Strategies and needs for training

Sudhir Malik
Univ. of Puerto Rico Mayaguez

Peter Elmer
Princeton University

Latin American Workshop on Software and Computing challenges in High-Energy Particle Physics (LAWSCHEP 2019), Mexico City, 20-23 Nov, 2019

Current HEP collaborations are huge enterprises

- **Big Collaborations, many, continents, countries**
  - **CMS- 5000 particle physicists, engineers, technicians, students and support staff from 200 institutes in 50 countries (September 2019)**
  - **ATLAS** - 3000 scientists from 174 institutes in 38 countries work on the ATLAS experiment (February 2012)What are some examples of training from individual experiments
  - **LHCb** - 700 scientists from 66 different institutes and universities make up the LHCb collaboration (October 2013).
  - **ALICE** - 1000 scientists from over 100 physics institutes in 30 countries.
  - **DUNE** - 1000 collaborators from over 180 institutions in over 30 countries plus CERN
  - ***Others***
    - (Past - **DZero** 540 members, 90 institutions in 18 **countries**. **CDF** - 600 physicists, including 30 US institutions and labs 12 countries
    - (**LIGO** - 1200 scientists from over 100 institutions in 18 different countries)

Sun never sets on a collaboration

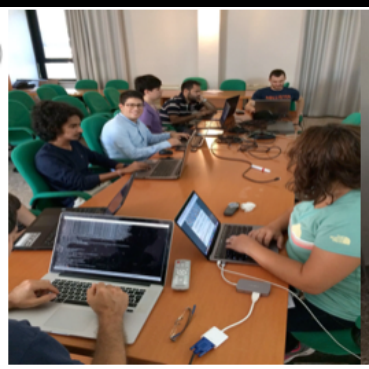
New Paradigm for Users

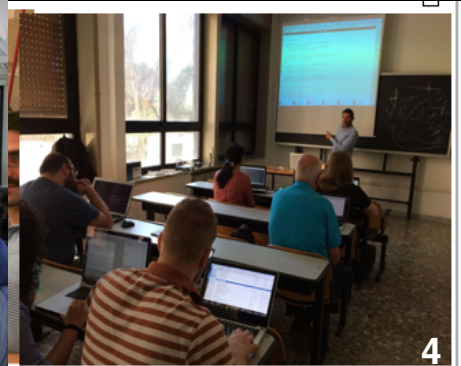
- In the setting of Big Collaborations
  - Jump start physics - computing tools and physics analysis strongly intertwined
  - meet experimental challenge
- Enormous resources, manpower
- Long life span of the experiment ~ 30 years
- Enormous data rate - ~ 10,000 copies/sec of *Encyclopaedia Britannica*
- Most users not resident at host laboratory
  - Possible financial and logistic constraints to be at CERN
- Highly distributed environment for
  - Computing (Grid)
  - Physics analysis
- Physics and Computing Support
  - Should reach every user wherever they may be
  - Be taken up in organized and central way

Need for an organized training

Training in Experiments (hands-on)

- **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 for sustainability
- **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)



CMS Training example

CMS Data Analysis Schools

PROGRAM	GOAL	TIMETABLE	TOPICS	MATERIAL
Pre-School Exercises	Beginners to Experienced who want to jump start Physics Analysis	The preparatory Exercises start a month before the school, Exercises prepared and checked before that by a team of facilitators	**CMS Basics - software, access to code data, run Grid jobs, Github, ROOT, Python, PyRoot, Fitting**	twikis, espace to answer questions
5-Day Hands on sessions, Students work in class-like settings with student/teacher ration of ~5:1 **Held at several places typically 2-3 times per year - Fermilab, Taiwan, CERN, Italy, DESY, India, Korea** **Survey and Feedback from users**	Common interface to the algorithms developed by physics objects groups, single entry point to information associated to physics objects, Approved algorithms and sensible default, configure ones analysis in python language **1000 students trained so far**	Lectures - 1/2 day	**LHC Machine, CMS Physics, CMS Detector, CMS Software Tools, Physics Analysis Design**	Indico (talks)
		Short Exercises - 2 days Each Exercise - 2hrs, Can take up to 6 exercises	**Roostats, Generators, Tracking, Vertexing, Electrons, Muons, Jets, b-tagging, PFlow, Pileup, Event**	hands-on, twikis Indico (talks)
		Long Exercises - 2.5 days Physics Analysis: 6-8 students per analysis	**Examples - Dark matter (with Higgs boson to four-leptons), Mono-Photons, B2G Boosted Z'->ttbar semileptonic, SUSY hadronic, Z to tau-tau, Top mass measurement etc**	hands-on, twikis Indico (talks),
		Mini-symposium	The student groups present their work and compete for the "first prize" judged by panel of senior CMS physicists	Indico talks




ATLAS Software Training Program

NAME	AUDIENCE	PROGRAM	TOPICS	MATERIAL
Software Tutorial	Beginner (+Refresher)	5 days @CERN 4x/year (since 2012 900 people trained)	ATLAS-specific software (formats, frameworks, analysis model), the Grid, git, cmake, statistical tools	Indico (talks) twiki git-style docs recorded videos
Migration Tutorials	Experienced	1-2 days (as needed)	Examples: svn→git, cmt → cmake, Run 1→Run 2 analysis model	Indico (talks) twiki git-style docs recorded videos
Developer Tutorial	Intermediate/ Advanced	5-days @CERN ~1/year	code quality, writing code in ATLAS, multi-threading	Indico (talks)
Workbooks	Beginner	updated by dedicated responsible person as needed	Computing Physics Analysis Software Developers	twiki

Career Guidance Activities

- CMS started career guidance early on and that model was picked
- Career guidance sessions are part of events at CERN and FNAL

CMS Career Guidance (non-academia)

- CMS Collaboration Board endorsed the composition of the **CMS Career Committee** (members from all geographic regions) in 2012 - four working groups:
  - **Networking** with CMS alumni and to the non-academic job market in general,
  - **Collecting and providing information** on academic jobs
  - Reflecting on the **recognition of individual achievements**
  - Organizing **information sessions** on career related topics

- Our skills, especially computing, are much sought after in industry
- Committee organizes career events **bringing CMS alumni from companies** in a diverse range of fields (industry, finance, IT) at CERN, Fermilab etc.
- Lately this idea was recognized liked by other LHC experiments and since then we have been **organizing career events with ALICE, ATLAS, CMS and LHCb**

  - https://indico.cern.ch/event/561880
  - https://indico.cern.ch/event/440616/

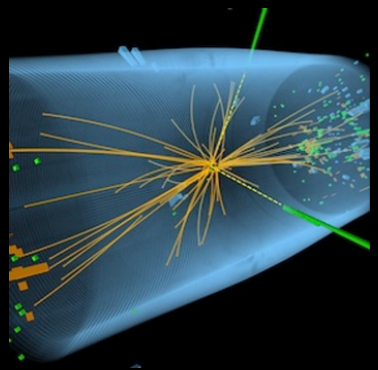
- We also maintain a **CMS job twiki** - academic and industry jobs and guidance, highly popular, jobs advertised free of cost, only for HEP community - https://twiki.cern.ch/twiki/bin/view/CMSPublic/JobOpportunities

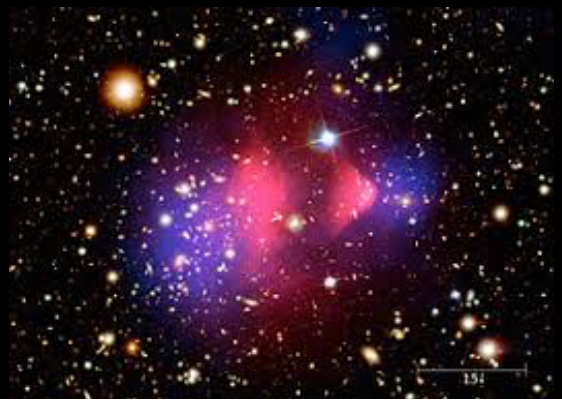
Current Training - Limitations

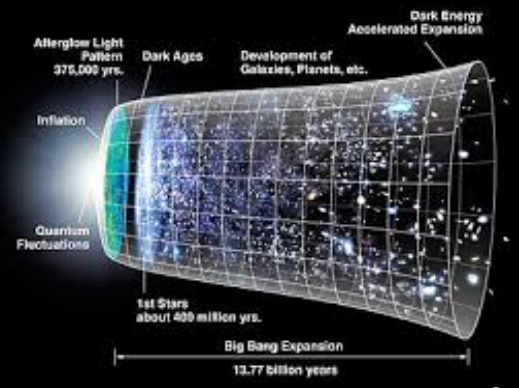
- Training activities today are fragmented and partially redundant
- Each project (experiment, laboratory, etc.) is left to reinvent most aspects of research software training from scratch, the result is duplicative and often incomplete
- Most training activities are carried out “locally”, with specific objectives in the context of a specific experiment, university or laboratory
- The modest effort devoted to training is not always positioned for maximum impact
- The resulting activities are also quite difficult to sustain over time
- They are 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

Science Drivers

- Use the Higgs boson as a new tool for discovery
- Pursue the physics associated with neutrino mass
- Identify the new physics of dark matter
- Understand cosmic acceleration: dark matter and inflation
- Explore the unknown: new particles, interactions, and physical principles
- Bases of operation of many current experiments as well as the design of the large, next-generation, facilities in HEP HL-LHC, LBNF at Fermilab, Super KEK-B and experiments - ALICE, ATLAS, CMS, LHCb, DUNE, Belle- II

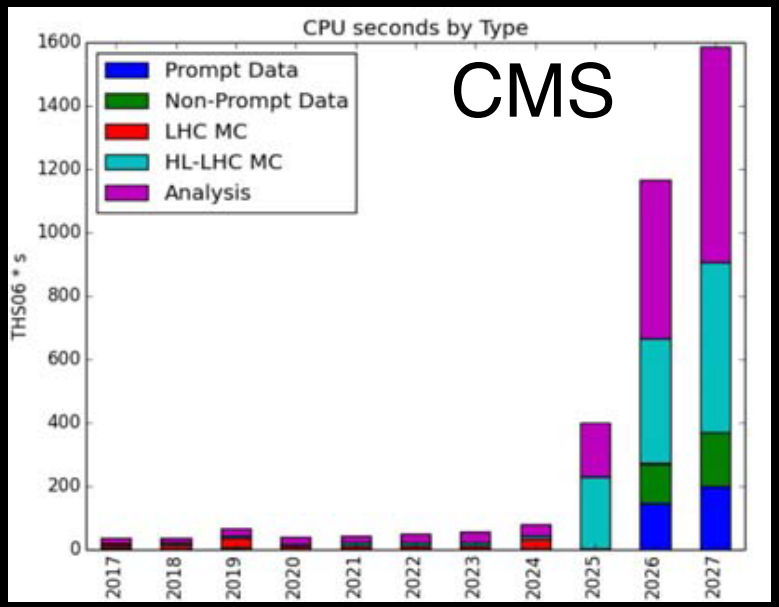


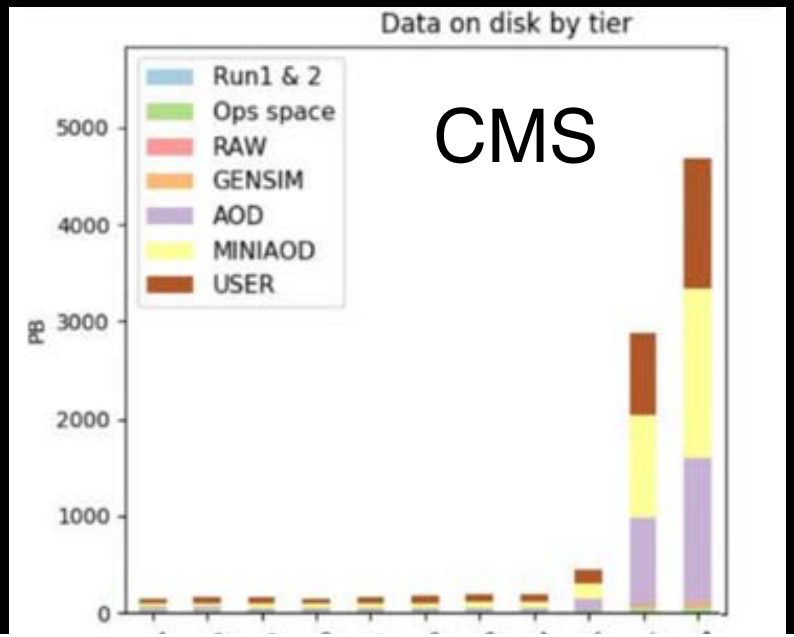




Computing Challenges in HEP

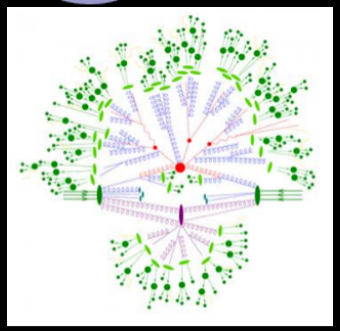
- Large data-intensive HEP experiments rely on
  - Significant data storage
  - High throughput computing
  - LHC Experiments - ~170 computing centers, nearly an exabyte of disk and tape storage and 750,000 CPU cores
  - HL-LHC
    - 100 billion proton-proton collisions per year
    - 10x or more computing needs, 100 times the data of (unto 2030s)
    - Other HEP facilities are planning similar increases in data volume





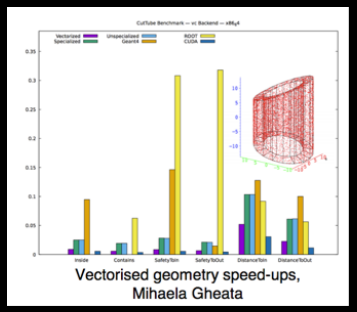
HEP software ecosystem

Software and Physics analysis are intertwined



Physics Event Generators

Data, Software,
Analysis Preservation

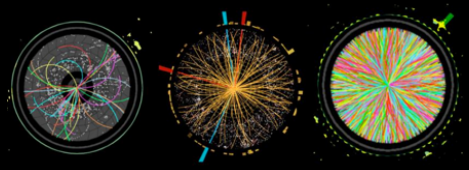


Detector Simulation

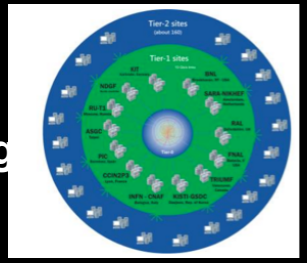
Security

Software Development

Trigger,
Event Reconstruction



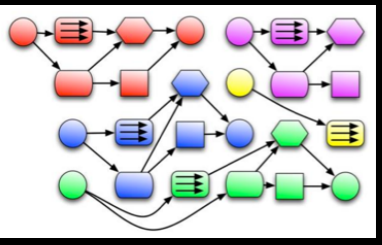
Facilities,
Distributed Computing



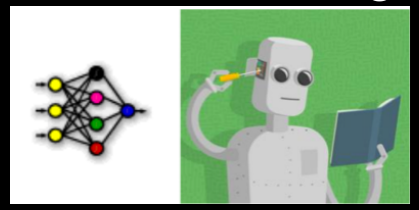
Data Analysis
Interpretation

Visualization

Data Processing
Frameworks



Machine Learning



Data Management
Organisation Access

Sustainability

- Successful evolution of this ecosystem to meet the challenges, requires new tools and a workforce HEP domain knowledge and advanced software skills
- Investment in SW critical to match HL-LHC requirements of “flat budget” scenario
- Investing in training leads to preservation and propagation of knowledge
- Investing 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
- Computation is a central element of 21st century science, and clearer career paths will provide a virtuous cycle of feedback to enhance the vibrancy of the training and workforce development activities

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
- Despite this many individuals do get enthusiastic about training others, but often only in specific career phases and often as a side “hobby” project. How do the activities then scale up?
- Technology evolution means that training materials need to evolve, too. Separating “local” specifics (e.g. computing environments or experiment-specific bits) from generally usable material is important, but doesn’t always happen
- 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

Training - motivation moving forward

- People are the key to successful software
  - Working together across disciplines, experiments, and generations, they are the real *cyberinfrastructure* underlying sustainable software
- Developing, maintaining, and evolving the algorithms and software implementations for HEP experiments will continue for many decades
- The HEP community is currently planning hardware upgrades for the HL-LHC era which will start collecting data 8 or 9 years from now, and then acquire data for at least another decade
- Building the necessary software requires a workforce with a mix of HEP domain knowledge, advanced software skills, and strong connections to other related disciplines
- The investments to grow this workforce must begin today
- The HEP community planning process during 2017 triggered numerous discussion regarding training
- Training is central to building the community skills needed to address the computing challenges of the HL-LHC era
- One key insight is the need to think of training not as a set of individual, disconnected activities, but as part of a larger framework

CWP to IRIS-HEP

In Software for High Energy Physics

A Roadmap for HEP Software and Computing R&D for the 2020s

HEP Software Foundation[1]

Abstract: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, eith
Similarly, it re
manage, proce
for the HL-LH
agree on the so
In this spirit, t
this software u

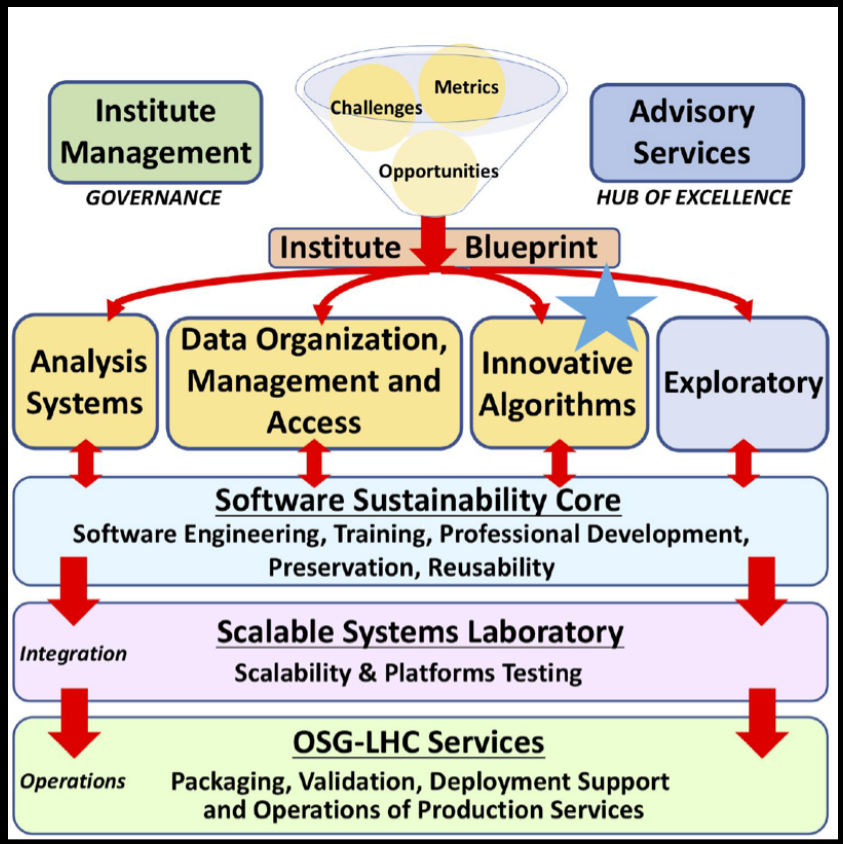
Contents

CWP Roadmap

CWP: https://arxiv.org/pdf/1712.06982.pdf

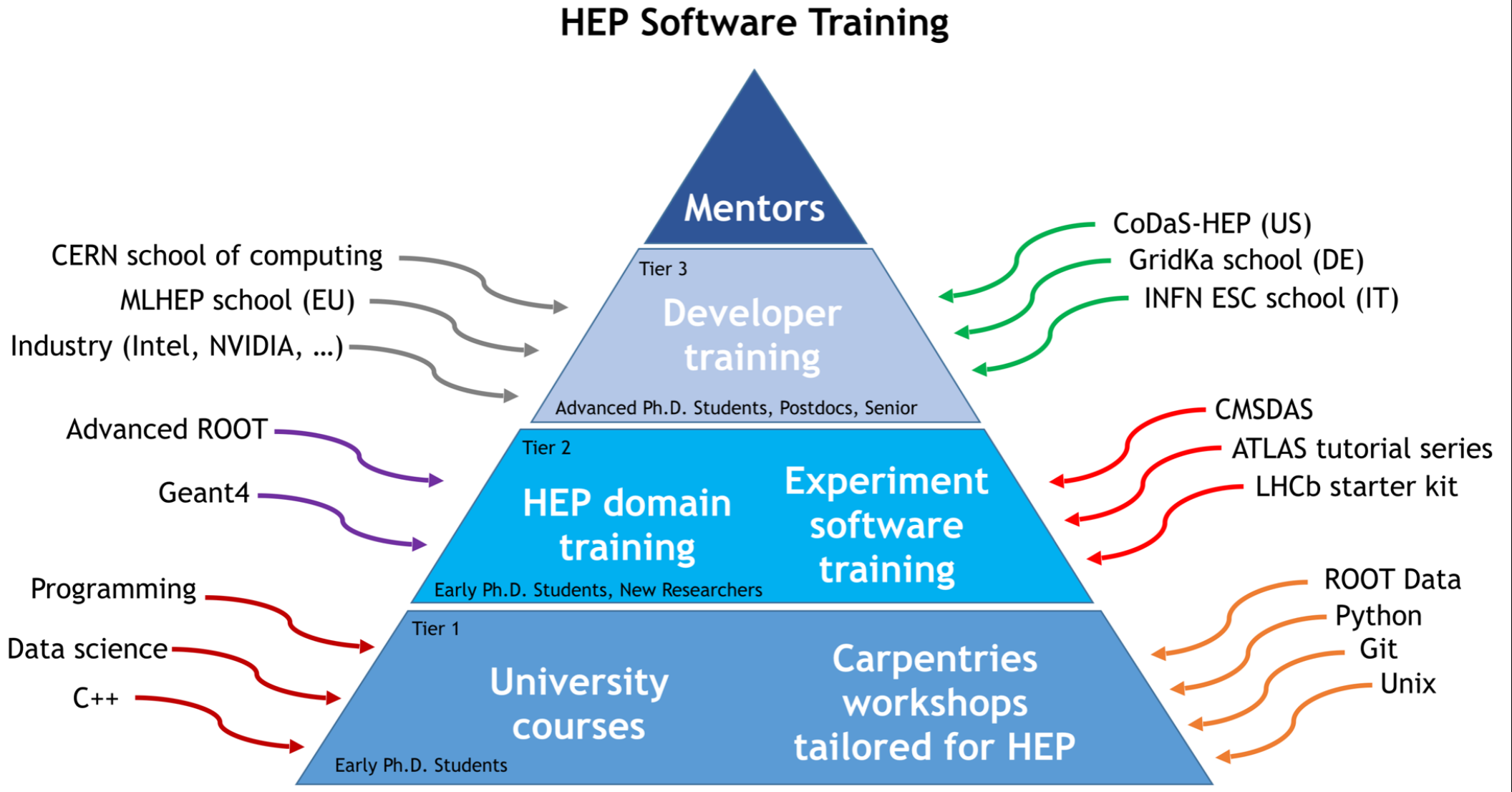
Careers & Training

https://arxiv.org/abs/1807.02875



Gordon's Talk - this workshop

Training Vision



Funded by NSF *via* iris hep *and* *FIRST-HEP*

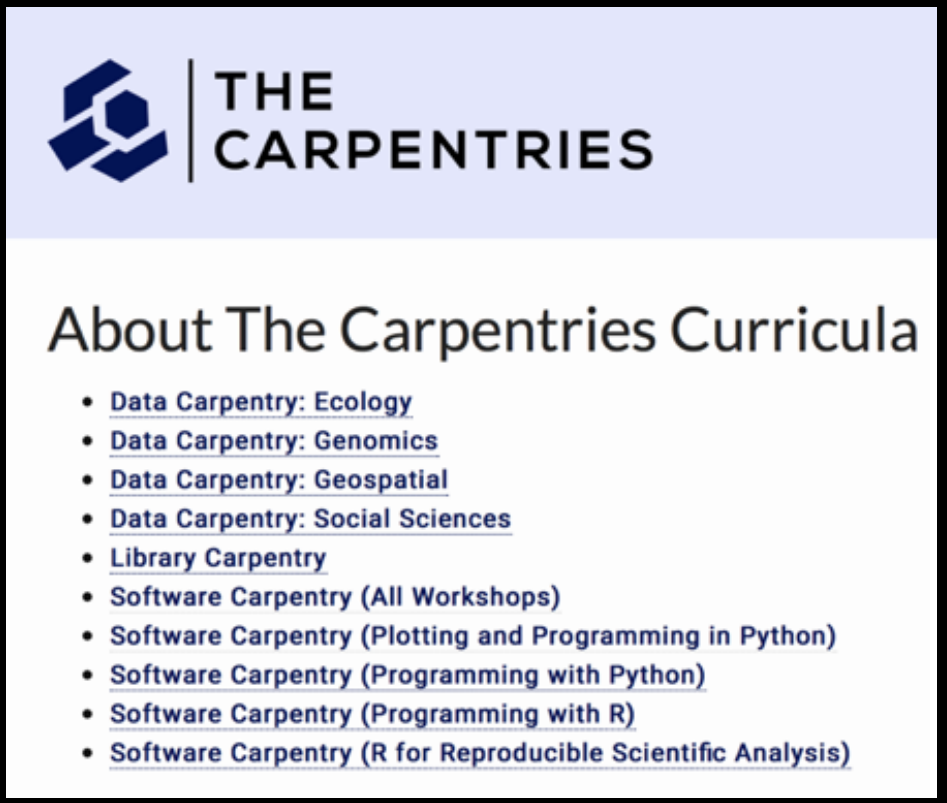
Training Vision - Input

- Establish a *community framework for software training* in order to prepare the scientific and engineering workforce required for the computing challenges of HEP experiments
- For sustainability , *build "users" community* as well as *"developer" community*
- Instead of new curriculum development *leverage and build upon existing material* from the HEP and larger research community
- *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*

Inspiration

https://carpentries.org

https://lhcb.github.io/starterkit/

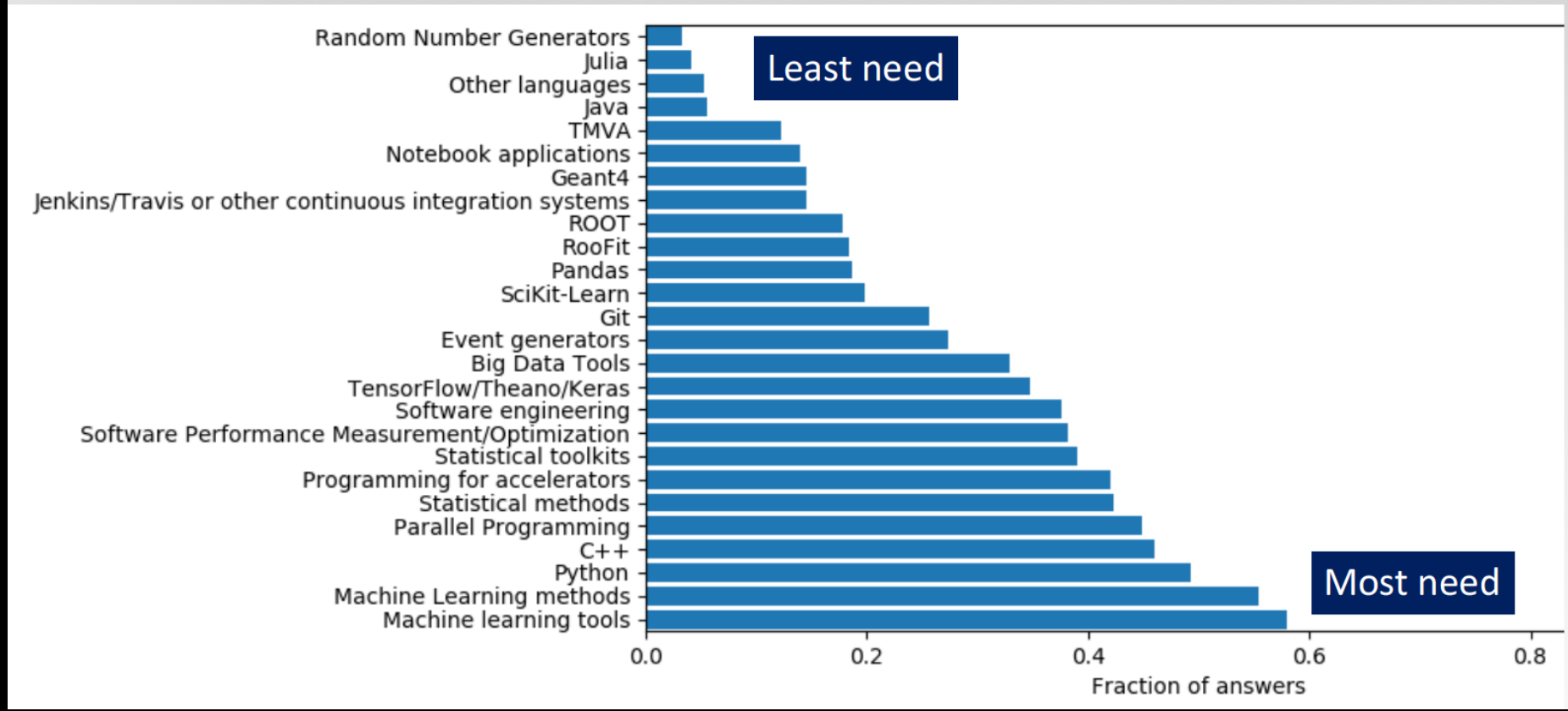


Key insight: thinking of training as a community building exercise. And not only for the “student” participants, but also for the “instructors”.

Training Survey (Jan-Feb 2019)

- Input on training needs and current training practices for various topics in HEP related to software/computing and related software centric areas
- ~ 350 users (mostly postdocs, students, mostly LHC

Areas that need for more training materials/courses to help your research



For more info - HSF Training survey - by David Lange

FIRST-HEP

- Stands for - *Framework for Integrated Research Software Training in High Energy Physics*
- Collaborative proposal (Princeton+UPRM)
  - http://first-hep.org
- **Funding from NSF**
  - Training Workshops/Participant Support/Brain Storming Session

OAC-1829707 OAC-1829729

IRIS-HEP

- "*Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)*" - http://iris-hep.org
  - Sharing vision and co-supporting/organizing FIRST-HEP activities
  - Funding for Training Activities/Participant Support
  - Fellowships to work on Software and Computing Projects
  - Job Opportunities

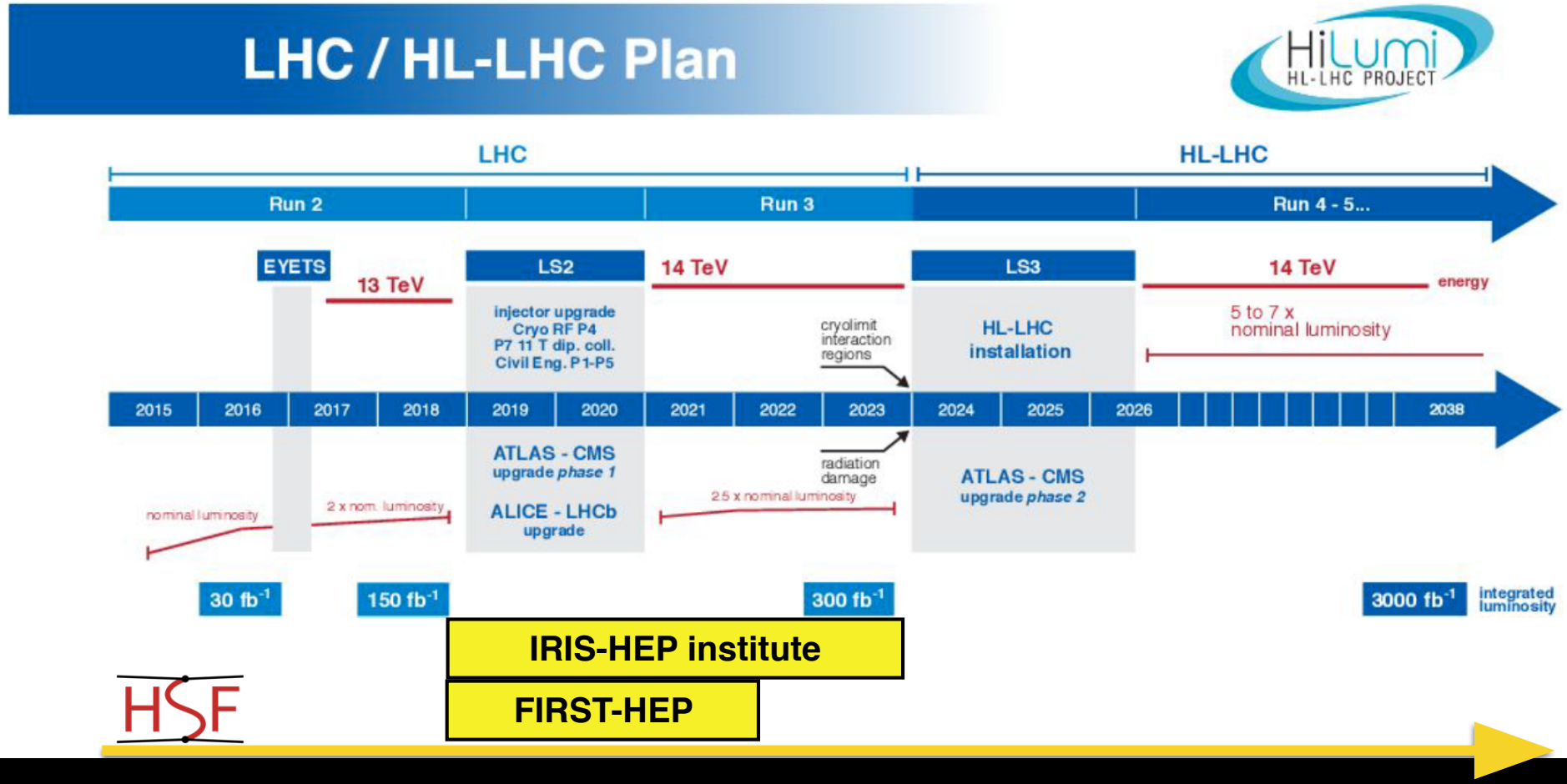
OAC-1836650

International Vision

- The HEP Software Foundation (HSF)
  - formed in 2015 facilitates cooperation and common efforts in High Energy Physics software and computing internationally
  - strong training component to build HEP domain knowledge + advanced software skills+ strong connections to other related disciplines
  - https://hepsoftwarefoundation.org

IRIS-HEP/FIRST-HEP/HSF vs HL-LHC

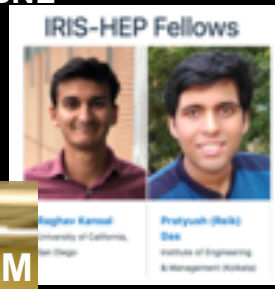
High Luminosity LHC

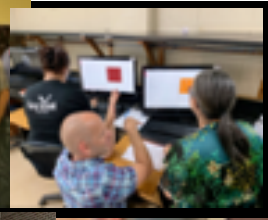


Training for HEP community - highlights

- Software Carpentry Workshops
  - Github/Unix/Python/Plotting
  - Universities, National Labs
  - 180 people trained
  - Female participants highly encouraged
- Outreach
  - Programming for STEM teachers
  - Underrepresented Communities
  - Scientific Software Club at UPRM
  - Female participants highly encouraged
- Careers and Jobs
  - IRIS-HEP Fellows (HEP/non-HEP)
  - Graduates
  - Undergraduates
- Collaboration
  - FIRST-HEP (http://first-hep.org)
  - HEP Software Foundation
    - (https://hepsoftwarefoundation.org/
  - The Carpentries (https://carpentries.org)

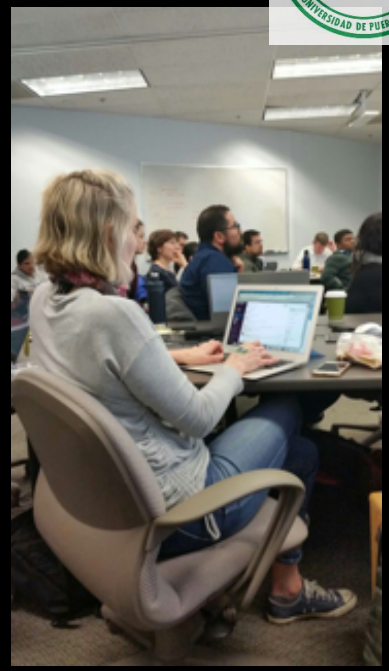
LBNL

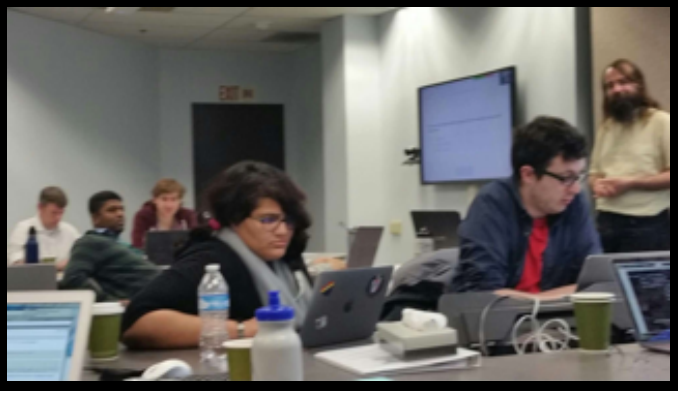




Software Carpentry Workshop (Fermilab)

(1-2 April 2019)





- Agenda: https://indico.fnal.gov/event/20233/
- Version Control with Git
- Python Foundations
- Building Programs with Python
- Data analysis - Numpy, Pandas
- Data analysis Cont. and Graphs
- Advanced Python and PyROOT, uproot
- Post-workshop Survey

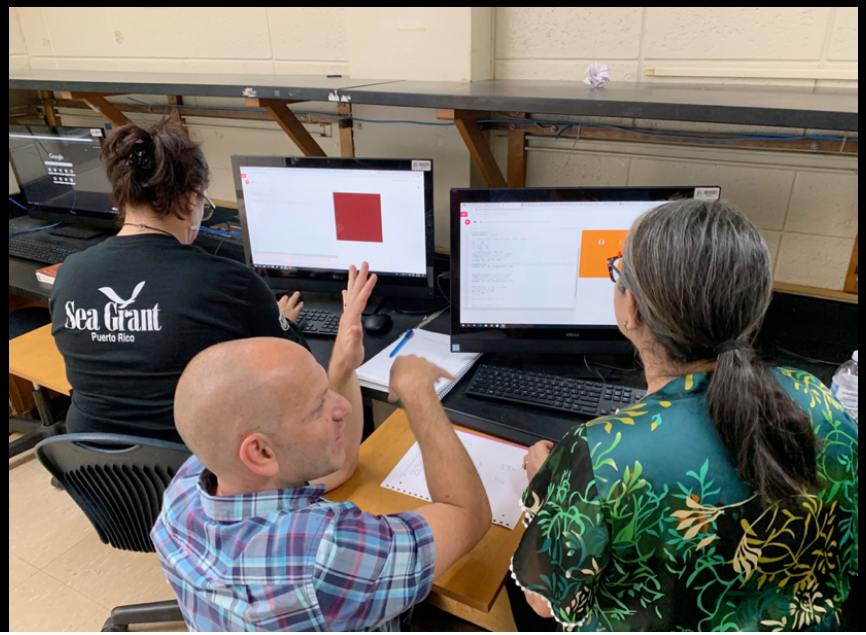
ML Hackathon (Puerto Rico)

(24-26 April, 2019)

https://indico.fnal.gov/event/20233/

- Agenda: https://indico.cern.ch/event/809812
- Talks ML/Data science Applications - Physics, Computer Science, Math, Engineering
- Machine Learning, Deep Learning, ANN
- Hands-on exercises
- Hackathon on LHC Physics

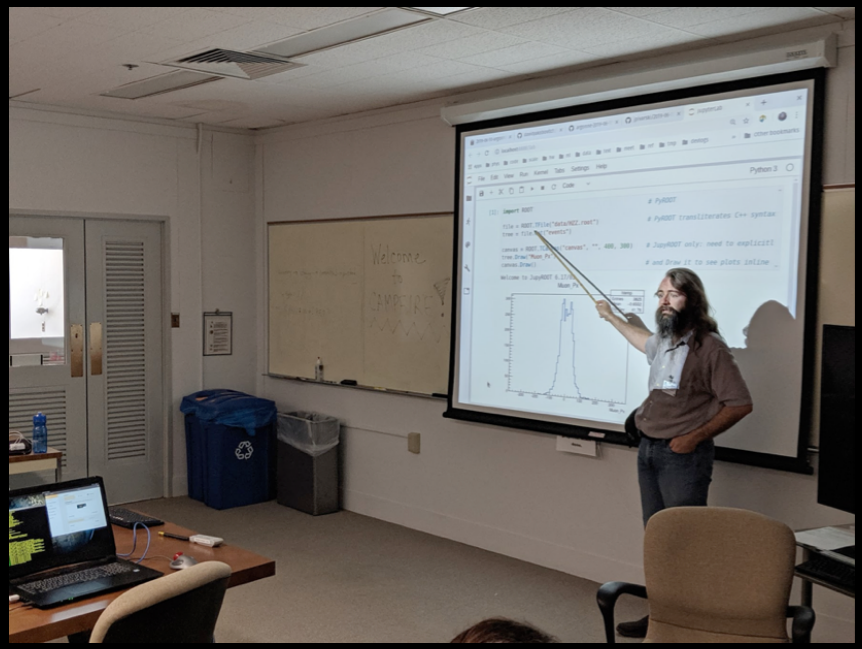
An introduction to programming for STEM teachers (Puerto Rico) (4-5 June, 2019)



- Agenda: https://indico.cern.ch/event/817539
- Introduction to Processing and p5js
- Python and Colab
- Basics of python, Jupyter notebooks, and Colab (hands-on)
- Data analysis with python

FIRST-HEP/ATLAS Training (Argonne)

(10 June, 2019)



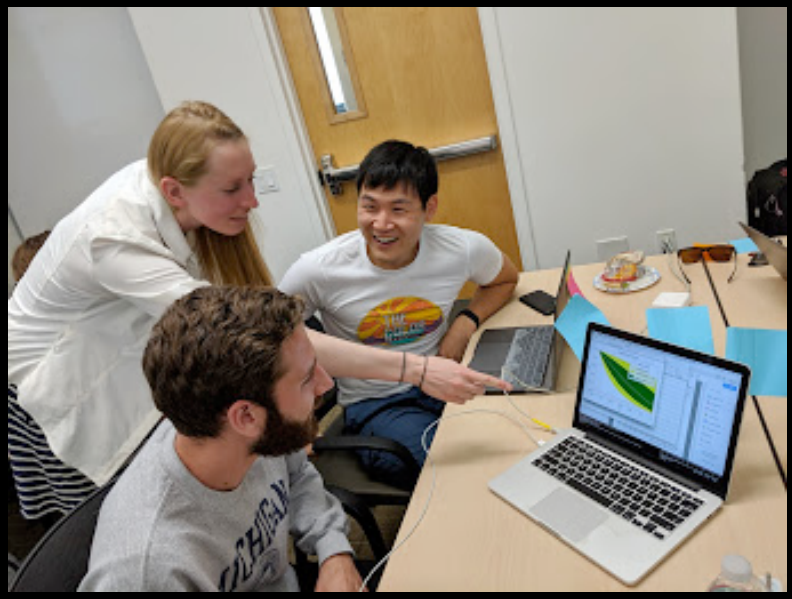
- Agenda: https://indico.cern.ch/event/827231/
- A little Unix
- A lot of Git
- Numpy introduction
- PyROOT
- uproot

3rd CoDaS-HEP School (Princeton)
(22-26 July, 2019)

- Agenda: https://codas-hep.org/
- Parallel Programming
- Scientific Python Ecosystem
- Big Data Tools and Techniques
- Machine Learning
- Practical skills like performance evaluation, use of git

FIRST-HEP/ATLAS Training (LBNL)

(19-21 August, 2019)



- Agenda: https://indico.cern.ch/event/816946/
- Version Control Essentials
- Jupiter
- Build Systems : From gcc to cmake
- Continuous Integration : Why and how?

IRIS-HEP Fellows Program - mentoring

- **Key Insight:** we need to provide incentivized and explicit paths forward for enthusiastic students from the more advanced training schools (ESC/Bertinoro, CoDaS-HEP, MLHEP, etc.) or for people who become engaged with our software projects in other ways

- **Project focused:** bring students into contact with "mentors" to work on a specific, pre-defined project, allowing them to grow their software skills and project experience. The fellow supports, when possible, travel and subsistence for a 3 month extended stays in the mentor's institution

- More info: http://iris-hep.org/fellows.html

Current IRIS-HEP Fellows

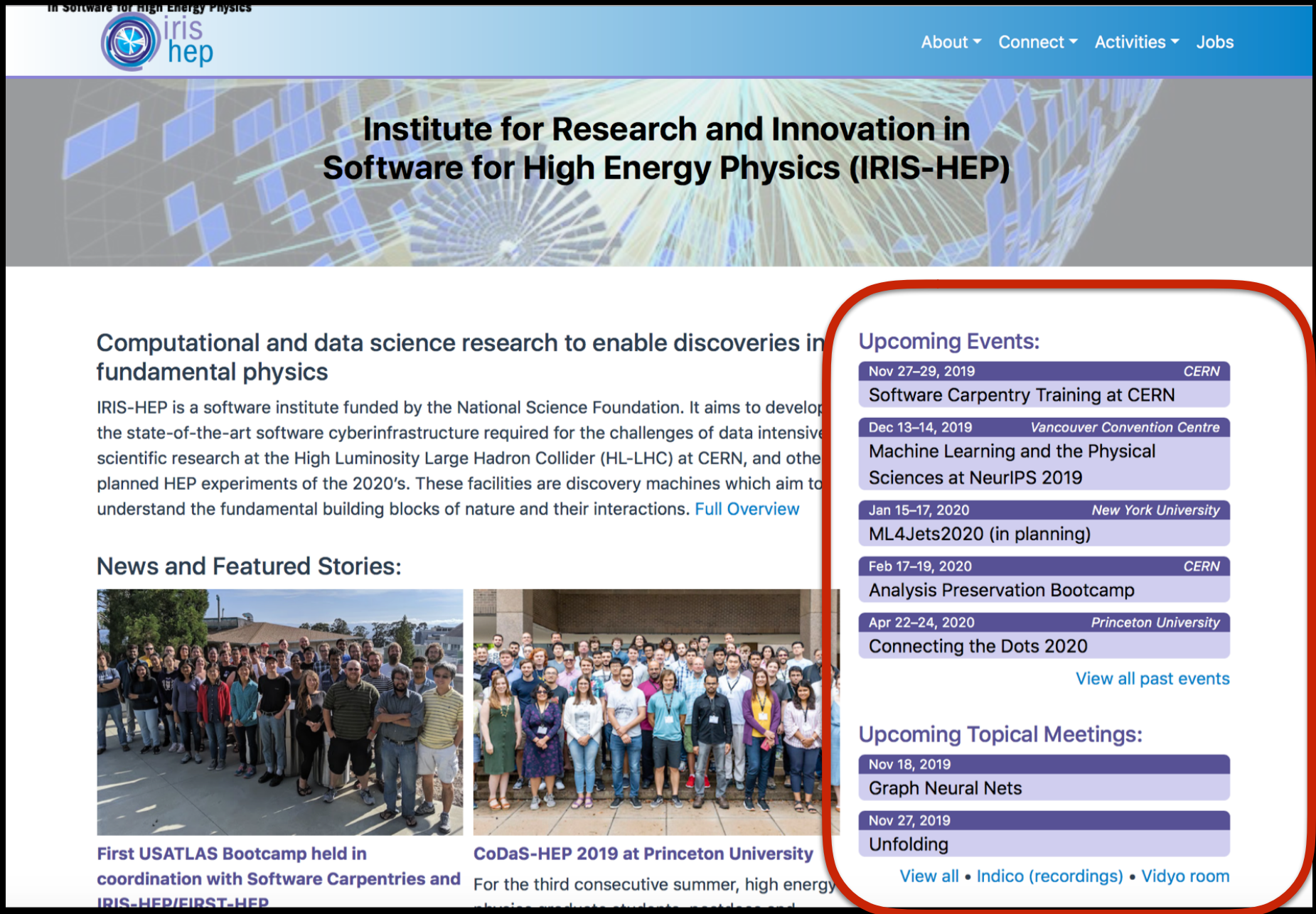
Raghav Kansal
University of California, San Diego

IRIS-HEP Fellow
Jun-Aug 2019

Pratyush (Reik) Das
Institute of Engineering & Management (Kolkata)

IRIS-HEP Fellow
Jun-Sep 2019

IRIS-HEP website



Next Steps

- Training has picked up momentum
- Strong partnership with Software Carpentries: https://software-carpentry.org/
- First training event outside US (at CERN, Nov 2019)
- Early 2020 do a checkpoint on the curricula we are using (including also the LHCb/Alice efforts) and attempt to define some new training modules with The Carpentries that are appropriate for natural sciences and/or engineering students
- Then during 2020 we continue to run similar workshops using the new curriculum and look more closely how to scale them up and integrate them with planned experiment meetings and experiment-specific training activities
- Sometime during 2020, we would also like to hold a workshop or BoF session on the more advanced training schools (ESC, CSC, CoDaS-HEP, …) and a community-wide, federated, approach to facilitating student projects (“mentoring”)

Connecting with IRIS-HEP/HSF

- IRIS-HEP website http://iris-hep.org/
  - Jobs on IRIS-HEP and Collaborating Projects http://iris-hep.org/jobs
  - General public announcement mailing list for IRIS-HEP events, talks, meetings, workshops, opportunities for training and job opportunities (subscribe to)
  - announcements@iris-hep.org

- HSF (HEP Software foundation) - https://hepsoftwarefoundation.org
  - Weekly training meeting hsf-training-wg@googlegroups.com
  - General Information about HSF (subscribe to): hsf-forum@googlegroups.com
  - Discussions and activities in the HEP Software Foundation mailing lists can be found here (General and Dedicated Forums): https://hepsoftwarefoundation.org/forums.html
  - You can contribute https://hepsoftwarefoundation.org/cwp/cwp-working-groups.html
  - HSF Events/Workshops - https://hepsoftwarefoundation.org/events.html

Thank you for the invitation

FIRST-HEP

HSF
HEP Software Foundation