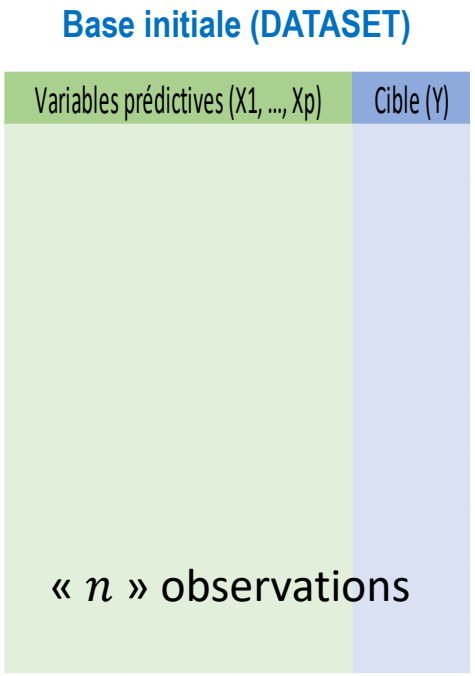
Classement, classification supervisée



Comment construire un modèle prédictif $M(n\,;\alpha)$
(ex. régression logistique, arbre de décision, SVM, etc.)

Comment en évaluer les performances (ex. taux d'erreur)

Evaluation en « resubstitution »

Base initiale (DATASET)

Variables prédictives (X1, ..., Xp)	Cible (Y)
« n » observations	

Données

Modélisation

Entraîner le modèle sur l'ensemble des observations disponibles

$$M(n\ ;\ \hat{\alpha})$$

Prédiction sur l'échantillon même qui a servi à la construction du modèle

Cible (Y)	Y^

Mesures de performances

Matrice de confusion

Y		Y^ +^	Y^ -^
	+	a	b
	-	c	d

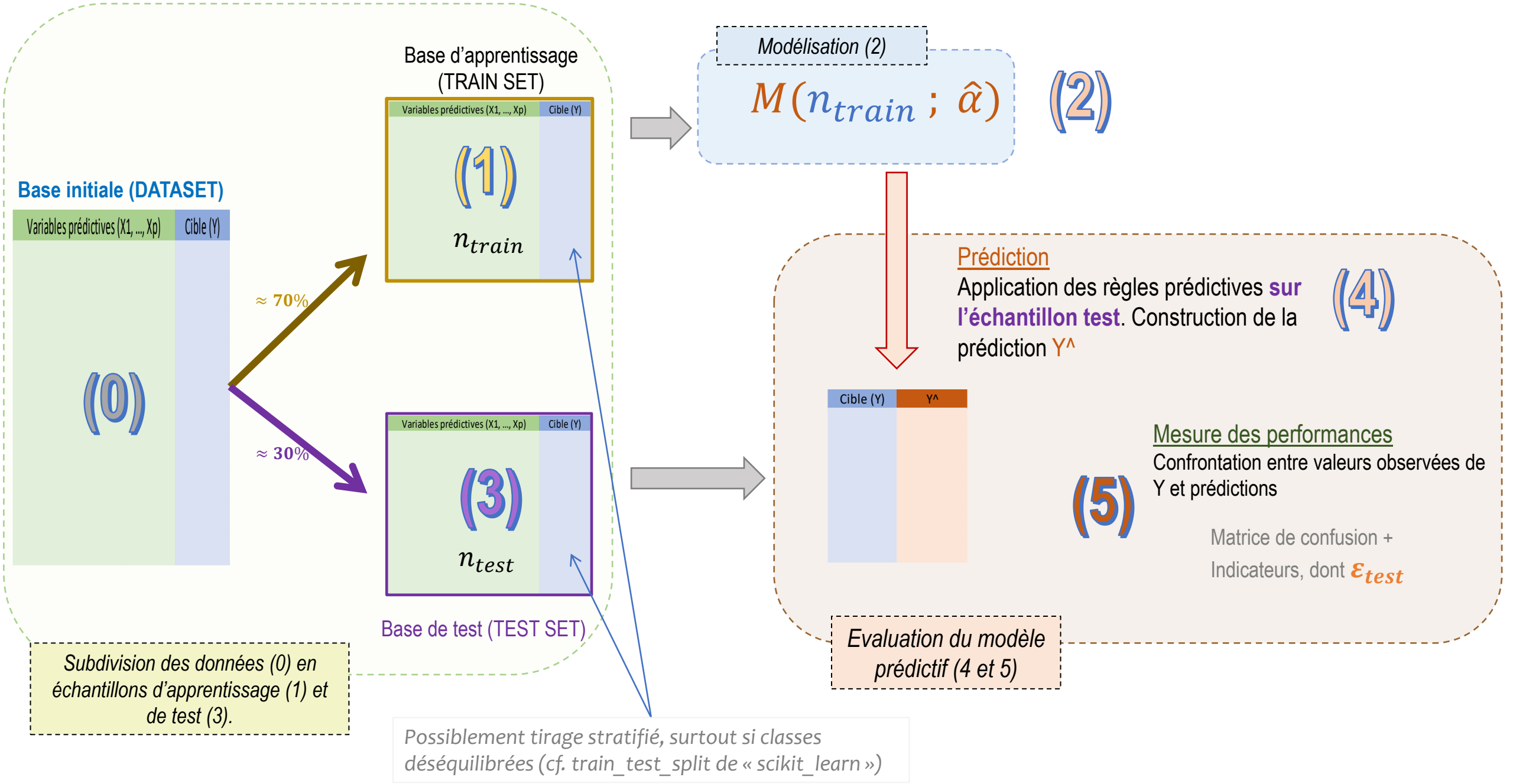
Indicateurs

- Taux d'erreur -- ε_{resub}
- Rappel (sensibilité)
- Précision

! Les mêmes données sont juges et parties, les estimations des performances sont faussées (trop optimistes)

Evaluation « faussée » du modèle prédictif

Schéma « holdout » (apprentissage – test)



Validation croisée (K-Fold cross-validation)

Base initiale (DATASET)

Variables prédictives (X1, ..., Xp)	Cible (Y)
« n » observations	

Données

Modélisation

Entraîner le modèle sur l'ensemble des observations disponibles

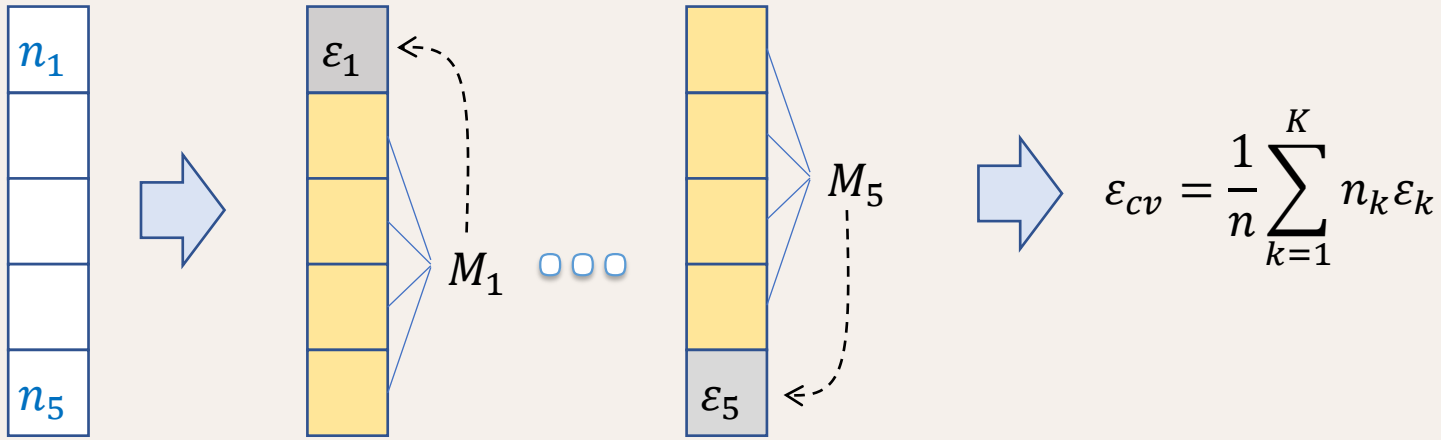
$M(n\ ;\ \hat{\alpha})$

* ε_{cv} est un estimateur des performances du modèle construit sur la totalité des données

* Les modèles M_k ne servent plus par la suite

Subdivision en « K » blocs des données

Succession de « K » apprentissage-test



Validation croisée (*la procédure peut être répétée pour plus de précision*)

* « K » est un paramètre (ex. K = 5)

* Possiblement tirage stratifié

Leave-one-out (« K = n » cross-validation)

Base initiale (DATASET)

Variables prédictives (X1, ..., Xp)	Cible (Y)
« n » observations	

Données

Modélisation

Entraîner le modèle sur l'ensemble des observations disponibles

$M(n\ ;\ \hat{\alpha})$

* ε_{loo} *est un estimateur des performances du modèle construit sur la totalité des données*

* *Les modèles* M_i *ne servent plus par la suite*

Succession de « n » apprentissage-test

(1) Modélisation sur les (n-1) dernières observations, classement du 1er → $e_1 \in \{0, 1\}$

...

...

(n) Modélisation sur les (n-1) premières observations, classement du nème → e_n

$$\varepsilon_{loo} = \frac{1}{n}\sum_{i=1}^{n} e_i$$

Leave-one-out (un seul essai possible)