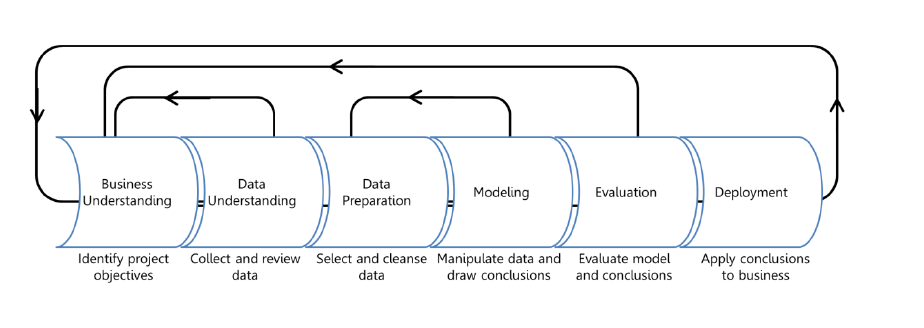
## Intro

Large data sets, often referred to as “**big data**,” provide opportunities and challenges for extracting information and knowledge that are not present within smaller data sets. Big data often contains information which cannot be found by a human simply looking at the data set but can be found through the use of computer algorithms. The analysis of these data sets can lead to very important **knowledge discoveries**. Techniques for analyzing big data often include **machine learning** algorithmswhich are capable of using a few simple rules applied to the whole data set to find patterns. The analysis of big data is valuable almost anywhere big data is present and is a quickly growing professional field. Medical, commercial, and governmental agencies often make use of big data to find new discoveries that improve their efficiency in their respective fields.

## Body of the Article → Examples and Facts



<https://docs.microsoft.com/en-us/archive/blogs/azuredev/the-data-science-process-with-azure-machine-learning>

Machine learning process steps explained:

* Gain a basic understanding of the field where the big data is being drawn from.
  + Computer science experts don’t always know much about the field they’re analyzing since their expertise is in using algorithms so the first step is to learn more about the field where the algorithms need to be applied and what conclusions would be useful to try to find.
* The data understanding phase of the machine learning process consists of finding data that currently exists and seeing what sort of data is within those data sets.
  + For example, the data could be numbers, words, images, videos, or any number of things. In order to start thinking about which algorithms can be used to draw conclusions the data must be understood by the machine learning expert.
  + In order to start thinking about which algorithms can be used to draw conclusions, the data must be understood by the machine learning expert.
* The data preparation phase of the machine learning process includes cleaning of the data using a myriad of techniques.
  + Data preparation will often include filling in missing data values through **data imputation** techniques and turning numbers or symbols into a uniform format that the targeted algorithm for use can understand.
* The modeling step of the data mining process is where machine learning algorithms begin to come into this process.
  + During this step the machine learning expert has already picked an algorithm to use and cleaned the data so that the algorithm can be applied to the data.
  + After running one or multiple algorithms on the data set during this step the machine learning expert should begin to find some patterns and conclusions that could not be found without machine learning algorithms.
* During the evaluation of the machine learning process the machine learning expert can begin to analyze conclusions and determine their potential usefulness.
  + The machine learning expert should take these conclusions and present them to experts in the field of the data to have them verify usefulness and, if the conclusions are useful, the machine learning expert can move onto the deployment step of the machine learning process.
* During the deployment step the machine learning expert provides their model to experts in the field who can begin to use the conclusions drawn by the new model to increase the efficiency of their work.
  + During this final step the machine learning expert should be carefully monitoring their model to ensure it is working correctly and not drawing false conclusions which could lead to catastrophic consequences in some cases.

During the machine learning process there are several points at which the machine learning expert can revert to previous steps in order to avoid failure and ensure the most accurate conclusions are found. At the end of this process the machine learning expert will be likely to find a conclusion or multiple conclusions which experts in the data field will have never found. These conclusions could lead to lives being saved, more products being sold to people who want them, less money being wasted on useless government programs, or many more extraordinarily useful ideas to advance humanity’s knowledge base.

## Try it

Machine learning is complicated but you don’t have to start with the world’s largest data set and most complicated process. You can start by simply finding an excel spreadsheet with a lot of data that you don’t want to analyze manually. Then apply an algorithm to it to see if you can find any patterns. Then, if you have used a small data set for practice, you can check your results manually. There are several good guides on YouTube and the internet in general for getting into machine learning. A good place to start would be this article from geeksforgeeks: <https://www.geeksforgeeks.org/introduction-machine-learning-using-python/>. Another easy way to gain an understanding of machine learning is by downloading rapidminer which is a program that helps to easily implement machine learning algorithms on large data sets. Rapidminer is usually free for students and you can try to download it here: <https://rapidminer.com/>

## Resources

What is machine learning?:

<https://www.youtube.com/watch?v=HcqpanDadyQ>

Python practice project:

<https://www.geeksforgeeks.org/introduction-machine-learning-using-python/>

More advanced Python practice project:

<https://towardsdatascience.com/simple-machine-learning-model-in-python-in-5-lines-of-code-fe03d72e78c6>

Azure Machine Learning Process:

<https://docs.microsoft.com/en-us/archive/blogs/azuredev/the-data-science-process-with-azure-machine-learning>

Rapidminer:

<https://rapidminer.com/>

## Vocabulary

* **Big data**
* **Machine learning**
* **Knowledge discovery**
* **Data imputation**