

# Linear Regression

---

Read Chapter 7 (Regression) of the Textbook

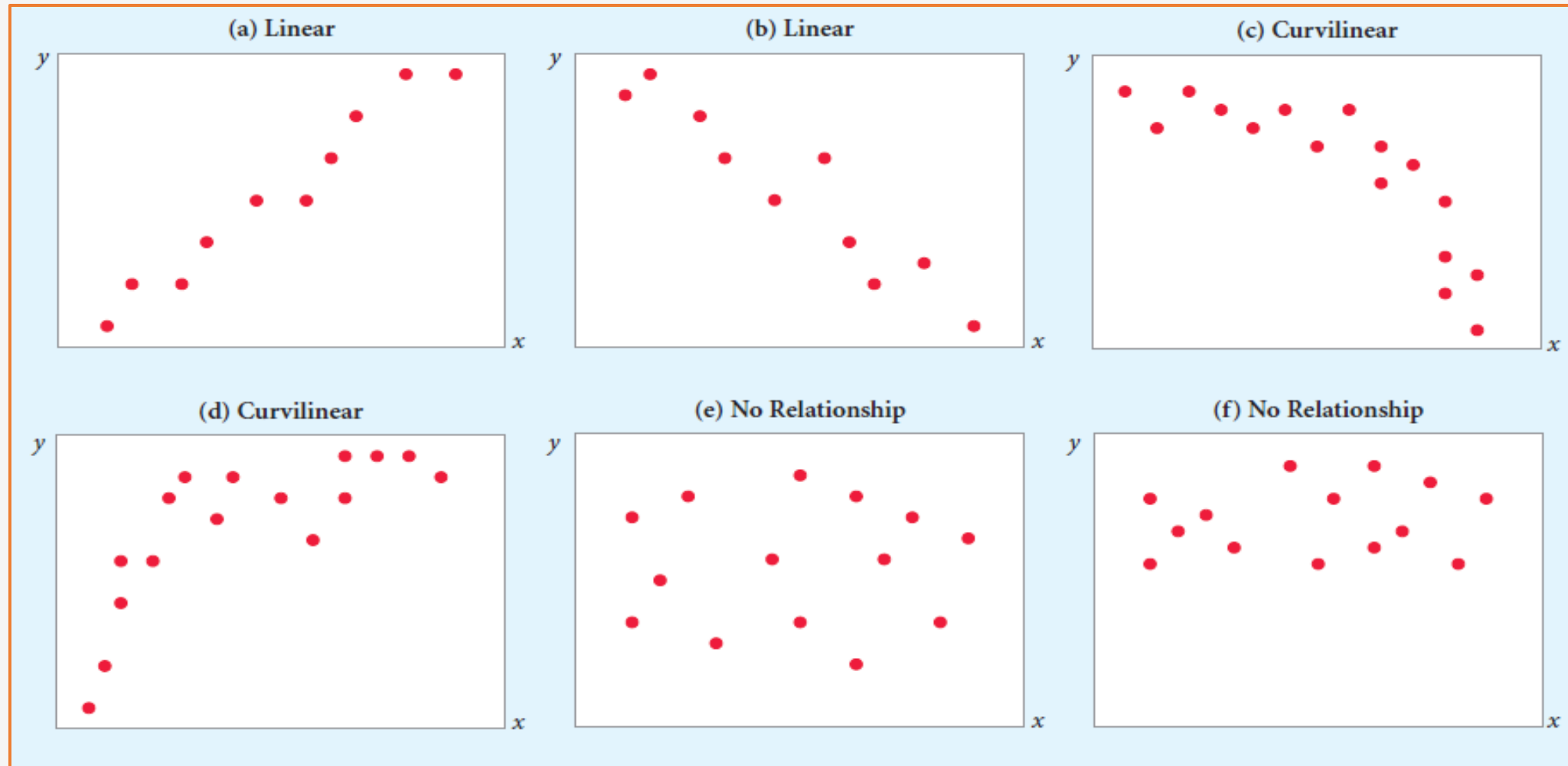
# Regression Models

---

- To understand the application of regression analysis in data mining
  - Linear/nonlinear
  - Logistic (Logit)
- To understand the key statistical measures of fit

# Relationships between variables

---



# When the data shows linear relationship

---

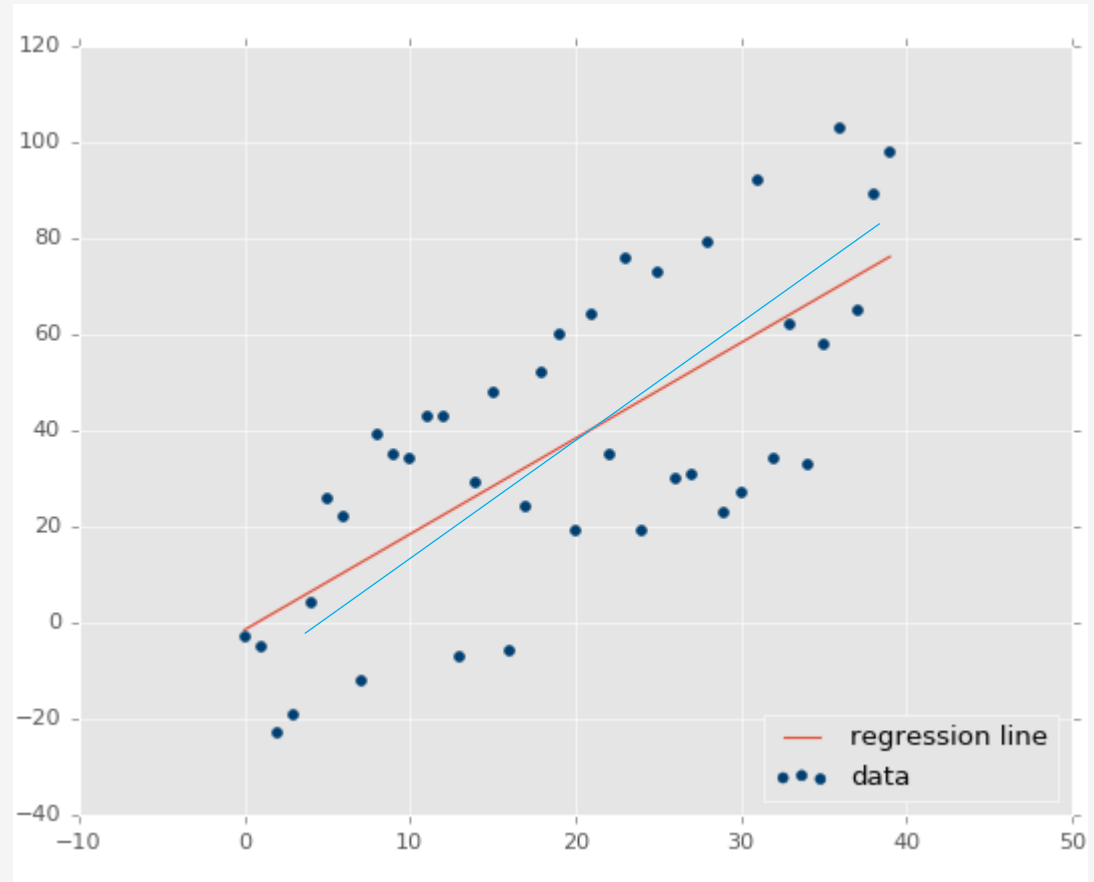
Correlation is high (positive or negative) and Scatter plots display a linear relationship

First model come to mind is

$$Y = m X + b$$

But still, there can be many lines that can “kind of” fit the data as well

Question: How to pick the “best-fit” line?



## How to find the best fitting line?

---

Define Mean Squared Error (MSE)

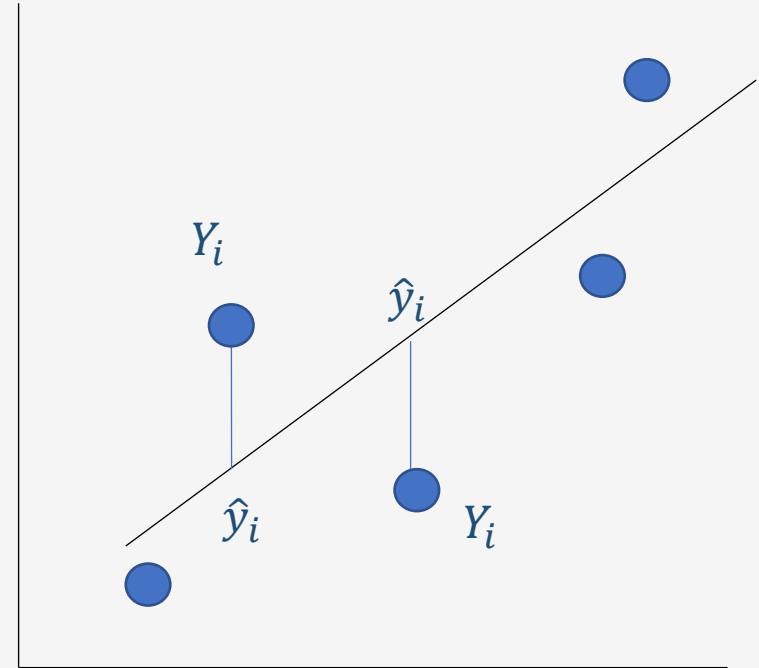
To be the square of the distance between actual and predict Y values

$\hat{y}_i$  = prediction,  $Y_i$  = *actual value*

$$\text{MSE} = \frac{1}{N} \sum_i^n (y_i - \hat{y}_i)^2$$

Best fitted line is the line that minimize the MSE =>

Least Square Methods



## R-square as metrics for determining “goodness” of the fit

---

- Determining the relationship between predictor & outcome
- Relationship Among SST, SSR, SSE

$$r^2 = SSR/SST$$

$$SST = SSR + SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Higher R-square =>  
Lower SSE => Better  
Model

R-square is 0% to  
100%, anything >  
70% is great

# Common Theme, Toolbox and Research workflow in Data Science

---

Apply different algorithms to solve different problems based on the same  
<Theme> and <Research Workflow>

## Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP



## Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

## Common Theme, Toolbox and Research workflow in Data Science

---

Will use Linear Regression for many of the general practices in building models, some of them are

- Split the dataset into training set and a testing set
- Use standard metrics to judge model performance
- K-fold cross validation



# Linear Regression

---

Learning by doing

# Linear Regression Continued

---

## Challenges Number 1 multi-linear regressions

$$Y = \text{beta\_0} + \text{beta\_1} X_1 + \text{beta\_2} X_2 + \dots$$

- Collinearity
  - Pick the factors with highest correlation first, but what about the second factors?
    - Second highest correlation coefficients or lowest correlation with the first factor, but with high enough correlation with the dependent variable
- Solution is: find an Orthogonal independent vectors
  - PCA (Principal Components Analysis)

=> Features Engineering

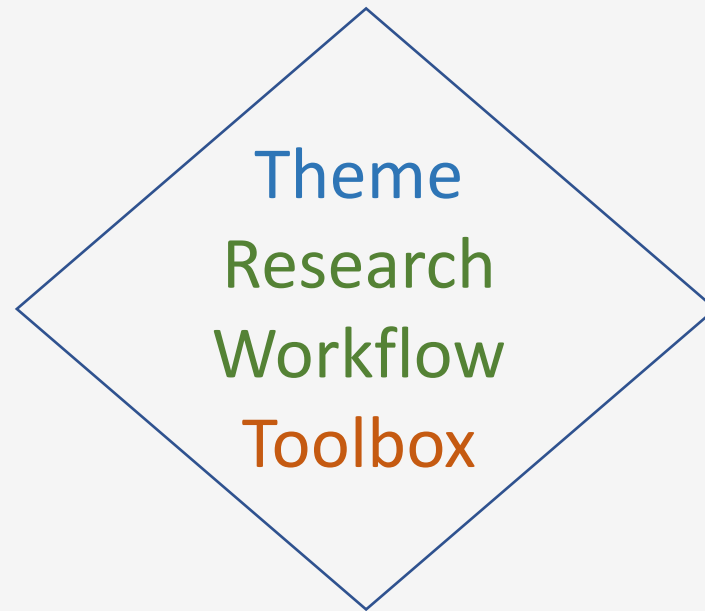
# Common Theme, Toolbox and Research workflow in Data Science

---

Apply different algorithms to solve different problems based on the same  
<Theme> and <Research Workflow>

## Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP



## Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

---

Common  
Theme in  
Machine  
Learning

confusion matrix      bias  
bias vs variance      train test split  
train test split      precision vs recall  
features engineering  
regularization      **overfitting**      cross validation  
encoding categorical var      traintestsplit  
features engineering      cost function  
data normalization      r-square  
model performance  
type ii error      type i error

# Linear Regression

---

## Challenge Number 2:

- Relationship is NOT linear
- Solution: may become linear after transformation

$$Y = a X^2 + b X + c \Rightarrow Y = b_1 Z_1 + b_2 Z_2 + b_3$$

where  $Z_1 = X^2$  and  $Z_2 = X$

$$N = N_0 \exp(-\lambda * t) \Rightarrow \ln(N/N_0) = -\lambda * t + c$$

$$\Rightarrow Y = m X + b$$

where  $Y = \ln(N)$   
 $X = t$

## Polynomial Regression

Learning by doing

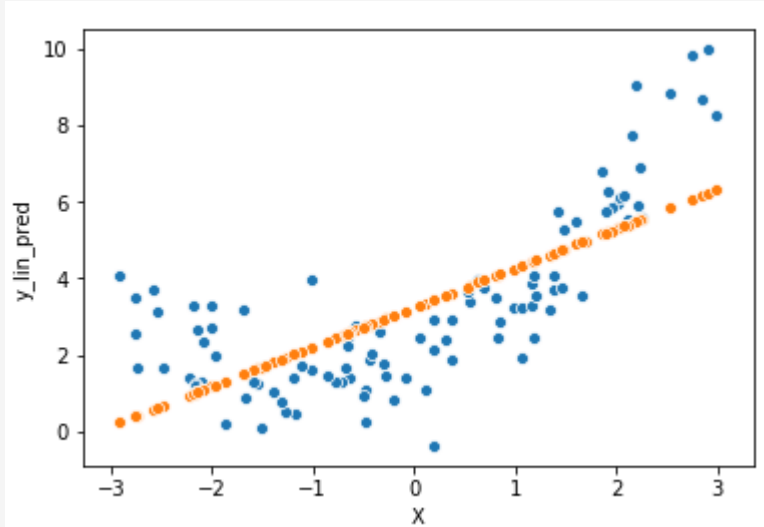
## Simple vs More complicated model

---

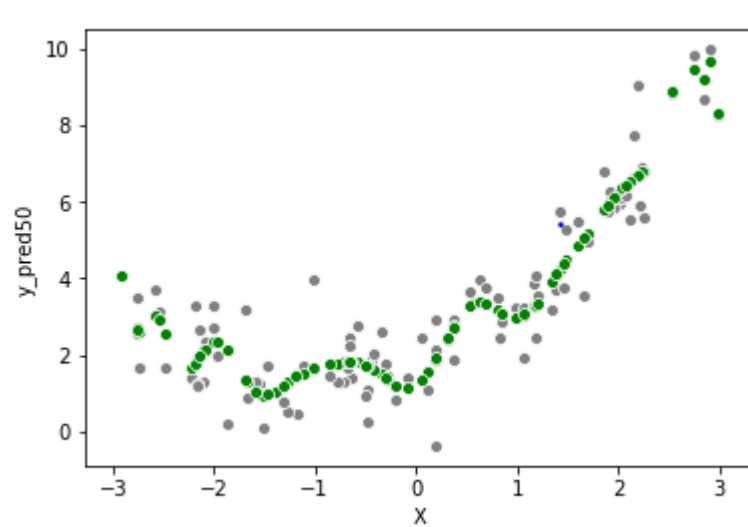
- Using a model with more parameters (more features, more predictors), you are guaranteed to fit your in-sample data (training data) better
  - More parameters  $\Rightarrow$  R-squares always increases
- BUT it doesn't mean you have a better model
  - Adjusted R-squares ( R-squares adjusted by penalizing models with more parameters)

# Lesson Learned from Polynomial Regression

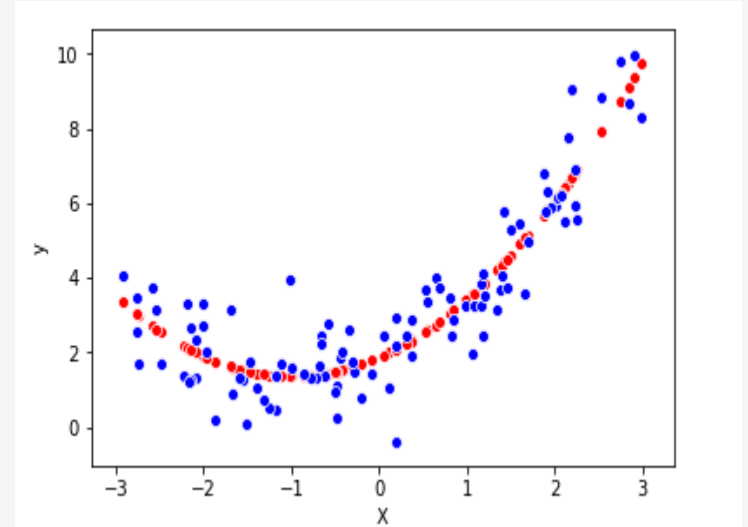
Underfit



Overfit



Good fit



A more sophisticated model tends to have smaller errors in the training set, but can perform worse in testing dataset because it overfit

A too simplistic model will never be able to fit well on both the training set as well as the testing dataset

## Bias vs Variance

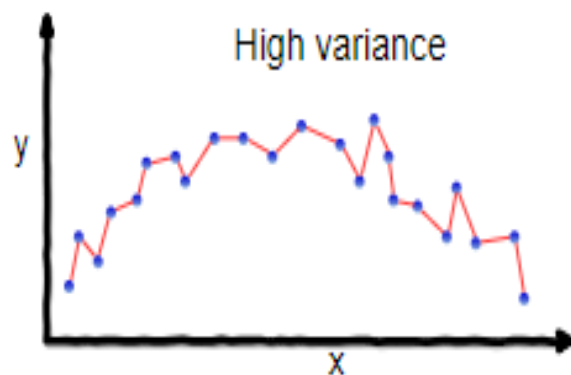
---

- Bias means your model is intrinsically wrong (off, biased) that you will not fit the data well. If you use a too simplistic model, you will have high bias.
- On the other hand, using a more complicated model, you will have low bias. However, your model will not generalize well to testing dataset (out-of-sample data). The “variance” of your prediction will be high
- We call this the Bias vs Variance trade-off

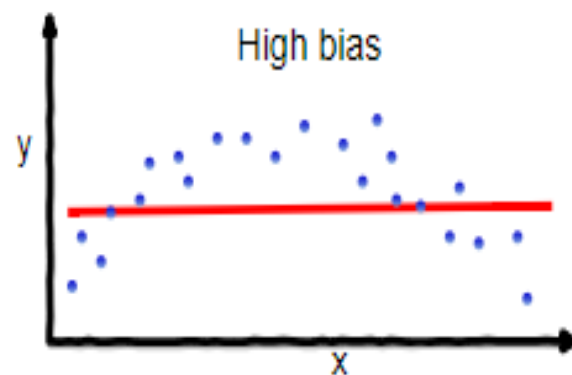


# Bias vs Variance Tradeoff

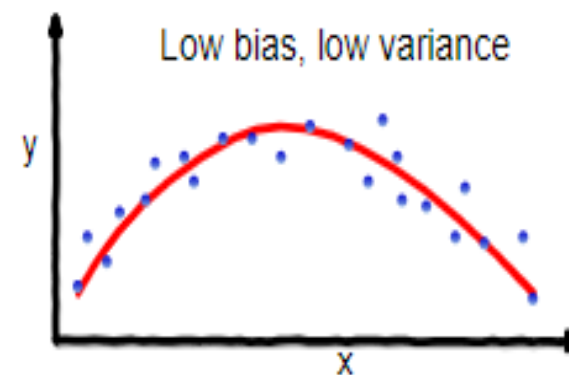
---



**overfitting**

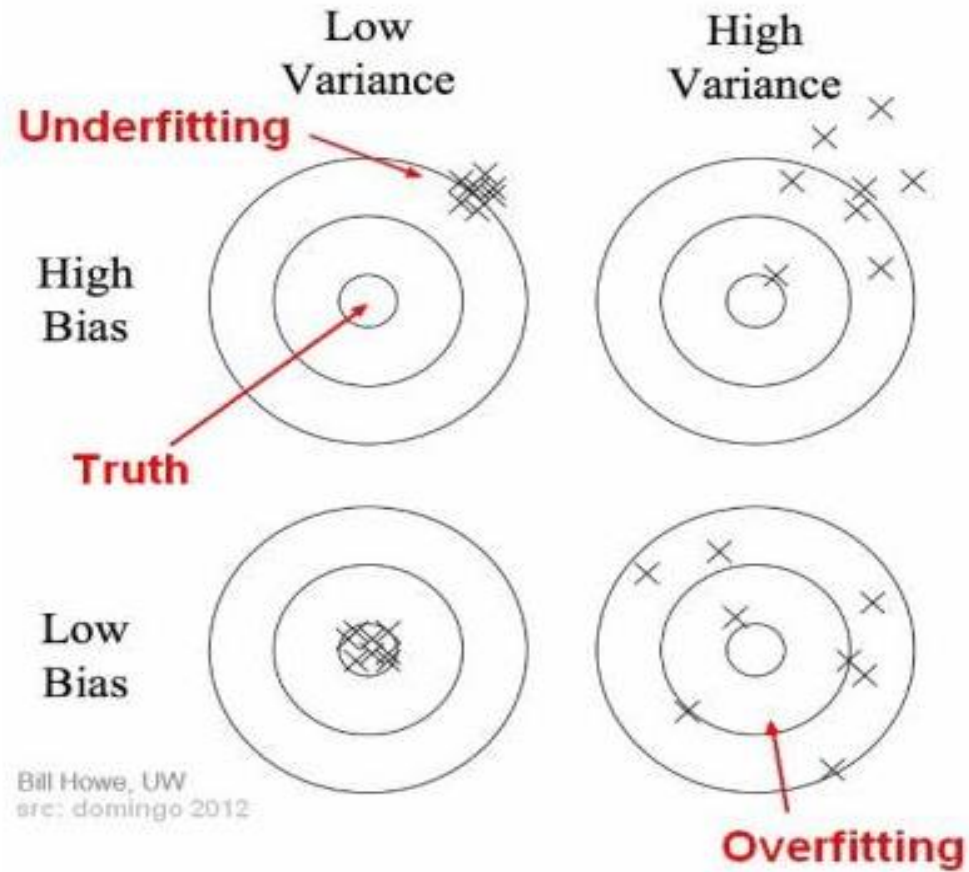


**underfitting**

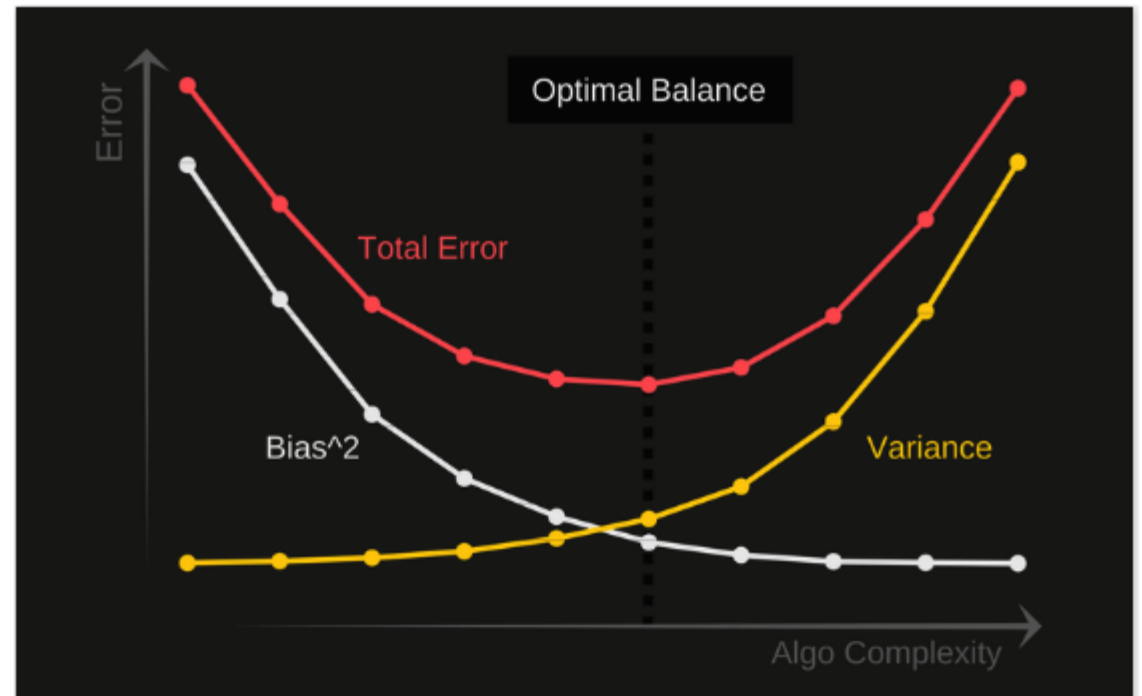


**Good balance**

# Bias vs Variance Tradeoff



$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



## Recall Linear Regression can still apply to non-linear relationship

---

$$Y = a X^2 + b X + c \Rightarrow Y = b_1 Z_1 + b_2 Z_2 + b_3$$

where  $Z_1 = X^2$  and  $Z_2 = X$

$$N = N_0 \exp(-\lambda * t)$$
$$\Rightarrow \ln(N/N_0) = -\lambda * t + c$$
$$\Rightarrow Y = m X + b \text{ where } Y = \ln(N) \text{ and } X = t$$

$$\text{If } Y = \log(P / (1-P)) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots \beta_N + X_N$$

where  $P$  is the probability of something happens

It is called Logistic Regression, which we will cover next

# Classification Problem

---

Linear Regression: Target variable can take any numeric value

Binary Classification Problem: Target variable is either 1 or 0, Yes or No

Multi-class Classification Problem: Target variable is a list of possible values  
(such as classify a picture of animal as a cat, dog, bird, fish picture)

⇒ NEXT TOPICS

⇒ Classification Problem and Logistics Regression