**Training and Testing Data**

From my previous GitHub repository, we already know what are machines what is machine learning and what are algorithms and how ml models can be trained to form decisions based on past experience. However, we still don't know what is training and testing data so whenever we get a data it is always split into training data and testing data. Training data is generally the first split of data that is the initial reserve of data you used to develop your model provides the training data. For example: let's say our model is trained to block emails with suspicious email id, subject line and keywords like dear friend, free, PayPal, casino, payment, bankruptcy, winner, invoice etc. But what if it comes like “PayPal has received your payment for casino royale purchased on eBay”. But since we already discussed that it’s a false positive case hence new rules or modifications must be added in the model so that it can be trained to form better decisions. After successfully training the data you can now test the model on the remaining data known as testing data

So we always need to remember that whenever we have a data we need to split it into training data and testing data. For example, if we take training data is equal to 0.70 then automatically the testing data would be 0.30. If you just mention any one of the parameter then the other parameter will by default be calculated. In the above example we see that the data we got we will train the 70% of the data and test on the remaining 30%.

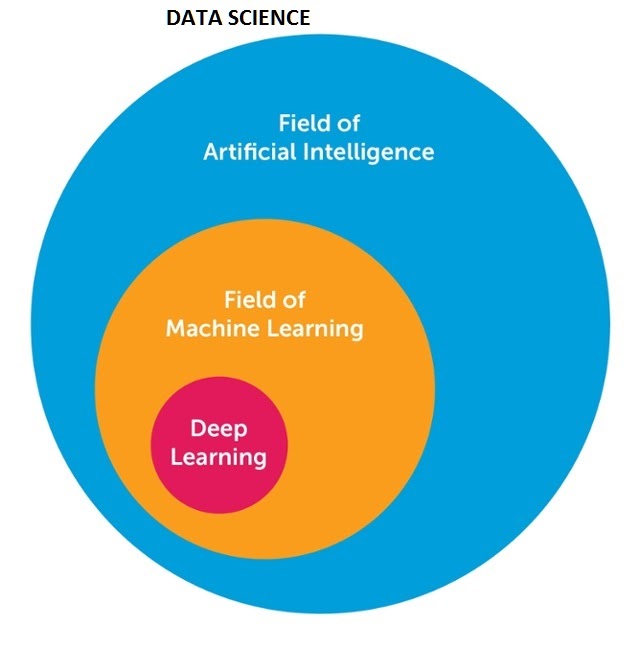
However now the question arises why do we need to divide the data into training and testing? Why can't we train and test on the same data?

This is because at the end of the day machine learning is all about generalizing and being able to apply the patterns in training data to new types of data a system might encounter. If you train your model with the training set and test it with the same set of cost your model will do well! You want to evaluate performance of your model on a set of data it has never seen before.

For example, when we learn and study for a test. Our teacher gives us sample questions and practice question so that we get familiar to the questions that might come in exam. This is because obviously if the teacher will give the same questions that he/she wants to give in the paper then all the students will obtain full marks.

This is the same reason why we can't have the same train set and test set.

Now, let’s get familiar with some more terms:



**AI** (Artificial Intelligence) encompasses the ability of machines to perform intelligent and cognitive tasks. AI is driving the development of machines capable of simulating cognitive abilities (mental capability that involves the ability to reason, plan, solve problems, think, comprehend complex ideas, learn quickly and learn from experience)

**Data mining** is just cleaning large datasets to glean valuable insights from the past.

And we already know what is **machine learning**. It focusses on the incremental process of self-learning and data modeling to form predictions about the future from past results and experience, understanding variance and patterns and developing new strategies so as to reduce error and improve accuracy.