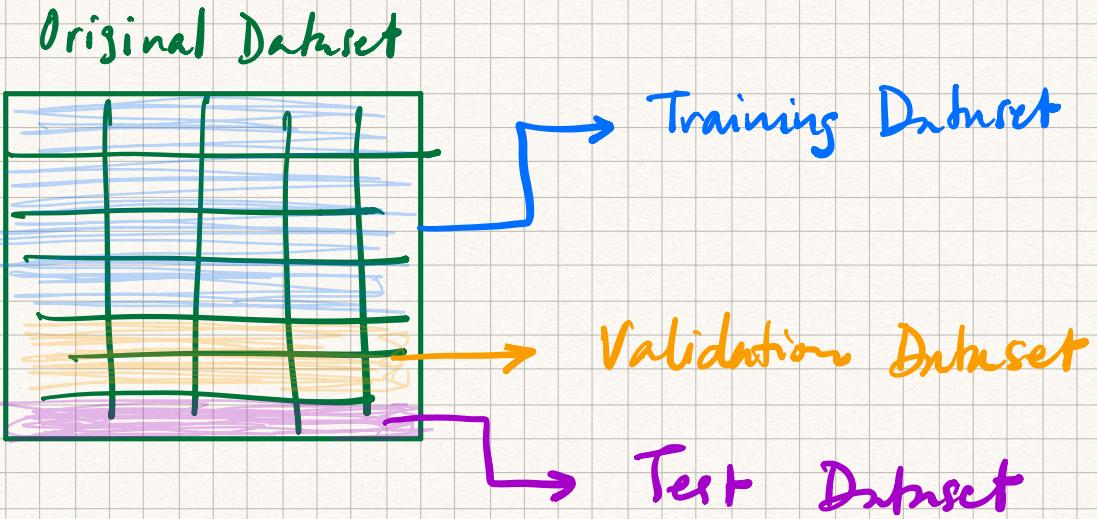


TRAINING, VALIDATION,

AND

TEST DATASETS

Why split up the data
this way and what happens
if you don't?

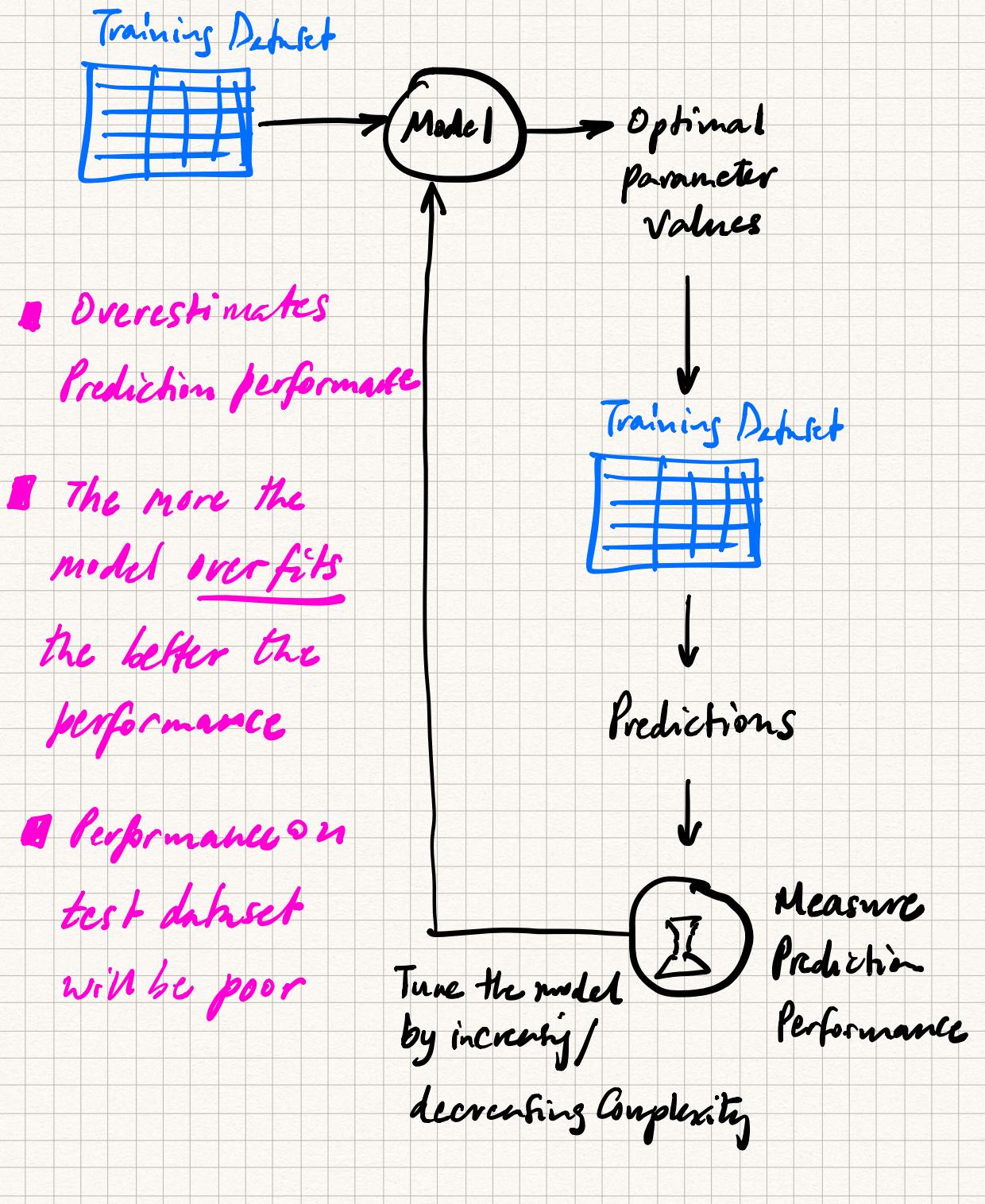


We're going to set the Test dataset aside and work with just the Training and Validation datasets.

To act to the chase

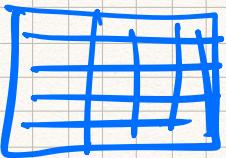
The validation dataset is used to tune the model to the right complexity.

Finding the Best Model - Scenario 1



Finding the Best Model - Scenario 2

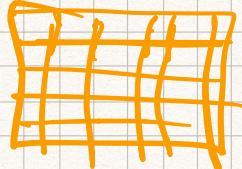
Training Dataset



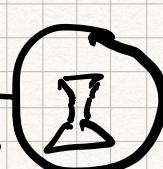
Optimal
parameter
values

- More accurate measure of performance
- However, because the model repeatedly "Sees" the same validation dataset, it becomes less novel/ new to the model.
- After a few rounds of tuning, the model will start to overfit.

Validation Dataset



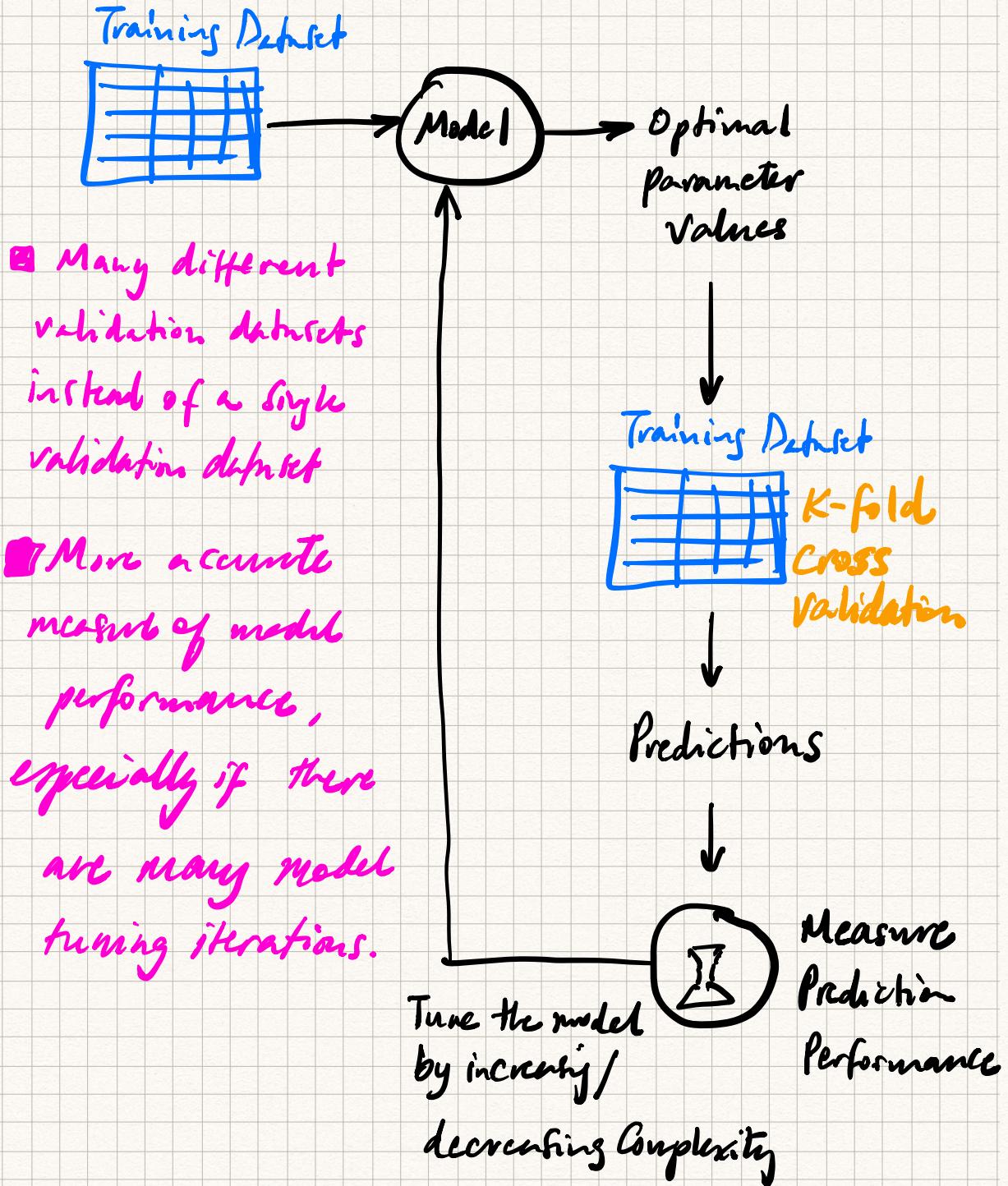
Predictions



Tune the model by increasing / decreasing Complexity

Measure
Prediction
Performance

Finding the Best Model - Scenario 3



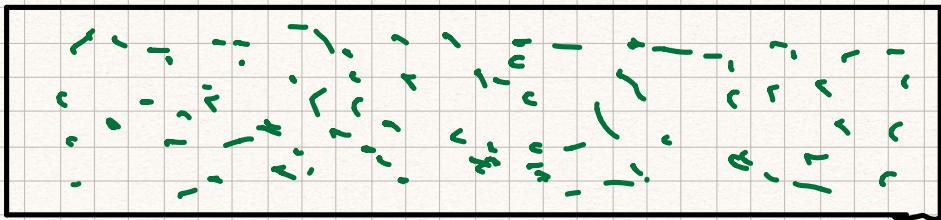
What is k-fold cross validation and
why do we need it?

It's a technique for tuning the hyper-
parameters of a model without overfitting
the model.

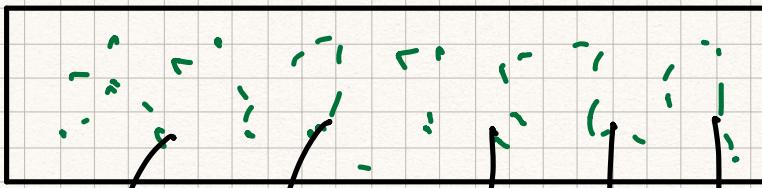
Remember: When a model is tweaked
repeatedly based on its performance as measured
on the same data set, e.g., the validation
data set, it will start to "memorize"
this validation data and start to
overfit - and become less able to
generalize.

How K-fold Cross Validation Works

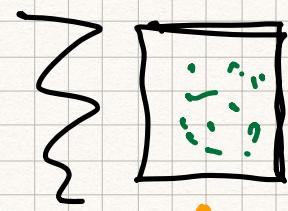
The Complete Data Set



The Training Data Set

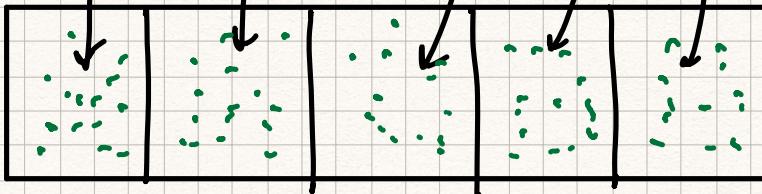


Test Data



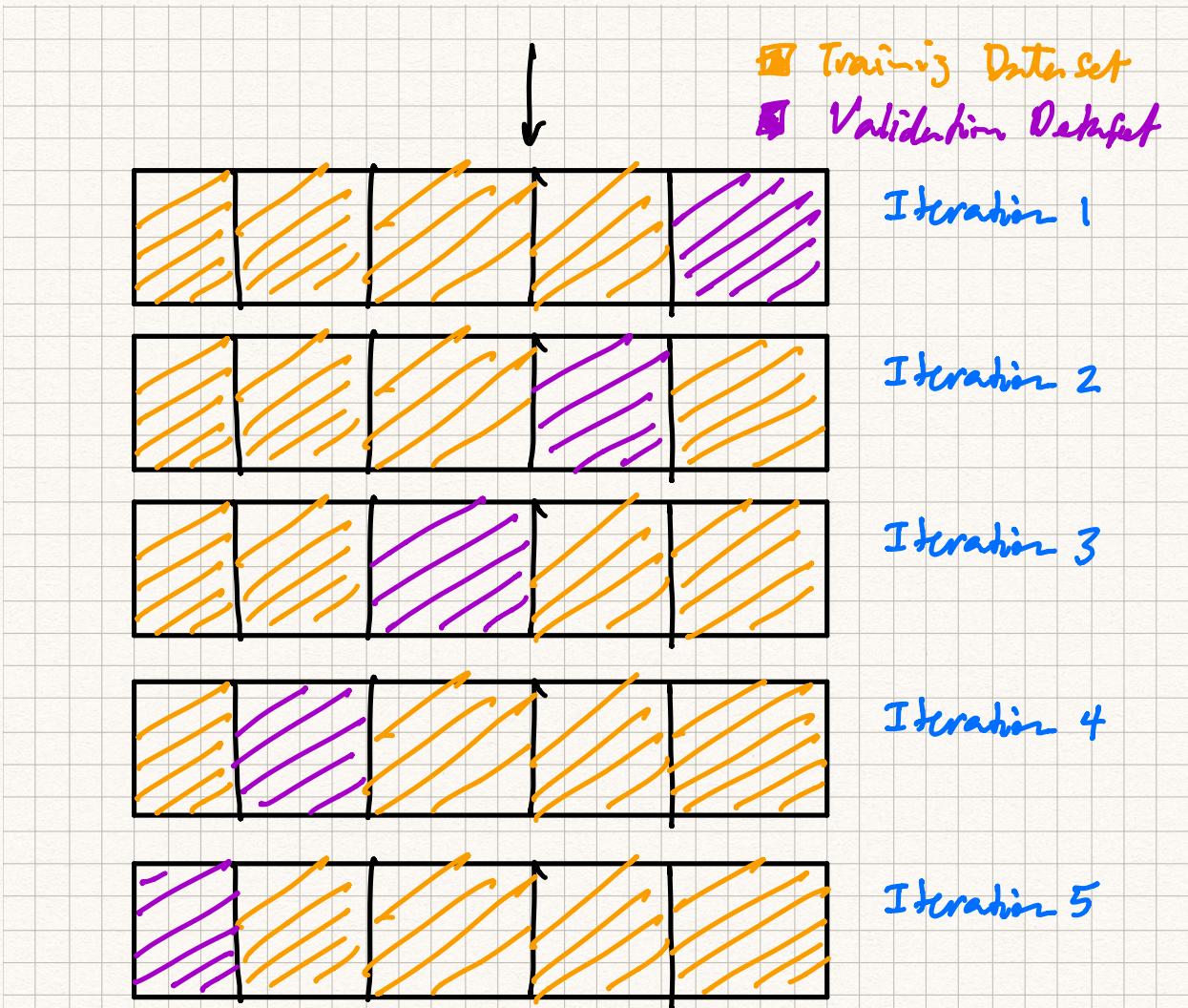
Randomly assign
data to one of 5 folds

↑
Set aside
the test
data



fold 1 fold 2 fold 3 fold 4 fold 5





To measure the performance of a model,
do it 5 times (5-fold Cross Validation).

- Train it on the Training data set
- Measure performance on the validation dataset.
- Take the average performance value as the performance of the model.

Once the model is tuned to the right complexity via k-fold cross validation, measure the performance of the model on the test dataset as a sanity check.

Typically, performance on the test dataset will be worse than the performance achieved through k-fold cross validation on the validation dataset.

Even when the performance
of the model(s) on the training
and test datasets are in accord,
the model may start to
predict less well as more
and more predictions are
made.

This means the model needs
re tuning.