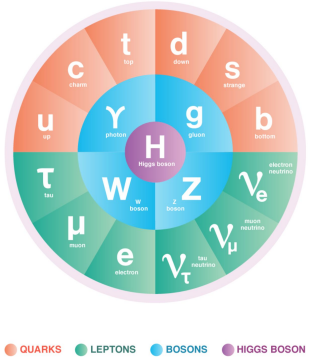
Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) Introduction and Context

PI: Peter Elmer (Princeton), co-PIs: Brian Bockelman (Morgridge Institute), Gordon Watts (U.Washington) with UC-Berkeley, University of Chicago, University of Cincinnati, Cornell University, Indiana University, MIT, U.Michigan-Ann Arbor, U.Nebraska-Lincoln, New York University, Stanford University, UC-Santa Cruz, UC-San Diego, U.Illinois at Urbana-Champaign, U.Puerto Rico-Mayaguez and U.Wisconsin-Madison

http://iris-hep.org

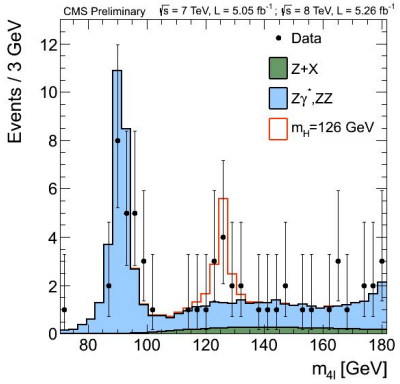
OAC-1836650

Science Driver: Discoveries beyond the Standard Model of Particle Physics

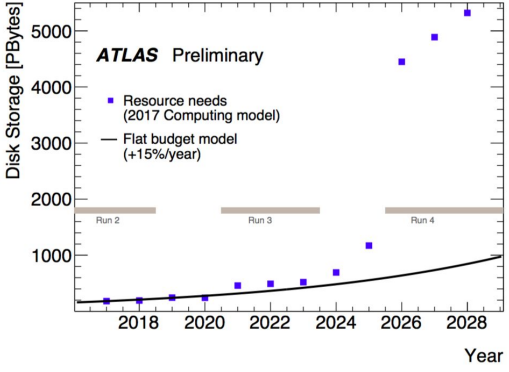


From "Building for Discovery - Strategic Plan for U.S. Particle Physics in the Global Context" - Report of the Particle Physics Project Prioritization Panel (P5):

1) Use the Higgs boson as a new tool for discovery
2) Pursue the physics associated with neutrino mass
3) Identify the new physics of dark matter
4) Understand cosmic acceleration: dark matter and inflation
5) Explore the unknown: new particles, interactions, and physical principles

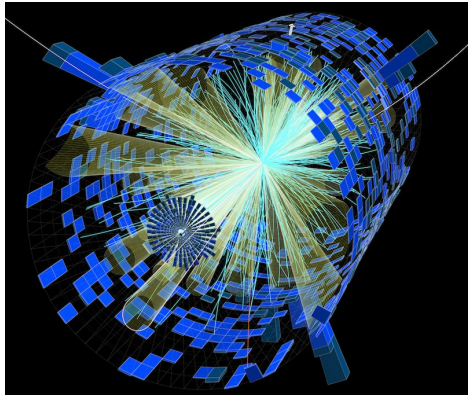


Computational and Data Science Challenges of the High Luminosity Large Hadron Collider (HL-LHC) and other HEP experiments in the 2020s



The HL-LHC will produce exabytes of science data per year, with increased complexity: an average of 200 overlapping proton-proton collisions per event.

During the HL-LHC era, the ATLAS and CMS experiments will record ~10 times as much data from ~100 times as many collisions as were used to discover the Higgs boson (and at twice the energy).



Timeline

CTDR

S2I2-HEP

IRIS-HEP Institute

Design | Execution

Snowmass

Institute Conceptualization and Community White Paper Process

CERN HL-LHC Planning - Computing Technical Design Reports (CTDR) - ATLAS/CMS

U.S. HEP Community Planning Process

HEP Software Foundation (HSF)

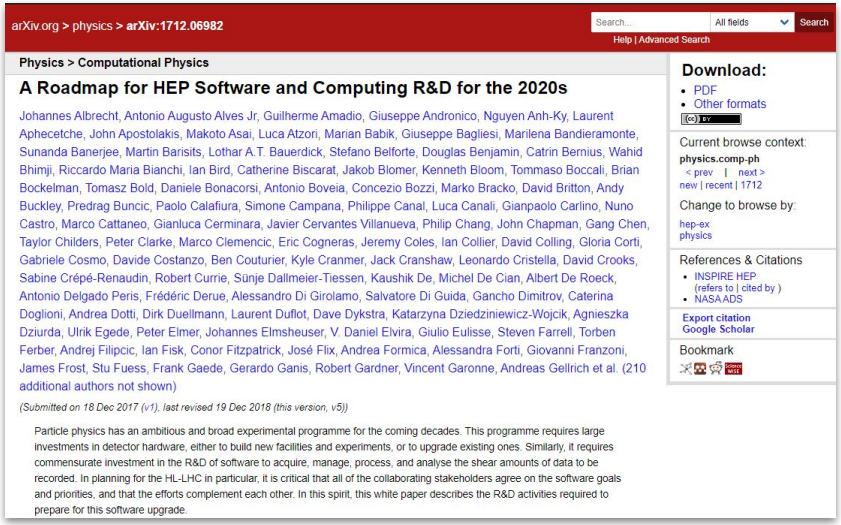
The HSF (http://hepsoftwarefoundation.org) was created in early 2015 as a means for organizing our community to address the software challenges of future projects such as the HL-HLC. The HSF has the following objectives:

- 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

Community White Paper

January 2017
UCSD

June 2017
Annecy



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/

<u>Individual Papers on the arXiv</u>:

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

<u>Community White Paper</u> & the <u>Strategic Plan</u>

<u>arXiv 1712.06982</u> <u>arXiv 1712.06592</u>

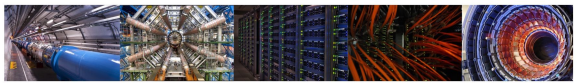
IRIS-HEP

U.S. S2I2-HEP Conceptualization: Additional Criteria

Impact - Physics: Will efforts in this area enable new approaches to computing and software that maximize, and potentially radically extend, the physics reach of the detectors?

Impact - Cost/Resources: 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 within their computing budgets?

Impact - Sustainability: Will efforts in this area significantly improve the long term sustainability of the software through the lifetime of the HL-LHC?



Strategic Plan for a
Scientific Software Innovation Institute (S^2I^2)
for High Energy Physics

arXiv 1712.06592
Dec. 2017

Interest/Expertise: Does the U.S. university community have strong interest and expertise in the area?

Leadership: Are the proposed focus areas complementary to efforts funded by the US-LHC Operations programs, the DOE, and international partners?

Value: Is there potential to provide value to more than one HL-LHC experiment and to the wider HEP community?

Research/Innovation: Are there opportunities for combining research and innovation as part of partnerships between the HEP and Computer Science/Software Engineering/Data Science communities?

US-ATLAS and US-CMS Ops were integral partners in developing this strategic plan

IRIS-HEP

Intellectual Hub for the HEP Community

Sustainable Software R&D objectives

1) Development of **innovative algorithms** for data reconstruction and triggering;

2) Development of highly performant **analysis systems** that reduce “time-to-insight” and maximize the HL-LHC physics potential; and

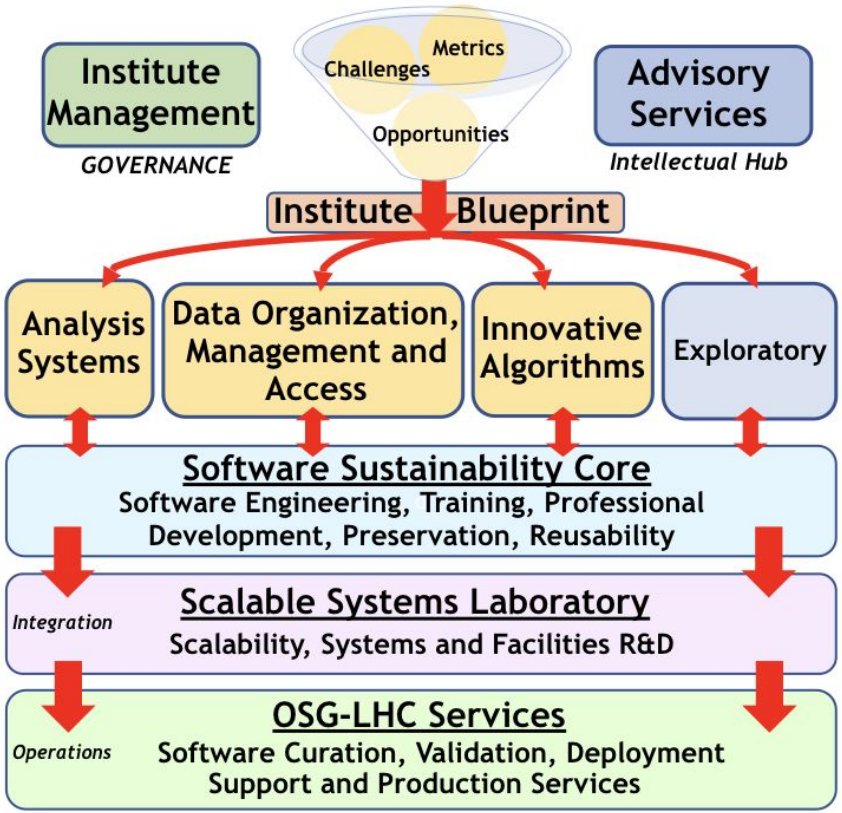
3) Development of **data organization, management and access systems** for the community’s upcoming Exabyte era.

4) Integration of software and scalability for use by **the LHC community on the Open Science Grid**, the Distributed High Throughput Computing infrastructure in the U.S.

The plan for IRIS-HEP reflects a community vision developed by an international community process organized by the HEP Software Foundation (https://hepsoftwarefoundation.org). The S2I2-HEP conceptualization project (http://s2i2-hep.org) derived a Strategic Plan from the community roadmap which would leverage the strengths of the U.S. university community. IRIS-HEP aims to function as an **intellectual hub** for the national and international HEP community, through training, community workshops and the development of wider collaborations with the larger computer and data science communities.

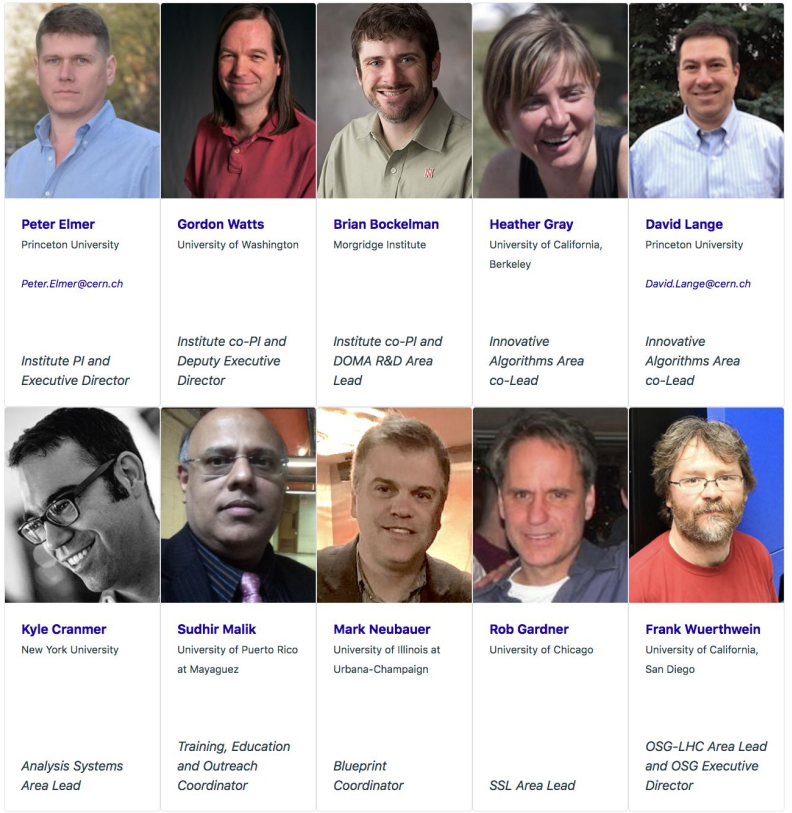
IRIS-HEP funded as a 5 year project from 1 Sep, 2018

IRIS-HEP Structure and Executive Board



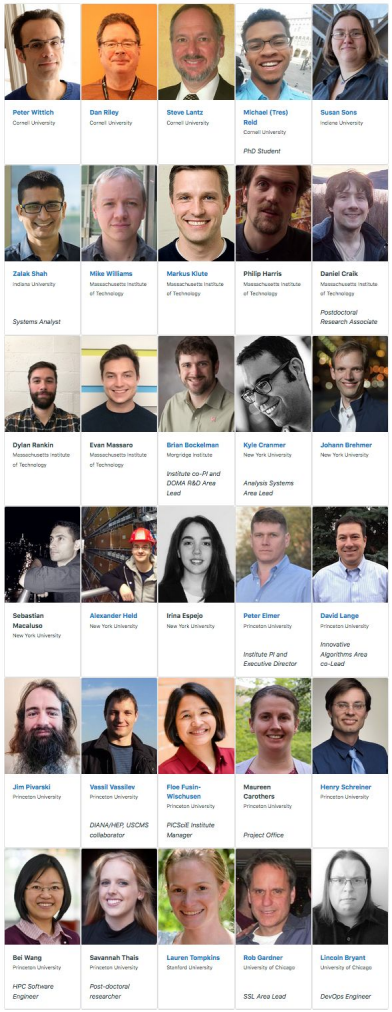
Executive Board

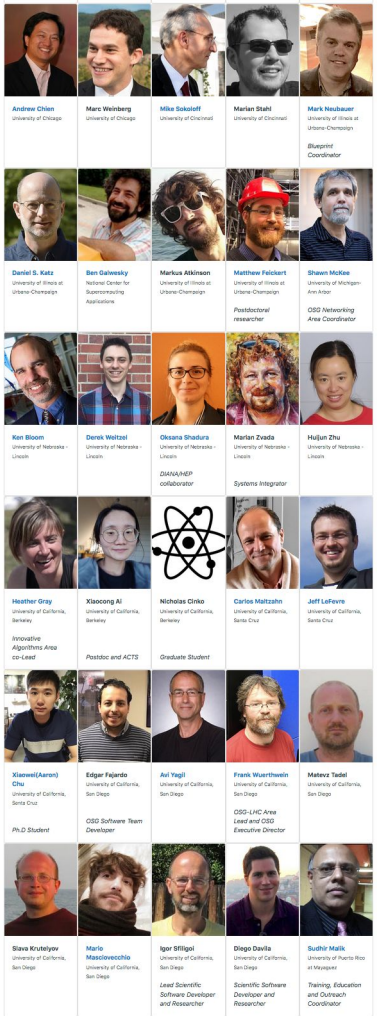
The IRIS-HEP Executive Board manages the day to day activities of the Institute.



Peter Elmer Princeton University *Peter.Elmer@cern.ch* *Institute PI and Executive Director*	**Gordon Watts** University of Washington *Institute co-PI and Deputy Executive Director*	**Brian Bockelman** Morgridge Institute *Institute co-PI and DOMA R&D Area Lead*	**Heather Gray** University of California, Berkeley *Innovative Algorithms Area co-Lead*	**David Lange** Princeton University *David.Lange@cern.ch* *Innovative Algorithms Area co-Lead*
Kyle Cranmer New York University *Analysis Systems Area Lead*	**Sudhir Malik** University of Puerto Rico at Mayaguez *Training, Education and Outreach Coordinator*	**Mark Neubauer** University of Illinois at Urbana-Champaign *Blueprint Coordinator*	**Rob Gardner** University of Chicago *SSL Area Lead*	**Frank Wuerthwein** University of California, San Diego *OSG-LHC Area Lead and OSG Executive Director*

The Executive Board meets weekly.

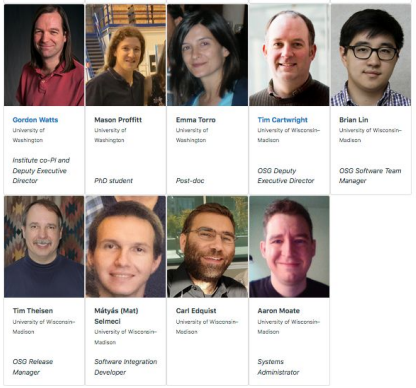




IRIS-HEP Team

http://iris-hep.org/about/team

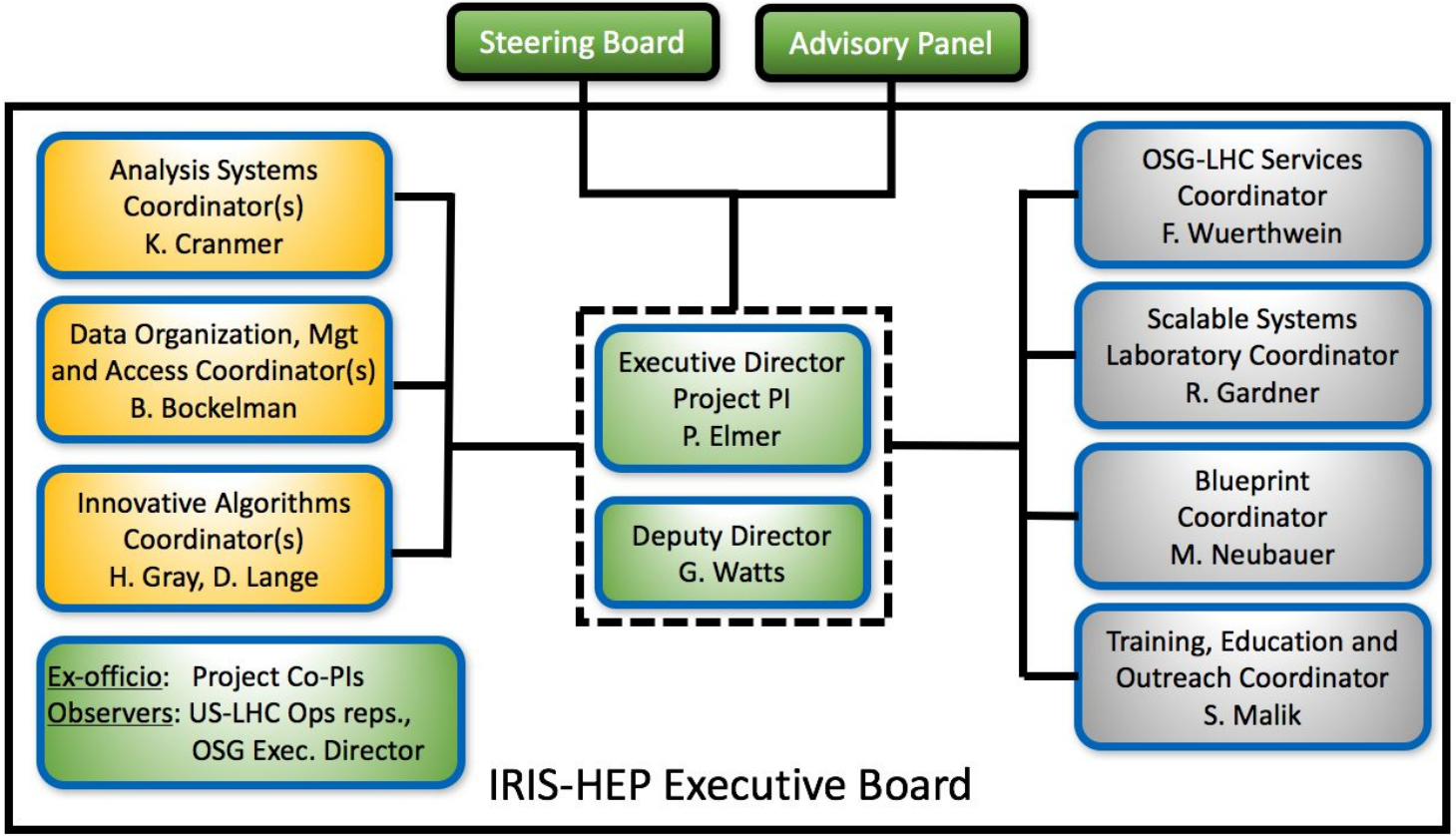
About 28 FTEs of funded effort spread over a larger number of people from 18 universities/institutions



Gender Diversity

Exec Board: 10%
Subaward PIs: 16.7%
Full Team: 17%
For comparison:
CoDaS-HEP 2019: 25.9%
US-CMS Physicists 2017: 16%
US-CMS Grad Students 2017: 17%

Management and Coordination



Steering Board

Represents the major stakeholders and partners for the IRIS-HEP project. Will meet quarterly with the IRIS-HEP Executive Board to learn the status of the project and **provide feedback on the large scale priorities** and current strategy of the Institute.

The steering board meets quarterly with the executive board:

https://indico.cern.ch/category/10989/

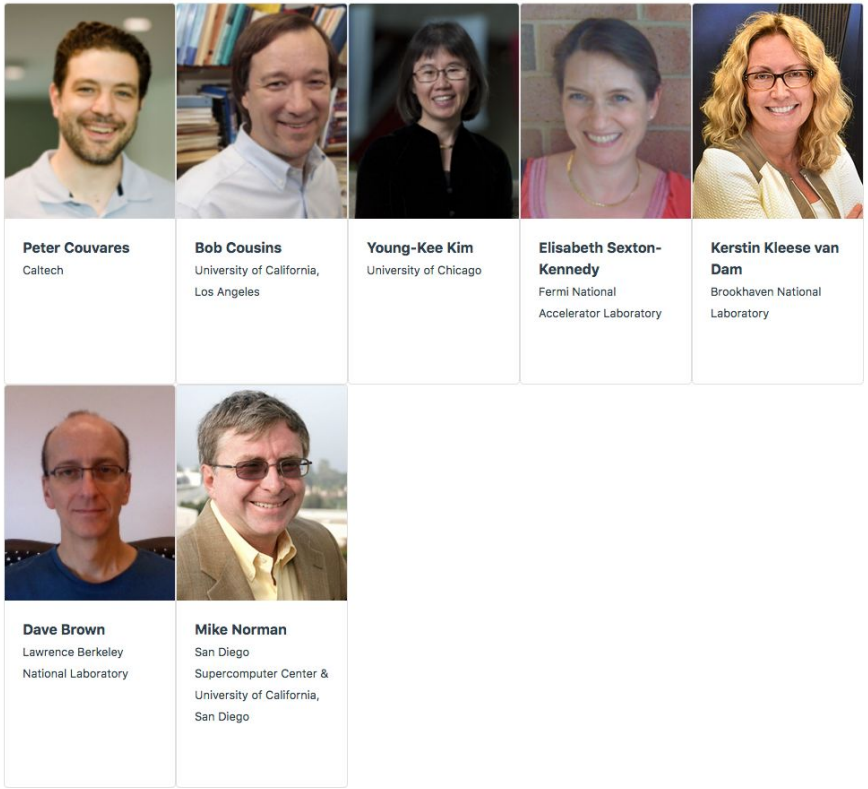
https://iris-hep.org/about/steering-board

Advisory Panel

Provides annual non-stakeholder feedback on the goals and evolving project plans, and evaluates how well the institute is achieving its overall mission as defined with NSF. The Advisory Panel consists of 7 fixed members with an option of inviting ad-hoc additional members as needed for particular topics.

The first in-person meeting with the Advisory Panel took place on 9 September, 2019:

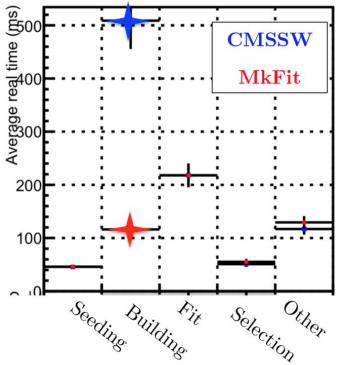
https://indico.cern.ch/event/840467/



IRIS-HEP Innovative Algorithms Highlights

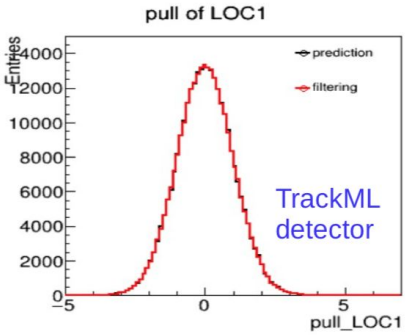
Parallel tracking contributions to MkFit

- Develop track finding/fitting implementations that work efficiently on many-core architectures (vectorized and parallelized algorithms):
- 4x faster track building w/ similar physics performance in realistic benchmark comparisons



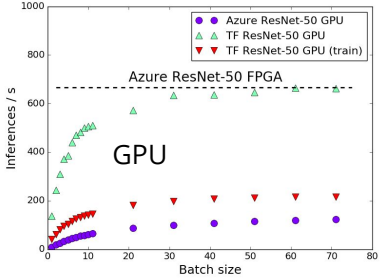
Tracking contributions to ACTS

- Development of the Kalman Filter
- Porting ACTS seeding code to run on GPUs
- Developing connections with other experiments (e.g. Belle-2, JLAB) who may be interested in using ACTS



ML on FPGAs contributions to HLS4ML/FastML

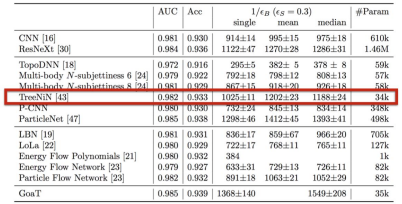
- identifying specific use cases and operational scenarios for use of FPGA-based algorithms in experiment software trigger, event reconstruction or analysis algorithms



https://arxiv.org/pdf/1904.08986.pdf

ML4Jets establishing and curating common metrics and data sets

- Aim to connect with diverse segments of machine learning community. Strong connections with theoretical community interested in jet physics
- Tree Neural network approach demonstrated on reference dataset



	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [30]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k

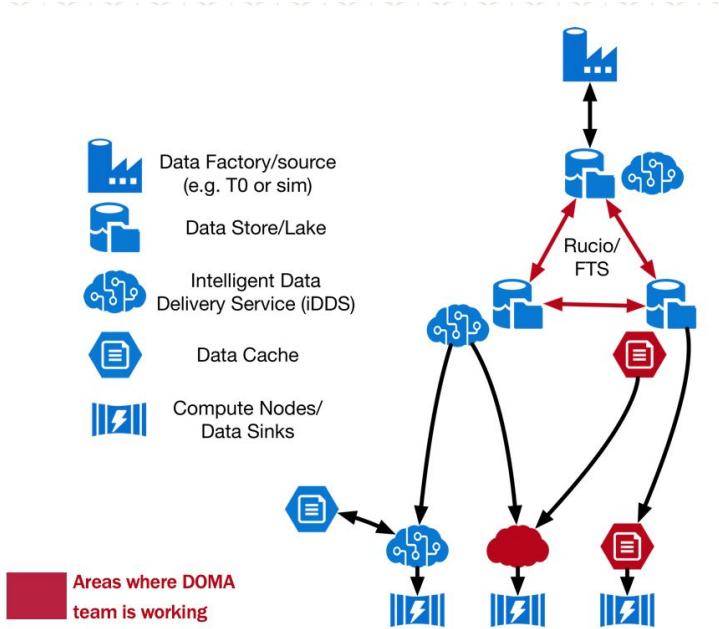
https://arxiv.org/pdf/1902.09914.pdf

Prototype Phase – Used in analysis by early adopters

DOMA (Data Organization, Management, Access)

Fundamental R&D related to the central challenges of organizing, managing, and providing access to exabytes of data from processing systems of various kinds.

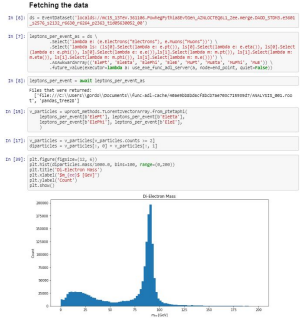
- Data Organization: Improve how HEP data is serialized and stored.
- Data Access: Develop capabilities to deliver filtered and transformed event streams to users and analysis systems.
- Data Management: Improve and deploy distributed storage infrastructure spanning multiple physical sites. Improve inter-site transfer protocols and authorization.



ServiceX / Intelligent Data Delivery

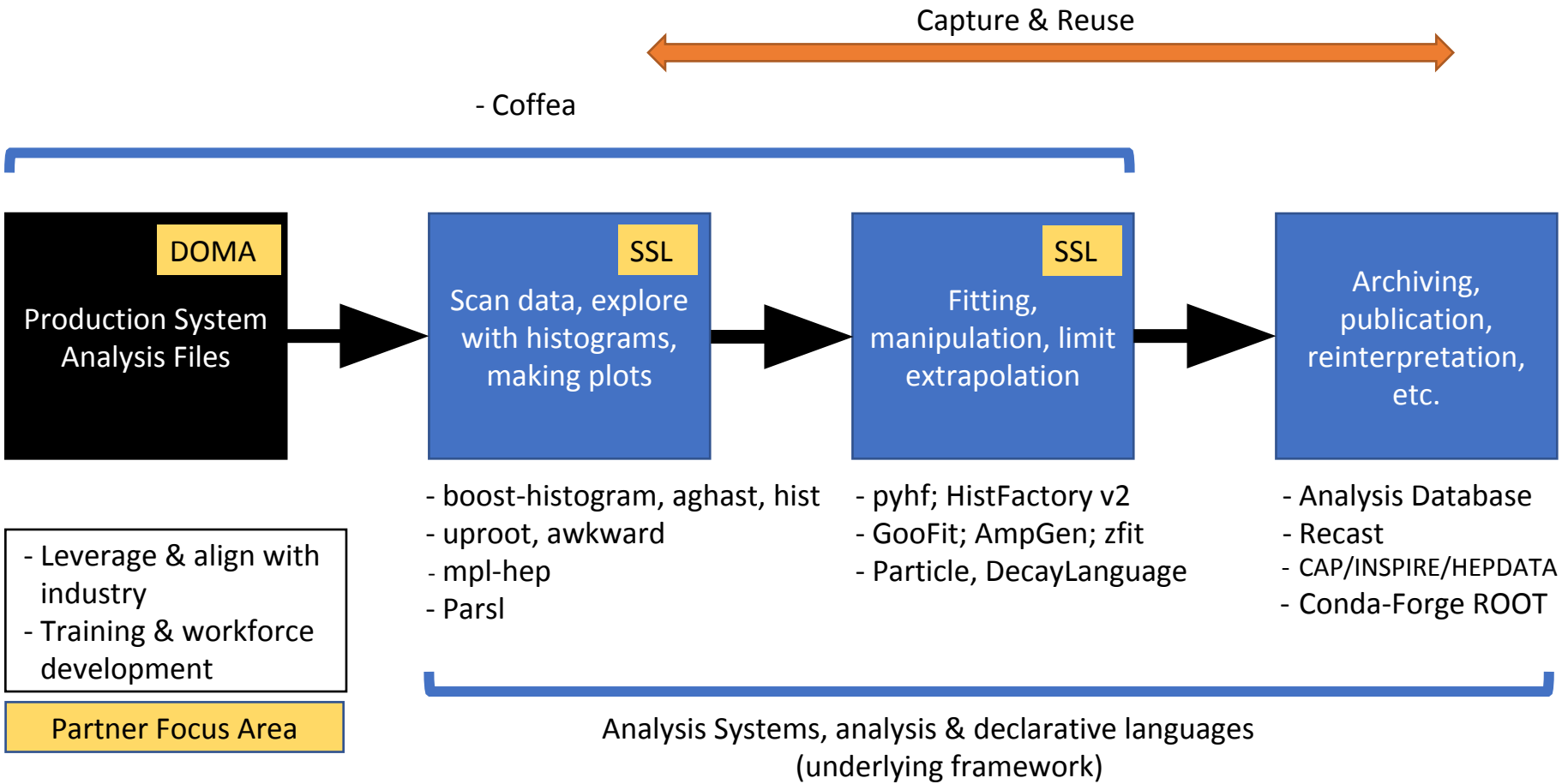
Low-latency delivery of numpy-friendly data transformed from experiment custom formats enabling the use of community supported data science tools.

(joint effort with Analysis Systems)

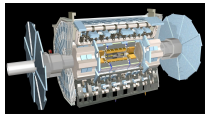


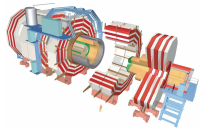
Jupyter Notebook

Analysis Systems Data Flow and Projects



Analysis Systems - Data Query





Production & Simulation

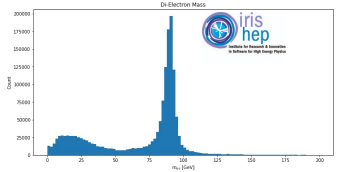
IDDS & ServiceX

Designed for high speed data delivery

Awkward Array

Spark

Coffea



- **ServiceX (IDDS)** - Collaboration with DOMA to serve analysis data in a “columnar form”.
- **Awkward Array** - HEP-style column-wise data in the python ecosystem for manipulating the data
- **Coffea** - column-oriented framework for analysis (developed initially at FNAL in the US CMS context)
    - Builds on top of other backends allowing execution on Spark- or HTCondor-based resources.

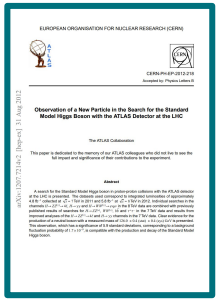
Full chain to make a Z mass peak in electron data!

Analysis Systems

Develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments.

- create greater functionality to enable new techniques,
- reducing time-to-insight and physics,
- lowering the barriers for smaller teams, and
- streamlining analysis preservation, reproducibility, and reuse.

Experiment's Production System



Data Query, histogramming, plotting, statistical models, fitting, archiving, reproducibility, publication

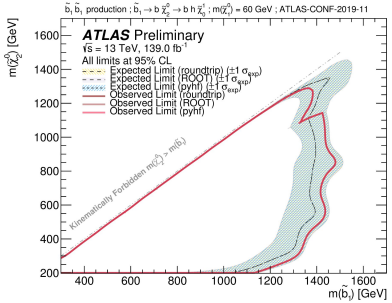
pyhf
∂ifferentiable
ℒikelihoods

Statistical Modeling Language and Tool Limit Extraction

Rewritten from C++ in Python to use TensorFlow or PyTorch as back end.

C++: 10+ hours
pyhf: 30 minutes

GPU acceleration comes for "free"



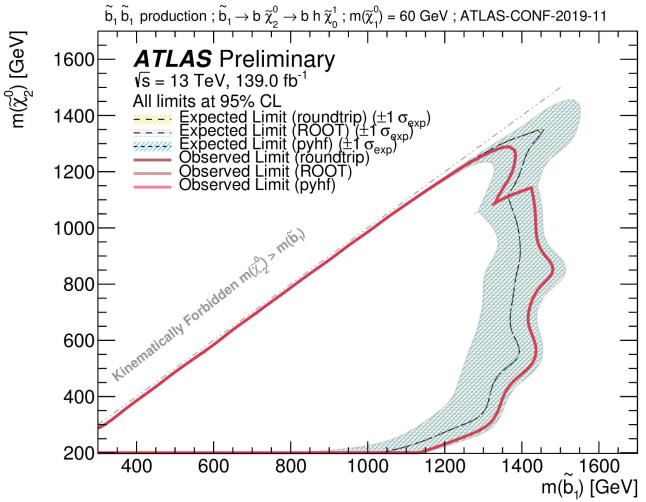
Just released and being incorporated into Analyses Now

Built into SciKit-HEP, a suite of packages that are being adopted by the community

Analysis Systems - Statistical Models

- Build statistical models from binned distributions and data
  - Common last step in analysis to statistically characterize discovery or determine limits
- Already used by ATLAS
- Python library published in the native Python ecosystem.
- Leverages open source libraries as backends for efficient vectorized computation
  - NumPy, TensorFlow, PyTorch
  - Allows external experts do the “heavy lift” of implementing hardware acceleration (on GPUs, TPUS), not physicists.
- Enhances reproducibility of statistical model
  - Allows publications to include full likelihood data on HEPData.

pyhf
∂ifferentiable
ℒikelihoods



(ATL-PHYS-PUB-2019-029)

Shown to reproduce results but faster!
ROOT: 10+ hours **pyhf:** < 30 minutes

Preservation & Reinterpretation

The field is at a tipping point, DIANA/DASPOS/IRIS-HEP contributions have been transformational.

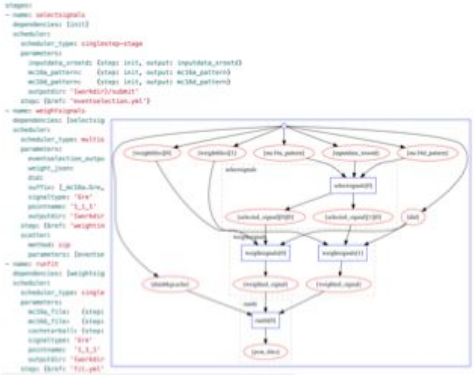
Archiving Real ATLAS Analyses

Using **Industry Standard** Software Packaging to archive analysis:

- Linux Containers ("Docker")
- Integrated into existing analysis infratructure (revision control, continous intergration, grid computing)

Plain-text JSON formats to capture commands and workflows

Close coordination with CERN Analysis Preservation / Reuse Projects



reana

CERN Analysis Preservation

nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

Corrected: Publisher Correction

OPEN

Open is not enough

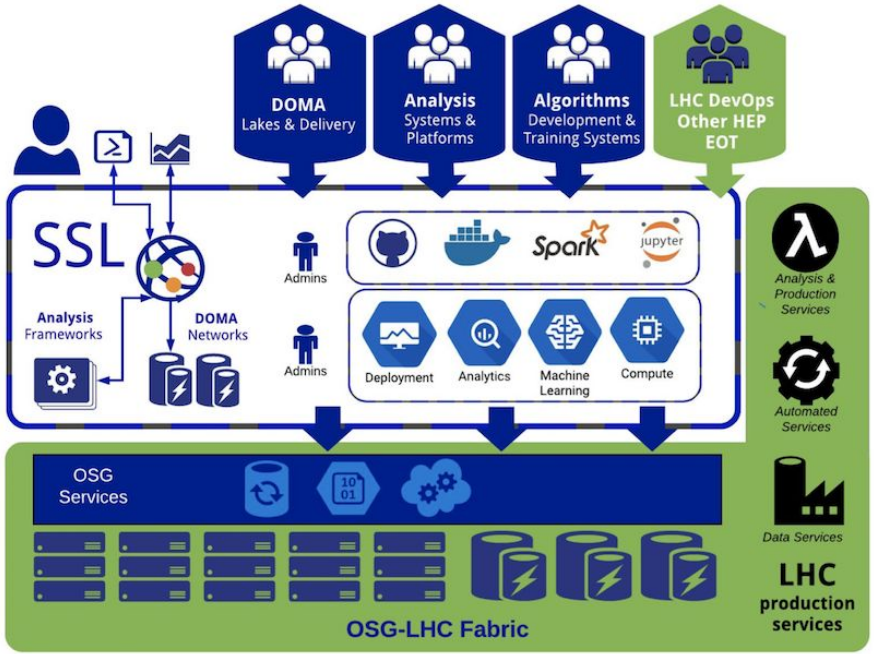
Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1*], Robin Dasler[1,11], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1*], Tim Smith[1], Ana Trisovic[1,5*], Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassila-Perini[4] and Sebastian Neubert[10]

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

Scalable Systems Laboratory (SSL)

Goal: Provide the Institute and the HL-LHC experiments with scalable platforms needed for development in context, perform facilities and systems R&D

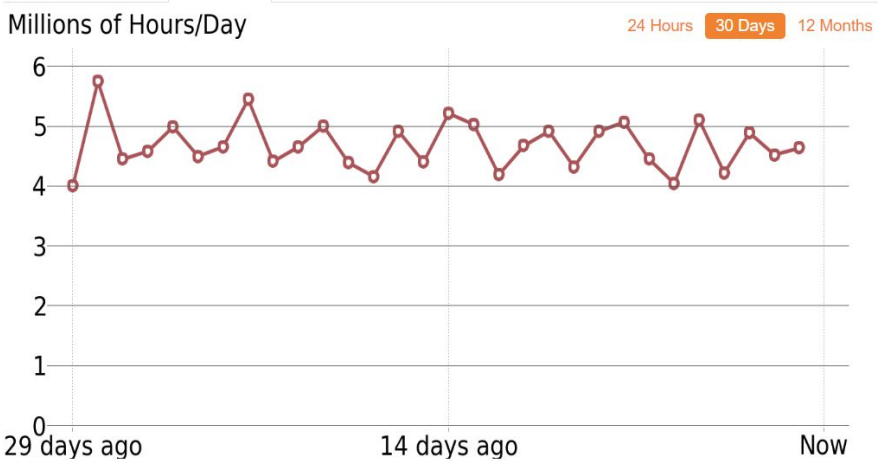
- Provides access to infrastructure and environments
- Organizes software and resources for scalability testing
- Does foundational systems R&D on accelerated services
- Provides the integration path to the OSG-LHC production infrastructure



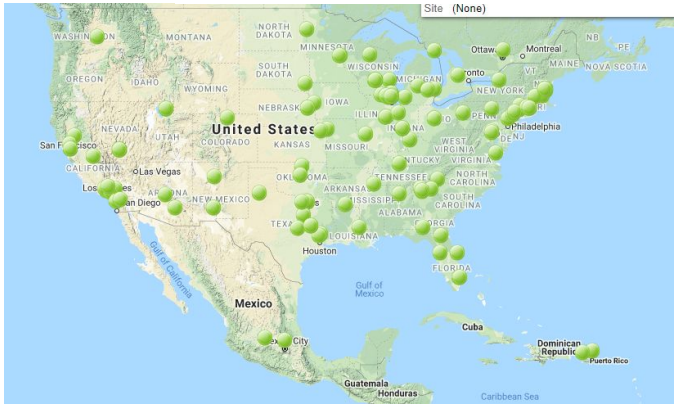
Open Science Grid - LHC

The OSG is a consortium dedicated to the advancement of all of open science via the practice of Distributed High Throughput Computing, and the advancement of its state of the art.

- IRIS-HEP supports LHC operations and development of the consortium.



Open Science Grid



- Work to separate local site hardware and software support by moving services into containers.
- Transitioning security service to use tokens

Particle physicists all over the world depend on these services and scheduling of processing hours (~10,000)

Intellectual Hub - Building Community & Vision

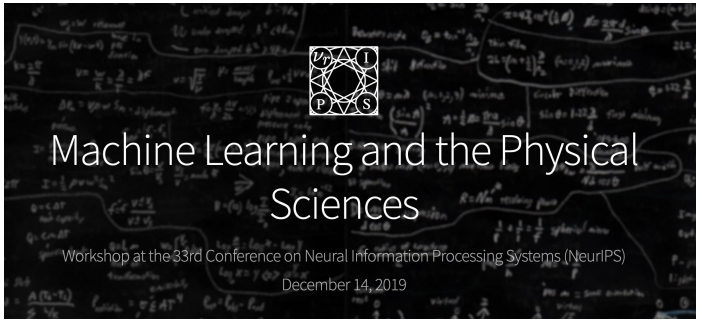
Sponsorship and/or (co-)organization (HSF, etc.) of relevant community workshops and

https://indico.cern.ch/event/759388/

https://indico.cern.ch/event/831165/

https://indico.cern.ch/event/813759/

https://indico.cern.ch/event/769263/



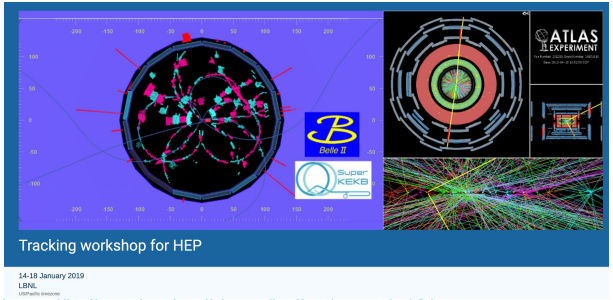
https://ml4physicalsciences.github.io/

[PyHEP 2019](#)

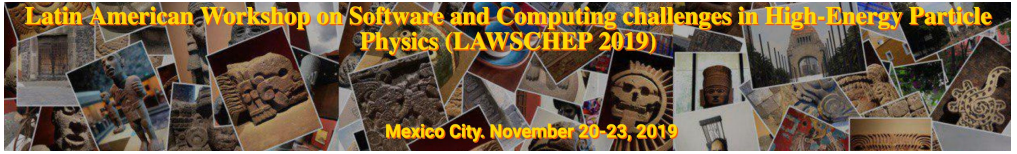
(16-18 October, 2019)

[ML4Jets](#)

(15-17 January, 2020)



https://indico.physics.lbl.gov/indico/event/712/



https://indico.cern.ch/event/813325/

Full list: https://iris-hep.org/events.html

PyHEP Workshop Series

PyHEP is a series of workshops started in 2018 to discuss and promote the usage of Python in the HEP community at large. It has been supported by DIANA/HEP, and now IRIS-HEP, in collaboration with HSF.

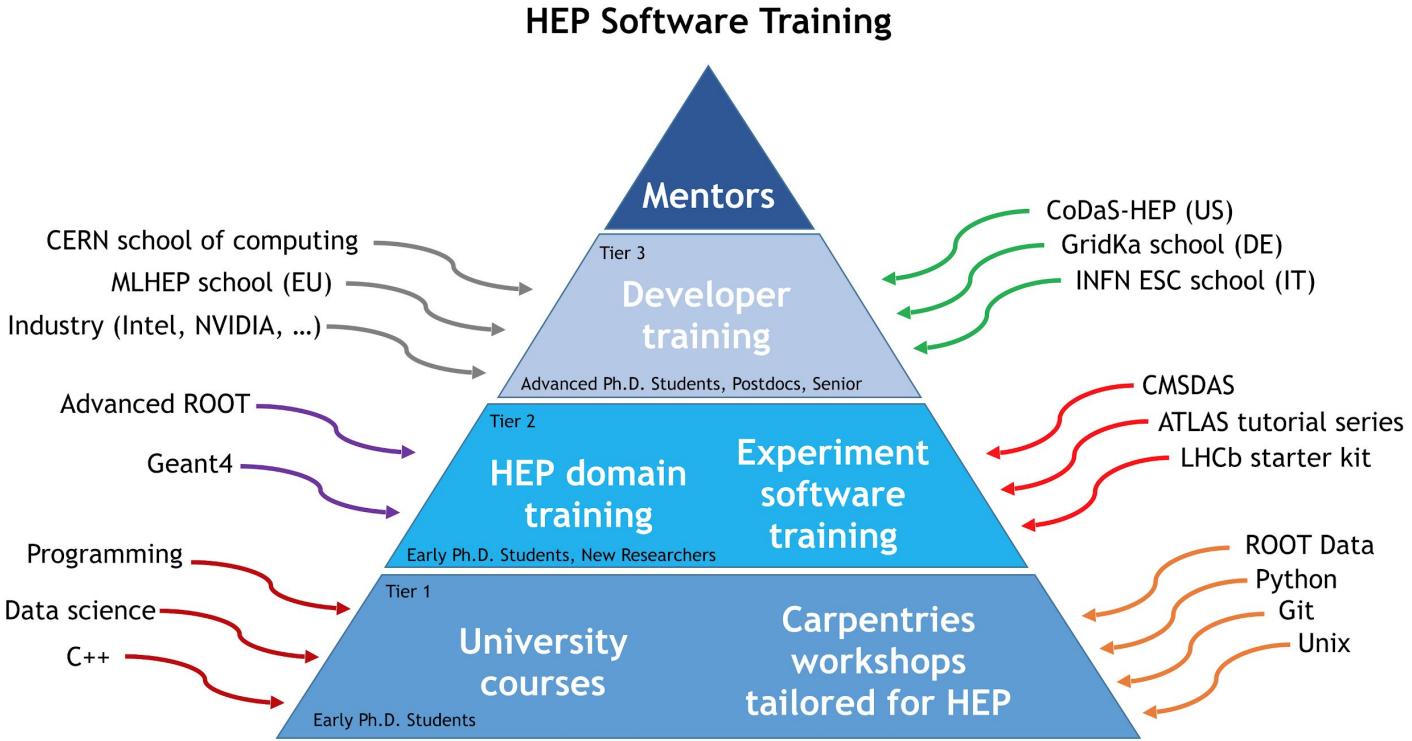
PyHEP 2020 will soon be announced: 11-13 July, 2020 in Austin, TX, partially overlapping with the SciPy 2020 conference, also in Austin, TX

This is not just a “programming language” issue, it is a key place where HEP can explore how to interact with, learn from, contribute to, and perhaps lead areas in the larger scientific, data science and ML communities. (Including use of open data, experimentalist - theorist interactions, etc.)

A consistent message from our students and postdocs who transition to industry and other fields is that we teach them great skills, but they are limited initially by only knowing HEP-only tools.

A growing community: 38 participants at PyHEP 2018, 55 participants at PyHEP 2019, aiming for 80-100 participants at PyHEP 2020

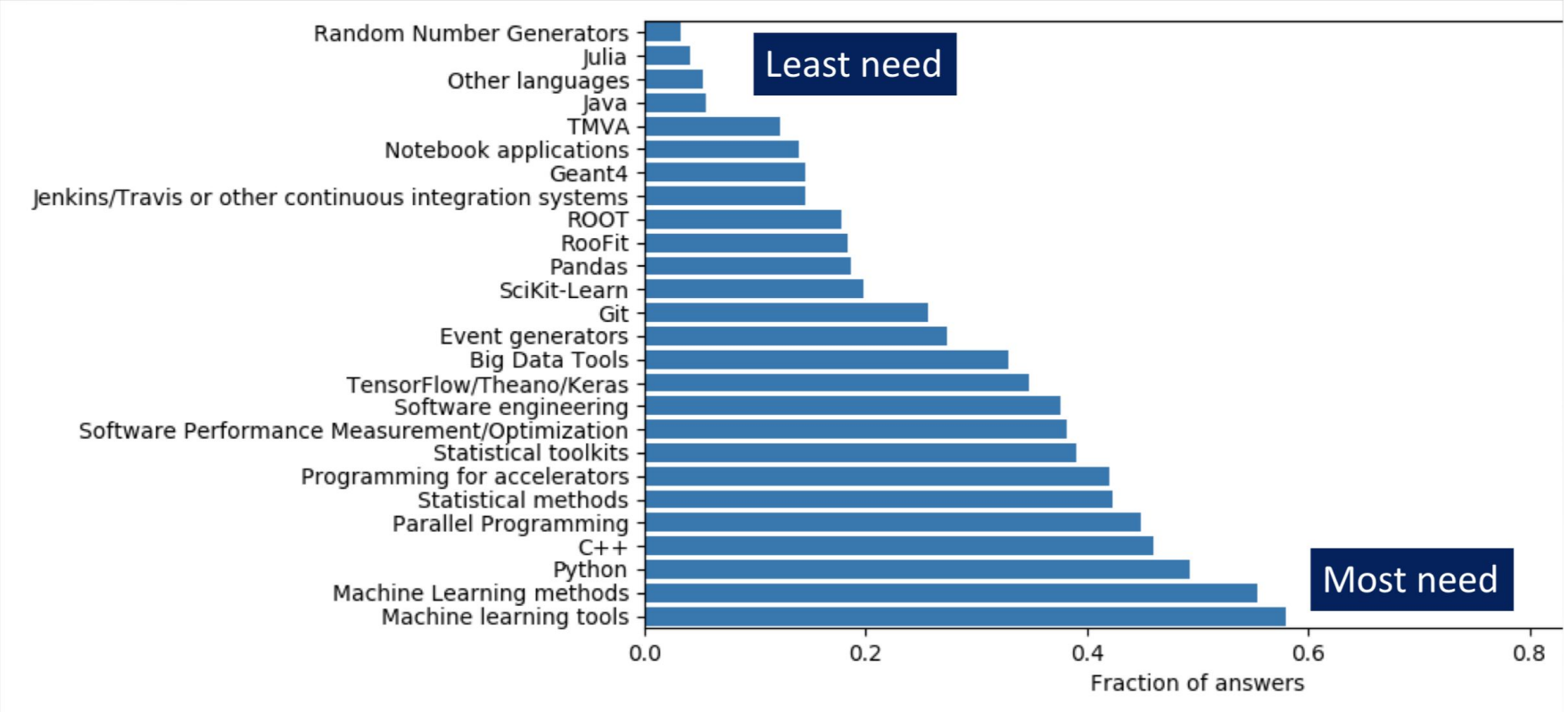
Training and Education - Sustainability/Scalability



This is a general framework for training, but from the NSF we have funds from both IRIS-HEP (OAC-1836650) and a separate project FIRST-HEP (OAC-1829707, OAC-1829729, http://first-hep.org) which are working towards implementing this model.

Training Survey

In early 2019, we did a survey of training needs (link for results summary), 334 people responded!



Inspirations

https://carpentries.org

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

THE CARPENTRIES

About The Carpentries Curricula

- Data Carpentry: Ecology
- Data Carpentry: Genomics
- Data Carpentry: Geospatial
- Data Carpentry: Social Sciences
- Library Carpentry
- Software Carpentry (All Workshops)
- Software Carpentry (Plotting and Programming in Python)
- Software Carpentry (Programming with Python)
- Software Carpentry (Programming with R)
- Software Carpentry (R for Reproducible Scientific Analysis)

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

Training, Education and Outreach Events

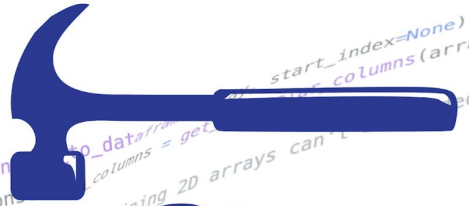
Upcoming events:

- 27-29 Nov, 2019 - Software Carpentry at CERN
  - CERN, Geneva, Switzerland
  - Indico page

Past events:

- 9-21 Aug, 2019 - ATLAS Software Carpentries Training
  - Lawrence Berkeley National Laboratory
  - Indico page
- 22-26 Jul, 2019 - Computational and Data Science for High Energy Physics (CoDaS-HEP) 2019 School
  - Princeton University
  - Webpage
- 10 Jun, 2019 - FIRST-HEP/ATLAS Software Training
  - Argonne National Laboratory
  - Indico page
- 3-4 Jun, 2019 - An introduction to programming for STEM teachers
  - University of Puerto Rico at Mayaguez
  - Indico page
- 24-26 Apr, 2019 - Machine Learning Hackathon for UPRM Students
  - University of Puerto Rico at Mayaguez
  - Indico page
- 1-2 Apr, 2019 - Software Carpentry Workshop
  - Fermi National Accelerator Laboratory
  - Indico page

In collaboration with FIRST-HEP (http://first-hep.org), the Carpentries (https://carpentries.org) and others



software carpentry

iris hep

HSF
HEP Software Foundation

iris hep

@CERN
24-27 March

Confused by code? Troubled by tests? Pull requests got stuck?

Software Carpentry will get you up and running fast, teaching key skills in a friendly, fun, supported environment

- Python
- Git and GitHub
- Shell
- PyROOT and UpRoot
- Jupyter Notebooks
- SWAN

LHCb
ATLAS
CMS
ALICE

https://indico.cern.ch/e/sc-cern

Summary

- IRIS-HEP was funded on September 1st, 2018
  - We are approaching the end of the design phase
  - Projects in all phases (design, prototype, and production) exist.
  - We are fully staffed, ~30 FTE's
  - Full description of projects available on our website, http://iris-hep.org
- Community Impact
  - Software is being adopted by others, in some cases dramatically.
  - Facilities work in SSL and OSG is leading the international field
- Community Outreach
  - We've reached almost 1000 people with our workshops, and another 300 with our training efforts
  - We continue to organize Blueprint workshops to build community consensus.
- Next
  - Start "Execution Phase" September 2020
  - Work on integrating projects in prototype stage into coherent and scalable software for the community
  - The "Snowmass Process-2021" provides an opportunity for us to update the Community White Paper/Roadmap.