Meta-analysis of the machine learning operations open source ecosystem

Florida Polytechnic University, Spring 2022

Isabel Zimmerman   Reinaldo Sanchez-Arias, Ph.D., Advisor

11/28/22

Abstract

Machine learning operations, or MLOps, is a set of practices to deploy and maintain machine learning models in production reliably and efficiently. This paper will explore the open source MLOps ecosystem as a whole and highlight tools of interest. MLOps tools will be compared and classified by task and maturity. Finally, there will be an overview of key players in the space. It seeks to highlight the knowledge necessary for practitioners and academics to select an open source machine learning operations tool (or tools) to suit their data science projects and machine learning systems.

Literature review

With the proliferation of machine learning tools, a secondary ecosystem of frameworks to bring machine learning models to production and support them in production has emerged. Machine learning operations, or MLOps, was first introduced with the realization of the massive technical debt that occurs quickly in real-world ML systems [1]. MLOps is a set of practices to deploy and maintain machine learning models in production reliably and efficiently [2]. These practices seek to shorten the time from developing a new machine learning model to bringing it into production [3]. Once in production, MLOps seeks to give practitioners agility in updating or rolling back a model, monitoring model performance, and otherwise maintain the model deployment.

There are many choices to make when building an MLOps system. MLOps systems formed from open source solutions generally require multiple open source projects to do tasks such as: versioning a model, deploying a model, monitoring deployment, orchestrating tasks, and managing infrastructure. To manage these projects, practitioners can use scorecards to rate their MLOps systems, gaining points for reduction of technical debt, using best practices, and leveraging continuous integration when possible [4].

Each open source project falls into one (or sometimes multiple) different categories to fulfill these different tasks: all-in-one, data pipeline, infrastructure, modeling and training, monitoring, and serving [5]. Current literature on MLOps often focuses on case studies for a specific company or industry's MLOps tools [6]. While some papers propose state-of-the-art of MLOps solutions, the MLOps ecosystem is constantly changing, with new tools and frameworks being released regularly, in an attempt to address the complex concerns of monitoring and managing ML systems [7].

Knowing the maturity of a project is crucial for adoption. Open source projects follow a life cycle, tracked as a maturity model cycle [8]. As shown in Figure 1, maturity cycles start with an unknown open source project that then goes into rapid adoption (technology trigger), a peak of activity beyond early adopters (peak of inflated expectations), a by rapid decrease as users lose interest or are unable to use the project as intended (trough of disillusionment), a generation of best practices (slope of enlightenment), and a final general plateau for users who will continue to use this product going forward (plateau of productivity). This cycle applies to not only an individual project, but also entire ecosystems as a whole.

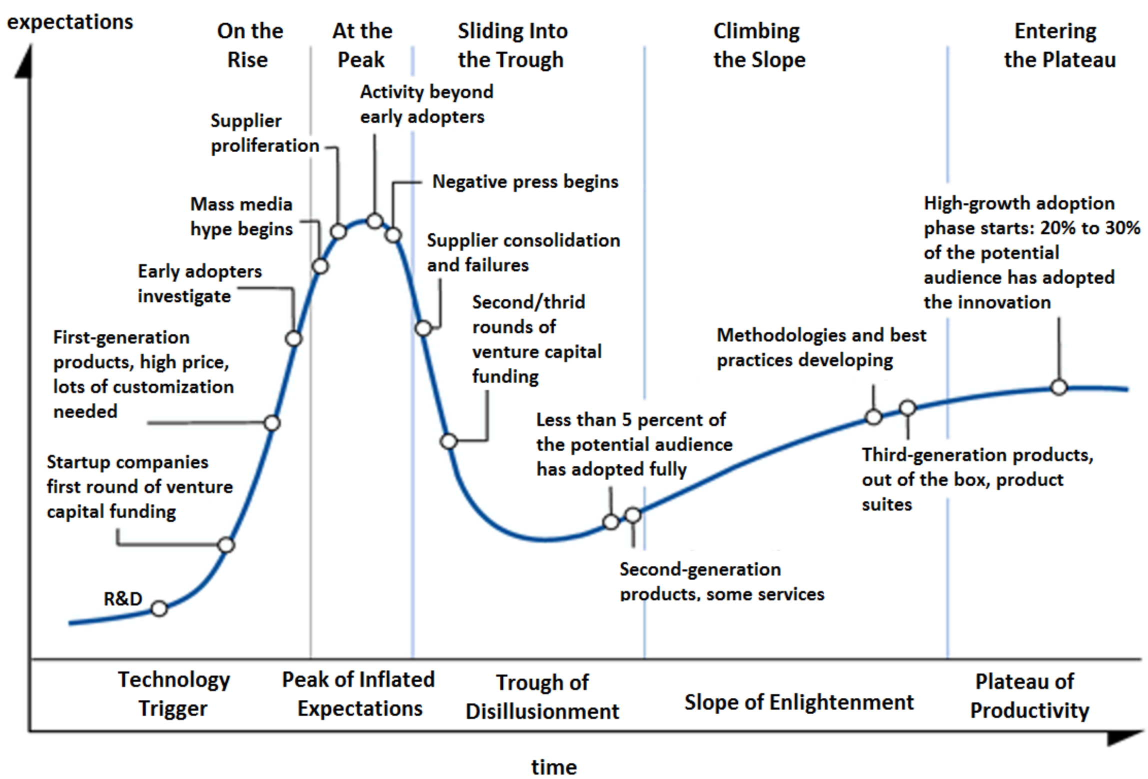

Figure 1: Hype cycle chart from [8]

It is difficult to determine where open source ecosystems lie on the maturity model as download rates are not accurately available. It is possible to proxy this information by looking at other metrics such as stars, contributions, and other repository activity.

In all, machine learning operations in the open source world is an emerging field. It is clear that tools exist and span a variety of tasks to help practitioners bring models into production. Knowing the maturity of this ecosystem can help contextualize machine learning operations tools as a whole.

The interdisciplinary nature of proposed research

Machine learning operations have a naturally interdisciplinary nature. In fact, many of the reasons effectively implementing MLOps solutions is so difficult is due to the span of knowledge necessary for successful operationalization of models. In researching this topic, it is necessary to have knowledge of data wrangling and statistical modeling to create the model (as taught in *CAP 5320* - Data Wrangling and Exploratory Data Analysis and *CAP 5765* - Computational Data Analysis), along with knowledge of software engineering best practices to have a stable deployment (*CEN 5035* - Advanced Software Engineering) and visualization skills to continually monitor model performance (*CAP 5735* - Data Visualization and Reproducible Research).

Deliverables

The intended deliverable of this project is a paper that both lays out an overview of the open source MLOps ecosystem as a whole and highlights specific tools of interest. This paper evaluates and compares open source tools based on their functionality and maturity, and provides an overview of key players in the space.

Expected outcomes

- Research paper
- Interactive dashboard

Through the expected outcomes, the implications of machine learning operations in open source software will be clear. The paper will lay out the types of tools available for operationalization of models, clustering of types of tools, and an analysis of the maturity of the ecosystem. The dashboard will give an interactive way to explore the data and maturity of the machine learning operations open source ecosystem.

Objective timeline

As seen in Figure 2, this project will be broken up into the following parts: literature review, prepare project proposal, gather data, clean and explore data, build maturity models, build unsupervised learning models, build interactive dashboard, and write final report. The literature review and project proposal will be finished by December 2023. Starting in December 2022, data collection of many open source projects will begin. The first task will be to curate a list of projects, sources such as blog posts and GitHub tags will be used. From this list of repositories, further data on these tools can be scraped from the GitHub API. This data will give information such as: description, stars, number of contributors, forks, language, open issues, etc. Once gathered, exploratory data analysis is expected to begin January 2023e. By mid-February, it is expected to be starting on building unsupervised learning models for categorizing tools, maturity models for determining maturity, and a dashboard for an interactive experience to explore the data and models. Starting in March, the final report will be written, to be presented late-April.

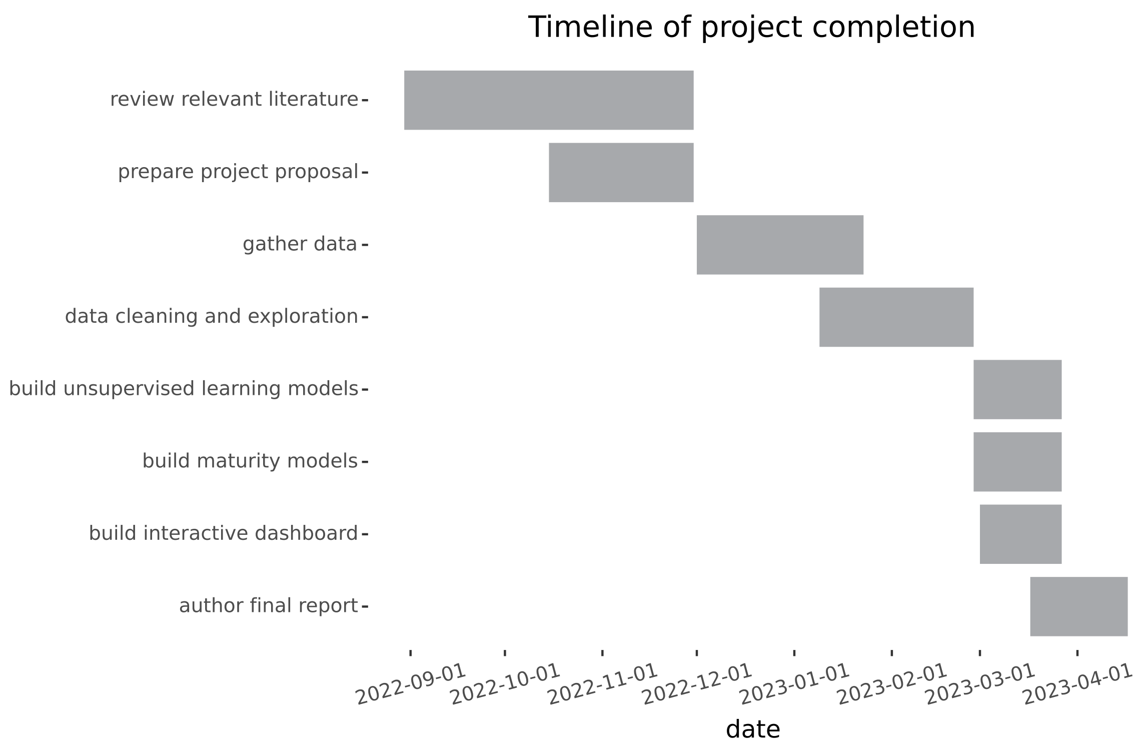

Figure 2: Project Gantt chart

```
<ggplot: (309344464)>
```

Further work

The discoveries from this work could easily be extended further. One possible outcome could be an opinionated workflow for a specific task built with a selection of the tools analyzed; this workflow would contribute to these open source tools through documentation and tutorials, if not also code contributions. It is possible that this software stack could be used to build an accompanying case study, either to be included inside this paper or as part of another work. Another outcome could be a second paper in the style of “Ten simple rules for X”, where the topic would be related to machine learning operations.

References

[1] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.F. Crespo, and D. Dennison. “Hidden technical debt in machine learning systems.” *Advances in neural information processing systems*, 2015, pp. 2502-2511. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

[2] C. Breuel. “ML Ops: Machine Learning as an Engineering Discipline.” Towards Data Science. Accessed: Nov. 20, 2022. [Online]. Available: https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f

[3] S. Islam. “MLOps vs. DevOps: What is the Difference?” pdData. Accessed: Nov. 20, 2022. [Online]. Available: https://www.phdata.io/blog/mlops-vs-devops-whats-the-difference/

[4] C. Huyen, *Designing Machine Learning Systems*, 1st ed. Sebastopol, CA, USA: O’Reilly, pp. 293–327.

[5] E. Breck, S. Cai, E. Nielsen, M. Salib and D. Sculley, “The ML test score: A rubric for ML production readiness and technical debt reduction,” *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1123-1132, doi: 10.1109/BigData.2017.8258038.

[6] N. Muralidhar, S. Muthiah, P. Butler, M. Jain, Y. Yu, K. Burne, W. Li, D. Jones, P. Arunachalam, H.S. McCormick, and N. Ramakrishnan. “Using AntiPatterns to avoid MLOps Mistakes.” *arXiv*, 2021. [Online]. Available: https://arxiv.org/pdf/2107.00079.pdf

[7] P. Ruf, M. Madan, C. Reich, and D. Ould-Abdeslam, “Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools,” *Applied Sciences*, vol. 11, no. 19, p. 8861, Sep. 2021, doi: 10.3390/app11198861. [Online]. Available: http://dx.doi.org/10.3390/app11198861

[8] “Open-source Maturity Model.” Open-source Quality Research Project. Accessed: Nov. 20, 2022. [Online]. Available: https://open-source-quality.github.io/2018/02/01/open-source-maturity-model.html