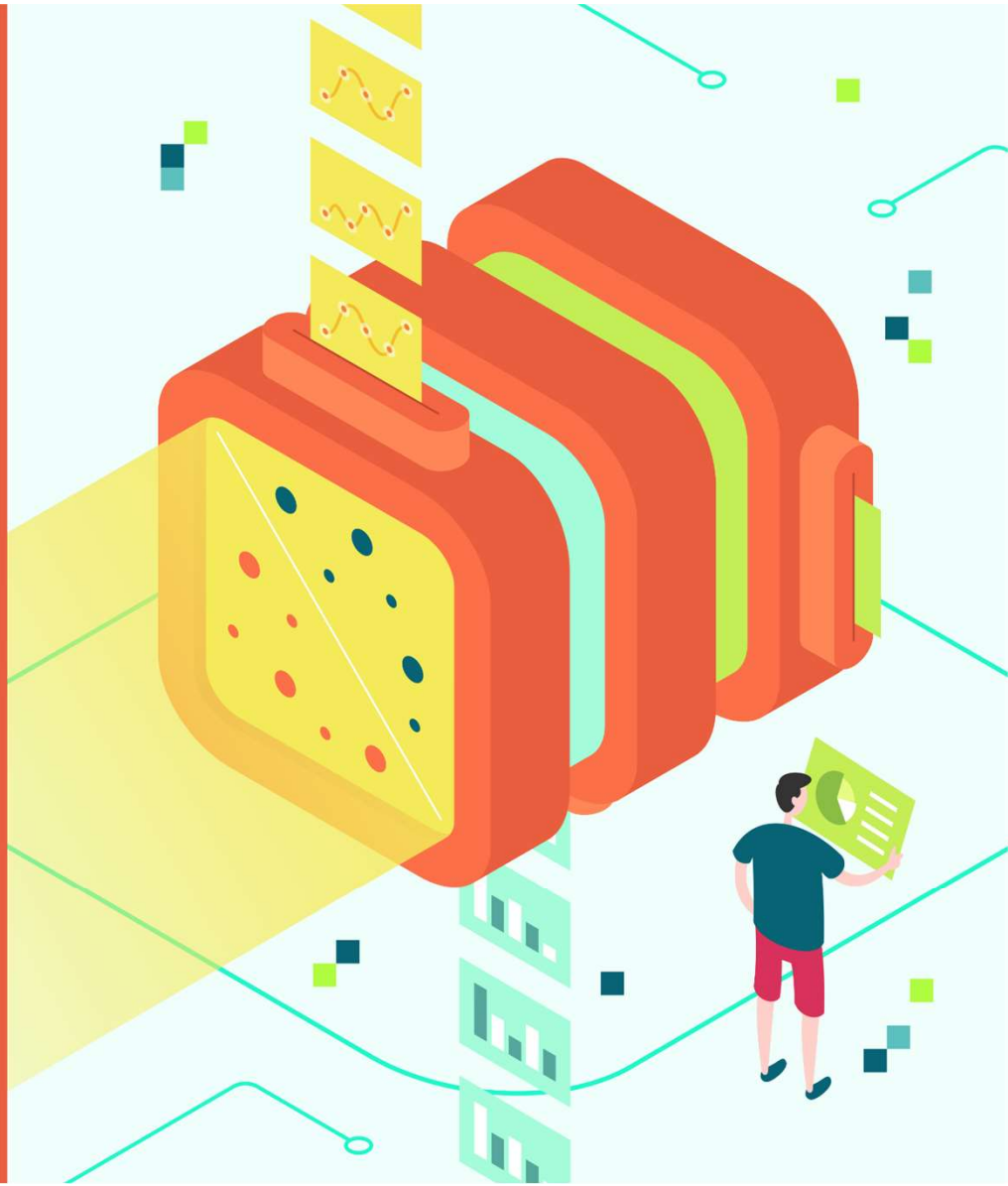
MACHINE LEARNING

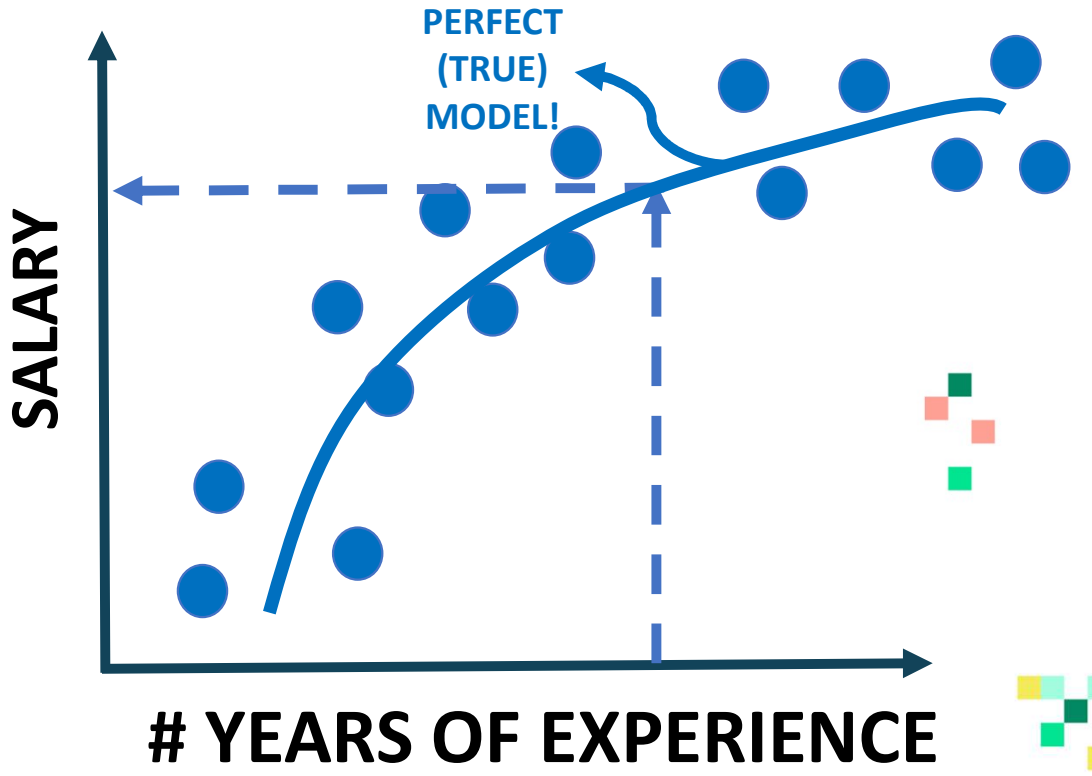
REGRESSION

RIDGE AND LASSO REGRESSION



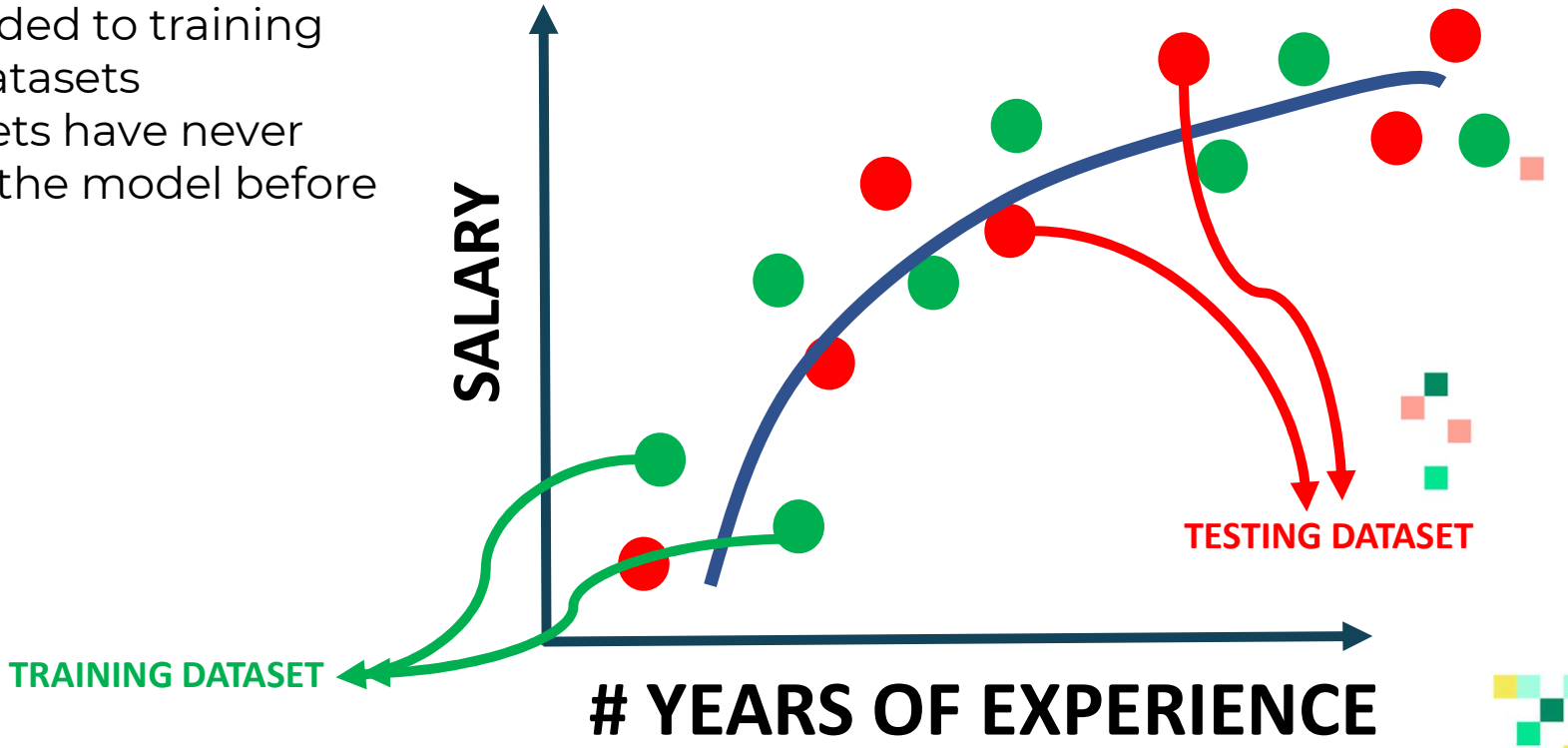
BIAS AND VARIANCE: INTUITION

- Let's assume that we want to get the relationship between the employee salary and number of years of experience
- Fresh graduates tend to have low salaries
- As years of experience increase, the salaries tend to increase as well.
- As number of years go beyond a certain limit, salaries tend to plateau and they do not increase anymore



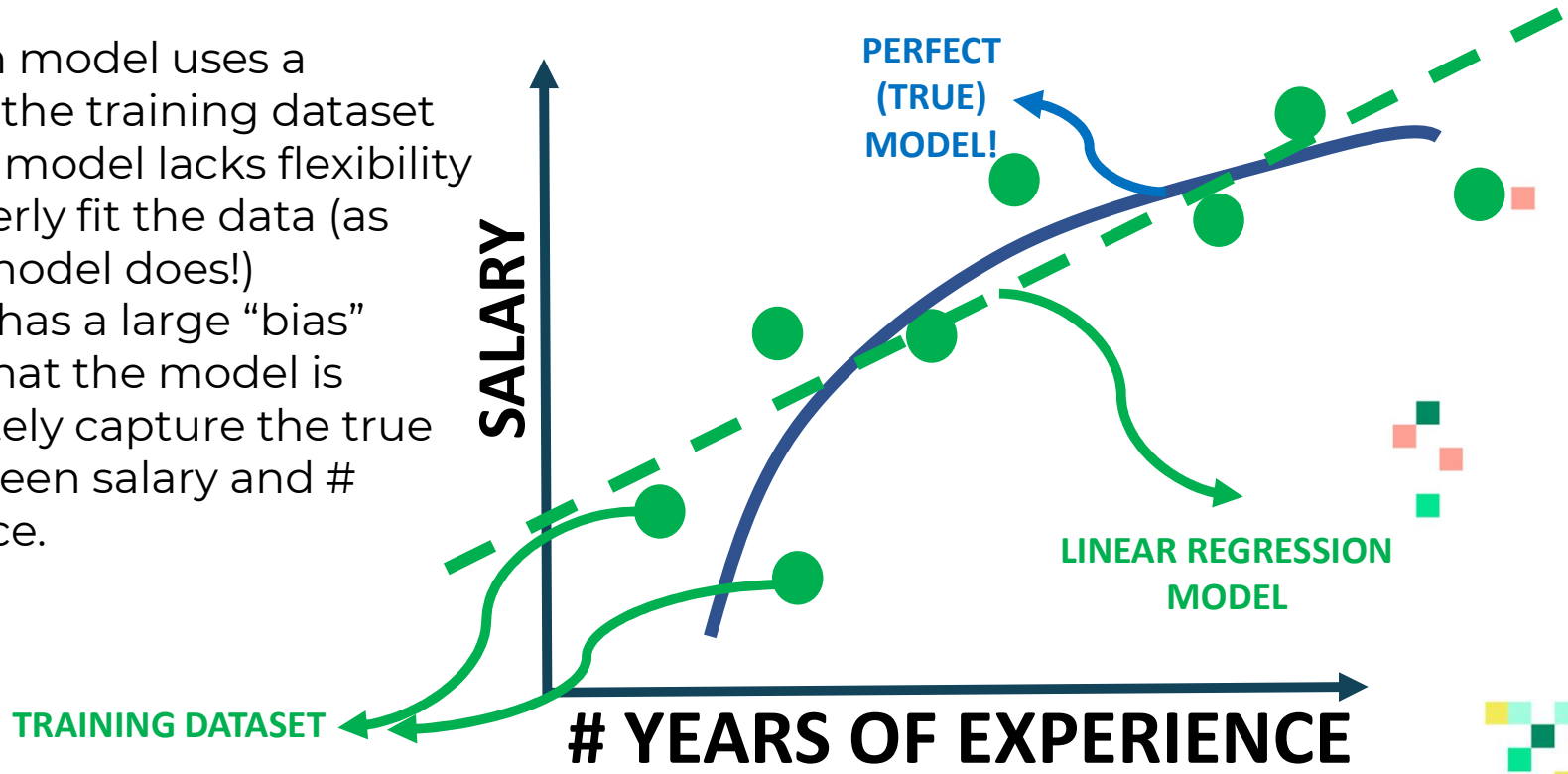
BIAS AND VARIANCE: TRAINING VS. TESTING DATASETS

- Dataset is divided to training and testing datasets
- Testing datasets have never been seen by the model before



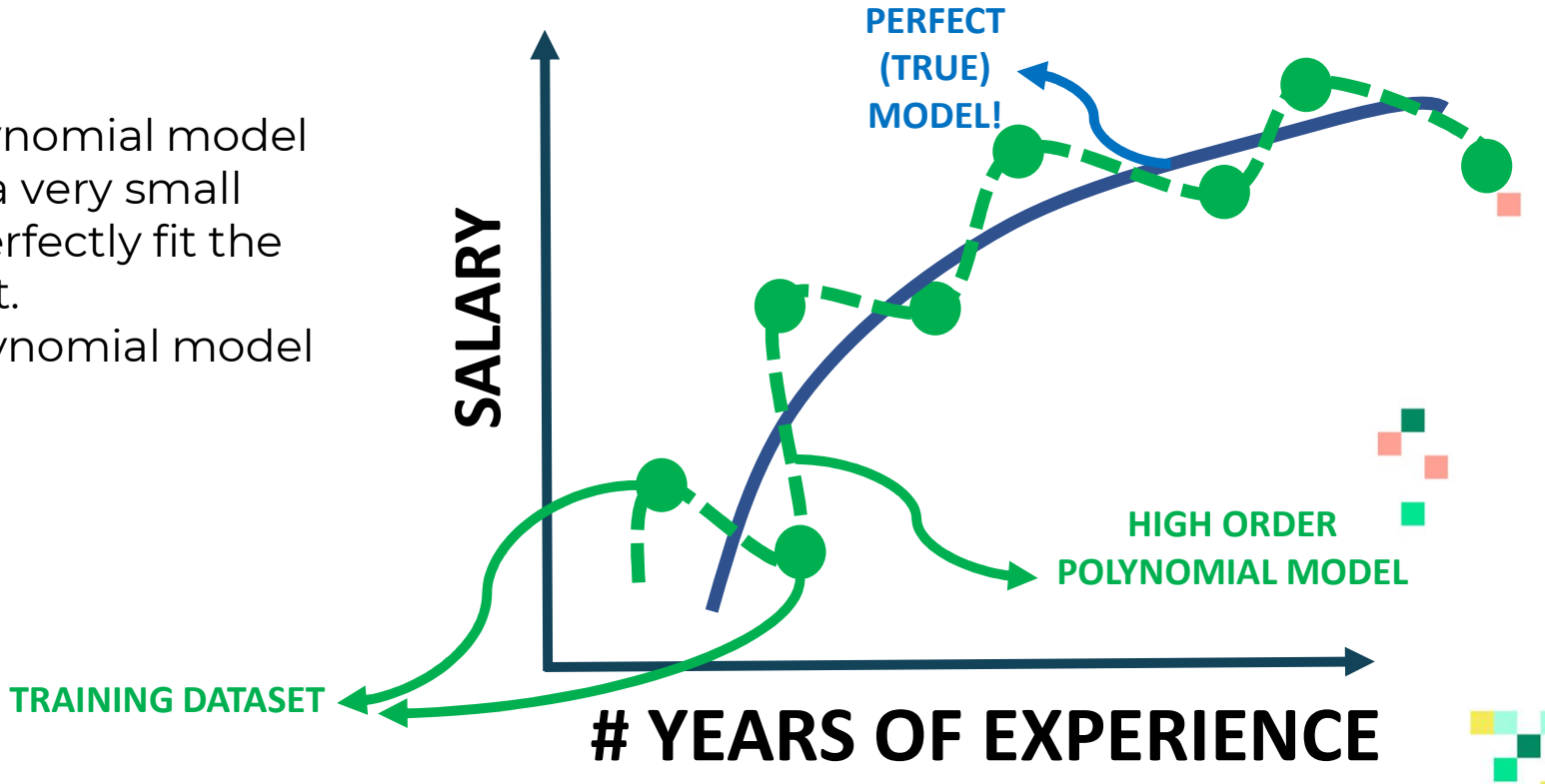
BIAS AND VARIANCE: MODEL #1– LINEAR REGRESSION (SIMPLE)

- Linear Regression model uses a straight line to fit the training dataset
- Linear regression model lacks flexibility so it cannot properly fit the data (as the true perfect model does!)
- The linear model has a large “bias” which indicates that the model is unable to accurately capture the true relationship between salary and # years of experience.

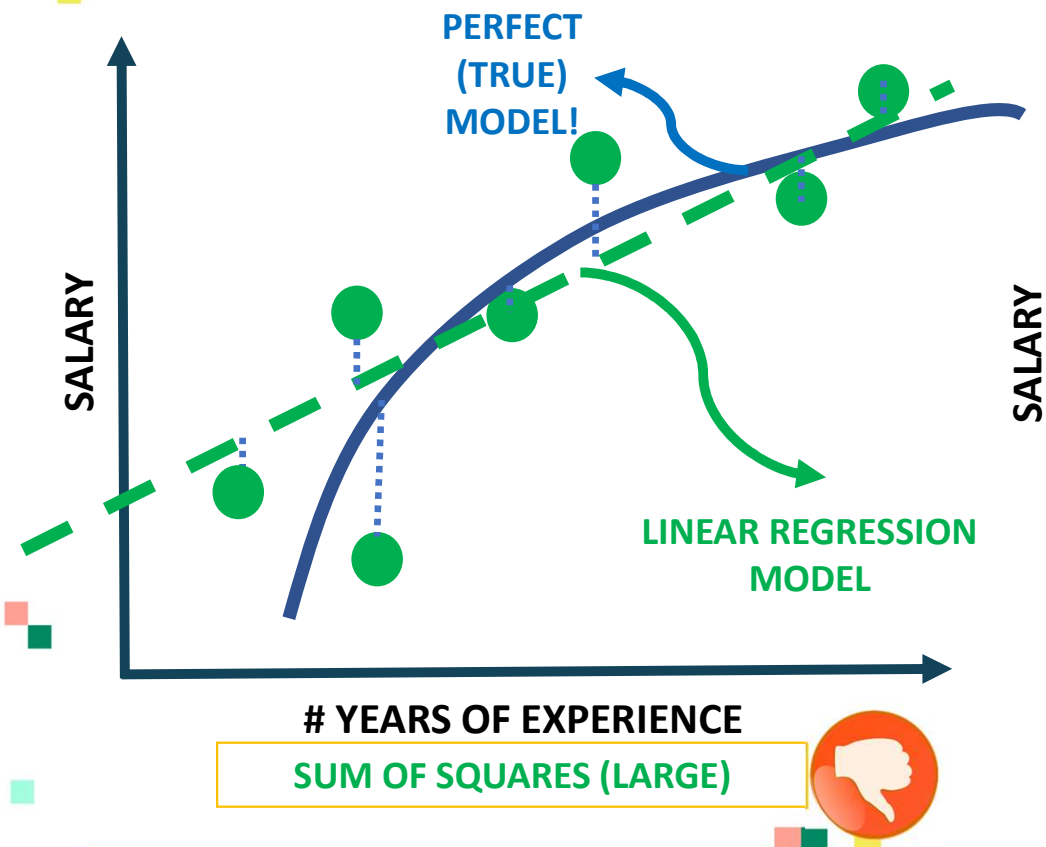


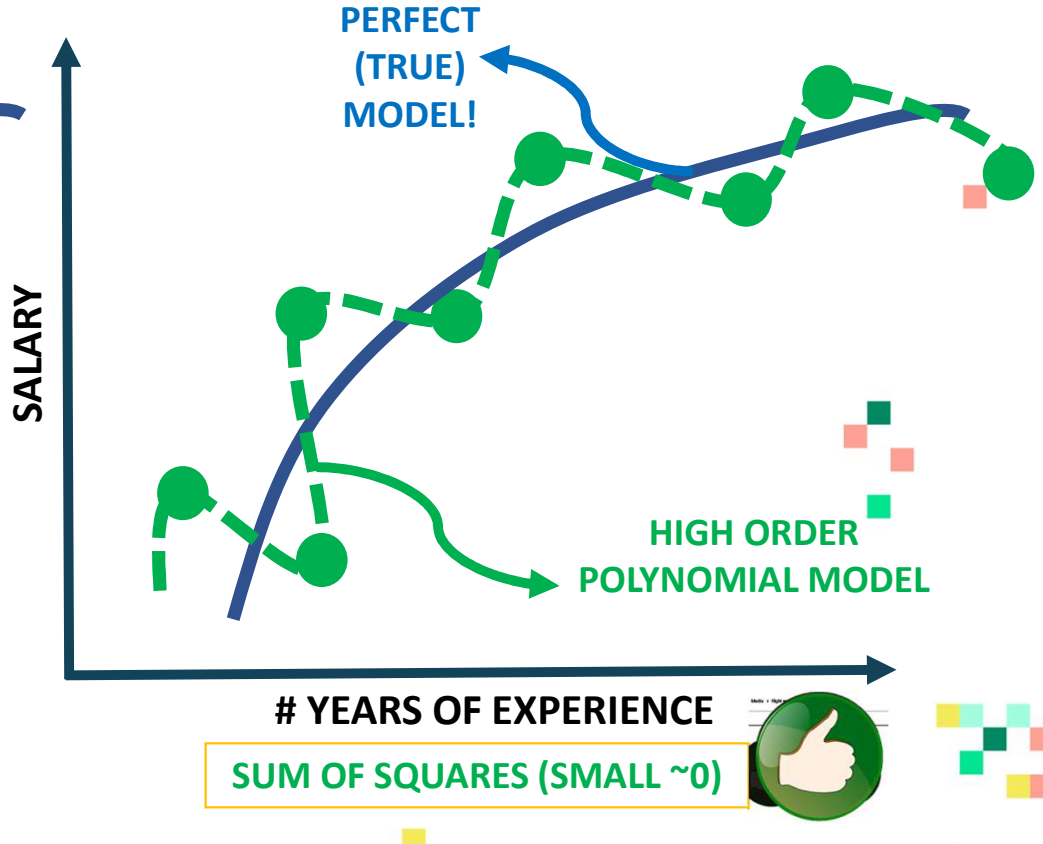
BIAS AND VARIANCE: MODEL #2 – HIGH ORDER POLYNOMIAL REGRESSION (COMPLEX)

- High order polynomial model is able to have a very small bias and can perfectly fit the training dataset.
- High-order polynomial model is very flexible



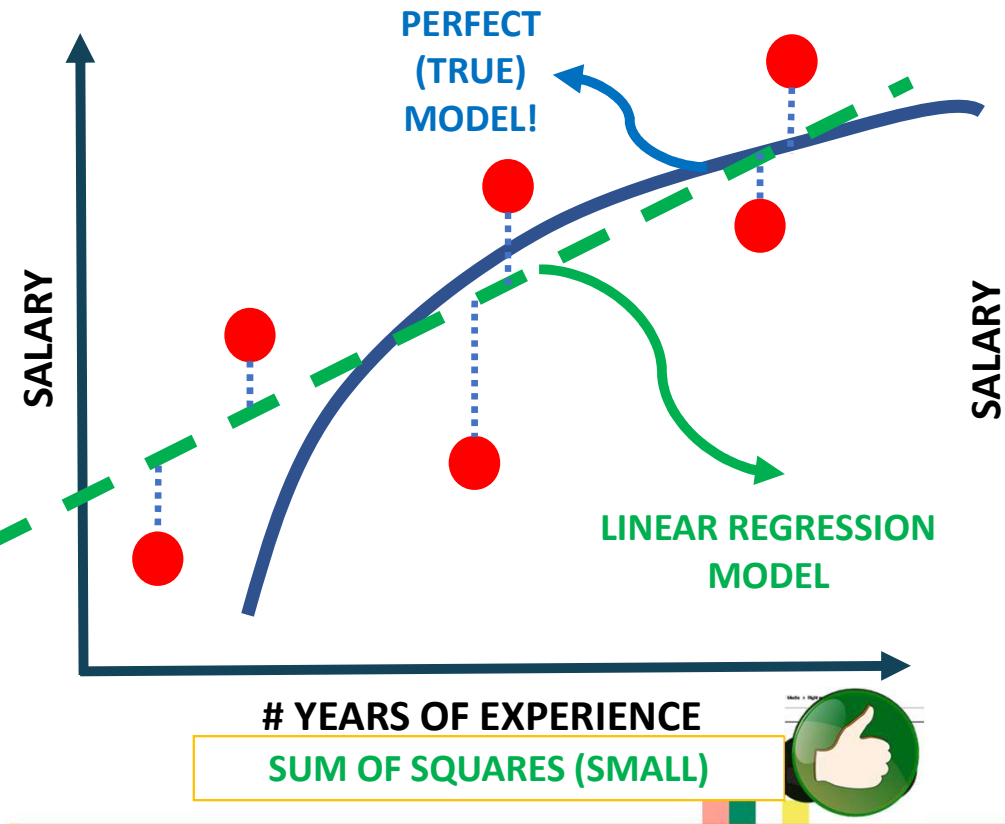
BIAS AND VARIANCE: MODEL #1 Vs. MODEL #2 DURING TRAINING

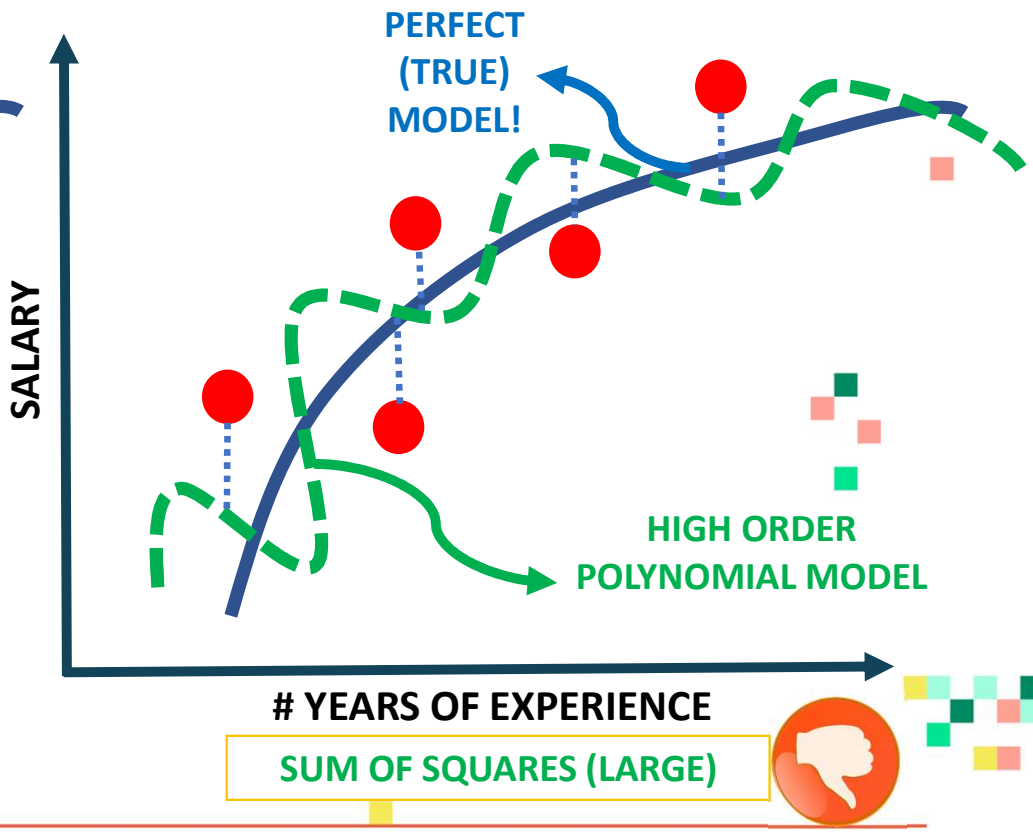




THIS IS NOT THE WHOLE STORY!!

BIAS AND VARIANCE: MODEL #1 Vs. MODEL #2 DURING TESTING

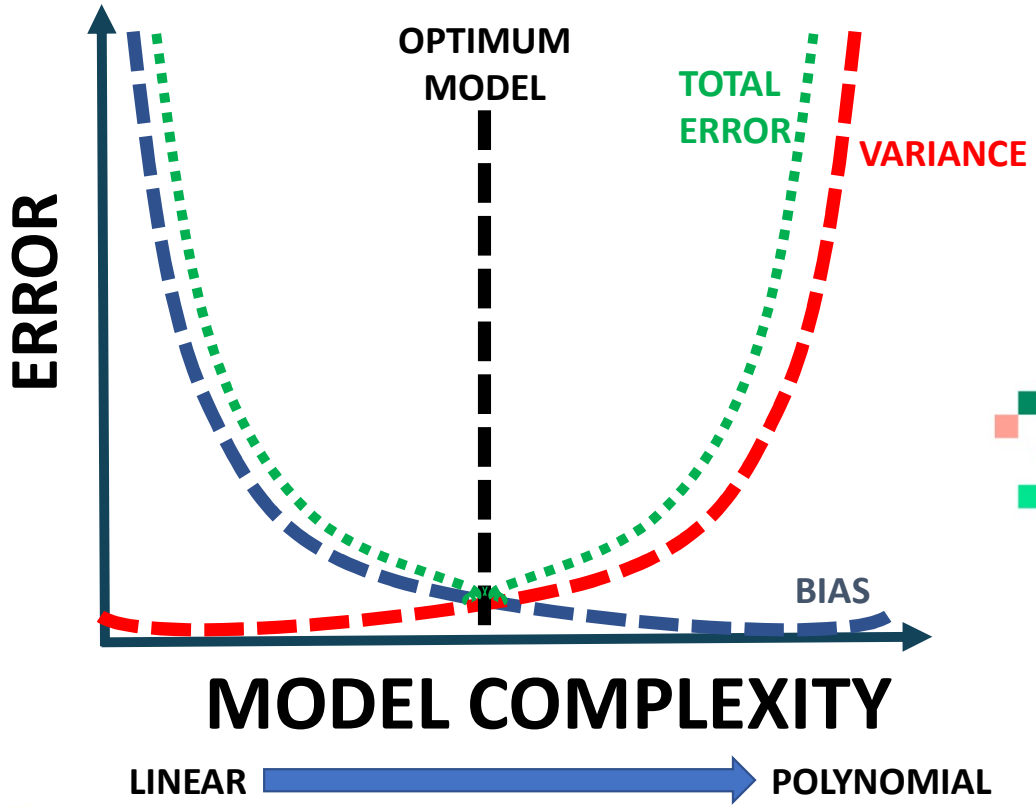




ıe polynomial model performs poorly on the testing dataset and therefore it has large variance

MODEL COMPLEXITY VS. ERROR

- Regularization works by reducing the variance at the cost of adding some bias to the model.
- A trade-off between variance and bias occurs



MODEL COMPLEXITY VS. ERROR

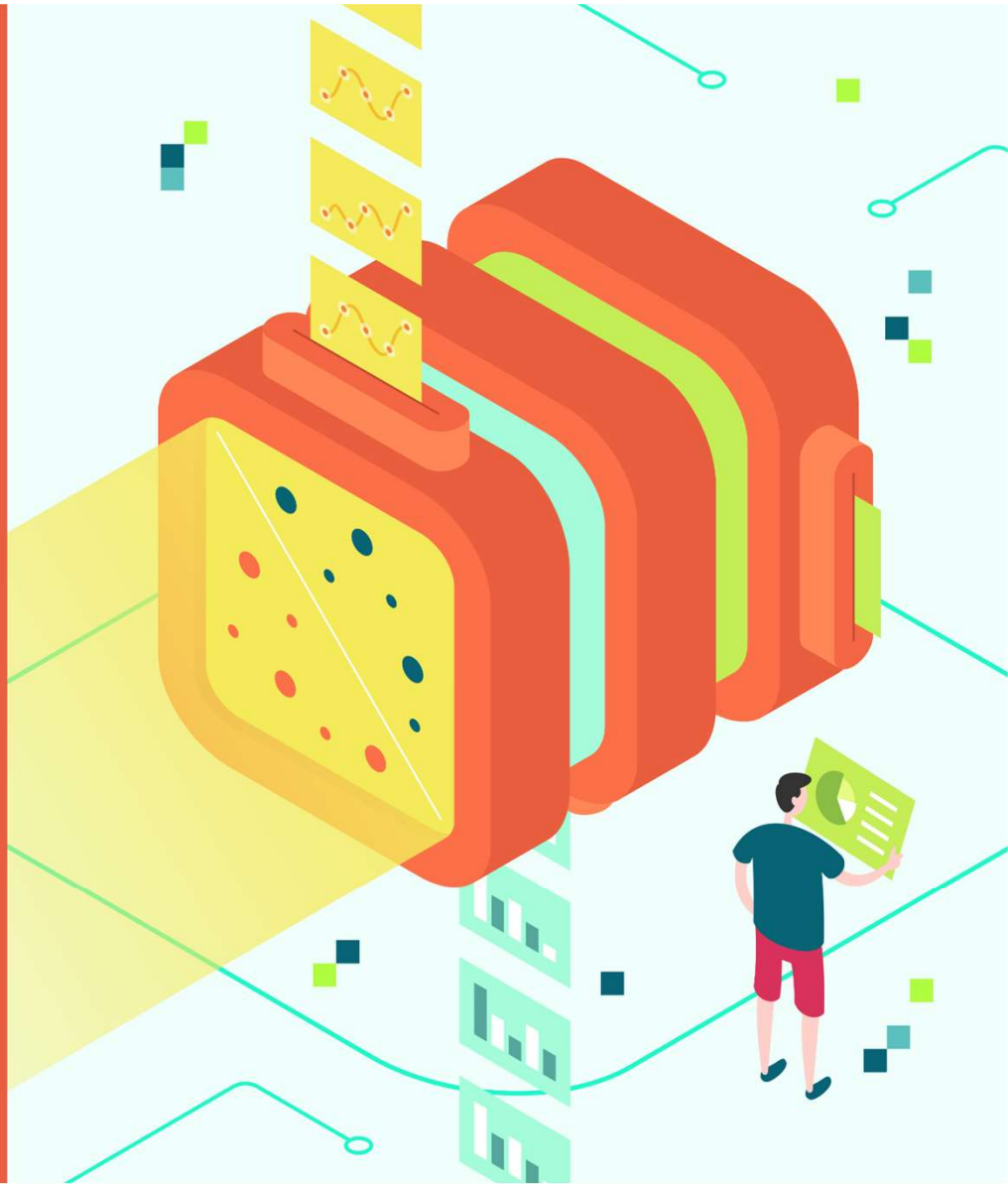
MODEL #1 (LINEAR REGRESSION) (SIMPLE)	MODEL #2 (HIGH ORDER POLYNOMIAL) (COMPLEX)
Model has **High bias** because it is very rigid (not flexible) and cannot fit the training dataset well	Model has **small bias** because it is flexible and can fit the training dataset very well.
Has **small variance (variability)** because it can fit the training data and the testing data with similar level (the model is able to generalize better) and avoids overfitting	Has **large variance (variability)** because the model over fitted the training dataset and it performs poorly on the testing dataset
Performance is consistent between the training dataset and the testing dataset	Performance varies greatly between the training dataset and the testing dataset (high variability)
Good generalization	Over fitted

- *Variance measures the difference in fits between the training dataset and the testing dataset*
- *If the model generalizes better, the model has small variance which means the model performance is consistent among the training and testing datasets*
- *If the model over fits the training dataset, the model has large variance*

PERFECT REGRESSION MODEL SHALL HAVE SMALL BIAS AND SMALL VARIABILITY!
A TRADEOFF BETWEEN THE BIAS AND VARIANCE SHALL BE PERFORMED FOR ULTIMATE RESULTS

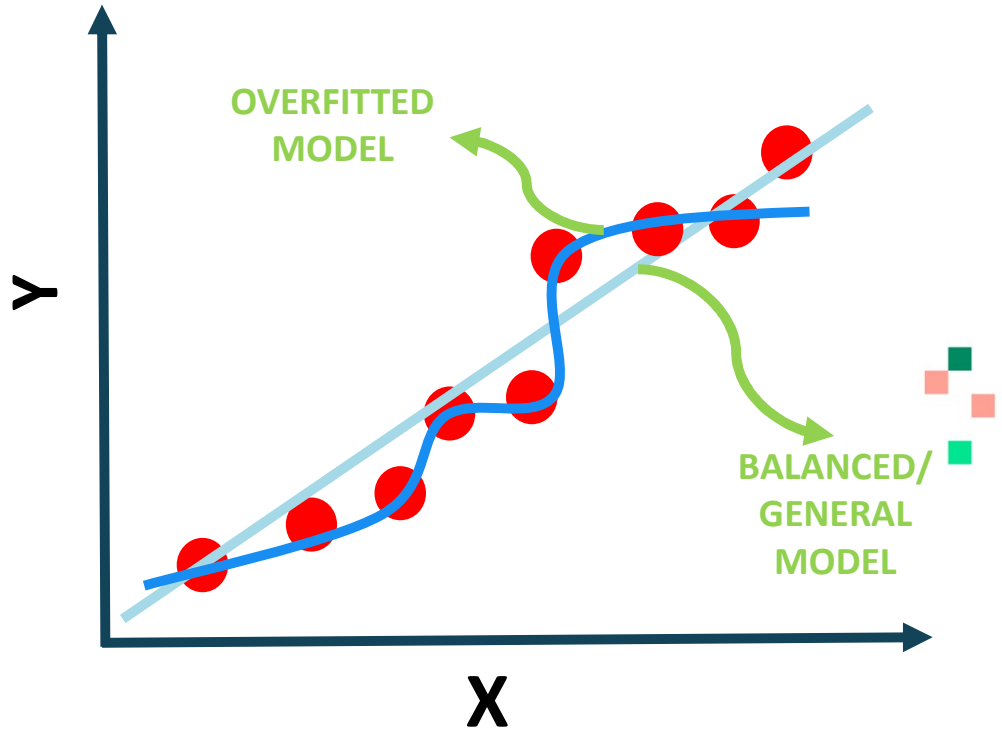
MACHINE LEARNING REGRESSION

RIDGE REGRESSION



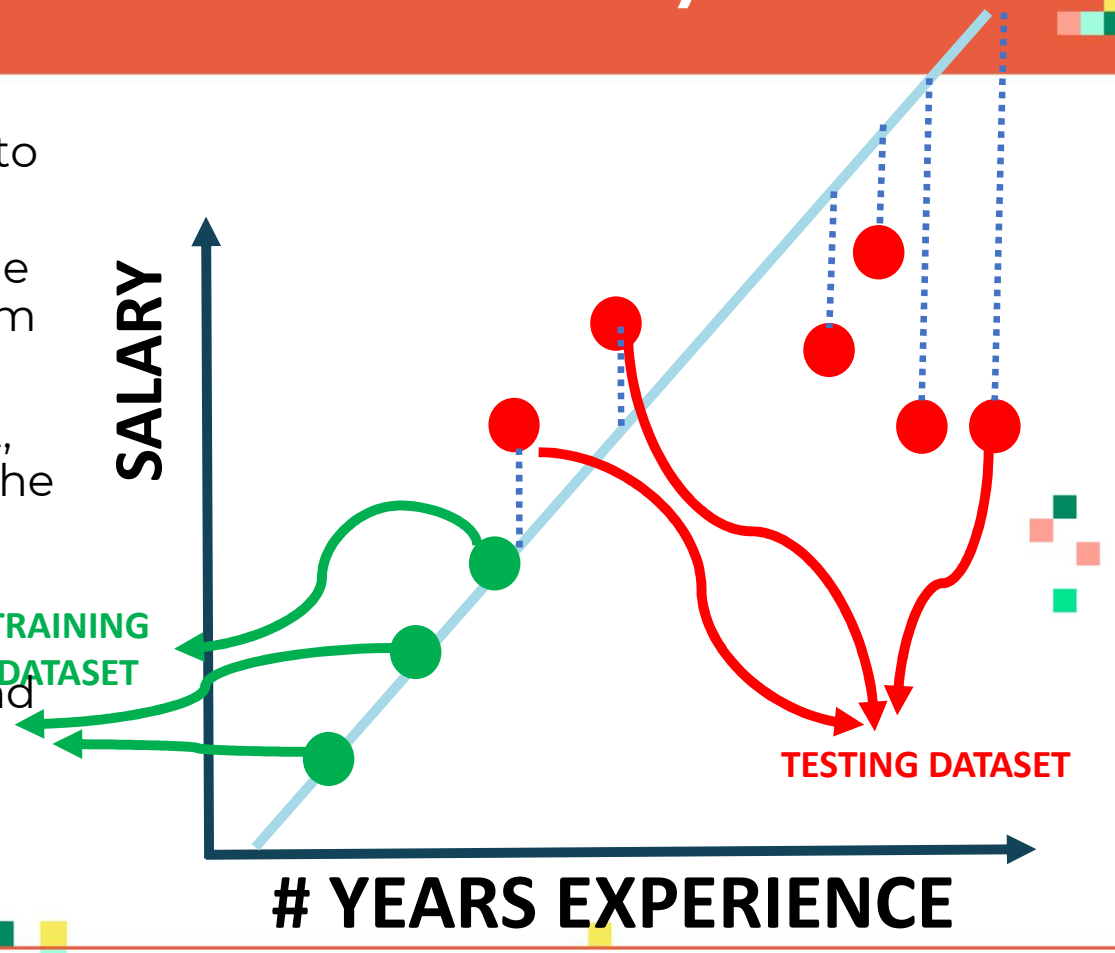
RIDGE REGRESSION (L2 REGULARIZATION): INTUITION

- Ridge regression advantage is to avoid overfitting.
- Our ultimate model is the one that could generalize patterns; i.e.: works best on the training and testing dataset
- Overfitting occurs when the trained model performs well on the training data and performs poorly on the testing datasets
- Ridge regression works by applying a penalizing term (reducing the weights and biases) to overcome overfitting.



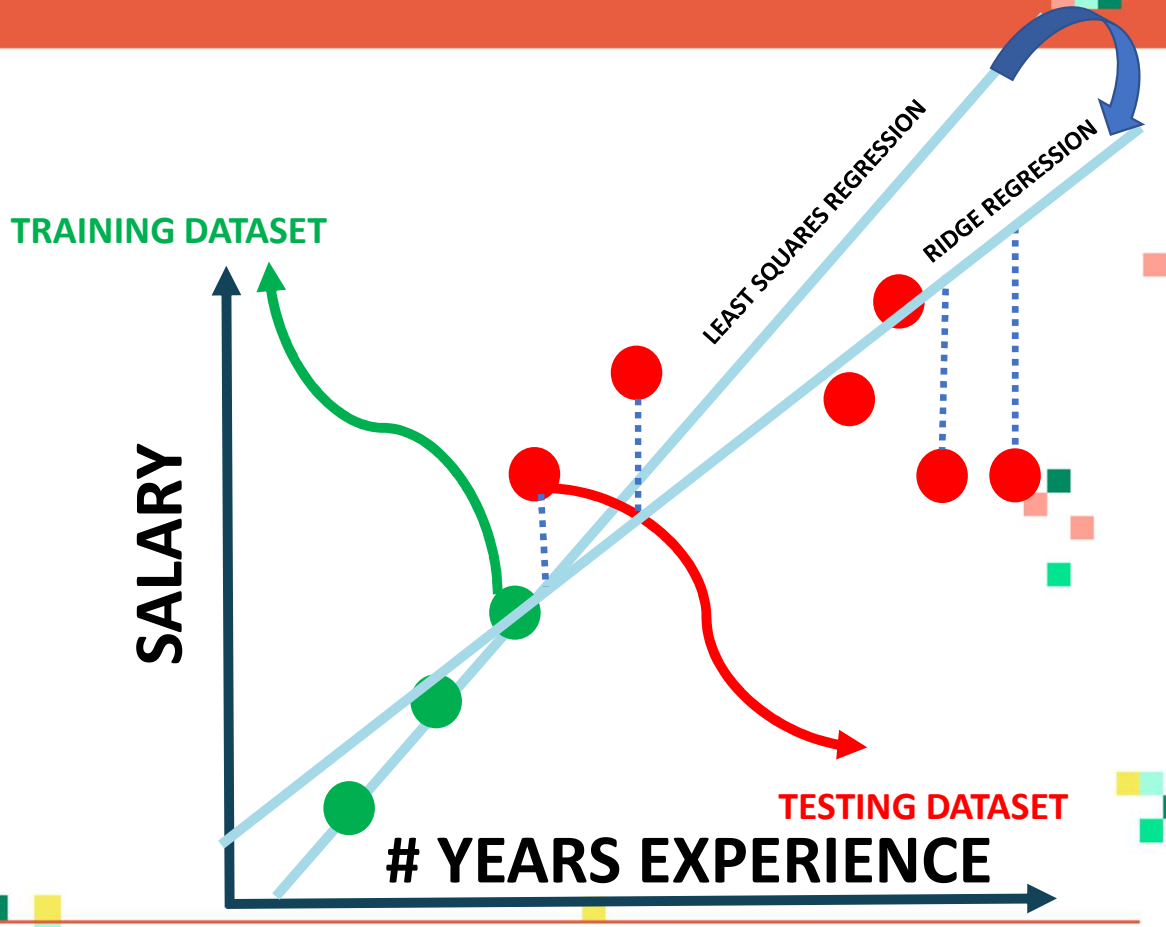
RIDGE REGRESSION (L2 REGULARIZATION): INTUITION

- Least sum of squares is applied to obtain the best fit line
- Since the line passes through the 3 training dataset points, the sum of squared residuals = 0
- However, for the testing dataset, the sum of residuals is large so the line has a high variance.
- Variance means that there is a difference in fit (or variability) between the training dataset and the testing dataset.
- This regression model is overfitting the training dataset



RIDGE REGRESSION (L2 REGULARIZATION): INTUITION

- Ridge regression works by attempting at increasing the bias to improve variance (generalization capability)
- This works by changing the slope of the line
- The model performance might be little poor on the training set but it will perform consistently well on both the training and testing datasets.



RIDGE REGRESSION (L2 REGULARIZATION): MATH

- Slope has been reduced with ridge regression penalty and therefore the model becomes less sensitive to changes in the independent variable (#Years of experience)

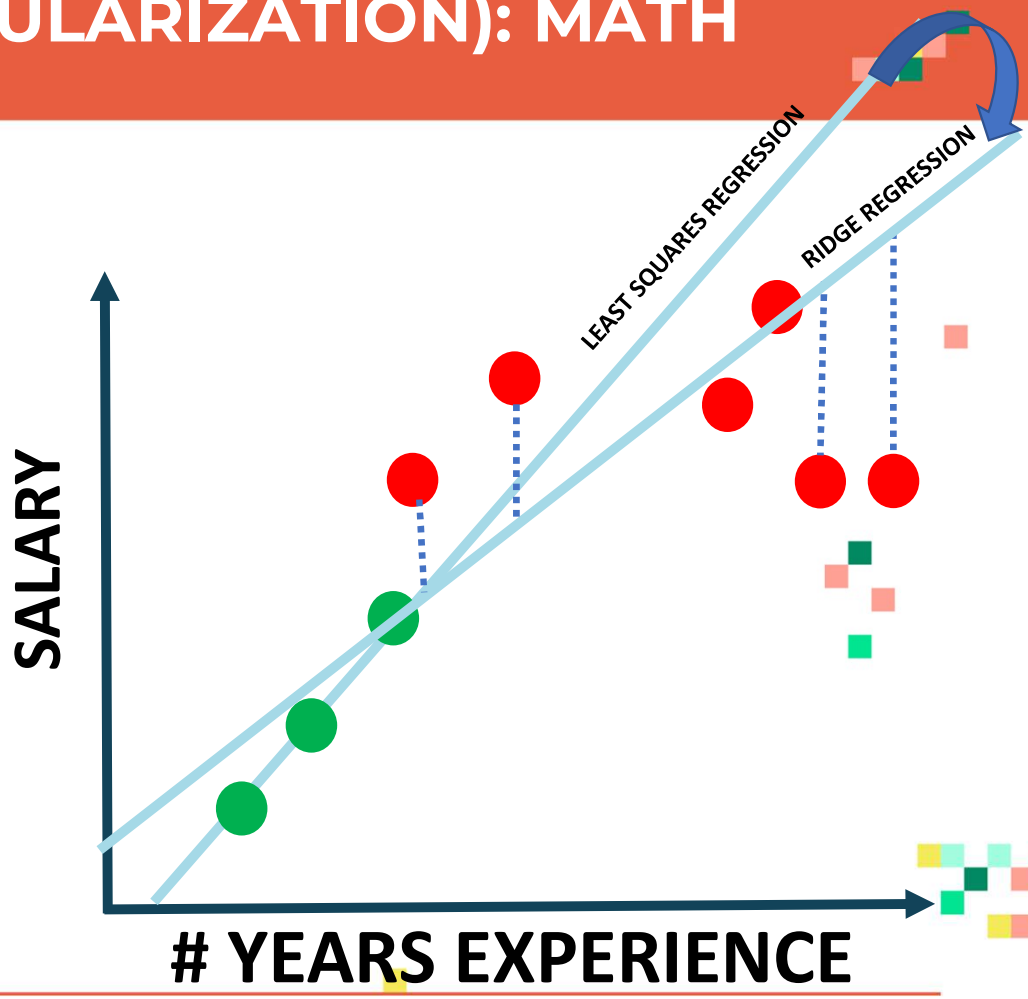
PENALTY TERM

Least Squares Regression:

$$Min(sum\ of\ the\ squared\ residuals)$$

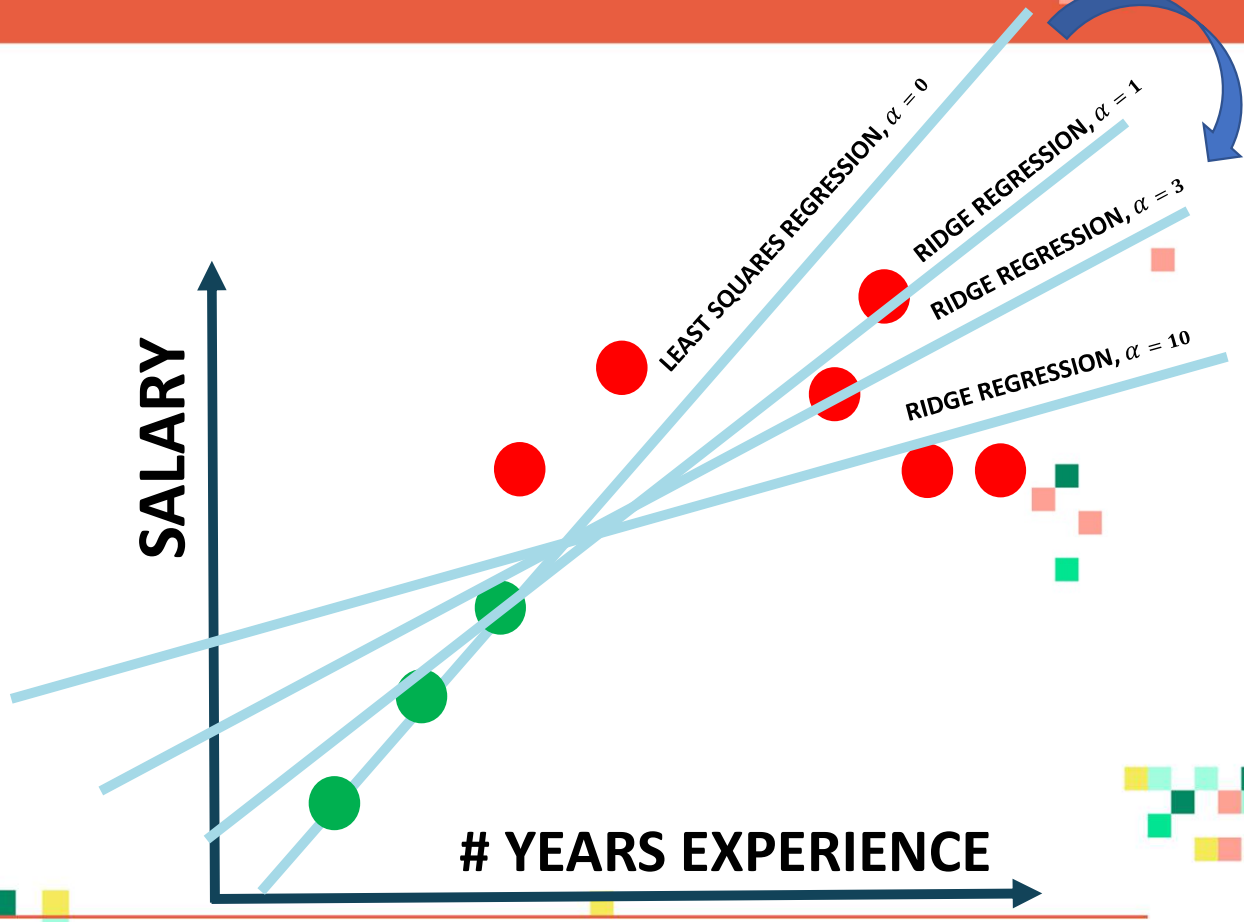
Ridge Regression:

$$Min(sum\ of\ squared\ residuals + \alpha * slope^2)$$



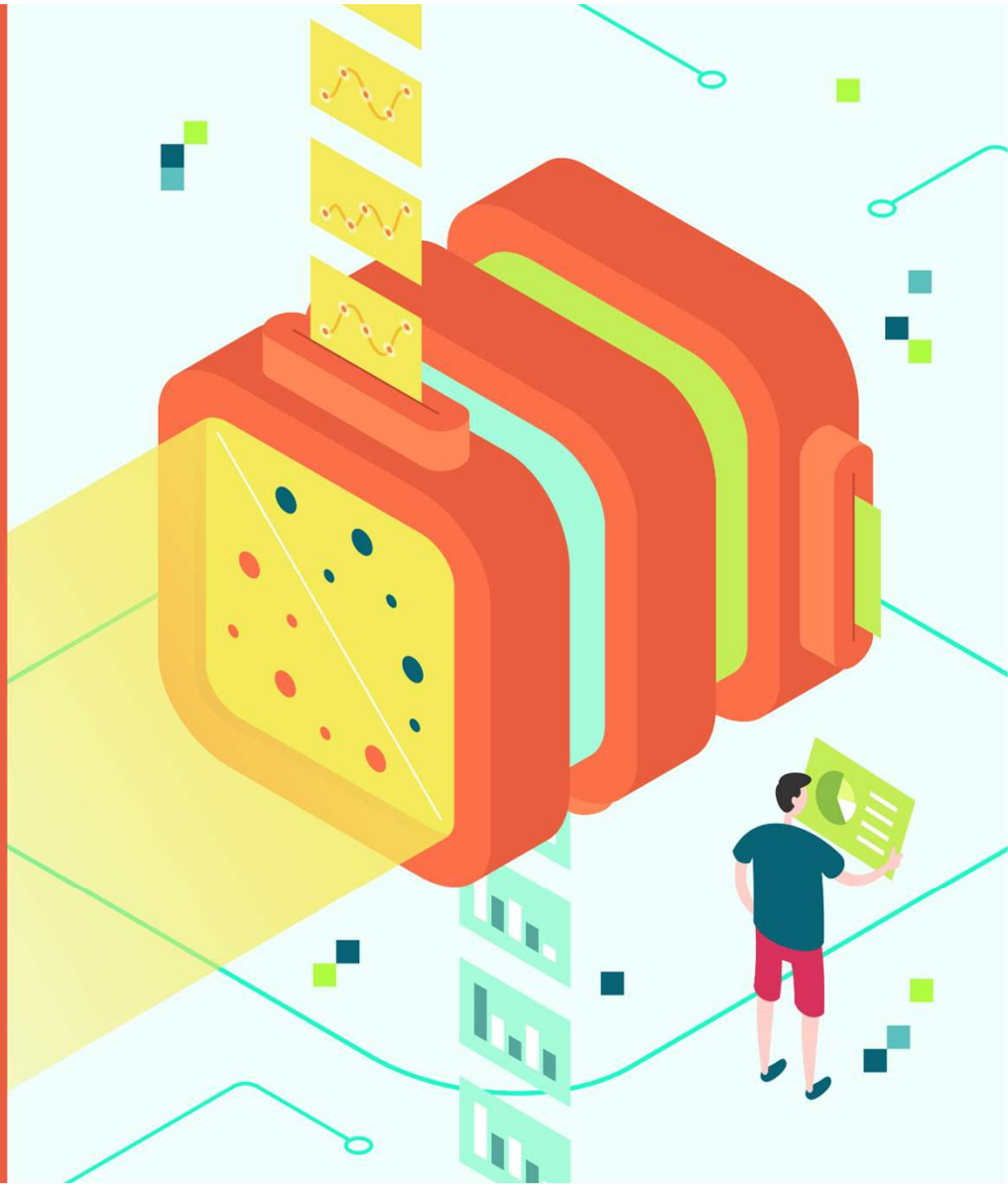
RIDGE REGRESSION (L2 REGULARIZATION): ALPHA EFFECT

- As Alpha increases, the slope of the regression line is reduced and becomes more horizontal.
- As Alpha increases, the model becomes less sensitive to the variations of the independent variable (# Years of experience)



MACHINE
LEARNING
REGRESSION

LASSO REGRESSION



LASSO REGRESSION: MATH

- Lasso Regression is similar to Ridge regression
- It works by introducing a bias term but instead of squaring the slope, the absolute value of the slope is added as a penalty term

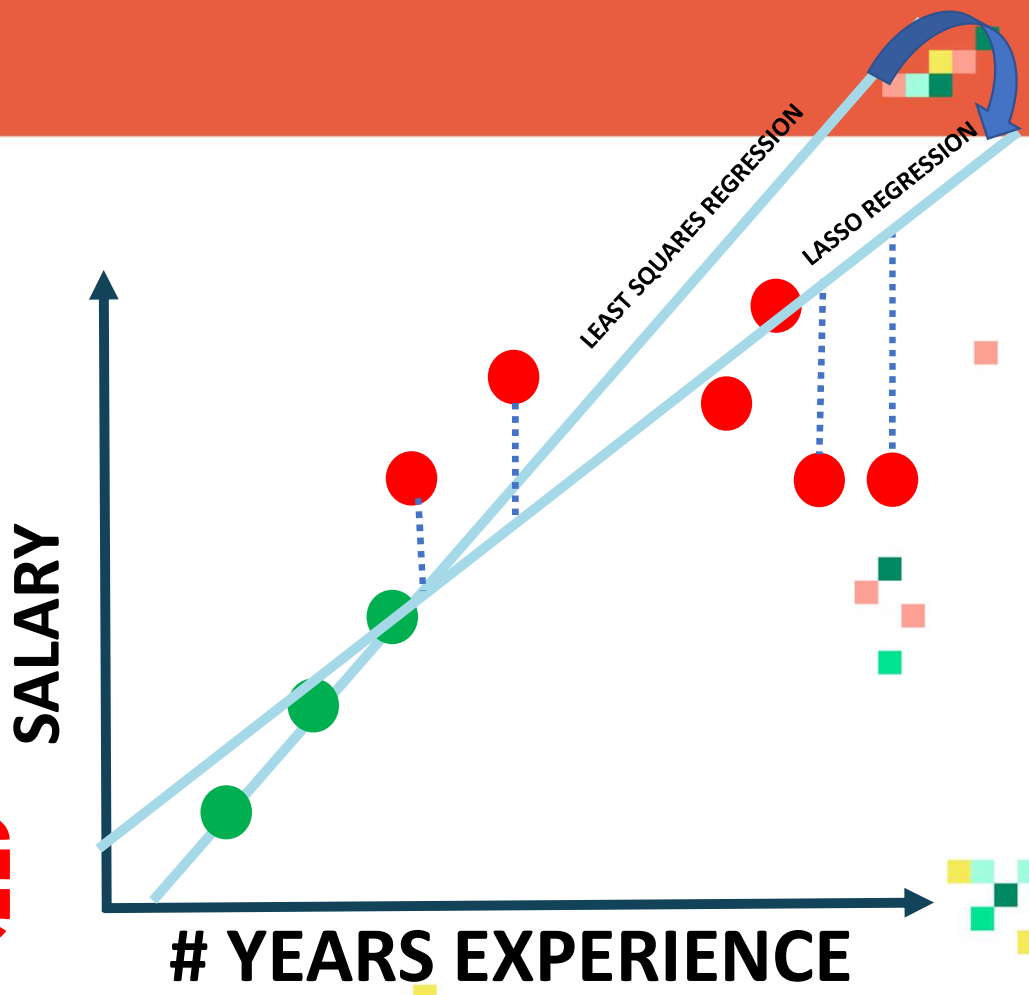
PENALTY TERM

Least Squares Regression:

$$Min(sum\ of\ the\ squared\ residuals)$$

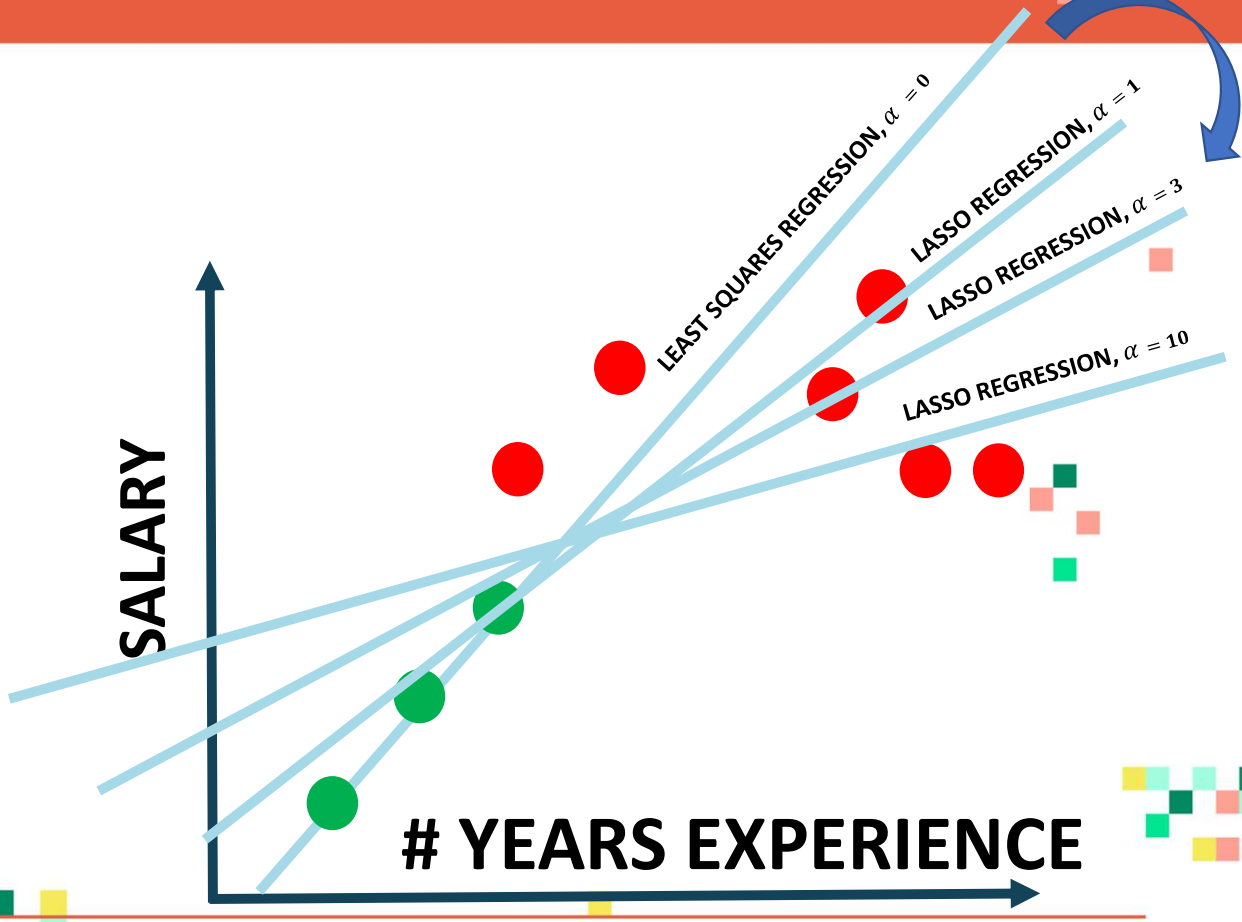
Lasso Regression:

$$Min(sum\ of\ squared\ residuals + \alpha * |slope|)$$



RIDGE REGRESSION (L2 REGULARIZATION): ALPHA EFFECT

- The effect of Alpha on Lasso regression is similar to its effect on ridge regression
- As Alpha increases, the slope of the regression line is reduced and becomes more horizontal.
- As Alpha increases, the model becomes less sensitive to the variations of the independent variable (# Years of experience)



LASSO REGRESSION: MATH

- Lasso regression helps reduce overfitting and it is particularly useful for feature selection
- Lasso regression can be useful if we have several independent variables that are useless
- Ridge regression can reduce the slope close to zero (but not exactly zero) but Lasso regression can reduce the slope to be exactly equal to zero.

Least Squares Regression:

$$Min(sum\ of\ the\ squared\ residuals)$$

Ridge Regression:

$$Min(sum\ of\ squared\ residuals + \alpha * slope^2)$$

Lasso Regression:

$$Min(sum\ of\ squared\ residuals + \alpha * |slope|)$$