HSF Training

Sudhir Malik
University of Puerto Rico Mayaguez
(on behalf of HSF Training Group)

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

Scientific Collaborations are big

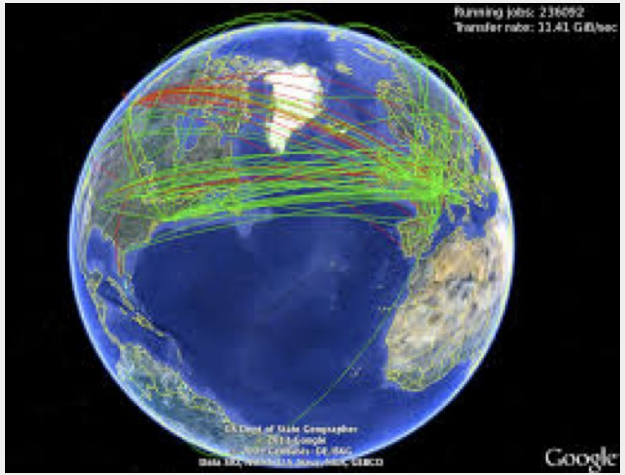
- Current
  - **CMS** - 4000 collaborators/200 institutes/50 countries
  - **ATLAS** - 3000 collaborators/174 institutes/38 countries
  - **LHCb** - 1200 collaborators/76 institutes/16 countries
  - **ALICE** - 1000 collaborators/100 institutes/30 countries.
  - **DUNE** - 1000 collaborators/180 institutions/30 countries
  - **LIGO** - 1200 collaborators/100 institutions/18 countries
- Past
  - **DZero** - 540 collaborators/90 institutions /18 countries
  - **CDF** - 600 collaborators, 30 institutions/12 countries

CMS Collaboration Map

Collaborations are big enterprise and multinational

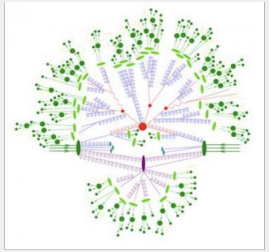
New Paradigm for Users

- Big Scientific Drivers and Computing Challenges
- Computing tools and physics analysis are intertwined
- Experimental challenge
- Enormous resources, person power
- Long life span of the experiment ~ 30 years
- Enormous data rate - ~ 10,000 copies/sec of Encyclopaedia Britannica
- Most users not resident at host laboratory
  - Financial and logistic constraints to be at host lab (e.g. CERN)
- Highly distributed environment for
  - Computing (Grid)
  - Physics analysis
- Physics/Computing Support
  - Should reach every user wherever they may be
  - Should be taken up in organized and central way



Need organized training for Big Science

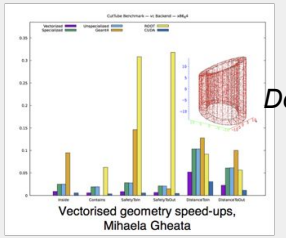
HEP software ecosystem



Software and Physics analysis are intertwined

Physics Event Generators

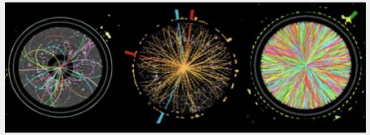
Data, Software, Analysis Preservation



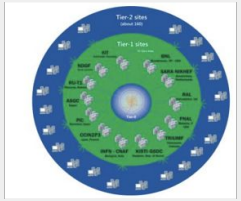
Detector Simulation

Security

Software Development



Trigger,Event Reconstruction



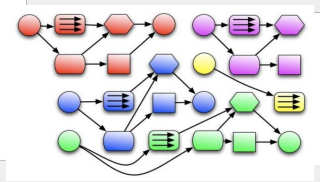
Facilities, Distributed Computing

Lots of challenges

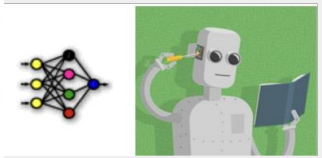
Data Analysis, Interpretation, Simulation

Visualization

Data Processing Frameworks



Machine Learning



Data Management Organisation, Access

Training benefits

- Computation is a central element of 21st century science
- Successful evolution of this ecosystem to meet the challenges
  - Requires new tools and a workforce HEP domain knowledge and advanced software skills
  - Enhances the vibrancy of the training and workforce development activities
- Investing in SW critical to match HL-LHC requirements of “flat budget” scenario
- Investing in training leads to preservation and propagation of knowledge
- Investing in software skills is not only important to actually build the requisite software infrastructure, but will also change community norms, create role models and promote career paths
- Transfer skills to next generation of our students and postdocs
- Training leads to sustainability

Investing in Training is key to Big Science

Trainings in HEP Experiments

- **CMS** - CMS Physics Analysis Toolkit, CMS Data Analysis Schools, CMS Physics Object School, CMS Upgrade School, Documentation, WorkBook
- **ATLAS** - Software Tutorials, Migration Tutorials, Developer Tutorials, WorkBook
- **LHCb/ALICE** Analysis tutorial week/Impactkit/StarterKit (shared), documentation
- **Belle II** - StarterKit, documentation,
- **Virgo/LIGO** - working towards organized training, documentation
- **Neutrino (FNAL)** - common S/W stack documentation, online cookbook, “101” kind training
- For more details, sustainability challenges about above, please have a look at “*Training and Careers” - lightning talks (Joint WLCG & HSF Workshop 2018, Naples, 2018)*
  - https://indico.cern.ch/event/658060/timetable/?view=standard
- Advanced software and computing topics - CERN School of Computing, GridKa school (Karlsruhe) , ESC School (INFN), recent CoDaS-HEP (Princeton)

Lots of specific (experiment) software trainings

Current Training - Limitations

- Training activities - fragmented and partially redundant
- Experiment specific flavor of software training despite commonality across HEP
- Modest effort devoted to training is not always positioned for maximum impact
- Training support is difficult to sustain over time
- Too often critically dependent on specific individuals whose careers evolve
- The effort to keep training materials up-to-date is too often lacking
- No single entity has a mandate to organize these disparate efforts into a collective effort whose impact would be much greater than the sum of its parts

Need for common (software) training

Training challenges

- **Not all funding agencies**, institutions and funded projects **have the same priority for training and education** (e.g. DOE vs NSF in the US) relative to other goals like building/operating experiments, physics analysis, etc

- **Training activities are not always valued** relative to other activities in making career steps

- **Will efforts** in this area **lead to** improvements in **software efficiency, scalability** and performance and make use of the advances in CPU, storage and network technologies, that allow the experiments to maximize their physics reach?

- **New tools requires** investing in training, separating “local” specifics (e.g. computing environments or experiment-specific bits) from **common usable material** is important, but doesn’t always happen

- Technology evolution means that **training materials need to evolve**, too. Are training materials a common good or an individual product? Even if individuals do want to contribute to a common good, how do they do so

This is where HSF steps in

HSF Training - vision, mission, philosophy

- **Vision** - People are the key to successful software, are the real cyberinfrastructure underlying sustainable software
- **Mission** - “to help the research community to provide training in the computing skills needed for researchers to produce high quality and sustainable software”
- **Philosophy** - Build training activities and material into a “common good” with a strong community of both instructors and participants, and with a feeling of community ownership
- **Training style** - Hands-on, Student-centric, Experiment Agnostic, Reuseable, Open and Accessible
- **Goal** - Sustainability ⟷ Scalability
- Trained about 700 participants

Training Partners

iris hep
Institute for Research & Innovation in Software for High Energy Physics

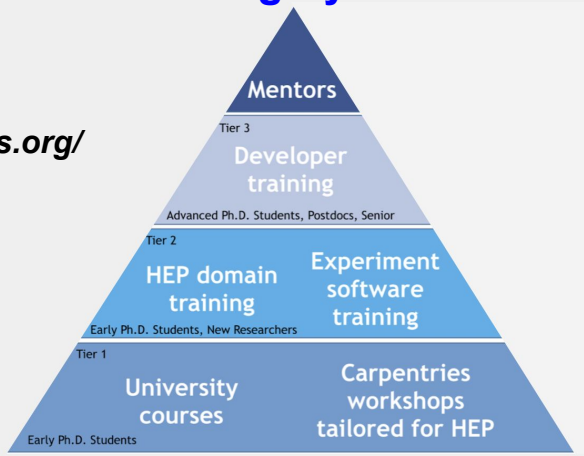
https://iris-hep.org/
OAC-1836650

FIRST-HEP
http://first-hep.org/
OAC-1829707
OAC-1829729

Training Motivation

software carpentry
https://carpentries.org/

Training Pyramid



Training Material

- **Training events**: https://indico.cern.ch/category/11386/
- **Material**: All the training material developed so far resides: https://hepsoftwarefoundation.org/training/curriculum.html
- **Community**: Our training community is listed here: https://hepsoftwarefoundation.org/training/community.html
- **Procedure**: how to request and organize a training: https://hepsoftwarefoundation.org/training/howto-event.html
- **Funding**: Funding for training events is provided by the IRIS-HEP/FIRST-HEP
- **Blueprint**: First blueprint on training https://indico.cern.ch/event/889665/

Training Lessons Wishlist

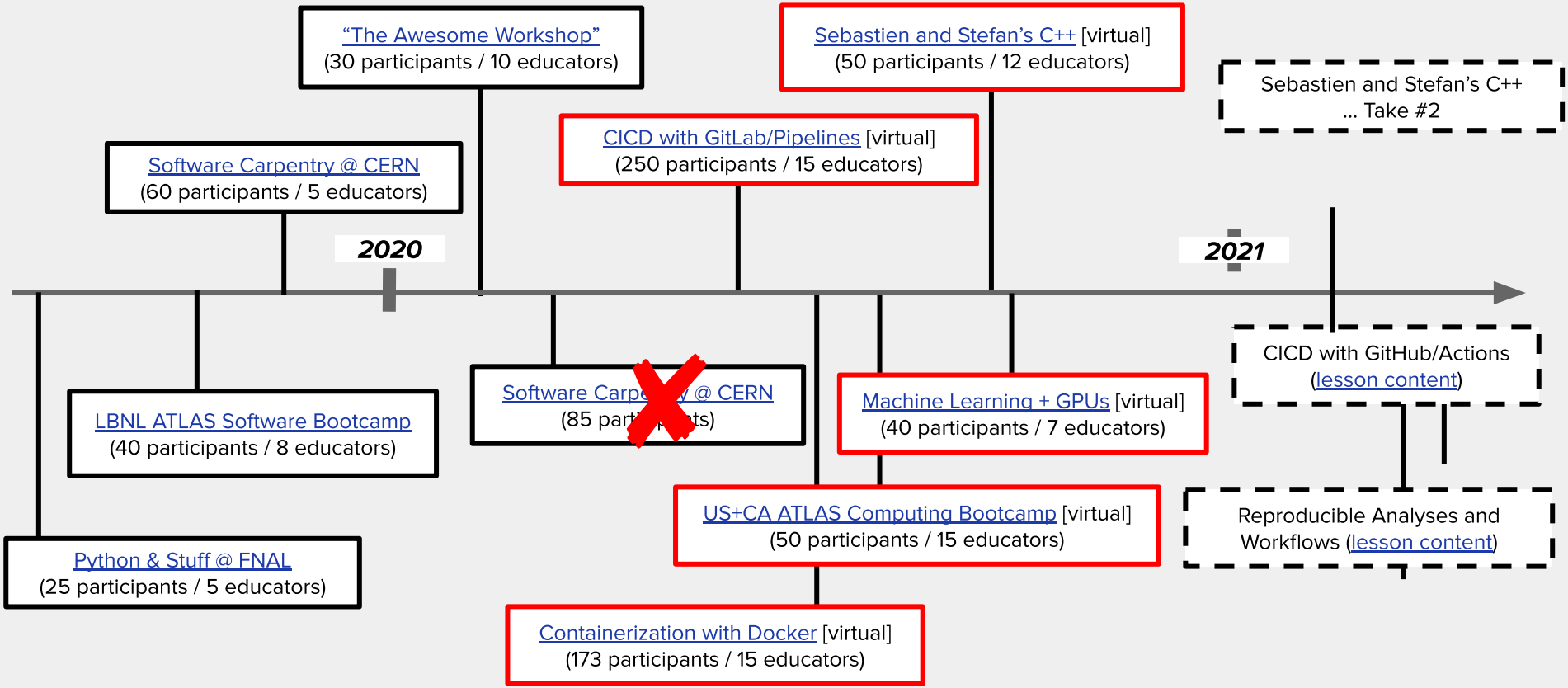
From the SWC Curriculum
Production Ready
In (various stages of) Development

1. **Git/vcs essentials/github (“How to”)**
2. Advanced module for git
3. **Python foundations**
4. Building programs with python
5. Data analysis: numpy, pandas
6. Advanced data analysis
7. Advanced python and pyroot, **uproot**
8. **Build systems: from gcc to cmake**
9. **Continuous Integration/Development**
10. **Docker and Containerization**
11. **Unix (shell, bash, scripting, ...)**
12. Advanced unix (shell, bash, scripting, ...)
13. Suggestion: Advanced Unix/terminal
14. Jupyter notebooks and Binder/SWAN
15. ROOT
16. **C++**
17. Package managers and RPMs
18. Distributed file systems (mounting, access protocols)
19. Batch systems (common scheduler concepts):
20. Distributed computing
21. Best practices and “software engineering”
22. Text editors (vim/emacs/...?) and IDEs
23. **Authentication in general; SSH; keys; ssh config; tunneling**
24. **Machine Learning**
25. Debuggers (gdb)
26. Parallel programming
27. **Workflows (e.g. yadage) & Reproducibility (e.g REANA)**
28. Monte Carlo (pythia, sherpa, madgraph, …)
29. Simulations (e.g. GEANT)
30 . Documentation (doxygen, sphinx …)

……..

……..

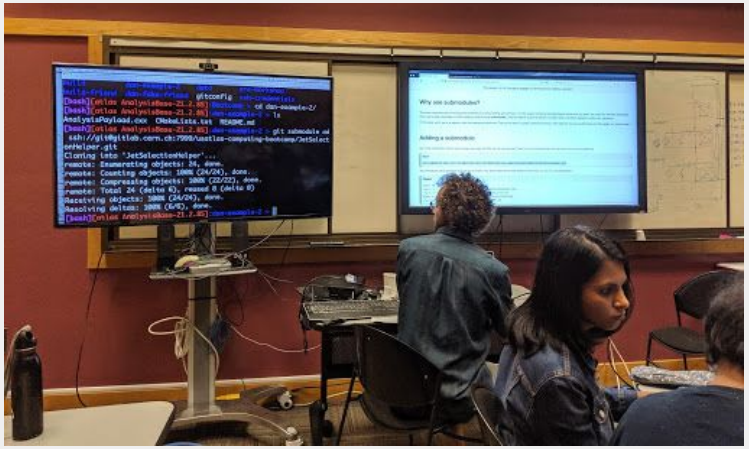
Trainings to Date



You want to have a high impact and advance HEP? - Training might be your most effective choice!

In person training

- **Attendance** : few ***dozen***
- **Positives**
  - Active/efficient engagement of participants
  - Professional networking and additional “events”
- **Negatives**
  - Travel costs (education should not be exclusive)
  - Long lead time for planning logistics
    - Related to travel/room booking
  - Requires participant “sacrifice”
- **Important things**
  - Room setup is crucial
    - Two projects/screens
    - Not an auditorium
    - Ample power
- **Suggested Ratio of Participant** : Educator <= 5
    - This is *essential* to allow for the “hands on” aspect of the workshop to be successful
- **Large time commitment** on behalf of the educators
    - Can’t just “do your talk” and then leave



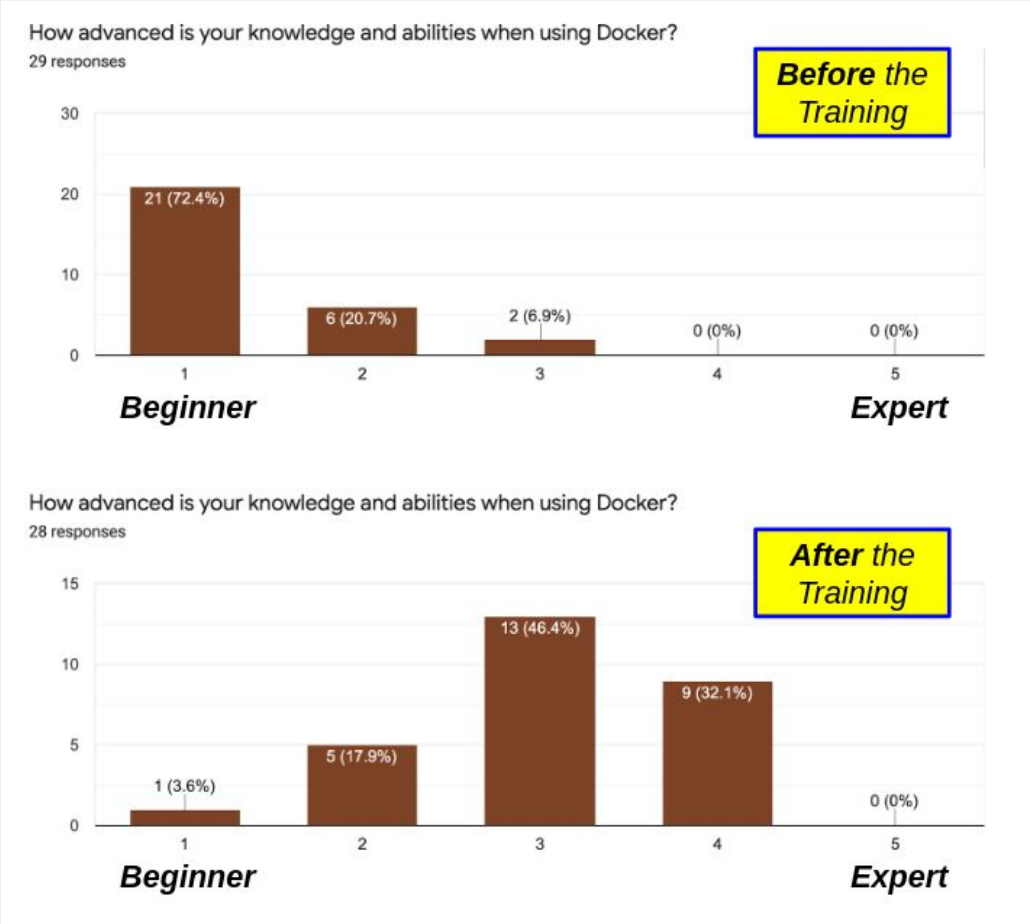
Virtual training (Covid Enforced)

- Attendance : few ***hundred***
- Positives
  - Broader reach : >100 registrants possible
  - 2 times greater likelihood to participate
  - No travel costs → critical for some supervisors
  - Don’t need to plan in as much advance
  - Materials are more fully preserved (i.e. videos)

- Negatives
  - Difficult educator/participant interactions
  - Need mentors spaced in (potentially) different time zones
  - Challenging to keep everyone on same page
- Important things
  - Have well defined roles
  - Effective chat application is essential
    - e.g. mattermost/discord/slack

Training Works !!

- We do our best to diligently collect before/after data via surveys
  - Pre-survey
    - Demographics
    - How much do you know?
  - Post-survey
    - How much do you **now** know?
    - What can we do better next time?
  - Would like to have further out “follow up” surveys (takes more work …)
- Self-reported learning *does* happen!



HSF Training Journey - special mention

Student in *Awesome Workshop*
Interested in helping in further events → **yes!**

Mentor in *Virtual Pipelines Training*
(asked last minute because of lack of mentors)

Mentor in *Virtual Docker Training*
(also helping out with surveys)

Joined weekly **meetings**
(heard that someone was interested in delivering a Machine Learning workshop…)

Instructor for ML Training
(developing & improving material)

Co-convenor for HSF Training
(leader)

Meirin Oan Evans

Future Plans

- **Focus on** building two groups -
  - **Community of individuals** - facilitators, learners, instructors, experts and hosts
  - **Core team to support** the overall mission of training
- **Build regional and local capacity** to empower sustainable HEP communities by creating local mentorship and leadership that is guided and supported by the core team. We need to engage HEP labs and R1 universities to achieve this goal.
- Mindful and proactive in **equity, diversity, inclusion and accessibility in participation** across HEP communities and serve as a role model while being mindful of under-resourced institutions and communities in different geographical regions.
- A workflow **mechanism to get feedback** from our communities and improve as we scale
- Our community is our strongest resource and we need to ensure that our **core team and volunteers are afforded opportunities** to grow professionally and have career paths.
- **Brainstorm sessions** for experts, **improve documentation**
- Explore ways for a **long term financial support model**
- **Expand collaboration** to Nuclear Physics,Neutrino and other related communities

Back up

In addition have a look at talks - HSF WLCG Virtual Workshop

19 -24 Nov 2020 https://indico.cern.ch/event/941278/