Building Effective Machine Learning Teams

Why Visibility, Reproducibility, and Collaboration are Required for ML & AI Success

Introduction

It seems like everyday there are new headlines about how innovations in artificial intelligence (AI) and machine learning are being used to tackle some of the world's most complicated problems. The caveat? A number of these innovations and stories are *forward looking*, making a bet on the promise that AI and machine learning can deliver. If only it was that simple.

According to a recent study, **55% of companies never take their models to production.** And that's only if projects are successful - **another study showed that 87% of projects fail.**

Why is this? Despite massive innovations in technology and tools, the workflows, team structures, and collaborative processes required to make machine learning and AI projects successful haven't caught up to the technology.

WHAT THE OPS?

If you've been following machine learning for a while (or really any tech topics), chances are you've heard of the "ops" trend. DevOps, DataOps, MLOps. All of these trends attempt to tackle the same problems, albeit from different angles.

The thread that runs across all of the topics is simple: how do I "shift left" via people, process, and technology, to be more agile and more iterative, all while embracing continuous integration and continuous delivery (CI/CD) (see *Table 1*).

What the Ops?	Defines the...	Goal
DevOps	People, Process, & Technology to...	Shorten application and system development lifecycles to deliver features and applications faster
DataOps		Shorten the time it takes to access, model, and deliver the right data, to the right place, at the right time
MLOps		Shorten the machine learning development lifecycle to successfully get to production faster

Table 1: Though focused on different disciplines, the "Ops" movement is driven by a need to deliver value faster.

The "why" is pretty straightforward - IT, data, and AI leaders are under increasing pressure to deliver value to their organization. It's the traditional "better, faster, and less expensive" challenge that every engineer and IT team member has experienced.

WHAT THIS MEANS FOR MACHINE LEARNING

Machine learning and AI, by their nature, are iterative and complex. This complexity isn't limited to the challenges inherent in data science. It also extends to the processes, teamwork, and communication required for machine learning and AI projects to be successful.

Taken from an MLOps perspective, one of the strategic goals for machine learning and AI teams should be figuring out how to drive efficiencies and collaboration across the entire machine learning pipeline (see *Figure 1*).

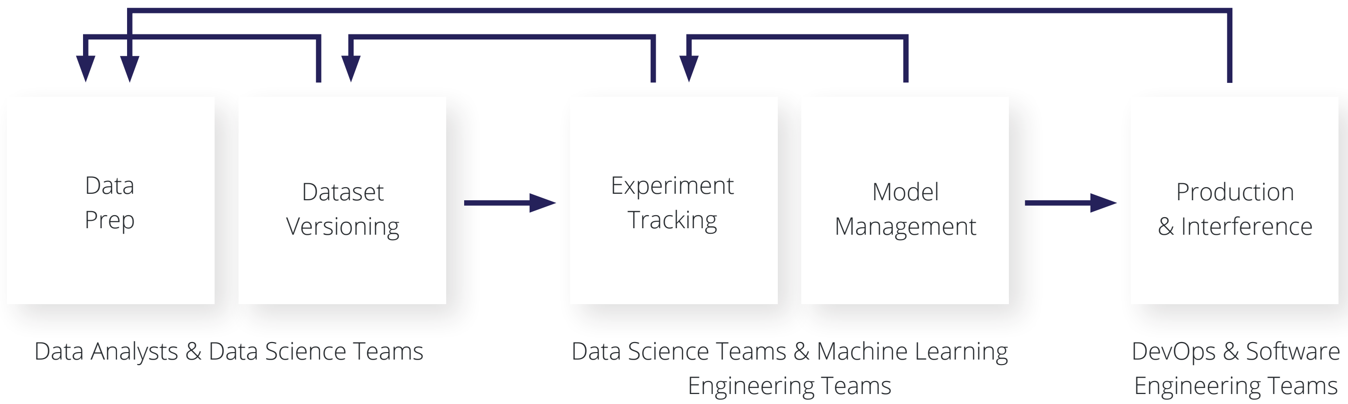

Figure 1: ML & AI teams are dispersed across multiple functions and disciplines.

And for teams that are successful, the results are staggering. According to a **recent study by Deloitte,** the majority of AI adopters believe that it will substantially transform both their own organization and industry within the next three years. In terms of the advantage this drives, "Looking at competitiveness as a measure, 26 percent of all respondents said that AI technologies enable them to establish a significant lead over their competitors. For Seasoned adopters, this rises to 45 percent."

That means the 45% of seasoned adopters are already pursuing AI, getting it into production, and are getting further and further ahead of the pack. So do you become one of those organizations?

The Critical Requirements for an Effective Machine Learning Team

Effective machine learning teams can be calibrated across three vectors: visibility, reproducibility, and collaboration. Addressing these three vectors will impact your team's productivity, success with future candidates, and ultimately your competitive advantage -- after all, we all want to be part of the 45% getting ahead of the pack.

- **Visibility** - ability to view, access, and react to machine learning processes and deliverables
- **Reproducibility** - one of the cornerstones of science: the ability to take an experiment, repeat the steps, and reach the same conclusion
- **Collaboration** - communications and teamwork across multiple, often disjointed units

Because machine learning is iterative, all three of these must be achieved across the MLOps pipeline in order to consistently deliver machine learning models that successfully get from development to production.

VISIBILITY

Here's a hypothetical situation: imagine you're a manager of a high-performing machine learning team, and your lead data scientist one day decides to quit. Do you:

- **Have a clear line-of-sight** into the models that were being trained?
- **Know which models** are being used in production?
- **Know which hyperparameters** were used to train a model?
- **Know the performance of a model,** how it was optimized or the data that was used to train it?

If you answered no to any of the above (or don't know the answer), you may have a visibility issue.

Visibility is fundamentally about sharing processes, information, and deliverables across your team. The reason this is so challenging in machine learning and AI is due to the diverse set of skills and teams required for success (see *Figure 2*).

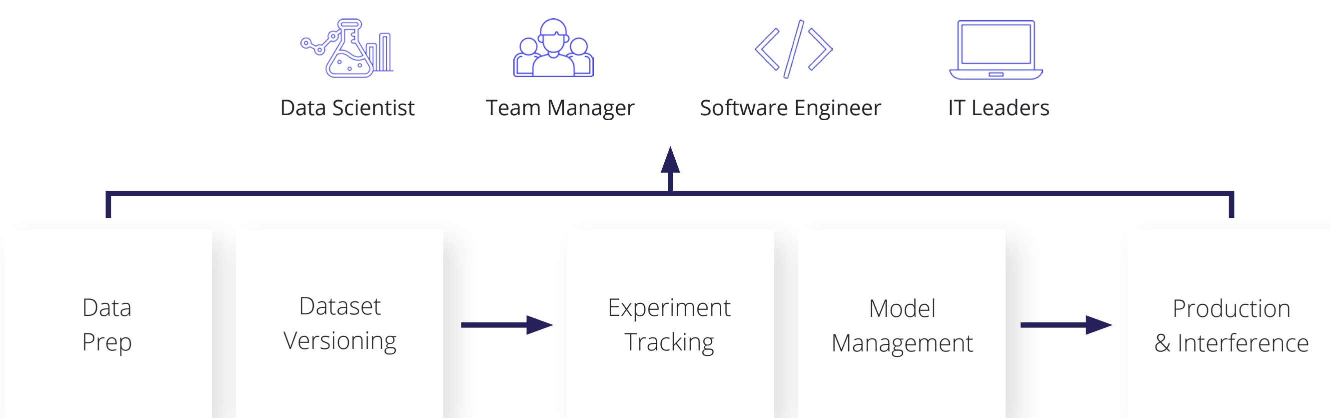

Figure 2: How do you get visibility across multiple teams and disparate functions?

These barriers can be grouped into four categories:

- **Decoupled pieces** -- data, code, configuration, and results are generated at different steps in the process
- **Siloed functions** -- the ML workflow is spread across multiple teams and functions (DBAs, research, software engineers, DevOps, etc.)
- **Different tools** -- Multiple tools, libraries, and infrastructure providers
- **Documentation** -- Often outdated, sparse, and managed separately from actual workflow artifacts

Given that machine learning and AI is a combination of disciplines, it makes aligning these functions -- much less getting the visibility you need -- extremely challenging. But at a minimum, teams that are able to monitor, track, and report on progress cross-functionally across projects will benefit from information sharing and protect themselves from unexpected turnover.

REPRODUCIBILITY

Machine learning and AI are built on data science, and as with any science, being able to reproduce or replicate findings is critical.

With regards to building effective machine learning teams, this focus is all about being able to run the same procedure for the same outcome. But like visibility, this can be difficult for a variety of factors:

Challenges to Reproducibility

- Changes in incoming/connected data
- Inconsistent hyperparameters
- Changes in ML frameworks
- Noisy hidden layers in NN architectures
- Variant computer architectures for CPUs and GPUs

There are tangible benefits to addressing reproducibility:

- Meta Analysis (think "run experiments on top of experiment data" - it's meta!)
- Easy Comparison
- Efficient Debugging
- Reduced Risk

Reproducibility vs. Replicability

One conversation you may hear in machine learning circles is the distinction between reproducibility and replicability. For the purposes of this paper, we're using both interchangeably.

Broadly speaking, this breaks down to:

- **Reproducibility:** Perform two separate independent or different experiments that prove the same thesis
- **Replicability:** Use the same procedures and get the same outcome

COLLABORATION

The final vector is collaboration. It's not uncommon for machine learning teams to essentially be lone wolves sitting in the same room. And it's not because of a lack of desire to collaborate, but simply due to the difficulty of collaboration. Teams are often forced to rely on meetings or emails to simply share results. This gets even harder, given machine learning function spans multiple teams.

The Software Engineering vs. Machine Learning Process Dilemma

Most people coming from the engineering community would likely say, "But wait - isn't this a problem that's already been solved?"

This thought process makes sense, given that the principal tool for both teams is code. However, the type of work being done on the machine learning and AI side differs drastically from software engineering. The iterative nature and variety of skill sets required for machine learning mean that it's not possible to simply reuse existing tools to drive this collaboration.

Given that software engineering and machine learning are different disciplines, different processes are necessary. Here's a simplified example of how the processes vary (see Figure 3):

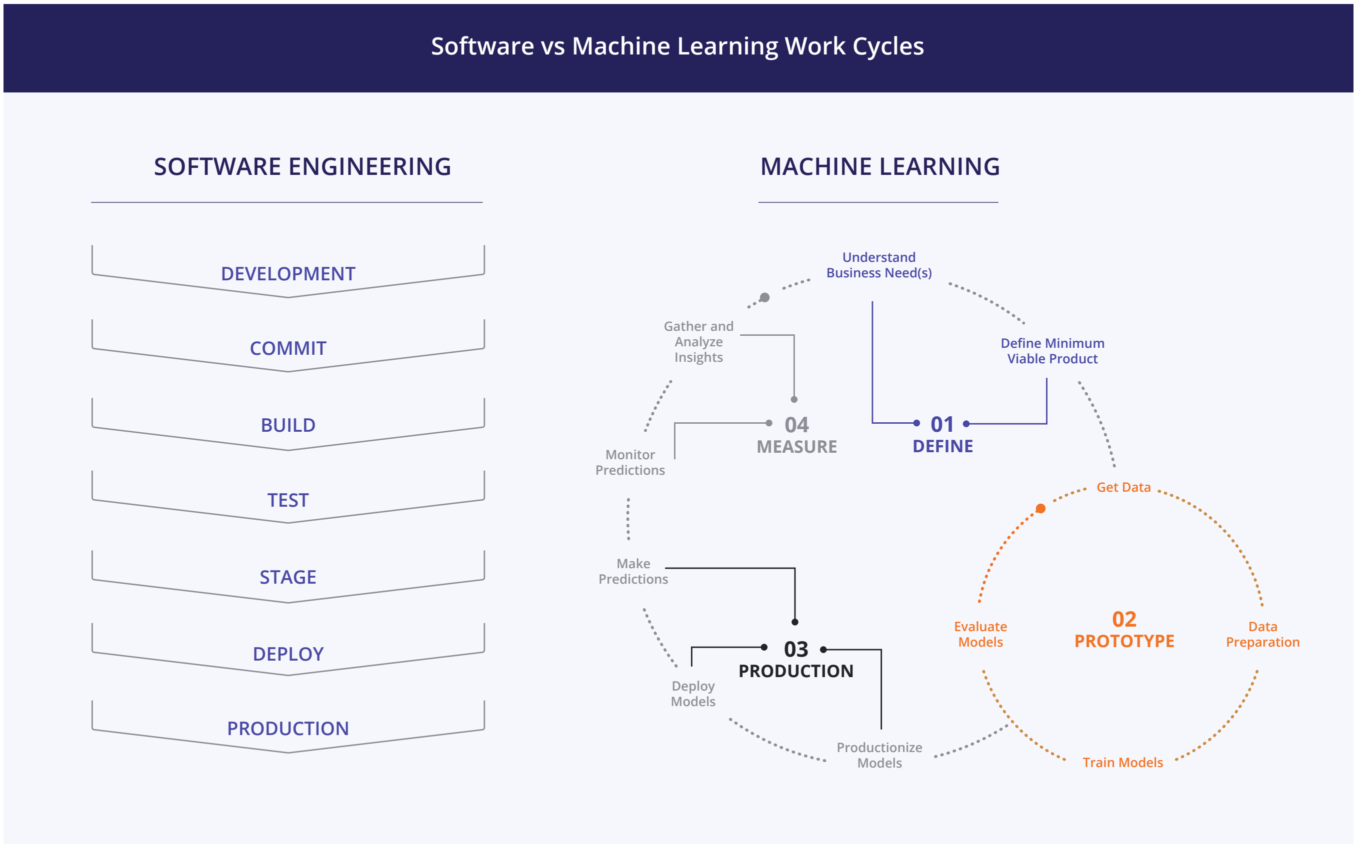

Figure 3: Though code-based functions, the processes for software engineering and machine learning are different.

Collaboration Across Machine Learning Teams

Machine learning is inherently an uncertain discipline. And that uncertainty is multiplied by the variety of teams, technologies, and functional processes that are required for each individual function. Even when you consider existing disciplines, such as application development which requires smooth processes between data and engineering teams, collaboration is a challenge, despite tools and processes having existed for decades.

Embracing an MLOps methodology, for example, requires alignment across the disparate teams and functions that are responsible for getting ML and AI projects into production (see *Figure 4*).

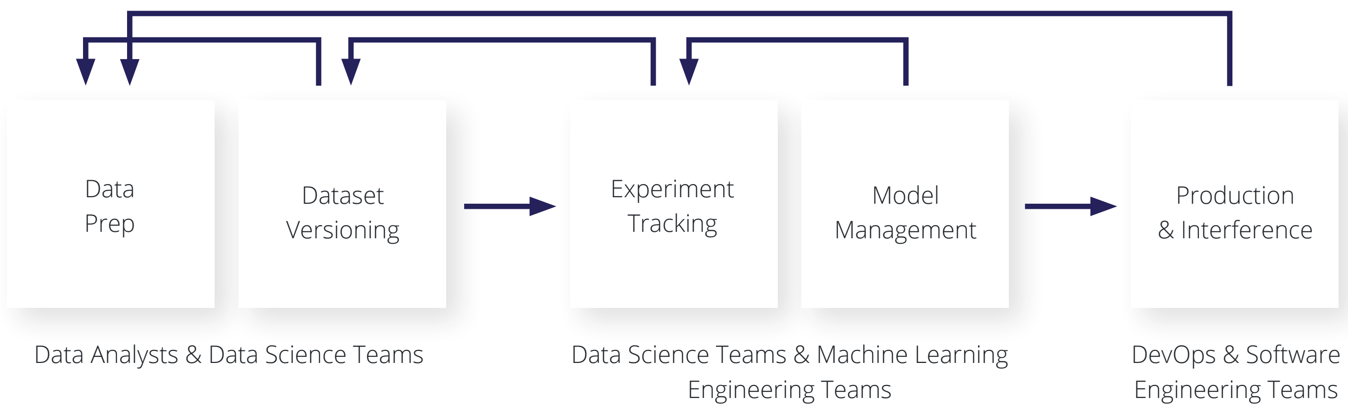

Figure 4: Driving collaboration across teams can drive significant value.

Conclusion

Despite significant progress in the machine learning and AI space, implementing scalable teams, frameworks, and processes still presents critical challenges. Even the best teams struggle with effective iteration of models, reproducible work, and maintaining institutional knowledge when teammates leave. Leaning only on traditional software engineering practices and tools contribute to these blockers and impact production model performance.

Remember the 45% of seasoned adopters that are innovating and getting ahead of the competition?

The critical takeaway is that the teams that figure out how to attain visibility, reproducibility and collaboration across their machine learning and AI teams and processes will not just join that exclusive club, but exceed beyond it.

Get started driving visibility, reproducibility, and collaboration for free today.
www.comet.ml/site/data-scientists/

Comet provides a self-hosted and cloud-based meta machine learning platform allowing data scientists and teams to track, compare, explain and optimize experiments and models.

Backed by thousands of users and multiple Fortune 100 companies, Comet provides insights and data to build better, more accurate AI models while improving productivity, collaboration and visibility across teams.

Learn more at www.comet.ml