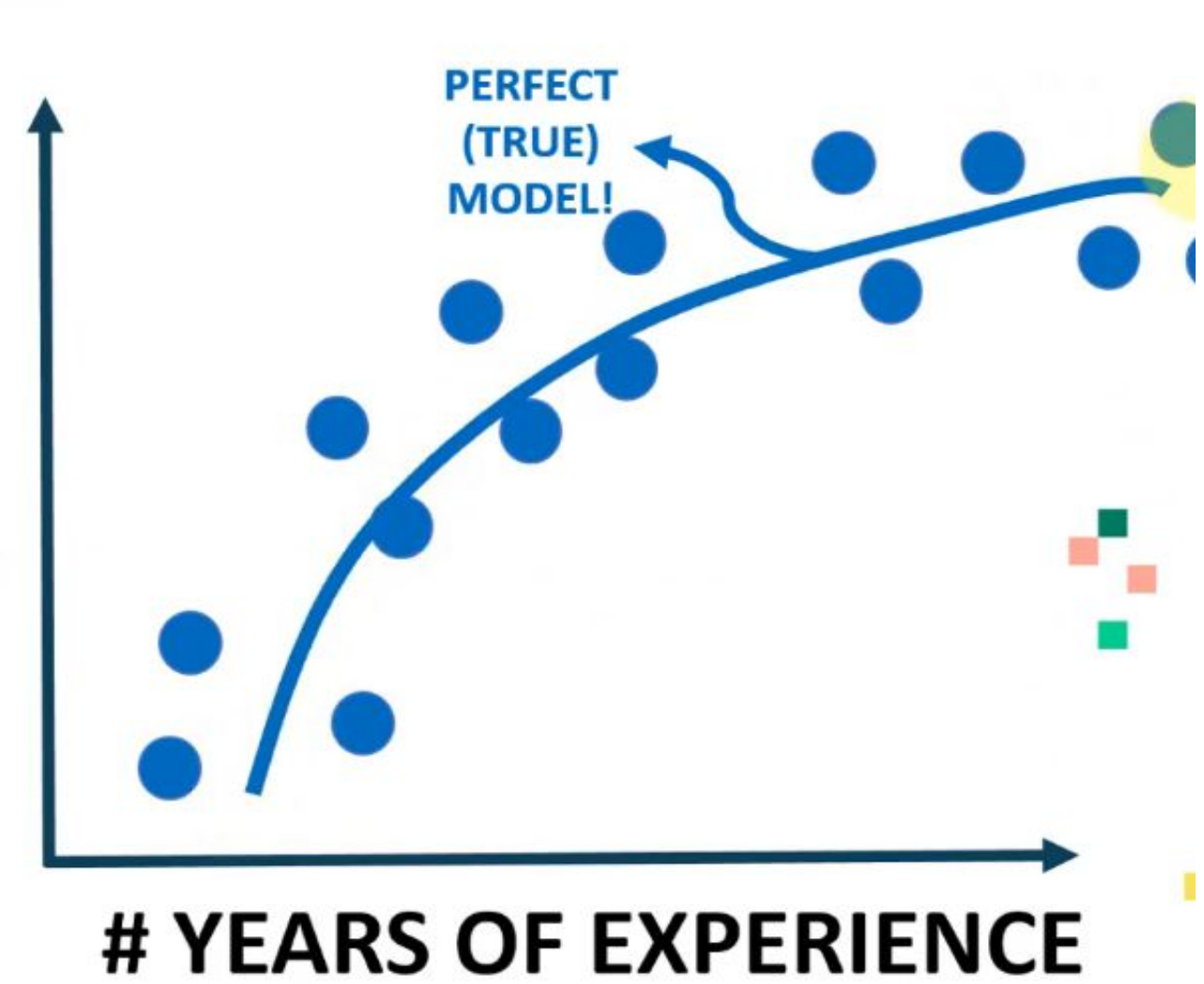
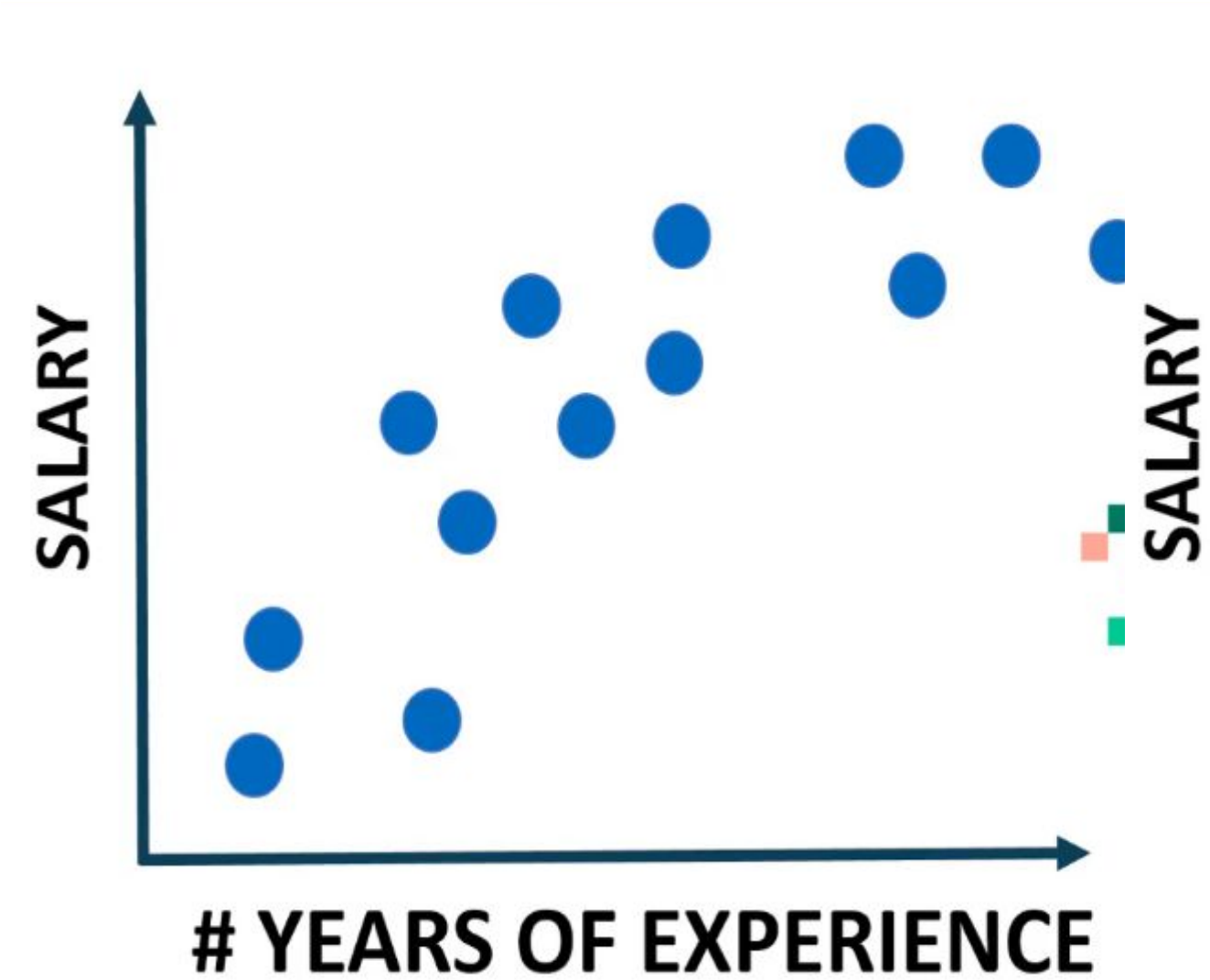


# BIAS AND VARIANCE: INTUITION

- 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





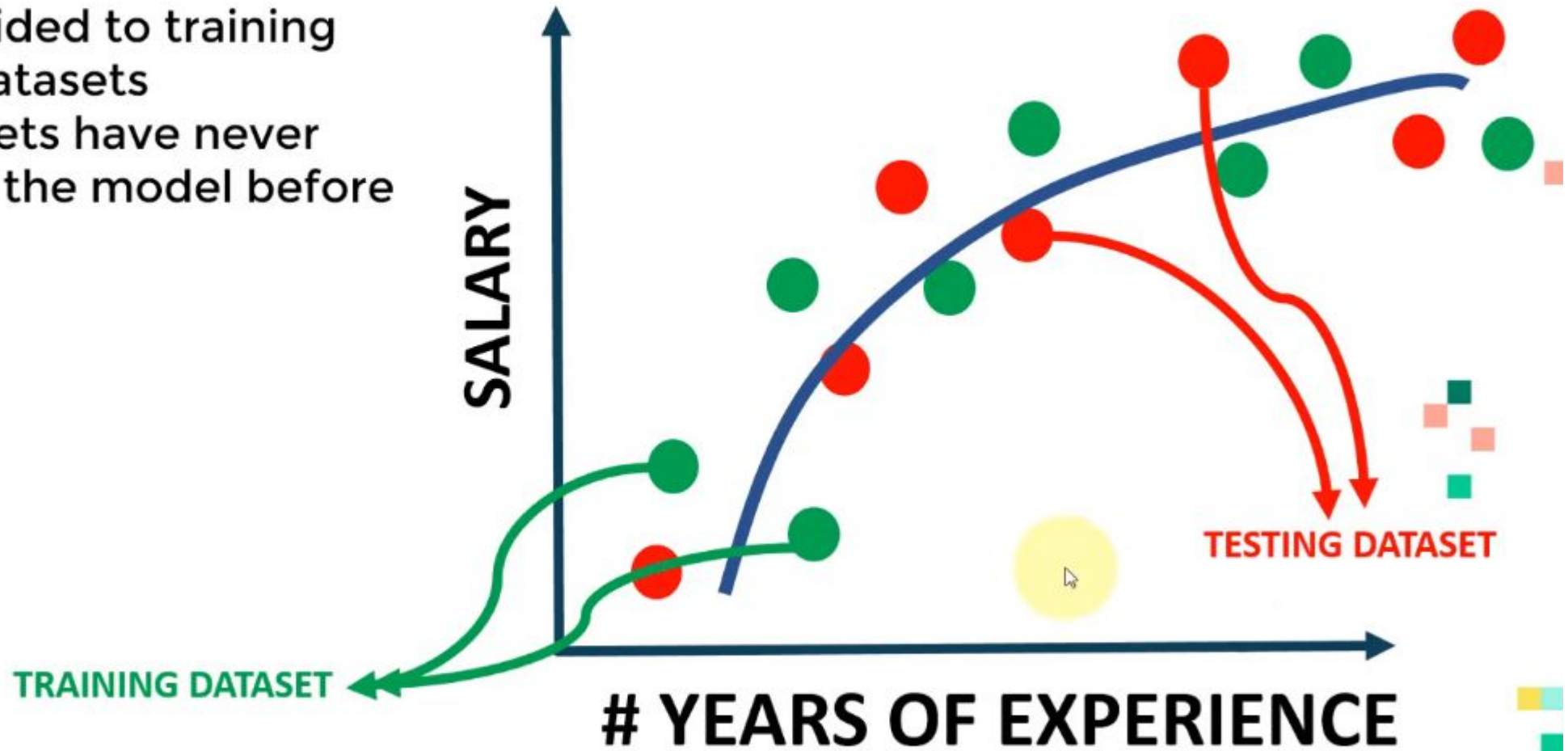
# BIAS AND VARIANCE: TRAINING VS. TESTING DATASETS

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



# BIAS AND VARIANCE: TRAINING VS. TESTING DATASETS

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





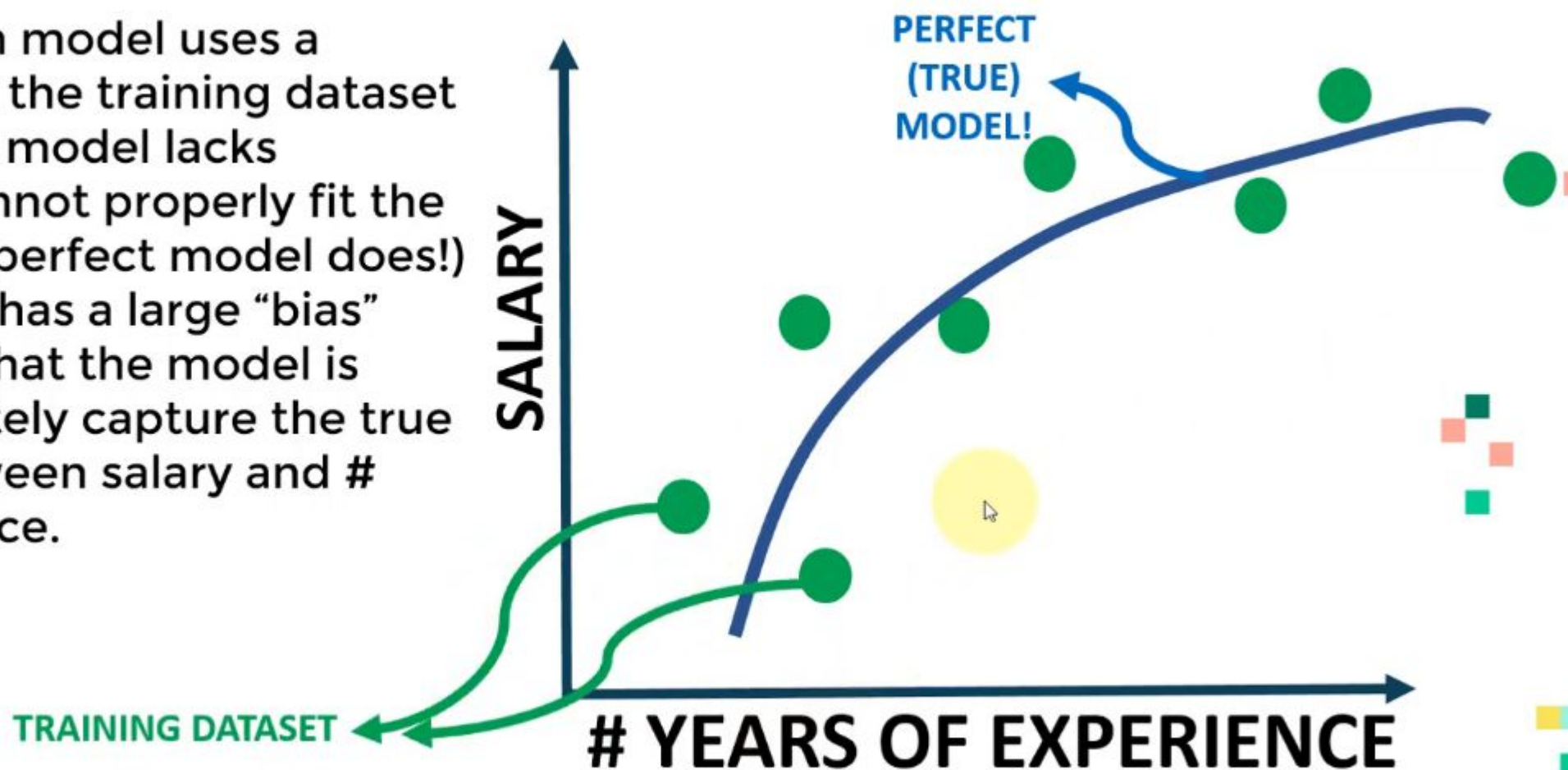
# BIAS AND VARIANCE: MODEL #1- LINEAR REGRESSION (SIMPLE)

- 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.



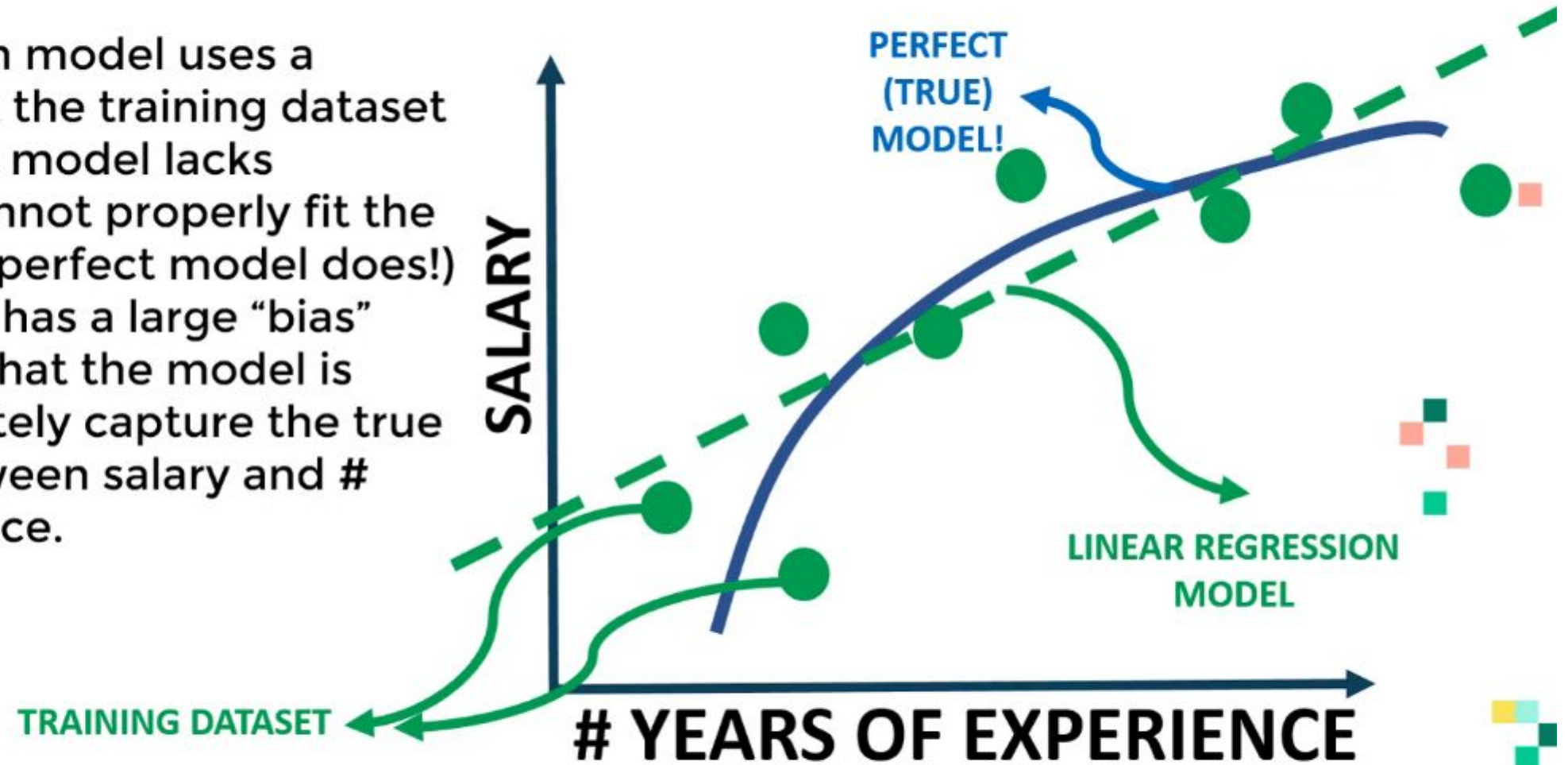
# BIAS AND VARIANCE: MODEL #1- LINEAR REGRESSION (SIMPLE)

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 #1- LINEAR REGRESSION (SIMPLE)

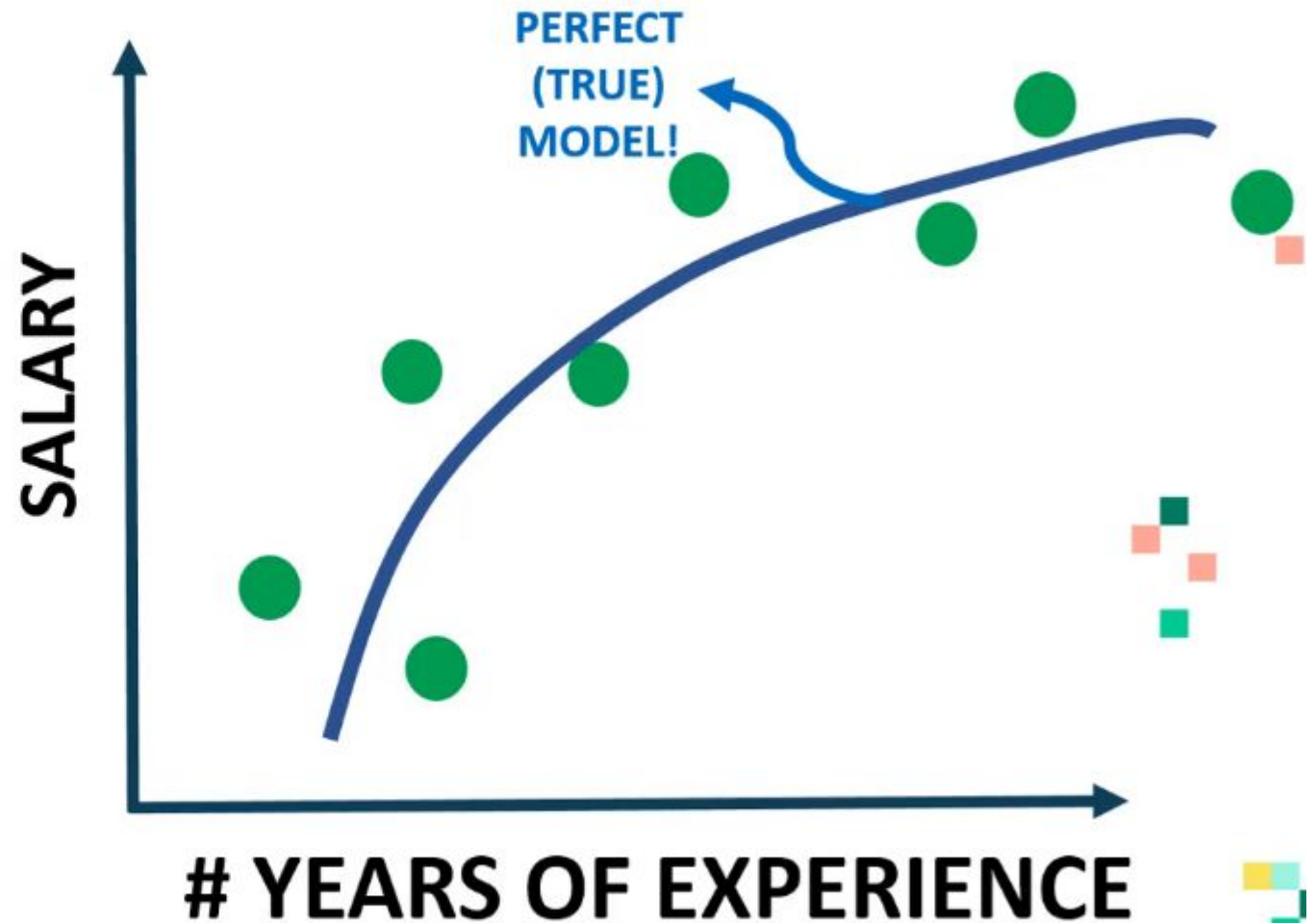
- Linear Regression model uses a straight line to fit the training dataset
- Linear regression model lacks flexibility so it cannot properly fit the data (as the true perfect model does!)
- The linear model has a large “bias” which indicates that the model is unable to accurately capture the true relationship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





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

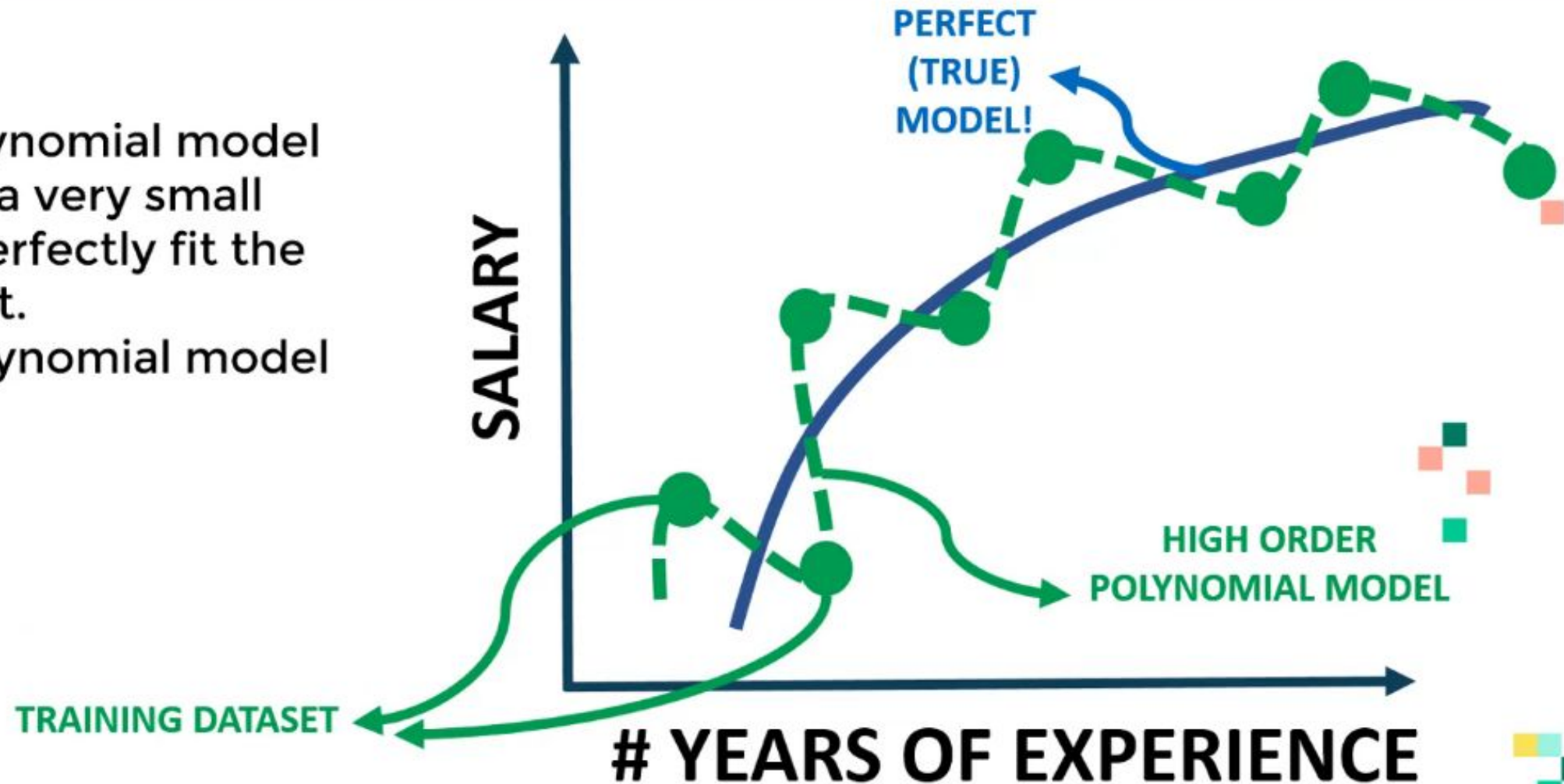
- High order polynomial model is able to have a very small bias and can perfectly fit the training dataset.
- High-order polynomial model is very flexible



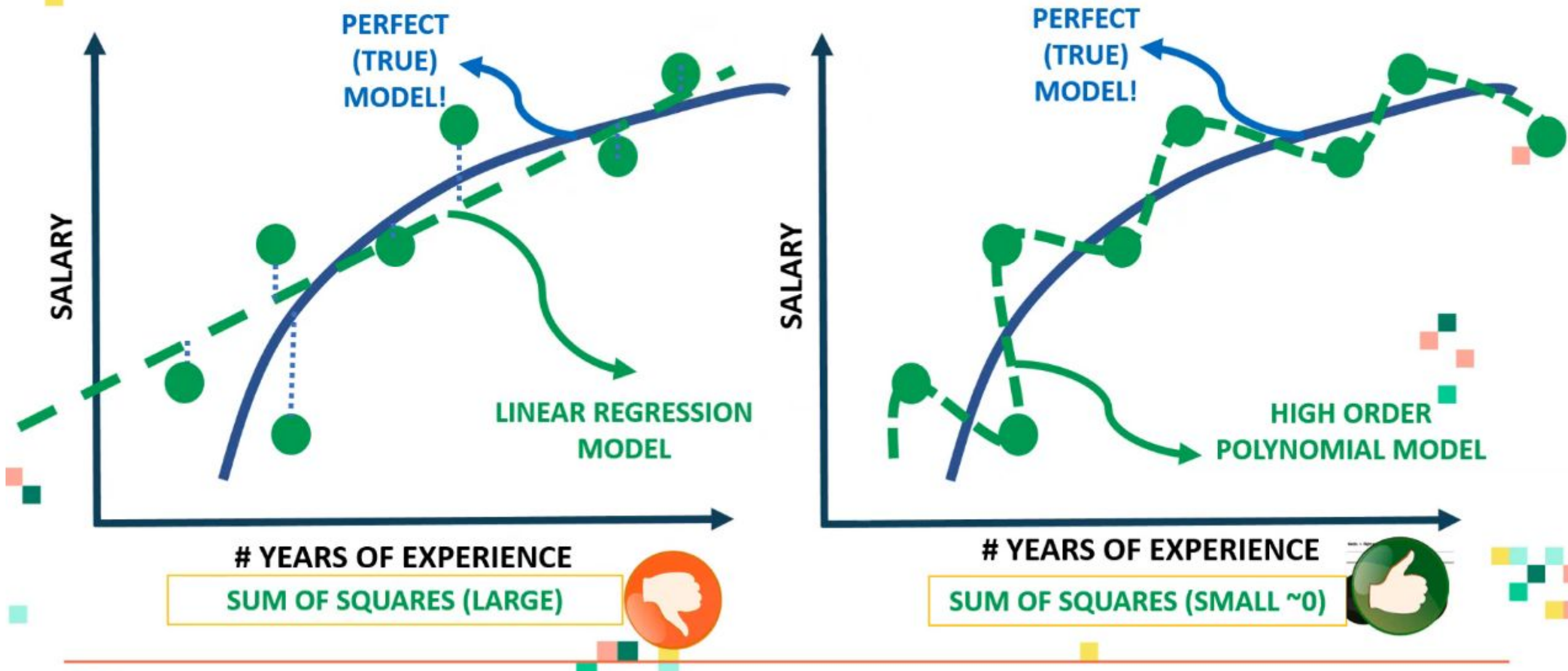


# BIAS AND VARIANCE: MODEL #2 – HIGH ORDER POLYNOMIAL REGRESSION (COMPLEX)

- 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

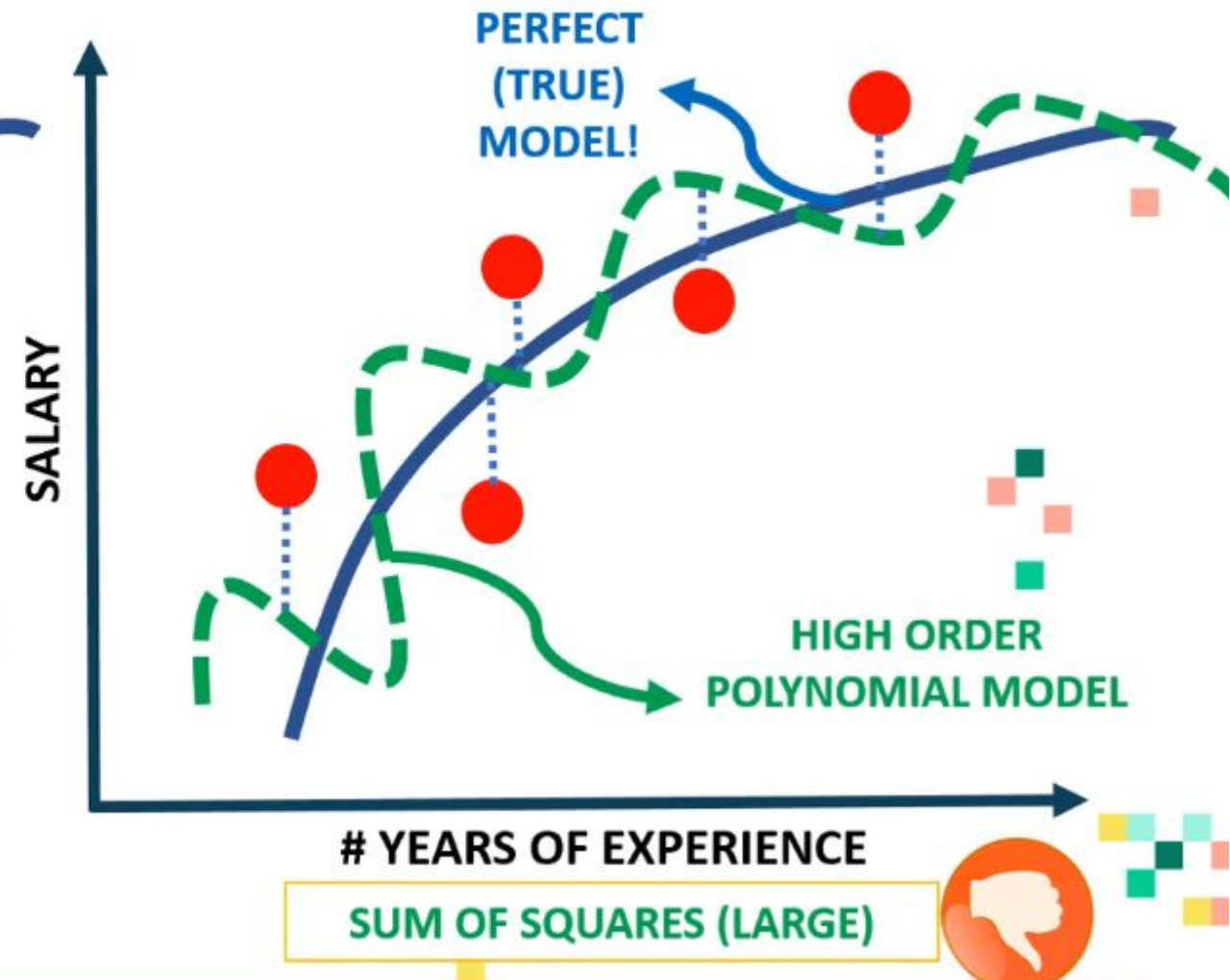
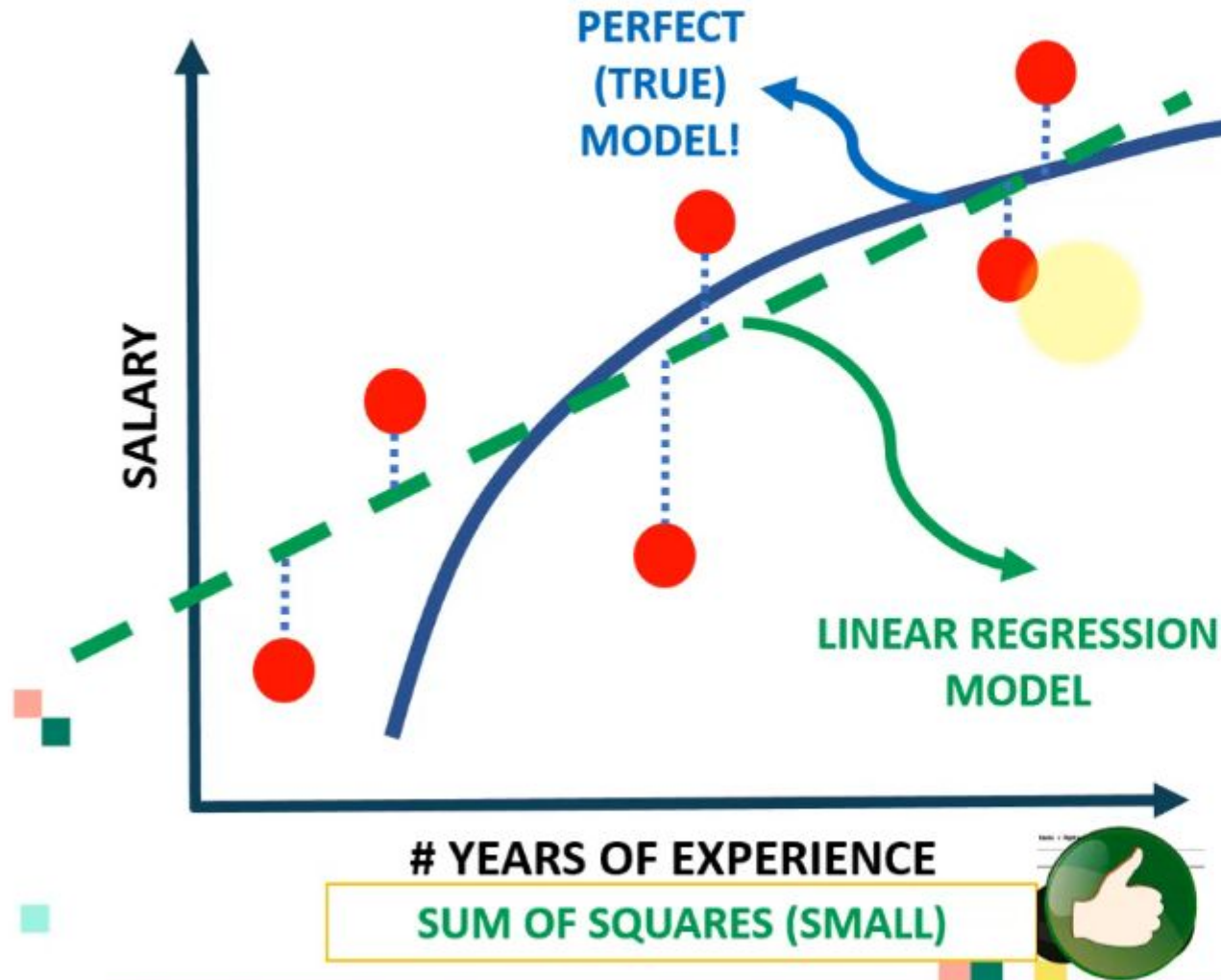


# BIAS AND VARIANCE: MODEL #1 Vs. MODEL #2 DURING TRAINING





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





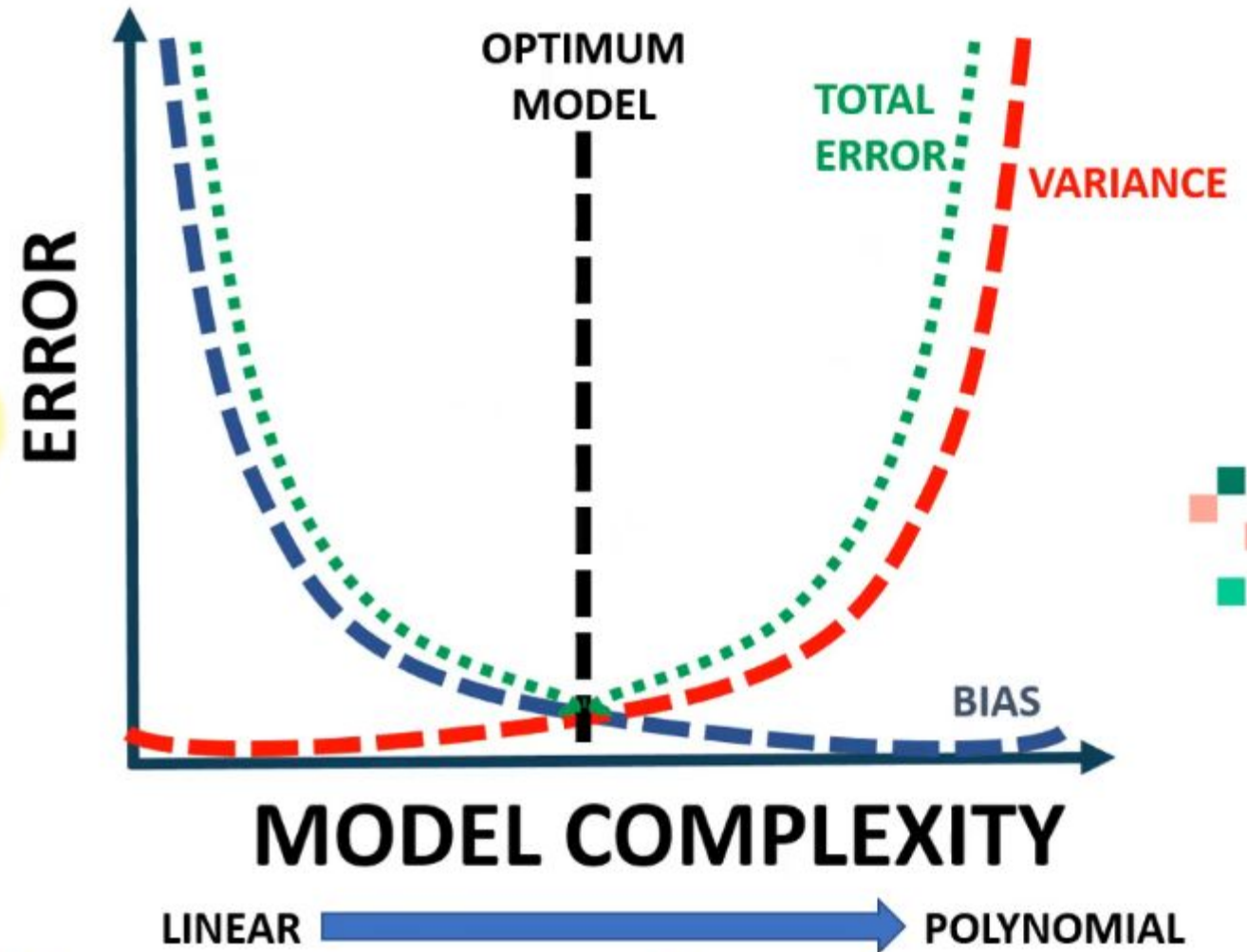


The polynomial model performs poorly on the testing dataset

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 is needed



# MODEL COMPLEXITY VS. ERROR

MODEL #1 (LINEAR REGRESSION) (SIMPLE)	MODEL #2 (HIGH ORDER POLYNOMIAL) (COMPLEX)
Model has <b>High bias</b> because it is very rigid (not flexible) and cannot fit the training dataset well	Model has <b>small bias</b> because it is flexible and can fit the training dataset very well.
Has <b>small variance (variability)</b> 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 <b>large variance (variability)</b> 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**



# What is bias?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Model with high bias pays very little attention to the training data and oversimplifies the model.

It always leads to error on training and test data.

# What is variance?

- Variance is the variability of model prediction for a given data point.
- Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.
- As a result, such models perform very well on training data but has high error rates on test data.

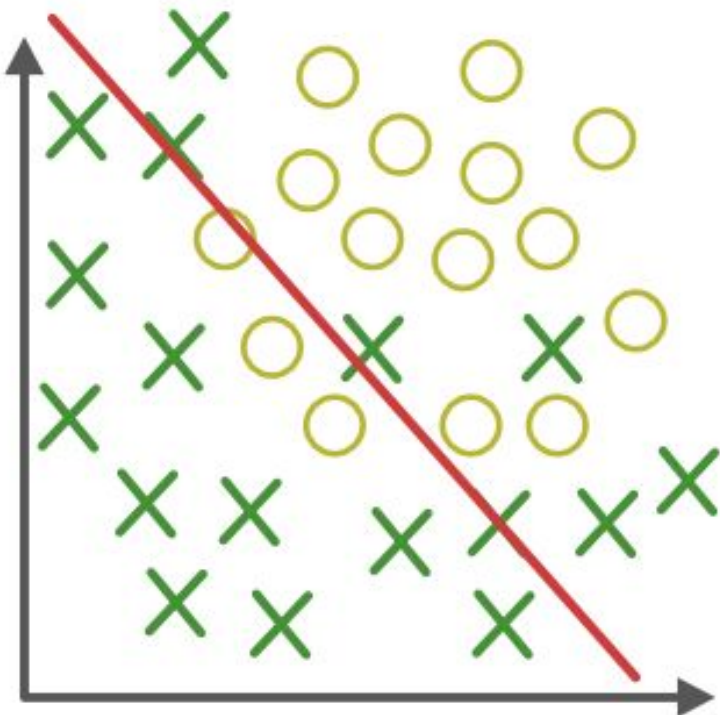
## **Overfitting:**

Good performance on the training data, poor generalization to other(test) data.

## **Underfitting:**

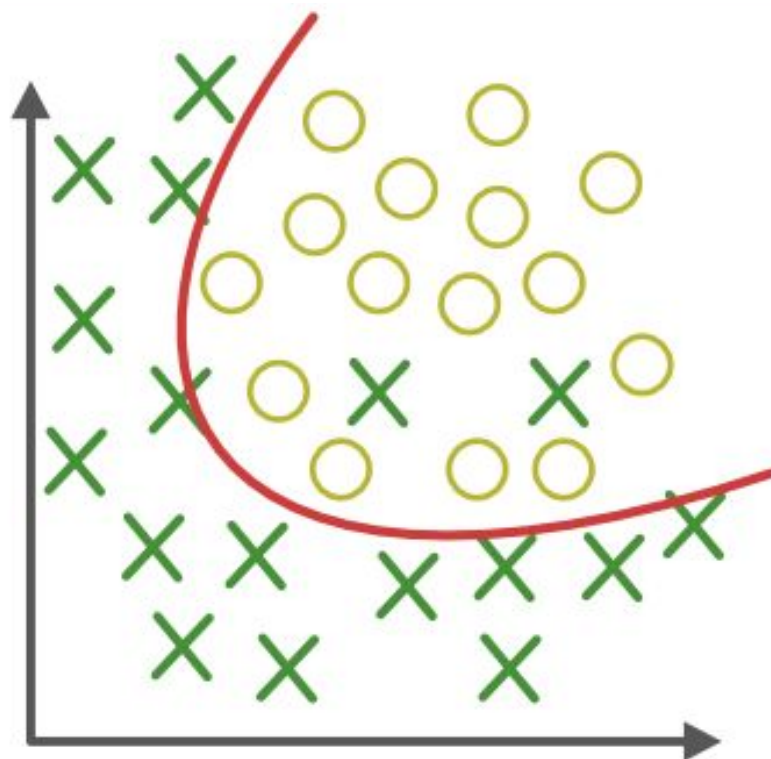
Poor performance on the training data and poor generalization to other(test) data.



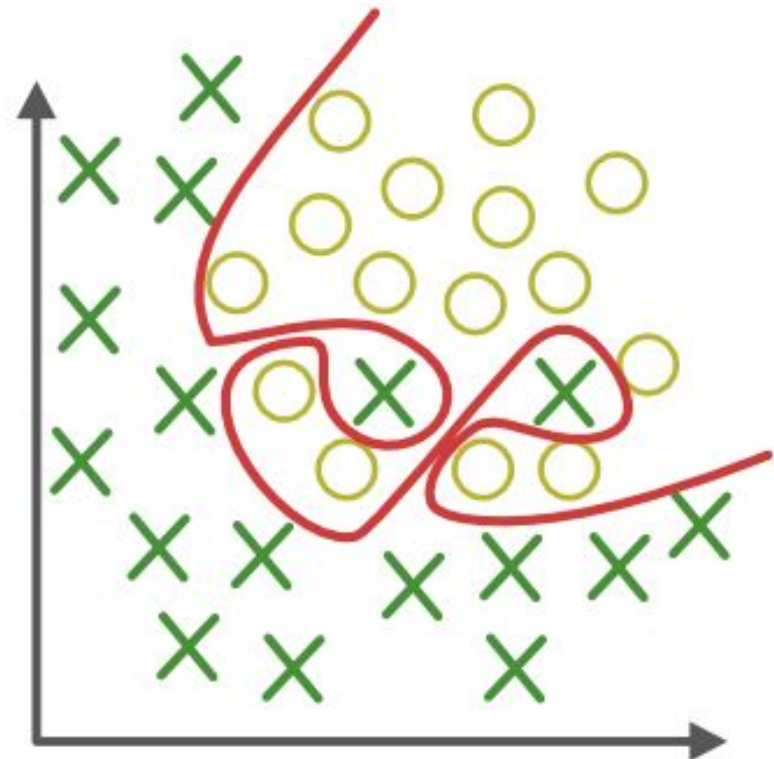


### Under-fitting


(too simple to  
explain the variance)

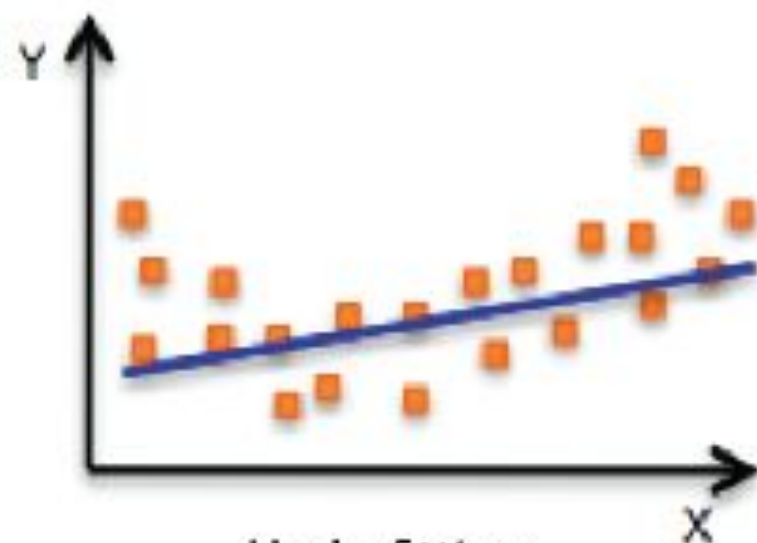


### Appropriate-fitting

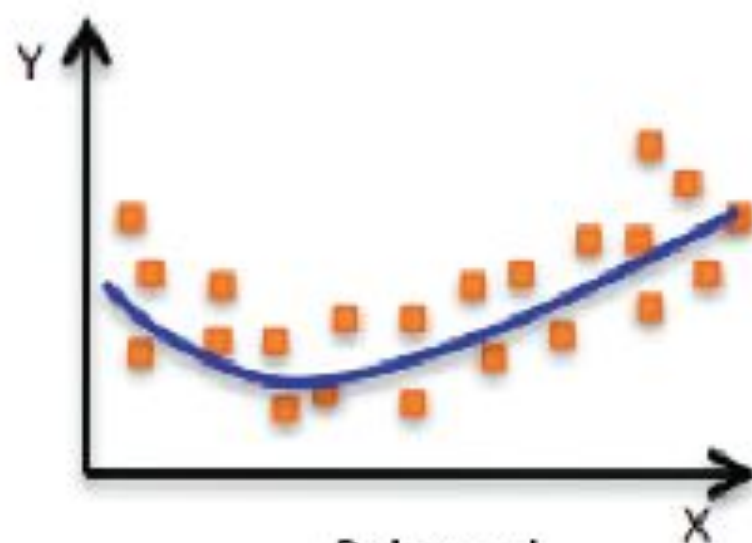


### Over-fitting

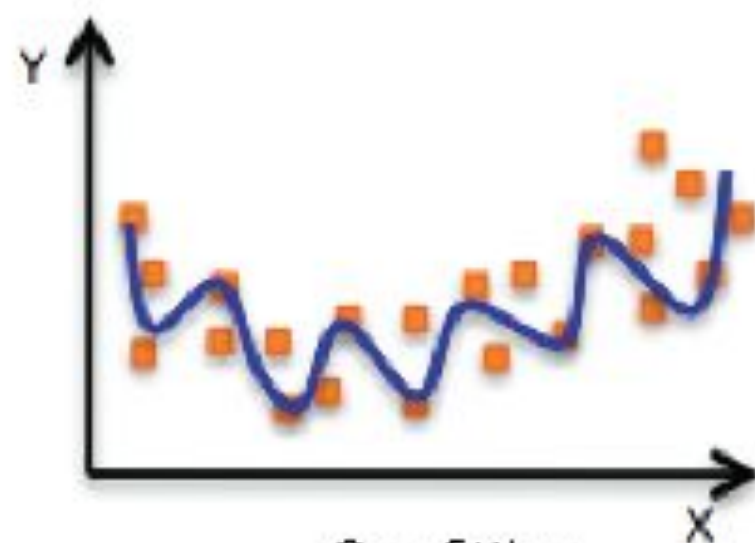
(forcefitting--too  
good to be true) 



Underfitting



Balanced



Overfitting