Model validation

Lecture 8 of “Mathematics and AI”

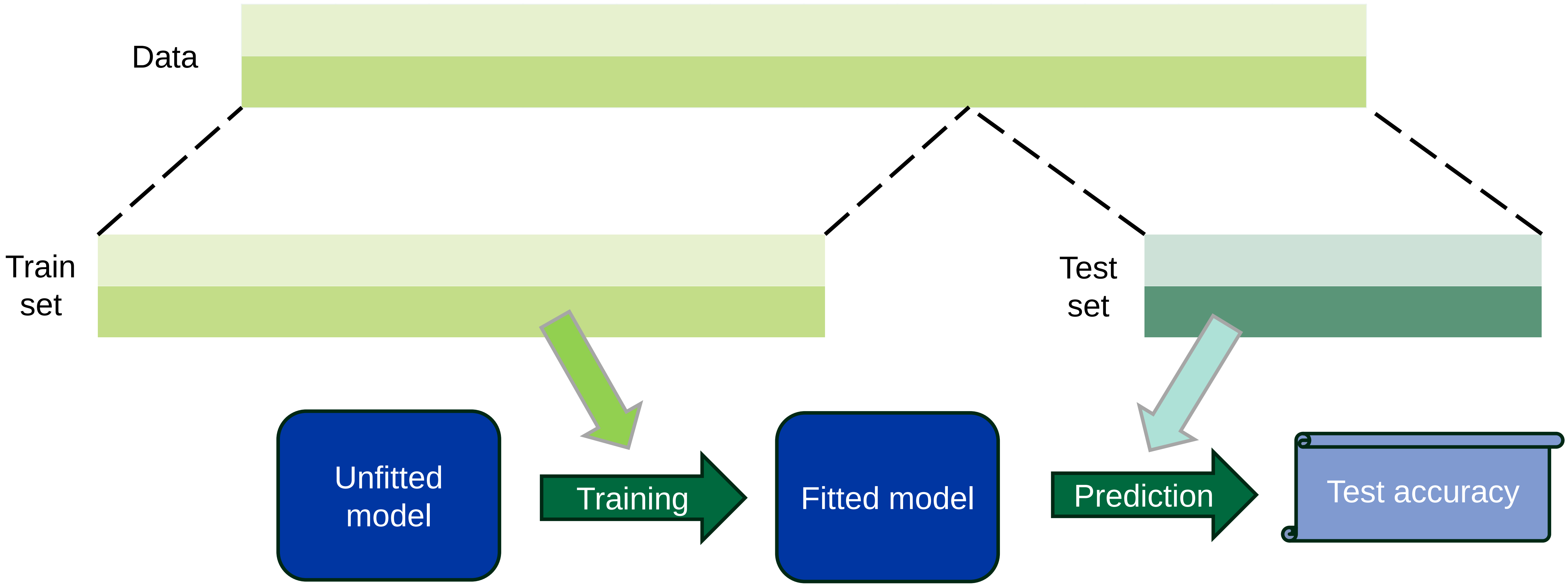
Outline

1. Why validate models?
2. Single-validation set approach
3. Crossvalidation
4. Bootstrap
5. Data leakage

Why validate models?

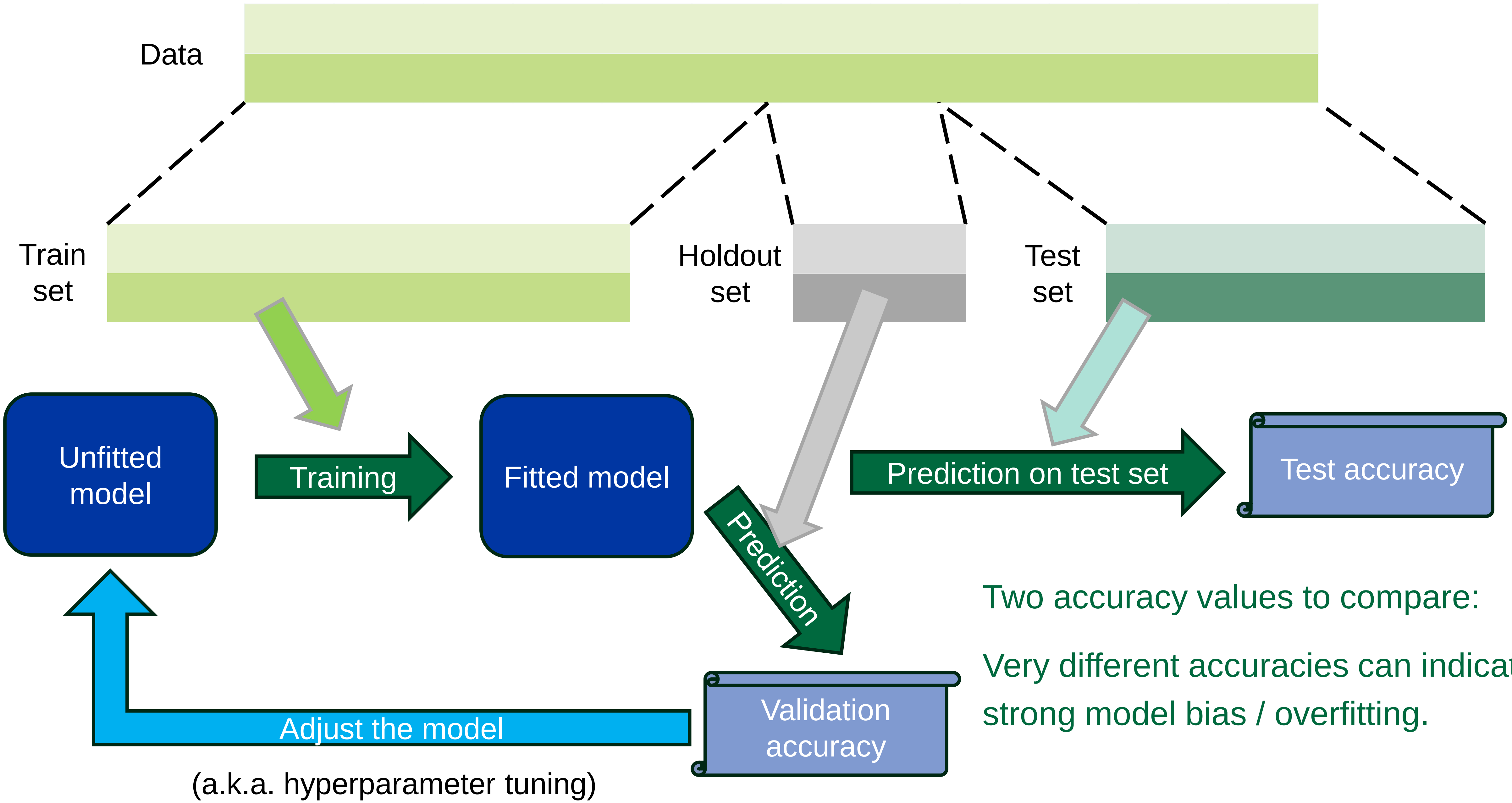
Data

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	…	x_{n-1}	x_n
y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	…	y_{n-1}	y_n



- High test accuracy: Do we actually have a good model or did we pick a very lucky train-test split?
- Low test accuracy Did we pick a bad model of did we just pick a very unlucky train-test split?

Single-validation set approach



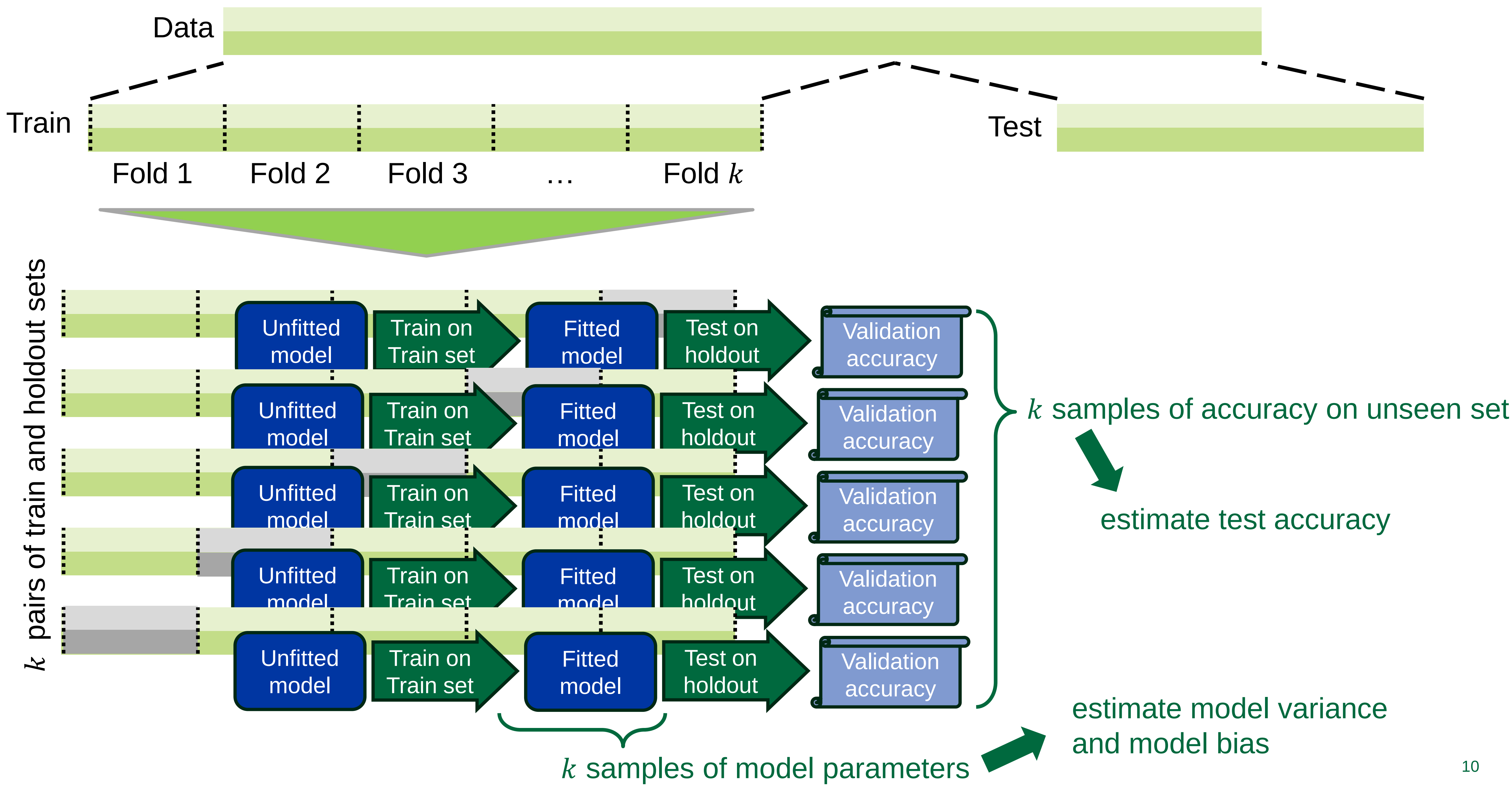
Two accuracy values to compare:

Very different accuracies can indicate strong model bias / overfitting.

Model validation can be used to assess **model bias**.

It would be nice if we could assess **model variance** too.

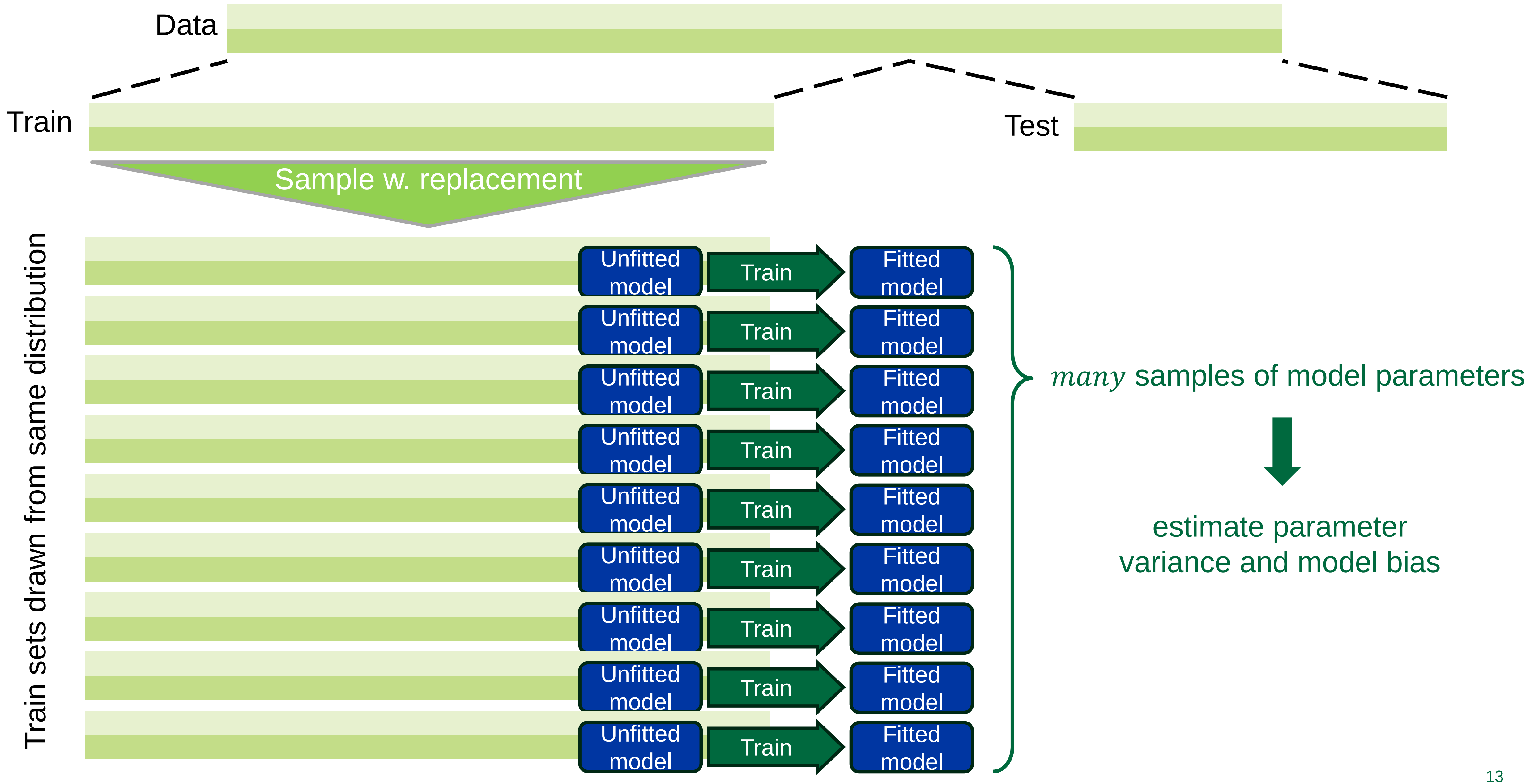
Crossvalidation


Data
Train
Test
Fold 1
Fold 2
Fold 3
…
Fold k
k pairs of train and holdout sets
Unfitted model
Train on Train set
Fitted model
Test on holdout
Validation accuracy
k samples of accuracy on unseen set
estimate test accuracy
k samples of model parameters
estimate model variance and model bias

k-fold Crossvalidation

- Leave-one-out crossvalidation (LOOCV) = n-fold validation
- Benefits of large k :
  - Improved estimate of model variance and bias through large number of samples
- Benefits of small k :
  - Greater independence among the k train sets
  - Fewer models to train

Bootstrap


Data
Train
Test
Sample w. replacement
Train sets drawn from same distribution
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
Unfitted model
Train
Fitted model
$many$ samples of model parameters
estimate parameter
variance and model bias

Bootstrap

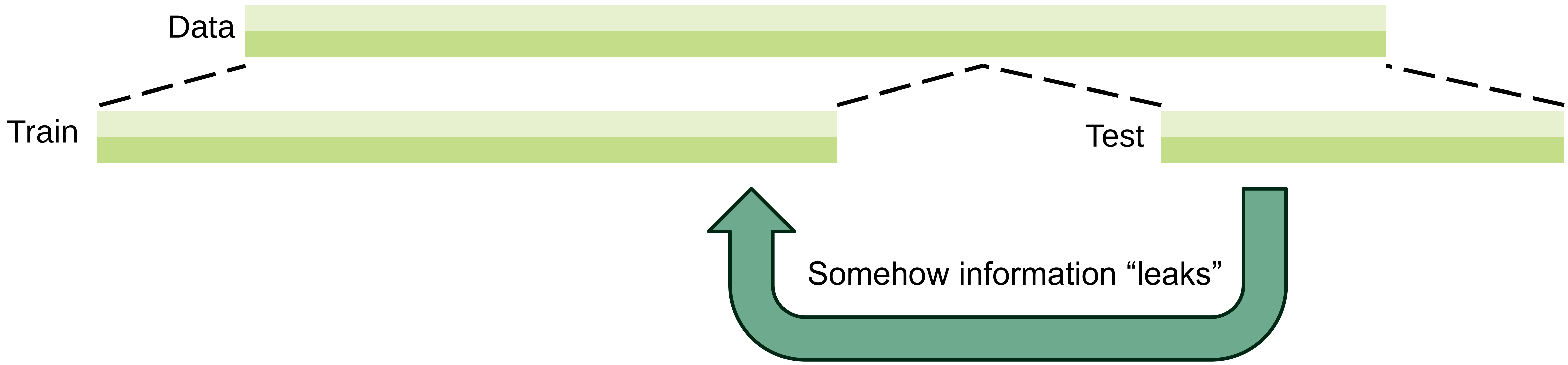
Useful when:

- Accurate assessment of parameter variance is necessary
- Too few training samples for good crossvalidation

Data leakage

“Data leakage can be [a] multi-million dollar mistake in many data science applications.”

Dan Becker (Kaggle Instructor)



Examples:

- Observations of the same test subject in train and test set (compare “eigenfaces” example)
- Conducted feature selection or data imputation on the full data set