

TU1BCO01

A workflow for training and deploying Machine learning models in EPICS

Mateusz Leputa

ICALEPCS 2023 – Cape Town, South Africa

10th of October 2023



Overview

- Motivation
- Workflow development to deployment
 - Development workspaces
 - Model and data archiving
 - Deployment and serving
- Examples
- Workflow summary
- Future development



www.isis.stfc.ac.uk



[@isisneutronmuon](https://twitter.com/isisneutronmuon)



uk.linkedin.com/showcase/isis-neutron-and-muon-source

Motivation

Developing machine learning systems for accelerator controls.

However, we have to deal with:

- Shared responsibilities.
- Frequent shelving and “reheating” of projects.
- Turnover

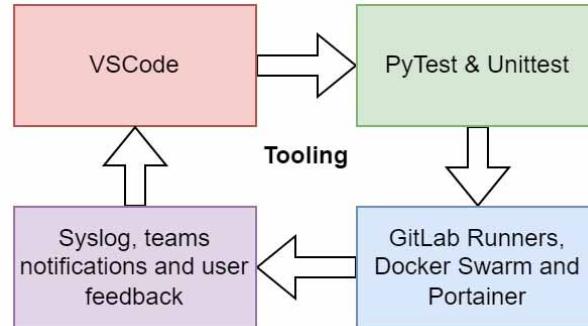
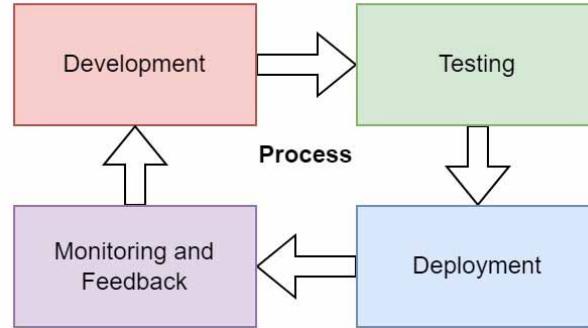
Adapt existing DevOps to machine learning operations, MLOps.

Thing we want to adopt:

- CI/CD – Continuous Integration and Continuous Deployment
- Modular Architecture with majority “off-the-shelf” components.
- Version control systems for models and data

Prevent:

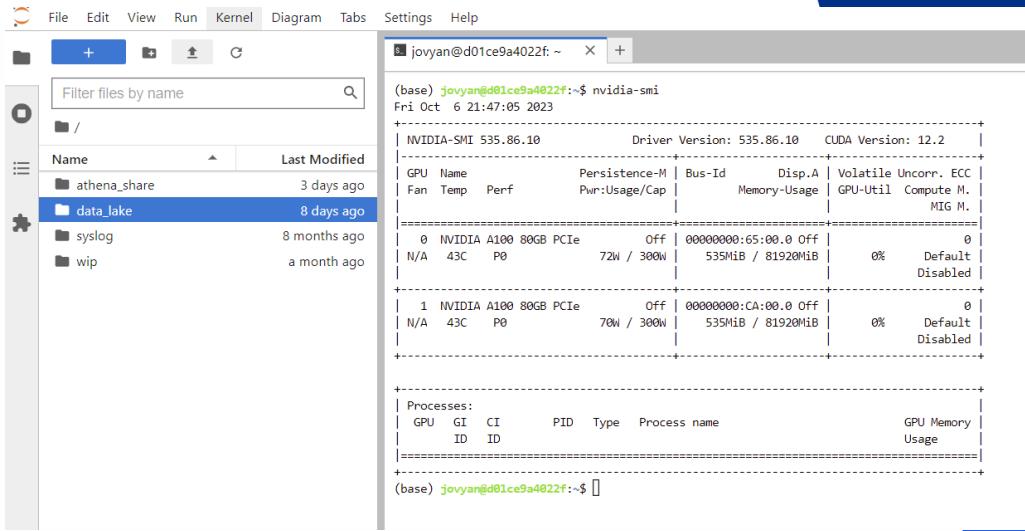
- Knowledge siloing
- Recreating the same code in every project



Remote Workspaces/Development Environment

Stack: JupyterLab and Hub

- Developers are already familiar with JupyterLab
- NFS – facilitates data transfers and collaboration spaces
- High spec servers – GPUs, high spec CPUs, RAM etc.
- 24/7 uptime – no need to leave PC on or wait for jobs to finish.
- Centralised!



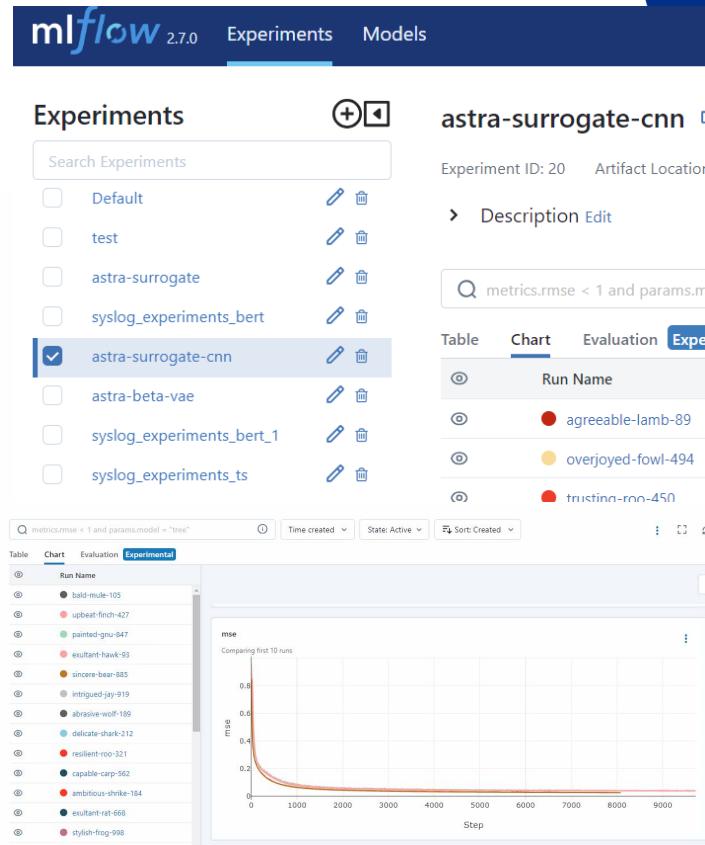
The screenshot shows a JupyterLab interface. On the left is a file browser with a search bar and a list of files: athena_share (3 days ago), data_lake (8 days ago), syslog (8 months ago), and wip (a month ago). The 'data_lake' folder is selected. On the right is a terminal window titled 'jovyan@d01ce9a4022f:~'. It displays the command 'nvidia-smi' and its output, which includes information about two NVIDIA A100 80GB PCIe GPUs. The output table shows details like Persistence-M, Bus-Id, Disp.A, and Memory-Usage. Below the GPU table is a section for Processes, showing a table with columns for GPU ID, GI ID, CI ID, PID, Type, and Process name. At the bottom of the terminal window, it says '(base) jovyan@d01ce9a4022f:~\$ []'



Model and Data Archiving

Stack: MLflow, MINIO, PostgreSQL

- Comes with a web GUI.
- Saves experiment setup, performance metrics, datasets and model artifacts.
- Provides an API to programmatically upload and download models, query experiment results and charts etc.
- Comes with its own model serving utilities.
- Can tag models e.g. experimental or production!



www.isis.stfc.ac.uk



@isisneutronmuon



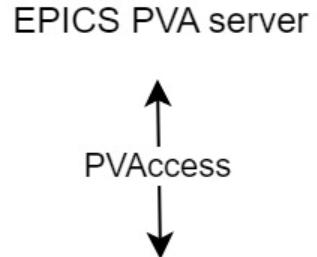
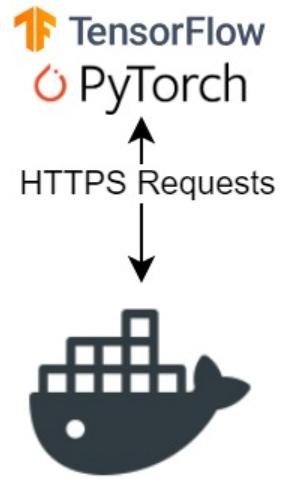
uk.linkedin.com/showcase/isis-neutron-and-muon-source

Model Serving and Deployment

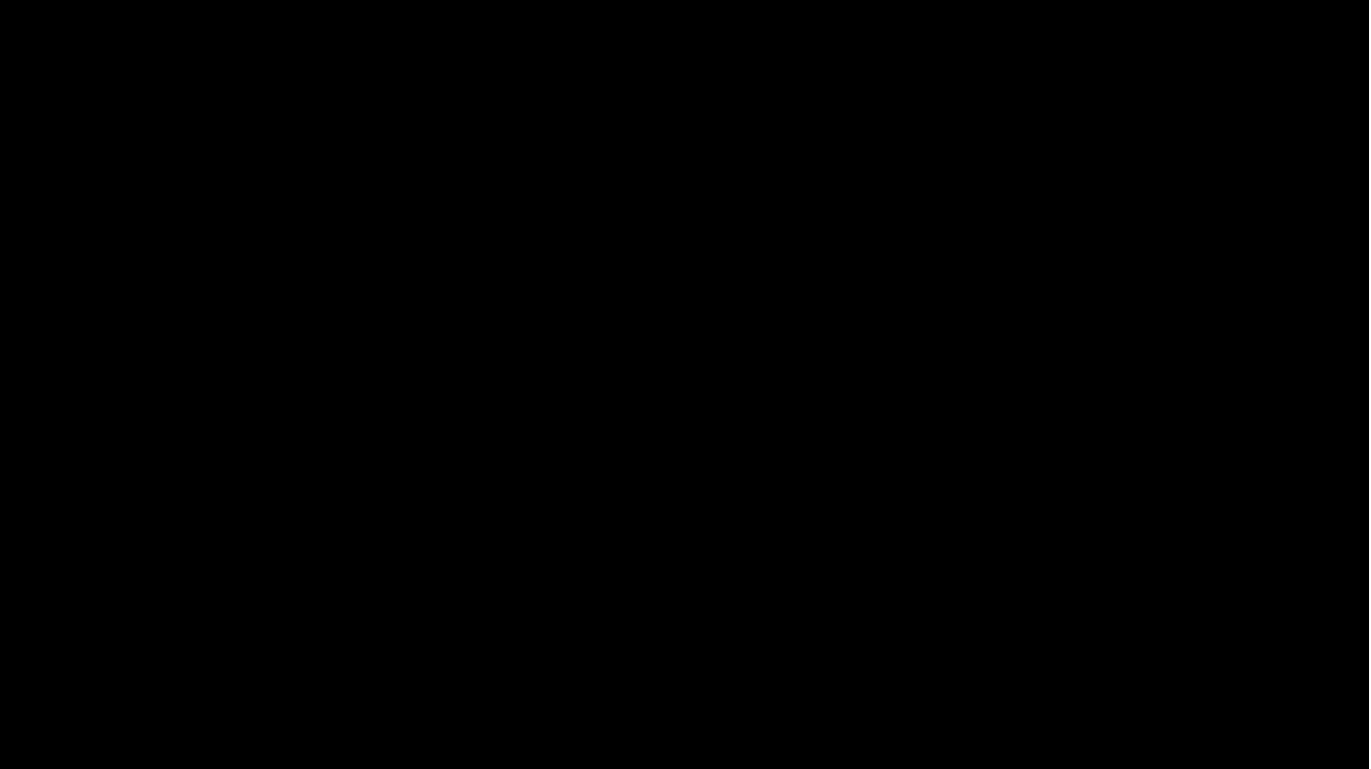
Stack: TF-Serve, TorchServe & p4p python library

- Model deployment is facilitated through serving frameworks such as TF-Serve or TorchServe. Deployed as containers
- Both serve models as HTTPS endpoints and can run inference remotely on GPUs
- Latency of 16-40 ms for small models (mostly attributed to network latency)
- HTTPS to EPICS PVA server deployed as a service.

Frontend application can then connect to the EPICS PVs and interact with the model ...



Example 1 – ASTRA Surrogate – ISIS MEBT



www.isis.stfc.ac.uk

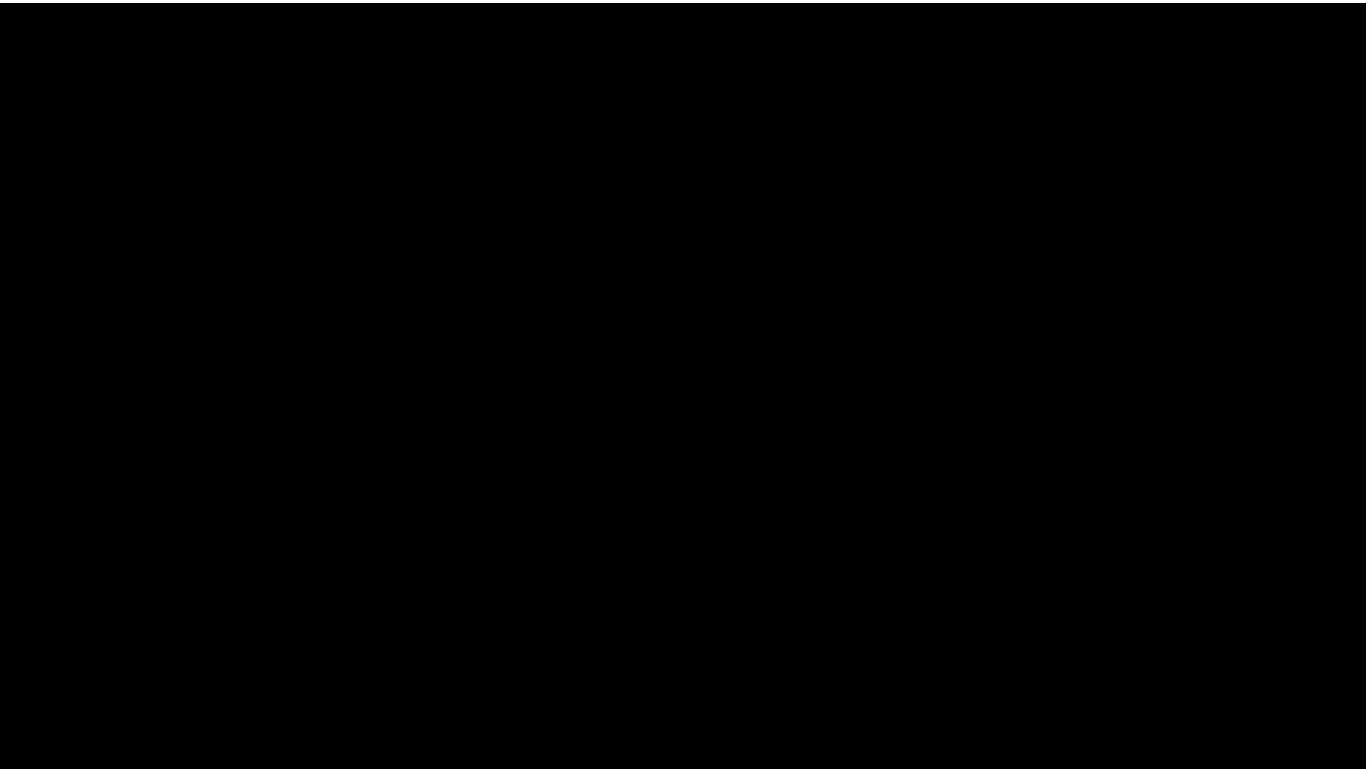


[@isisneutronmuon](#)

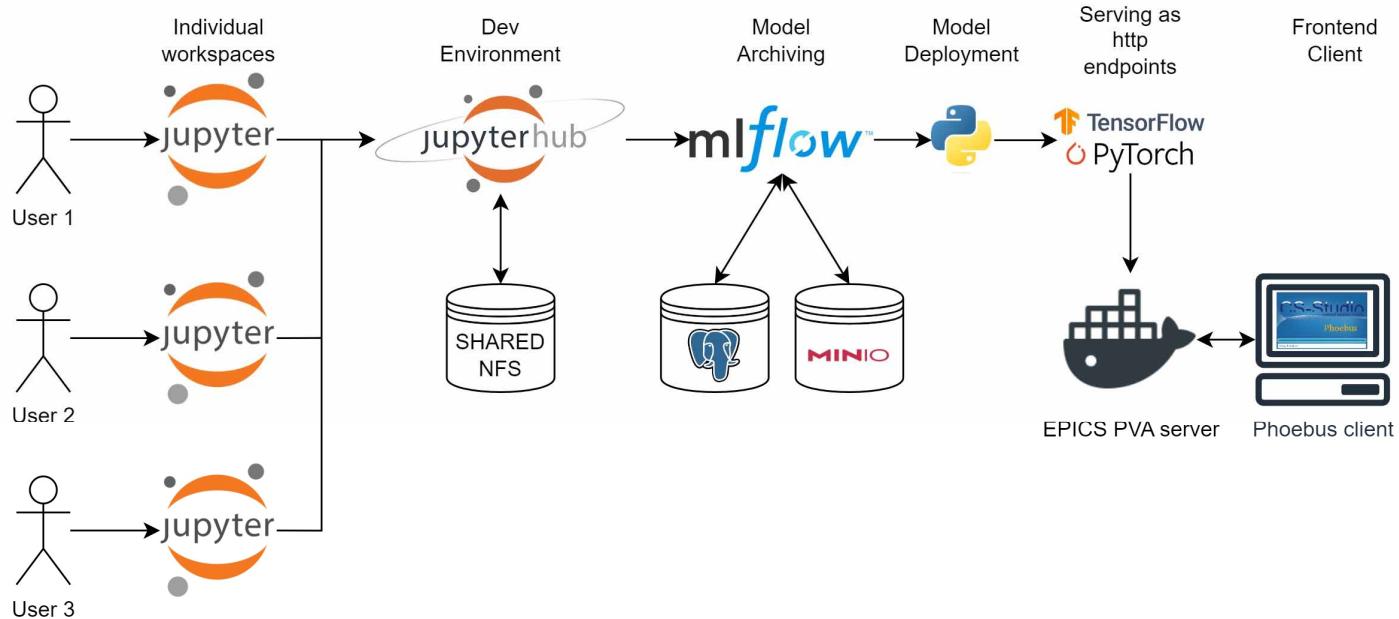


uk.linkedin.com/showcase/isis-neutron-and-muon-source

Example 2 - LEBT



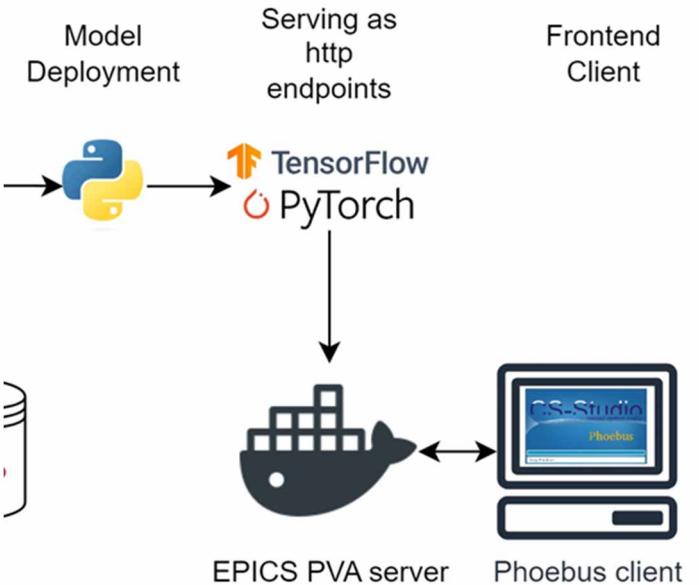
Workflow Architecture - Summary



Further Developments

- Standardise model HTTPS to EPICS serving.
- Model monitoring and retraining for continual learning.
- Model profiling.

Mostly concerning the later parts of the MLOps workflow



www.isis.stfc.ac.uk



[@isisneutronmuon](https://twitter.com/isisneutronmuon)



uk.linkedin.com/showcase/isis-neutron-and-muon-source

Questions?

mateusz.leputa@stfc.ac.uk



www.isis.stfc.ac.uk



[@isisneutronmuon](https://twitter.com/isisneutronmuon)



uk.linkedin.com/showcase/isis-neutron-and-muon-source