Training on Analysis Pipelines
Containerization + CI/CD

26th Feb - 1st Mar, 2024

Instructors (in the recordings):	Mentors:	Local organising committee:
• Podman (Docker): Michel Hernández Villanueva • Apptainer (Singularity): Marco Mambelli • GitHub CI/CD: Andrés Ríos Tascón • GitLab CI/CD: Guillermo Fidalgo	• Roy Cruz Candelaria • Lera Lukashenko • Marco Mambelli • Jim Pivarski • Richa Sharma • Michel Hernández Villanueva • Alexander Moreno Briceño	• Lera Lukashenko • Jim Pivarski • Holly Szumila-Vance • Alexander Moreno Briceño

If you aren't recording this on Zoom, enable captioning and start the recording ...

(just a reminder)

Everyone is Welcome

- You are physicists working in international collaborations. All of you should know this page:
  - The CERN code of conduct
- Built on a set of core CERN values →
- Taken together, provide the basis for respect: respect for others, respect for the organization and respect for its mission.

- We encourage a culture of openness where all contributors feel free to engage in the discussion.

What is an Analysis Pipeline?

- **A data analysis pipeline**
  - After an exploratory phase, computations that were found to be useful are formalized as reusable programs that convert input data into final results, and these programs are run over and over, with updates, as new corrections and considerations for come to mind.

- **It is very important for a data analysis pipeline to be reproducible**.
  - You want to draw conclusions about your data by running the pipeline under different conditions and seeing how the results change, but they would not be valid conclusions if running it under the same conditions also yields different results!
  - A clean workbench is an essential part of the scientific method, and your data analysis code is part of your scientific workbench.

What is an Analysis Pipeline?

- **Scientific results need to be reproducible after your experiment is done**. Ensuring reproducibility during your analysis simplifies the process of preserving your analysis for future research.

- Reproducibility is a concern for software developers as well, and many of the tools that have been developed for the software industry can be applied to data analysis.

- This training event is for data analysts who want to learn how to make their analysis pipelines robust using continuous testing (CI/CD) and containerization (Podman, Docker, and Apptainer).

What are we learning this week?

- We will take a quick tour learning the basic functionality of tools popular in analysis preservation and reproducibility.

  - **Containerization technologies**
    Podman (Docker)
    Apptainer (Singularity)

  - **Continuous Integration/Deployment (CI/CD)**
    GitLab pipes
    GitHub actions

Monday

Welcome

Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

Help with Setup
[16:10-17:00 CET]

Tuesday to Thursday

Work on your own, when you want

Watch and work through tutorials:
Indico Agenda

Friday

Hands-on sessions

Block 1:
[10-12 CET]

Block 2:
[13-15 CET]

Block 3:
[17-19 CET]

Block 4:
[21-23 CET]

Monday

Welcome

Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

Help with Setup
[16:10-17:00 CET]

Get on the same page with logistics and debug **initial setups/installations.**

Tuesday to Thursday

Work on your own, when you want

Watch and work through tutorials:
Indico Agenda

Friday

Hands-on sessions

Block 1:
[10-12 CET]

Block 2:
[13-15 CET]

Block 3:
[17-19 CET]

Block 4:
[21-23 CET]

Monday
Welcome

Tuesday to Thursday
Work on your own, when you want

Friday
Hands-on sessions

Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

Help with Setup
[16:10-17:00 CET]

Watch and work through tutorials:
Indico Agenda

Work through all of the content here and learn/work at your own pace with **our virtual support on Slack**.

Channel: #analysis-pipelines

Block 1:
[10-12 CET]

Block 2:
[13-15 CET]

Block 3:
[17-19 CET]

Block 4:
[21-23 CET]

Monday

Welcome

Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

Help with Setup
[16:10-17:00 CET]

Tuesday to Thursday

Work on your own, when you want

Watch and work through tutorials:
Indico Agenda

Sign up for mentoring sessions

Deadline: Wed. 4 pm (CERN), 10 am (ET), 11pm (Peking) We will assign you to one of the sessions afterward

Join the room indicated for your specific hands-on session.

Friday

Hands-on sessions

Block 1:
[10-12 CET]

Block 2:
[13-15 CET]

Block 3:
[17-19 CET]

Block 4:
[21-23 CET]

If you haven't done yet...

[1] Join the Slack channel: invite
If you have troubles to join, let us know now.

[2] Follow the setup pages
- Podman (Docker) (setup)
- Apptainer (Singularity) (setup)
- CI/CD with Github (here) or GitLab (here).

[3] Take a look at the Analysis example with CMS open data.

[4] Sign up for mentoring sessions
Deadline: **Wed. 4 pm (CERN), 10 am (ET), 11pm (Peking)**
We will assign you to one of the sessions on Wednesday afternoon (ET)

Meet your Instructors!

Guillermo Fidalgo Rodríguez

Masters Student in Physics
University of Puerto Rico - Mayagüez

My research:
Analysis on Emerging Jets.
Machine Learning for Tracker DQM

My expertise is:
Using python for ML Studies and python training.

A problem I'm grappling with:
My Thesis

I've got my eyes on:
Pursuing a PhD.

I want to know more about:
C++, Arduino

Marco Mambelli

Software Developer at Fermilab in Batavia, IL

My research:
I work on workflows and distributed computing system.
In particular the GlideinWMS and HEPCloud projects that are used to run all analyses and simulations for CMS, and most Fermilab experiments.

My expertise is:
Distributed scientific computing, coding and system engineering

A problem I'm grappling with:
Efficient use of GPUs on supercomputers

I've got my eyes on:
New tools to ease collaboration

I want to know more about:
Quantum computing

Fermilab

Michel Villanueva
(he/him)

Research Staff, BNL
Working in tau lepton physics
and Distributed Computing at Belle II

My research: *Precision measurements with tau leptons*

My expertise is: Data analysis in distributed computing environments.

A software and computing problem I'm grappling with: Scalability of the Belle II analysis workflow in the high-luminosity scenario.

I've got my eyes on: Sustainable operation of the grid. Training newcomers and get fresh ideas!

I want to know more about: Machine learning pipelines.

Belle II

Brookhaven National Laboratory

Andres Rios-Tascon

Research Software Engineer, Princeton University

My research: *Innovative algorithms for High-Energy Physics*

My expertise is:
Designing fast algorithms to solve difficult problems

A software and computing problem I'm grappling with:
Learning how to use the CMSSW framework

I've got my eyes on:
Automation tools to ease development workflows

I want to know more about:
Rust and FPGAs

Richa Sharma

Postdoctoral Research Associate
University of Puerto Rico - Mayagüez

My research:
Search for Dark Matter with Emerging Jets

My expertise is:
Data analysis to search for new physics using C++ and Python

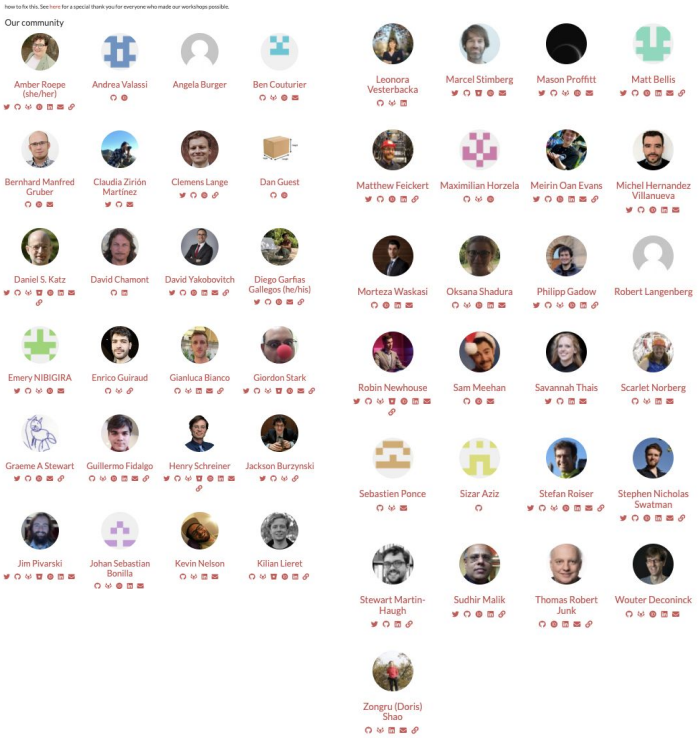
A problem I'm grappling with:
Using Machine Learning to develop tools for tracker data quality monitoring

I've got my eyes on:
Training a universal domain adaptation algorithm to classify CMS data

I want to know more about:
GPU programming

HSF Educator/Mentor community

https://hepsoftwarefoundation.org/training/educators.html



- **Join our hackathons**
  - **In 1 or more topics**
- **Join our community**
- **Become training Educator**
  - **a mentor, facilitator, instructor**
- **Open a PR for any of our modules**

Group picture!