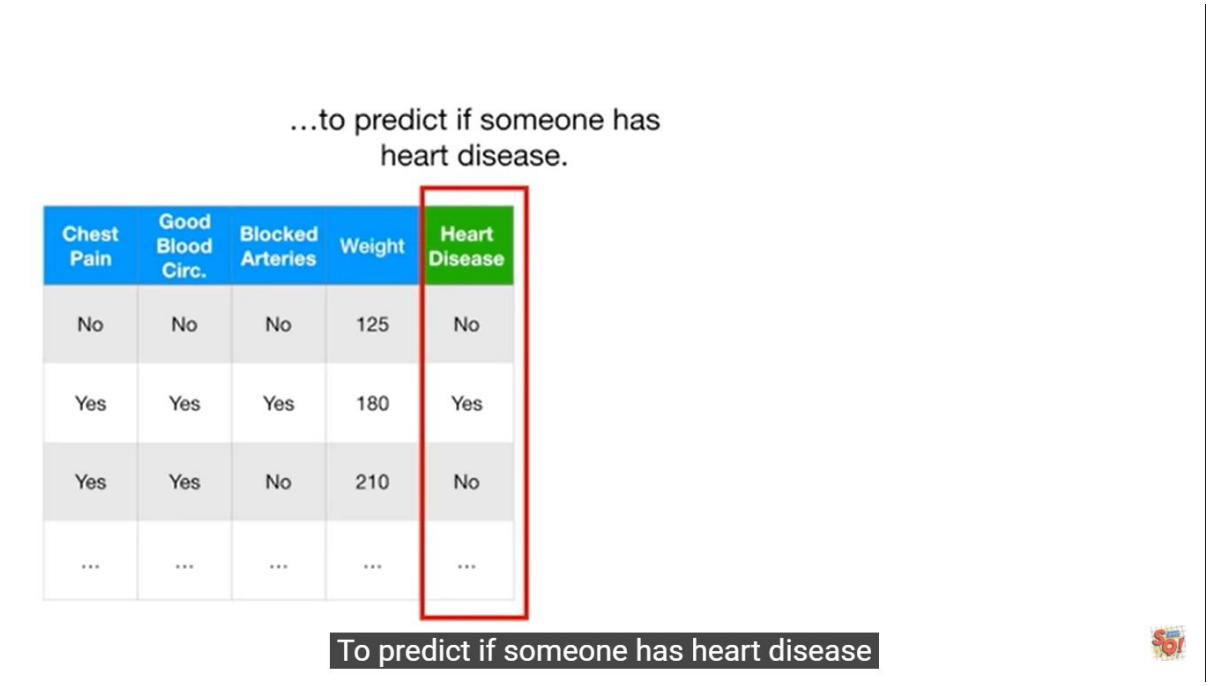
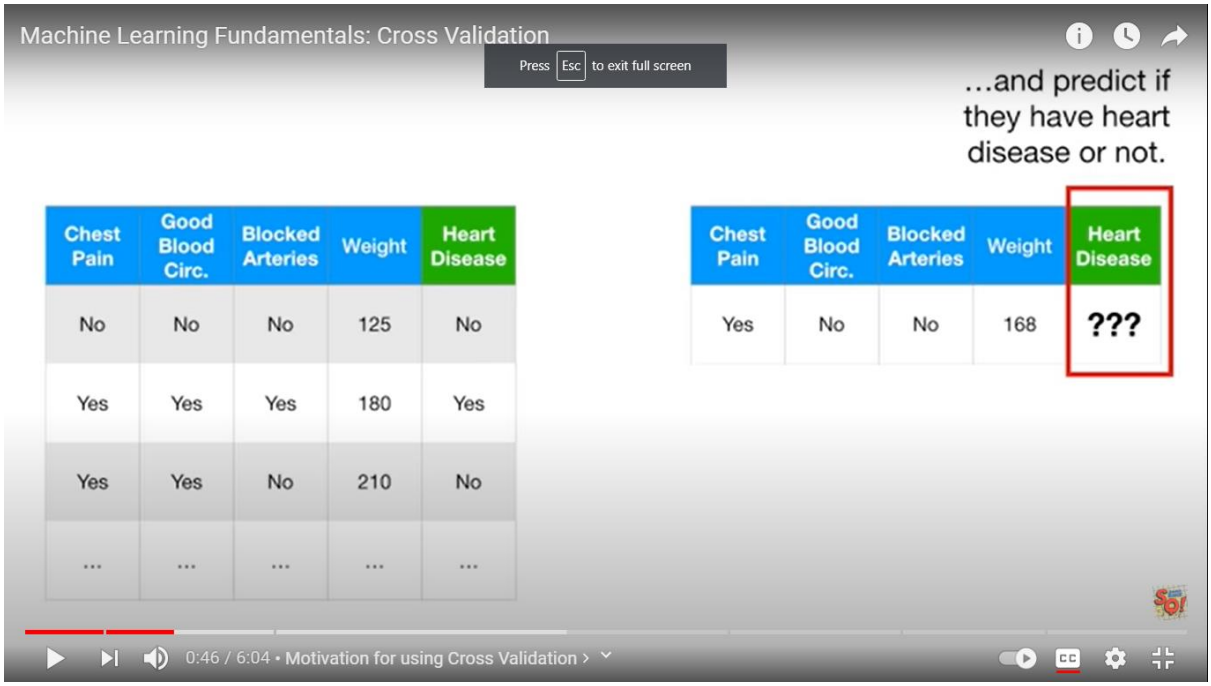
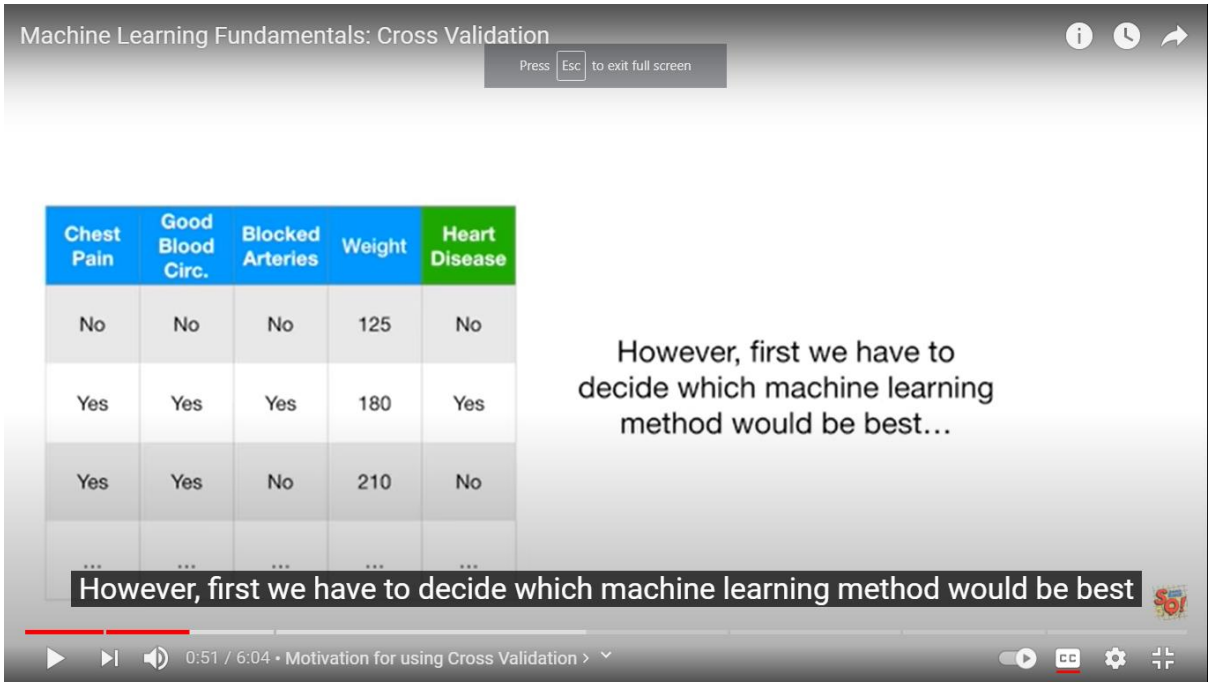
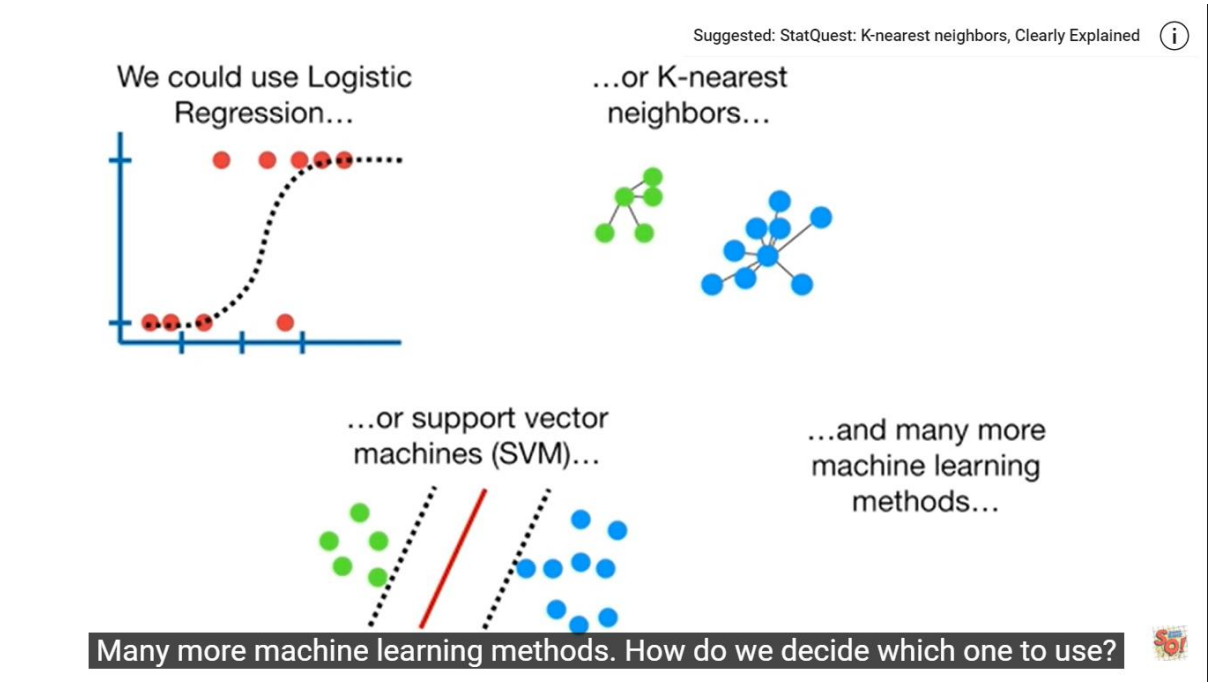
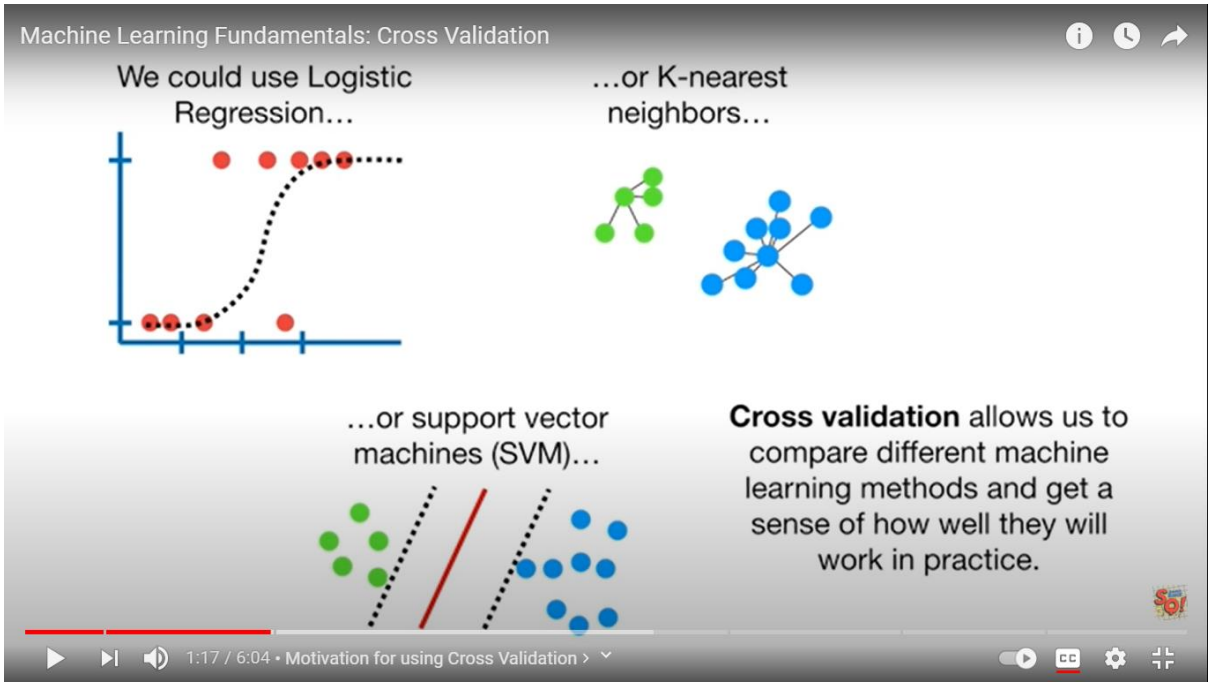


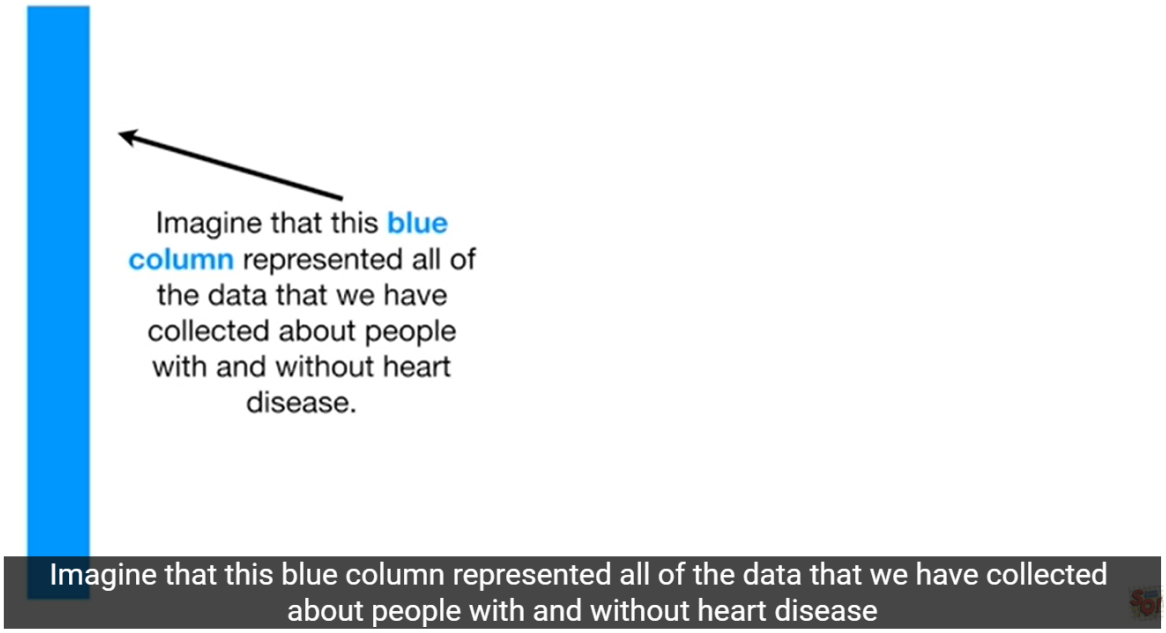
When new patient comes up

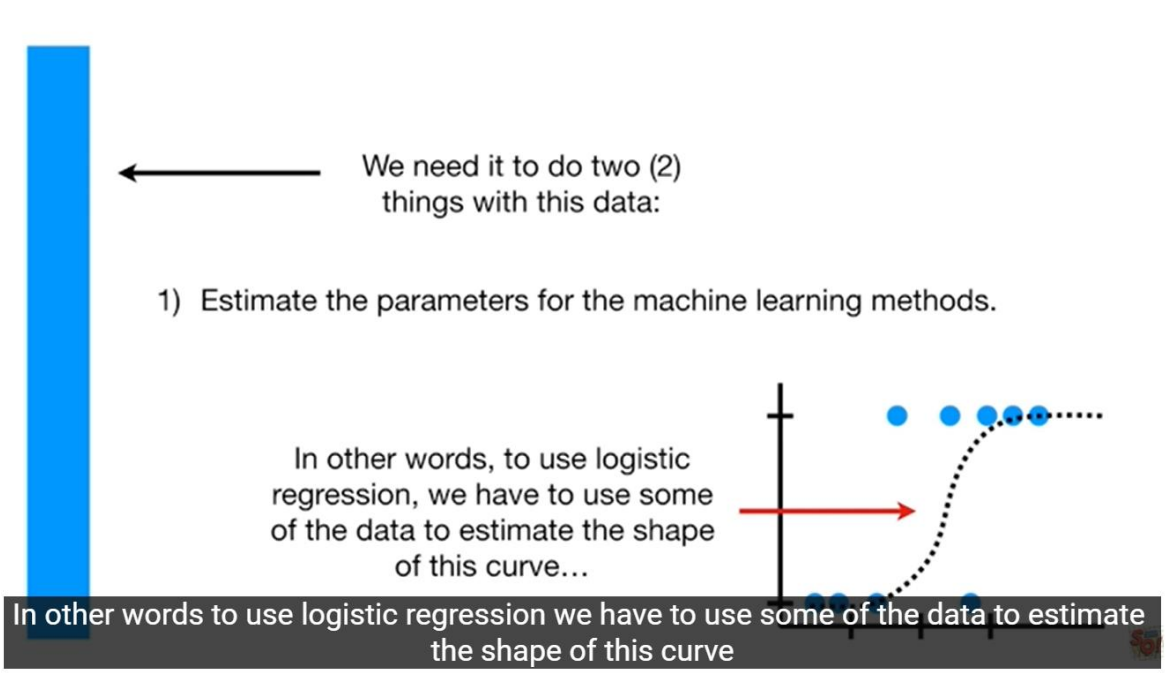


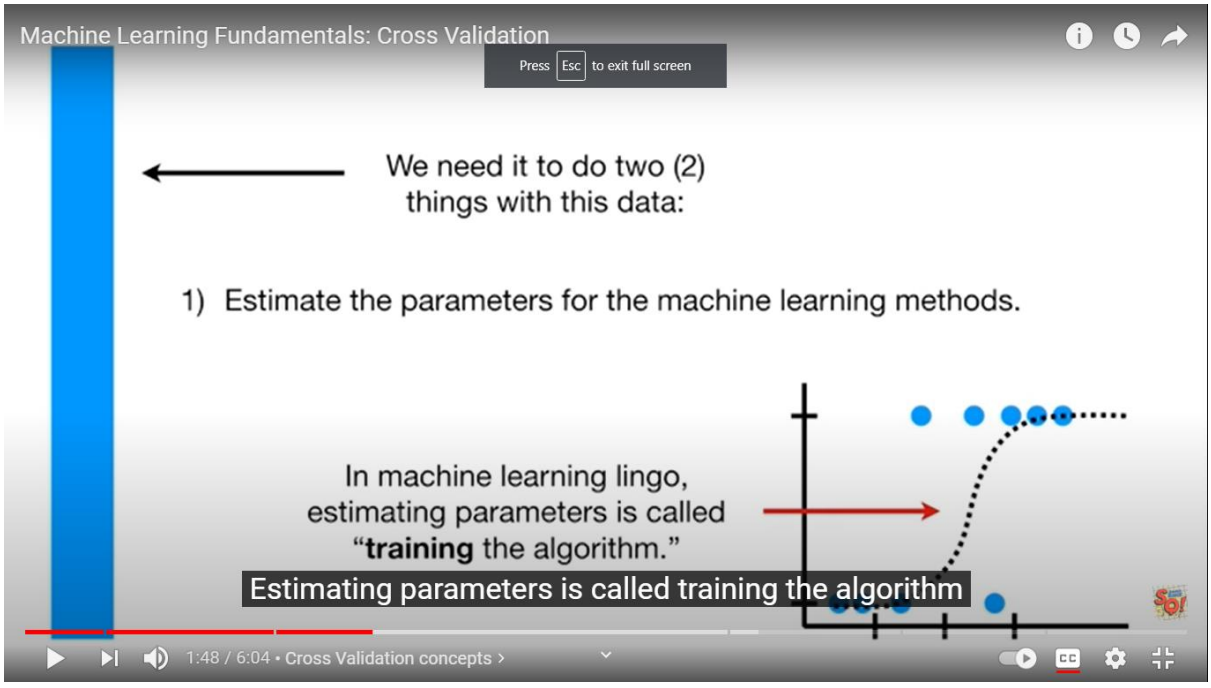

Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
Chest Pain | Good Blood Circ. | Blocked Arteries | Weight | Heart Disease
No | No | No | 125 | No
Yes | Yes | Yes | 180 | Yes
Yes | Yes | No | 210 | No
However, first we have to decide which machine learning method would be best…
However, first we have to decide which machine learning method would be best
0:51 / 6:04 • Motivation for using Cross Validation


Suggested: StatQuest: K-nearest neighbors, Clearly Explained
We could use Logistic Regression…
…or K-nearest neighbors…
…or support vector machines (SVM)…
…and many more machine learning methods…
Many more machine learning methods. How do we decide which one to use?


Machine Learning Fundamentals: Cross Validation
We could use Logistic Regression…
…or K-nearest neighbors…
…or support vector machines (SVM)…
Cross validation allows us to compare different machine learning methods and get a sense of how well they will work in practice.
1:17 / 6:04 • Motivation for using Cross Validation ›


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


We need it to do two (2) things with this data:
1) 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 other words to use logistic regression we have to use some of the data to estimate the shape of this curve


Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
We need it to do two (2) things with this data:
1) Estimate the parameters for the machine learning methods.
In machine learning lingo, estimating parameters is called "training the algorithm."
Estimating parameters is called training the algorithm
1:48 / 6:04 • Cross Validation concepts ›

We need it to do two (2) things with this data:

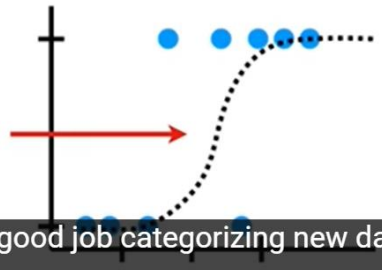
1) Estimate the parameters for the machine learning methods.
2) Evaluate how well the machine learning methods work.

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

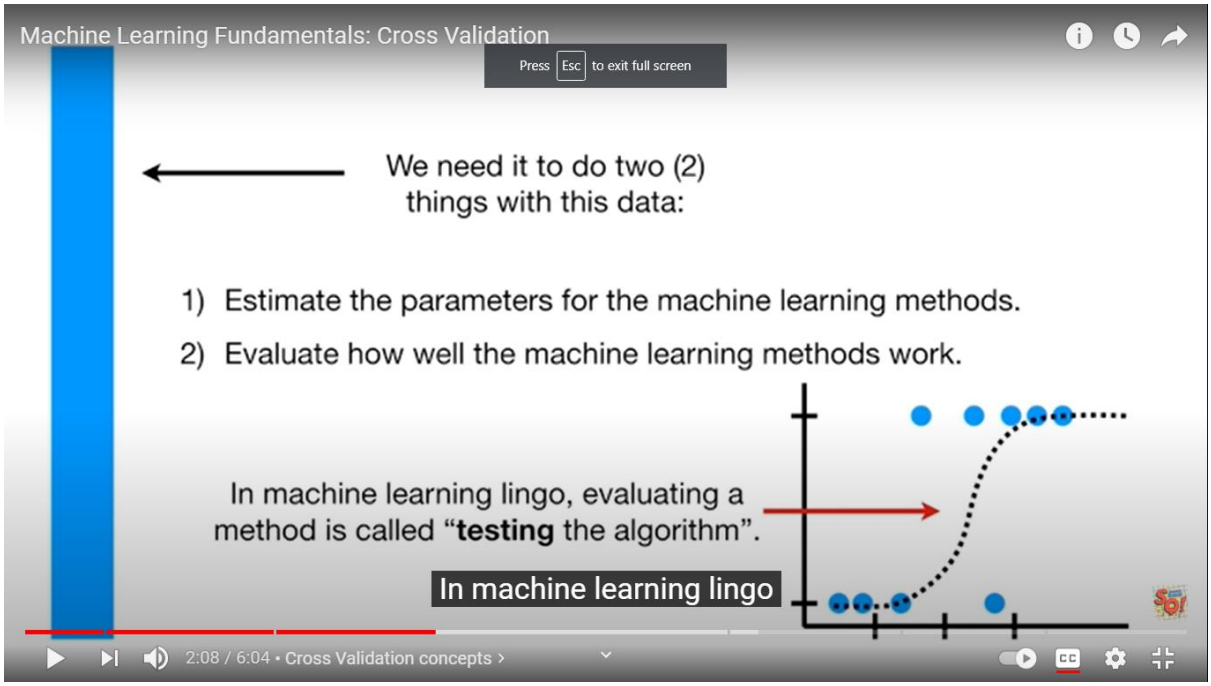
We need it to do two (2) things with this data:

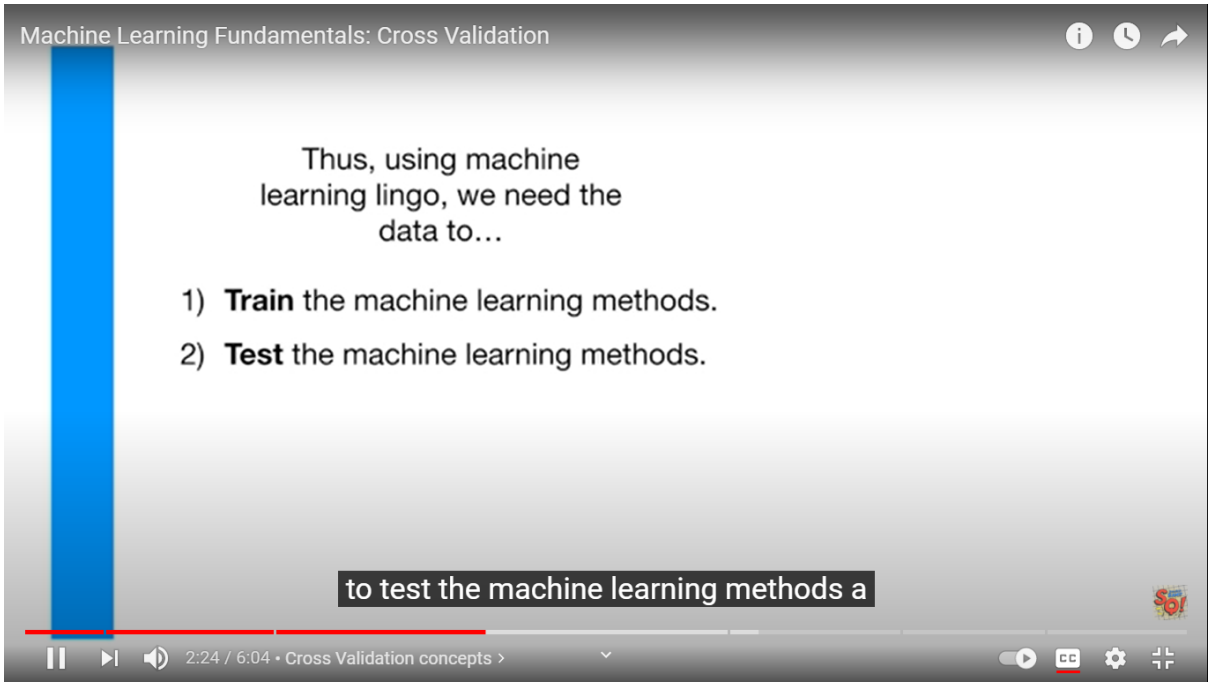
1) Estimate the parameters for the machine learning methods.
2) Evaluate how well the machine learning methods work.

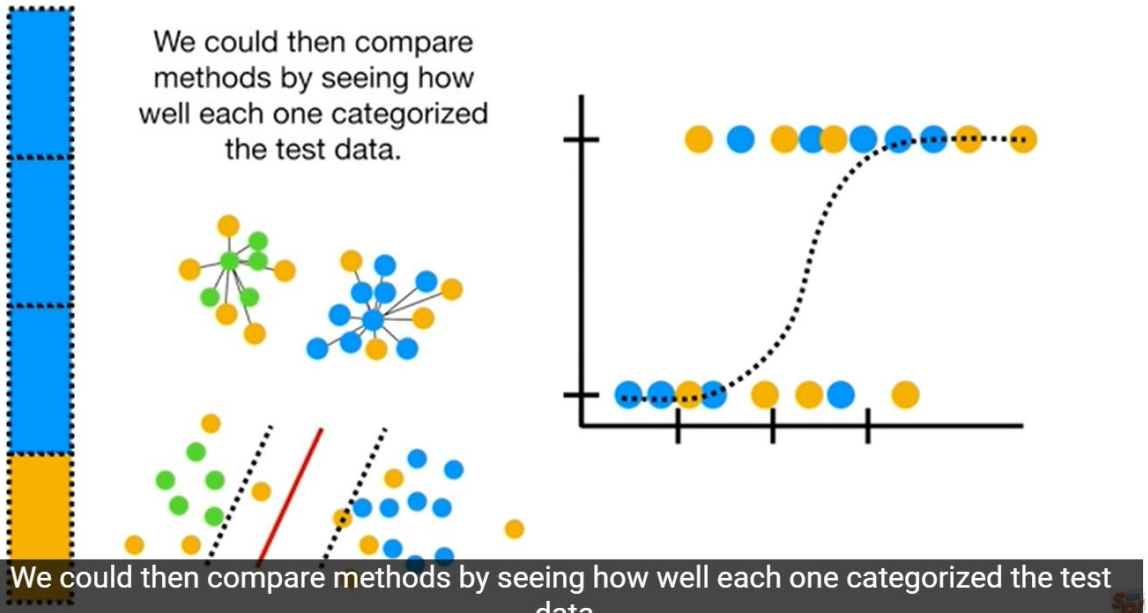
In other words, we need to find out if this curve will do a good job categorizing new data.



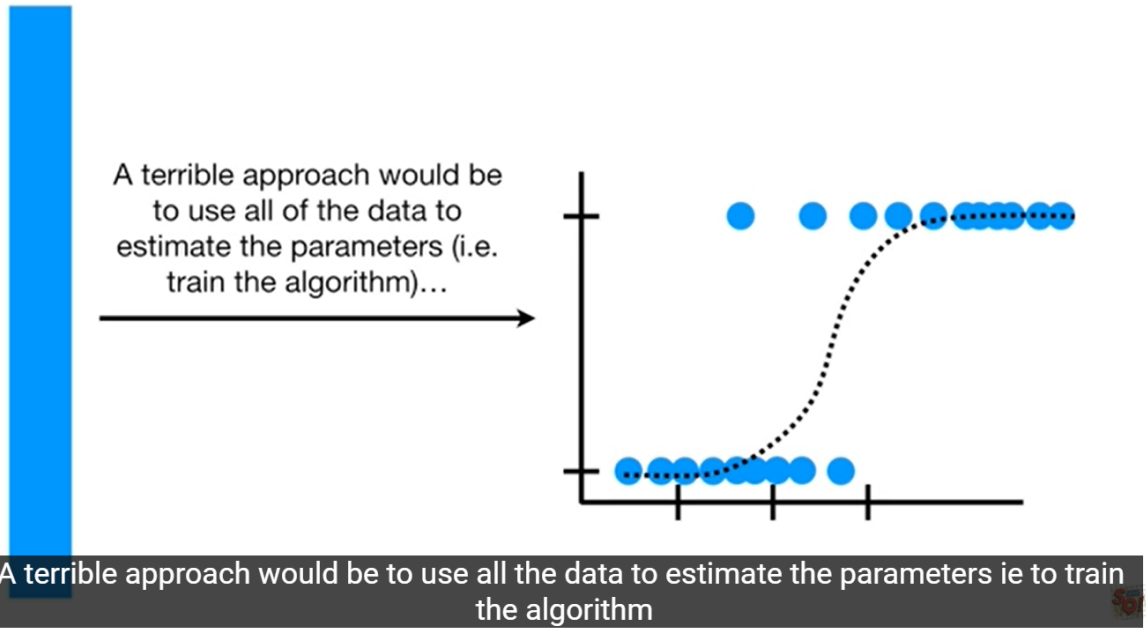
Other words we need to find out if this curve will do a good job categorizing new data in


Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
We need it to do two (2) things with this data:
1) Estimate the parameters for the machine learning methods.
2) Evaluate how well the machine learning methods work.
In machine learning lingo, evaluating a method is called "testing the algorithm".
In machine learning lingo
2:08 / 6:04 • Cross Validation concepts ›


Machine Learning Fundamentals: Cross Validation
Thus, using machine learning lingo, we need the data to…
1) Train the machine learning methods.
2) Test the machine learning methods.
to test the machine learning methods a
2:24 / 6:04 • Cross Validation concepts ›



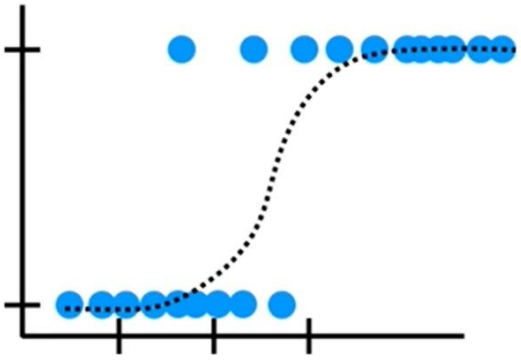
We could then compare methods by seeing how well each one categorized the test data



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

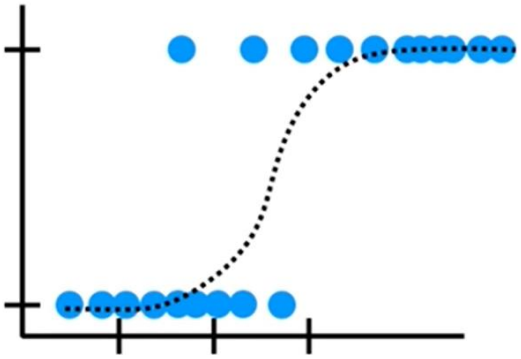
A terrible approach would be to use all of the data to estimate the parameters (i.e. train the algorithm)…

…because then there wouldn't be any data left to test the method.

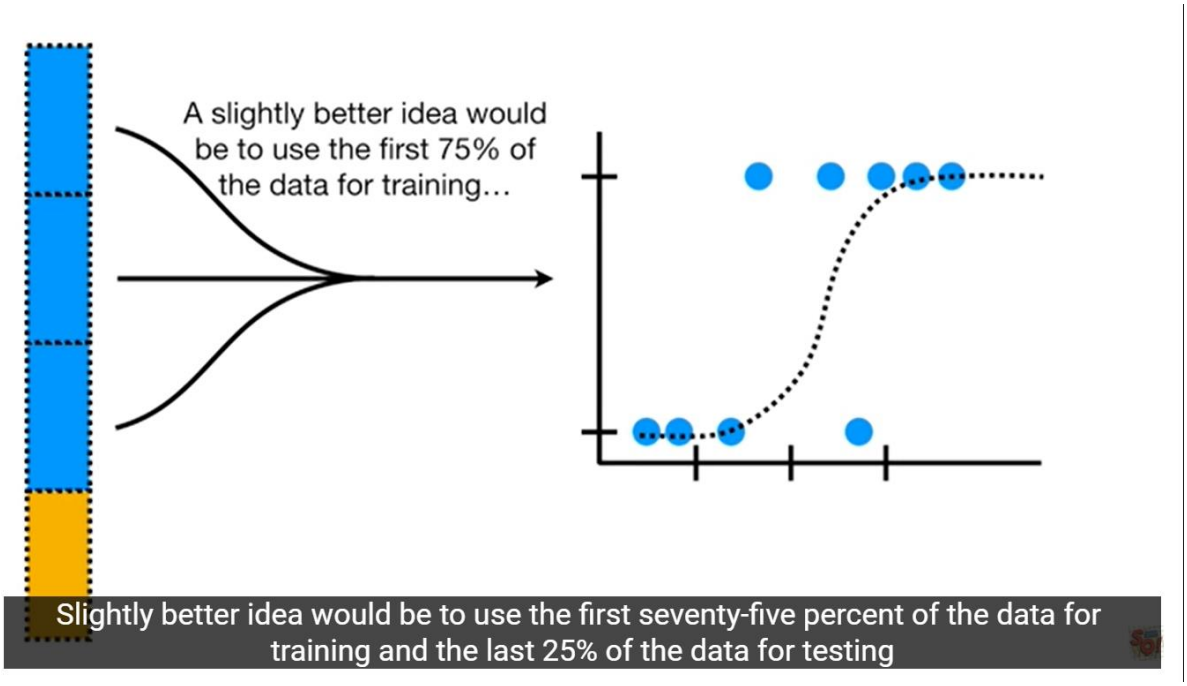


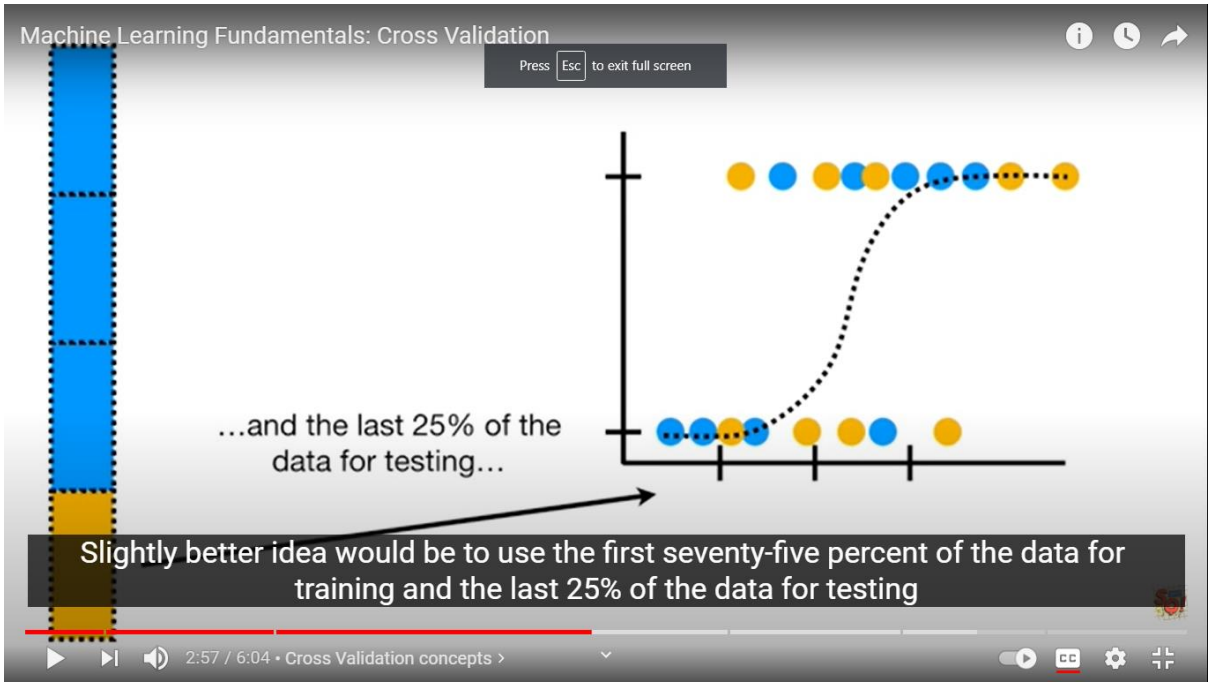
Because then we wouldn't have any data left to test the method

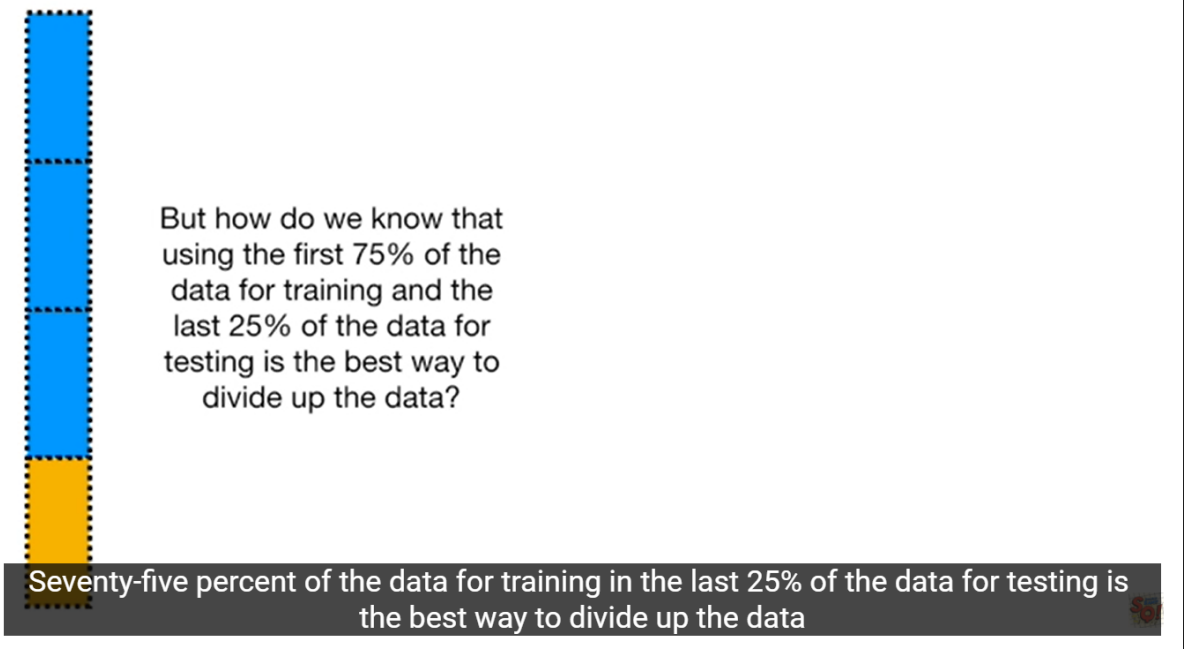
Reusing the same data for both training and testing is a bad idea because we need to know how the method will work on data it wasn't trained on.

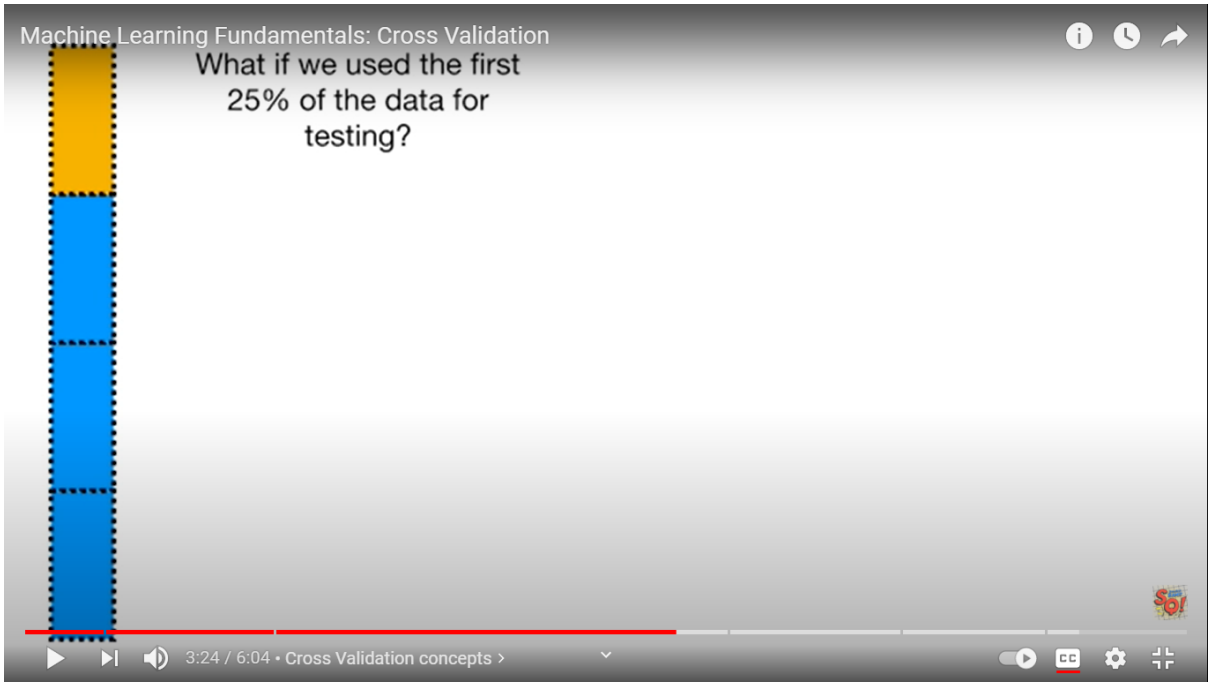


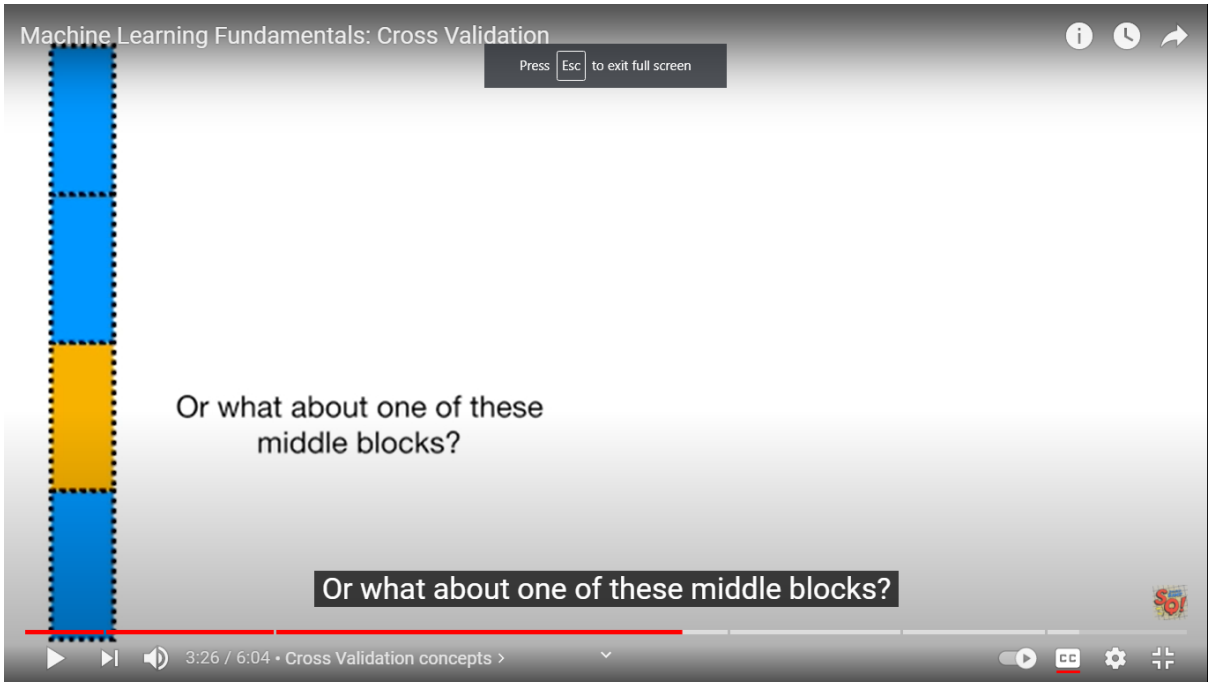
Testing is a bad idea because we need to know how the method will work on data. It wasn't trained on a

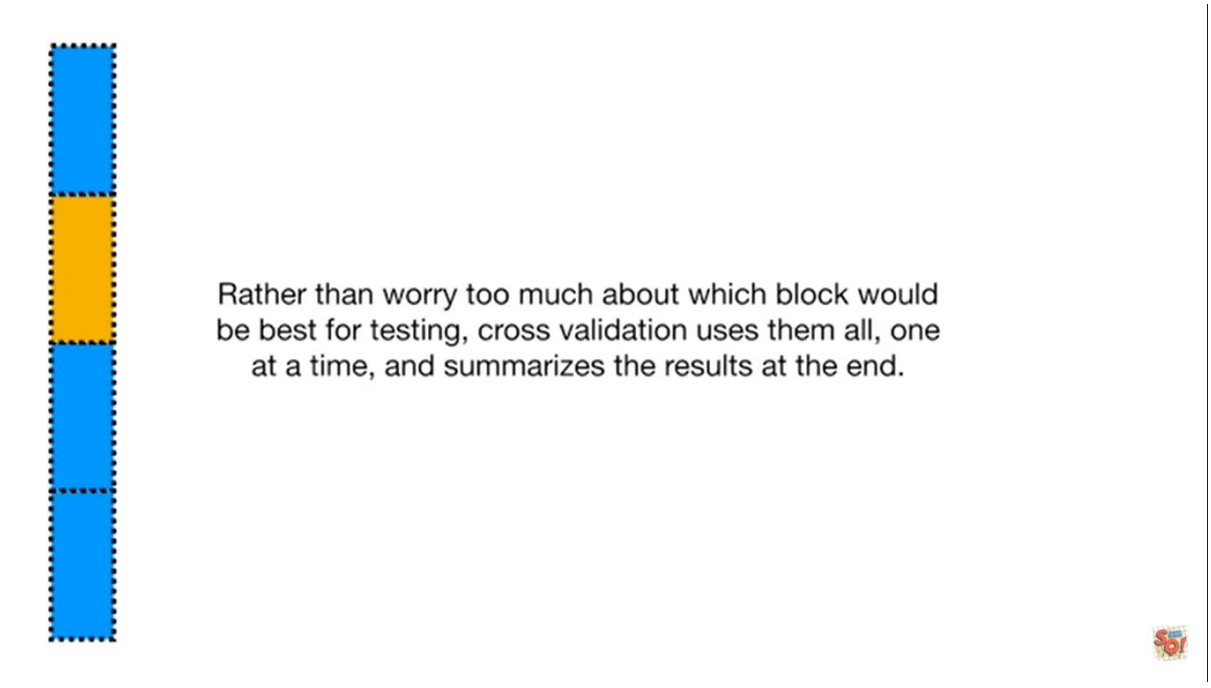

A slightly better idea would be to use the first 75% of the data for training…
Slightly better idea would be to use the first seventy-five percent of the data for training and the last 25% of the data for testing

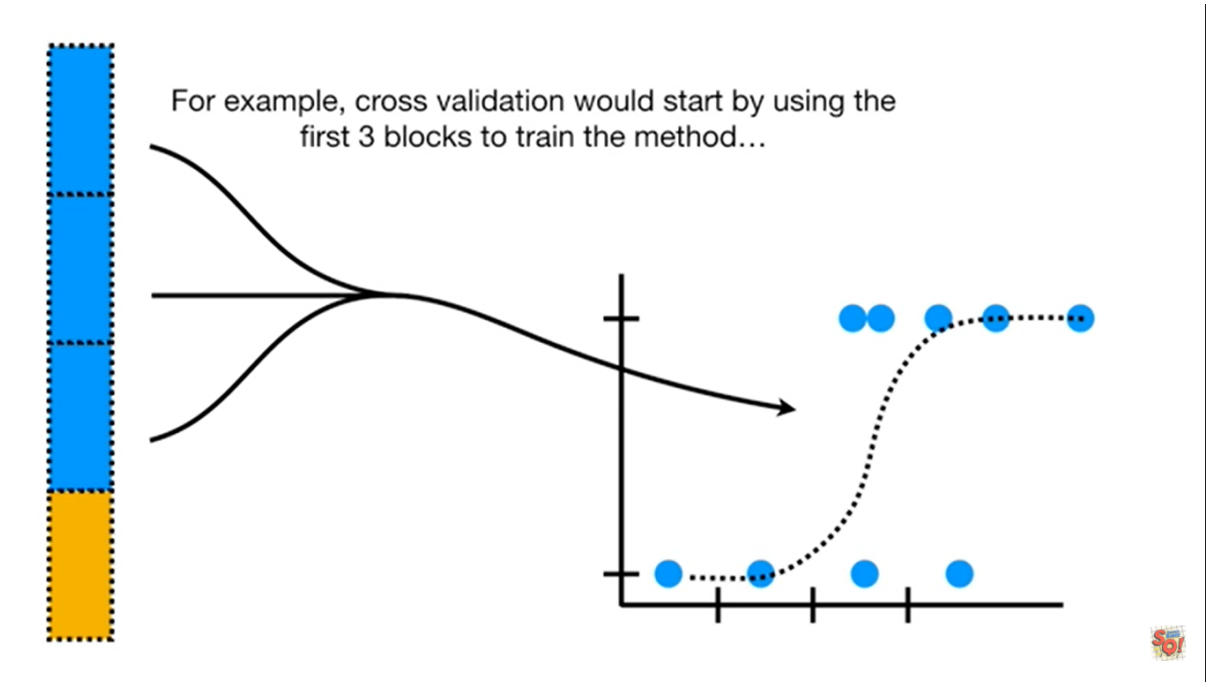

Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
…and the last 25% of the data for testing…
Slightly better idea would be to use the first seventy-five percent of the data for training and the last 25% of the data for testing
2:57 / 6:04 • Cross Validation concepts ›

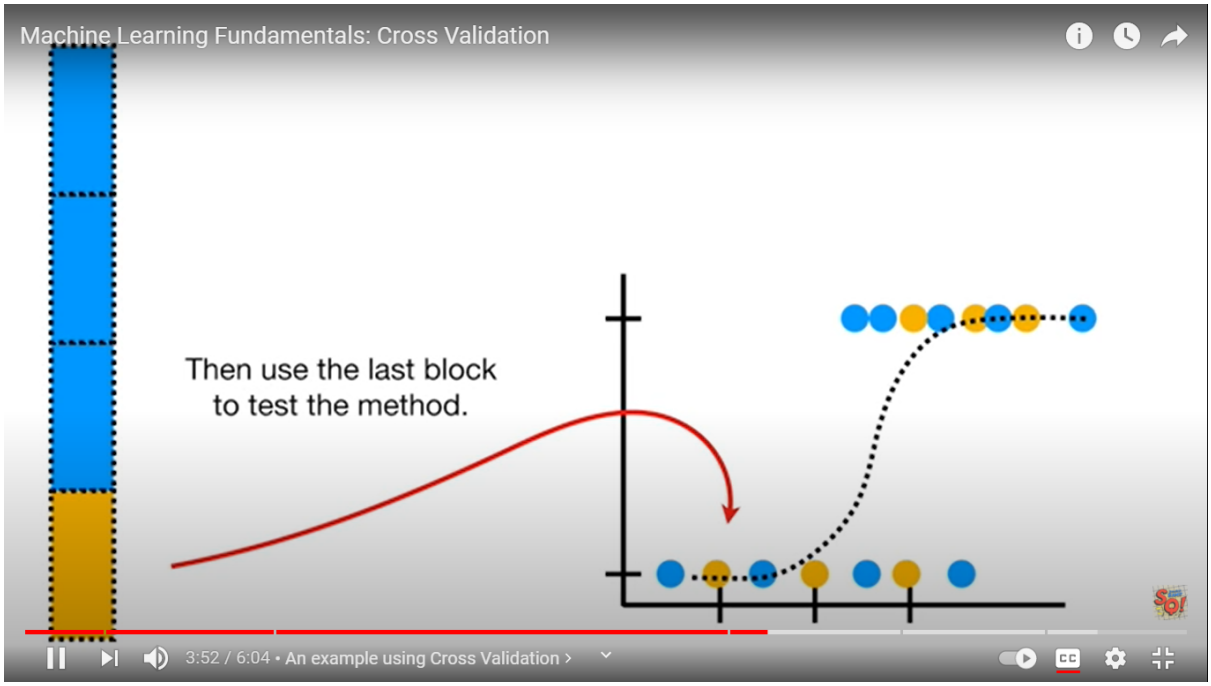

But how do we know that using the first 75% of the data for training and the last 25% of the data for testing is the best way to divide up the data?
Seventy-five percent of the data for training in the last 25% of the data for testing is the best way to divide up the data

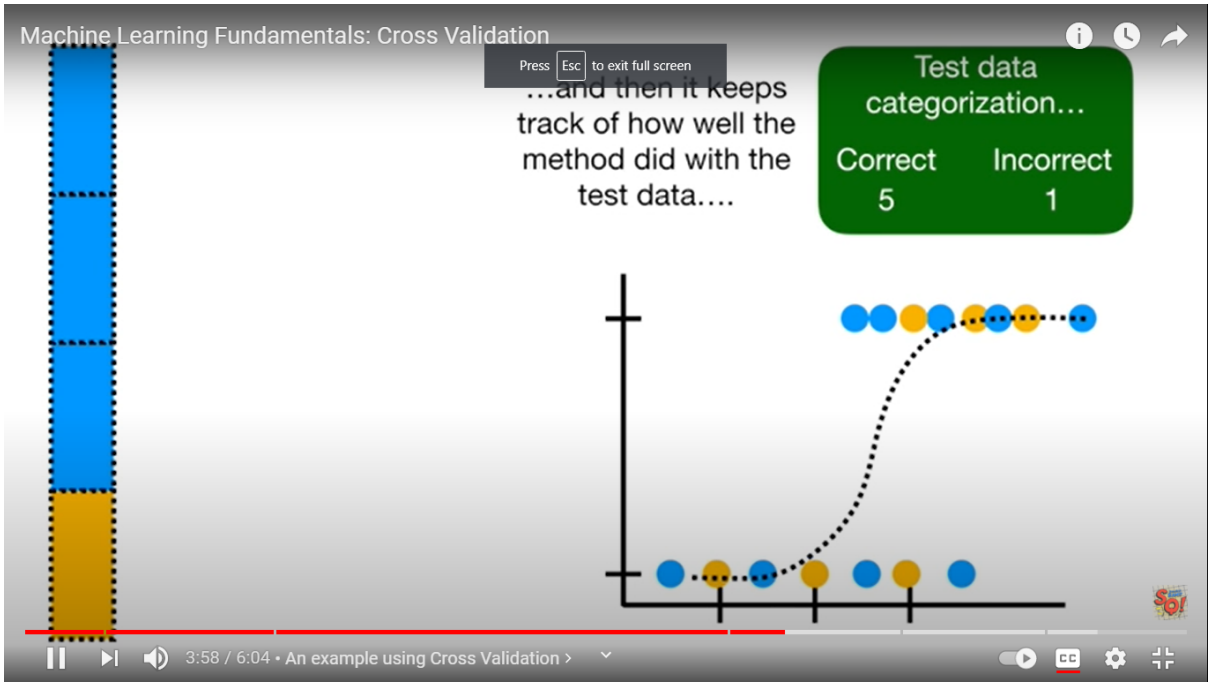

Machine Learning Fundamentals: Cross Validation
What if we used the first 25% of the data for testing?
3:24 / 6:04 • Cross Validation concepts ›

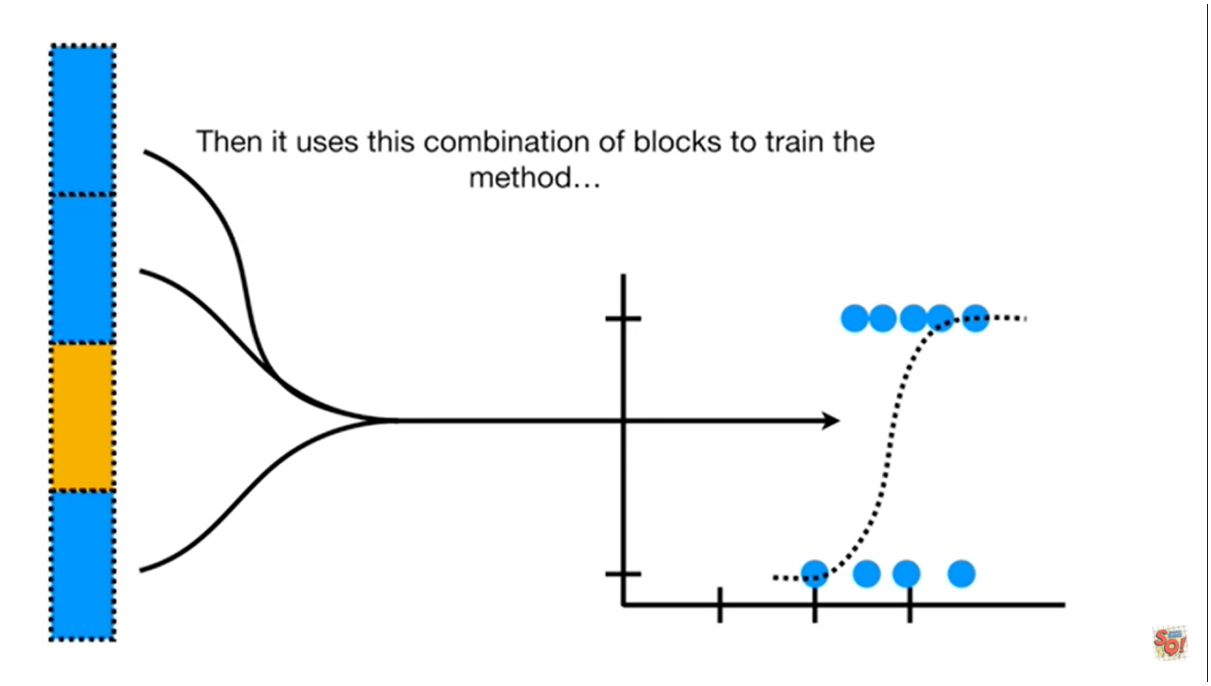

Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
Or what about one of these middle blocks?
Or what about one of these middle blocks?
3:26 / 6:04 • Cross Validation concepts ›

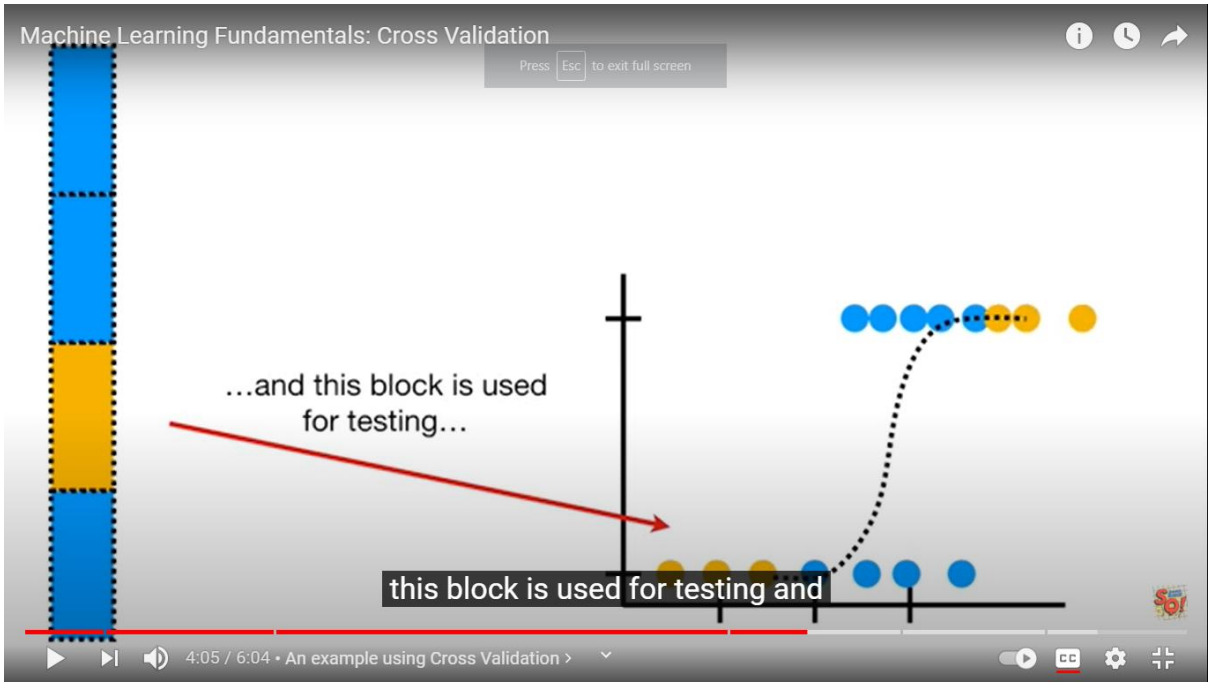

Rather than worry too much about which block would be best for testing, cross validation uses them all, one at a time, and summarizes the results at the end.

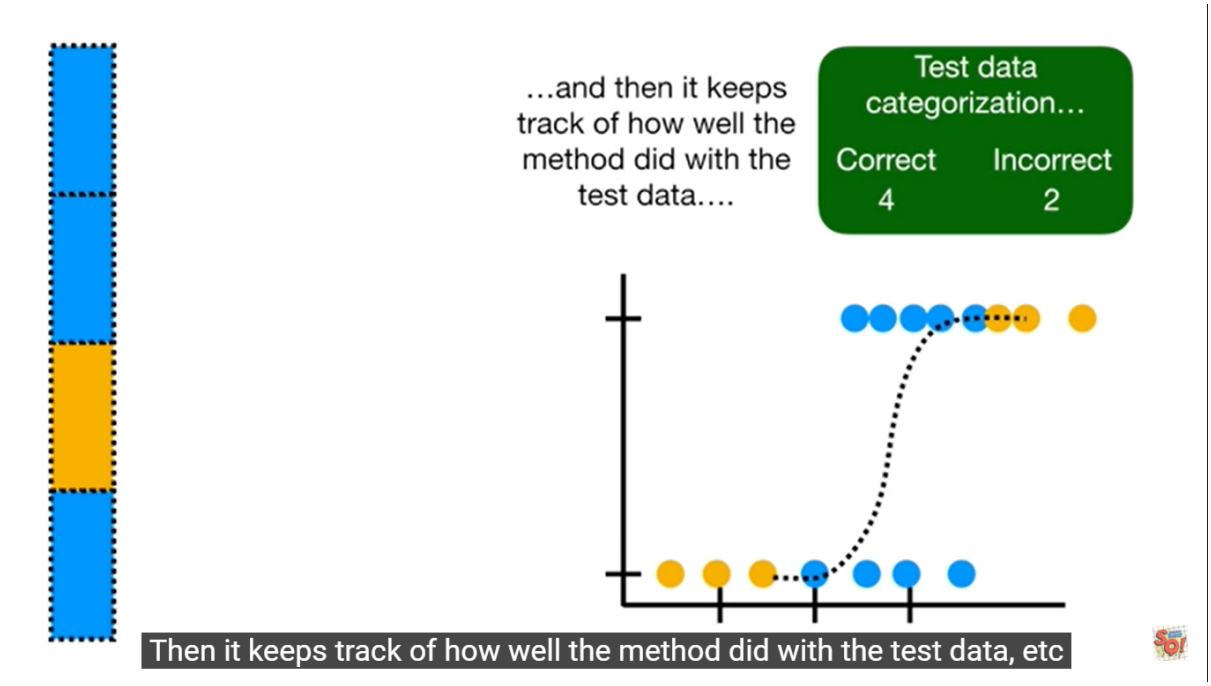

For example, cross validation would start by using the
first 3 blocks to train the method…

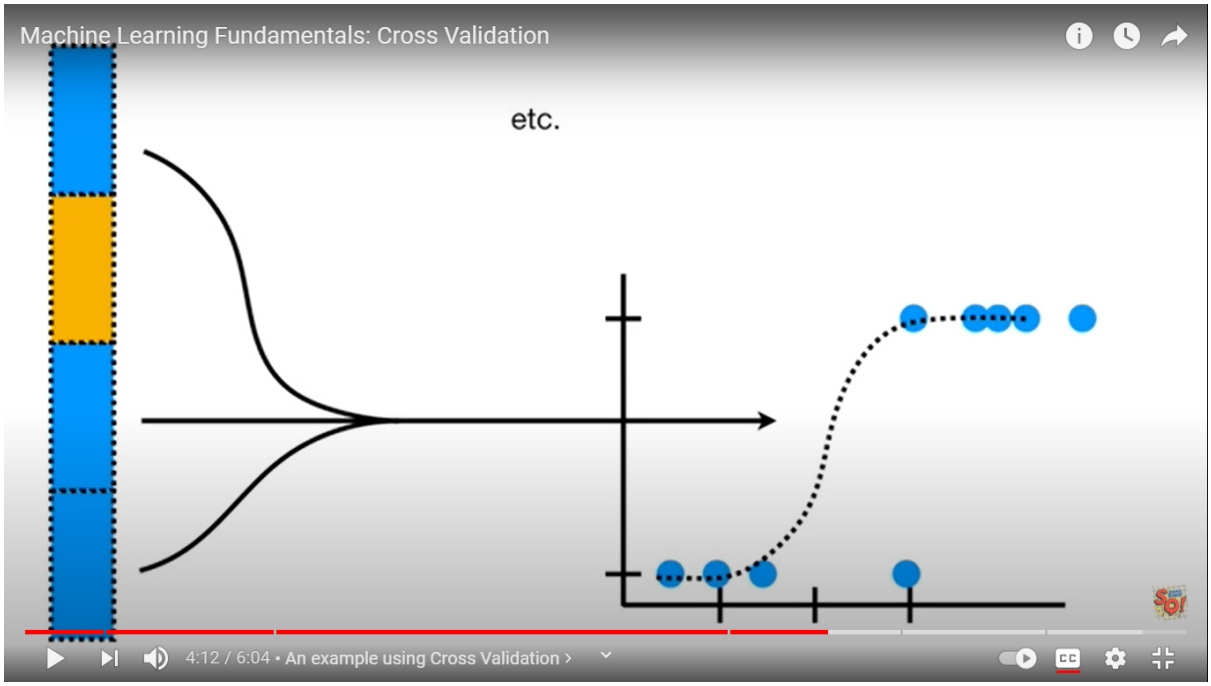

Machine Learning Fundamentals: Cross Validation
Then use the last block
to test the method.
3:52 / 6:04 • An example using Cross Validation ›

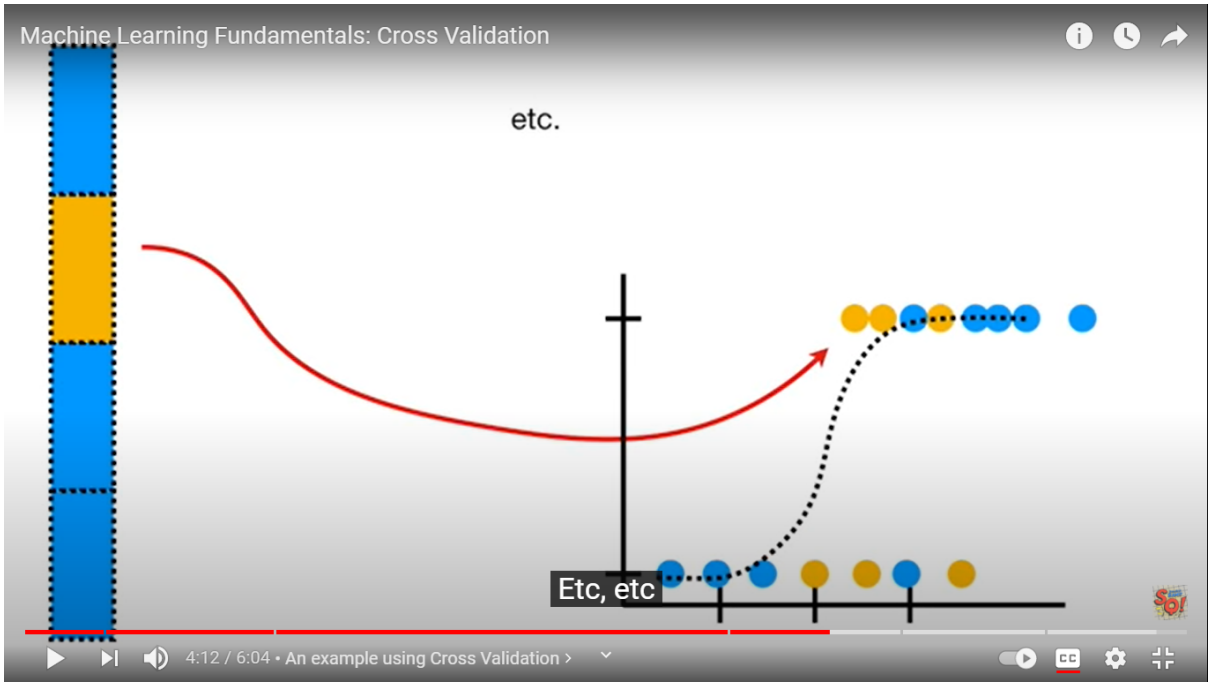

Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
track of how well the
method did with the
test data....
Test data
categorization...
Correct
5
Incorrect
1
3:58 / 6:04 • An example using Cross Validation ›

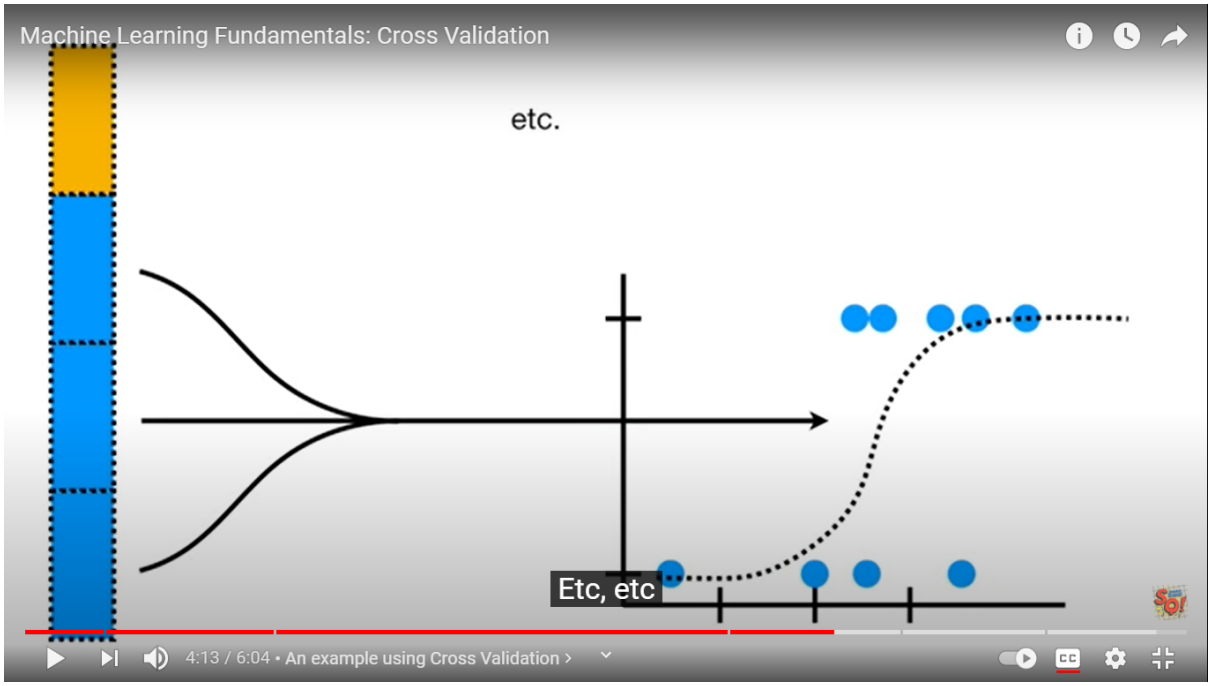

Then it uses this combination of blocks to train the
method...

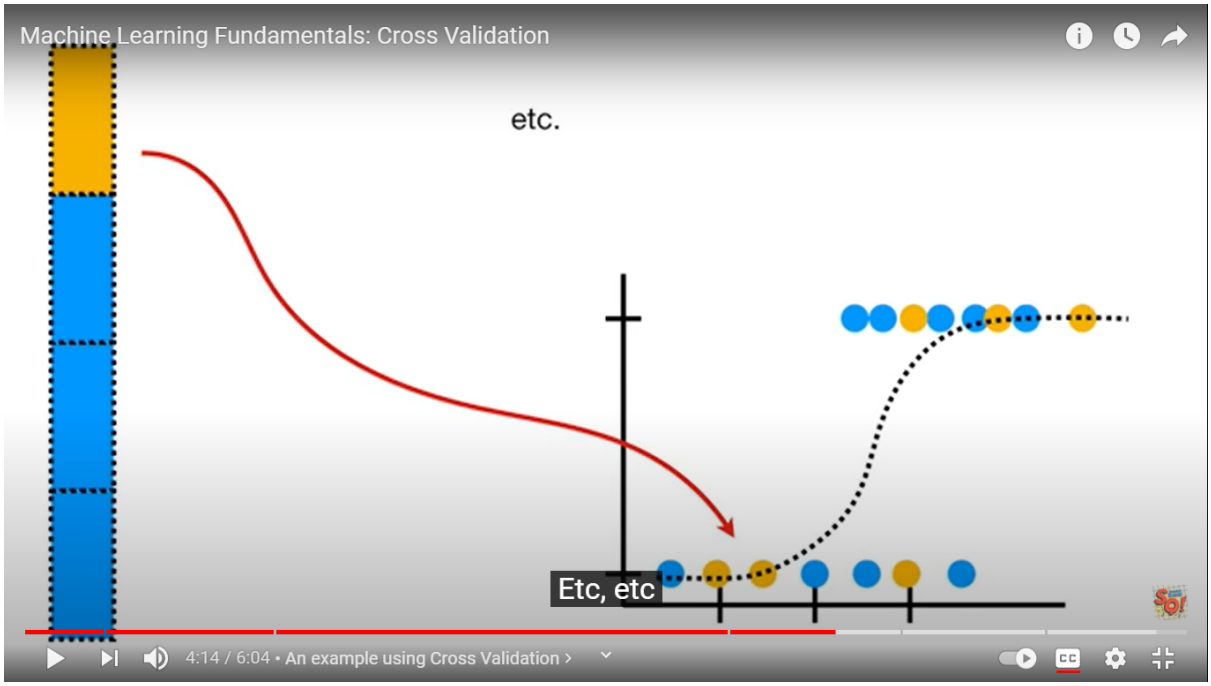

Machine Learning Fundamentals: Cross Validation
Press Esc to exit full screen
…and this block is used for testing…
this block is used for testing and
4:05 / 6:04 • An example using Cross Validation ›


…and then it keeps track of how well the method did with the test data….
Test data categorization…
Correct 4
Incorrect 2
Then it keeps track of how well the method did with the test data, etc

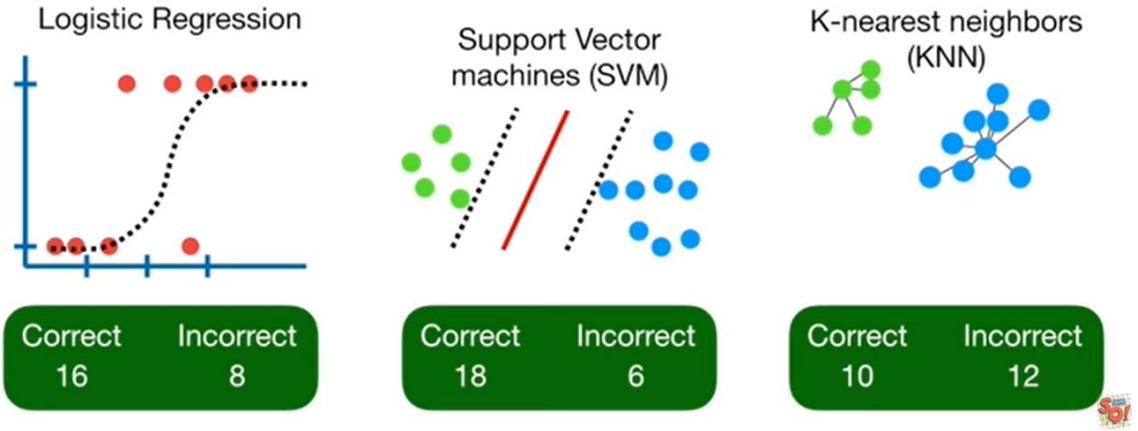

Machine Learning Fundamentals: Cross Validation
etc.
4:12 / 6:04 • An example using Cross Validation ›

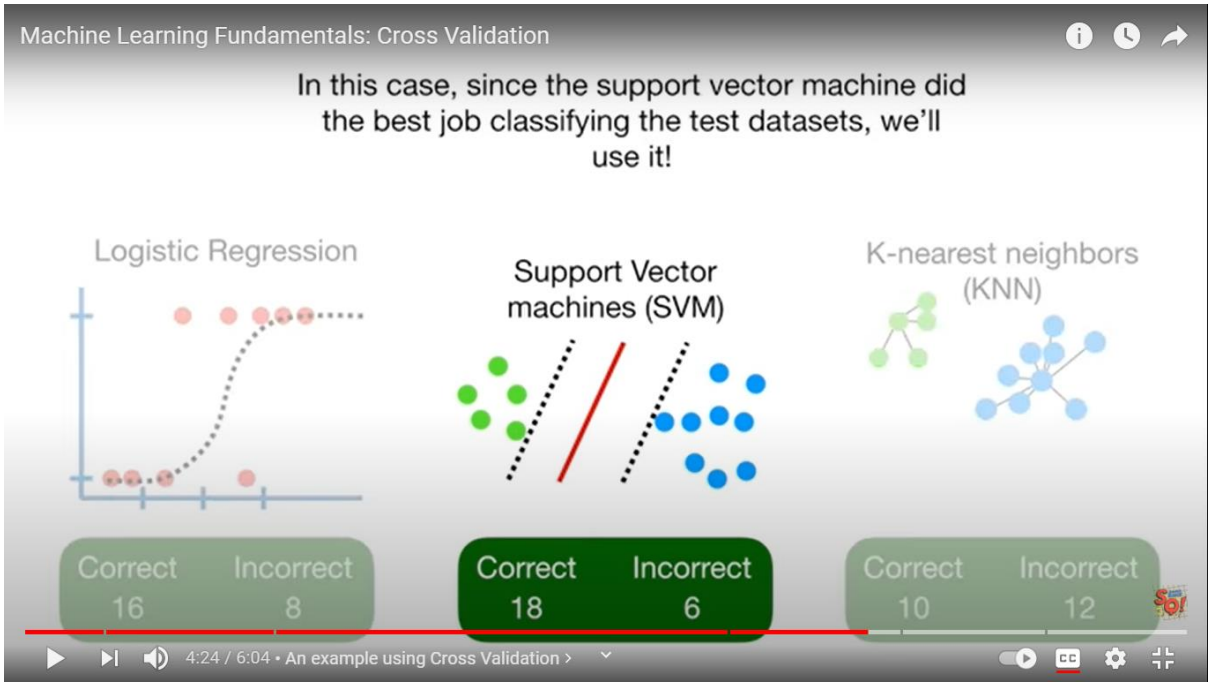

Machine Learning Fundamentals: Cross Validation
etc.
Etc, etc
4:12 / 6:04 • An example using Cross Validation ›


Machine Learning Fundamentals: Cross Validation
etc.
Etc, etc
4:13 / 6:04 • An example using Cross Validation ›


Machine Learning Fundamentals: Cross Validation
etc.
Etc, etc
4:14 / 6:04 • An example using Cross Validation ›

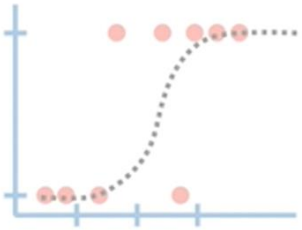
In the end, every block of data is used for testing and we can compare methods by seeing how well they performed.





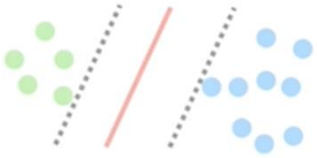
BAM!!!

Logistic Regression



Correct	Incorrect
16	8

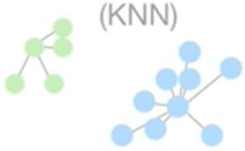
Support Vector machines (SVM)



Correct	Incorrect
18	6

BAM!!!

K-nearest neighbors (KNN)



Correct	Incorrect
10	12

NOTE: In this example, we divided the data into 4 blocks. This is called **Four-Fold Cross Validation**.

However, the number of blocks is arbitrary.

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

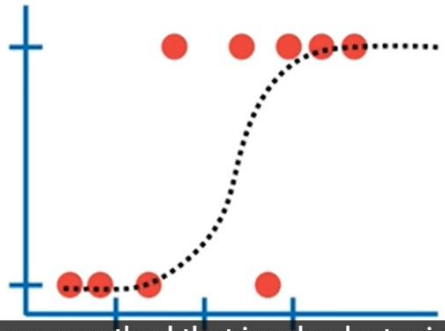
This is called "**Leave One Out Cross Validation**"

This is called "Leave One Out Cross Validation"

That said, in practice, it is very common to divide the data into 10 blocks. This is called **Ten-Fold Cross Validation**.

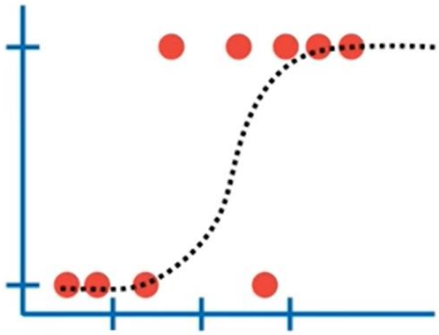
That said in practice it is very common to divide the data into ten blocks. This is called 10-fold cross-validation

Say like we wanted to use a method that involved a "**tuning parameter**" - a parameter that isn't estimated, but just sort of guessed. (For example, Ridge Regression, has a tuning parameter)…



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

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



The End!!!

Hooray we've made it to the end of another exciting StatQuest if you like this StatQuest and want to see more please subscribe