Model complexity and generalization

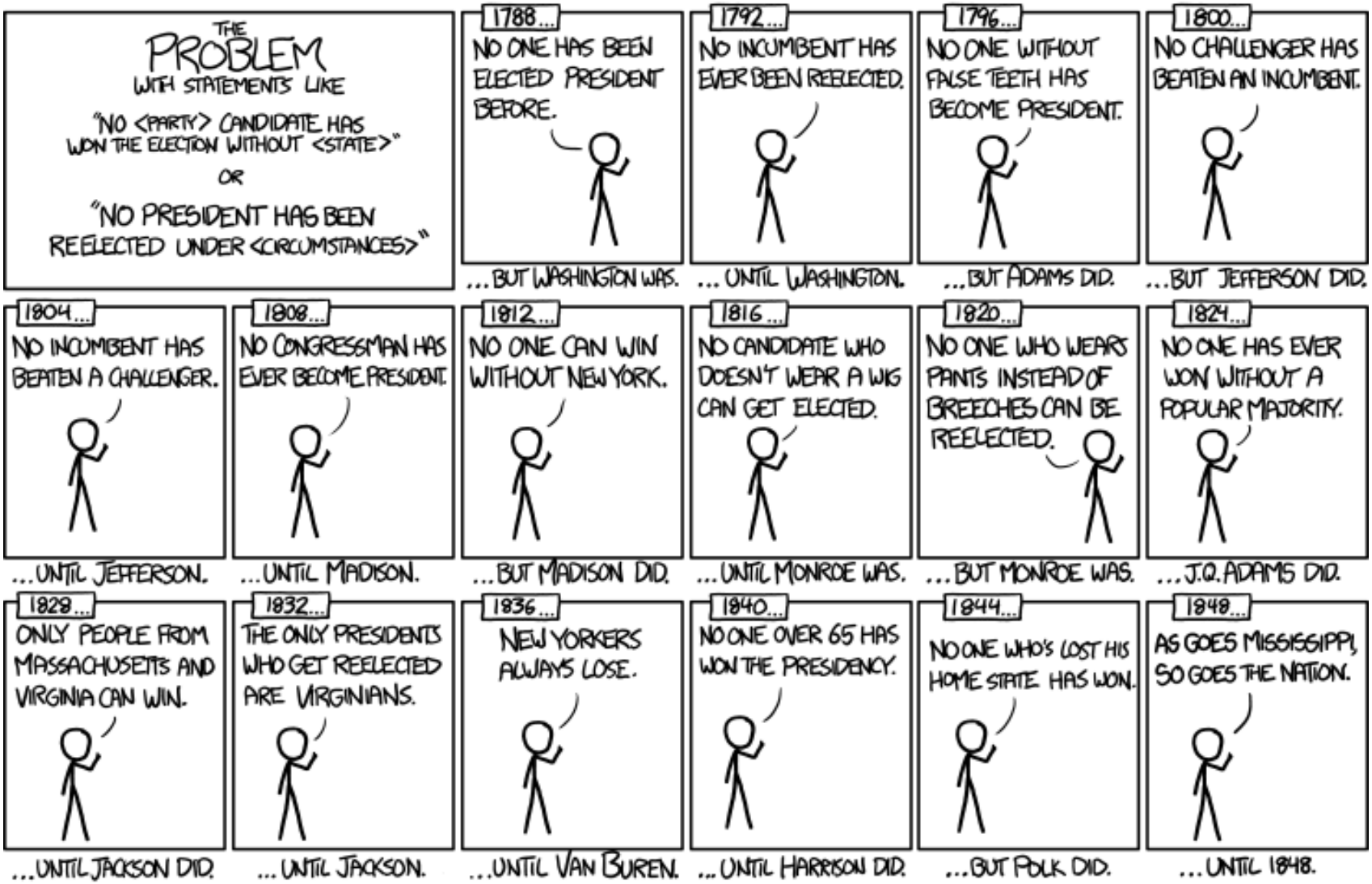
APAM E4990
Modeling Social Data

Jake Hofman

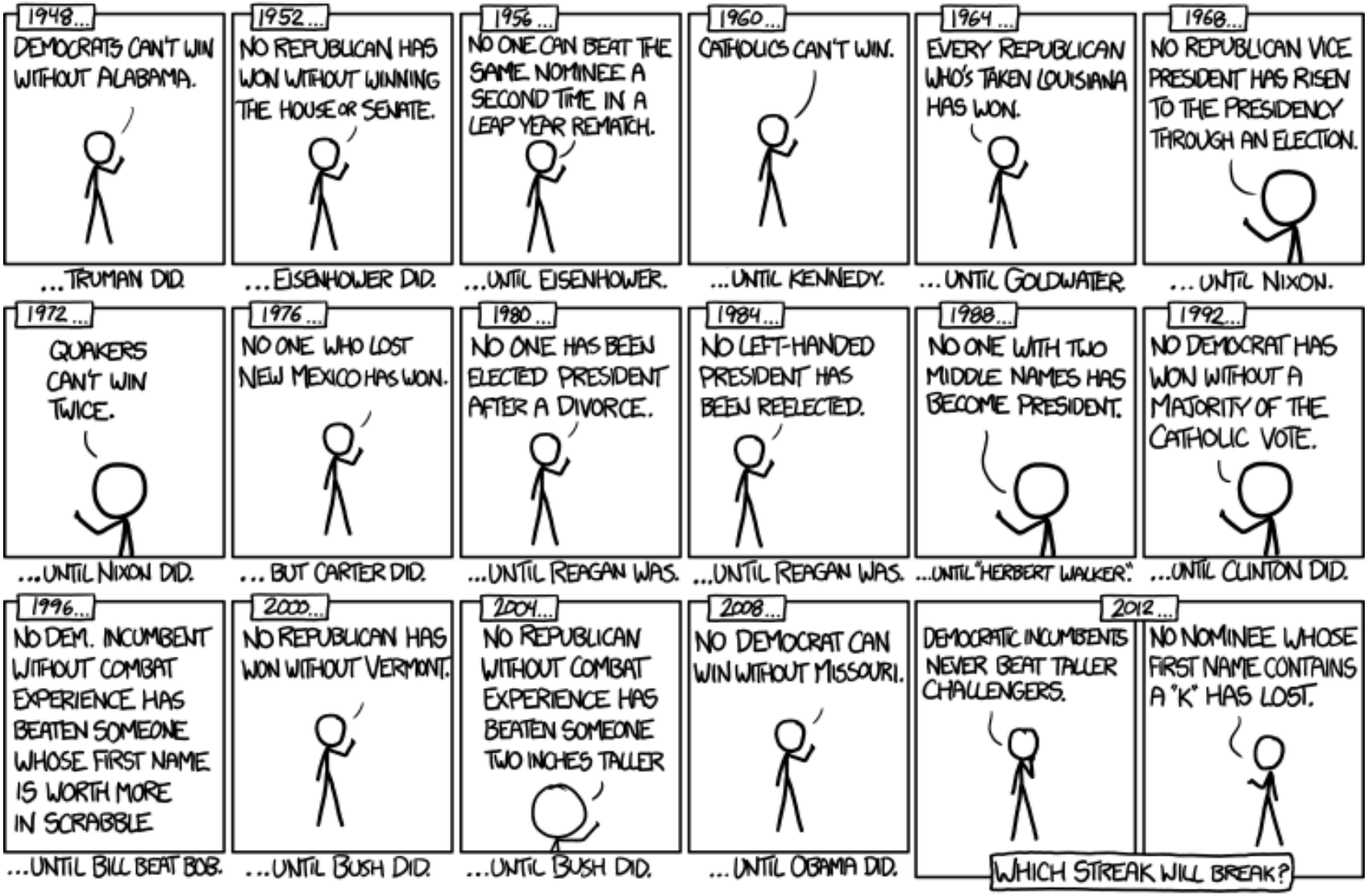
Columbia University

March 15, 2019

Overfitting (a la xkcd)

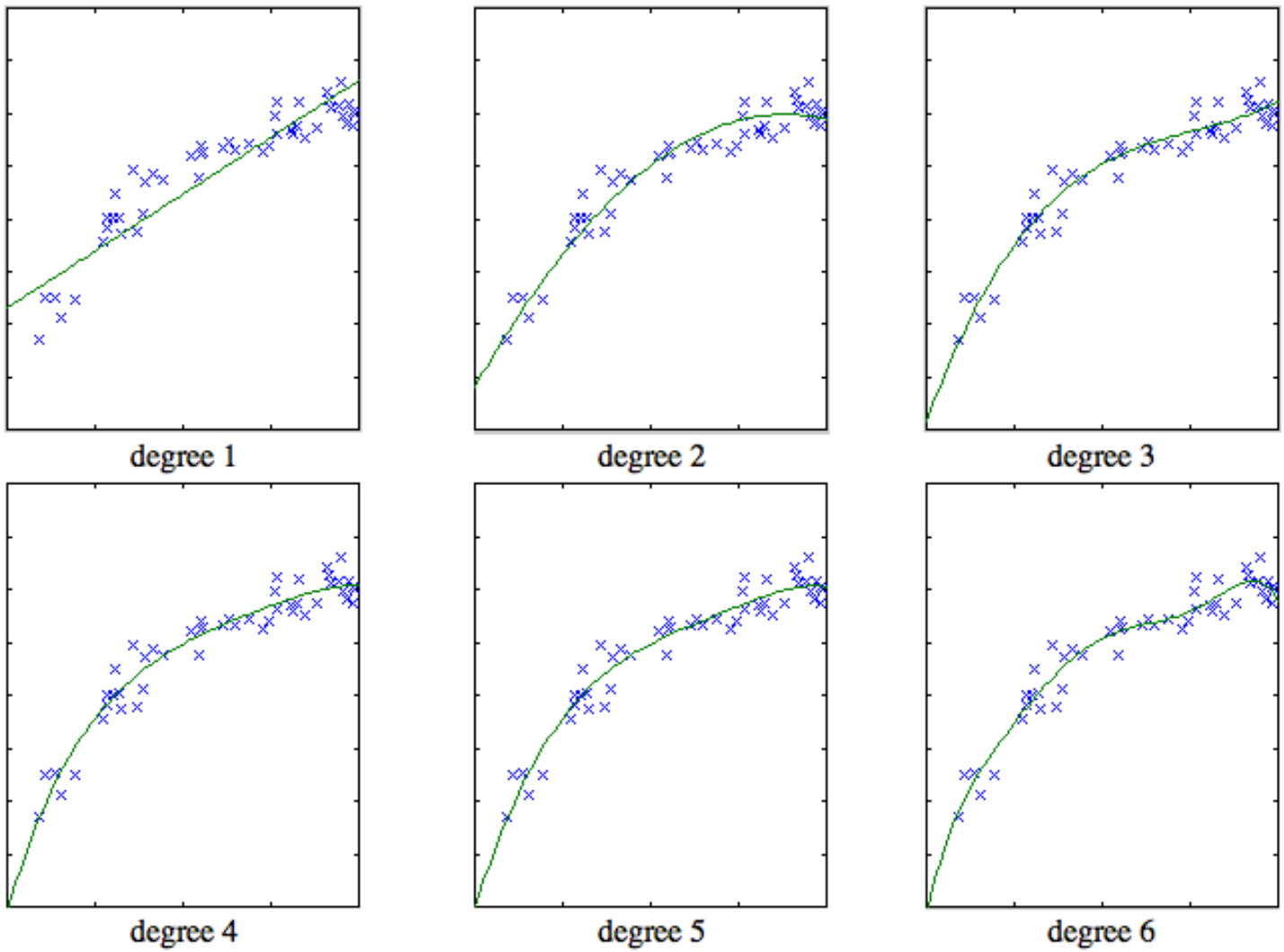


Overfitting (a la xkcd)

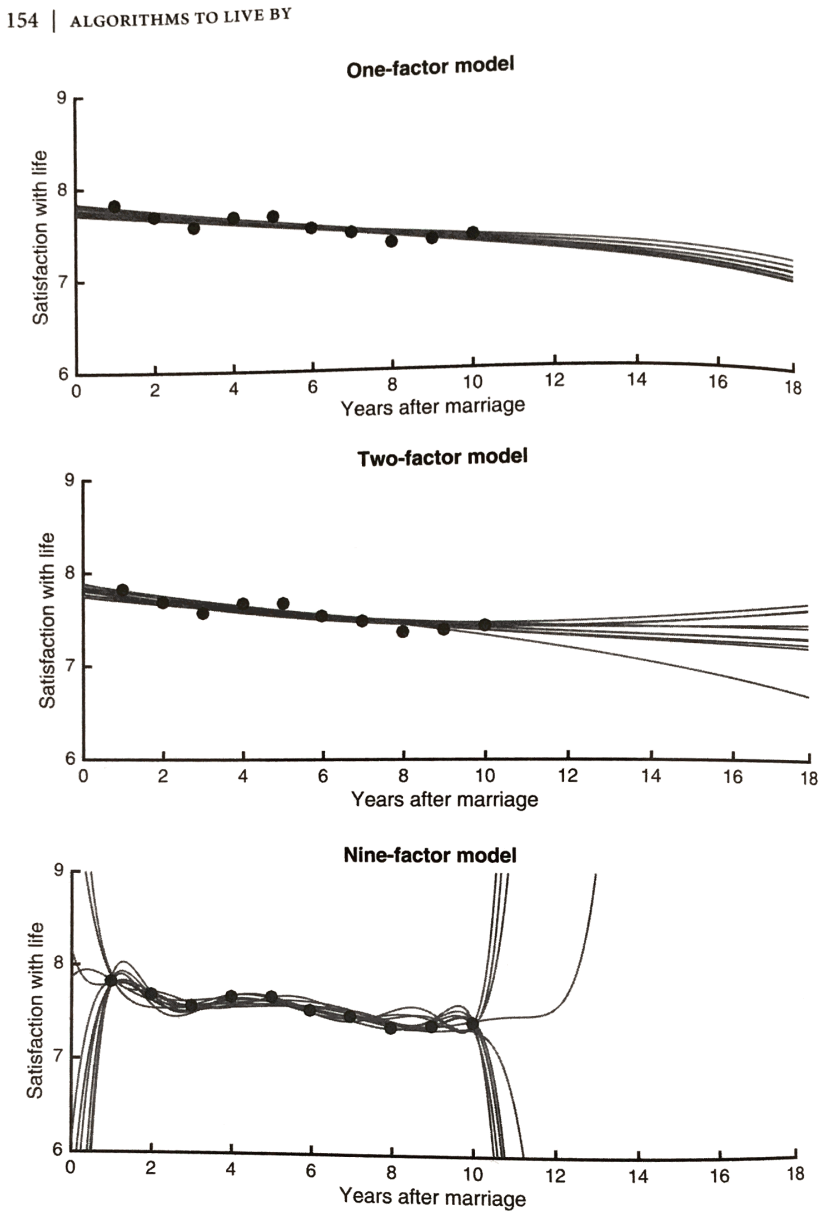


Complexity

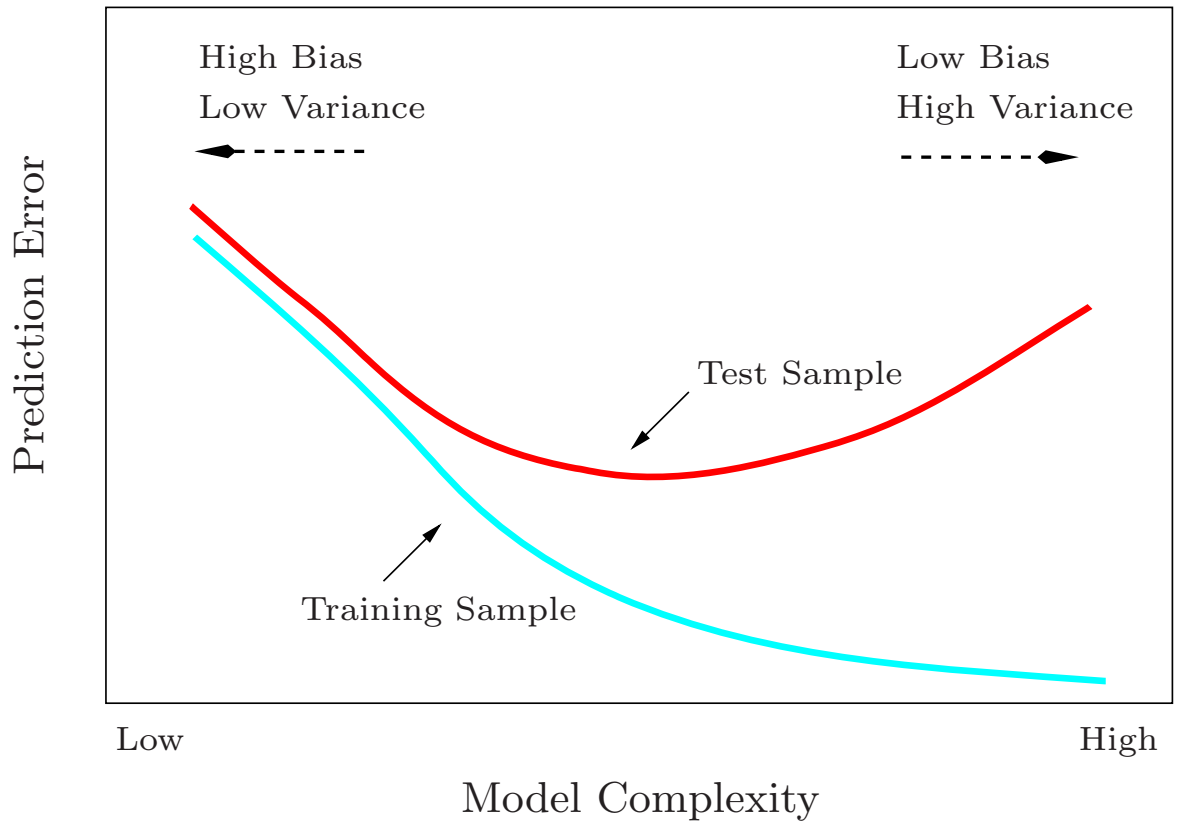
Our models should be complex enough to explain the past, but simple enough to generalize to the future



Bias-variance tradeoff

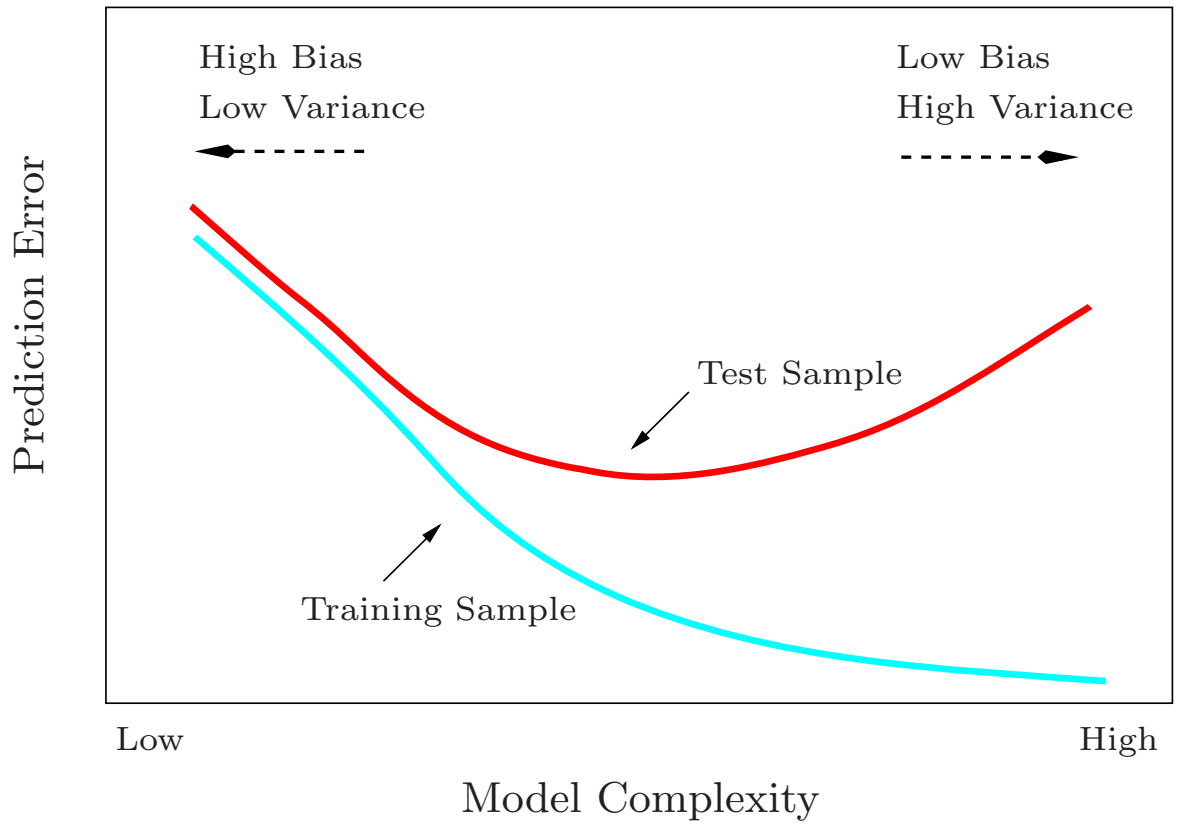


Bias-variance tradeoff



Simple models may be "wrong" (high bias), but fits don't vary a lot with different samples of training data (low variance)

Bias-variance tradeoff



Flexible models can capture more complex relationships (low bias), but are also sensitive to noise in the training data (high variance)

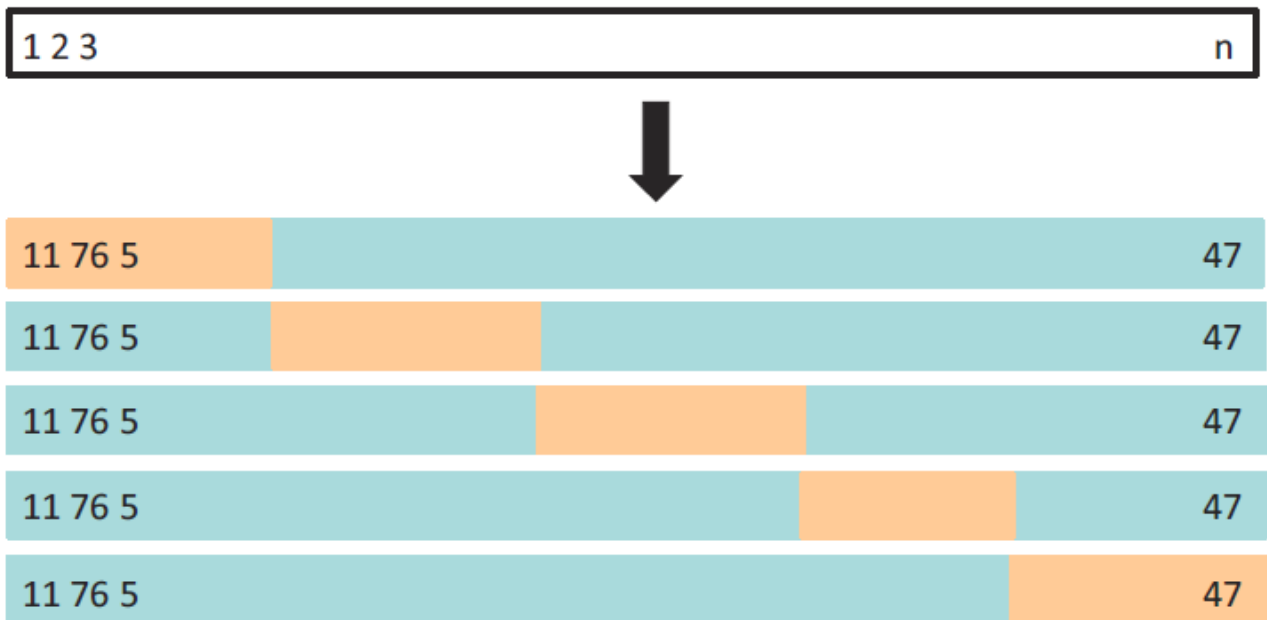
Bigger models $\neq$ Better models

Cross-validation

- Randomly split our data into three sets
- Fit models on the training set
- Use the validation set to find the best model
- Quote final performance of this model on the test set

K-fold cross-validation

Estimates of generalization error from one train / validation split can be noisy, so shuffle data and average over K distinct validation partitions instead



K-fold cross-validation: pseudocode

```
(randomly) divide the data into K parts

for each model

  for each of the K folds

    train on everything but one fold
    measure the error on the held out fold
    store the training and validation error

  compute and store the average error across all folds

pick the model with the lowest average validation error

evaluate its performance on a final, held out test set
```