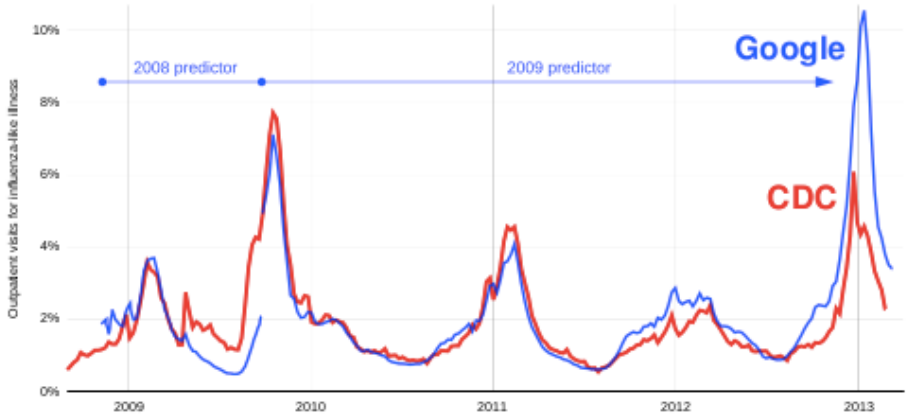
Training and Validation

Professor Kartik Hosanagar

Over-fitting the Data: Google Flu Trends

Second divergence in 2012–2013 for U.S.



Wharton
University of Pennsylvania

Agenda

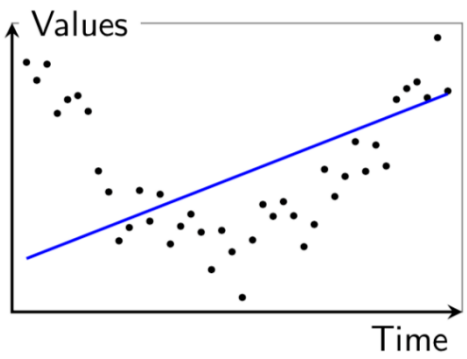
- **Training and Validation**
  - Overfitting
  - Training and Validation
  - Bias-variance trade-off
- **Validation Strategies**
  - Split Training and Testing
  - K-fold Cross Validation
  - Iterative Cross Validation

Training and Validation

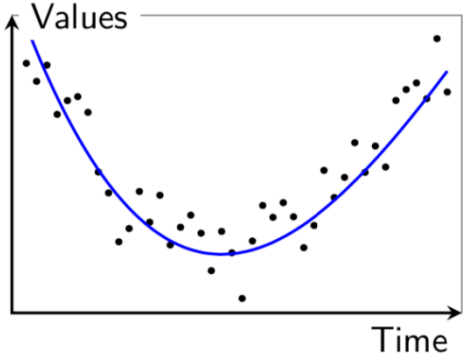
- What is overfitting?
- Why overfitting is critical?
- How to avoid overfitting?
  - Training and validation
- How to find the optimal point?
  - Bias-variance trade-off

Overfitting

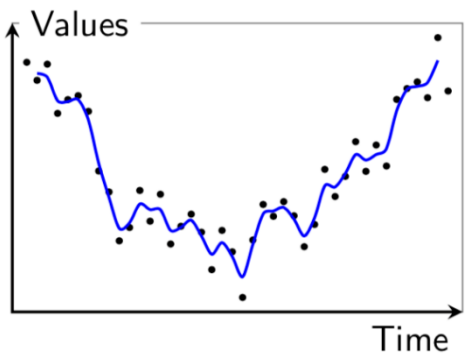
- “the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably".



Underfitted



Good Fit

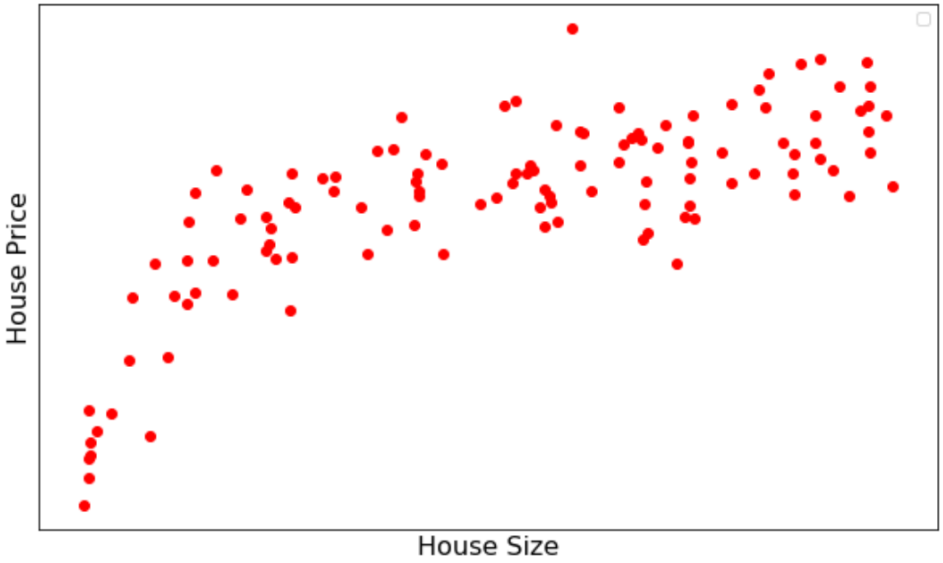


Overfitted

https://tex.stackexchange.com/questions/485900/plotting-over-underfitting-graph-with-tikz-package
Definition of "overfitting" at OxfordDictionaries.com: this definition is specifically for statistics.

Example: Real Estate Price

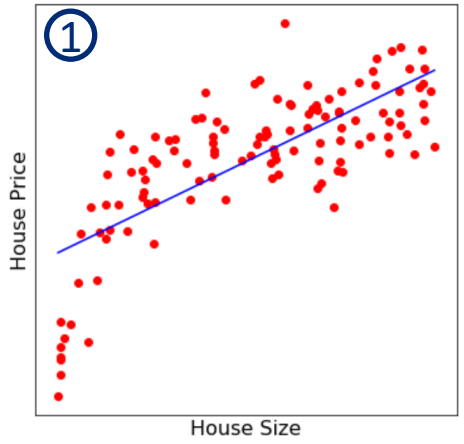
- Real estate company investigated the price of houses depending on size.

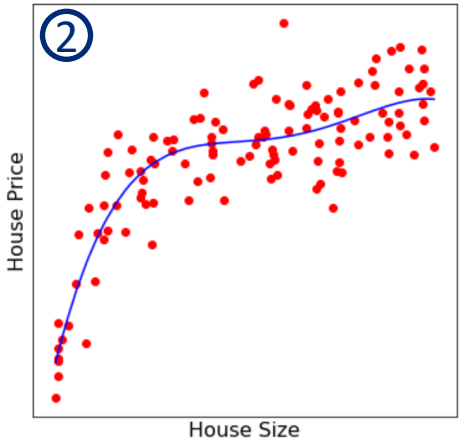


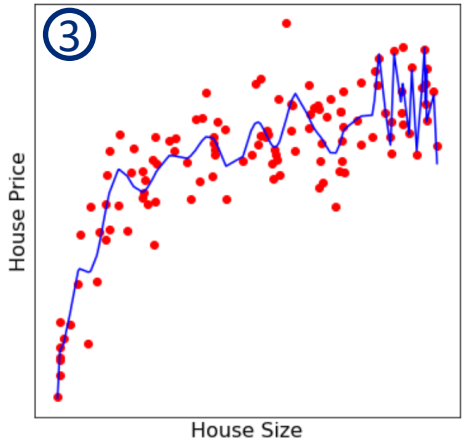
Wharton
University of Pennsylvania

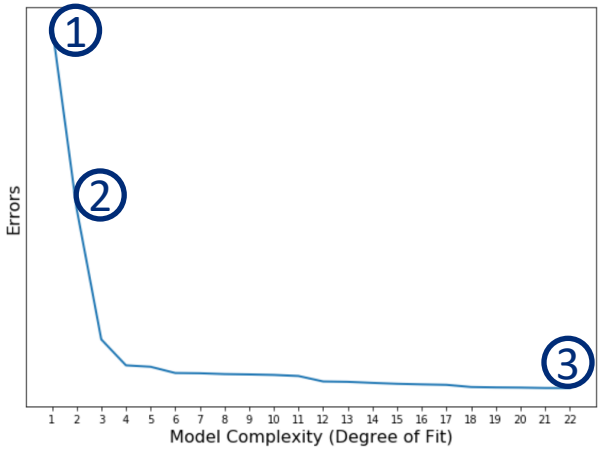
Example: Real Estate Price

- The company found that more complex models show better performance.









The error of their Polynomial Regression model decreases as the number of dimension increases.

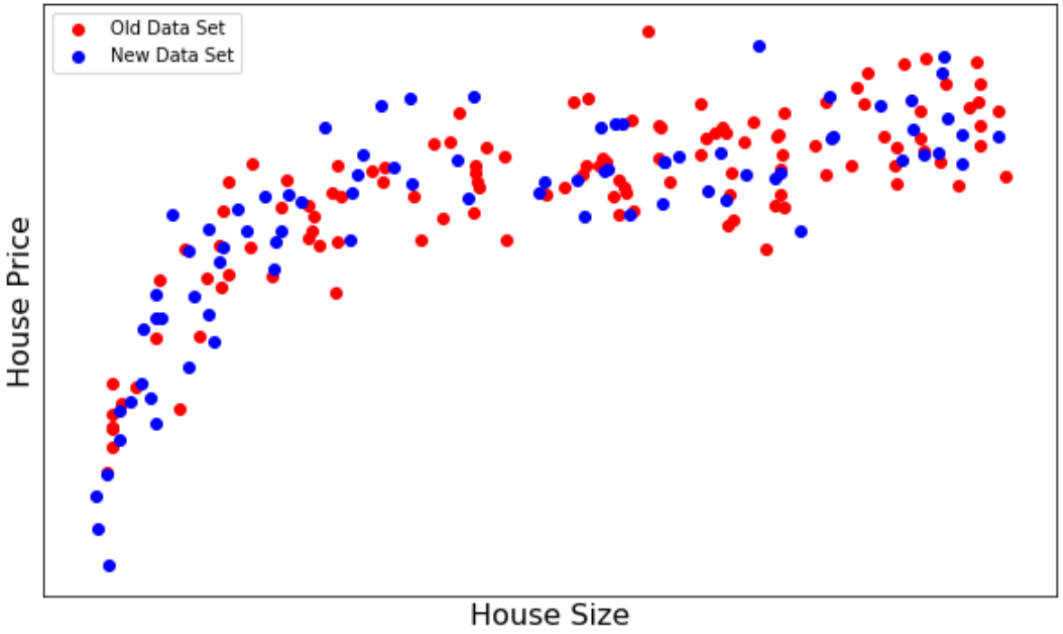
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

So, can we use the most complexed model ③?

Wharton
University of Pennsylvania

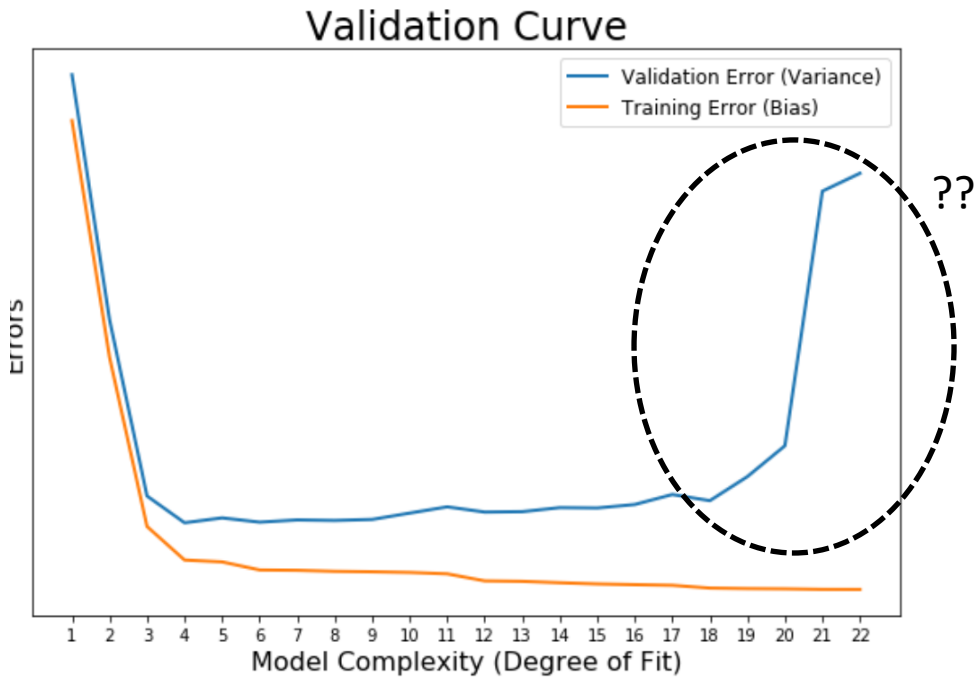
Example: Real Estate Price

- The company collected more data to validate the fitted models.
- The fitted models with old dataset will be validated with new dataset.



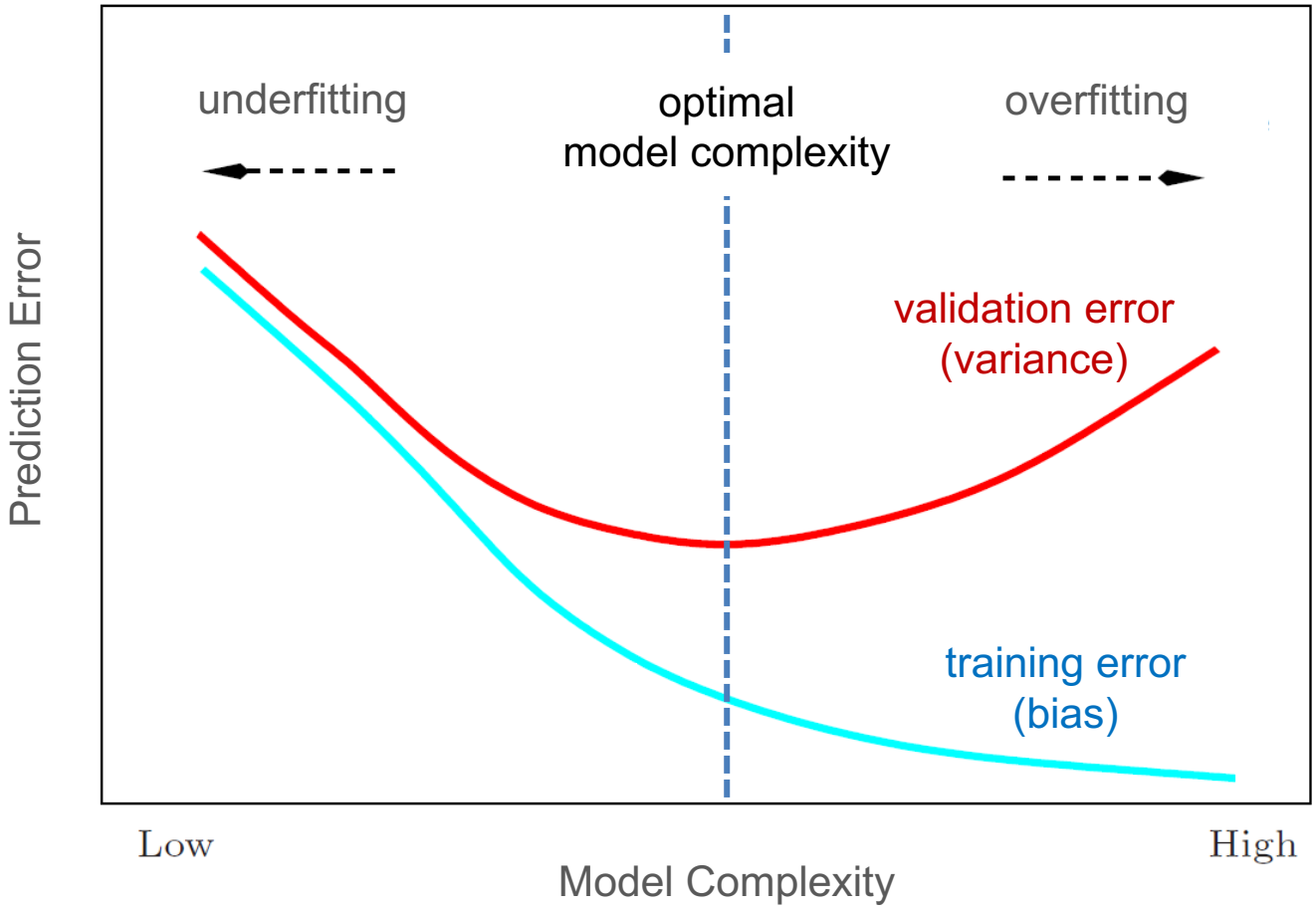
Example: Real Estate Price

- However, their best model could not predict the new dataset properly.
- It's because of 'Overfitting'.



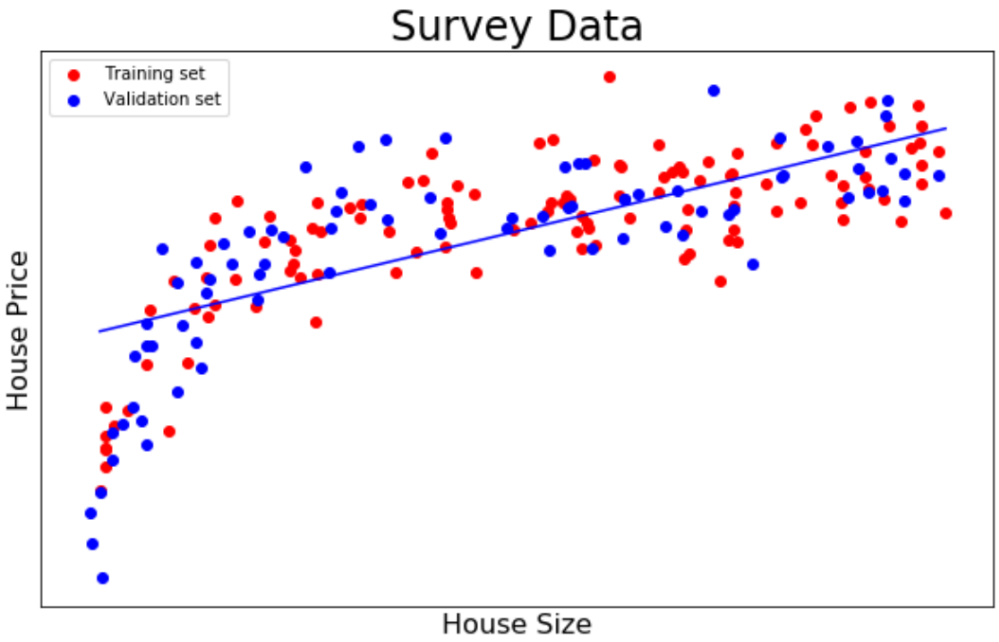
Wharton
University of Pennsylvania

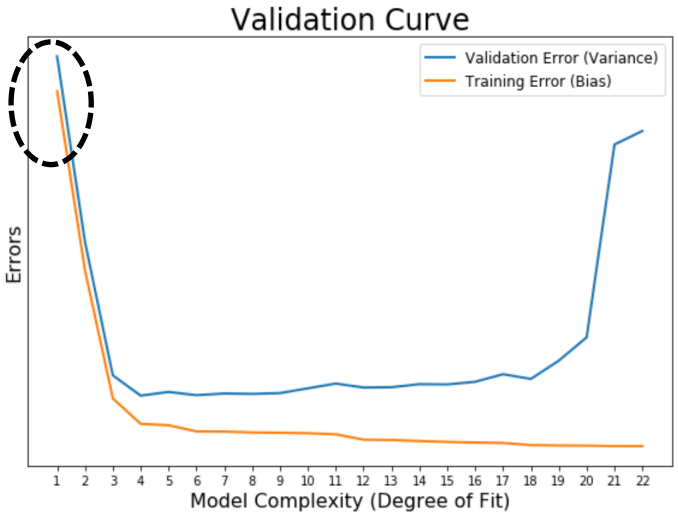
* Bias-Variance Trade-off



Underfitted Model (d=1)

- High training error (Bias)
- High validation error (Variance)
- Highly generalizable

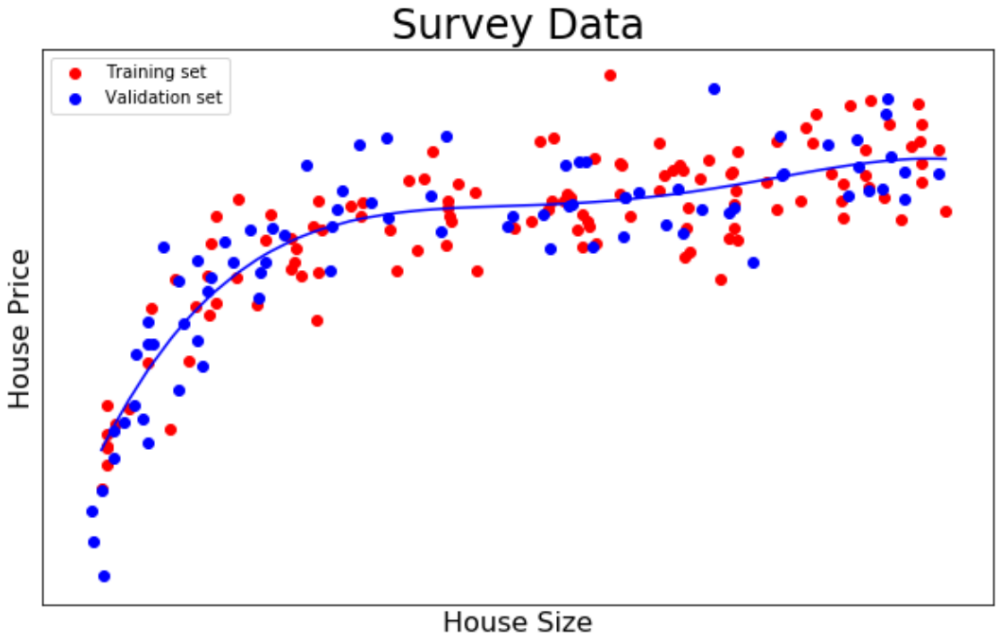


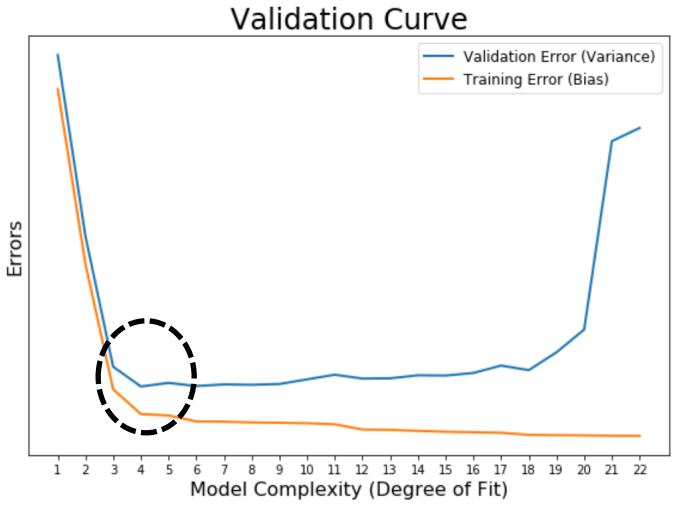


Wharton
University of Pennsylvania

Properly Fitted Model (d=4)

- Low training error (Bias)
- Low validation error (Variance)
- Properly generalizable

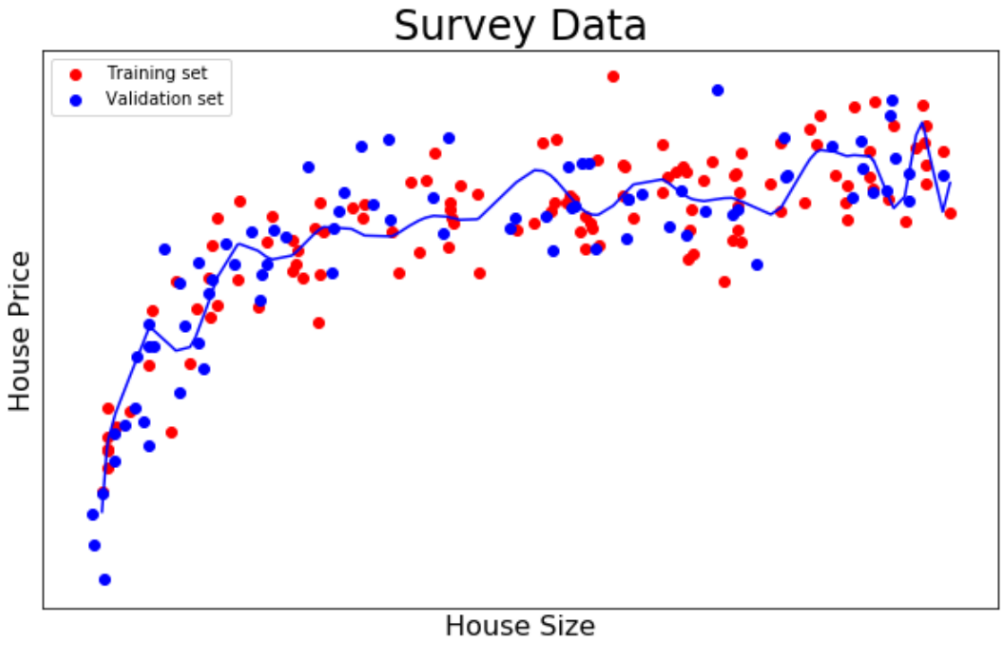


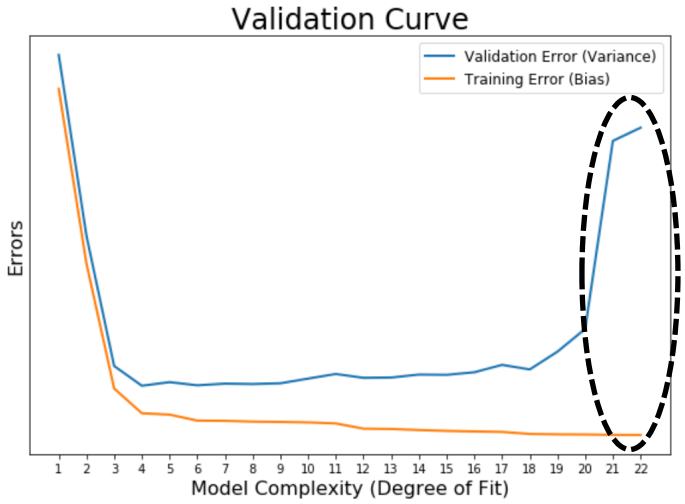


Wharton
University of Pennsylvania

Overfitted Model (d=22)

- Low training error (Bias)
- High validation error (Variance)
- Poorly generalizable





Wharton
University of Pennsylvania

Validation Strategies

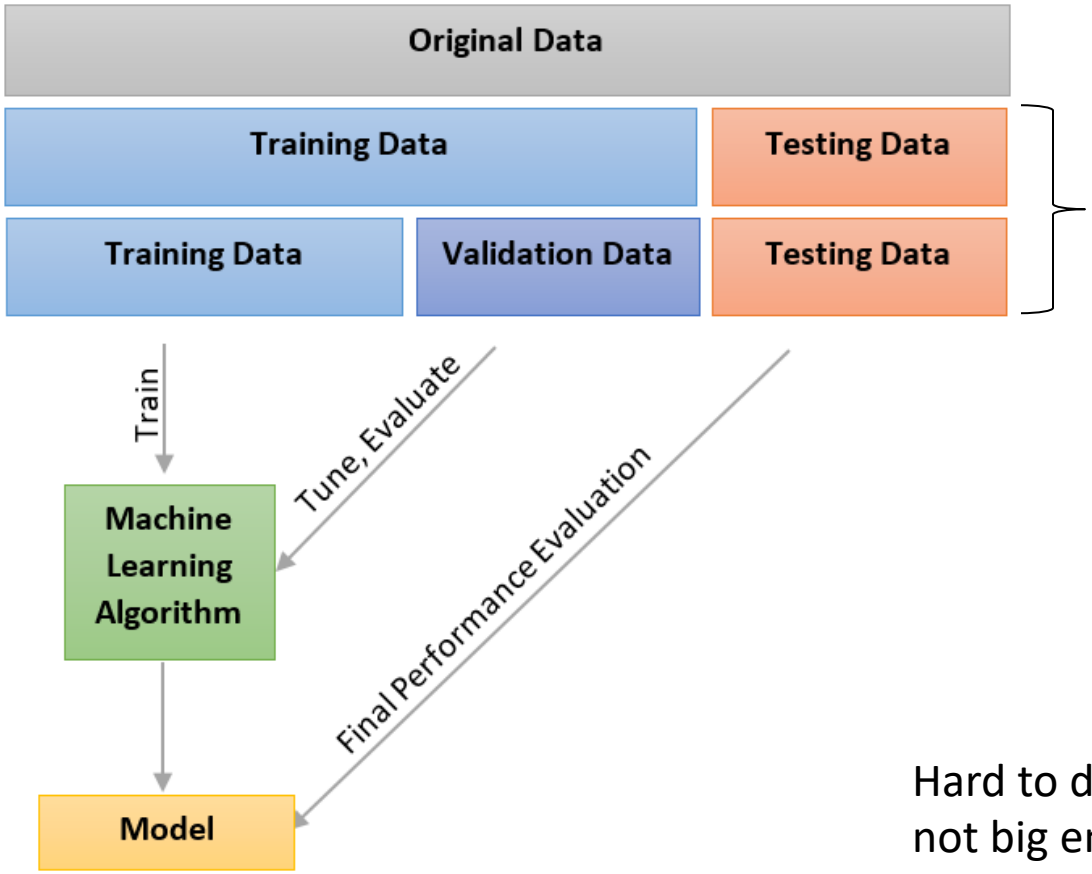
How to validate the model?

- Split Training and Testing
- K-fold Cross Validation
- Iterative Cross Validation

Model Evaluation

- Model evaluation is done by evaluating model performance on a validation dataset
  - Holdout validation: Partition available data into a training dataset and a holdout; evaluate model performance on holdout
  - Cross-validation: create a number of partitions (validation datasets) from the training dataset; fit model to the training dataset (sans the validation data); evaluate model against each validation dataset; repeat with each validation set and average results to obtain the cross-validation error.

Split Training and Testing



Hard to do when dataset is not big enough

https://www.freecodecamp.org/news/how-to-get-a-grip-on-cross-validations-bb0ba779e21c/

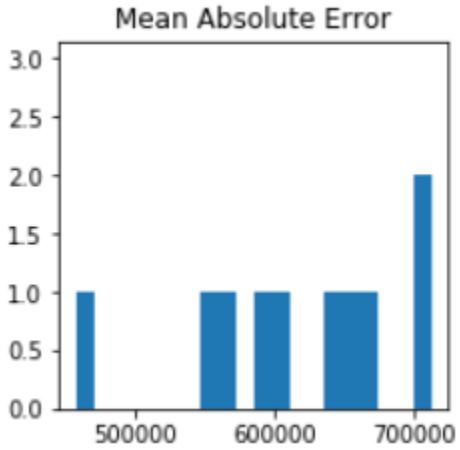
K-Fold Cross Validation

http://karlrosaen.com/ml/learning-log/2016-06-20/

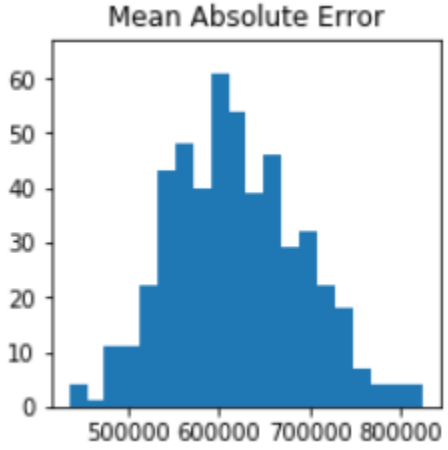
Wharton
University of Pennsylvania

Iterative CV

- Control unstable results (= large variance among training results) due to the poor quality of dataset.
- Use bootstrapping or iterative random splits.
- From ‘Deterministic’ to ‘Stochastic’.



Split Data into 7:3
randomly 10 times



Split Data into 7:3
randomly for 1,000 times

We can use
Mean and Variance
of the distribution

Wharton
University of Pennsylvania

ML in Practice

- We will do a code walk-through on Google Colab
- Colab is a Jupyter notebook environment to run ML code in the cloud
  - Jupyter notebook: open-source application to create documents with live code in them


Wharton
University of Pennsylvania