Speeding Up the Machine-learning Process with MLOps and Creating a Mechanism to Continuously Provide Service Value

Ei Yamaguchi

Abstract

Machine-learning operations (MLOps) is the machine-learning version of DevOps (development and operations) and represents the concept of how the people in charge of developing machine learning for a system and the people in charge of operating the system can collaborate to ensure smooth progress from implementation to operation of a commercial system. Recently, MLOps has been gaining in popularity; however, each vendor has a different definition of MLOps and there is no unified view. With that issue in mind, this article explains the background and basic concepts of MLOps as well as latest investigations and concrete means of implementing MLOps.

Keywords: machine learning, MLOps, AI

1. Background

The number of projects for launching new services or improving existing operations by using artificial intelligence (AI) is increasing yearly; however, there are many cases in which AI is not fully implemented in actual business.

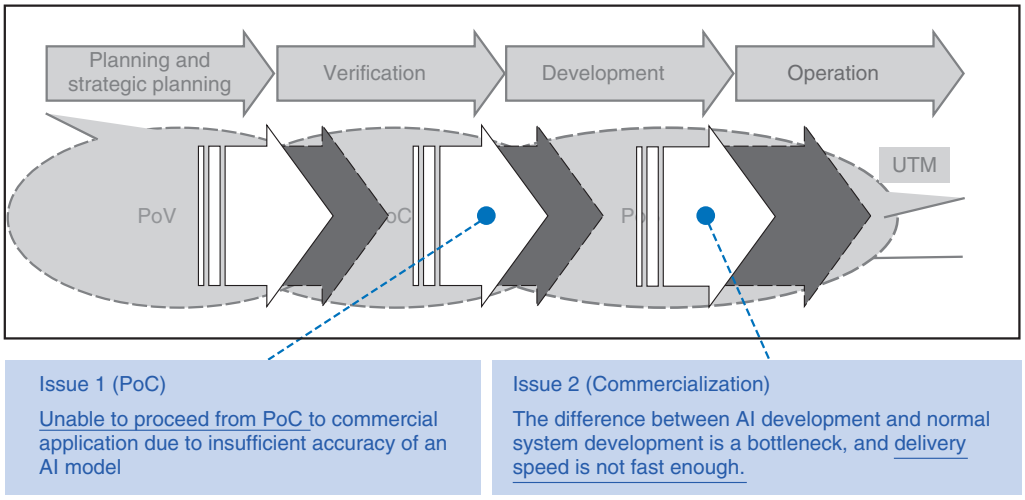
There are two main reasons for this state of affairs: (i) the accuracy of AI cannot be improved to a level that is sufficient for actual operations within a limited proof of concept (PoC) period and (ii) when the system developer takes over an AI model created by a machine-learning engineer and deploys it in a commercial system, it takes time to do it owing to the lack of communication and job splitting between them (**Fig. 1**).

Issues after service deployment include dealing with the phenomena of concept drift and data drift. For example, due to the spread of novel coronavirus (COVID-19) infections over the last year or so, people's behavior has been changing on a weekly, or even daily, basis. Consequently, the accuracy of AI models fine-tuned by data scientists before the change has deteriorated over time, making the models useless. Such a phenomenon is referred to as concept (or data) drift.

As a means of addressing the two issues shown in Fig. 1, a set of practices called machine-learning operations (MLOps) has recently been attracting attention. In this article, the approaches to use MLOps to address these issues are explained.

1.1 Issue 1: Unable to improve accuracy of an AI model to a level that can handle actual operations within the PoC period

To address the first issue, it is considered effective to increase the number of tunings of an AI model within a limited PoC period by streamlining the tuning process. It seems intuitively correct that increasing the number of tunings will improve the accuracy of the model because doing so will incorporate many measures that may contribute to improved accuracy. Two steps are considered effective to make the tuning process, which is a mundane task, more efficient:



PoV: proof of value
UTM: Unified Threat Management

Fig. 1. Issues concerning the machine-learning development process.

first, standardize the process and give it a common language; second, introduce tools that can streamline each step of the process.

For the first step, the analytical framework used in the data-analytics field is considered to be effective. For example, an analytical framework called CRISP-DM (cross-industry standard process for data mining) is available. This framework's process is different from the waterfall development process in that it permits movement between processes in an agile manner (**Fig. 2**). CRISP-DM is compatible with the machine-learning development process in which it is common to improve accuracy of an AI model through trial and error by examining questions such as "What happens if we use such feature values?" and "What happens if we change the algorithm?"

For the second step, open source software (OSS), cloud-service providers, and third-party vendors have released tools to improve the efficiency of machine-learning development. NTT DATA is providing the following services for introducing MLOps to help customers having problems with efficiency of developing machine-learning models introduce tools that can improve the efficiency of each process (**Fig. 3**).

A tool called AutoML is a typical example of a very effective tool that can directly contribute to improving accuracy of machine learning. AutoML selects the most-accurate machine-learning model by simultaneously running multiple algorithms necessary in the machine-learning development process, namely, feature design, model design, and model tuning.

However, even if these tools are used to improve the accuracy of AI, doing so will be meaningless unless the benefit to a customer's actual business can be clearly shown. For that purpose, it is necessary to translate indicators of AI-model accuracy, namely, accuracy and recall rate, into the words used in the customer's actual business so that the impact on their business is demonstrated.

1.2 Issue 2: Development of machine learning for commercial systems requires the cooperation of experts in various roles

An overview of the machine-learning development process and roles of the participating experts is given in **Fig. 4**. Even at the overview level, the participation of a variety of experts is necessary. Many projects in which AI engineers and other experts participate face the following issues:

- AI engineers tend to focus on improving model accuracy; however, the business and data-engineering side want AI engineers to make proposals concerning data generation and business processes required for reporting data quality and improving the data quality itself.
- AI engineers are not necessarily outstanding software developers, so the quality of the code they create may be poor, and the cost of rewriting that code by delivery-side developers to improve it to the high quality required for commercial use

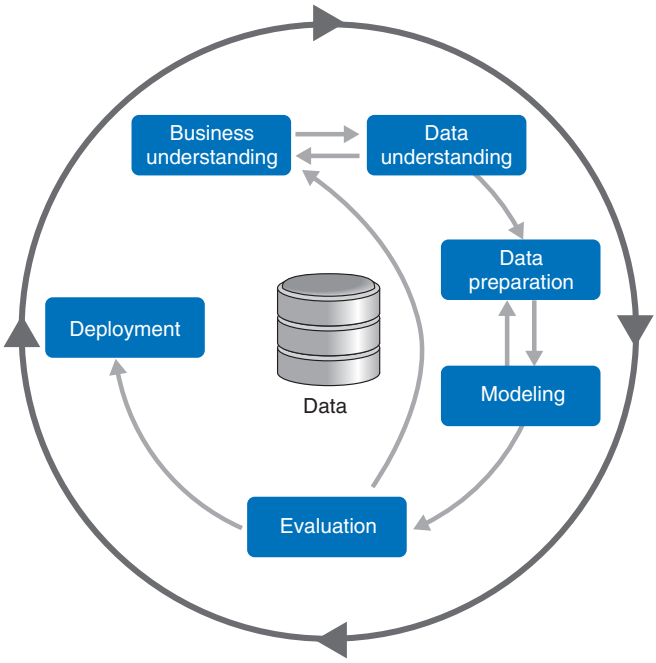


Fig. 2. CRISP-DM.

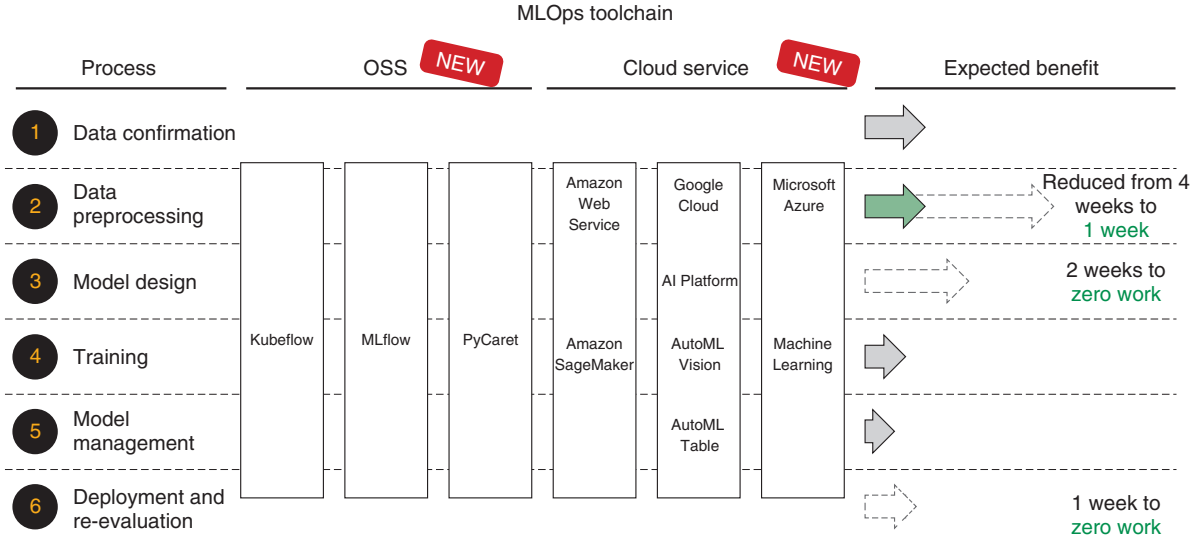


Fig. 3. MLOps toolchain of NTT DATA.

is very high.

In light of these issues, it is clear that the definition of the role of AI engineers needs to be broadened in regard to the commercial-development phase. In addition to technical knowledge to improve model accuracy, which is essential in the PoC phase, there is a need for AI engineers who can create high-quality code for implementing commercial systems, understand data and customer operations, and make proposals for improving data quality.

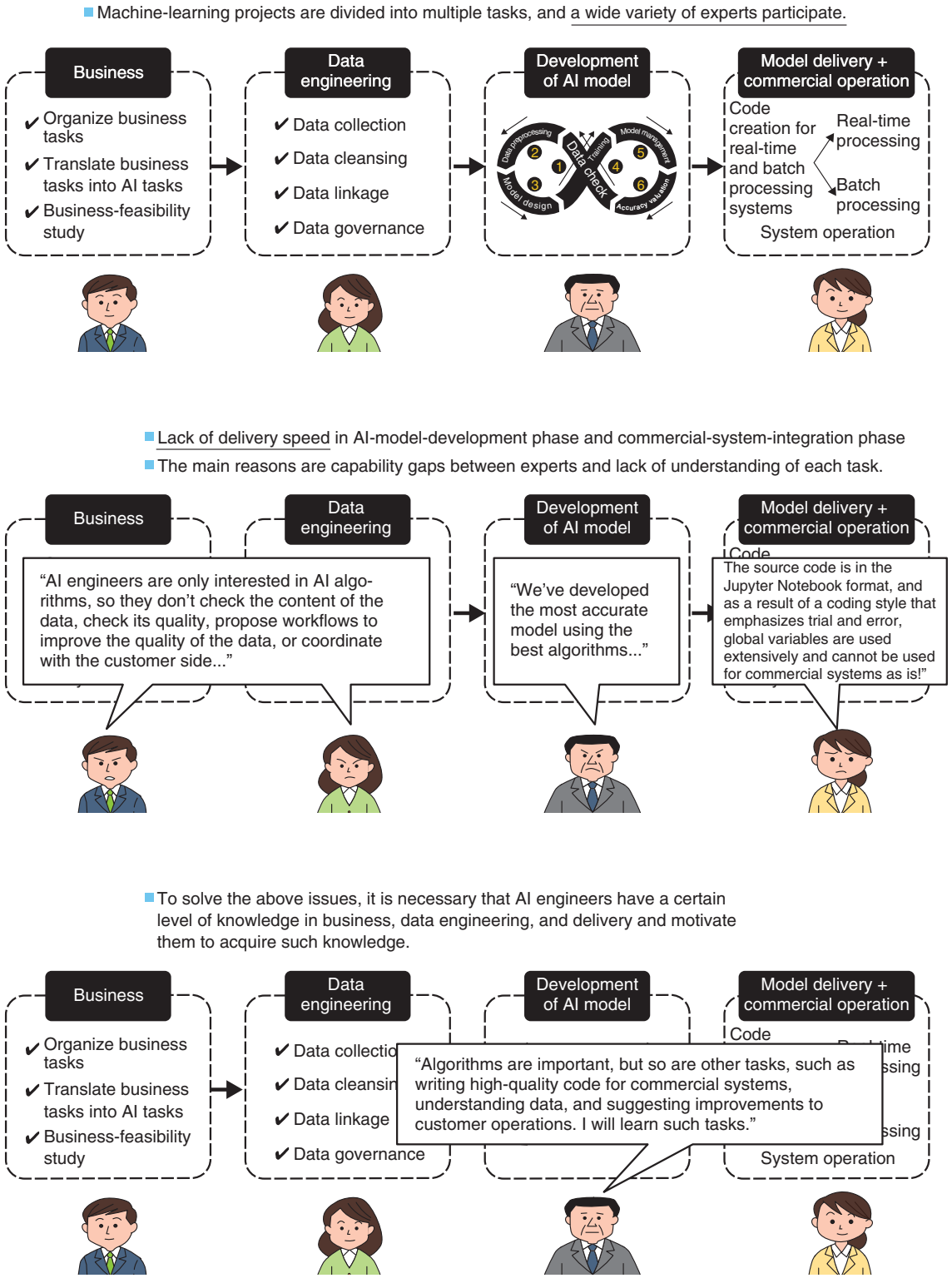


Fig. 4. Machine-learning development process and roles of experts.

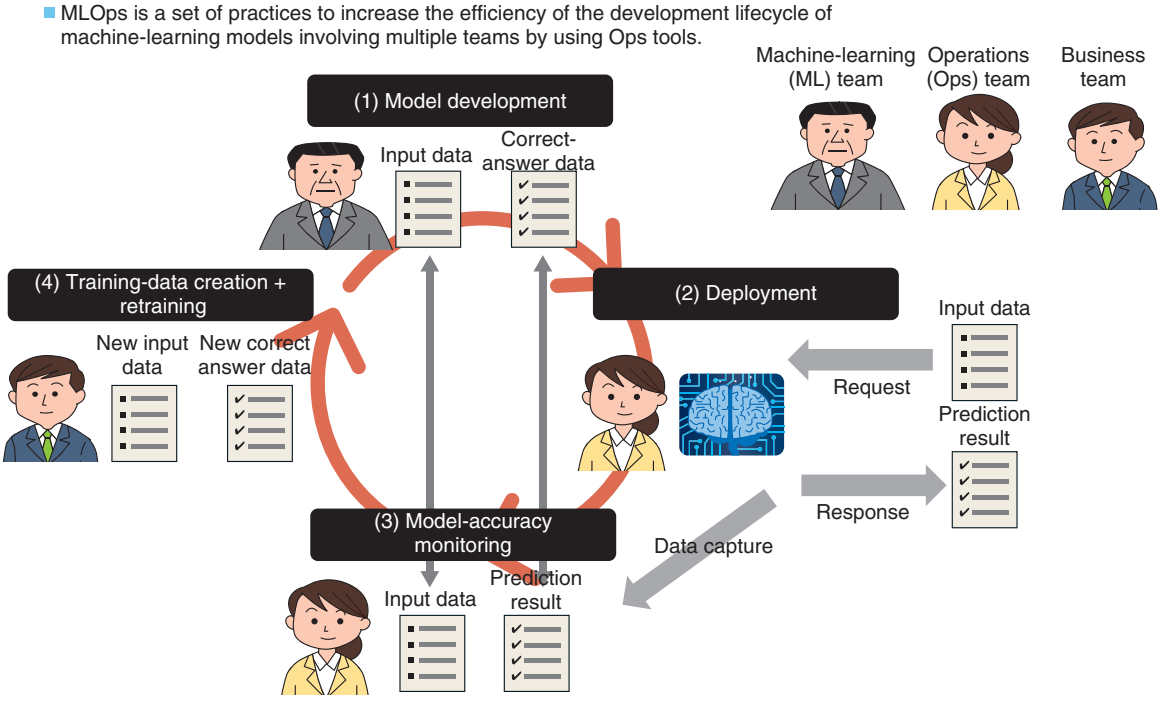


Fig. 5. Process flow of MLOps.

1.3 Issue 3: When an event that changes people's behavior occurs, such as a pandemic, machine-learning models that were tuned before the change become useless

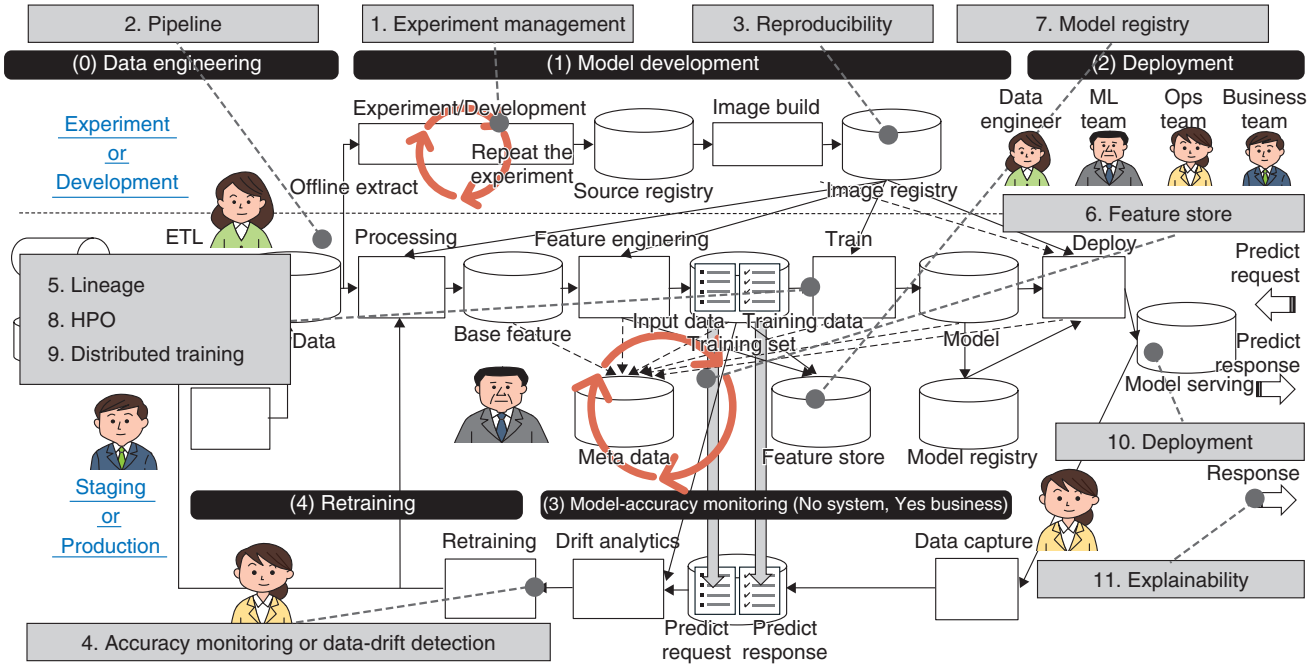
COVID-19 has spread rampantly around the world, and that situation has led to a change in people's purchasing habits and the need for fingerprint authentication instead of face recognition on smartphones because people wear masks. This situation is an example of the phenomena known as concept drift and data drift, and it represents the problem that AI models created thus far have become useless because the statistical properties of the data generated change due to changes in people's behavior. To deal with these phenomena, it is necessary to collect new data and rebuild the model with the collected data. However, it is inefficient to carry out that task manually every time this phenomenon arises; thus, it is necessary to automate the model-rebuilding process as a functional requirement and mechanism of the machine-learning system. In particular, for machine-learning models as well as general applications, the operation phase is crucial, and a mechanism for monitoring systematic error is not sufficient, that is, it is also necessary to monitor the accuracy of the machine-learning model.

The overall process of MLOps is shown in **Fig. 5**. Not only (1) model development and (2) deployment but also (3) model-accuracy monitoring and (4) training-data creation + retraining are required. If these processes are introduced together with the automation mechanism, it will be possible to adapt to concept drift or data drift as a system.

2. Future developments

MLOps is a recent technology trend, and the technology stack that underpins it, including definitions, is still immature. However, providers of OSS and cloud services as well as third-party vendors are working hard on developing MLOps tools and technologies, so it is necessary to continuously monitor these trends.

Although an industry-wide consensus definition of MLOps has not yet been established, the information provided by vendors, including those overseas, on MLOps can be organized into a system structure (**Fig. 6**) and 11 functional groups (**Table 1**). Due to space limitations, we cannot explain each function listed in the table, but it is expected that MLOps tools will continue to mature, and the social implementation of AI will accelerate in proportion to that



ETL: extract, transform, load
HPO: hyper parameter optimization

Fig. 6. Relationships between machine-learning system structure and MLOps tools.

maturing. As such a trend continues, various pieces of expertise about social implementation of MLOps and AI will continue to be accumulated, and by using that expertise, NTT DATA wants to contribute to the development of the machine-learning industry and the social implementation of AI.

Table 1. Eleven functional groups of MLOps.

Function name	Description	Expected benefits	
		Increase speed	Ease of use
1. Experiment management	After conducting a machine-learning experiment, the experiment can be traced later. Multiple experiments can be compared and visualized.	✓	
2. Automation of the pipeline	Multiple machine-learning processes can be stitched together into a job.	✓	
3. Reproducibility	It is possible to ensure packaging and portability with technologies such as containers so that machi-learning processing does not carried out in other environments due to the library on which machine-learning algorithms depend.		✓
4. Accuracy monitoring or data-drift detection	To be able to detect the deterioration in accuracy from the deviation in the distribution of the original data of the model, it is possible to compare the deviations between the baseline data and data acquired through Request.		✓
5. Lineage (tracking)	It is possible to trace the connection between multiple machine-learning processes (preprocessing, training, etc.) and the input and output of each process at a later phase by using an API (application programming interface), etc.		✓
6. Feature store	Feature-value data created by a data scientist in preprocessing can be shared with other data scientists and managed in a special storage area and portal.		✓
7. Model registry	Machine-learning models created by data scientists can be registered in a model-specific registry service and deployed to the environment later by “one click” of the approver.		✓
8. Hyper-parameter tuning	It is possible to execute a job dedicated to hyper-parameter tuning and select the job that generated the most accurate model.	✓	
9. Distributed training	Data partitioning or model decomposition can be used, and training of machine-learning models can be scaled out.	✓	
10. Multi-framework deployment	Generation of serving code for multiple machine-learning algorithms can be semi-automated, and advanced deployment methods (such as canary release) can be implemented.	✓	
11. Explainability	From the prediction results by machine learning, it is possible to determine what explanatory variables contribute to the results and explain the results.		✓

Trademark notes

Author: Ei Yamaguchi, NTT DATA Corporation