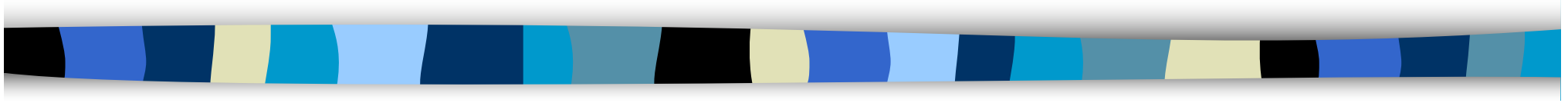
Évaluation de l'apprentissage : méthodes



Antoine Cornuéjols

AgroParisTech

(basé sur Sebastian Thrun CMU class
et sur tutoriel Padraic Cunningham ECML-09)

Questions

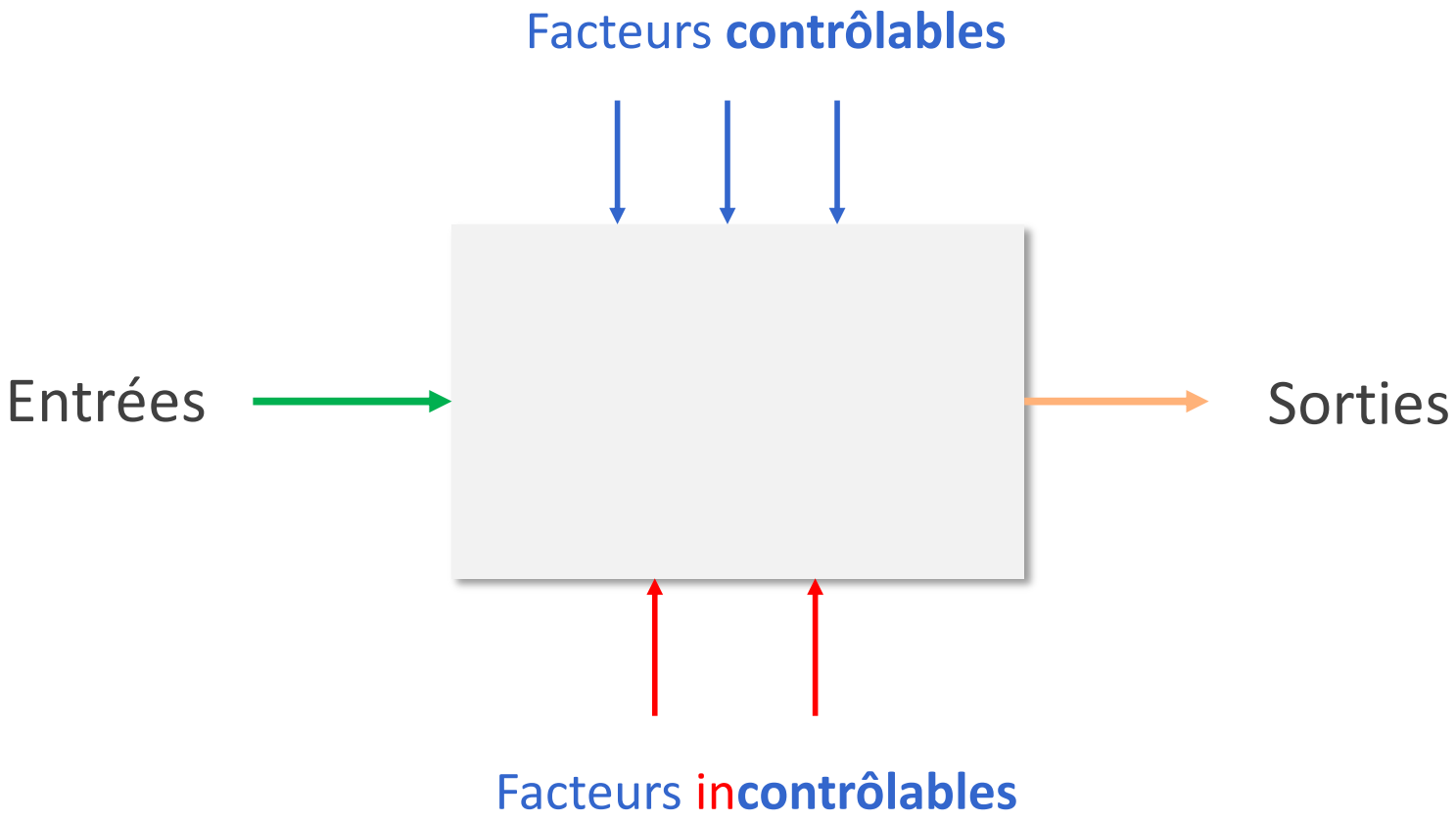
L'induction est une forme d'inférence faillible, il faut donc savoir évaluer sa qualité

- Questions types:
  - Quelle est la performance d'un système sur un type de tâche ?
  - Est-ce que mon système est meilleur que l'autre ?
  - Comment dois-je régler mon système ?

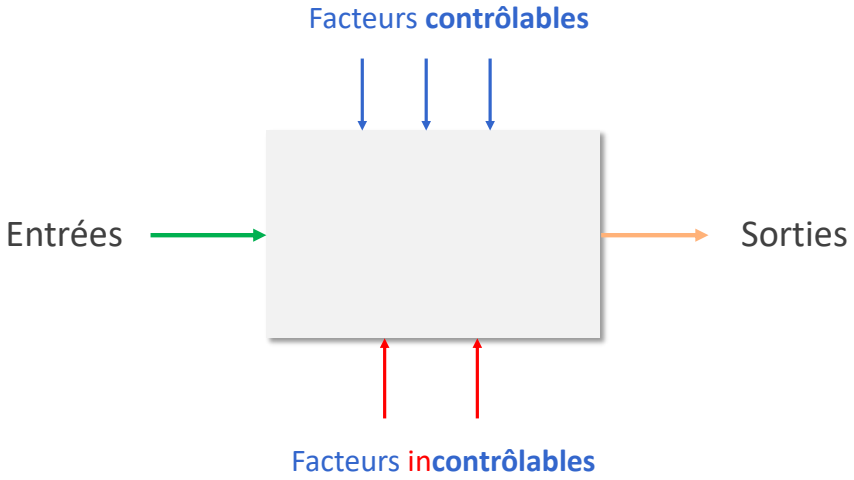
Précautions

1. Attention : les résultats obtenus ne fournissent par la performance (absolue) d'un algorithme
   - Ils **dépendent** du jeu de **données**
   - Le nfl : pas de meilleur algorithme dans l'absolu

2. Le plus souvent, on compare les algorithmes par le **taux d'erreur**
   - Mais, ce n'est qu'un critère d'évaluation
   - Qui **nous aveugle peut-être** et nous trompe dans le type d'algorithme à inventer (cf. Léon Bottou, Jean-Louis Dessalles)

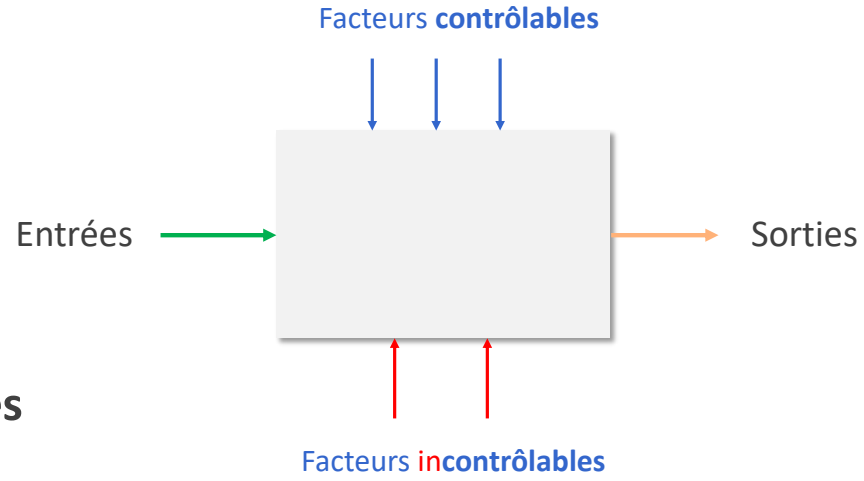

Facteurs **contrôlables**
Entrées
Sorties
Facteurs **incontrôlables**

...



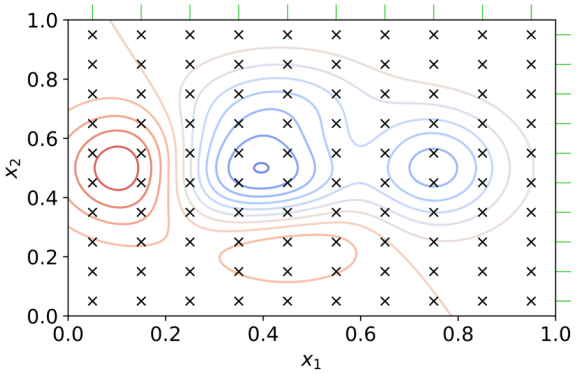
- Facteurs **contrôlables**
  - L'**algorithme**
  - Ses **méta-paramètres**
    - **Architecture** du RN
    - **Nombre** de voisins et **distance** si k-ppv
    - Type de **prétraitements**
- Facteurs **in**contrôlables
  - **Bruit** dans les données
  - Le **tirage aléatoire** des données
  - Le caractère éventuellement **aléatoire de l'algorithme**
    - E.g. Initialisation

Il faut **répéter** les expériences pour obtenir une distribution correspondant aux effets des facteurs incontrôlables



- **Optimisation** des facteurs **contrôlables**
  - Par **grid search**
  - Optimisation **bayésienne**
  - Optimisation par **gradient**
  - Optimisation par **algorithmes évolutionnaires**

Attention : l'optimisation des hyper-paramètres est réalisée sur l'ensemble de validation et donc sujette à sur-ajustement (over-fitting)



Ici, optimisation de deux hyper-paramètres.
En **bleu** : des zones de performance élevée
En **rouge** : de faible performance
(From Wikipedia)

Plan

1. Que mesurer
2. Comment le mesurer
3. La courbe ROC
4. Autres mesures de performances

Types de mesures de performance

```
Correctly Classified Instances         117               70.9091 %
Incorrectly Classified Instances        48               29.0909 %
Kappa statistic                          0.3071
Mean absolute error                      0.2909
Root mean squared error                  0.5394
Relative absolute error                 62.6804 %
Root relative squared error            112.1168 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.617      0.718     0.895     0.797      0.639    good
                 0.383     0.105      0.676     0.383     0.489      0.639    bad
Weighted Avg.    0.709     0.431      0.703     0.709     0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

SVM

```
Correctly Classified Instances         103               62.4242 %
Incorrectly Classified Instances        62               37.5758 %
Kappa statistic                          0.1995
Mean absolute error                      0.3793
Root mean squared error                  0.5316
Relative absolute error                 81.7353 %
Root relative squared error            110.5048 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.686     0.483      0.713     0.686     0.699      0.674    good
                 0.517     0.314      0.484     0.517     0.5        0.674    bad
Weighted Avg.    0.624     0.422      0.63      0.624     0.627      0.674

=== Confusion Matrix ===

  a  b   <-- classified as
 72 33 |  a = good
 29 31 |  b = bad
```

Naive Bayes

Types de mesures de performance

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.895     0.617      0.718     0.895      0.797      0.639    good
                   0.383     0.105      0.676     0.383      0.489      0.639    bad
Weighted Avg.      0.709     0.431      0.703     0.709      0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

Indicateurs de performances

m exemples au total

- ***Sensibilité*** TP-rate $\frac{VP}{FN + VP}$
- ***Rappel*** $\frac{VP}{VP + FN}$
- ***Spécificité*** **TN**-rate $\frac{VN}{VN + FP}$
- ***Précision*** $\frac{VP}{VP + FP}$

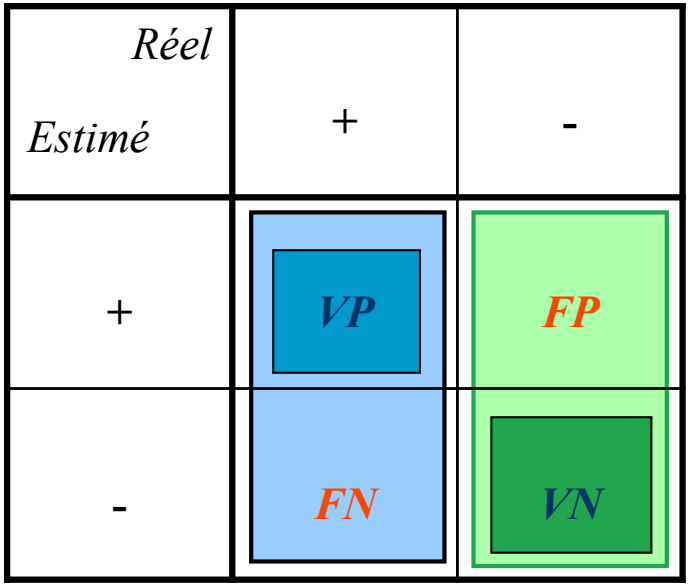
	P	N
Réel / *Estimé*	+	-
+	*VP*	*FP*
-	*FN*	*VN*

- ***Taux d'erreur*** $\frac{FP + FN}{m}$
- ***Accuracy*** = ***1 – Taux d'erreur***

FP-rate = $\frac{FP}{FP + VN}$

Indicateurs de performances

- **FN-rate** $\dfrac{FN}{VP + FN}$
- **FP-rate** $\dfrac{FP}{FP + VN}$
- **F-measure** $\dfrac{2 \times rappel \times précision}{rappel + précision} = \dfrac{2\,VP}{2\,VP + FP + FN}$



Types de mesures de performance

$$F\text{-}measure = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

harmonic mean of precision and recall

Types de mesures de performance

- Test set: 105 good, 60 bad
  - NB Accuracy 62.4%
  - SVM Accuracy 70.1%

SVM

Classified as

good	bad	
94	**11**	good
37	**23**	bad

Act. Class

Naive Bayes

Classified as

good	bad	
72	**33**	good
29	**31**	bad

Act. Class

Types de mesures de performance

- Test set: 105 good, 60 bad
  - NB Accuracy 62.4%
  - SVM Accuracy 70.1% ← Apparent best

SVM

Classified as

	good	bad	
105	**94**	**11**	good
60	**37**	**23**	bad
	131	34	

Act. Class

SVM biased toward majority class

Naive Bayes

Classified as

	good	bad	
105	**72**	**33**	good
60	**29**	**31**	bad
	101	64	

Act. Class

What if this is important?

Types de mesures de performance

- Hold-out validation - 33% holdout set

```
Correctly Classified Instances         117               70.9091 %
Incorrectly Classified Instances        48               29.0909 %
Kappa statistic                          0.3071
Mean absolute error                      0.2909
Root mean squared error                  0.5394
Relative absolute error                 62.6804 %
Root relative squared error            112.1168 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.617      0.718     0.895      0.797      0.639    good
                 0.383     0.105      0.676     0.383      0.489      0.639    bad
Weighted Avg.    0.709     0.431      0.703     0.709      0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

SVM

	TPR	FPR
SVM	0.89	0.62
NB	0.69	0.48

```
Correctly Classified Instances         103               62.4242 %
Incorrectly Classified Instances        62               37.5758 %
Kappa statistic                          0.1995
Mean absolute error                      0.3793
Root mean squared error                  0.5316
                                        81.7353 %
                                       110.5048 %
                                       165

                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                  0.686     0.483      0.713     0.686      0.699      0.674    good
                  0.517     0.314      0.484     0.517      0.5        0.674    bad
                  0.624     0.422      0.63      0.624      0.627      0.674

  29 31 |  b = bad
```

Naive Bayes

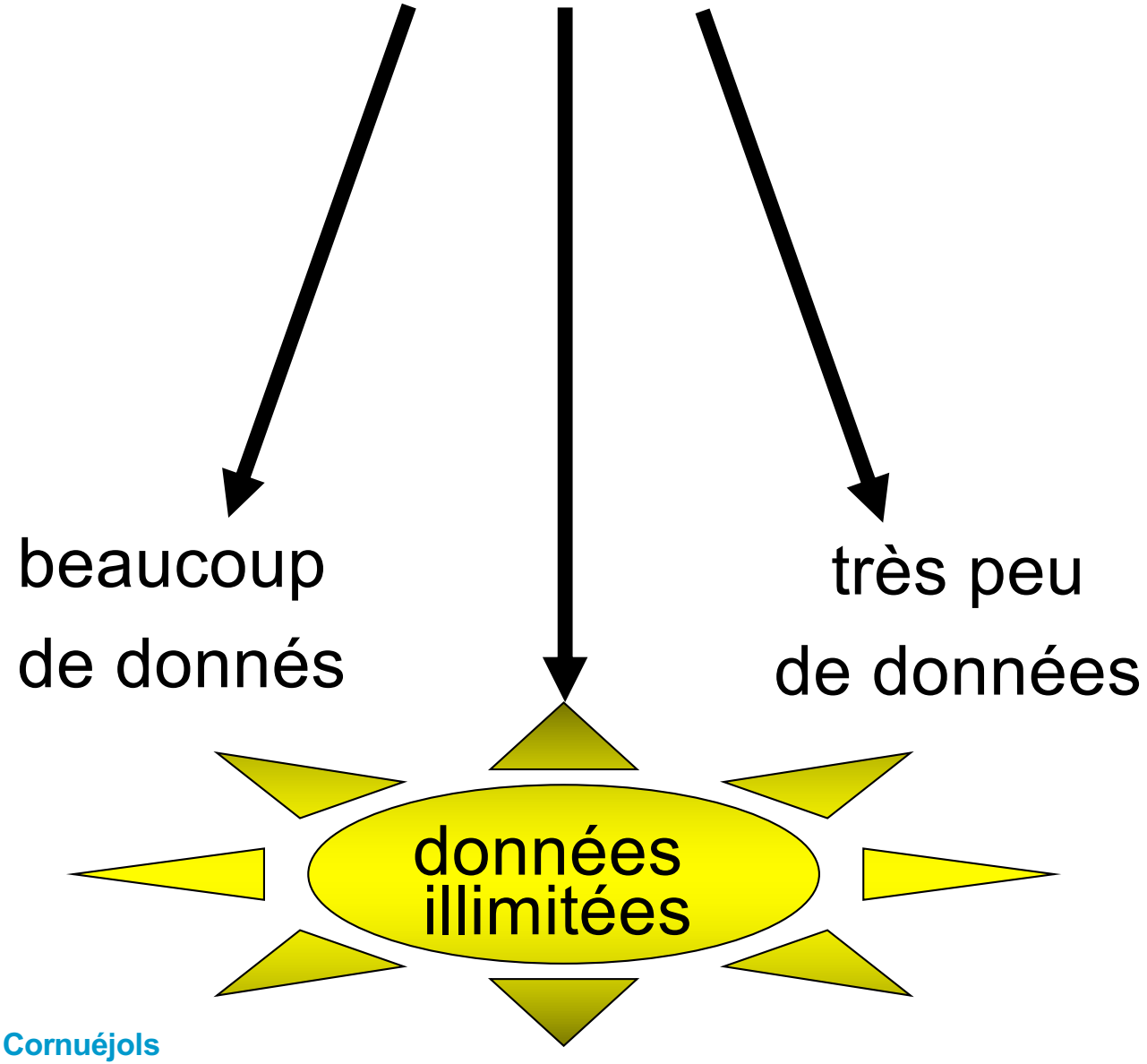
Réel / Estimé	good	bad
good	0.895 = 94/105	0.617 = 37/60
bad	0.105 = 11/105	0.383 = 23/60

Précision(good) = 94 / 131 = 0.718

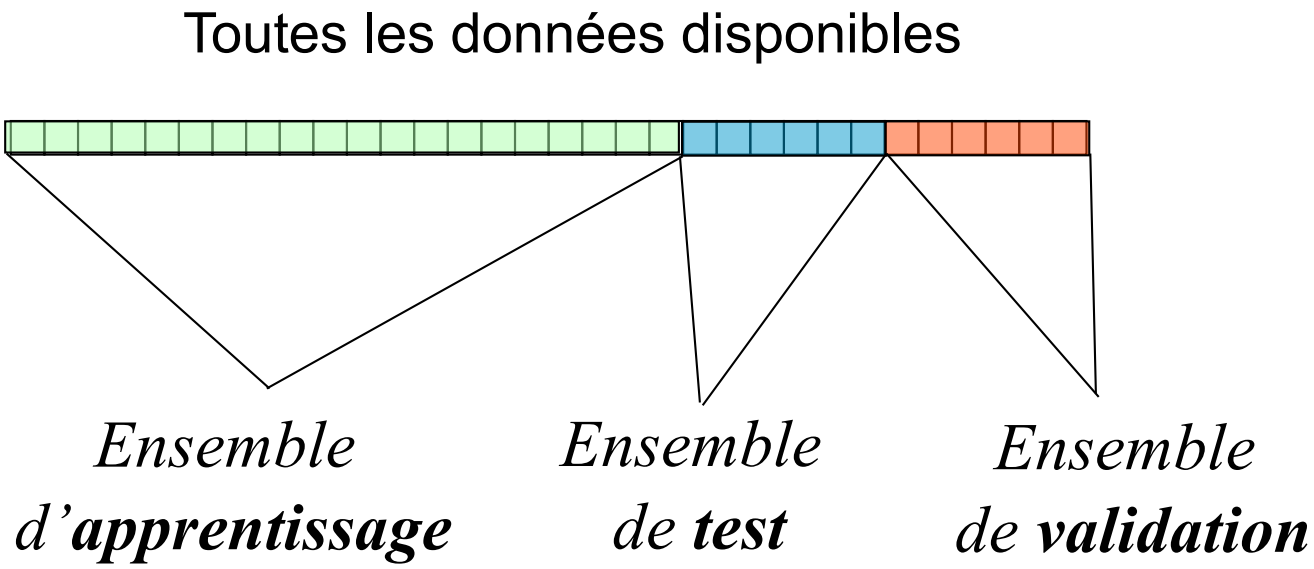
Plan

1. Que mesurer
2. Comment le mesurer
3. La courbe ROC
4. Autres mesures de performances

Évaluation des hypothèses produites

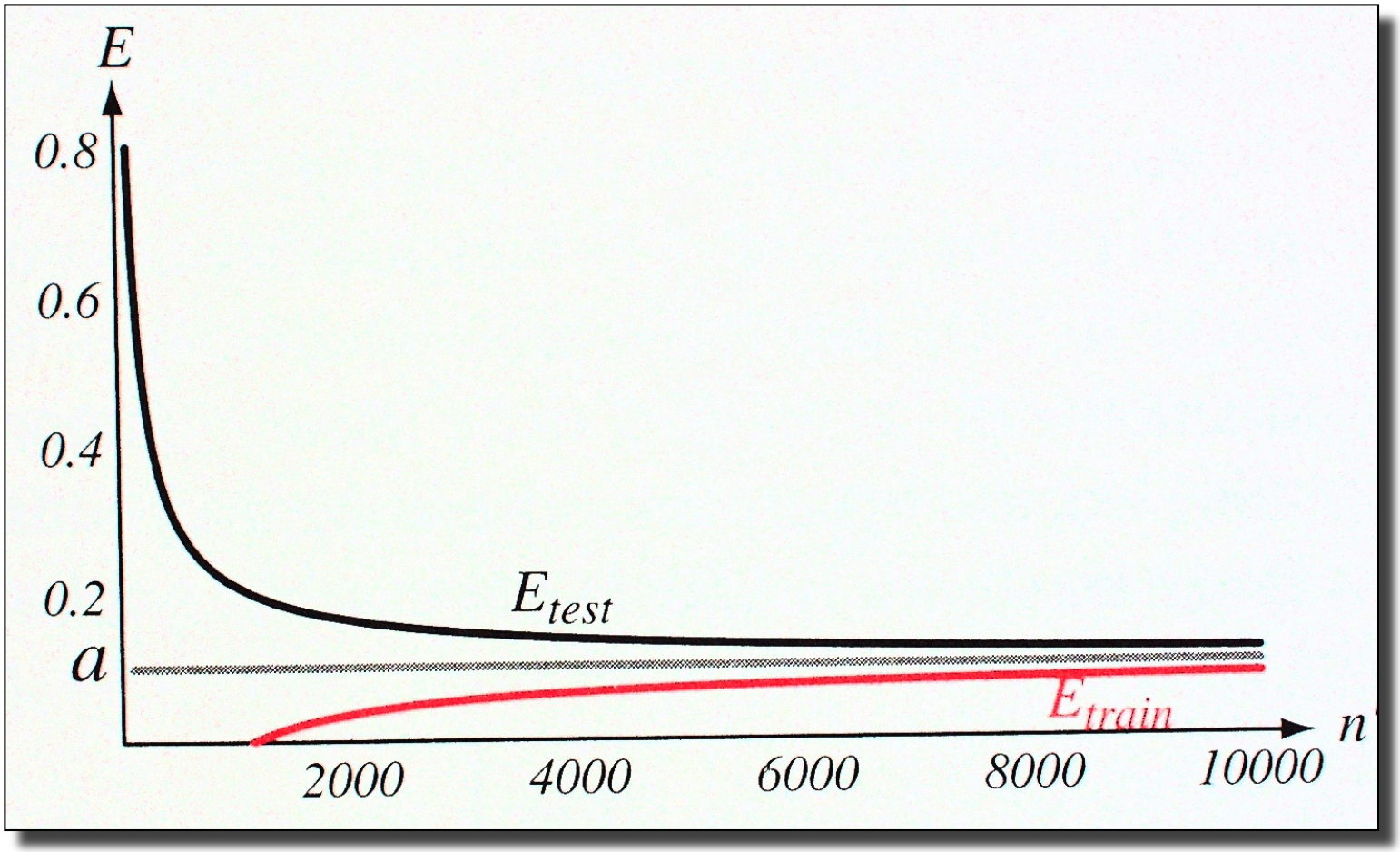


Ensembles de données (collections)

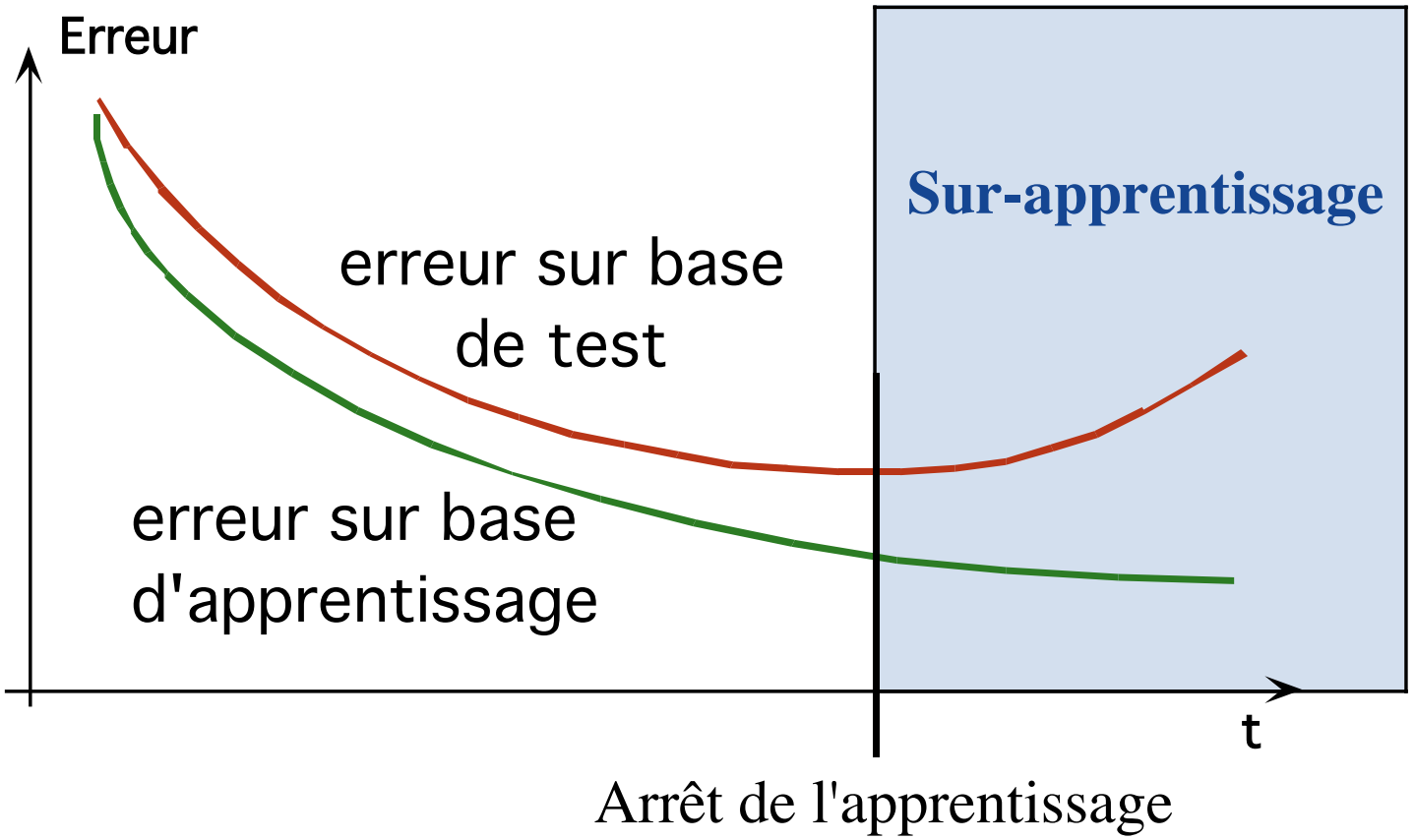


Prédiction asymptotique

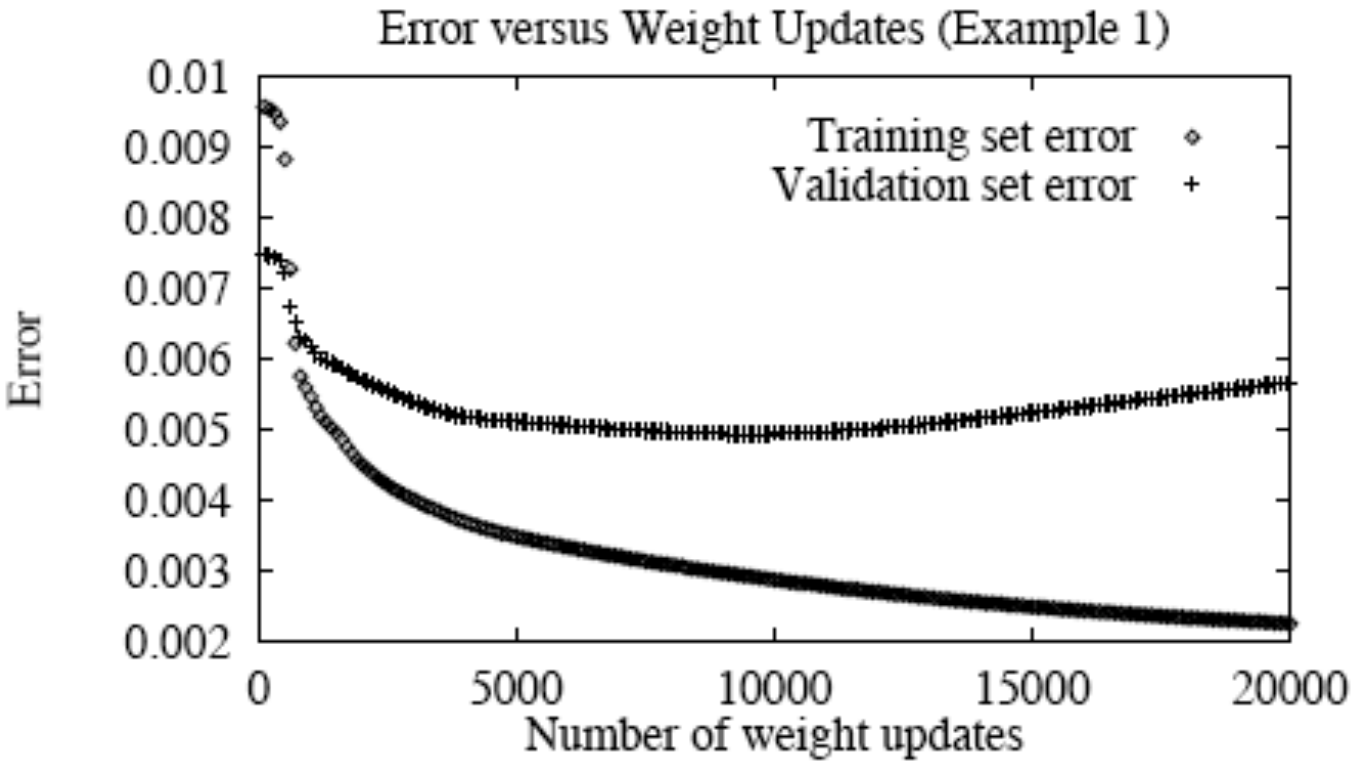
(le cas idéal)



Le **sur-apprentissage** (*over-learning*)



Sur-apprentissage (RN)

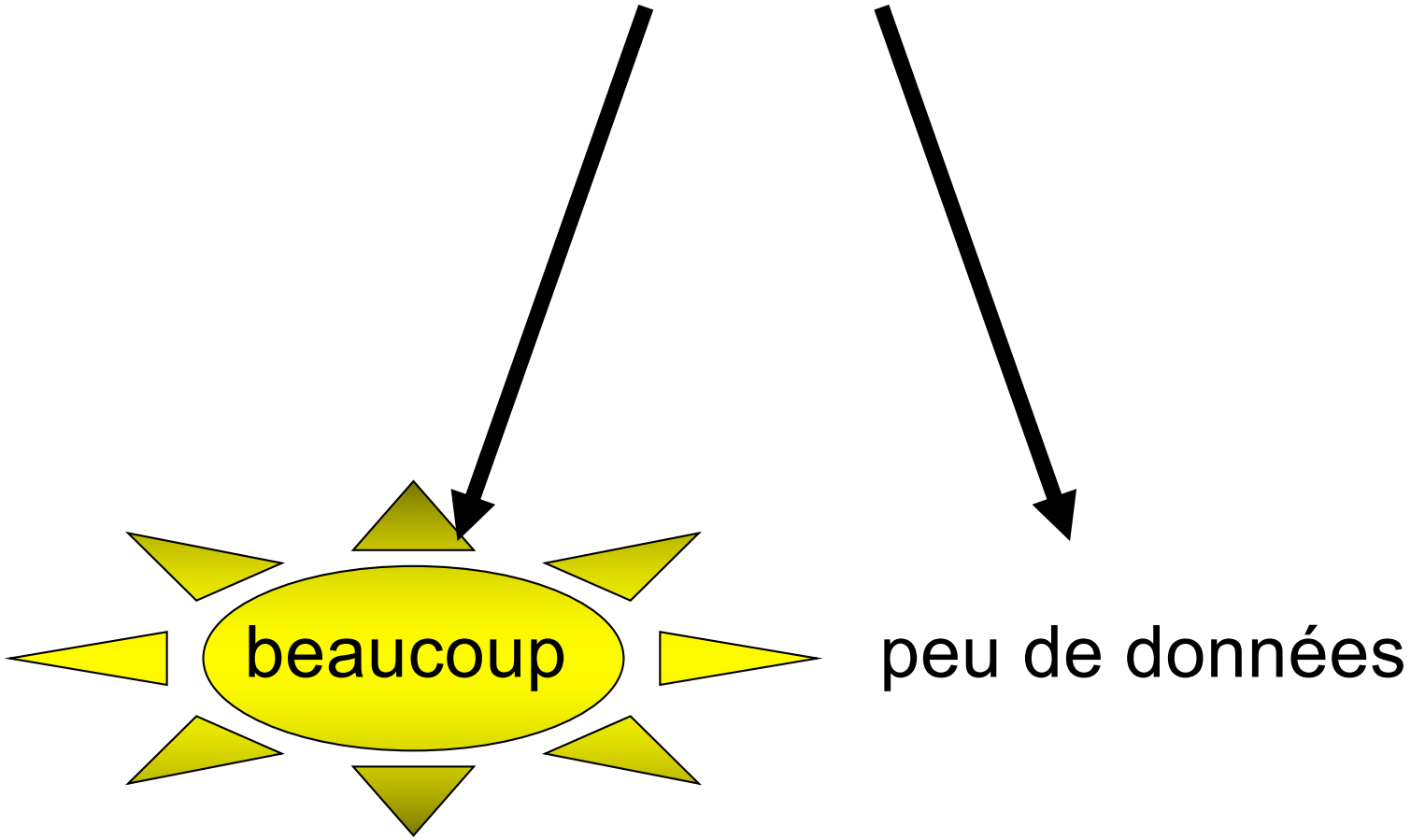


- Courbes pour 1 000 exemples
- ***Courbes pour 2 000 exemples ?***

Utilisation de l'ensemble de validation

- On règle les paramètres de l'algorithme d'apprentissage
  - E.g. : nb de couches cachées, nb de neurones, ...
  - en essayant de réduire l'erreur de test
- Pour avoir une estimation non optimiste de l'erreur, il faut recourir à une base d'exemples non encore vus : la ***base de validation***

Évaluation des hypothèses produites



Évaluation de l'erreur

- Erreur vraie:

(Risque **réel**)

$$e_D = \int_D \left| y - f(x,\theta\ ) \right|\ p(x,y)\ dx, y$$

D = toutes les données possibles

- Erreur de test:

(Risque **empirique**)

$$\hat{e}_S = \frac{1}{m} \sum_{\langle x,y \rangle \in T} \left| y - f(x,\theta\ ) \right|$$

T = données test

m = # de données test

Exemple:

- L'hypothèse classe mal 12 des 40 exemples dans l'ensemble de test T.
- Q : Quelle sera l'erreur sur des exemples non vus ?
- R : ???

Intervalle de confiance (1)

- *Définition* : un ***intervalle de confiance*** à *N*% pour une variable *p* est l'intervalle dans lequel sa valeur est attendue avec une probabilité de *N*%
- Soit une probabilité d'erreur (pour 2 classes) de *p*, la **probabilité d'avoir *r* erreurs sur *n* évènements** est :

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

(loi binomiale)

Espérance du nombre d'erreurs $E[X] = np$

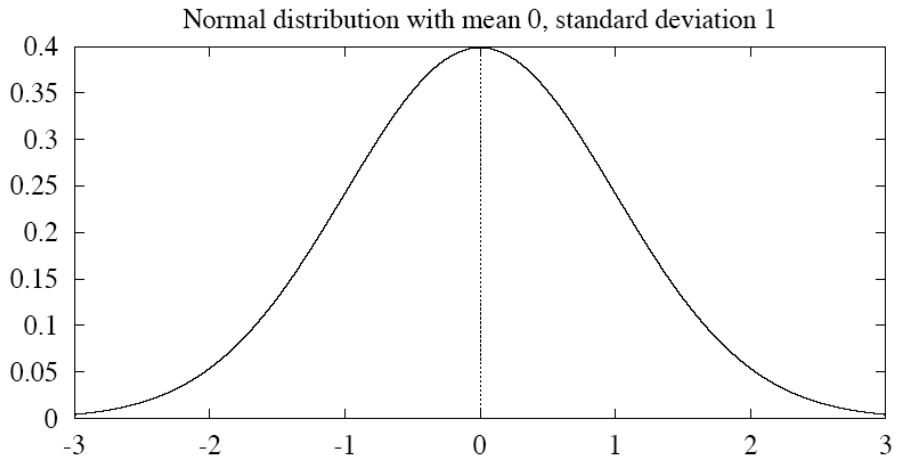
Variance $Var(X) = np(1-p)$

Ecart-type $\sigma_X = \sqrt{np(1-p)}$

Intervalles de confiance (2)

- La ***loi binomiale*** peut être estimée par la ***loi normale***
  si $n\,p\,(1 - p) \geq 5$ de même moyenne μ et même variance σ

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Intervalles de confiance (3)

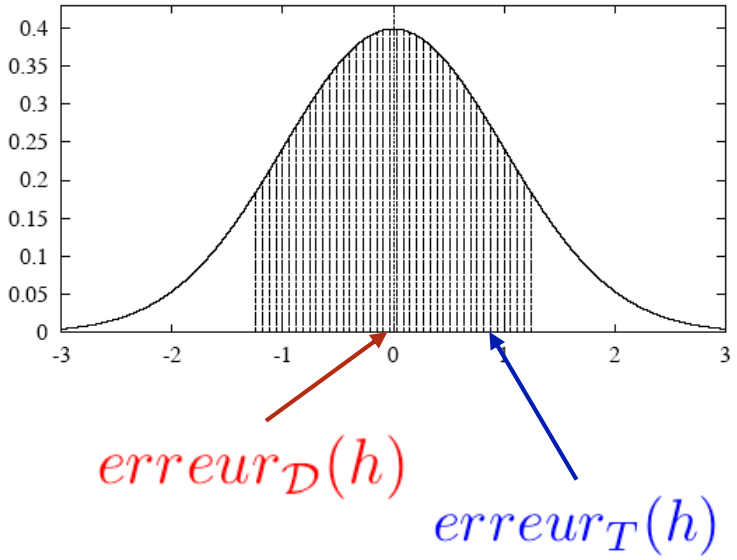
- Je voudrais évaluer $erreur_{\mathcal{D}}(h)$.
- Je l'estime en utilisant $erreur_T(h)$ qui est régie par une loi binomiale
  - De moyenne $\mu_{erreur_{\mathcal{D}}}(h) = erreur_{\mathcal{D}}(h)$
  - D'écart-type $\sigma_{erreur_{\mathcal{D}}}(h) = \sqrt{\frac{erreur_{\mathcal{D}}(h)\,(1 - erreur_{\mathcal{D}}(h))}{n}}$
- **Que l'on estime par** la loi normale
  - De moyenne : $\mu_T(h) = erreur_T(h)$
  - D'écart-type : $\sigma_T(h) \approx \sqrt{\frac{erreur_T(h)\,(1 - erreur_T(h))}{n}}$

Intervalles de confiance (4)

- Loi normale

$$\mu_{erreur_{\mathcal{D}}}(h) = erreur_{\mathcal{D}}(h)$$

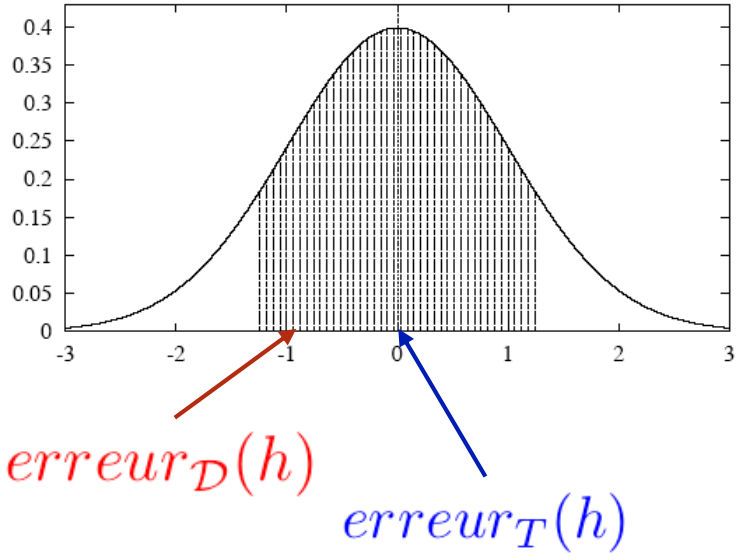
$$\sigma_{erreur_{\mathcal{D}}}(h) = \sqrt{\frac{erreur_{\mathcal{D}}(h)\,(1 - erreur_{\mathcal{D}}(h))}{n}}$$



- Loi normale

$$\mu_T(h) = erreur_T(h)$$

$$\sigma_T(h) \approx \sqrt{\frac{erreur_T(h)\,(1 - erreur_T(h))}{n}}$$



Intervalles de confiance (5)

Avec une probabilité de *N*%, l'erreur vraie $erreur_D$ est dans l'intervalle :

$$erreur_T(h) \pm Z_N \sqrt{\frac{erreur_T(h)\,(1 - erreur_T(h))}{n}}$$

N%	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.0	1.28	1.64	1.96	2.33	2.58

Intervalles de confiance (cf. Mitchell 97)

Si

- T contient m exemples tirés indépendamment
- $m \geq 30$

Alors

- Avec une probabilité de 95%, l'erreur vraie e_D est dans l'intervalle :

$$\hat{e}_S \pm 1.96 \sqrt{\frac{\hat{e}_S\ (1-\hat{e}_S)}{m}}$$

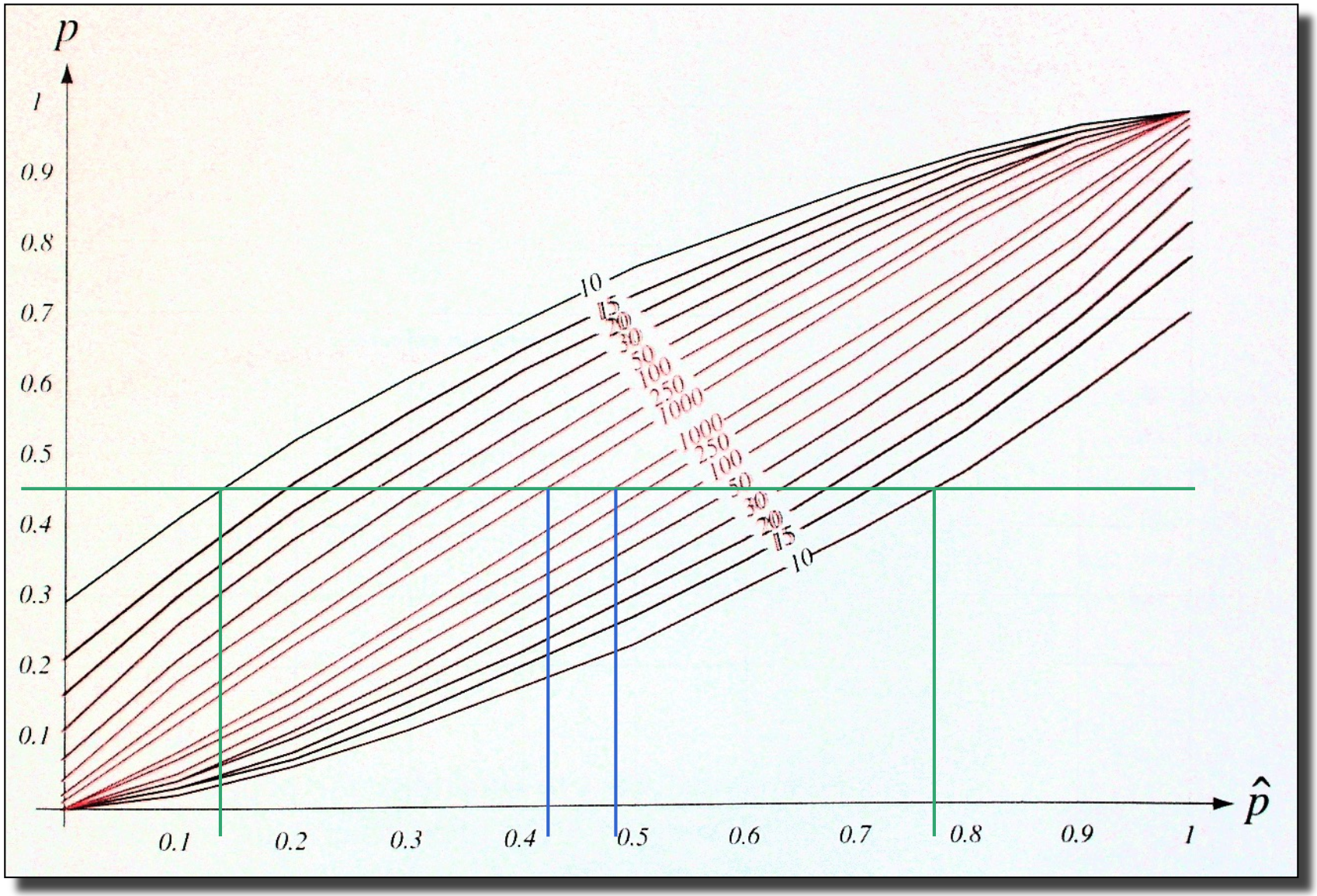
Exemple:

- L'hypothèse classe mal 12 des40 exemples dans la base de test *T*.
- Q: Quelle sera l'erreur vraie sur les exemples non vus ?
- A: Avec 95% de confiance, l'erreur vraie sera dans l'intervalle :

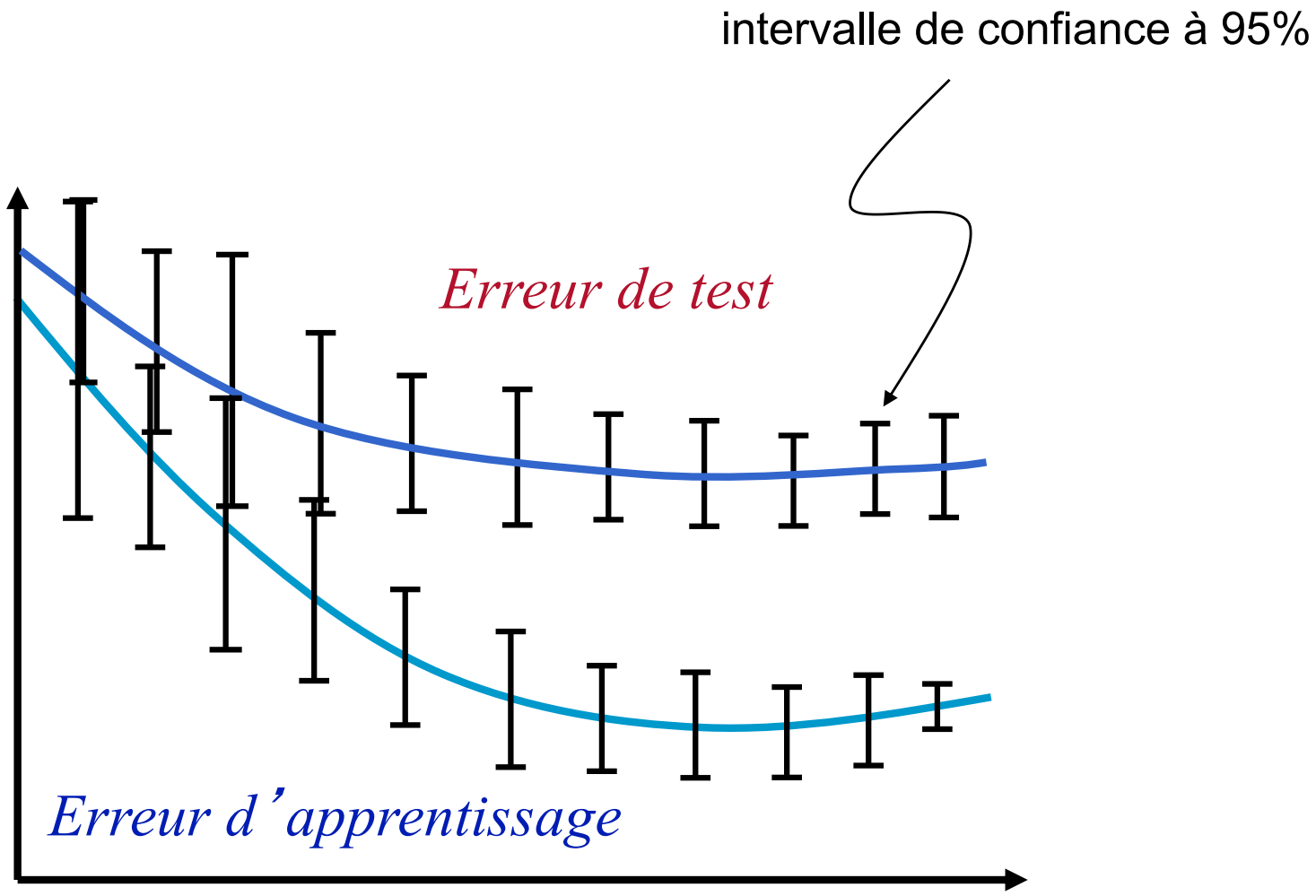
$$[0.16;0.44] \approx \hat{e}_S \pm 1.96\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}}$$

$$m = 40 \qquad \hat{e}_S = \frac{12}{40} = 0.3 \qquad 1.96\sqrt{\frac{\hat{e}_S\,(1-\hat{e}_S)}{m}} \approx 0.14$$

Intervalles de confiance à 95%



Courbes de performance



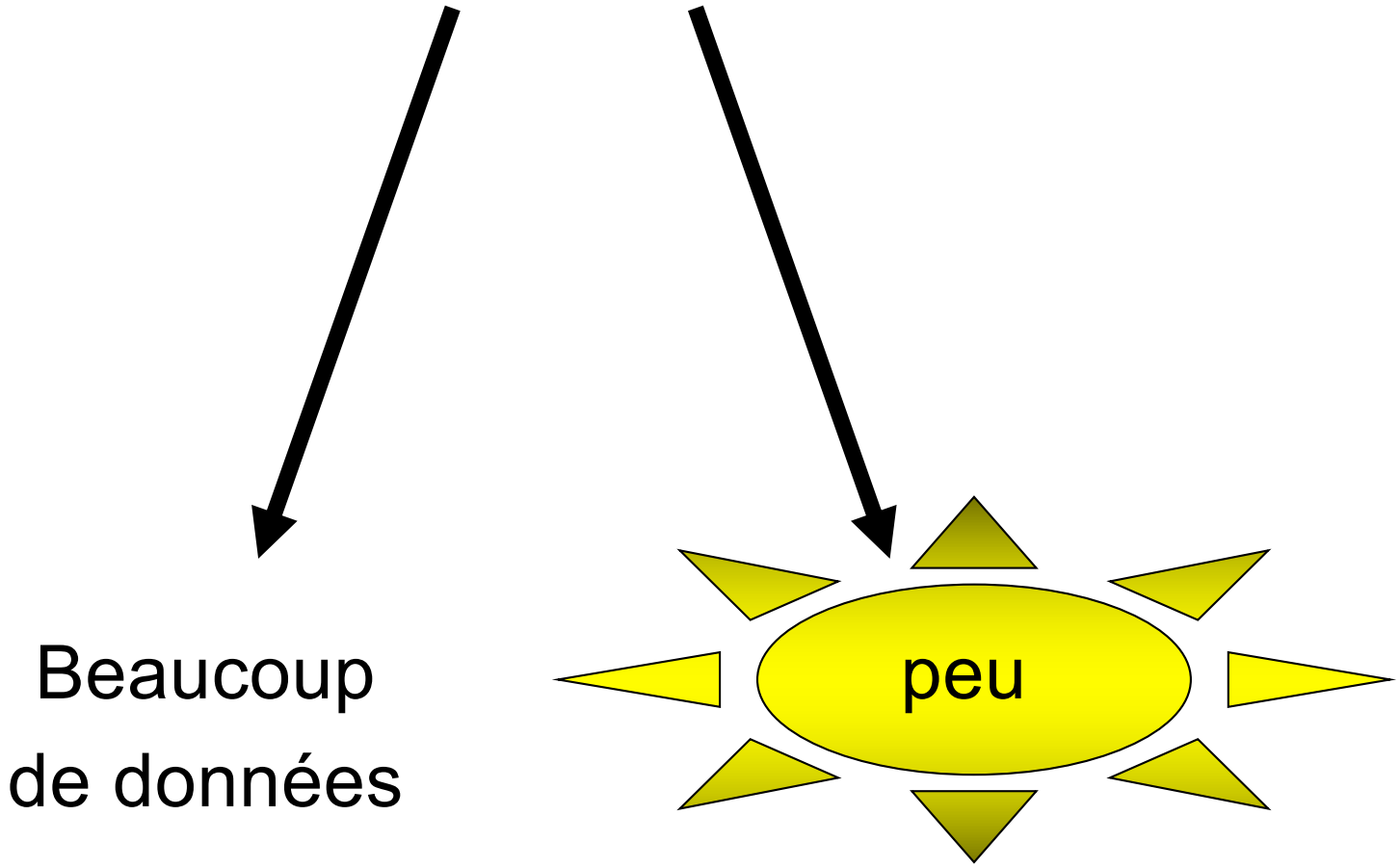
Comparaison de différentes hypothèses

- On cherche la différence vraie: $d = e_D(\theta_1) - e_D(\theta_2)$
- On estime par : $\hat{d} = \hat{e}_S(\theta_1) - \hat{e}_S(\theta_2)$
- Qui est une loi normale différence de 2 lois normales
- Intervalle de confiance à 95% :

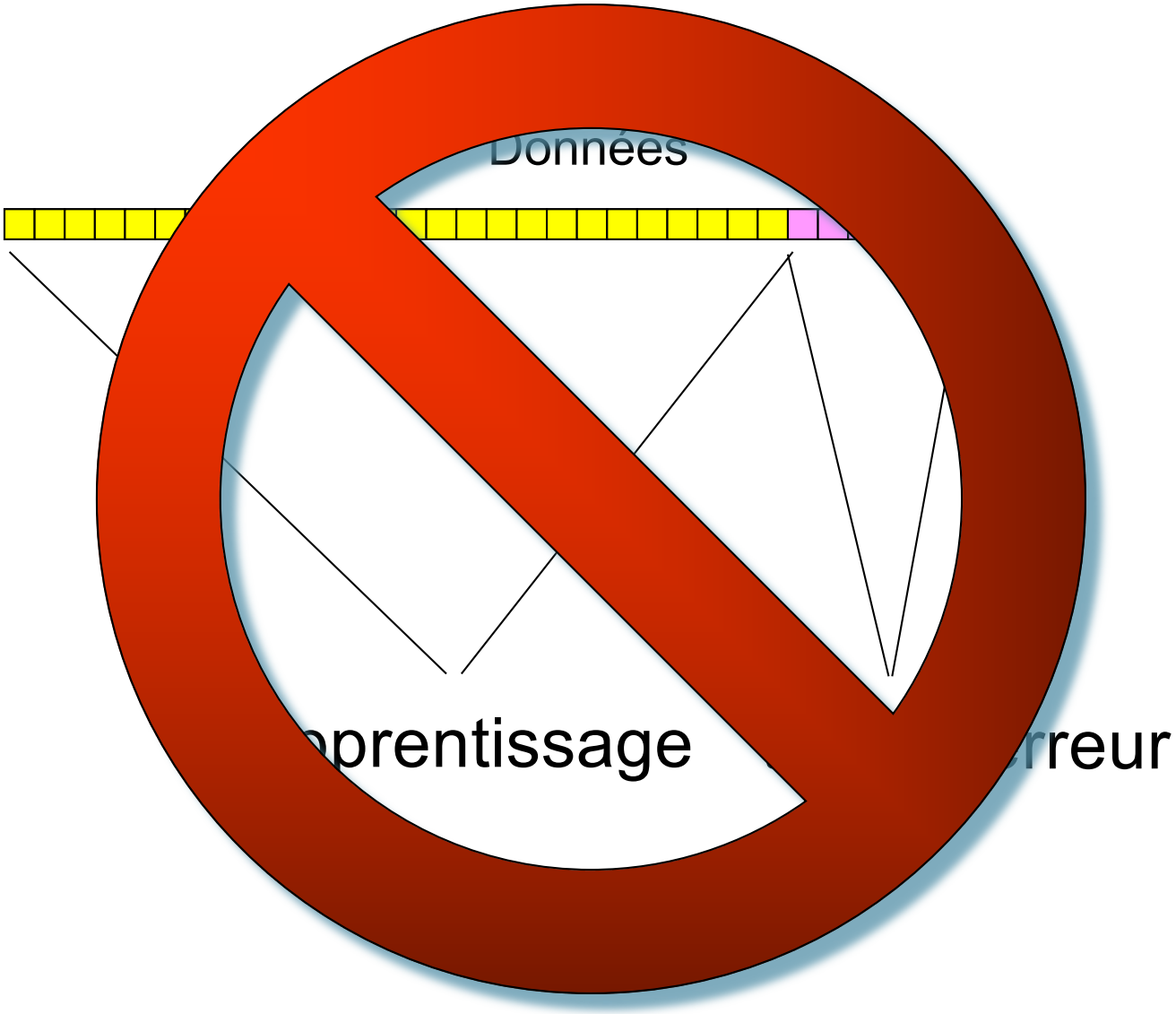
$$\hat{d} \pm 1.96 \sqrt{\frac{\hat{e}_S(\theta_1)\,(1-\hat{e}_S(\theta_1))}{m_1} + \frac{\hat{e}_S(\theta_2)\,(1-\hat{e}_S(\theta_2))}{m_2}}$$

Rq : il faudrait normalement ne pas tester les deux hypothèses sur le même ensemble de test. La variance obtenue avec un même ensemble de test est un peu plus fiable (cf. paired *t* tests).

Évaluation des hypothèses produites

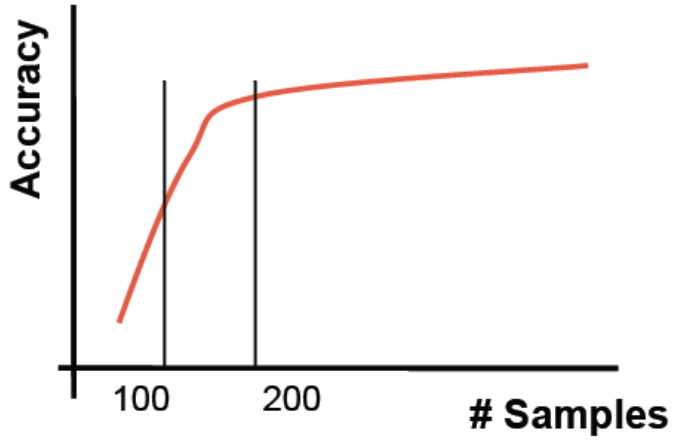


Différents ensembles



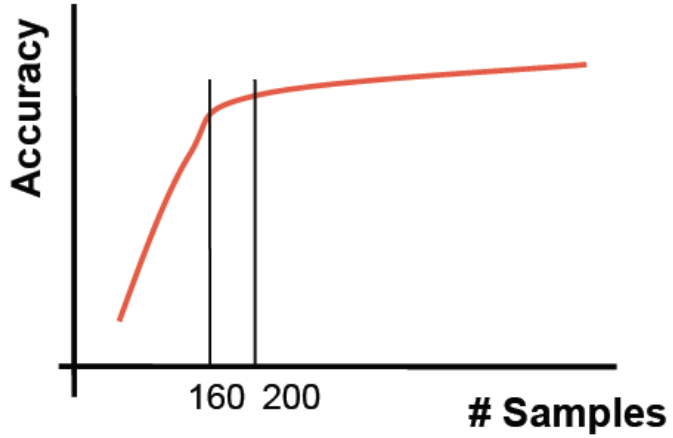
Peu de données

- Imagine 200 samples are available for training:
  - 50:50 split underestimates generalisation acc.

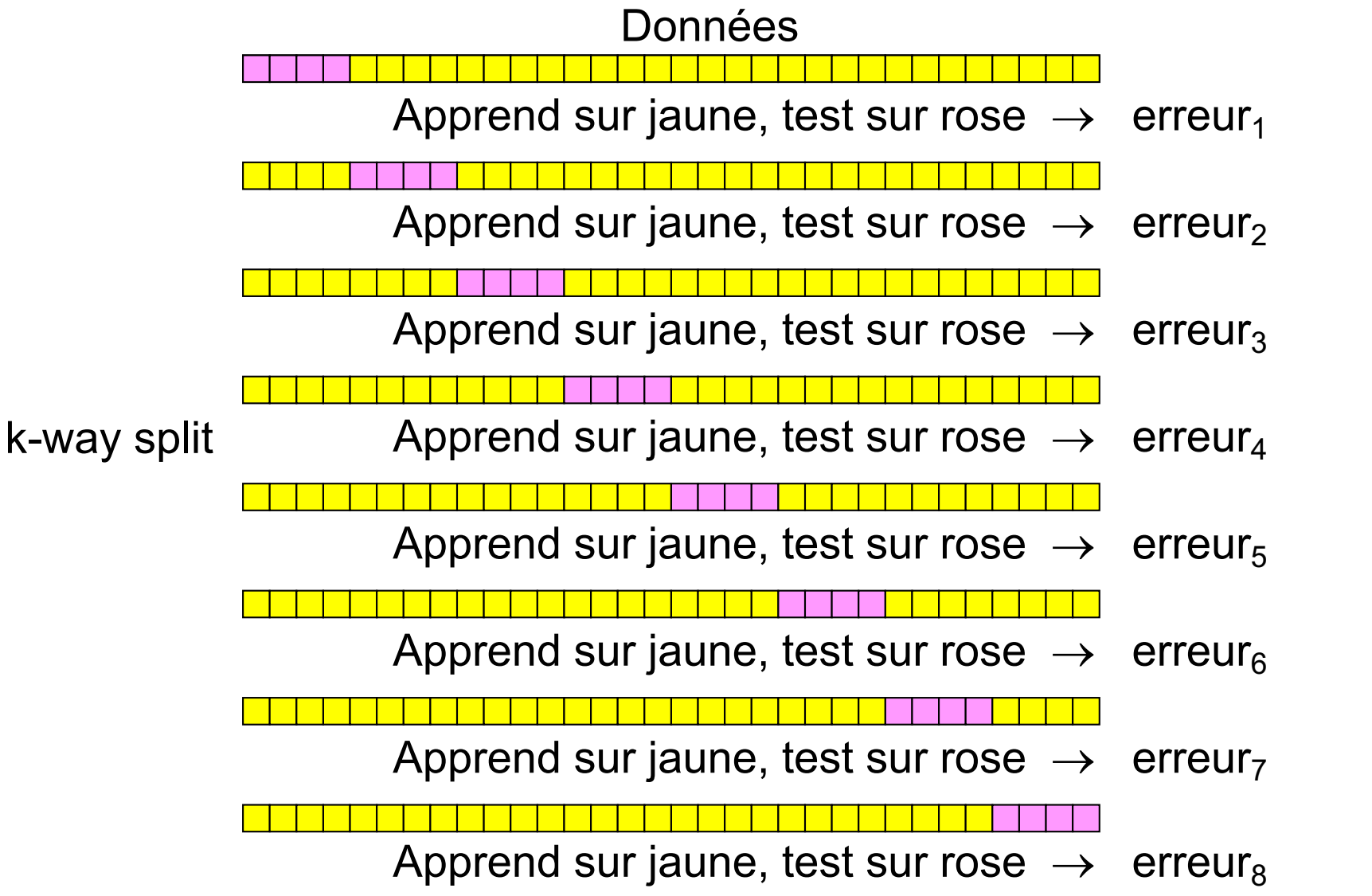


Peu de données

- Imagine 200 samples are available for training:
  - 50:50 split underestimates generalisation acc.
  - 80:20 estimate based on a small sample (40)
    - Different hold-out sets - different results

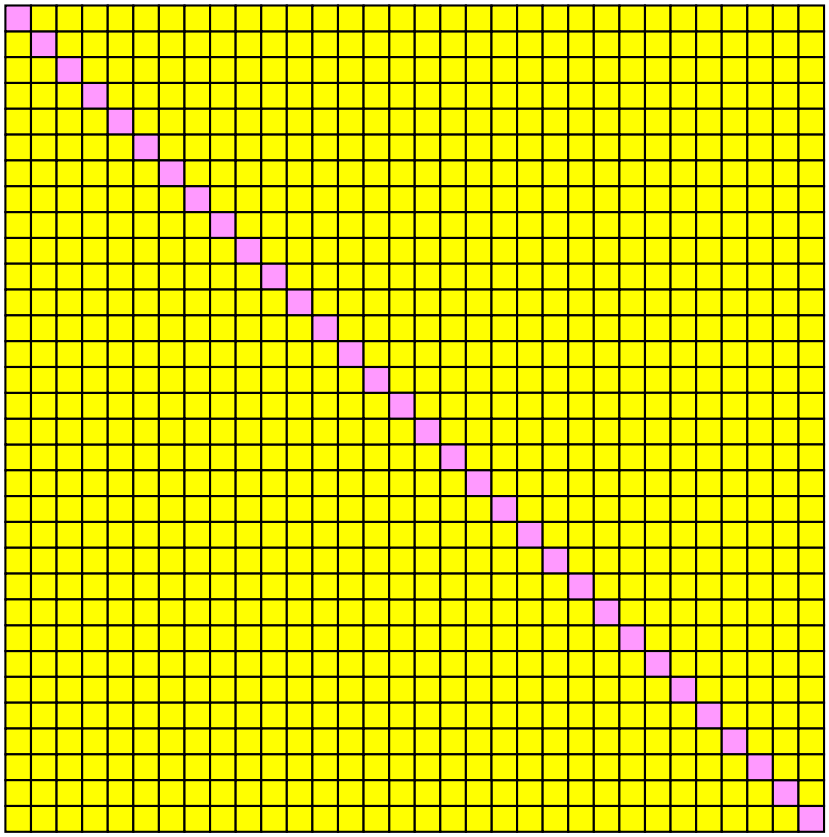


Validation croisée à k plis (k-fold)



Procédure “leave-one-out”

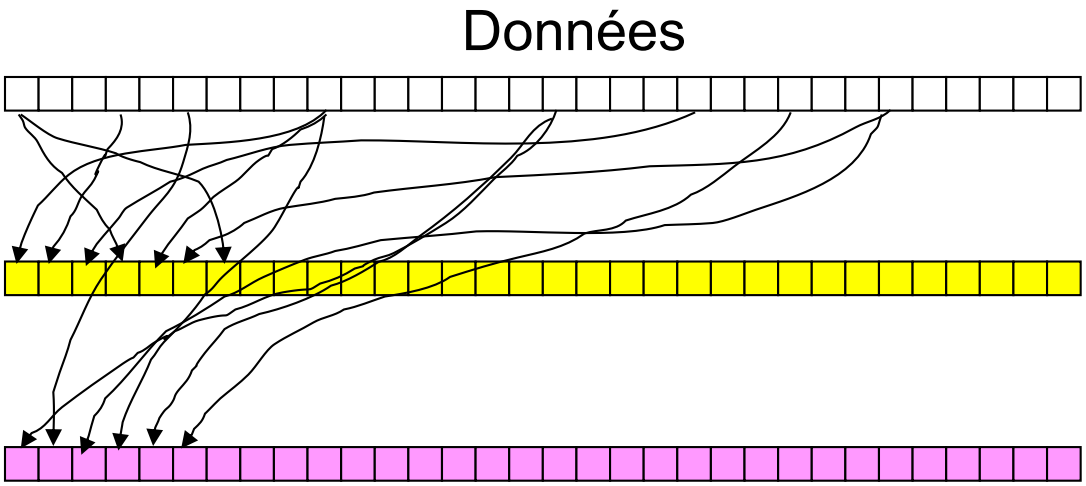
Données



- Faible biais
- Haute variance
- Tend à sous-estimer l’erreur si les données ne sont pas vraiment i.i.d.

[Guyon & Elisseeff, jMLR, 03]

Le Bootstrap



- Apprend sur jaune, test sur rose $\rightarrow$ erreur
- Répéter et faire la moyenne

Problème

- Le calcul des intervalles de confiance suppose l'indépendance des estimations.
- Mais nos estimations sont dépendantes. ☹

$$\hat{R}_{Réel}(h) = 0.636\,\bar{P}_1 + 0.368\,\bar{P}_2$$

Estimation du risque réel pour *h* finale

Moy. du risque sur les *k* ens. de test

Moy. du risque sur l'ens. des données

Plan

1. Que mesurer
2. Comment le mesurer
3. La courbe ROC
4. Autres mesures de performances

Types d'erreurs

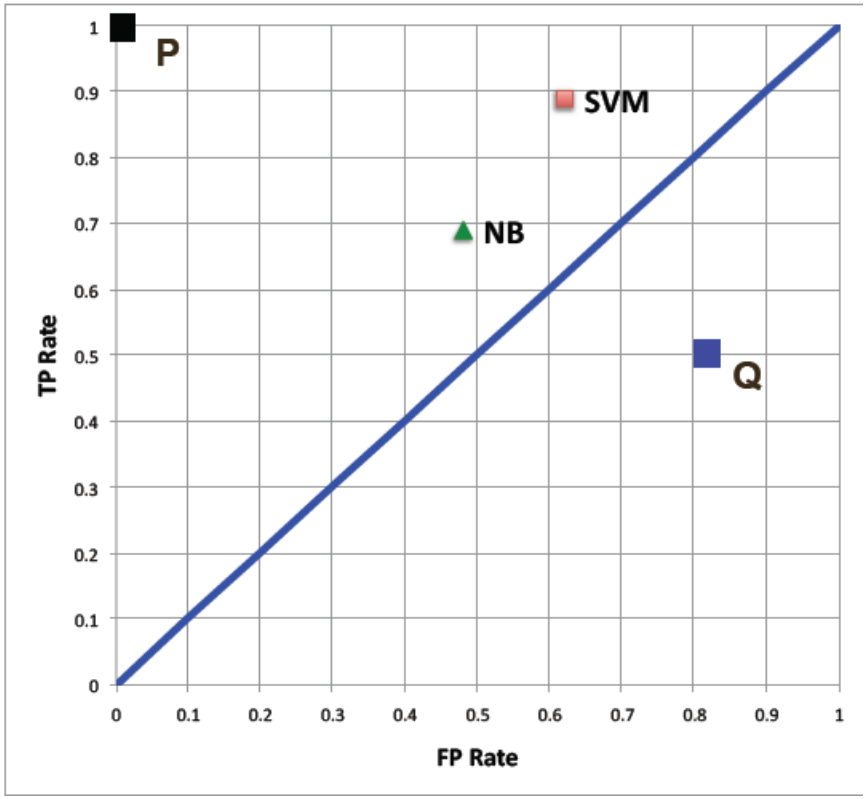
- Erreur de **type 1** (alpha) : ***faux positifs***
  - Probabilité d'accepter l'hypothèse alors qu'elle est fausse

- Erreur de **type 2** (beta) : ***faux négatifs***
  - Probabilité de rejeter l'hypothèse alors qu'elle est vraie

→ Comment **arbitrer** entre ces types d'erreurs ?

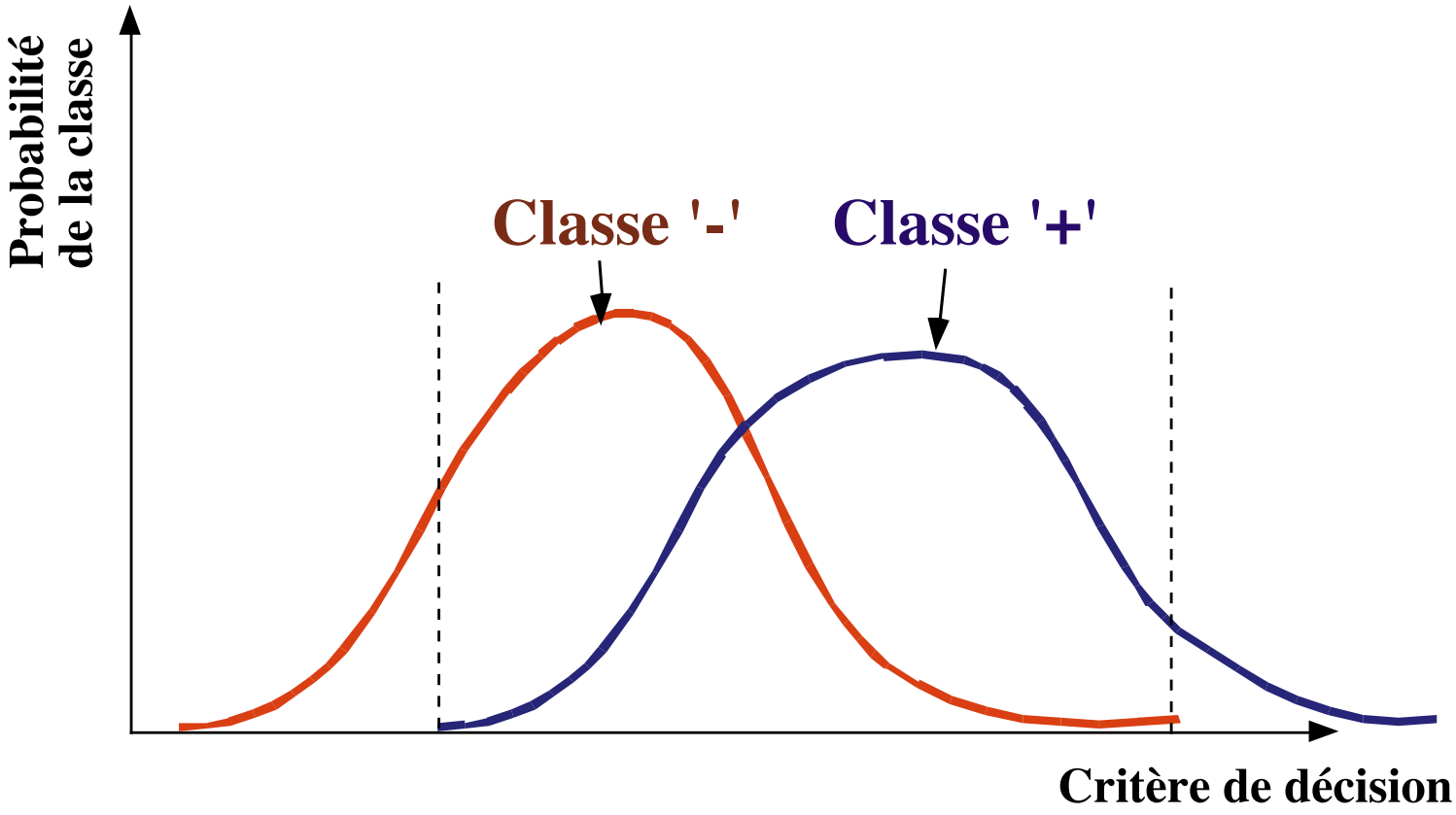
Courbe ROC

	TPR	FPR
SVM	0.89	0.62
NB	0.69	0.48

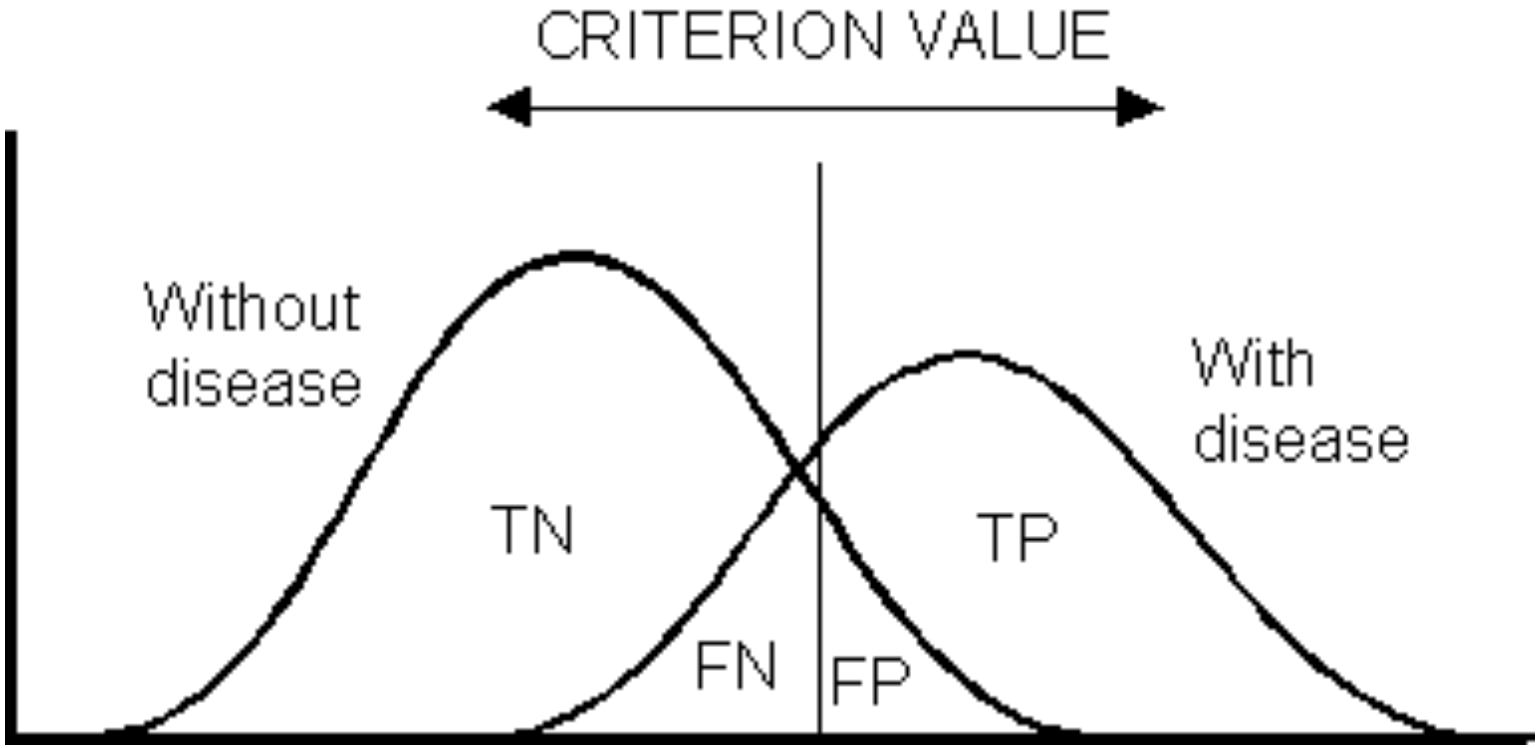


Courbe ROC

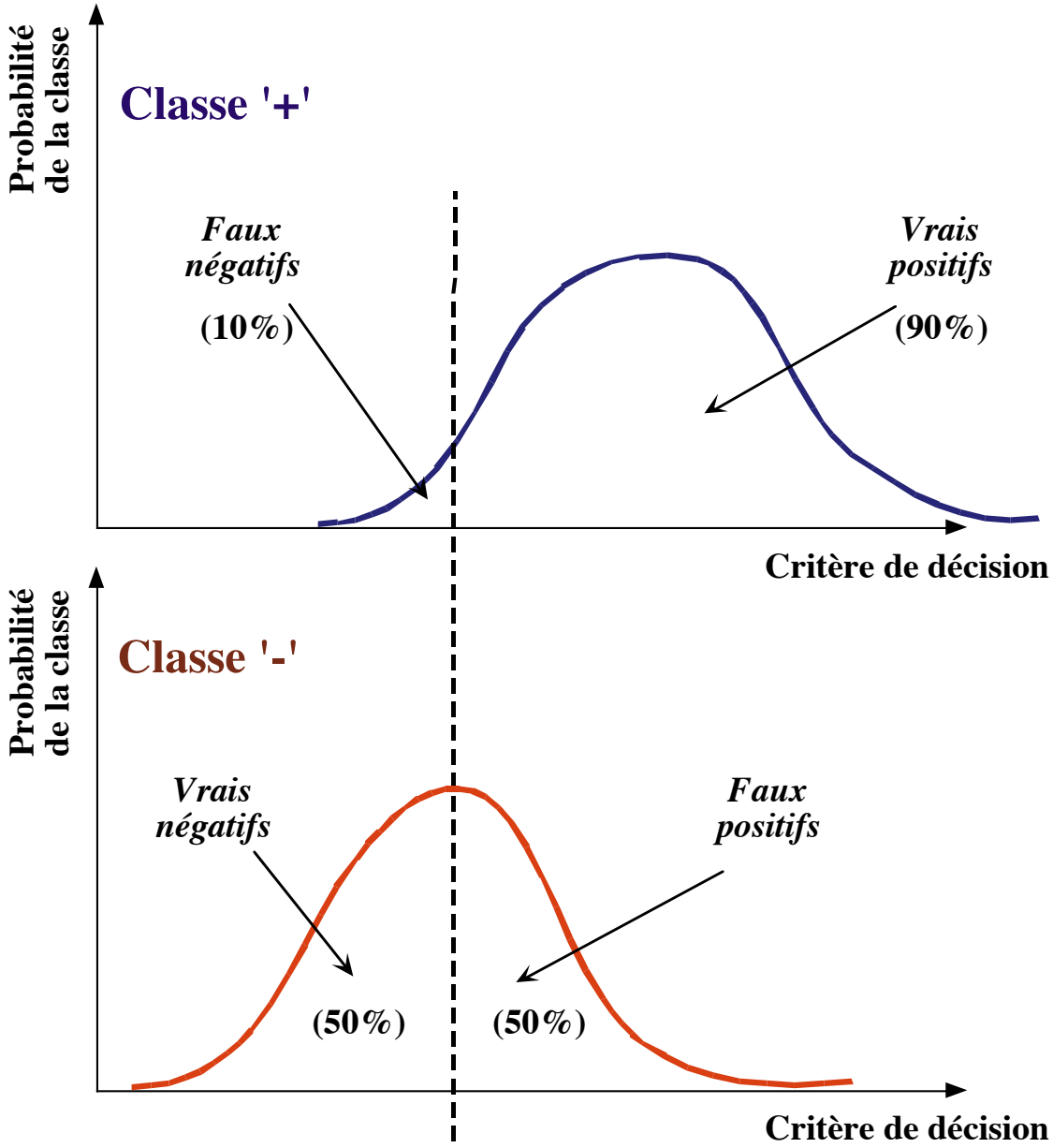
ROC = Receiver Operating Characteristic



Les types d'erreurs



La courbe ROC

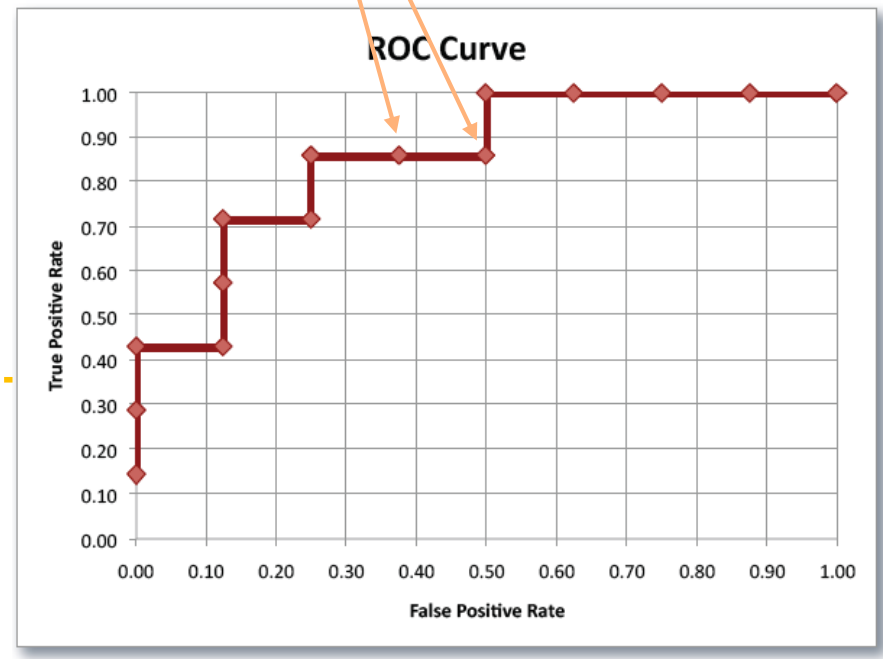


Courbe ROC

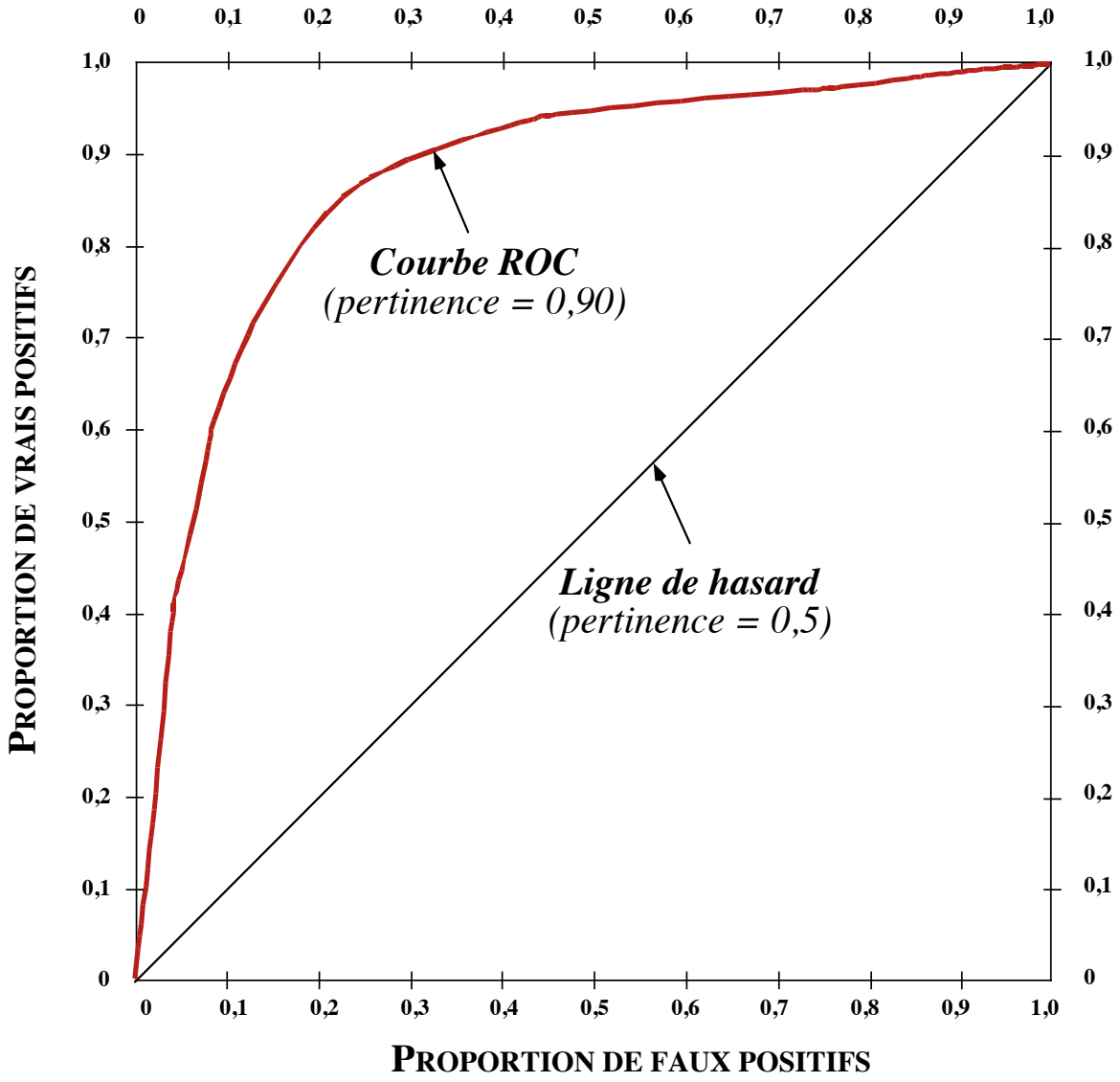
On fait évoluer T de 0 à 1

Score	T=0.5	Label	FP	TP	FPR	TPR
0.99	1	1	0	1	0.00	0.14
0.9	1	1	0	2	0.00	0.29
0.8	1	1	0	3	0.00	0.43
0.85	**1**	0	1	3	0.13	0.43
0.7	1	1	1	4	0.13	0.57
0.7	1	1	1	5	0.13	0.71
0.65	**1**	0	2	5	0.25	0.71
0.6	1	1	2	6	0.25	0.86
0.45	0	0	3	6	0.38	0.86
0.45	0	0	4	6	0.50	0.86
0.4	**0**	1	4	7	0.50	1.00
0.3	0	0	5	7	0.63	1.00
0.2	0	0	6	7	0.75	1.00
0.2	0	0	7	7	0.88	1.00
0.2	0	0	8	7	1.00	1.00

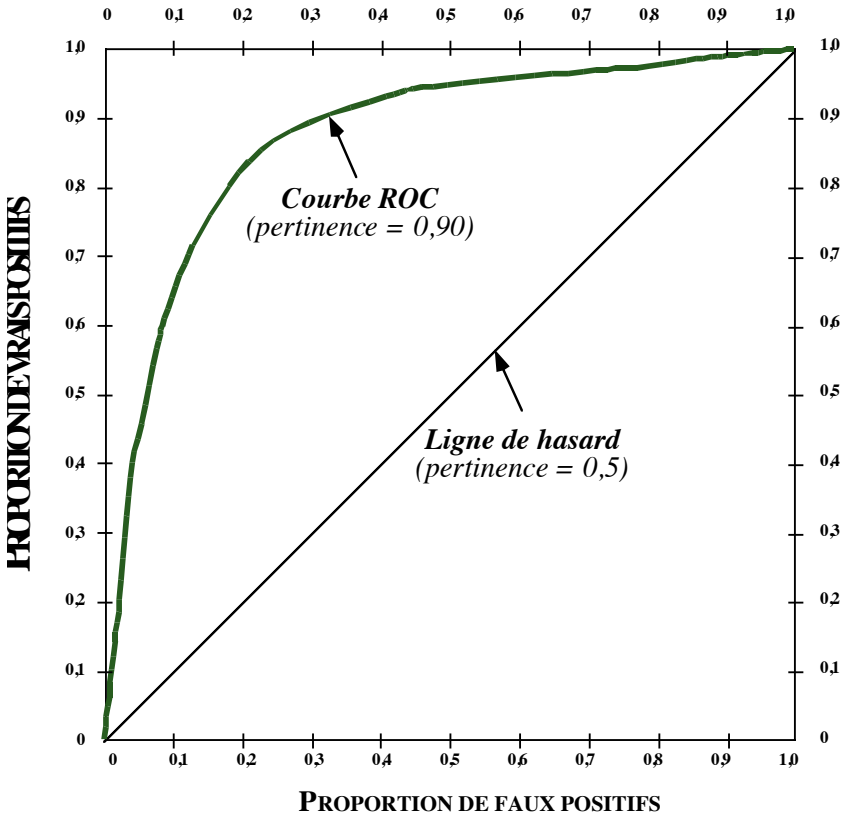
T = 0.5

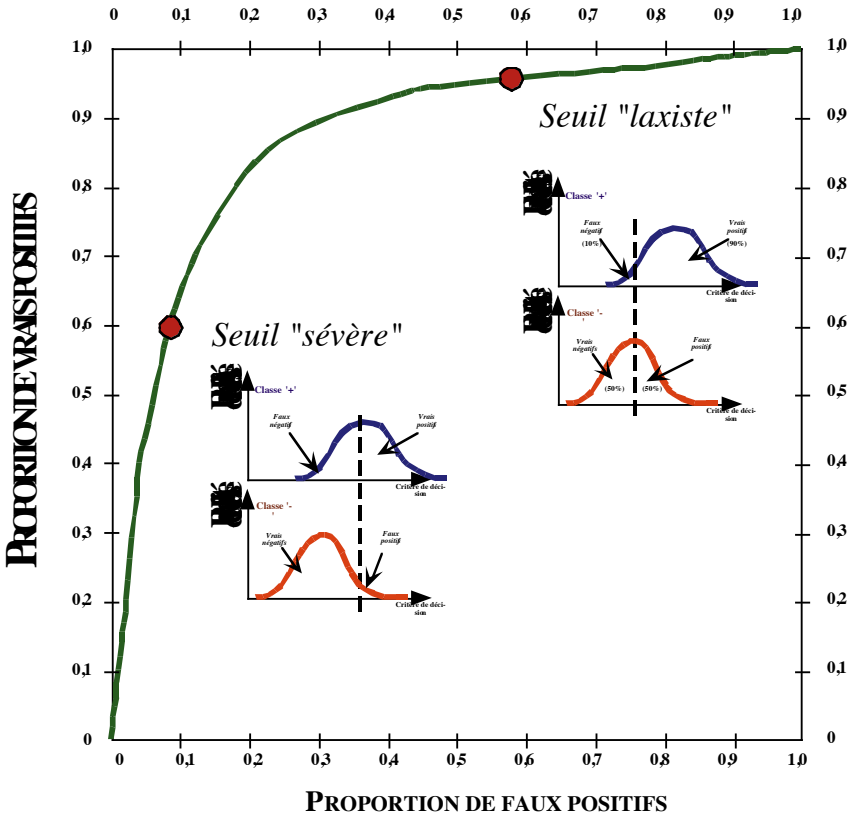


La courbe ROC

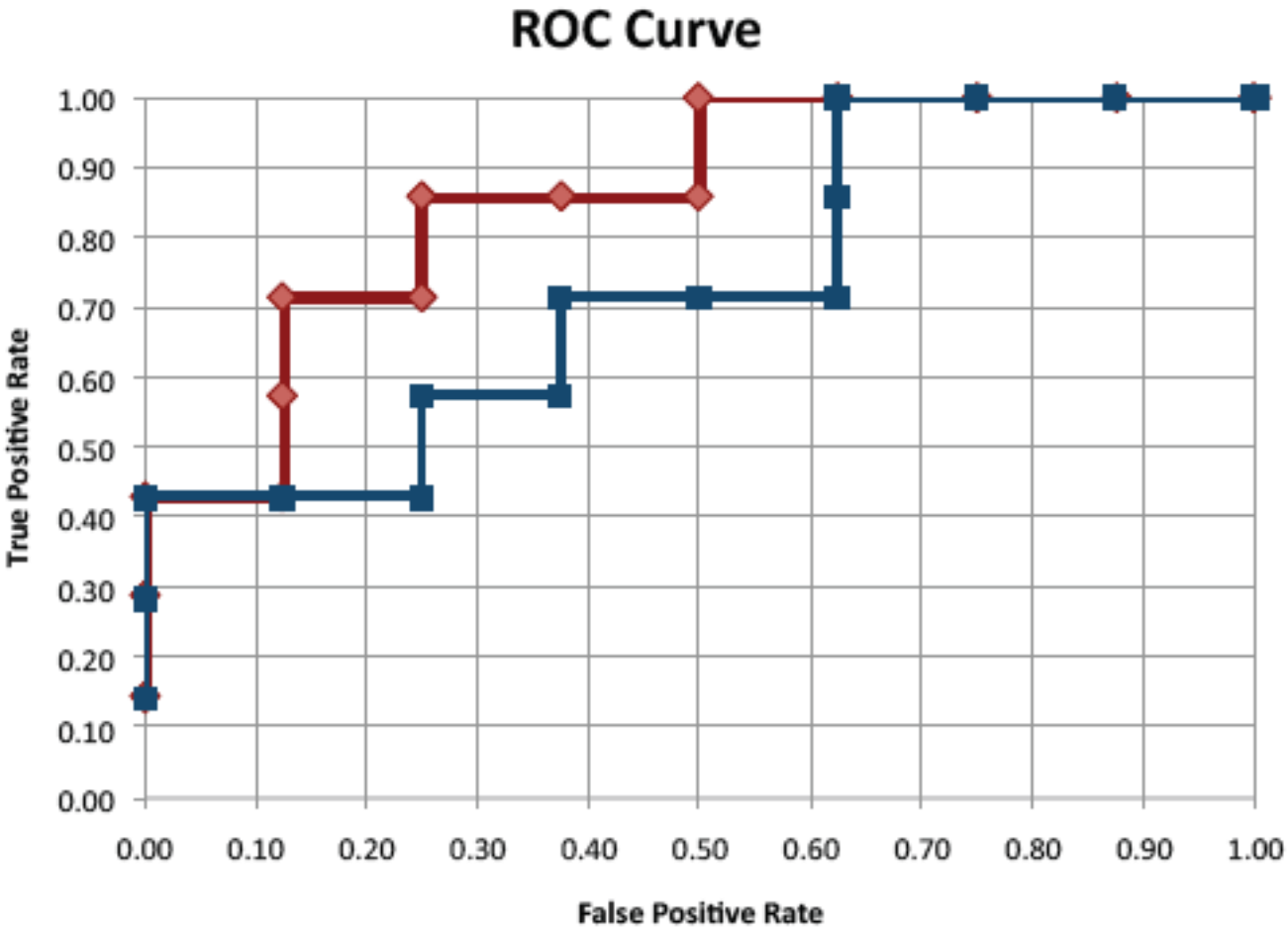


La courbe ROC

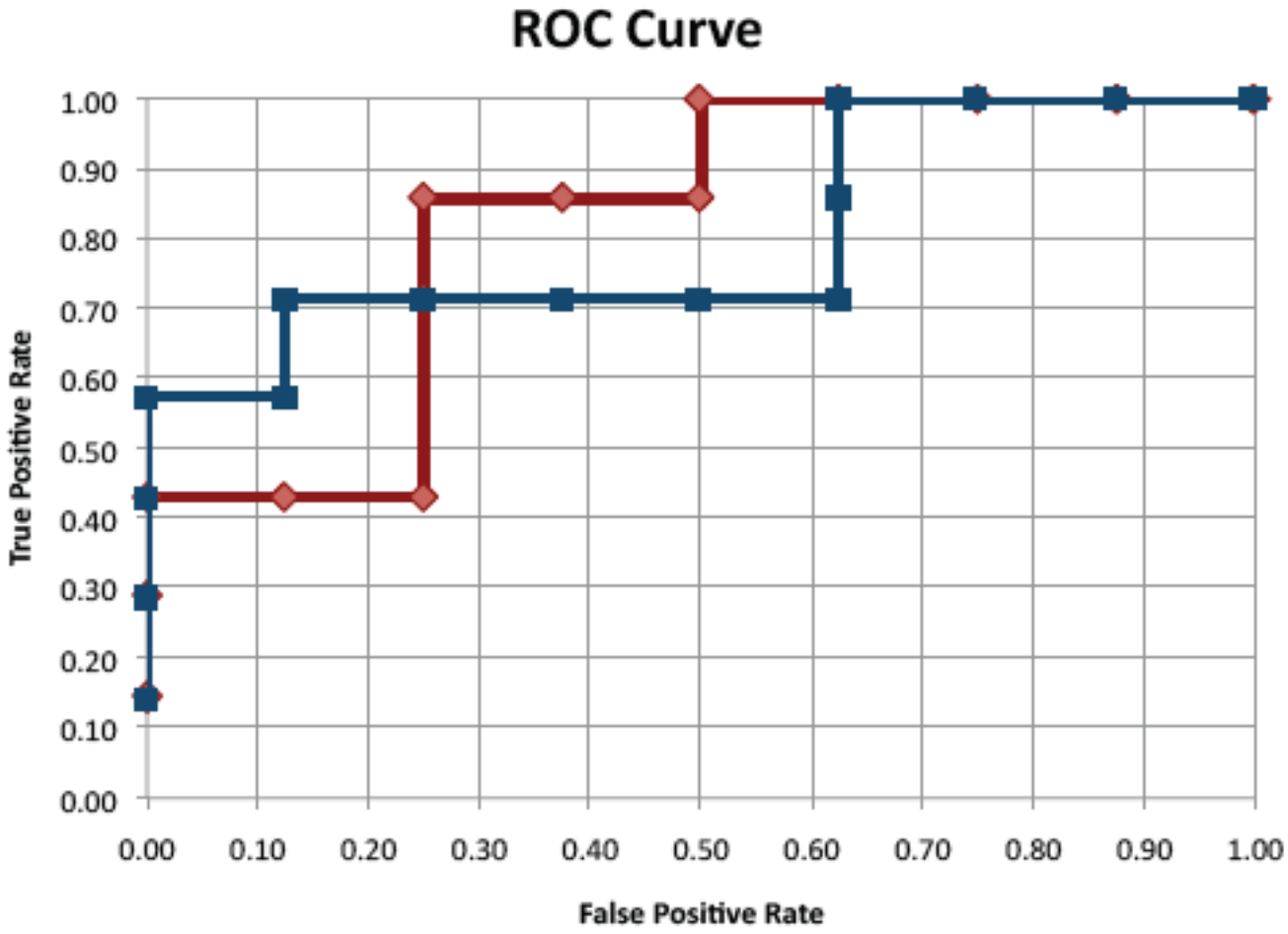




Courbe ROC


ROC Curve
True Positive Rate
False Positive Rate
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

Courbe ROC


ROC Curve
True Positive Rate
False Positive Rate
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

Matrice de confusion

14% des papillons sont pris pour des poissons

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl
Avi	**67%**	2%	-	-	2%	2%	10%	10%	4%	2%	-	-
Pla	-	**21%**	-	2%	7%	19%	10%	12%	5%	-	19%	5%
Uta	17%	-	**33%**	-	7%	-	-	3%	10%	10%	13%	7%
Min	-	-	-	**100%**	-	-	-	-	-	-	-	-
Chi	26%	5%	7%	-	**14%**	9%	12%	9%	5%	2%	12%	-
Poi	5%	13%	3%	8%	-	**13%**	18%	21%	-	3%	10%	8%
Ver	2%	2%	-	-	10%	7%	**43%**	-	21%	5%	7%	2%
Pap	6%	6%	-	-	2%	14%	14%	**35%**	6%	-	12%	4%
Por	2%	2%	-	-	-	2%	-	12%	**70%**	10%	-	2%
Fig	-	-	-	-	-	-	6%	-	24%	**70%**	-	-
Voi	21%	6%	-	-	4%	4%	8%	4%	4%	29%	**19%**	-
Fle	2%	9%	-	-	-	9%	21%	14%	-	-	16%	**28%**

Plan

1. Que mesurer
2. Comment le mesurer
3. La courbe ROC
4. Autres mesures de performances

Autres critères d'évaluation

- *Intelligibilité* **des résultats (hypothèses produites)**
  - **E.g. exit les réseaux de neurones**
- *Performances* **en généralisation**
  - **Pas toujours en adéquation totale avec le point précédent**
- *Coûts*
  - **de préparation** (des données)
  - **coût computationnel** (e.g. coût d'une passe et nombre de passes nécessaires, …)
  - **coût de l'expertise en apprentissage**
  - **coût de l'expertise sur le domaine**

Résumé

- Attention à votre fonction de coût :
  - qu'est-ce qui importe pour la mesure de performance ?
- Données en nombre fini:
  - calculez les intervalles de confiance
- Données rares :
  - Attention à la répartition entre données d'apprentissage et données test. Validation croisée.
- N'oubliez pas l'ensemble de test

- **L'évaluation est très importante**
  - Ayez l'esprit critique
  - Convainquez-vous vous même !

Références

- Nathalie Japkowicz & Mohak Shah (2011) Evaluating Learning Algorithms. A classification perspective. Cambridge University Press.