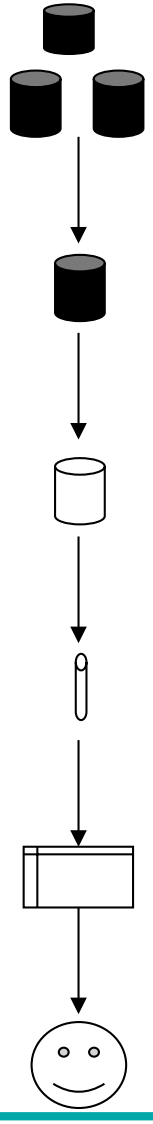
Processus d'extraction de connaissances à partir de données

Laurent.Bougrain@loria.fr
Université de Lorraine
Centre de recherche INRIA Nancy – Grand Est

FAYYAD U., PIATETSKY-SHAPIRO G. et SMYTH P. : *From data mining to knowledge discovery in databases*, Ai Magazine, vol.17, pp37-54, 1996

Processus de découverte de connaissances
Extraction de connaissances à partir de base de données (ECBD)
(Knowledge Discovery in Database (KDD), Piatetsky-Shapiro, IJCAI'89)



Totalité des Données disponibles
Bases de données relationnelles et/ou entrepôts de données

1. Sélection

Données sélectionnées

2. Pré-traitement

Données pré-traitées

3. Transformation

Données Transformées

4. Fouille de données

Résultats

5. Interprétation

connaissances

Laurent.Bougrain@loria.fr

Préparation des données

1. Sélectionner un sous-ensemble des données disponibles

Quels **attributs** sont intéressants pour notre problématique ?
Combien d'attributs peut-on retenir ?
Quels **individus** sélectionner ?
Combien d'individus prend-on pour constituer l'échantillon ?

2. Pré-traitements

Traitement des valeurs manquantes
Traitement des valeurs aberrantes
Recensement des contradictions

3. Transformations

Compression de l'information (réduction de la dimensionnalité)
Recodage de l'information
Regroupement par ensembles de valeurs

Sélection

A priori de variables

- ID
- Nom
- Adresse
- …

D'individus

- Tranche d'âge
- Genre
- …

Plus le nombre d'individus sera élevé, moins les modèles ferons de surapprentissage.

Certaines méthodes de modélisation possèdent une complexité proportionnelle au nombre de variables, d'autres au nombre d'individus.

Sélection : disproportion de classes

De nombreuses discriminations présentent une disproportions de classes

- Prédiction d'usure: 97% ok, 3% usé (dans un mois)
- Diagnostic médical: 90% sains, 10% malades
- eCommerce: 99% ne pas acheter, 1% acheter
- Sécurité: >99.99% des connexions ne sont pas des attaques
- Géologie : gisements vs sites stériles

La présence parmi les individus d'une classe dominante risque d'aboutir à un modèle avec de bonnes performances mais inutile.

Il peut être nécessaire de rééquilibrer les classes.

- En supprimant une partie des individus de la classe majoritaire
- Ou en dupliquant les individus des classes minoritaires
- Ou en biaisant le tirage aléatoire des individus

Pré-traitements: Valeurs Manquantes

1. Supprimer les individus incomplets
2. Supprimer les variables fortement non renseignées

Les résultats des méthodes suivantes sont assez décevants

1. Prendre la valeur moyenne pour un attribut quantitatif

2. - Prendre la valeur la plus fréquente pour un attribut qualitatif
   - Prendre la valeur la plus fréquente pour un attribut qualitatif sachant la classe de l'individu par la formule de Bayes

$$P\left(A = a_i \middle| C = c_j\right) = \frac{P\left(A = a_i, C = c_j\right)}{P\left(C = c_j\right)}$$

   - Prendre la valeur "Inconnue" pour un attribut qualitatif

Méthode laborieuse

1. Prédire la valeur manquante sachant la classe et la valeur des autres attributs de l'individu par un arbre de décision

Transformations : Normalisation

Lorsque les domaines des variables ne sont pas homogènes, les variables ayant les plus grandes variables peuvent influencer plus fortement la réponse du modèle.

Afin d'équilibrer ces influences, il faut homogénéiser les domaines des variables en effectuant une transformation des variables quantitatives :

Exemple : poids en kg ([0,250]) VS taille en m ([0,2.5])

- interpolation linéaire

$x' = (x - min)/(max - min)$

- centrer-réduire

$x' = (x - moyenne)/ecart\text{-}type$

Laurent.Bougrain@loria.fr

Transformations : Recodage (numérisation, discrétisation)

Certaines techniques de fouille de données, voire certains algorithmes d'une même technique, ne peuvent pas utiliser directement les variables quantitatives (règles d'association) ou qualitatives (réseaux de neurones, régressions statistiques, machine à vecteurs supports)

Qualitatif -> Quantitatif

- variable qualitative à **2 modalités : 1 variable quantitative**

Ex : Genre = {homme, femme} -> Homme={1,0} ou {1,-1}

- variable qualitative à **n modalités : n variable quantitative**

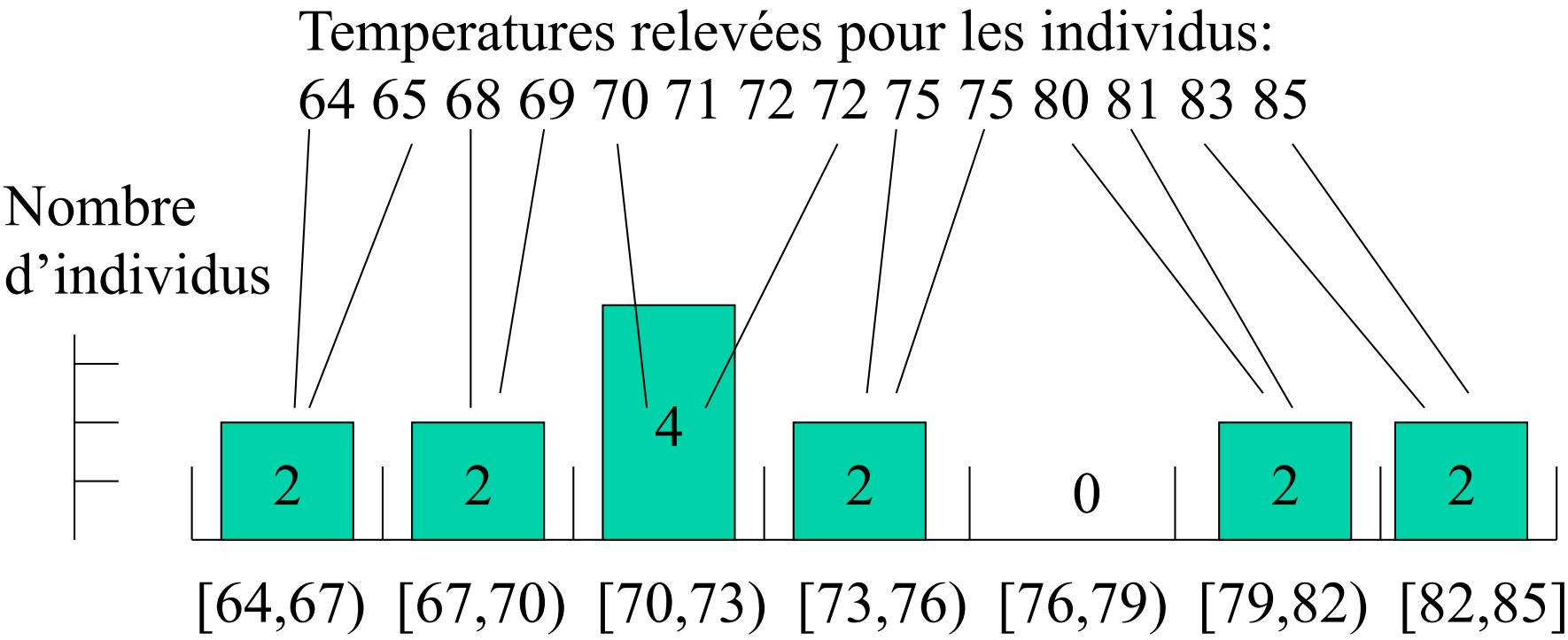
Ex : Statut = {marié, célibataire, veuf} -> Marié={1,0} ou {1,-1}
Célibataire={1,0} ou {1,-1}
Veuf={1,0} ou {1,-1}

Statut=marié -> Marié = 1, Célibataire = 0, Veuf = 0

Transformations : Discrétisation

Quantitatif -> Qualitatif : Regroupement par ensembles de valeurs

- intervalle régulier

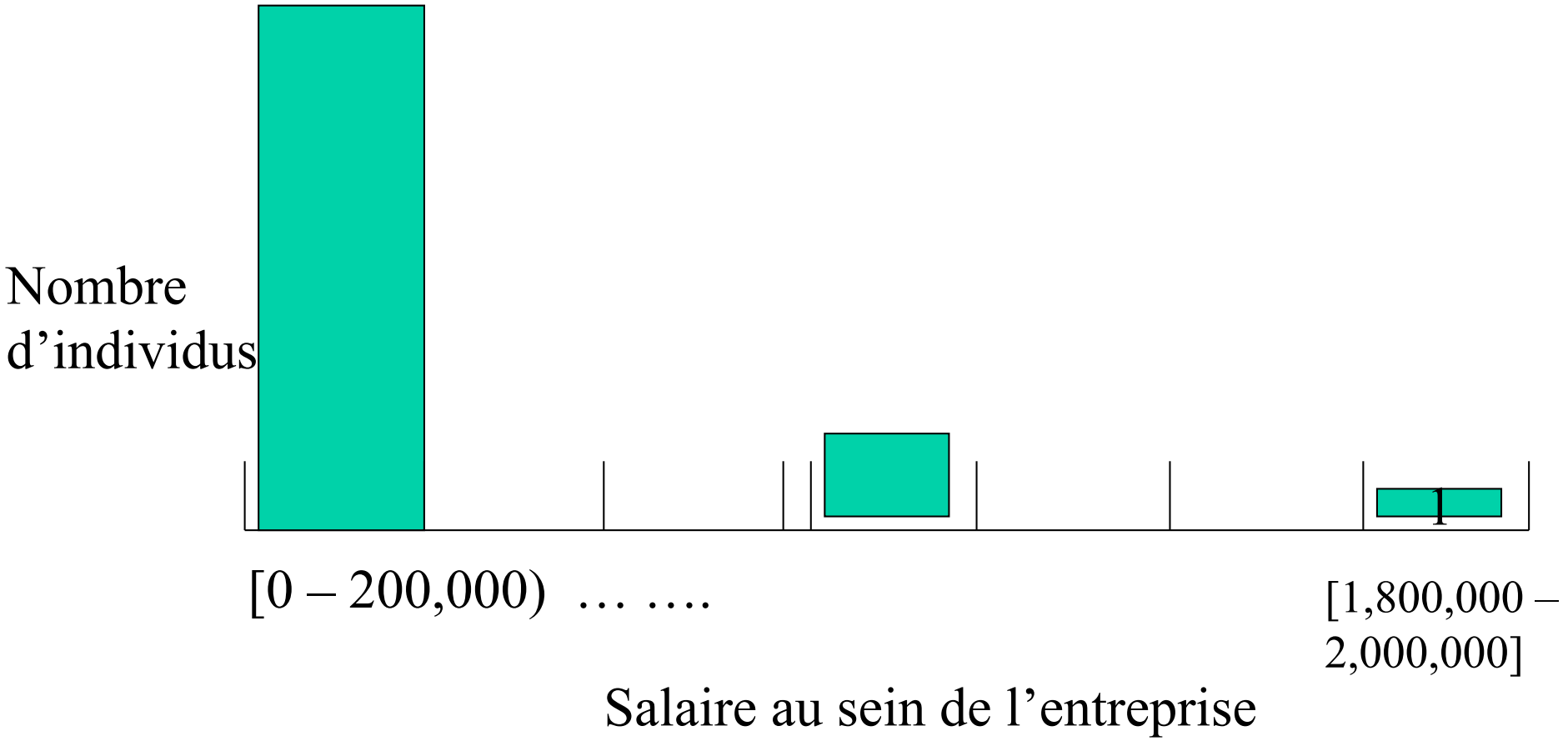


Laurent.Bougrain@loria.fr

Transformations : Discrétisation

Regroupement par ensembles de valeurs :

- intervalle régulier : Attention à ne pas perdre trop d'informations



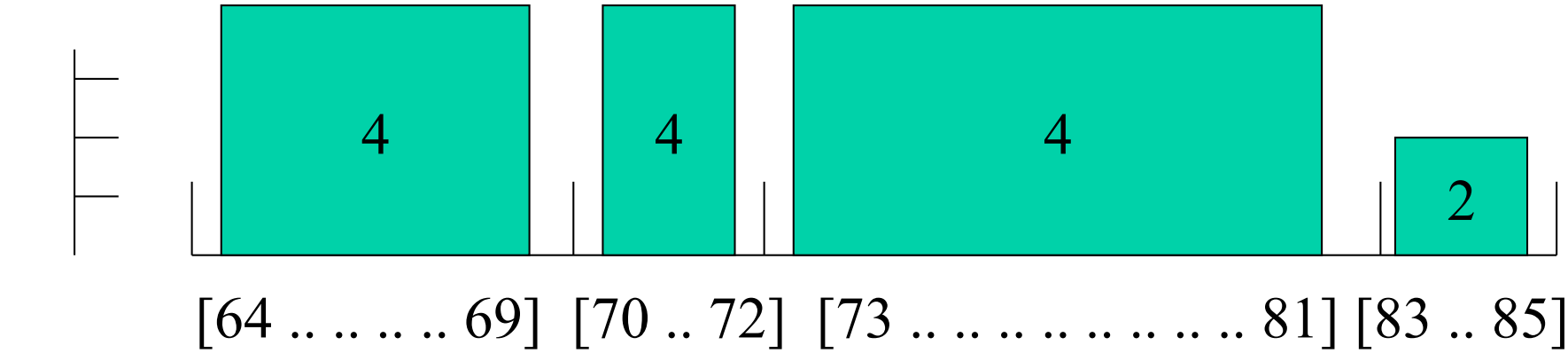
Laurent.Bougrain@loria.fr

Transformations : Discrétisation

Regroupement par ensembles de valeurs :

- nombre d'individus régulier

Températures relevées pour les individus :
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Laurent.Bougrain@loria.fr

4. Fouille de données

Application des techniques statistiques ou de l'apprentissage automatique pour découvrir de nouvelles connaissances potentiellement utile à partir des données (régression, discrimination, catégorisation, association, caractérisation).

5. Interprétation et validation

Interprétation des graphismes et des tableaux issus de la fouille de données.

Estimations du risque réel à partir du risque empirique.

Calcul de l'intervalle de confiance des résultats.

6. Exploitation

Reste à exploiter cette connaissance pour proposer des solutions marketing, de protection, de gestion, etc.

Analyse des résultats : motivations

1. **Le modèle est-il significatif ?**

Au regard de sa complexité, du nombre d'exemples et des performances observées

2. **Le modèle est-il performant en déploiement ?**

Sur de nouveaux cas

3. **Le modèle est-il meilleur qu'un autre ?**

Analyse des résultats : les mesures d'erreurs (régression)

- **Erreur quadratique moyenne :** $E = \frac{1}{N} \sum_{n=1}^{N} \left(t^n - y^n \right)^2$
- **Erreur quadratique relative :** $E = \frac{\sum_{n=1}^{N} \left(t^n - y^n \right)^2}{\sum_{n=1}^{N} \left(t^n - \bar{t} \right)^2}$
- **Erreur absolue moyenne :** $E = \frac{1}{N} \sum_{n=1}^{N} \left| t^n - y^n \right|$
- **Erreur absolue relative :** $E = \frac{\sum_{n=1}^{N} \left| t^n - y^n \right|}{\sum_{n=1}^{N} \left| t^n - \bar{t} \right|}$

…

Analyse des résultats : matrice de confusion (classement)

Mon modèle classe bien 90% des exemples. Est-ce bien ?

La prise en compte du taux de bonne classification n'est pas suffisante.

1. Il faut comparer ce taux avec la performance que l'on obtiendrait si on choisissait la classe au hasard (1 / nombre de classes).
2. Il faut regarder la proportion de chaque classe dans l'ensemble d'apprentissage.
3. Il faut regarder la matrice de confusion

Classe prédite

	Malade	**Sain**
Malade	**9**	**1**
Sain	**2**	**88**

Classe réelle

Classe prédite

	Oui	Non
Oui	**Vrais positifs True Positive (TP)**	**Faux négatifs False Negative (FN)**
Non	**Faux positifs False Positive (FP)**	**Vrais négatifs True Negative (TN)**

Classe réelle

Un "faux positif" n'a pas la même valeur qu'un "faux négatif" en fonction du problème.

Laurent.Bougrain@loria.fr

Analyse des résultats : mesures de performances (classement)

Taux de bonnes classifications	(TP+TN)/(TP+FP+TN+FN)
Taux d'erreurs	$(C_{21}*FP+C_{12}*FN)/(TP+FP+TN+FN)$ Où C_{ij} sont les éléments de la matrice de coût de mauvaises affectations (par défaut la matrice est symétrique et unitaire)
Rappel (sensibilité)	TP/(TP+FN)
Précision (spécificité)	TP/(TP+FP)
F-mesure	(2×rappel×précision)/(rappel+précision)
Taux de faux positifs (1-spécificité)	FP/(FP+TN)

Analyse des résultats : Courbes de ROC (Receiver Operating Characteristic)

- **Indépendante des matrices de coûts de mauvaise affectation**
  Il permet de savoir si M1 sera toujours meilleur que M2 quelle que soit la matrice de coût

- **Opérationnel même dans le cas des distributions très déséquilibrées**
  Sans les effets pervers de la matrice de confusion liés à la nécessité de réaliser une affectation

- **Résultats valables même si l'échantillon test n'est pas représentatif**
  Tirage prospectif ou tirage rétrospectif : les indications fournies restent les mêmes

- **Un outil graphique qui permet de visualiser les performances**
  Un seul coup d'œil doit permettre de voir le(s) modèle(s) susceptible(s) de nous intéresser

- Un **indicateur synthétique** associé : la surface sous la courbe ROC

Aisément interprétable

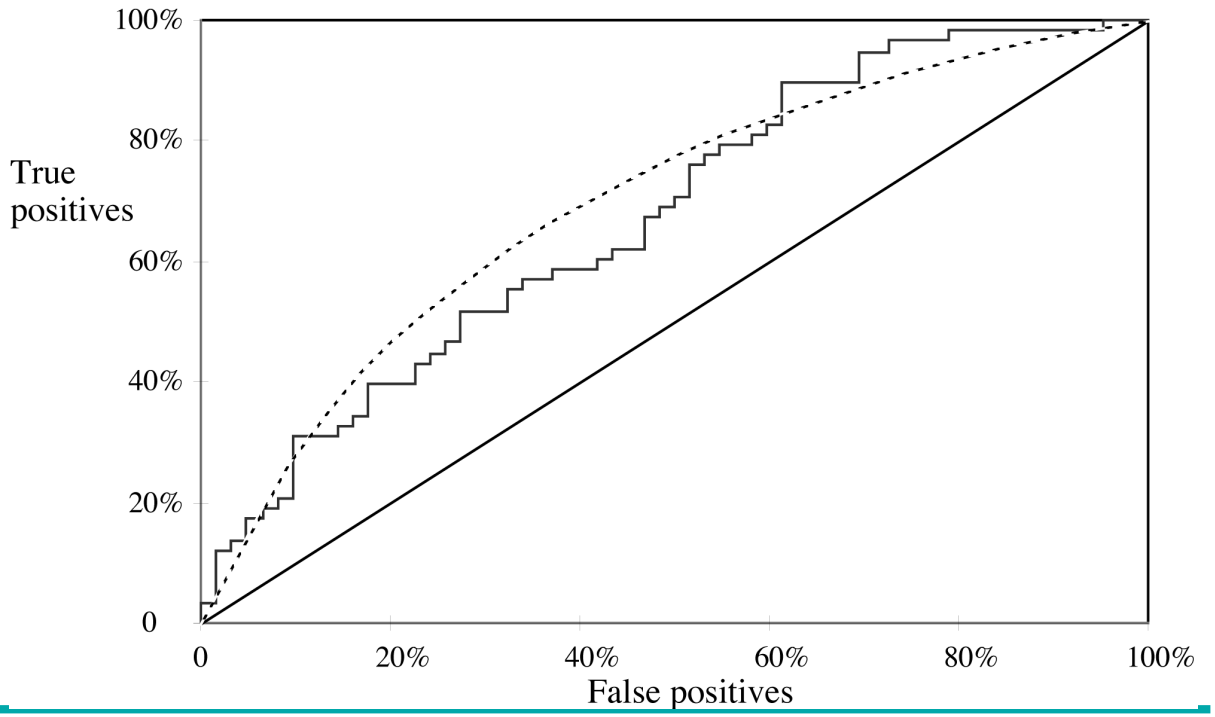
Laurent.Bougrain@loria.fr

[R. Rakotomalala]

Analyse des résultats : Courbes de ROC

- Utilise un mesure de type probabilité d'appartenance à la classe (ex probabilité à posteriori, distance à la frontière de décision)
- Les exembles sont classés par ordre croissant d'appartenance à la classe
- On calcule les FPR et TPR pour uniquement le premier exemple, puis les deux premiers, etc.

- La surface sous la courbe vaut 0.5 pour le hasard
1 pour un modèle parfait



Validité des résultats : apprentissage et généralisation

L'erreur obtenue sur l'ensenble d'apprentissage n'est pas un bon indicateur.

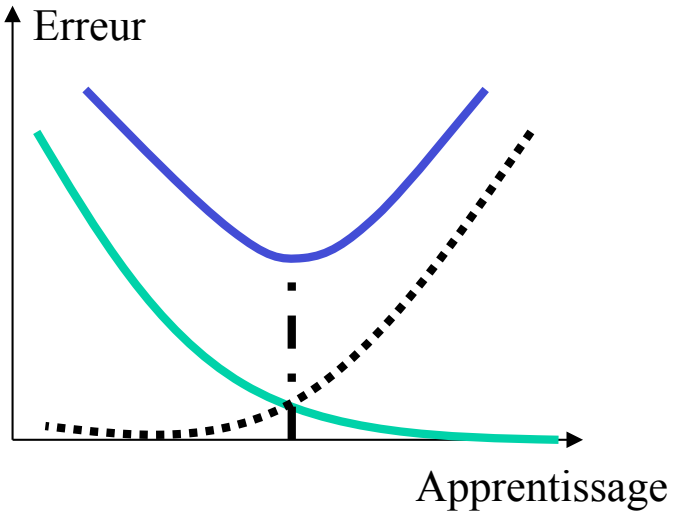
Problème du surapprentissage :

Décomposition biais-Variance

Apprentissage sur un ensemble de données

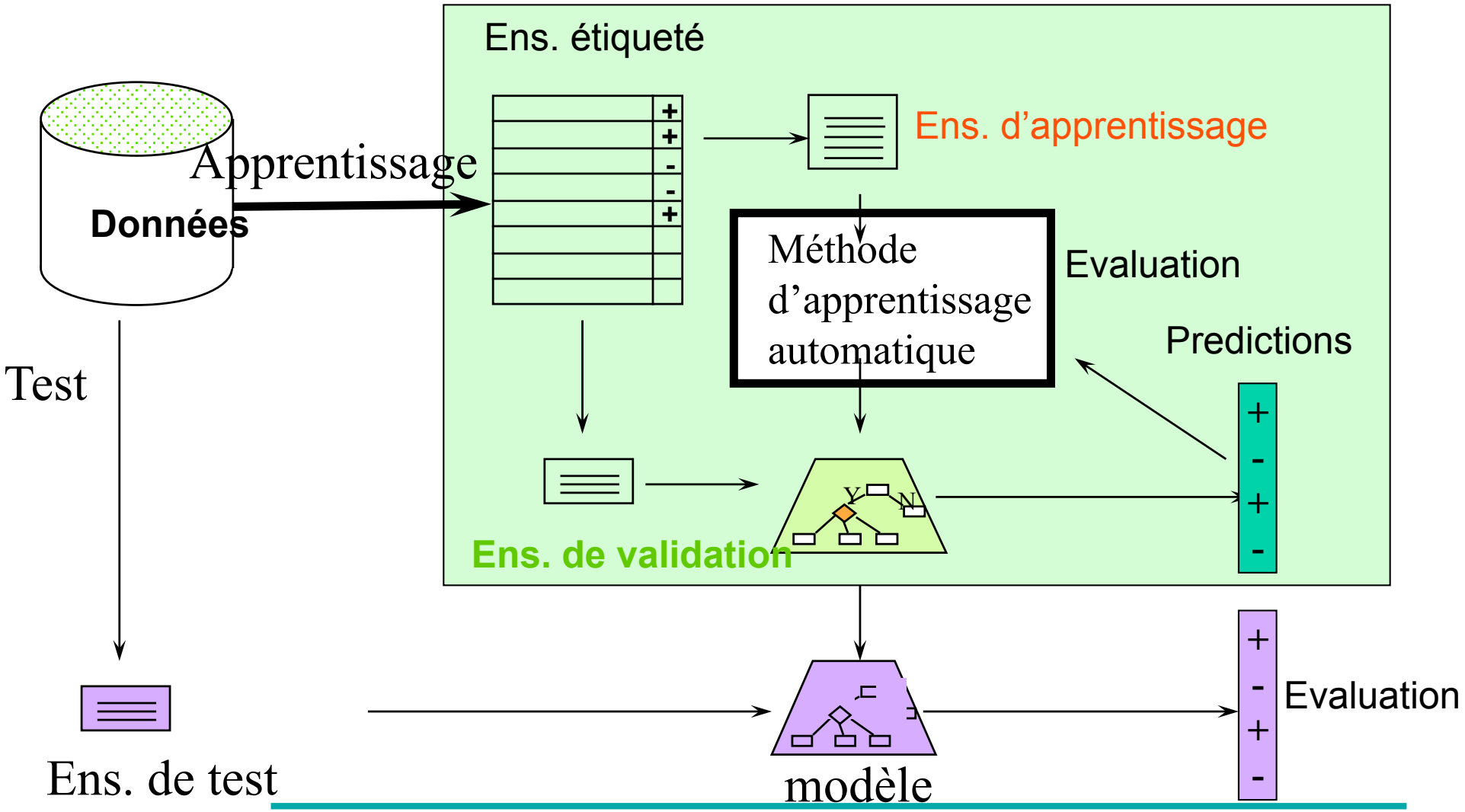
Propriété d'approximation universelle
⇒ biais de l'échantillon
⇒ variance du modèle

Généralisation à de nouvelles données



Validité des résultats : nombreux individus

- séparation ens. d'apprentissage, de validation et de test



Laurent.Bougrain@loria.fr

Validité des résultats : peu d'individus

- Validation croisée (10-fold cross validation)

Partitionnez les individus en k groupes de taille semblable

Construisez k modèles en utilisant (k-1) groupes et évaluez-le en utilisant le dernier groupe *(en rouge)*

Modèle 1

Modèle k

L'estimation de l'erreur est basée sur la moyenne des erreurs observées

- **leave-one-out (« n-fold cross validation »)**

Validité des résultats : peu d'individus

- bootstrap

Si on possède n individus, on constitue l'ens. d'apprentissage en tirant n fois un exemple au hasard avec remise. Ceux qui n'ont pas été sélectionnés forment l'ens. de test. On recommence plusieurs fois et on moyenne l'erreur.

- **0.632 Bootstrap**

Si on possède n individus, chaque individu a une probabilité de 1-1/n de faire partie de l'ens. de test. Donc environs 36,8% des individus feront partie de l'ens. de test (et 63,2% de l'ens. d'apprentissage) :

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

L'erreur est calculée de la manière suivante : $err = 0.632 \cdot err_{\text{test}} + 0.368 \cdot err_{\text{app}}$

Comme on n'apprend que sur 63,2% d'un petit ens d'individus, le modèle ne sera certainement pas aussi bon que si on les utilise tous. L'erreur sur l'ens de test serait très inférieure à la vérité.

On recommence plusieurs fois et on moyenne l'erreur.

Laurent.Bougrain@loria.fr

Validité des résultats : intervalle de confiance

Une autre solution est de calculer, à partir de l'erreur empirique, l'intervalle [a,b] dans lequel se trouve l'erreur réelle avec une probabilité de α%.

De nombreux nombreuses estimations sont possibles.

Ex: pour une discrimination entre deux classes :

Quand le nombre d'individus n est assez grand (>50)

$$a = \frac{T + \frac{Z_\alpha^{\,2}}{2n} - Z_\alpha \sqrt{\frac{T(1-T)}{n} + \frac{Z_\alpha^{\,2}}{4n^2}}}{1 + \frac{Z_\alpha^{\,2}}{n}}$$

$$b = \frac{T + \frac{Z_\alpha^{\,2}}{2n} + Z_\alpha \sqrt{\frac{T(1-T)}{n} + \frac{Z_\alpha^{\,2}}{4n^2}}}{1 + \frac{Z_\alpha^{\,2}}{n}}$$

$$\Delta = 2Z_\alpha^{\,2}\left(\frac{T(1-T)}{n} + \frac{Z_\alpha^{\,2}}{4n^2}\right)$$

T: taux de reconnaissance (bonne classification)

$$P(|X| < Z_\alpha) = 1 - \alpha$$

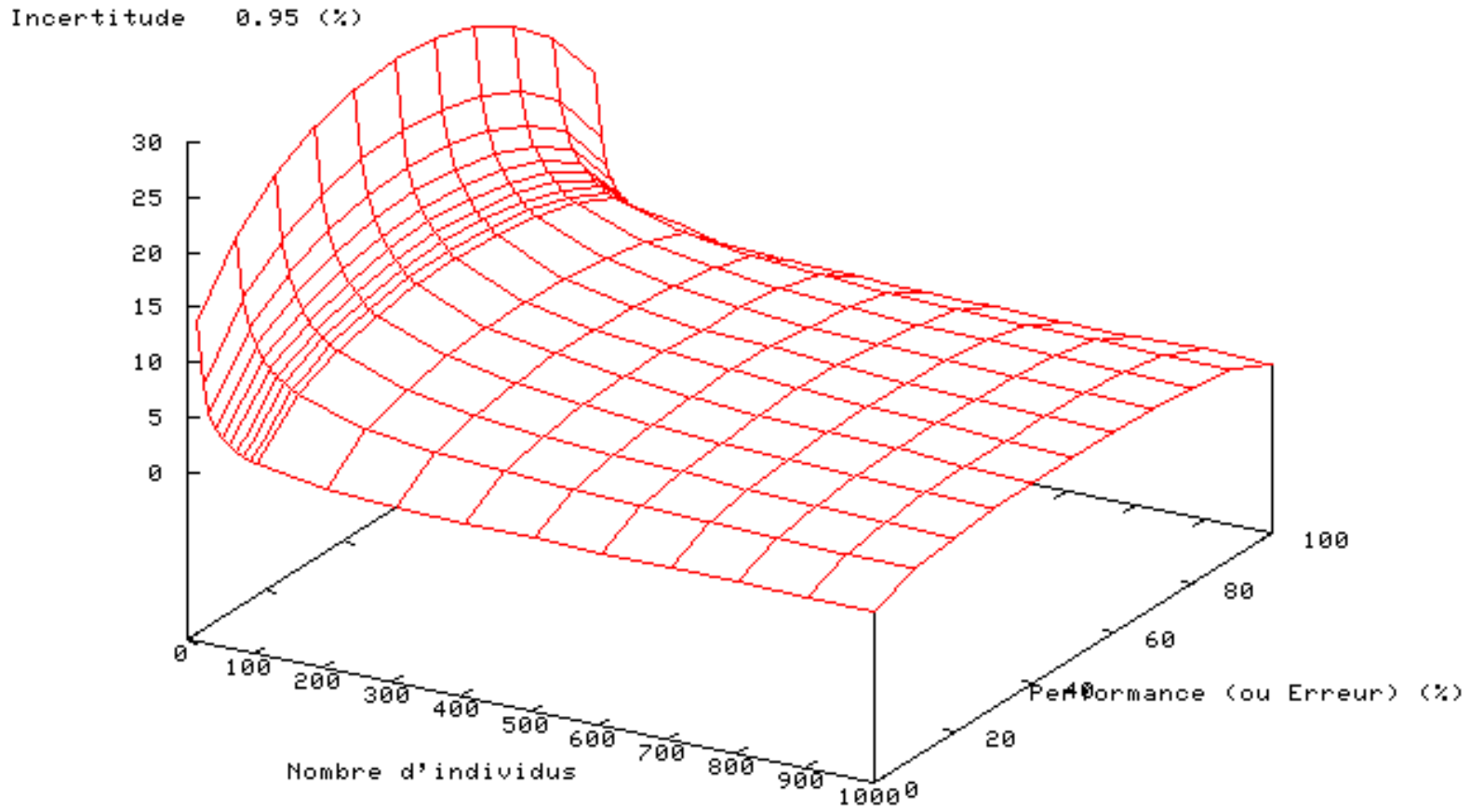
Z_α : déduit de la table de la distribution de X (loi normale)

n : nombre d'exemples de l'ensemble de test

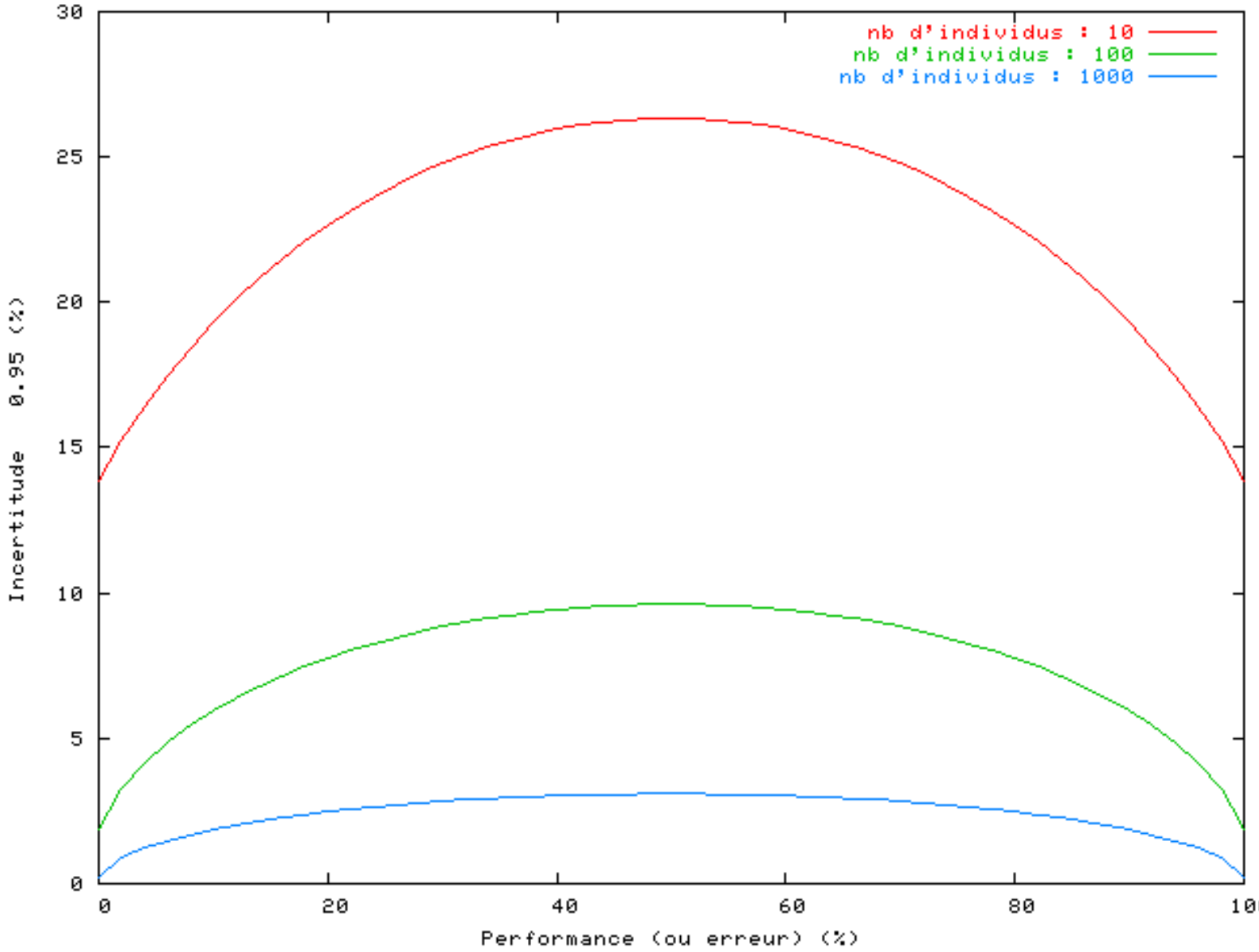
Quand le nombre d'individus n est très grand : $I_\alpha = T \pm Z_\alpha \sqrt{\frac{T(1-T)}{n}}$

Laurent.Bougrain@loria.fr

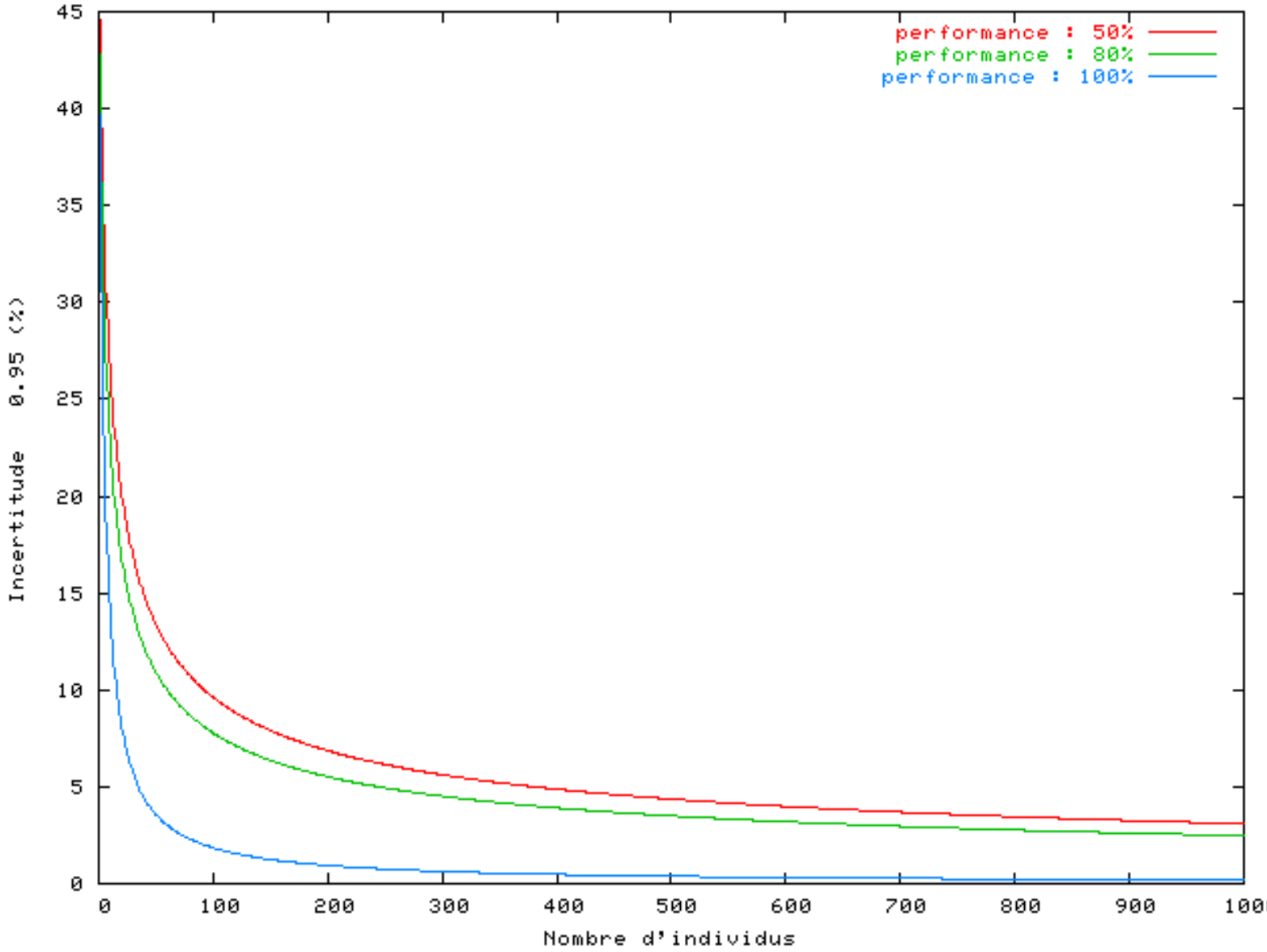
Validité des résultats : intervalle de confiance



Validité des résultats : intervalle de confiance



Validité des résultats : intervalle de confiance



Evaluation des résultats : résumé

- **Equilibrez les classes**
- **Etudiez la matrice de confusion, le rappel, la précision, tracer la courbe ROC…**

- **Utilisez la séparation ens d'app, de validation et de test pour les grands ensembles d'individus**
- **Utilisez la validation croisée pour les petits ensembles d'individus**

Attention au sur-apprentissage !

Méthodes de fouilles de données

Analyse factorielle discriminante
Regression linéaire
Classification hiérarchique ascendante
K-means
…

Arbres de décision
Réseaux de neurones artificiels
Règles d'association
Raisonnement à partir de cas
Machines à vecteurs supports
K plus proches voisins
…