

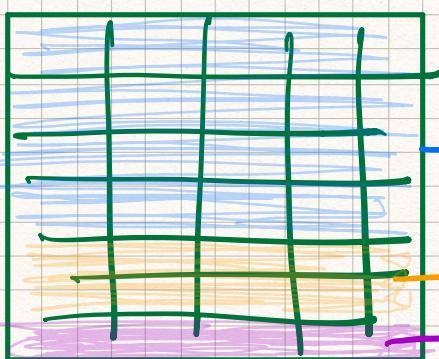
TRAINING, VALIDATION,

AND

TEST DATASETS

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

Original Dataset

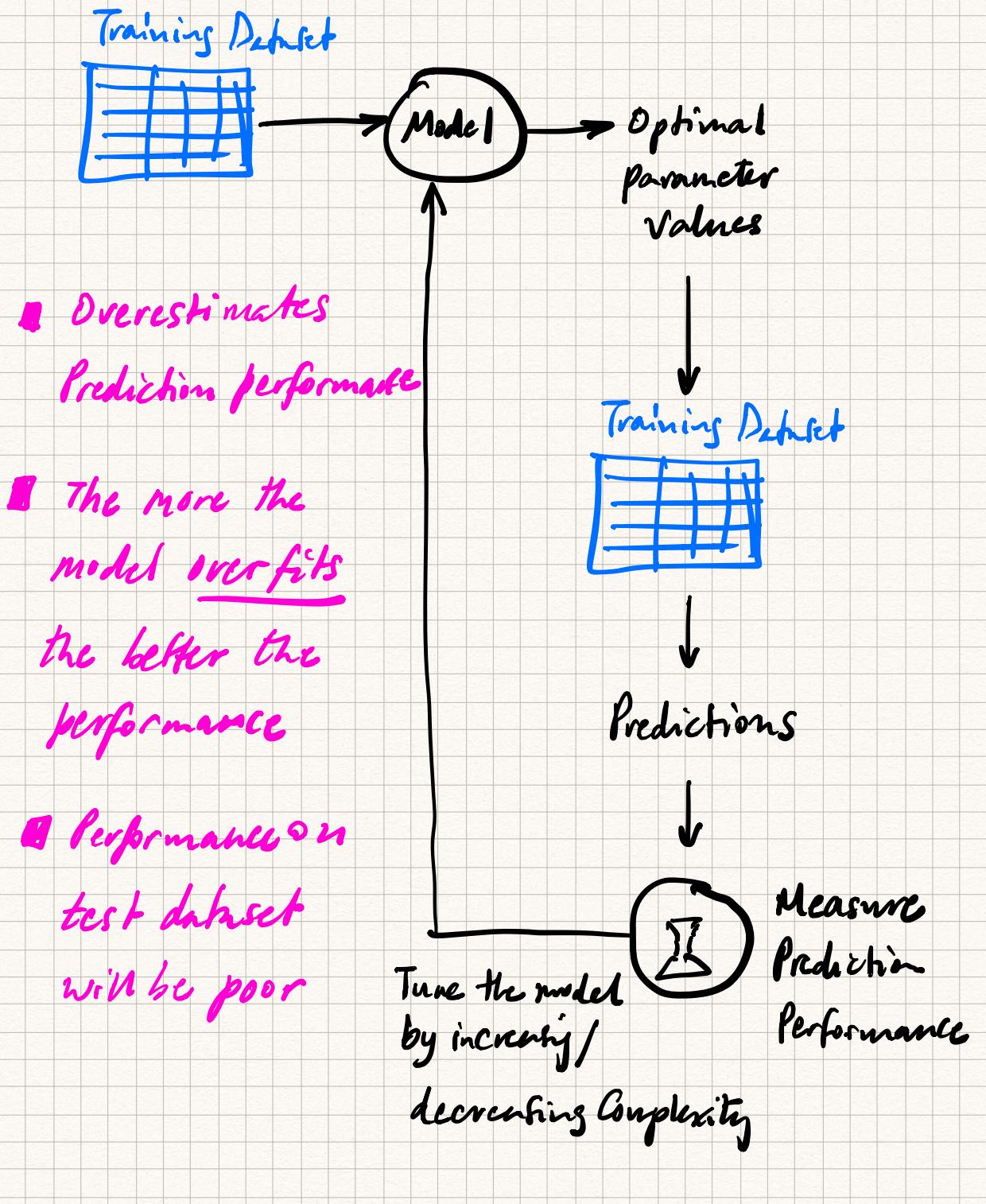


Training Dataset

Validation Dataset

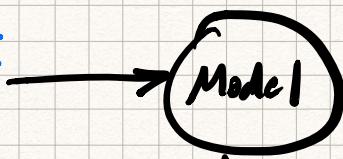
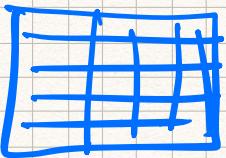
Test Dataset

Finding the Best Model - Scenario 1



Finding the Best Model - Scenario 2

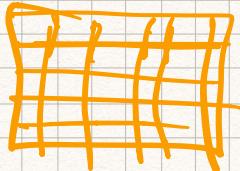
Training Dataset



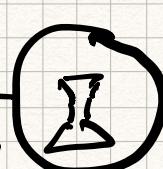
Optimal
parameter
values

- 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



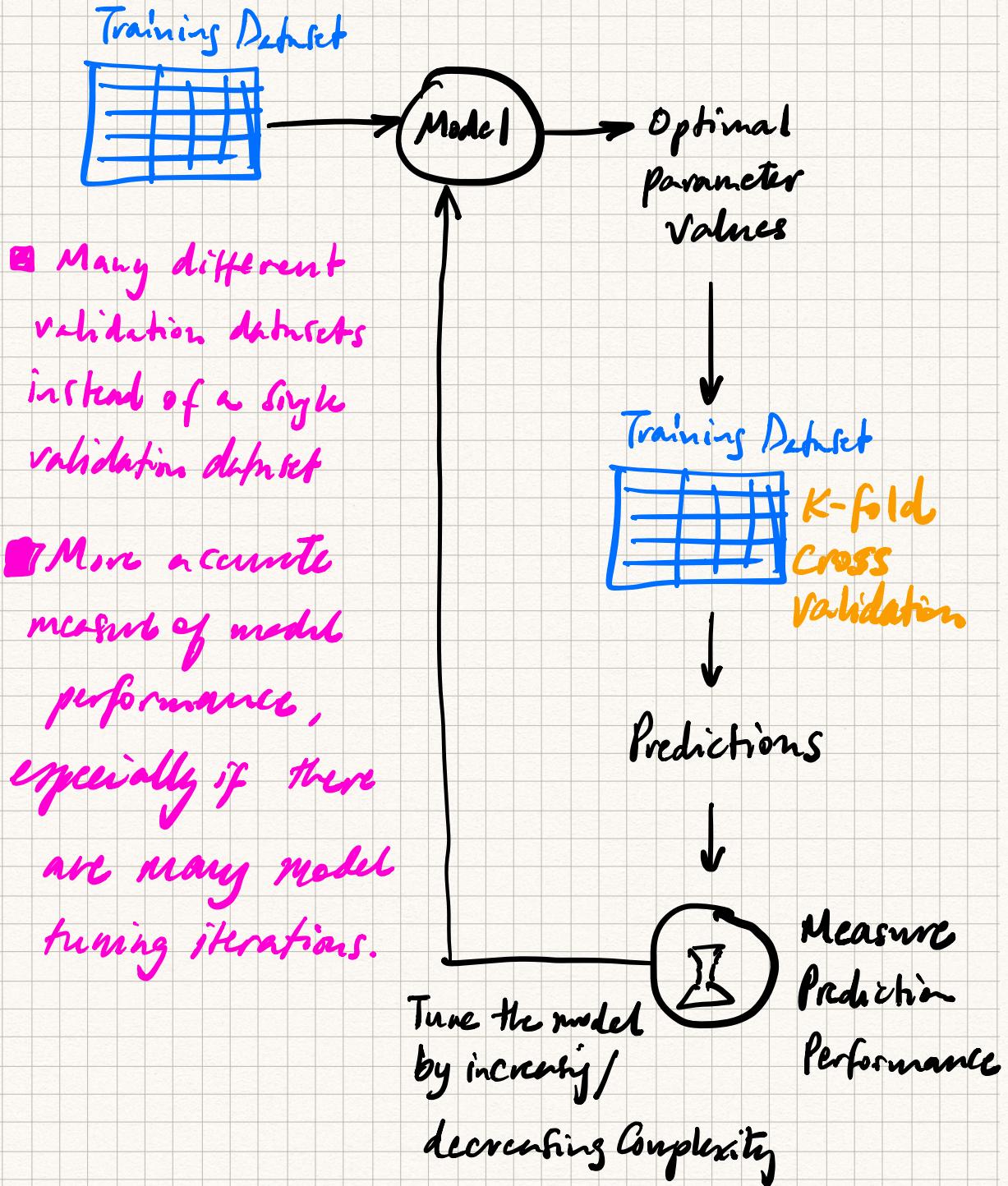
Predictions



Tune the model by increasing / decreasing Complexity

Measure
Prediction
Performance

Finding the Best Model - Scenario 3



Once the best model is found using k-fold cross validation on the training dataset, 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 training dataset.

If performance on the test dataset differs a great deal from the performance on the training dataset,

Something is wrong.

Usually, it means that the test dataset is not representative
- it differs in some ways from the training 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.