Machine Learning

Session 3

24th Feb 2018

Rama Krishna Bhupathi
ramakris@gmail.com

Agenda

- Terminologies
- Linear Regression MultiVariable
- Descriptive Statistics
- Predicting House Prices using Regression

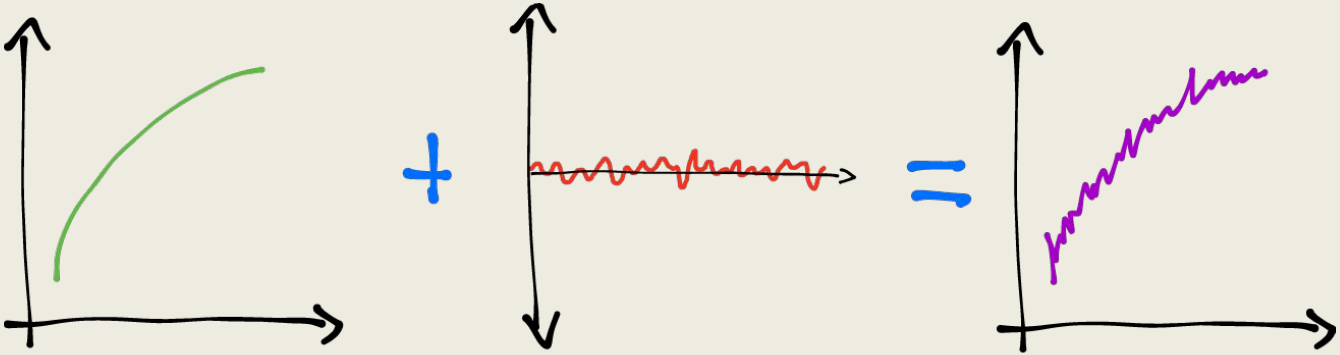
Noise Vs. Signal?

What is Signal?

In predictive modeling, you can think of the “signal” as the true underlying pattern that you wish to learn from the data.

What is Noise? ?

“Noise,” on the other hand, refers to the irrelevant information or randomness in a dataset.



© Machine Learning @ Berkeley

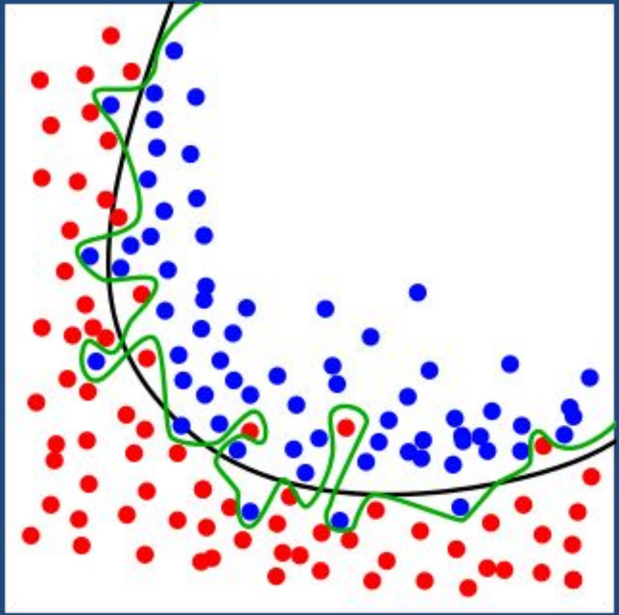
Our “perfect” function added to noise is what we end up measuring in the real world.

What is *Overfitting/Underfitting?*

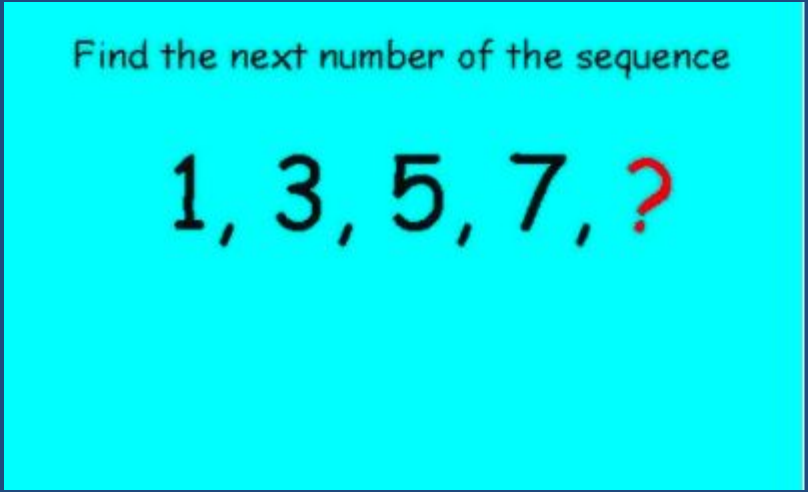
Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data. The model doesn't generalize well from our training data to unseen data.

Intuitively, overfitting occurs when the model:

- ***algorithm fits the data too well.***
- ***shows low bias but high variance.***
- ***complicated model***



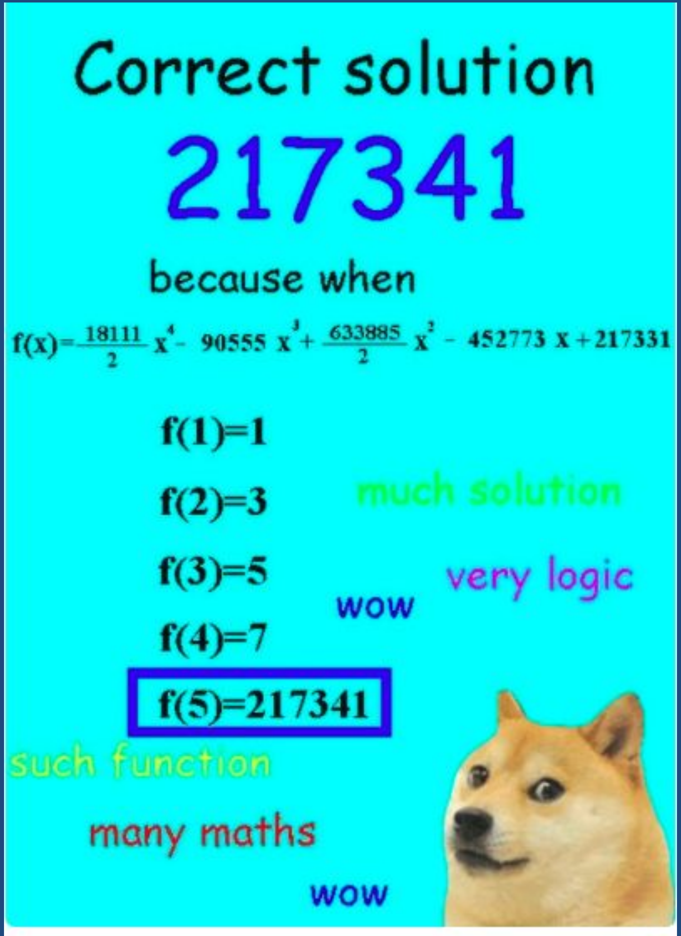
Quiz?



Find the next number of the sequence

1, 3, 5, 7, ?

Solution...



Correct solution

217341

because when

$$f(x)=\frac{18111}{2}x^4-90555\,x^3+\frac{633885}{2}x^2-452773\,x+217331$$

$f(1)=1$

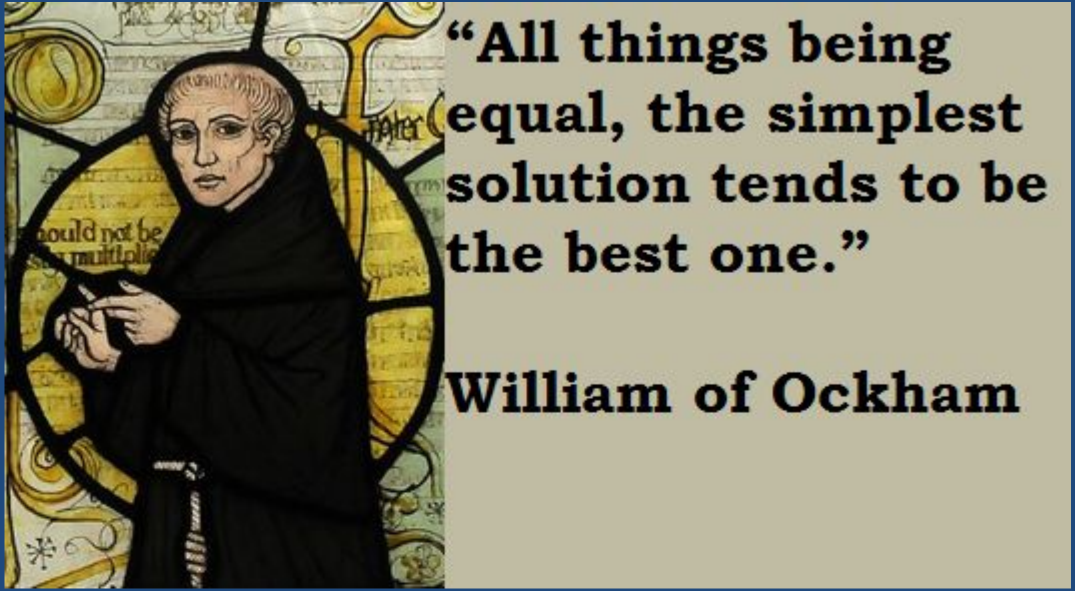
$f(2)=3$

$f(3)=5$

$f(4)=7$

$\boxed{f(5)=217341}$

Occam’s Razor



“All things being equal, the simplest solution tends to be the best one.”

William of Ockham

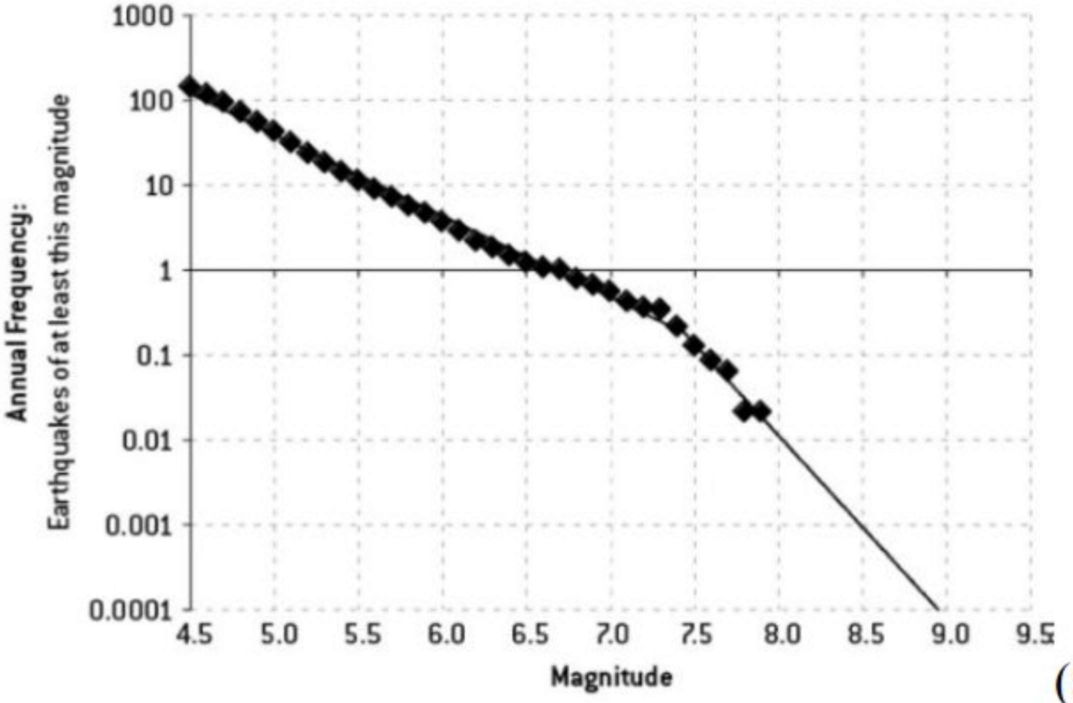
Fukushima Nuclear Disaster

The Fukushima Daiichi nuclear disaster was an energy accident at the Fukushima Daiichi Nuclear Power Plant in Ōkuma, Fukushima Prefecture, initiated primarily by the tsunami following the Tōhoku earthquake (magnitude 9 .1)on 11 March 2011.

What has this to do with overfitting?

Fukushima Nuclear Disaster

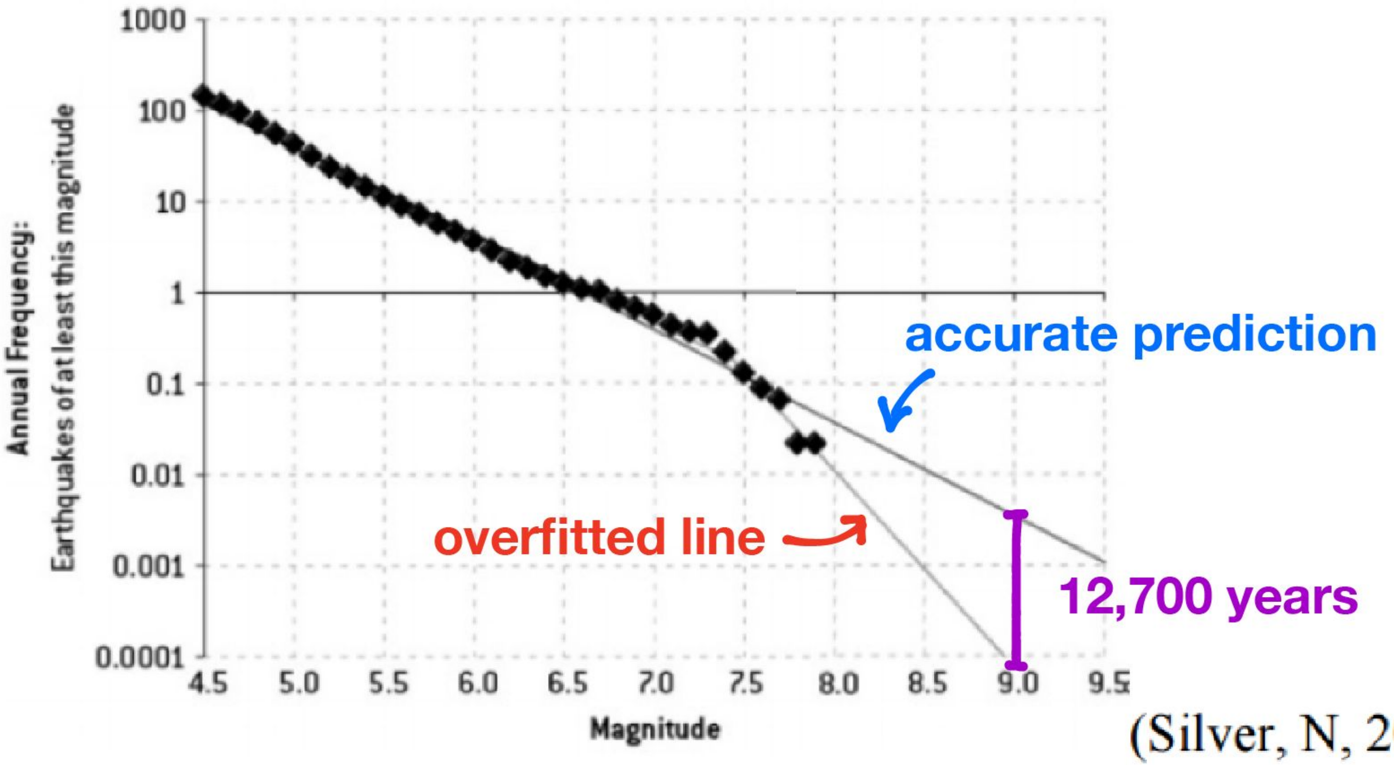
FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
CHARACTERISTIC FIT



(Silver, N, 2012)

Overfitting

FIGURE 5-7C: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
CHARACTERISTIC FIT



accurate prediction

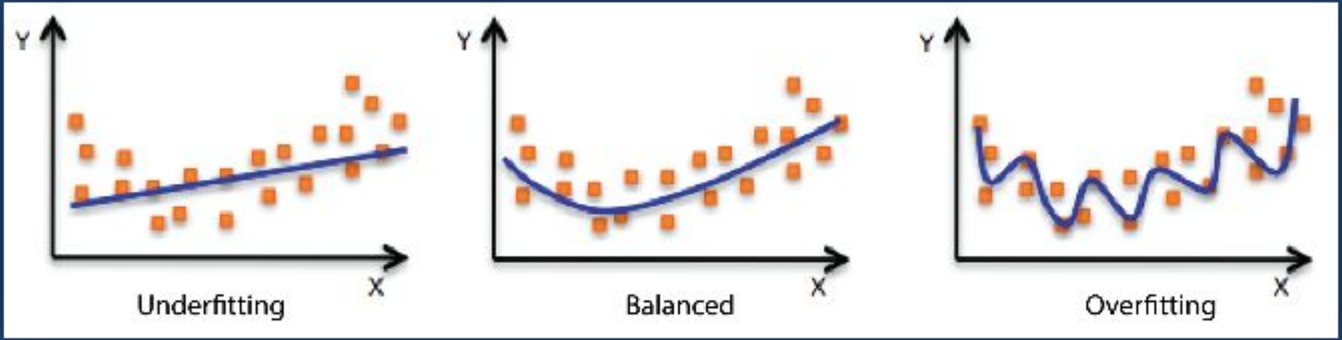
overfitted line

12,700 years

(Silver, N, 2012)

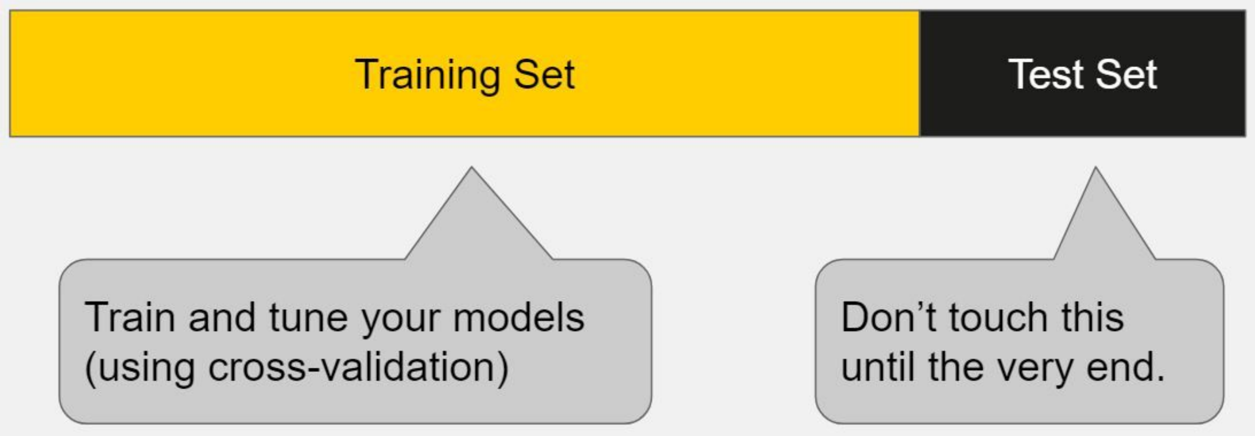
Underfitting?

Underfitting occurs when a model is too simple – informed by too few features or regularized too much – which makes it inflexible in learning from the dataset.



How do you know your model is overfitted?

If our model does much better on the training set than on the test set, then we’re likely overfitting.


Training Set
Test Set
Train and tune your models (using cross-validation)
Don’t touch this until the very end.

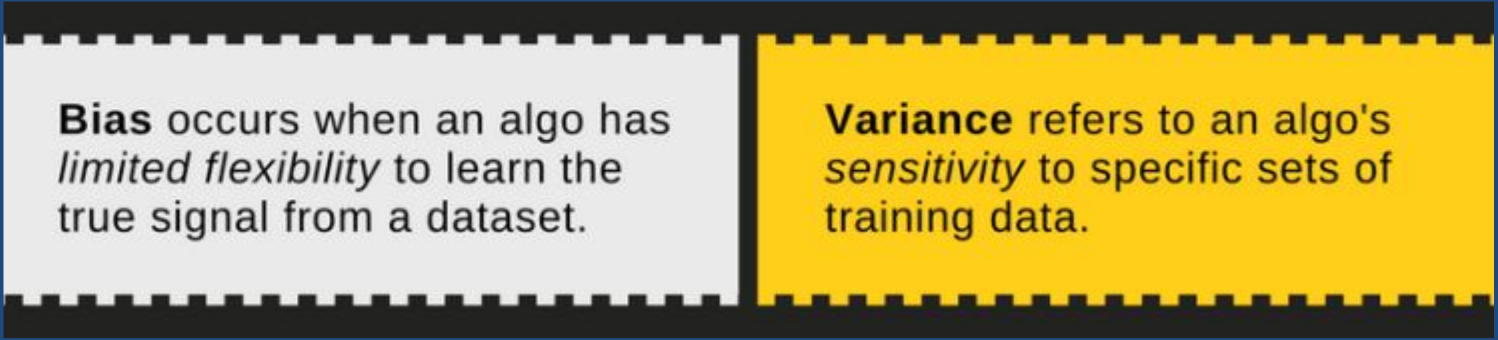
Prevent overfitting/underfitting?

***Underfitting*: Get More Data.**

***Overfitting*:**

- ***Cross Validation (K-Fold Cross Validation)***
- ***Train with more data***
- ***Regularization Techniques***
- ***Remove features that are irrelevant. Most algorithms will help you find that.***
- ***Early stopping or reduce complexity***
- ***Boosting Algorithms***
- ***Ensembling (putting multiple models together)***

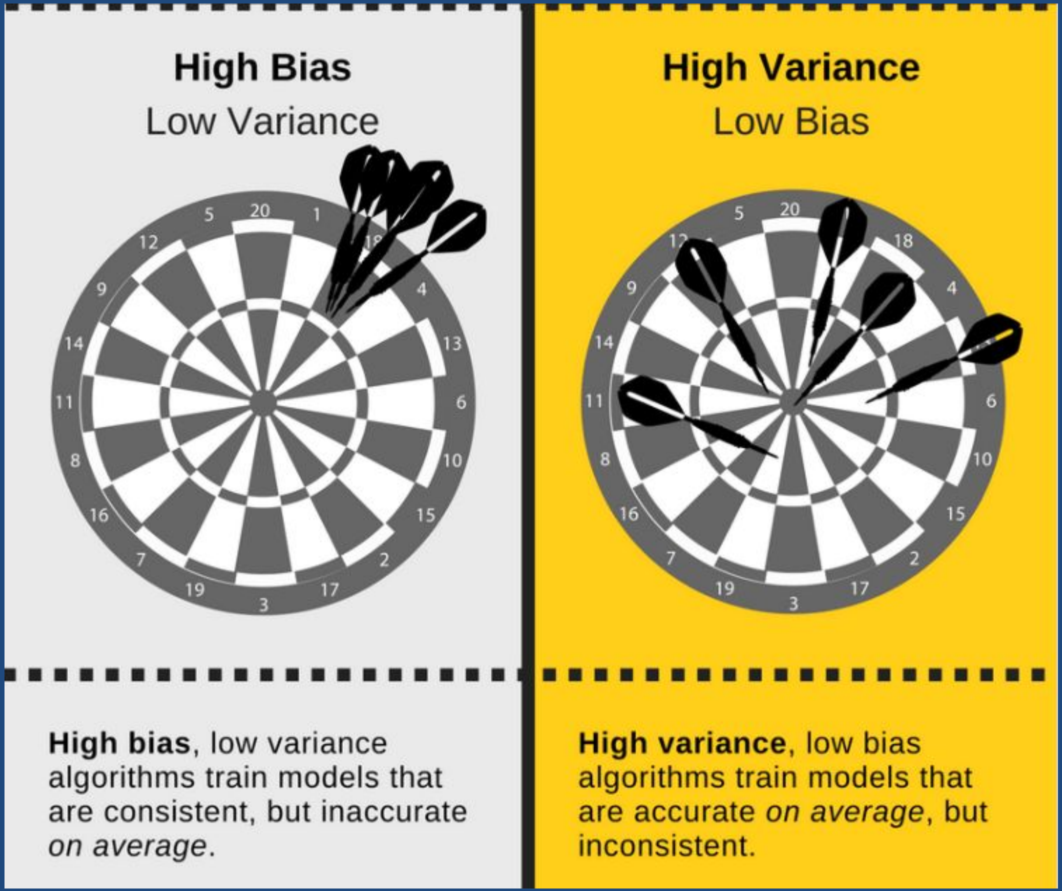
Bias Vs. Variance?



Bias occurs when an algo has *limited flexibility* to learn the true signal from a dataset.

Variance refers to an algo's *sensitivity* to specific sets of training data.

Bias Vs. Variance?



High Bias

Low Variance

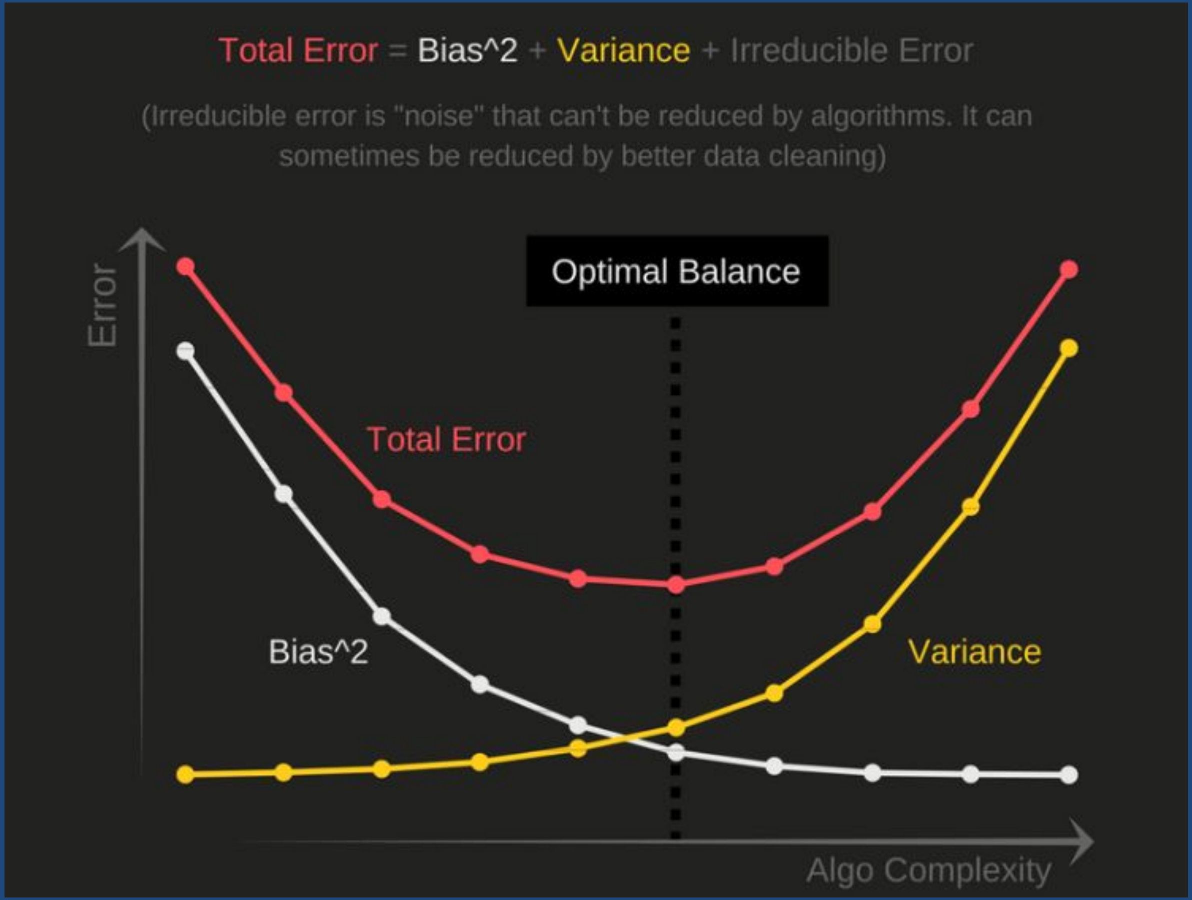
High Variance

Low Bias

High bias, low variance algorithms train models that are consistent, but inaccurate *on average*.

High variance, low bias algorithms train models that are accurate *on average*, but inconsistent.

Bias Vs. Variance?



Total Error = Bias^2 + Variance + Irreducible Error

(Irreducible error is "noise" that can't be reduced by algorithms. It can sometimes be reduced by better data cleaning)