IRIS-HEP Training Challenge

Sudhir Malik
University of Puerto Rico Mayaguez
Training Accomplishments

● **Software modules**
  ○ Basic software curriculum
    - One introductory software training curriculum that serves HEP newcomers
  ○ Intermediate modules, some specific to HEP data analysis
  ○ Survey Software Training (Feb 2019)

● **Training events**
  ○ 13 software training
  ○ 6 in-person, 7 online trainings
  ○ 1000 participants
  ○ Feedback

● **Outreach events**
  ○ Training material for outreach
  ○ 5 workshops (2 in-person and 3 online)
  ○ 75 teachers (Puerto Rico)
  ○ Feedback
Lessons Learnt

- **Technical and Pedagogical lessons for In-person and Virtual Training**
  - In-person and Virtual Trainings have their specific positive and negative aspects

- **Organization**
  - Build Community of individuals - facilitators, learners, instructors, experts and hosts
  - Incentivise facilitators, instructors
  - Core team to support the overall mission of training

- **Documentation**
  - Good documentation of training material is essential
Training Challenge

● **2021-2022 focus**
  ○ Scalability
  ○ Develop new and improve existing training material
    ■ a continuous process
  ○ Scale up efforts
  ○ Expand training

● **2022-2023 focus**
  ○ Sustainability

● **Training Scope**
  ○ Our [HEP training community](#)
  ○ ~2000 in US (Postdocs + PhD students + Undergrads) (not including faculties and scientists)
  ○ ~8000 worldwide (Postdocs + PhD students + Undergrads) (not including faculties and scientists)
    ■ All are linked via some international collaboration (e.g. CMS/ATLAS) or international Lab (e.g. FNAL/CERN)
Training Scope

● **Who are our customers?**
  ○ HEP Postdocs PhD students, Master/Undergrad students
    ■ Some from other areas like Astro/CS
  ○ Located at Universities/ National Labs
    ■ domestic/international

● **What trainings we offer (and they need)**
  ○ HEP Community
  ○ Modules - [Beginner + Intermediate + Advanced](#)
    ■ HEP intensive : Intermediate + Advanced
  ○ For non-HEP, related fields or anyone
    ■ Modules - Beginner
New Training Formats

- To scale training efforts we develop new formats while working on core format
- Our core format - we organise and teach
  - We know how to do this
  - In-person, Online
  - All trainings so far are in this format
  - To scale up, need to expand to other formats
  - We fund instructors to travel and teach
- DIY (Do-it-yourself)
  - Minimal help from us (no expense involved), using training material
  - In-person, Online
- Asynchronous (Anytime/Anywhere)
  - Flipped classroom style, Coursera type - small professional videos ~10 min, then Q/A
    - We can use current material to extend training to this style
Building community of Training Instructors

● This is key to scalability and sustainability, key to success

● For development of training material and training instructors:
  ○ Need to shift from current volunteer basis style to more committed style
    ■ Creative work of module development and its teaching (recording videos etc) is one week work of time,
    ■ incentivise making curriculum lessons, upgrade and participation
    ■ Financial rewards for the development
    ■ A detailed plan called “Pay to Teach and Learn”,
      ● [https://docs.google.com/document/d/1Bcl0iS_SWsQdeYZAt02cilDtW-ATT4WapOhuIPx2B7M/edit#](https://docs.google.com/document/d/1Bcl0iS_SWsQdeYZAt02cilDtW-ATT4WapOhuIPx2B7M/edit#)
  ○ Involve time of IRIS-HEP Fellows
    ● Training is everyone’s responsibility towards the community
    ● Collaborate with US and non-US projects and communities ([HSE, SWIFT-HEP](https://documentation.hepforge.org/))
  ○ Annual awards for developing material and for doing training
    ■ We have detector building awards in HEP but none for Software development and training
Resources Needed to reach the target

- How many FTEs needed for the trainings per year during scalability phase and thereafter sustainability
  - Number of instructors
- Start up costs
  - Paying instructors for materials (online or in-person training)
  - Cost of making Videos
Plan 2021- 2022 (Scalability)

● Develop new and improve existing training material
  ○ Brainstorm sessions for experts

● Focus on building:
  ○ Community of individuals - facilitators, learners, instructors, experts and hosts
  ○ Core team to support the overall mission of training
  ○ Expand collaboration to related communities (Nuclear Physics, Neutrino etc.)

● The auto-solution to above is to increase and expand our training
  ○ More online workshops for beginners (2-monthly)
    ■ Evaluate curriculum receptibility
  ○ Build a course around basic curriculum for HEP beginners
    ■ Give course credits/certificate as incentive
  ○ More in-person workshops: Intermediate/advanced (3-monthly)
    ■ Evaluate curriculum receptibility
    ■ Build a course around that for advance HEP users
  ○ Bootcamps and brainstorming among experts (4-monthly)
Plan 2022-2023 (Sustainability)

- **Long term training model**
  - Minimal set of people are needed to keep the training infrastructure running
  - Identify what are additional costs for additional events
  - Explore Long-term Financial Model

- **Build regional and local capacity**
  - Empower sustainable HEP communities
    - Creating local mentorship and leadership (guided and supported by the core team)
    - Engage HEP labs and R1 universities to achieve this goal.

- **Mechanism of feedback**
  - from our communities and improve as we scale

- **Opportunities to grow professionally**
  - have career paths for core team and volunteers

- **Equity, diversity, inclusion and accessibility**
  - participation across HEP communities, under-resourced institutions, communities in different geographical regions, serve as a role model

- **Carpentry workshops a core offering across Physics departments**
## Timeline

<table>
<thead>
<tr>
<th>Month</th>
<th>Event Description</th>
<th>Repeat Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 2021</td>
<td>Workshop for beginners</td>
<td>Repeat every 2 months</td>
</tr>
<tr>
<td>September 2021</td>
<td>Advance Workshop (intermediate/advanced)</td>
<td>Repeat every 3 months</td>
</tr>
<tr>
<td>August 2021</td>
<td>Bootcamps and brainstorming among experts</td>
<td>Repeat every 4 months</td>
</tr>
<tr>
<td>December 2021</td>
<td>Video recordings ready for Beginner’s training for DIY</td>
<td>To be used for DIY and Asynchronous style (Beginner’s)</td>
</tr>
<tr>
<td>January 2022</td>
<td>First DIY mode training</td>
<td>First DIY mode training</td>
</tr>
<tr>
<td>January 2022</td>
<td>Advertise Asynchronous material</td>
<td>Continuous access, monitor people watching the videos</td>
</tr>
</tbody>
</table>
Accomplishment Outreach

- Develop new and improve **existing material**
  - So far events are in PR
  - One in-person and 4 online
  - Mostly we taught what we think is useful (Python via Google Colabs, Physics problems, HEP data preview, CMS Open Data)
  - [Arduino programming](#) workshop next month is on teacher’s demand (helps in their Robot activities)
  - UPRM students taught students at CROEM how to apply programming to their Astronomy course
  - High School teacher (from CROEM school) in Puerto Rico helped reach out to Physics teachers
    - Facebook, Society of Physics teachers
Outreach Challenge 2021-2023

- **Focus on more participation**
  - Need a core group of committed people
    - Key to build new material and scale activities
  - Engaging Quarknet organisers
  - Identify and expand collaboration with other HEP universities who are doing HEP outreach
  - Survey of what teachers would like to be taught that is related with their education curriculum
    - To engage teachers interest

- **Outreach workshops**
  - Have 10 HEP universities do outreach once a year in their community with us
    - Winter or Summer
    - 20 workshops in 2 years
    - Teachers are mostly available for a week at the end of their semesters or summer
  - Develop short video modules for outreach material for High School teachers and students to learn our software anytime (in line with software training)
    - Mostly to supplement workshops (in-person or online)
Metrics/Milestones

- How many basic trainings per year?
- How many intermediate training per year trainings per year?
- How many people trained?
- How many mentors added?
- How many universities/institutions participated in hosting trainings
Back up
# Software Training Events

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Name</th>
<th>Level</th>
<th>Participants/Tutors</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021</td>
<td>Feb</td>
<td>GitHub CI/CD Training</td>
<td>(virtual) - Basic</td>
<td>200/10</td>
</tr>
<tr>
<td>2020</td>
<td>Nov</td>
<td>ML + GPU Training</td>
<td>(virtual) - Intermediate</td>
<td>40/7</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>Virtual Docker Training</td>
<td>(virtual) - intermediate</td>
<td>173/15</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>Virtual Pipelines Training</td>
<td>(virtual) - Intermediate</td>
<td>250/15</td>
</tr>
<tr>
<td></td>
<td>Feb</td>
<td>Analysis Preservation Bootcamp</td>
<td>(virtual) - Intermediate</td>
<td>70/12</td>
</tr>
<tr>
<td></td>
<td>Jan-Jul</td>
<td>CSU Summer Student Computing/Analysis Training 2020</td>
<td>(virtual) - Basic</td>
<td>bi-weekly</td>
</tr>
<tr>
<td>2019</td>
<td>Nov</td>
<td>Software Carpentry at CERN</td>
<td>- B</td>
<td>60/5</td>
</tr>
<tr>
<td></td>
<td>Aug</td>
<td>FIRST-HEP/ATLAS Training (LBNL)</td>
<td>- I</td>
<td>40/8</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>CoDaS School at Princeton University</td>
<td>- Intermediate</td>
<td>50/10</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>OSG User School University of Wisconsin Madison</td>
<td>- ??</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>FIRST-HEP/ATLAS Training (Argonne)</td>
<td>- Intermediate</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>April</td>
<td>Software Carpentry Workshop - Fermilab</td>
<td>- Basic</td>
<td>25/5</td>
</tr>
<tr>
<td>Year</td>
<td>Month</td>
<td>Name</td>
<td>Participants/Tutors</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-----------------------------------------------------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>Feb</td>
<td>Machine Learning Basics for STEM teachers</td>
<td>8/3</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>July</td>
<td>Data Analysis for STEM teachers</td>
<td>16/3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>Data Camp for STEM teachers</td>
<td>11/3</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>June</td>
<td>An introduction to programming for STEM teachers</td>
<td>16/1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>April</td>
<td>Machine Learning Workshop/Hackathon</td>
<td>25/1</td>
<td></td>
</tr>
</tbody>
</table>
Training Information

- **Training events**: [https://indico.cern.ch/category/11386/](https://indico.cern.ch/category/11386/)
- **Material**: All the training modules developed so far resides: [https://hepsoftwarefoundation.org/training/curriculum.html](https://hepsoftwarefoundation.org/training/curriculum.html)
- **Community**: Our training community is listed here: [https://hepsoftwarefoundation.org/training/community.html](https://hepsoftwarefoundation.org/training/community.html)
- **Procedure**: how to request and organize a training: [https://hepsoftwarefoundation.org/training/howto-event.html](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/](https://indico.cern.ch/event/889665/)
In person training (lessons learnt)

- 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 (lessons learnt)

- Covid Enforced
- Attendance: few \textit{hundred}
- Positives
  - Broader reach: >100 registrants possible
  - 2 times greater likelihood to participate
  - No travel costs $\rightarrow$ 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
# Current Curriculum modules

## Beginner level

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Status</th>
<th>Authors</th>
<th>Repo</th>
<th>Site/Mate</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Unix Shell</td>
<td>Introduction to the <a href="#">unix command line/shell</a></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSH</td>
<td>Introduction to the Secure Shell (SSH)</td>
<td>❔</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version controlling with git</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced git</td>
<td></td>
<td>❔</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Programming with python</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEP C++ Course</td>
<td>Sebastien Ponce</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic Modern C++</td>
<td></td>
<td>❔</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Build systems: cscope</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Distributed file systems and grid computing**

**ROOT**

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Status</th>
<th>Authors</th>
<th>Repo</th>
<th>Site/Mate</th>
</tr>
</thead>
<tbody>
<tr>
<td>uproot</td>
<td>Reading and writing ROOT files without having to install ROOT.</td>
<td>🗿️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Intermediate

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Status</th>
<th>Authors</th>
<th>Repo</th>
<th>Site/Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel programming</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Docker</td>
<td>Introduction to the <a href="https://docs.docker.com">docker</a> container image system</td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workflows &amp; reproducibility</td>
<td>E.g. yadage and reana</td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machine learning on GPU</td>
<td></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI/CD</td>
<td>Continuous integration and deployment with <a href="https://gitlab.com">gitlab</a></td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI/CD github</td>
<td>Continuous integration and deployment with <a href="https://github.com">github</a> actions</td>
<td>❔</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Advanced

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Status</th>
<th>Authors</th>
<th>Repo</th>
<th>Site/Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation</td>
<td><a href="https://sphinx-doc.org">sphinx</a> , <a href="https://www.doxygen.nl">doxygen</a> , etc.</td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event generation and MC</td>
<td><a href="https://pythia.org">pythia</a> , <a href="https://sherpa.uchicago.edu">sherpa</a> , <a href="https://madgraph.hepforge.org">madgraph</a> , etc.</td>
<td>✔️</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpaka</td>
<td><a href="https://gitlab.com/lhc-aark/jupyter-extensions">alpaka</a> is a header-only C++ abstraction library for accelerator development</td>
<td>❔</td>
<td>authors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Module Status

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)
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 ...)

......
......
Target Community

- 1000 PhDs - ⅓ in Particle and Fields = 300 graduating in two years means 150/year (2011-2012), assume not much change

- CMS has 1037 registered PhD doctoral students, US size is ⅓ means 300 US students, PhD takes 6 years => 50 PhDs from US per year on CMS, assume same from US ATLAS and smaller fraction from ALICE/LHCb => graduating PhDs is 110 PhDs from US-LHC
  - [https://cms.cern/collaboration/people-statistics](https://cms.cern/collaboration/people-statistics)

- CMS has 1906 PhD Physicist (scientists/postdocs) (1569 men, 337 women), 1037 doctoral students (796 men, 240 women), ratio of Physicist to Student is 2:1 => CERN (CMS~ATLAS=3*ALICE/LHCb plus) has 2500 doctoral students, 400 PhDs graduating per year
  - Over 2,400 PhD students are registered at CERN, and 600 PhD theses are completed every year

- US has about 600 enrolled PhD students in HEP
- Postdocs ~ 200, Undergrads ~ 1000 (HEP aspirants) at any time
- US number is ~ 200+600+1000 = 1800, HEP worldwide without US ~ 6000
- Our training community ~ 8000 (not including faculties and scientists)