

Training Activities and Experiences in the HSF

Delivering sustainable software training

Michel Hernandez Villanueva
Brookhaven National Laboratory

... And many more! (see slide 20)

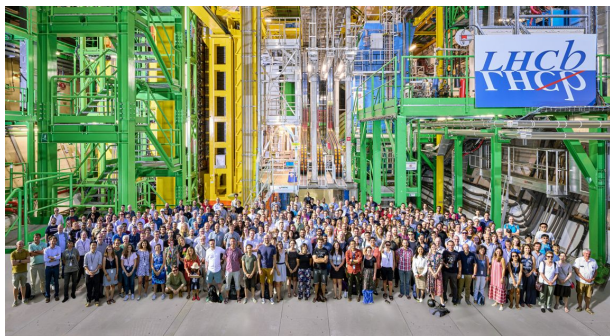
PUNCH Lunch Seminar
June 26, 2025



Software Development in HEP

As a key for a sustainable scientific program

- Scientific collaborations are **big and growing**.
 - $O(1K)$ collaborators in hundreds of institutes around the world.



Software Development in HEP

As a key for a sustainable scientific program

- Scientific collaborations are **big and growing**
 - O(1K) collaborators in hundreds of institutes around the world
- High Energy Physics (HEP) and Nuclear Physics (NP) are **computationally intensive** and **data driven** fields
 - A full physics potential requires investment into the software used to collect, process, and analyse data
- **Developers with strong foundation** are critical resources in the success of the current and future experiments
 - The researchers must be brought up to date with new software technologies, concurrent programming, and artificial intelligence
 - They must maintain, improve, and sustain the software

Training and Onboarding Initiatives in HEP

How do experiments teach software?

Virtual

Hybrid

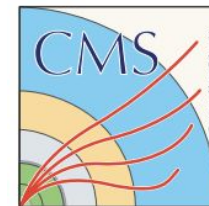
In person



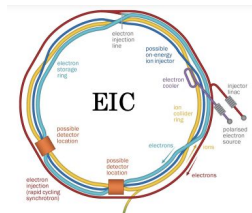
Online book



Starter Kit



Data Analysis Schools



Online tutorials



Software tutorials



Synchronous tutorials
“Carpentries-style”

“Software is different, but challenges are common”

[Front. Big Data 8:1497622. doi: 10.3389/fdata.2025.1497622](https://doi.org/10.3389/fdata.2025.1497622)

HEP Software Training

Why not leave everything to the experiments?

- **O(10k) HEP 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**
 - Training activities are not the most popular in making career steps and by funding agencies
 - Material for training is a moving target as technology evolves (e.g., Deep NNs, GPUs, FPGAs, ...)
- **This should be a community effort!**



HSF Software Training

Organization

- Established in 2018
- An active community of members supporting training on voluntary basis 🤝
 - Anyone is welcome to join and contribute!
- Develops material for an introductory software curriculum
 - And teaches this curriculum to scientists
- Focuses on common software material across HEP
 - From basic core software skills to advanced training
- Remote weekly public meetings (via Zoom) to plan and assess progress
 - Proposals are discussed and events are planned
- 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!

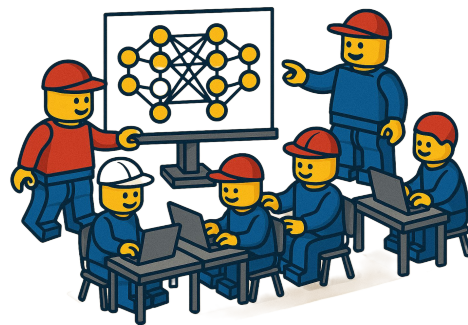
Self study!

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>

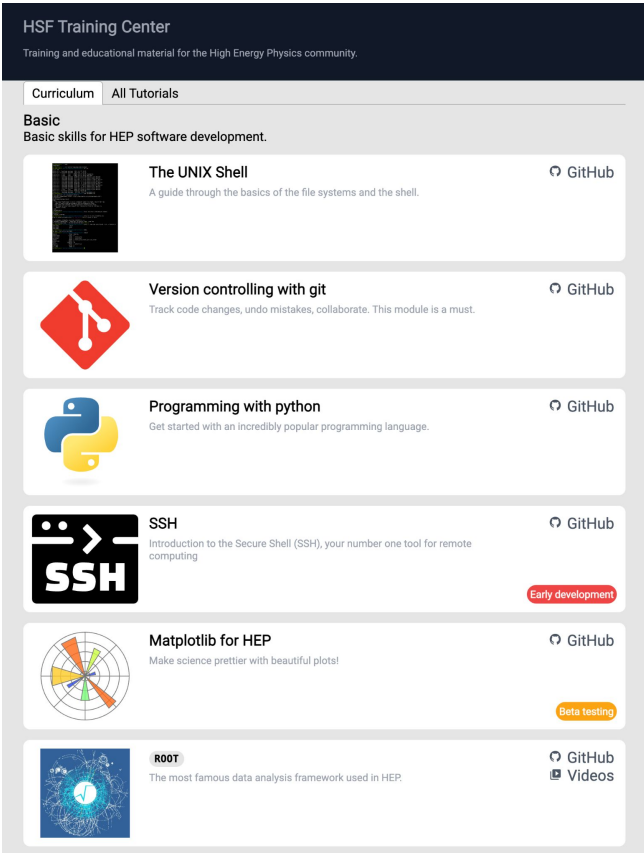
The HSF-Training Philosophy

- **Hands-On**
 - Learning is achieved by doing
- **Experiment Agnostic**
 - Teach tools and techniques that are independent of a specific context
- **Reusable**
 - If resources already exist, use them
 - If they do not exist, develop them
- **Open and Accessible**
 - If someone wants to learn, they should be able to learn



HSF Training Center

- Most experiments have similar training needs and material
- We are aiming to compile the available material in a single entry point
- Students can focus on what is most relevant to them
 - Basics as Unix shell, Python, Git
 - Intermediate modules as Docker, Apptainer, CMake, CI/CD
 - Some specific to HEP data analysis: ROOT, Scikit-HEP, REANA



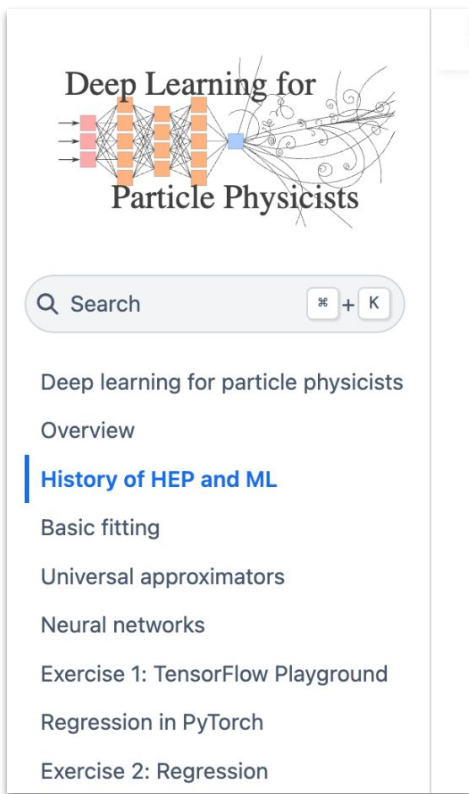
The screenshot displays the HSF Training Center website. At the top, the header reads "HSF Training Center" with the subtitle "Training and educational material for the High Energy Physics community." Below the header, there are two tabs: "Curriculum" and "All Tutorials". The "Basic" section is highlighted, with the description "Basic skills for HEP software development." A list of tutorials follows, each with an icon, title, description, and a GitHub link. The tutorials are: "The UNIX Shell" (terminal icon), "Version controlling with git" (git logo), "Programming with python" (python logo), "SSH" (SSH logo), "Matplotlib for HEP" (matplotlib logo), and "ROOT" (ROOT logo). The "SSH" and "Matplotlib for HEP" entries have additional labels: "Early development" and "Beta testing" respectively. The "ROOT" entry also includes a "Videos" link.

Tutorial	Description	Status	Links
The UNIX Shell	A guide through the basics of the file systems and the shell.		GitHub
Version controlling with git	Track code changes, undo mistakes, collaborate. This module is a must.		GitHub
Programming with python	Get started with an incredibly popular programming language.		GitHub
SSH	Introduction to the Secure Shell (SSH), your number one tool for remote computing	Early development	GitHub
Matplotlib for HEP	Make science prettier with beautiful plots!	Beta testing	GitHub
ROOT	The most famous data analysis framework used in HEP.		GitHub, Videos

<https://hsf-training.org/training-center/>

The Training Modules

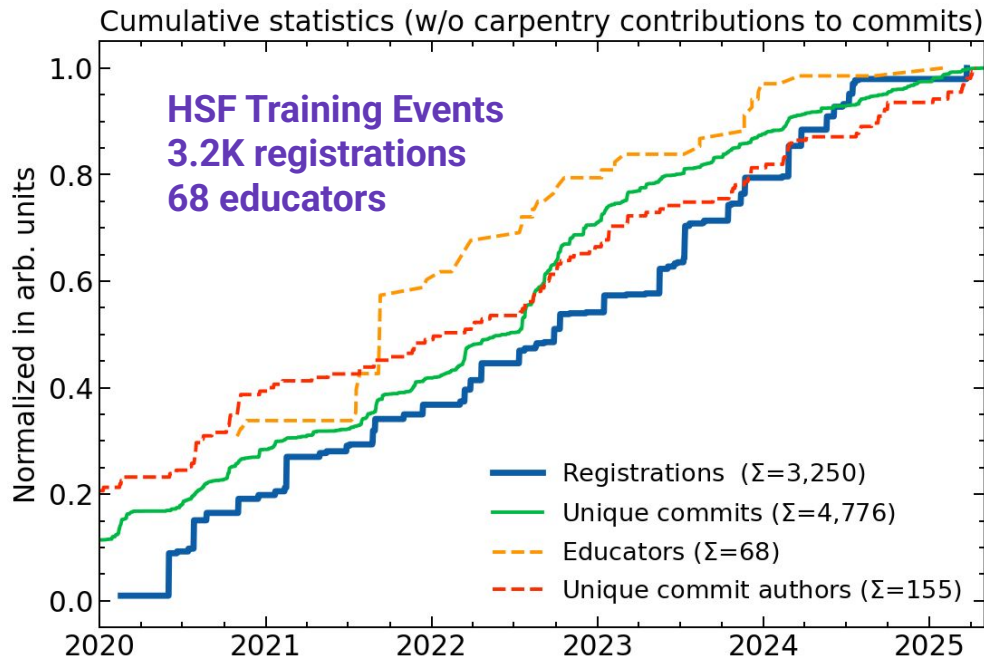
- The style and pedagogy of the HSF training is heavily inspired by [The Carpentries](#)
- Reusable study material that is open source and open access
 - Many licensed under [CC BY 4.0](#): free to share, remix, transform, and build upon the material for any purpose
- Hosted in the [HSF's Training repositories](#) on GitHub
 - Using Carpentries styles and [Jupyter Book](#)
 - Anyone is welcome to open issues and pull requests



hsf-training.github.io/deep-learning-intro-for-hep/

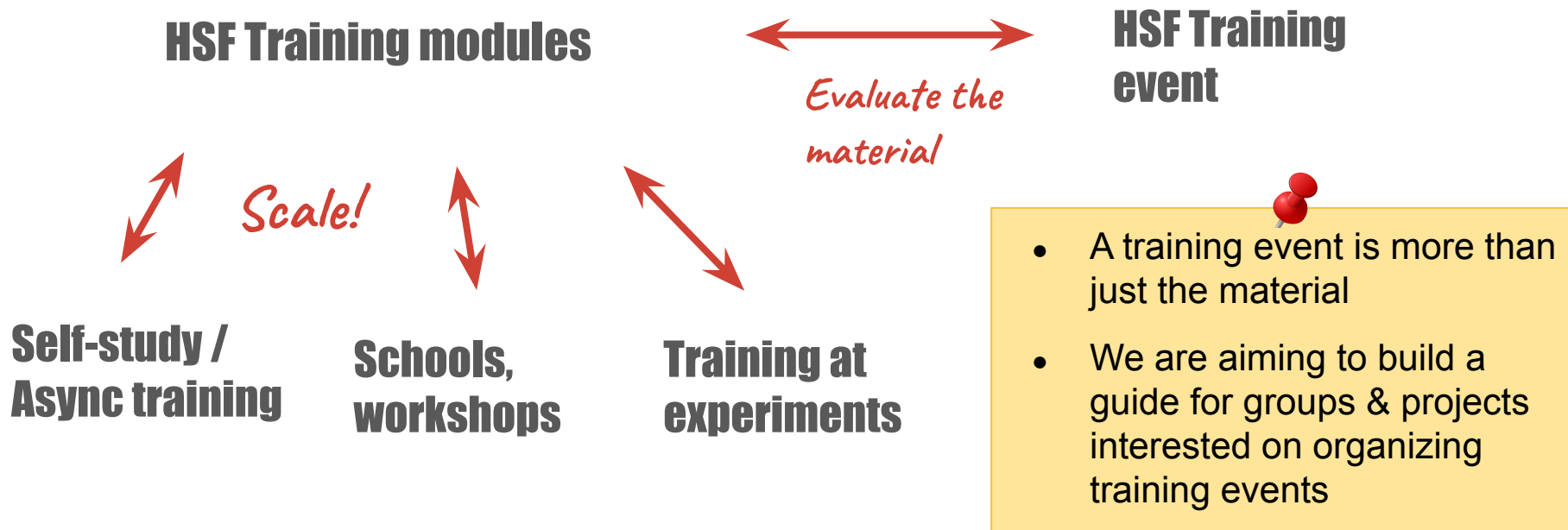
The Training Challenge

- Active discussion on how to expand the training 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?



Coping with The Training Challenge

- $O(10K)$ trainees across HEP and Nuclear Physics



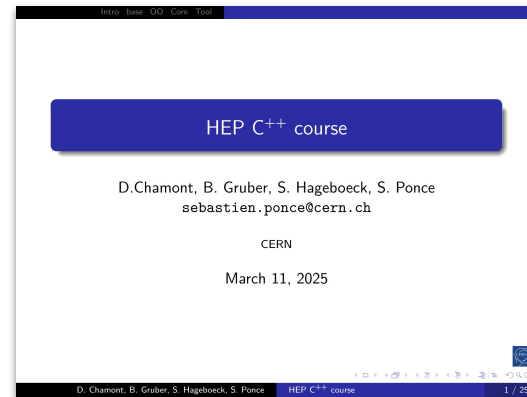
HSF Software Basics (online)

- Since 2021 we have established a training event for newcomers
 - COVID-motivated, but this training modality is here to stay
 - Examples: [here](#), [here](#) and [here](#)
- Two days of fast-track with software fundamentals: Bash, Git, Python
- One day for HEP Software: [ROOT](#) , [Scikit-HEP](#) (analysis in Python)
- Material and recordings are preserved on the page of the event
- Logistics are easier. Recordings available
 - Plus, minimum environmental impact
- But it also has disadvantages:
 - Lower engagement, distractions
 - Harder to enable meaningful interactions and networking



Back to In-Person Training

- Post-COVID-19, in-person activities ramped up slowly
 - A notable example is the [C++ training events](https://indico.cern.ch/event/1516608/)
- HSF Software Basics Training at CERN
 - 18–20 June 2025, CERN
<https://indico.cern.ch/event/1516608/>
- Impact on social and ecological sustainability:
 - Travel limits the accessibility to research groups with sufficient funding
 - Intra-continental travel + hotel: ~ 0.5 t CO₂e per person
 - 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%



HSF Training software tutorials

“Async” Training: Analysis Reproducibility

Monday

Group picture!

CMS Open Data simplified analysis example



<https://opendata.cern.ch/research/5900>

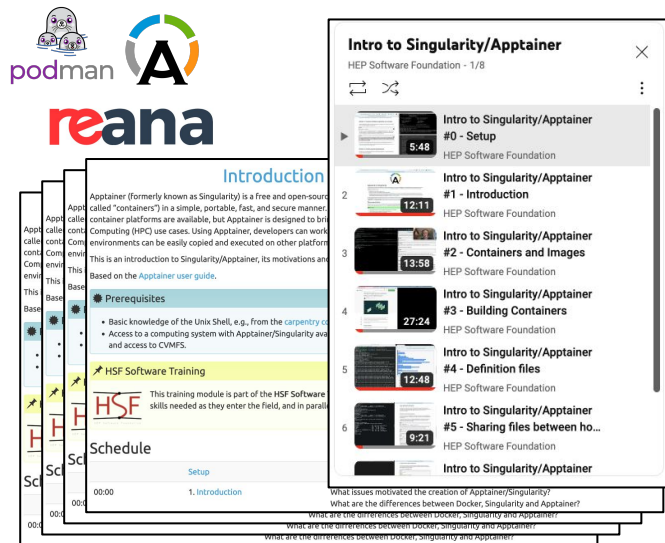


A kick-off session

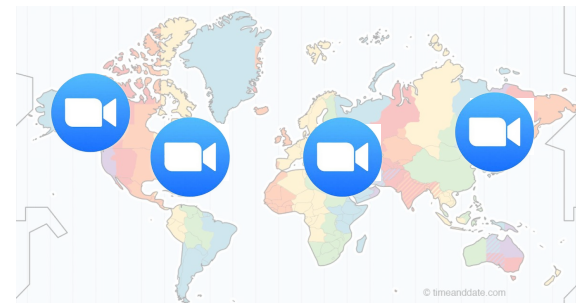


Experts available via chat

During the week



Friday



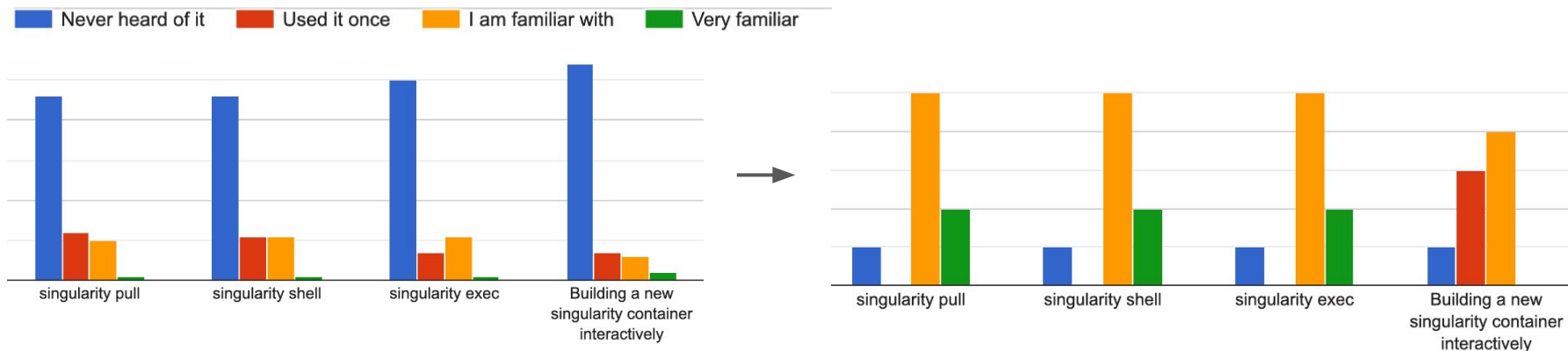
Mentoring sessions

67 registered participants

<https://indico.cern.ch/event/1508102>

Lessons with Async Training

“How comfortable are you with the following Apptainer topics”?



- Comparison between pre and post-surveys show that the training “works”
- However, low activity in the chat during the week and in the mentoring sessions
 - It affects future events, as experts find mentoring a not good use of their time
- We are talking with other communities about how to make async events more effective

Building a Shared Vision

- We have used the same document for several years to store meeting notes and link resources. Anyone joining can look at the past
- A shared folder enable all members of the HSF training group to shared resources
 - Folders organized by year/training event

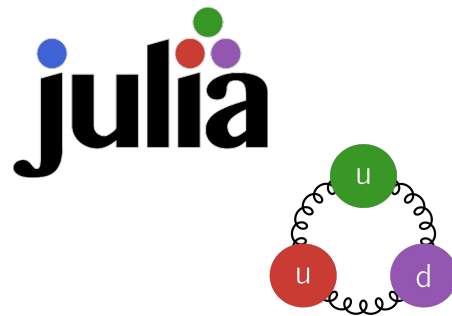
The screenshot shows a document editor with a sidebar on the left containing 'Document tabs' and a list of items including 'IRIS-HEP/FIRST...', 'TODO [keep on top]:', 'All upcoming/planned/...', 'Conferences where we ...', 'Meeting 2025-04-28', 'Attendance', 'Training Events', 'Training Center', 'JOSE paper on Appta...', 'AOB', and 'Meeting 2025-04-21'. The main content area is titled 'Meeting 2025-04-28' and contains sections for 'Attendance' (listing Richa, Callum, Andres, Michel, Lera, Alex, Marco) and 'Training Events' (listing HSF/WLCG Workshop, Software basics training at CERN, and various tasks related to the June 18-20 workshop).

The screenshot shows a Google Drive folder view titled 'Shared with me > ... > 2025 > our_workshops'. It includes a search bar, a 'Ask Gemini' section with buttons for 'Create an FAQ', 'Summarize this folder', and 'Ask about this folder', and a table of shared folders.

Name	Owner	Last modified	File size
25_06_swb	alexander.moreno	Feb 17, 2025 alexander.mor...	—
25_05_WLCG/HSF_Workshop	alexander.moreno	Apr 7, 2025 me	—
25_03_Analysis_Reproducibility	alexander.moreno	Jan 27, 2025 alexander.mor...	—

Next Training Events

- **Deep learning “Train-the-trainer”, with [ErUM-Data-Hub](https://indico.desy.de/event/47263/)**
 - 15–19 Sept 2025, Berlin
 - Practical knowledge on how to teach deep learning
 - <https://indico.desy.de/event/47263/>
- **Tutorials in JuliaHEP 2025 Workshop**
 - 28 - 31 July 2025, Princeton
 - <https://indico.cern.ch/event/1488852/timetable/>
- Join us to organize the next one!



Summary

- A successful experience organizing training events
 - However, challenges arise when considering the scalability of the training activity
- Reaching as many trainees as the field require a community effort
 - Relying on self-study, on boarding at the experiments, schools and workshops
 - Our HSF Training material is freely available, free to use and adapt
- We need your help!
 - Writing and maintaining training modules
 - As mentor in a HSF training event
 - Or organizing your own events with our material, sharing extremely valuable feedback
 - Reach us via the channels shown in [the HSF webpage!](#)

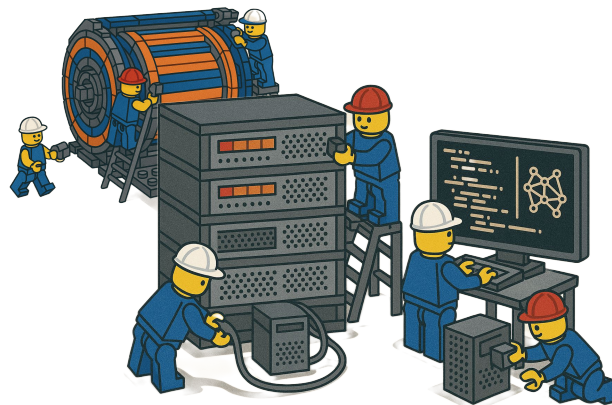
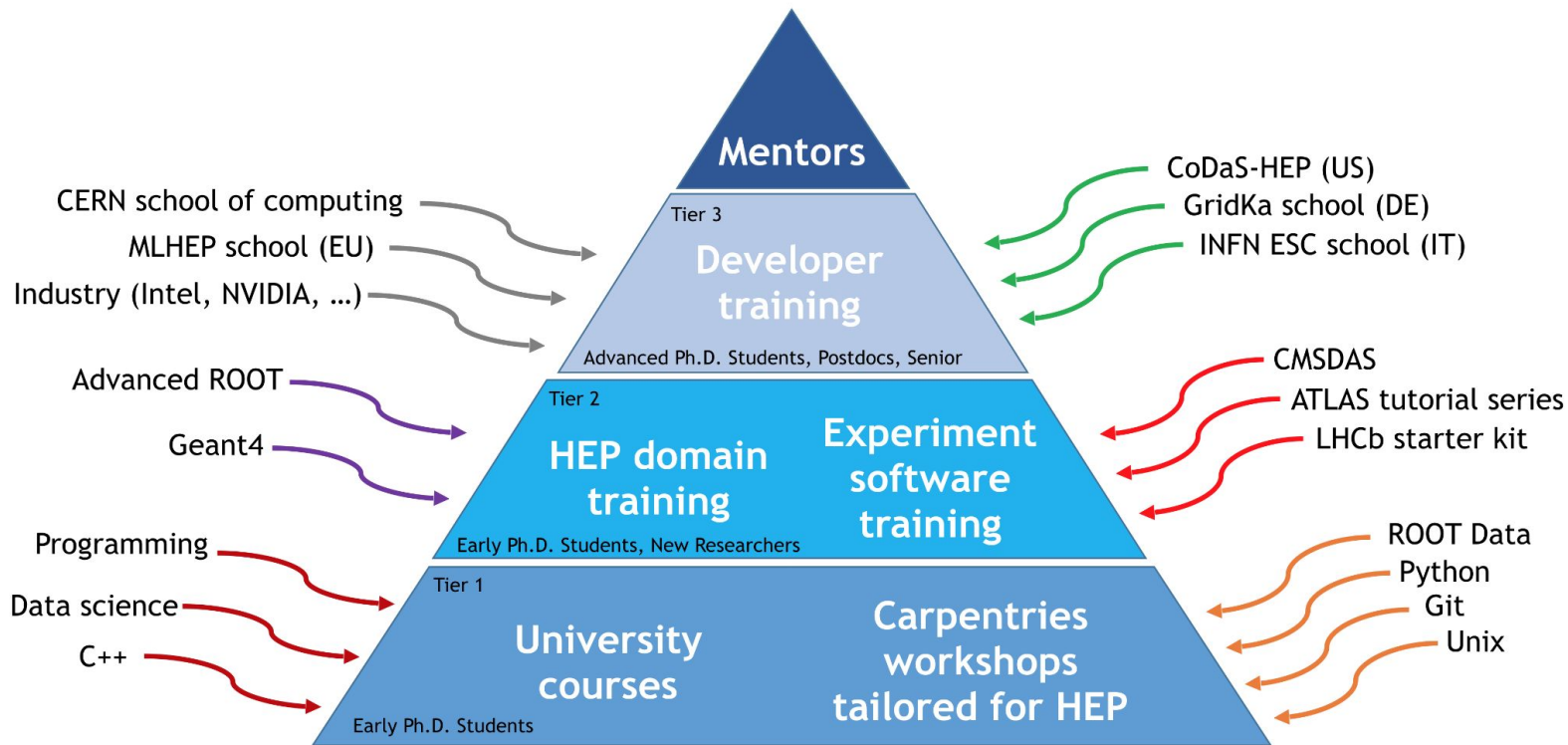


Figure: generated with DALL·E 3

Backup


HEP Software Training

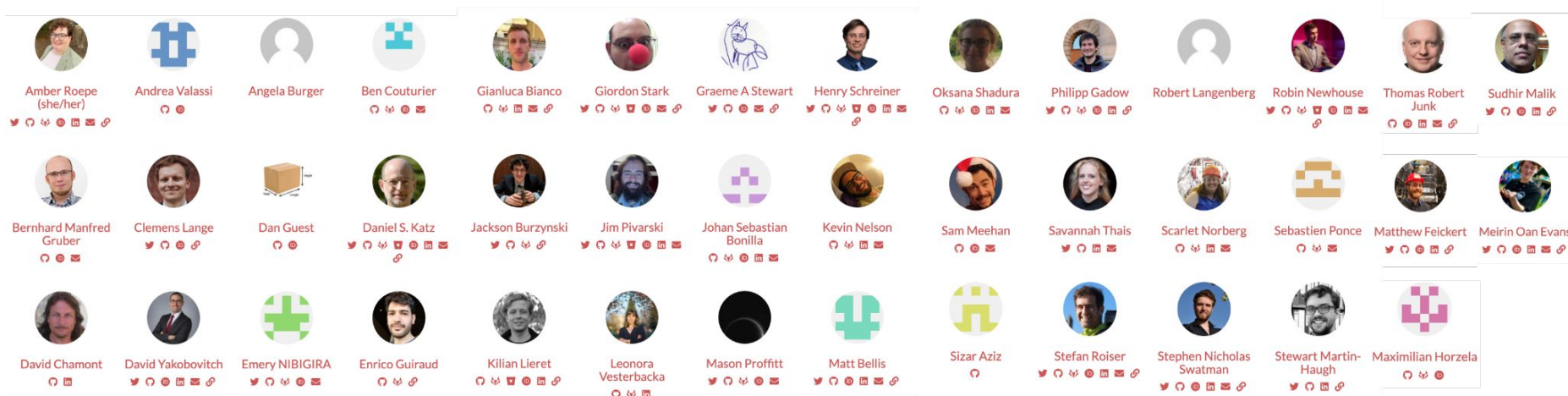
We can cover more ground together!



HSF Software Training

The community

- An active community of members supporting training on voluntary basis 
 - Coming from multiple collaborations, adding value to the training from different environments



<https://hepsoftwarefoundation.org/training/community.html>

The Roles for a Training Event

Three broad “Educator” roles are necessary

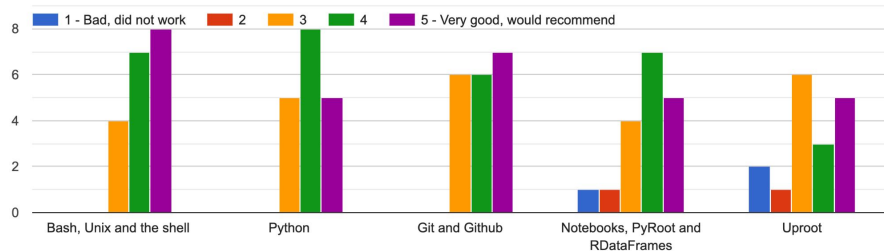
- **Facilitator**
 - Sends all the emails, organizes the agenda
 - makes sure everyone is on the same page
- **Instructors**
 - The experts on one of the given training modules
 - In an in-person event, they are in the front of the room
- **Mentors**
 - Energetic and eager individuals who know the material to varying degrees and are willing to help teach others
- Once we know the target attendance, we found that having ~1 mentor for every 5-10 participants works well



Surveys

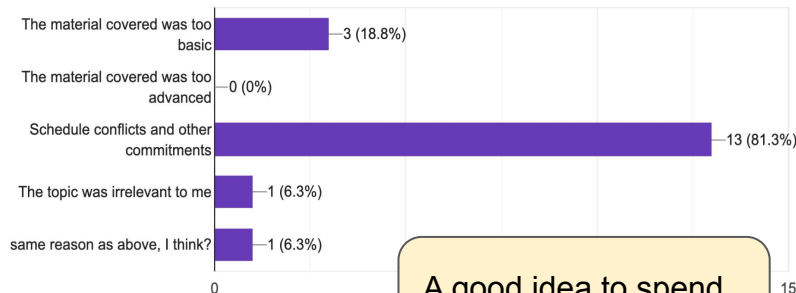
Some examples

Please rate how successfully each topic was covered (1 - 5)



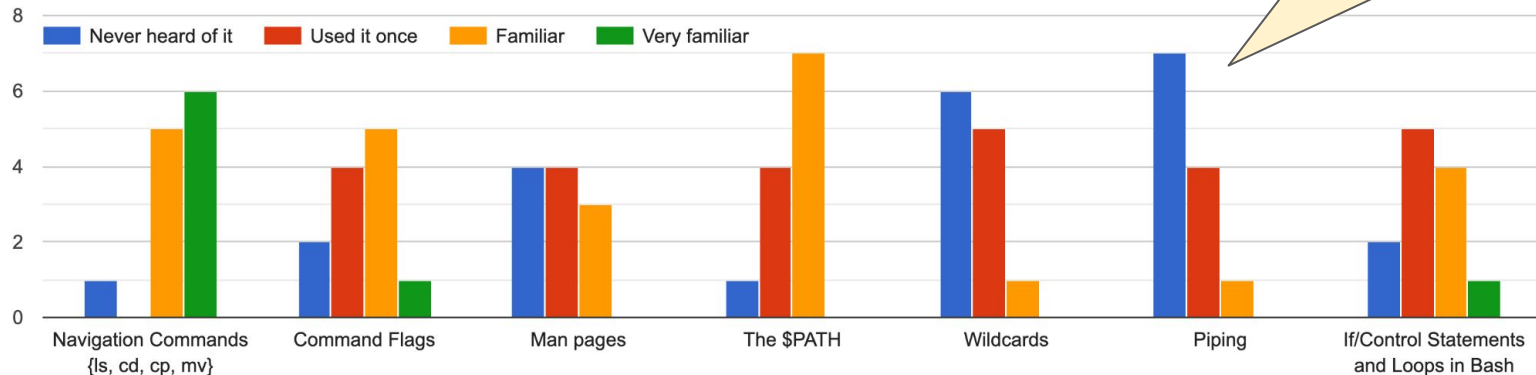
If you only attended FEW sessions, why did you skip the other sessions?

16 responses



A good idea to spend more time on this

“How comfortable are you with the following topics in the shell?”



The Setup

- A proper setup provides the environment where participants feel secure enough to experiment and make mistakes
- Minimize friction: pre-configured environments reduce time spent on setup, keeping focus on the actual learning. A few useful tools
 - GitHub Codespaces: full setup in a few clicks (example: [GitHub codespace with MySQL](#))
 - Containers: provide an image with the required environment
 - Conda environments, etc.
- We found effective as well
 - Pre-recorded videos explaining how to setup ([examples](#))
 - Booking an optional session on Indico to help with the setup

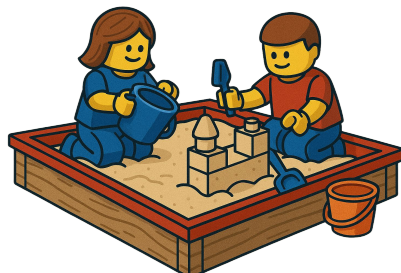


Figure: generated with DALL·E 3

Examples and Exercises

- Must be relevant and ideally have an immediate impact in the work of the trainee
 - Focus on (real!) problems in the HEP & Nuclear Physics context
 - Challenge on making them experiment-agnostic
- Some examples that we have found
 - Training on Apptainer:
 - A [definition file](#) for starting a Jupyter notebook with ROOT or Scikit-HEP tools available
 - Introduction to Databases (in development)
 - A [metadata catalog](#) with information of files like recorded in collisions
 - An introduction to [Conditions Databases](#)