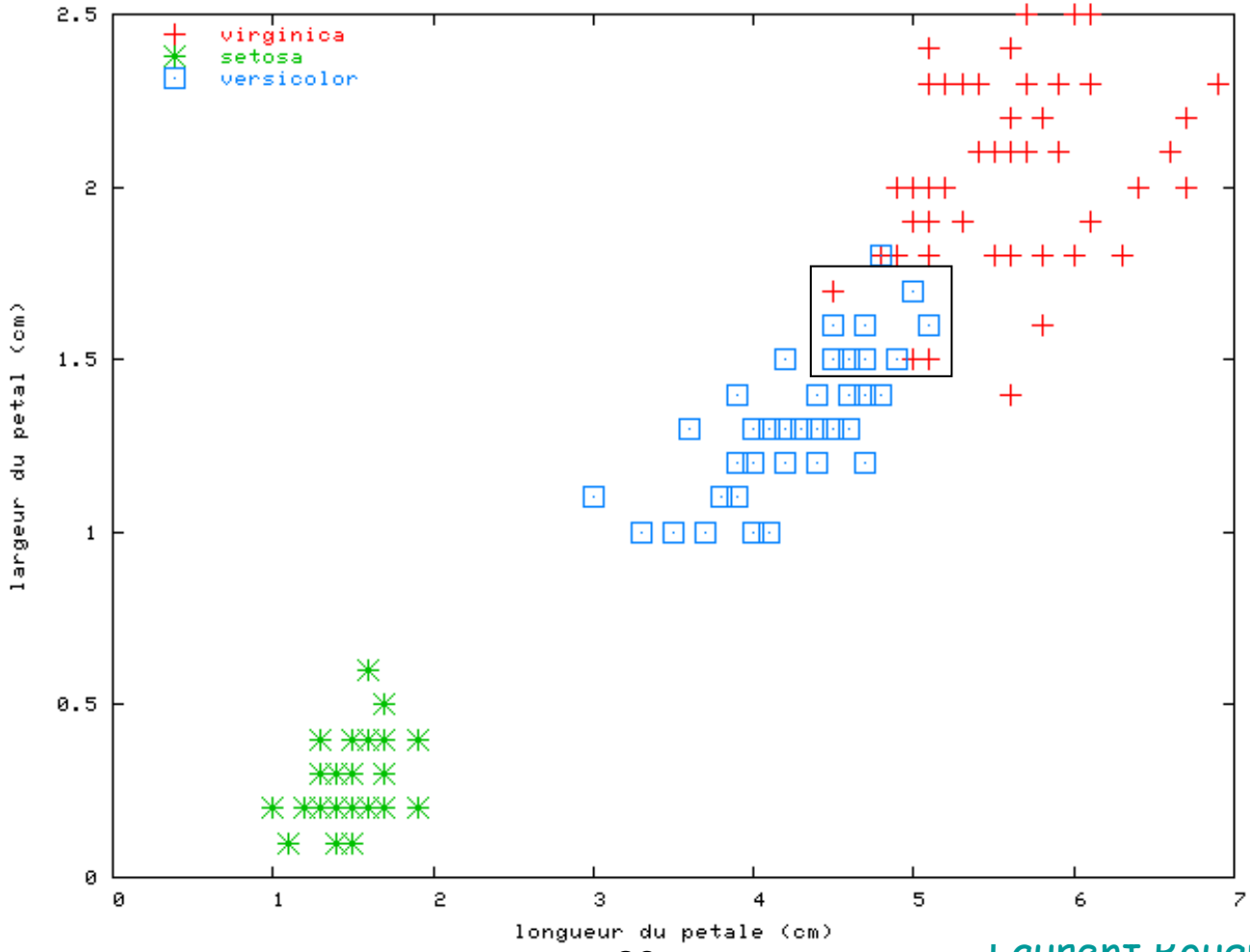
- L'arbre n'est pas développé complètement si certains critères d'arrêt sont vérifiés.

- L'arbre est développé entièrement puis il est simplifié en l'élaguant progressivement en remontant des feuilles vers la racine.

Laurent.Bougrain@loria.fr

Pourquoi élaguer ?

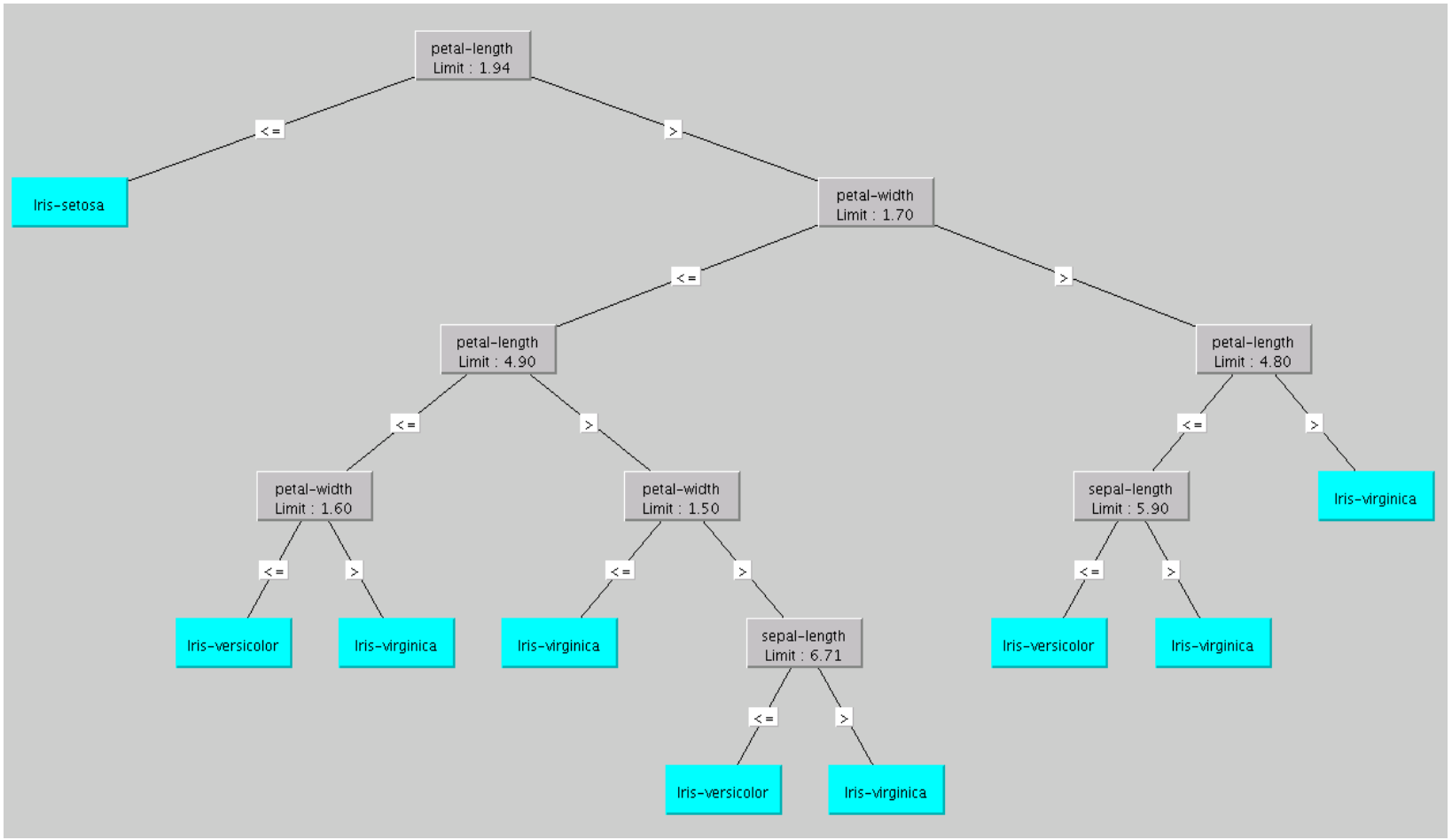
Visualisation des individus de la base de données des Iris sur les 2 attributs continus décrivant la largeur et la longueur du pétale.



Laurent.Bougrain@loria.fr

Pourquoi élaguer ?

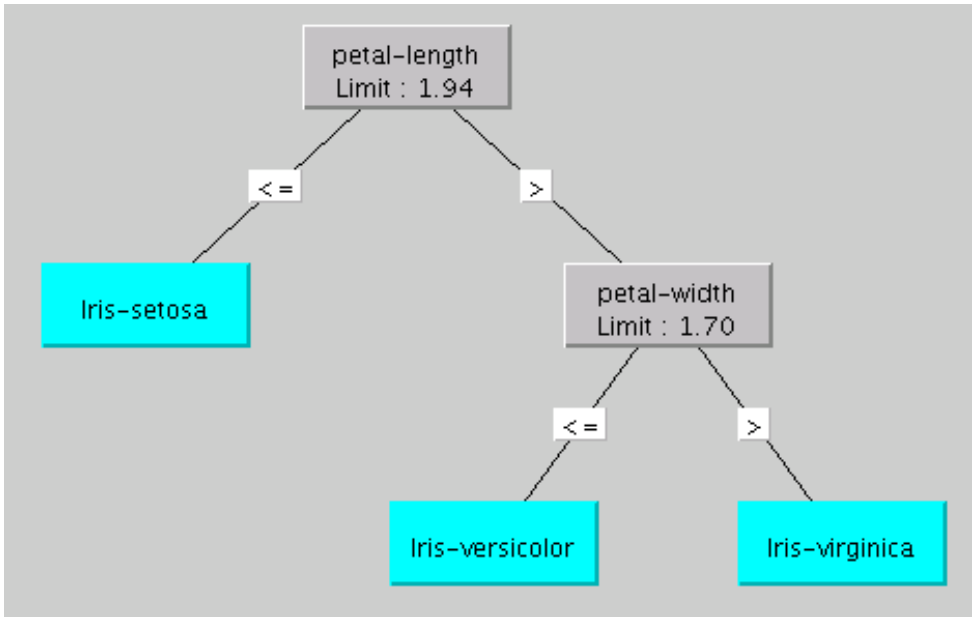
Visualisation d'un arbre de décision parfait



Laurent.Bougrain@loria.fr

Pourquoi élaguer ?

Visualisation d'un arbre de décision élagué



Laurent.Bougrain@loria.fr

Limiter l'expansion de l'arbre

On cesse de diviser les exemples lorsque un critère d'arrêt est vérifier.
Le nœud reçoit alors l'étiquette de la classe majoritaire.

Par exemple :

- **l'impureté du nœud est inférieure à un seuil.**
- **Le nombre d'individus est inférieure à un seuil.**
- **Le gain est inférieure à un seuil.**
- **…**

Laurent.Bougrain@loria.fr

Elagage

On élague progressivement en remontant des feuilles vers la racine.

Soit $\{T_{max}, T_1, \ldots, T_n\}$ cette séquence d'arbres où T_{max} est l'arbre complet et T_n l'arbre constitué uniquement de la racine.

Pour passer de T_k à $T_{k+1,}$ Il faut transformer un nœud de T_k en feuille.

Le bénéfice de cet élagage est mesuré par un critère qui indique un compromis entre la complexité de l'arbre et l'erreur obtenu sur un nouvel ensemble de données, le corpus de test. Ce critère à minimiser permet de décider quand arrêter l'élagage.

Si le coût de T_{k+1} dépasse le coût de T_k, on s'arrête.

Parmi les différents critères possibles prenons $c(T_k, s) = \dfrac{MC_{élag}(T_k, s) - MC(T_k)}{n(T_k)(nt(T_k, s) - 1)}$

où

$MC_{élag}(T_k, s)$ est le nombre d'exemples de l'ensemble d'apprentissage mal classés par le nœud s de T_k si on transforme s en feuille.

$MC(T_k)$ est le nombre d'exemples de l'ensemble d'apprentissage mal classés sous le nœud s de T_k.

$n(T_k)$ est le nombre de feuilles de T_k

$nt(T_k, s)$ est le nombre de feuilles du sous-arbre situé sous le nœud s de T_k.

Laurent.Bougrain@loria.fr

Elagage : pseudo-algorithme

Procédure : **élaguer(T_{max})**

$k \leftarrow 0$

$T_k \leftarrow T_{max}$

tant que T_k possède plus d'un noeud

pour chaque nœud s de T_k **faire**

calculer le critère $c(T_k, s)$ sur l'ensemble d'apprentissage

fin pour

choisir le nœud s_m pour lequel le critère est minimium

T_{k+1} se déduit de T_k en y remplaçant s_m par une feuille

$k \leftarrow k+1$

fin tant que

choisir dans l'ensemble des arbres $\{T_{max}, T_1, \ldots, T_k, \ldots, T_n\}$ celui qui a la plus petite erreur de classification sur l'ensemble de validation

Laurent.Bougrain@loria.fr

Arbre de régression

Certaines méthodes ont été étendues pour créer un arbre de régression c'est-à-dire pour prédire la valeur d'une variable numérique.

L'indication de mélange est donnée par la variance.

$$var\left(C \middle| A = a_i\right) = \frac{1}{n_i} \sum_{k=1}^{n} \left(y_k - \bar{y}_i\right)^2$$

Erreur est évaluée en sommant l'erreur de chaque feuille

$$E = \sum_{a_i \in D_A} E_{a_i} \qquad \textbf{avec} \quad E_{a_i} = p\left(A = a_i\right) var\left(C \middle| A = a_i\right)$$

Laurent.Bougrain@loria.fr

Arbres de décision standard

Nom	Propriétés	Test de discrimination
THAID	Discrimination	Erreur de discrimination
CHAID Chi-square Automatic Interaction Detection	Discrimination et régression Variables nominales et continues	Chi2 (discrimination) Fisher (régression)
ID3 [Quinlan, 79] Induction of Decision Tree	Discrimination Variables nominales	Gain d'information Ou Entropie
C4.5 [Quinlan, 86]	Discrimination et régression Variables nominales et continues Valeurs manquantes	Ratio du gain d'information
CART [Brieman, 84] Classification And Regression Tree	Discrimination et régression Variables nominales et continues Arbre binaire Elagage	Impureté de Gini (discrimination) Variance (régression)

Laurent.Bougrain@loria.fr

En résumé

Les arbres de décisions sont historiquement et essentiellement des systèmes de classification (au sens de discrimination) bien que certains d'entre eux aient été étendus pour la régression.

Ils permettent de prédire la valeur d'une variable à expliquer (nominale dans le cas de la classification, continue dans le cas de la régression) par une suite de tests sur les valeurs des variables explicatives d'un objet.

Le modèle est construit à partir d'un ensemble d'exemples (d'apprentissage).

Il se représente sous la forme d'un arbre.

Il peut être simplifié par élagage

Les intérêts des arbres de décision sont :

- la possibilité de les utiliser pour de la discrimination et de la régression
- leurs capacités à traiter des attributs de nature variée (binaire, nominal, continu)
- la facilité à comprendre les raisons de leur décision (en effet, un chemin dans un arbre de décision représente une formule logique propositionnelle).
- Ils effectuent automatiquement une sélection des variables pertinentes
- Ils peuvent gérer les données manquantes lors la construction et lors de l'utilisation du modèle

Laurent.Bougrain@loria.fr