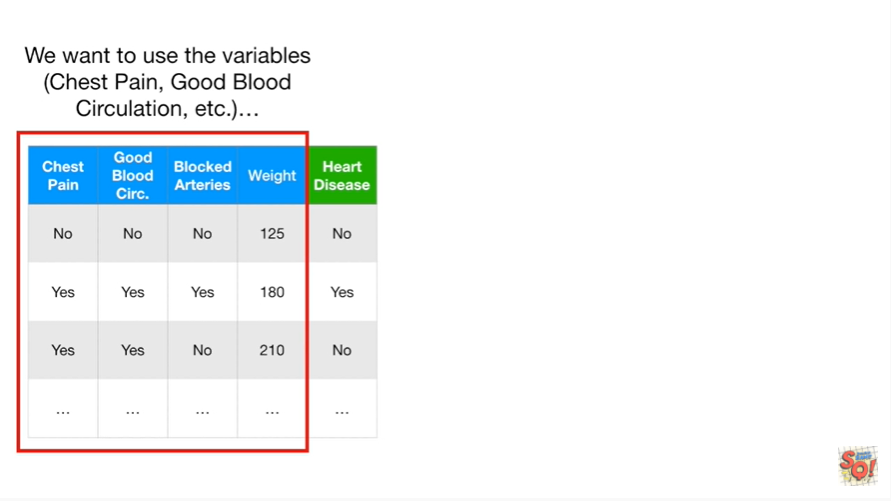
<https://www.youtube.com/watch?v=fSytzGwwBVw&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=2>



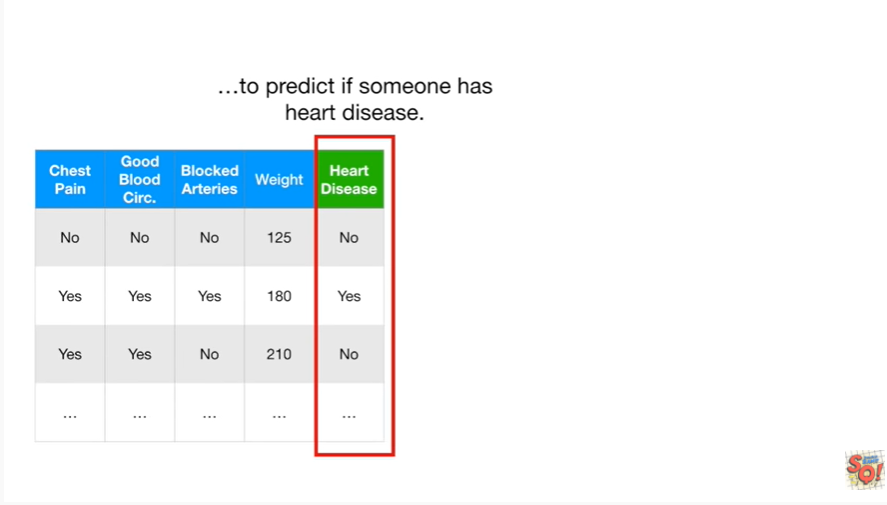
Today we're going to talk about cross validation and it's gonna be clearly explained.



Okay let's start with some data.



We want to use the variables chest pain good blood circulation etc



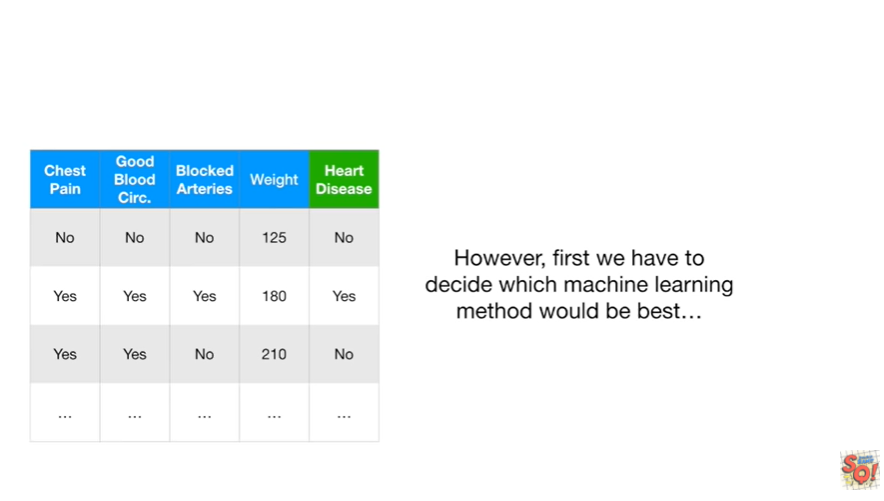
to predict if someone has heart disease.



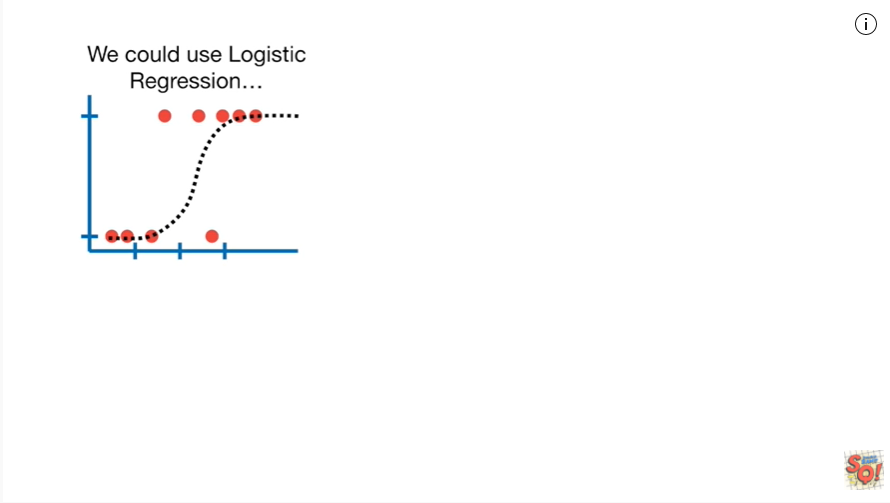
Then when a new patient shows up we can measure these variables



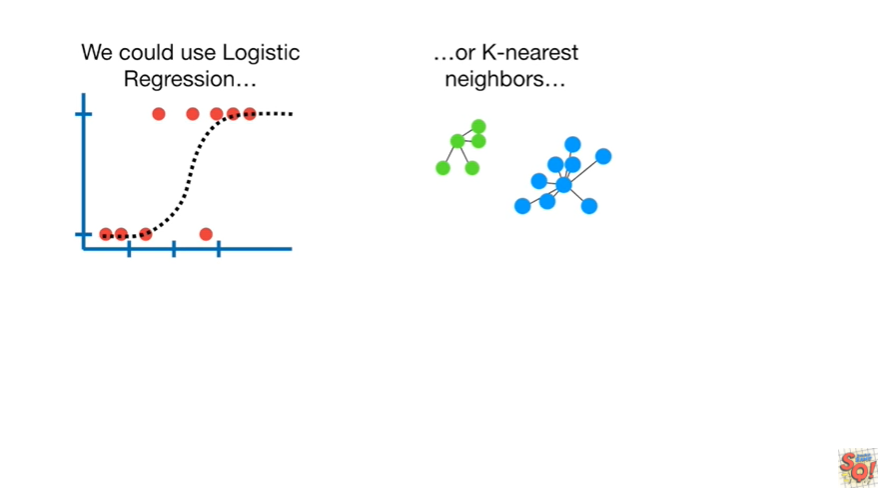
and predict if they have heart disease or not.



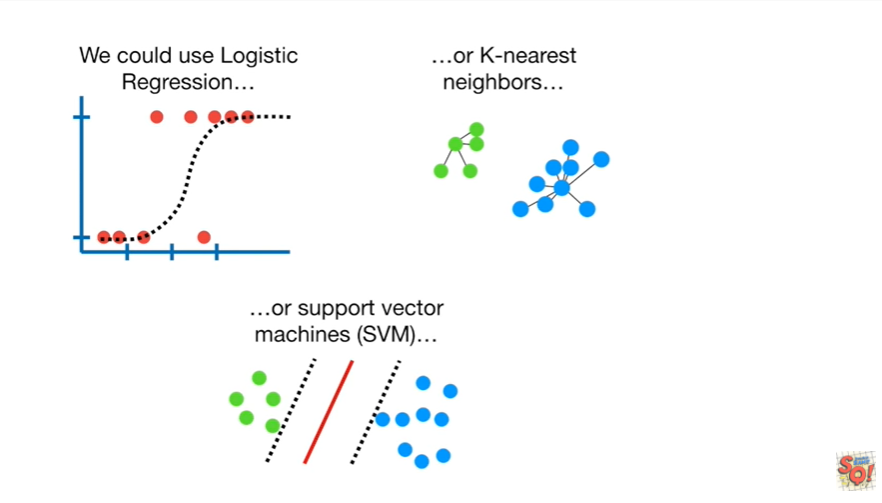
However first we have to decide which machine learning method would be best.



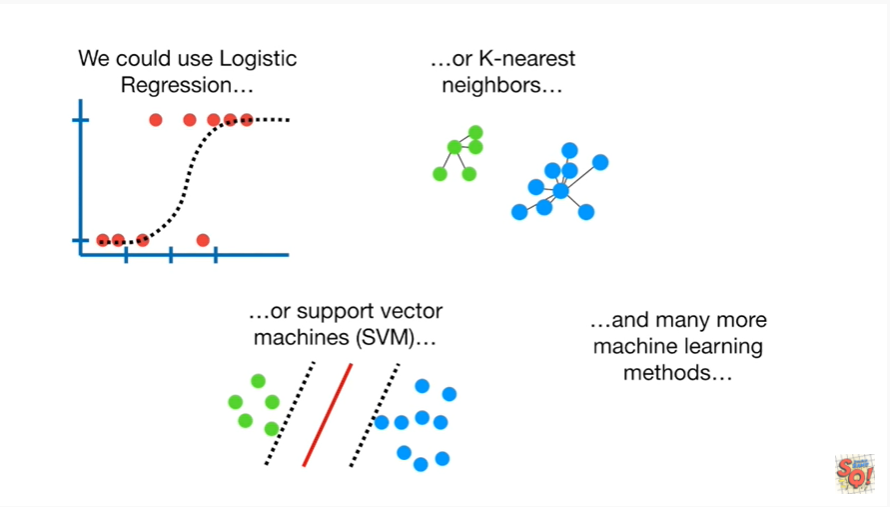
We could use logistic regression



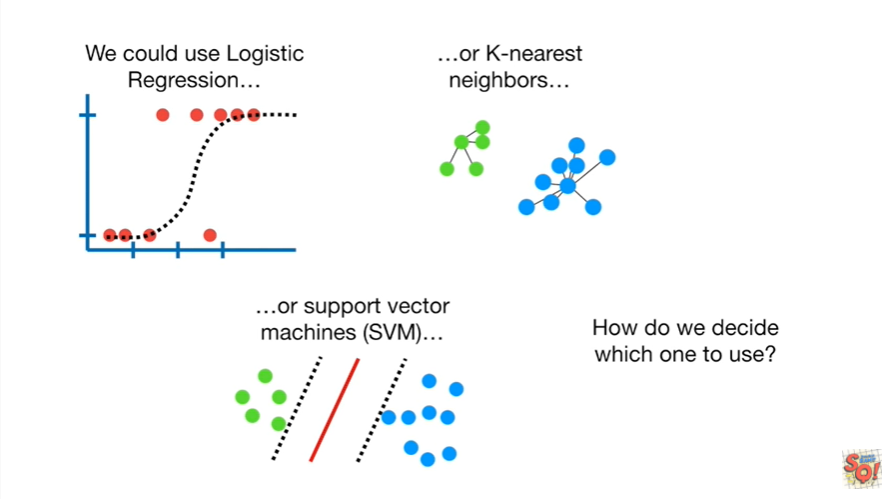
or K nearest neighbors



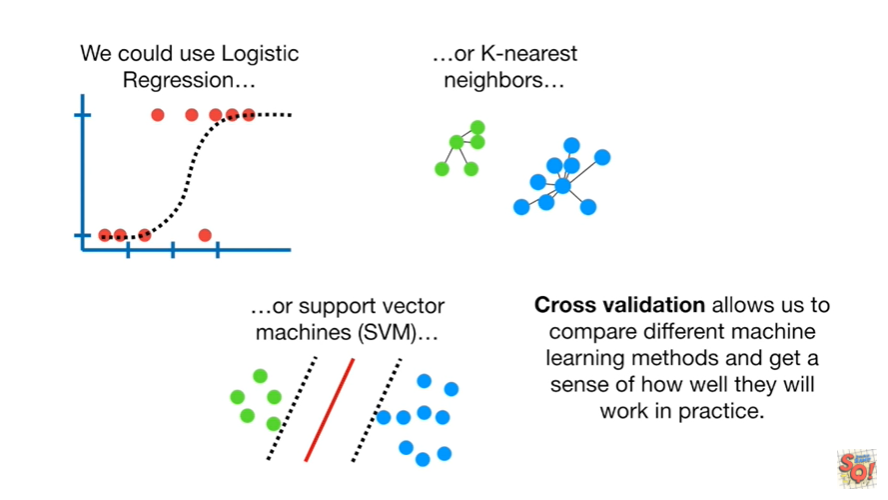
or support vector machines



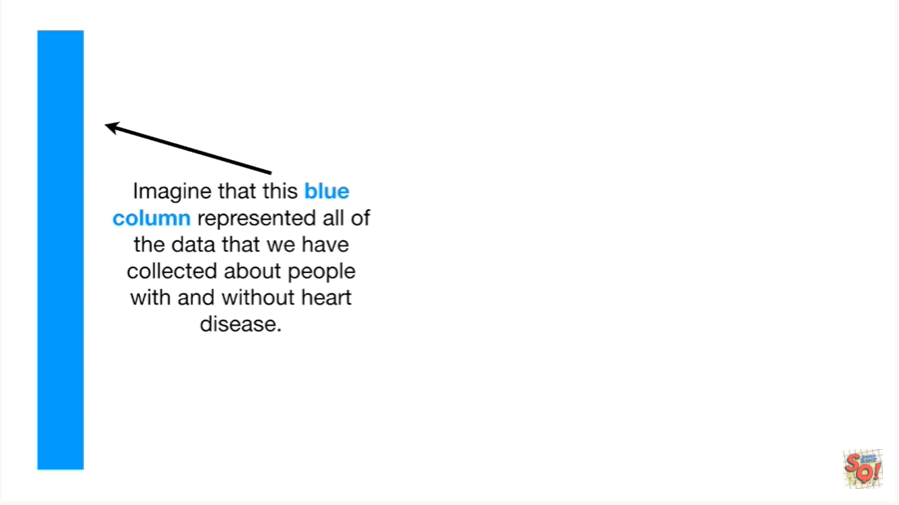
and many more machine learning methods.



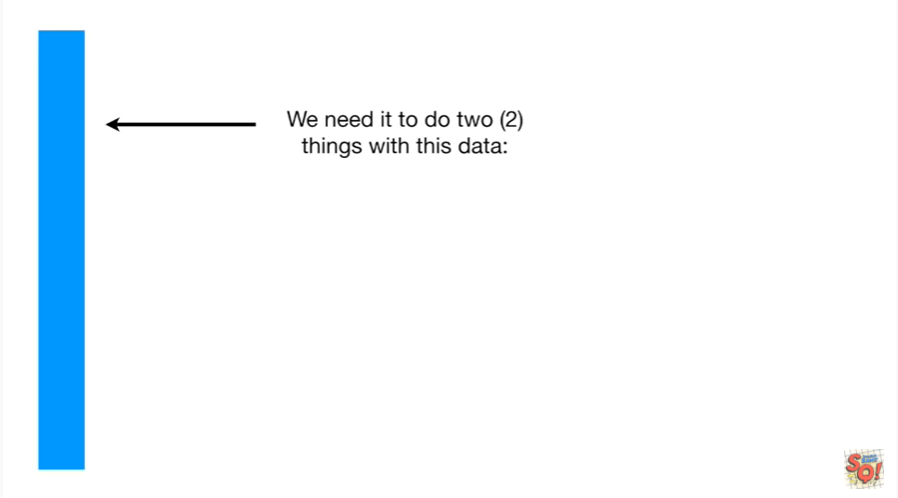
How do we decide which one to use ?



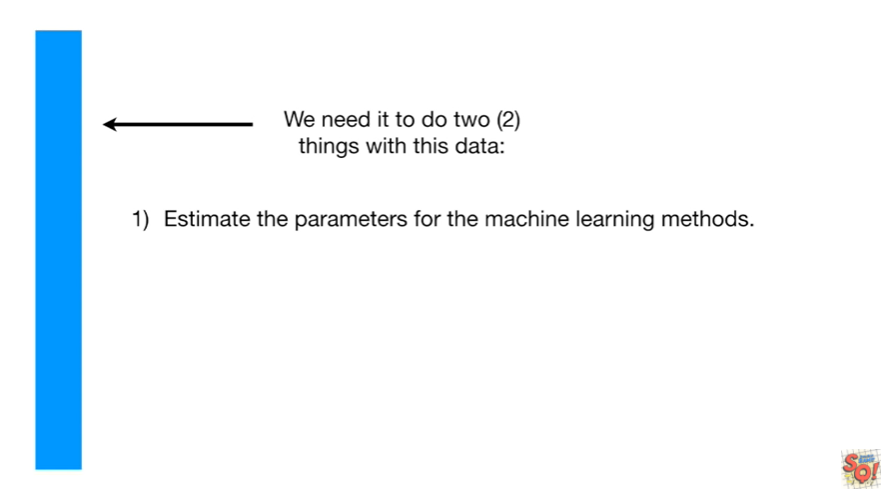
Cross-validation allows us to compare different machine learning methods and get a sense of how well they will work in practice.



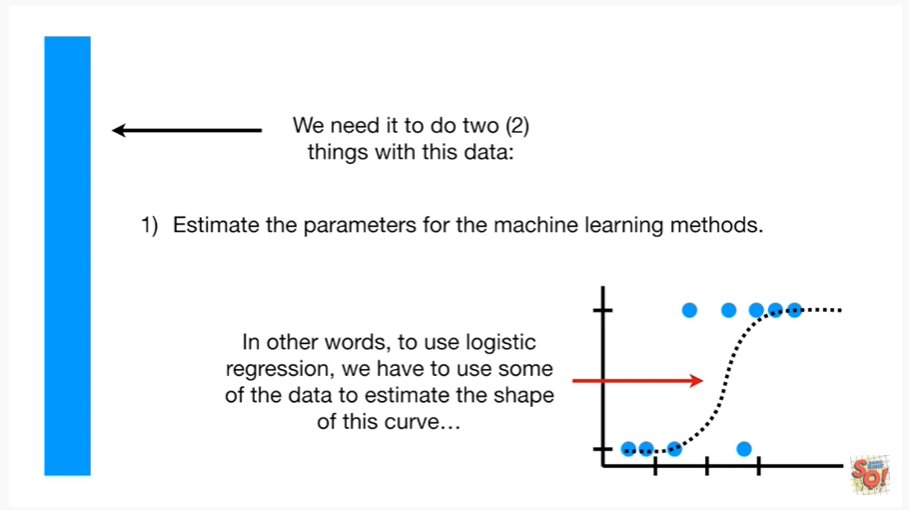
Imagine that this blue column represented all of the data that we have collected about people with and without heart disease.



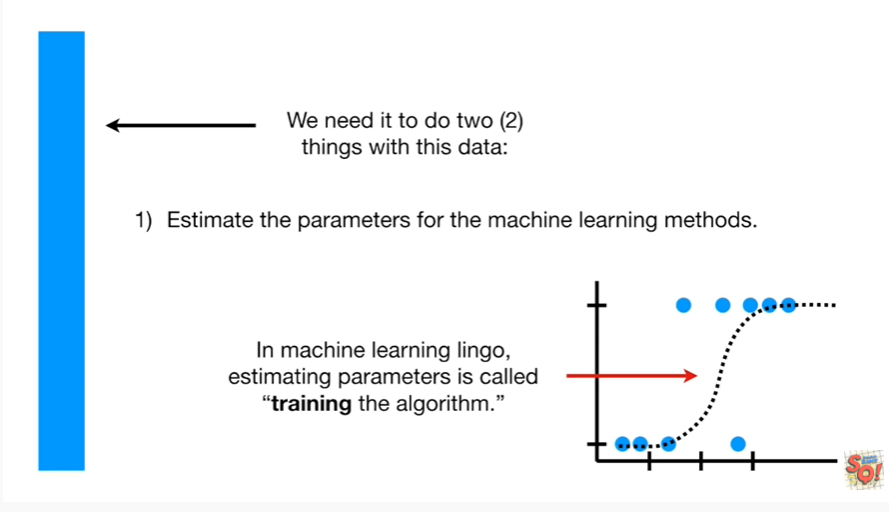
We need to do two things with this data :



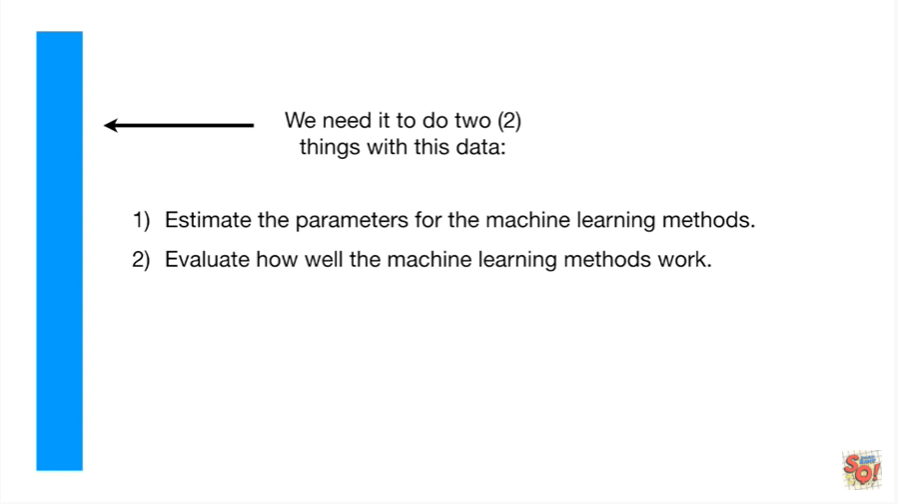
one we need to estimate the parameters for the machine learning methods.



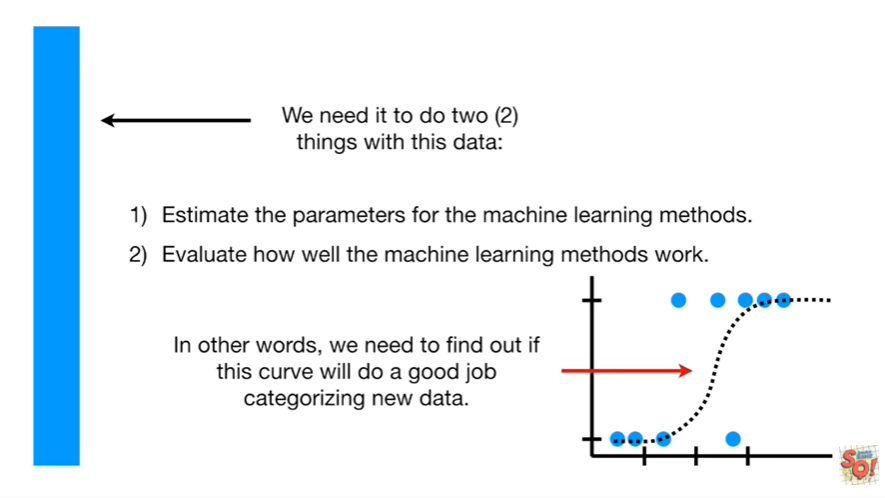
In other words to use logistic regression we have to use some of the data to estimate the shape of this curve.



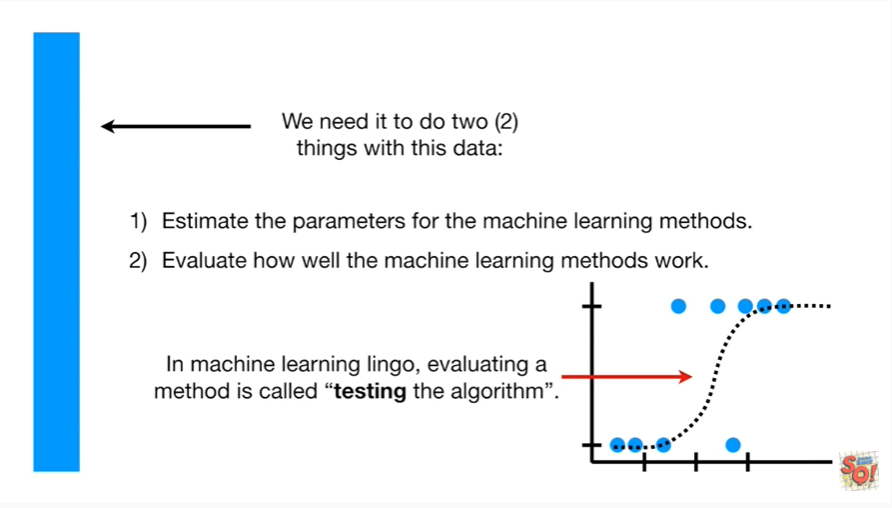
In machine learning lingo estimating parameters is called training the algorithm.



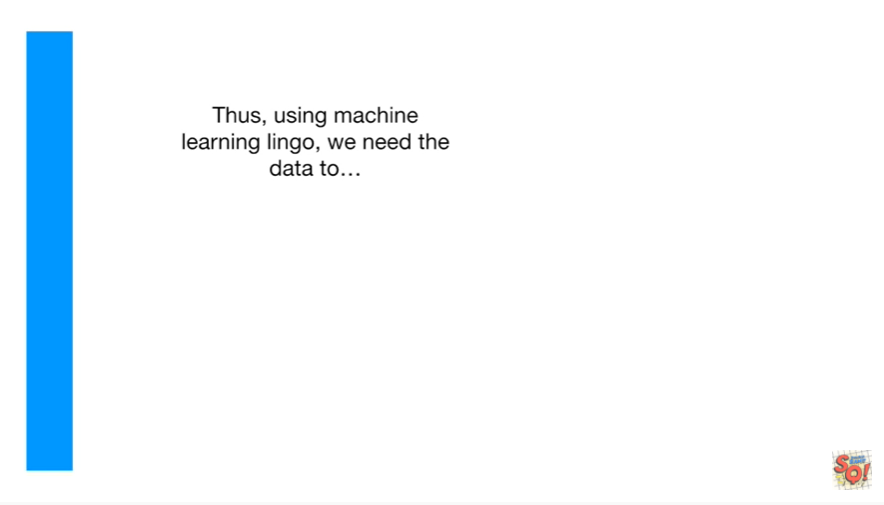
The second thing we need to do with this data is evaluate how well the machine learning methods work.



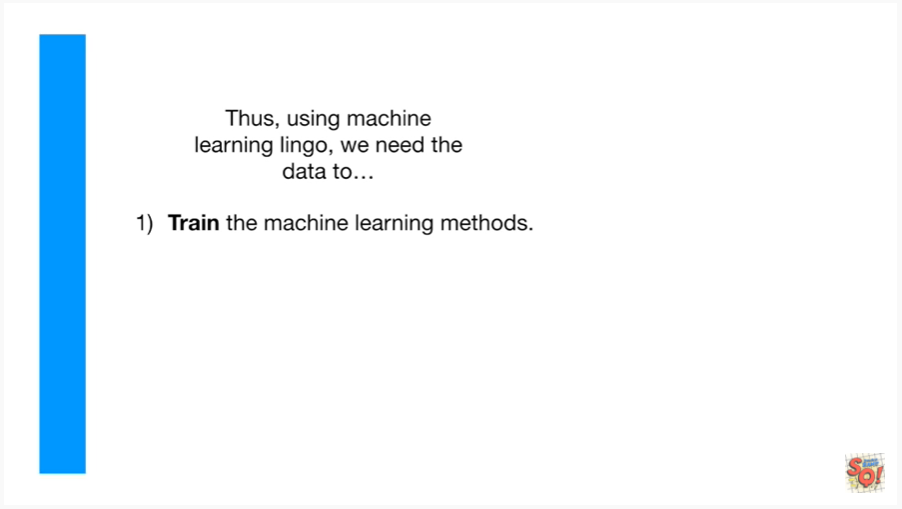
In other words we need to find out if this curve will do a good job categorizing new data.



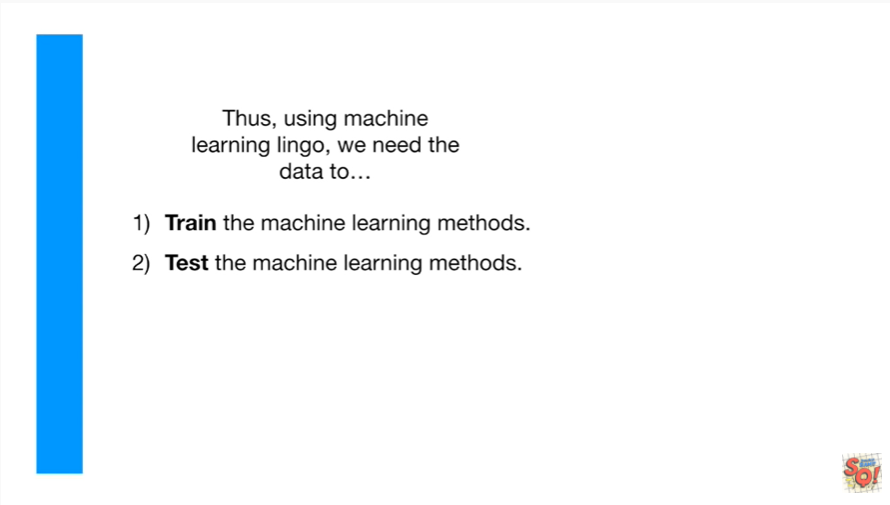
In machine learning lingo evaluating a method is called testing the algorithm.



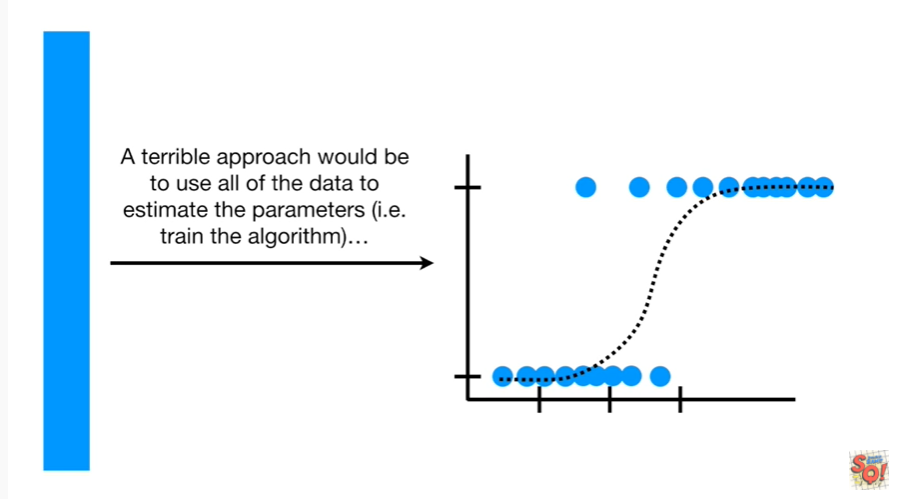
Thus using machine learning lingo we need the data to :



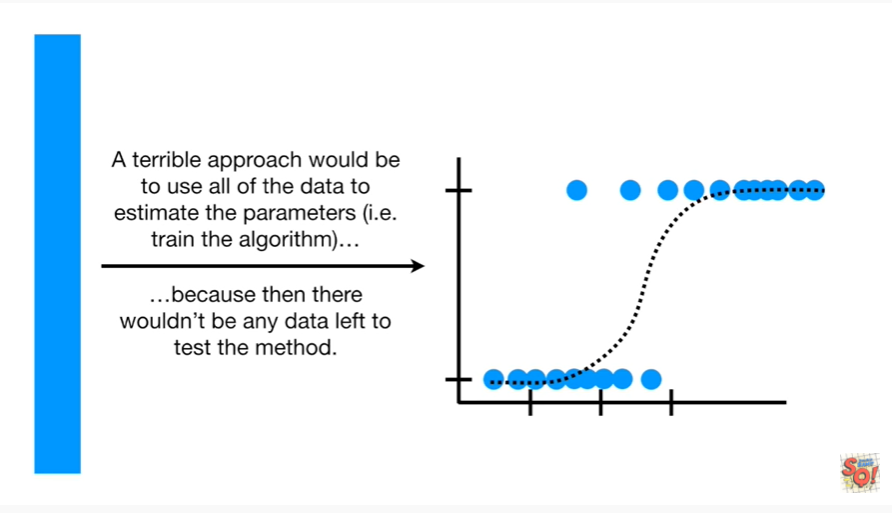
one train the machine learning methods



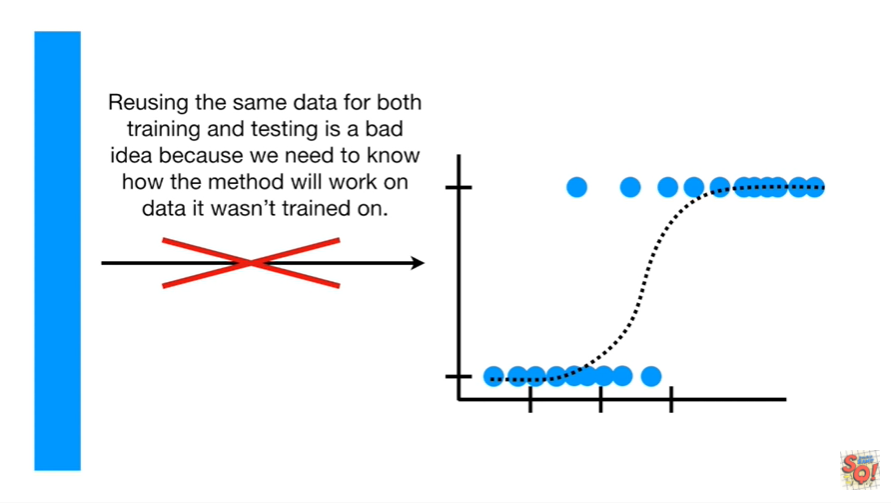
and to test the machine learning methods.



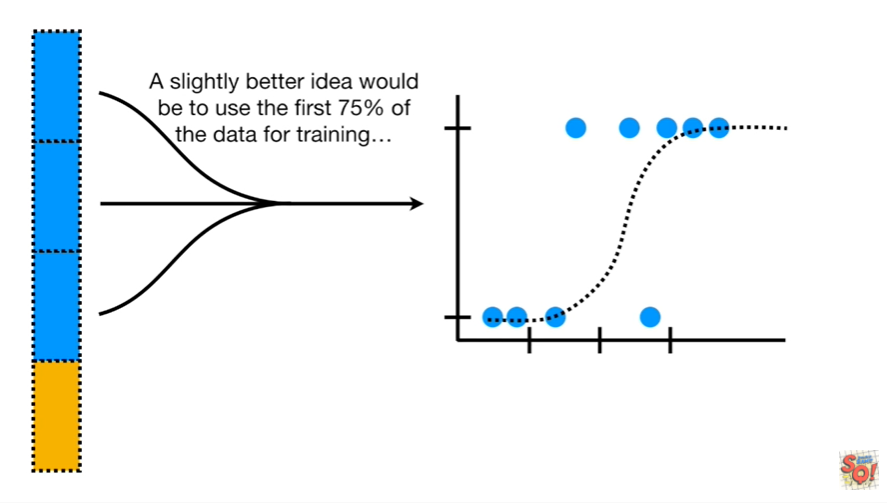
A terrible approach would be to use all the data to estimate the parameters ie to train the algorithm



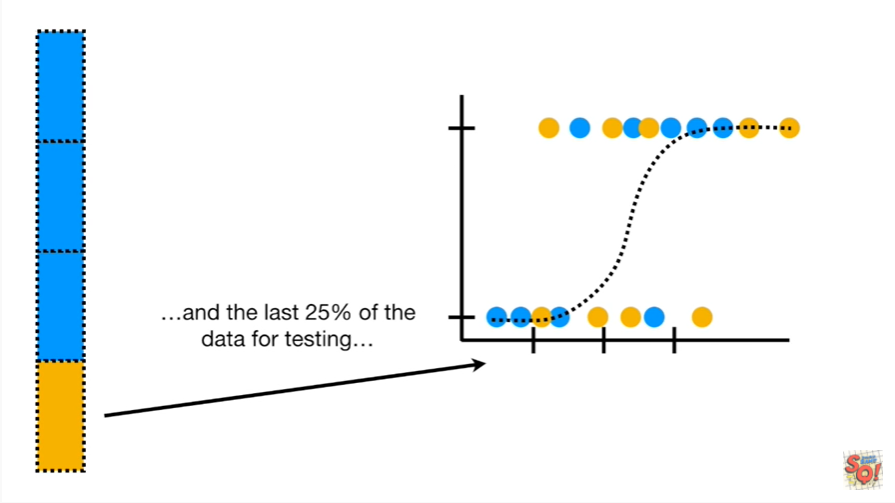
because then we wouldn't have any data left to test the method.



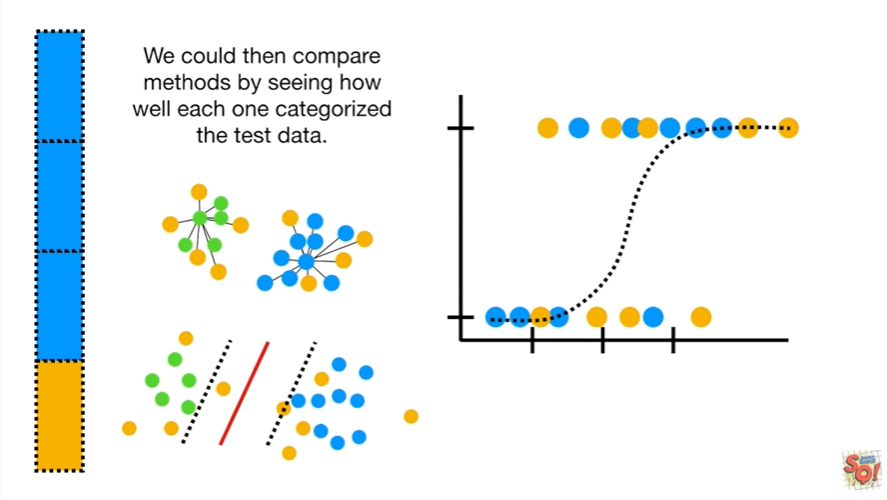
Reusing the same data for both training and testing is a bad idea because we need to know how the method will work on data it wasn't trained on.



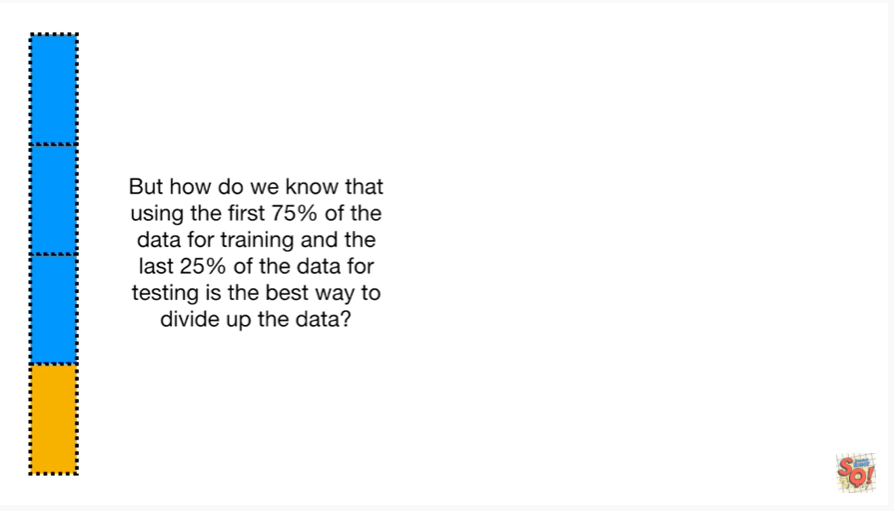
A slightly better idea would be to use the first seventy-five percent of the data for training



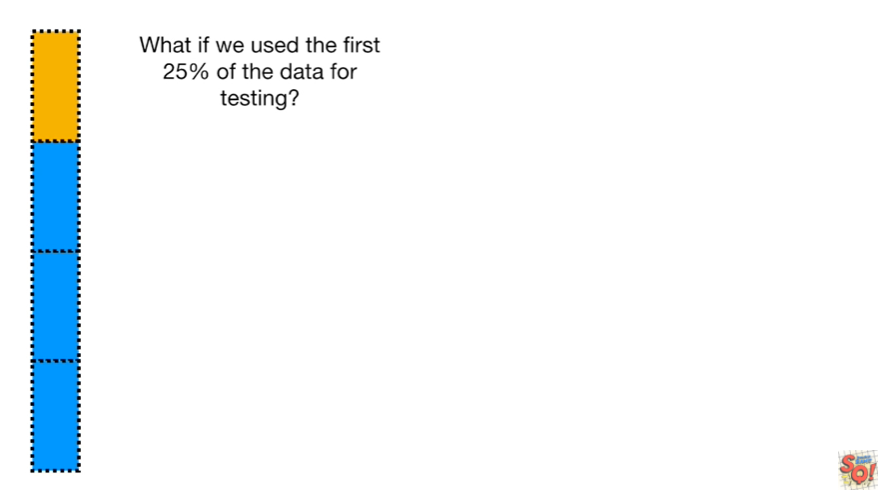
and the last 25% of the data for testing.



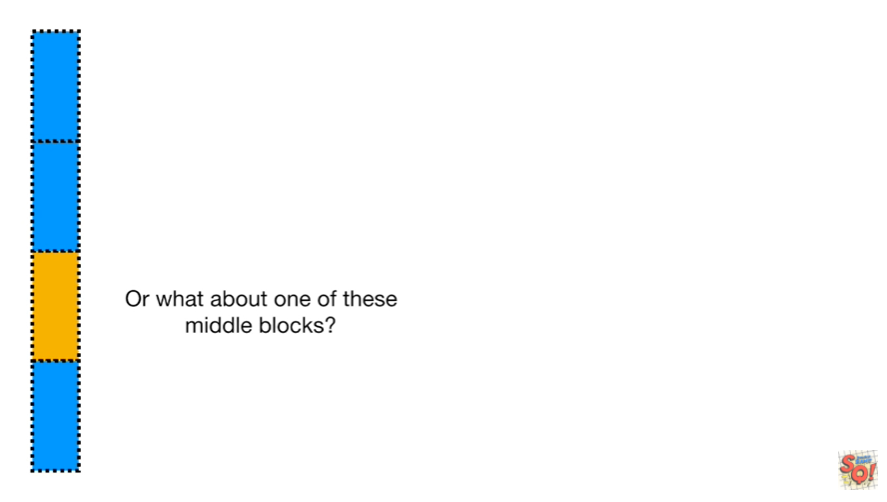
We could then compare methods by seeing how well each one categorized the test data.



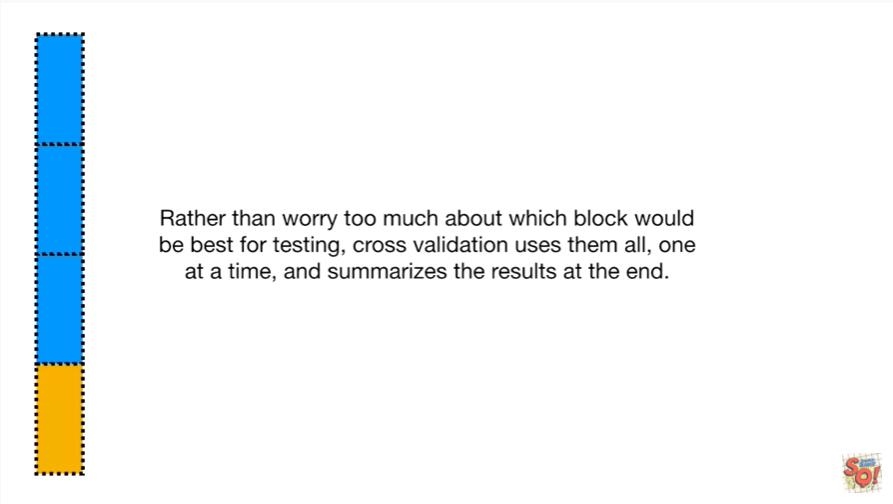
But how do we know that using the first seventy-five percent of the data for training in the last 25% of the data for testing is the best way to divide up the data ?



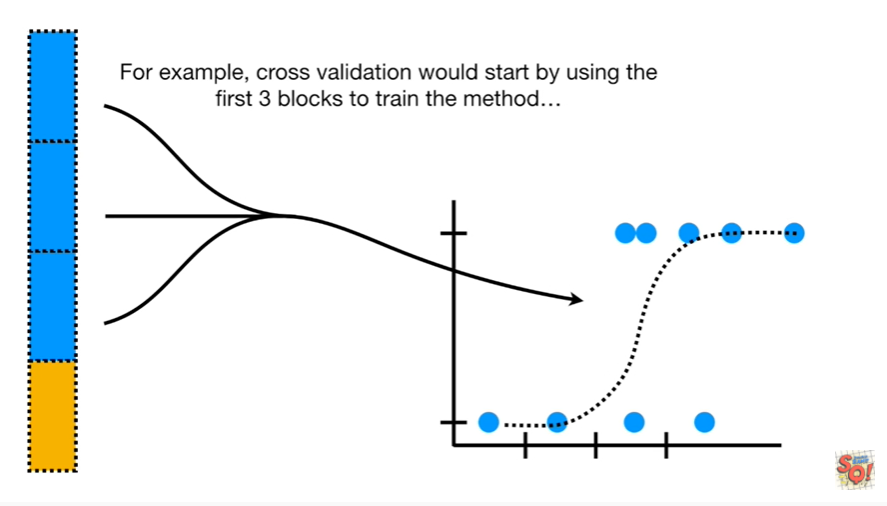
What if we use the first 25% of the data for testing ?



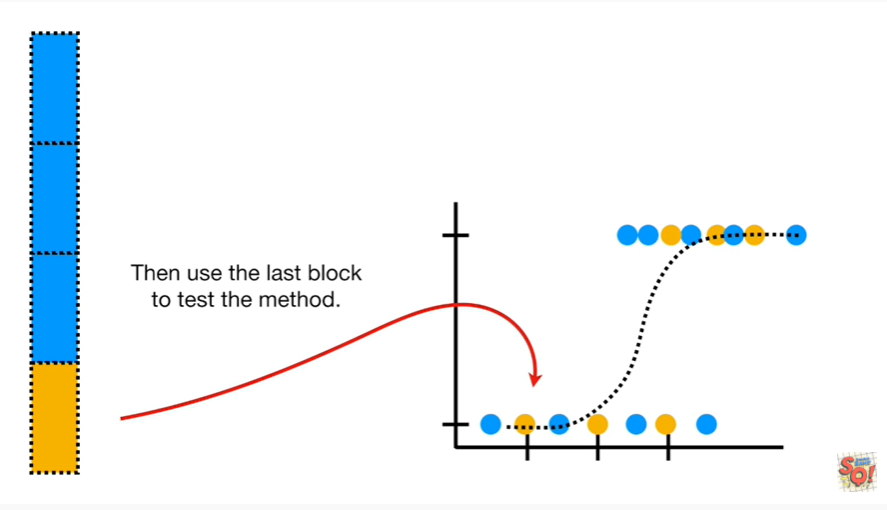
Or what about one of these middle blocks ?



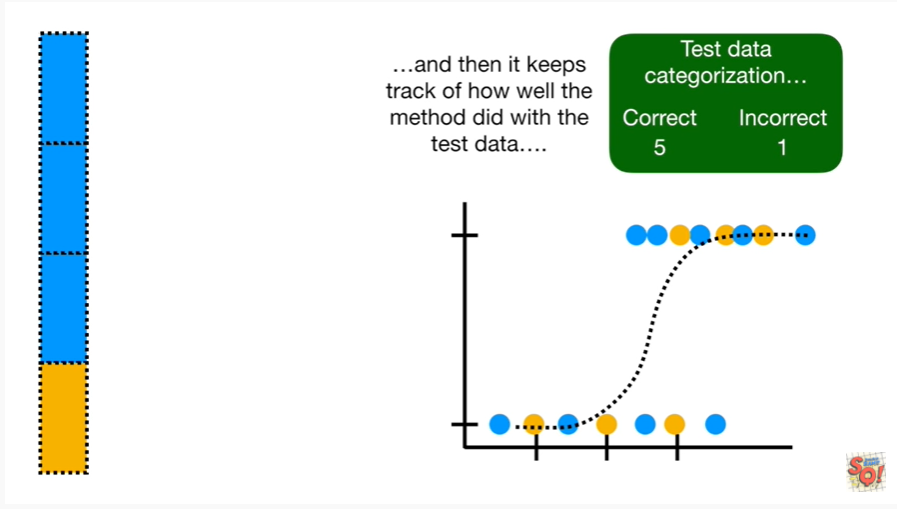
Rather than worry too much about which block would be best for testing cross-validation uses them all, one at a time, and summarizes the results at the end.



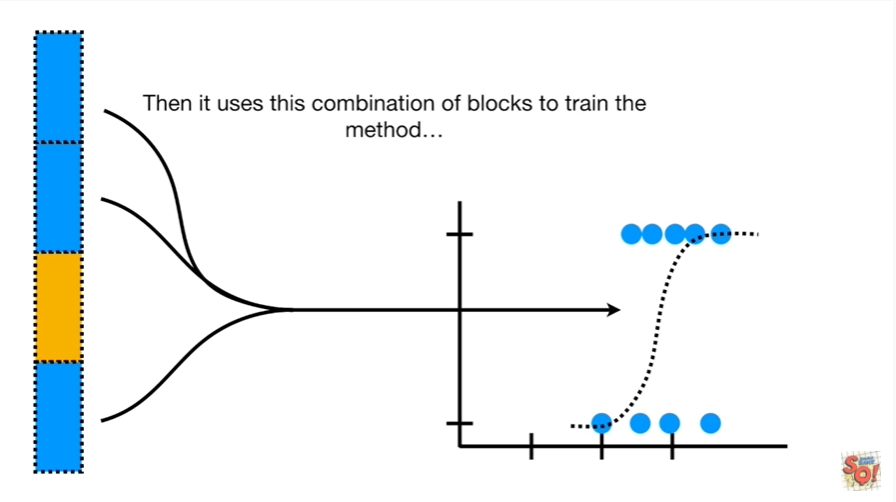
For example cross-validation would start by using the first three blocks to train the method



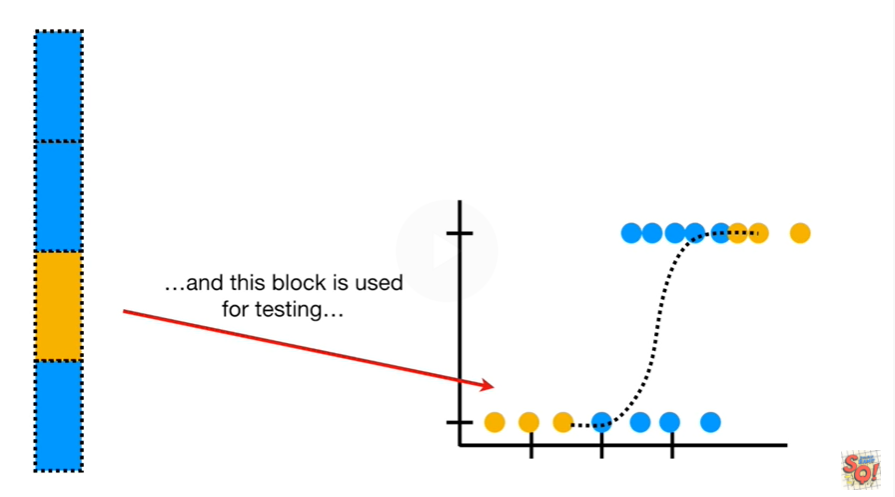
and then use the last block to test the method



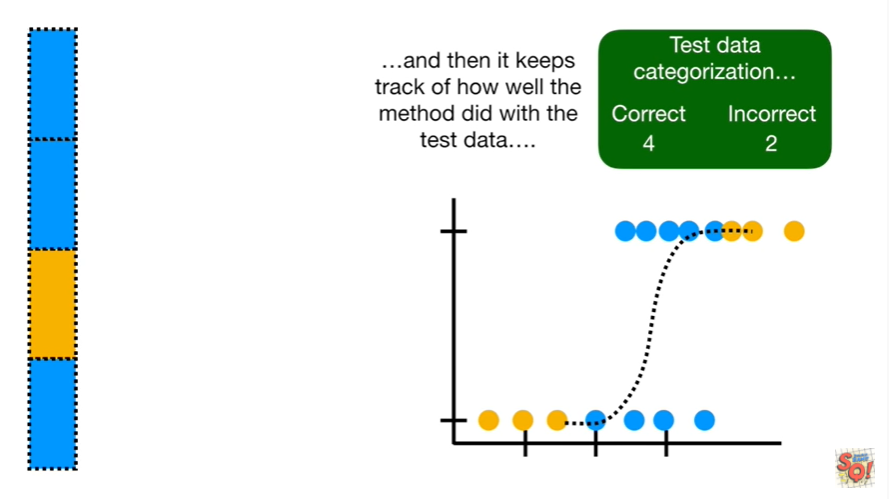
and then it keeps track of how well the method did with the test data.



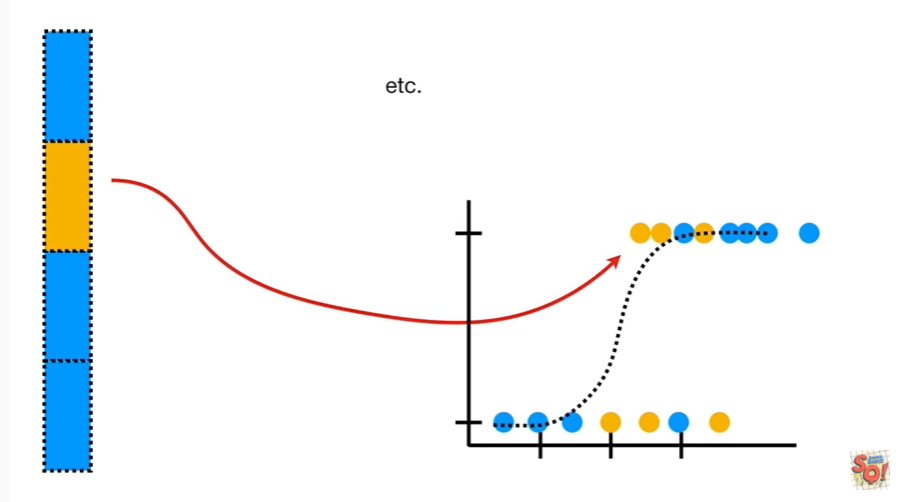
Then it uses this combination of blocks to train the method

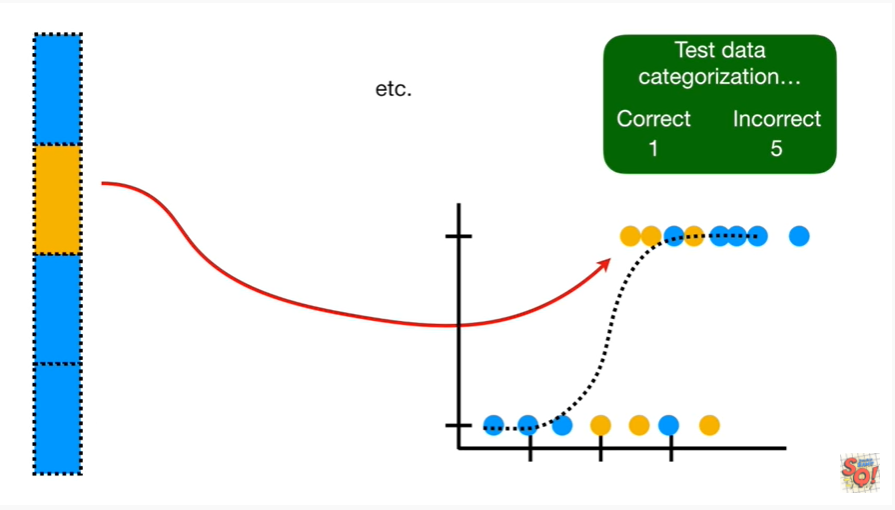


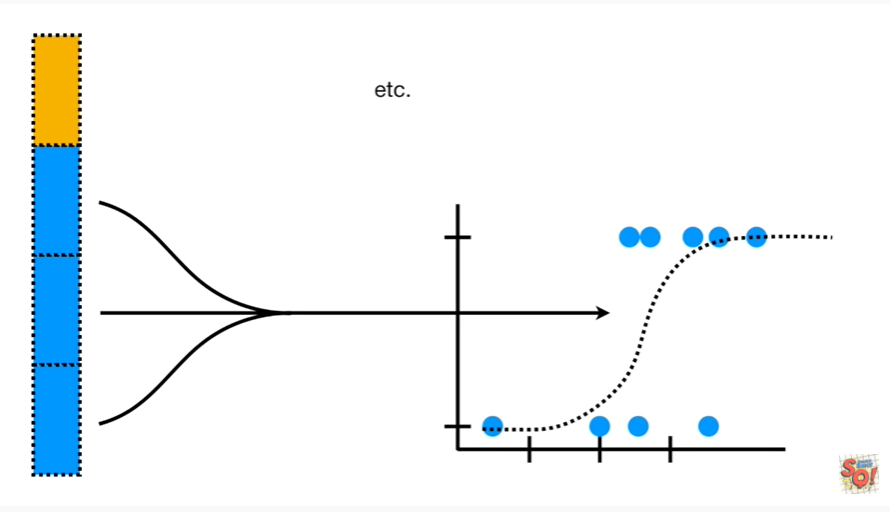
and this block is used for testing

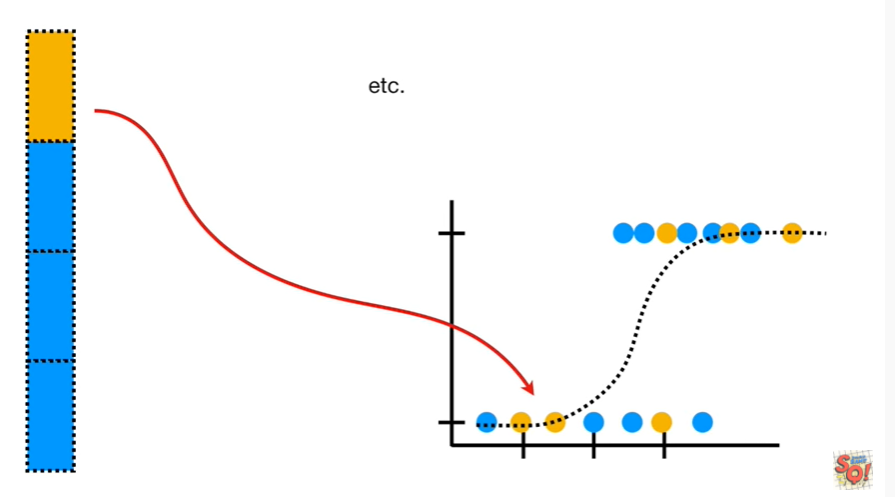


and then it keeps track of how well the method did with the test data.

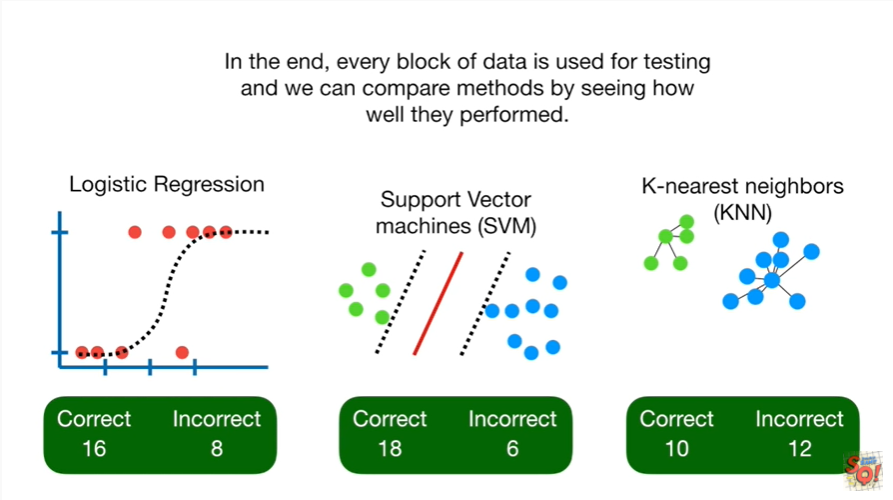




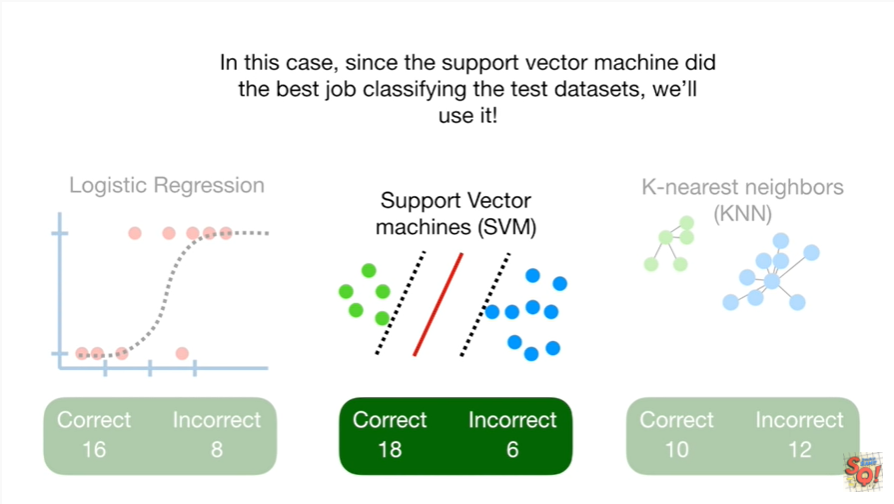




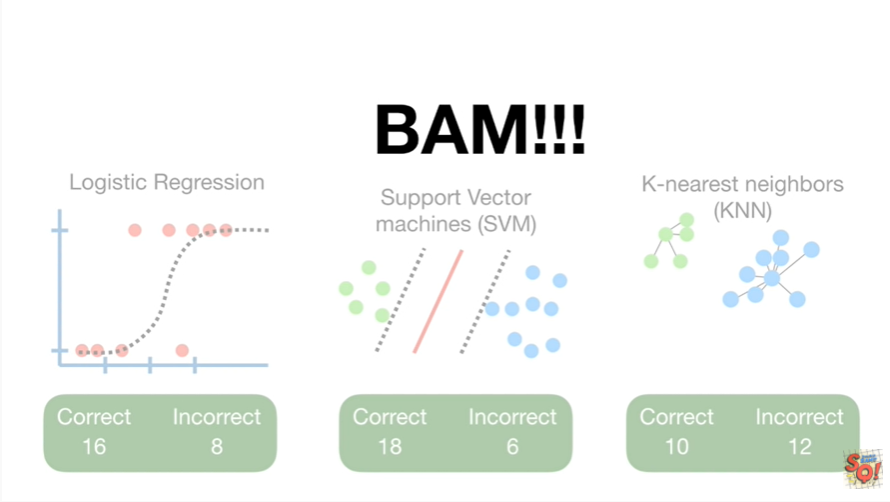
Etc etc etc



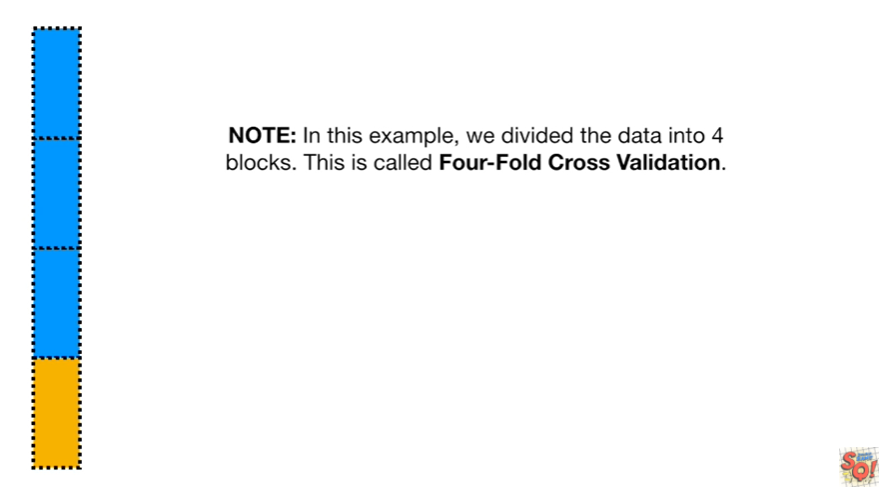
In the end every block of data is used for testing and we can compare methods by seeing how well they performed.



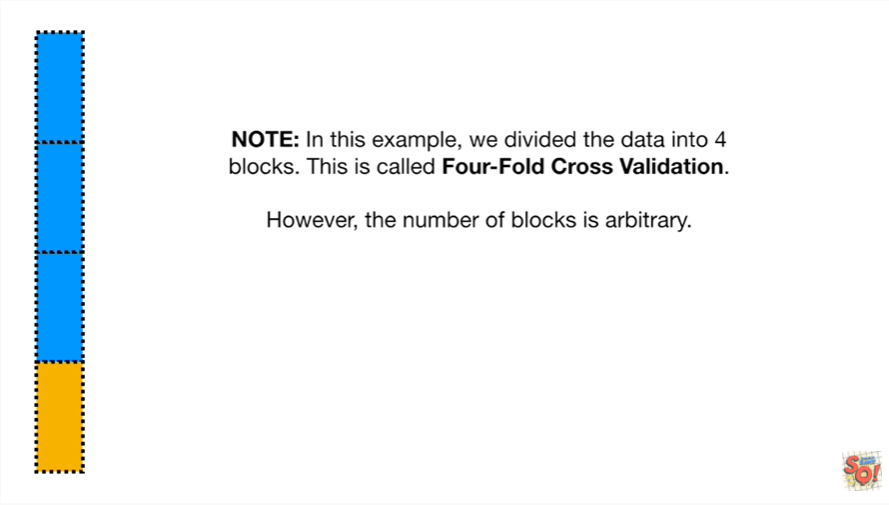
In this case since the support vector machine did the best job classifying the test data sets we'll use it.



BAM !!!



Note : in this example we divided the data into four blocks this is called fourfold cross-validation.

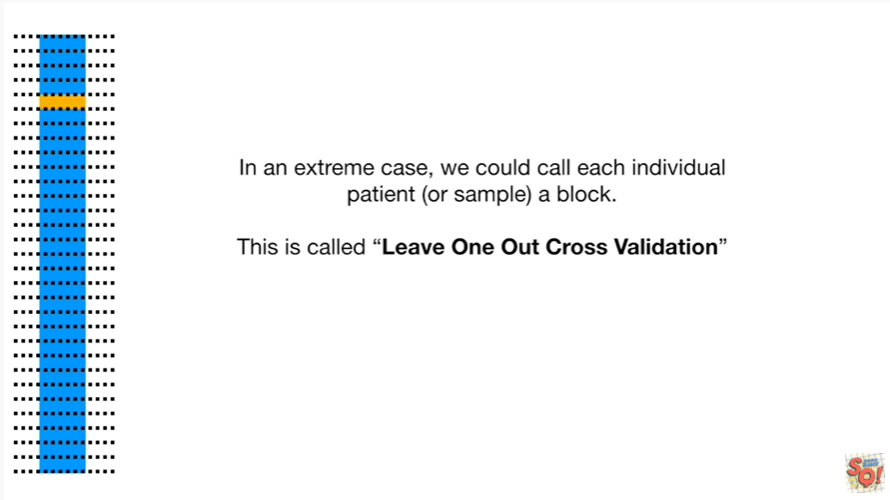


However the number of blocks is arbitrary.

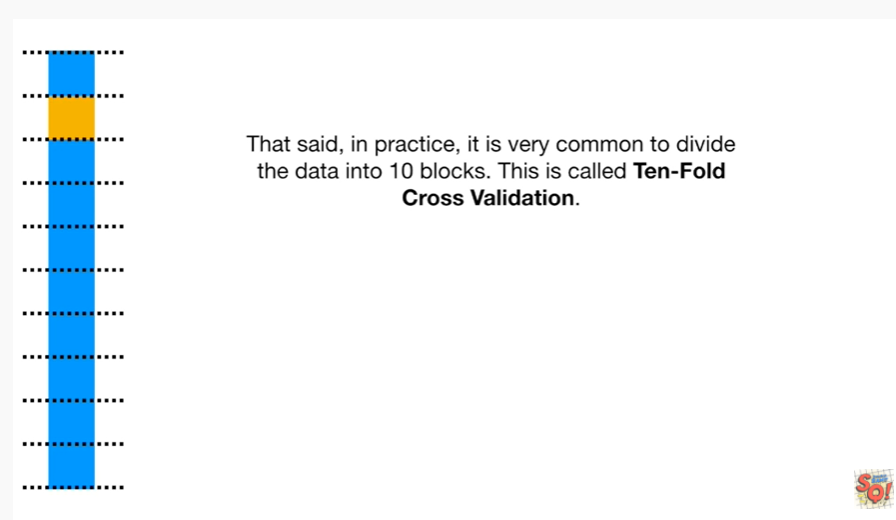


In an extreme case we could call each individual patient or sample a block.

This is called leave one out cross validation.

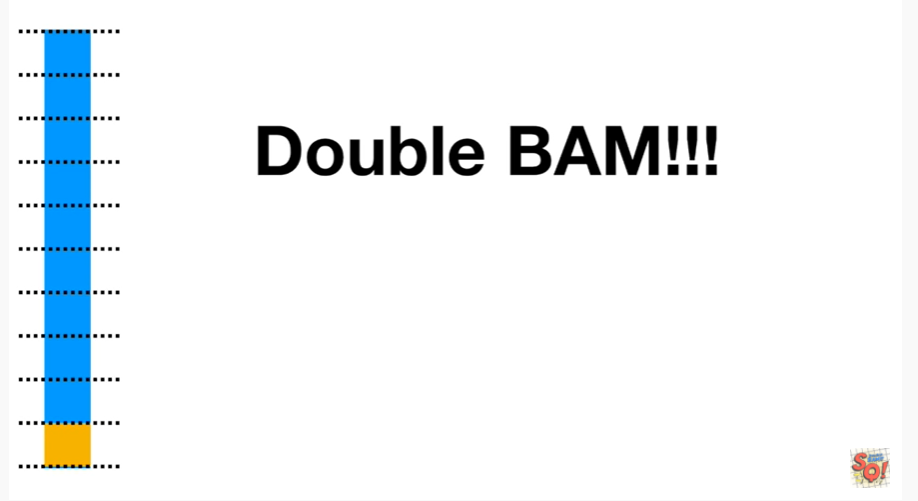


Each sample is tested individually.

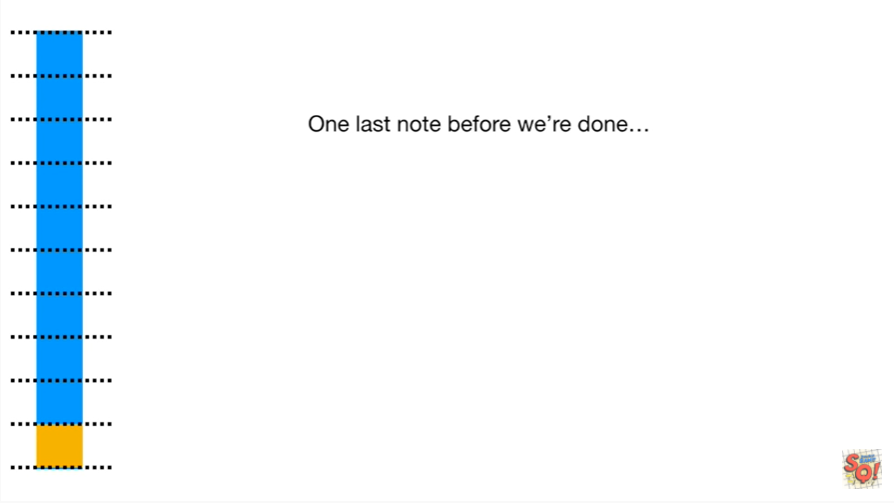


That said in practice it is very common to divide the data into ten blocks.

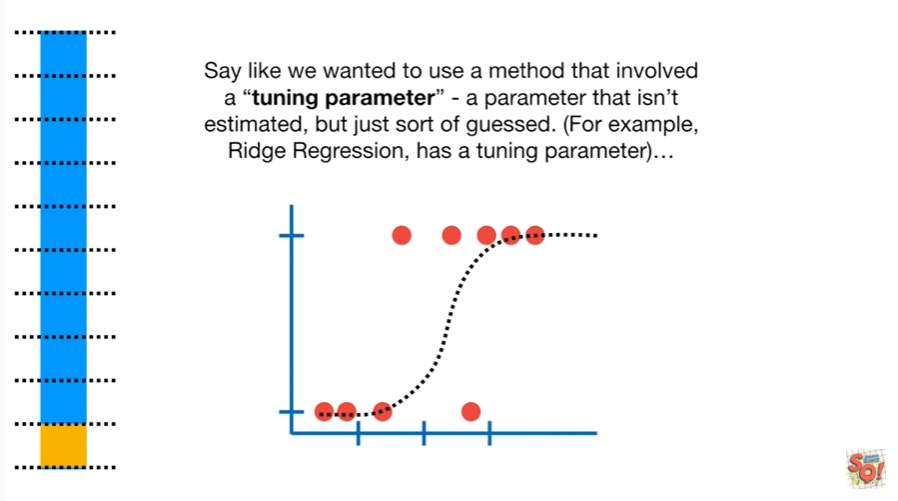
This is called 10-fold cross-validation.



Double BAM.

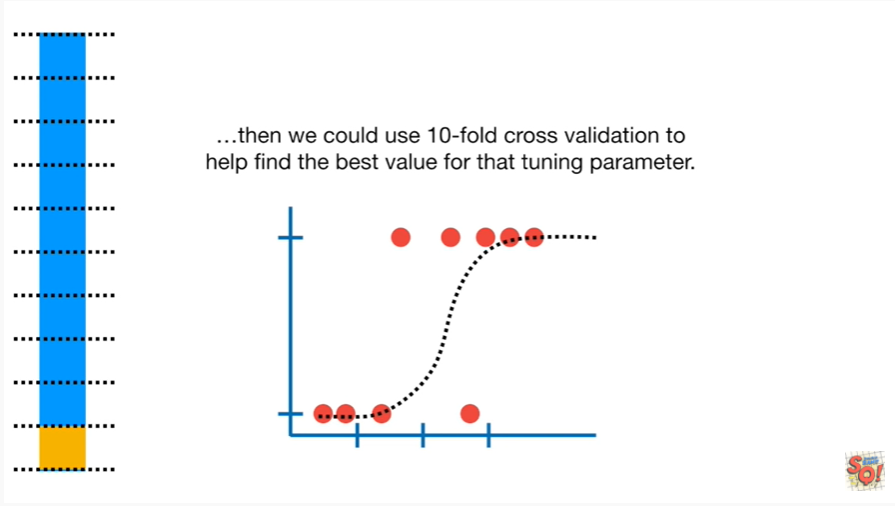


One last note before we're done



say like we wanted to use a method that involved a tuning parameter - a parameter that isn't estimated, but is just sort of guessed.

For example Ridge regression has a tuning parameter.



Then we could use 10-fold cross-validation to help find the best value for that tuning parameter.



Tiny BAM !