Tags	MLOps, Machine Learning, Introducing MLOps, MLOps process, Machine Learning problems, implement MLOps
Meta Description	Machine Learning problems may cost your company time, effectiveness and leave you behind the competition. Find out what they are.
Intro KB	From Google to tech startups, everyone is rushing to use Machine Learning to expand its position in the market. What problems should be avoided to count in this race?
Autor	Jakub Jurczak
LINK do grafik	

MLOps: 5 Machine Learning problems resulting in ineffective use of data

In recent times, Machine Learning has seen a surge in popularity. From Google to tech startups, everyone is rushing to use Machine Learning to expand its position in the market. However, not all organizations have the know-how and resources to deploy machine learning at a large scale. Furthermore, as we know from Uncle Ben (the uncle of Spiderman), “Great power is a great responsibility”, so Big Data is a great opportunity to increase business, yet also comes with a huge responsibility in dealing with the risks of not reaching its full potential. So today I want to highlight and explain 5 Machine Learning problems resulting in the ineffective use of data.

Introducing MLOps for business

One of the most significant areas of society's progress has been the increased amount of available data, resulting in the expansion in applying artificial intelligence methods in practice. In this way, data has become fuel for the modern economy and

initiated the fourth industrial revolution. **Accessing this fuel on demand is critical to any organization that relies heavily on automated data-driven decision-making business processes**. As a result, companies using big data technologies rush to find new ways to extract business values from their data.

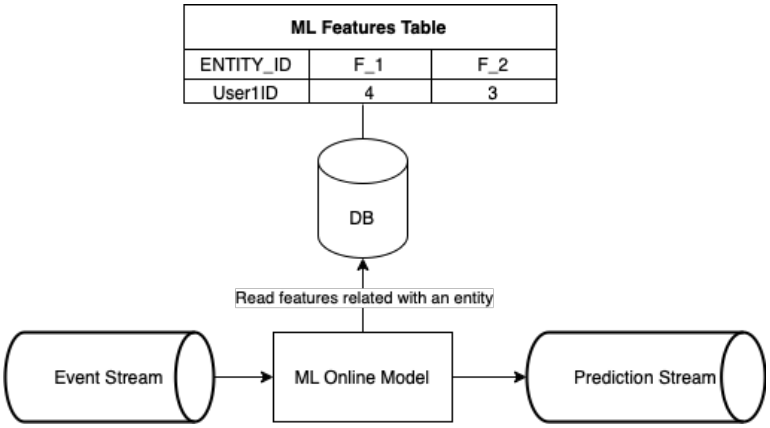
AI opened up a spectrum of opportunities for companies that they previously could only dream of. Therefore, companies are racing to create faster and more accurate ML models, which allow them to **optimize their business processes and gain a competitive advantage**. However, the constant need to train, implement, monitor and improve models **creates a big challenge for engineering teams, especially when processing a large amount of data and making automated decisions in real-time**.

MLOps is responsible for optimizing and maintaining the maximum effectiveness of Machine Learning. MLOps is a set of practices which are the solution to ML challenges.

So, to the point. What are the 5 areas in the Machine Learning process where the implementation of MLOps is crucial?

1. (In)Accurate online predictions in Machine Learning

The first area at risk of loss of effectiveness is the process of making online predictions.
Let's assume that we are building a system to detect financial fraud for a bank. One of our system's elements is a service that uses an ML model to analyze loan requests and decide whether or not we are dealing with fraud. When the request comes, the service must ask the database for the current data of a given user such as "scoring", "number of loan requests in the past seven days", "average income", etc. We say that this data is our user's features in the Machine Learning terminology.



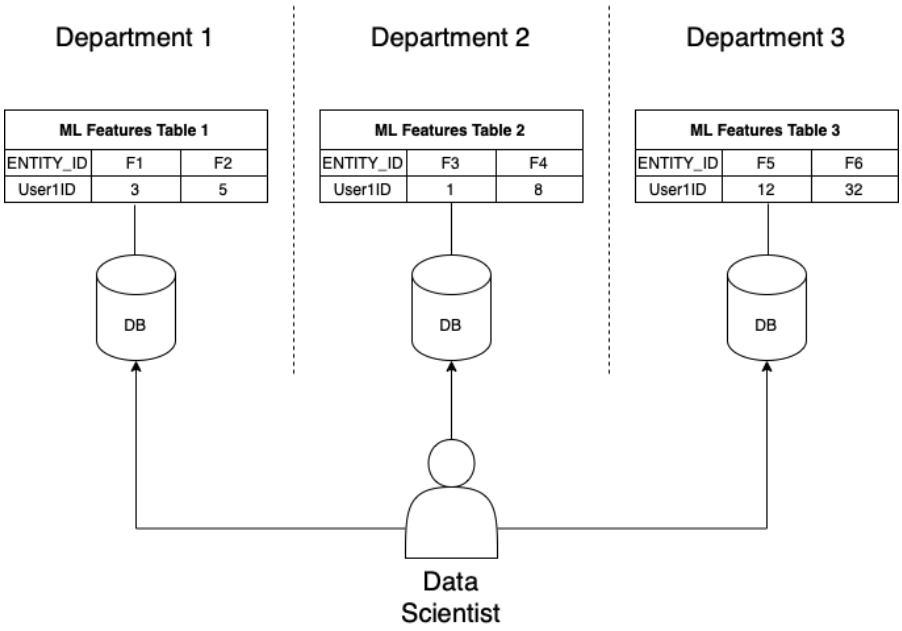
As we can see, the ML model requires knowledge of the latest facts about the user to make the correct prediction. **The seemingly simple activity of enriching data raises several Machine Learning problems that we have to deal with**:

- **How can we guarantee a few millisecond response time from the database** so that it doesn't affect the decision process?
- **How should we scale the database to deal with the peak of user events?**
- What if the Machine Learning process uses the same database as the legacy application? **How can we ensure that the prediction process doesn't affect the performance of the rest of the system?**

Certainly, when designing a real-time, AI-driven decision process in the big data world, architects should ask these questions.

2. Falling into the data silos trap

Any organization that wants to implement a data-driven approach in their business processes knows how vital data warehousing is in the enterprise architecture. Unfortunately, **many companies keep their data in warehouses scattered around different departments**. This way of data storage creates data silos, which have a negative impact on the daily work of the IT department.



When data scientists want to retrieve data of interest, they must search through datasets located in many warehouses in different locations. This is very inconvenient and **causes problems such as:**

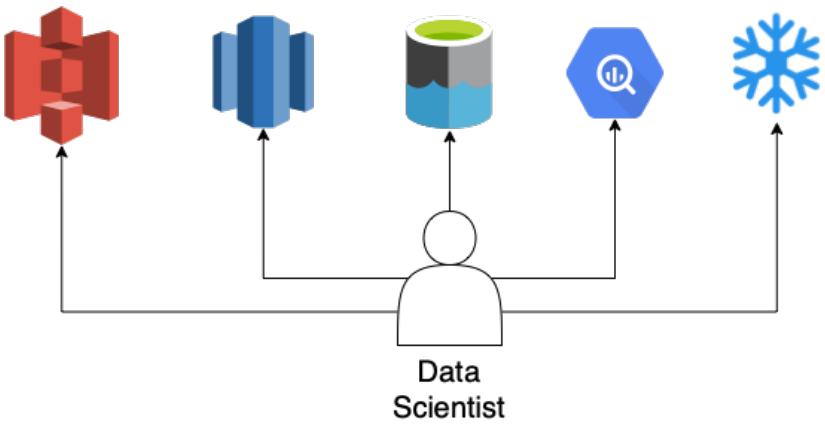
- **Mismatching IDs between warehouses** can make joining data between different sources difficult and **sometimes even impossible**.
- The data scientist needs to dig and understand data structures from many different databases to find data of interest. **Searching for data wastes employees' time which could otherwise be devoted to creative work.**

- Finding features of an entity at a specific time can be challenging if the time between data refresh in warehouses is different.

As a result, your organization is less efficient, you don't innovate quickly enough, and you can't make the most of your data. Certainly, simple and universal access to data is crucial for all enterprises that want to keep up with the rapidly changing world.

3. Multi-cloud, multi-risk

In the modern approach to application design, it is common practice to **select the database that best suits the problem being solved. In Machine Learning we call this approach a polyglot database pattern**. The polyglot persistence pattern uses two or more data storage technologies for different data types. For example, you use MySQL for relational data and Redis for fast caching. When using polyglot persistence, you can choose the database that best suits your performance, scalability, and security needs.



However, a polyglot database poses new challenges in supporting applications that use different database technologies. Primarily when working in a multi-cloud environment, where each cloud provider offers its warehouse as a service:

- **Data scientists should focus on building high-quality models** instead of delving into the intricacies of the various dialects of the query languages.

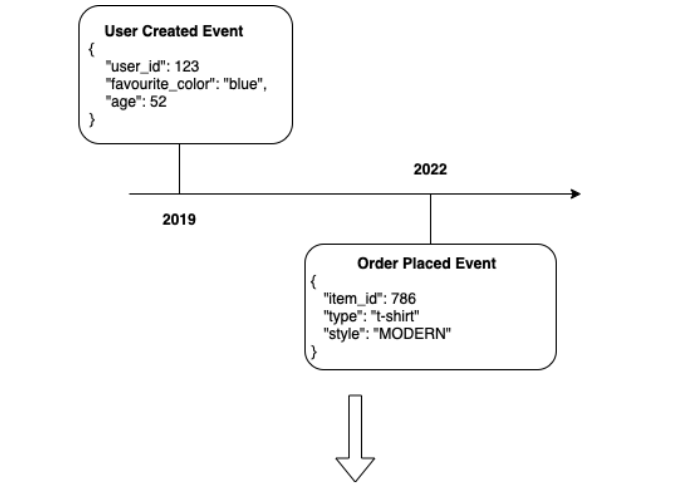
- **Data access management is becoming a real challenge**, especially in multi-cloud environments.

- **Joining data between two different databases might be impossible**. Mainly when working on products from different vendors or mixing SQL and NoSQL technologies.

I think I managed to draw your attention to the most common problems of using different technologies for storing data. However, it is essential to remember that each tool has its strengths and weaknesses, so we should always choose the right tool for the problem.

4. Time goes by. So does data.

When it comes to data analysis and prediction in Machine Learning, one of the most important tasks is creating a dataset used for training and testing. Unfortunately, it isn't as easy as many people think. Especially when we need to track changes in the domain in time. For example, let's assume we are working on an ML model for targeted email campaigns for online store users. We have a stream of events with information about the user, such as age, favorite color and events related to orders placed in the past.



Features					
user_id	age	item_id	type	style	favourite_color
123	52	786	t-shirt	MODERN	blue

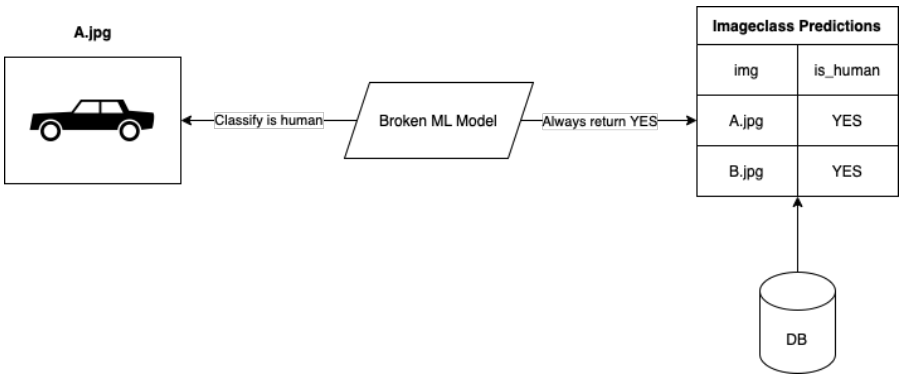
We want to create a view of our domain to combine information about users with their orders. The seemingly simple operation turns out to be more complicated. **We can't join events because our domain is changing over time**. We need to know the domain's state from the exact moment each event occurred, in order to join events. For example, a user who was 52 in 2019 is 55 years old in 2022. **When building ML models on past data, it is essential to remember that the world has changed since that data was collected**:

- **You can never know if data that is being processed is new or stale data, so there is a need for some TTL (time to live) information** that says how long old data is good.
- Because the state of the domain changes all the time, **access to the latest data for the training process is needed** so that models can perform better in a production environment.
- **It's important to check the ML model performance over time to ensure that it is still performing well**, so we need to evaluate the model on data from different periods and see if the performance is consistent.

Extracting features for ML models from the data streams is what enterprises face today. As data volume and velocity continue to increase, enterprises need to find ways to manage their growing volumes efficiently on a big scale.

5. Skewing data

For each feature used for training ML models, its expected range of values and distribution should be recorded. Then, if someone builds a model that uses this feature as an input, they should also store information about how much each feature influences this model output. This information can be used to monitor features for unexpected changes in value or distribution that may invalidate assumptions made during modeling.



If the value of a feature changes significantly over time, then the model performance could suffer. As an extreme case, if a corrupted ML model generates this feature, this model will no longer work.

- As ML models have become ubiquitous, **companies have to develop systems to monitor their quality and performance**.

- The training data needs to be monitored and detected when changes have occurred because this model will probably need to be retrained to ensure that the model predictions are accurate.

- Features monitoring allows to detect a bug and revert features to the last correct version.

The first step for a**ny company that wants to implement MLOps into its business process successfully should be developing a monitoring system**. Only then should they start developing algorithms and integrating them into their systems. Understanding the importance of data monitoring will allow you to **save a lot of resources and successfully implement AI-based solutions.**

Where there are Machine Learning problems there is also potential…

As you see, there are many areas where Machine Learning problems can cause a decrease in the company's efficiency and a slowdown in expansion in the market in which the company is competing. Big Data gives fuel to gain an advantage as long as we are able to not lose efficiency. **MLOps role is to find the area causing ineffectiveness and apply solutions and practices to eliminate the problem**. Therefore, **the company can reach its full potential on the market by being able to make data-driven accurate decisions and defining trends at once.**

— usunięto: So

We have proven MLOps ways to improve each of these critical areas. You can talk to us about them now >> contact <<. In the near future, we also plan to publish a deeper analysis of the topic with solutions, and a step by step guide. In order not to miss the publication, subscribe to our >>newsletter<<

—---------- VERSION FOR MEDIUM —----------

Where there are Machine Learning problems there is also potential…

As you see, there are many areas where Machine Learning problems can cause a decrease in the company's efficiency and a slowdown in expansion in the market in which the company is competing. Big Data gives fuel to gain an advantage as long as we are able to not lose efficiency. **Without sealing the critical areas for performance, it's like driving an F1 car in neutral gear. Because that's what Machine Learning is - a machine to accelerate (right, data based on) the decision making process and take part in the race in the competition market. And MLOps is like the F1 team that keeps the racecar at maximum efficiency.**

We have proven MLOps ways to improve each of these critical areas. You can talk to us about them now >> contact <<. In the near future, we also plan to publish a deeper analysis of the topic with solutions, and a step by step guide. In order not to miss the publication, subscribe to our >>newsletter<<

I hope that I made it clear which areas of Machine Learning need to be checked for effectiveness and that it's time for the MLOps pitstop.