Machine Learning for management

Why AI/ML?

- Companies across every industry are using AI to make their products or services more **predictive**, **personalized** and **automated**
- AI is also creating the ability to solve previously unsolved problems
- Successfully bringing AI products to market requires a team effort
- Everyone needs to speak the same language and have the same fundamental understanding

What is Machine Learning?

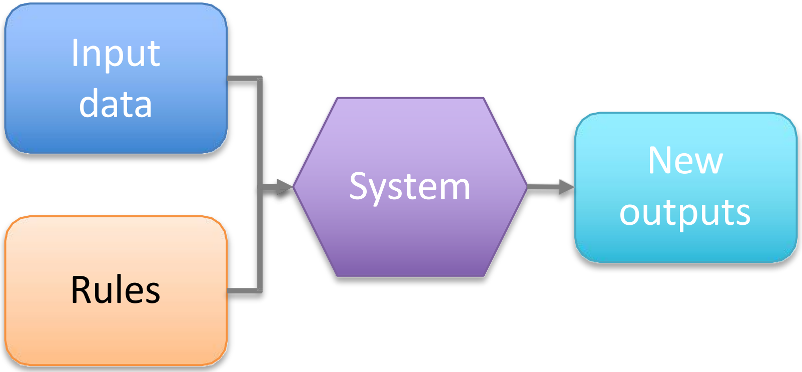
- "Field of study that gives computers the ability to learn **without being explicitly programmed**" – Arthur Samuel, IBM, 1959

- Instead of providing a computer with exact instructions to solve a problem, we show it examples of the problem to solve and let it figure out how to solve it itself

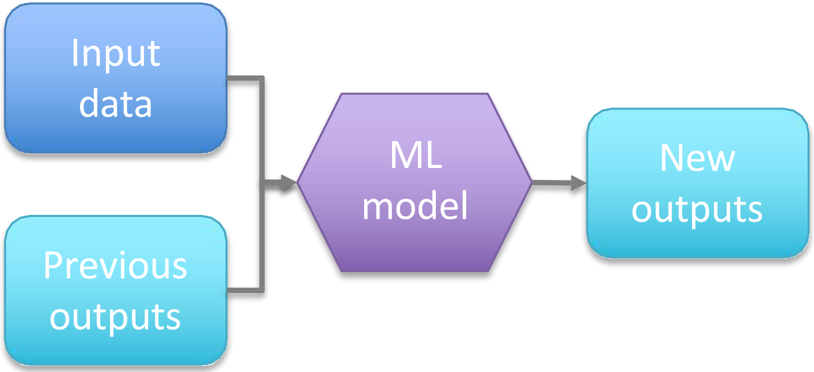
By Hongreddotbrewhouse - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=33551162

ML vs. traditional software

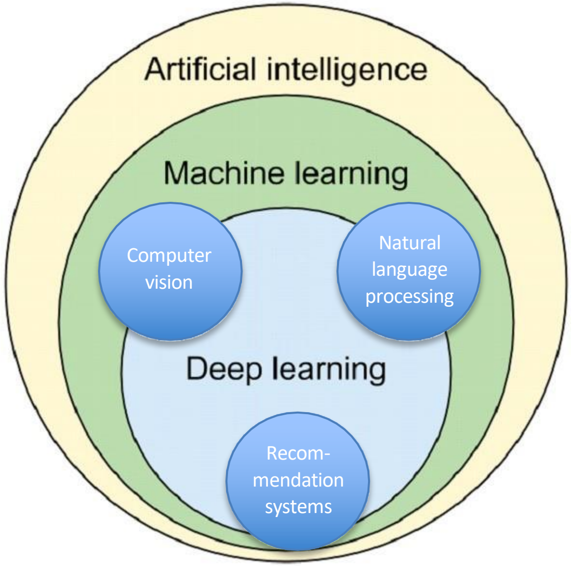
How traditional software generates predictions



How machine learning generates predictions



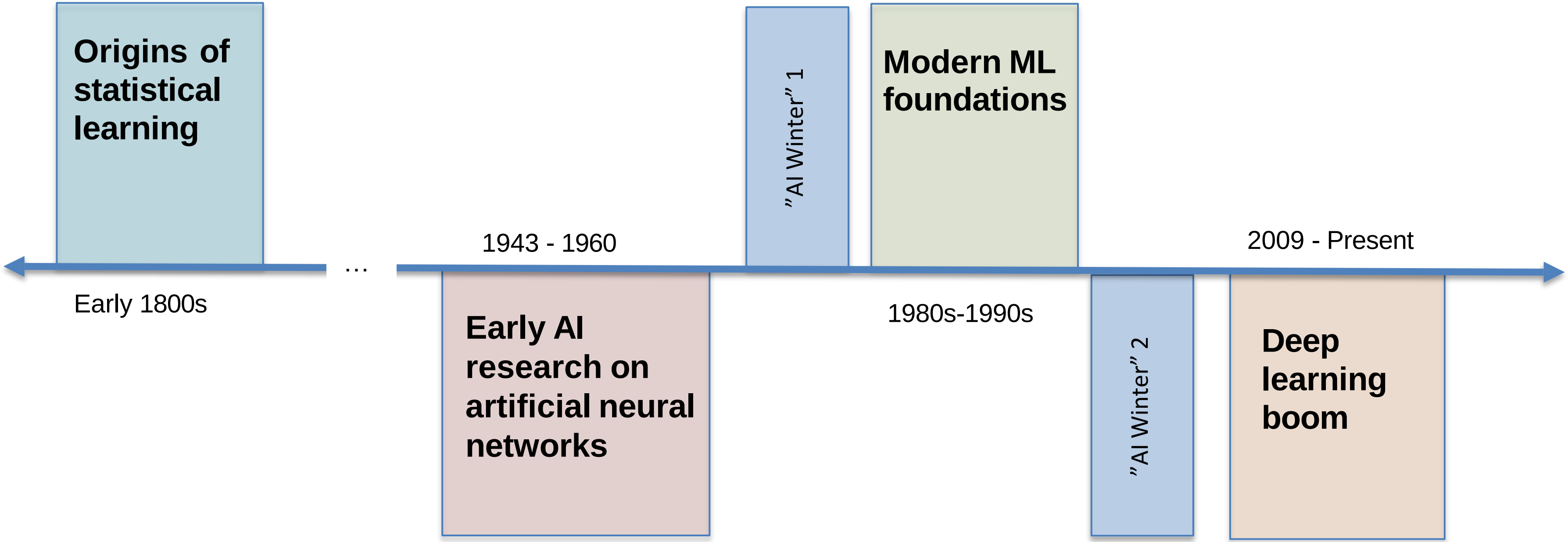
AI vs. Machine Learning



https://commons.wikimedia.org/wiki/File:Fig-X_All_ML_as_a_subfield_of_AI.jpg

- **Machine learning** is a set of methods & tools which help realize the goal of the field of **artificial intelligence**
- **Deep learning**, or the use of neural networks containing many layers, is a sub-field of machine learning
- **Computer vision, natural language processing, recommendation systems** etc. are sub-fields of AI which rely on machine learning methods

Brief History of AI/ML

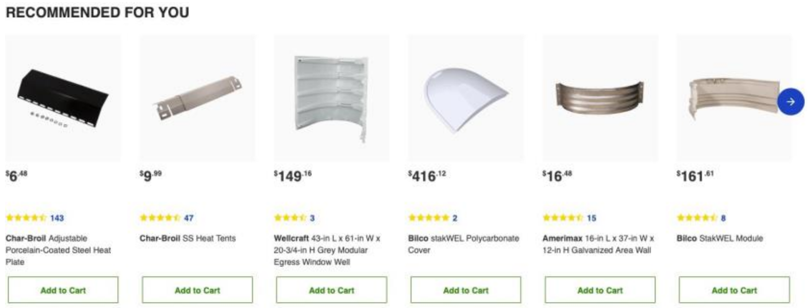


Machine Learning Today

- **Explosion in data**
  - Ubiquitous internet connectivity
  - Advances in sensor technology
  - Smart connected devices
- **Deep learning** has made what was impossible, possible
  - Massive increase in computational power – GPUs
  - Huge sets of labeled data for training
  - Algorithmic advances
- **Pervasiveness** of machine learning models in products and systems we interact with daily

Where Do We Find ML?

Product recommendations



Spam filters

Where Do We Find ML?

Mail routing via OCR

Credit card fraud detection

Data Comes in Many Forms

"**Data** are characteristics or information, usually numerical, that are collected through observation." [OECD Glossary of Statistical Terms]

Almost anything can be turned into numbers:

- Measurements
- Text
- Images
- Sound
- Video

Data may have different relationships:

- Spatial relationships
- Temporal relationships

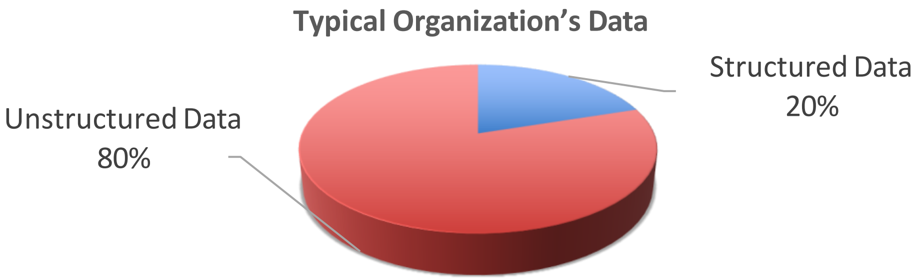
Structured vs. Unstructured Data

Structured data

- Set structure based on pre-defined fields for each record
- Often stored in relational databases
- Easy to enter, search and analyze
- Works well with common tools

Unstructured data

- Does not follow a defined format of fields
- Many types – images, videos, sounds, text
- Requires specialized tools to work with



Continuous vs Categorical Data

Continuous

- Numeric variable that has an infinite number of values between any two values
- E.g. length of a part, temperature, height, time

Categorical

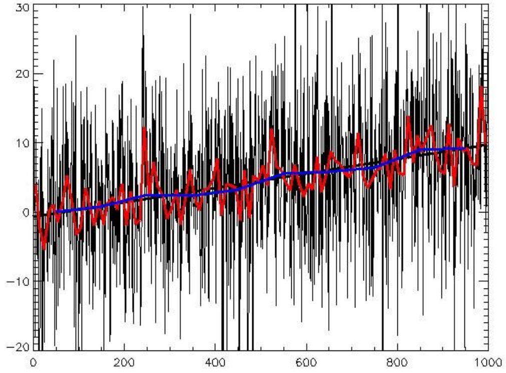
- Finite number of categories / distinct groups
- May or may not have a logical order
- E.g. gender, student major, material type, color

Discrete

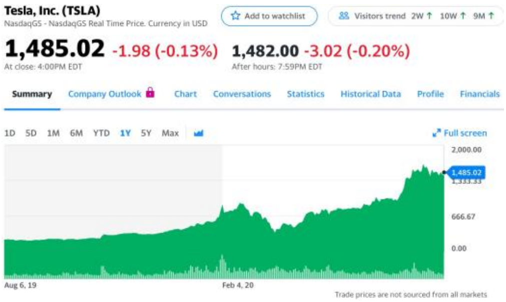
- Numeric variable that has a countable number of values between two values
- E.g. age, number of parts, year made
- Rule of thumb – if number of possible values small (e.g <10), treat as categorical

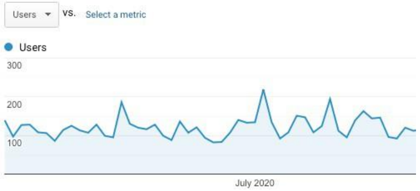
Time series data

- Series of data points organized in time order
- Points are usually equally spaced by time
- Assumptions:
  - Time is considered one-way
  - Points close together in time are more related than points further apart



[Wikipedia]





Terminology

Targets / Labels / Annotations / Response / Y Variable / Dependent Variable

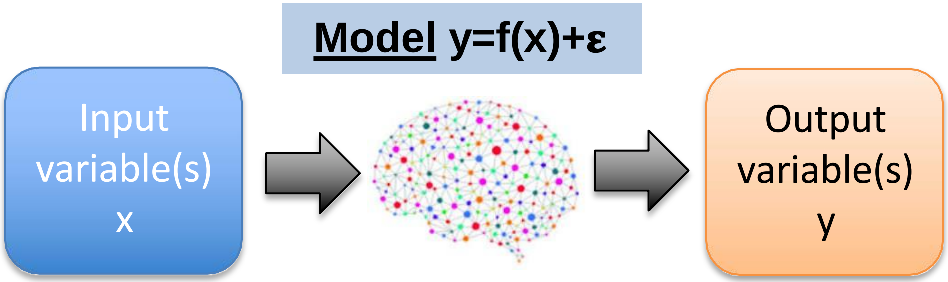
Features / Factors / Predictors / X Variables / Independent Variables / Attributes / Dimensions

Observations / Instances / Examples / Feature Vectors

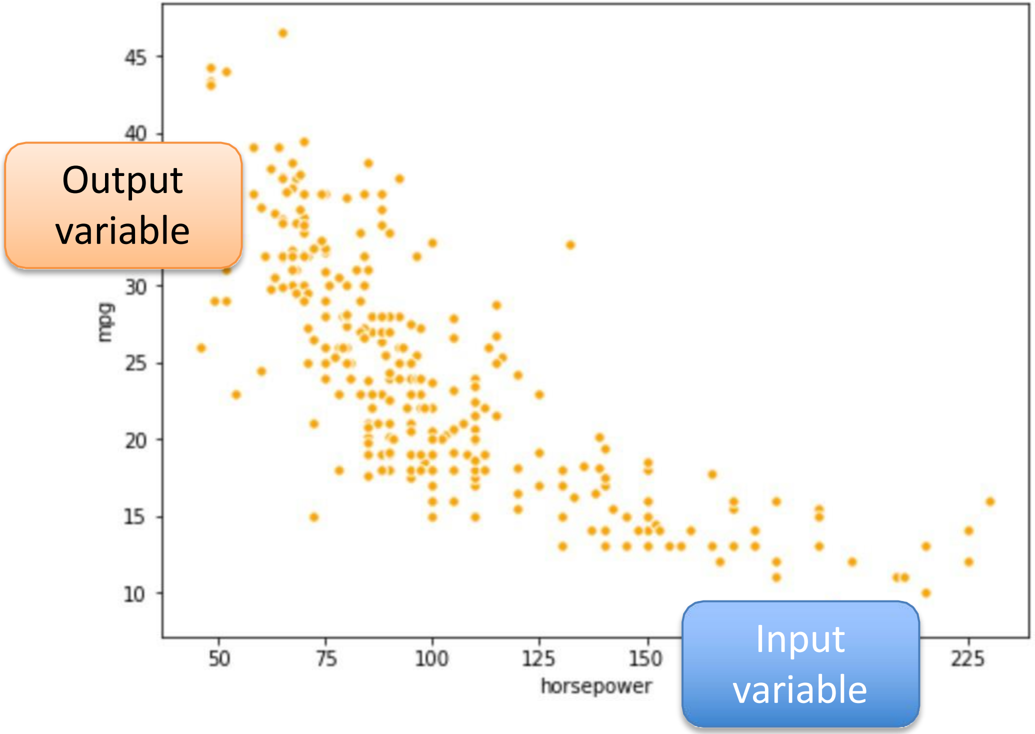
	Neighbor-hood	School district	Square footage	Number of bedrooms	Year built	Market sale price
House 1	Weycroft	Wake	3400	4	2010	$612,000
House 2	Horton Creek	Wake	4200	5	2008	$675,000
House 3	Cary Park	Chatham	3250	4	2012	$520,000
…	…	…	…	…	…	…

What is a model?

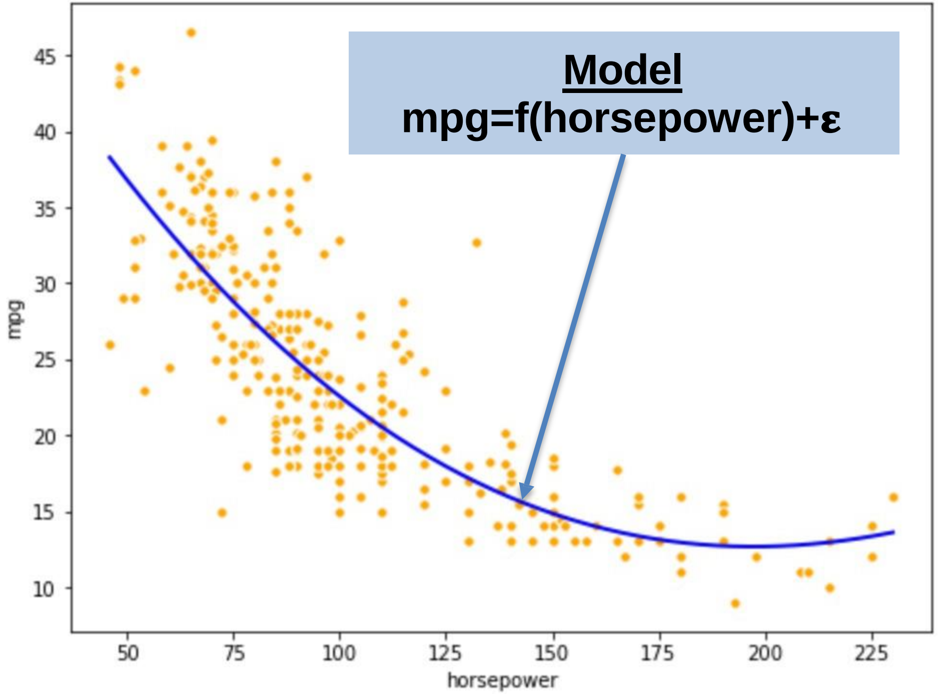
A **model** is an approximation of the relationship between two variables



What is a model?



What is a model?



What is a model?

Observations of input data **(X)**

	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
…	…	…	…	…	…

Model y=f(X)+ε

Predictions of **target (y)**

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

Building a model

To create a model we define four things:

1. **Features** to use
2. **Algorithm** – acts as a form/template for model
3. **Hyperparameter** values for algorithm
4. **Loss function** to optimize

We **train** our model using historical data:

- Algorithm & hyperparameters provide overall model form
- "Learn" values for the model which minimize loss function

Types of Machine Learning

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Objective	Prediction of a target variable	Organize data by inherent structure	Learn strategies via interaction
Learning Task(s)	Classification Regression	Clustering Anomaly detection	Achieve a goal
Target Data Required?	Yes	No	Yes, but delayed
Examples	• Identifying pneumonia from xray images • Predicting real estate prices	• Market segmentation • Identifying fraudulent activity	• AlphaZero • Autonomous vehicles

Supervised vs. Unsupervised Learning

Supervised learning

At least some past observations of the features (X_i) and targets (y_i) are known and used to build a model

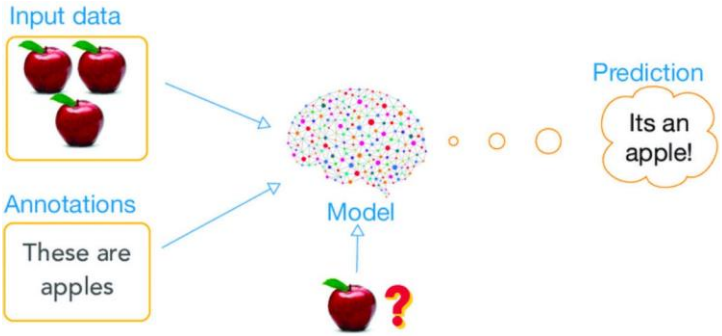


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

Supervised vs. Unsupervised Learning

Unsupervised learning

We only have observations of the features (X_i). We need to use the observations to guess what the targets (y_i) would have been and build a model from there

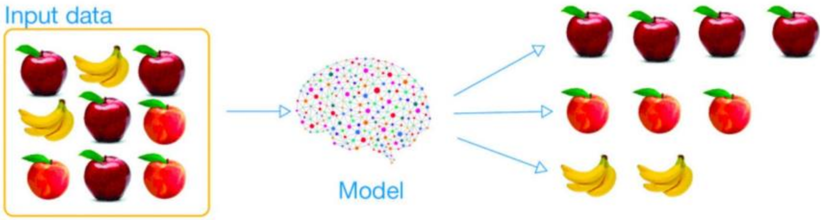
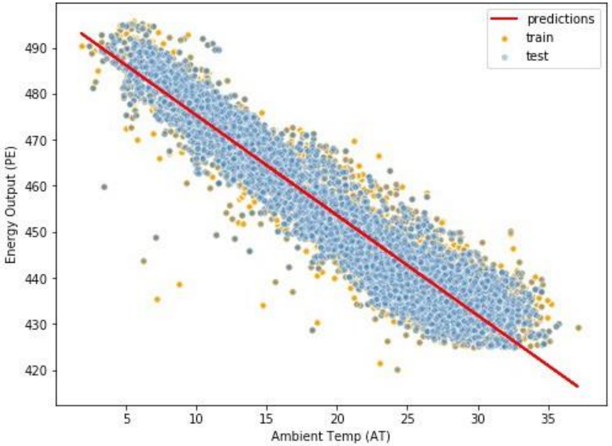


Image source: https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

Regression vs. Classification

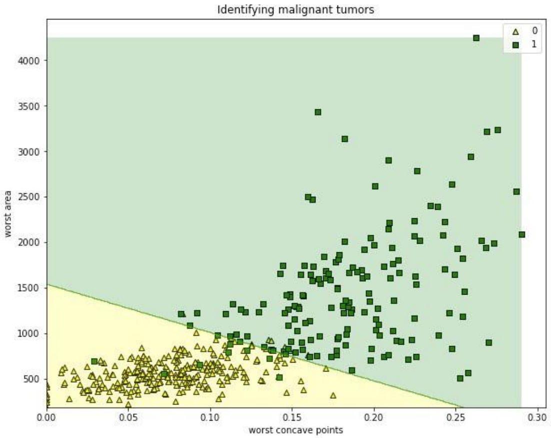
Regression

- Predict one or more **numerical** target variables
- E.g. home price, number of power outages, product demand



Classification

- Predicts a **class / category** – either binary or out of a set
- E.g. lung disease detection, identifying types of plants, sentiment analysis, detecting spam



What ML can do well*

- Automate straightforward tasks
- Make predictions by learning input-output relationships
- Personalize for individual users

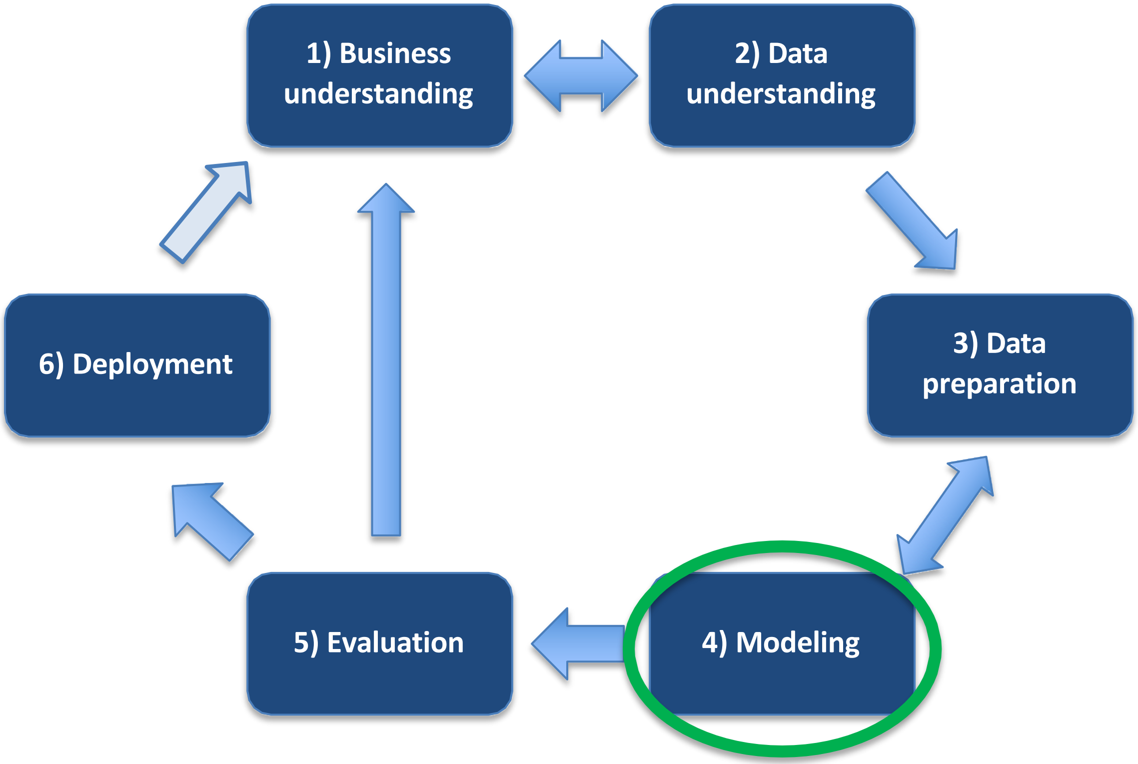
* Given sufficient quantity and quality of data

What ML cannot do well

- Understand context
- Determine causation
- Explain “why” things happen
- Determine the impact of interventions / find solutions

Building a Model

CRISP-DM Process



Creating a Model

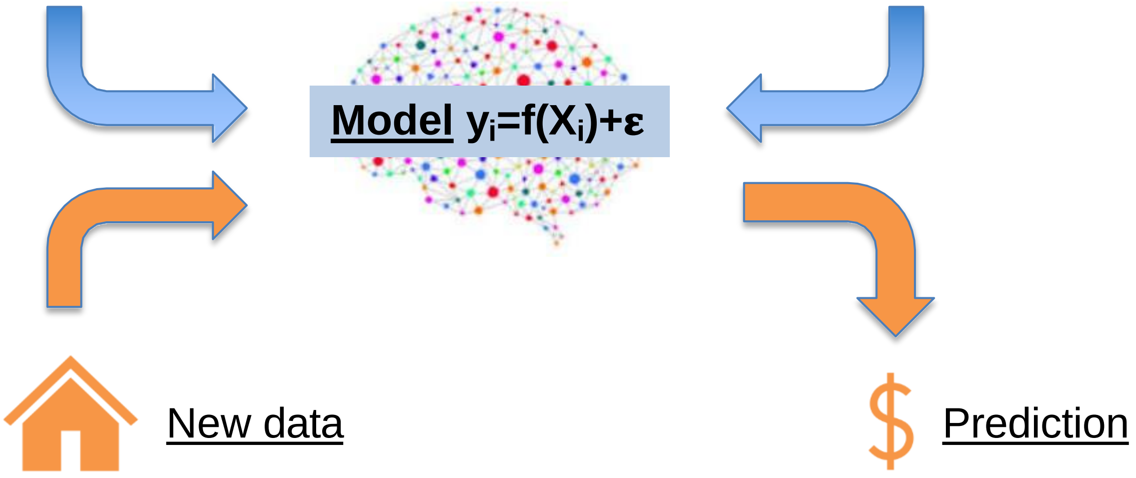
Past Observations

	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
…	…	…	…	…	…

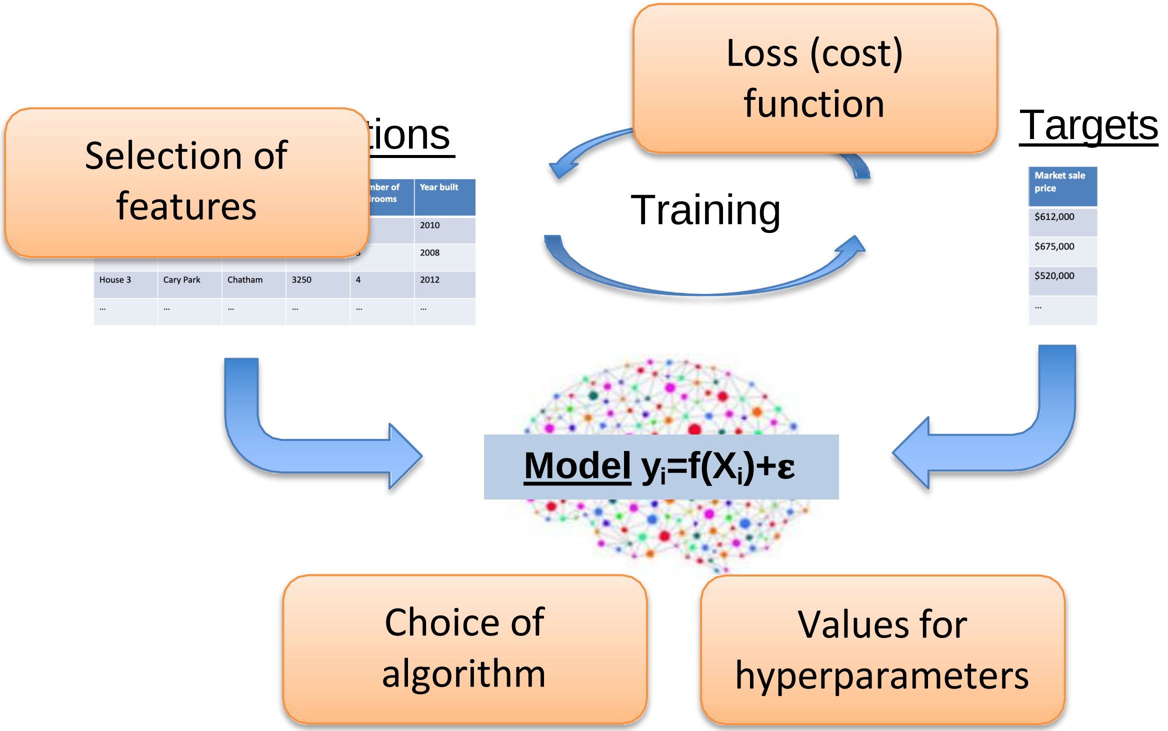
Training

Targets

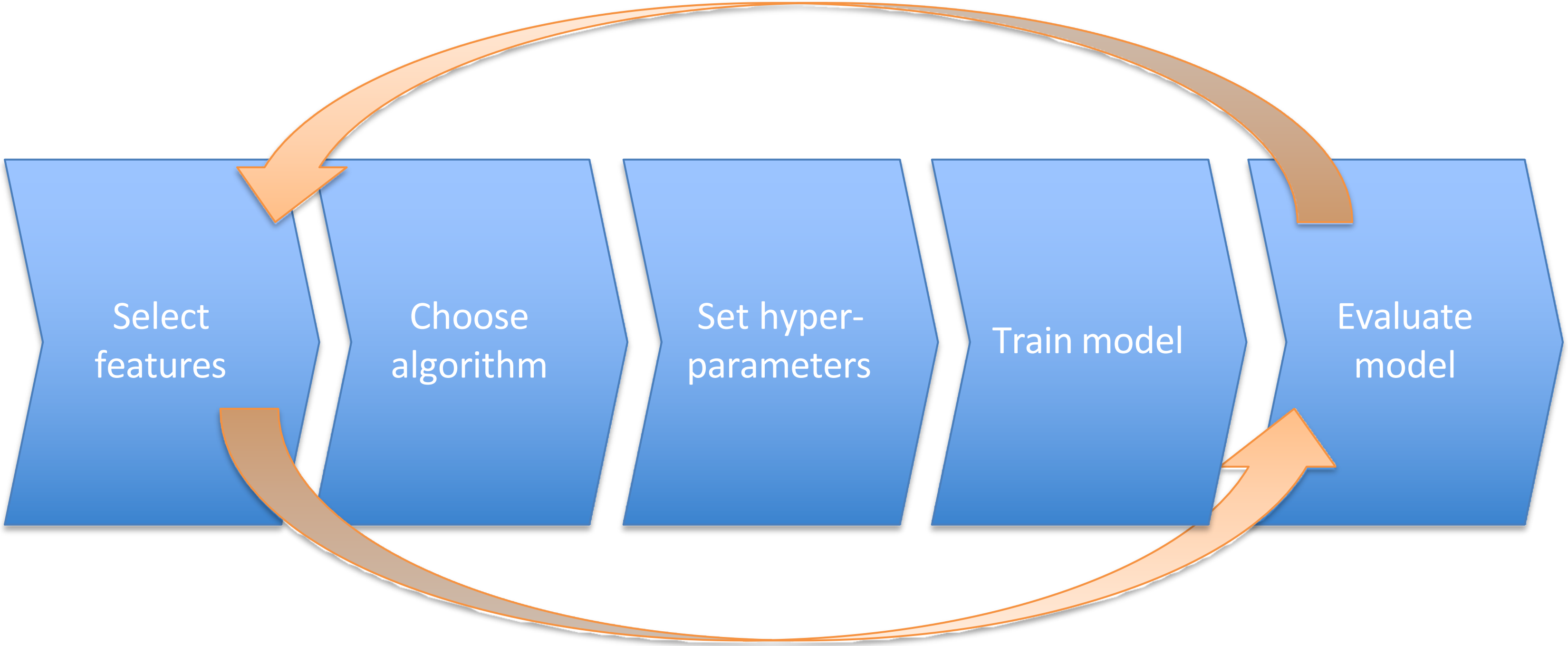
Market sale price
$612,000
$675,000
$520,000
…



Components of a Model



Modeling Process

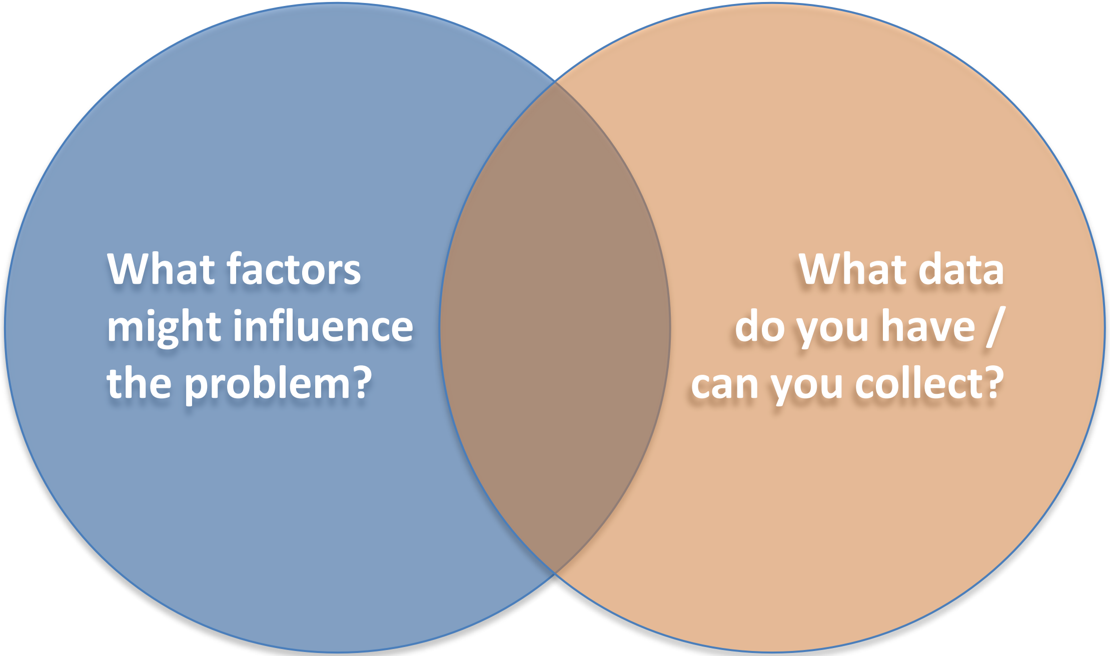


What are Features?

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



Methods of Feature Selection

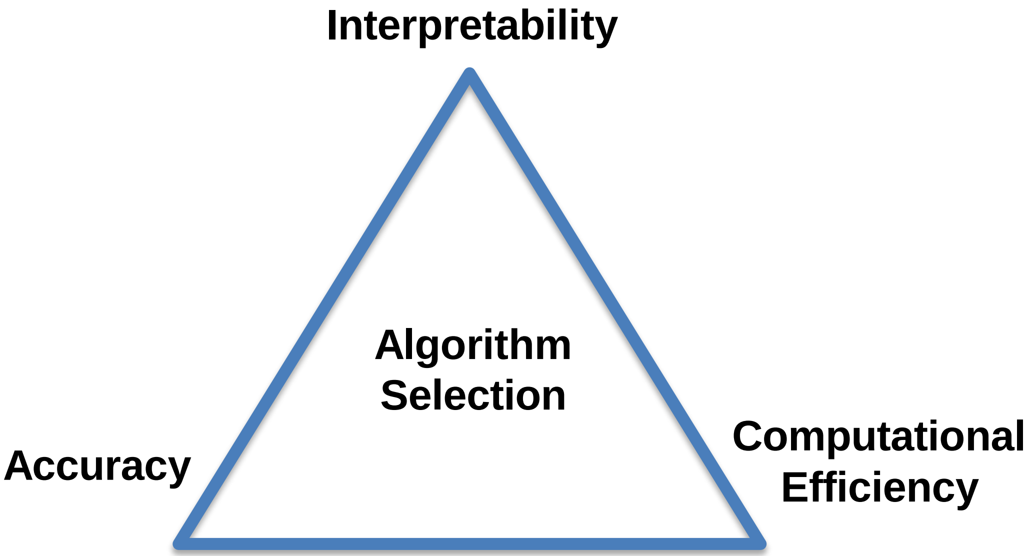
- Domain expertise
- Visualization
- Statistical correlations
- Modeling

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

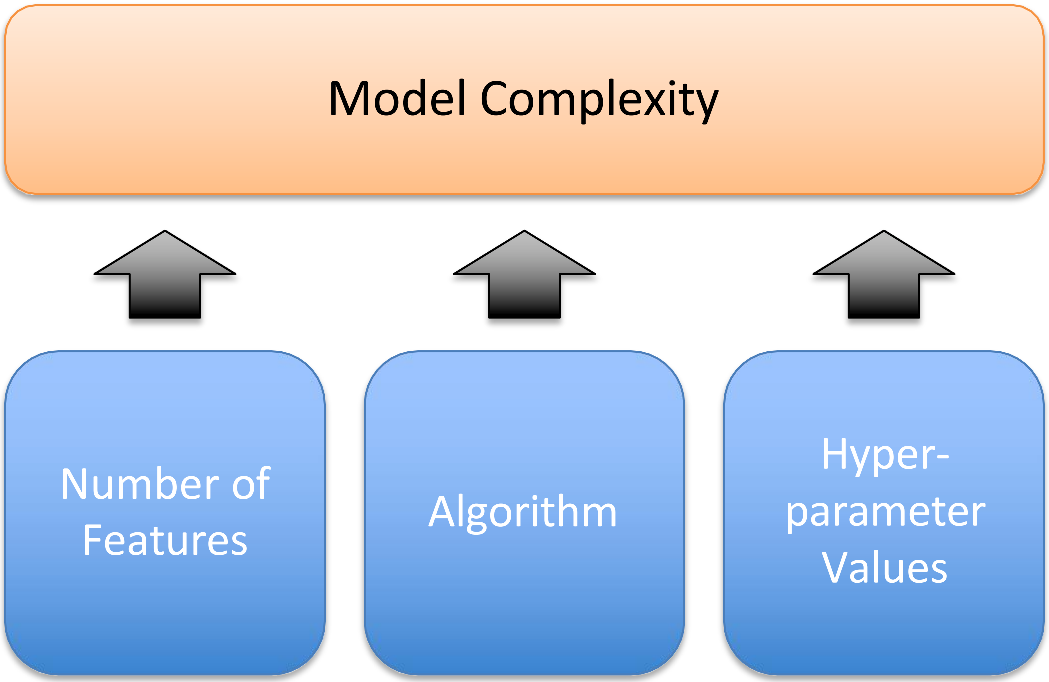
Algorithm Selection

"No free lunch theorem"

Algorithm Selection



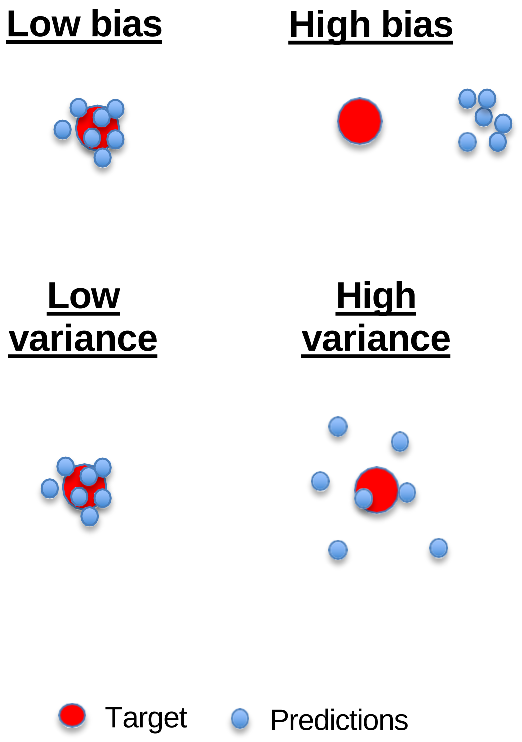
Model Complexity



Bias and Variance

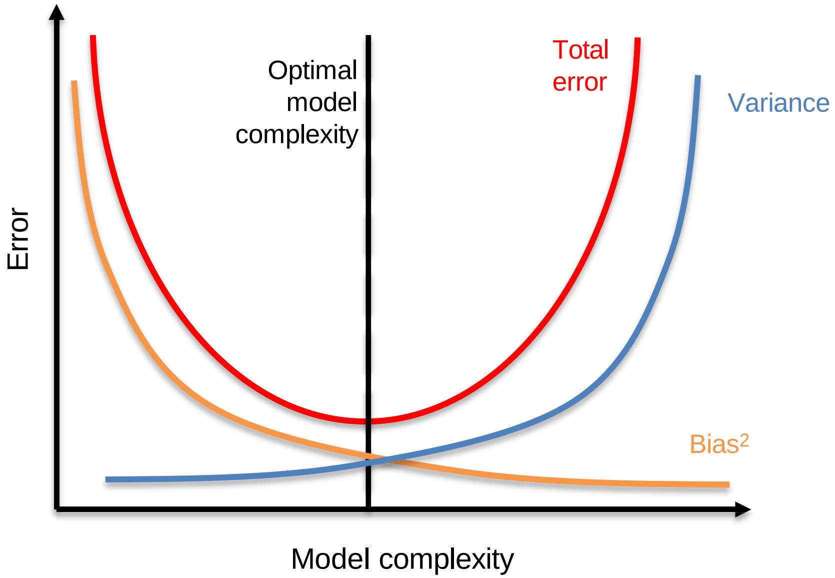
- **Bias** is error introduced by modeling a real life problem using a simpler model that is unable to fully capture the underlying patterns in data

- **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
- $\text{Total Error} = \text{Bias}^2 + \text{Var} + \sigma^2$



Underfitting vs. Overfitting

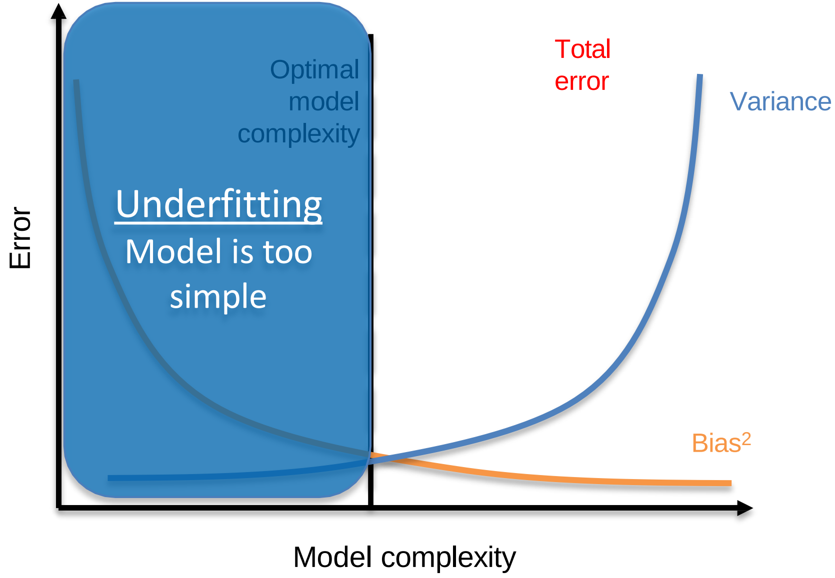


Image source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Underfitting vs. Overfitting

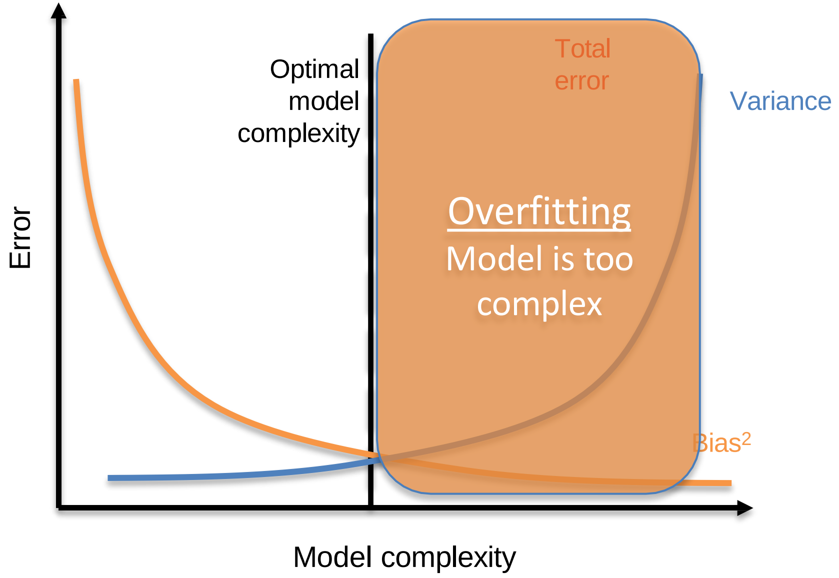
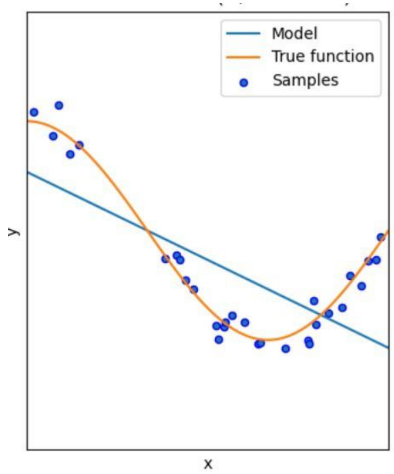


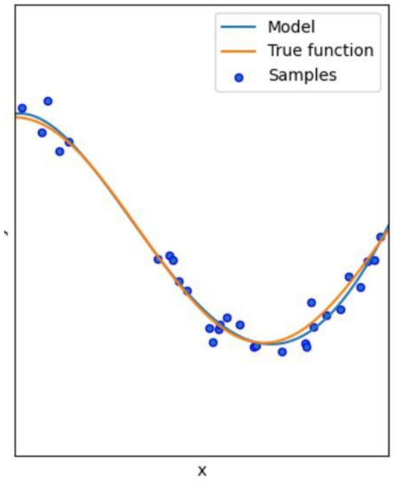
Image source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Underfitting vs. Overfitting

Underfitting
Model is too simple



Good Fit
Model fits well, with some error



Overfitting
Model is too complex

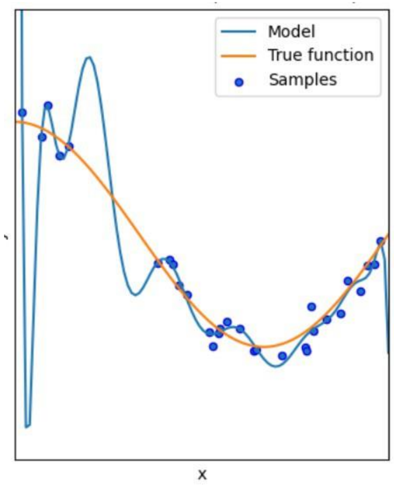
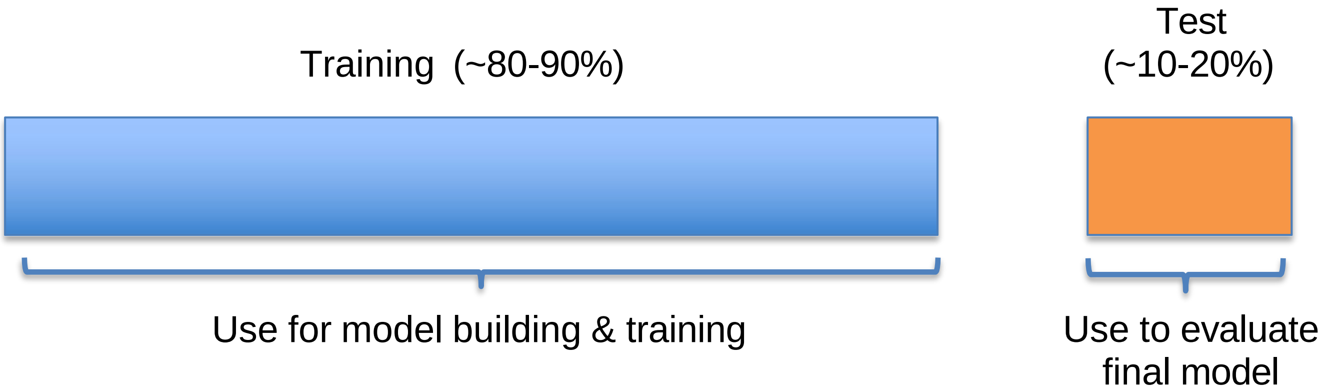


Image source: Scikit Learn documentation

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

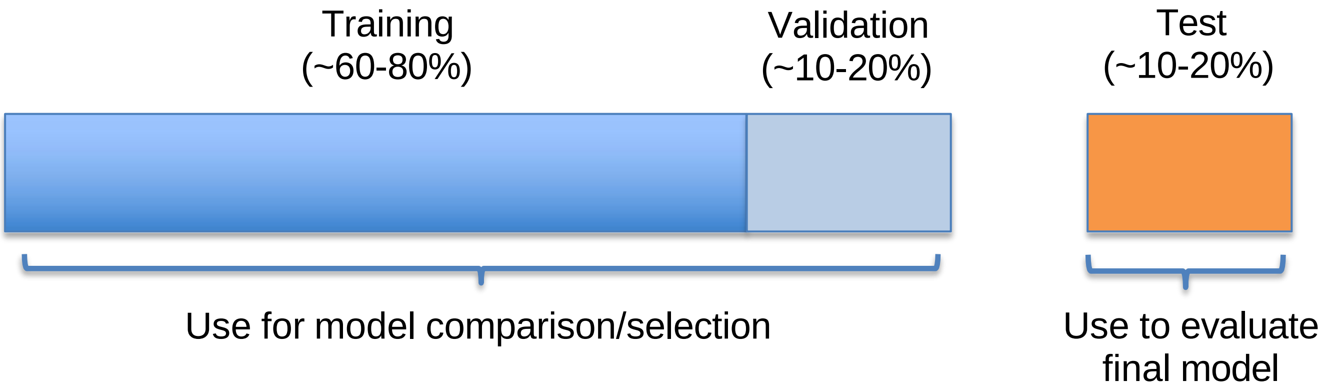


Data Leakage

- **“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

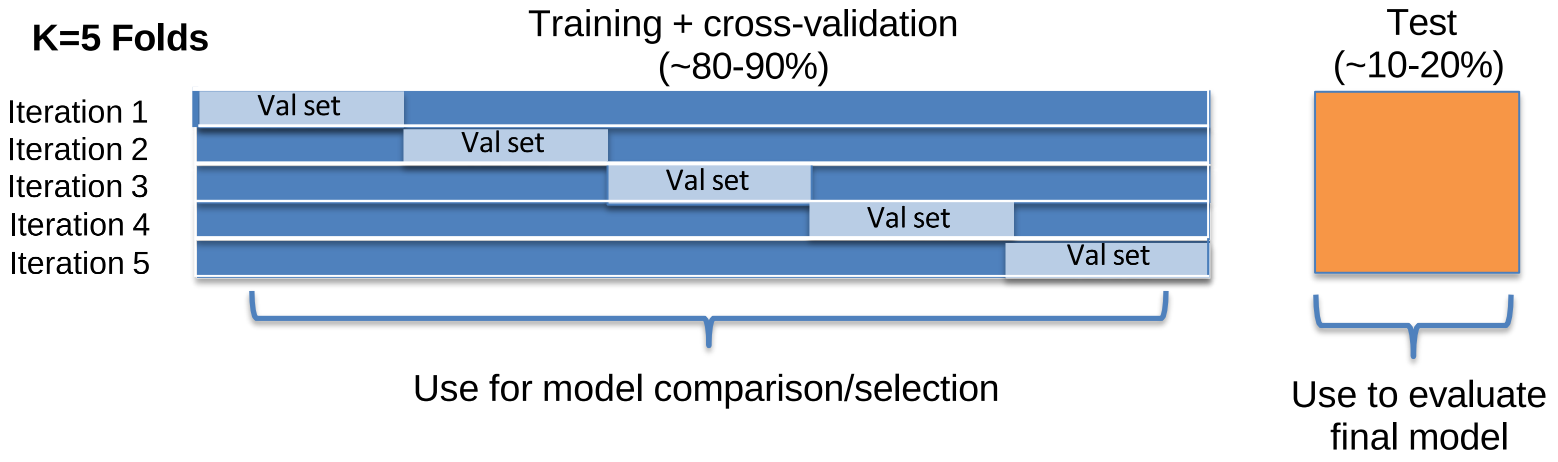
Validation Sets

- 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



K-Folds Cross Validation

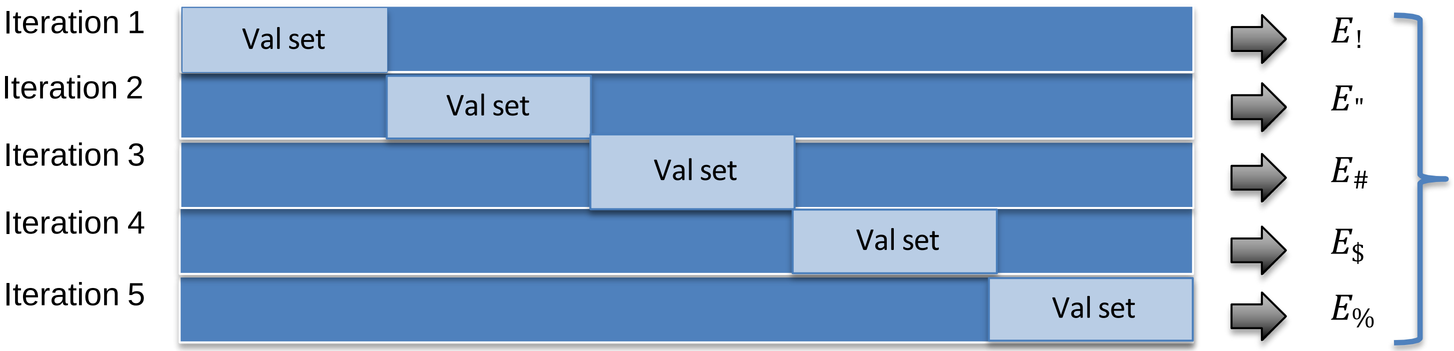
Rather than using a fixed validation set, we train and run the model(s) multiple times, each time using a different subset ("fold") as the validation set



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



$$E = \frac{1}{5}\sum_{k=1}^{5} E_k$$

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