Machine Learning Process Lifecycle

Talat Iqbal, Luke Kumar, Shazan Jabbar, Sankalp Prabhakar
University of Alberta, Applied Scientists, Alberta Machine Intelligence Institute
talat@amii.ca, luke@amii.ca, shazan@amii.ca, sankalp@amii.ca

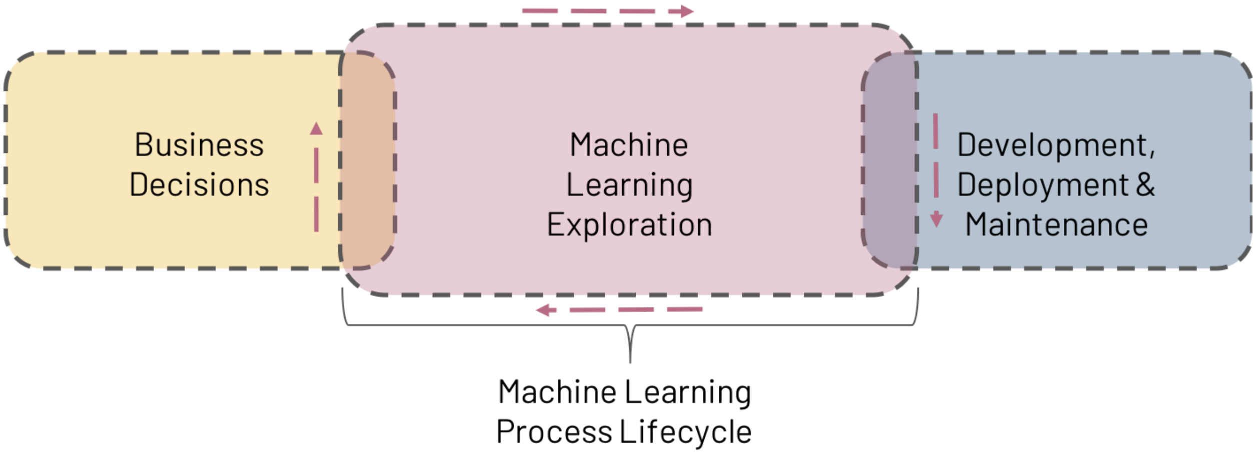
Machine Learning has been gaining a lot of popularity due to its usefulness and its ability to learn from the data. Unlike regular software development, where the development lifecycle of a software has been well studied, the machine learning solution development is a fairly new and less understood process. Developing a machine learning solution is usually exploratory, and closely tied with the problem that it addresses and the associated data.

What is MLPL ?

MLPL stands for 'Machine Learning Process Lifecycle' and is a framework that captures the iterative process of developing a Machine Learning solution for a specific problem. Defining and understanding the business domain and the data related to that problem plays a key role in coming up with a good ML solution. As previously mentioned, ML solution development itself is very much an exploratory and experimental process where different learning algorithms and methods are tried before arriving at a satisfiable solution. And almost always, back and forth passes between different stages of this whole process from understanding the problem to coming up with an ML solution is required to meet the business expectations. MLPL tries to capture this dynamic workflow between different stages and the sequence in which these stages are carried out.

Where does MLPL fit ?

When business organizations develop new software systems or introduce new features to existing systems, they go through two major phases: 1) Business analysis - making assessments and business decisions regarding the value and feasibility of a new software product or feature; and 2) Product development - develop the solution, usually following one of the existing software development methodologies, and put it in the production. However, when an organization thinks about adopting machine learning to complement their current software products or services and to address a fresh business problem, there is an additional exploration phase between the business analysis and the product development phases. This exploration phase is usually ad-hoc and to streamline this process, we propose an iterative methodology called Machine Learning Process Lifecycle.



Machine Learning Process Lifecycle is an iterative methodology to execute this Machine Learning exploration task. This is an attempt to generalize the process so that it is flexible and modular enough to be applied to different problems in different domains and at the same time have enough modules to arrive at a solution.

What does an ML Exploration Process entail ?

In an ideal scenario, the organization would want to know what are the different answers that I can get before introducing the ML solution into a system. Although all questions cannot be answered, there are a few questions that this exploration process should be able to answer. The ML solution might not always be the best solution and sometimes, traditional methods that you were using to solve the exact problem might be a better fit but at least you will be able to answer the following questions.

1. Can Machine Learning address my business problem?
2. Is there a supporting data ?
3. Can algorithms take advantage of the data ?
4. What is the value added by introducing Machine Learning ?
5. What is the technical feasibility of arriving at a solution with Machine Learning ?

At the end of this exploration, if one is able to answer the above questions, then the exploration process is in the right direction.

What MLPL does not capture ?

When the organization thinks about Machine Learning, the very first thing that an organization does is to perform some sort of business analysis. This involves, identifying business workflows, business problems, resource assessment, identifying tasks and decision points which an ML solution would fit right in and bring business value. This part of the overall adaptation of ML into a business process is something that MLPL doesn't capture. Furthermore, there is another phase that happens after the exploration is done, where one goes into the phase of developing the solution as a tangible component into your product or service, deploying it into production and maintaining it. This phase is also not captured on MLPL. MLPL only deals with the exploration phase where different methods are tried to arrive at a proof-of-concept solution which can be later adapted to develop a fully-fledged productionable ML-system.

Why do we need MLPL?

We have seen an overview of what MLPL captures and what it does not. But why do we need a process to capture an exploration task that is usually deemed ad-hoc. There are a few important reasons of why we need MLPL.

Risk Mitigation: The MLPL standardize the stages of an ML project and defines standard modules for each of those stages, thereby minimizing the risk of missing out on important ML practices. A good to have check-list is always handy to identify if some of the modules have been implemented or at least not-missed.

Standardization: Standardizing the workflow across teams through an end-to-end framework enables the users to easily build and operate machine learning systems while being consistent and allows the inter-team tasks to be carried out smoothly.

Tracking: This probably might be one of the most important motivations to introduce MLPL into your ML exploration workflow. MLPL allows you to track the different stages and the modules inside each of the stages. This being an exploration task, there are a lot of throwaways in different stages that will never be used in the final ML solution but have been invested in. These throwaways are required to be tracked to document the resources that has been spent on them and to know the lessons learnt for any future iteration.

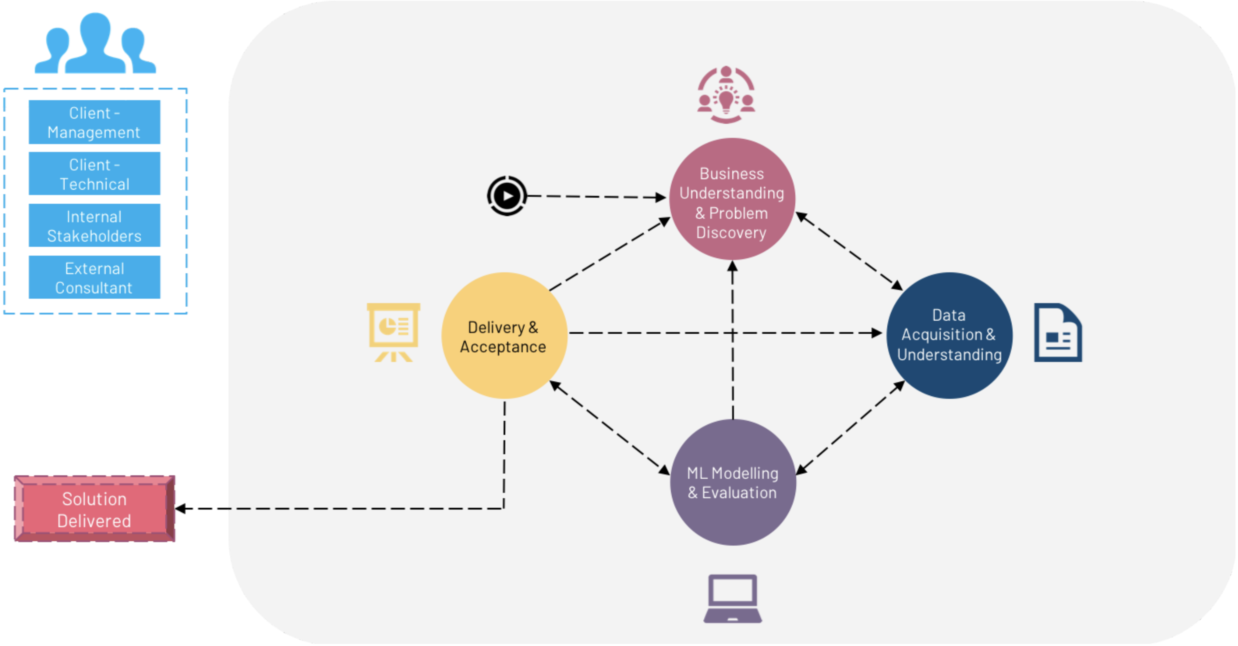
Reproducibility: Having a standardized process allows us to build pipelines for creating and managing experiments that can be compared and reproduced for future.

Scalability: A standard workflow will also allow us to manage multiple experiments simultaneously.

Governance: Well defined stages and modules for each of these stages will help in better audits to assess if the ML systems are designed appropriately and operating effectively.

Communication: A standard guideline helps in setting the expectations and effectively facilitate communication between teams about the workflow of the projects.

MLPL - AN OVERVIEW



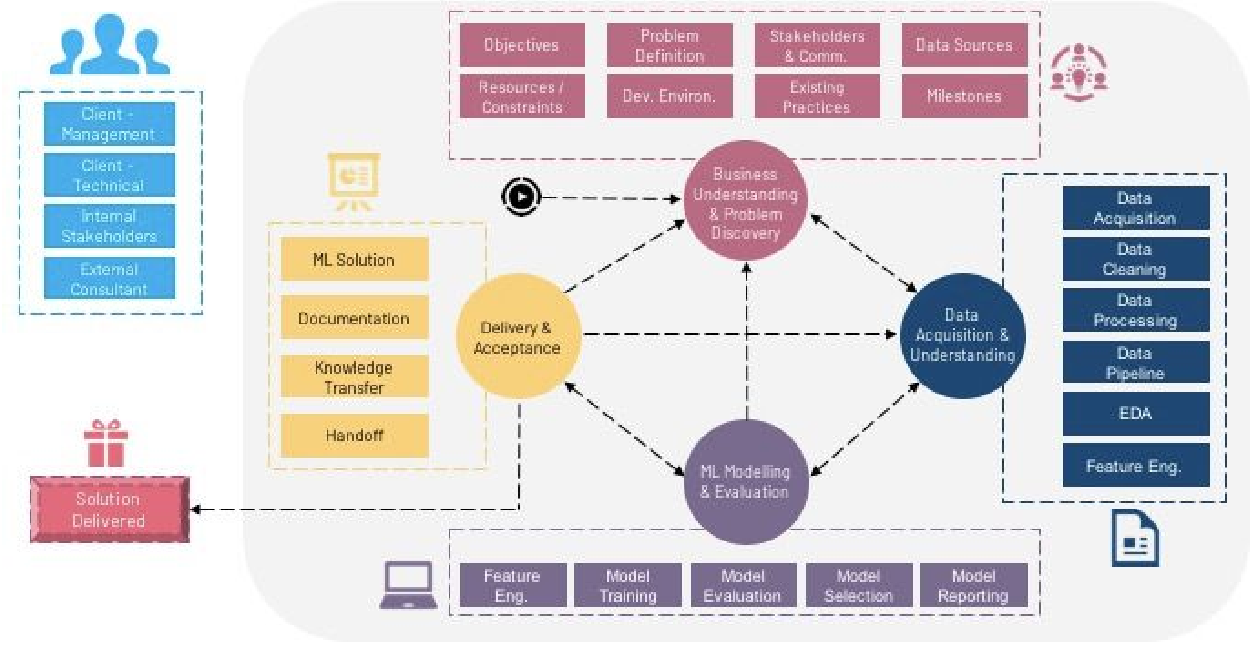
There are 4 stages in the exploration task.

1. **Business Understanding and Problem Discovery (BUPD)** - This stage identifies a business problem by understanding the business requirements and defines a machine learning problem that would help in addressing the business problem. For example, the business problem is to make existing customer consume more content on my streaming methods. One possible way is to recommend them with smart choices about what to consume next.
2. **Data Acquisition and Understanding (DAU)** - The second stage would be to explore the data is available and identifying the abilities and restrictions with the data to use it for machine learning. This would involve an in-depth analysis of the data and its potential.
3. **Machine Learning Modelling and Evaluation (MLME)** - The third stage is the stage where the magical machine learning algorithms come into the process. A

lot of the time, organizations start with this stage assuming this is the only portion of the process that needs to be done to arrive at a solution, but at some point they must go back to the previous two stages as both these stages are critical before thinking about what machine learning algorithm(s) to use and their configurations.

4. **Delivery and Acceptance (D&A)** - This is the final stage where we validate if the machine learning problem is addressing the initial business problem. This stage is not encountered frequently since, especially if there is good communication among all the stakeholders and a clarity in defining the problem. An ideal project should arrive at this only once, but given how quickly a project evolves for various reasons, there is a possibility that this stage might have to be revisited many times.

MLPL IN DEPTH

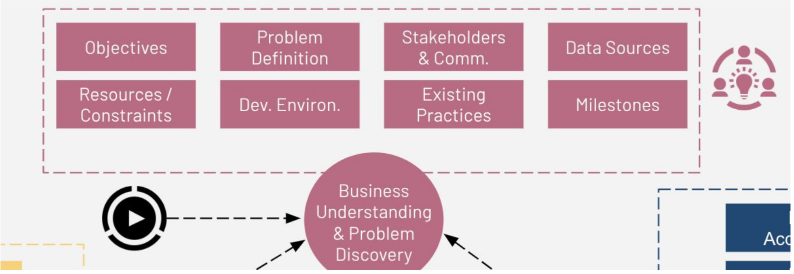


Business Understanding and Problem Discovery

Few key aspects to be taken care of during this stage are:

- **Objectives**: Identify business objectives that machine learning techniques can address.
- **Problem Definition:** Discover the machine learning problem that would help solve the business problem. Sometimes, there is one exact problem to address one machine learning problem and sometimes and bunch of machine learning problems together address the business problem.
- **Data sources:** Identify existing data resources. In the real world, the data typically comes from different sources and has been combined from these sources. Identifying what are the different data sources will help in better narrowing down the data that can be useful. Data sources can be proprietary in-house data, publicly available data or the data that can be bought from third parties.
- **Current practices:** Identify what business process or practices are in place that are addressing the business problem in the current setting. The business problem could be completely a new one or it could be an existing one.

- **Development Environment:** Define development and collaborative environment (code/data repos, programming languages, etc.).



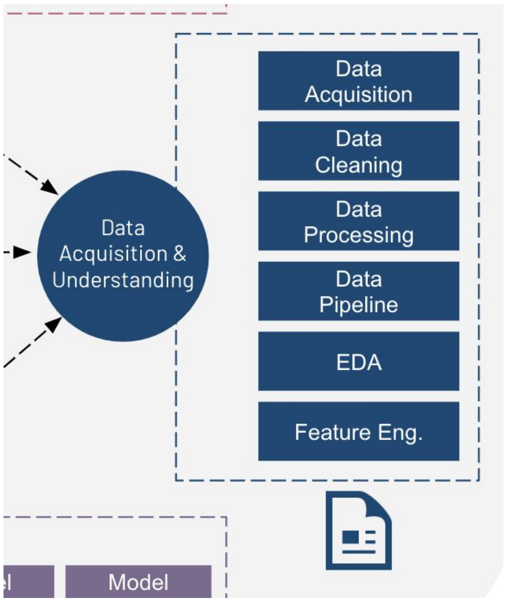
- **Communication**: Agree on methods of communication and the frequency of communication.
- **Milestones**: Define milestones, timelines and deliverables. Sometimes it's not feasible to arrive at definite milestones, given this is an exploration task. But thinking in that direction will help in more structural approach.
- **Resources:** Identify the resources that will be required. The resources can be time, money, data engineers / analysts/ scientists or computational resources.
- **Stakeholders**: Identify internal/external stakeholders and their roles. There are usually multiple stakeholders who should be a part of this process continuously. Management team that decided that ML approach should be tried, a technical team that is actively exploring the solution, teams that would own the different stages of exploration and third parties associated with the development and deployment of the final solution. All the teams involved in each of the stages of MLPL should be on the same page.
- **Constraints**: Identify the constraints that are acceptable for the problem. Do we need machine learning solutions that are interpretable? Is there any part of data that should be removed due to privacy concerns?

By the end of this stage, we would have identified and defined our goals and other aspects that will help us understand the problem better and dive deeper into subsequent stages of MLPL. Any form of tools, worksheets that can be helpful can be used at this stage.

Data Acquisition and Understanding

- **Acquisition**: Acquiring the data is an important task. After all, the whole concept of using machine learning is because of data. At this point, we should be in a situation to have identified the data sources and gathered the data. Once the data sources have been identified, we would want to combine the data sources that help answer the questions defined in the objectives and then consolidate the raw data. In some cases, combining the data sources might not be trivial and requires in-depth domain knowledge and expertise to find out how to align the data and combine them into one data source.

- **Cleaning**: In the real world, the data is usually corrupt due to various reasons. Inaccurate readings from sensors, inconsistencies across various readings and invalid data are some of the data issues that you might find in data. A thorough analysis on how to fix these values with the help of a domain and data expert should be carried out.
- **Processing**: Data that is read is still not consumable by the machine learning process. Data pre-processing might involve various techniques such as normalization, standardization or scaling of the data.
- **Pipeline**: Set up a process to score new data or refresh the data regularly as part of an ongoing learning process
- **Exploratory Data Analysis:** Engage in exploratory data analysis to gain understanding about the data. Understanding the data is very important and could lead to better design and selection of machine learning process. Also, it gives an in-depth understanding of what could be useful information for further steps.

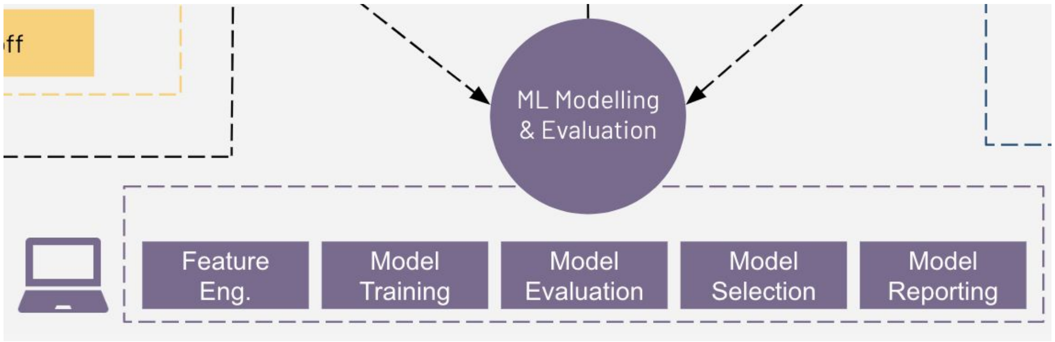


- **Feature Engineering:** Feature Engineering is a continuous process and would occur in various stages of a machine learning process. In the Data Understanding phase, feature engineering might be to identify the features that are irrelevant and do not add any information. For example, in high dimensional data, are there any signals whose variance is close to 0?

- **Data Split:** The data should be split in such a way that there is a portion of data called ‘training data’. Training data is used for training the QuAM and there is a separate portion of data called ‘test data’ to evaluate how good the QuAM is.

MLME

- **Algorithm Selection:** Algorithm Selection is a process of narrowing down a suite of algorithms that would suit the problem and data. With many algorithms across various domains in machine learning, narrowing down helps us to focus on certain selected algorithms and working with them to arrive at a solution.
- **Feature engineering:** This part of feature engineering focuses on preparing the dataset to be compatible with the machine learning algorithm. Data transformation, dimensionality reduction, handling of outliers, handling of categorical variables are some examples of feature engineering techniques.
- **QuAM/Model Training:** Once an algorithm has been selected and data prepared for the algorithm, we need to build the QuAM. A QuAM is the combination of an algorithm and data. In the machine learning world, a QuAM is also referred to as a model. The training part includes using the training data to generalise the QuAM. Hyperparameter tuning involves identifying the best parameters suited for the algorithm and finding the optimal model.



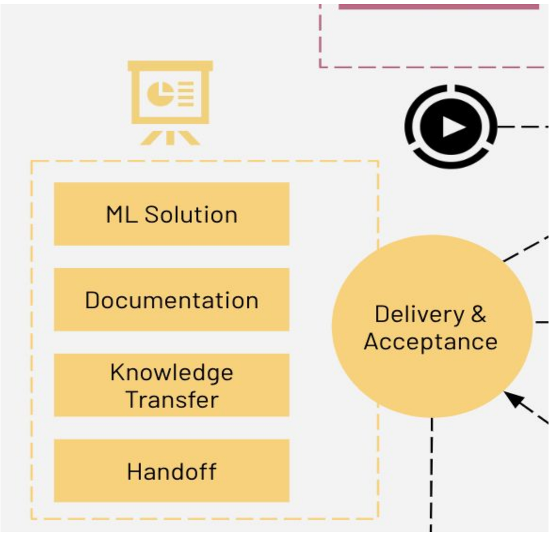
- **Evaluation**: Identifying the evaluation criteria is an important task. How we define the success criteria and the tolerance of making mistakes will determine what the success criteria will look like. If identifying the matching values is the only criteria, then accuracy is a good enough measure. If there is a cost associated with identifying false positive or false negatives, then other measures such as precision and recall can be used. These will be discussed further in this specialization.

- **Refinement**: Refine the model by identifying the best parameters for each of the algorithms on which you have trained the QuAM. This step is called hyper-parameter tuning.

Delivery and Acceptance:

This is the stage where we confirm if the machine learning problem is addressing the business problem. Having a conversation with the client is very important to understand if the business problem is addressed. From the delivery perspective, an ML solution is to be delivered to the client. The solution could be in one or all three of the forms below.

1. **Prototype:** Source code of the prototype is provided along with readme and dependency files on how to use the prototype. The prototype need not necessarily be a production level code but should be clean enough with comments and relatively stable so that the engineering teams can take it and use it to build a product around it.
2. **Documentation:** A good documentation always accompanies a prototype. Some of the technical details should be listed and explained.
3. **Project Report:** This is a complete list of methodologies used and decisions taken along the lifetime of the project and the reason behind those decisions. This gives a high level idea of what was achieved in the project.

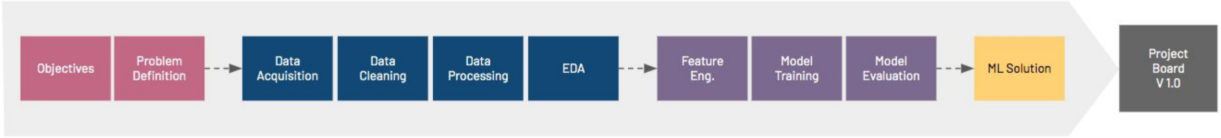


- **Knowledge Transfer:** Identify in-house training required for understanding ML solution and present it to the client. This is the appropriate time to clarify all the questions that would arise regarding the ML solution and a feedback checkpoint with the business prior to incorporating the ML solution into full operation.

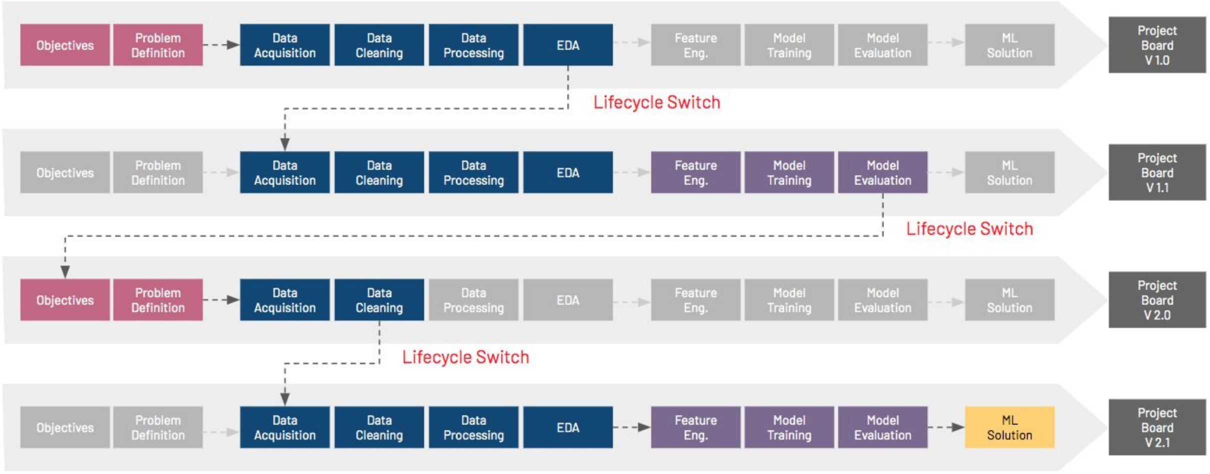
- **Handoff:** Turn over all materials to client, so that the client could take it from here and execute it.

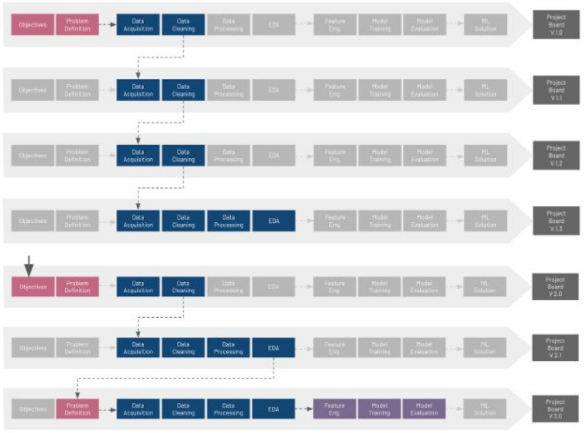
ML Exploration - Expectation vs Reality

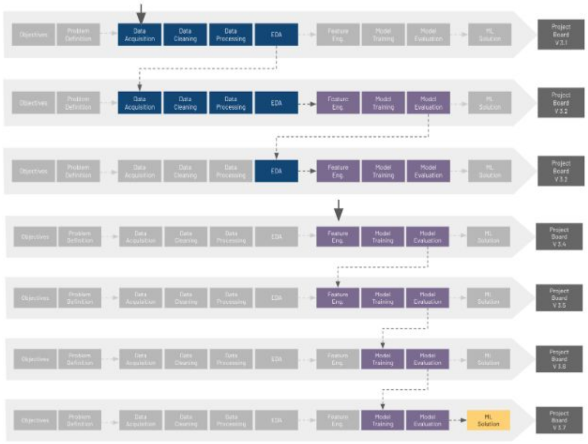
Whenever we do a project, irrespective of the domain or field, all the stakeholders involved would want their projects to tread smoothly and with little obstacles. But this is usually not the case. Specially, if your project is an exploration task, then definitely we should expect a lot of such obstacles.



What we expect is shown above but what typically happens is shown in the below picture. The picture shows how frequently that we would have to apply the brakes on what we do and go back to one of the previous stages or modules in the same stage. This switch to a different stage/module is what we call as a lifecycle switch. A lifecycle switch forces you to revisit some of the modules that you have already done, because remember, it's all about the original business problem. If any changes happen midway, there is a high chance that the other components already visited before might change. One example shown below is that, while doing the EDA, it was identified that the data for only one season was available while trying to develop a universal model for the whole year. In such a case, we would have to acquire more data but there is a possibility that the addition of more has changed a lot of properties of the data. How we compute the missing values, the assumptions we made earlier for data processing and cleaning would also change. Therefore, it is required that we move sequentially without skipping some of the modules.







Reasons for a Life-cycle Switch

- Business objectives change
- Problem defined does not address the business goals
- Data does not address the ML problem
- Data is insufficient
- ML exploration goes in the wrong direction
- ML exploration doesn't give sufficiently good results

Final words

We introduced MLPL, our framework to execute an iterative process when developing machine learning solutions. We discussed different stages of the MLPL that one should go through and introduced sub-components/modules to pay attention to in each of the stages before arriving at a final ML solution. As explained previously, development of an ML solution to a business problem is an iterative and exploratory process and requires a few back-and-forth hops between different stages of the MLPL. This understanding will be key to set the right expectations among the stakeholders involved in the project and to execute a successful ML project in your organization.