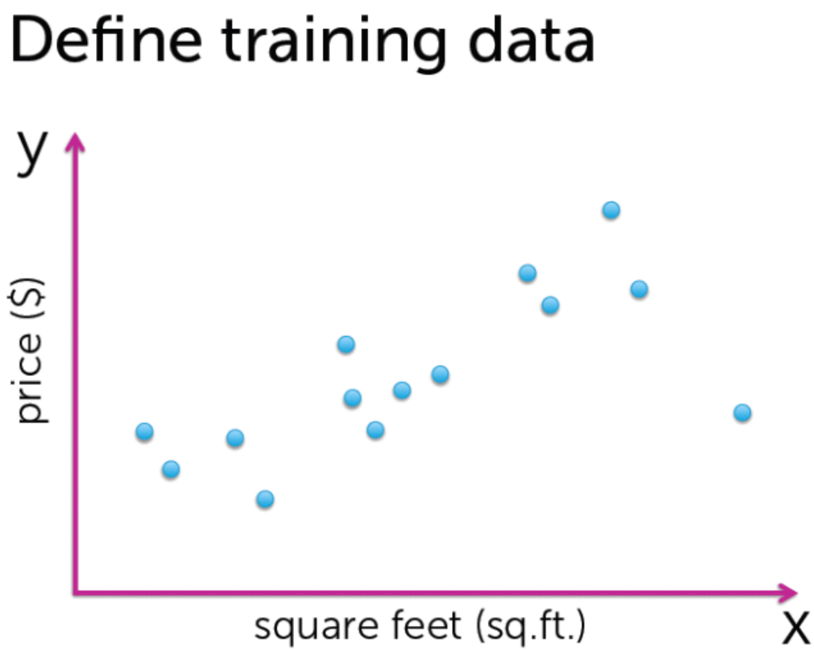
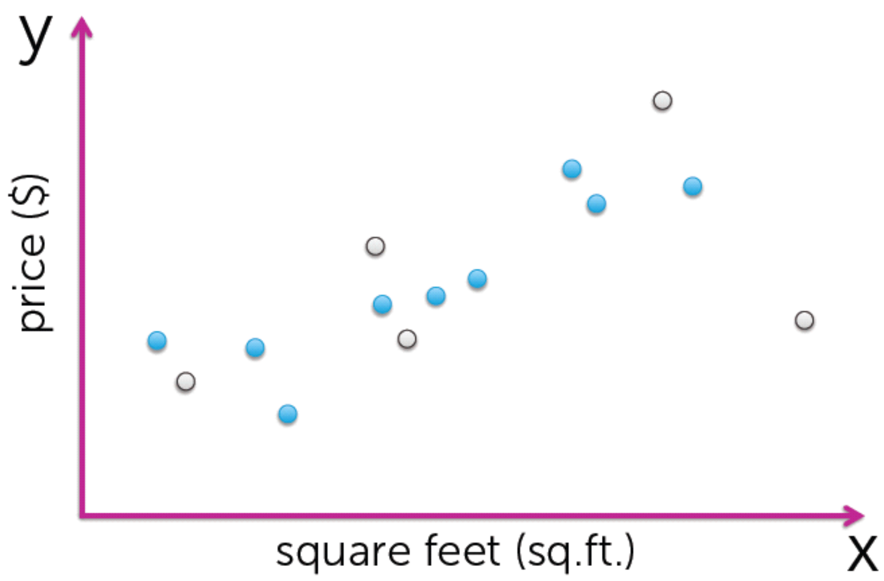
16 January 2018
23:00

We already known that error is nothing but Actual data point - Predicted data point
i.e $y - y = e$

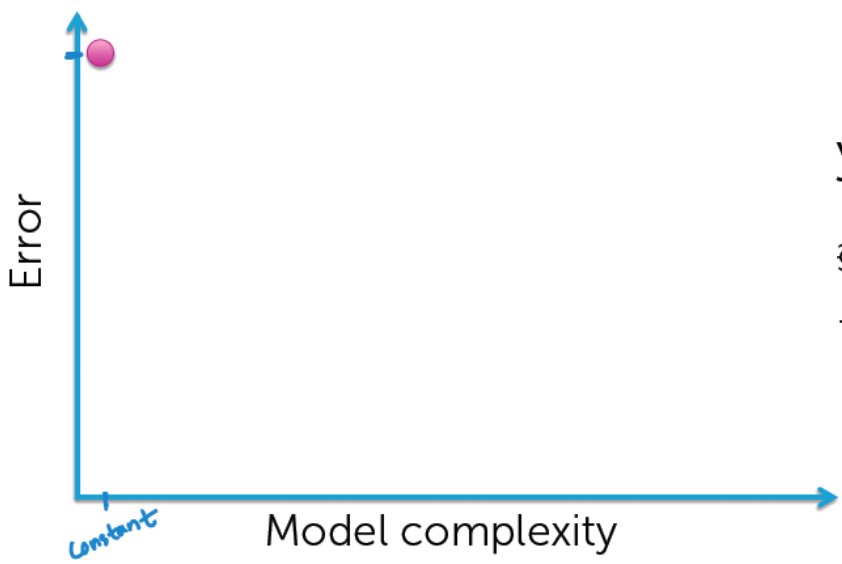
Suppose, if we have 100 observations in the data set and we can find out 100 predicted points from developed model, So we can also calculate 100 errors as well.
The average of all that e's are considered as **Training error** if we work on training data. If we work on test data, we will call it as **Test error**.

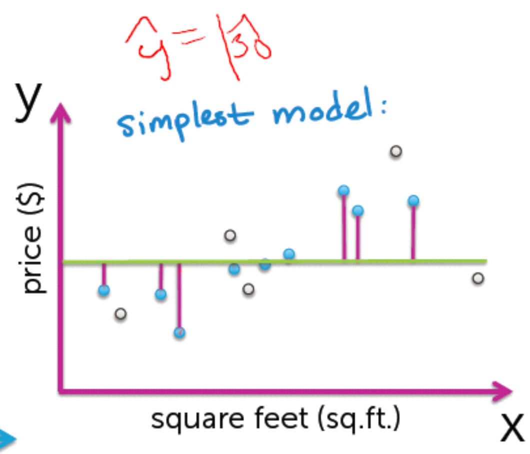
Let us go with an example





Training error vs. model complexity



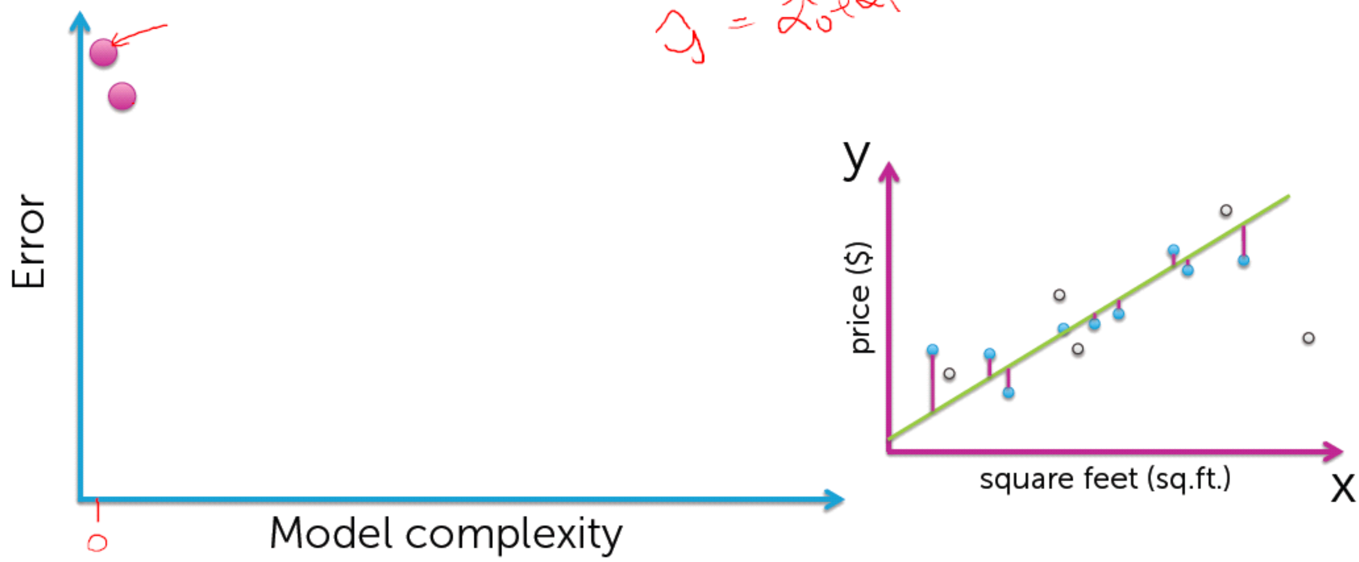


⭐ When I started with a Null Model with zero complexity then the error will be high.

Training error vs. model complexity

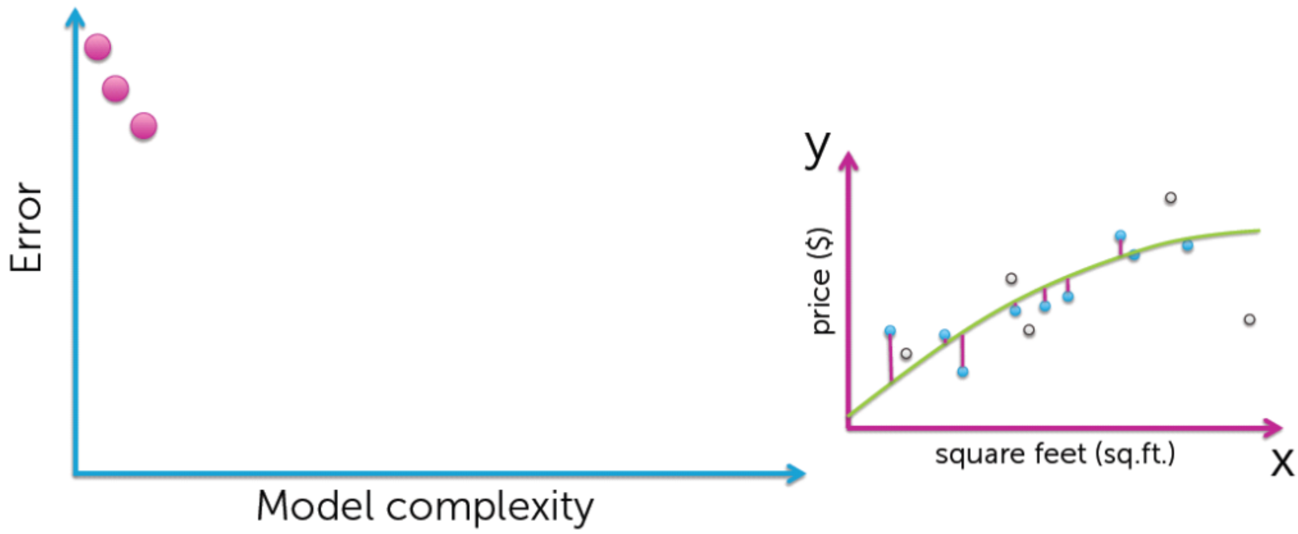
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x + \hat{\alpha}_2 x^2$

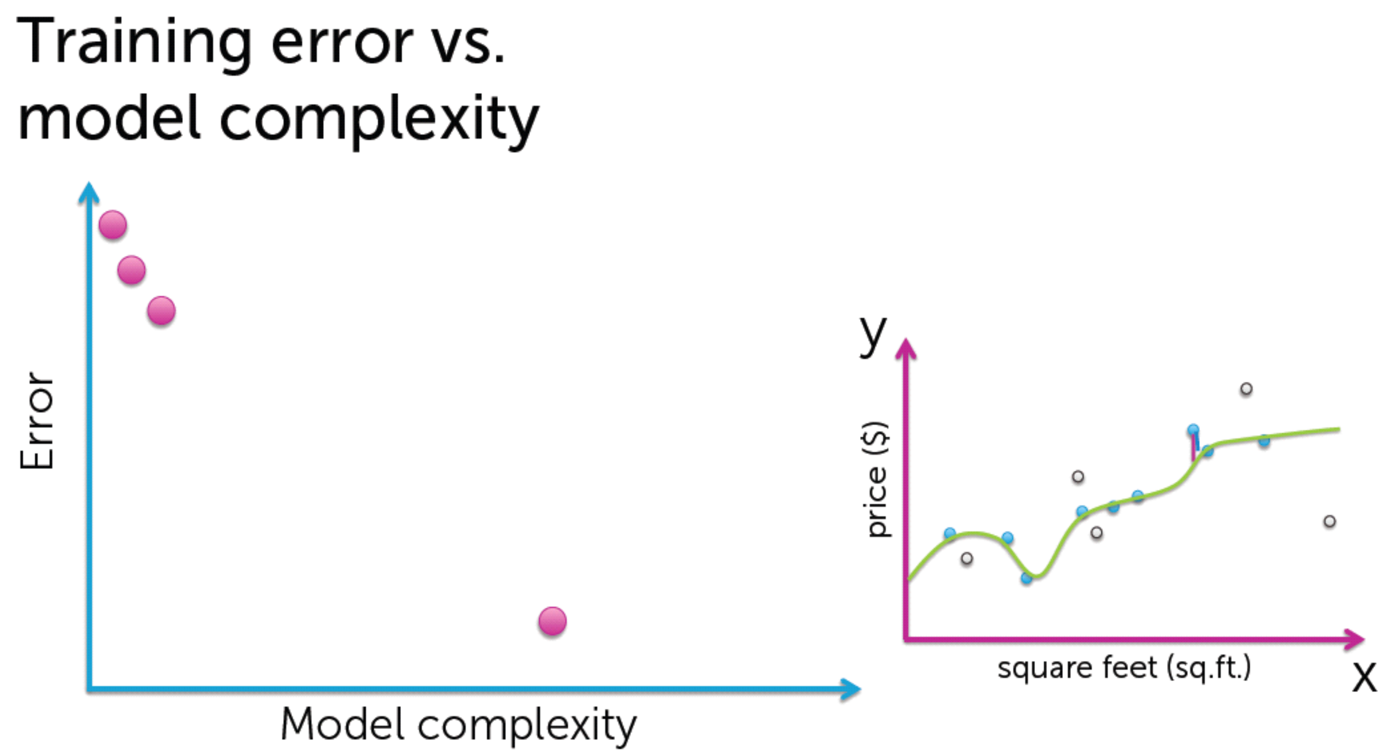


⭐ When I added some variable in my model, the complexity increases and error slowly coming down.

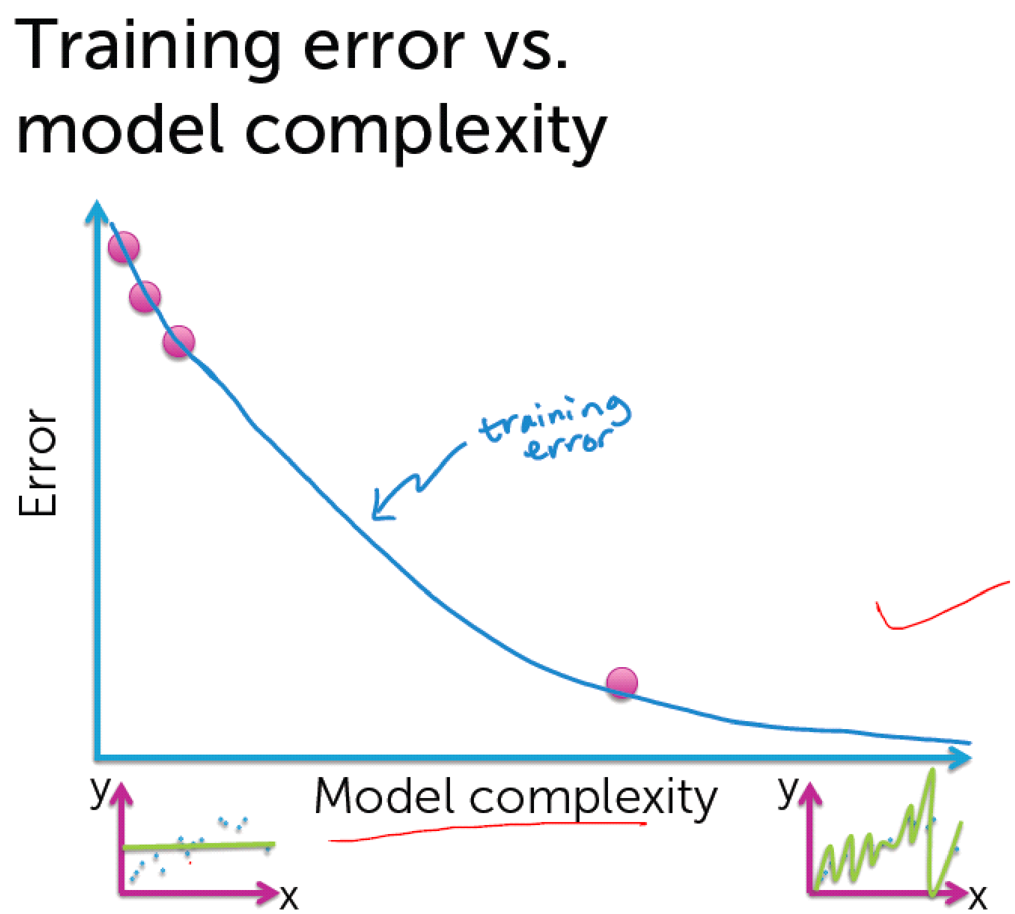
Training error vs. model complexity



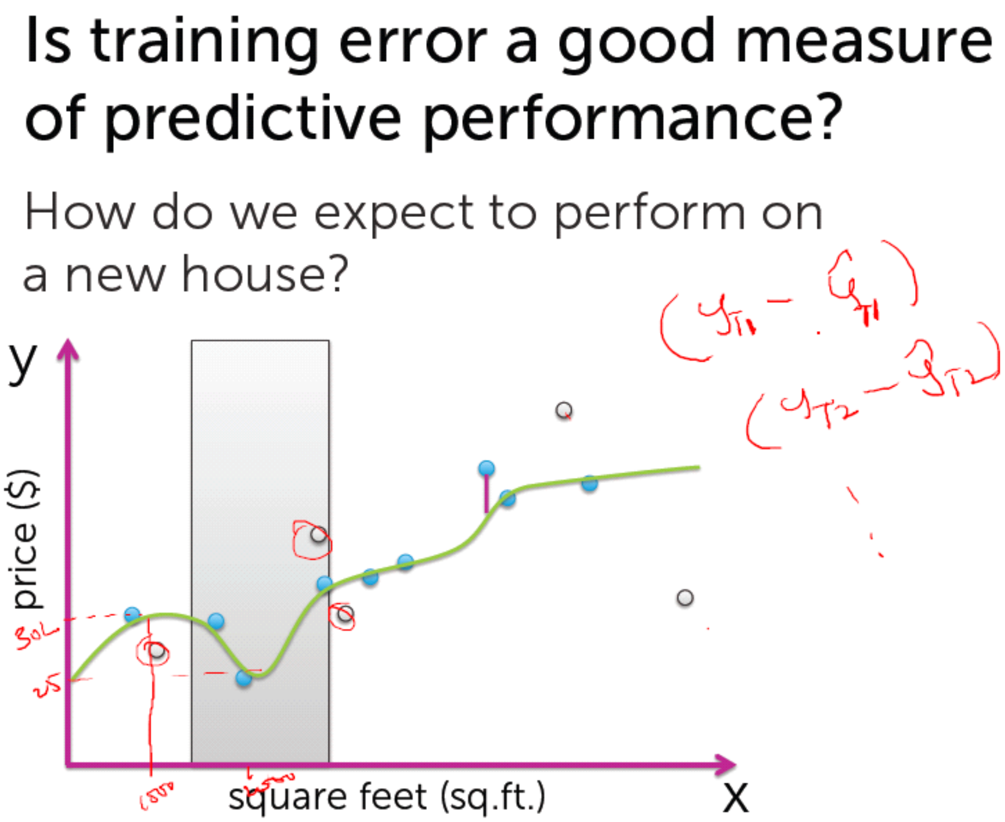
⭐ When I added one more variable to the model the error some more goes down but my linearity is becomes little curvature kind.



⭐ We see that if model increases complexity training error decreases



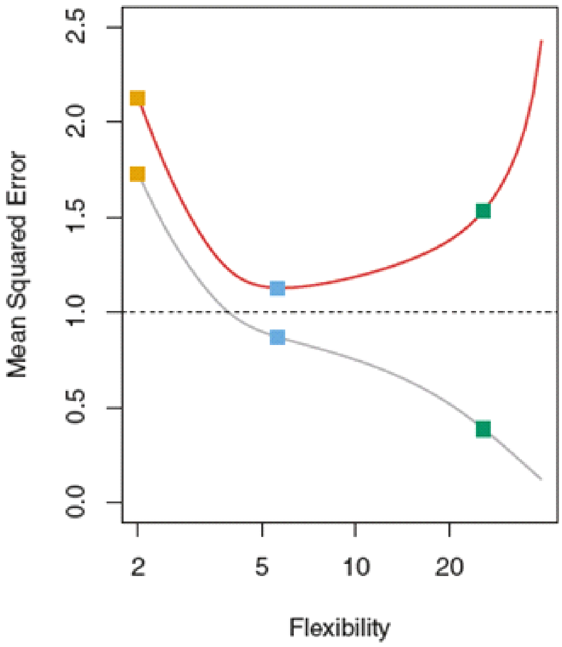
- Assuming that if my training error is less in my model it is going to predict well in the future and we are testing on test data with an assumption as it will be an Right model.



- But when you see the above graph domain knowledge sense is not matching with represented graph.
- Now, you can calculate the test error of test data of one model, Such that you can calculate the test errors of all possible models. Out of them, which ever model has minimum test error that is going to be your Best fit model.

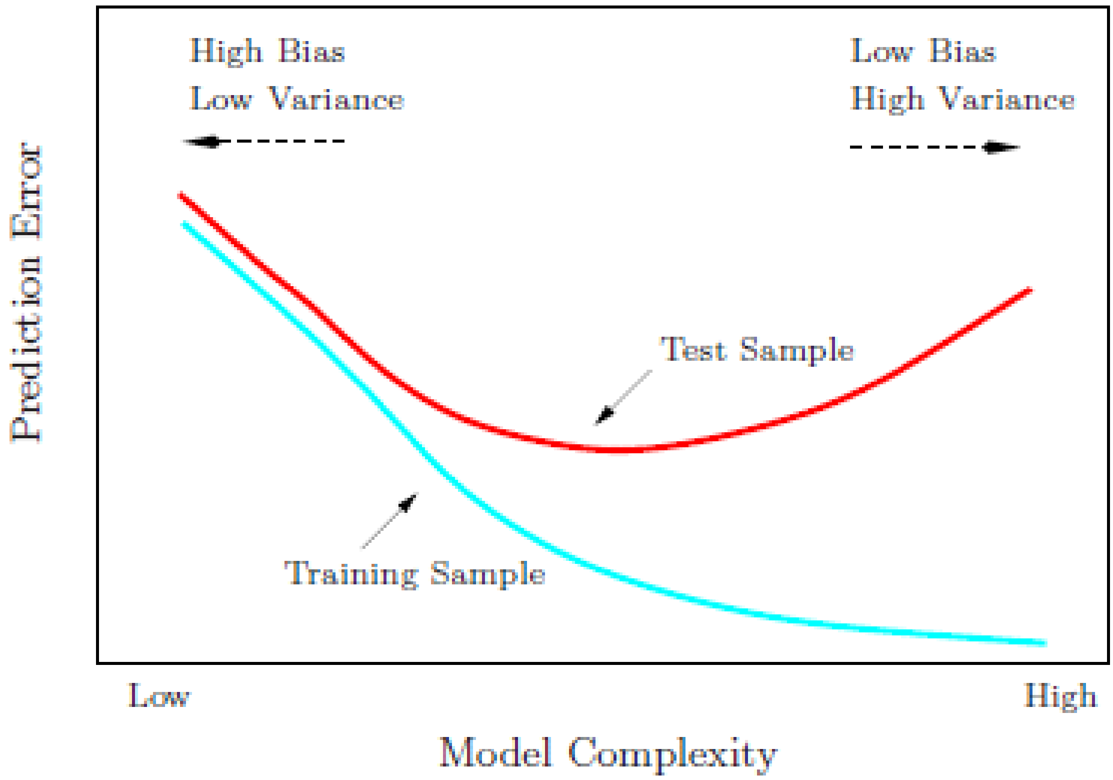
To make under stand well you need to approach

"Variance baised Trade-off"



- In the above figure, As with the training MSE, the test MSE initially declines as the level of flexibility increases. However, at some point the test MSE levels off and then starts to increase again .
- We know that from our previous knowledge

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$



- As the model becomes more and more complex, it uses the training data more and is able to adapt to more complicated underlying structures. Hence there is a decrease in bias but an increase in variance.

- Training error consistently decreases with model complexity, typically dropping to zero if we increase the model complexity enough. However, a model with zero training error is over fit to the training data and will typically generalize poorly.