Cross-validation

STAT 471

September 23, 2021

Where we are

Unit 1: Intro to modern data mining

Unit 2: Tuning predictive models

Unit 3: Regression-based methods

Unit 4: Tree-based methods

Unit 5: Deep learning

Lecture 1: Model complexity

Lecture 2: Bias-variance trade-off

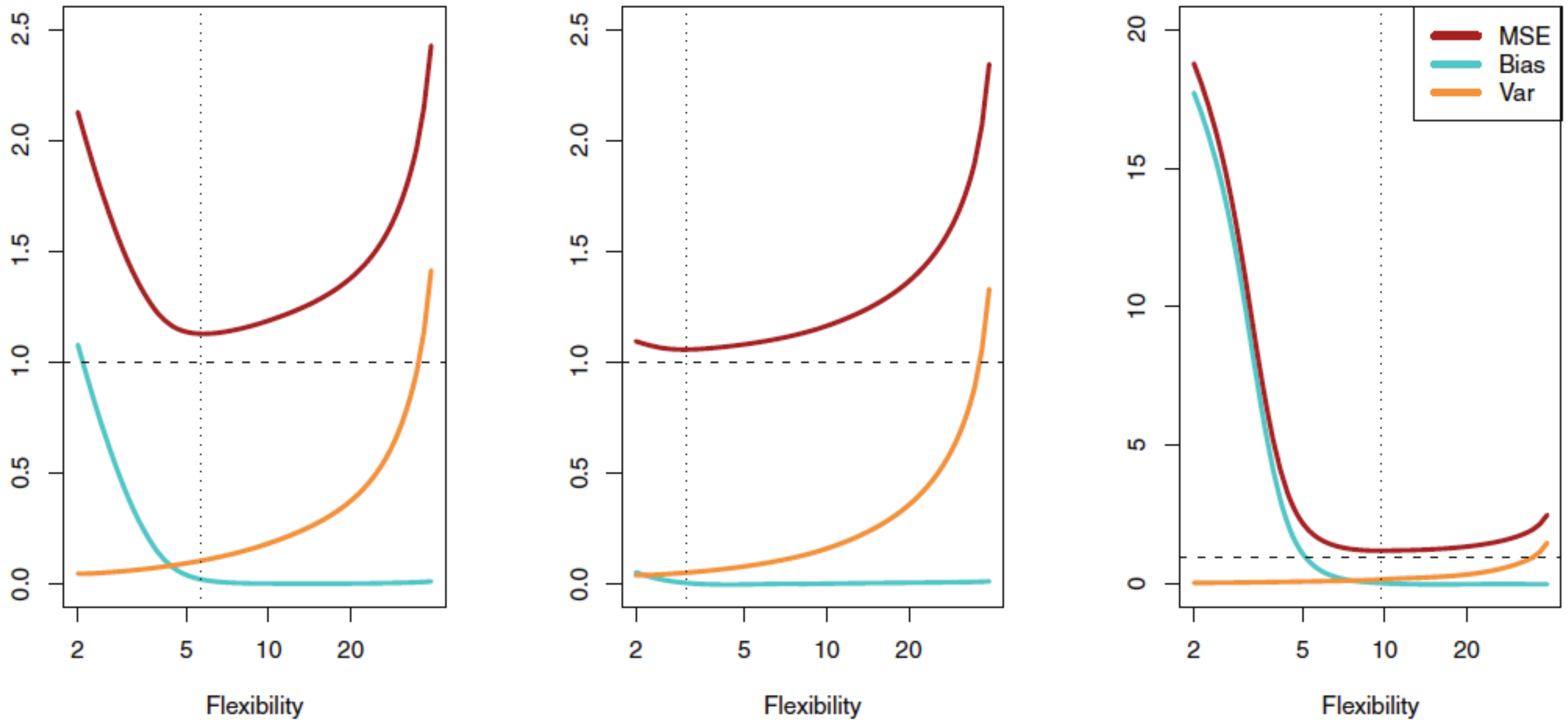
Lecture 3: Cross-validation

Lecture 4: Classification

Lecture 5: Unit review and quiz in class

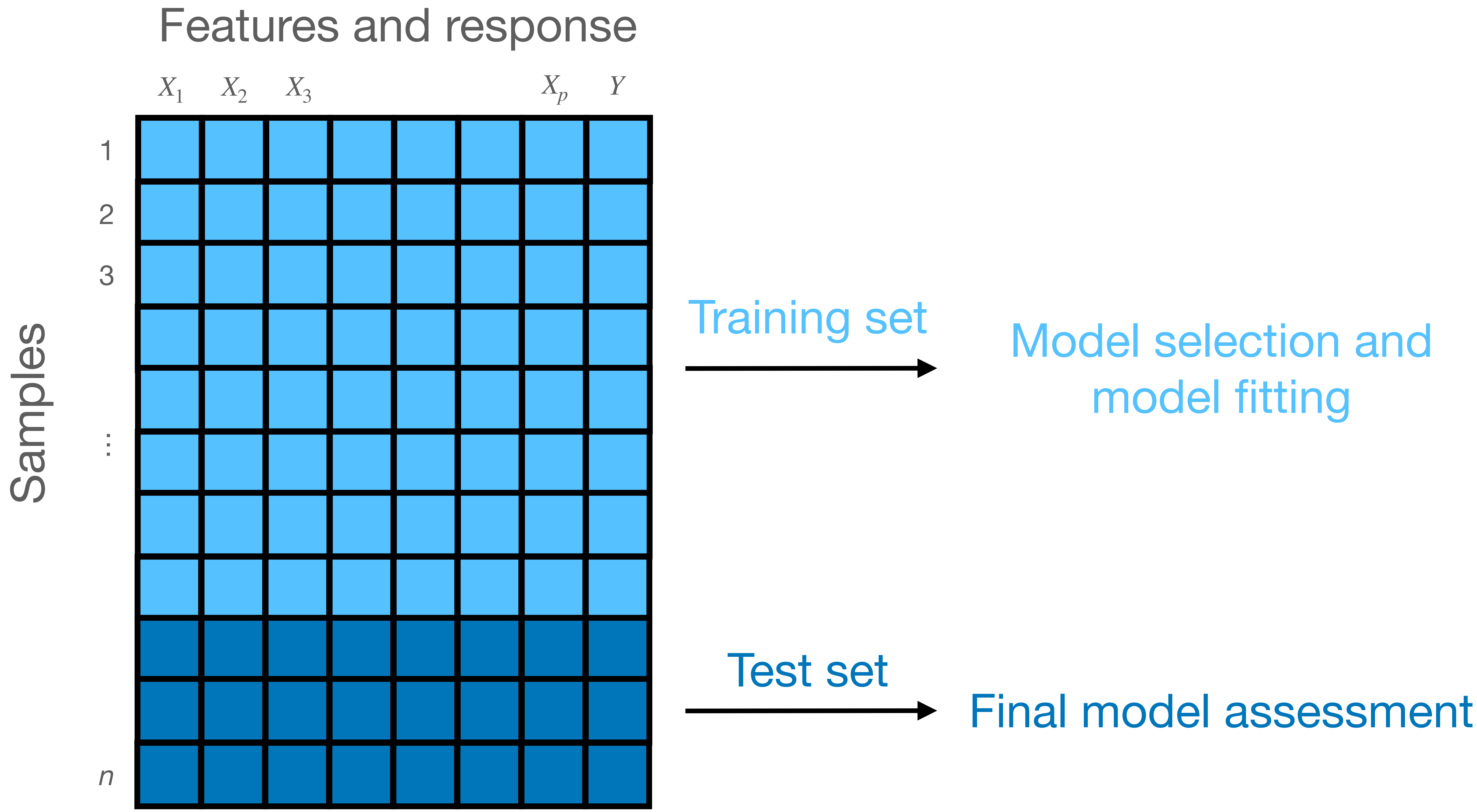
Homework 1 due the following Monday.

Estimating test error for model selection and assessment



- How do we estimate the test error for model selection?
- How do we estimate the test error for final model assessment?

Estimating test error for final model assessment

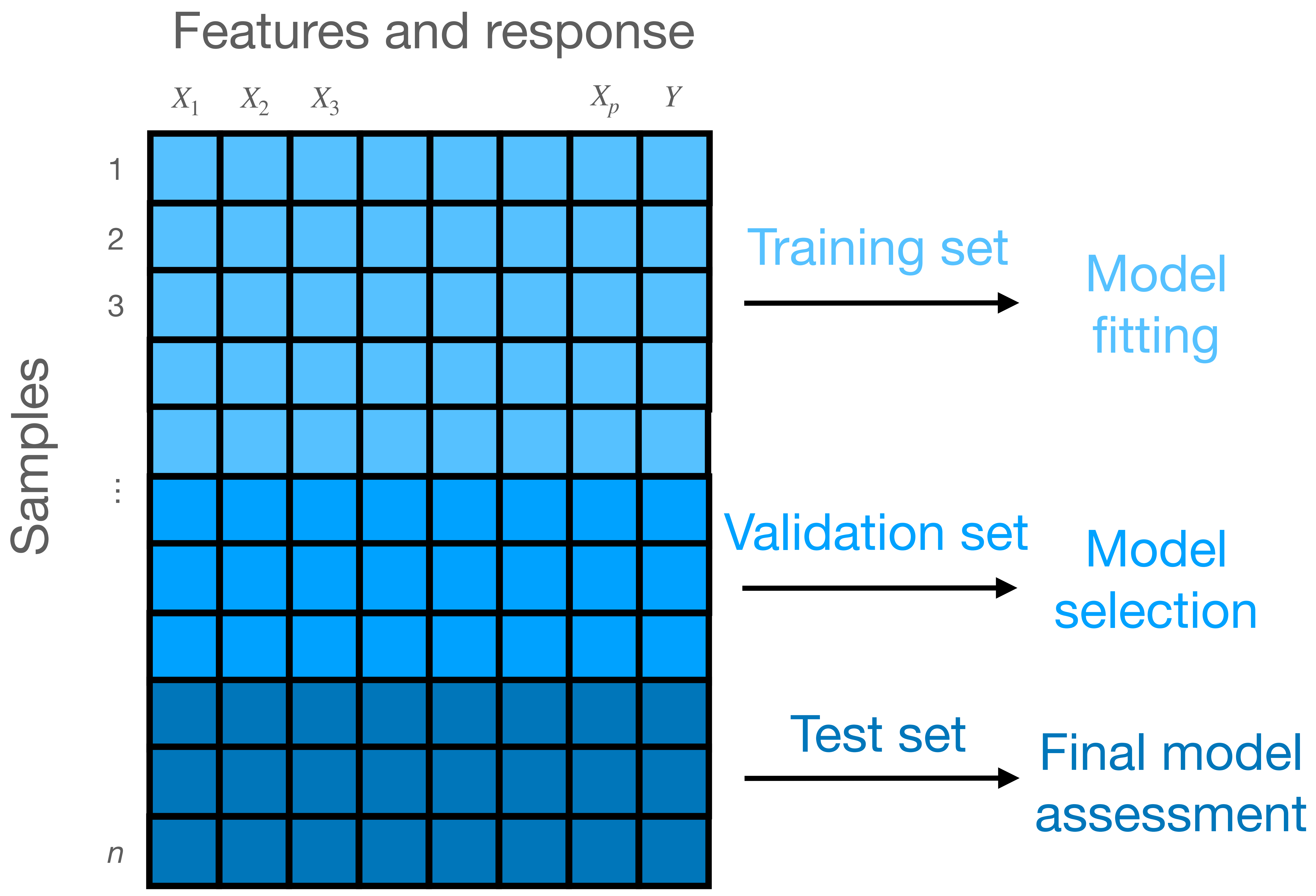


More samples for training: better fitted model.

More samples for testing: better estimate of test error.

Common splits range from
50% training / 50% testing
to
80% training / 20% testing.

Validation set approach for model selection



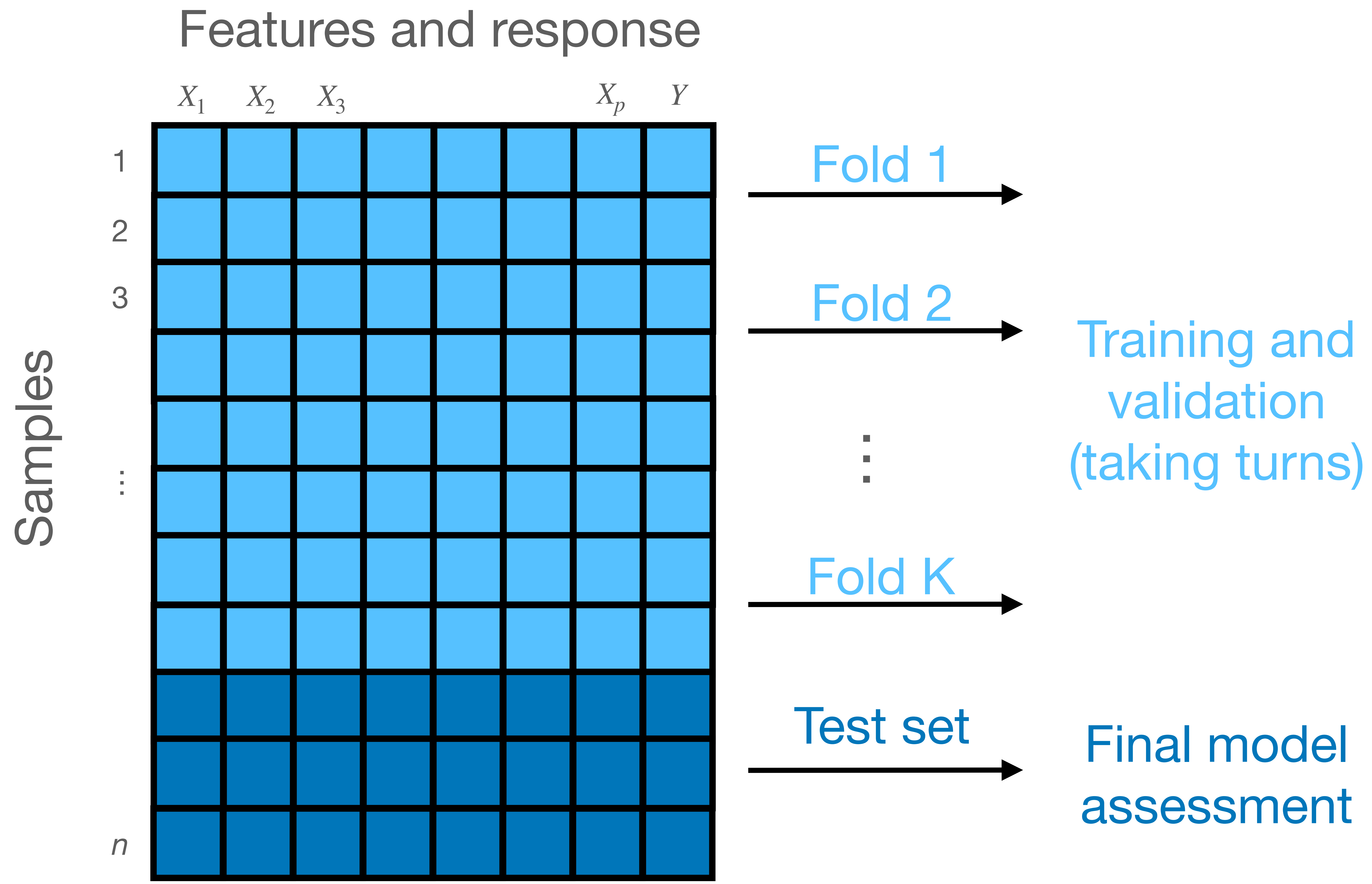
1. Fit models of varying complexity to training set
2. Estimate test error for each model on validation set
3. Choose model complexity to minimize validation error
4. Refit this model on combined training and validation sets
5. Evaluate the final model on the test set

Drawback of validation set approach

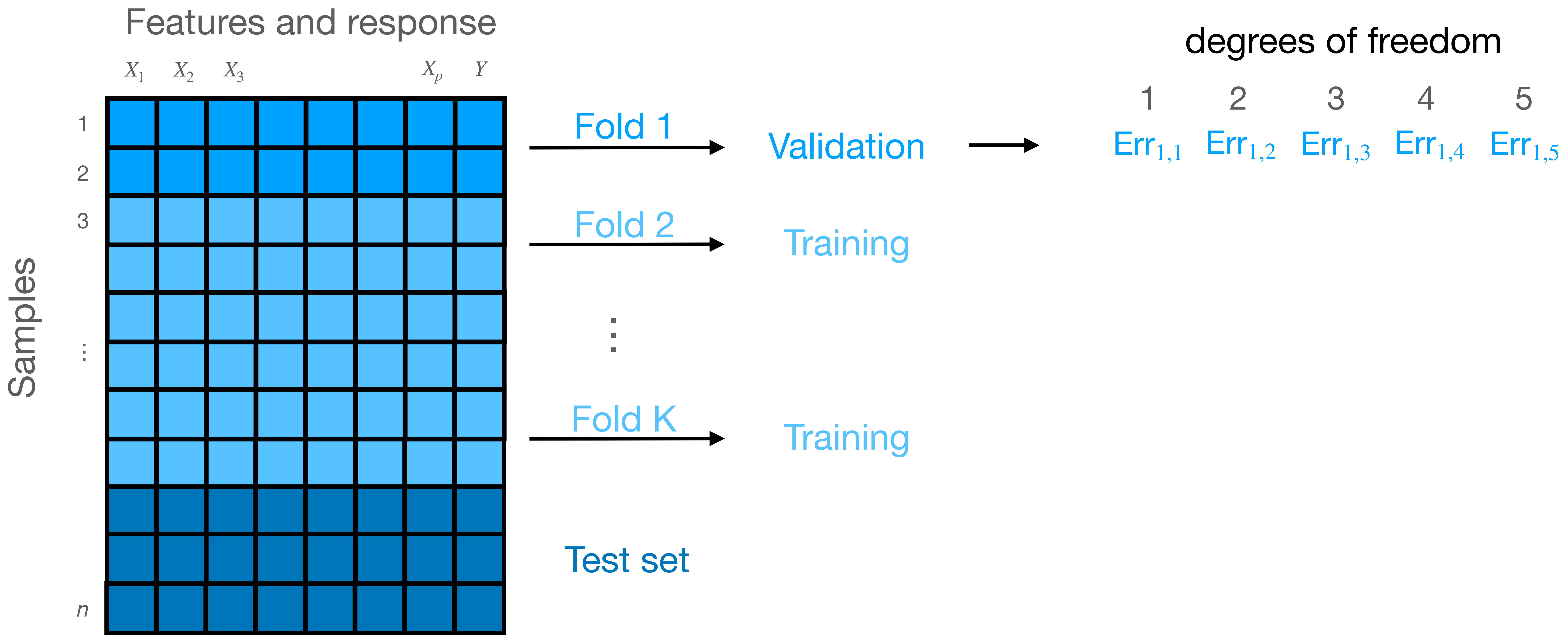
The validation set approach does not make efficient use of samples.

Using a small validation set can lead to a suboptimal model complexity choice.

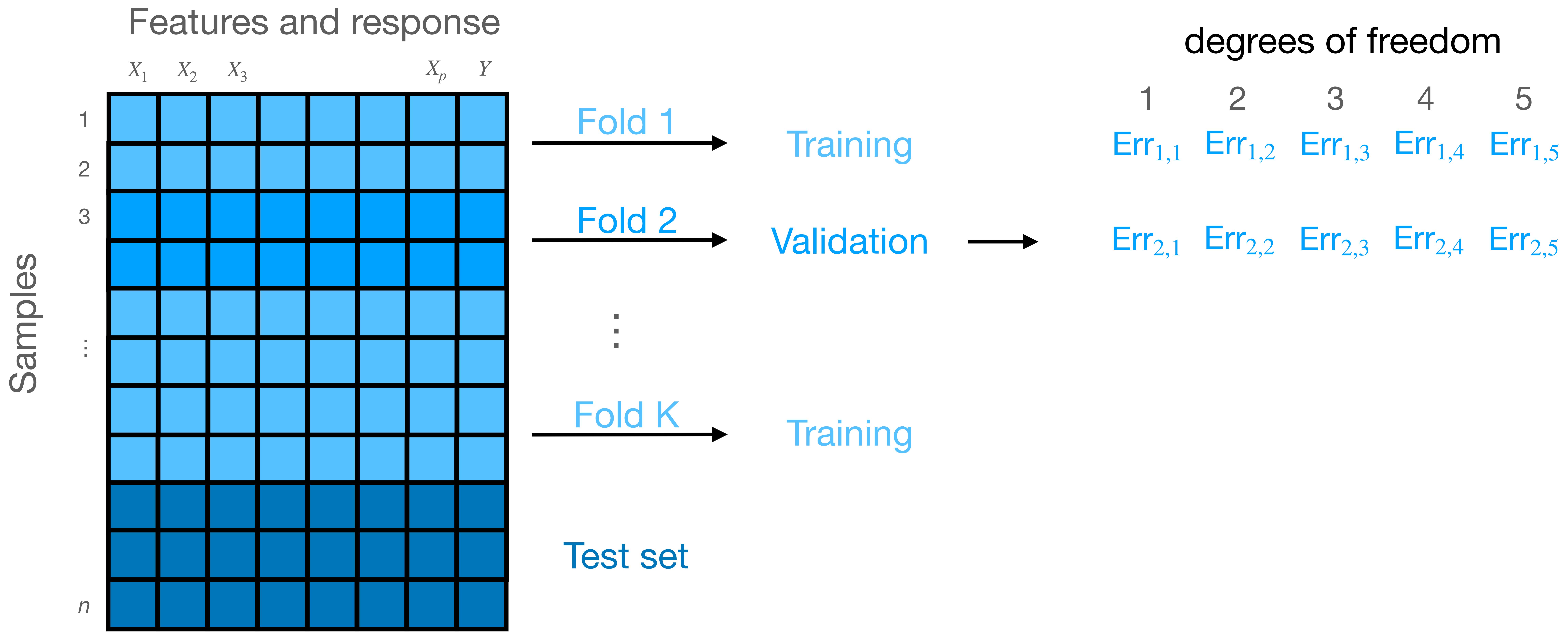
Cross-validation for model selection



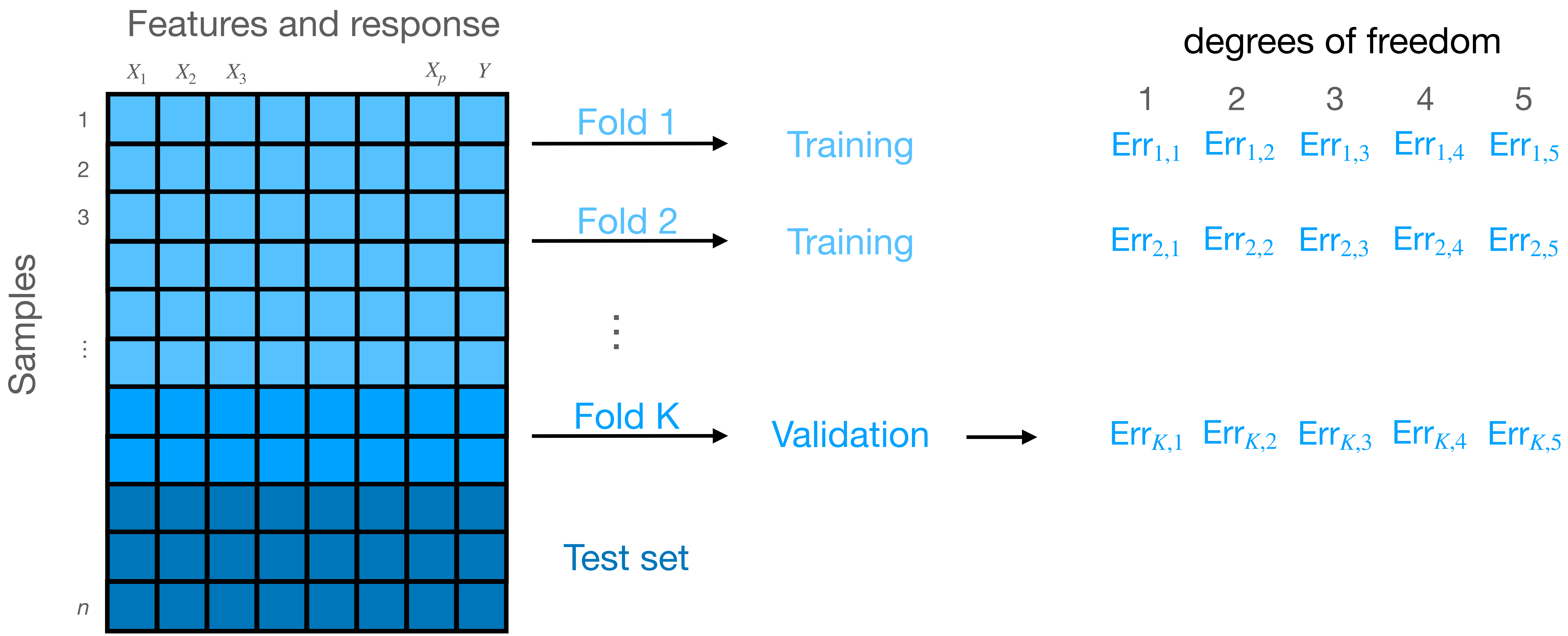
Cross-validation for model selection



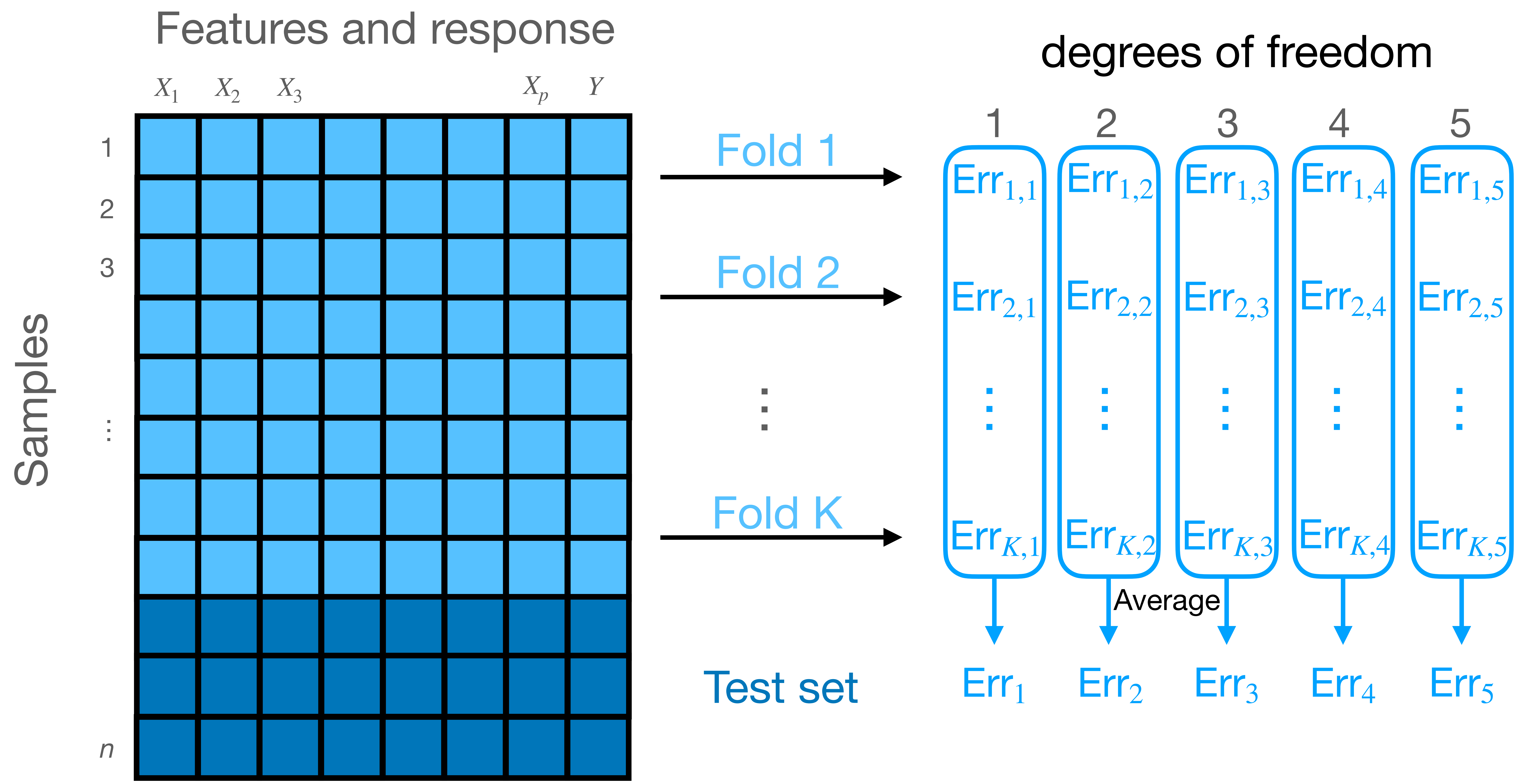
Cross-validation for model selection



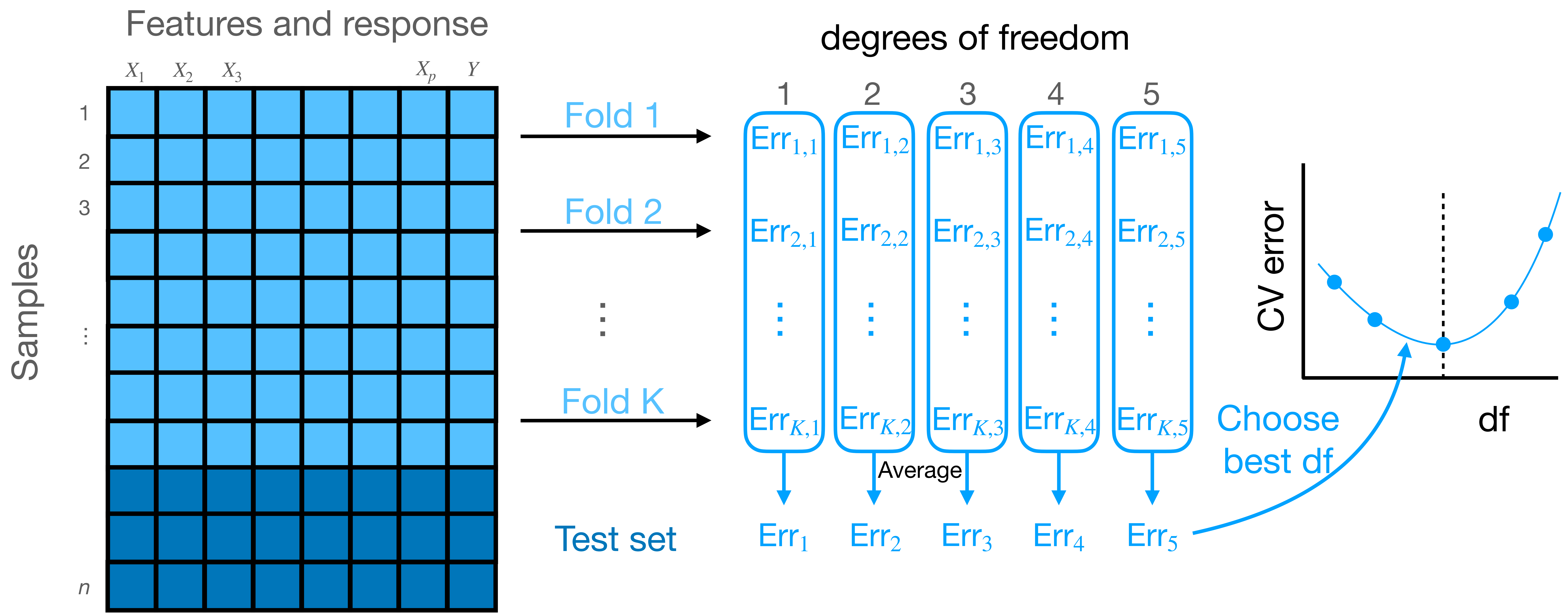
Cross-validation for model selection



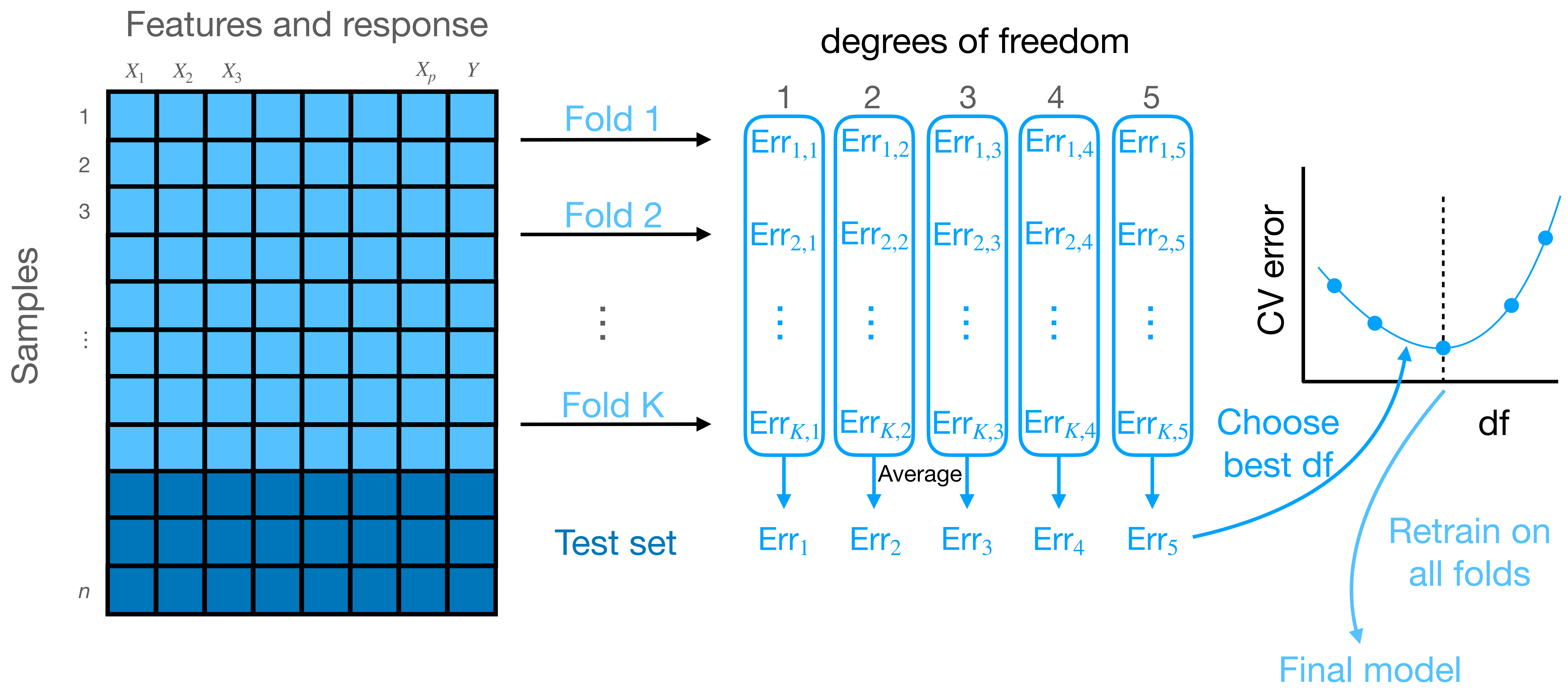
Cross-validation for model selection



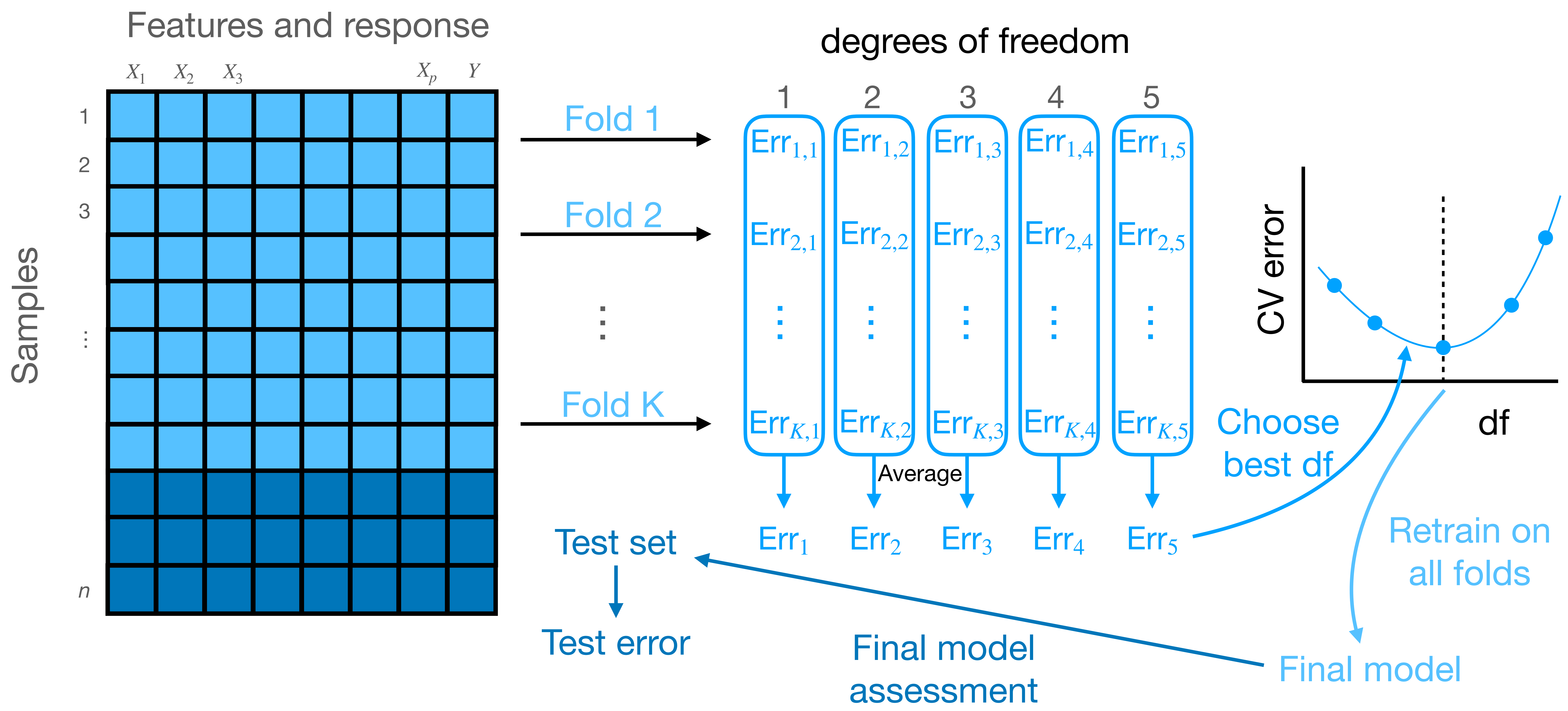
Cross-validation for model selection



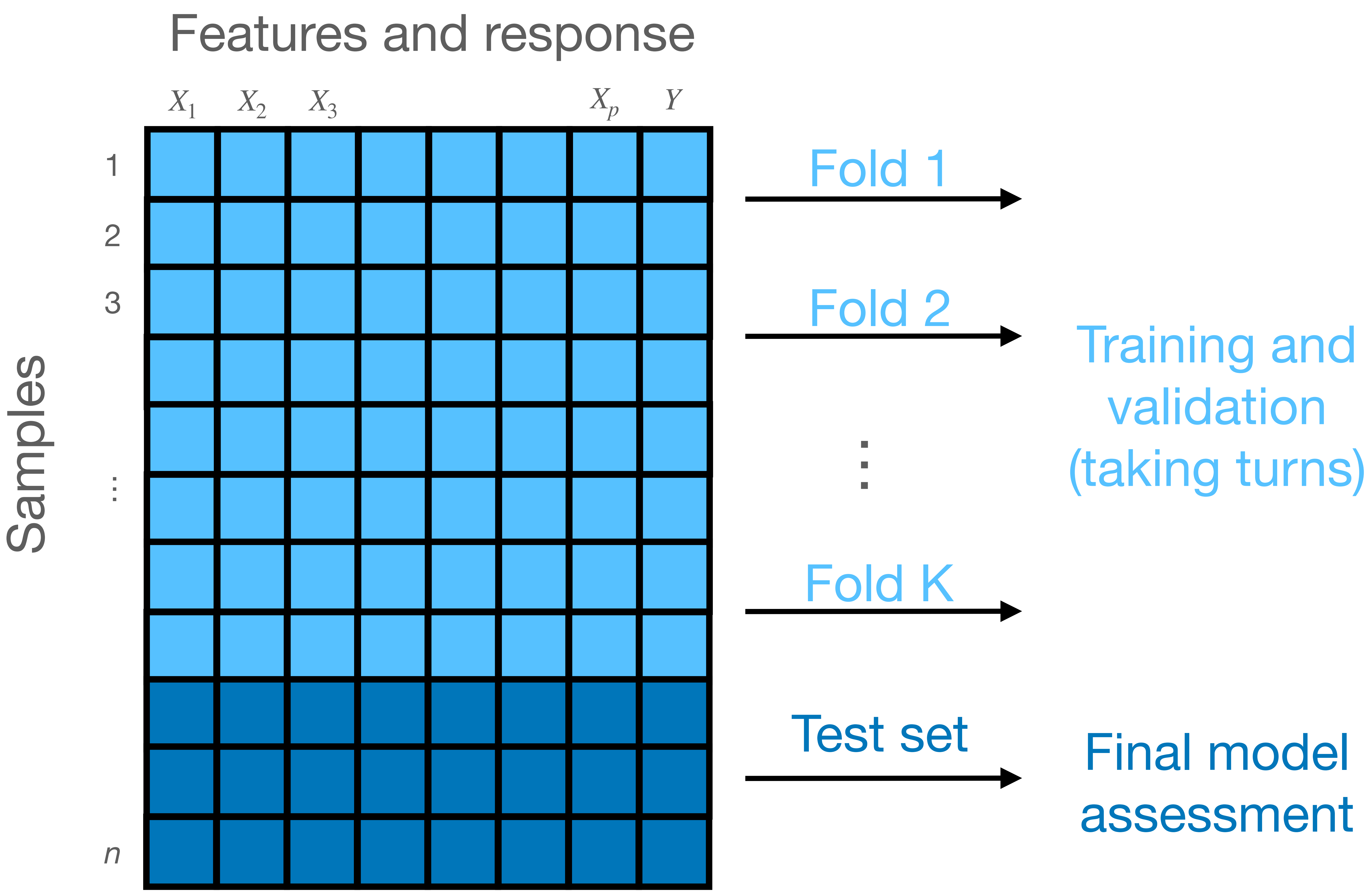
Cross-validation for model selection



Cross-validation for model selection



Cross-validation (summary)



1. Split data into K folds
2. For each fold k,
   - Fit models of varying complexity to training data, holding out fold k
   - Evaluate validation error for each model on fold k
3. Average across folds to get CV error for each model complexity
4. Choose model complexity to minimize CV error
5. Refit this model on all folds
6. Evaluate final model on the test set

Different kinds of test error

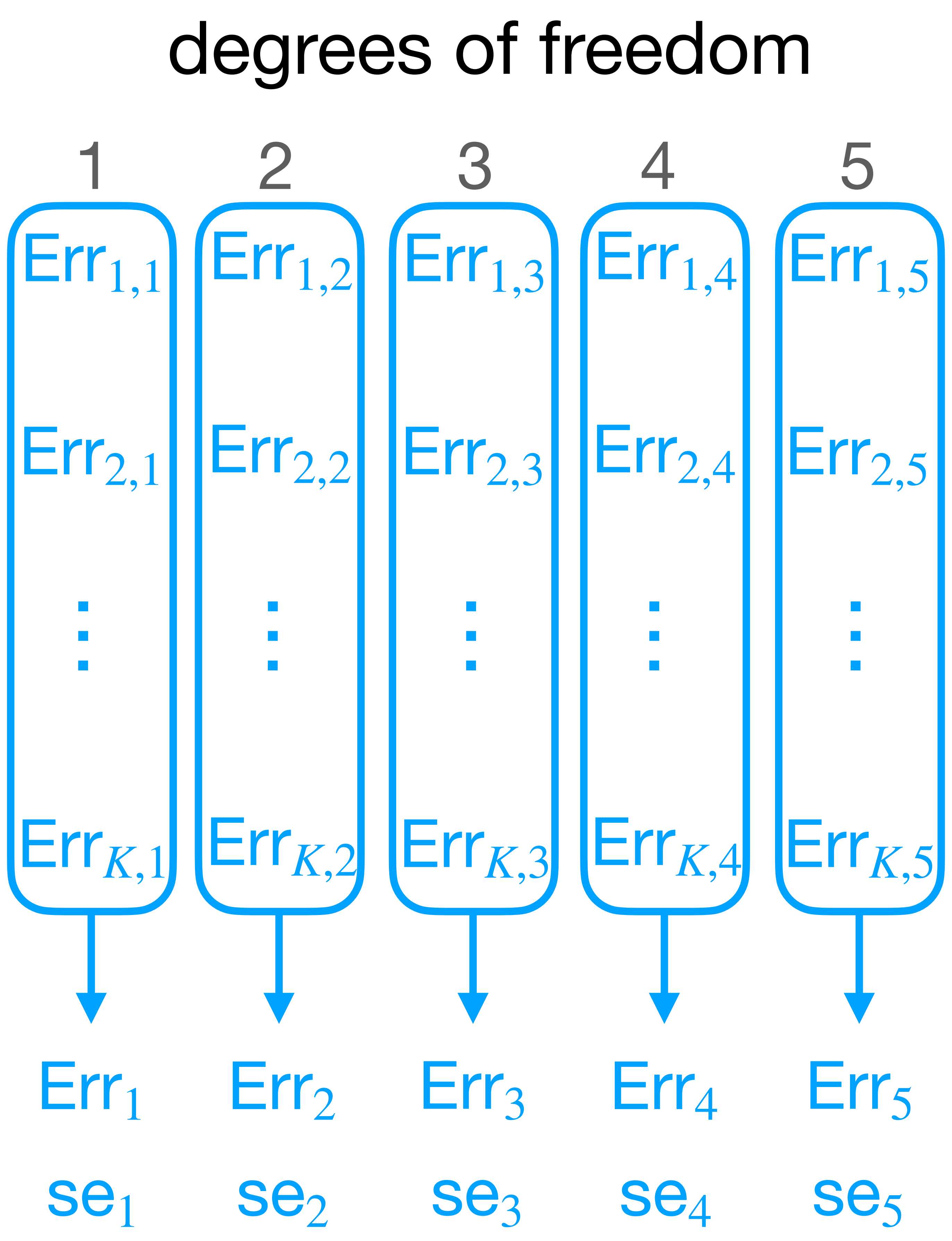
Cross-validation is compatible with any definition of test error (e.g. mean squared error or misclassification error).

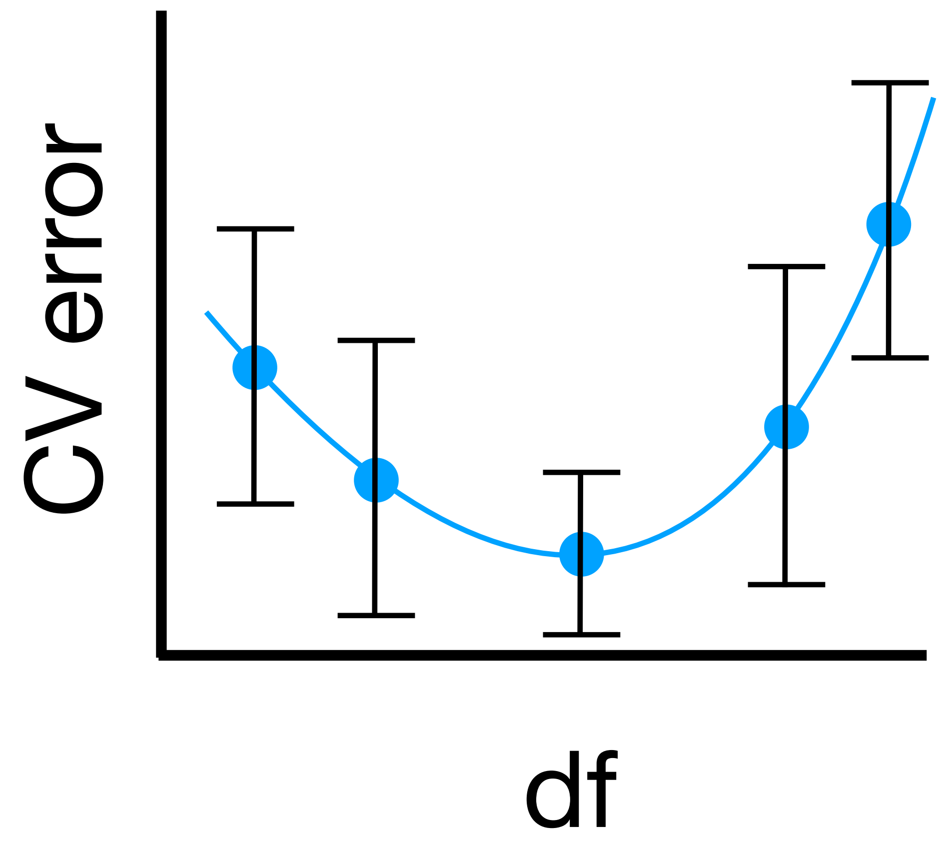
Cross-validation should be used with the same error metric as will be used in the final model evaluation, even if different from error used for training.

Choosing the number of folds

- More folds means more computation
- Fewer folds means the training sets used for model selection are much smaller than the actual training set
- In practice, $K = 5$ or $K = 10$ are common choices

Cross-validation standard error





$$\mathrm{se}_{\mathrm{df}} = \frac{1}{\sqrt{K}} \times \mathrm{s.d.}(\mathrm{Err}_{1,\mathrm{df}}, \ldots, \mathrm{Err}_{K,\mathrm{df}})$$

One standard error rule

Occam's razor:

Select the smallest model for which the CV error is within one standard error of the lowest point on the curve.