Data and Validation

Machine Learning in Molecular Science

Prof. Michael Shirts

July 10th, 2023

The importance of good data

- GIGO
  - Garbage in, Garbage out
- Sometimes (often?) data is erroneous
- Need to be careful about what you include in your data set
- For LARGE data sets
- Anomaly detection
  - A problem in unsupervised learning

Missing Data?

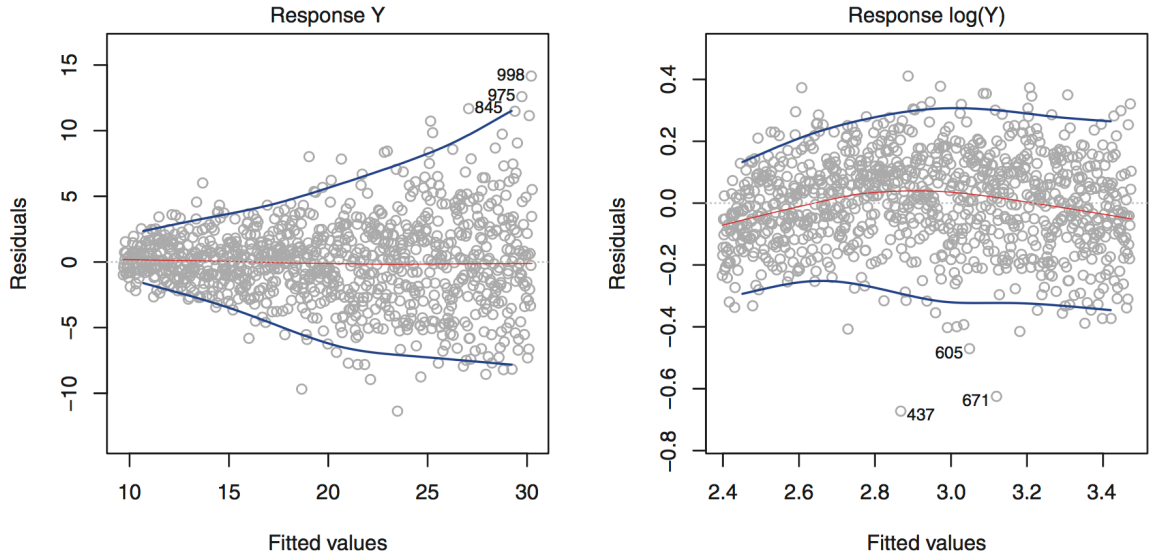
- Data imputation
  - Let's say we have 10,000 data points, and 100 features
  - 3% chance a given feature is missing for every point
  - Only 5.3% of the points have all features, despite 97% of the data being complete.
  - That's leaving a lot of data if we only use ones with all features!
- Strategy
  - Fill in missing data with using "around" it
  - mean of all data
  - Random selection or mean of "close" data
  - Regress from other inputs
  - multiple imputation generates multiple guesses for each missing data point, which can help improve the statistics
- Tools in pandas and scikit-learn

Data Leakage

- Does training data get into the testing set?
- Are some of your features actually surrogates of your labels?
- What is going to happen with ChatGPT 8 when it ends up getting trained on outputs of ChatGPT 4 through 7?

Nonconstant Variance

- This phenomena is known as heteroscedasticity
- One solution: transform data to logarithmic form



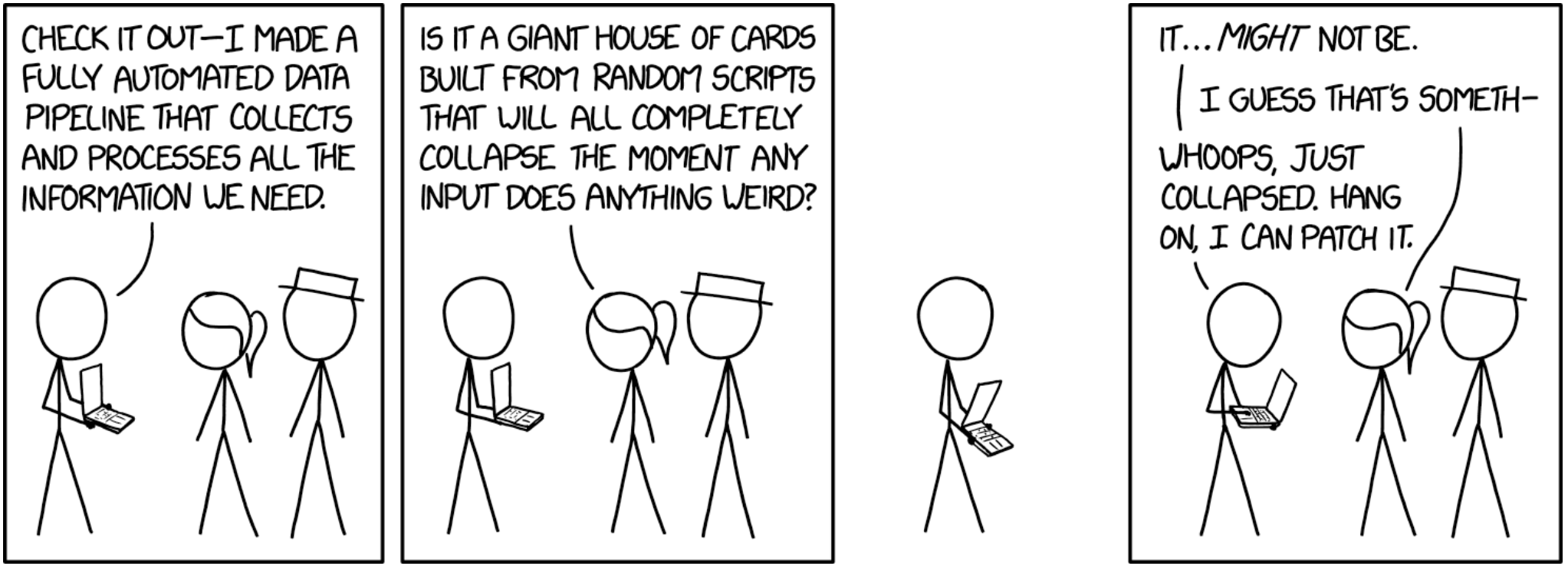
- (Sometimes) Weight the residuals

$$\text{RSS} = \sum_i w_i(y_i - \beta_0 - \beta_1 x)^2$$

- Question: should high variance x's be higher in weight or lower in weight?

Be careufi

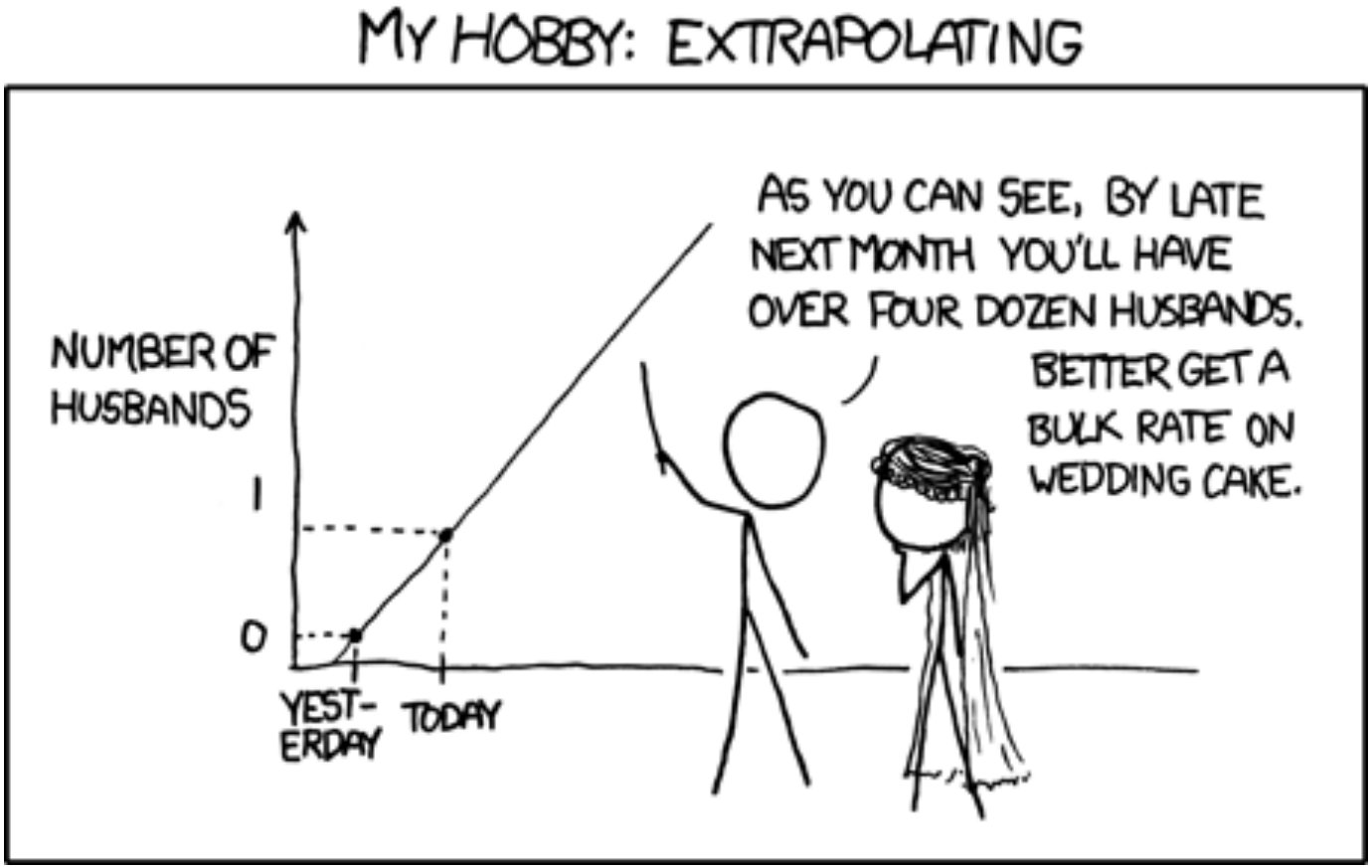
- Two needs:
  - Automating data processing to ensure consistent treatment over the entire data set.
  - Carefully human curation of the data to make sure there's nothing the automation didn't handle.



https://xkcd.com/2054/

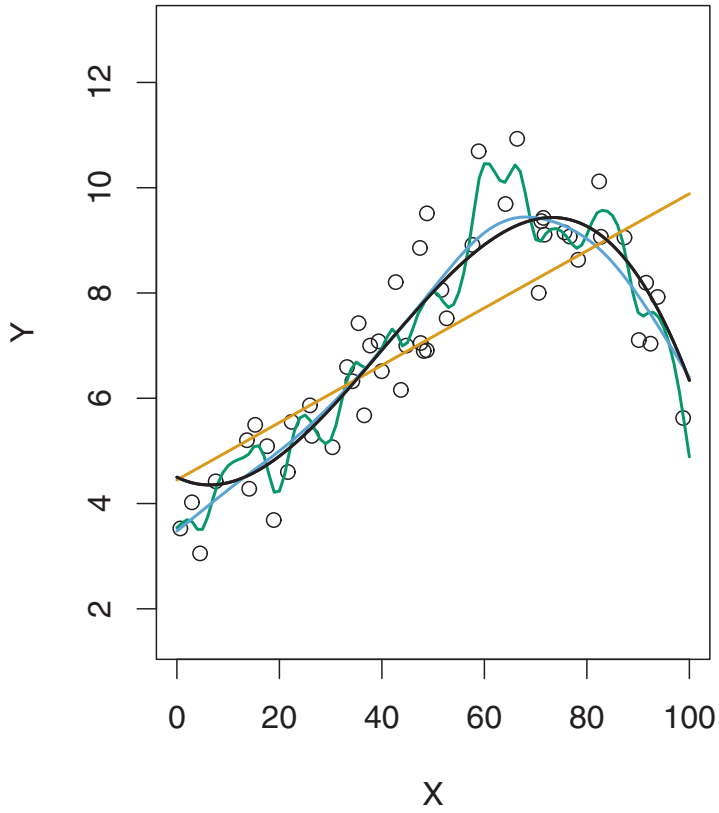
Go to notebook!

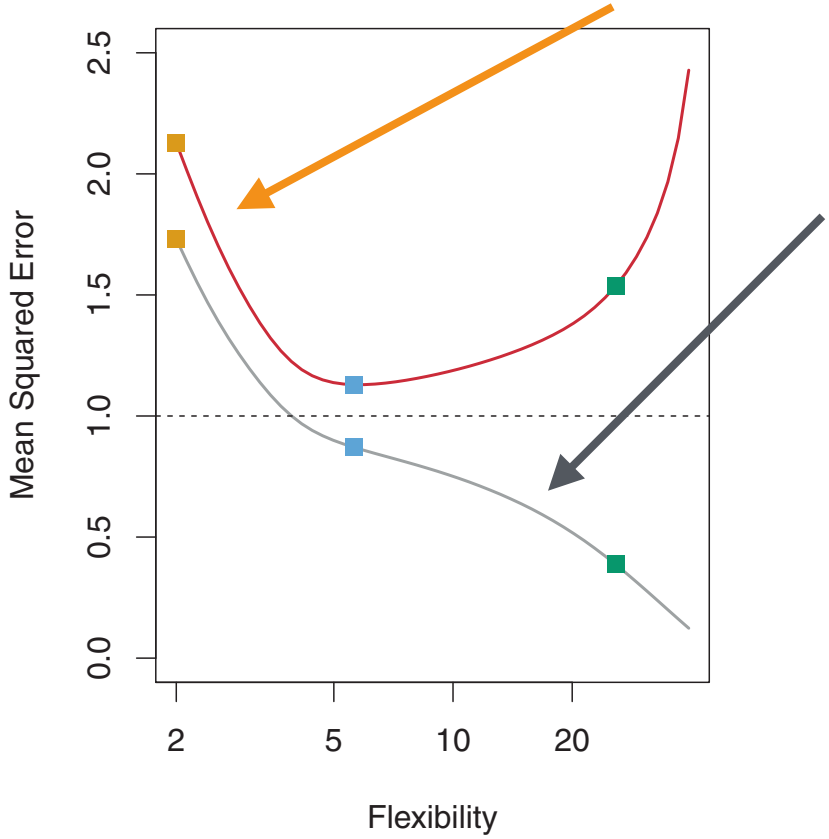
Importance of validating models



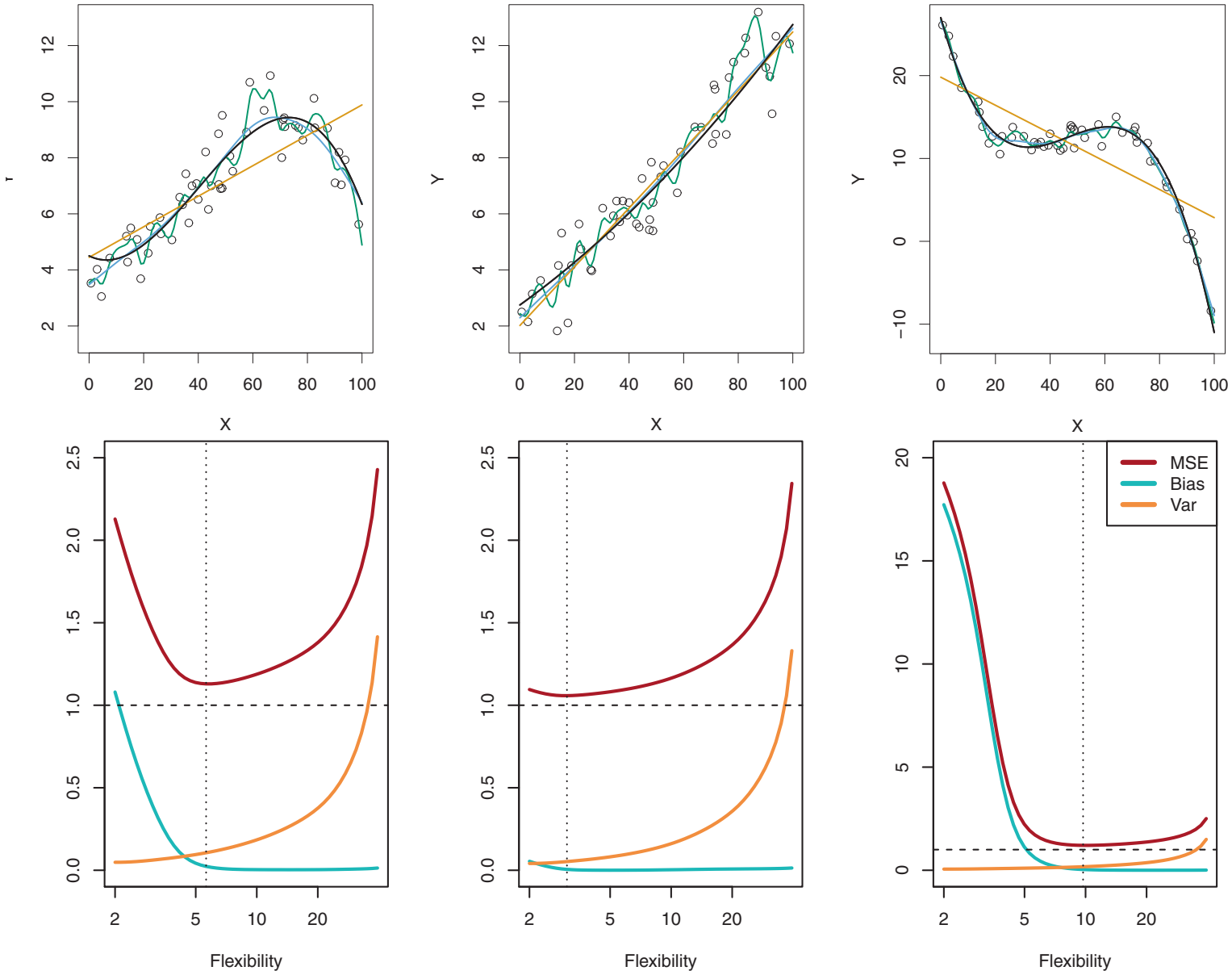
https://xkcd.com/605/

Goal is to minimize OUT OF DATASET MSE, not TRAINING MSE





Different types of data have different bias/ variance tradeoffs



Validation

- Basic idea: treat some of your training data as validation data instead.

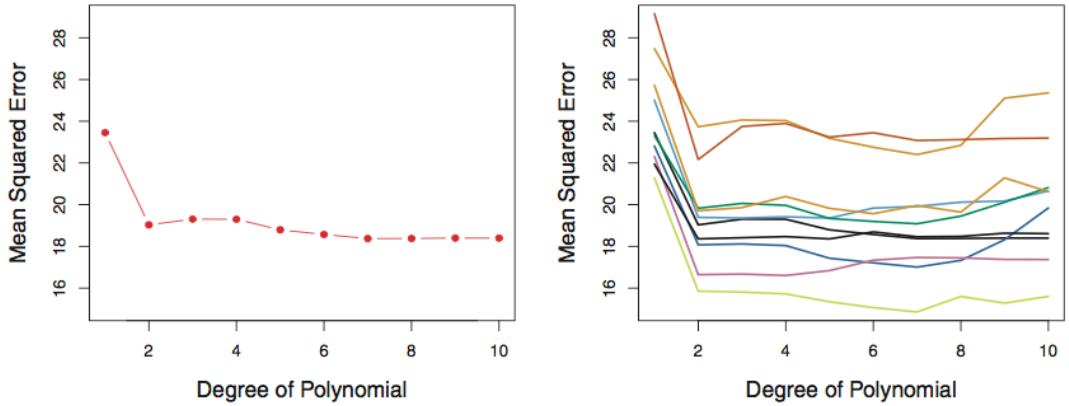
- Three strategies
  - Train on some, validate on some
  - Leave one out cross-validation
  - K-fold cross-validation

Train some, test some

Train

Test

- Problems:
  - We are leaving some data out of the training
  - How do we make the decisions of how much to do?
  - How do we decide which set of data to reserve for testing?
  - It's a noisy process!



Leave One Out Cross Validation (LOOCV)

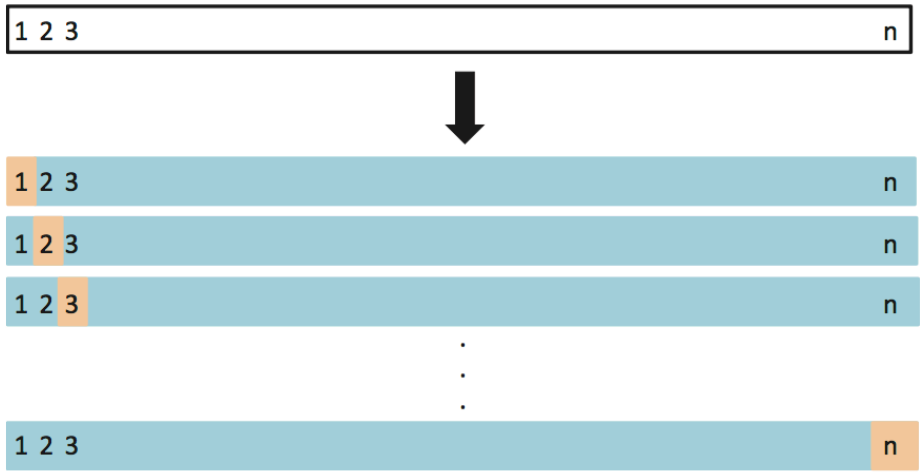
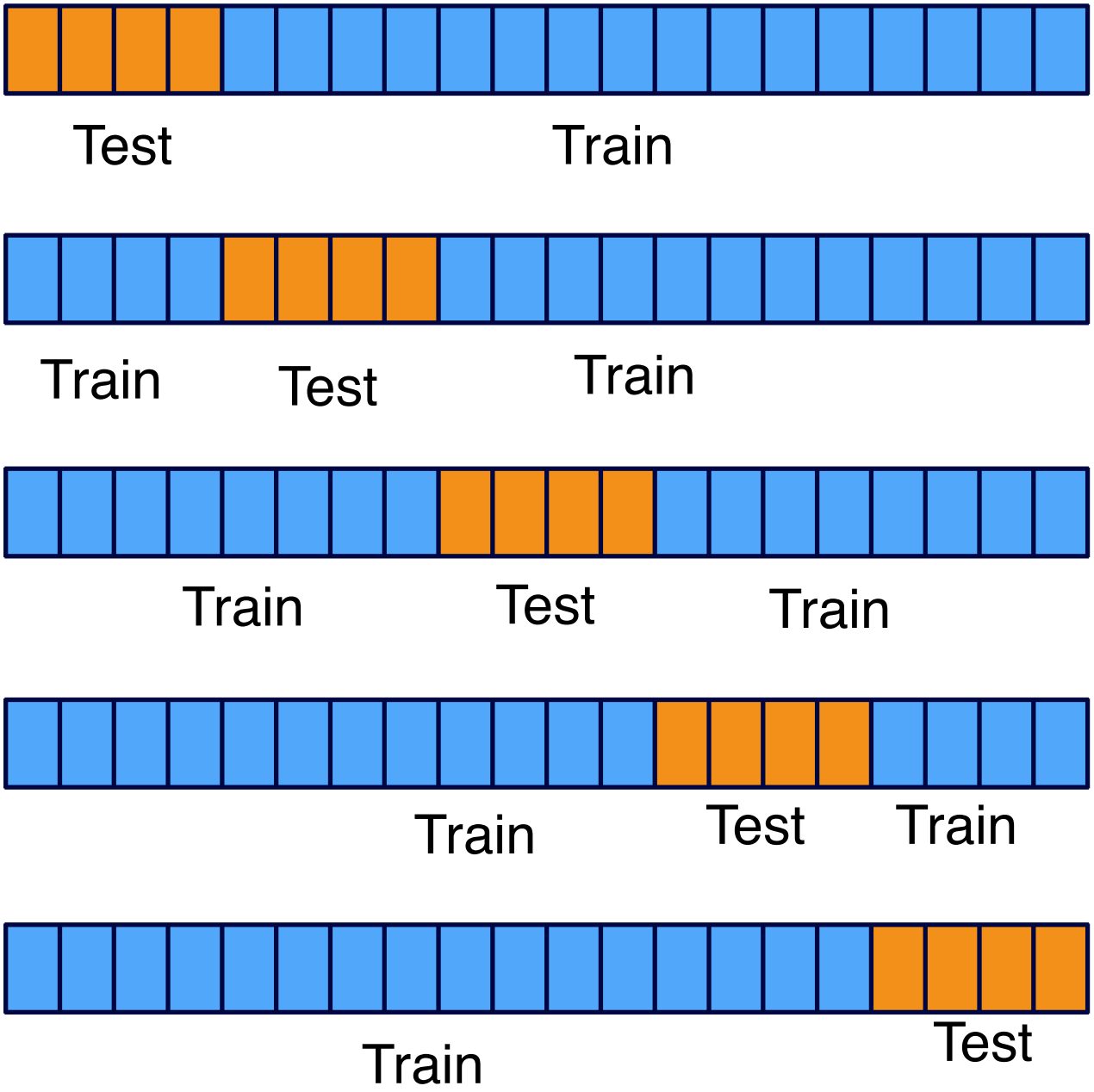


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

Each test is ONE prediction
Report the average over the N tests

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

K-fold Cross Validation



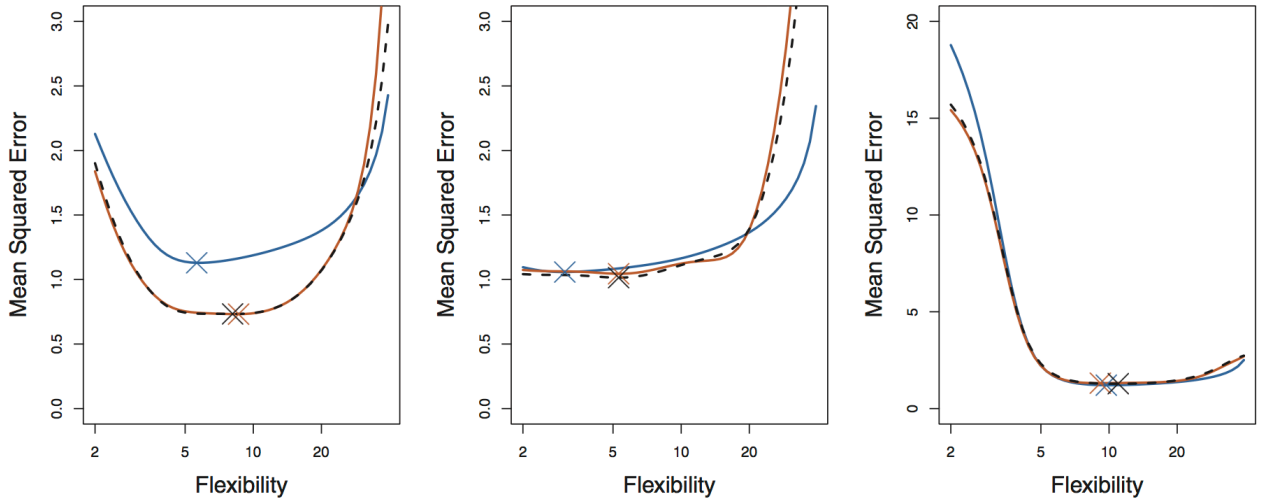
Example: 5 fold cross validation

- Average the five training and test MSEs.
- Report the average

$$\mathrm{CV}_{(k)} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{MSE}_i$$

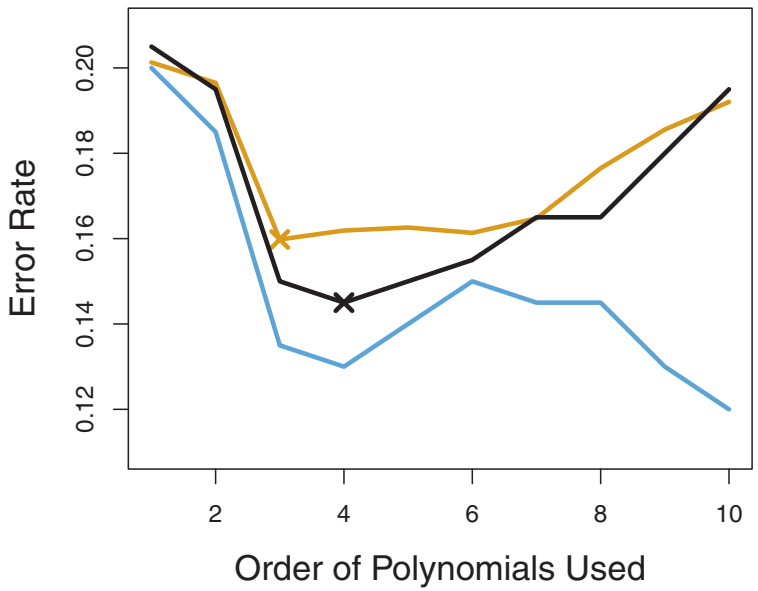
Bias/variance tradeoff: use 5 or 10 folds

- Leave-one-out cross validation is N-fold
  - Always gives the same answer
  - Can be very expensive for large data sets!
- Empirically, 5 fold and 10 fold provide a good bias/variance tradeoff.
- 10 fold is ~N/10 times cheaper than LOOCV!

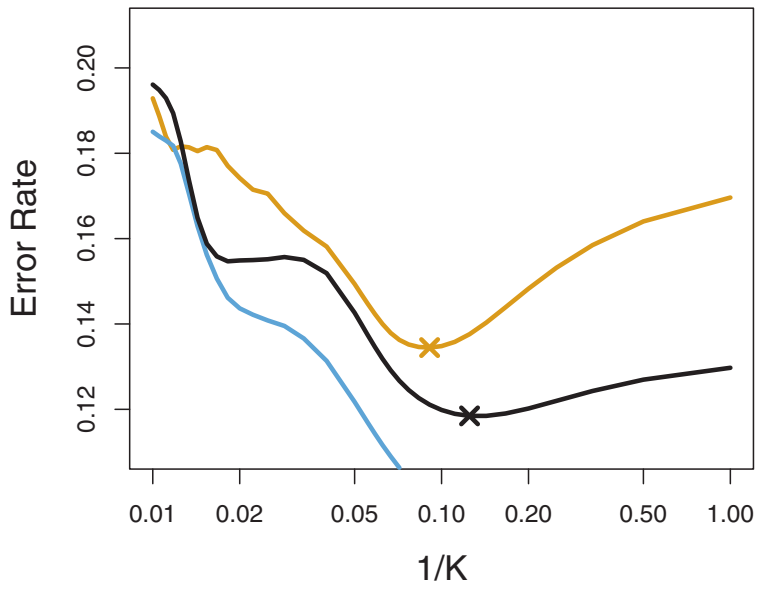


Works for classification as well:
Works for all supervised learning

Logistic regression with polynomial boundary



K-nearest neighbors



Blue: training set error
Orange: validation set error
Black: 10-fold cross validation

Reminder:

- Training: Used to train your model
- Validation: Used to estimate out-of-data set performance, in order to tune hyperparameters
- Test: FRESH data, separate from validation to see how good tuned model is.

- If you are iteratively changing your model based on the test data set, it's actually the validation data set.

To the workbook!