Lesson3. Data Science Lifecycle

Lesson Objectives:

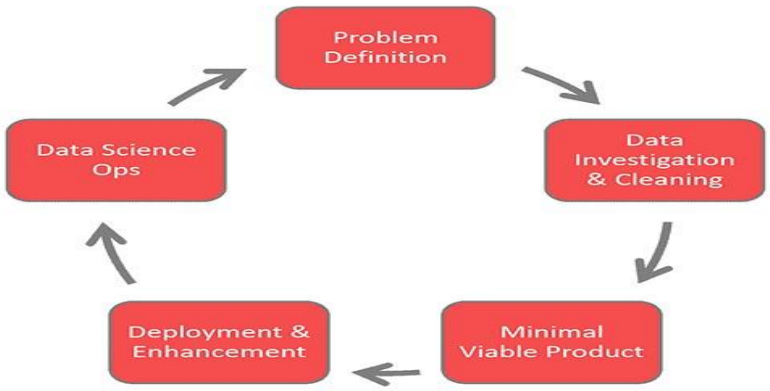
At the end of this lesson, you will be able to:

1. To understand the data science lifecycle
2. To learn the role of virtualization in a cloud environment

---

Data Science Lifecycle

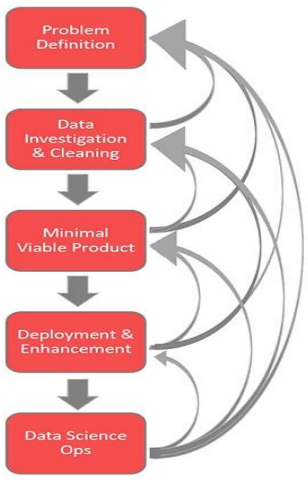
The datascience-pm.com explained that a **data science life cycle** is an iterative set of steps you take to deliver a data science project or product. Because every data science project and team are different, every specific data science life cycle is different. However, most data science projects tend to flow through the same general life cycle.



Towardsdatascience.com states that it is important to understand each section well and distinguish all the different parts. Specifically is very important to understand the difference between the *Development stages* versus the *Deployment stage*, as they have different requirements that need to be satisfied as well the business Aspect.

Most Data Science projects have similar work-flow/ structure that you can use to structure your projects.

Some data science life cycles narrowly focus on just the data, modeling, and assessment steps. Others are more comprehensive and start with business understanding and end with deployment.



And the one we'll walk through is even more extensive to include operations. It also emphasizes agility more than other life cycles.

This life cycle has five steps:

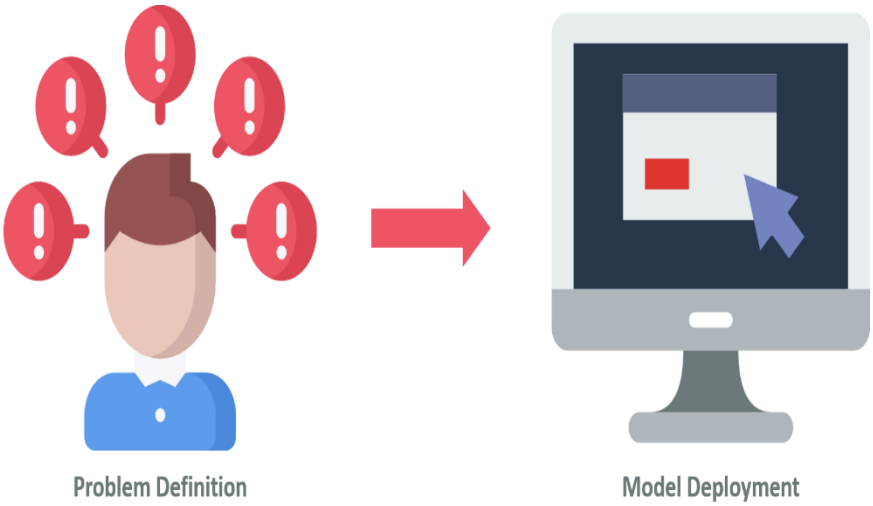
1. Problem Definition
2. Data Investigation and Cleaning 3. Minimal Viable Model
4. Deployment and Enhancements
5. Data Science Ops

I. Problem Definition

Just like any good business or IT-focused life cycle, a good data science life cycle starts with "why". If you're asking "Why start with why?"

Generally, the project lead manages this phase. Regardless, this initial phase should:

- State clearly the problem to be solved and why
- Motivate everyone involved to push toward this why
- Define the potential value of the forthcoming project
- Identify the project risks including ethical considerations
- Identify the key stakeholders
- Align the stakeholders with the data science team
- Research related high-level information
- Assess the resources (people and infrastructure) you'll likely need
- Develop and communicate a high-level, flexible project plan
- Identify the type of problem being solved*
- Get buy-in for the project



Identify the type of problem being solved. Many assume that advanced data science methods are the solution. This is often not the case. Therefore a key question you should ask throughout the project (and especially in the early phases) is: "Is the problem best solved by machine learning or something else?"

1. Clearly identify the root cause of a problem, by identifying the true, underlying problem.
2. Develop a detailed problem statement that includes the problem's effect on the targeted client /customer.

We identify the root cause of the problem *by collecting information and then talking with stakeholders and industry experts*. Then followed by *Merging existing research and information from your stakeholders* can offer some insight into the problem and possible high-level solution. Consider data sources that could help you more clearly define the problem. Starting by doing a **literature review**, and if necessary, surveys in the data science community.

II. Data Investigation and Cleaning

Without data, you've got nothing. Therefore, the need to identify what data is needed to solve the underlying problem. Then determine how to get the data:

- Is the data internally available? -> Get access to it
- Is the data readily collectable? -> Start capturing it
- Is the data available for purchase? -> Buy it
- Once you have the data, start exploring it. Your data scientists or business/data analysts will lead several activities such as:
  - ✦ Document the data quality
  - ✦ Clean the data
  - ✦ Combine various data sets to create new views
  - ✦ Load the data into the target location (often to a cloud platform) ✦ Visualize the data
  - ✦ Present initial findings to stakeholders and solicit feedback

This where the Business problem is defined, understanding the business problem is key in coming up with a good model because it makes you understand the Business objective. Now that we have a business understanding, we can define the success criteria of the project, the success criteria can be based on what we are currently doing. You need to see if the project is viable in longer-term and does it give you a business advantage to really to take it forward.

Now one thing very obvious, but still extremely important to denote is that the low barriers to entry internet marketplace where anyone has great information available have made the whole market extremely competitive in a business point of view. Every one of us is facing competition, therefore only by continually measuring and **tracking the right metrics and thinking about how to move the metrics to improve your performance we can increase our chances to come out on top.**

III. Minimal Viable Model

All data science life cycle frameworks have some sort of modeling phase. However, I want to emphasize the importance of getting something useful out as quickly. This concept borrows from the idea of a Minimal Viable Product.

What is a Minimal Viable Product?

According to Eric Ries viable product is the minimum viable product is that version of a new product which allows a team to collect the maximum amount of validated learning about customers with the least effort.

Extending this concept to modeling, the minimal viable model is the version of a new model which allows a team to collect the maximum amount of validated learning about the model's effectiveness with the least effort

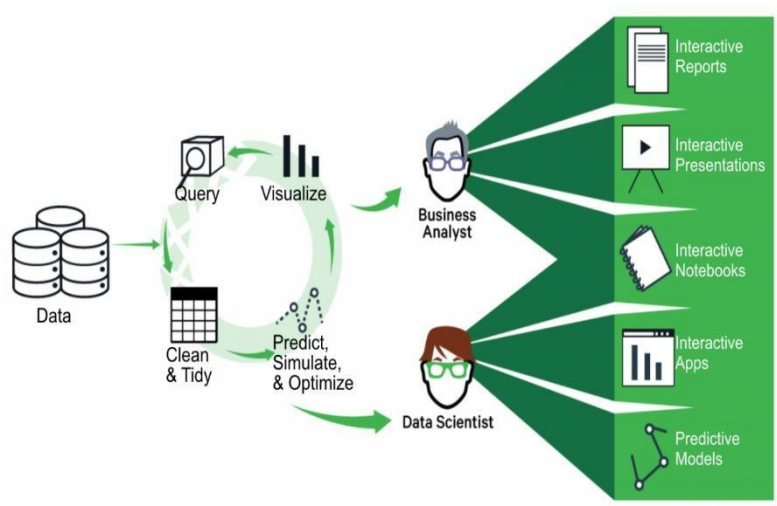
Minimal: The model is narrowly focused. It is not the best possible model but is sufficient enough to make a measurable impact on a subset of the overall problem. Collect the maximum amount of validated learning about the model's effectiveness: Develop a hypothesis and test it. This validated learning confirms or denies your team's initial hypotheses. It has two main parts:

- Is the model technically performing better than the baseline?
- Is the model able to make a meaningful impact to the underlying business problem?
- **Least effort:** Full-fledged deployments are typically costly and timeconsuming. Therefore, find the simplest way to get the model out.

IV. Deployment and Enhancements

Deployment

Many data science life cycles include "Deployment" or a similar term. This step creates the delivery mechanism you need to get the model out to the users or to another system. It's key because no machine learning model is valuable, unless it's deployed to production



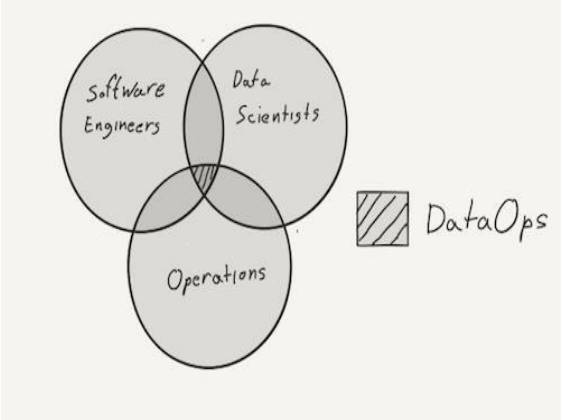
This step means a lot of different things for different projects. Any short-cuts taken in earlier the minimal viable model phase are upgraded to production-grade systems. Typically the more "engineering-focused" team members such as data engineers, cloud engineers, machine learning engineers, application developers, and quality assurance engineers execute this phase.

Enhancements

This is portion is not as common in other data science life cycle frameworks. The importance of getting something basic out and then improving it. While great as a starting point, the model probably isn't as good as it should be. So use the time that the engineers need to deliver the model to improve the models. Conceptually, this "Enhancements" phase means to:

- Extend the model to similar use cases (i.e. a new "Problem Definition" phase)
- Add and clean data sets (i.e. a new "Data Investigation and Cleaning" phase)
- Try new modeling techniques (i.e. developing the next "Viable Model")

V. Data Science Ops



Most other data science life cycles end with a Deployment phase or even before that with Assessment. However, as data science matures into mainstream operations, companies need to take a stronger product focus that includes plans to maintain the deployed systems long-term.

There are **three major overlapping facets of management** to this.

1. Software Management

A productized data science solution ultimately sits as part of a broader software system. And like all software systems, the solution needs to be maintained. Common practices include:

- Maintaining the various system environments
- Managing access control
- Triggering alert notifications for serious incidents
- Executing test scripts with every new deployment
- Meeting service level agreements (SLAs)
- Implementing security patches

To help bridge the development to production gap, organizations are rapidly adopting a DevOps culture which includes general principles that can also guide many aspects of data science operations. However, software system maintenance is a necessary but not sufficient for data science. There's a broader (and trickier) set of considerations

2. Model and Data Management

Data science product operations have additional considerations beyond standard software product maintenance:

- **Monitor the Data:** The data comes from the "real world" which is beyond your control and presents unique challenges. Therefore, validate that the incoming data sets are of expected format and that the data comes in acceptable ranges.
- **Monitor Model Performance:** Software functionality tends to be binary — it works or it doesn't. However, models are probabilistic. So you often can't say definitively whether the model "is working". However, you can get a good feel by monitoring model performance to check against unacceptable swings in core metrics such as standard deviation or mean average percent error.
- **Run A/B Tests:** Models can drift to become worse than random noise. They can also (nearly) always be improved. Therefore, during operations, continue to routinely hold-out small portions of your population as a control group to test performance against the running model. Occasionally, develop and deploy new test models to measure their performance against the incumbent production model.
- **Ensure Proper Model Governance:** Regulations in certain industries require companies to be able to explain why a model made certain decisions. And even if you're not in one of these regulated industries, you will want to be able to trace the specific set of data and the specific model used to evaluate specific outcomes.

3. On-going Stakeholder Management

On-going stakeholder management is critical to your product's success. Continue to educate your stakeholders and set expectations that the model isn't magic. To drive adoption, communicate realistic benefits and if needed, provide training to end users. Likewise, warn stakeholders of the risks and shortcomings of the models and how to mitigate these.

Lesson4. The Application of Data Science

Lesson Objectives:

At the end of this lesson, you will be able to:

1. To understand application of data science

Data Science Applications

According to simplilearn.com Data science has found its applications in almost every industry.

Healthcare

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

Gaming

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

Image Recognition

Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.

Recommendation Systems

Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.

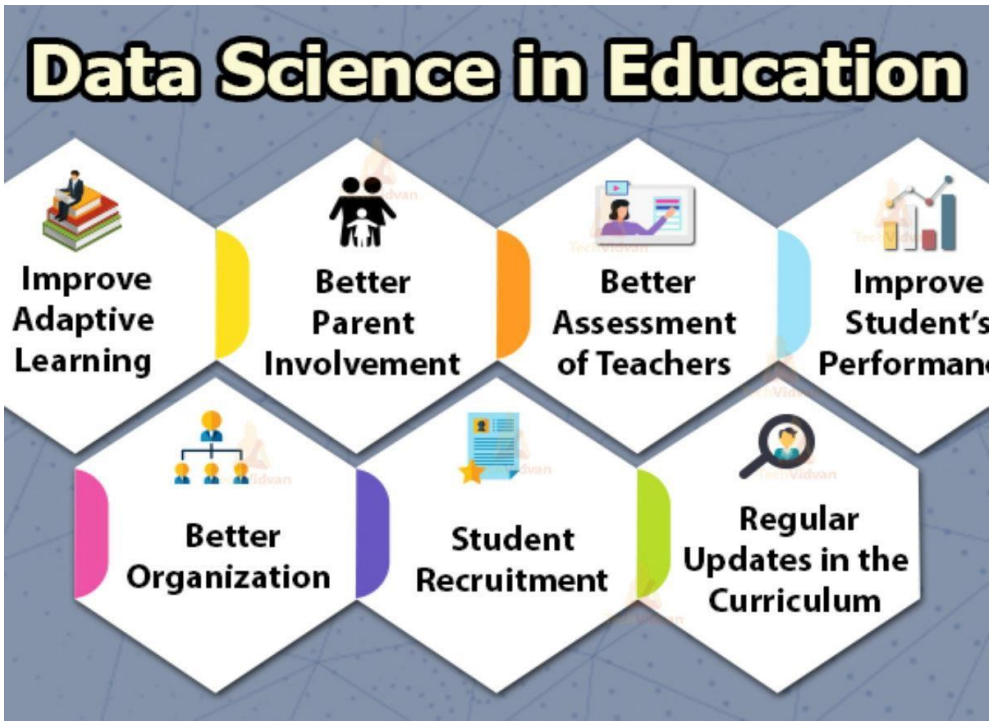
Logistics

Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

Fraud Detection

Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

The Use of Data Science for Education



According to devsaran.com data science is the use of scientific methods to collect and process data. It is an interdisciplinary field that makes use of all forms of data. Data science machine learning makes use of mathematics, computer science, information science, and statistics. It is a growing and promising field of knowledge that has already affected many industries, including education.

The application of data science tools is immense, especially when it comes to numerical representation of academic results and outcomes. Big data analytics may help in resolving problems in education through the ability to focus on specific groups of students. However, we have to admit that we do not have enough big data in education yet. The opportunities it provides can be assessed judging from its successful use in other fields. Moving towards having more data analytics in education is helpful at least in determining how effective certain policies and initiatives are.

The use of data analytics has increased over the past years, and the best thing about big data is that it can be highly useful in any field, including education. Schools and universities have stored enormous amounts of data, including academic records, tests results, grades, etc. However, little has been done to this amassed data due to obsolete privacy policies or limited technical capabilities.

However, big data and analytics of this information could do a big favor to the student learning enhancement. The adoption of technologies could promote better data gathering and improve the effectiveness of test score data analysis. For example, modern data science tools could analyze the level of knowledge of algebra across the entire country, splitting data into federal, state, regional, and local levels. Here are some of the advantages of using big data in education:

1. Improve Adaptive Learning

Each student learns differently. This is a bit of a challenge when it comes to selecting which methods to use in a classroom, but big data can help teachers implement adaptive learning techniques. Big data tools help teachers to teach students according to the individual abilities of students.

2. Better Parent Engagement

Teachers can use big data analytics to track and evaluate a student's performance. Parents can use this information to help resolve any issues that may be affecting their child's performance. Most importantly, educators and families should understand that there is nothing more informative than the data collected over the years. It may trigger significant positive changes in the educational system that will lead to a decrease in public costs and an improvement in the overall levels of knowledge children have. Data analytics can provide valuable insight into the efficiency of specific academic initiatives and the factors that enhance learner's success.

3. Better Assess Teachers

Big data makes it easier for administrators to monitor and assess teachers by helping you to determine which methods and teachers are most effective. The information you gather can also be used to highlight the strengths and weaknesses of teachers during performance evaluations. School archives usually stored a broad range of learner data, including previous learning activities, grades, attendance, parental income, family, health concerns, extracurricular, talents, etc. Universities also collected data on the scientific interests of students and their institutional statistics. Most of the schools worldwide do very little with the wealth of data they have at hand. For example, these aggregated data could serve as a fundamental basis for a comprehensive social, economic or demographic research at the local level.

4. Improve Student Performance

Because you can track the test scores of each student, it is easy to assess a student's performance. With this information, you can then try to make changes that will benefit the student and find out if the student ever asked that desperate question "Does anyone can do my assignment for me?" to help them with projects. If a student's performance deteriorates, big data can help the teachers determine the cause of the problem. The existing data can be integrated into more advanced data research if we employ social networks. Learning management systems could be turned into something similar to social networks, where analysis of educational outcomes will occur as naturally as Facebook analyses out preferences. However, it takes time to develop and implement such big data analytics solutions.

5. Better Organization

From an organizational point of view, big data is equally useful, because it can help schools to become better organized. Big data and analytics can use the help summary to improve how an educational institution organizes logistics, human resources, and business operations. Education is a complex field, and before taking any big data analytics action, it is important to evaluate the analytics models and understand in which context they do not work or provide invalid information. However, challenges and problems should not prevent the adoption of learning analytics. Instead, they should shape the way for the use of data science in Education.

6. Student Recruitment

All educational facilities, from elementary schools to universities, can use big data in their enrollment efforts by using it to find out which educational programs are best suited for incoming children. To learn big data and advance in Education analytics should become a top priority for new Data Science graduates. Young and experienced specialists should be interested in promoting the idea of learning big data in Education. This field is so new that it promises a wide range of scientific breakthroughs and meaningful conclusions. Moreover, the youth interested in data science will have an opportunity to provide a dramatic structural change in education that will affect thousands of data science graduates in the future.