



A guide to MLOps

[Website](#) · [GitHub](#)

Swiss AI Center contributors

This work is licensed under the [CC BY-SA 4.0](#) license.

INTRODUCTION

SWISS AI CENTER

Five HES from the HES-SO (HEIG-VD, HEIA-FR, HE-Arc, HEVS and HEPIA) work on a project called Centre Suisse d'Intelligence Artificiel à destination des PME (CSIA-PME), also known as the Swiss AI Center.

The Swiss AI Center's mission is to **accelerate the adoption of artificial intelligence in the digital transition of Swiss SMEs.**

HEIG-VD is responsible for **setting up tools to manage ML experiments from code to production.**

OUR TEAM

**Bertil
Chapuis**
Professor



[Mail](#) · [LinkedIn](#)

**Ludovic
Delafontaine**
aR&D Associate



[Mail](#) · [LinkedIn](#)

**Rémy
Marquis**
aR&D Associate



[Mail](#) · [LinkedIn](#)

**Leonard
Cseres**
Assistant

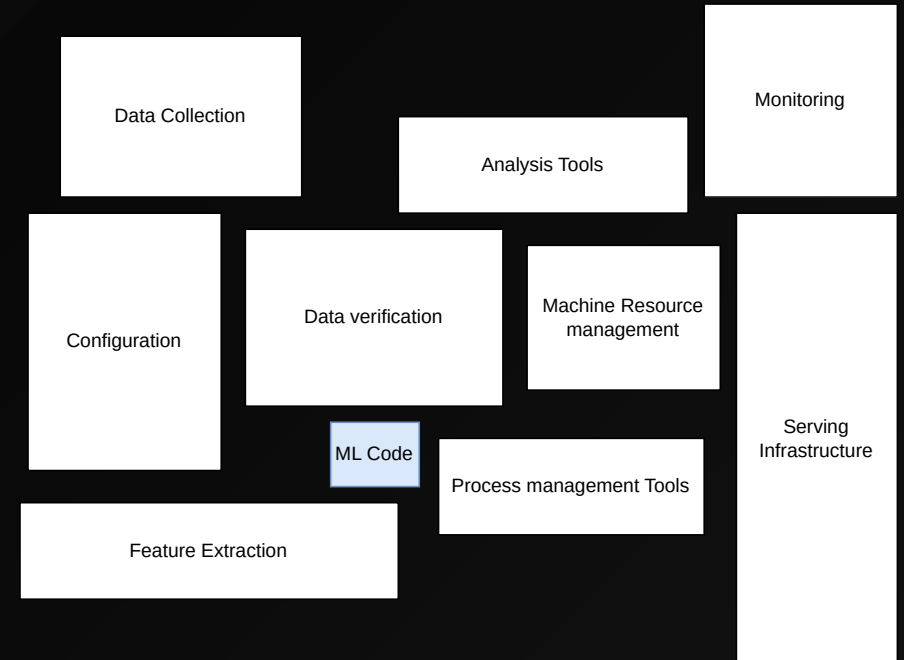


[Mail](#) · [LinkedIn](#)

ML CODE VS ML SYSTEM

Only a small fraction of real-world ML systems is composed of the ML code.

The required surrounding infrastructure is vast and complex.



DIFFICULTIES WITH ML PROJECTS

Get out of the context of the experience

“ I ran the experiment but didn't get the same results, can you check my way of running the experiment? ”

Make sure you can build the model at all times

“ I tried to build the model on my machine but it doesn't work... Are you sure it builds on yours? ”

Monitor the evolution of the model over time

“ I’m not sure my changes really help the model’s performances... I hope it still works in production. ”

Move to production quickly, efficiently and in a semi-automated way

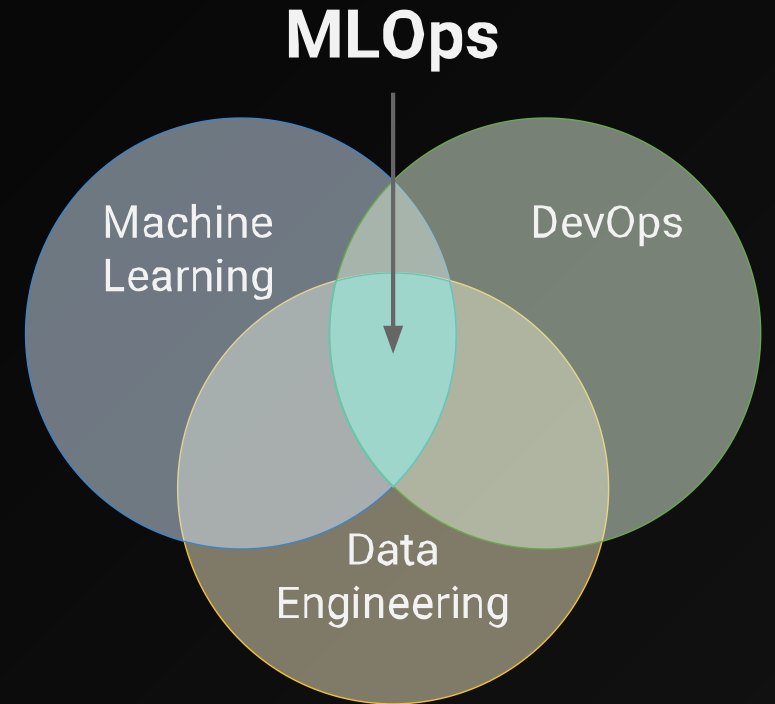
“ Is your model available in production? Can I use it with my mobile app/website? How can I do so? ”

SMALL AND MEDIUM-SIZED ENTERPRISES (SMES) FACE THE SAME PROBLEMS

A SOLUTION

MLOps

- ➔ Draw inspiration from Software and DevOps best practices
- ➔ Adapting these practices to the world of machine learning
- ➔ Improve the management and quality of machine learning projects



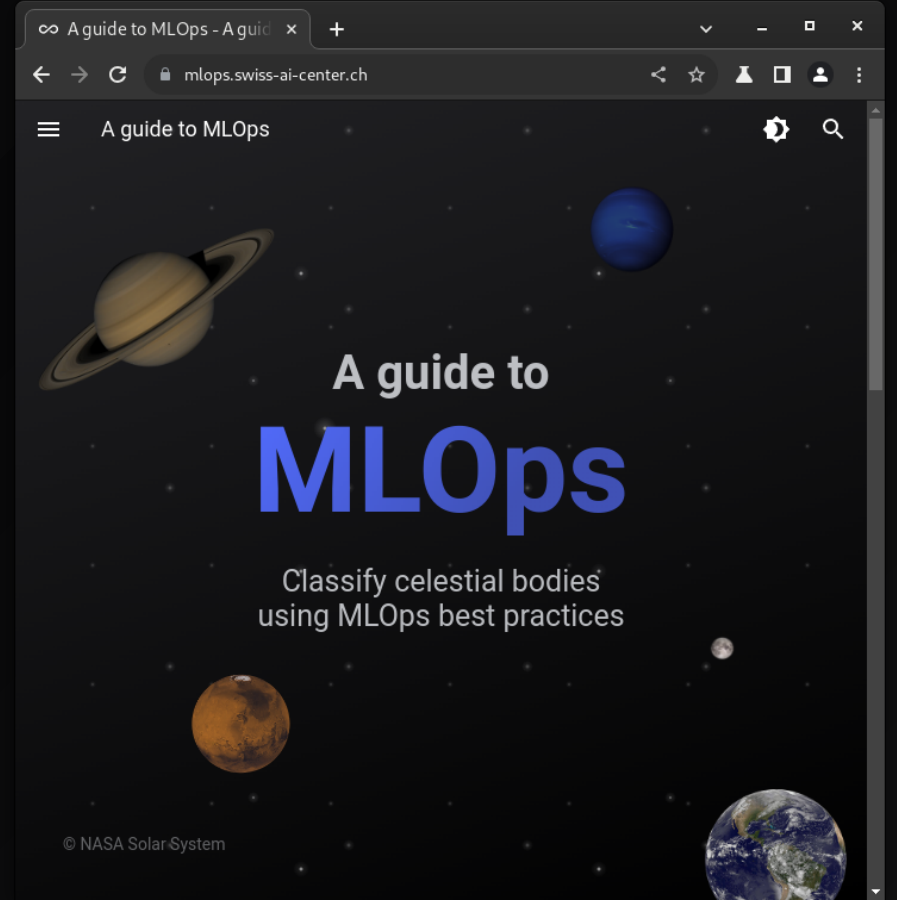
OUR PROPOSAL

A guide to MLOps

🔧 Switch from a Jupyter Notebook to production using state-of-the-art MLOps tools

🚀 Go from experience to production on the Cloud

📖 Use the best practices for ML



A GUIDE TO MLOPS

A quick presentation of the guide

"WELCOME TO THE TEAM!"

You just have joined a team of data scientists and machine learning (ML) engineers (*welcome!*).

The team is working on a model capable of visually identifying planets or moons within our solar system from images in a Jupyter Notebook.

The team is facing difficulties to move the model to production.

Your mission is to help the team to improve the model and deploy it to the cloud using MLOps best practices.

THE BIG PICTURE

Codebase



Git

Data + Reproduce



DVC

Tracking



DVC

CML

Serving + publishing



BentoML



Docker

Storage



Google Cloud

Labellisation



Label Studio

Orchestration



GitHub

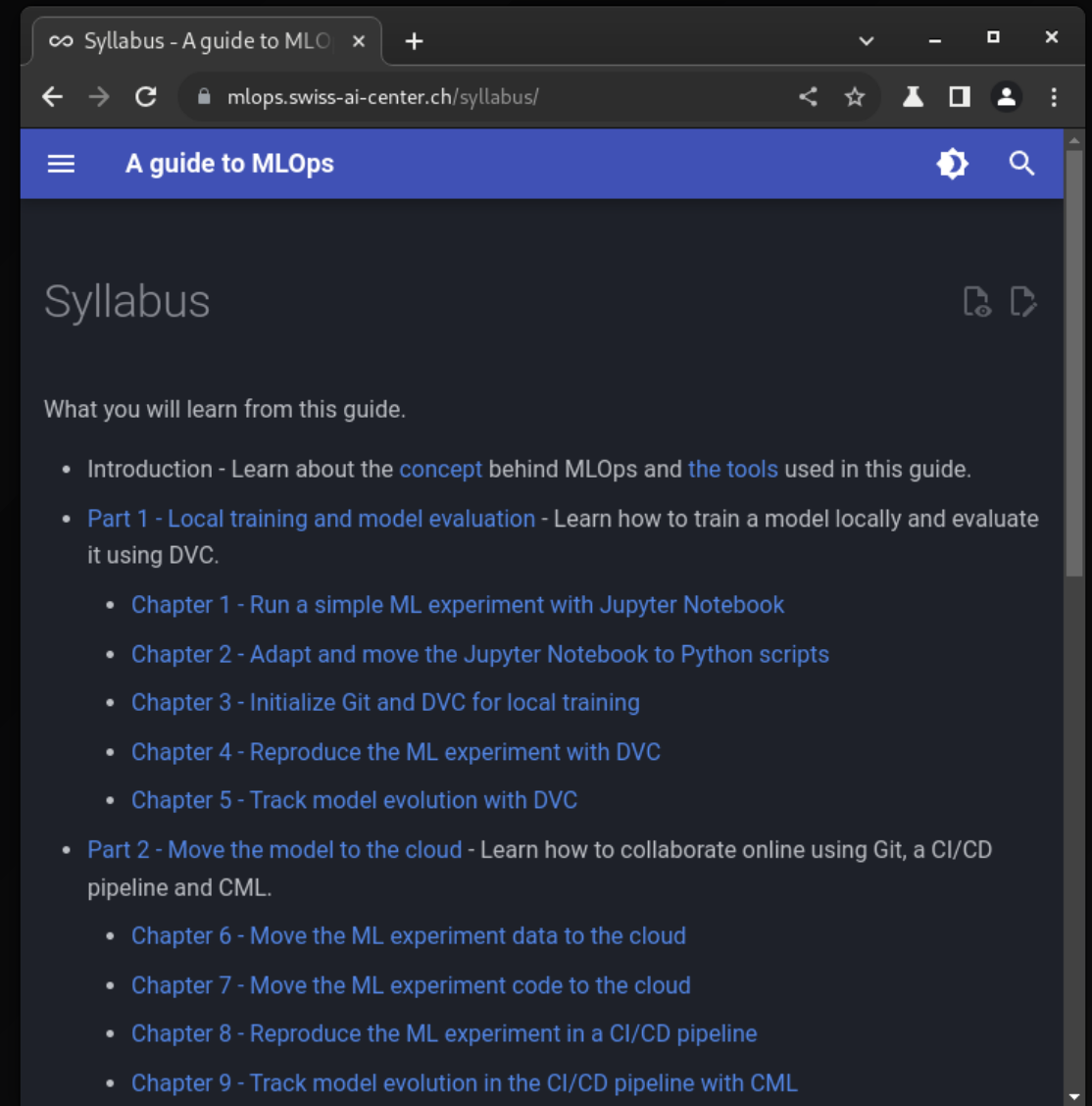
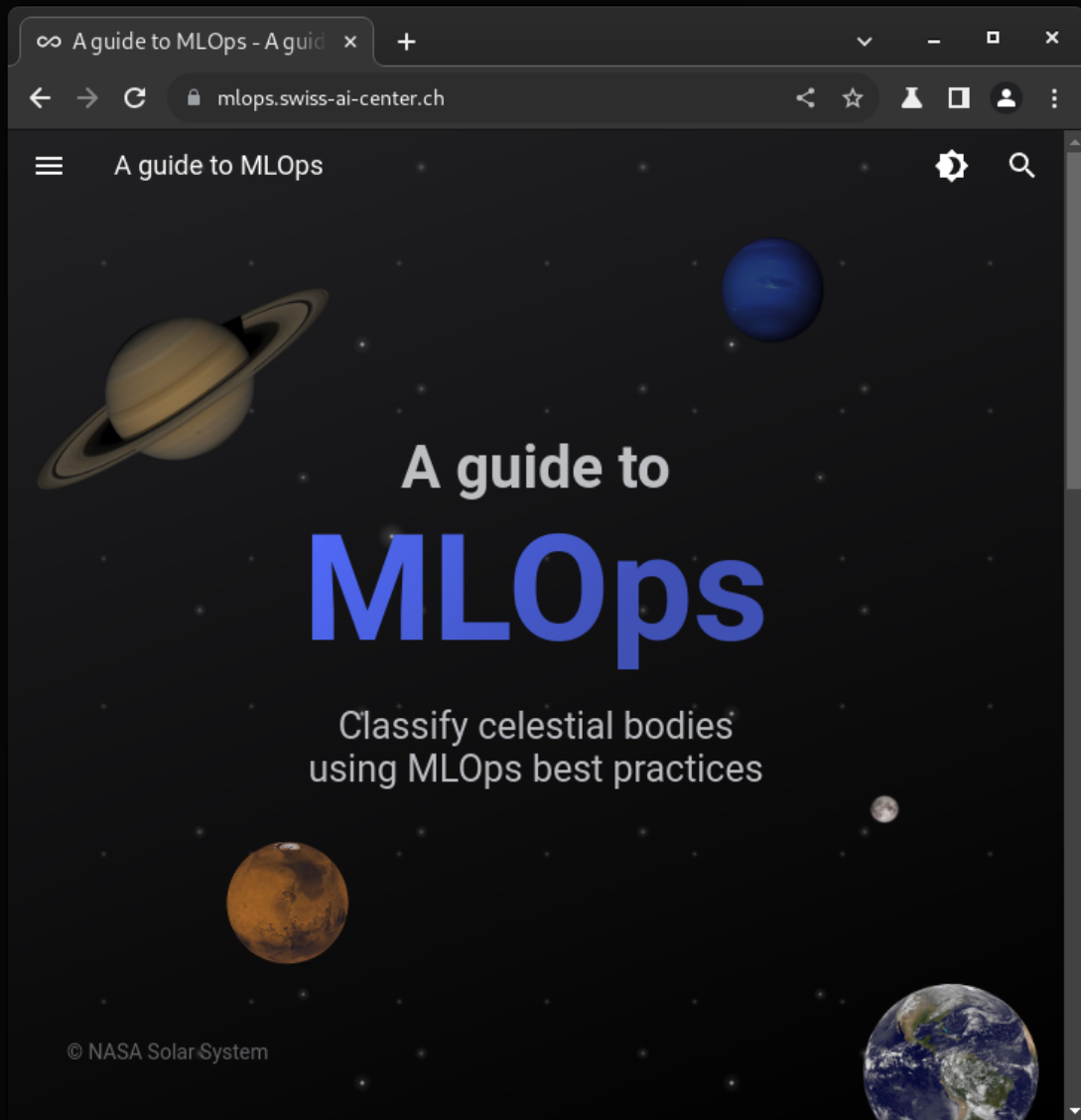


GitLab

Deployment



Kubernetes



The screenshot shows a GitHub issue page for the repository 'ludelafo / a-guide-to-mlops'. The issue title is 'Made some changes to the model #5'. It is marked as 'Open' and was opened 32 minutes ago. The issue description is 'No description provided.' Below the description is a comment from 'ludelafo' with the same text. At the bottom, there is a 'Add a comment' section with a rich text editor.

The screenshot shows a GitHub pull request page for the repository 'ludelafo / a-guide-to-mlops'. The pull request title is 'Made some changes to the model #6'. It is marked as 'Open' and is intended to merge 1 commit into the 'main' branch from the '5-demonstrate-model-evolution-tracking' branch. The pull request includes a commit titled 'Change parameters' with commit hash '8bde18e'. A comment from 'ludelafo' states 'No description provided.' and links to issue #5, 'Made some changes to the model #5', which is marked as 'Open'.

github.com/ludelafo/a-guide-to-mlops/actions/runs/...

← MLOps

✓ Made some changes to the model #5

✓ train-and-report ▾

train-and-report Beta Give feedback 🔄 ⚙️
succeeded last week in 3m 37s

- > ✓ Set up job 3s
- > ✓ Checkout repository 5s
- > ✓ Setup Python 20s
- > ✓ Install dependencies 1m 30s
- > ✓ Login to Google Cloud 0s
- > ✓ Train model 53s
- > ✓ Setup CML 31s
- > ✓ Create CML report 11s
- > ✓ Post Login to Google Cloud 0s
- > ✓ Post Setup Python 0s
- > ✓ Post Checkout repository 0s

github.com/ludelafo/a-guide-to-mlops/actions/runs/...

train-and-report Beta Give feedback 🔄 ⚙️
succeeded last week in 3m 37s

✓ Train model 53s

```

25 Stage 'evaluate' is cached - skipping run, checking out outputs
26
27 To track the changes with git, run:
28
29     git add data/raw.dvc
30
31 To enable auto staging, run:
32
33     dvc config core.autostage true
34 Use `dvc push` to send your updates to remote storage.

```

> ✓ Setup CML 31s

✓ Create CML report 11s

```

1 ▶ Run # Fetch all other Git branches
2 From https://github.com/ludelafo/a-guide-to-mlops
3 * [new branch]      main      -> main
4 * [new branch]      main      -> origin/main
5 file:///home/runner/work/a-guide-to-mlops/a-guide-to-mlops/dvc_plots/index.html
6 https://github.com/ludelafo/a-guide-to-mlops/pull/4#issuecomment-1964463218

```

> ✓ Post Login to Google Cloud 0s

> ✓ Post Setup Python 0s

> ✓ Post Checkout repository 0s

github.com/ludelafo/a-guide-to-mlops/pull/4

Jump to bottom

Made some changes to the model #4

Merged

ludelafo merged 3 commits into `main` from `3-demonstrate-model-evolution-tracking` last week

Conversation 1 | Commits 3 | Checks 1 | Files changed 3

ludelafo commented last week

No description provided.

ludelafo linked an issue [last week](#) that may be closed by this pull request

Demonstrate model evolution tracking #3 Closed

github-actions bot commented last week • edited

Experiment Report (7a44cb8)

Params workflow vs. main

Path	Param	main	workspace
params.yaml	train.conv_size	32	64
params.yaml	train.dense_size	64	128
params.yaml	train.lr	0.0001	0.001

Metrics workflow vs. main

Path	Metric	main	workspace	Change
evaluation/metrics.json	val_acc	0.68536	0.96885	0.28349
evaluation/metrics.json	val_loss	1.32753	0.16357	-1.16396

Plots

Training History

github.com/ludelafo/a-guide-to-mlops/pull/4

Merged

Made some changes to the model #4

ludelafo merged 3 commits into `main` from `3-demonstrate-model-evolution-tracking` last week

github-actions bot commented last week • edited

Experiment Report (7a44cb8)

Params workflow vs. main

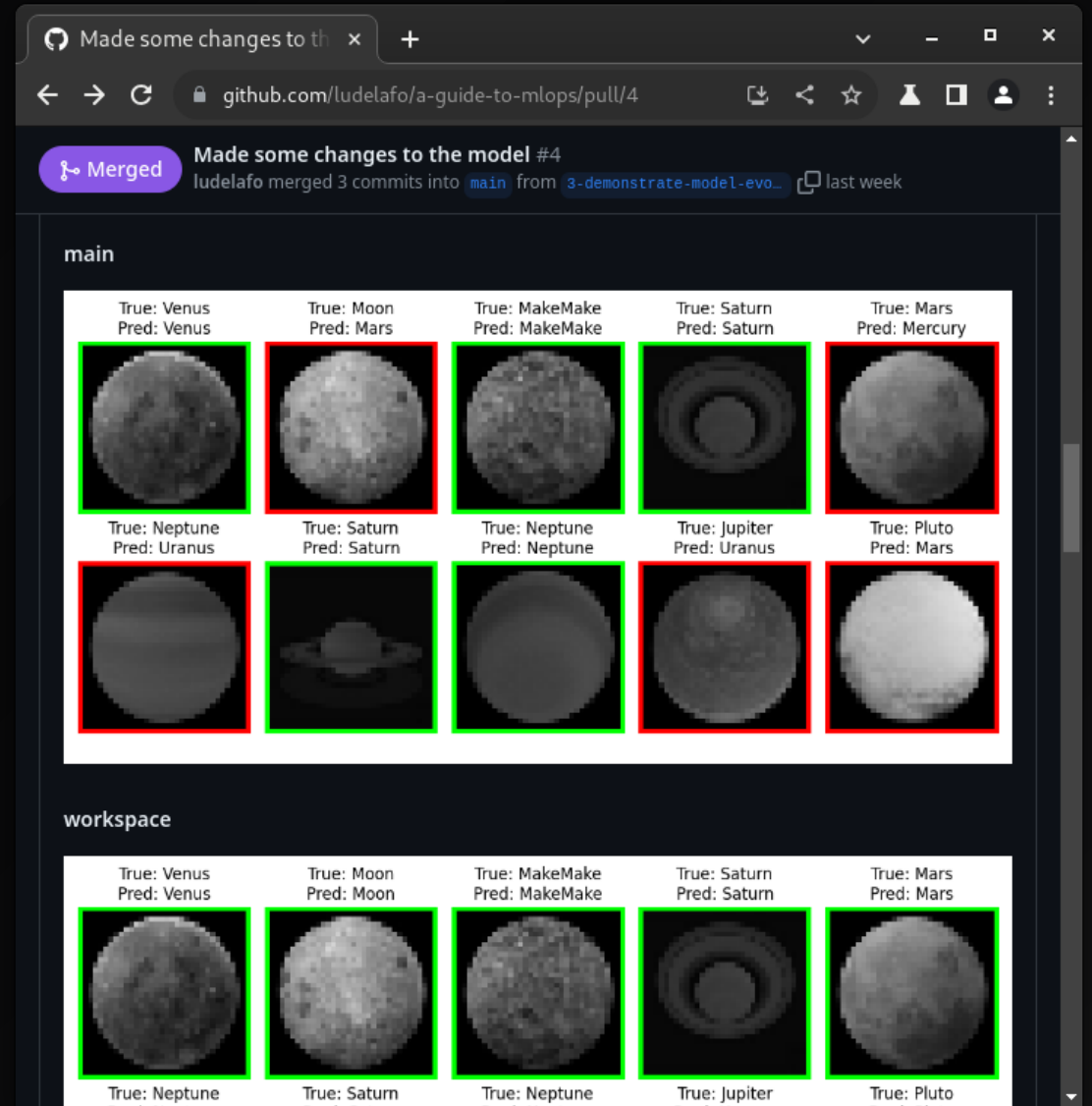
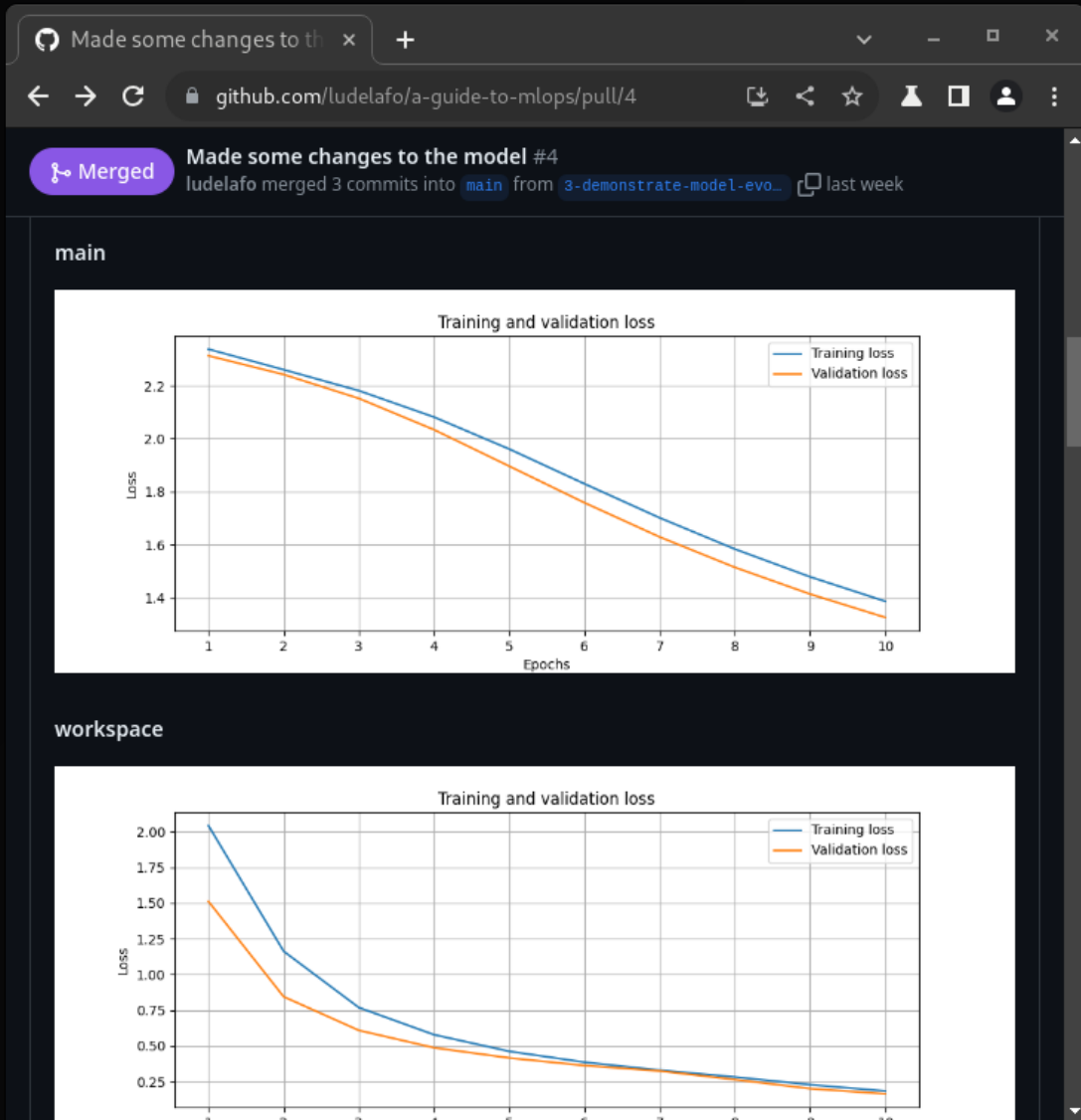
Path	Param	main	workspace
params.yaml	train.conv_size	32	64
params.yaml	train.dense_size	64	128
params.yaml	train.lr	0.0001	0.001

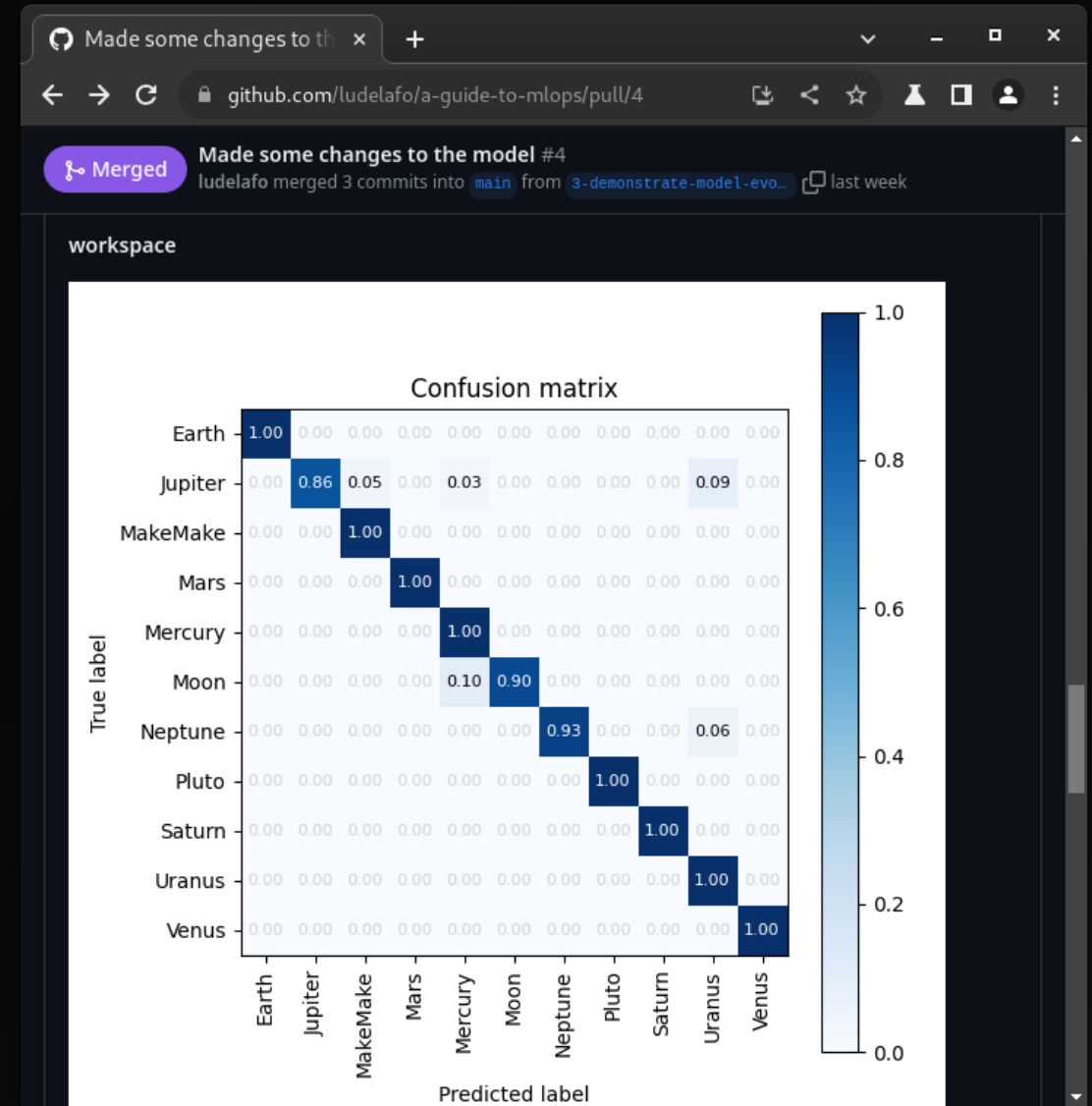
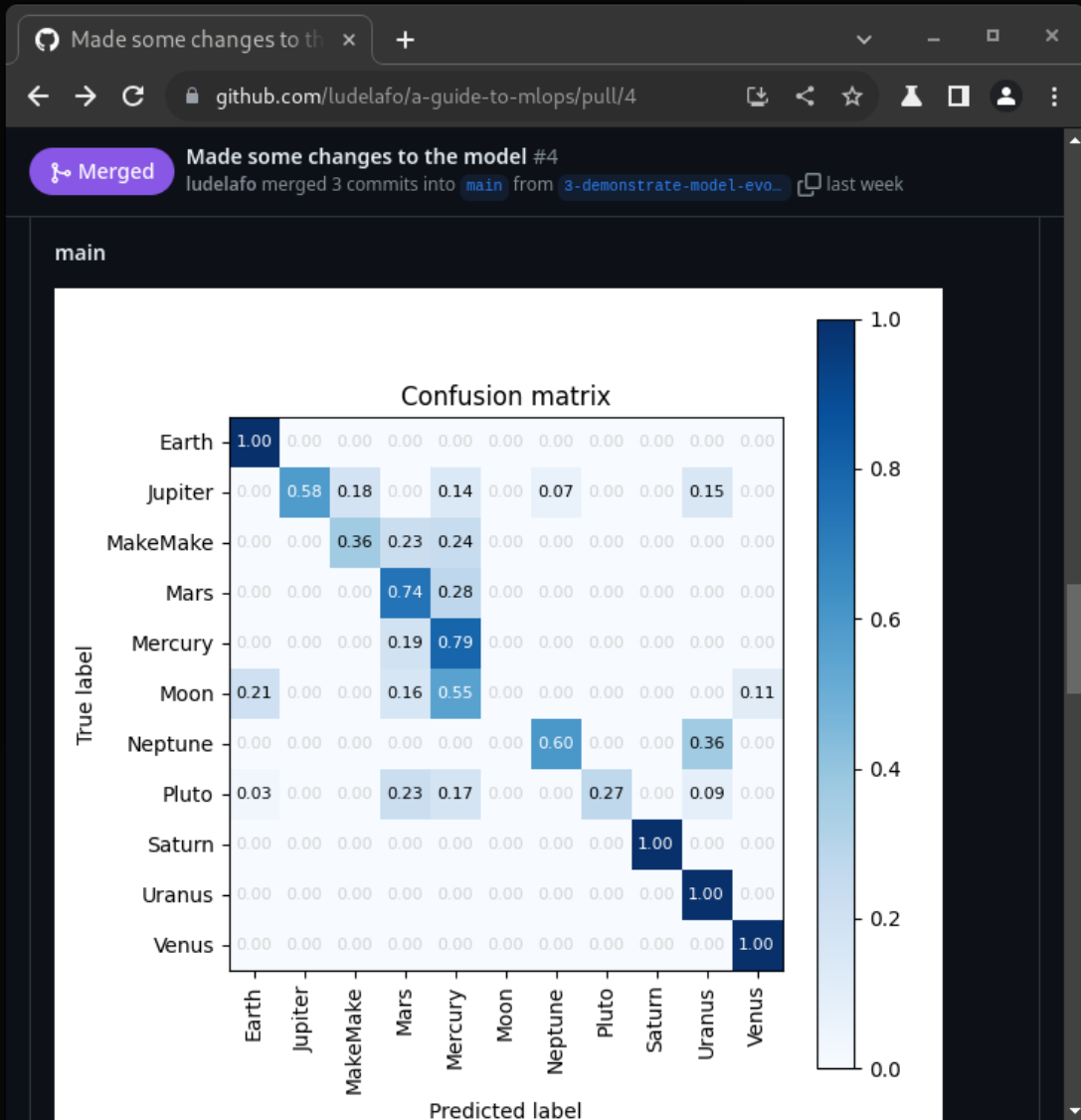
Metrics workflow vs. main

Path	Metric	main	workspace	Change
evaluation/metrics.json	val_acc	0.68536	0.96885	0.28349
evaluation/metrics.json	val_loss	1.32753	0.16357	-1.16396

Plots

Training History





Bump docker actions vers x +

github.com/ludelafo/a-guide-to-mlops/actions/runs/...

train-report-publish-and-deploy

Beta Give feedback

succeeded last week in 8m 31s

- > ✓ Set up job 4s
- > ✓ Checkout repository 2s
- > ✓ Setup Python 13s
- > ✓ Install dependencies 1m 36s
- > ✓ Login to Google Cloud 0s
- > ✓ Train model 1m 27s
 - ⊗ Setup CML 0s
 - ⊗ Create CML report 0s
- > ✓ Log in to the Container registry 2s
- > ✓ Import the BentoML model 2s
- > ✓ Build the BentoML 'Bento' 41s
- > ✓ Containerize and publish the Bento 4m 11s
- > ✓ Get Google Cloud's Kubernetes credentials 1s
- > ✓ Update the Kubernetes deployment 0s
- > ✓ Deploy the model on Kubernetes 3s

Bump docker actions vers x +

github.com/ludelafo/a-guide-to-mlops/actions/runs/...

train-report-publish-and-deploy

Beta Give feedback

succeeded last week in 8m 31s

- ✓ Containerize and publish the Bento 4m 11s


```

1024 ba473bdfd54e: Layer already exists
1025 40c5e3152548: Layer already exists
1026 latest: digest:
      sha256:e908db62950d449c1df8823c1d4e536051b9eb8ceab72cb13eb117d4b2cbfb6a size: 3457


```
- > ✓ Get Google Cloud's Kubernetes credentials 1s
- > ✓ Update the Kubernetes deployment 0s
- ✓ Deploy the model on Kubernetes 3s



```


1 ▶ Run kubectl apply \
23 deployment.apps/celestial-bodies-classifier-deployment unchanged
24 service/celestial-bodies-classifier-service unchanged

```
- > ✓ Post Log in to the Container registry 0s
- > ✓ Post Login to Google Cloud 0s
- > ✓ Post Setup Python 0s
- > ✓ Post Checkout repository 0s
- > ✓ Complete job 0s

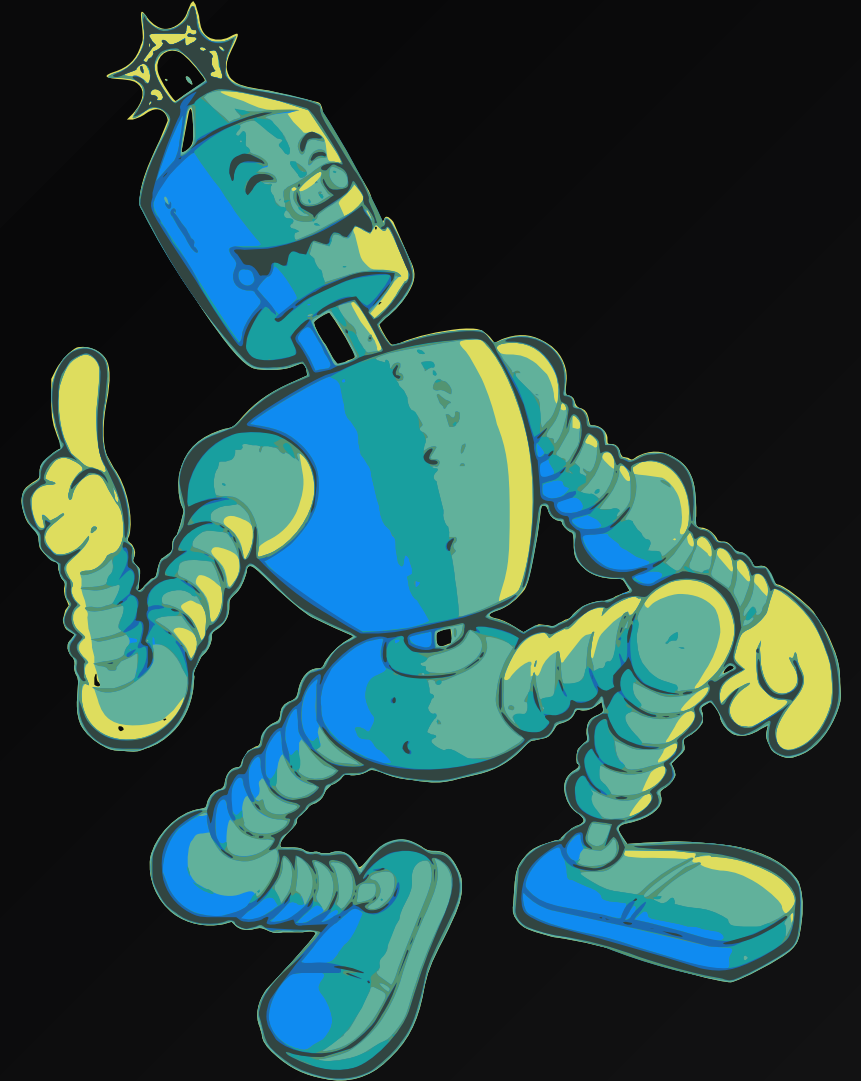
TARGET AUDIENCE

 You regularly work with machine learning projects

 You want to improve processes to ensure quality

 You want to consolidate your current infrastructure

 You want to move to the Cloud



PREREQUISITES

- ♿ Accessible to everyone!
- 🧠 Basic knowledge of Python and terminal is sufficient
- 💳 A valid credit card for cloud deployment
- 🤝 You will be accompanied throughout the guide!



TECHNICAL CHECKS

Before we start:

 macOS, Linux, Windows with WSL2

 Editor and Terminal (VS Code recommended)

 Python 3.11, pip, git, wget, unzip, docker

 GitHub account, Google Cloud account

ACCESS THE GUIDE

👉 Access the guide at mlops.swiss-ai-center.ch.

💪 Feel free to open an issue on [GitHub](#) if you encounter any difficulties or want to contribute.

🙏 Leave us a star if you like the guide!



NOW IT'S YOUR TURN!

Feel free to ask questions, share your feedback and contribute to the guide!

We are here to help.

CLEAN UP

Now that you have completed the guide, it is important to properly manage and remove the resources and environments you have created.

This is necessary to avoid:

- unnecessary incurring costs
- potential security concerns

CONCLUSION

Congratulations! You have completed the guide to MLOps!

You have learned how to improve the management and quality of machine learning projects.

You are now able to switch from a Jupyter Notebook to production using state-of-the-art MLOps tools.

You can go from experiment to production on the Cloud, using the best practices for ML. 🚀

BONUS SLIDES

USUAL ML WORKFLOW

Each member of the team manages their own codebase, their own dataset and their own models.

The reproducibility of the model creation is difficult and cannot be guaranteed over time.

Improvements made to the model are hard to track.

Models are hard to share and deploy in production.

HIGH FLEXIBILITY FOR THE TEAM...

...but hard to maintain.

...hard to reproduce in the future.

...time consuming.

We can do better.

CODEBASE (1/2)

Current situation

- Each developer has its own codebase
- Sharing the code with peers is difficult

CODEBASE (2/2)

What we are trying to improve

- Allow each developer to improve a common codebase
- Quickly benefit of the work from others



DATA (1/2)

Current situation

- The dataset must be manually downloaded and put in the right place
- Different datasets are being used at the same time
- Datasets are hard to improve

DATA (2/2)

What we are trying to improve

- Allow the usage of a common and up-to-date dataset
- Efficiently share new revisions to train the model
- Datasets can be stored anywhere



REPRODUCE (1/2)

Current situation

- Steps to create the model can be complex
- Intermediate commands should not be skipped
- Hyperparameters are hard to track from one run to another

REPRODUCE (2/2)

What we are trying to improve

- Document the steps to reproduce the experiment
- Ensure it can be run anytime in the future
- DVC can improve time efficiency



TRACKING (1/2)

Current situation

- Changes to a model are difficult to track
- Visualize the differences are hard
- Cannot guarantee the changes are beneficial

TRACKING (2/2)

What we are trying to improve

- Have a visual way to identify the consequences of the changes made to a model
- Errors/anomalies are easily identified



SERVING AND PUBLISHING (1/2)

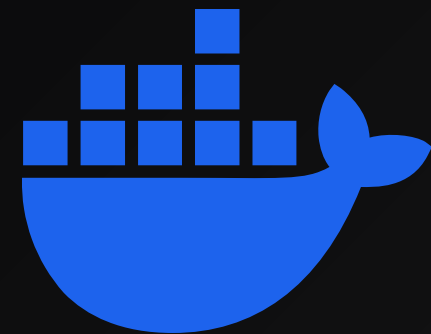
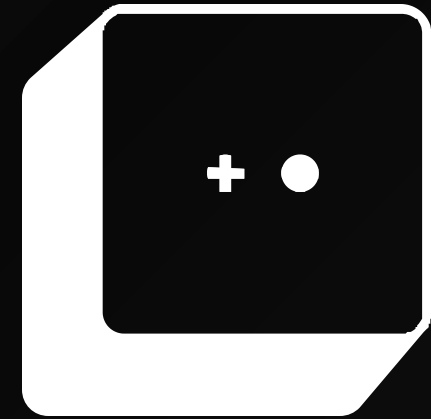
Current situation

- The model is hard to use outside the experiment context
- The model is hard to deploy in production
- The model is hard to share with others

SERVING AND PUBLISHING (2/2)

What we are trying to improve

- The model can be used outside the experiment context
- The model can be deployed in production
- The model can be shared with others



DEPLOYMENT (1/2)

Current situation

- An experiment can run on one machine but can fail on another
- Models must be prepared to be run outside its experiment context
- Exposing the model to the outside world is hard

DEPLOYMENT (2/2)

What we are trying to improve

- Run the experiment in a clean state to ensure it works everywhere
- Package the model with all its dependencies
- The model can be used over the Internet by other applications
- Automate the process



LABELLISATION (1/2)

Current situation

TODO

LABELLISATION (2/2)

What we are trying to improve

TODO



SOURCES

- MLOps Venn diagram by Cmbreuel on [Wikipedia](#)
- ML system diagram by [D. Sculley et. al. NIPS 2015: Hidden technical debt in Machine learning systems](#)
- Robot illustration by [OpenClipart-Vectors](#) on [Pixabay](#)