HSF/IRIS-HEP Training on Analysis Reproducibility

24th - 28th Mar, 2025

Instructors (in the recordings):

- Podman (Docker): Michel Hernández Villanueva
- Apptainer (Singularity): Marco Mambelli
- GitHub CI/CD: Andrés Ríos Tascón
- GitLab CI/CD: Guillermo Fidalgo
- REANA: Tibor Simko

Mentors:

- Giordon Stark
- Lera Lukashenko
- Marco Mambelli
- Roy Cruz Candelaria
- Richa Sharma
- Michel Hernández Villanueva
- Alexander Moreno Briceño
- Tibor Simko
- Anil Panta
- Emery Nibigira
- Mateo Elisondo
- Callum McCracken
- Leonid Didukh
- Alp Tuna
- Tetiana Mazurets

Local organising committee:

- Lera Lukashenko
- Michel Hernández Villanueva
- Richa Sharma
- 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 Analysis Reproducibility?

- **The value of data analysis workflows**
  - Typically after an exploratory phase, computations that were found to be useful are formalized as **reusable** programs (workflows) that convert input data into results
  - 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 workflow to be reproducible**
  - You want to draw conclusions about your data by running the analysis under different conditions, 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 Analysis Reproducibility?

- **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 workflows robust using
  - Continuous testing (CI/CD)
  - Containerization (Podman, Docker, and Apptainer)
  - REANA: reproducible analysis platform

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
  - **REANA**

Monday

Welcome

Kickoff/Orientation
[14:00 CERN time]

Analysis Preservation@CMS
[14:10 CERN time]

Analysis Preservation with REANA
[14:40 CERN time]

Help with Setup
[15:10-16:00 CERN time]

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 CERN time]

Block 2:
[13-15 CERN time]

Block 3:
[17-19 CERN time]

Block 4:
[21-23 CERN time]

Monday

Welcome

Kickoff/Orientation [14:00 CERN time]

Analysis Preservation@CMS [14:10 CERN time]

Analysis Preservation with REANA [14:40 CERN time]

Help with Setup [15:10-16:00 CERN time]

Tuesday to Thursday

Work on your own, when you want

Watch and work through tutorials: Indico Agenda

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

Friday

Hands-on sessions

Block 1: [10-12 CERN time]

Block 2: [13-15 CERN time]

Block 3: [17-19 CERN time]

Block 4: [21-23 CERN time]

Monday
Welcome

Tuesday to Thursday
Work on your own, when you want

Friday
Hands-on sessions

Kickoff/Orientation
[14:00 CERN time]

Analysis Preservation@CMS
[14:10 CERN time]

Analysis Preservation with REANA
[14:40 CERN time]

Help with Setup
[15:10-16:00 CERN time]

Watch and work through tutorials:
Indico Agenda

Block 1:
[10-12 CERN time]

Block 2:
[13-15 CERN time]

Block 3:
[17-19 CERN time]

Block 4:
[21-23 CERN time]

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

Channel: #analysis-reproducibility2025

Monday

Welcome

Kickoff/Orientation
[14:00 CERN time]

Analysis Preservation@CMS
[14:10 CERN time]

Analysis Preservation with REANA
[14:40 CERN time]

Help with Setup
[15:10-16:00 CERN time]

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), 11 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 CERN time]

Block 2:
[13-15 CERN time]

Block 3:
[17-19 CERN time]

Block 4:
[21-23 CERN time]

If you haven't done yet...

[1] Join the Slack channel: #analysis-reproducibility2025
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)
- REANA (setup)

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

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

Meet the HSF mentors!

Guillermo Fidalgo Rodríguez

PhD Student in Physics
University of Alabama

My research:
Quantum Entanglement with tops

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

A problem I'm grappling with:

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

I want to know more about:
Automation
Quantum Machine Learning

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

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 and analysis tools for High-Energy Physics*

My expertise is:
Designing fast algorithms to solve difficult problems

A software and computing problem I'm grappling with:
Learning the depths of the CMSSW framework

I've got my eyes on:
AI 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:
Improving simulation models for cross-talk in pixel detectors

I want to know more about:
GPU programming

Callum McCracken

PhD Student in Physics
TRIUMF + University of British Columbia

My research:
Precision Higgs measurements, detector upgrades

My expertise is:
Python/C++ for physics analysis

A problem I'm grappling with:
"It'll be fine, we'll fix it in software"

I've got my eyes on:
You. No funny business 👀

I want to know more about:
Firmware

Tibor Simko

Department of Information Technology
CERN

My research:
Analysis preservation and reuse, containerised computational workflows, open and reproducible science

My expertise is:
Software architecture and development, computational workflows, containers and cloud computing

A problem I'm grappling with:
Running 1k concurrent ATLAS pMSSM analysis workflows as fast as possible

I've got my eyes on:
Encouraging "preproducibility" from early analysis phase to facilitate future reuse

I want to know more about:
Variety of analysis techniques and user experience; what can IT do together with researchers to facilitate reproducible science

Emery Nibigira

Research Associate,
University of Tennessee, Knoxville

My research: *Search for Physics Beyond the Standard Model and Silicon Tracking Detectors*

My expertise is: Data analysis and interpretation, silicon detectors, CI/CD, containers (Docker/Podman).

A software and computing problem I'm grappling with: Columnar Analysis.

I've got my eyes on: Next-generation particle collider!

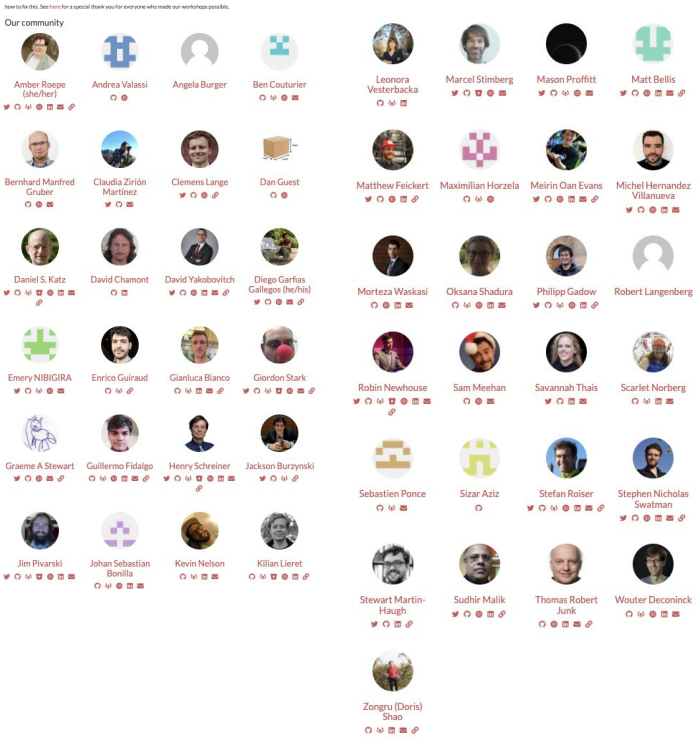
I want to know more about: AI.

CMS Compact Muon Solenoid

THE UNIVERSITY OF TENNESSEE KNOXVILLE

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!