

Training and The HSF-India Project

David Lange

Princeton University



**PRINCETON
UNIVERSITY**

Towards making this a global collaboration: The HSF-India project

- HSF-India is a 5 year project funded by the US National Science Foundation (Princeton, UMass-Amherst, Oregon State) that aims to build international research software collaborations between US, European, and India based researchers to reach the science goals of experimental particle, nuclear and astroparticle research.
- Given the growing complexity of our scientific data and collaborations, these collaborations are increasingly important to raise the collective productivity of our research community.

Intended as a long-term investment in international team science.

HSF-India is different from a “typical” research project

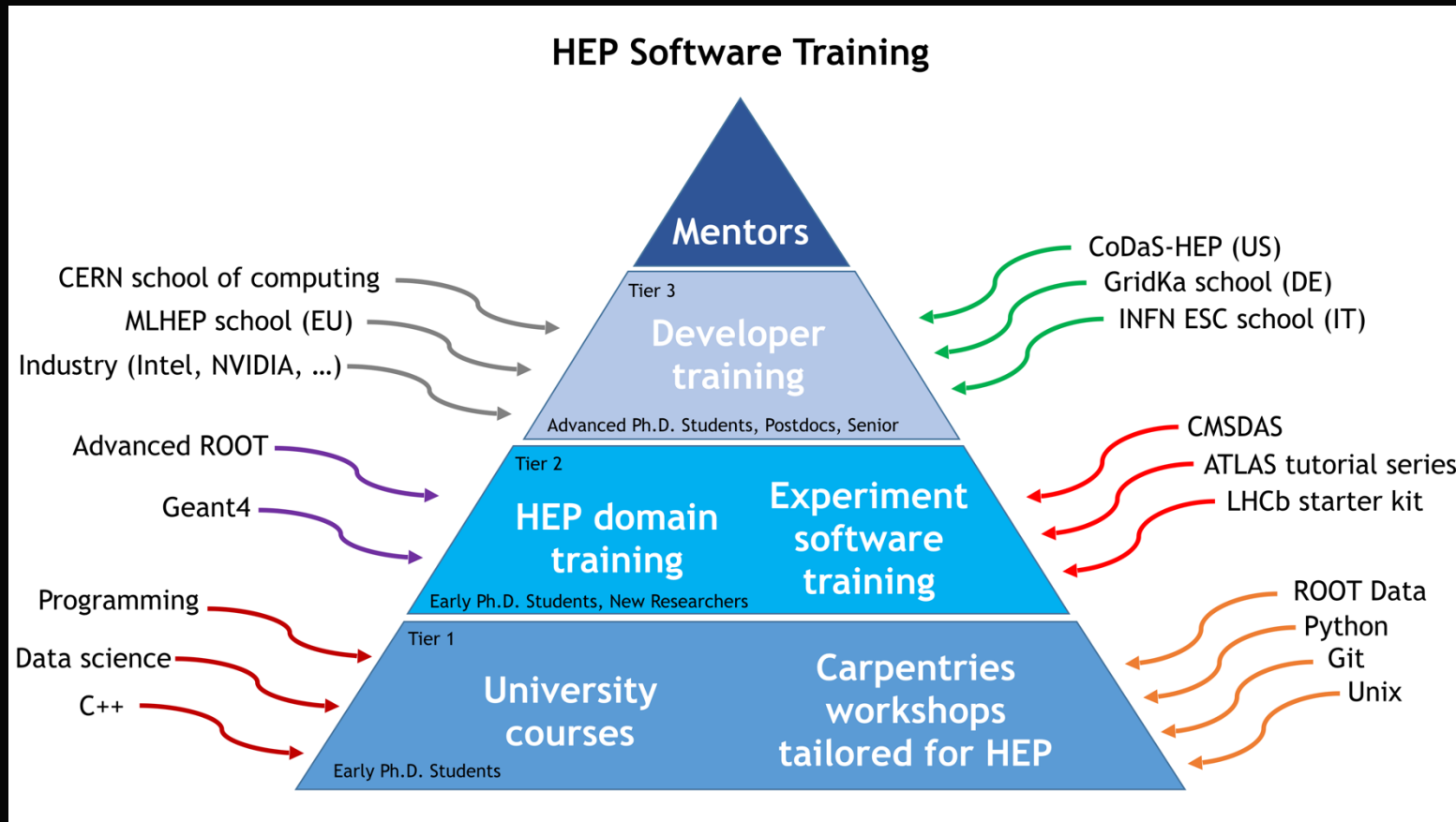
Much of our funding is to facilitate research collaborations, rather than directly fund a specific research activity

Around half of the funding is for supporting participant activities

- Training in research software skills
- Bidirectional research exchanges
- Summer or semester student programs



Bootstrap collaboration through software training



- A vision for training in HEP: researchers progress (vertically) from basic skills training, through user training in existing software to training in skills needed to develop new research software.

Training events in a nutshell

- Who:
 - O(50) participants [typically spanning undergrad, masters, PhD]
 - O(7) experts/instructors
- What:
 - Scientific Python
 - ML/AI
 - GPU programming
 - Simulation techniques
 - Topical lectures
- Where:
 - Regionally across India

All courses use Jupyter-based materials saved in GitHub for post-workshop followup and offline learning

Training events in a nutshell

- Why:
 - Benefit to participants
 - In person, interactive training opportunity helpful for their research (hopefully 😊)
 - Exposure to topics essential for contributing to distributed/collaborative research software
 - Visiting someplace new (for those not local...)
 - Benefit to hosts
 - Cross-institute, cross-experiment engagement builds community
 - Networking and collaboration with international researchers
 - Recruitment opportunity for perspective researchers
 - Benefit to instructors
 - Networking and collaboration with international researchers
 - Recruitment opportunity for perspective researchers
 - Leveraging existing course materials reduces total time commitment
 - Visiting someplace new

Our first software workshop at TIFR in Mumbai (April 2023)

- ~50 students registered ahead, growing registration during the week. Most students came from one of the universities in the Mumbai region.
- Mix of local and US instructors
- Materials derived from/patterned after [HSF training courses](#)



NISER software workshop in December 2023(Bhubaneswar)



- Curriculum built around high-level (but hands-on) overviews of scientific python, machine learning, modern C++, GPU programming and topics of local interest (eg, simulation techniques)



University of Delhi – May 2024



- We missed the heat wave that made international news by a few days... better luck next time?



Upcoming HSF-India workshops


- Regional events help reduce travel burden (within India) and help to meet and collaborate with new groups
- Upcoming events
 - VECC – Kolkata - <https://indico.cern.ch/event/1461967/>
 - Planned for December 18-22
 - University of Hyderabad - <https://indico.cern.ch/event/1394564/>
 - Planned for January 13-17

So who what compute
infrastructure do we rely on?

Infrastructure Components

- Room infrastructure
 - Reliable but not high throughput wifi needed
 - Power (essential) and table (ideal) for each student
 - Configured to allow course leads to work with students (questions, debugging, etc)
- Flexible software configuration and setup
 - Inevitable that course leads want to make last moment updates
 - Interactivity is essential
- Sufficient hardware able to support 50-100 students
 - Adequate GPU resources
 - Flexible authentication mechanism required

Infrastructure choices for interactive sessions

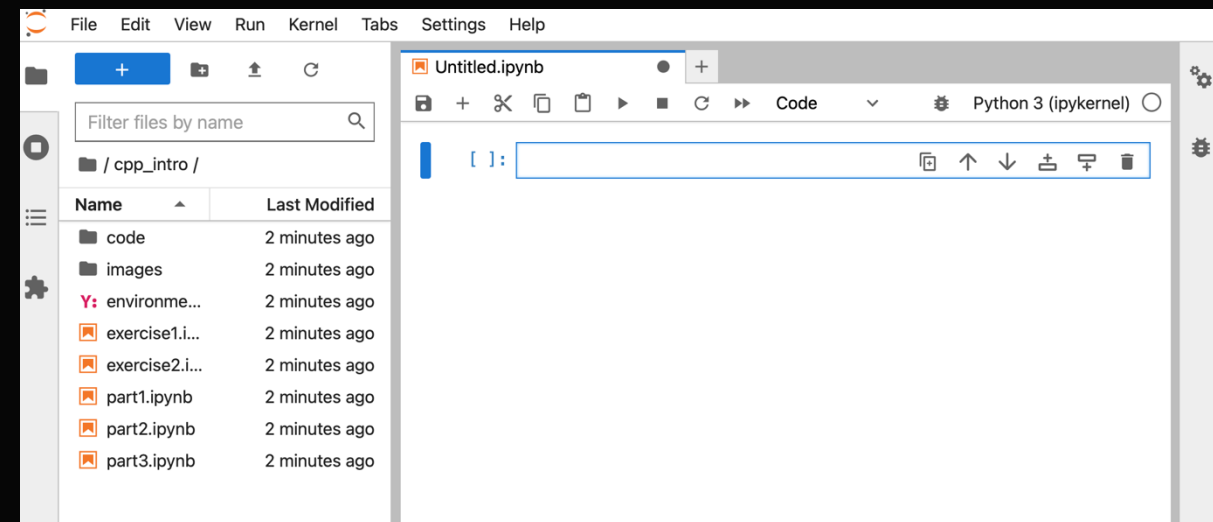
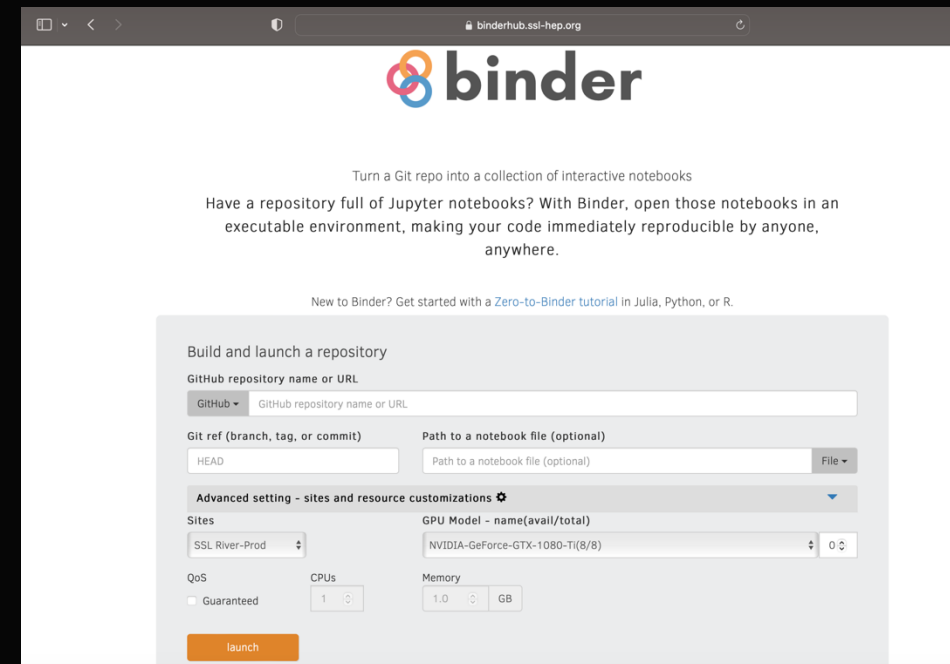
Method	Failure modes	P(works for everyone)	Reusable afterward
Have students install everything on their own laptops: venv, conda-forge, Docker	1. Windows. 2. Not having the software to install the software. 3. 	$1 - 0.9^N$	yes-ish
Public cloud-based Binder (mybinder.org)	1. Stuck loading image. 2. Crashes without persistence.	0.8	yes
GitHub Codespaces	1. Images too large. 2. Boots in VSCode, not Jupyter.	0.95	yes
Google Colab (with GPUs!)	1. Persistence. 2. Fake Jupyter.	0.95	yes
CERN Swan	1. CERN accounts.	0.8	yes
Paid cloud solution: AWS, SaturnCloud	1. Authentication (slips of paper!)	0.95	no
In-browser JupyterLite	1. Not all packages can use it.	1	yes
Self-hosted JupyterHub/BinderHub	1. Authentication. 2. GPUs.	0.9	depends

Infrastructure choices for interactive sessions

Method	Failure modes	P(works for everyone)	Reusable afterward
Have students install everything on their own laptops: venv, conda-forge, Docker	1. Windows. 2. Not having the software to install the software. 3. <code>~_(\ツ)_/</code>	$1 - 0.9^N$	yes-ish
Public cloud-based Binder (mybinder.org)	1. Stuck loading image. 2. Crashes without persistence.	0.8	yes
GitHub Codespaces	1. Images too large. 2. Boots in VSCode, not Jupyter.	0.95	yes
Google Colab (with GPUs!)	1. Persistence. 2. Fake Jupyter.	0.95	yes
CERN Swan	1. CERN accounts.	0.8	yes
Paid cloud solution: AWS, SaturnCloud	1. Authentication (slips of paper!).	0.95	no
In-browser JupyterLite	1. Not all packages can use it.	1	yes
Self-hosted JupyterHub/BinderHub	1. Authentication. 2. GPUs.	0.9	depends

We use remote infrastructure based on BinderHub developed/maintained by the University of Chicago Atlas/IRIS-HEP group

- Using remote resources (SSL-River) gives us a working environment that we can reuse from workshop to workshop.
 - A laptop+browser gives everyone the same environment
 - Available to our program thanks to our connection with IRIS-HEP
 - Reproducibility, reproducibility, reproducibility....



Things to consider...

- Images with CUDA and/or ML frameworks are large and complex
 - They take time to setup and use.
 - However - the ahead of time environment configuration is much easier after the first 1-2 times
- Infrastructure allocates a full GPU (or slice of one) per binder instance. This means you need more GPUs than students
 - No one likes to look over the shoulder of their neighbor



Things to consider...

- Authentication via “CILogon”
 - CILogon has support for university accounts and widely used general purpose accounts (eg, gmail).
 - Perhaps too open? Anyone going to the binderhub logon page could successfully authenticate without additional requirements.
 - We use an additional approval step once workshop attendees are known
- Timezones do not always facilitate real-time support for remote resources...
- Teaching environment management techniques (eg, conda, pip) is still useful for facilitating post-workshop work and teaching best practices
 - There is always seems to be someone with a Nvidia GPU on their laptop...

The UChicago team found it useful to make substantial developments on top of the Binderhub infrastructure

- Ideal goal: Launching user notebooks should be instantaneous
- Reality: Scientific programming images are large and complex.
- An iterative process of deploying a service and seeing how it performs in an environment where lots of users connect simultaneously has been necessary for building a reliable platform

Binderhub development lessons learned : Authentication

- To authenticate, we asked users for their “id”:
 - The identity provided by users (e.g. john.smith@university.edu) is NOT necessarily the identity asserted by their institution (e.g. 'jsmith1@mail.university.edu')
 - Means resolving issues case-by-case
- Beyond the openness of CILogin, we found other reasons to make it easier to handle authentication requests
 - Better: Have users tell us their preferred identity in some programmatic way. Solution developed gives the user a page where they can request to join a group, and approvers can accept them into the group.
 - A “group” might correspond to “HSF-India Delhi May 2024 workshop”
 - This also facilitates management of post-event access..

Binderhub: The “Large Image” problem

- Binder images with CUDA and associated GPU libraries are large
- GPU resources at the scale of 50+ GPUs are distributed (eg, NRP resource in US)
- To pull the image to nodes on a distributed cluster takes 10+ minutes.
 - Even worse during a workshop when many users pull images concurrently.
 - Unacceptable for interactive computing.
- Solution developed: “Continuous Image Puller” Continuously pull images to nodes where notebook will be scheduled. (Puller is a Kubernetes daemonset)

Thanks for listening

- HSF-India is still a new project. We hope can catalyze global collaboration in research software in Physics.
- We have organized week long training events engaging more than 150 students and around 15 US/EU researchers in the last 18 months.
 - Two events scheduled soon and others are being planned.
- Courses rely on SSL (University of Chicago) resources for compute. This is being developed as a customized binderhub configuration to support authentication and flexible software configuration.
- We are here to discuss ways that this project work with you to help further HSF goals and initiatives
 - <http://research-software-collaborations.org>

