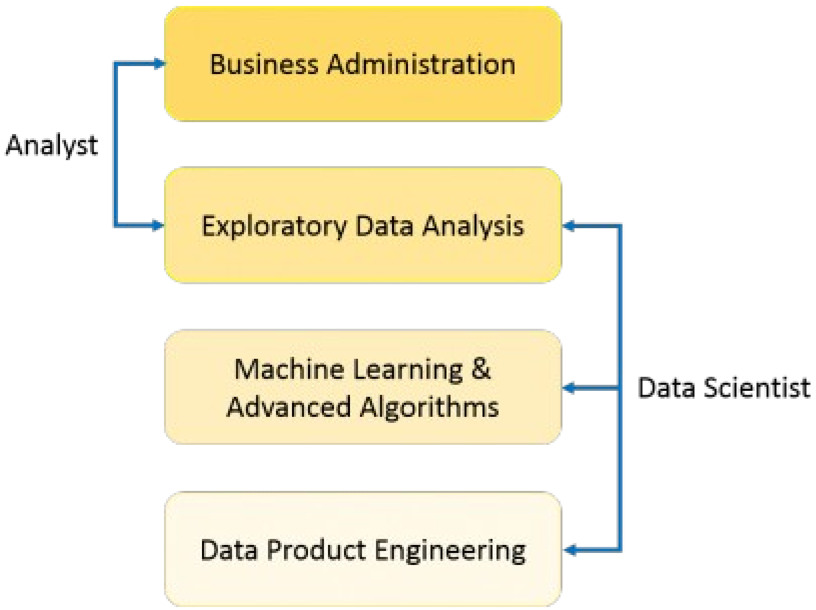
DATA SCIENCE EAST AFRICA

UNDERSTANDING DATA SCIENCE

Day __/20

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.



Analyst

Business Administration

Exploratory Data Analysis

Machine Learning & Advanced Algorithms

Data Product Engineering

Data Scientist

As illustrated in the figure above, a Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future.

A Data Scientist will look at the data from many angles, sometimes angles not known earlier. So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

Who is a Data Scientist?

There are several definitions available on Data Scientists. In simple words, a Data Scientist is one who practices the art of Data Science. The term “Data Scientist” has been coined after considering the fact that a Data Scientist draws a lot of information from the scientific fields and applications whether it is statistics or mathematics.

What does a Data Scientist do?

Data scientists are those who crack complex data problems with their strong expertise in certain scientific disciplines. They work with several elements related to mathematics, statistics, computer science, etc (though they may not be an expert in all these fields). They make a lot of use of the latest technologies in finding solutions and reaching conclusions that are crucial for an organization’s growth and development. Data Scientists present the data in a much more useful form as compared to the raw data available to them from structured as well as unstructured forms.

Differences Between Business Intelligence and Data Science

Business Intelligence (BI) basically analyzes the previous data to find hindsight and insight to describe business trends. Here BI enables you to take data from external and internal sources, prepare it, run queries on it and create dashboards to answer questions like quarterly revenue analysis or business problems. BI can evaluate the impact of certain events in the near future.

Data Science is a more forward-looking approach, an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the aim of making informed decisions. It answers the open-ended questions as to “what” and “how” events occur.

Data Science Lifecycle

A common mistake made in Data Science projects is rushing into data collection and analysis, without understanding the requirements or even framing the business problem properly. Therefore, it is very important for you to follow all the phases throughout the lifecycle of Data Science to ensure the smooth functioning of the project.

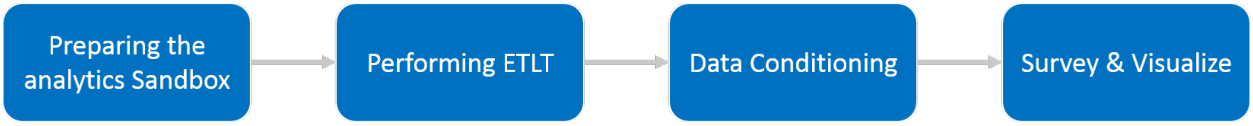
Here is a brief overview of the main phases of the Data Science Lifecycle:

Phase 1 – **Discovery**: Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the project. In this phase, you also need to frame the business problem and formulate initial hypotheses.

Phase 2 – **Data preparation**: In this phase, you require analytical sandbox in which you can perform analytics for the entire duration of the project. You need to explore, pre-process and condition data prior to modeling.

Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox.

Let's have a look at the Statistical Analysis flow below.



You can use R or Python programming languages for data cleaning, transformation, and visualization. This will help you to spot the outliers and establish a relationship between the variables. Once you have cleaned and prepared the data, it's time to do exploratory analytics on it.

Phase 3 – **Model planning** : Here, you will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which you will implement in the next phase. You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

Phase 4 – **Model building**: In this phase, you will develop datasets for training and testing purposes. Here you need to consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing). You will analyze various learning techniques like classification, association and clustering to build the model.

Phase 5 – **Operationalize**: In this phase, you deliver final reports, briefings, code and technical documents. In addition, sometimes a pilot project is also implemented in a real–time production environment. This will provide you a clear picture of the performance and other related constraints on a small scale before full deployment.

Phase 6 – **Communicate results**: Now it is important to evaluate if you have been able to achieve your goal that you had planned in the first phase. So, in the last phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase.

Use Case:

Data Science has also changed the way in which students interact with teachers and evaluate their performance. Instructors can use data science to analyze the feedback received from the students and use it to improve their teaching. Data Science can be used to create predictive modeling that can predict the drop–out rate of students based on their performance and inform the instructors to take necessary precautions.

IBM analytics has created a project for schools to evaluate student's performance based on their performance. Universities are using data to avoid retention supplement the performance of their students.

For example, the University of Florida makes in United State of America use of IBM Cognos Analytics to keep track of student performance and make necessary predictions. Also, MOOCs and online education platforms are using data science to keep track of the students, to automate the assignment evaluation and to better the course based on student feedback.

More Case studies :

https://bigdata-madesimple.com/6-of-my-favorite-case-studies-in-data-science/

Best Wishes,

Regrads Data Science East Africa.