outraggeously AMBITIOUS

Module 2: The Modeling Process

Duke
PRATT SCHOOL of ENGINEERING

Module 2 Objectives:

At the conclusion of this week, you should be able to:

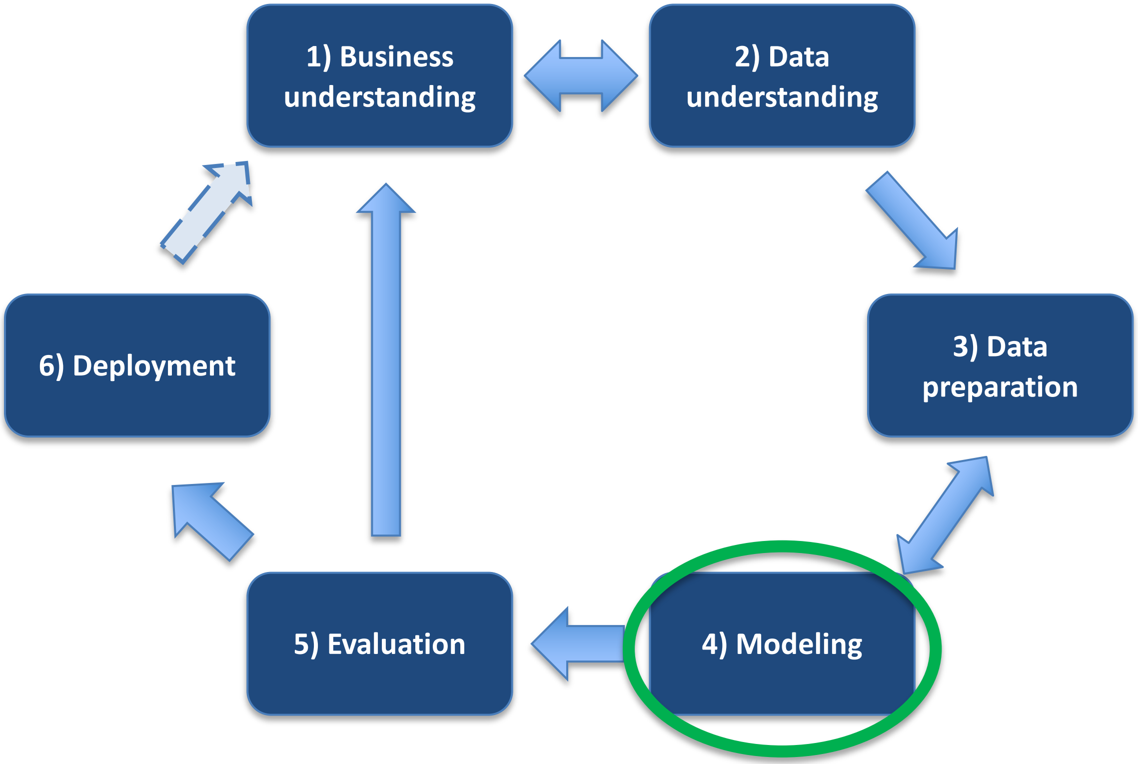
1) Describe the steps to develop a ML model
2) Explain the bias-variance tradeoff
3) Identify possible sources of data leakage and strategies to prevent it

outrageously
AMBITIOUS

Building a Model

Duke
PRATT SCHOOL *of*
ENGINEERING

CRISP-DM Process



Creating a Model

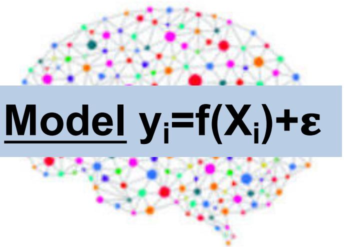
Past Observations

	Neighbor-hood	School district	Square footage	Number of bedrooms	Year built
House 1	Weycroft	Wake	3400	4	2010
House 2	Horton Creek	Wake	4200	5	2008
House 3	Cary Park	Chatham	3250	4	2012
…	…	…	…	…	…

Training

Targets

Market sale price
$612,000
$675,000
$520,000
…

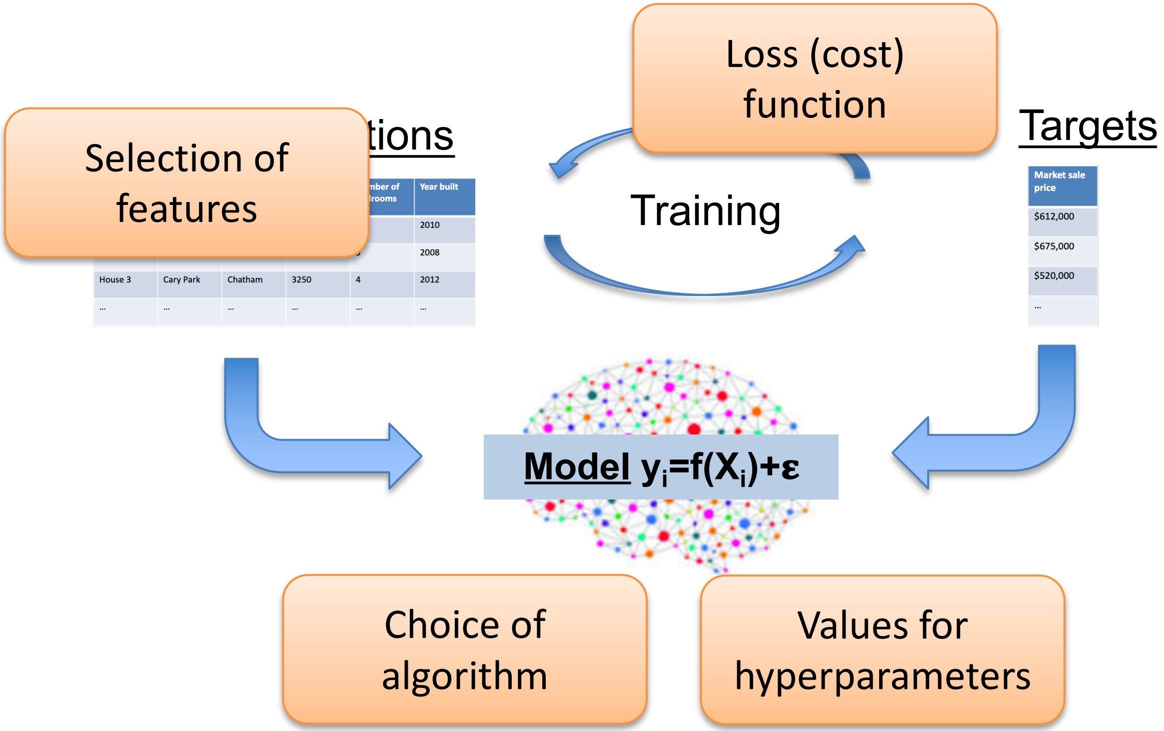


Model $y_i = f(X_i) + \varepsilon$

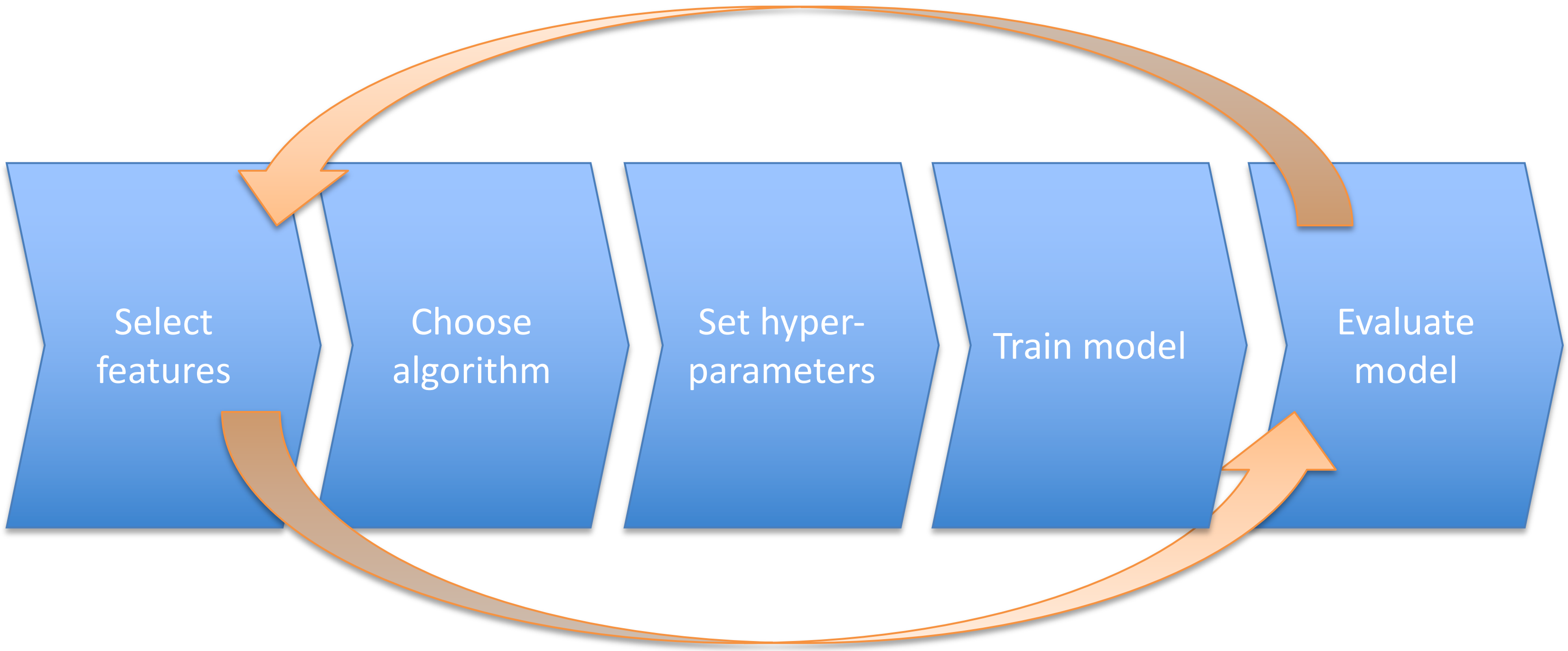
New data

Prediction

Components of a Model



Modeling Process



outrageously AMBITIOUS

Feature Selection

Duke

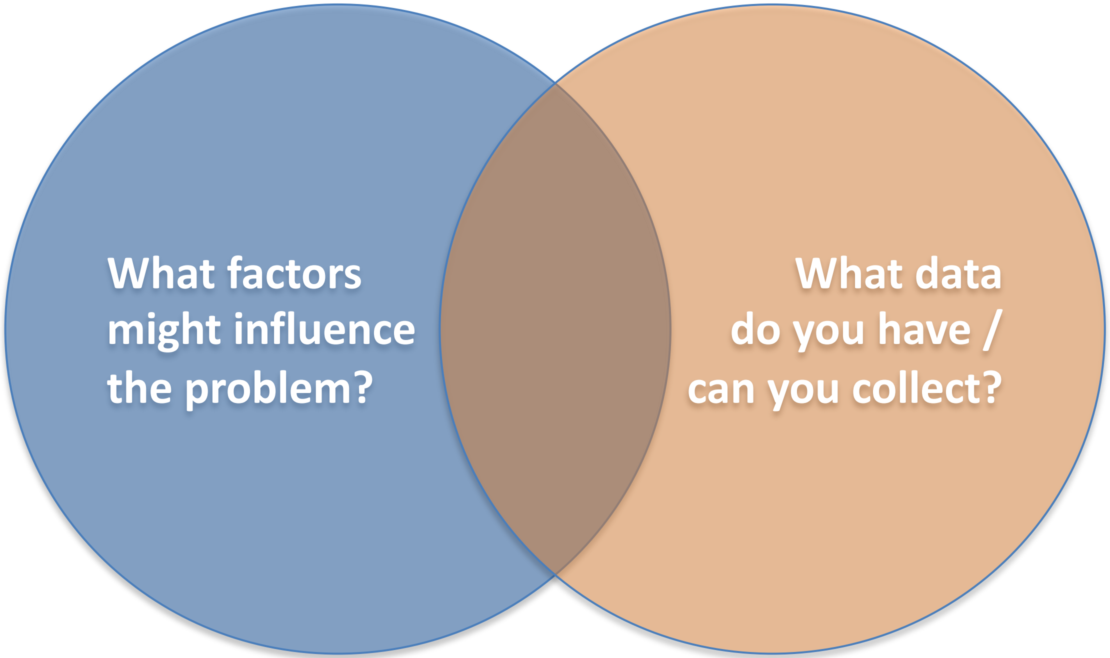
PRATT SCHOOL of ENGINEERING

What are Features?

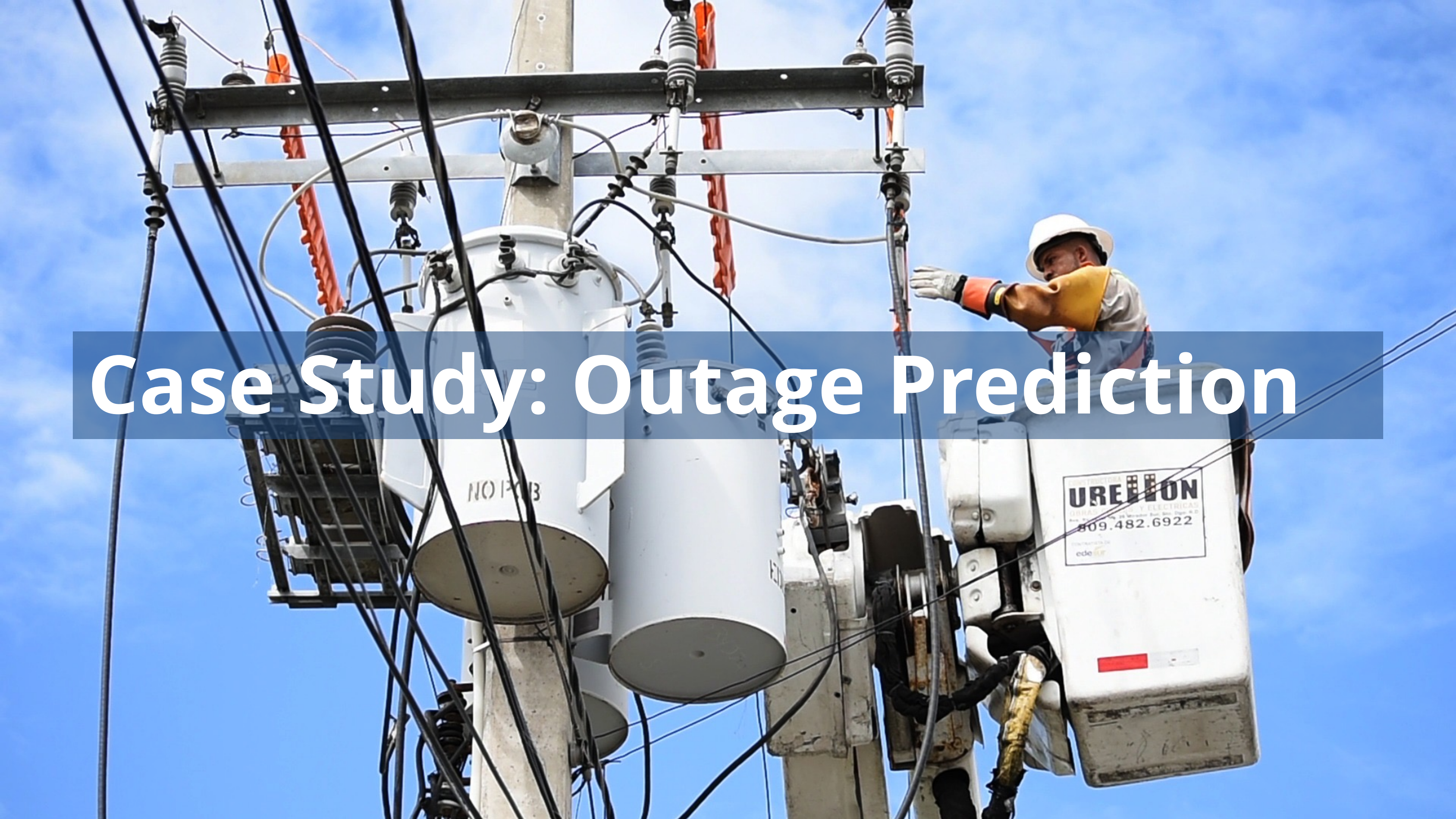
$$\overbrace{\hspace{10em}}^{\text{Features}}$$

	Neighbor -hood	School district	Square footage	Number of bedrooms	Year built
House 1	Weycroft	Wake	3400	4	2010
House 2	Horton Creek	Wake	4200	5	2008
House 3	Cary Park	Chatham	3250	4	2012
…	…	…	…	…	…

How to Define Features



Case Study: Outage Prediction



Methods of Feature Selection

- Domain expertise
- Visualization
- Statistical correlations
- Modeling

Including too few features is usually much worse than including too many!

Algorithm Selection

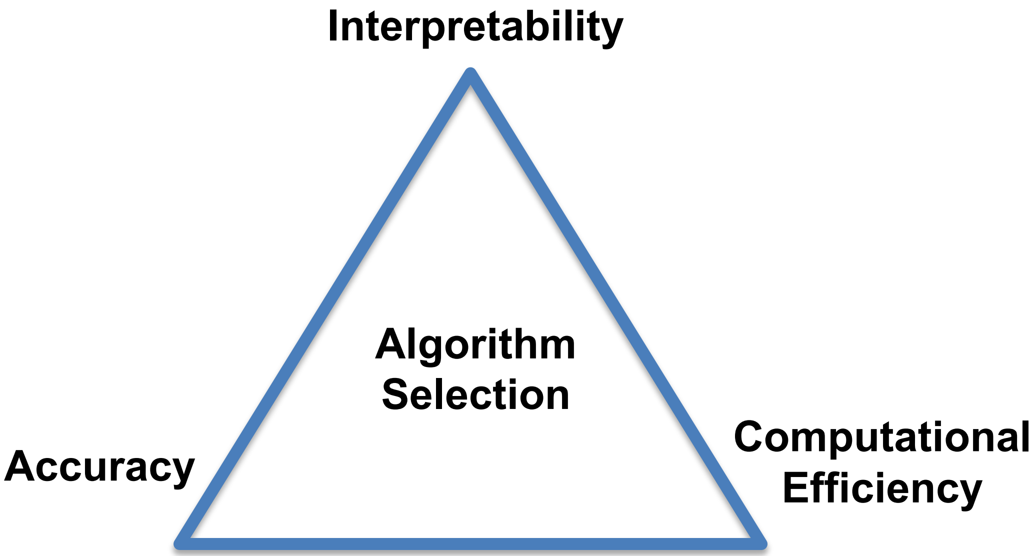
outraggeously AMBITIOUS

Duke
PRATT SCHOOL of ENGINEERING

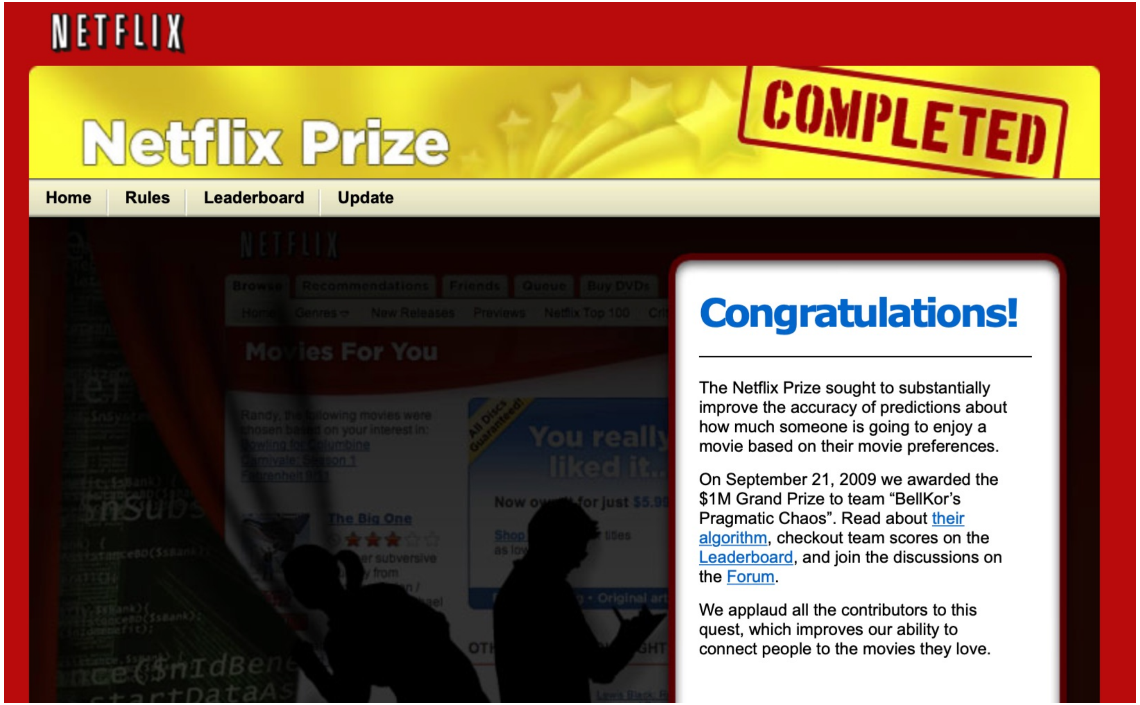
Algorithm Selection

“No free lunch theorem”

Algorithm Selection



Netflix Example



https://netflixprize.com

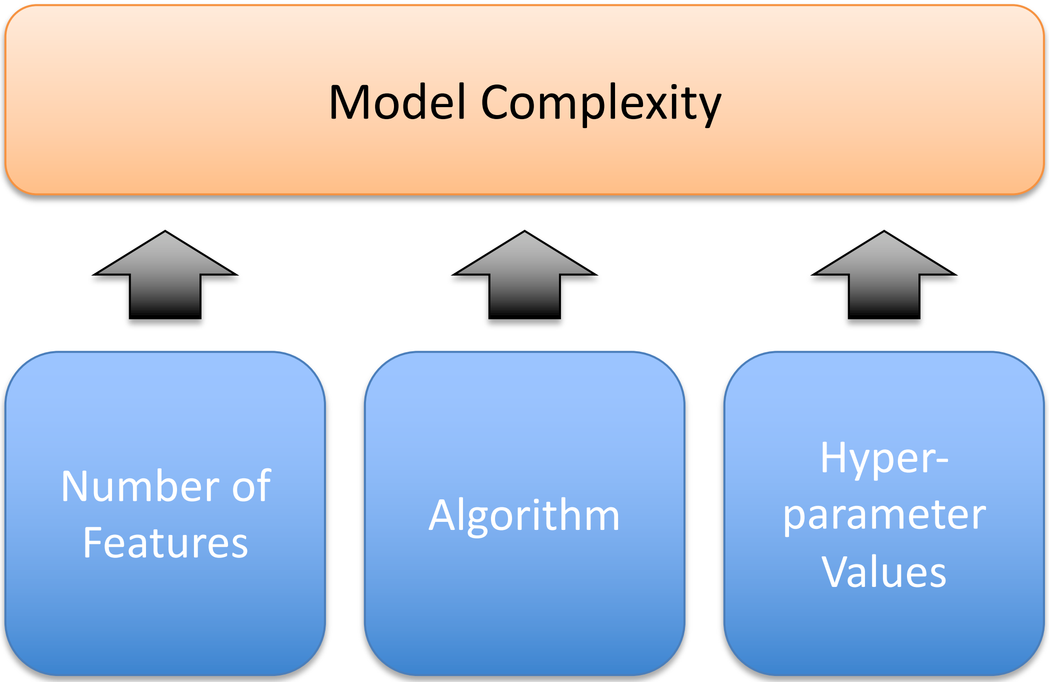
outrageously AMBITIOUS

Bias – Variance Tradeoff

Duke

PRATT SCHOOL of ENGINEERING

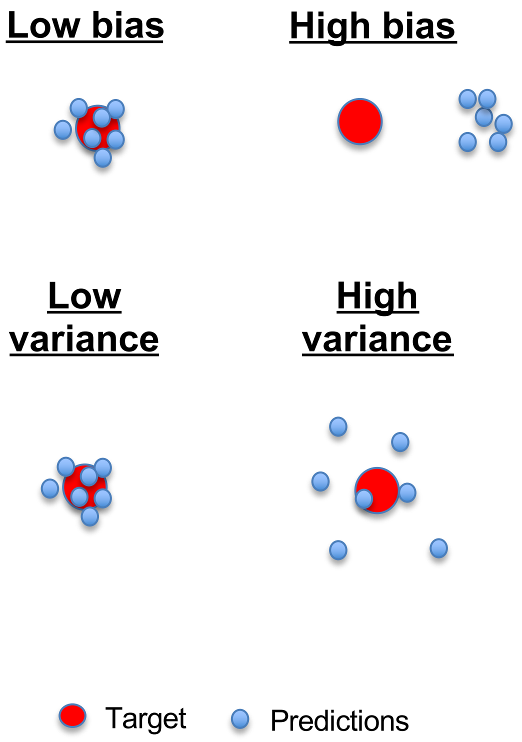
Model Complexity



Bias and Variance

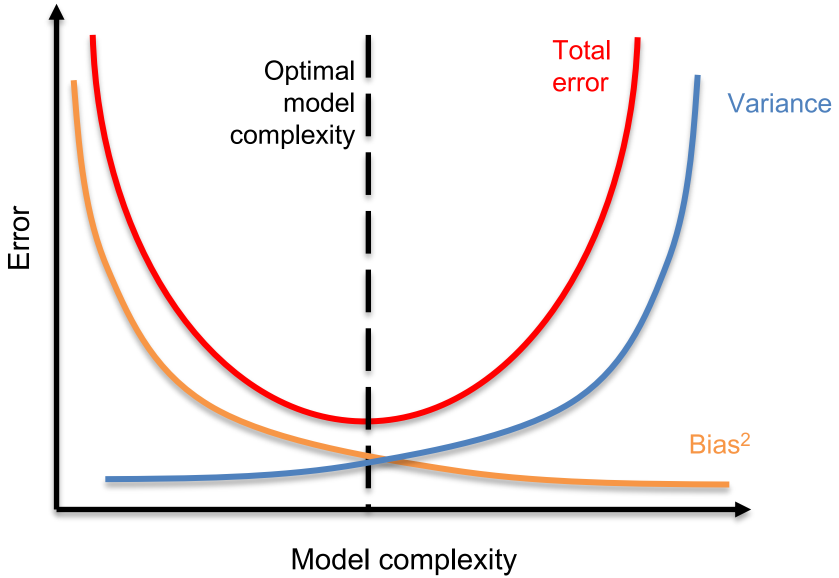
- **Bias** is error introduced by modeling a real life problem using a simpler model that is unable to fully capture the underlying patterns in data

- **Variance** refers to the sensitivity of the model to small fluctuations in the data, because it models fine patterns which may just be noise



Bias – Variance Tradeoff

- Simpler models often have **higher** bias and **lower** variance
- Complex models typically have **lower** bias but **higher** variance
- $\text{Total Error} = \text{Bias}^2 + \text{Var} + \sigma_e^2$



Underfitting vs. Overfitting

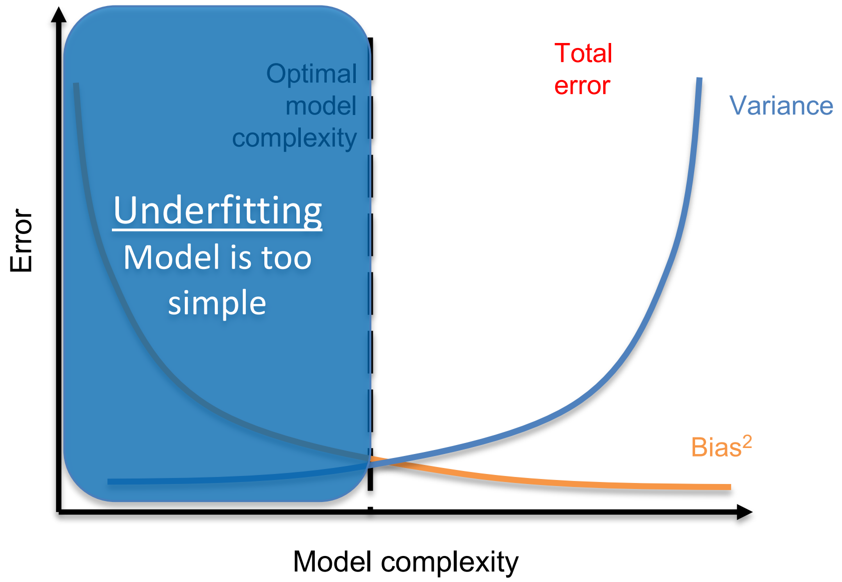


Image source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Underfitting vs. Overfitting

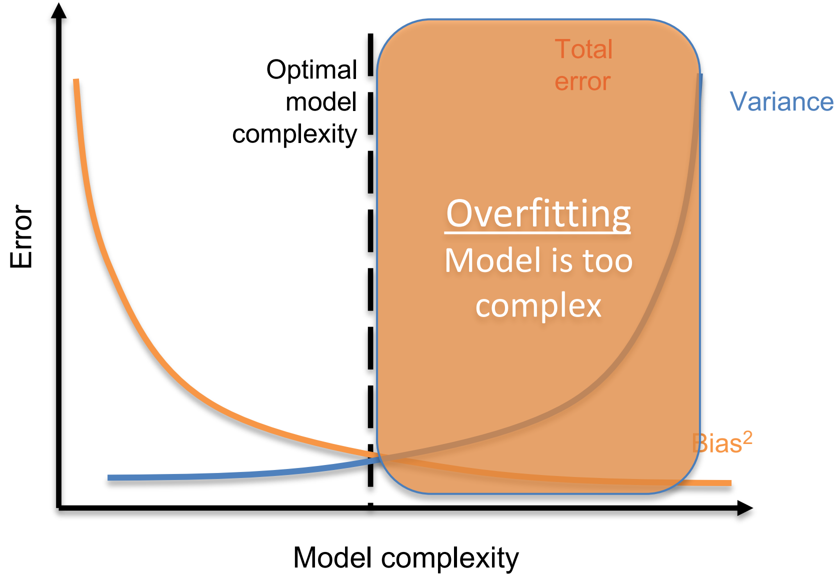
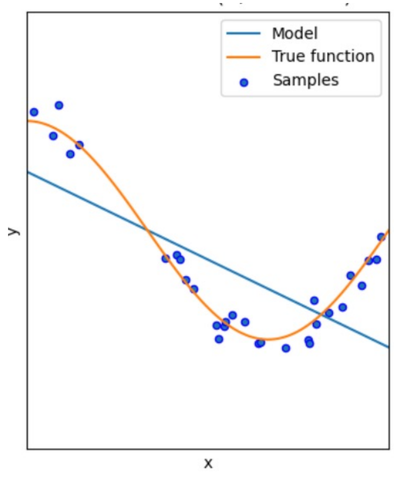


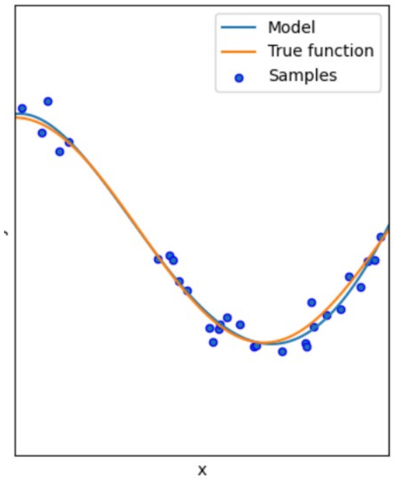
Image source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Underfitting vs. Overfitting

Underfitting
Model is too simple



Good Fit
Model fits well, with some error



Overfitting
Model is too complex

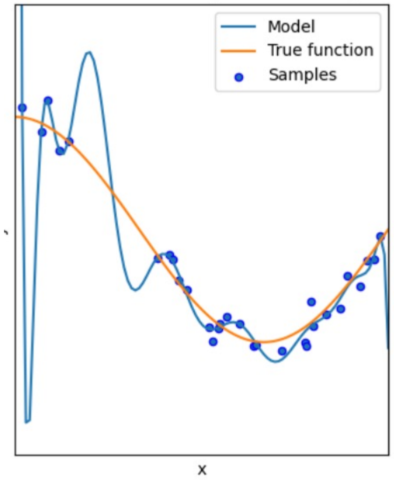


Image source: Scikit Learn documentation

Test & Validation Sets

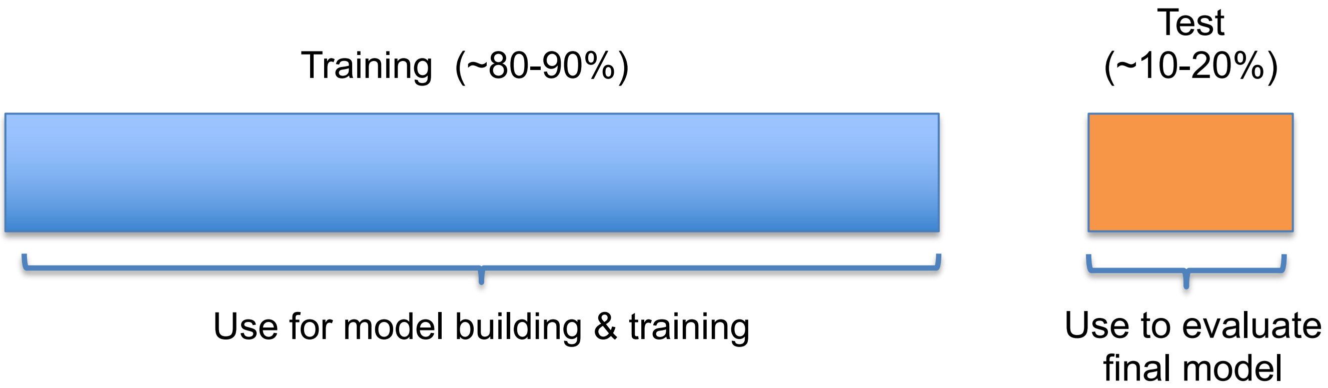
outrageously AMBITIOUS

Duke

PRATT SCHOOL *of* ENGINEERING

Training & Test Sets

- Goal of predictive modeling is to create a model that makes accurate predictions on new unseen data
- We cannot estimate performance on data we do not have, so instead we split our data into two sets
  - **Training set** - build and train the model
  - **Test set** – Evaluate model performance performance

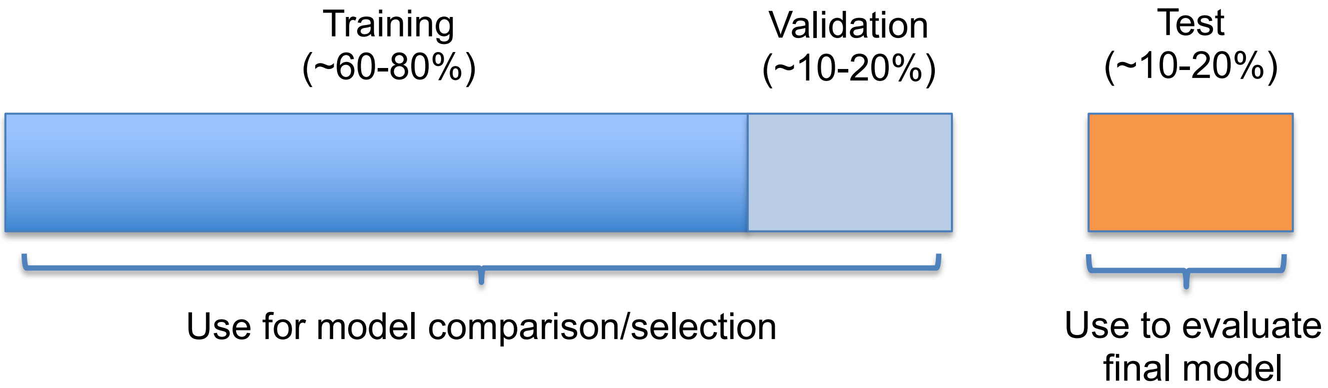


Data Leakage

- **“Data leakage”** occurs when some of our test set data “leaks” into model building and influences the development of the model
- For example, if we use all of our data to select our features, or compare algorithms
- This **invalidates the estimated performance** of the model and causes it to be overoptimistic

Validation Sets

- Often we want to compare models to select the optimal model
- If we use the test set to compare model performance, it is not longer an unbiased indicator of performance
- Instead, we split our training set further into training and validation sets
- We use the validation set for model selection, and report performance on the test set



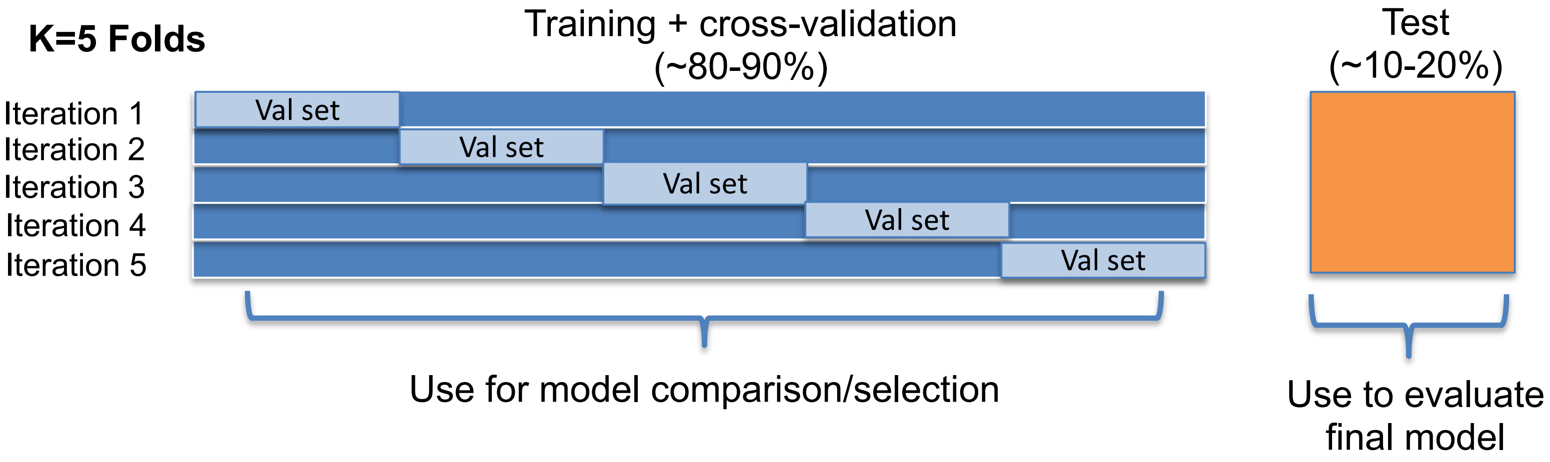
outrageously
AMBITIOUS

Cross Validation

Duke
PRATT SCHOOL *of*
ENGINEERING

K-Folds Cross Validation

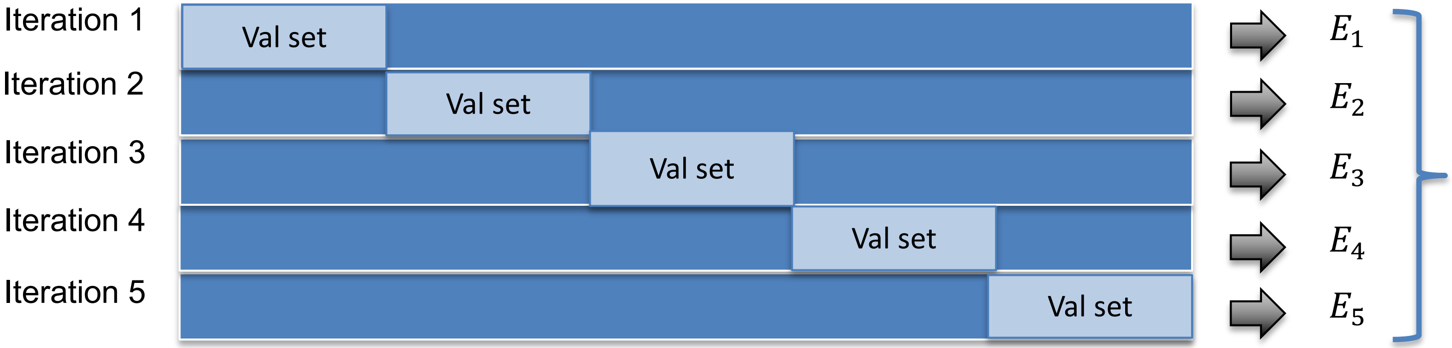
Rather than using a fixed validation set, we train and run the model(s) multiple times, each time using a different subset ("fold") as the validation set



K-Folds Cross Validation

We calculate the error on the validation fold for each iteration, and then average them together to get the average error

K=5 Folds



$$E = \frac{1}{5}\sum_{i=1}^{5} E_i$$

Benefits of Cross Validation

- Maximizes the data available for training the model – important for small datasets
- Provides a better evaluation of how well the model can generalize to new data – validation performance is not biased by choice of datapoints to use for validation

Wrap-up

outrageously
AMBITIOUS

Duke
PRATT SCHOOL of
ENGINEERING

Wrap Up

- Modeling process is just one piece of the CRISP-DM process
- Model complexity comes from features, algorithm and hyperparameters
- Underfitting and overfitting are common modeling issues
- Test sets and validation sets ensure we properly select and evaluate models