## [ML Ops: Key to Accelerated Business Outcomes and Increased Last Mile Adoption](https://www.tredence.com/blog/ml-ops-part-1/)

You have just completed a machine learning pilot, tackling a challenging business problem which has stymied your executives for the last few years. As the plaudits roll in, you start thinking about how to move your project from the pilot stage to a full-blown application which will drive business value to your stakeholders.

Chances are, you reach out to your IT team to help scale your pilot, and are subject to a barrage of questions around scalability, system performance, integrations, and access privileges, topics which you didn’t think about too much when you set up your pilot. All this while, your stakeholders are pressuring you to go full scale so they can start seeing a return on their investment.

We at Tredence understand your challenges and can help you overcome them and take your machine learning to the next level. This article is the first part of a series on our MLOps capabilities and accelerators. Here we will talk about the common challenges in deploying MLOps and demystify the Ops component which serves as the nemesis to many. We will then elucidate our offerings in this space.

Why is machine learning at scale such a big challenge?

Let us start by looking back to the world of software engineering, where large-scale enterprise applications follow a DevOps methodology to drive continuous integration (CI) and continuous deployment (CD) and ensure high software quality. The DevOps process deals with versioned code being pushed through the software engineering value cycle.

Now let’s come back to the world of machine learning. One or more data scientists in your team has built a set of ML models, which have been tested in a controlled environment. The models must be deployed in a production environment, as well as continuously monitored, managed and improved upon.

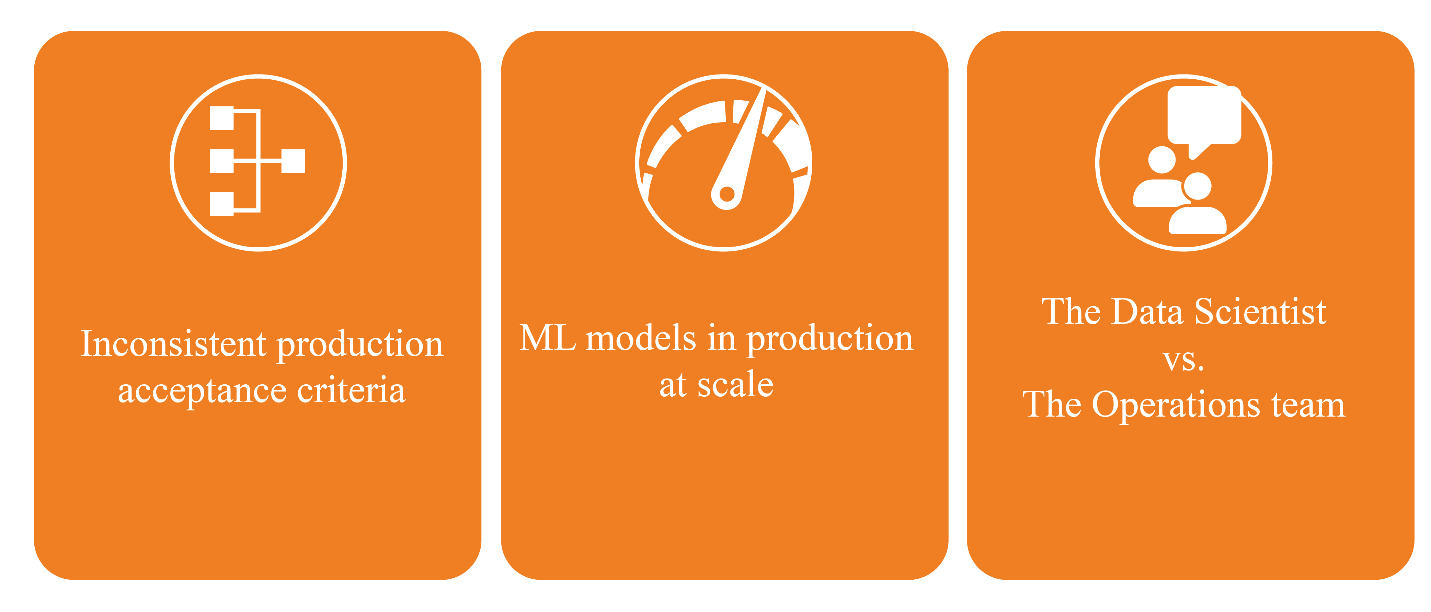
Which is how we come to MLOps. Quoting from Wikipedia, “MLOps (a compound of “machine learning” and “operations”) is a practice for collaboration and communication between data scientists and operations professionals to help manage production ML (or deep learning) lifecycle. Similar to the DevOps or DataOps approaches, MLOps looks to increase automation and improve the quality of production ML while also focusing on business and regulatory requirements.

***MLOps is not just about the moving of versioned code.***

***It is the moving of versioned code, data and models.***

MLOps is not just about the moving of versioned code. It is the moving of versioned code, data and models. But how good are we at versioning all of these? Data scientists focus on building a best in class model, and not in documenting their code and model parameters in a standard manner. Inconsistent document standards and incomplete information on codes and models, especially legacy codes which business requires be integrated now with new ML models – in short, inconsistent production acceptance criteria – present the first big challenge in setting up your ML models for scale.

Because ML models have been created in controlled environments, thought is often not given to the computational needs of deploying them in a production scale environment. This leads to an underestimation of computational needs, in turn leading to very high compute needs, drastically reducing the ROI of the project.

[](https://www.tredence.com/wp-content/uploads/2020/05/ML-ops-1-01.png)

We finally have the dichotomy of the data scientist and the Ops team. Not only do we have very diverse skill sets required to do the 2 sets of roles, neither group is (or wants to be, for that matter) involved in the work of the other. Ops teams are not involved in the life cycle of the model development and are loath to support any outcome which is defined by an SLA. Data scientists likewise tend to worry not too much about how their models will integrate with other systems, whether they will continue to work as accurately as they did during development, or who has access to their outputs. This also manifests in the form of governance challenges – for example, who is responsible for tracking model quality? What happens if end users report a dip in model performance? Who will fix it?

The Ops of MLOps

To answer these questions, we need to get deeper into the Ops of MLOps. This will help us break down the challenges described above into specific operations (or activities).

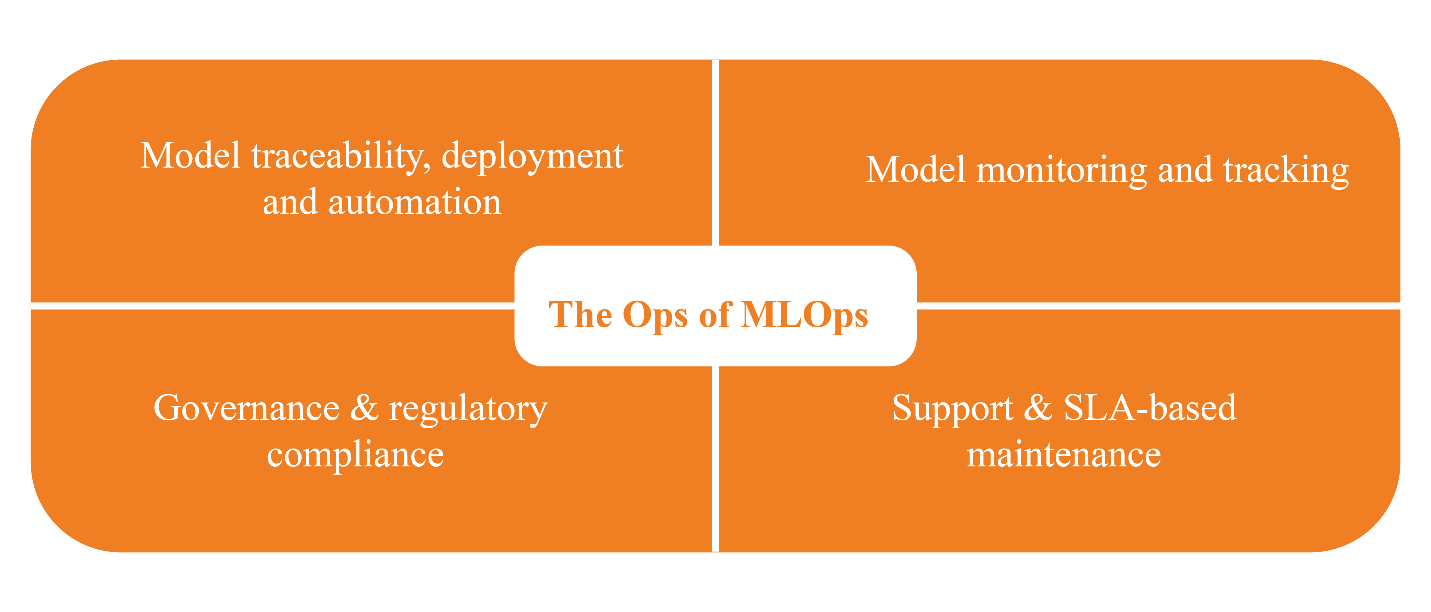
Your data scientists have built the models for you. The first step is to deploy and automate the models for production. This requires setting up the CI/CD pipelines for your ML models, and rigorously documenting (and tracking) the traceability of the models. In the software engineering world, Git most often played this role. Jenkins was often used along with Git to build, test and deploy your code. But in the MLOps world, you need something bigger – an ML pipeline which can ingest model parameters as well as data, validate both, train & validate the model, and then deploy it in production. The pipelines become even more important because in today’s world, ML models must be explainable. For the model results to be explainable and to make sense to business stakeholders, the traceability of both the models as well as the data (what we call data lineage) has to be established. The ML pipeline thus needs to track the complete traceability of the components – data, models, et al – to confirm their provenance and make the processes as well as results explainable.

***The ML pipeline needs to track the complete traceability of the components –***

***data, models, et al – to confirm their provenance and make the processes as well as results explainable.***

***.***Now that your ML models have been set up to deploy to production in an automated manner, you need to monitor them on a real-time basis to setup and track model specific metrics as well as the performance of all deployed models. Monitoring is not restricted to tracking model performance – accuracy, hyper parameters, etc. – items of interest to a data scientist, but also covers items like data load durations and data count from the world of data engineering, and model health score and completeness which are of interest to business users. We also have the system related metrics ranging from CPU usage to network throughput to application load times. Of course, monitoring is of no use without an efficient method to visualize and act upon the information, so monitoring dashboards are required.

With thousands of ML models now running on production, the need arises to govern the models as well as the underlying infrastructure, as well as access to the same. Governance ranges from managing security controls, both from an authorization as well as regulatory compliance standpoint, to monitoring and providing access to model metrics or user access logs. Deployment approvals must be built in with periodic approvals to identify gaps. Automated code and data quality checks generate quality reports which must be archived and made available to the right stakeholders. Needless to say, model traceability metrics, documentation standards and naming conventions for codes and components, and CI/CD monitoring all find a place here.

[](https://www.tredence.com/wp-content/uploads/2020/05/ML-ops-2-01.png)

We now have all the lights up and running. All the good work from your ML pilot is now ready to be received by stakeholders. But we still need a support team to keep the lights on for you. The support organization will ensure uptime and maintenance of all services, with an SLA driven support management process. This set of activities will cover maintenance, user access management, usage reporting, security audits, incident management, and executive reporting on the environment performance. Your ML pilot is now all set to run smoothly.

How does Tredence help organizations overcome MLOps challenges?

Overcoming MLOps challenges requires deep expertise and process understanding of both machine learning and operations. With robust experience in implementing MLOps projects, Tredence helps bridge the gap between your development and ops teams, helping you define the provenance of your models and data pipelines, allowing the setup of a one-click deployment process, with strong governance mechanisms and reporting to support and maintain the MLOps setup. Our strong experience enables us to respond much faster to L2 & L3 issues as well as overcome persona-based challenges.

Tredence’s accelerators are perfectly positioned to help organizations setup MLOps at scale. Look out for the next article in this series to know more about our accelerators.

Author: [**Sanat Pai Raikar**](https://in.linkedin.com/in/sanatpairaikar), *Associate Principal, Tredence*