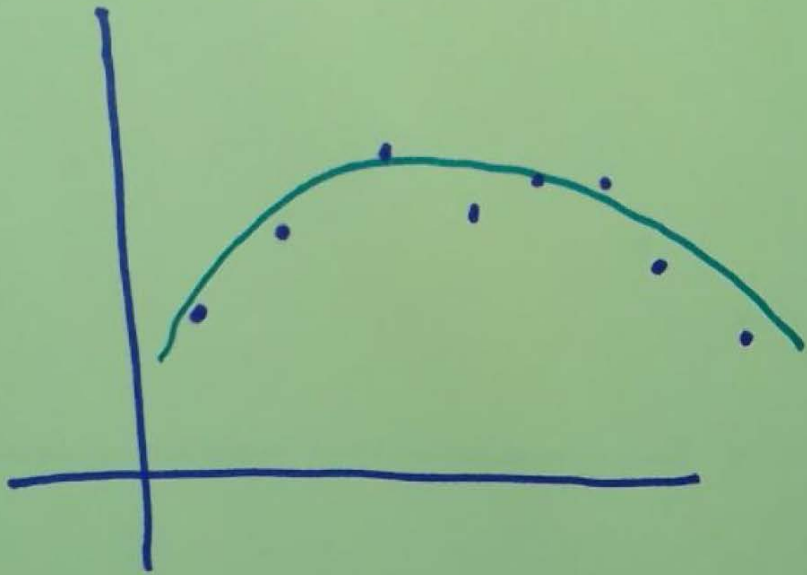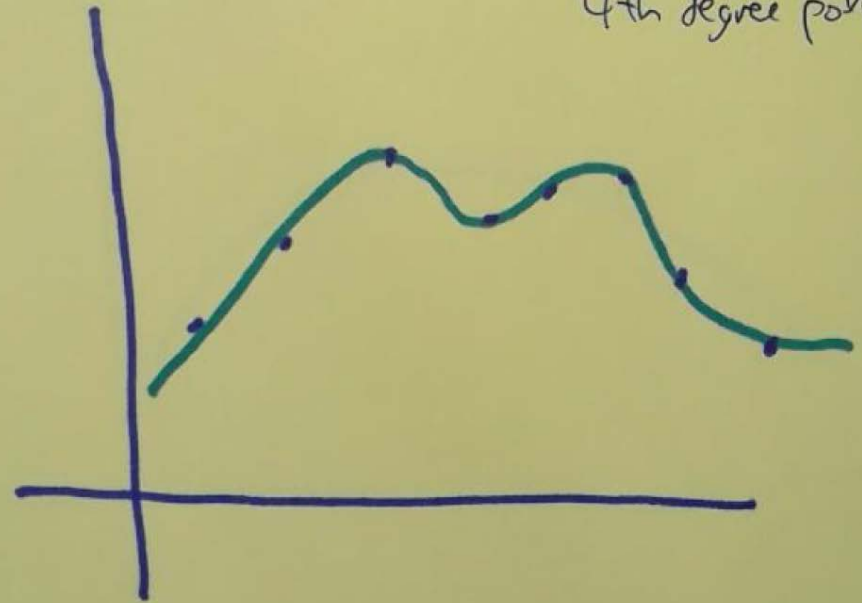# complexity, bias variance tradeoff, overfitting

TASK: GIVEN THESE TEST DATA POINTS, FIND A CURVE THAT BEST PREDICTS UNSEEN DATA

2nd degree polynomial

4th degree polynomial

WHICH DECISION TREE WILL BE MORE ACCURATE AT REPRESENTING TEST DATA? AT PREDICTING FUTURE DATA?
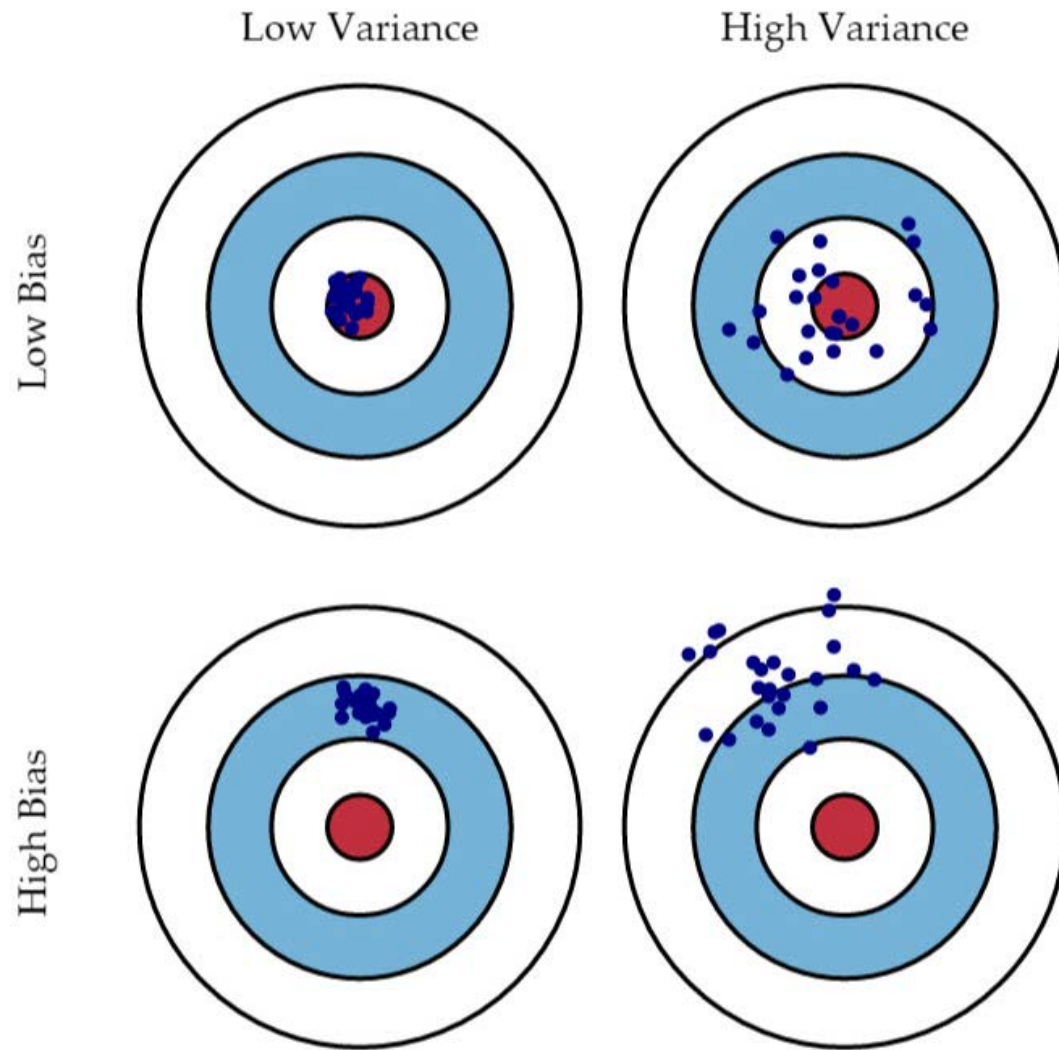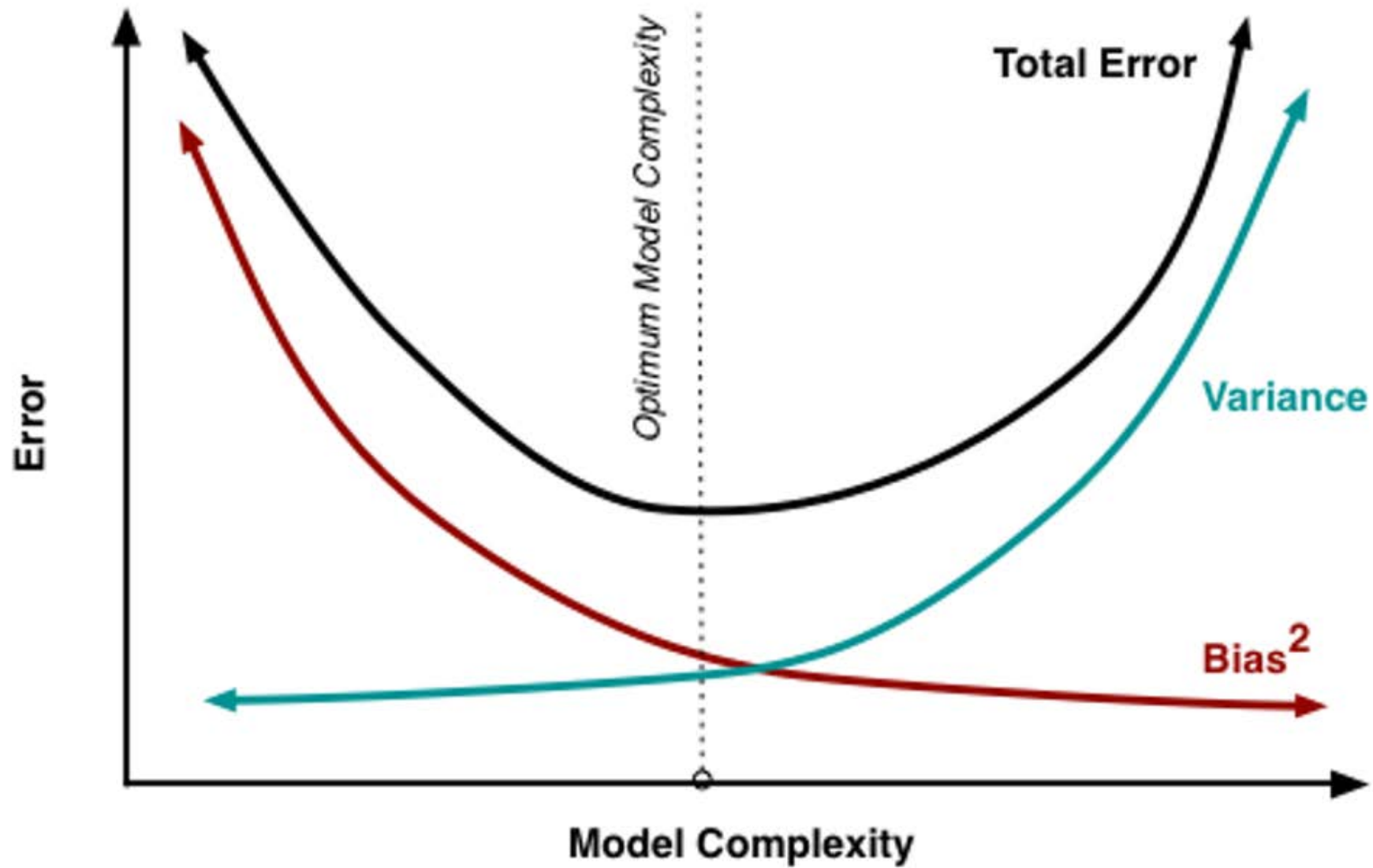
Low Variance   High Variance

Low Bias

High Bias

Fig. 1 Graphical illustration of bias and variance.

As the **flexibility** of f-hat (our estimate for y) increases, its variance increases and its bias decreases.

• **Variance** refers to the amount by which f-hat would change if we estimated it using a different training data set.
• **Bias** refers to the error that is introduced by approximating a real life problem, which may be extremely complicated, by a much simpler model.

Source: Bias variance tradeoff. http://scott.fortmann-roe.com/docs/BiasVariance.html]

Source: *Data Science for Business*, chapter 5