GENERAL ASSEMBLY

Model Validation (Train/Test Split & Cross-Validation)

Matt Speck Data Science Immersive

AGENDA

- Review: Bias-Variance Tradeoff
- Training, Validating, Testing
- Cross Validation
- Three-way Train/Test Split
- Coding Implementation

BIAS-VARIANCE REVIEW

- **Q:** What is error due to bias?

BIAS-VARIANCE REVIEW

- **Q:** What is error due to bias?

- **A:** Bias is error due to the difference between the correct model and our predicted value

BIAS-VARIANCE REVIEW

- **Q:** What is error due to bias?

- **A:** Bias is error due to the difference between the correct model and our predicted value

- **Q:** What is error due to variance?

BIAS-VARIANCE REVIEW

- **Q:** What is error due to bias?
- **A:** Bias is error due to the difference between the correct model and our predicted value
- **Q:** What is error due to variance?
- **A:** Variance is the error due to the variability of a model for a given data point

BIAS-VARIANCE REVIEW

- **Q:** What causes error due to bias?

BIAS-VARIANCE REVIEW

- **Q:** What causes error due to bias?

- **A:** Having a model that is too simple

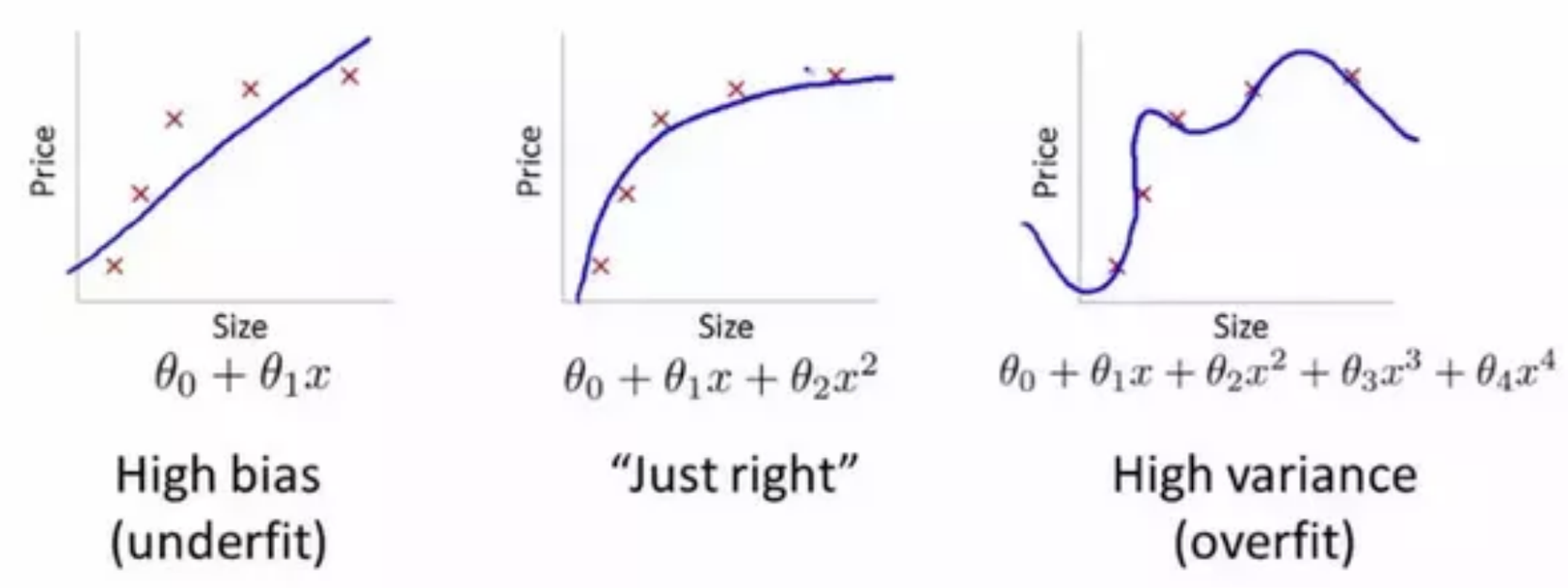
BIAS-VARIANCE REVIEW

- **Q:** What causes error due to bias?
- **A:** Having a model that is too simple
- **Q:** What causes error due to variance?

BIAS-VARIANCE REVIEW

- **Q:** What causes error due to bias?
- **A:** Having a model that is too simple
- **Q:** What causes error due to variance?
- **A:** Having a model that is too complex

BIAS-VARIANCE REVIEW



BIAS-VARIANCE → TRAIN/TEST SPLIT

- Total Error is this:

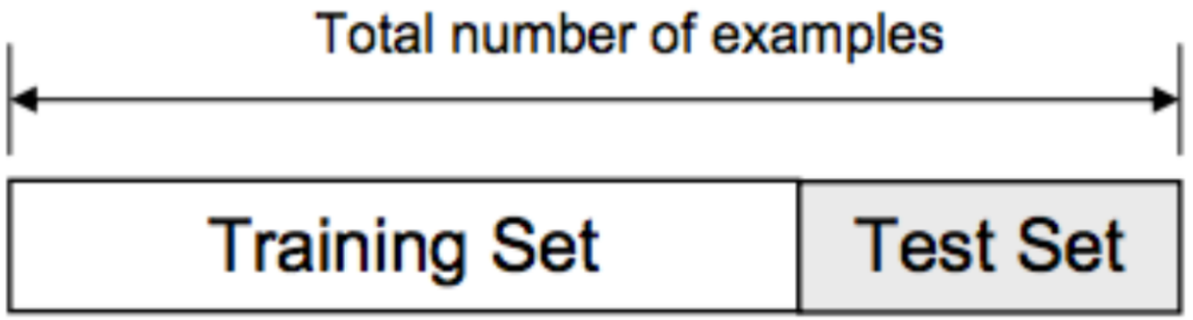
$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

  - *Err(x)* = Bias^2 + Variance + Irreducible Error

- We measure error, or accuracy, or some other metric, to **evaluate** our model.
- If we fit our model on our **entire** dataset, and then looked at the mean squared error, what problems might we face?

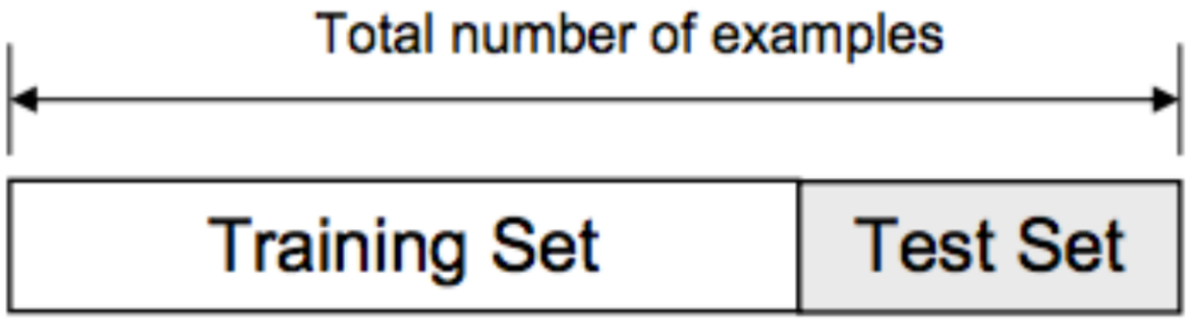
SPLITTING THE DATA: TRAIN/TEST SPLIT

- The Holdout Method: Train/Test Split
- **Training Set:** Used to train model
- **Testing Set:** Used to estimate the error of the model
- **Advantages?**
- **Disadvantages?**



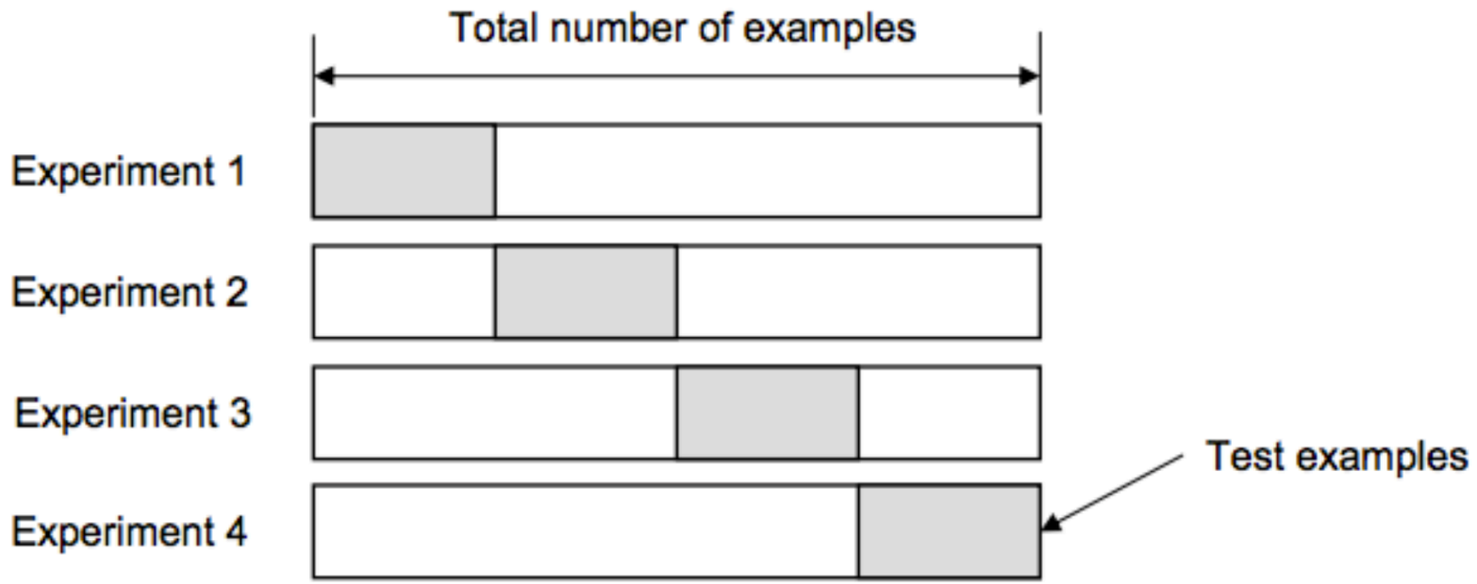
SPLITTING THE DATA: TRAIN/TEST SPLIT

- The Holdout Method: Train/Test Split
- **Training Set:** Used to train model
- **Testing Set:** Used to estimate the error of the model
- **Advantages?** Fast! Simple! Computationally inexpensive!
- **Disadvantages?** Lose data! Imperfect splits! Can't re-test!



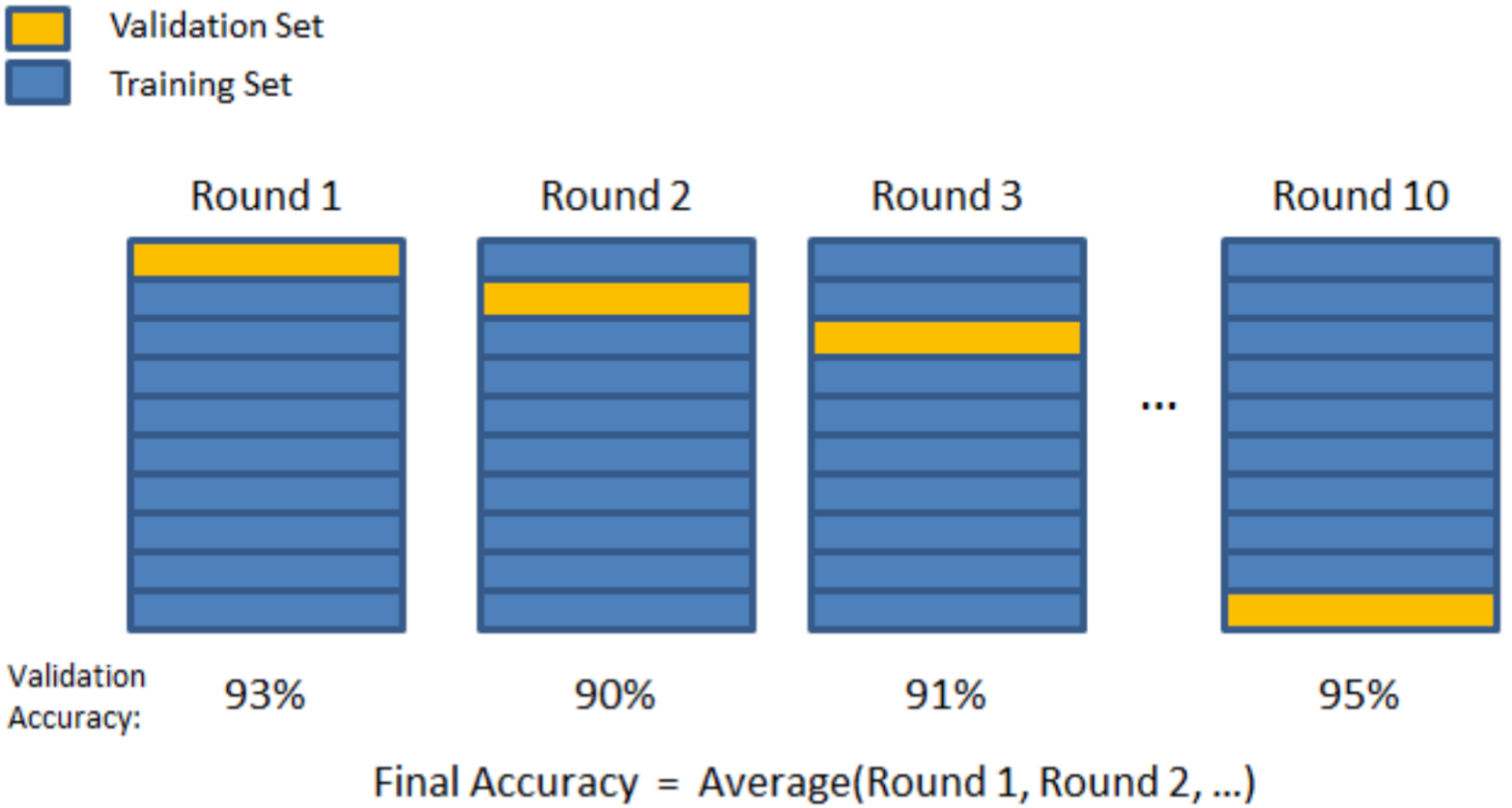
K-FOLDS CROSS VALIDATION

- Split our data into a number of different pieces (folds)
- Train using k-1 folds for training and a different fold for testing
- Average our model over each of those iterations
- Look at stdev of scores



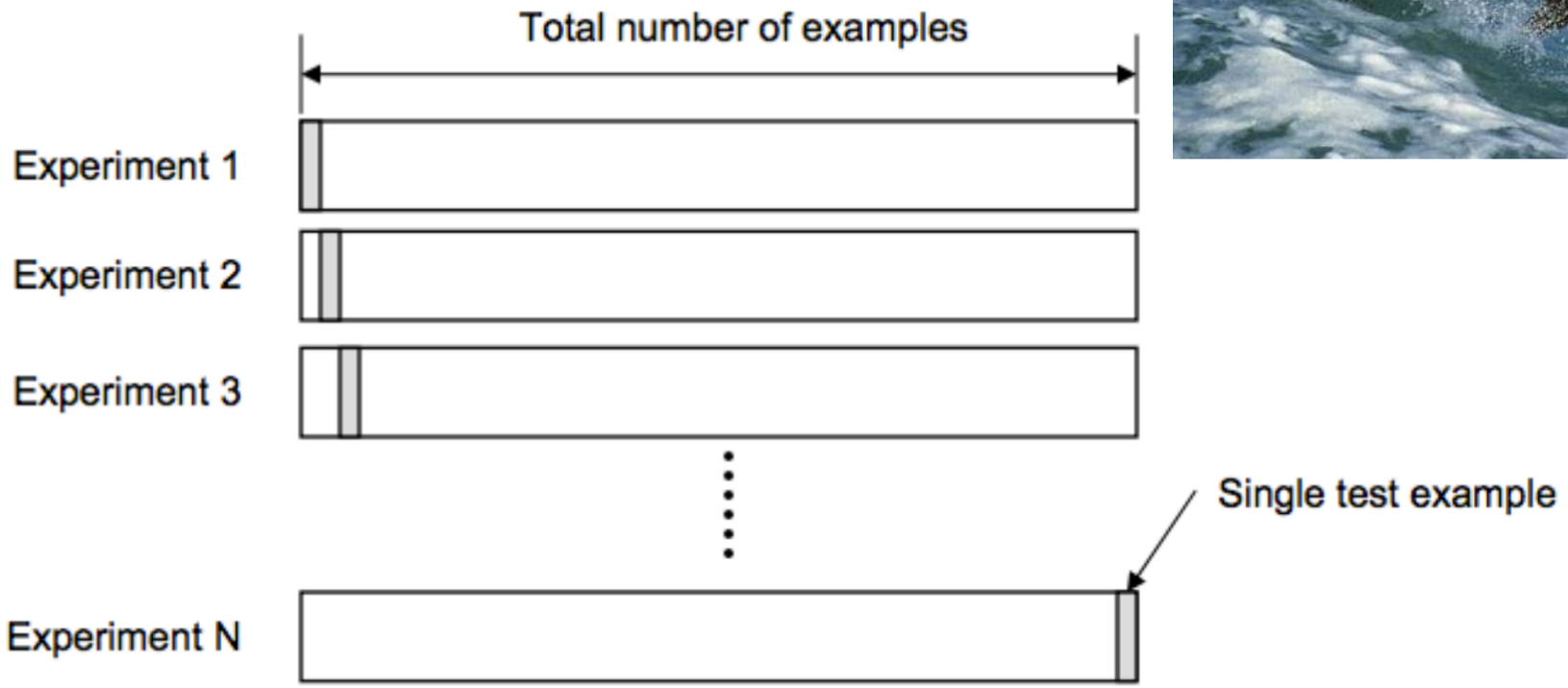
K-FOLDS CROSS VALIDATION

- K=10
- Round 1: Check 9 training sets against one validation set. . . Round 2. . .



LEAVE ONE OUT CROSS VALIDATION (LOOCV)

- K-folds to the EXTREME! K=N
- For dataset of N observations, perform N experiments
- Average our model over each of those iterations



K-FOLDS CV VS. LOOCV

- **K-Folds:**
- More bias
- Less variance
- Computationally cheap
- **LOOCV:**
- Less bias
- More variance
- Computationally expensive
- **In general, K-Folds has less overall test error**
- **In general, we use a K-Folds CV with K=10**

PROCEDURE

- 1. Divide data into training and testing sets
- 2. Select architecture (model type) and training parameters (k)
- 3. Train the model using the training set
- 4. Evaluate the model using the training set
- 5. Repeat 2-4 selecting different architectures (models) and tuning parameters
- 6. Select the best model
- 7. Assess the model with the final testing set

PARTING QUESTIONS

- The demo covers a basic test/train split as well as k-fold cross-validation Check: Is 2-fold cross-validation the same as a 50:50 test/train split?

- Will two different 50:50 (or x:y) splits produce the same model score?