Train In Data

Cross-Validation

Hyperparameter Tuning: Search

A search consist of:

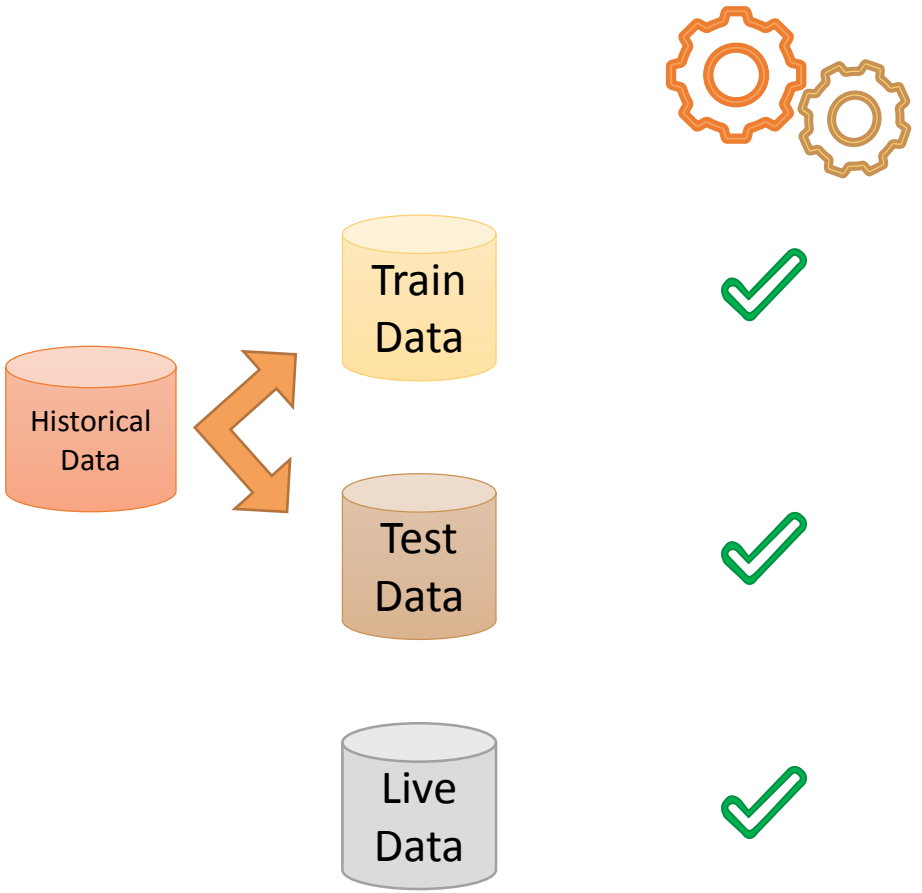
- Hyperparameter space **Parameter Grid**
- A method for sampling candidate hyperparameters
- A cross-validation scheme
- A performance metric to minimize (or maximize)

Hyperparameter Tuning: Search

A search consist of:

- Hyperparameter space **Parameter Grid**
- A method for sampling candidate hyperparameters
- **A cross-validation scheme**
- A performance metric to minimize (or maximize)

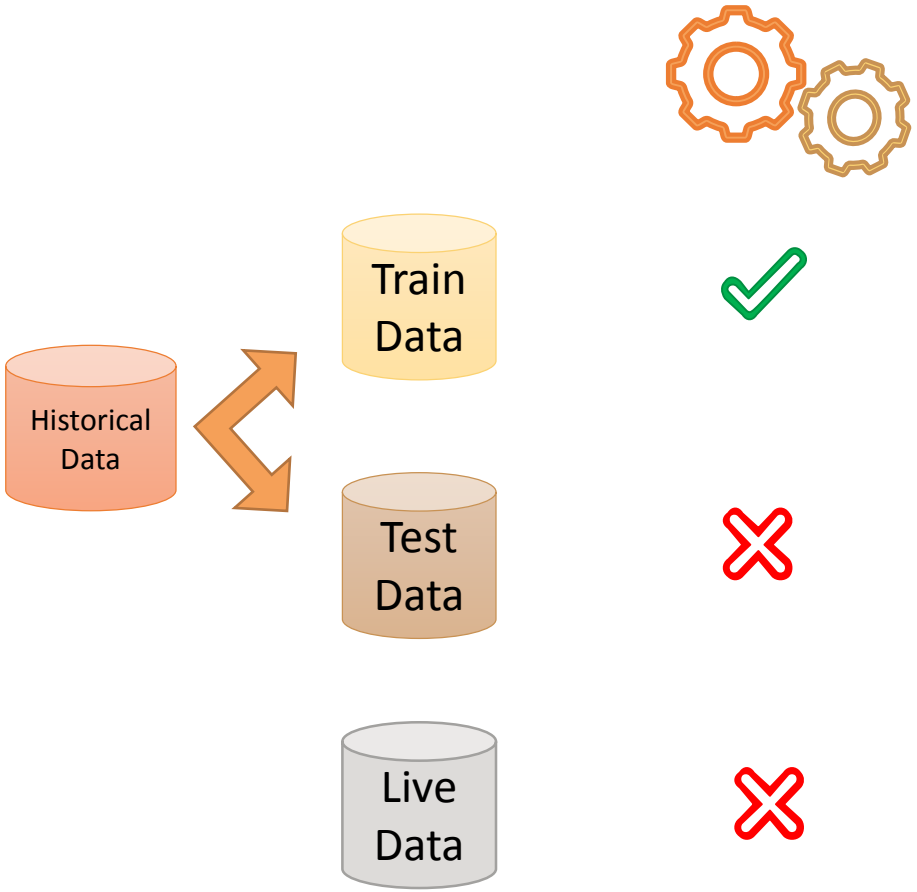
Generalization vs Over-fitting



Generalization is the ability of an algorithm to be effective across various inputs.

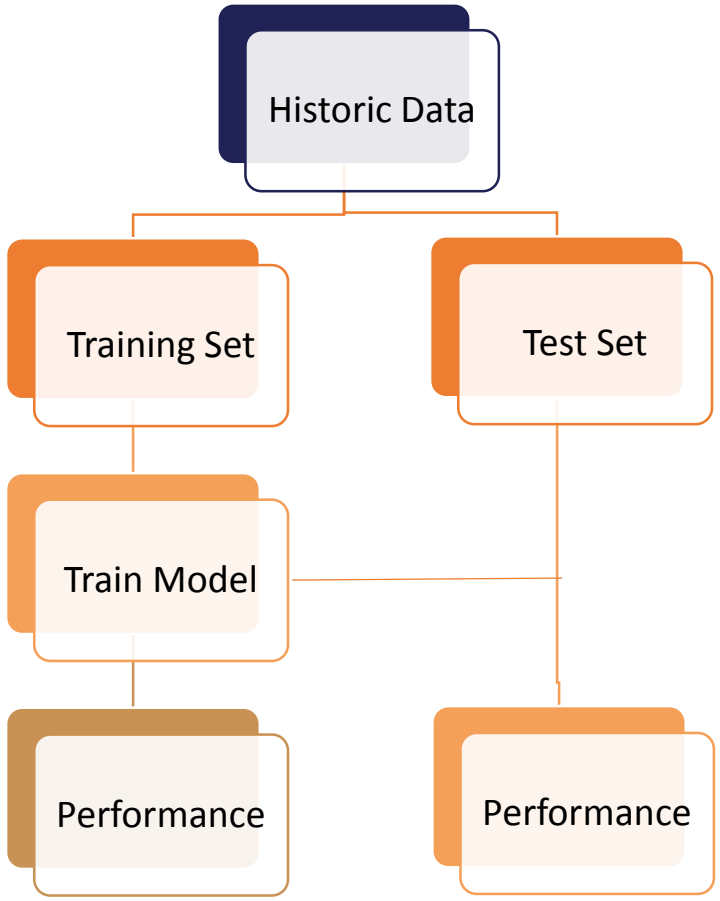
The performance of the machine learning model is constant across different datasets (with the same distribution of the training data).

Generalization vs Over-fitting



When a model performs well on the train set, but not on new / naïve data, the model over-fits to the training data

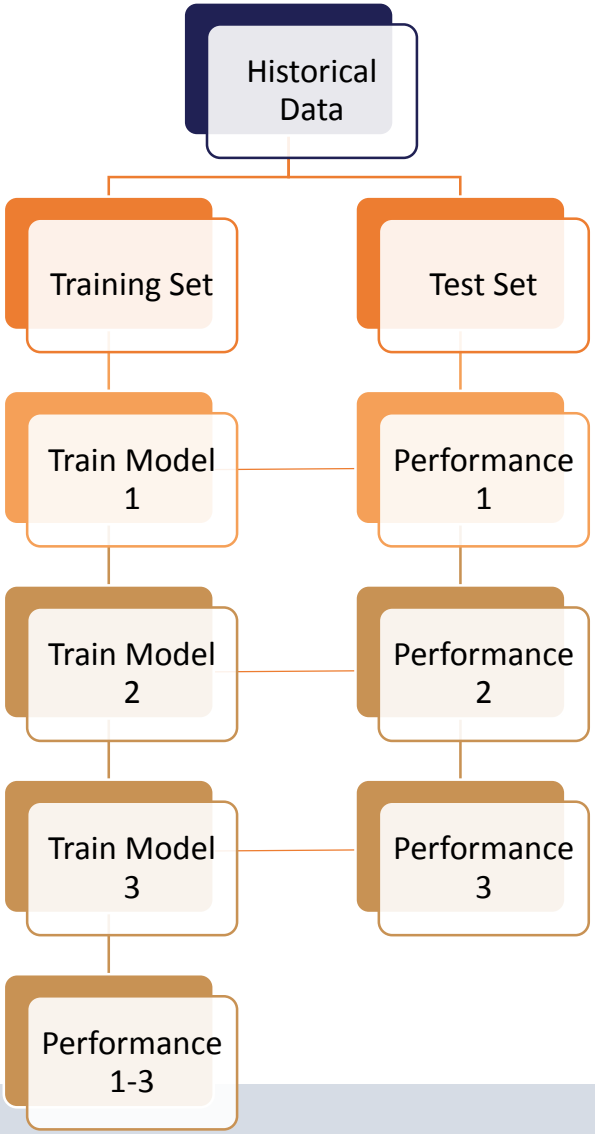
Training a Machine Learning Model



To prevent over-fitting, it is common practice to:

- Separate the data into a train and a test set.
- Train the model in the train set
- Evaluate in the test set

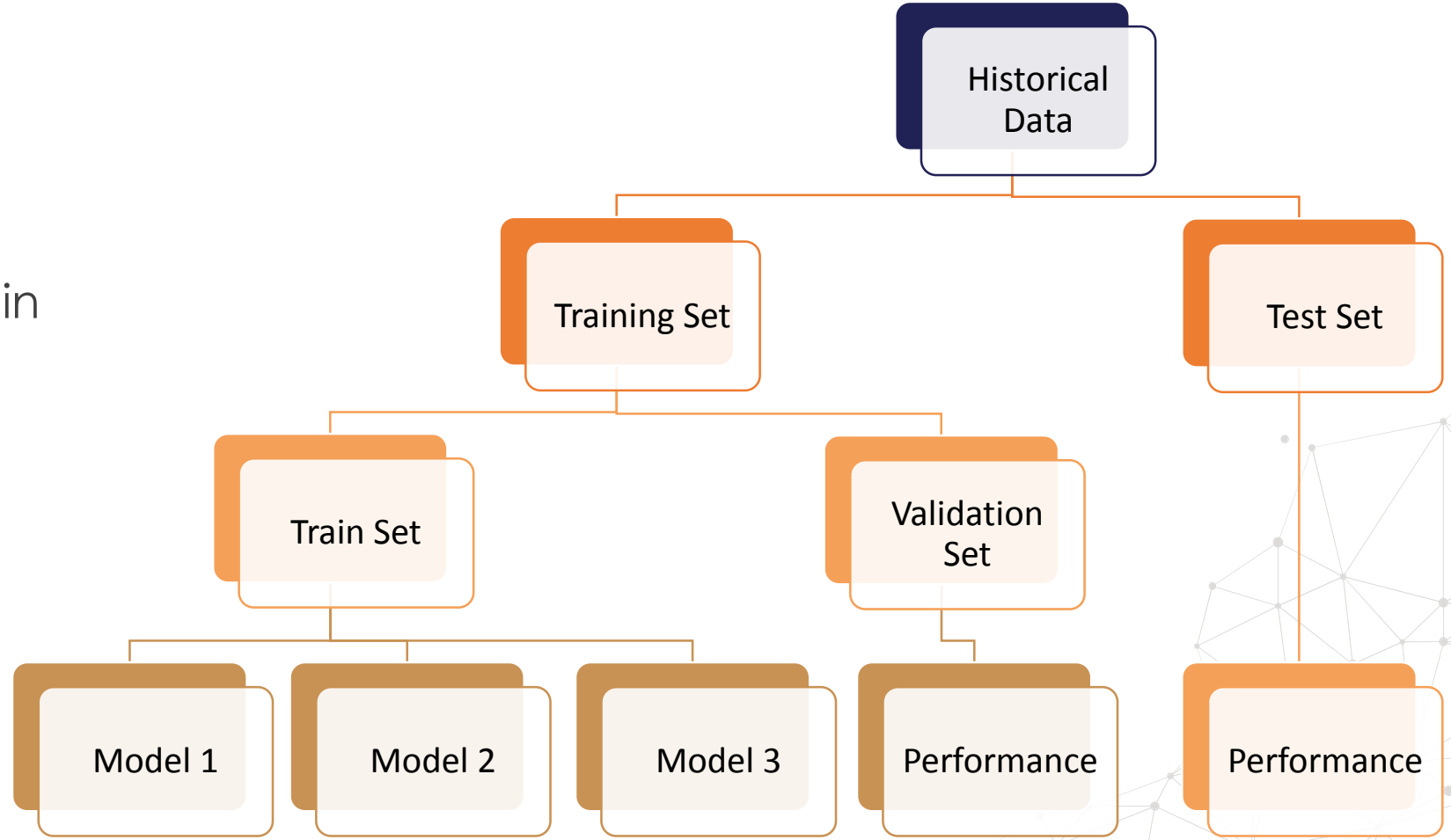
Tuning Hyperparameters



- When evaluating different hyperparameter spaces there is a risk of overfitting on the test.
- We select the best model based on performance over test set
- Knowledge about the test set can "leak" into the model ➔ lack or unknown generalization.
- Common mistake in Data Science Competitions

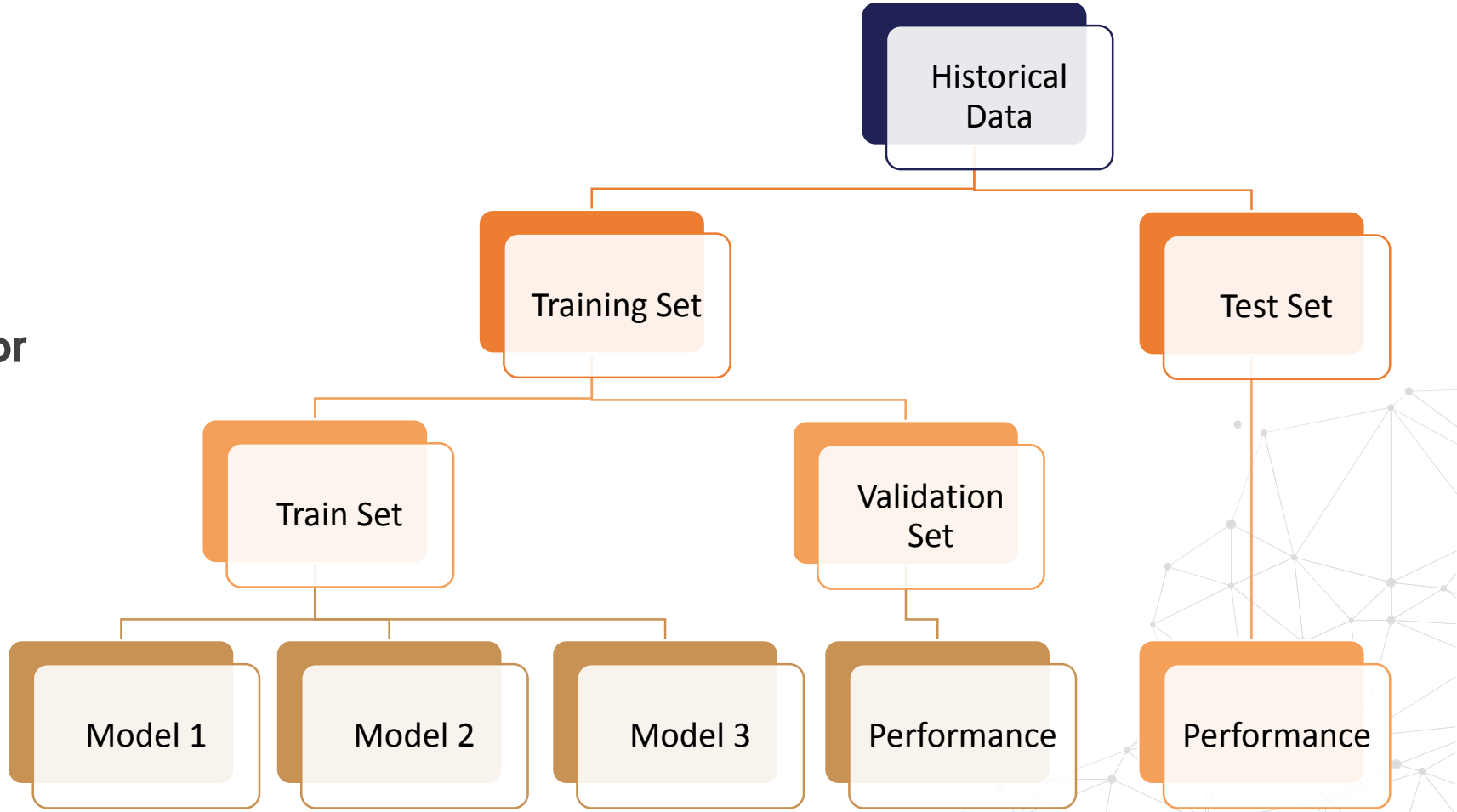
Another Hold-Out Sample

- Subsequently divide the train set in a train set and validation set
- Train model on most of train set
- Test Performance on validation Set
- Select best model
- Test best model's performance on test set



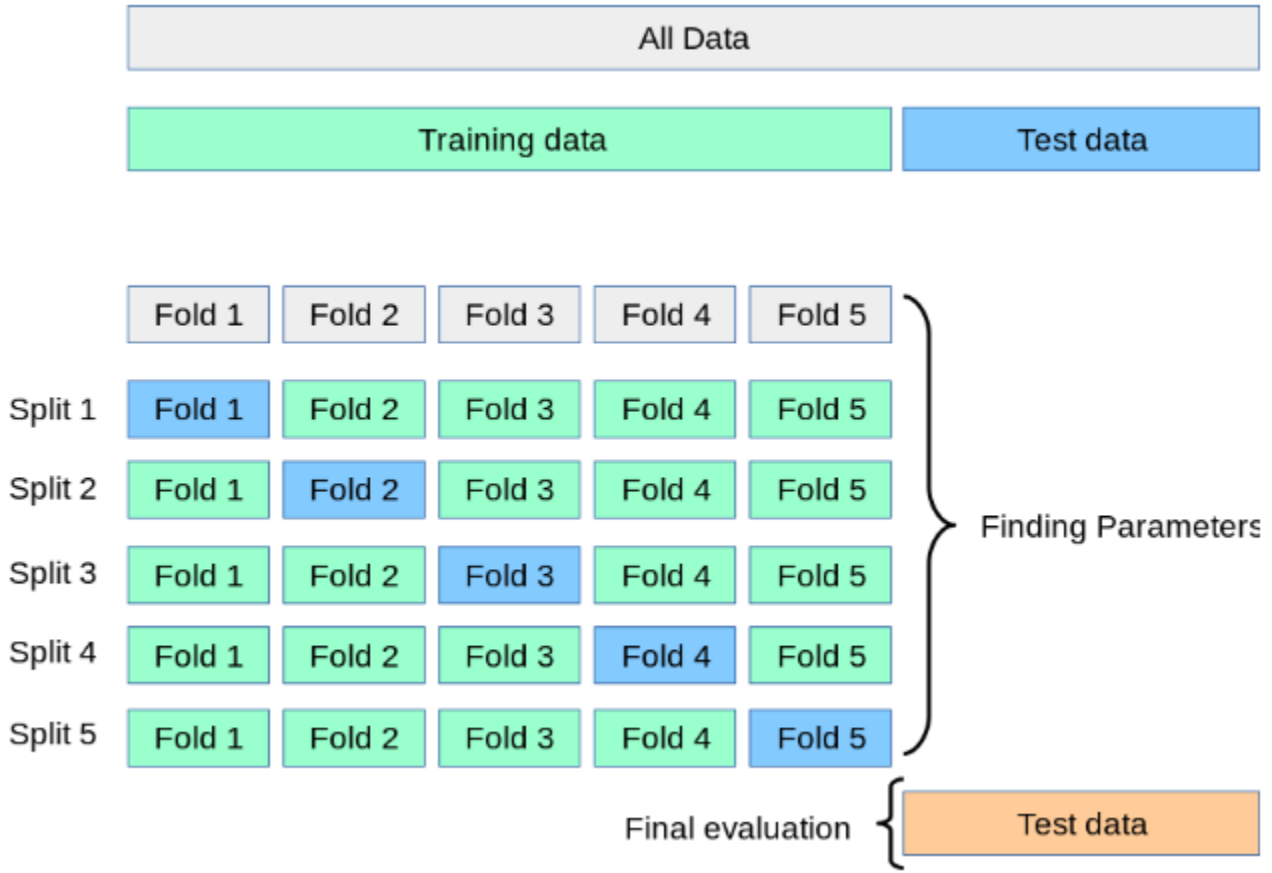
Another Hold-Out Sample

- We could be left out with very little data to train the model
- We have no metric of **error**
- **Metric ± error**



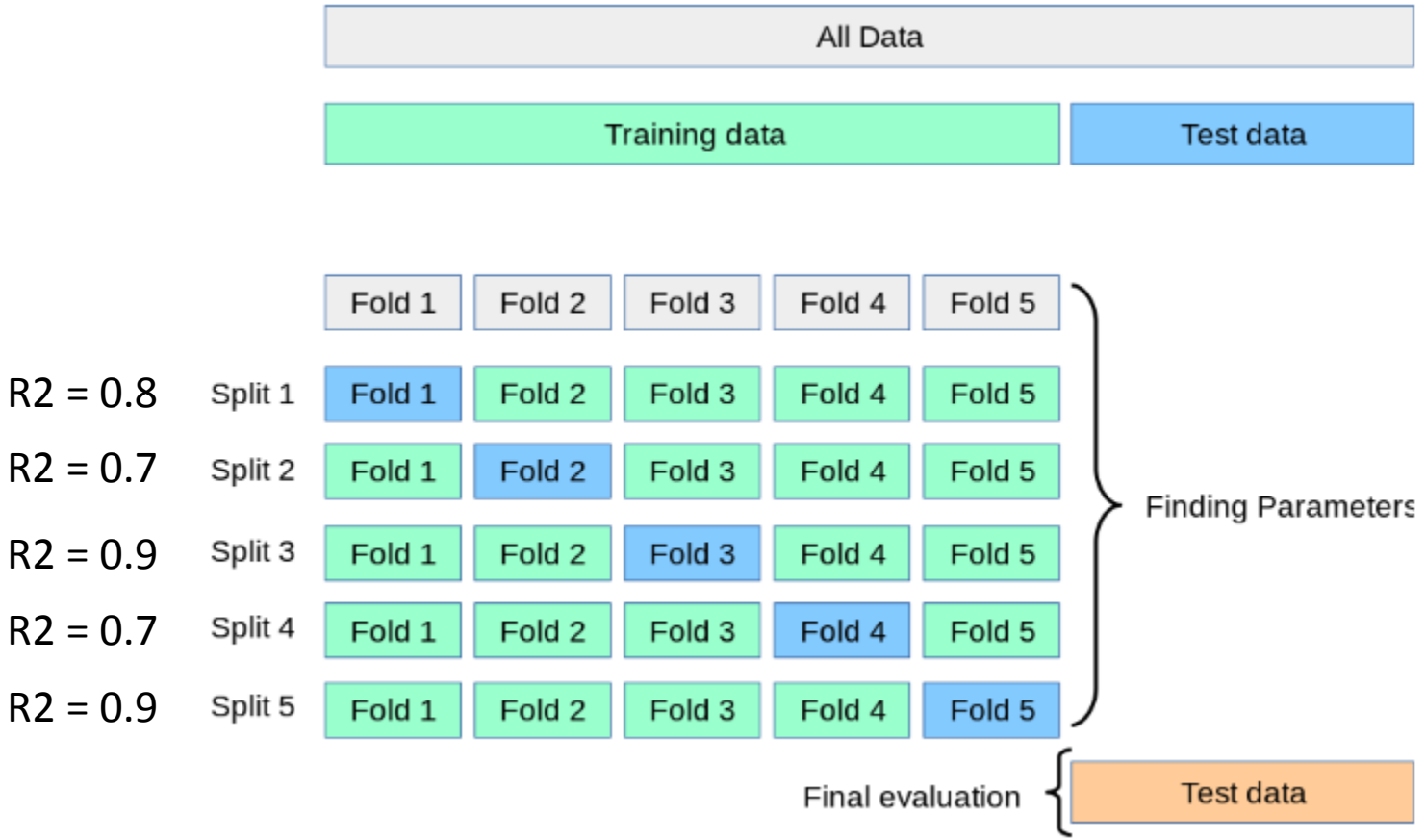
Cross-Validation

- Train set divided into k folds
- Model trained in k-1 fold
- Model tested in the k^{th} fold
- Repeat k times
- Final performance metric is the average
- Can determine an error



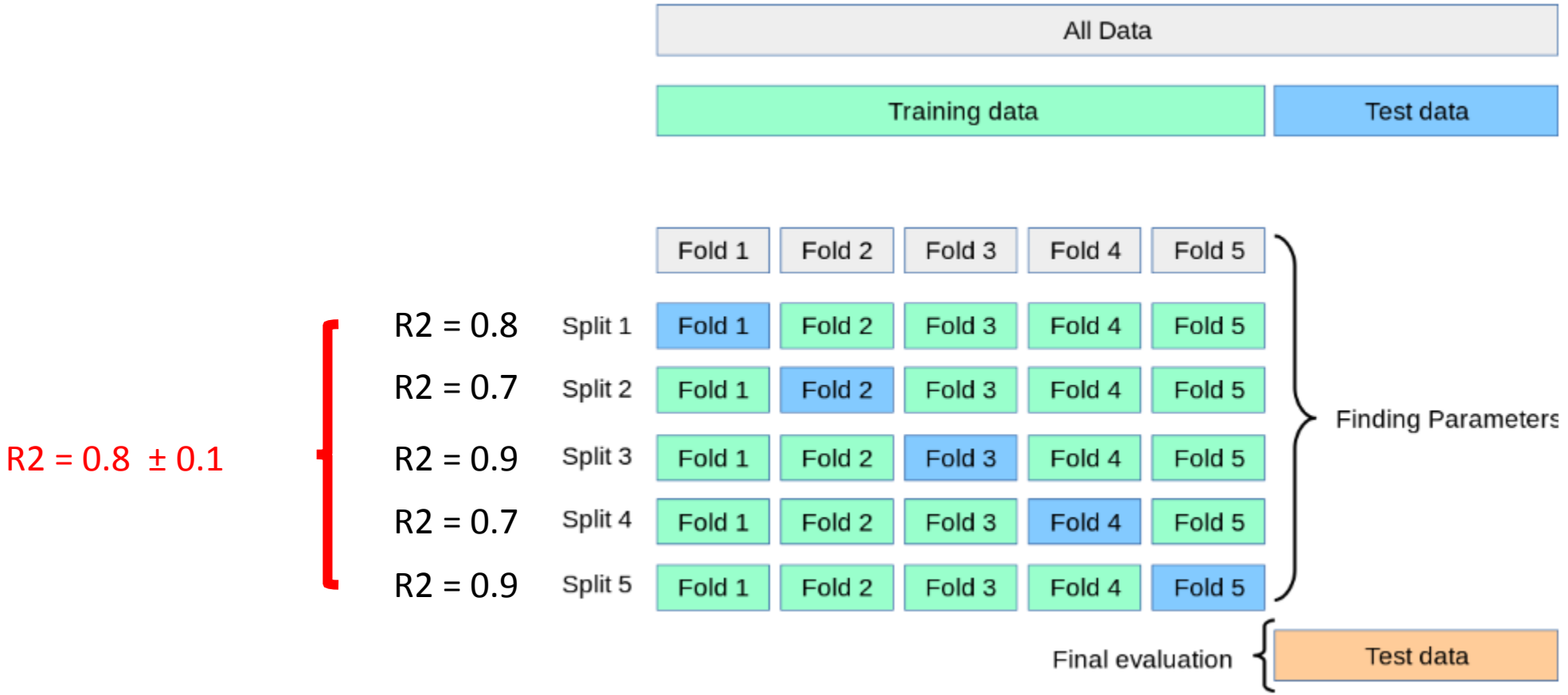
https://scikit-learn.org/stable/modules/cross_validation.html

Cross-Validation



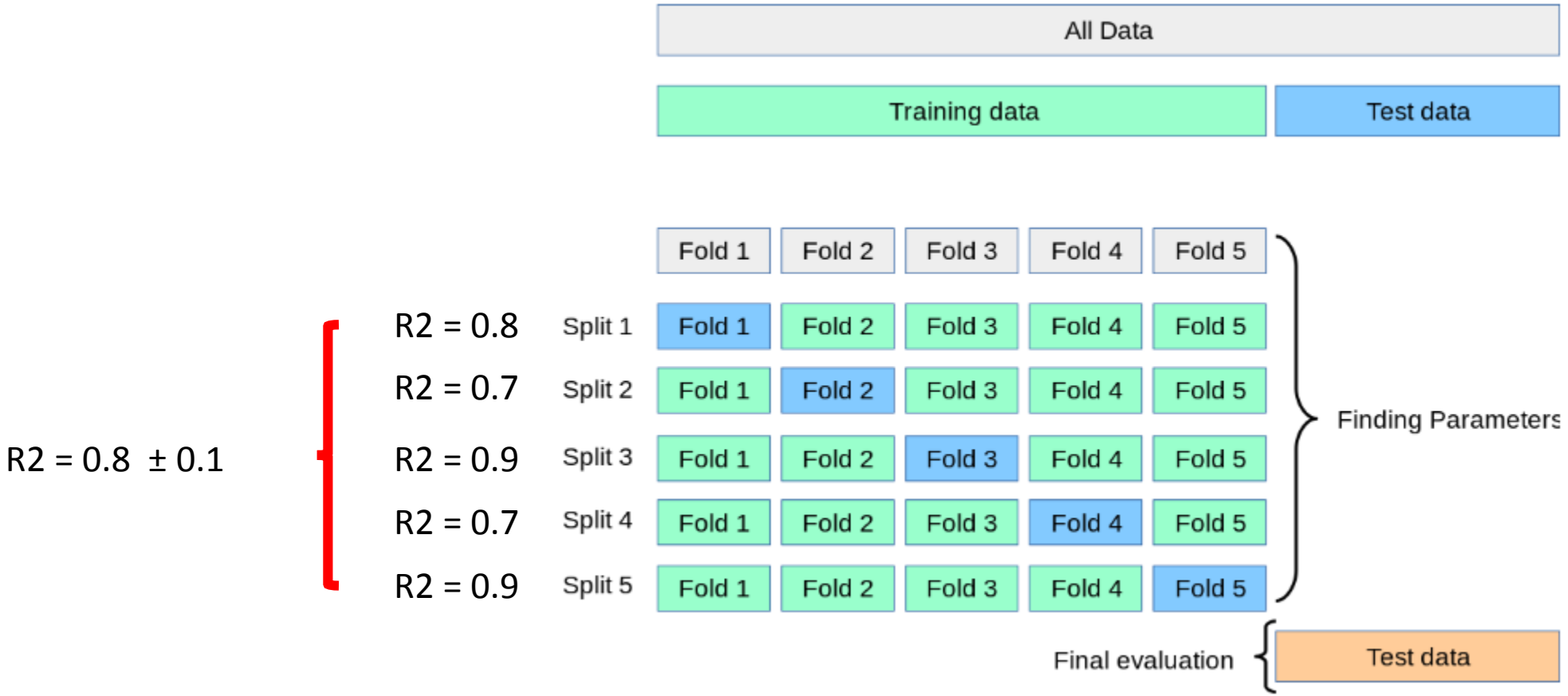
https://scikit-learn.org/stable/modules/cross_validation.html

Cross-Validation



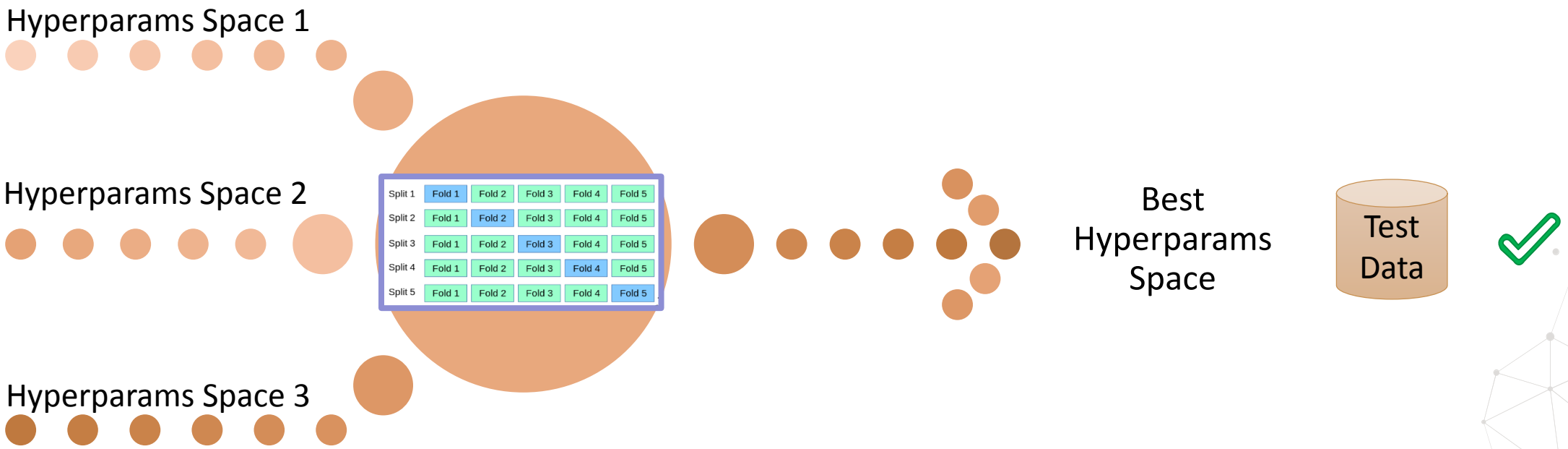
https://scikit-learn.org/stable/modules/cross_validation.html

Cross-Validation



https://scikit-learn.org/stable/modules/cross_validation.html

Hyperparams tuning with Cross-Val



Cross-Validation

- K-Fold
- Leave One Out (LOOCV)
- Leave P Out (LPOCV)
- Repeated K-Fold
- Stratified Cross-Validation
- Group Cross-Validation
- Nested Cross-Validation

Train In Data

THANK YOU

www.trainindata.com