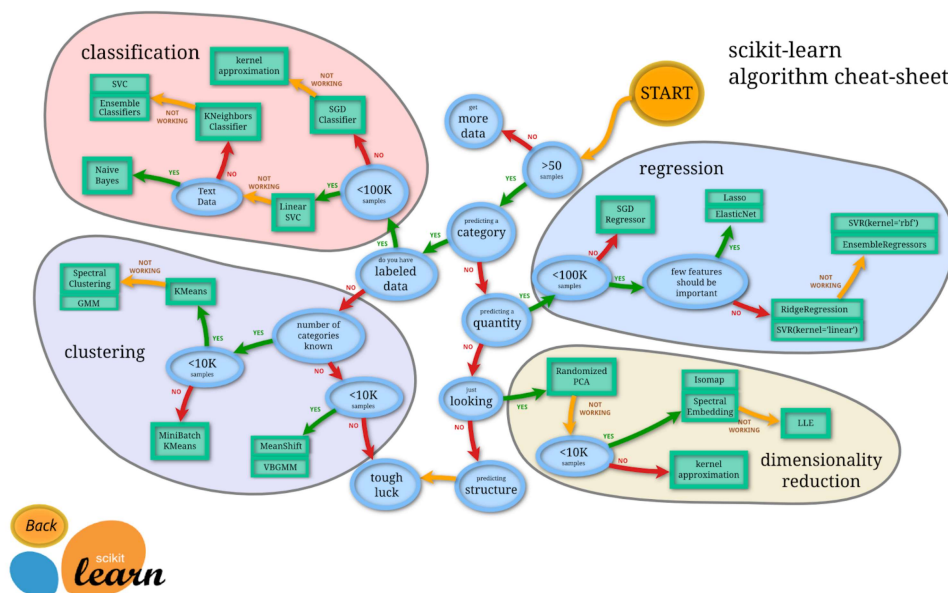


# Module 5-3: Everything About Machine Learning

Ever wonder what kind of machine learning technique to use? Not sure if you should be using clustering, classification, or regression? Well, have no fear! This handy scikit-learn algorithm cheat-sheet will answer all of your questions!

Simply begin at the orange circle labeled START, and let it take you down the path. Each blue circle contains a question that guides you eventually to a kind of machine learning technique. The machine learning techniques are represented by green boxes.



However, notice the first question that is asked.

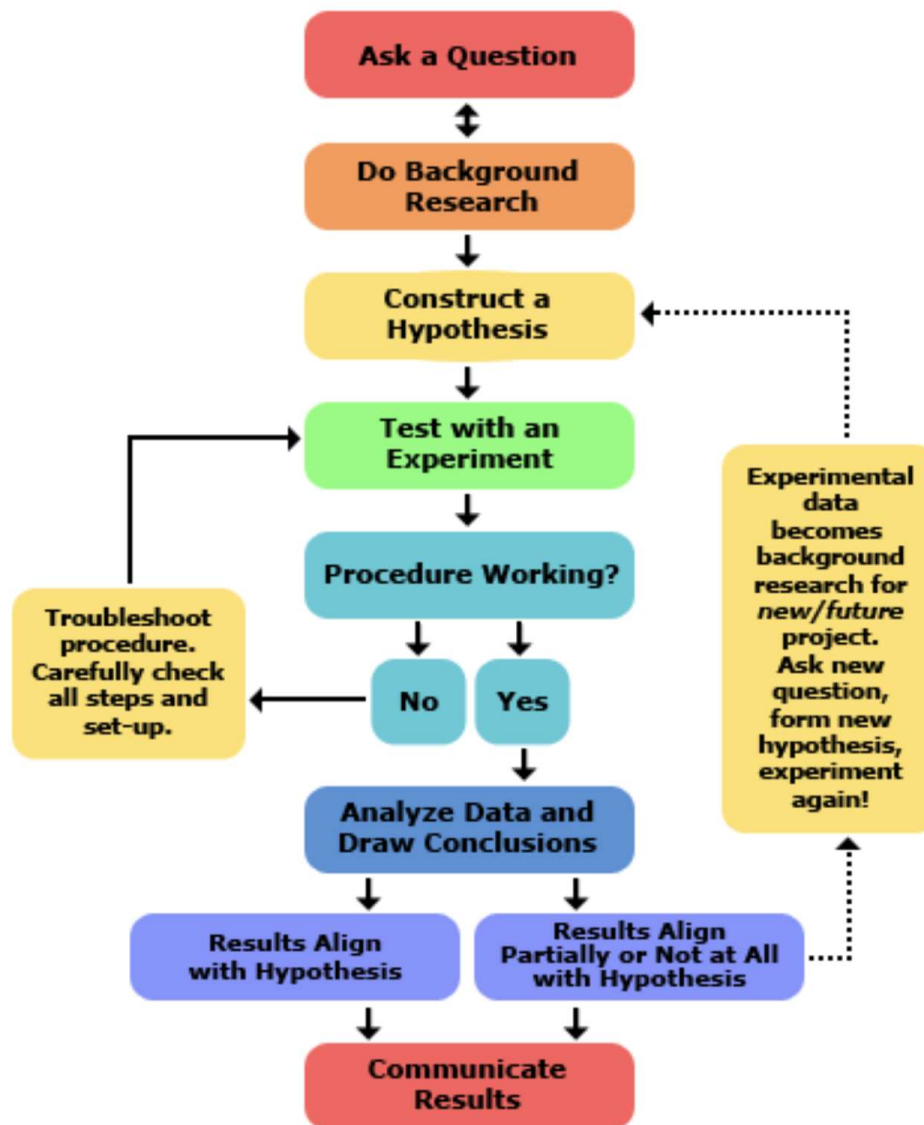


Send Issue

At this point, you may be thinking to yourself "Do I really need more than 50 samples to create a machine learning model?" Well, yes and no.

To answer this question, we need to think about the scientific method. The **scientific method** is a process for experimentation that is used to explore observations and answer questions. Additionally, the scientific method has eight important steps.

Basically, the scientific method is all about finding an answer to a question by repeatedly running experiments on a wide range of data.



If you were to run experiments on just a few pieces of data, there is a good chance it might not include all the cases that your hypothesis was covering.

For example, say that you wanted to find out if chocolate is the favorite flavor of ice cream for people under the age of ten. If you only ask your little brother and sister, you will have only received

feedback from two people under the age of ten. In the United States alone, there are over 35 million people under the age of ten, so by only asking two people for your study, you exclude many different opinions. Just because your siblings might like chocolate does not mean that the 35 million other children will enjoy it as well. You need to cast a wider net and ask more people about their ice cream preferences before you can draw a conclusion.

This same principle applies with machine learning. If you have less than fifty samples, there is a good chance that you will be excluding many different types of data points. It is essential to train the data with tons of data so that it will know how to accurately predict new data. If the machine learning model is not trained with a lot of data, then it may not be able to make the best predictions. Ideally, the model will be trained with a variety of data which will make it flexible and accurate.

Up until this point, we have looked at three different kinds of machine learning techniques. The first kind was regression, which is great for predicting unique values that the model has never seen before. The second kind was classification, which is great for predicting pre defined categorical data. If you already have categories that you defined and what the machine learning algorithm to give back those categories as output, then classification would be perfect. The third kind was clustering, which is great for placing data into groups that are not already defined. Similar to classification, clustering works on placing data into groups, but the group that the data is placed into is entirely dependent on the machine learning model.