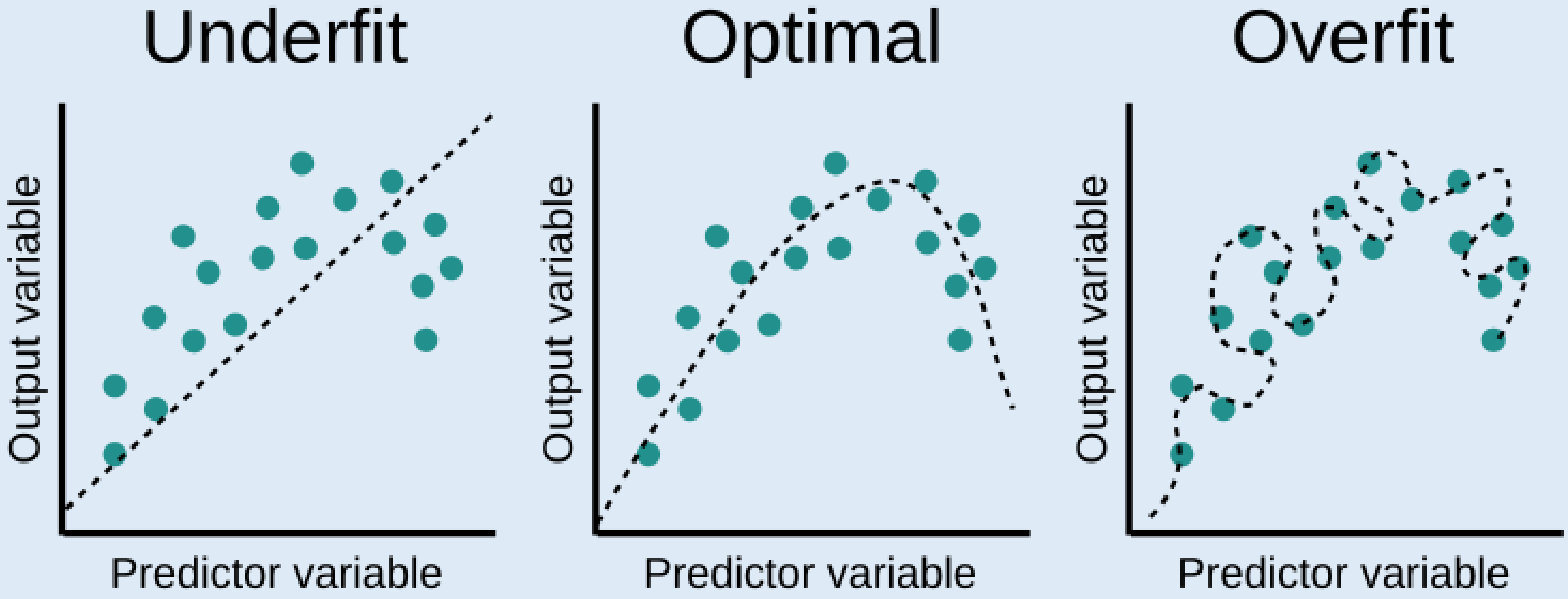
Overfitting and Underfitting



Underfit

High Bias
Low Variance

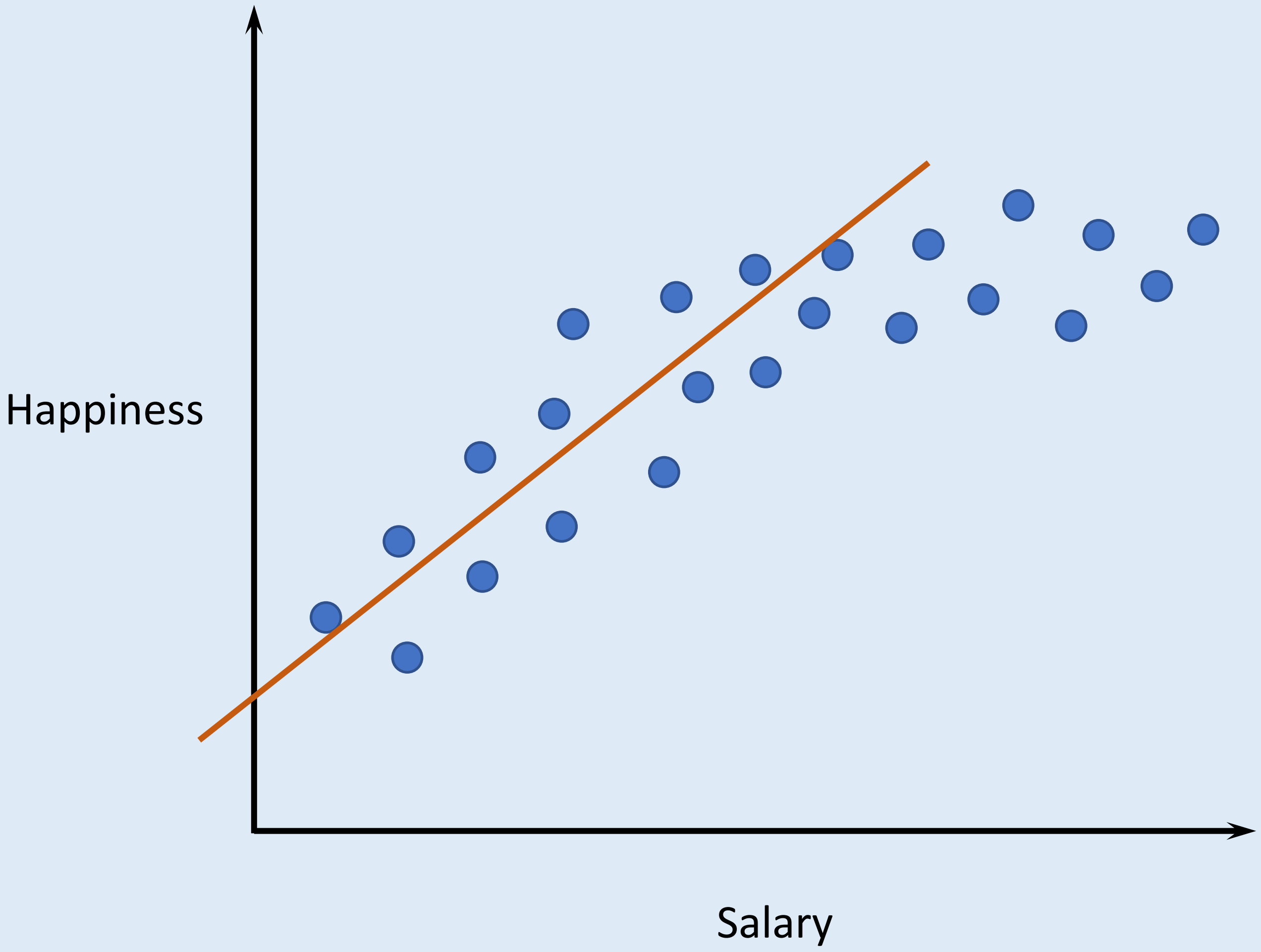
Optimal

Low Bias
Low Variance

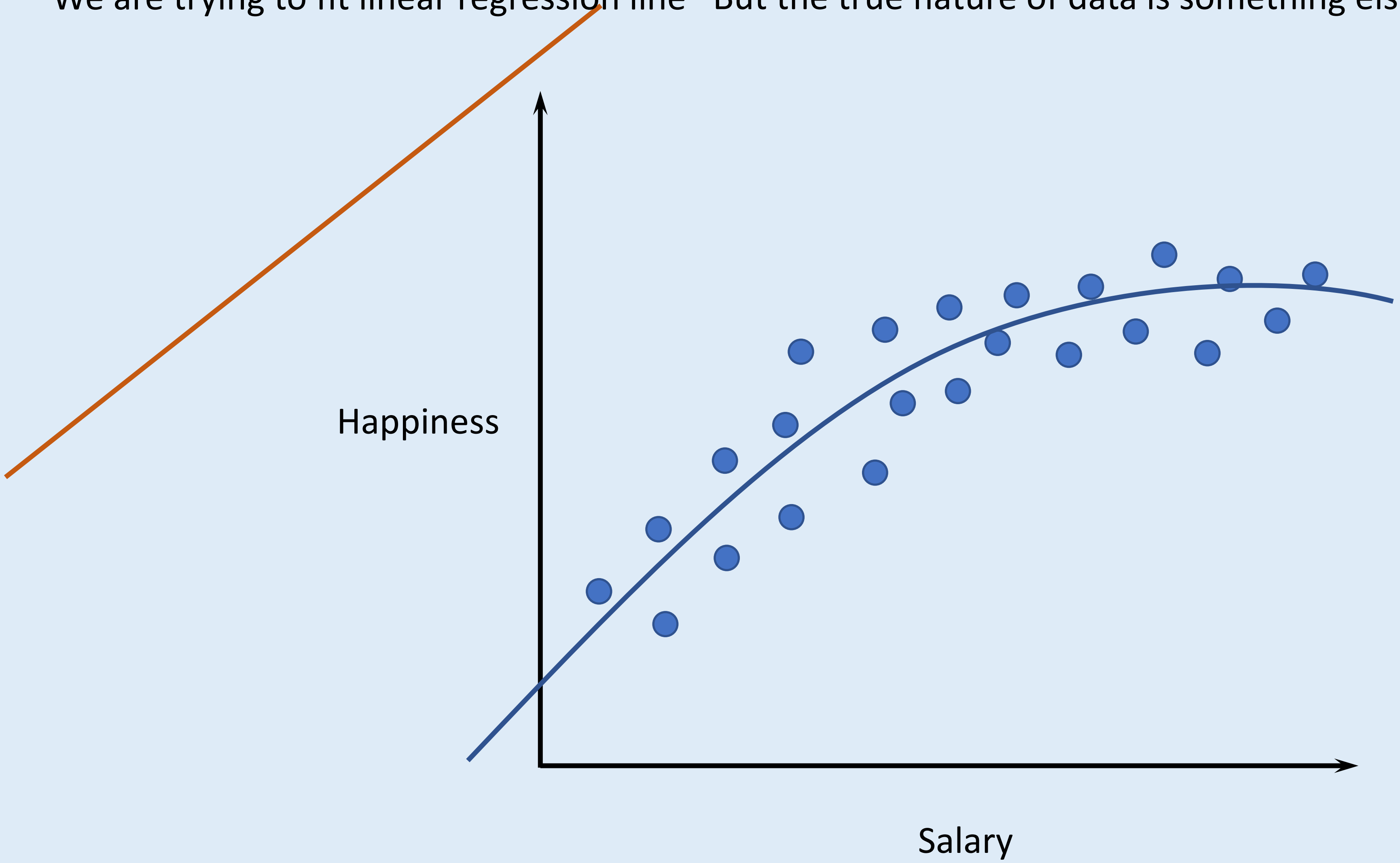
Overfit

Low Bias
High Variance

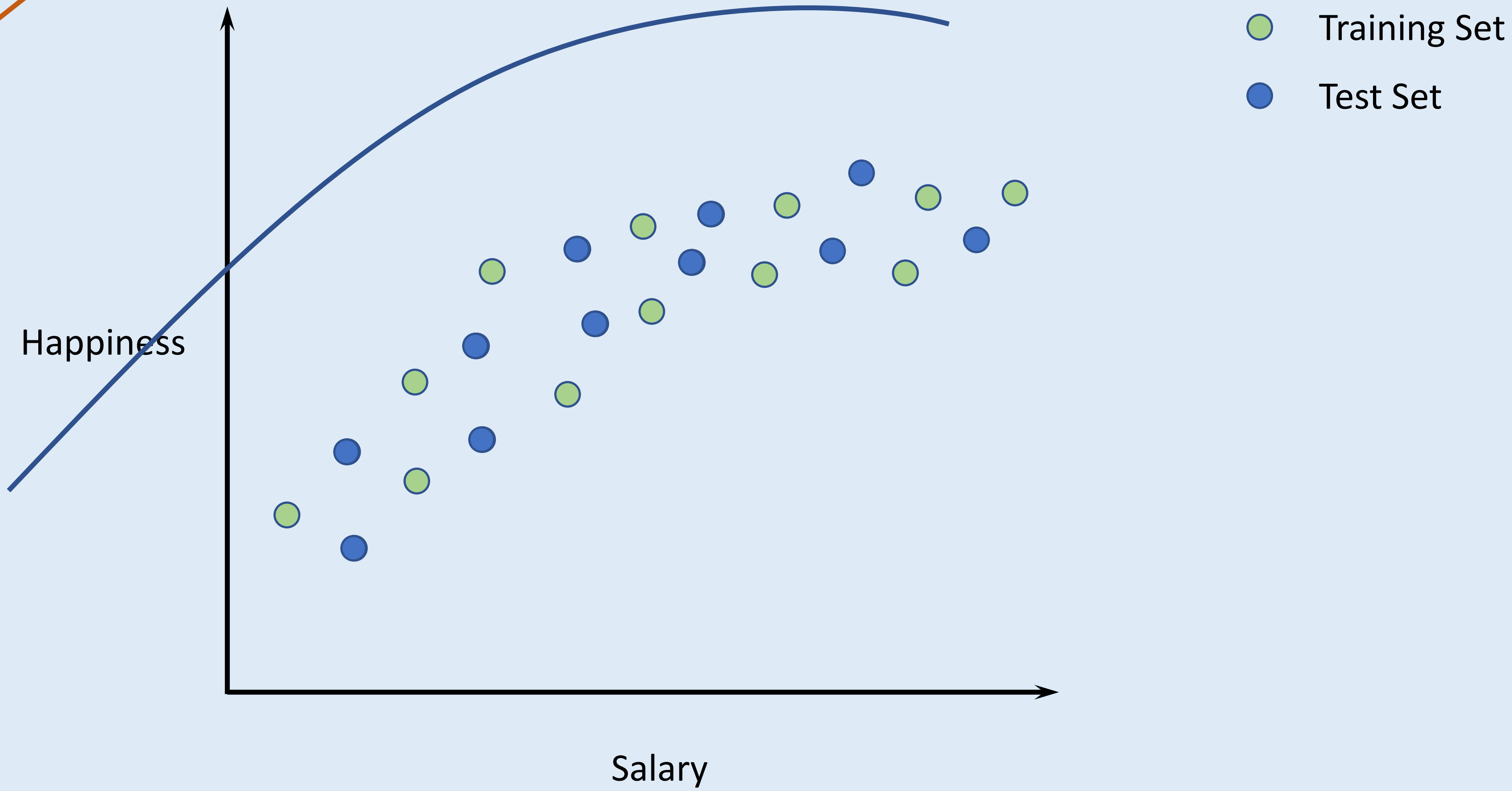
We are trying to fit linear regression line



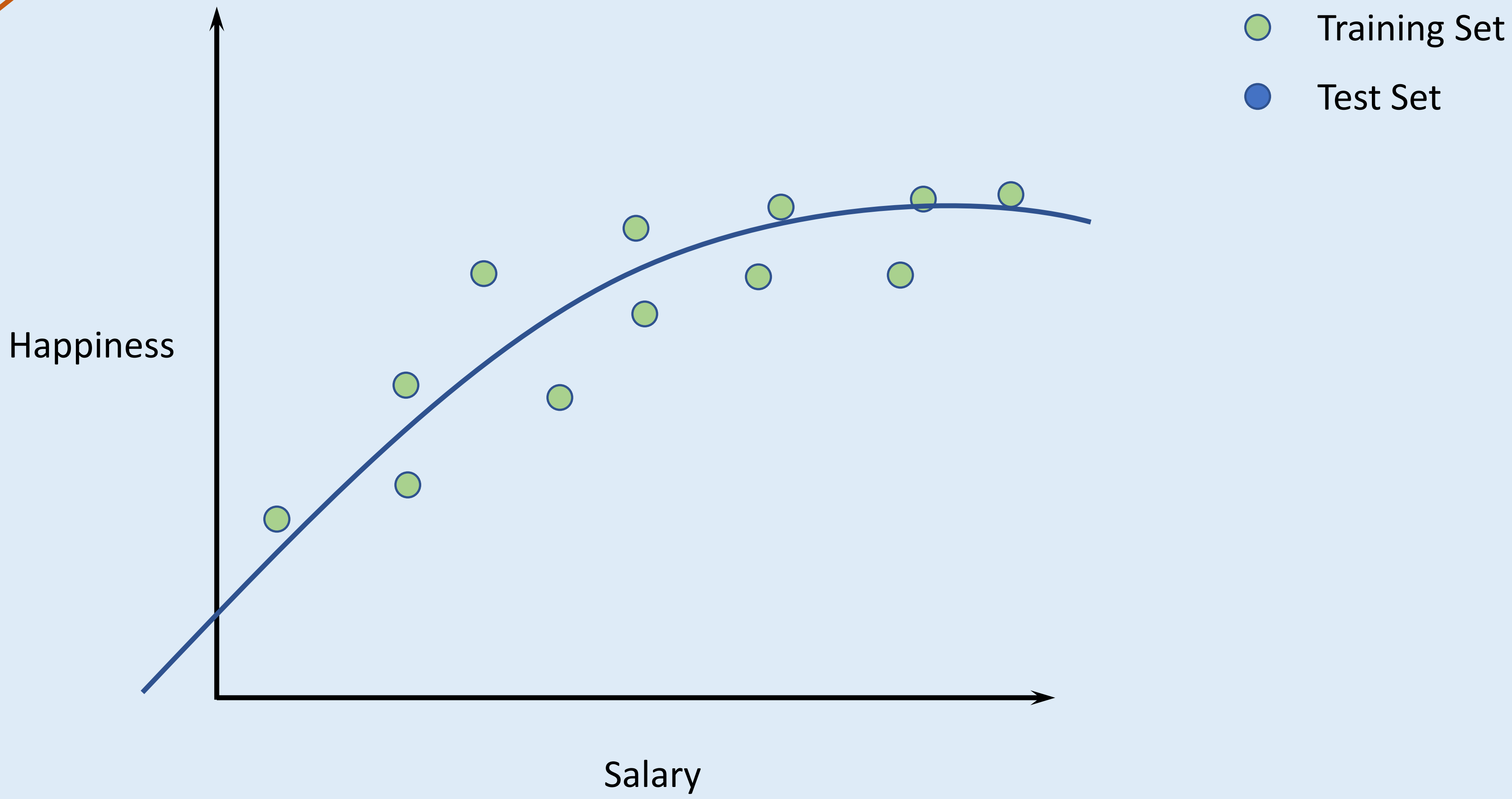
We are trying to fit linear regression line But the true nature of data is something else



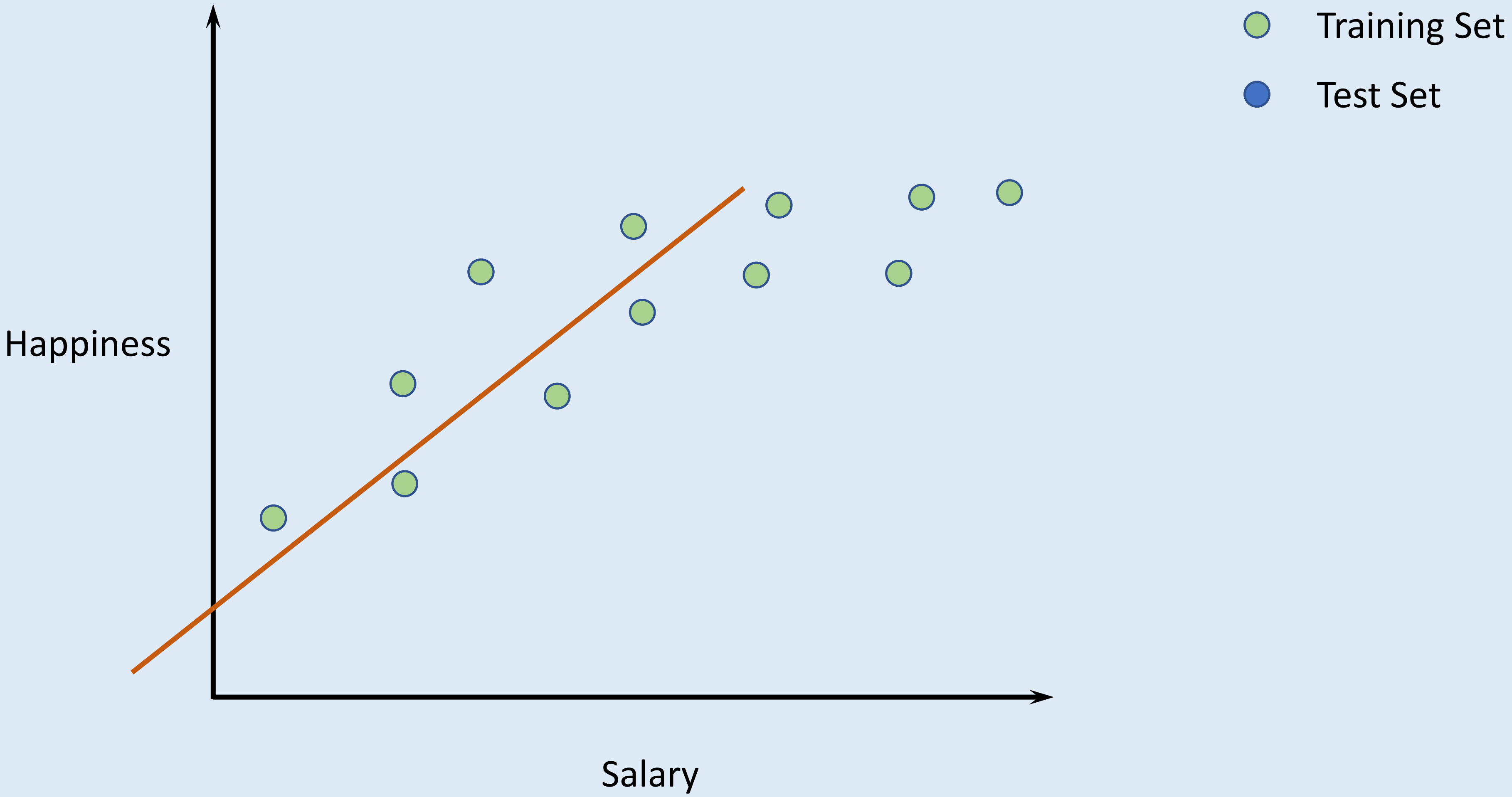
Let’s Split the data in training set and testing set



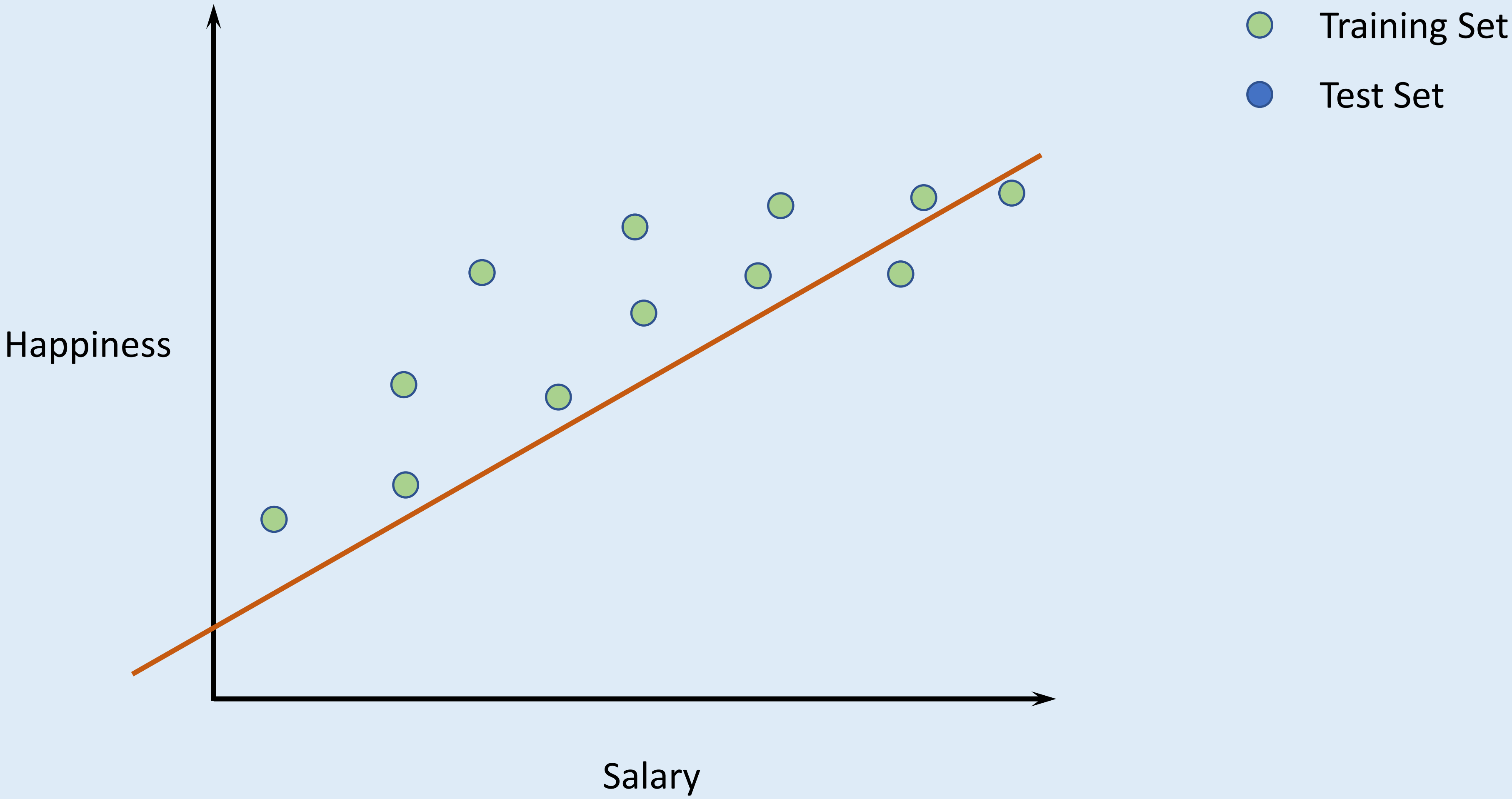
Let’s Split the data in training set and testing set



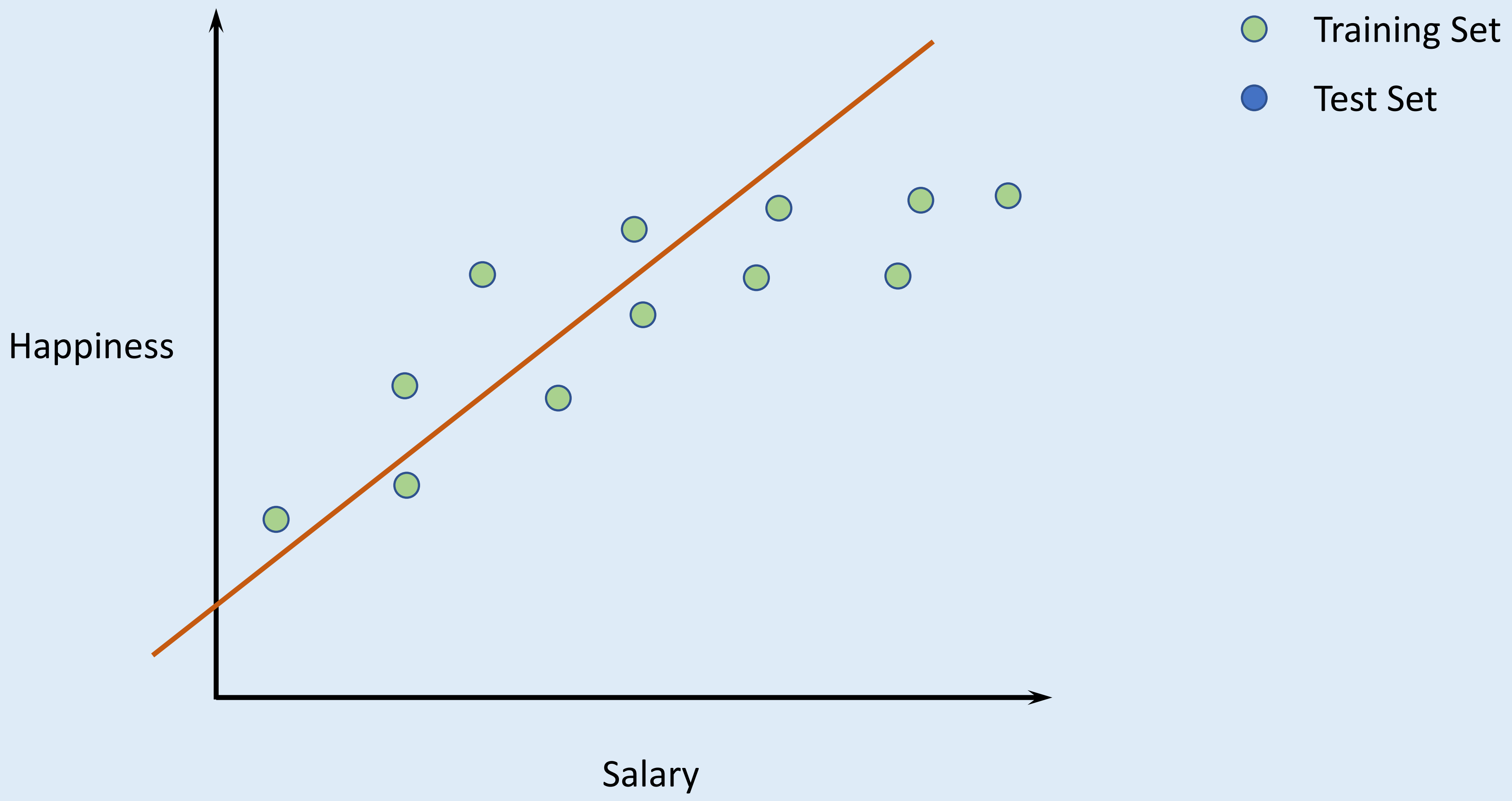
Let’s try to fit the regression line



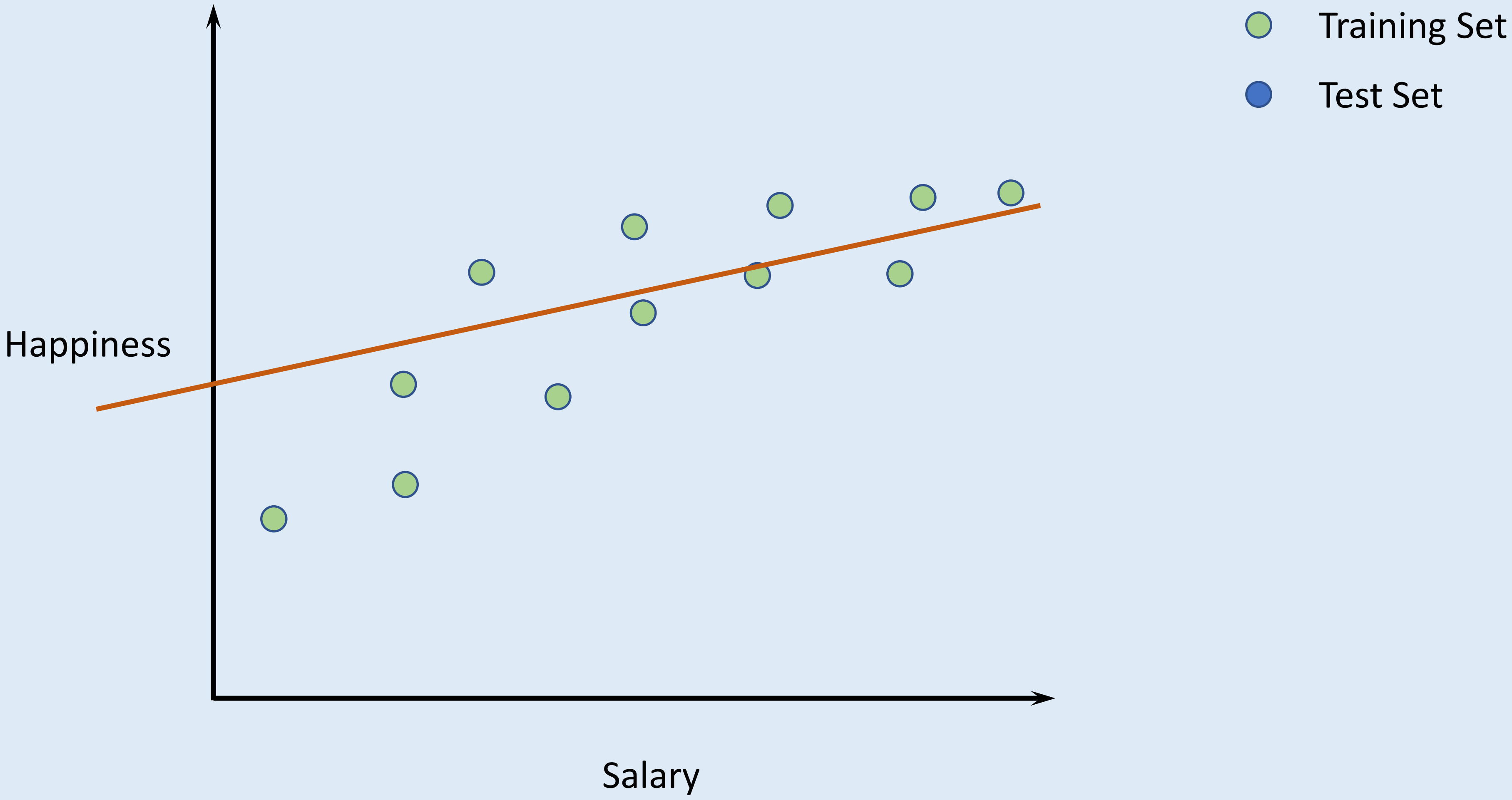
No matter how much we try



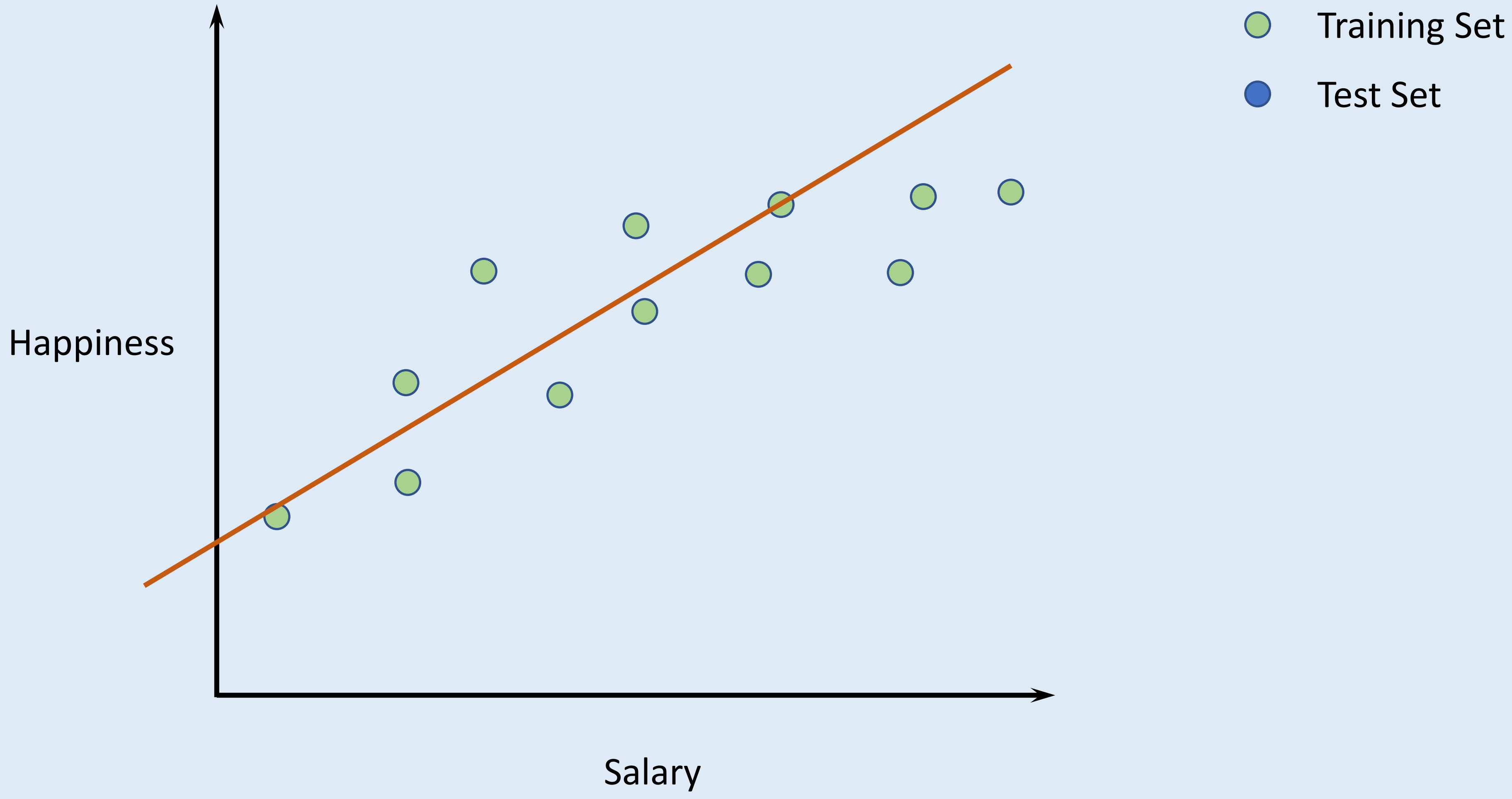
No matter how much we try Regression line fails to capture true relationship



No matter how much we try Regression line fails to capture true relationship

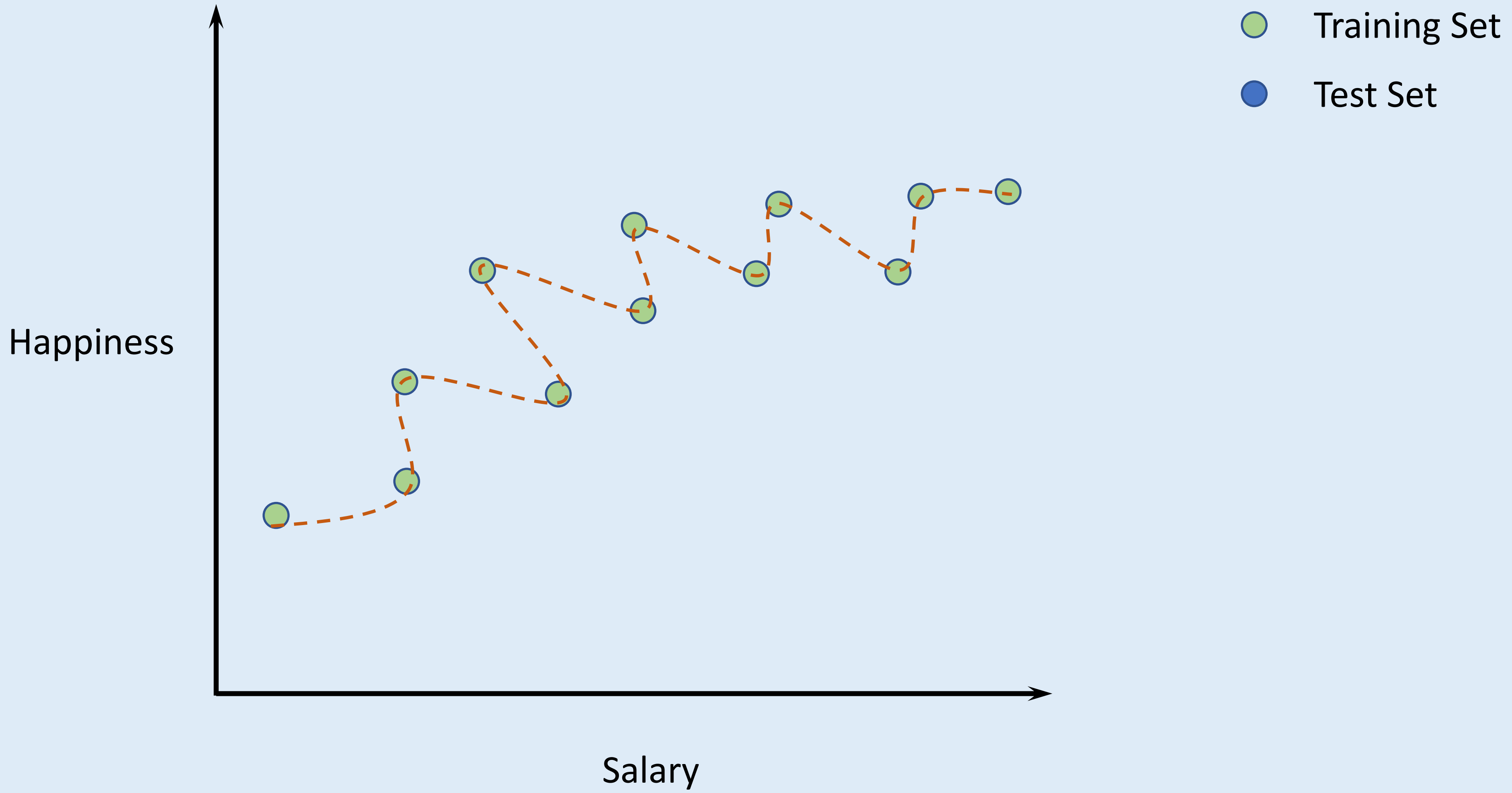


No matter how much we try Regression line fails to capture true relationship



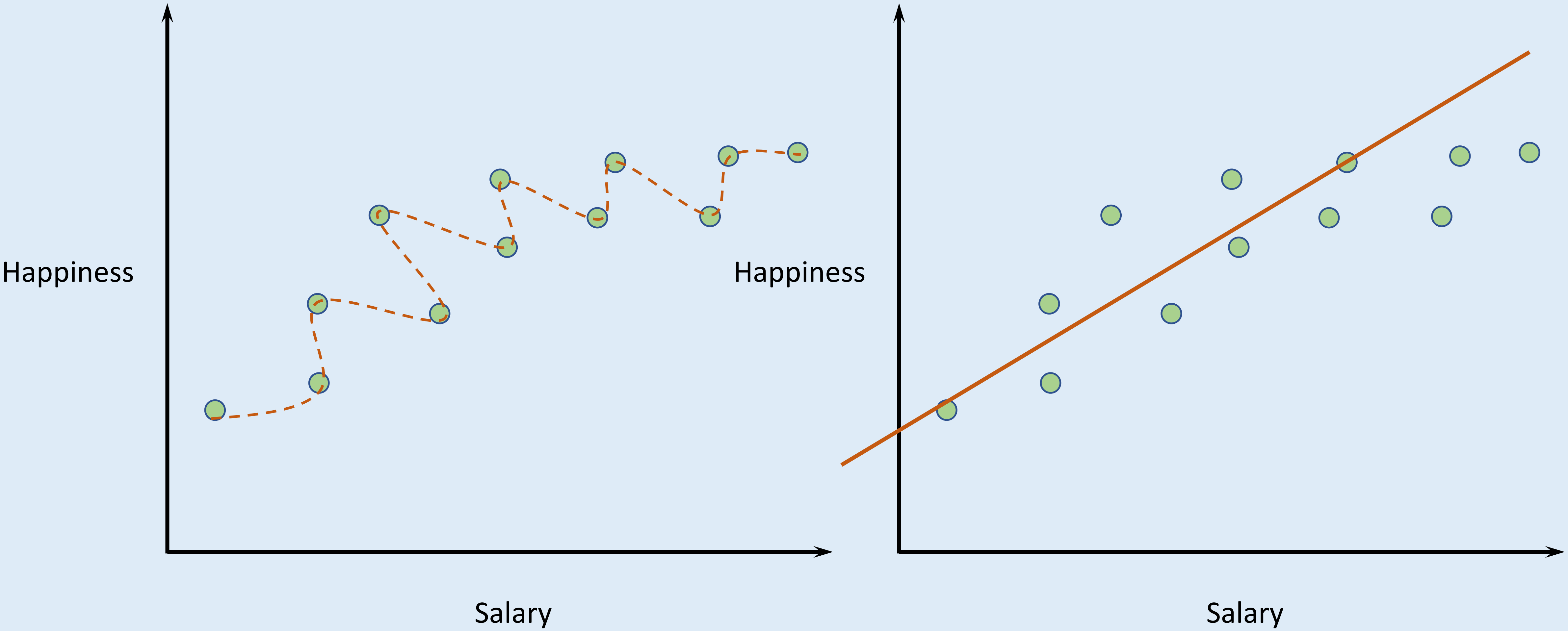
This inability of machine learning algorithm to capture true relationship is what we call bias of machine learning algorithm

Let’s fit line with polynomial regression on test set



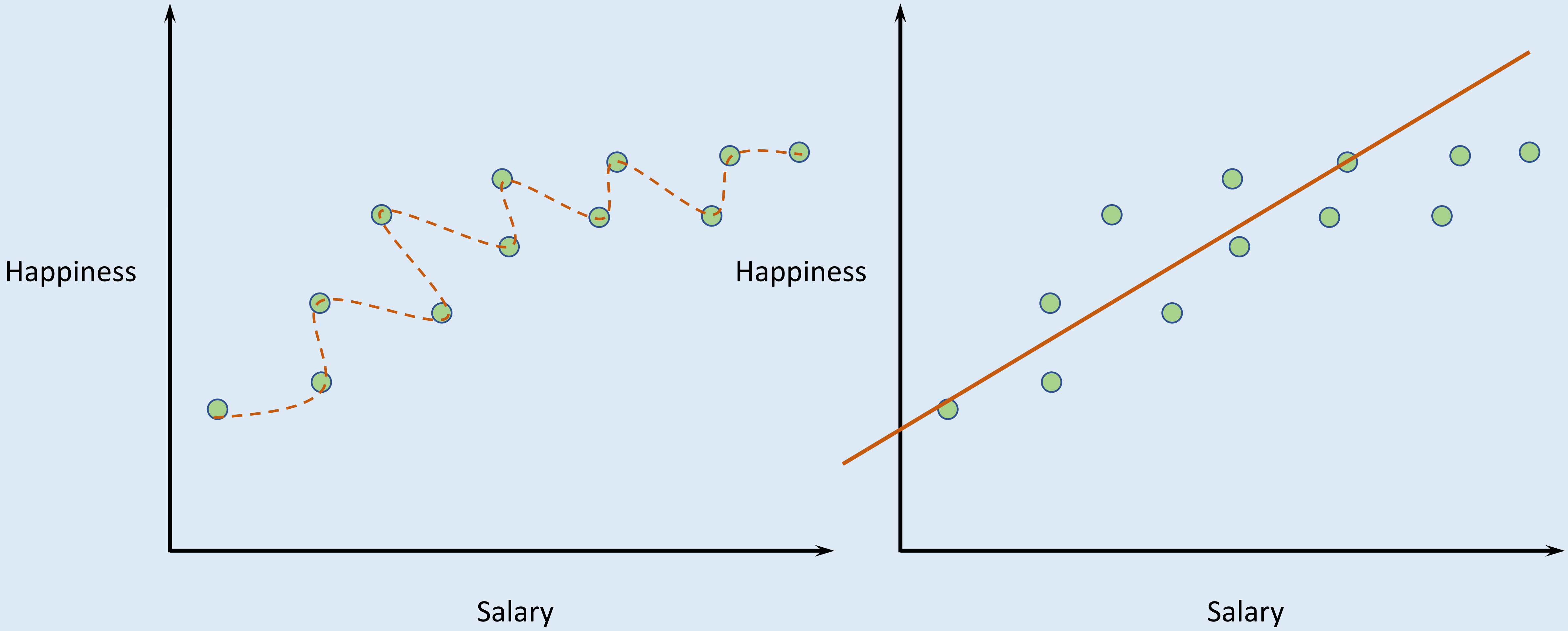
This line fits really well on test data and captures true relationship in test data really well, hence it has low bias

Let’s fit line with polynomial regression on test set



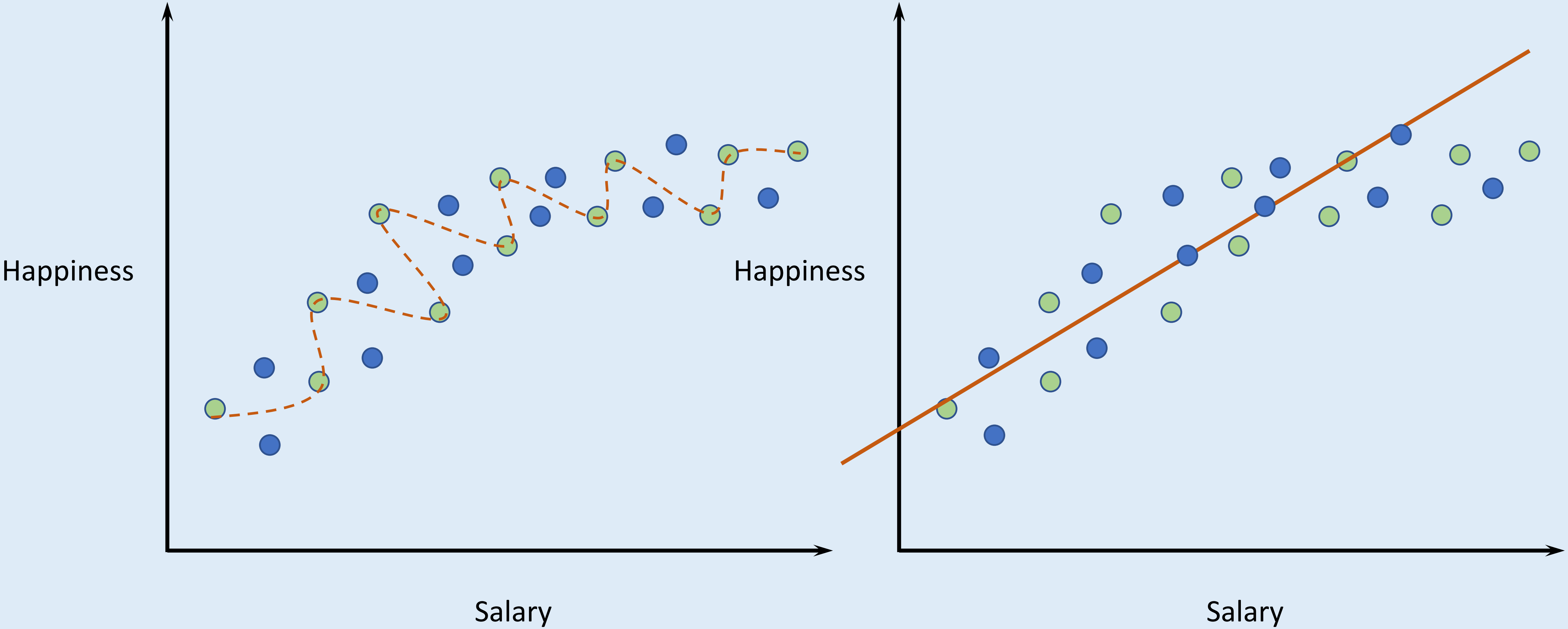
This line fits really well on test data and captures true relationship in test data really well, hence it has low bias, whereas the linear regression has high bias

Let’s fit line with polynomial regression on test set



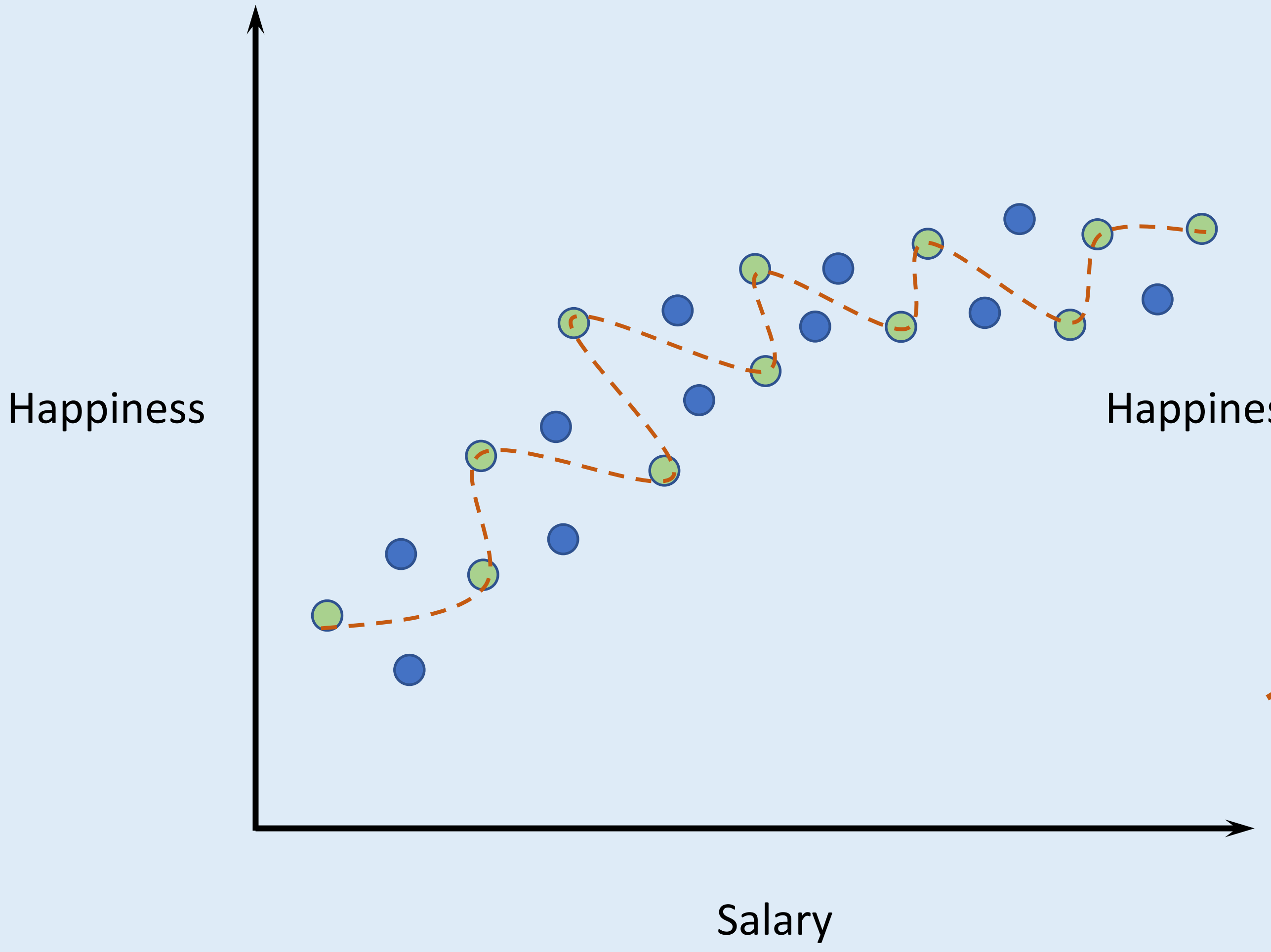
Calculating sum of squared error on polynomial regression will give zero error on train data
Where as the linear regression will have comparatively high error on train data

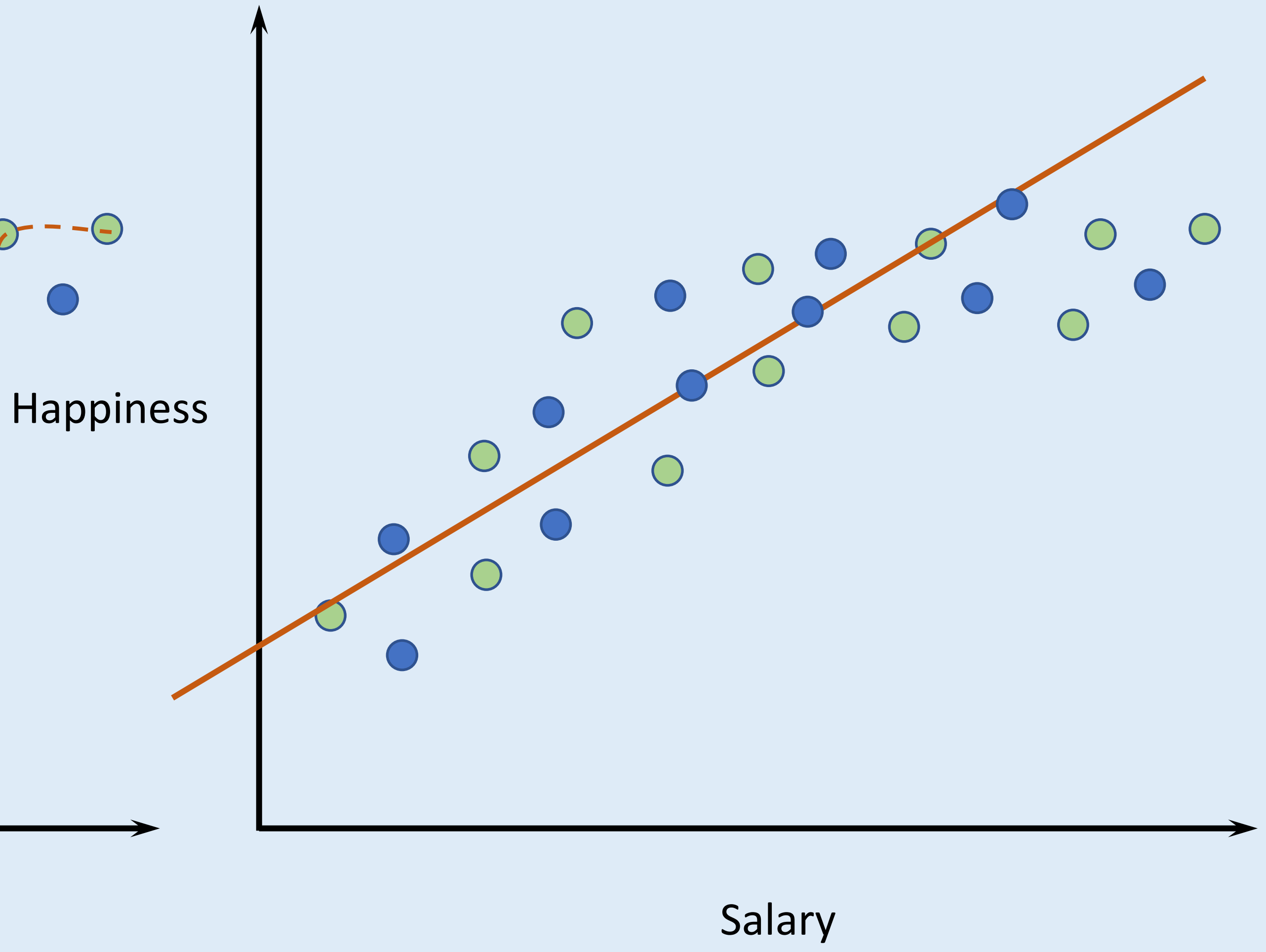
Let’s fit line with polynomial regression on test set



Whereas, the sum of squared error for testing data for polynomial regression is high and for Linear regression it won’t change much

Let’s fit line with polynomial regression on test set

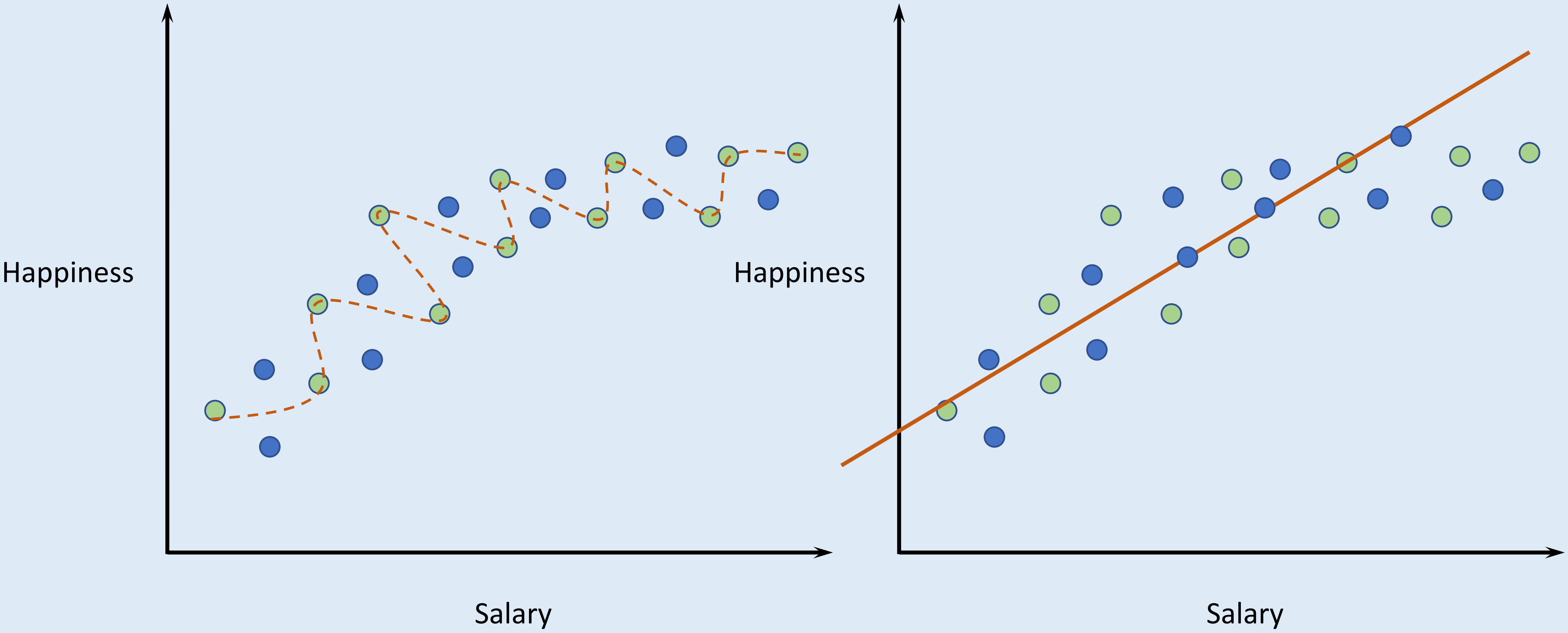




Such difference in change in sum of squared error in testing and training set is called as variance

Here Polynomial Regression has high variance whereas linear regression has low variance

Let’s fit line with polynomial regression on test set

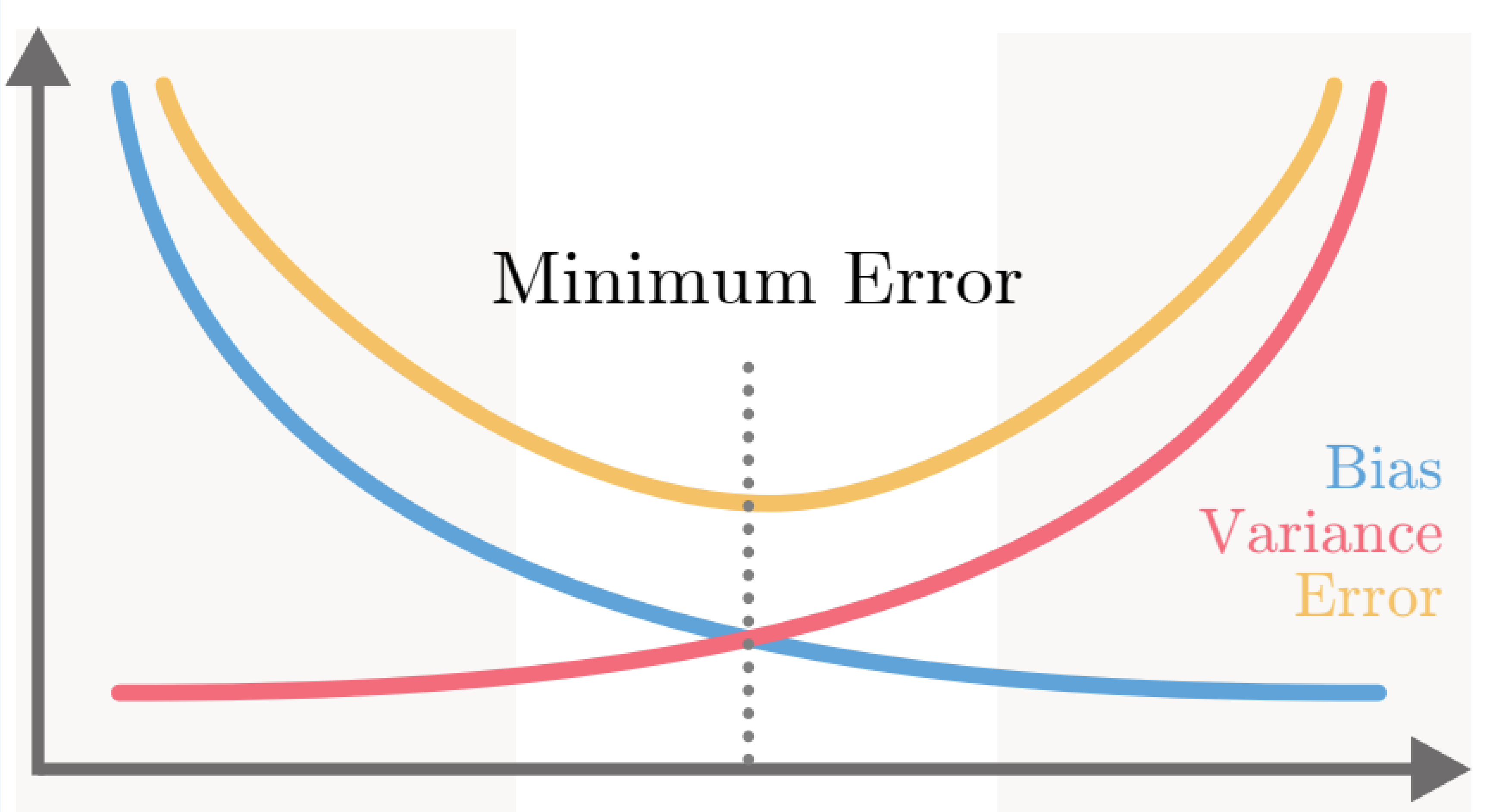


Finally,

The line which fits training data too well and gives high error on test data is called overfit line

Where as the line which neither give good result on training or testing set is call underfit line

Bias Variance Trade-off


Minimum Error
Bias
Variance
Error