AI4 | eosc

Machine Learning Operations (MLOps): from global landscape to practice in AI4EOSC

EGI Conference 2024. October 3d, 2024. Lecce, Italy

Valentin Kozlov[1),], Lisana Berberi[1)], Borja Esteban Sanchis[1)], Giang Nguyen[2)], Judith Sainz-Pardo Diaz[3)], Amanda Calatrava[4)], Germán Molto[4)], Viet Tran[2)], Alvaro Lopez Garcia[3)]*

*valentin.kozlov@kit.edu

1) KIT 2) IISAS 3) IFCA-CSIC 4) UPV

Funded by the European Union

Outline

- Machine Learning Operations (MLOps) definition(s)
- MLOps landscape of platforms & tools
- AI4EOSC MLOps practices

Funded by
the European Union

MLOps

Wikipedia: MLOps is a paradigm that aims to **deploy** and **maintain** machine learning models in **production** reliably and efficiently.

Google: MLOps is an ML engineering **culture** and **practice** that aims at unifying ML system **development** (Dev) and ML system **operation** (Ops).

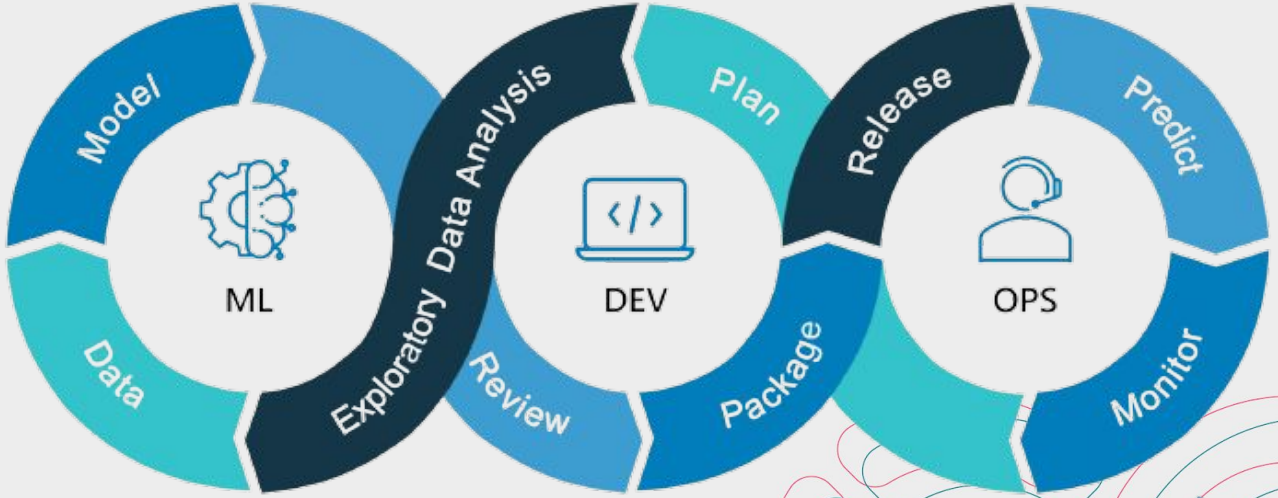
Databricks*): MLOps is the set of **processes** and **automation** for managing **data**, **code** and **models** to improve performance stability and long-term efficiency in ML systems

*) The Big Book of MLOps: Second Edition (Databricks)

Funded by
the European Union

MLOps

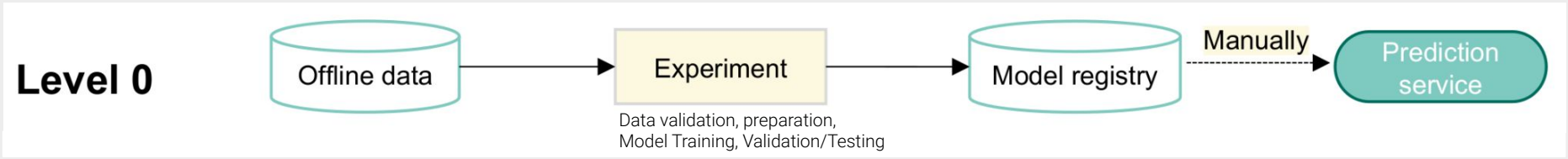
MLOps:
culture, practices, processes of **automation** for managing **data**, **development**, **models**, **operations**

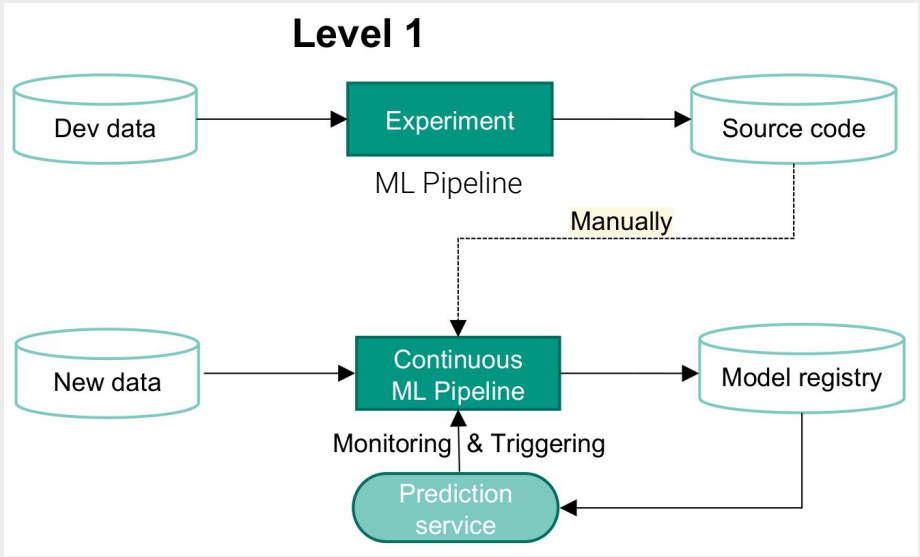


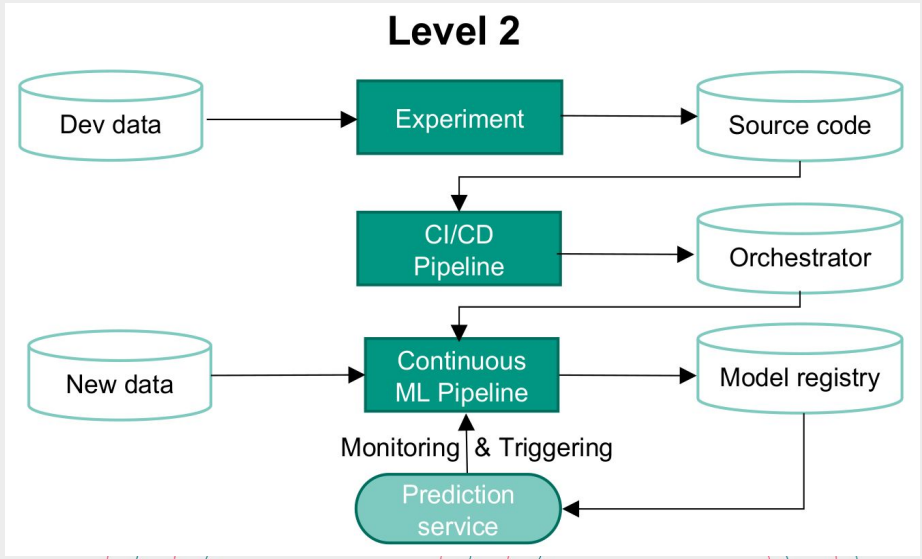
Credit: NealAnalytics

Funded by
the European Union

MLOps automation levels (Google)



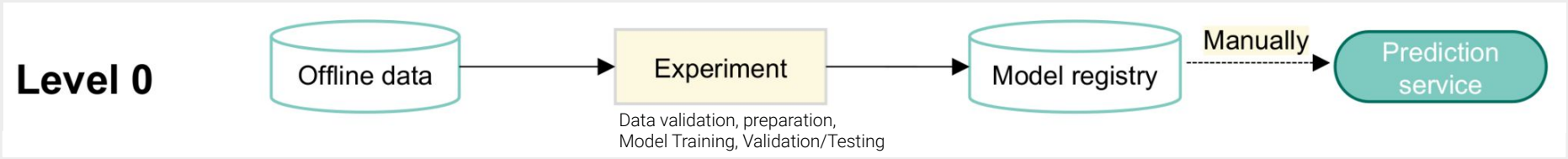


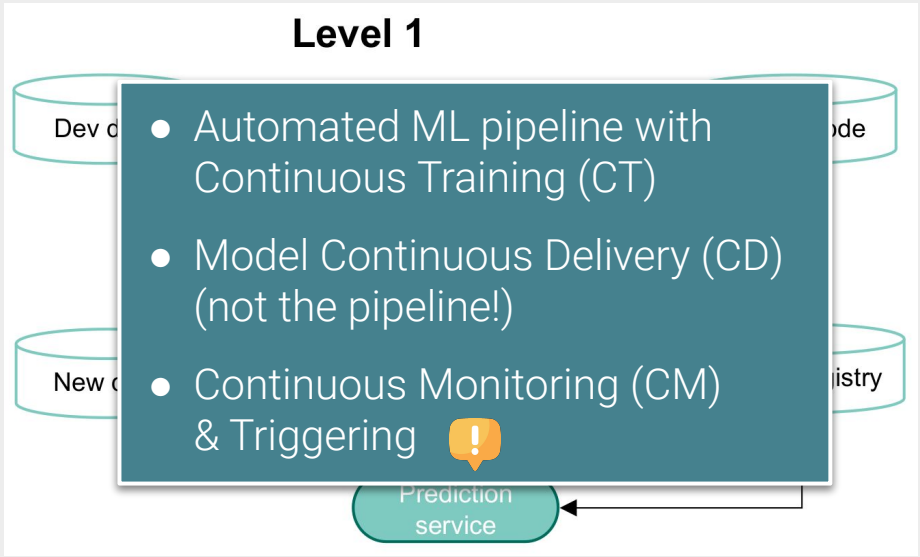


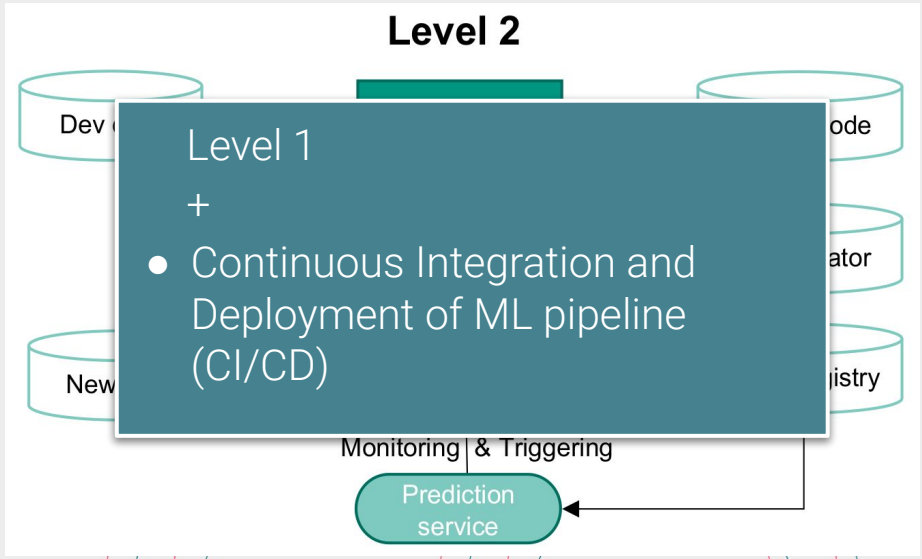
Funded by the European Union

manual
automated

MLOps automation levels (Google)



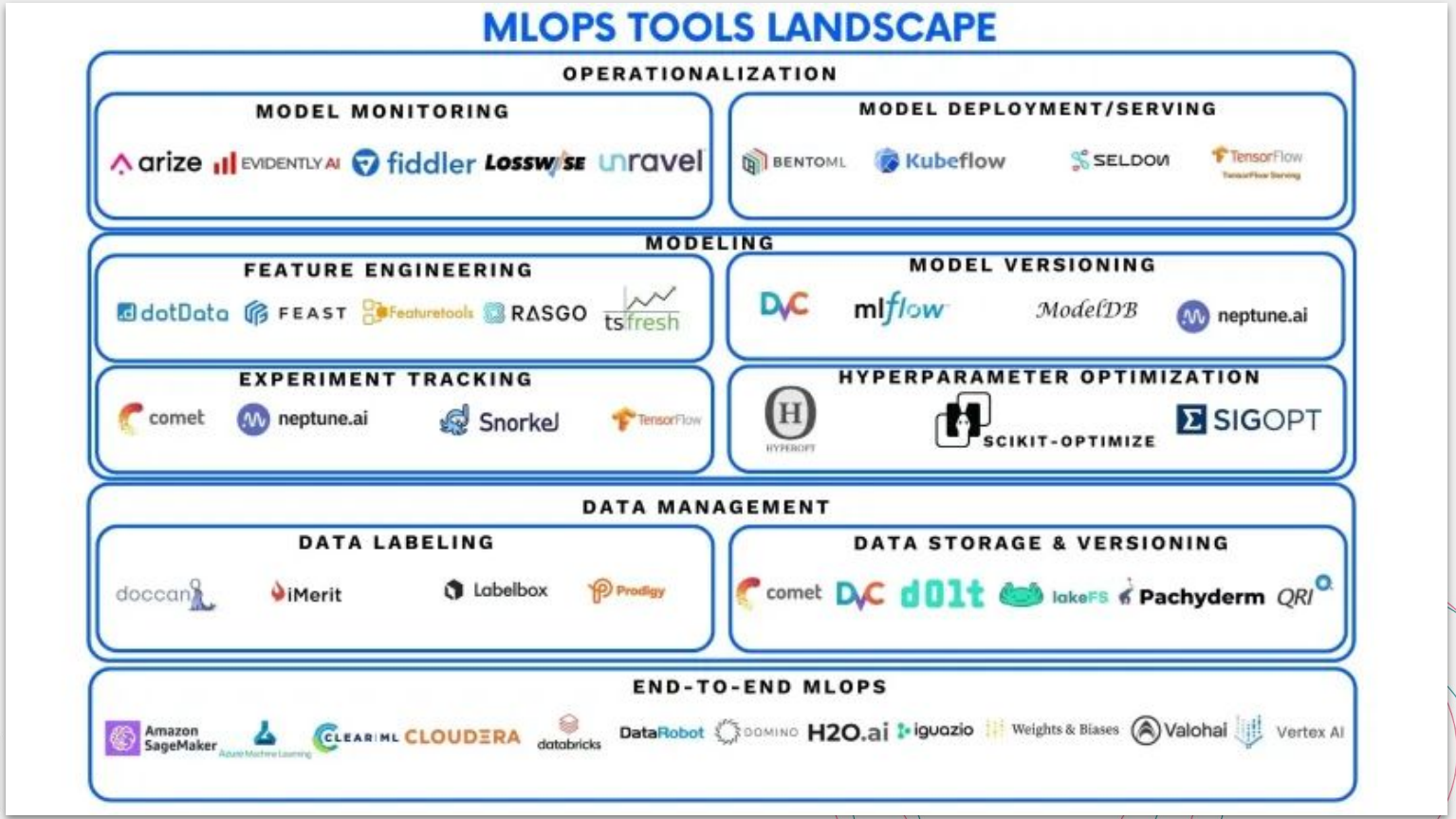




Funded by the European Union

manual
automated

MLOps platforms & tools (subset)



Credit: AIMultiple Research

MLOps Platforms

Criteria to assess MLOps Platforms*)

Only **open source** is considered

- *Orchestration* (O)
- *Distributed Training* (DT)
- *Code Management* (CM)
- *Model Development* (MDV)
- *Model Testing/Validation* (MTV)
- *Model Inference* (MI)
- *Model Deployment* (MDP)
- *Experiment Tracking and Metadata Store* (ETMS)
- *Data Versioning and Management* (DVM)
- *Model Performance Monitoring* (MPM)

Real-world implementation depends on your needs and other already implemented services

*) from L.Berberi “Machine Learning Operations Landscape: Platforms and Tools”, submitted

Funded by
the European Union

Product	**GitHub Stars**	**O** Orchestration	**DT** Distributed Training	**CM** Code Management	**MDV** Model Development	**MTV** Model Testing/-Validation	**MI** Model Inference	**MDP** Model Deployment	**ETMS** Experiment Tracking and Metadata Store	**DVM** Data Versioning and Management	**MPM** Model Performance Monitoring	**Full** Score	**Partial** Score
MLflow	17.1 K			✓			✓✓	✓✓	✓✓			30%	10%
Prefect	14.4 K	✓✓		✓						✓✓		20%	10%
Kubeflow	13.6 K	✓✓	✓✓		✓✓				✓✓			40%	0%
Dagster	10 K	✓✓						✓✓		✓✓		30%	0%
W&B (WB)	8.1 K	✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓✓	✓✓	70%	10%
MetaFlow	7.5 K	✓✓							✓✓	✓	✓	20%	20%
Mage	6.9 K						✓✓	✓	✓	✓✓	✓	20%	30%
Pachyderm	6.1 K	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓✓		60%	30%
ClearML	5.2 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Flyte	4.7 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓		80%	10%
Seldon core	4.2 K	✓✓				✓✓	✓✓	✓✓	✓		✓✓	50%	10%
ZenML	3.6 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Polyaxon	3.5 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓	90%	10%
TFX	2.1 K	✓✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓	✓✓		70%	20%
MLeap	1.5 K	✓✓						✓✓			✓✓	30%	0%
MLRun	1.2 K	✓✓	✓✓	✓	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓✓	80%	20%

from L.Berberi “Machine Learning Operations Landscape: Platforms and Tools”, submitted

Product	**GitHub Stars**	**O** Orchestration	**DT** Distributed Training	**CM** Code Management	**MDV** Model Development	**MTV** Model Testing/-Validation	**MI** Model Inference	**MDP** Model Deployment	**ETMS** Experiment Tracking and Metadata Store	**DVM** Data Versioning and Management	**MPM** Model Performance Monitoring	**Full** Score	**Partial** Score
MLflow	17.1 K			✓			✓✓	✓✓	✓✓			30%	10%
Prefect	14.4 K	✓✓		✓						✓✓		20%	10%
Kubeflow	13.6 K	✓✓	✓✓		✓✓				✓✓			40%	0%
Dagster	10 K	✓✓						✓✓		✓✓		30%	0%
W&B (WB)	8.1 K	✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓✓	✓✓	70%	10%
MetaFlow	7.5 K	✓✓							✓✓	✓	✓	20%	20%
Mage	6.9 K						✓✓	✓	✓	✓✓	✓	20%	30%
Pachyderm	6.1 K	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓✓		60%	30%
ClearML	5.2 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Flyte	4.7 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓		80%	10%
Seldon core	4.2 K	✓✓				✓✓	✓✓	✓✓	✓		✓✓	50%	10%
ZenML	3.6 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Polyaxon	3.5 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓	90%	10%
TFX	2.1 K	✓✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓	✓✓		70%	20%
MLeap	1.5 K	✓✓						✓✓			✓✓	30%	0%
MLRun	1.2 K	✓✓	✓✓	✓	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓✓	80%	20%

very limited open-source version

from L.Berberi “Machine Learning Operations Landscape: Platforms and Tools”, submitted

Product	**GitHub Stars**	**O** Orchestration	**DT** Distributed Training	**CM** Code Management	**MDV** Model Development	**MTV** Model Testing/-Validation	**MI** Model Inference	**MDP** Model Deployment	**ETMS** Experiment Tracking and Metadata Store	**DVM** Data Versioning and Management	**MPM** Model Performance Monitoring	**Full** Score	**Partial** Score
MLflow	17.1 K			✓			✓✓	✓✓	✓✓			30%	10%
Prefect	14.4 K	✓✓		✓						✓✓		20%	10%
Kubeflow	13.6 K	✓✓	✓✓		✓✓				✓✓			40%	0%
Dagster	10 K	✓✓						✓✓		✓✓		30%	0%
W&B (WB)	8.1 K	✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓✓	✓✓	70%	10%
MetaFlow	7.5 K	✓✓							✓✓	✓	✓	20%	20%
Mage	6.9 K						✓✓	✓	✓	✓✓	✓	20%	30%
Pachyderm	6.1 K	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓✓		60%	30%
ClearML	5.2 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Flyte	4.7 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓		80%	10%
Seldon core	4.2 K	✓✓				✓✓	✓✓	✓✓	✓		✓✓	50%	10%
ZenML	3.6 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Polyaxon	3.5 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓	90%	10%
TFX	2.1 K	✓✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓	✓✓		70%	20%
MLeap	1.5 K	✓✓						✓✓			✓✓	30%	0%
MLRun	1.2 K	✓✓	✓✓	✓	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓✓	80%	20%

very limited open-source version K8s bound

from L.Berberi “Machine Learning Operations Landscape: Platforms and Tools”, submitted

Product	**GitHub Stars**	**O** Orchestration	**DT** Distributed Training	**CM** Code Management	**MDV** Model Development	**MTV** Model Testing/-Validation	**MI** Model Inference	**MDP** Model Deployment	**ETMS** Experiment Tracking and Metadata Store	**DVM** Data Versioning and Management	**MPM** Model Performance Monitoring	**Full** Score	**Partial** Score
MLflow	17.1 K			✓			✓✓	✓✓	✓✓			30%	10%
Prefect	14.4 K	✓✓		✓						✓✓		20%	10%
Kubeflow	13.6 K	✓✓	✓✓		✓✓				✓✓			40%	0%
Dagster	10 K	✓✓						✓✓		✓✓		30%	0%
W&B (WB)	8.1 K	✓	✓✓	✓✓	✓✓	✓✓		✓	✓✓	✓✓	✓✓	70%	10%
MetaFlow	7.5 K	✓✓							✓✓	✓	✓	20%	20%
Mage	6.9 K						✓✓	✓	✓	✓✓	✓	20%	30%
Pachyderm	6.1 K	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓	✓✓		60%	30%
ClearML	5.2 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Flyte	4.7 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓		80%	10%
Seldon core	4.2 K	✓✓				✓✓	✓✓	✓✓	✓		✓✓	50%	10%
ZenML	3.6 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	100%	0%
Polyaxon	3.5 K	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓	90%	10%
TFX	2.1 K	✓✓	✓✓	✓✓	✓✓	✓	✓✓	✓✓	✓	✓✓		70%	20%
MLeap	1.5 K	✓✓						✓✓			✓✓	30%	0%
MLRun	1.2 K	✓✓	✓✓	✓	✓✓	✓	✓✓	✓✓	✓✓	✓✓	✓✓	80%	20%

very limited open-source version K8s bound TF-bound Builds on top of other tools

from L.Berberi “Machine Learning Operations Landscape: Platforms and Tools”, submitted

MLOps Drift tools

Criteria to assess Drift tools*)

Only **open source** is considered.

Types of drift detectors reviewed:

- *Change drift* : monitor single variables in the streaming context
- *Concept drift* : monitor the performance of the model, trying to identify shifts in the learned concept, i.e. between the data's feature values and their labels
- *Data drift* : monitor the distribution of the data features (model-agnostic)
- *Ensemble*: groups of detectors combined to draw the conclusion

Two data modes are reviewed: *streaming* and *batch*

*) from L.Berberi "Machine Learning Operations Landscape: Platforms and Tools", submitted

Funded by
the European Union

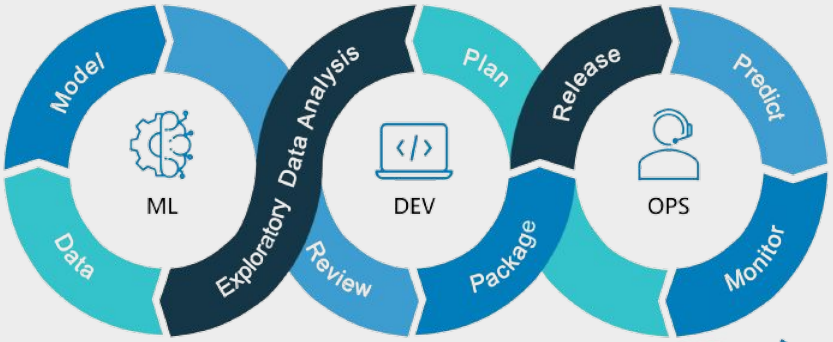
Drift Type	Drift Detector Algorithm	Abbrev. | Ref.	River		Alibi Detect		Torchdrift		Frouros		Menelaus		Evidently	
			Streaming	*Batch*	*Streaming*	*Batch*	Streaming	Batch	Streaming	Batch	Streaming	Batch	Streaming	Batch
Change Detection	Bayesian Online Change Detection	BOCD, (Adams and MacKay, 2007)	✓		✓				✓					
	Cumulative Sum Test	CUSUM, (Page, 1954)	✓						✓		✓		✓	
	Geometric Moving Average Detection	-, (Roberts, 1959)	✓		✓				✓					
	Page-Hinkley	PH, (Page, 1954)	✓						✓		✓		✓	
	ADaptive WINdowing	ADWIN, (Bifet and Gavaldà, 2007)	✓						✓		✓			
	Kolmogorov-Smirnov Windowing Detection	KSWIN, (Raab et al, 2020)	✓		✓		✓		✓					✓
Concept Drift	Drift Detection Method	DDM, (Gama et al, 2004)	✓						✓		✓			
	Early Drift Detection Method	EDDM, (Baena-Garcıa et al, 2006)	✓						✓		✓			
	EWMA Concept Drift Detection Warning	ECDDWT, (Ross et al, 2012)	✓						✓					
	Statistical Test of Equal Proportions to Detect concept drift	STEPD, (Nishida and Yamauchi, 2007)							✓		✓			
	Hoeffding's drift detection Method	HDDM, (Frías-Blanco et al, 2015)	✓						✓					
	Fast Hoeffding drift detection	FDDM, (Pesaranghader and Viktor, 2016)							✓					
	Reactive Drift detection Method	RDDM, (Barros et al, 2017)							✓					
Data Drift	Cramér-von Mises test Method	CVMTest, (Cramér, 1928)		✓		✓				✓		✓		✓
	Hellinger Distance Drift Detection Method	HIDDDM, (Hdlinger, 1909)		✓						✓		✓		✓
	Kullback-Leibler divergence Detection	KL, (Kullback and Leibler, 1951)				✓				✓	✓	✓		✓
	PCA-Based Change Detection	PCA-CD, (Cramér, 1928)	✓			✓		✓			✓			✓
	Earth Mover's Distance Detection Method	EMD, (Rubner et al, 2000)								✓				✓
	Maximum Mean Discrepancy Detection	MMD, (Gretton et al, 2012)	✓		✓		✓		✓	✓				✓
	Incremental Kolmogorov-Smirnov	IncrementalKSTest, (dos Reis et al, 2016)	✓		✓				✓					✓
Ensemble	Streaming Ensemble	-, (Maciel et al, 2015)	✓		✓				✓		✓		✓	
	Batch Ensemble	-		✓		✓						✓		✓
		Gitsource	https://github.com/online-ml/river		https://github.com/SeldonIO/alibi-detect		https://github.com/torchdrift/torchdrift		https://github.com/IFCA-Advanced-Computing/frouros		https://github.com/mitre/menelaus		https://github.com/evidentlyai/evidently	

from L.Berberi "Machine Learning Operations Landscape: Platforms and Tools", submitted

Drift Type	Drift Detector Algorithm	Abbrev. | Ref.	River		Alibi Detect		Torchdrift		Frouros		Menelaus		Evidently	
			Streaming	*Batch*	*Streaming*	*Batch*	Streaming	Batch	Streaming	Batch	Streaming	Batch	Streaming	Batch
Change Detection	Bayesian Online Change Detection	BOCD, (Adams and MacKay, 2007)	✓		✓				✓					
	Cumulative Sum Test	CUSUM, (Page, 1954)	✓						✓		✓		✓	
	Geometric Moving Average Detection	-, (Roberts, 1959)	✓		✓				✓					
	Page-Hinkley	PH, (Page, 1954)	✓						✓		✓		✓	
	ADaptive WINdowing	ADWIN, (Bifet and Gavaldà, 2007)	✓						✓		✓			
	Kolmogorov-Smirnov Windowing Detection	KSWIN, (Raab et al, 2020)	✓		✓		✓		✓					✓
Concept Drift	Drift Detection Method	DDM, (Gama et al, 2004)	✓						✓		✓			
	Early Drift Detection Method	EDDM, (Baena-Garcıa et al, 2006)	✓						✓		✓			
	EWMA Concept Drift Detection Warning	ECDDWT, (Ross et al, 2012)	✓						✓					
	Statistical Test of Equal Proportions to Detect concept drift	STEPD, (Nishida and Yamauchi, 2007)							✓		✓			
	Hoeffding's drift detection Method	HDDM, (Frías-Blanco et al, 2015)	✓						✓					
	Fast Hoeffding drift detection	FDDM, (Pesaranghader and Viktor, 2016)							✓					
	Reactive Drift detection Method	RDDM, (Barros et al, 2017)							✓					
Data Drift	Cramér-von Mises test Method	CVMTest, (Cramér, 1928)		✓		✓				✓		✓		✓
	Hellinger Distance Drift Detection Method	HIDDDM, (Hdlinger, 1909)		✓						✓		✓		✓
	Kullback-Leibler divergence Detection	KL, (Kullback and Leibler, 1951)				✓				✓	✓	✓		✓
	PCA-Based Change Detection	PCA-CD, (Cramér, 1928)	✓			✓		✓			✓			✓
	Earth Mover's Distance Detection Method	EMD, (Rubner et al, 2000)								✓				✓
	Maximum Mean Discrepancy Detection	MMD, (Gretton et al, 2012)	✓		✓		✓		✓	✓				✓
	Incremental Kolmogorov-Smirnov	IncrementalKSTest, (dos Reis et al, 2016)	✓		✓				✓					✓
Ensemble	Streaming Ensemble	-, (Maciel et al, 2015)	✓		✓				✓		✓		✓	
	Batch Ensemble	-		✓		✓						✓		✓
		Gitsource	https://github.com/online-ml/river		https://github.com/SeldonIO/alibi-detect		https://github.com/torchdrift/torchdrift		https://github.com/IFCA-Advanced-Computing/frouros		https://github.com/mitre/menelaus		https://github.com/evidentlyai/evidently	

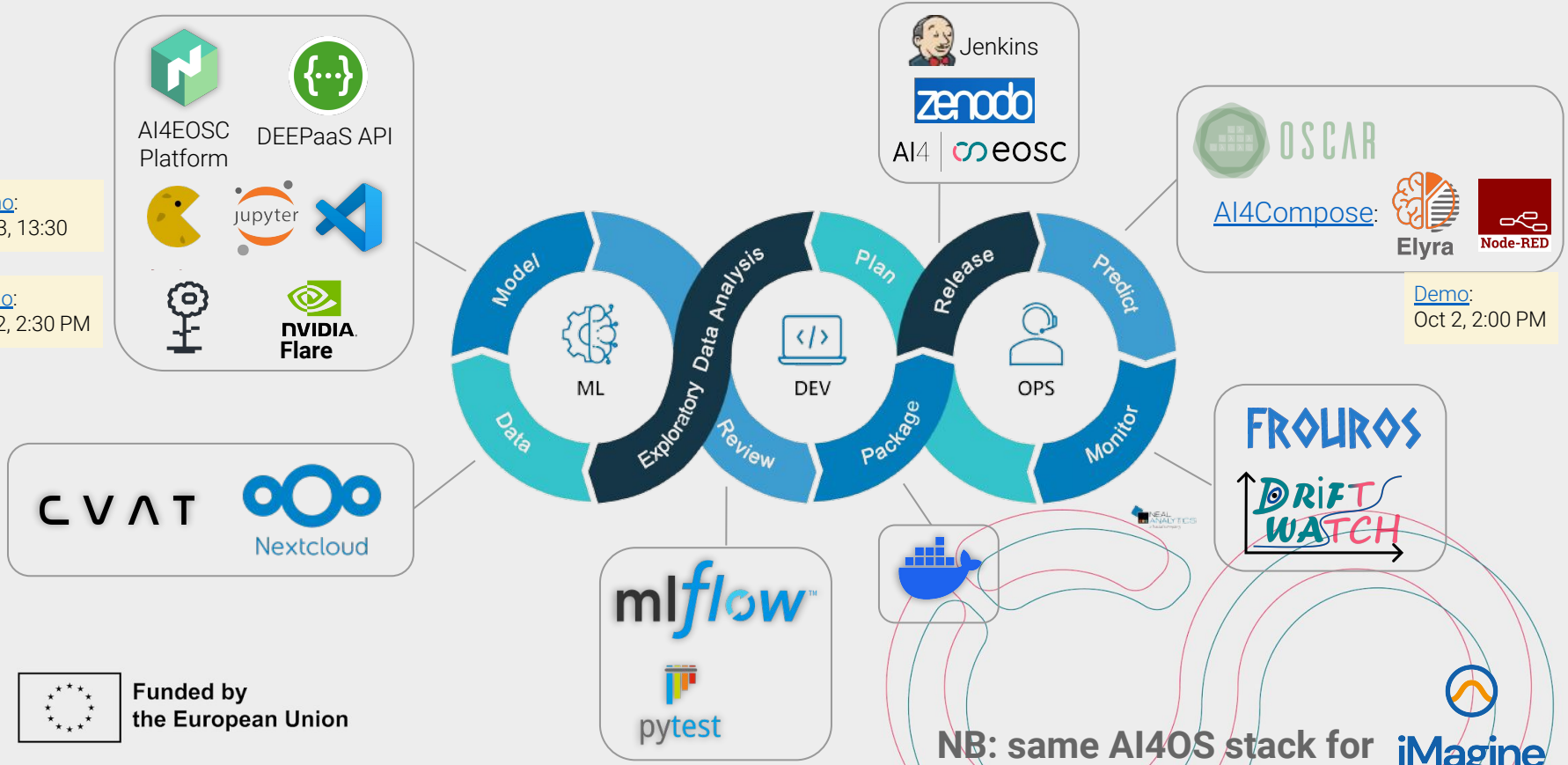
from L.Berberi "Machine Learning Operations Landscape: Platforms and Tools", submitted

AI4EOSC practices



Funded by
the European Union

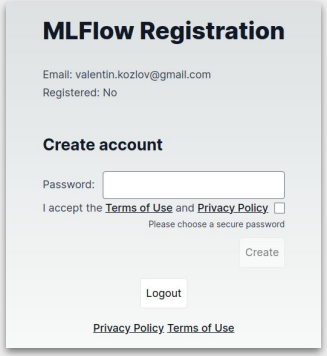
AI4EOSC practices

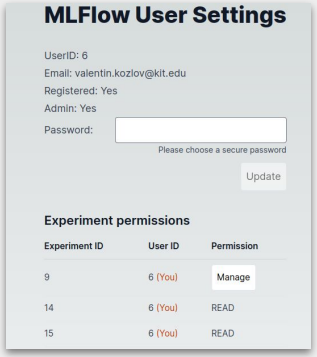


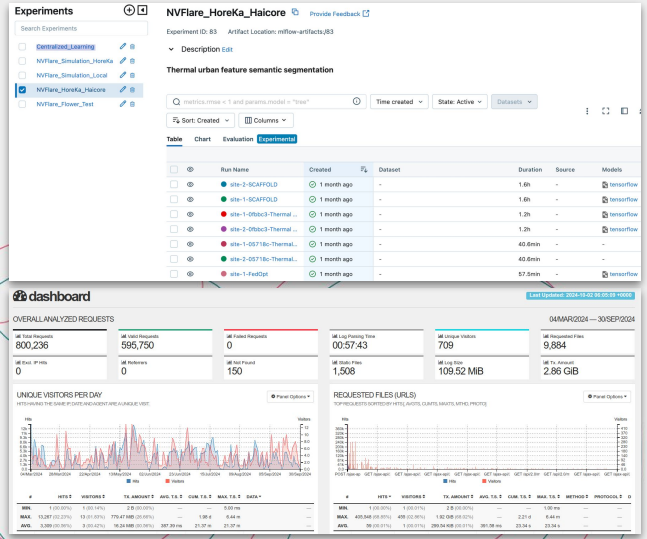
Funded by the European Union

AI4EOSC practices, MLflow

- Multi-user instances for AI4EOSC and iMagine
- Self-registrations service based on VO membership
  - Access update
  - Experiment & Model permissions
- MLflow's Support for LLMs (next talk by L.Berberi)
- Suitable for Federated Learning Experiment Tracking (after next talk by K.Alibabaei)
- Usage monitoring with Goaccess
- Regular Backups & Garbage collection







Funded by the European Union

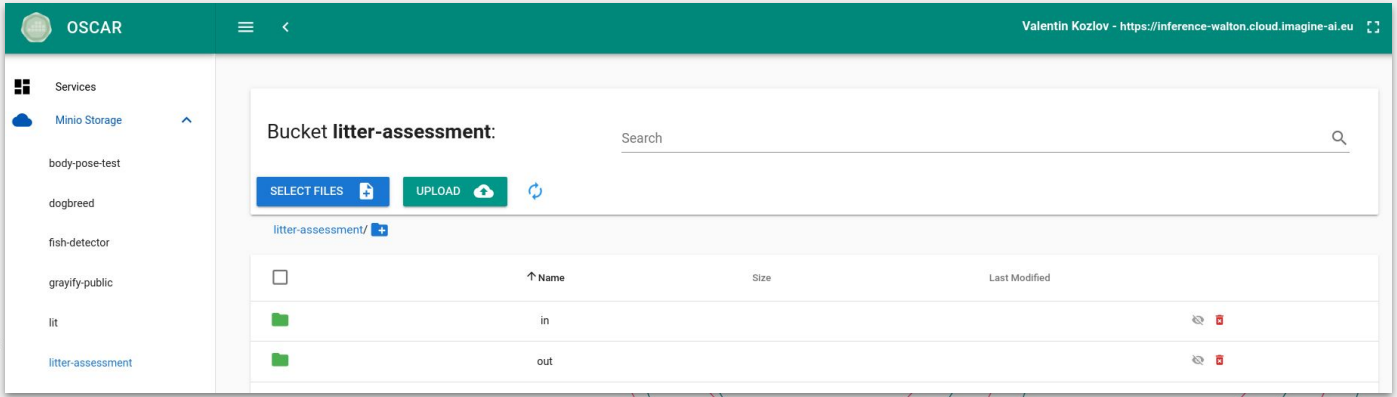
AI4EOSC practices, CI/CD

CI/CD and inference

Full project name: AI4OS-hub/litter-assessment/main

Stage View

	Declarative: Checkout SCM	Metadata tests	AI4OS Hub metadata V1 validation	AI4OS Hub metadata V2 validation (JSON)	AI4OS Hub metadata V2 validation (YAML)	License validation	Check if only metadata files have changed	User-defined module pipeline job	Docker build and delivery	Docker Variable initialization	AI4OS Hub Docker images build	AI4OS Hub Docker delivery to registry	Update OSCAR services	Declarative: Post Actions
Average stage times: (Average full run time: ~16min 56s)	2s	744ms	20s	0ms	24s	2s	1s	7min 14s	368ms	2s	5min 51s	2min 51s	14s	691ms
#20 Sep 23 15:36 — 8 commits	2s	744ms	20s		24s	2s	1s	7min 14s	368ms	2s	5min 51s	2min 51s	14s	691ms



Funded by the European Union

AI4EOSC practices, Drift monitoring

(Work-in-Progress)

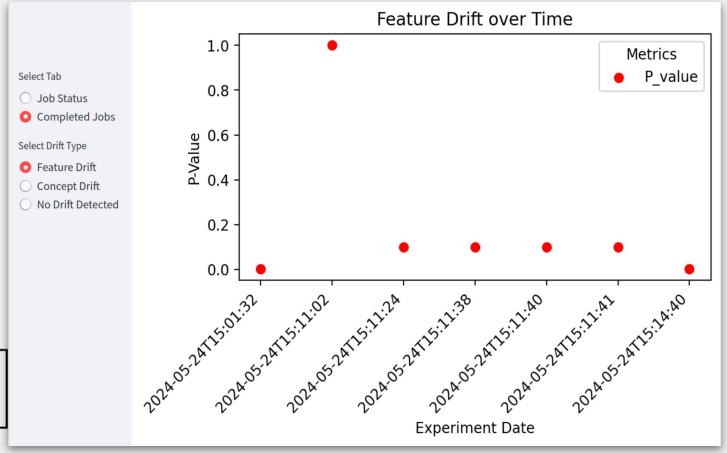
1. Release of a new dataset:
   => Before publishing, asses for the drift
   (e.g. DVC + CI/CD)

2. New data at the production service:
   => in parallel with inference, run a drift detector
   e.g. J. Sisniega et al, Fut.Gen.Comp.Sys.,161:174–188, 12 2024

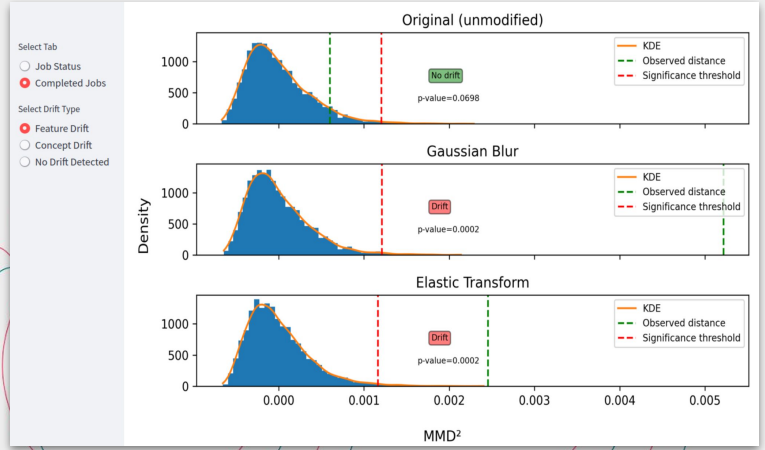
It is useful to have historical overview of drift detector runs with relevant metadata

DRIFT WATCH

Drift detection evolution



Max.Mean Descripancy (MMD) visualisation



Summary

- **MLOps** is a modern practice of **automation** for managing **data**, **development**, **models**, **operations** of AI/ML/DL-based services
- More MLOps tools and platforms become available, choose those you really need
- **AI4EOSC services** already **cover** a good part of **MLOps** processes, Level 1

Funded by
the European Union

AI4 | eosc

Co-funded by the European Union

AI4EOSC

ai4eosc-po@listas.csic.es

ai4eosc.eu

Reach us!

Thank you for your attention

Project Coordinator: Álvaro López García - aloga@ifca.unican.es

Funded by the European Union