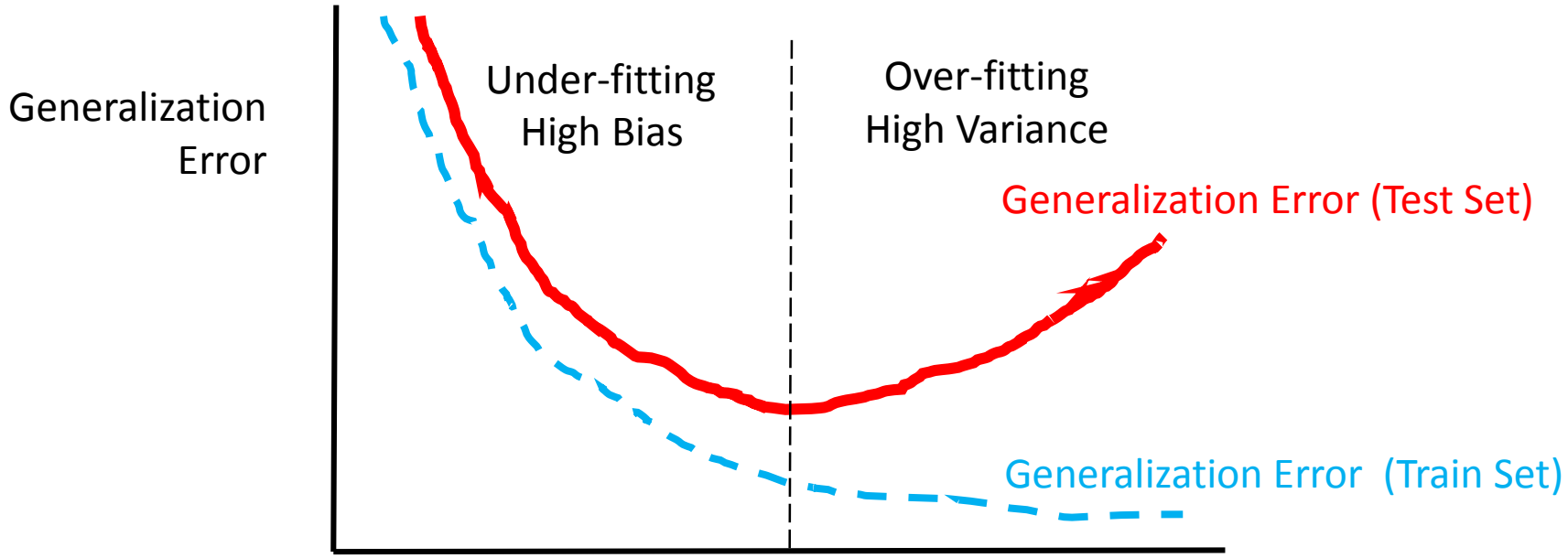
Train In Data

Cross-Validation Scheme

Cross-Validation Schemes

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

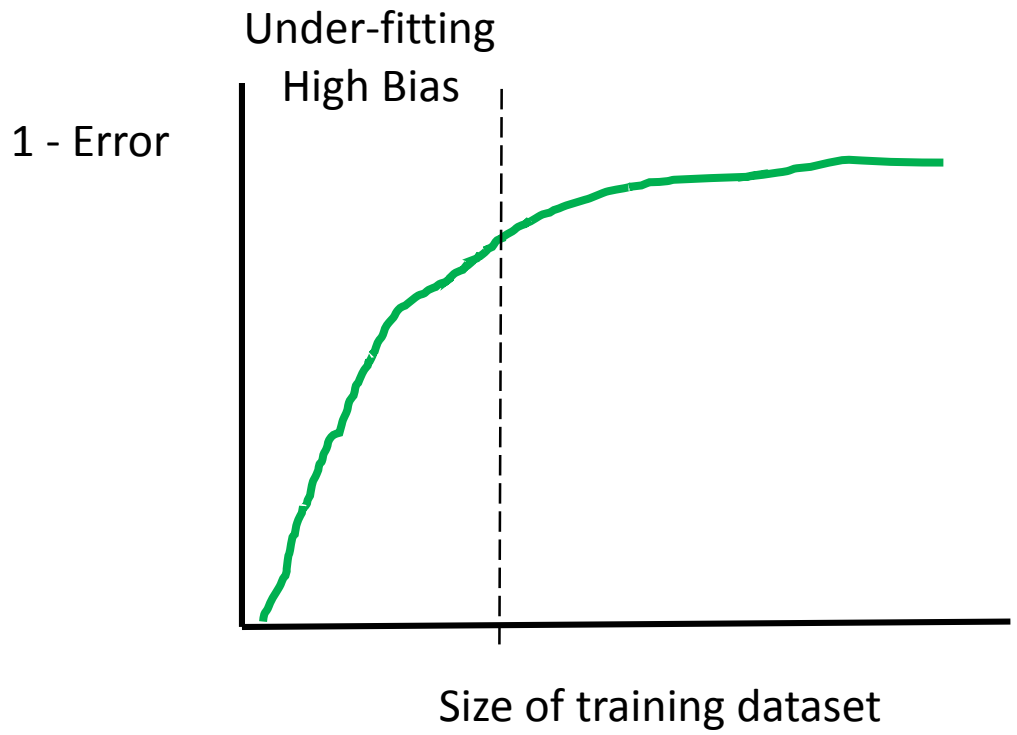
Bias vs Variance



Complexity of the model

- e.g., linear vs polynomial model
- Number of estimators in tree based algorithms, depth, etc.

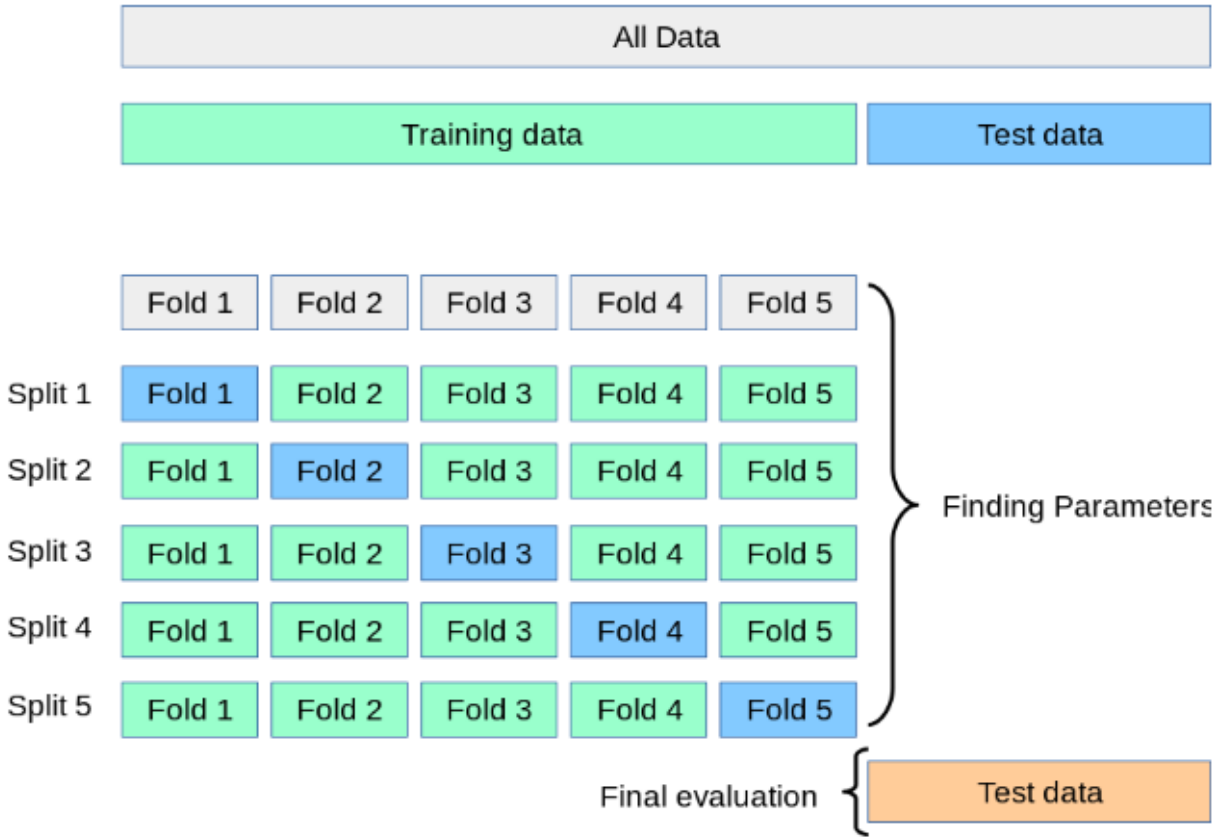
Train set size vs Bias (performance)



Smaller datasets may lead to under-fitted (highly biased) models

K-Fold Cross-Validation

- Divide Train set into k folds (of equal size)
- Train model in k-1 fold
- Test model in k^{th} fold
- Repeat k times ➔ train k models
- K performance values
- Final performance metric:
  - mean ± std



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

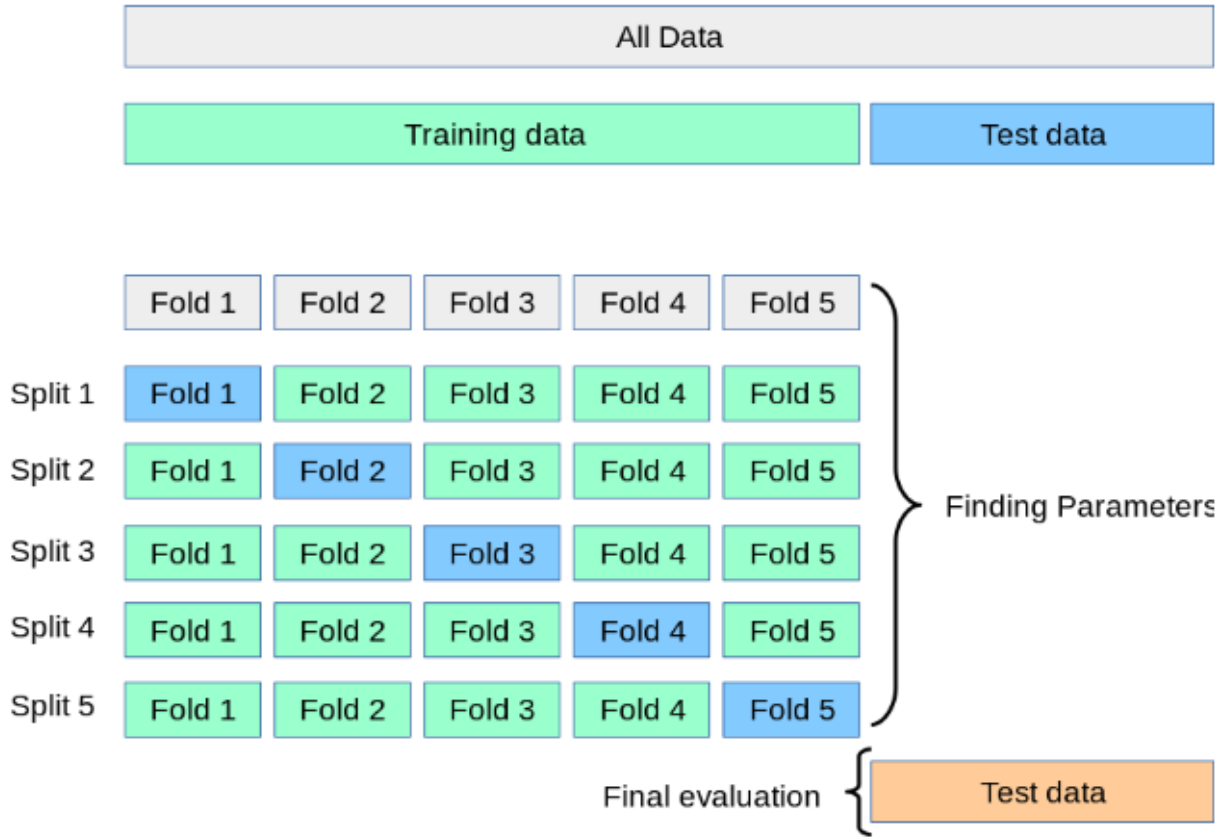
K-Fold Cross-Validation

Typical K is 5 or 10

Higher K:

- bigger train sets
- less model bias
- more variance

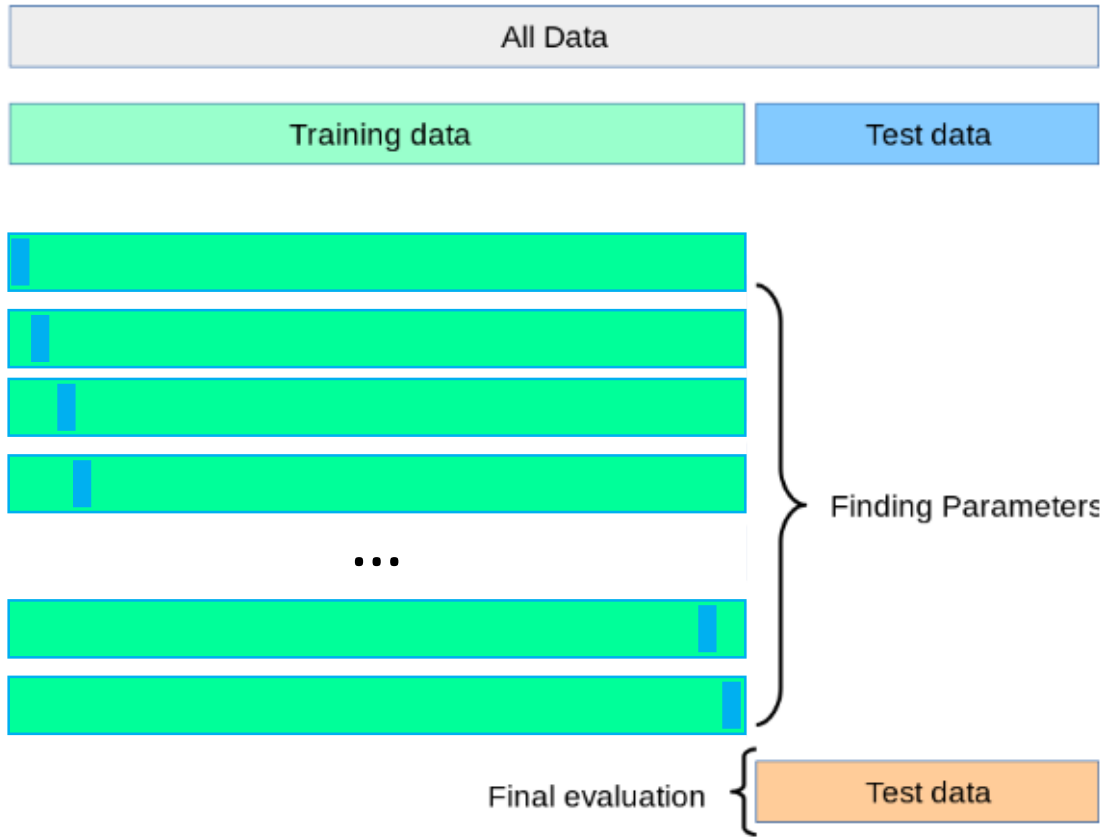
No overlap of tests sets in the different cross-validation rounds



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

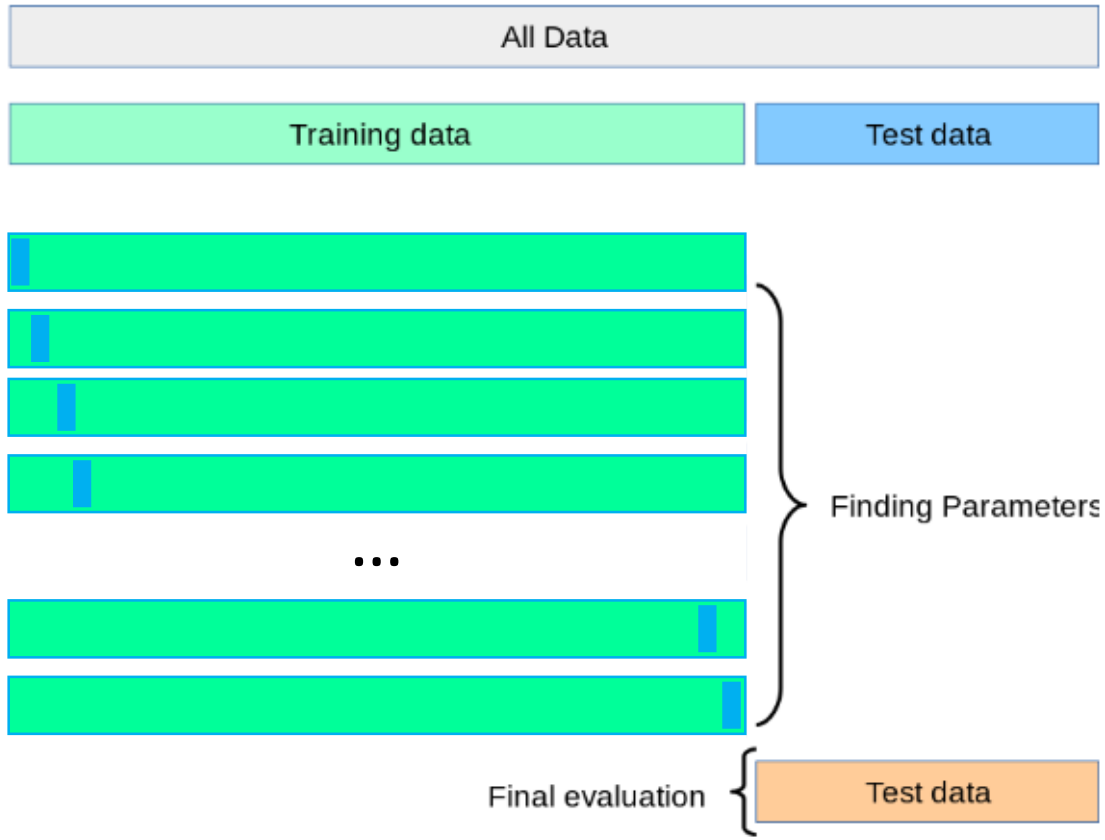
Leave One Out Cross-Validation

- K = n, where n is the number of observations
- Divide Train set into n folds
- Train model in n-1 fold
- Test model in n^{th} observation
- Repeat n times ➔ train n models
- n performance values
- Final performance metric:
  - mean ± std



Leave One Out Cross-Validation

- Computationally expensive
- Models are almost identical as they are trained on practically the same training dataset ➔ high variance
- No overlap of test sets in the different folds
- Some metrics can't be estimated, i.e., ROC-AUC, precision and recall.

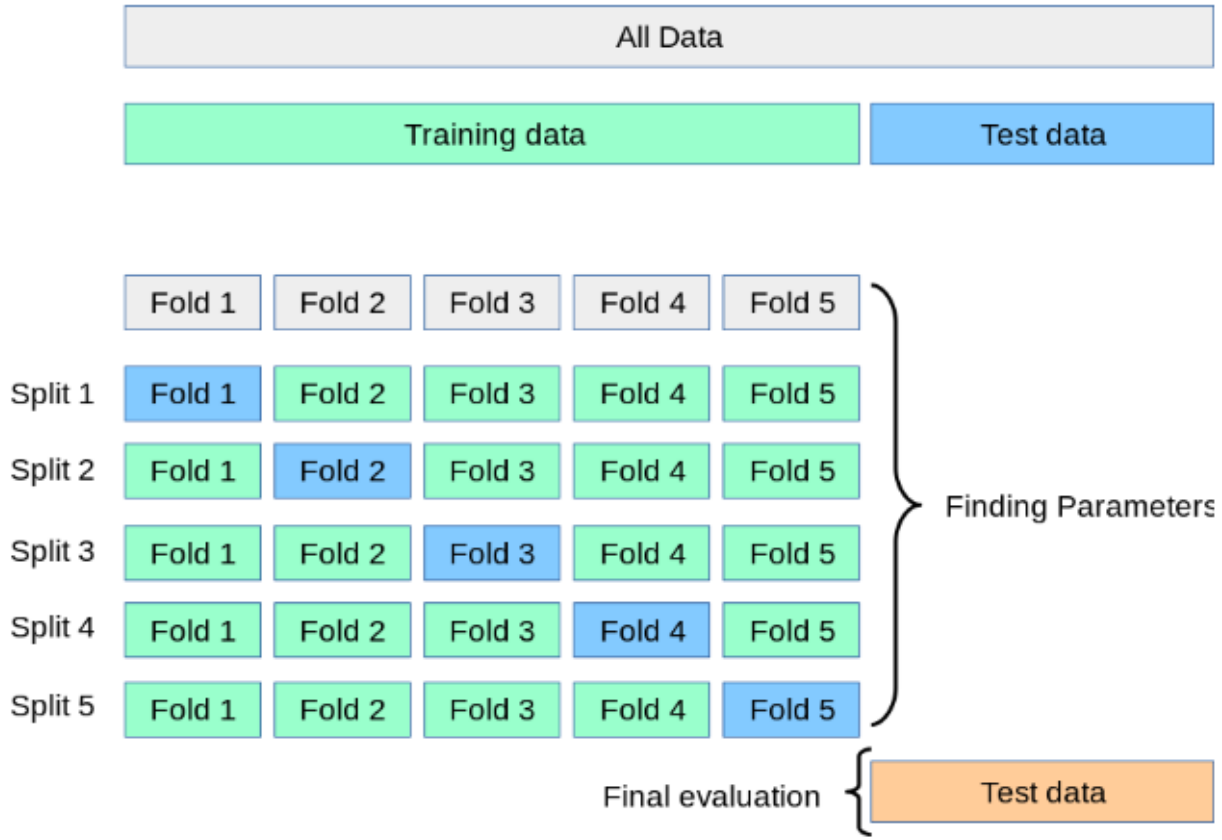


Leave P Out Cross-Validation

- Leaves out all possible subsets of p observations
- For n observations, this produces $\binom{n}{p}$train-test pairs
- There is overlap the different test sets
- We have bigger validation sets ➔ better measure of performance (than LOOCV)
- Very computationally expensive

Repeated K-Fold Cross-Validation

- Repeats K-Fold Cross-Validation, n times, each times making different data split
- The values of the training set are shuffled before making the split into the K fold
- Repeat n times:
  Shuffle data ➔ K-Fold CV
- K × n performance metrics
- There could be overlap between the tests sets in different repeats



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

Stratified K-Fold Cross-Validation

- Only for classification
- Procedure identical to K-fold Cross-Validation,
- Ensures that each fold has a **similar proportion of observations of each class**
- **Useful with (very) imbalanced datasets**
- K performance metrics
- No overlap of test sets

Uses of Cross-Validation

- Estimate the generalization error of a given model
- Select best performing model from a group of models
  - Different algorithms
  - Different feature subsets
- Select hyperparameters

To consider

- Generally use K-fold cross-validation with K equals 5 or 10
- Use Stratified K-fold if target class is imbalanced
- If K is too small, the error estimate is pessimistically biased because of the difference in training-set size between the original dataset and the cross-validation datasets.

- Leave-one-out cross-validation works well for estimating continuous error functions (e.g., mean squared error), but it may perform poorly for discontinuous error functions, (e.g., number of misclassified cases, precision and recall).

Train In Data

THANK YOU

www.trainindata.com