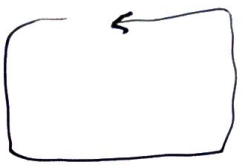
MODEL TRAINING: VALIDATION SET

It's rare that the first model you train will meet your business needs.

Model Training
Improve the model by optimizing parameters or data



Model Tuning
Analyze the model for generalization quality and source of underperformance such as overfitting

Try different models, features, alg selection, etc

hyperparam tweaks, features, etc

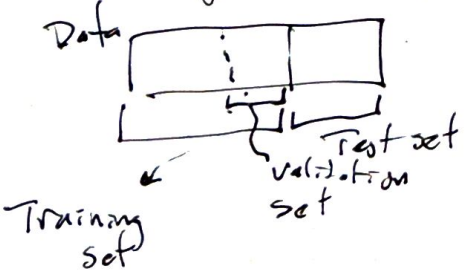
Problems with using test set

Motivation: model training and tuning involves comparing performance for different model or data settings

Problem:
when you use the test set for these comparisons, that effectively makes it part of a training set
- The model may learn the patterns from the test set during tuning

Training data: builds the model
Validation data: eval model perf during debugging/tuning
Testing data: generalizes the final data set

Solution
split training data into two parts: a training set and a validation set.
- Use the training set to train candidate models
- The validation set plays the role of test set during debugging and tuning (and model selection)
- Save the test set for measuring the generalization of your model



Issue: Splitting the training data into training and validation set may make it too small or representative

Solution: Use the holdout method to get the test set, then use K-fold cross-validation on the training set for debugging/tuning