Cross Validation

Machine Learning and Data Analytics using Python

Electronics & ICT Academy, IIT Roorkee

09-13th September, 2019

Prof. R. Balasubramanian
Presented by Himanshu B. and Puneet K.

Courtesy: YouTube StatQuest with Josh Starmer
"Machine Learning Fundamentals: Cross Validation"

Cross Validation

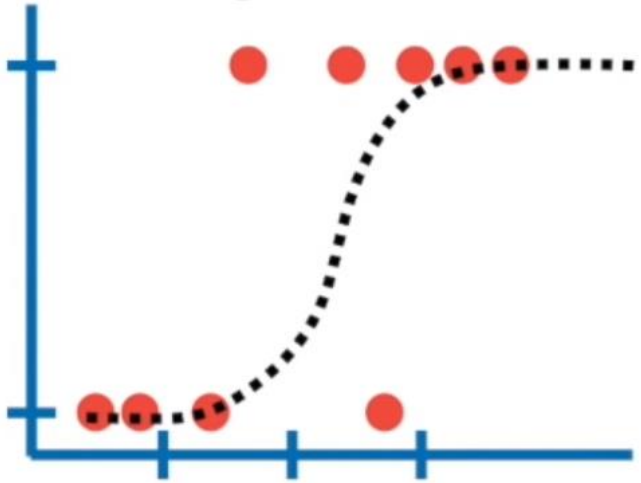
…and predict if they have heart disease or not.

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
…	…	…	…	…

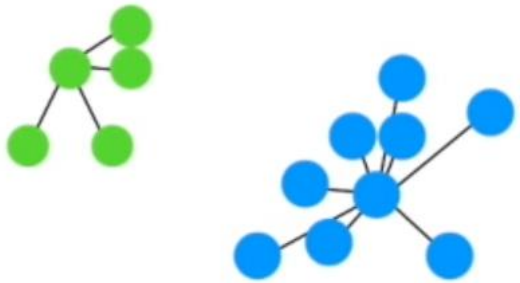
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	???

Cross Validation

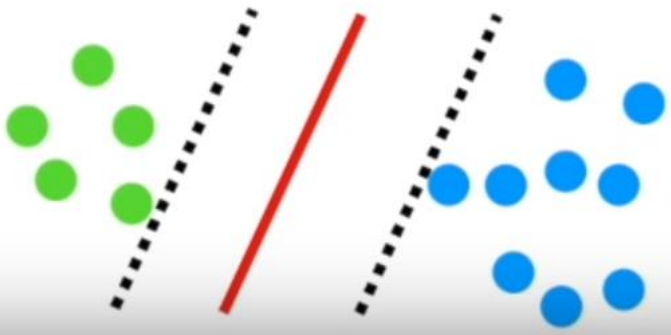
We could use Logistic Regression…



…or K-nearest neighbors…



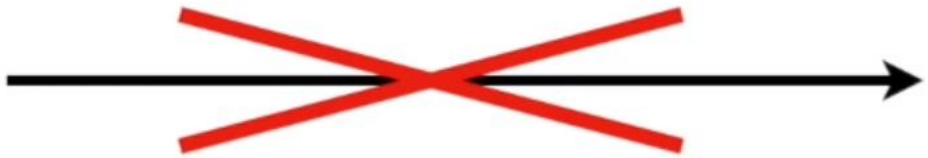
…or support vector machines (SVM)…

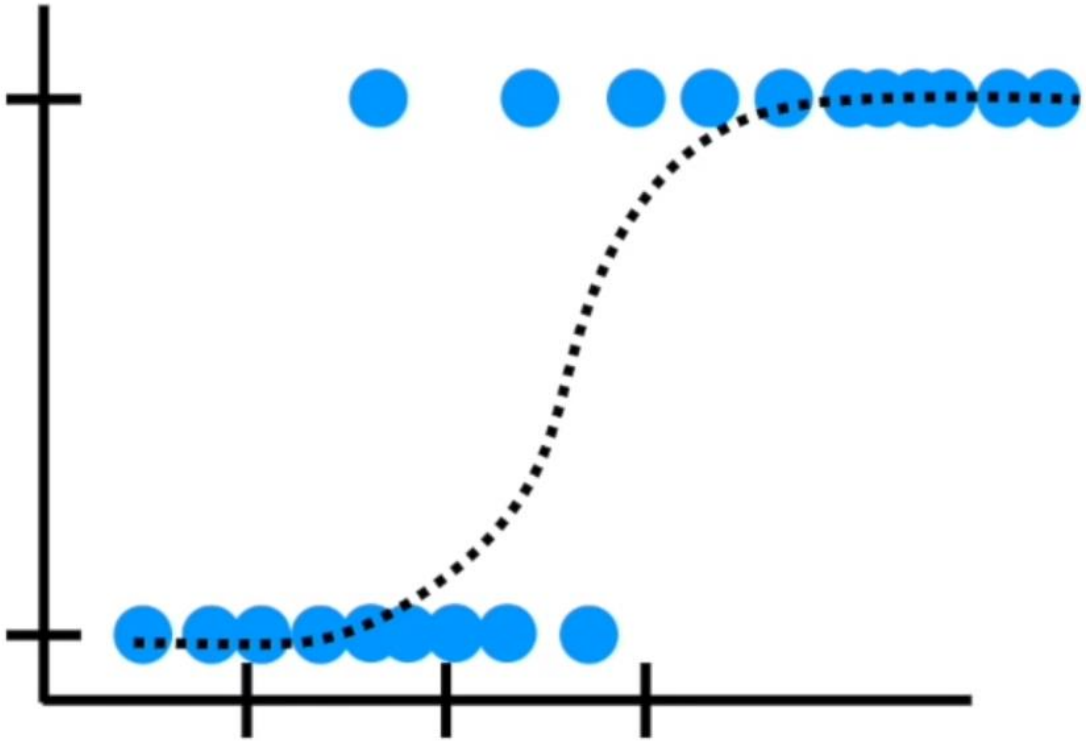


Cross validation allows us to compare different machine learning methods and get a sense of how well they will work in practice.

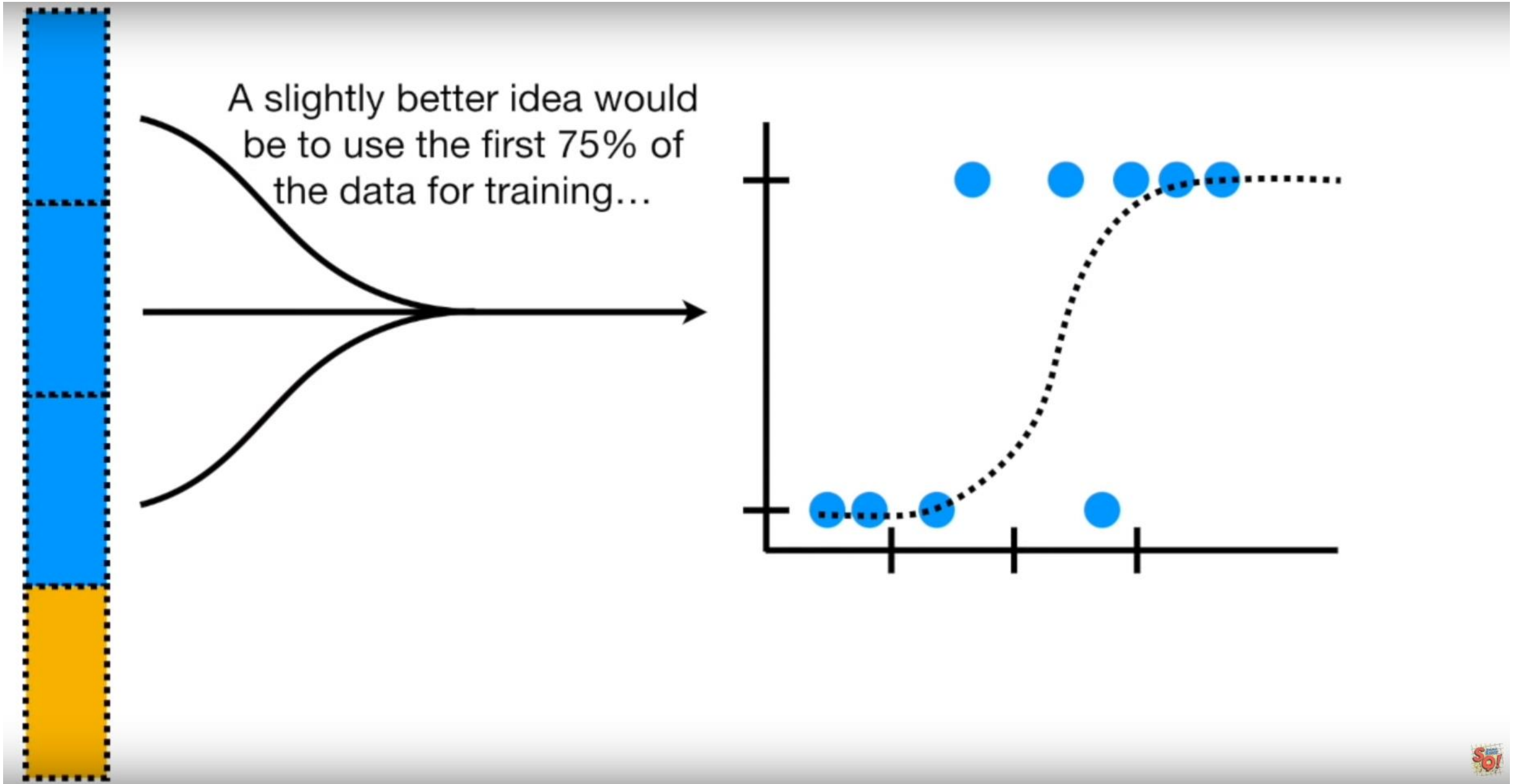
Cross Validation

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.

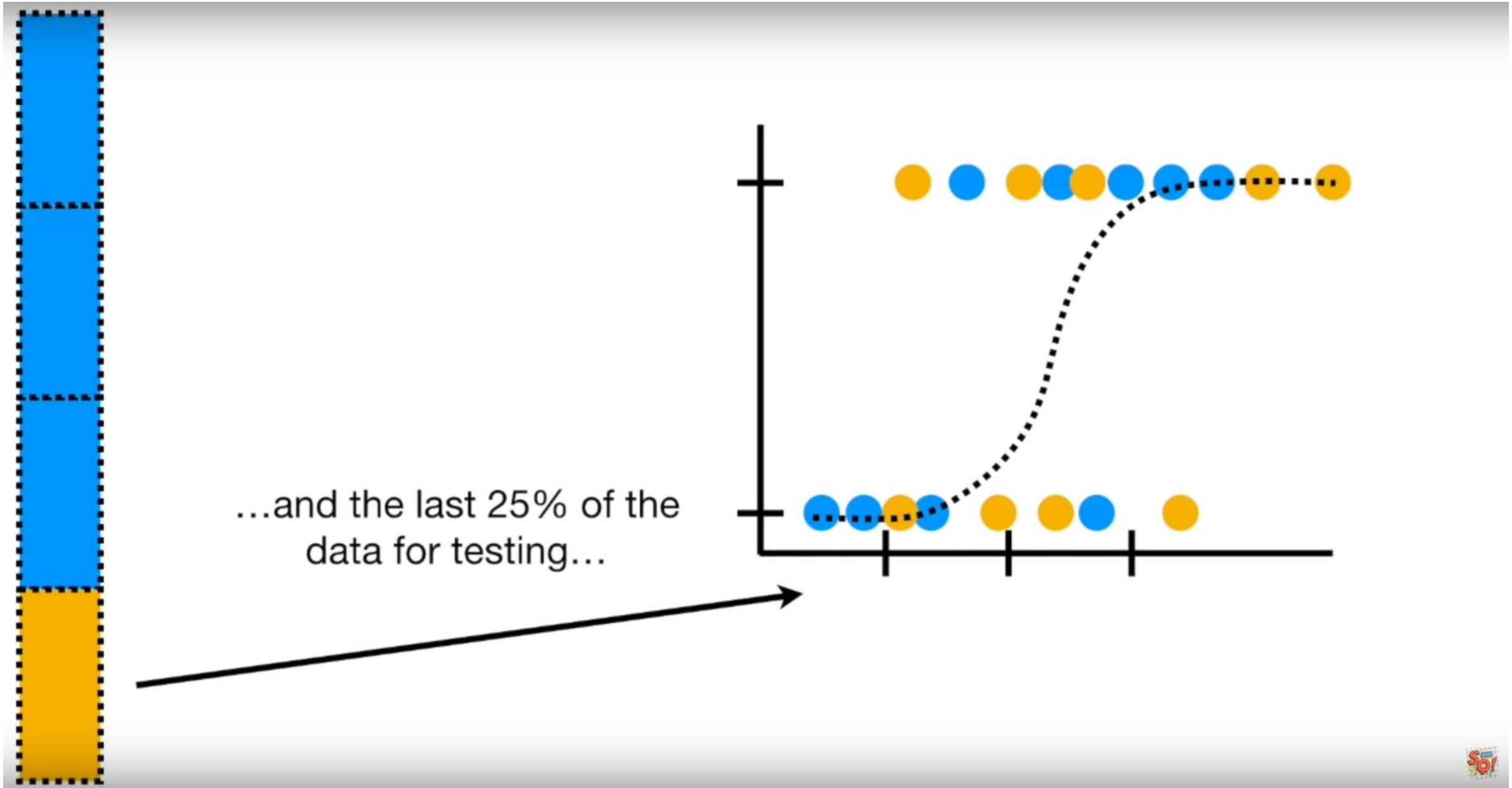




Cross Validation



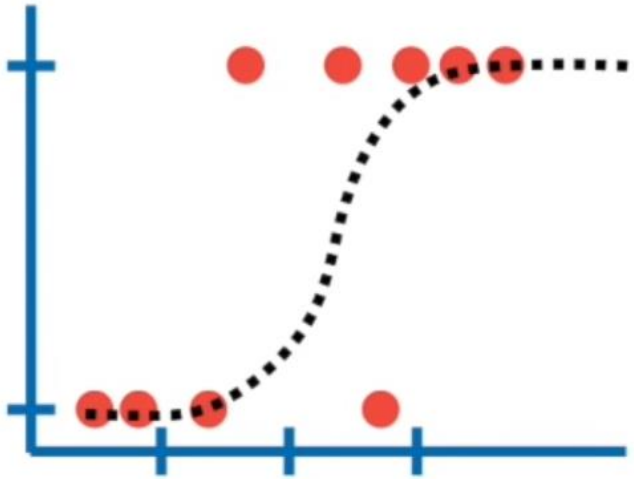
Cross Validation



Cross Validation

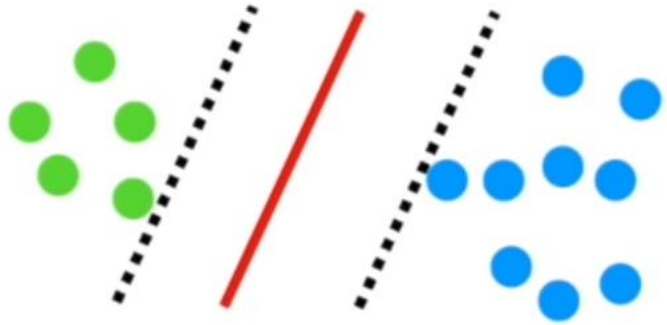
In the end, every block of data is used for testing and we can compare methods by seeing how well they performed.

Logistic Regression



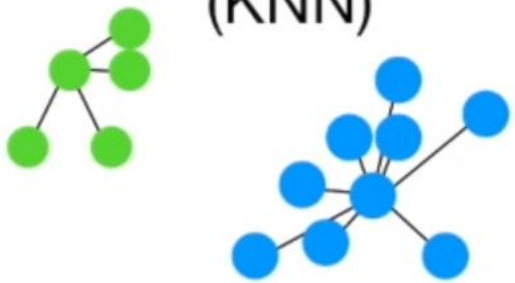
Correct	Incorrect
16	8

Support Vector machines (SVM)



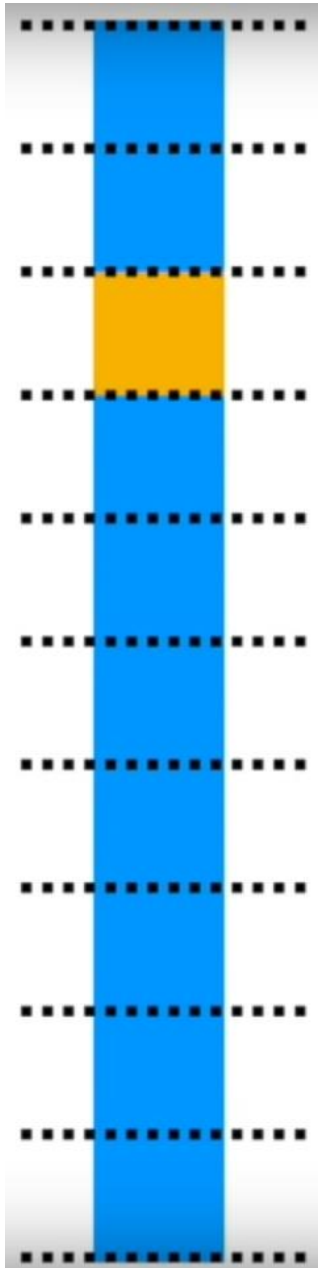
Correct	Incorrect
18	6

K-nearest neighbors (KNN)



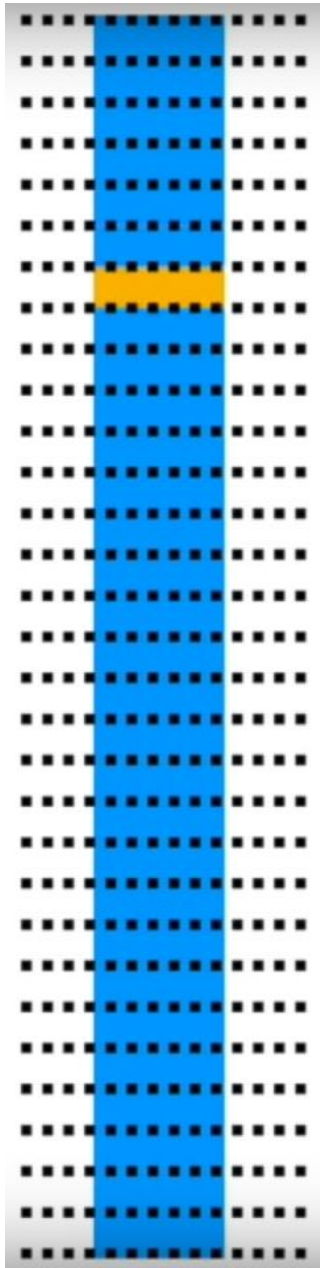
Correct	Incorrect
10	12

Cross Validation



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

Cross Validation



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

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

Thank you

Prof. R. Balasubramanian
Presented by Himanshu B. and Puneet K.