A guide to MLOps

<u>Website</u> · <u>GitHub</u>

Swiss AI Center contributors

This work is licensed under the <u>CC BY-SA 4.0</u> 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 PMEs (CSIA-PME)**, also known as the **Swiss Al 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 Ludovic Delafontaine aR&D Associate **Rémy Marquis** aR&D Associate

Leonard Cseres Assistant







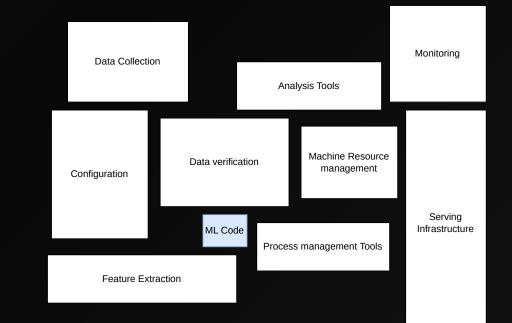


Mail · LinkedIn Mail · LinkedIn Mail · LinkedIn Mail · LinkedIn

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?

Swiss Al Center - A guide to MLOps 2024 - CC BY-SA 4.0

6

"

"

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

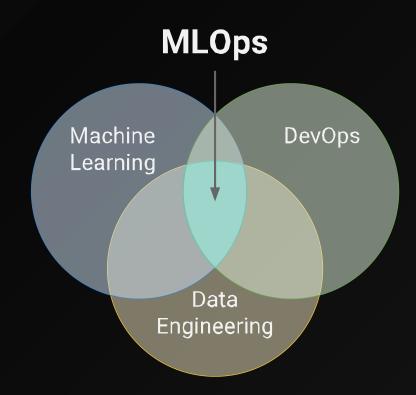
A SOLUTION

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



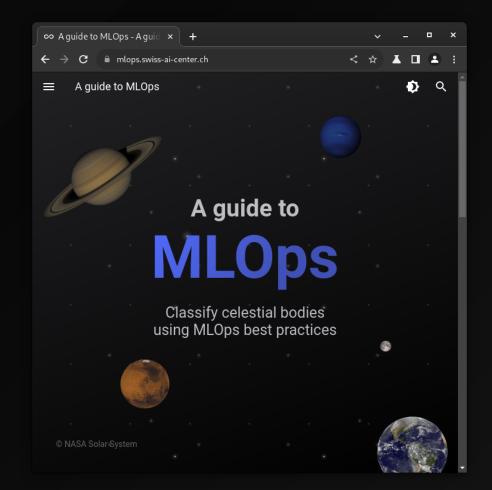
OUR PROPOSAL

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!"

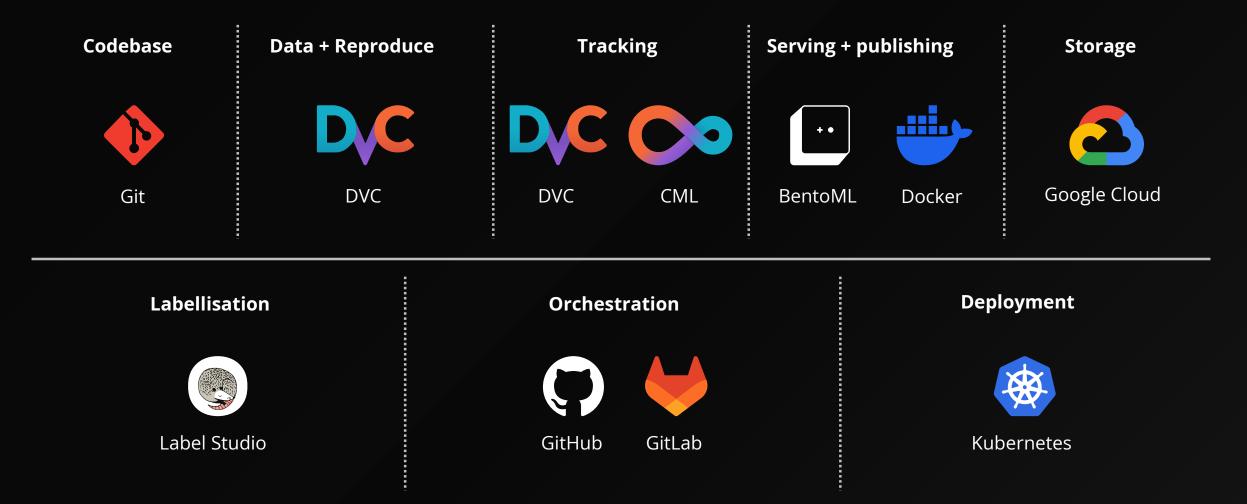
You just have joined a team of data scientists and machine learning (ML) engineers (*welcome!*).

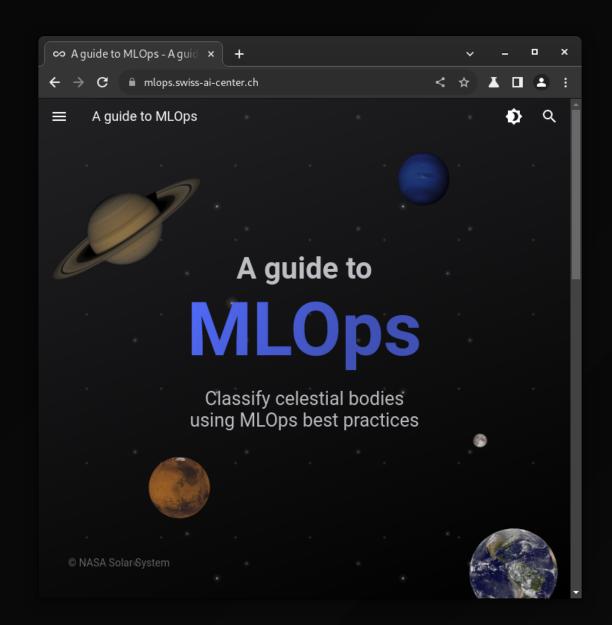
The team is working on a model capable of visually identifying planets or moons within our solar system from images in a Jupyter Notebook.

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





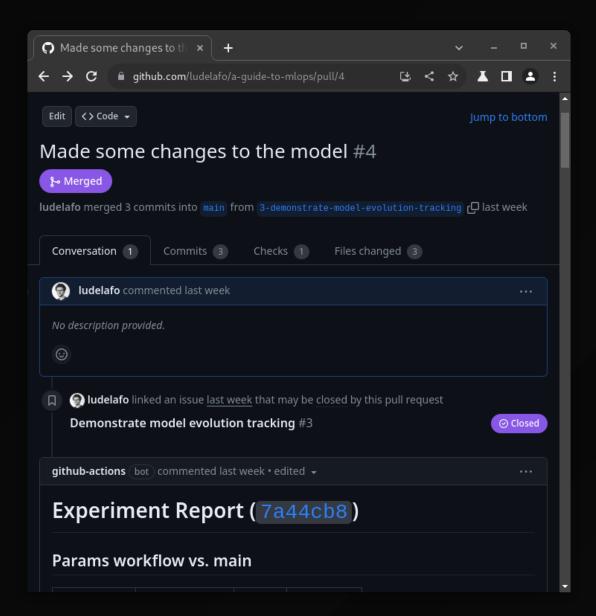
∞ Syllabus - A guide to MLO × +		~		-	•	×
↔ → C 🖬 mlops.swiss-ai-center.ch/syllabus/		☆	¥		٠	:
\equiv A guide to MLOps				٩	C	د د
Syllabus					ີລ []	
What you will learn from this guide.						
Introduction - Learn about the concept behind MLOps and the tools use	ed i	in th	is gı	uide.		
• Part 1 - Local training and model evaluation - Learn how to train a mode it using DVC.	el lo	ocal	ly ar	nd ev	aluat	e
Chapter 1 - Run a simple ML experiment with Jupyter Notebook						
Chapter 2 - Adapt and move the Jupyter Notebook to Python script	ts					
Chapter 3 - Initialize Git and DVC for local training						
Chapter 4 - Reproduce the ML experiment with DVC						
Chapter 5 - Track model evolution with DVC						
 Part 2 - Move the model to the cloud - Learn how to collaborate online uppeline and CML. 	usiı	ng G	it, a	CI/C	D	
Chapter 6 - Move the ML experiment data to the cloud						
Chapter 7 - Move the ML experiment code to the cloud						
Chapter 8 - Reproduce the ML experiment in a CI/CD pipeline						
Chapter 9 - Track model evolution in the CI/CD pipeline with CML						

😪 Made some changes to the 🗙 🕥 Made some changes to the 🗙 🕂	~	- •	×
← → C 🔒 github.com/ludelafo/a-guide-to-mlops/issues/5 🛛 🕒 <	☆ ∡		•
≡ ♀ ludelafo / a-guide-to-mlops Δ	Q	Ð	
<> Code O Issues 1 1 Pull requests 1 O Actions H Project	s 🖽 v	Wiki	
Edit New issue	Jum	p to bot	tom
Made some changes to the model #5			
⊙ Open Iudelafo opened this issue 32 minutes ago · 0 comments · May be fixe	ed by <mark>#6</mark>		
			- 1
ludelafo commented 32 minutes ago			
No description provided.			
Add a comment			
Write Preview H B I I \mathcal{O} $\frac{1}{2}$ I I	@ Ç	· ۲	
Add your comment here			
🖽 Markdown is supported 🛛 🕅 Paste, drop, or click to add files			

🚱 Made some changes to the x 🕞 🥵 Made some changes to the x 🕂 + 🗸 🗸	-	•	×
← → C 🔒 github.com/ludelafo/a-guide-to-mlops/pull/6 🛛 🗠 < ☆ 🛛			:
≡ ♀ ludelafo / a-guide-to-mlops Δ)	Î
<> Code 💿 Issues 1 1 Pull requests 1 🕟 Actions 🖽 Projects 🖽	Wiki		
Edit <> Code - Jur	np to l	bottom	
Made some changes to the model #6			
I'l Open			
ludelafo wants to merge 1 commit into main from 5-demonstrate-model-evolution-tracki	ng 🗘]	
Conversation 0 Commits 1 Checks 2 Files changed 1			
Judelafo commented 26 minutes ago			
No description provided.			
-O- 🎯 Change parameters	×	8bde18e	
 Iudelafo linked an issue <u>26 minutes ago</u> that may be <u>closed</u> by this pull request Made some changes to the model #5 	0) Open	
Add more commits by pushing to the 5-demonstrate-model-evolution-tracking branch on l	ludela	fo/a-	

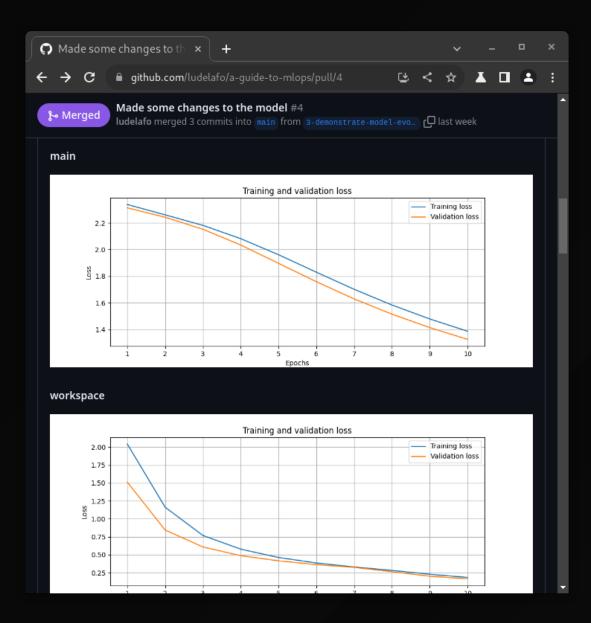
ົ	Mad	e some	e chang	es to th	×	+							~		-	•	×
÷	÷	G	🗎 gith	ub.com/lu	ıdelaf	fo/a-gu	ide-to-	-mlops	actions/	;/runs/	. 坣	<	☆	¥		۲	:
	Ma			:hange ort -	es to	o the	mo	del #	ŧ5								
		and-ro		37s							Beta	Give	feedba	ck	Ω	ŝ	
>	ø	Set up	o job														
>	ø	Check	out rep	ository												5s	
>	ø	Setup	Pythor													20s	
>	ø	Install	l deper	dencies											1r	n 30s	
>	ø	Login	to Goo	gle Cloud												Θs	
>	ø	Train I	model													53s	
>	ø	Setup	CML													31s	
>	ø	Create	e CML r	eport												11s	
>	ø	Post L	.ogin to	Google C	loud											Θs	
>	ø	Post S	Setup P	/thon												Θs	
>	ø	Post C	hecko	ut reposito	ory											0s	•

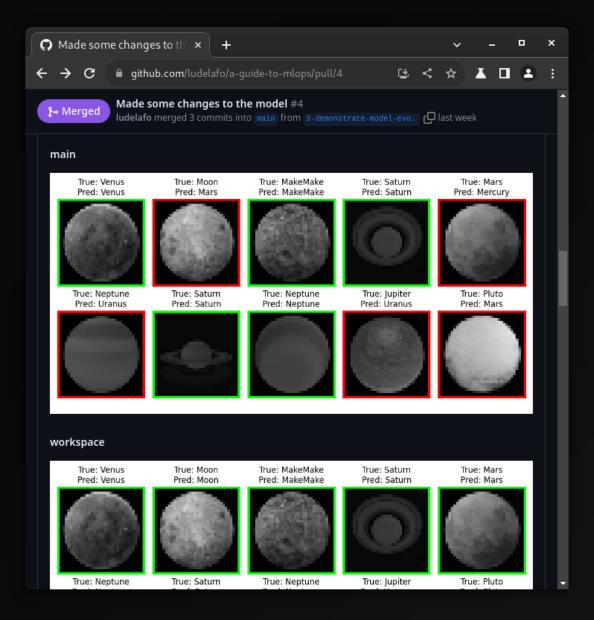
C	Mad	le some changes to th 🗙 🕂 🔶 -	•	×
←	→	C 🗎 github.com/ludelafo/a-guide-to-mlops/actions/runs/ 🔁 < 🛧 👗 🔳		:
		-and-report Beta Give feedback Carl last week in 3m 37s	ŝ	•
~		Train model	53s	
	25	Stage 'evaluate' is cached - skipping run, checking out outputs		
	26 27 28	To track the changes with git, run:		
	29 30	git add data/raw.dvc		
	31 32	To enable auto staging, run:		
	33	dvc config core.autostage true		
	34	Use `dvc push` to send your updates to remote storage.		
>	ø	Setup CML	31s	
~	0	Create CML report	11s	
	1 62 63 64 65 66	<pre> Run # Fetch all other Git branches From <u>https://github.com/ludelafo/a-guide-to-mlops</u> * [new branch] main -> main * [new branch] main -> origin/main file:///home/runner/work/a-guide-to-mlops/a-guide-to-mlops/dvc_plots/index.html <u>https://github.com/ludelafo/a-guide-to-mlops/pull/4#issuecomment-1964463218 } } </u></pre>		I
>	ø	Post Login to Google Cloud	0s	I
>	ø	Post Setup Python	0s	
>	ø	Post Checkout repository	0s	-

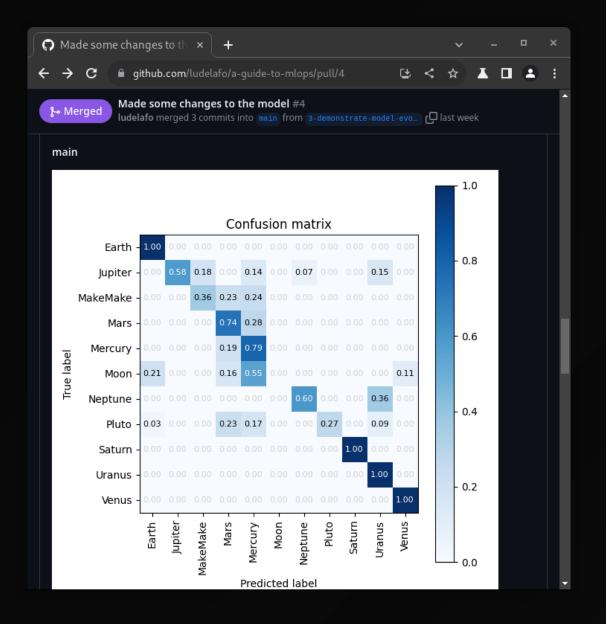


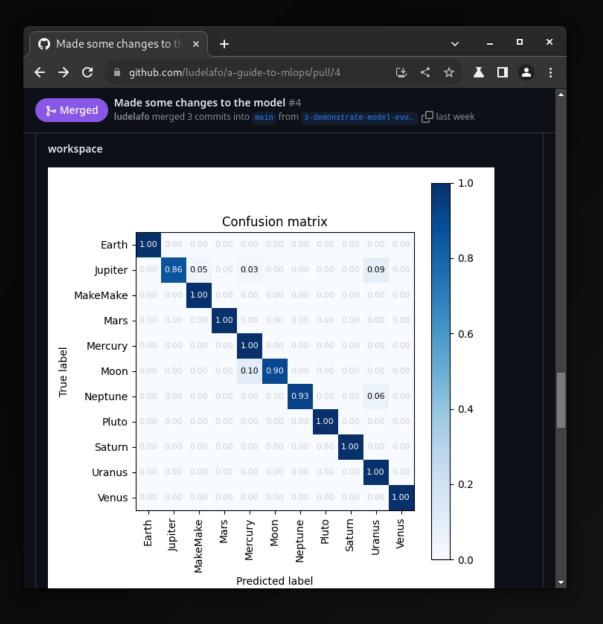
 → C ■ github.com/ludelafo/a-guide-to-mlops/pull/4 → Merged Made some changes to the model #4 Indelafo merged 3 commits into main from 3-demonstrate-model-evo □ last week github-actions bot commented last week • edited - ···· 	:
■ Merged Iudelafo merged 3 commits into main from 3-demonstrate-model-evo. □ last week github-actions (bot) commented last week • edited → ···	
Experiment Depart (70440b0)	
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	

Training History









Q Made some changes to th × +		~	-	•	×
← → C 🔒 github.com/ludelafo/a-guide-to-m	nlops/pull/4/files 🕒 <	: ☆	K 🗆		:
Changes from all commits - File filter - Conversation	ons + Jump to + ੴ +		Revie	ew 👻	
√ ‡ 30 ■■■■ dvc.lock 🖵		Uieweo	a p		
93 - size: 39706	93 + size: 37477				
∨ ∵∷ 6 ■■■■ params.yaml []		Vieweo			
6 7 train: 8 seed: 77 9 - lr: 0.0001 10 epochs: 10 11 - conv_size: 32 12 - dense_size: 64 13 output_classes: 11	6 7 train: 8 seed: 77 9 + lr: 0.001 10 epochs: 10 11 + conv_size: 64 12 + dense_size: 128 13 output_classes:				
Do not share my pe	Docs Contact Manage coo ersonal information ‡ GitHub, Inc.	kies			

Made some changes to th × +		~	-	•	×
← → C a github.com/ludelafo/a-guide-to-mlops/pull/4	<	☆ 2		٤	:
Made some changes to the model #4 ludelafo merged 3 commits into main from 3-demonstrate-model-evo_	٦	ist week			•
def1388 last week			Cor	npare	
ludelafo added 2 commits last week					
-O- 🗐 Revert GitHub Workflow				08eb1	f
👰 Add CML reporting to CI/CD pipeline			✓ 5	d1da7	
Iudelafo merged commit a199b52 into main last week 1 check passed I check passed I check passed I check passed I check passed I check passed I check passed I check passed I check passed I check passed I check passed		ew detail: st week		evert	
Add a comment					
Write Preview H B I III <> <> <> < <	Ø	@ (2 5		
Add your comment here					I
🖽 Markdown is supported 🛛 🖂 Paste, drop, or click to add files					
			Comm	ent	

C	Bur	np do	ocker	actio	ns vers	×	+								~			•	×
÷	\rightarrow	G		gith	Jb.com	/ludel	afo/a-	guide-to	o-mlop	ps/actior	ns/runs/	' (C±	<	☆	¥		٩	:
		1-rep			sh-ano ^{31s}	d-de	ploy					Beta	a) (ŝive fe	eedbao	:k	C	ŝ	^
>	•	Set	t up jo	ob														4s	
>	0	Che	eckou	ut rep	ository													2s	
>	ø	Set	up P	ython														13s	
>	•	Inst	tall d	lepen	dencies														
>	0	Log	gin to	o Goog	gle Clou	ıd												Θs	
>	ø	Trai	in me	odel													11	n 27s	
	\oslash	Set	tup C	ML														0s	
	\oslash	Cre	eate C	CML re	port													Θs	
>	0	Log	g in to	o the	Contain	ier re	gistry											2s	
>	ø	Imp	port	the Be	entoML	mod	el											2s	
>	ø	Bui	ild th	e Ben	toML 'B	ento												41s	
>	ø	Cor	ntain	erize	and pul	blish	the Be	nto									4n	n 11s	
>	ø	Get	t Goo	ogle C	oud's k	uber	netes	credent	tials									1s	
>	•	Upo	date	the K	ubernet	tes de	eployn	nent										0s	
>		Der	t vola	the m	odel on	Kub	ernete	S											-

	C I	Bum	p docker actions vers × + ~	-		•	×
÷	(→	C 🔒 github.com/ludelafo/a-guide-to-mlops/actions/runs/ ڬ < 🛠	¥		٩	:
			report-publish-and-deploy (Beta) Give feedback last week in 8m 31s	<	G	ŝ	•
	~	ø	Containerize and publish the Bento		4n	11s	
	10)24)25)26	ba473bfdf54e: Layer already exists 40c5e3152548: Layer already exists latest: digest: sha256:e908db62950d449c1df8823c1d4e536051b9eb8ceab72cb13eb117d4b2cbfb6a e	size:	345	57	
	>	ø	Get Google Cloud's Kubernetes credentials			1s	
	>	ø	Update the Kubernetes deployment			Θs	
	~	ø	Deploy the model on Kubernetes			3s	
		1 23 24	► Run kubectl apply \ deployment.apps/celestial-bodies-classifier-deployment unchanged service/celestial-bodies-classifier-service unchanged				
	>	ø	Post Log in to the Container registry			0s	
	>	ø	Post Login to Google Cloud			0s	
	>	ø	Post Setup Python			0s	
	>	ø	Post Checkout repository			0s	
		Ø	Complete job				
							-

🖃 BentoML Prediction Servi 🗙 🕂	~	-	-	•	×
← → C ③ 34.65.255.92		¥		٩	:
celestial_bodies_classifier:dev					
BentoML 1.2.2 docs passing # Join BentoML Slack Stars 6.4k Tollow BentoML					
This is a Machine Learning Service created with BentoML.					
Help					
 <u>Documentation</u>: Learn how to use BentoML. <u>Community</u>: Join the BentoML Slack community. <u>GitHub Issues</u>: Report bugs and feature requests. Tip: you can also <u>customize this README</u>. Contact BentoML Team					
Servers					
Service APIs BentoML Service API endpoints for inference.			/	\ 	
POST /predict				\sim	
Infrastructure Common infrastructure endpoints for observability.			/	~	

ſ	BentoML P	Prediction Servi × +	~	-	•	×
÷	- > C	34.65.255.92		▲ □		:
	image ¹ string(Dianot imago to analyzo				^
		Execute Clear				
					_	
	Responses	·				
	-H 'acce -H 'Con	/34.65.255.92/predict' \ apt: application/json' \ tent-Type: multipart/form-data' \ ge=@Moon_145.jpg;type=image/jpeg'		l	4	
		r-model-is-online:3000/predict				
	Server respo	nse				
	Code	Details				
	200	<pre>Response body { "prediction": "Moon", "probabilities": { "Earth": 1.3051907644945349e-12, ""inite": 0.00000000000000000000000000000000000</pre>				Ţ

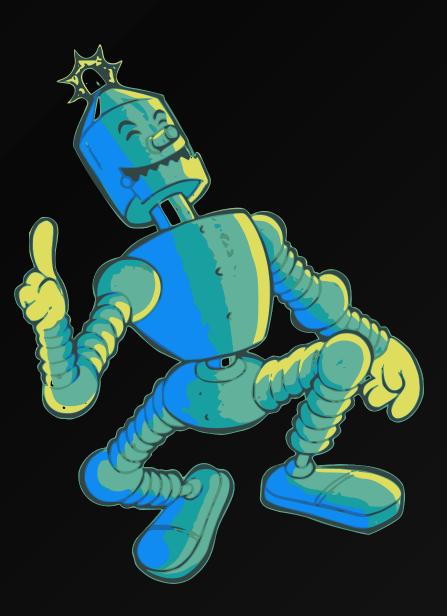
TARGET AUDIENCE

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

S 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)

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

GitHub account, Google Cloud account

ACCESS THE GUIDE

Access the guide at <u>mlops.swiss-ai-</u> <u>center.ch</u>.

Feel free to open an issue on <u>GitHub</u> 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.



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





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)

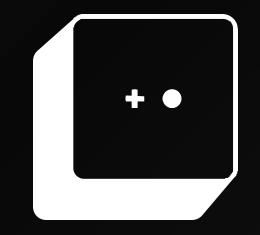
Current situation

- 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 <u>D. Sculley et. al. NIPS 2015: Hidden</u> technical debt in Machine learning systems
- Robot illustation by <u>OpenClipart-Vectors</u> on <u>Pixabay</u>