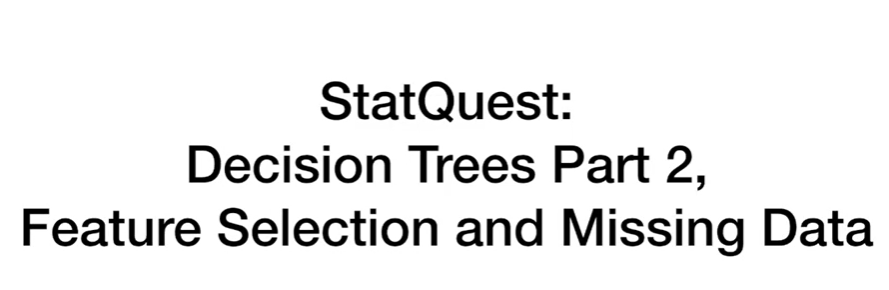
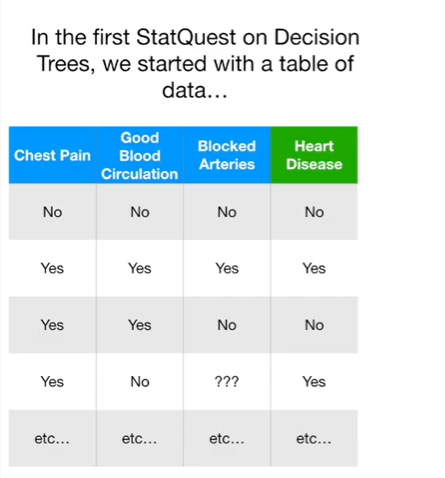
<https://www.youtube.com/watch?v=wpNl-JwwplA&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=39>

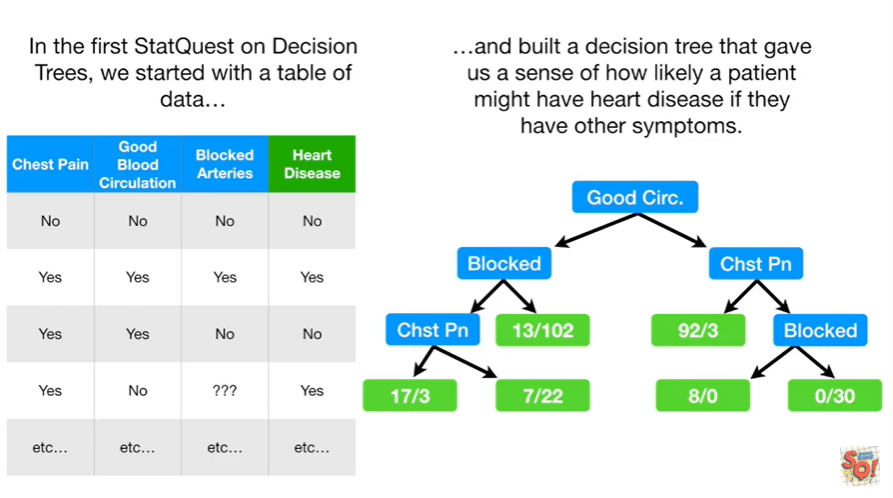


Today we're gonna be talking about decision trees part two feature selection and missing data.

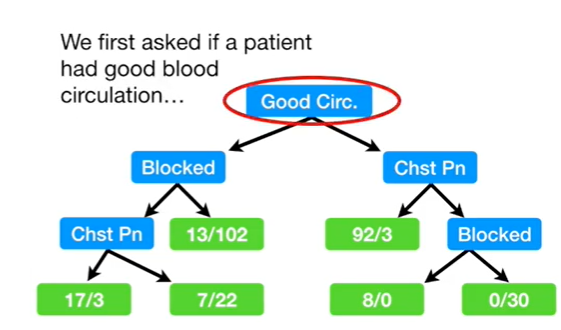
This is just a short and sweet stat quest to touch on a few topics we didn't get to in the original stat quest on decision trees.



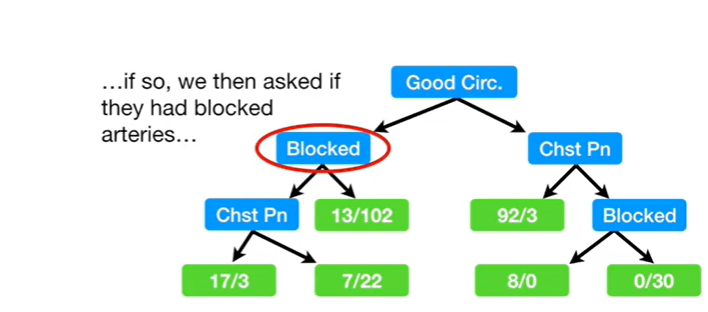
In the first stack quest on decision trees we started with a table of data



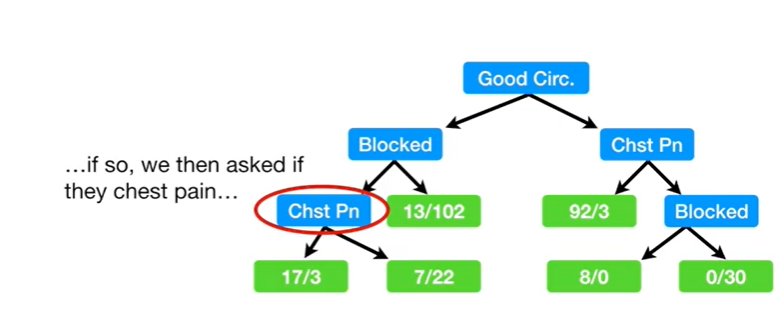
and built a decision tree that gave us a sense of how likely a patient might have heart disease if they have other symptoms.



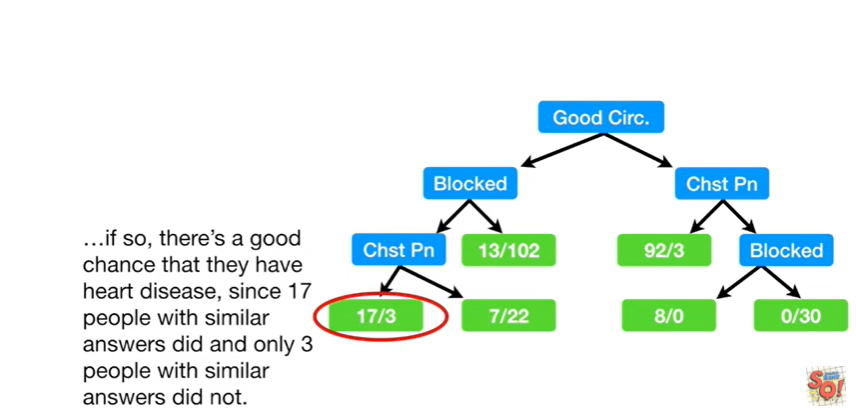
We first asked if a patient had good blood circulation



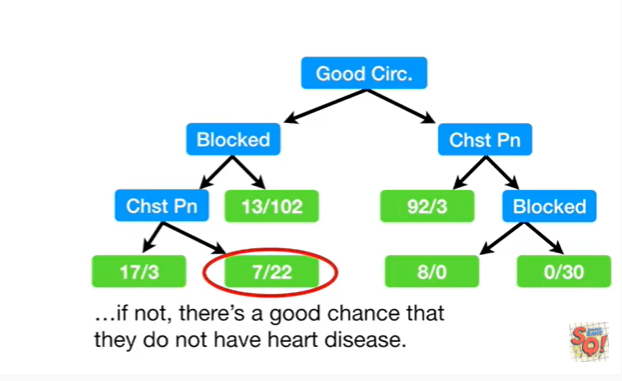
if so we then asked if they had blocked arteries



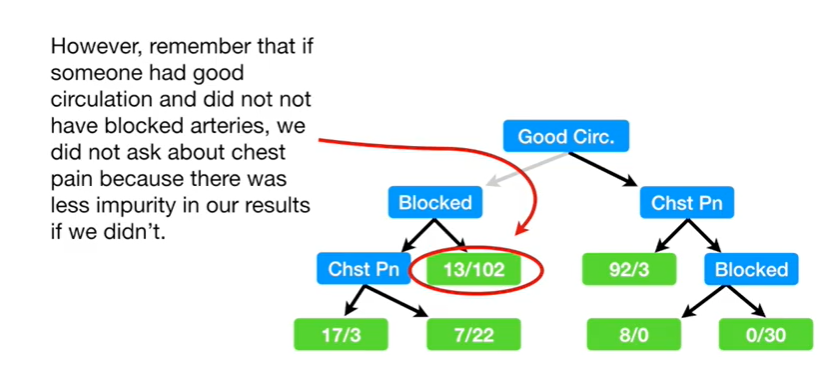
and if so we then asked if they had chest pain



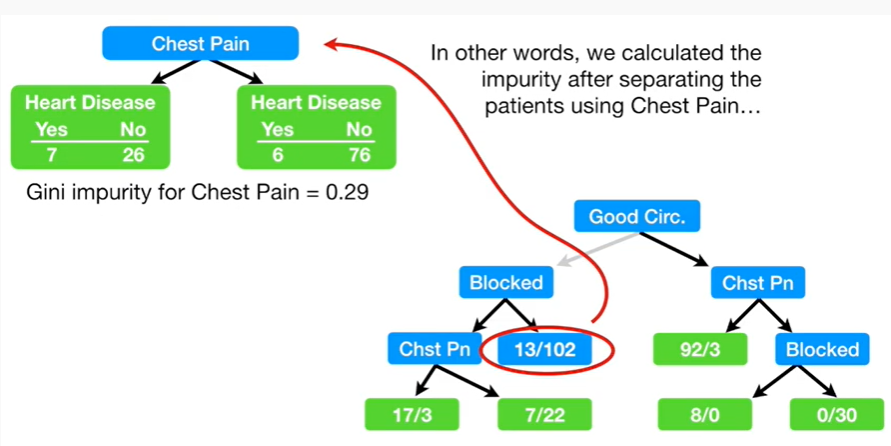
if so there's a good chance that they have heart disease since 17 people with similar answers did and only three people with similar answers did not.



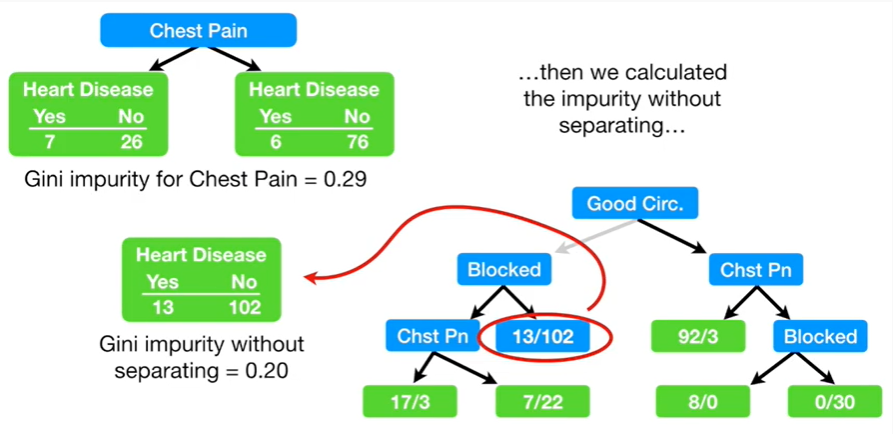
If they don't have chest pain, there's a good chance that they do not have heart disease.



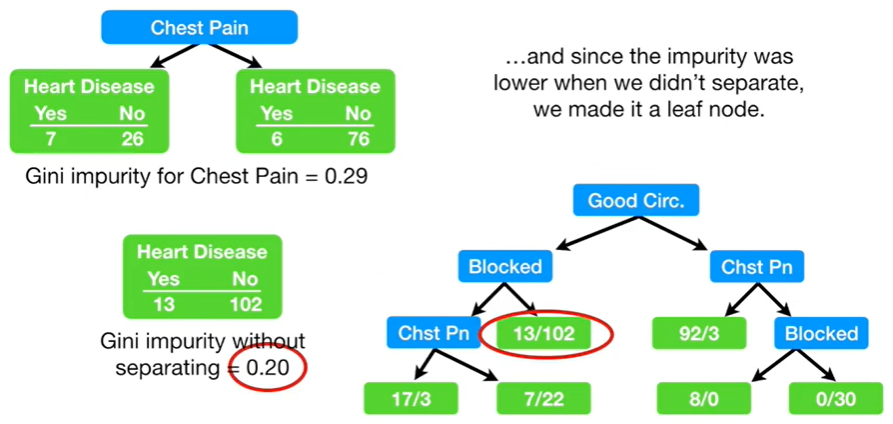
However, remember that if someone had good circulation and did not have blocked arteries we did not ask about chest pain because there was less impurity in our results if we didn't.



In other words we calculated the impurity after separating the patients using chest pain



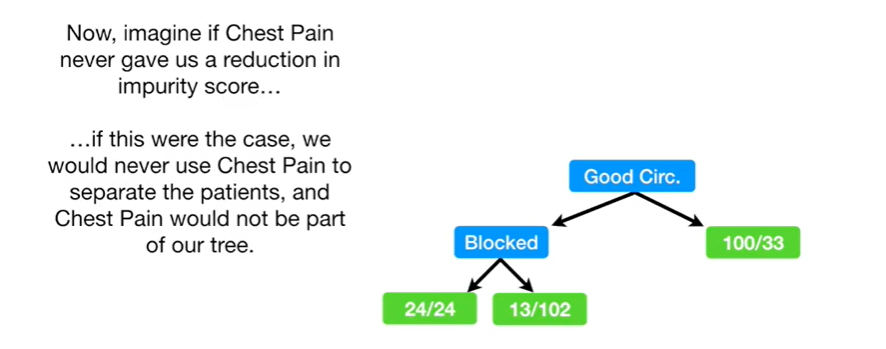
then we calculated impurity without separating



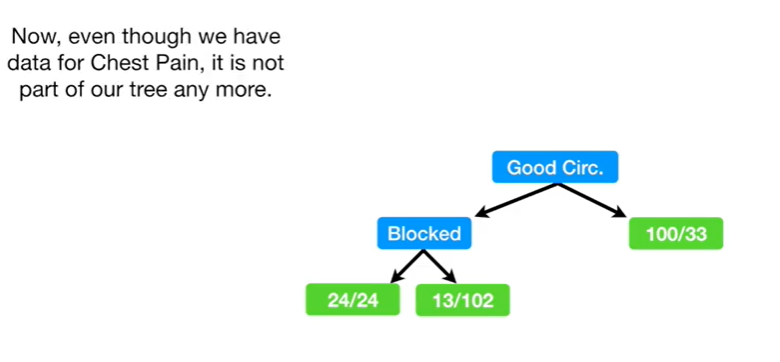
and since the impurity was lower when we didn't separate, we made it a leaf node.



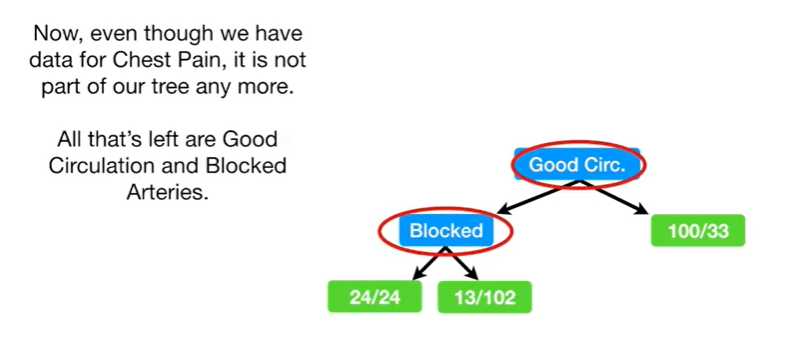
Now imagine of chest pain never gave us a reduction in impurity score



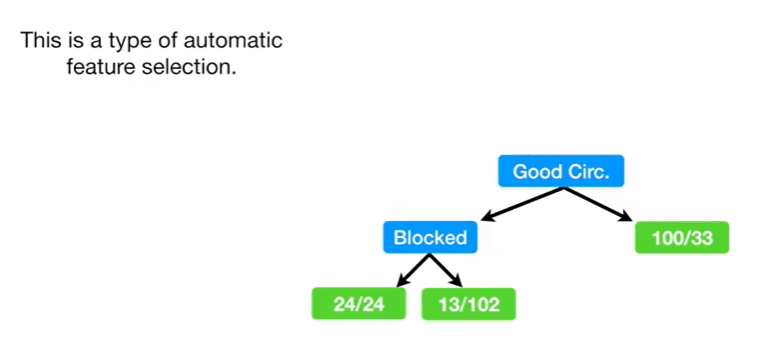
if this were the case we would never use chest pain to separate the patient,s and chest pain would not be part of our tree.



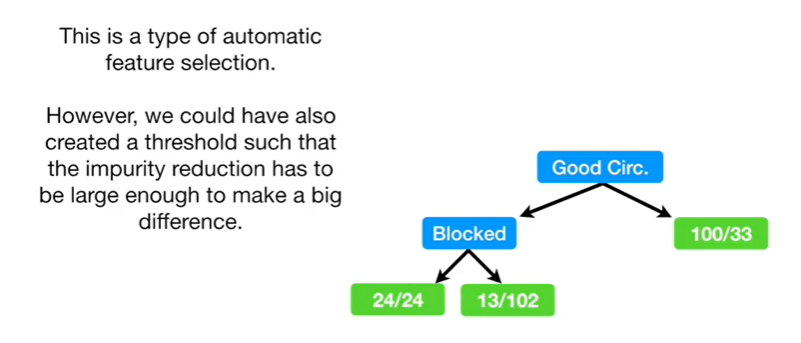
Now, even though we have data for chest pain, it is not part of our tree any more.



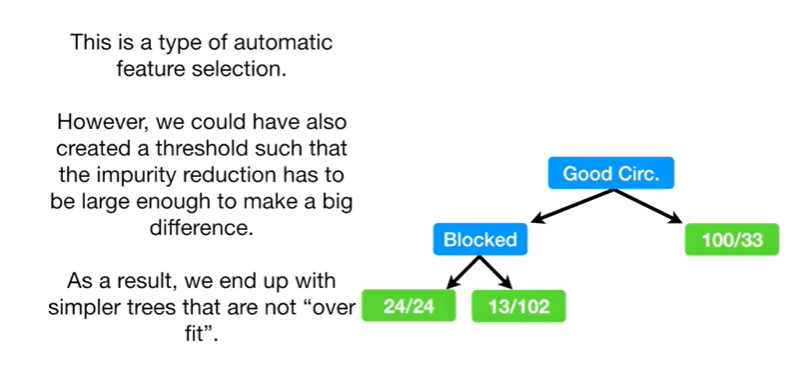
All that's left are good circulation and blocked arteries.



This is a type of automatic feature selection.



However we could also have created a threshold such that the impurity reduction has to be large enough to make a big difference.

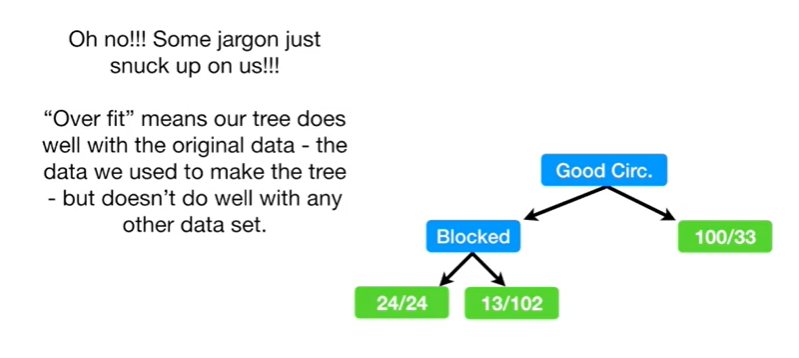


As a result we end up with simpler trees that are not over fit.

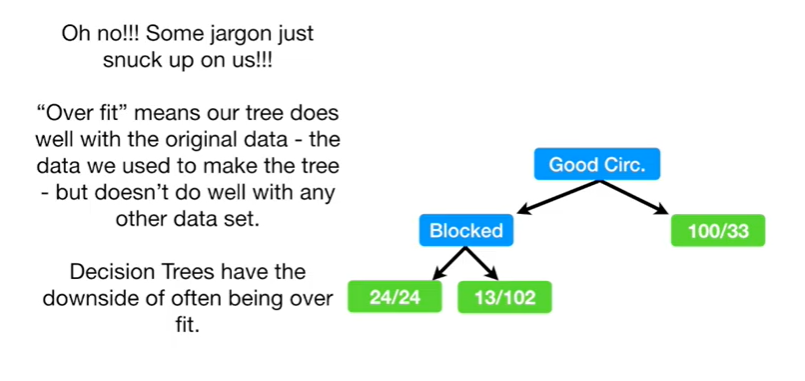


Oh no !!!

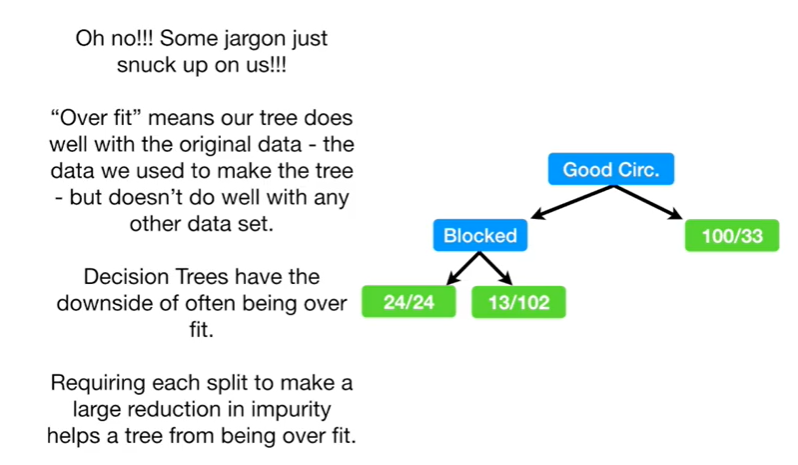
Some jargon just snuck up on us !!!



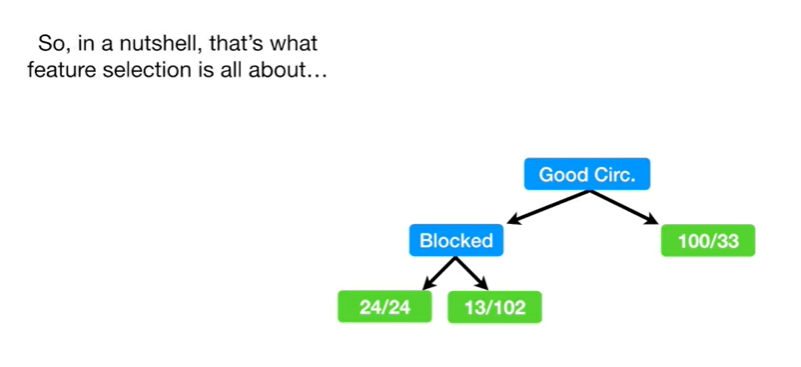
Overfit means our tree does well with the original data - the data we use to make the tree - but doesn't do well with any other data set.



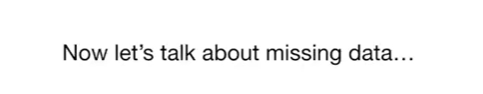
Decision trees have the downside of often being overfit.



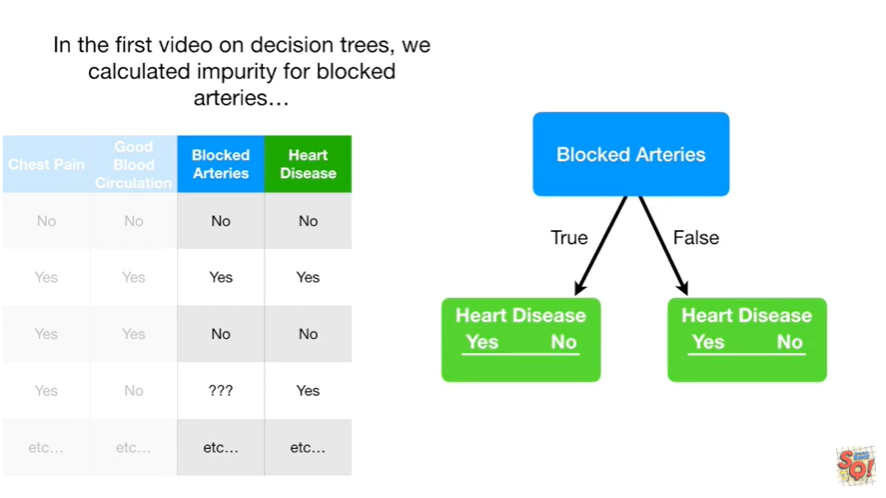
Requiring each split to make a large reduction in impurity helps a tree from being over fit.



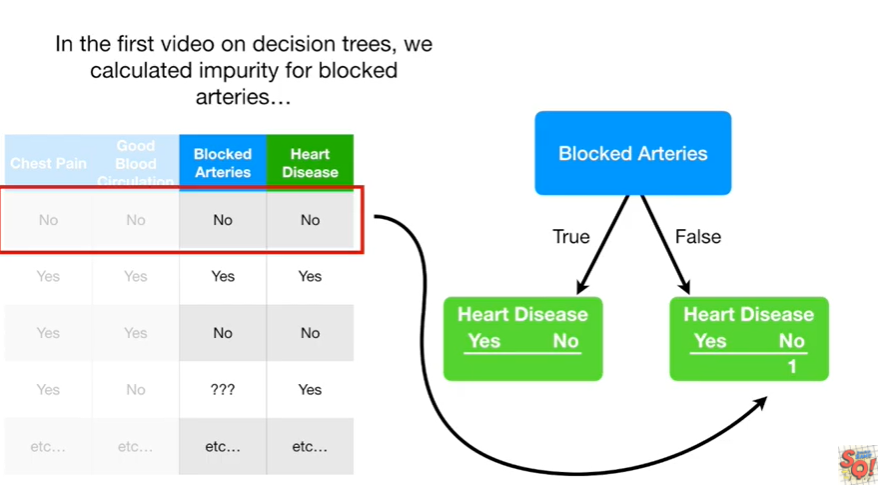
So in a nutshell, that's what feature selection is all about.

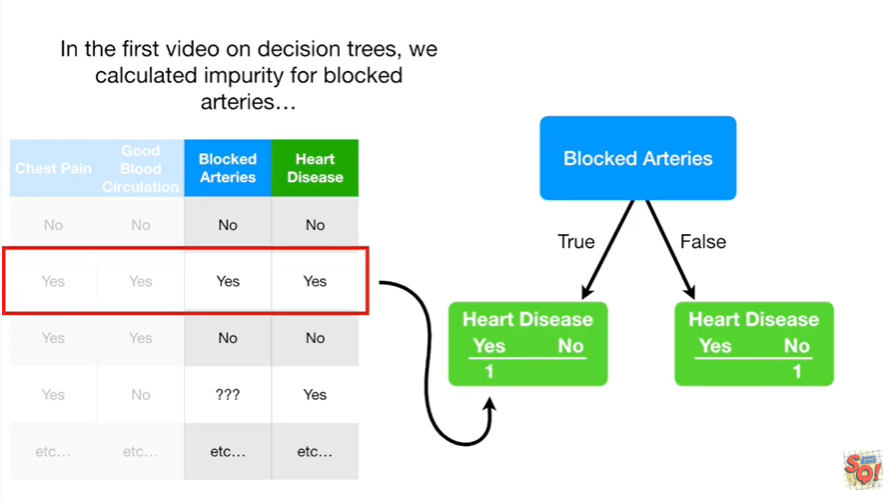


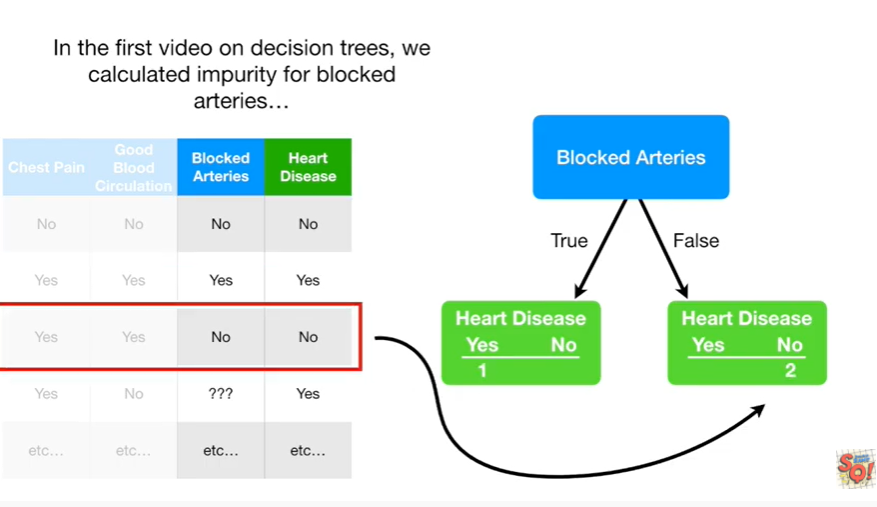
Now let's talk about missing data.



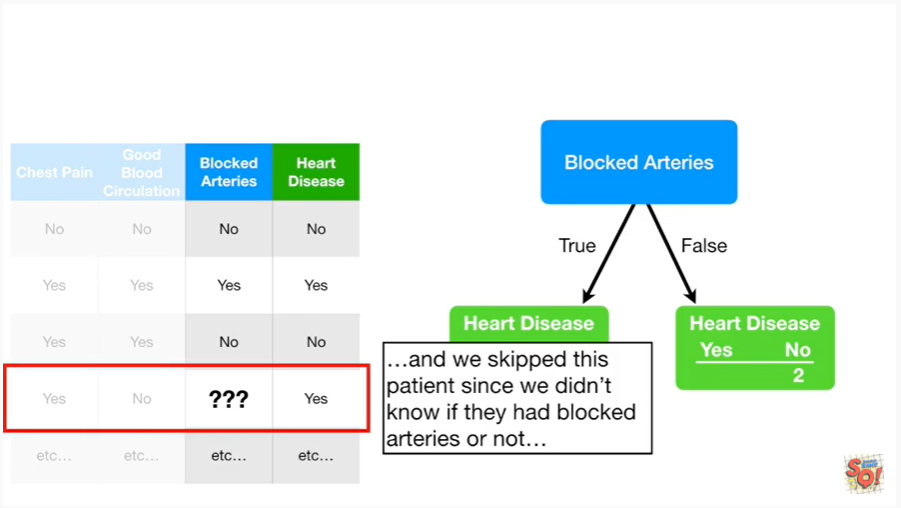
In the first video on decision trees, we calculated impurity for blocked arteries



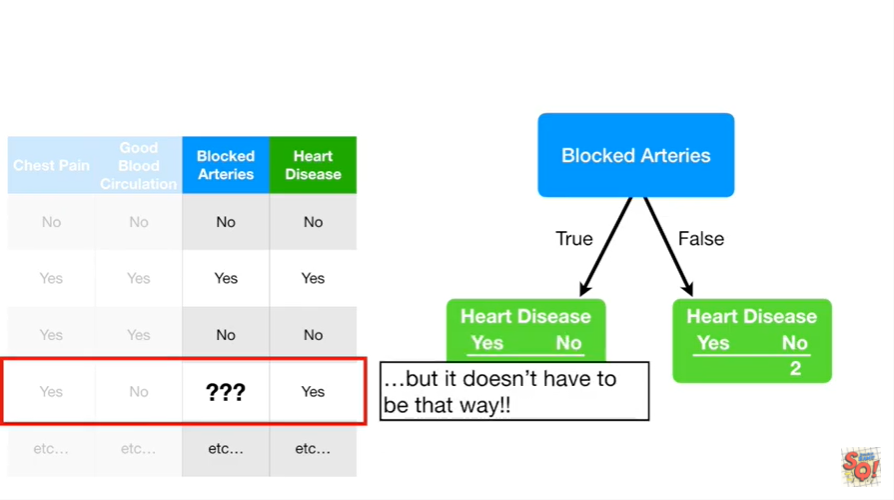




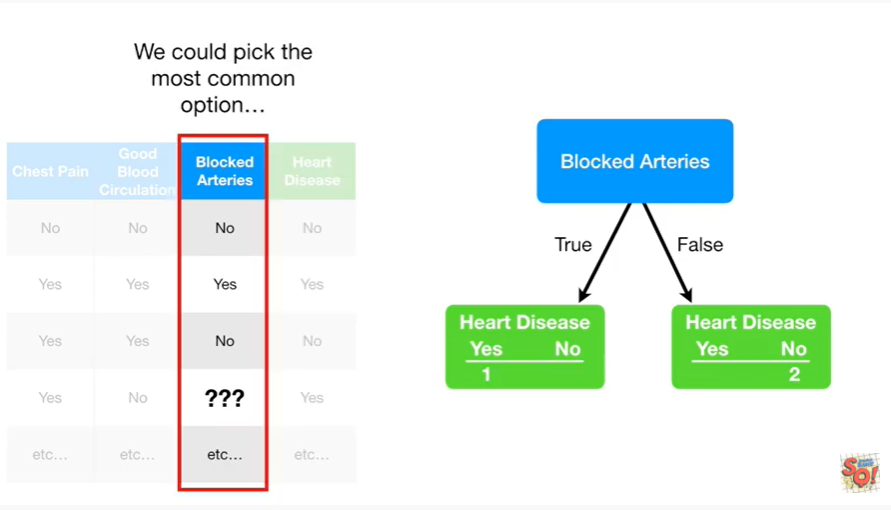
boota boota boota boota boota boota boo



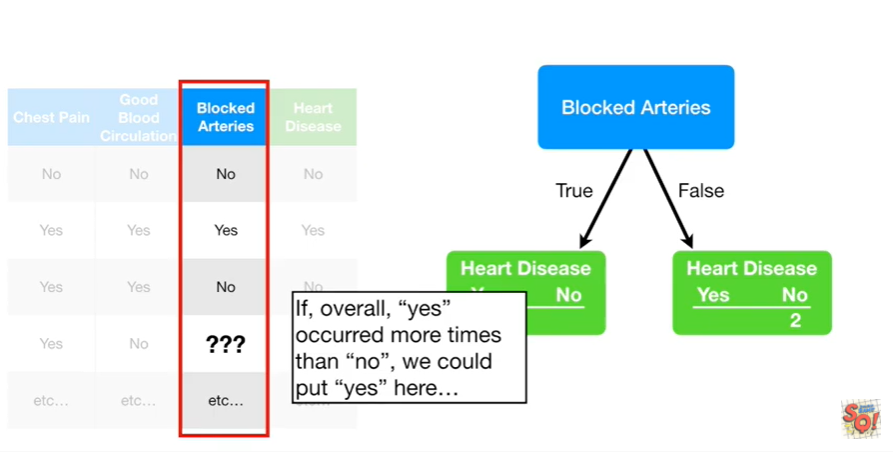
and we skipped this patient since we didn't know if they had blocked arteries or not



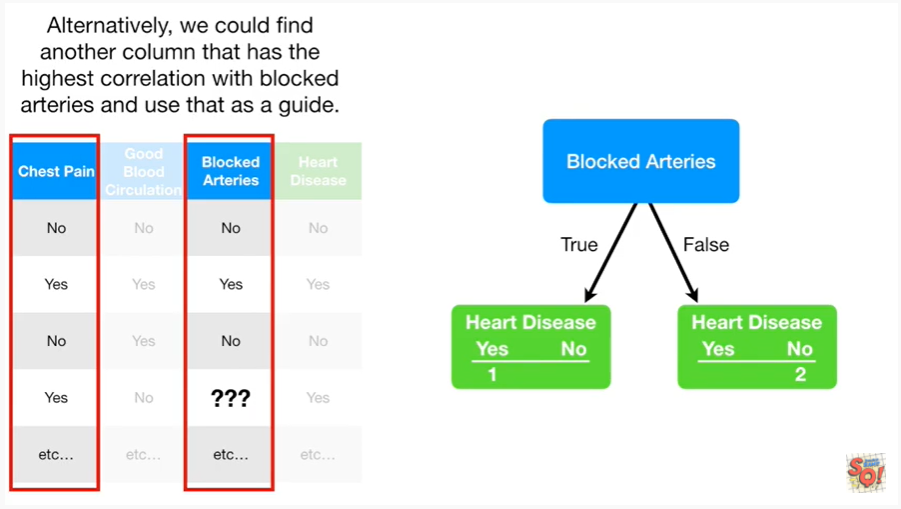
but it doesn't have to be that way.



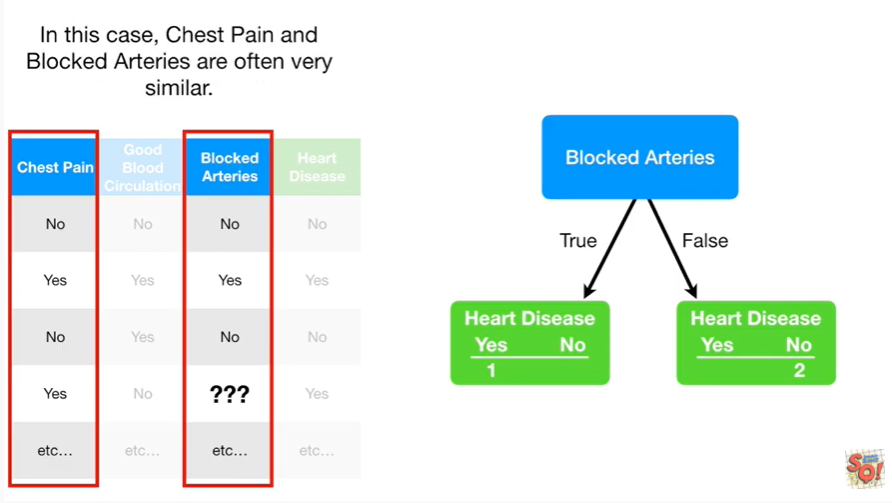
We could pick the most common option



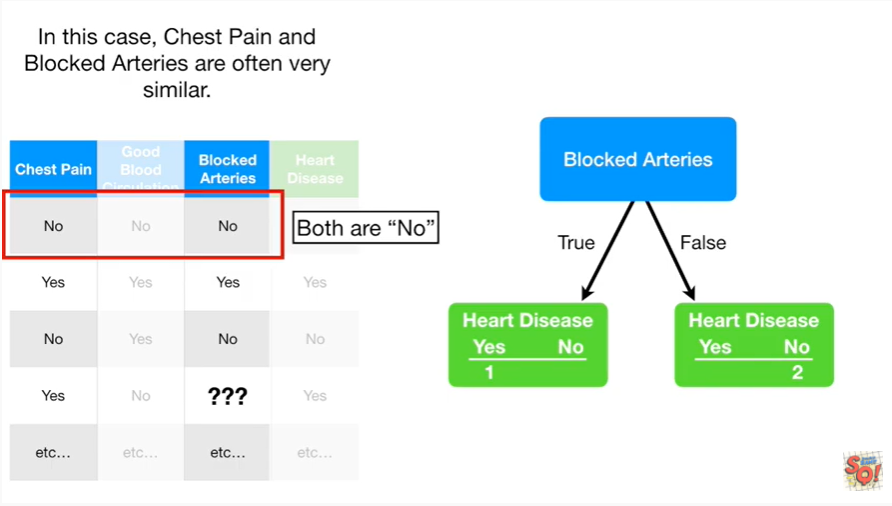
If, overall, yes occurred more times than no, we could put yes here.



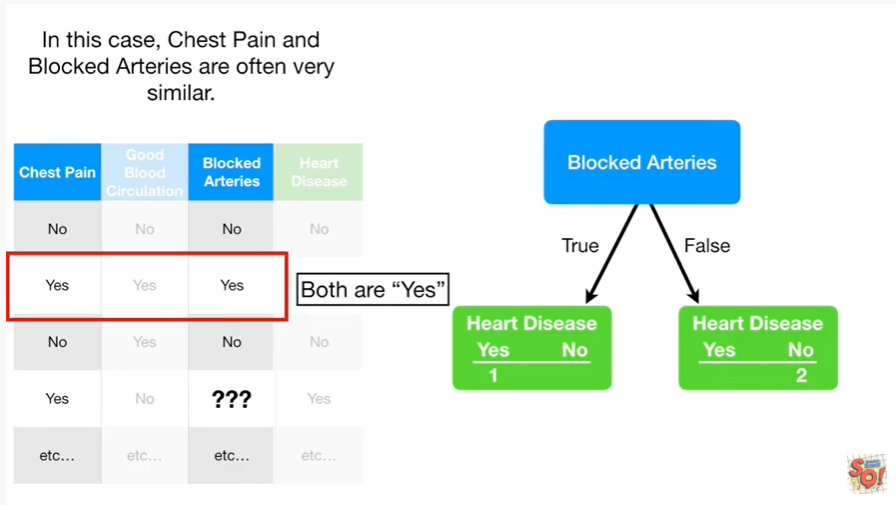
Alternatively we could find another column that has the highest correlation with blocked arteries and use that as a guide.



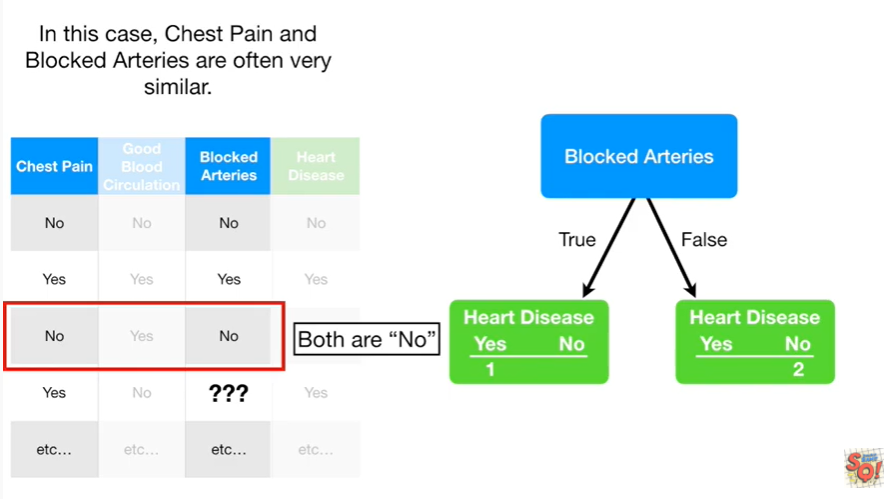
In this case chest pain and blocked arteries are often very similar.



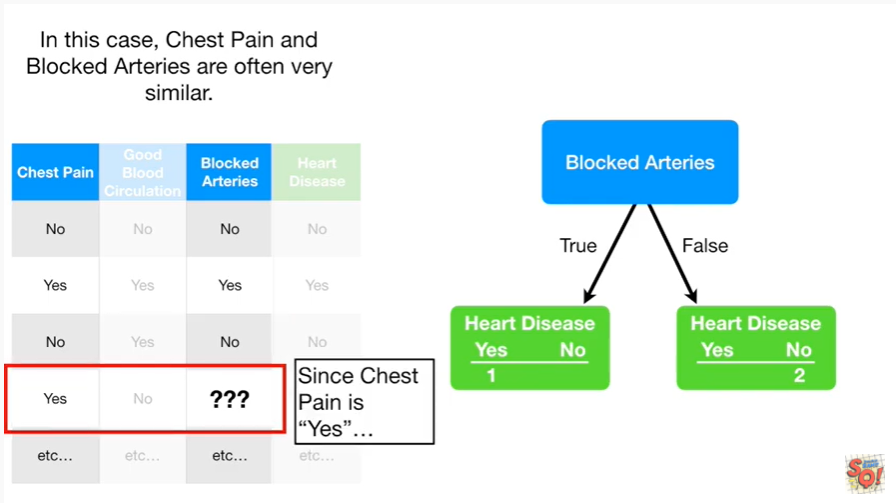
The first patient has no in both categories.



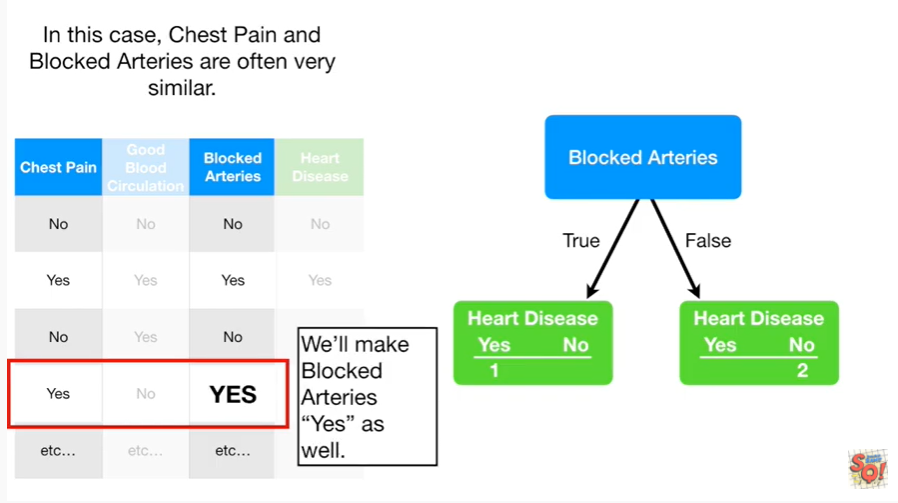
The second patient has yes in both categories.



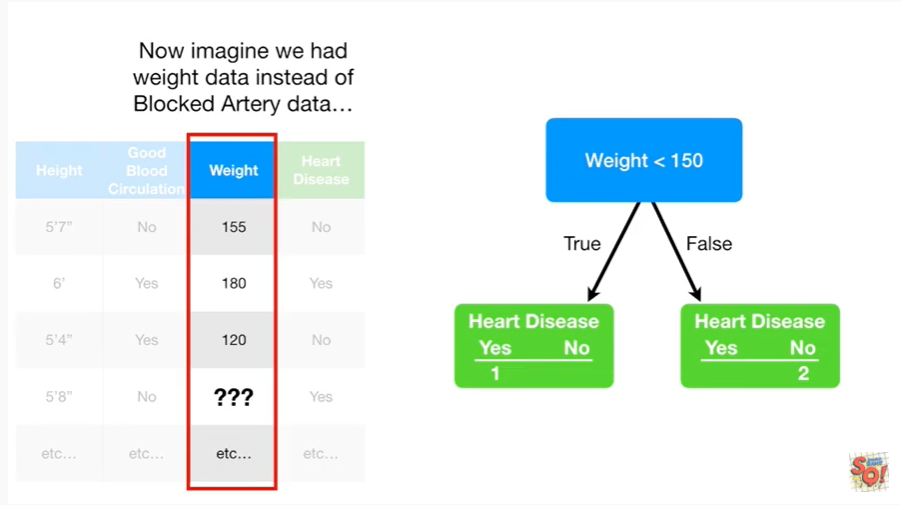
The third patient has no in both categories.



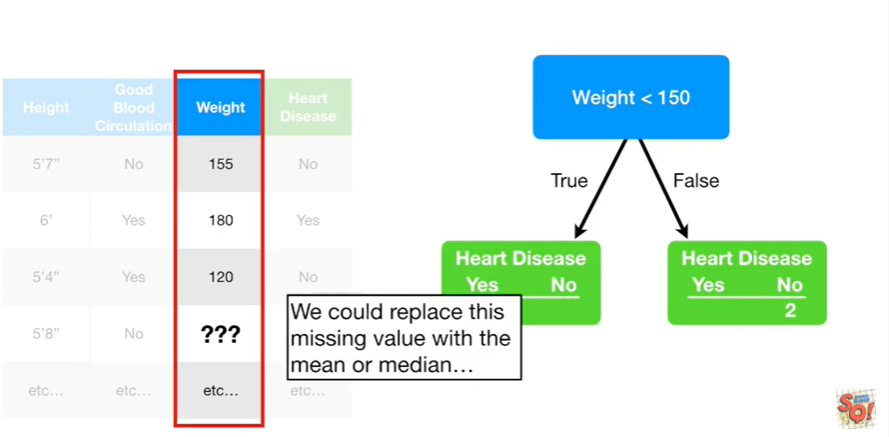
And so for the fourth patient since chest pain is yes



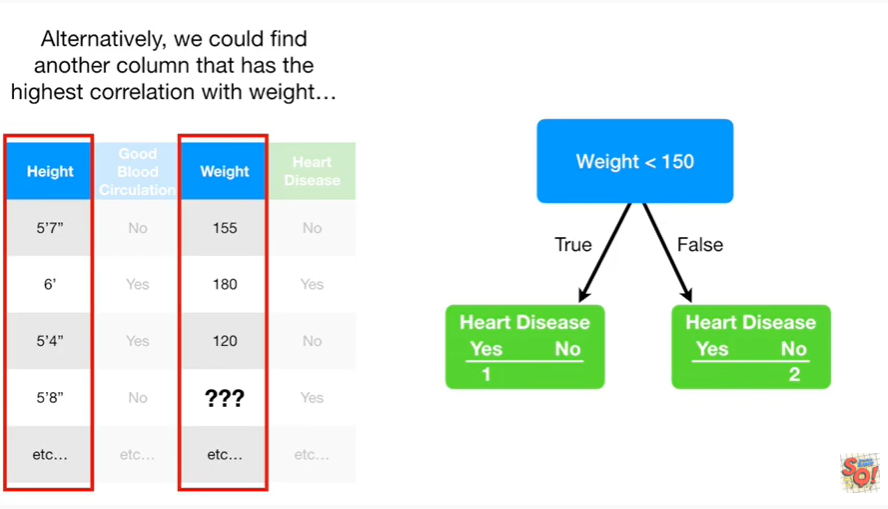
We'll make blocked arteries yes as well.



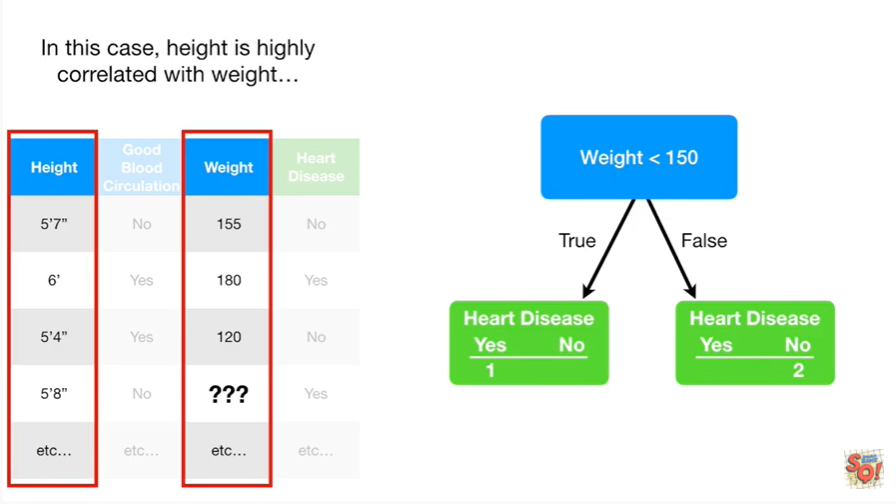
Now imagine we had weight data instead of blocked artery data.



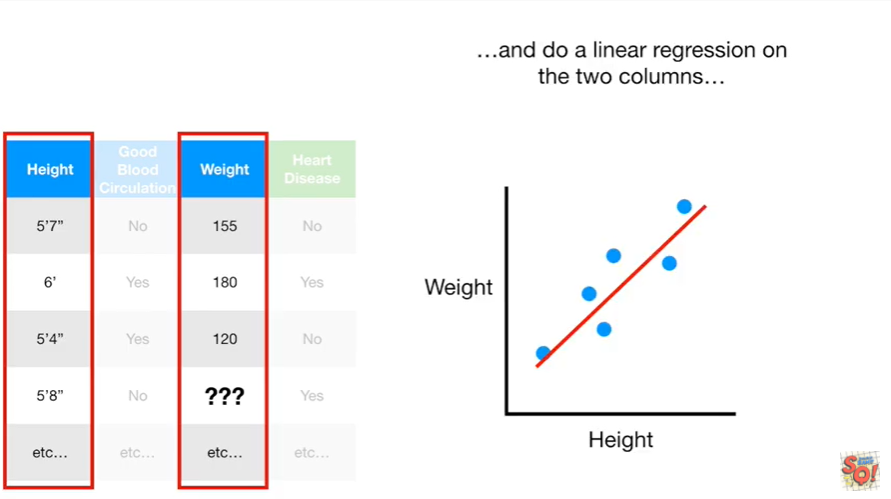
We could replace this missing value with the mean or the median.



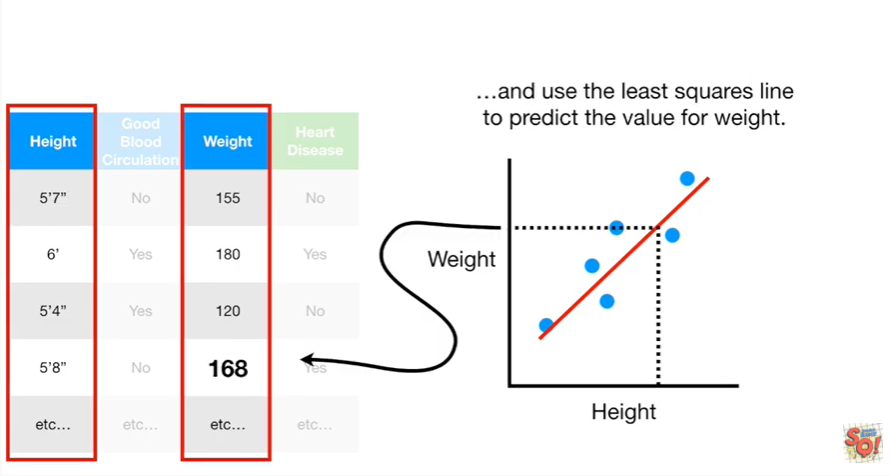
Alternatively we could find another column that has the highest correlation with weight.



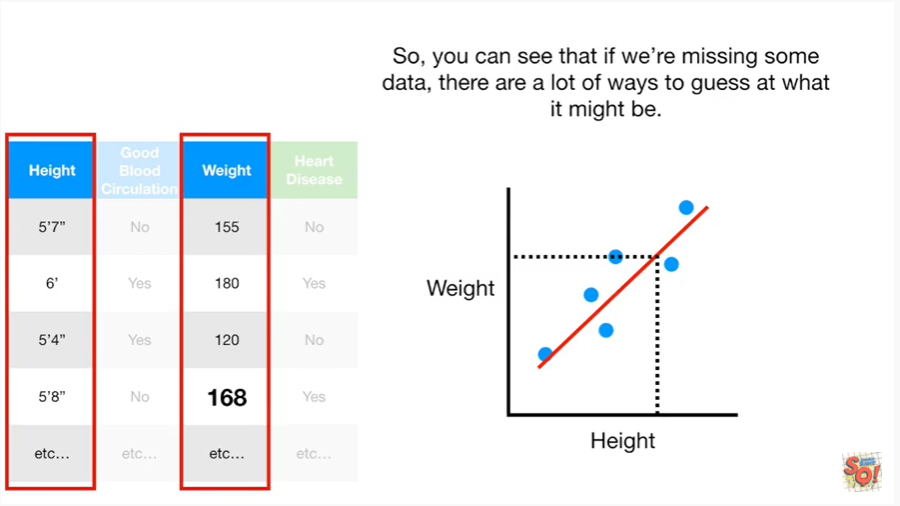
In this case height is highly correlated with weight



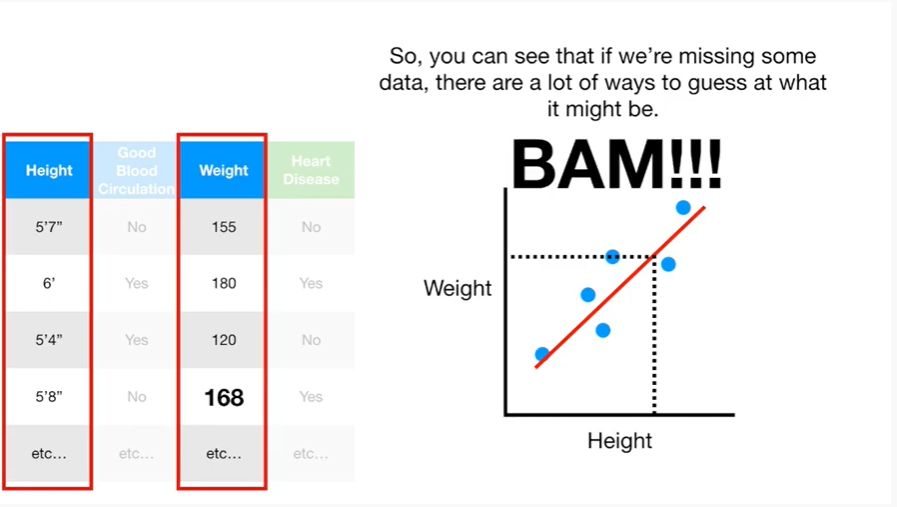
and then we can do a linear regression on the two columns



and use the least squares line to predict the value for weigh.



So, you can see that if we're missing some data there are a lot of ways to guess at what it might be.



BAM !!!