Training on analysis reproducibility in the HSF

Michel Hernandez Villanueva (BNL), and many more!

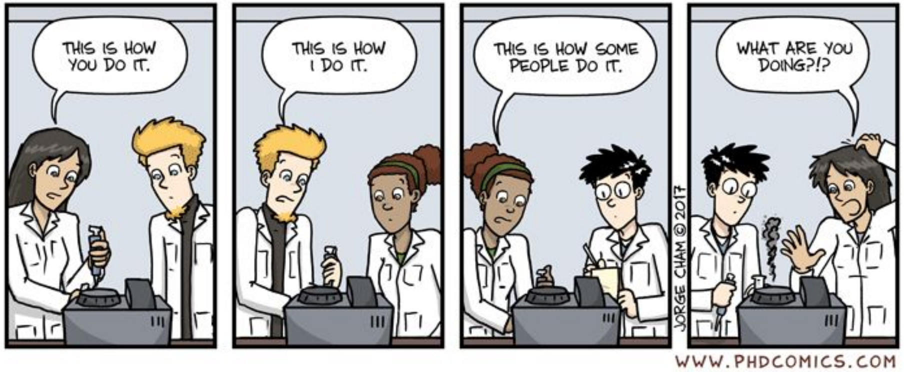
On behalf of the HSF Training WG

Workshop on ATLAS Computing and Software Activities at BNL
Mar 18, 2024

HEP and Nuclear Physics Software

As a key for a successful scientific program

- O(10k) HEP/NP people worldwide need to be trained in software engineering & computing
- Common challenges faced:
  - Most people developing code have non-permanent positions with contracts of 2 - 4 years
  - Material for training is a moving target as technology evolves (e.g., ML, GPUs, FPGAs, …)
  - Training activities not in the highest priority when making career steps
- **This should be a community effort!**



HSF Software Training

Organization

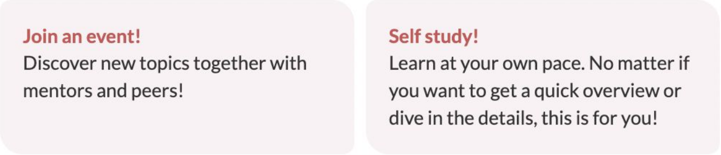
- The role of the HSF is to **coordinate common efforts** in software and computing across HEP in general

- HSF Training was established in 2018

- Develops material for an introductory software curriculum

- Focuses on **common software material across HEP and NP**
  - From basic core software skills, to advanced training required in software and computing
  - Experiment agnostic, minimal dependencies on having an account at X site

- Engages with different experimental collaborations and initiatives
  - **IRIS-HEP**, FIRST-HEP, and The Carpentries



Join an event!
Discover new topics together with mentors and peers!

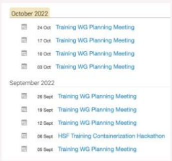
Self study!
Learn at your own pace. No matter if you want to get a quick overview or dive in the details, this is for you!

https://hepsoftwarefoundation.org/training

The HSF Software Training Platform

We can cover more ground together!

Weekly meetings



October 2022

- 24 Oct Training WG Planning Meeting
- 17 Oct Training WG Planning Meeting
- 10 Oct Training WG Planning Meeting
- 03 Oct Training WG Planning Meeting

September 2022

- 26 Sept Training WG Planning Meeting
- 19 Sept Training WG Planning Meeting
- 12 Sept Training WG Planning Meeting
- 06 Sept HSF Training Containerization Hackathon
- 05 Sept Training WG Planning Meeting

Monthly Hackathons

✨ Containerization Training Hackathon ✨

🚀 The big goal!

Training in software and computing is an essential ingredient for the success of any HEP experiment. As most experiments have similar basic prerequisites we want to join our efforts and create *one* introductory software training curriculum that serves HEP newcomers from all experiments. This curriculum is made up of independent training *modules* and should contain all software skills needed as they enter the field while instilling best practices for writing sustainable software.

We have started this work here and have completed and tested several of our modules to great success.

Platforms

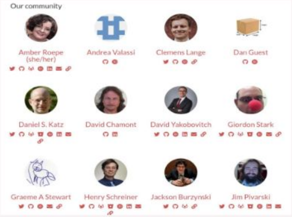
@Kilian Lieret, what all repos Friday, September 30th the hacktoberfest tag? Also, we're thinking of adding a few Scikit-HEP packages too, do you want to discuss during the next break?

14 replies Last reply 12 days ago

Kilian Lieret 2:21 PM

Welcome to HSF Training

Community pages



How-to guides

HSF Training Workshop Checklist ✅

Let's streamline our organization and make sure we don't forget anything!

Note: there's also a Hackathon checklist.

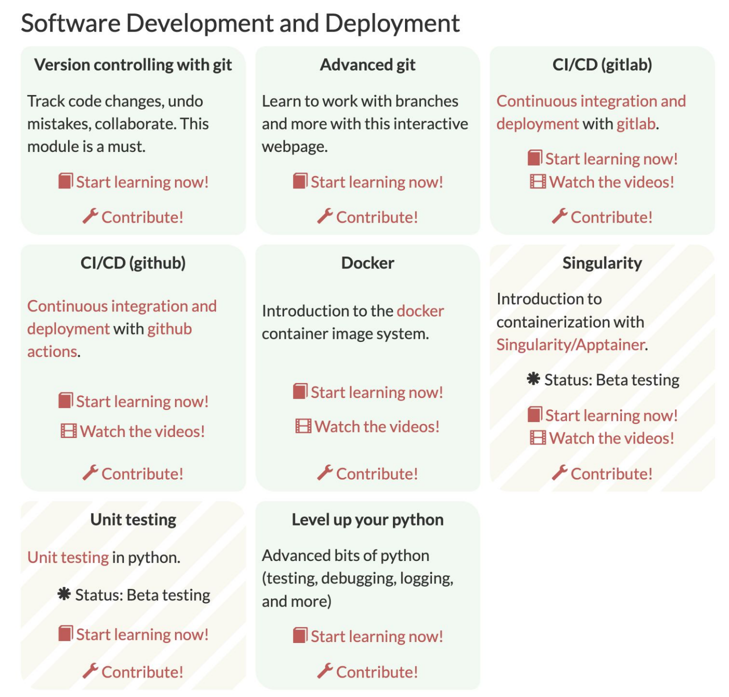
Before the workshop 📆

Setting up documents and more

- ☐ Create a new folder in our drive in the folder "20YY/our_workshops/YYMM_name_of_workshop"
- ☐ Copy this document there and call it "YYMM_name_checklist"
- ☐ Create an overview document "YYMM_name_planning": Use this for all the relevant planning info
- ☐ Create a document "YYMM_name_post_mortem": Collect everything that goes wrong in this document
- ☐ Create a slack channel for the event

Recruitment

- ☐ Announce on twitter
- ☐ Announce on mailing list



Software Development and Deployment

Version controlling with git

Track code changes, undo mistakes, collaborate. This module is a must.

Start learning now!

Contribute!

Advanced git

Learn to work with branches and more with this interactive webpage.

Start learning now!

Contribute!

CI/CD (gitlab)

Continuous integration and deployment with gitlab.

Start learning now!

Watch the videos!

Contribute!

CI/CD (github)

Continuous integration and deployment with github actions.

Start learning now!

Watch the videos!

Contribute!

Docker

Introduction to the docker container image system.

Start learning now!

Watch the videos!

Contribute!

Singularity

Introduction to containerization with Singularity/Apptainer.

✱ Status: Beta testing

Start learning now!

Watch the videos!

Contribute!

Unit testing

Unit testing in python.

✱ Status: Beta testing

Start learning now!

Contribute!

Level up your python

Advanced bits of python (testing, debugging, logging, and more)

Start learning now!

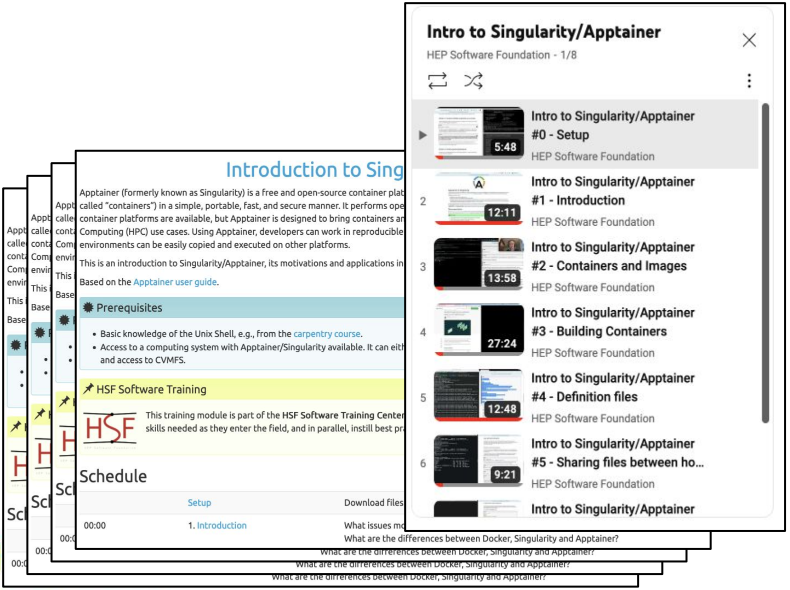
Contribute!

- We build a **unified**, **scalable**, and **sustainable** software training framework

Analysis Reproducibility

Train to Sustain

- **Data is expensive, experiments are unique**
  - Preserving the knowledge around them is a must
- We have developed modules to the learn the tools & consider analysis reproducibility right from the beginning



Introduction to Sing

Apptainer (formerly known as Singularity) is a free and open-source container plat called "containers") in a simple, portable, fast, and secure manner. It performs ope container platforms are available, but Apptainer is designed to bring containers an Computing (HPC) use cases. Using Apptainer, developers can work in reproducible environments can be easily copied and executed on other platforms.

This is an introduction to Singularity/Apptainer, its motivations and applications in

Based on the Apptainer user guide.

Prerequisites

- Basic knowledge of the Unix Shell, e.g., from the carpentry course.
- Access to a computing system with Apptainer/Singularity available. It can eith and access to CVMFS.

HSF Software Training

This training module is part of the **HSF Software Training Center** skills needed as they enter the field, and in parallel, instill best pr

Schedule

	Setup	Download files
00:00	1. Introduction	What issues mo What are the differences between Docker, Singularity and Apptainer?

What are the differences between Docker, Singularity and Apptainer?
What are the differences between Docker, Singularity and Apptainer?
What are the differences between Docker, Singularity and Apptainer?

Intro to Singularity/Apptainer
HEP Software Foundation - 1/8

1. 5:48 Intro to Singularity/Apptainer #0 - Setup — HEP Software Foundation
2. 12:11 Intro to Singularity/Apptainer #1 - Introduction — HEP Software Foundation
3. 13:58 Intro to Singularity/Apptainer #2 - Containers and Images — HEP Software Foundation
4. 27:24 Intro to Singularity/Apptainer #3 - Building Containers — HEP Software Foundation
5. 12:48 Intro to Singularity/Apptainer #4 - Definition files — HEP Software Foundation
6. 9:21 Intro to Singularity/Apptainer #5 - Sharing files between ho... — HEP Software Foundation

Intro to Singularity/Apptainer

- Developed by the HEP community during [Containerization & Analysis Preservation Hackathons]
- Using **[CMS Open Data]**
- **Analysis preservation is work-in-progress in most of the collaborations**
  - We do not focus in a particular reproducibility scenario
  - Instead, we review the commonly used tools

Analysis Reproducibility

Training modules

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

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

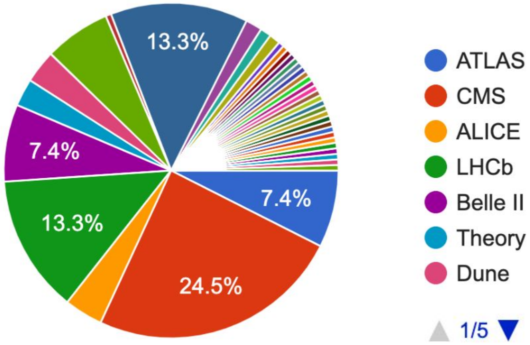
- **Continuous Integration/Deployment (CI/CD)**
  - GitLab pipelines
  - GitHub actions

- **Analysis platforms and workflows** (in preparation)
  - REANA
  - Snakemake
  - ...

Analysis Reproducibility

Virtual Events

- Organized in 2023 & 2024
  - Monday to Friday
  - **> 100 registrations per event**



Training on Analysis Preservation (Virtual)

16–21 Jan 2023
Virtual
Europe/Berlin timezone

Enter your search term

Learning the tools to make your analysis last to infinity and beyond!

Note: this tutorial used to be called: "Preservation" → "Pipelines" → "Reproducibility"

Brookhaven National Laboratory

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

Tuesday to Thursday

Work on your own, when you want

Friday

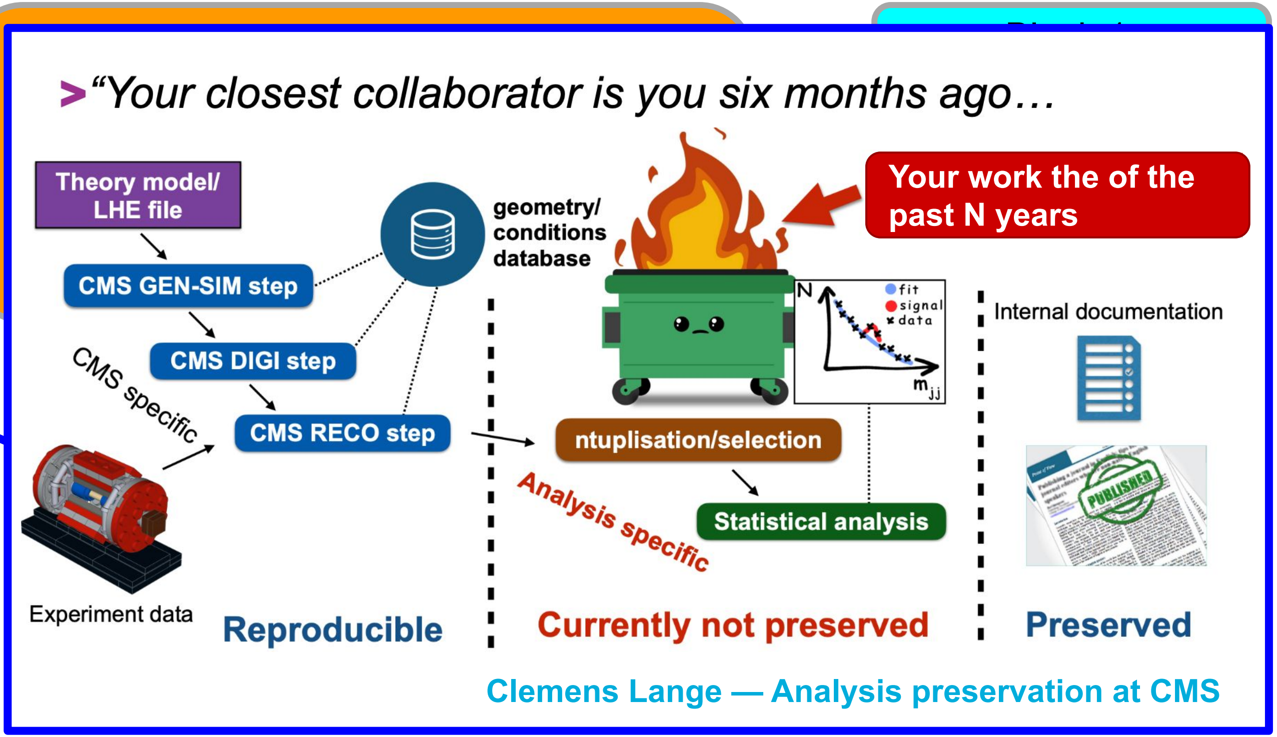
Hands-on sessions

Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

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



> "Your closest collaborator is you six months ago…

Clemens Lange — Analysis preservation at CMS

Brookhaven National Laboratory

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

Virtual support on Slack.

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

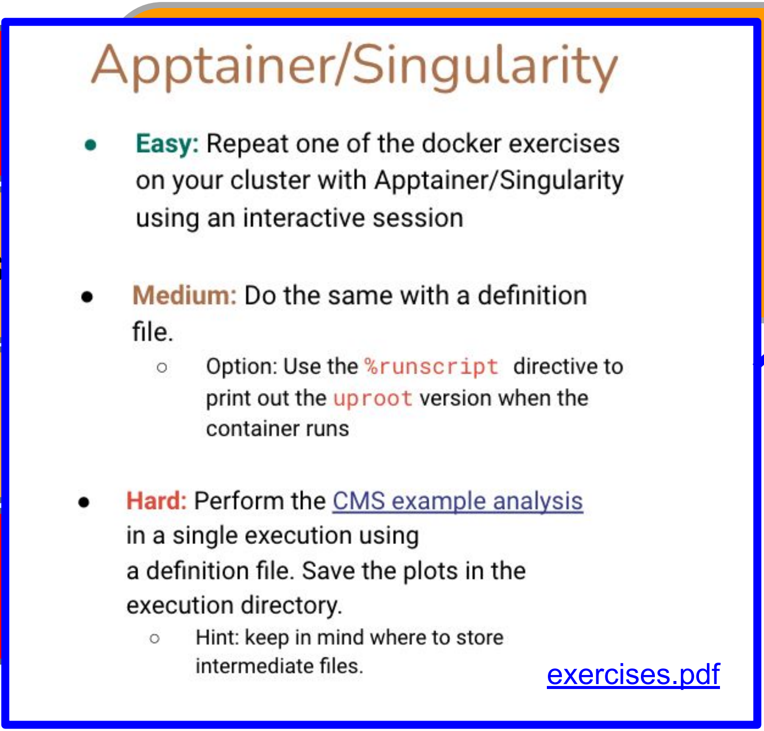
Kickoff/Orientation
[15:00 CET]

Analysis Preservation@CMS
[15:10 CET]

REANA
[15:40 CET]

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

rials:



Apptainer/Singularity

- **Easy:** Repeat one of the docker exercises on your cluster with Apptainer/Singularity using an interactive session
- **Medium:** Do the same with a definition file.
  - Option: Use the `%runscript` directive to print out the `uproot` version when the container runs
- **Hard:** Perform the CMS example analysis in a single execution using a definition file. Save the plots in the execution directory.
  - Hint: keep in mind where to store intermediate files.

exercises.pdf

Block 1:
[10-12 CET]

Block 2:
[13-15 CET]

Block 3:
[17-19 CET]

Block 4:
[21-23 CET]

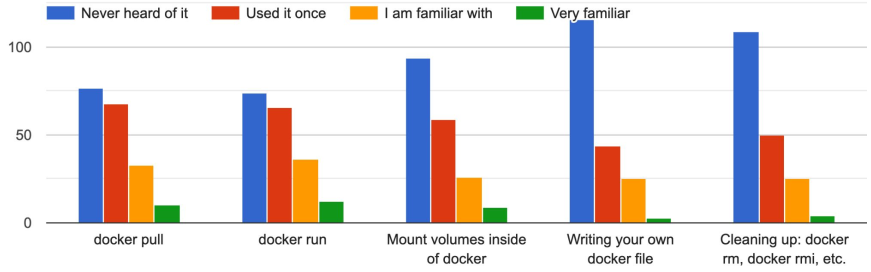
Brookhaven
National Laboratory

Surveys

Measuring how effective is our training

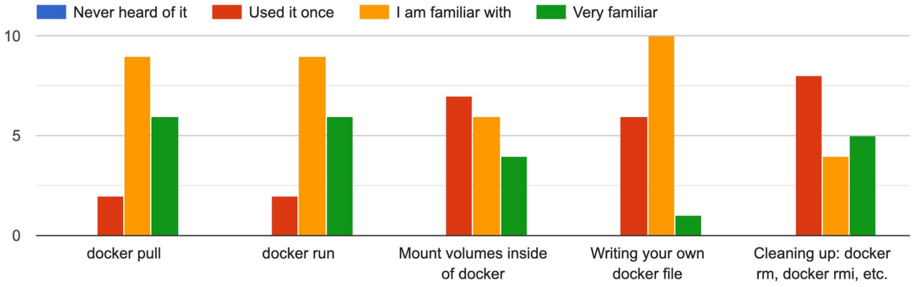
How comfortable are you with the following docker commands/tasks

Pre-workshop

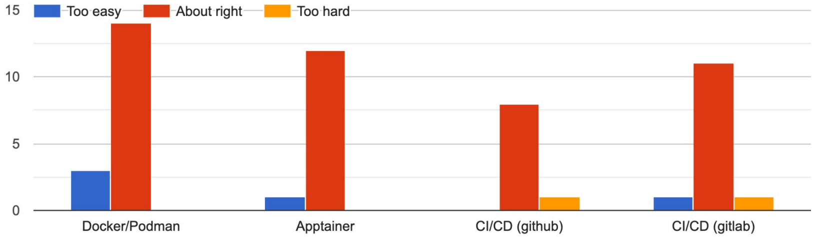


How comfortable are you with the following docker/podman commands/tasks

Post-workshop

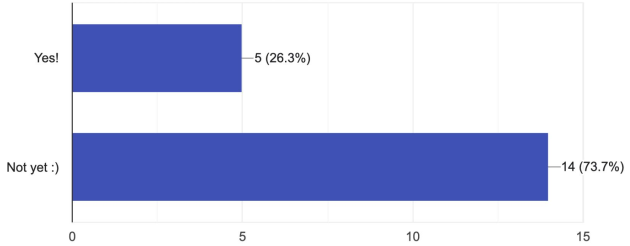


How difficult were the lessons/exercises?



Would you like to join the next events as a teacher or mentor? We will be happy to provide all the support necessary for your success.

19 responses



Analysis Reproducibility on ATLAS

Opportunities for collaboration

- The HEP/NP community would benefit greatly from learning about analysis reproducibility in ATLAS
- Many possibilities to share the ATLAS experience
  - An **introductory talk** on analysis reproducibility in the ATLAS collaboration
  - Include our material in ATLAS training events and **improve it based on the feedback**
    - Source code in GitHub, modules written in Markdown
    - A good example is the feedback from the Analysis Preservation BootCamp @ Valencia
  - Usage of **ATLAS open data** & analysis examples for HSF modules
  - New training modules with tools used at ATLAS

Summary

- The HSF Training is a community-driven effort, covering the software training requisites for a sustainable operation of physics experiments

- We have included a training event teaching Analysis Preservation: containerization & CI/CD with open data
  - **Extend to more related topics depends on the motivation of the community**

- Public weekly meetings: Mondays at 4pm CEST
  - https://indico.cern.ch/category/10294/
  - Everyone is welcome to join!

- **Reach us also via the channels shown in our webpage.**

Join us!

@hsf-training

HSF HEP Software Foundation hepsoftwarefoundation.org

indico

slack

Backup

Training and Onboarding Initiatives in HEP/NP

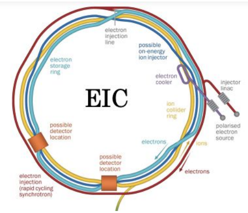
How do experiments teach software?

Virtual — Hybrid — In person

Belle II: Online book

LHCb: Starter Kit (As a consequence of covid19)

CMS: Data Analysis Schools


EIC
electron injection line
possible on-energy ion injector
electron storage ring
electron cooler
injector linac
ion collider ring
polarised electron source
possible detector location
electrons
ions
possible detector location
electrons
electron injection (rapid cycling synchrotron)

Online tutorials

ATLAS EXPERIMENT: Software tutorials

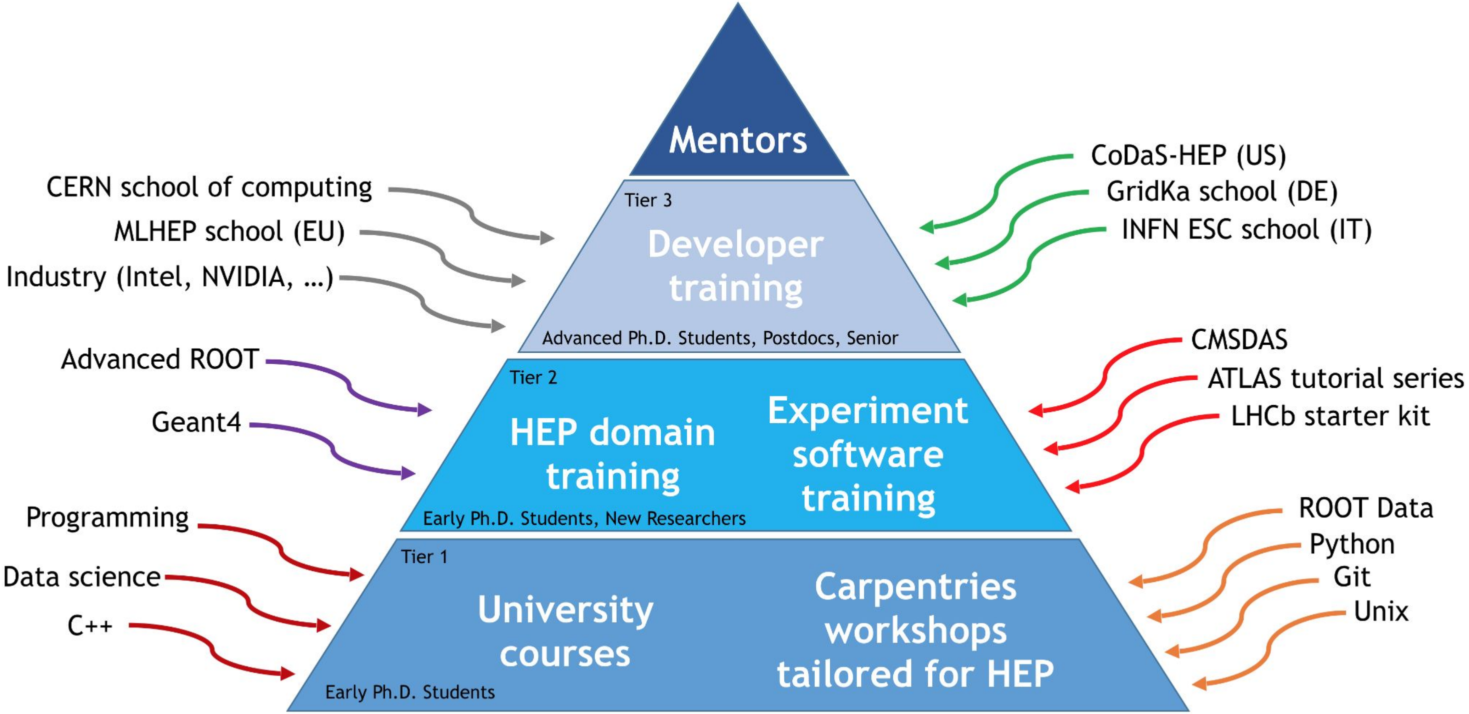
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT: Synchronous tutorials “Carpentries-style”

“Software is different, but challenges are common”

[Reinsvold Hall, Allison (US Naval Academy), CHEP 2023]

Software Training

The training pyramid



https://iris-hep.org/ssc.html

HSF Software Training

Principles

We need a **unified**, **scalable**, and **sustainable** software training framework

Unified

- Material and events should be **centrally listed** & **discoverable**
- Concentrate efforts by developing **cross-experiment** content
- A **community** must guide, support, and coordinate

Scalable

- Material must be teachable by **multiple instructors**
- **Self-study** must not be an afterthought

Sustainable

- Material must be **open source** and **maintained collaboratively**
- **Incentives & recognition** important motivators

HSF Training Center

- **The big picture: scientists with skills for delivering high-quality and reproducible analysis**

- Train our community with the best practices for sustainable software development
- A few examples:
  - Continuous Integration
  - Testing, testing, testing
  - Reproducibility, preservation
  - Project development methodologies
  - Green coding practices: efficient algorithms and data structures, reduce memory consumption and network traffic…

We are halfway on this list.

Reaching the bottom needs support from the community

- Large impact at computing centers ($$) in the long term!

Brookhaven National Laboratory

Training Events

In Person

- HSF Training software tutorials through 2020:
    - In-person participation only
    - Approximately 35 participants per workshop

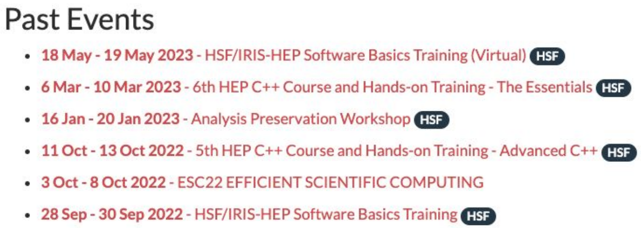
- Ecological and social impact:
    - Travel limits the accessibility to research groups with access to sufficient funding
    - Typical carbon footprint ~0.5 t CO2e / person:
        - Intra-continental travel: 0.4 t CO2e per person
        - Hotel stays: ~25 kg CO2e per person per night
    - Compare with estimated average EU (US) annual carbon footprint of 7 (16) t CO2e per person
    - **A workshop can increase one’s footprint by 5% to 10%**

Brookhaven National Laboratory

Training Events

Virtual

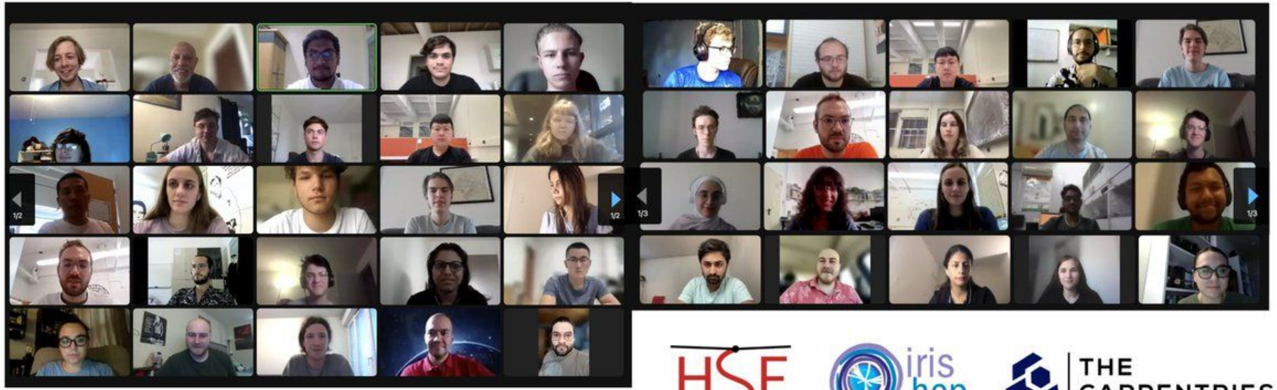
- Holding Virtual events since 2020.
  - COVID-motivated, but this training modality is here to stay
- 18 online software trainings, **1300+ participants trained**
  - Logistics are easier. Recordings available
  - Minimum environmental impact



Past Events

- 18 May - 19 May 2023 - HSF/IRIS-HEP Software Basics Training (Virtual) HSF
- 6 Mar - 10 Mar 2023 - 6th HEP C++ Course and Hands-on Training - The Essentials HSF
- 16 Jan - 20 Jan 2023 - Analysis Preservation Workshop HSF
- 11 Oct - 13 Oct 2022 - 5th HEP C++ Course and Hands-on Training - Advanced C++ HSF
- 3 Oct - 8 Oct 2022 - ESC22 EFFICIENT SCIENTIFIC COMPUTING
- 28 Sep - 30 Sep 2022 - HSF/IRIS-HEP Software Basics Training HSF

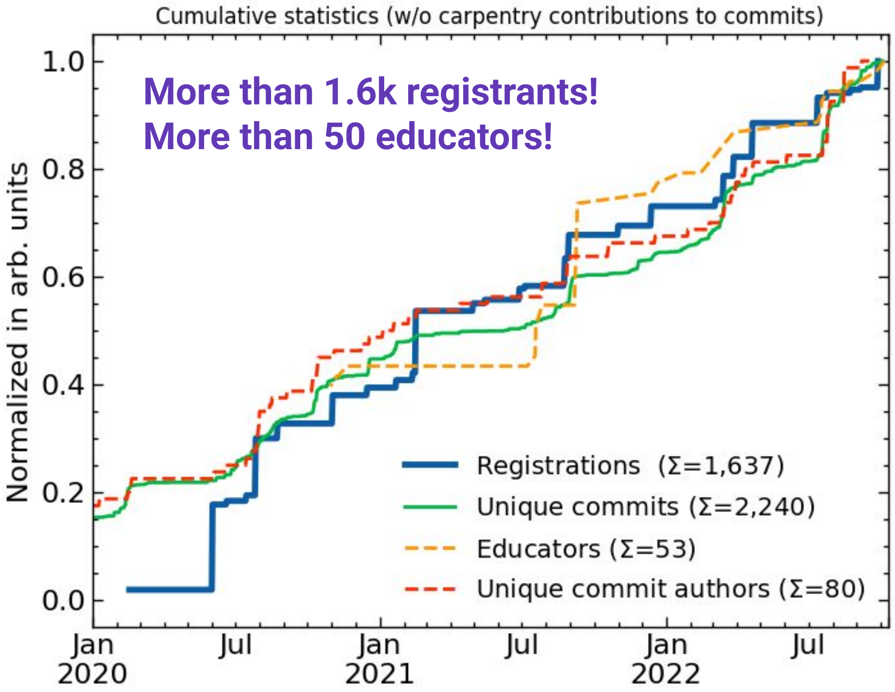
- But also disadvantages:
  - Lower engagement, distractions
  - Meaningful interactions harder



The Training Challenge

Scaling up

- Proposal to expand the effort in the long-term
  - **Scalability**: What is the number of students to reach? How many events does imply?
  - **Sustainability**: How to incentivize new trainers to continually join?
  - … **and Sustainability:** How to minimize the environmental impact, delivering effective training?
  - **Diversity and inclusion**: Everyone feels welcome to participate? How to standardize metrics?



More than 1.6k registrants!
More than 50 educators!