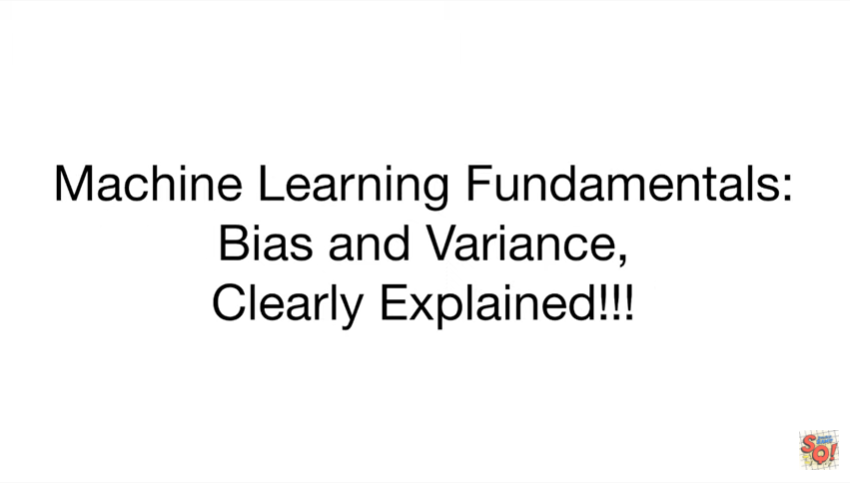
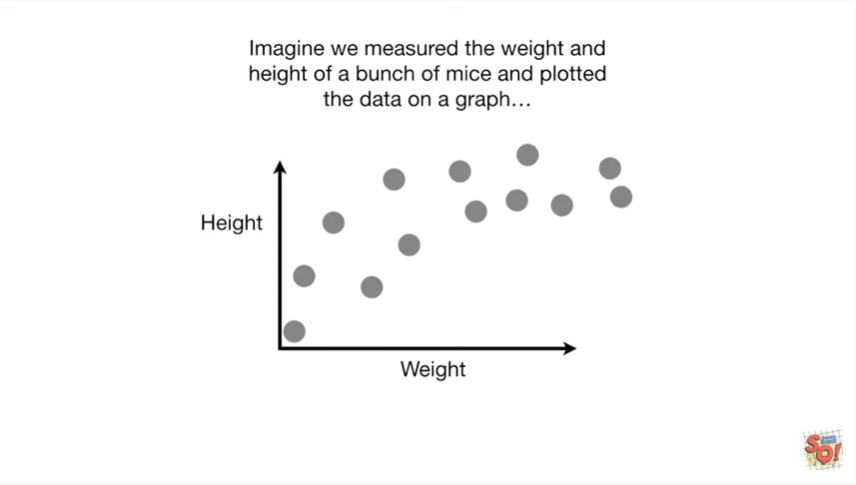
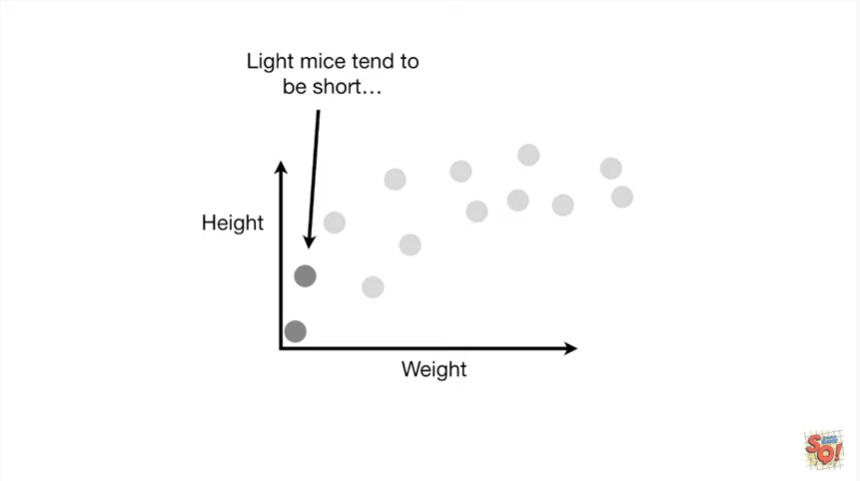
<https://www.youtube.com/watch?v=EuBBz3bI-aA&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=5>



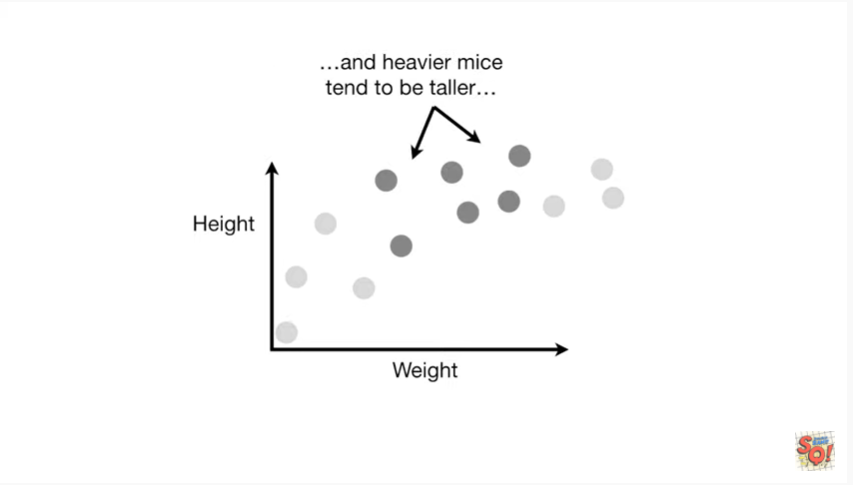
Today we're going to be talking about some machine learning fundamentals bias and variance and they're gonna be clearly explained.



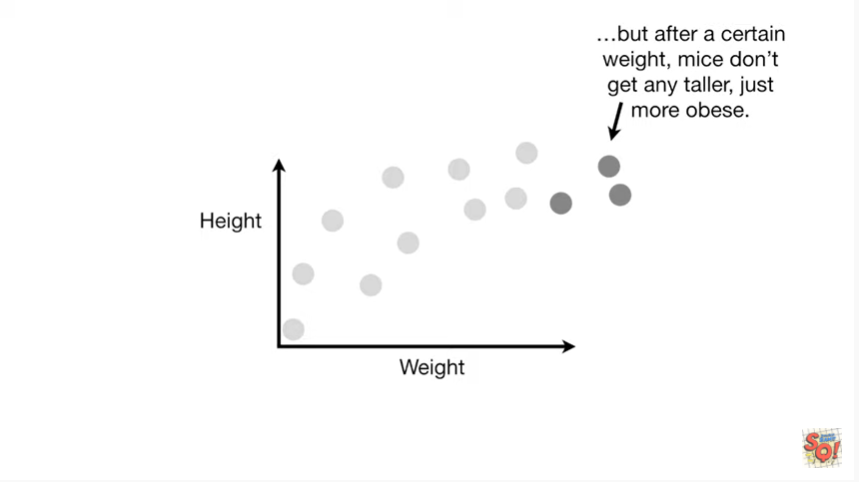
Imagine we measured the weight and height of a bunch of mice and plotted the data on a graph.



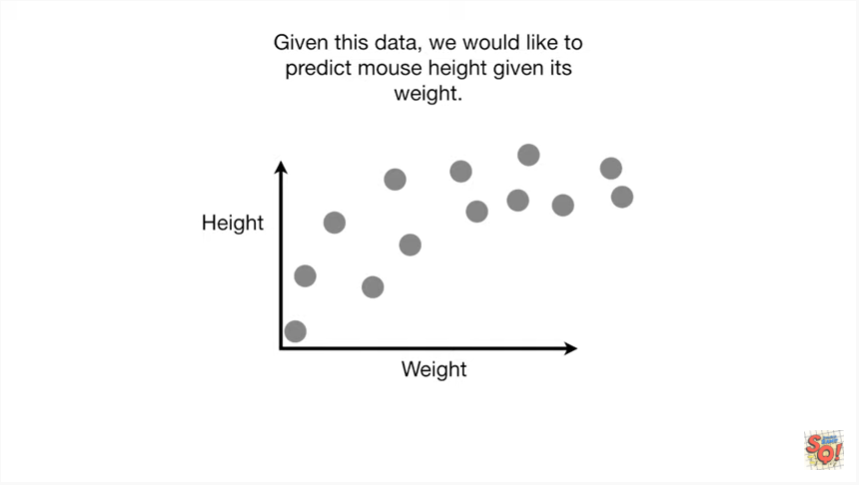
Light mice tend to be short



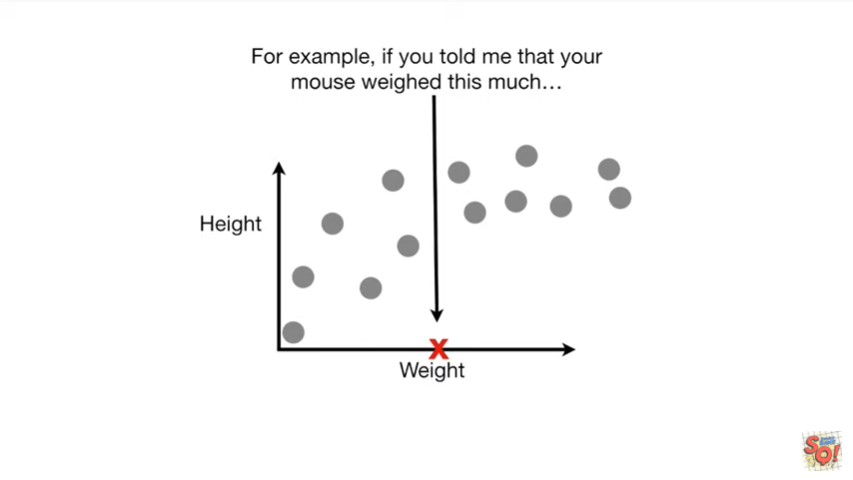
and heavier mice tend to be taller



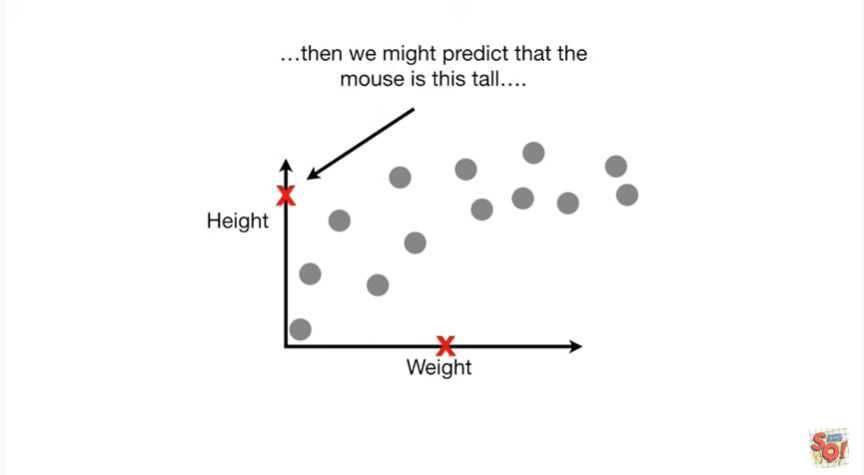
but after a certain weight mice don't get any taller just more obese.



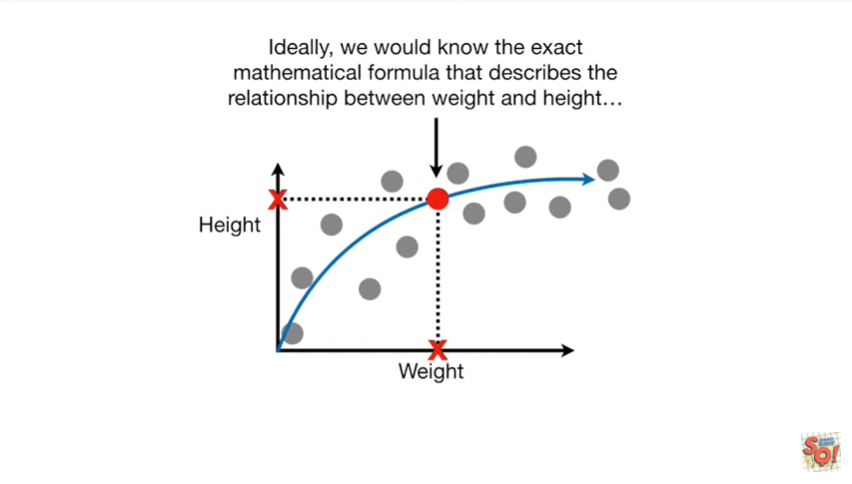
Given this data we would like to predict Mouse height given its weight.



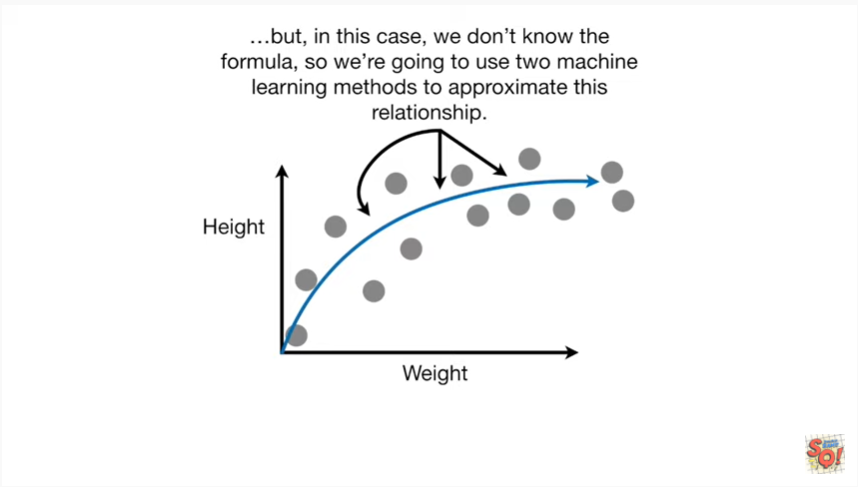
For example if you told me your mouse weighed this much



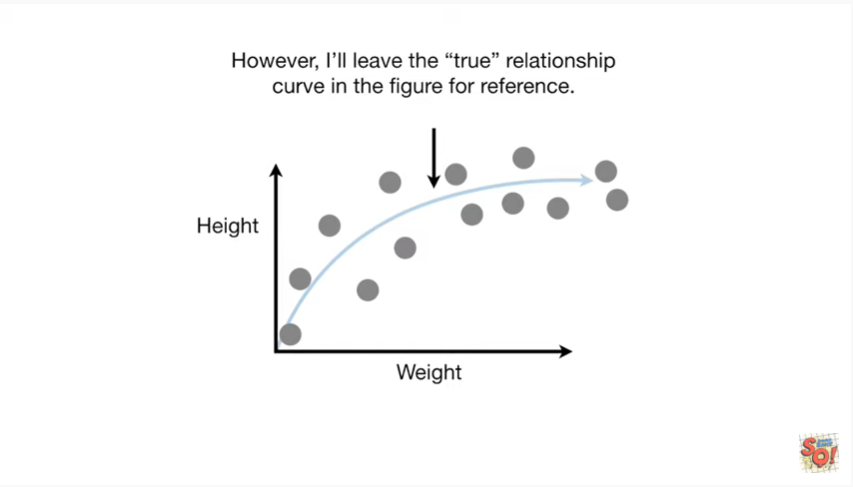
then we might predict that the mouse is this tall.



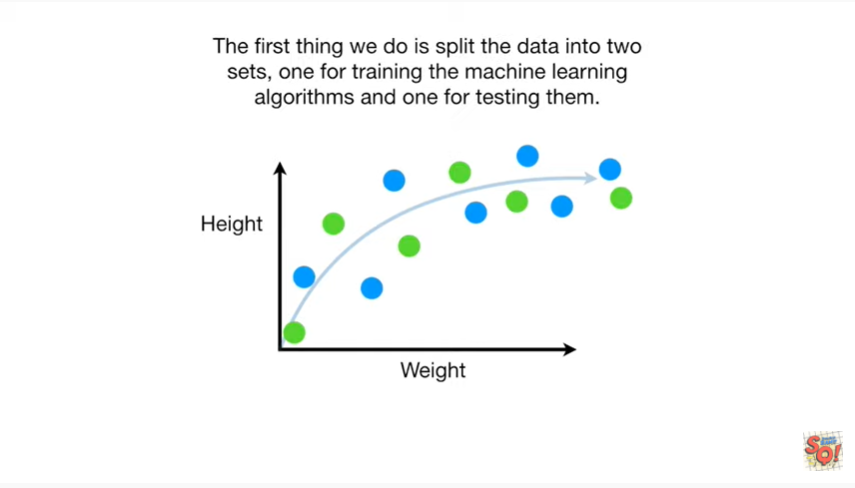
Ideally we would know the exact mathematical formula that describes the relationship between weight and height



but in this case we don't know the formula so we're going to use two machine learning methods to approximate this relationship.



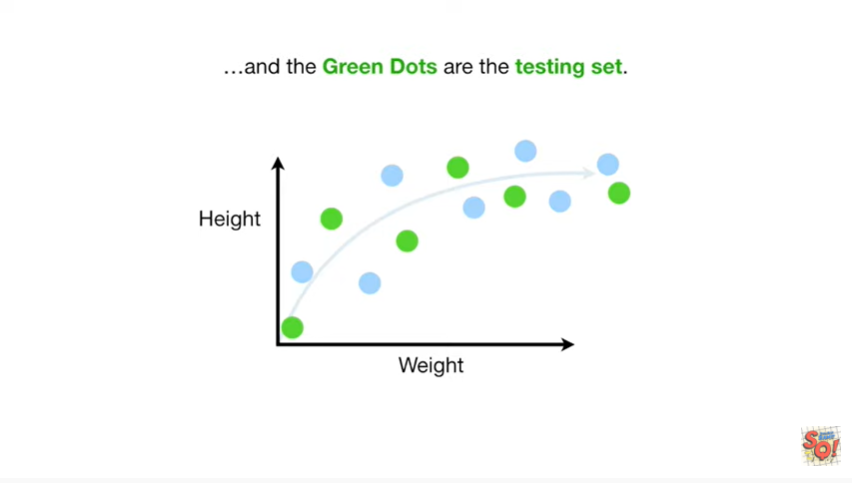
However I'll leave the true relationship curve in the figure for reference.



The first thing we do is split the data into two sets one for training the machine learning algorithms and one for testing them .



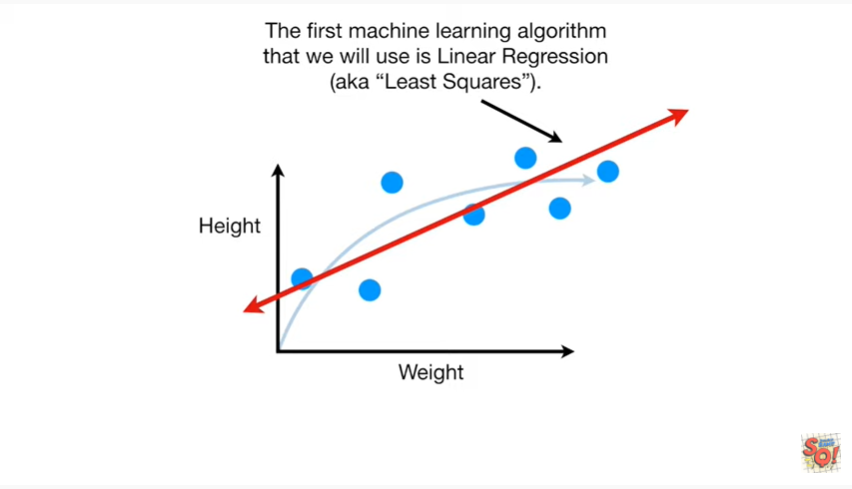
The blue dots are the training set



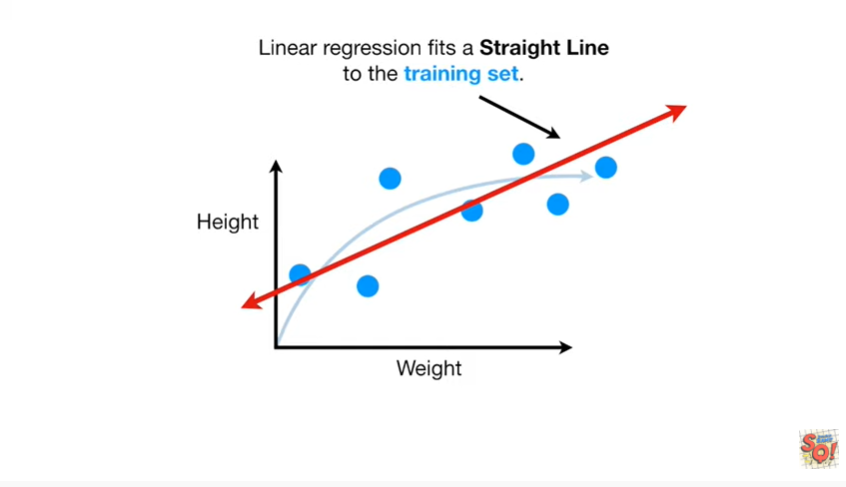
and the green dots are the testing set.



Here's just the training set.



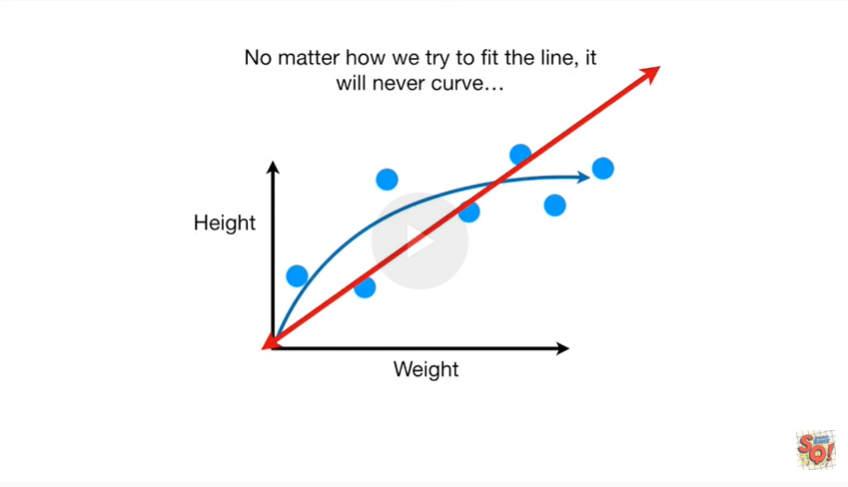
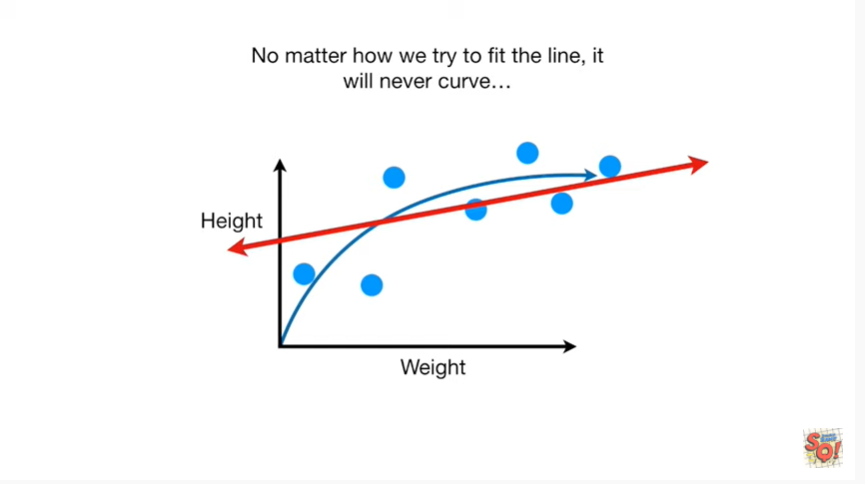
The first machine learning algorithm that we will use is linear regression aka least squares.



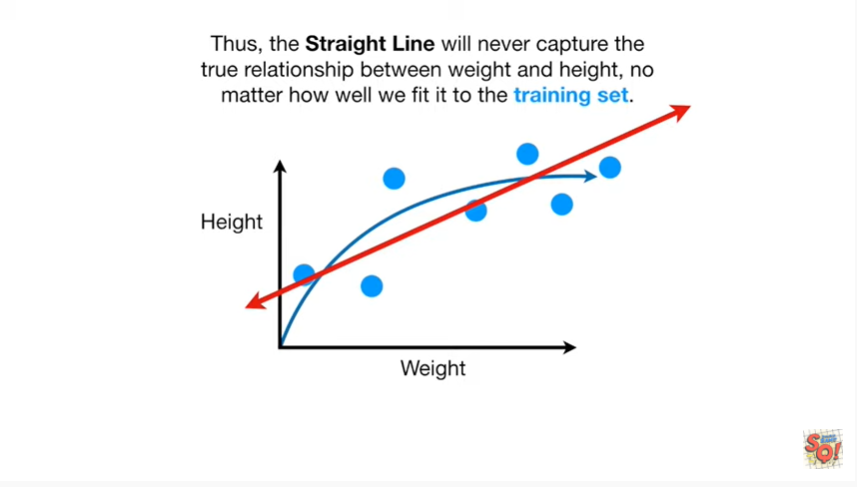
Linear regression it's a straight line to the training set.



Note : the straight line doesn't have the flexibility to accurately replicate the arc in the true relationship.

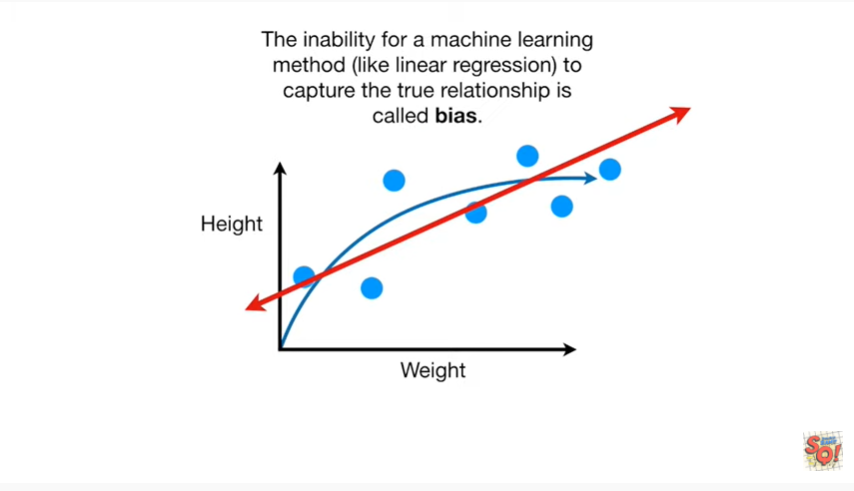


No matter how we try to fit the line it will never curve.

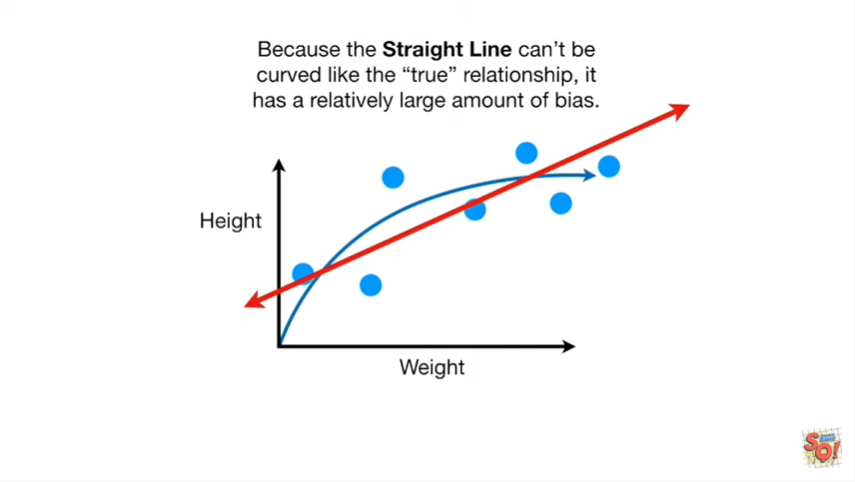


Thus the straight line will never capture the true relationship between weight and height.

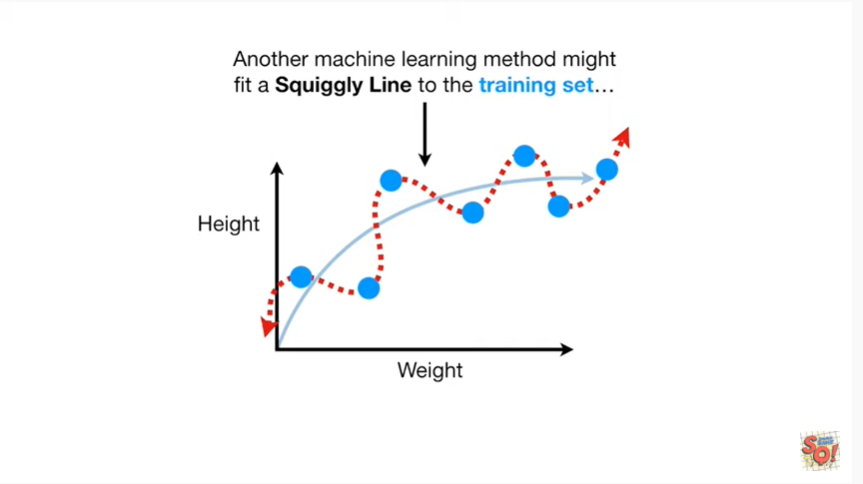
No matter how well we fit it to the training set.



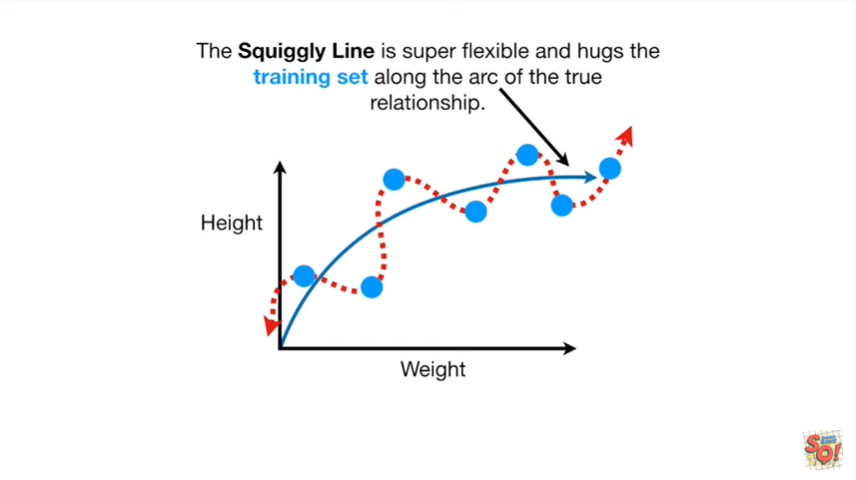
The inability for a machine learning method like linear regression to capture the true relationship is called bias.



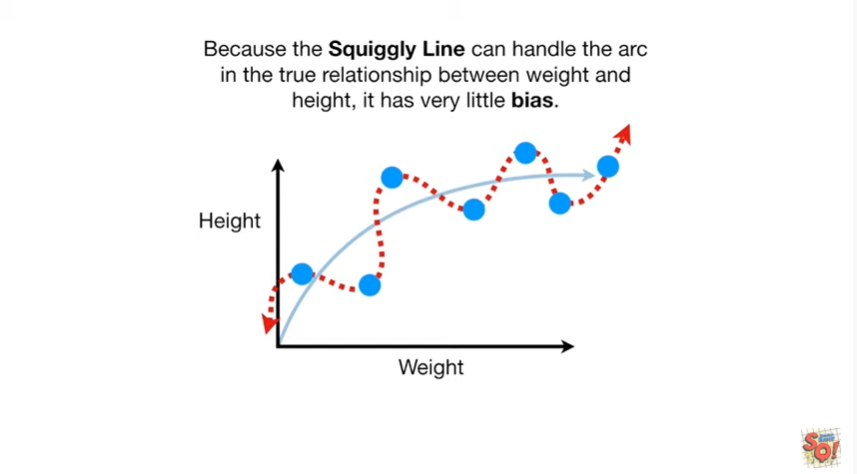
Because the straight line can't be curved like the true relationship it has a relatively large amount of bias.



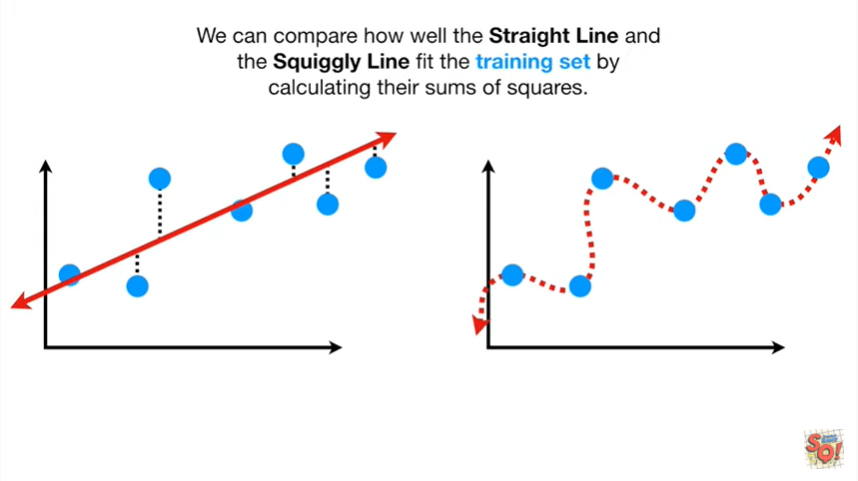
Another machine learning method might fit a squiggly line to the training set.



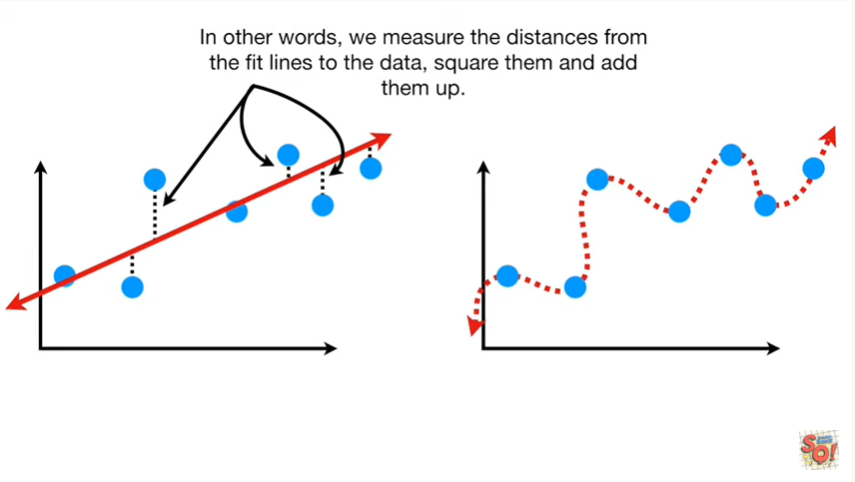
The squiggly line is super flexible and hugs the training set along the arc of the true relationship.



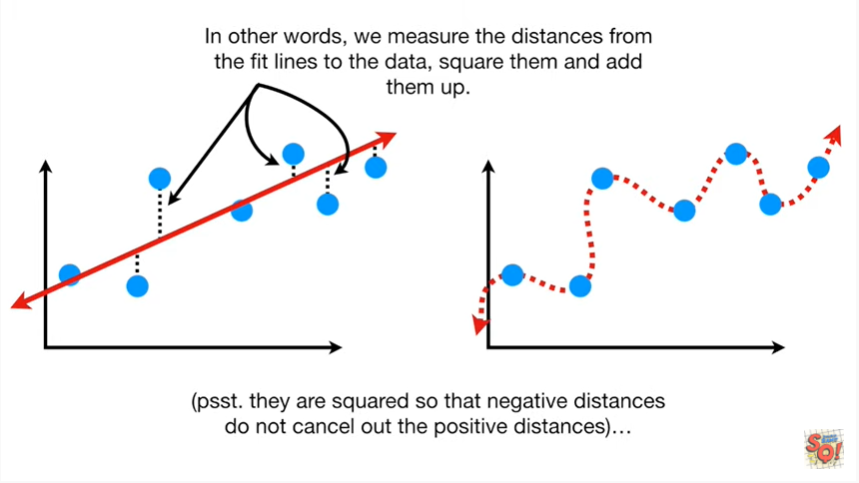
Because the squiggly line can handle the arc in the true relationship between weight and height it has very little bias.



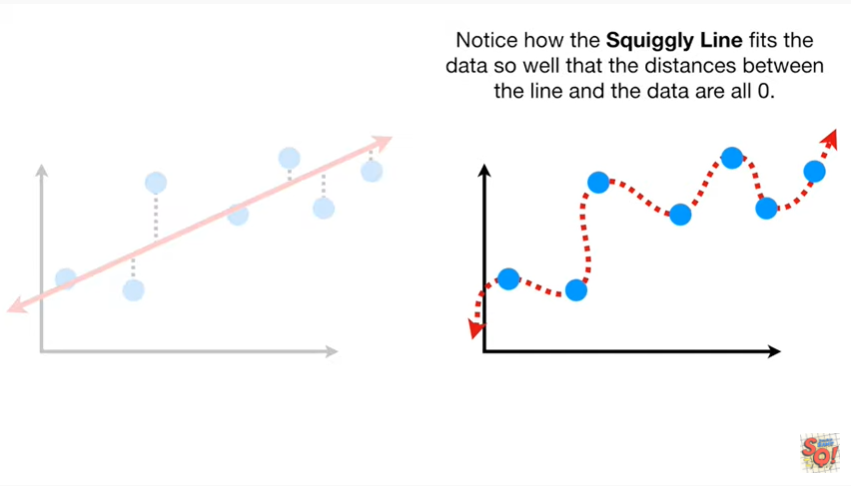
We can compare how well the straight line and the squiggly line fit the training set by calculating their sums of squares.



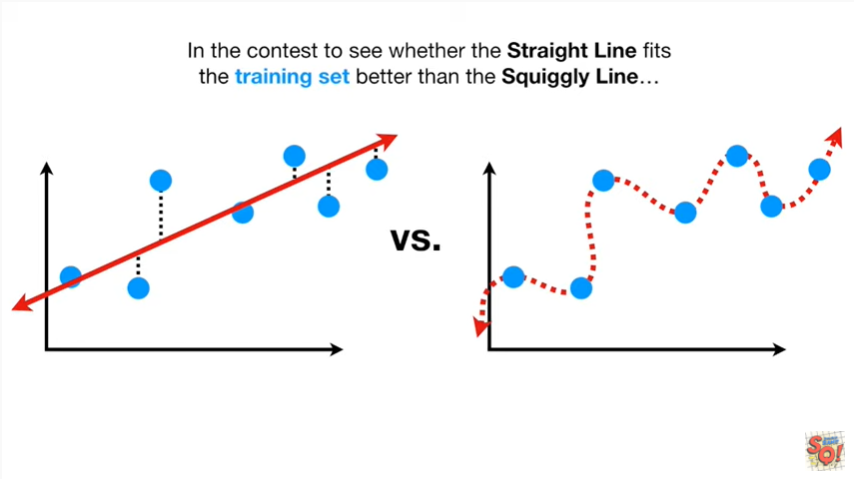
In other words we measure the distances from the fit lines to the data, square them and add them up.



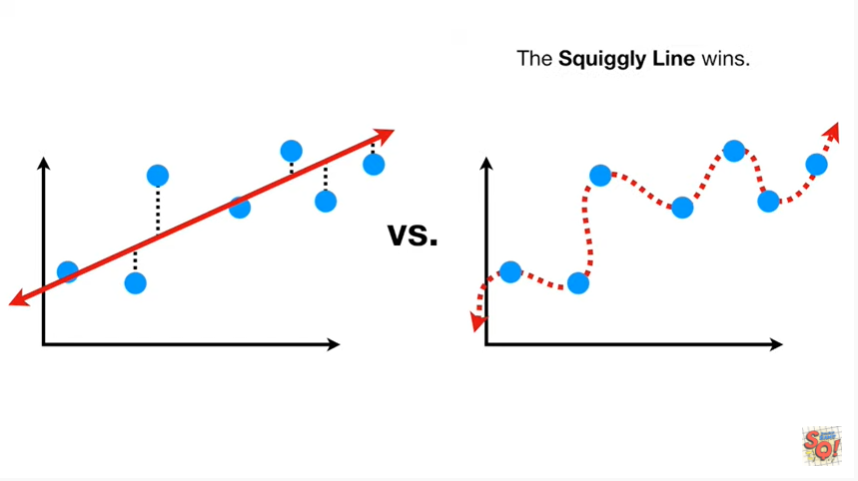
Psst they are squared so that negative distances do not cancel out positive distances.



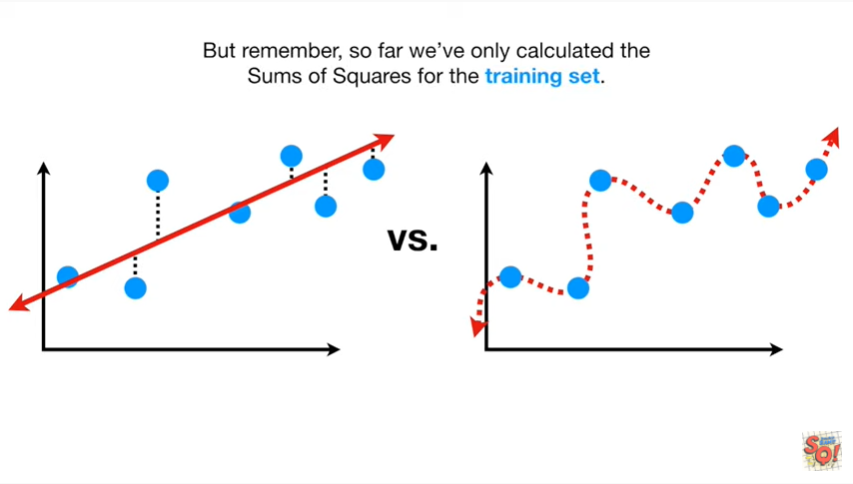
Notice how the squiggly line fits the data so well that the distances between the line and the data are all 0.



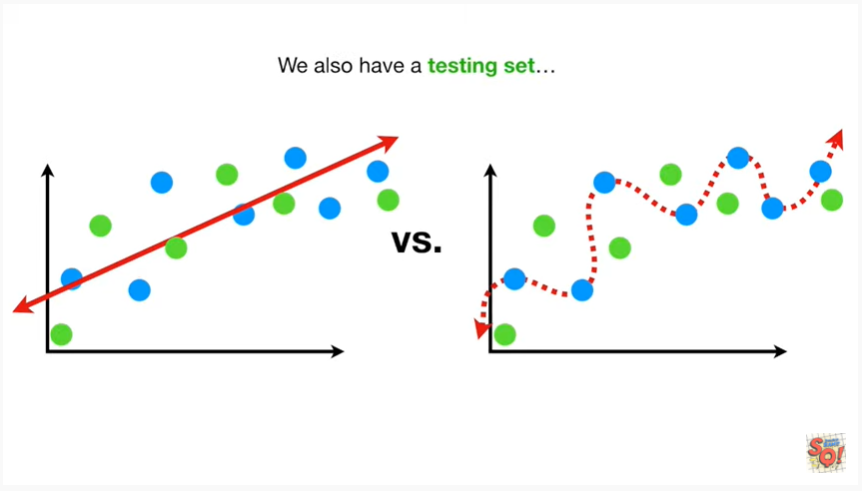
In the contest to see whether the straight line fits the training set better than the squiggly line.



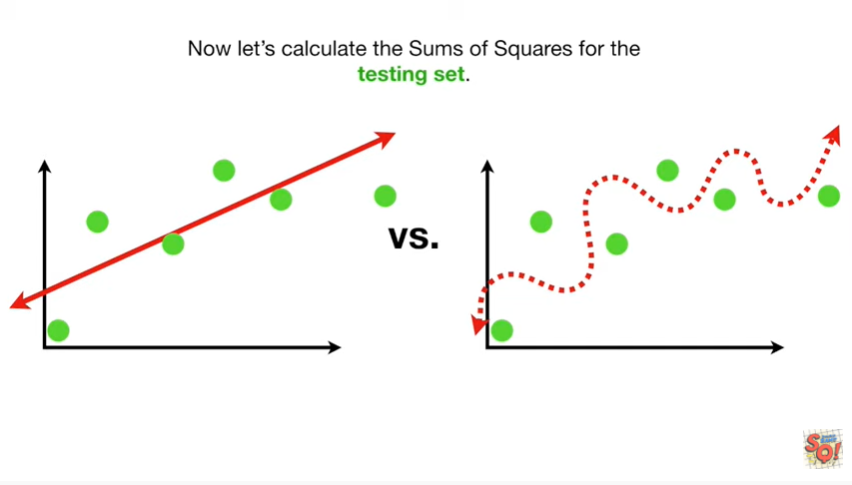
The squiggly line wins.



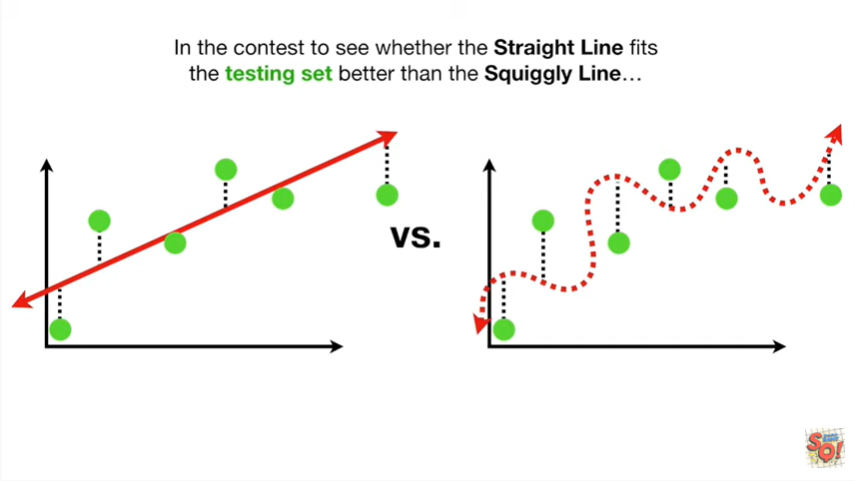
But remember so far we've only calculated the sums of squares for the training set.



We also have a testing set.



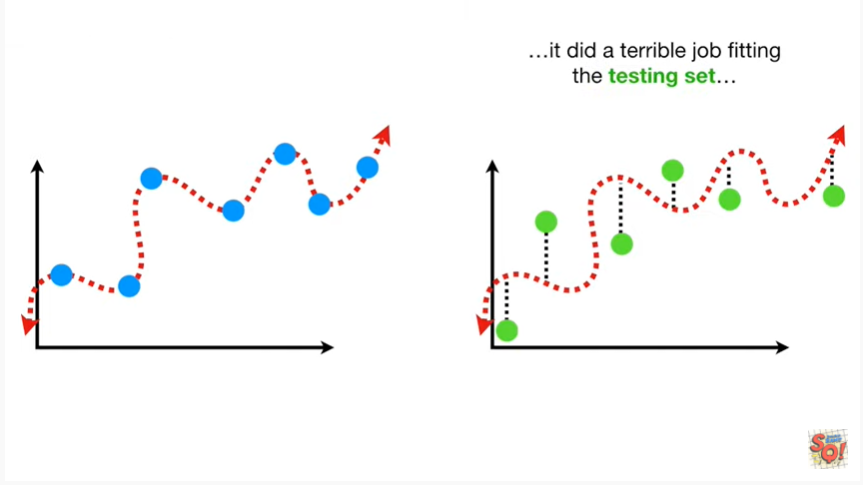
Now let's calculate the sums of squares for the testing set in the contest to see whether the straight line fits the testing set better than the squiggly line.



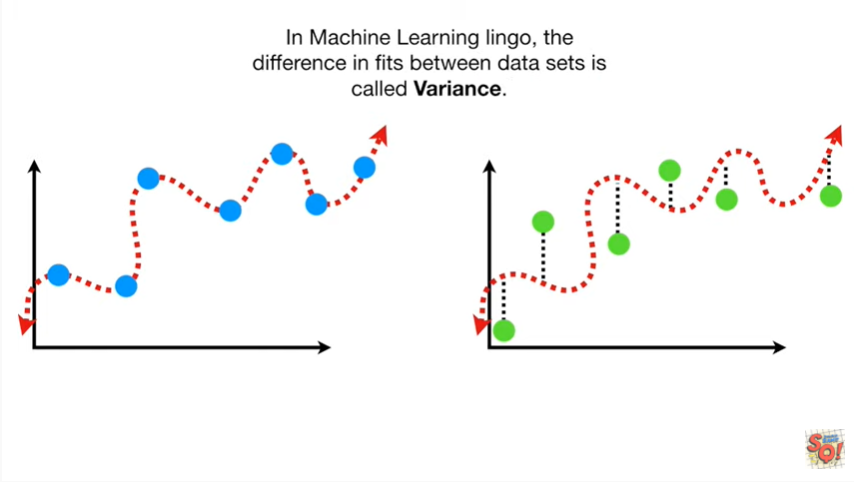
The straight line wins.



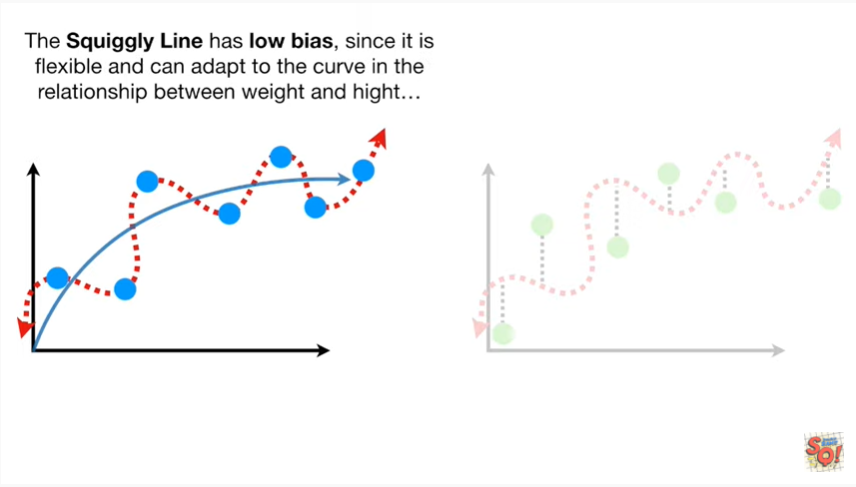
Even though the squiggly line did a great job fitting the training set



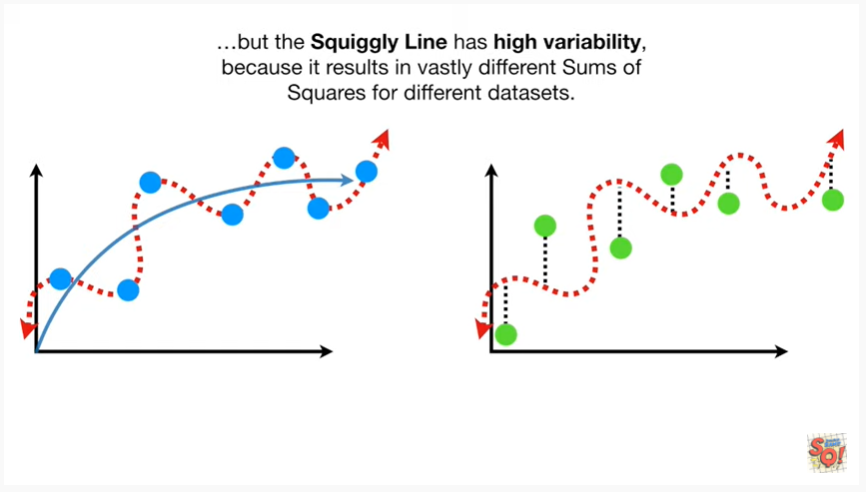
it did a terrible job fitting the testing set.



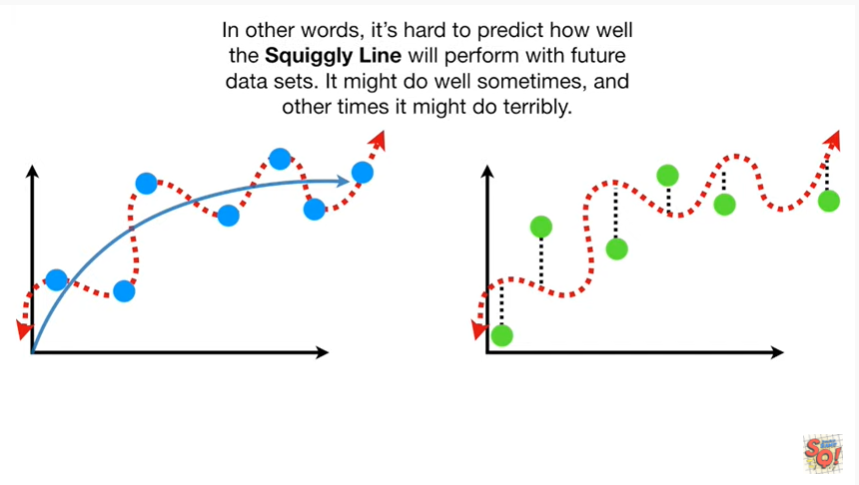
In machine learning lingo the difference in fits between data sets is called variance.



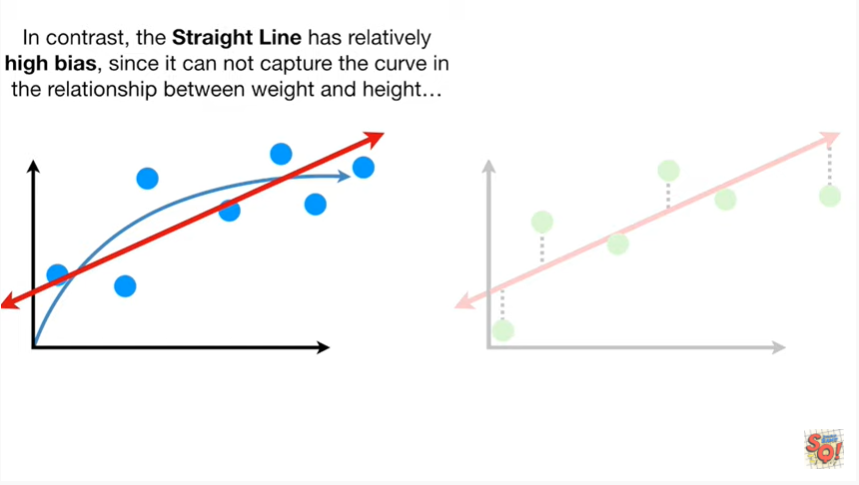
The squiggly line has low bias since it is flexible and can adapt to the curve in the relationship between weight and height



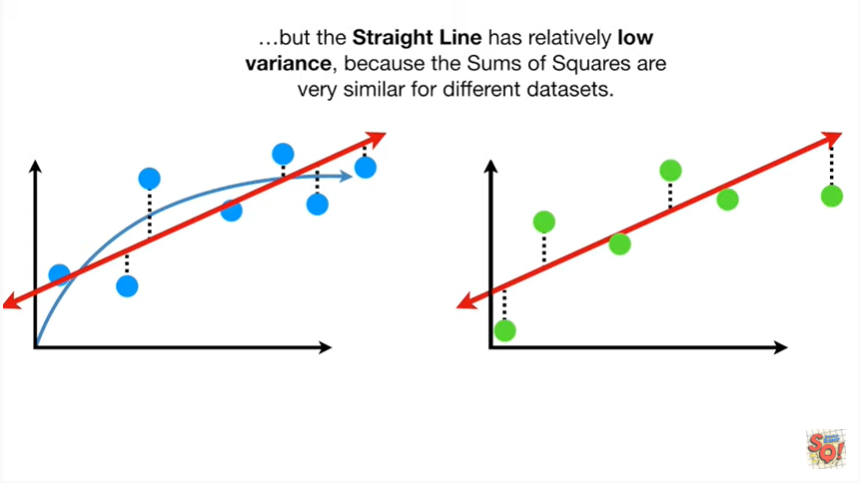
but the squiggly line has high variability because it results in vastly different sums of squares for different data sets.



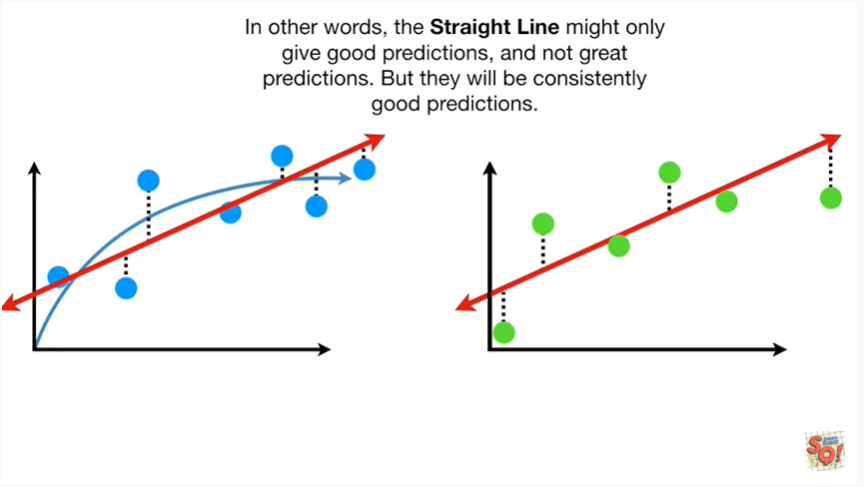
In other words it's hard to predict how well the squiggly line will perform with future data sets it might do well sometimes and other times it might do terribly.



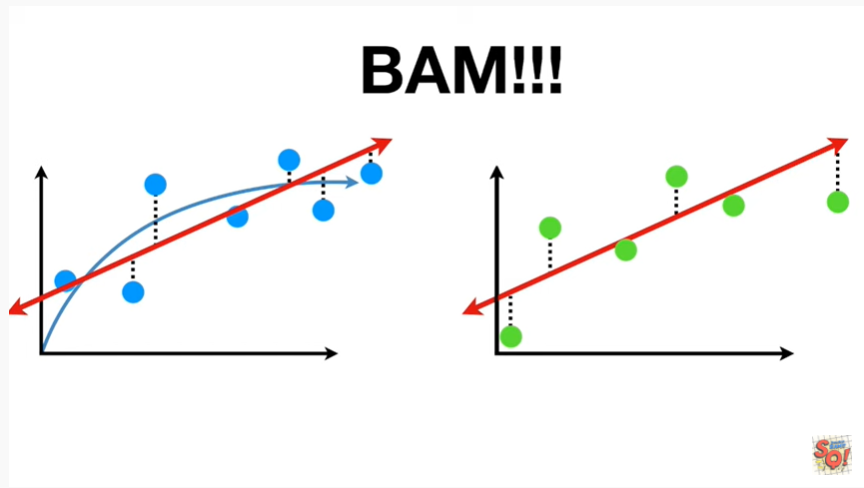
In contrast the straight line has relatively high bias since it cannot capture the curve in the relationship between weight and height



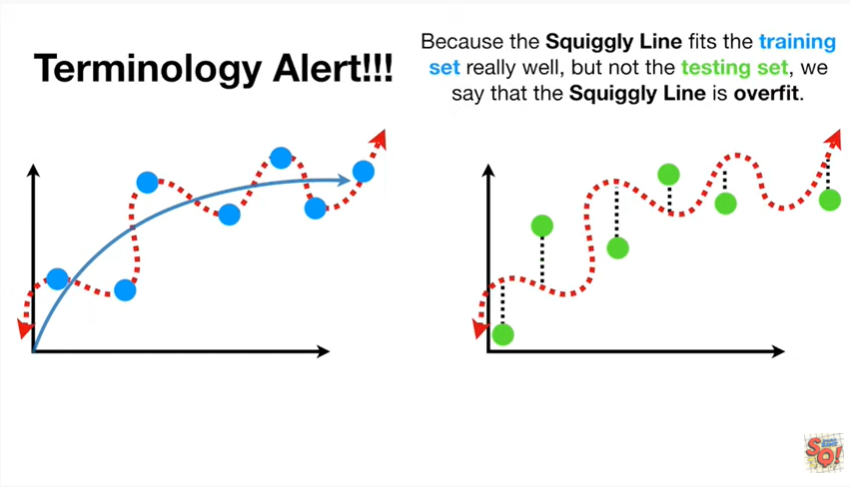
but the straight line has relatively low variance because the sums of squares are very similar for different data sets.



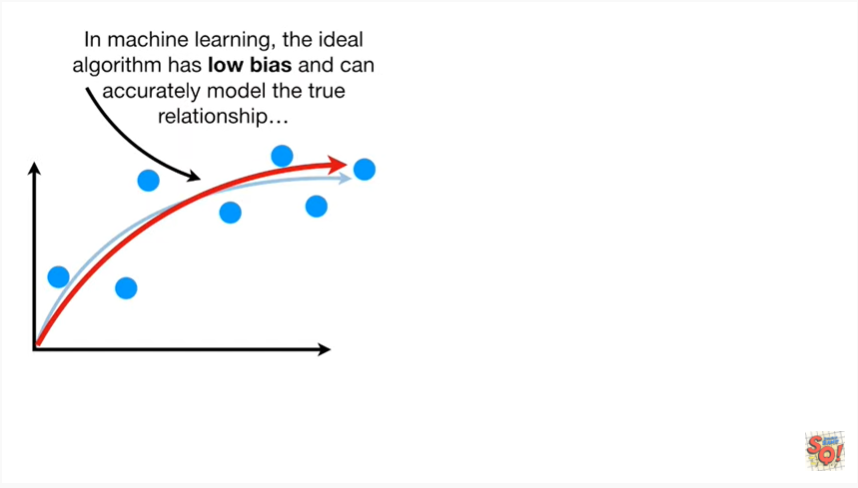
In other words the straight line might only give good predictions and not great predictions but they will be consistently good predictions.



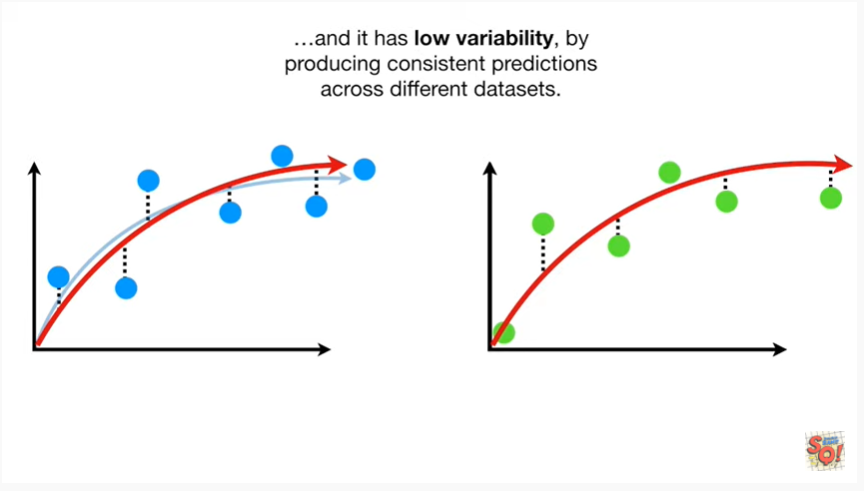
BAM !!!



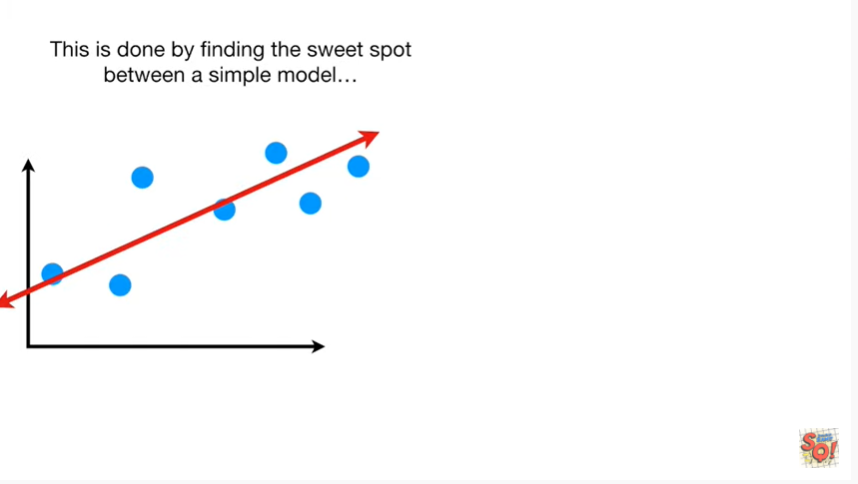
Oh No terminology alert because the squiggly line fits the training set really well but not the testing set we say that the squiggly line is over fit.



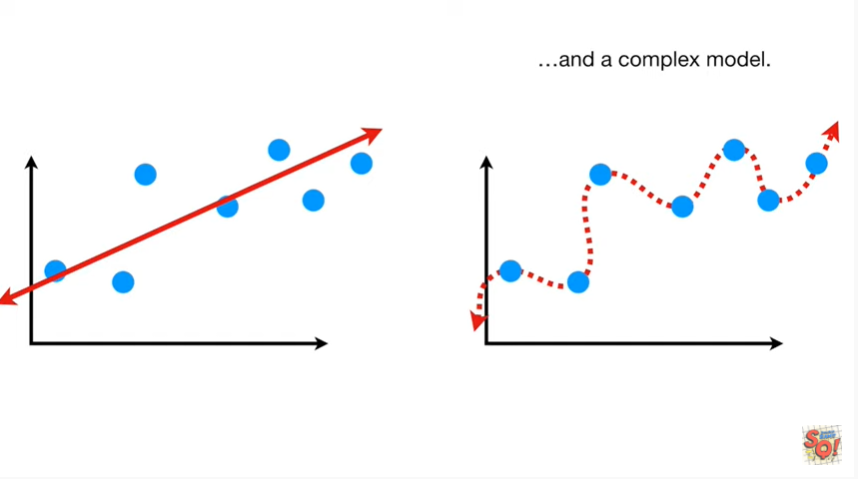
In machine learning the ideal algorithm has low bias and can accurately model the true relationship



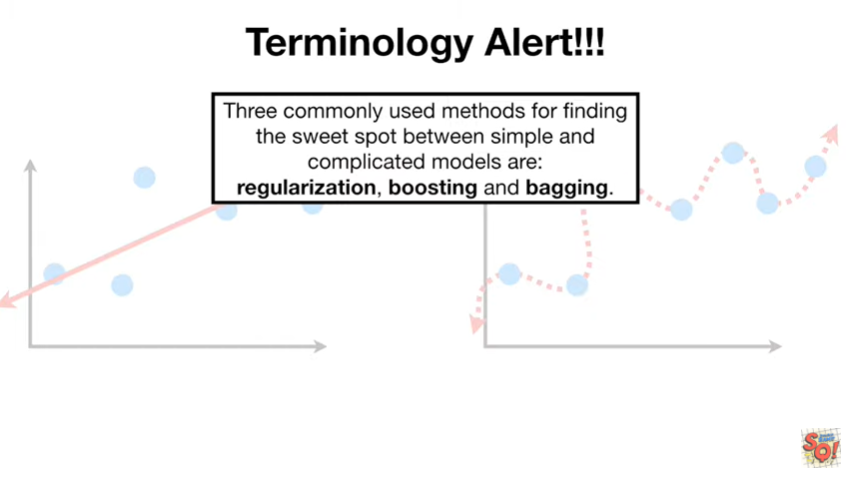
and it has low variability by producing consistent predictions across different data sets.



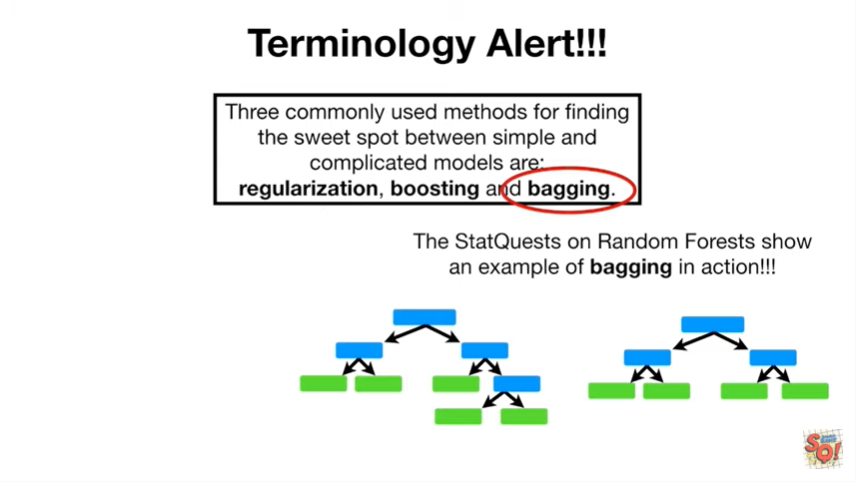
This is done by finding the sweet spot between a simple model



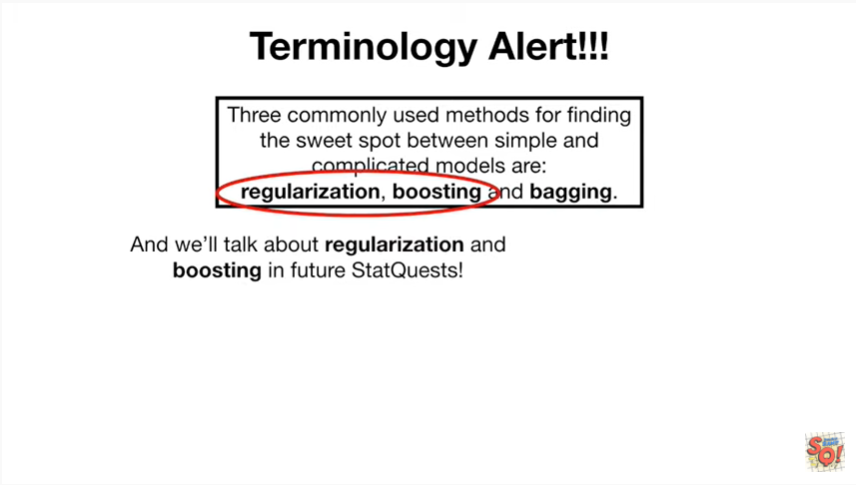
and a complex model.



Oh no another terminology alert 3 commonly used methods for finding the sweet spot between simple and complicated models are : regularization, boosting and bagging.



The stat quest on a random forest show an example of bagging in action.



And we'll talk about regularization and boosting in future stat quests.



Double bam !!!