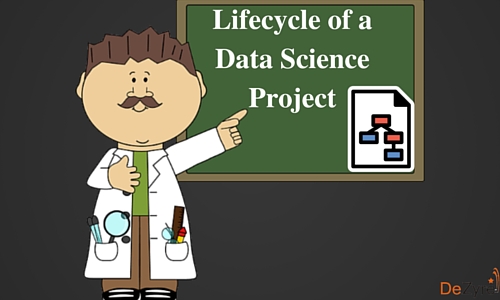
# Life Cycle of a Data Science Project

06 Jun 2016

Source: <https://www.dezyre.com/article/life-cycle-of-a-data-science-project/270>

When working with big data, it is always advantageous for data scientists to follow a well-defined data science workflow. Regardless of whether a data scientist wants to perform analysis with the motive of [conveying a story through data visualization](http://www.dezyre.com/article/why-data-scientists-need-to-be-good-data-storytellers/174) or wants to build a data model- the data science workflow process matters. Having a standard workflow for data science projects ensures that the various teams within an organization are in sync, so that any further delays can be avoided.

The end goal of any data science project is to produce an effective data product. The usable results produced at the end of a data science project is referred to as a data product. A data product can be anything -a dashboard, a recommendation engine or anything that facilitates business decision-making) to solve a business problem. However, to reach the end goal of producing data products, [data scientists](https://www.dezyre.com/article/data-engineer-vs-data-scientist-what-does-your-company-need/262) have to follow a formalized step by step workflow process. A data product should help answer a business question. The lifecycle of data science projects should not merely focus on the process but should lay more emphasis on data products. This post outlines the standard workflow process of data science projects followed by data scientists.



Data science projects do not have a nice clean lifecycle with well-defined steps like software development lifecycle(SDLC). Usually, data science projects tramp into delivery delays with repeated hold-ups, as some of the steps in the lifecycle of a data science project are non-linear, highly iterative and cyclical between the data science team and various others teams in an organization. It is very difficult for the data scientists to determine in the beginning which is the best way to proceed further. Although the data science workflow process might not be clean, data scientists ought to follow a certain standard workflow to achieve the output.

People often confuse the lifecycle of a data science project with that of a software engineering project. That should not be the case, as data science is more of science and less of engineering. There is no one-size-fits-all workflow process for all data science projects and data scientists have to determine which workflow best fits the business requirements. However, there is a standard workflow of a data science project which is based on one of the oldest and most popular-CRISP DM. It was developed for data mining projects but now is also adopted by most of the data scientists with modifications as per the requirements of the data science project.

According to a recent KDnuggets poll on – *“What main methodology are you using for your analytics, data mining, or data science projects?”* CRISP-DM remained the top methodology/workflow for data mining and data science projects with 43% of the projects using it.

Every step in the lifecycle of a data science project depends on various data scientist skills and data science tools. The typical lifecycle of a data science project involves jumping back and forth among various interdependent data science tasks using variety of data science programming tools. Data science process begins with asking an interesting business question that guides the overall workflow of the data science project.

## Standard Lifecycle of Data Science Projects

Data science project lifecycle is similar to the CRISP-DM lifecycle that defines the following standard 6 steps for data mining projects-

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

Lifecycle of data science projects is just an enhancement to the CRISP-DM workflow process with some alterations-

1. Data Acquisition
2. Data Preparation
3. Hypothesis and Modelling
4. Evaluation and Interpretation
5. Deployment
6. Operations
7. Optimization

## 1. Data Acquisition

For doing Data Science, you need data. The primary step in the lifecycle of data science projects is to first identify the person who knows what data to acquire and when to acquire based on the question to be answered. The person need not necessarily be a data scientist but anyone who knows the real difference between the various available data sets and making hard-hitting decisions about the data investment strategy of an organization – will be the right person for the job.

Data science project begins with identifying various data sources which could be –logs from webservers, social media data, data from online repositories like US Census datasets, data streamed from online sources via APIs, web scraping or data could be present in an excel or can come from any other source. Data acquisition involves acquiring data from all the identified internal and external sources that can help answer the business question.

A major challenge that data professionals often encounter in data acquisition step is tracking where each data slice comes from and whether the data slice acquired is up-to-date or not. It is important to track this information during the entire lifecycle of a data science project as data might have to be re-acquired to test other hypothesis or run any other updated experiments.

## 2. Data Preparation

Often referred as data cleaning or data wrangling phase. Data scientists often complain that this is the most boring and time consuming task involving identification of various data quality issues. Data acquired in the first step of a data science project is usually not in a usable format to run the required analysis and might contain missing entries, inconsistencies and semantic errors.

Having acquired the data, data scientists have to clean and reformat the data by manually editing it in the spreadsheet or by writing code. This step of the data science project lifecycle does not produce any meaningful insights. However, through regular data cleaning, data scientists can easily identify what foibles exists in the data acquisition process, what assumptions they should make and what models they can apply to produce analysis results. Data after reformatting can be converted to JSON, CSV or any other format that makes it easy to load into one of the [data science tools](https://www.dezyre.com/data-science-in-r-programming-tutorial/ggvis).

Exploratory data analysis forms an integral part at this stage as summarization of the clean data can help identify outliers, anomalies and patterns that can be usable in the subsequent steps. This is the step that helps data scientists answer the question on as to what do they actually want to do with this data.

*“Exploratory data analysis” is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there. — said John Tukey, an American Mathematician*

## 3. Hypothesis and Modelling

This is the core activity of a data science project that requires writing, running and refining the programs to analyse and derive meaningful business insights from data. Often these programs are written in languages like [Python](https://www.dezyre.com/data-science-in-python-online-training/36), [R](https://www.dezyre.com/data-science-in-R-programming-online-training/37), MATLAB or Perl. Diverse machine learning techniques are applied to the data to identify the machine learning model that best fits the business needs. All the contending machine learning models are trained with the training data sets.

## 4. Evaluation and Interpretation

There are different evaluation metrics for different performance metrics. For instance, if the machine learning model aims to predict the daily stock then the RMSE (root mean squared error) will have to be considered for evaluation. If the model aims to classify spam emails then performance metrics like average accuracy, AUC and log loss have to be considered. A common question that professionals often have when evaluating the performance of a machine learning model is that which dataset they should use to measure the performance of the machine learning model. Looking at the performance metrics on the trained dataset is helpful but is not always right because the numbers obtained might be overly optimistic as the model is already adapted to the training dataset. Machine learning model performances should be measured and compared using validation and test sets to identify the best model based on model accuracy and over-fitting.

All the above steps from 1 to 4 are iterated as data is acquired continuously and business understanding become much clearer.

## 5. Deployment

Machine learning models might have to be recoded before deployment because data scientists might favour Python programming language but the production environment supports Java. After this, the machine learning models are first deployed in a pre-production or test environment before actually deploying them into production.

## 6. Operations/Maintenance

This step involves developing a plan for monitoring and maintaining the data science project in the long run. The model performance is monitored and performance downgrade is clearly monitored in this phase. Data scientists can archive their learnings from a specific data science projects for shared learning and to speed up similar data science projects in near future.

## 7. Optimization

This is the final phase of any data science project that involves retraining the machine learning model in production whenever there are new data sources coming in or taking necessary steps to keep up with the performance of the machine learning model.

Having a well-defined workflow for any data science project is less frustrating for any data professional to work on. The lifecycle of a data science project mentioned above is not definitive and can be altered accordingly to improve the efficiency of a specific data science project as per the business requirements.