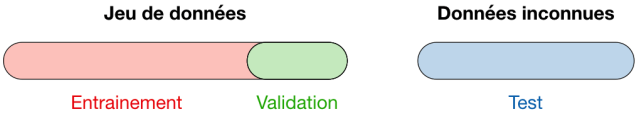
Training set	Validation set	Testing set
- Modèle est entrainé - Normalement 80% du dataset	- Modèle est évalué - Normalement 20% du dataset - Aussi appelé hold-out ou development set	- Modèle donne des prédictions - Données jamais vues

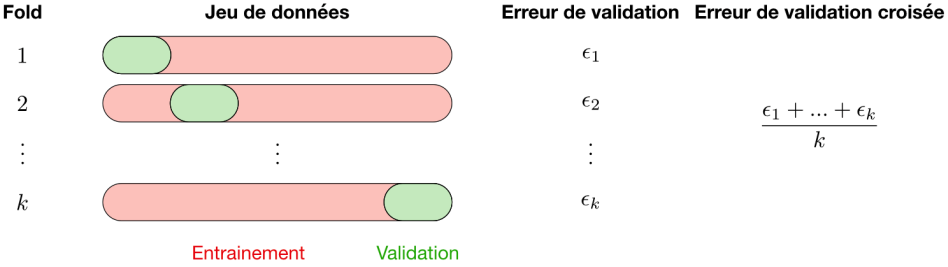
Une fois que le modèle a été choisi, il est entrainé sur le jeu de données entier et testé sur test set (qui n'a jamais été vu). Ces derniers sont représentés dans la figure ci-dessous :



❒ **Validation croisée** – La validation croisée, aussi notée CV (de l'anglais *Cross-Validation*), est une méthode qui est utilisée pour sélectionner un modèle qui ne s'appuie pas trop sur le training set de départ. Les différents types de validation croisée rencontrés sont resumés dans le tableau ci-dessous :

k-fold	Leave-p-out
- Entrainement sur $k-1$ folds et évaluation sur le fold restant - Généralement $k = 5$ ou 10	- Entrainement sur $n-p$ observations et évaluation sur les p restantes - Cas $p = 1$ est appelé *leave-one-out*

La méthode la plus utilisée est appelée validation croisée k-fold et partage le jeu de données d'entrainement en k folds, de manière à valider le modèle sur un fold tout en trainant le modèle sur les $k-1$ autres folds, tout ceci k fois. L'erreur est alors moyennée sur k folds et est appelée erreur de validation croisée.



❒ **Régularisation** – La procédure de régularisation a pour but d'éviter que le modèle ne surapprenne (en anglais *overfit*) les données et ainsi vise à régler les problèmes de grande variance. Le tableau suivant récapitule les différentes techniques de régularisation communément utilisées.