TRAINING AND TEST SETS

- Randomize first
- 10-15% for test.
- DO NOT TRAIN ON TEST DATA!!!
  → check for dupes, etc.

VALIDATION

Repeatedly evaluating against test data and tuning each time, choosing model that performs best vs. test can overfit to test data.

Solution: A third segment, validation set, is used for tuning. *After* tuning, then test vs. test set. This will catch overfitting.

★ Can also wear out / overfit like this! Best way to avoid is to collect more data / start over.

Common issues:

- Validation / test data different from training
- Consider data set: sorted? etc.