Workforce Development, Connection to Industry, Impact Beyond HEP (Software and Computing)

Peter Elmer - Princeton University

Computing Frontier: Software Development, Staffing and Training

Conveners: David Brown[1], Peter Elmer[2]

Observer: Ruth Pordes[3]

Contributing Authors: David Asner[4], Gregory Dubois-Felsmann[5], V. Daniel Elvira[3], Robert Hatcher[3], Chris Jones[3], Robert Kutschke[3], David Lange[6], Elizabeth Sexton-Kennedy[3], Craig Tull[1]

[1] *Lawrence Berkeley National Laboratory (LBNL)*
[2] *Department of Physics, Princeton University*
[3] *Fermi National Accelerator Laboratory (FNAL)*
[4] *Pacific Northwest National Laboratory (PNNL)*
[5] *SLAC National Accelerator Laboratory (SLAC)*
[6] *Lawrence Livermore National Laboratory (LLNL)*

Snowmass 2013 Working Group

Two goals from that report:

- **Goal:** We must insure that our developers and users will have the **training** needed to create, maintain, and use the increasingly complex software environments and computing systems that will be part of future HEP projects.

➡️ **Build a Sustainable and Scalable Model for Workforce Development**

- **Goal:** To maximize the scientific productivity of our community in an era of reduced resources, we must use **software development strategies and staffing models that will result in products that are generally useful for the wider HEP community**.
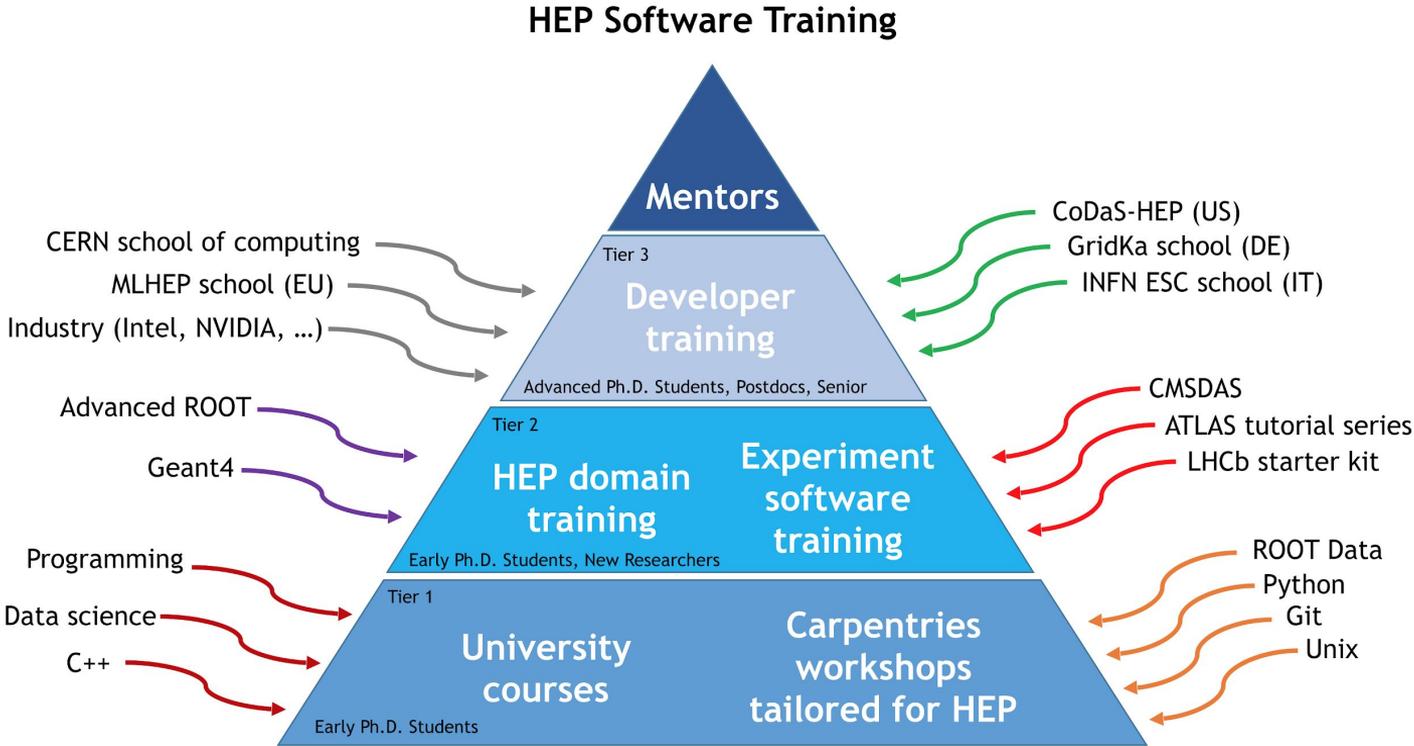
➡️ **Build wider Research Software Collaborations**

Workforce Development

Sustainable & Scalable Workforce Development Model?

HEP Software Training

2017-2018

Mentors

Tier 3
Developer training
Advanced Ph.D. Students, Postdocs, Senior

CERN school of computing
MLHEP school (EU)
Industry (Intel, NVIDIA, …)

CoDaS-HEP (US)
GridKa school (DE)
INFN ESC school (IT)

Tier 2
HEP domain training — Experiment software training
Early Ph.D. Students, New Researchers

Advanced ROOT
Geant4

CMSDAS
ATLAS tutorial series
LHCb starter kit

Tier 1
University courses — Carpentries workshops tailored for HEP
Early Ph.D. Students

Programming
Data science
C++

ROOT Data
Python
Git
Unix

Inclusive of transferable skills outside of HEP!

HSF Software Training Center

Central repository of training materials + instructor community - modeled on The Carpentries

Idea

Training in software and computing are essential ingredients for the success of any HEP experiment. As most experiments have similar basic prerequisites (Unix shell, Python, C++, ...) we want to join our efforts and create one introductory software training curriculum that serves HEP newcomers the software skills needed as they enter the field, and in parallel, instill best practices for writing software.

The curriculum is comprised of a set of standardized *modules*, so that students can focus on what is most relevant to them.

The modules

Basics

The UNIX Shell

A guide through the basics of the file systems and the shell.

📕 Start learning now!
🔧 Contribute!

Version controlling with git

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

📕 Start learning now!
🔧 Contribute!

Programming with python

Get started with an incredibly popular programming language.

📕 Start learning now!
🔧 Contribute!

SSH

Introduction to the Secure Shell (SSH)

⚠ Status: Early development

📕 Start learning now!
🔧 Contribute!

Machine learning

Get behind the buzzword and teach machines to work for you intelligently!

📕 Start learning now!
🎬 Watch the videos!
🔧 Contribute!

Matplotlib for HEP
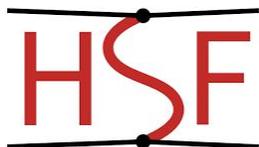
Make science prettier with beautiful plots!

❋ Status: Beta testing

📕 Start learning now!
🔧 Contribute!

ROOT

The most famous data analysis
~~framework used in HEP~~

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!

C++ corner

HEP C++ Course

A full introduction to C++ based on a series of slides and exercises.

Build systems: `cmake`

Building code is hard. `CMake` makes it easier.

Plus more...
And growing...

1500 students/50 instructors to date.

Computational and Data Science (CoDaS-HEP) Summer School

http://codas-hep.org

In-person summer school with Lectures & hands-on exercises:

- Parallel Programming
- Data Science Tools and Techniques
- Machine Learning - Technology and Methods
- Practical skills: performance evaluation, use of git for version control

CoDaS-HEP 2017

CoDaS-HEP 2018

CoDaS-HEP 2019

CoDaS-HEP 2022

Summer research internship program for undergraduates, focused on the **underrepresentation of women and minoritized students** in STEM fields such as Physics. New in 2022 (Meenakshi).

The program cover areas in instrumentation, technology, and computing projects, connecting students with mentors.

Leverages HSF software training material as part of using computational tools and data-science methods by analyzing data from CMS.

DOE CompHEP Traineeships - New in Fall, 2022

2 year graduate traineeships with dedicated coursework & training modules
(+ curriculum development) and R&D projects in collaboration with lab staff

- **Training to Advance Computational High Energy Physics in the Exascale Era (TAC-HEP)** - Wisconsin, UMass-Amherst, Princeton, FNAL, BNL

- **Western Advanced Training for Computational High-Energy Physics (WATCHEP)** - UC-Santa Cruz, Oregon State University, UC-Berkeley, UC-Irvine, UC-San Diego, U.Washington, BNL, FNAL, LBNL

- **Chicagoland Computational Traineeship in High Energy and Particle Physics (C2 The P2)** - UIC, NIU, FNAL, ANL

Significant limitation to recruit talent: only US citizens?

Multiple software, computing, data science and AI/ML programs aiming at different career stages

- Undergraduate summer programs: US-CMS PURSUE program, US-ATLAS SUPER program, IRIS-HEP Fellows program 🏅🏅
- HSF/IRIS-HEP Training activities (materials and events) - 1600 students and 50 educators in the past few years 🏅
- A3D3 PostBac program 🏅
- Summer schools: IAIFI, CoDaS-HEP 🏅
- US-ATLAS and US-CMS postdoc R&D programs 🏅🏅
- DOE CompHEP Traineeship projects: TAC-HEP, WATCHEP, C2-The-P2 🏅
- International: Fellows/mentoring in NSF-funded HSF-India project, HSF Google summer of code, CERN summer students, etc. 🏅🏅🟧

Workforce Development Coordination Opportunities

- <u>We have all of the elements of a **training pipeline** from undergraduate through postdocs, including opportunities for "cohort building"</u>

- We have a growing **central repository of shared training curriculum materials** (HSF Training area) - basic material is now well covered, and several projects will be developing more advanced material (IRIS-HEP/HSF, DOE CompHEP Traineeships)

- **Mentoring:** that pipeline can be connected to the R&D projects in **integrated teams** with computing professionals (e.g. GPU expertise)

- Equally an opportunity to **coordinate on DEI questions** (e.g. managing Code of Conduct violations), broadening participation, mentoring of mentors, effective training and evaluation.

- A de-facto implementation of Snowmass Community Engagement Frontier CEF01 Recommendation 6 (regarding lab/university collaboration on creating such pipelines) for software/computing

Wider Research Software Collaborations

Wider Research Software Collaborations

Are a platform for doing bigger things, including impact beyond HEP

Wider Research Software Collaborations

Are a platform for doing bigger things, including impact beyond HEP

Recognize that software is both a type of intellectual product of our research and a key "infrastructure" spanning projects/experiments.

HEP Software Foundation (HSF)

The HSF (http://hepsoftwarefoundation.org) was created in early 2015 as a means for organizing our community to address the software challenges of future projects such as the HL-HLC. The HSF has the following objectives:

- Catalyze new common projects
- Promote commonality and collaboration in new developments to make the most of limited resources
- Provide a framework for attracting effort and support to S&C common projects (new resources!)
- Provide a structure to set priorities and goals for the work

The HSF is an unfunded, volunteer organization, with a "bottoms-up" structure.

Community White Paper

January 2017
UCSD

June 2017
Annecy

"*The result: a Programme of Work for the field as a whole, a multifaceted approach to addressing growing computing needs on the basis of existing or emerging hardware.*"

Eckhard Elsen (CERN Director of Research and Computing), editorial published with CWP/Roadmap

Many workshops, involving a diverse group
- International participants
- Computing Management from the Experiments and Labs
- Individuals interested in the problems
- Members of other compute intensive scientific endeavors
- Members of Industry
- http://s2i2-hep.org/
- https://hepsoftwarefoundation.org/

Individual Papers on the arXiv:

Careers & Training, Conditions Data, DOMA, Data Analysis & Interpretation, Data and Software Preservation, Detector Simulation, Event/Data Processing Frameworks, Facilities and Distributed Computing, Machine Learning, Physics Generators, Security, Software Development, Deployment, Validation, Software Trigger and Event Reconstruction, Visualization

Community White Paper & the Strategic Plan

arXiv 1712.06982   arXiv 1712.06592

IRIS-HEP

NSF
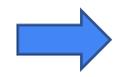
"Coordinated Ecosystem" Workshops (2017, 2019, 2022)

2022 Workshop: https://indico.cern.ch/event/1203733/timetable/

Organized by IRIS-HEP as a "blueprint" workshop (in Washington DC) with relevant U.S. experiment and R&D project leads to discuss:

1. How does the **ensemble of US Software R&D efforts fit togethe**r to implement the HL-LHC Software/Computing roadmap and meet the challenges of the HL-LHC?
2. How do the US Software **R&D efforts collaborate with each other and with international efforts**?
3. How should the **US R&D efforts be structured and evolved** in the coming years in order to achieve our goals between now and the HL-LHC era?
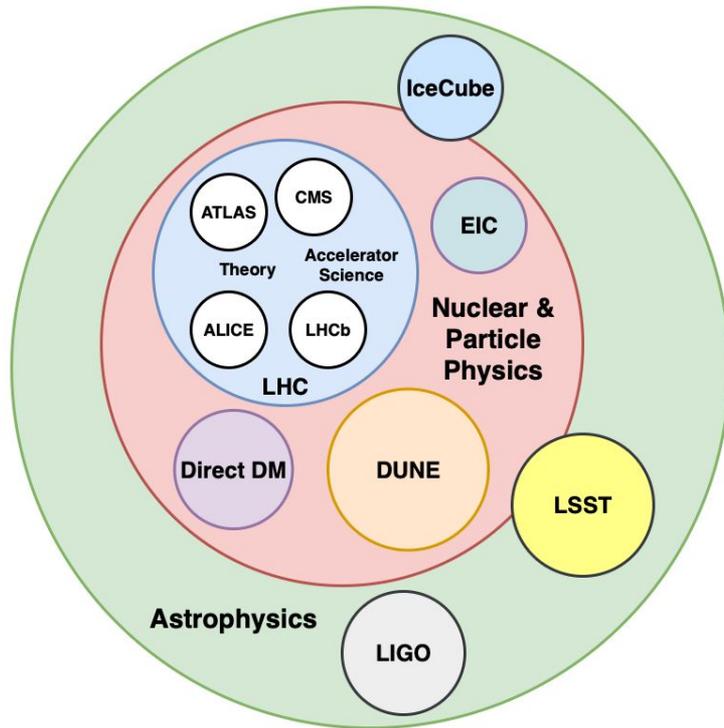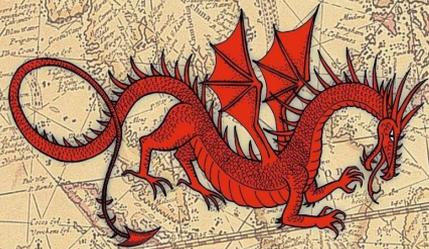
Evolving Ecosystem Workshops
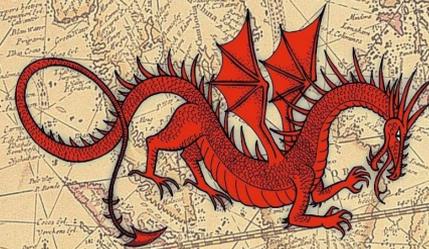

2022

2017

2019

Although initially focused on the HL-LHC Science Driver and needs, participation has grown to include:

- Experiments included DUNE, Dark matter, EIC, Vera Rubin.
- DOE Lab participation included BNL, LBNL, FNAL, ORNL, SLAC.
- NSF AI Institutes.
- … and more!

Strategies for Making an Impact Beyond HEP

Python in HEP (PyHEP) Ecosystem → Scikit-HEP

Our scientific Python development vision/ecosystem

Scientific Python / PyData vision/ecosystem

Jake Vanderplas 2017 keynote-style Python ecosystem, but with layer labeling a la "Array programming with NumPy", *Nature* , 585, 357-362 (2020)

<u>Intentional engagement with the larger data science community</u>: Awkward Array has users in health care, genomics, data science, radio astronomy, oceanography, …. IPython (Jupyter) came from Fernando Perez as particle physics PhD student.

In practice, any pretensions we have to "more general applicability" do not matter unless we build the developer and scientific relationships to make that happen. "If we build it, they will come" is insufficient to deliver widely used, impactful and sustainable software.

# Intentional strategies to tie everything together

Grow connections to multidisciplinary **University "Research Software Engineer" groups** with relevant expertise - already leveraged in some of our R&D projects - and for our expertise to make an impact of work beyond HEP, including career opportunities for the students/postdocs.

Greater engagement with **data science institutes and initiatives** at the universities to build true multi-disciplinary connections to Computer Science, Statistics/Applied Math, etc.
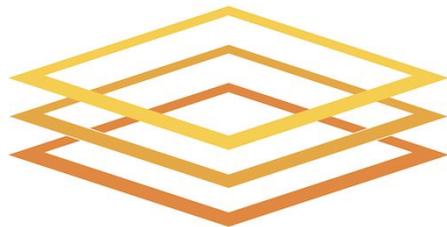
For example, qualitative difference in backgrounds of Summer Fellows applications:
    IRIS-HEP Fellows: >80% Physics (advertised on HEP lists)
    Princeton (general) RSE Fellows: >90% Computer Science (advertised via US-RSE)

Platforms for engaging/impacting the larger research community
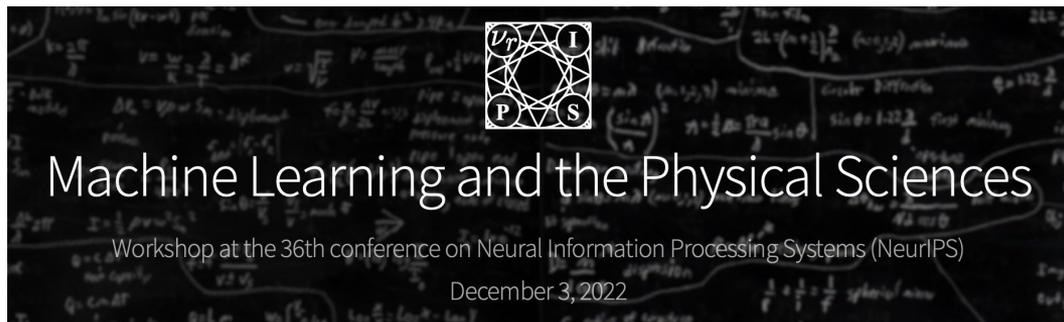
Machine Learning and the Physical Sciences

Workshop at the 36th conference on Neural Information Processing Systems (NeurIPS)

December 3, 2022

https://ml4physicalsciences.github.io/2022/

SciPy2022

Scientific Computing with Python
Austin, TX • July 11 - July 17

Track & Mini Symposia Chairs

Physics and Astronomy

Matt Craig

Mike Zingale

Matthew Feickert

OSG Consortium

iris hep

Institute for Research & Innovation
in Software for High Energy Physics

https://iris-hep.org/

Last thoughts

Our community "superpower" is that we know how to build distributed collaborations → We can build science-driven collaborations around things other than single experiments, in particular software, computing and data science.

The evolution of "workforce development" (software/computing/data-science) is now at a point where we can consider a larger "federated" collaboration over many funded projects to build a pipeline spanning career stages: emphasis on sustainability, scalability, (evolving) training scope & diversity/equity/inclusion

The labs are our backbone and our strength, but for engagement and impact "beyond HEP" there is a huge opportunity in University-led, but Lab-collaborative, projects which connect to other scientific domains through the growing research software engineering and data science activities on campuses across the US