IRIS-HEP

G. Watts (UW/Seattle)
For the IRIS-HEP Team

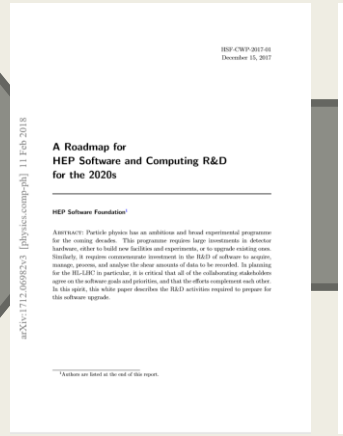
Guidance

1. How does the ensemble of US Software R&D efforts fit together to implement the HL-LHC Software/Computing roadmap and meet the challenges of the HL-LHC? Which areas are not covered by US R&D efforts and should have international coordination? Which areas present new challenges or new opportunities since the Community White Paper (CWP) process that was executed in 2017?
2. How do the US Software R&D efforts collaborate with each other and with international efforts? How do these efforts align with and leverage national exascale, national NSF OAC priorities and trends in the broader community?
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?

Guidance

- Project Summaries
  - (done) Brief descriptive summary of the R&D project goals, level of effort, institutions/personnel involved and timeline
  - (done) Is it generic work, or experiment specific?
- Where do the efforts fit into the HL-LHC software & computing ecosystem?
  - (done) To be used as part of reconstruction, part of the analysis pipeline, facilities, underlying infrastructure, etc.
- What other projects are you collaborating with (currently, planning)?
  - (done) List the connections with other projects (national and international)
  - What the connections are bringing to the table or you are giving them, and if they are healthy
- What connections are there with US-LHC Operations programs? With the experiments?
  - (done) US ATLAS and US CMS have large software and computing groups - how aware of the work are they, are they collaborating?
  - (done) Is there a delivery mechanism for the output of the R&D project to the experiments or the operations programs?
- What connections are you missing and would be good to build?
- Are there things around you in the ecosystem that do not exist, but if they were there it would make your project more impactful or increase its chances of success?

Community White Paper



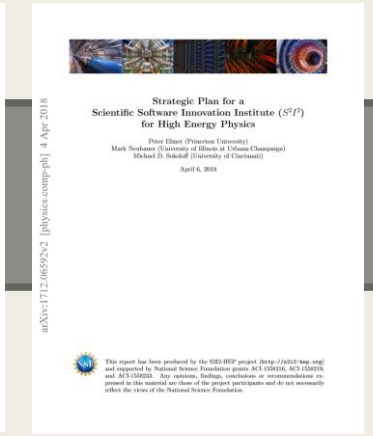
HSF-CWP-2017-01
December 15, 2017

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, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

[1]Authors are listed at the end of this report.



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

Peter Elmer (Princeton University)
Mark Neubauer (University of Illinois at Urbana-Champaign)
Michael D. Sokoloff (University of Cincinnati)

April 6, 2018

This report has been produced by the S2I2-HEP project (http://s2i2-hep.org) and supported by National Science Foundation grants ACI-1558516, ACI-1558519, and ACI-1558233. Any opinions, findings, conclusions or recommendations expressed in this material are those of the project participants and do not necessarily reflect the views of the National Science Foundation.

iris hep

Institute for Research & Innovation
in Software for High Energy Physics

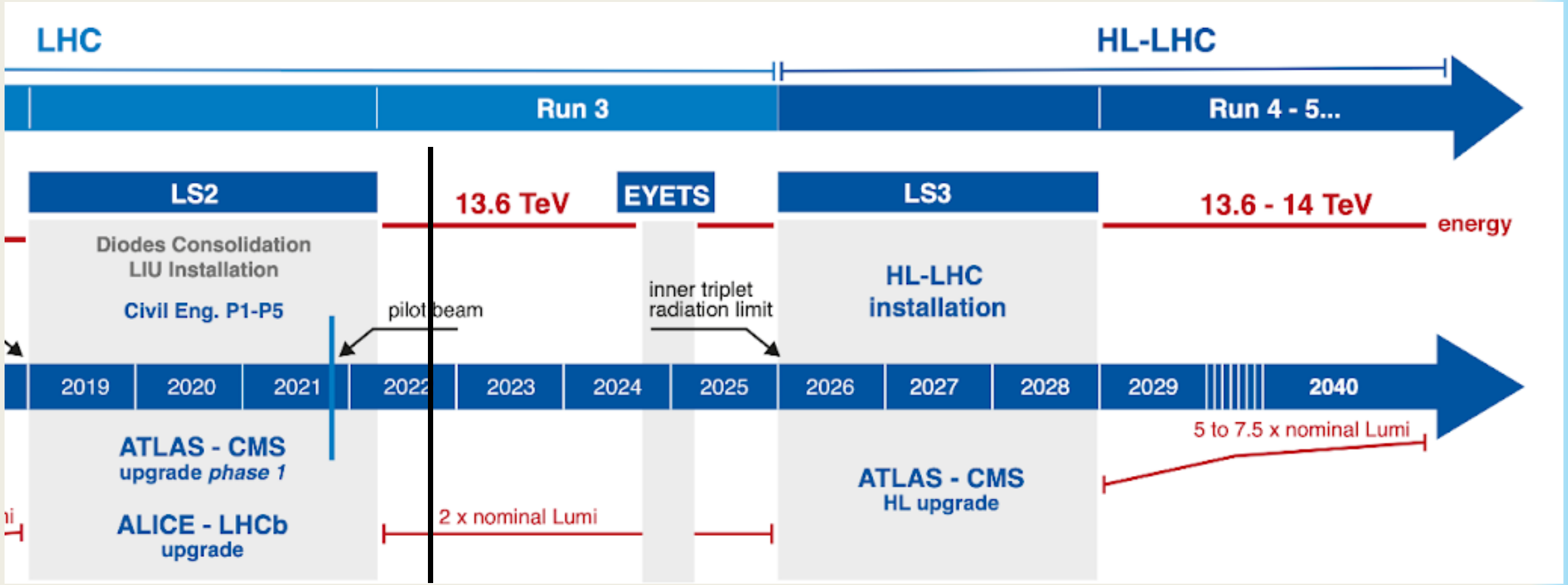
Goals

"IRIS-HEP aims to **develop** the **state-of-the-art software cyberinfrastructure** required for the challenges of **data intensive scientific research** at the High Luminosity Large Hadron Collider (**HL-LHC**) at CERN, and other planned HEP experiments of the 2020's.

These facilities are discovery machines which aim to understand the fundamental building blocks of nature and their interactions"

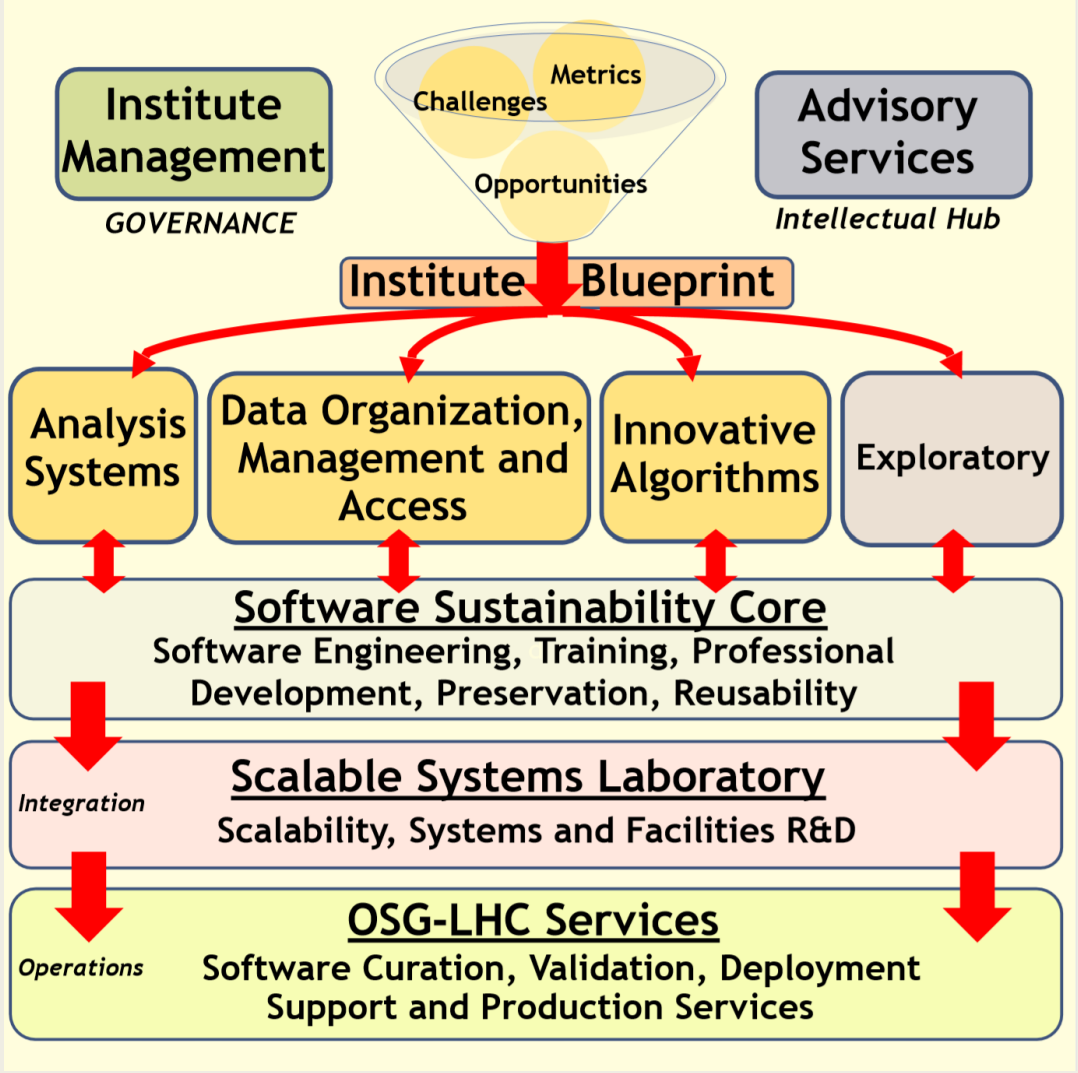
IRIS-HEP

Funded September 1, 2018, for 5 years



IRIS-HEP

Organization & Projects

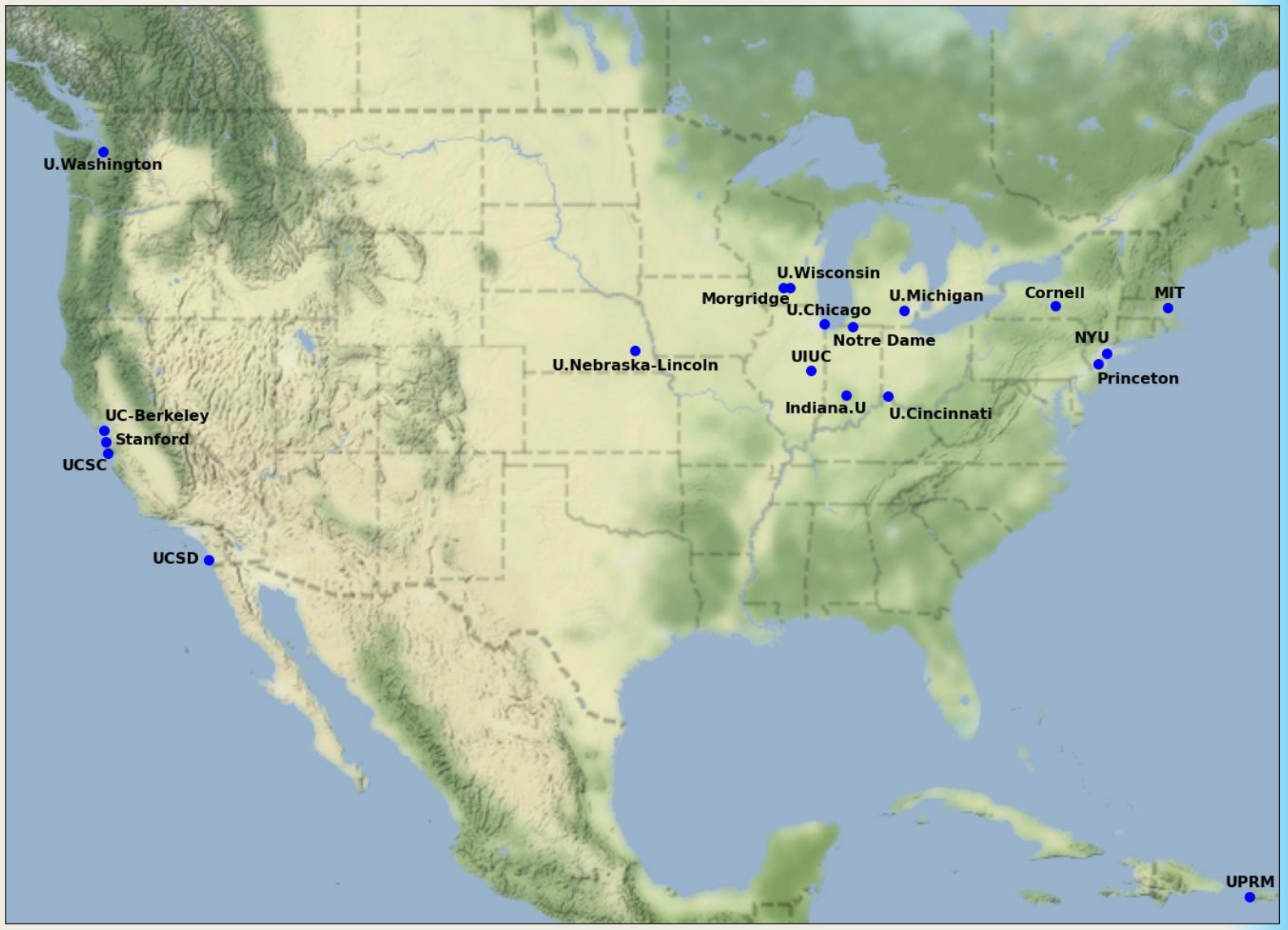


R&D work occurs at all phases

The further down, the closer to facilities and production

Institutions

Spread across the USA at
20 institutions



Effort Overview

Area	FTE
Mgmt/Project office	1.9
Analysis Systems	9.4
DOMA	2.9
Innovative Algorithms	8.7
Sustainability Core	1.4
SSL	1.3
OSG-LHC	5.2
Total	**30.8**

A mix of career stages and job categories:

- 12 students ~ 4.6 FTE
- 8.4 FTE postdocs
- 13.7 FTE staff/professionals
- 1.2 FTE faculty

The $5M/year budget goes almost entirely towards salaries (+ travel/M&S/tuition). There is only a very, very modest hardware budget. There are participant funds (~$80k/year) to support aspects of the intellectual hub activities, blueprint, training, etc.

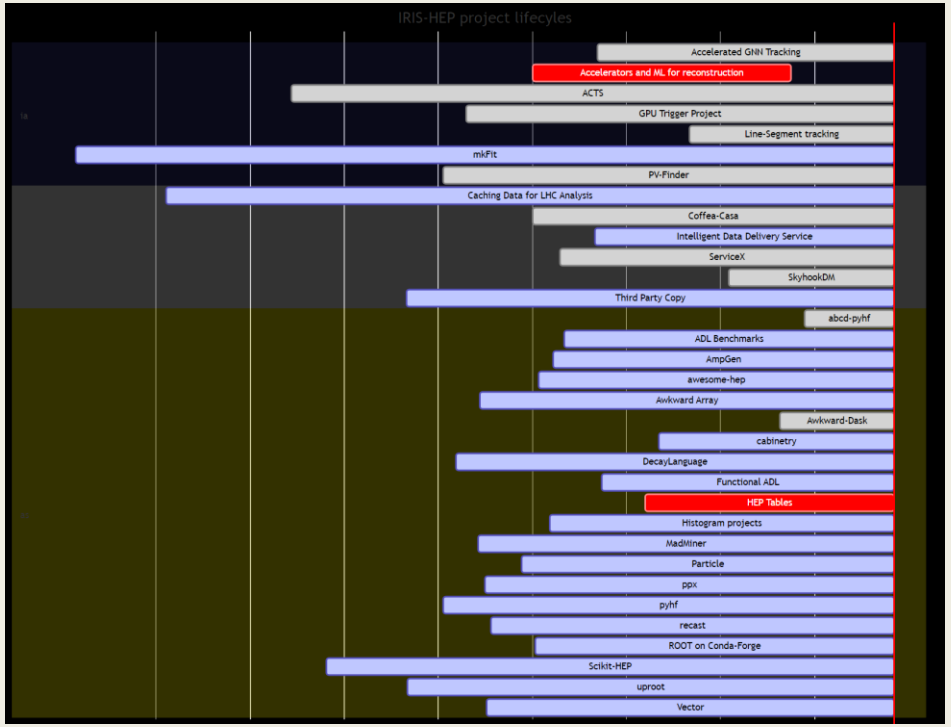
Organization & Projects

Name	Focus Area(s)	Maturity	Description
abcd-pyhf	AS	Development	Likelihood-based ABCD method with pyhf
Accelerated GNN Tracking	IA	Exploratory	accel-gnn-tracking
Accelerators and ML for reconstruction	IA	Archived	Accelerated calorimeter reconstruction using Machine Learning as a Service
ACTS	IA	Development	Development of experiment-independent, thread-safe track reconstruction.
ADL Benchmarks	AS	Deployed	Functionality benchmarks for analysis description languages
AmpGen	AS	Deployed	Generation and fitting for multibody hadron decays
Analysis Grand Challenge	DOMA, AS	Development	Analysis Grand Challenge
awesome-hep	AS	Deployed	A curated list of awesome high energy and particle physics software
Awkward Array	AS	Deployed	Manipulate arrays of complex data structures
Awkward-Dask	AS	Testing	Developing a new high-level Dask collection for Awkward Arrays
cabinetry	AS	Deployed	Building and steering template fits
Caching Data for LHC Analysis	DOMA, OSGLHC	Deployed	Cached-based placement of analysis datasets.
Coffea-Casa	DOMA, AS	Testing	A Prototype of Analysis Facility
DecayLanguage	AS	Deployed	Describe and convert particle decays
exploratory-ml	AS, IA	Development	Analysis Reinterpretation
Functional ADL	AS	Deployed	Functional Analysis Description Language
GPU Trigger Project	IA	Testing	Allen: a GPU trigger for LHCb
HEP Tables	AS	Archived	Heterogeneous Distributed Array Programming Environment
Histogram projects	AS	Deployed	Histogramming efforts
Intelligent Data Delivery Service	DOMA	Deployed	Delivering Data. Better.
Line-Segment tracking	IA	Development	Segment linking tracking for CMS
Machine Learning for jets	IA	Development	Machine learning for jets
MadMiner	AS	Deployed	Likelihood-free Inference
mkFit	IA	Deployed	Modernizing Kalman filter tracking for CMS
Modeling Data Workflows	DOMA, OSGLHC	Archived	Modeling HL-LHC Data flows
OSG Network Monitoring	OSGLHC	Deployed	Provide network monitoring for LHC and OSG sites
OSG Operations	OSGLHC	Deployed	Operate OSG-LHC services
OSG Security	OSGLHC	Deployed	OSG Cybersecurity team
OSG Software and Release	OSGLHC	Deployed	Provide integrated software for running dHTC services
Particle	AS	Deployed	Pythonic particle information
ppx	AS	Deployed	Cross-platform Probabilistic Programming eXecution protocol
PV-Finder	IA	Testing	CNNs to find primary vertices
pyhf	AS	Deployed	Differentiable likelihoods
recast	AS	Deployed	Analysis Reinterpretation
ROOT on Conda-Forge	AS	Deployed	Use ROOT in Conda through Conda-Forge
Scikit-HEP	AS	Deployed	Pythonic analysis tools
ServiceX	DOMA	Testing	Delivering columnar data on demand
SkyhookDM	DOMA	Development	Programmable Storage for Databases and Datasets
Third Party Copy	DOMA, OSGLHC	Deployed	Envisioning a new way to move LHC data
uproot	AS	Deployed	Read and write ROOT files in Python
Vector	AS	Deployed	Manipulate vectors

See https://iris-hep.org/projects for a comprehensive list of projects.

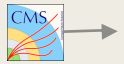
- About 90% complete
- OSG and SSL and training projects do not show up

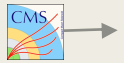
There is also a timeline...



Most projects built to be multi-experiment

Name	Focus Area(s)
abcd-pyhf	AS
Accelerated GNN Tracking	IA
Accelerators and ML for reconstruction	IA
ACTS	IA
ADL Benchmarks	AS
AmpGen	AS
Analysis Grand Challenge	DOMA, AS
awesome-hep	AS
Awkward Array	AS
Awkward-Dask	AS
cabinetry	AS
Caching Data for LHC Analysis	DOMA, OSGLHC
Coffea-Casa	DOMA, AS
DecayLanguage	AS
exploratory-ml	AS, IA
Functional ADL	AS
GPU Trigger Project	IA
HEP Tables	AS
Histogram projects	AS
Intelligent Data Delivery Service	DOMA

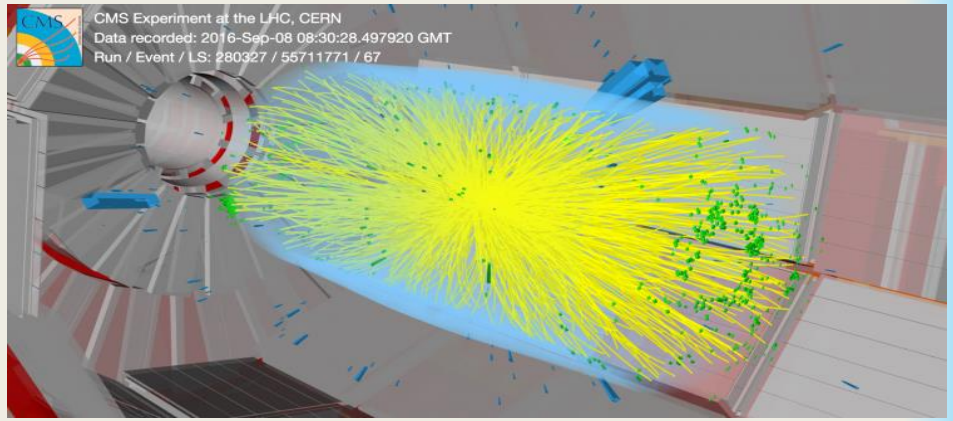




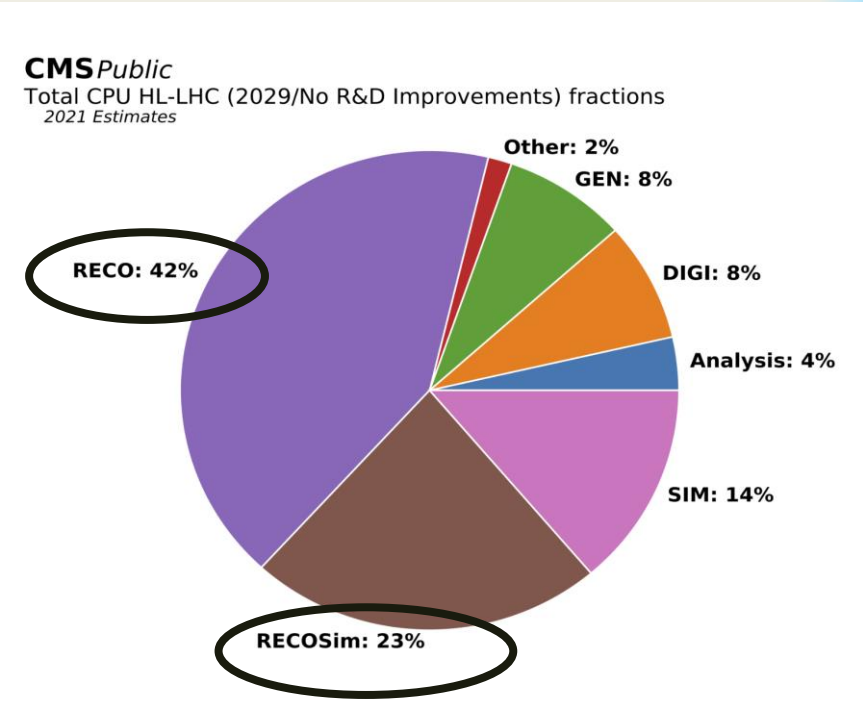
Line-Segment tracking	IA	Development
Machine Learning for jets	IA	Development
MadMiner	AS	Deployed
mkFit	IA	Deployed
Modeling Data Workflows	DOMA, OSGLHC	Archived
OSG Network Monitoring	OSGLHC	Deployed
OSG Operations	OSGLHC	Deployed
OSG Security	OSGLHC	Deployed
OSG Software and Release	OSGLHC	Deployed
Particle	AS	Deployed
ppx	AS	Deployed
PV-Finder	IA	Testing
pyhf	AS	Deployed
recast	AS	Deployed
ROOT on Conda-Forge	AS	Deployed
Scikit-HEP	AS	Deployed
ServiceX	DOMA	Testing
SkyhookDM	DOMA	Development
Third Party Copy	DOMA, OSGLHC	Deployed
uproot	AS	Deployed
Vector	AS	Deployed

HL-LHC Tracking Challenge

Identifying and estimating charged particle trajectories (**tracking**) is the largest component of reconstructing analysis objects (“RECO”) from HL-LHC detector data (real or simulated)



IRIS-HEP is modernizing tracking following a **processor-centric development approach**. We exploit modern CPU vector units and/or accelerators in a heterogeneous environment

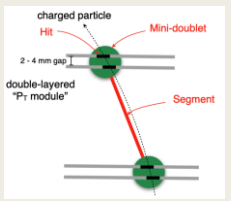


Modernized tracking codes unlocks the potential of accelerators for HEP, and **allows HL-LHC experiments to utilize a broader range of NSF facilities**.

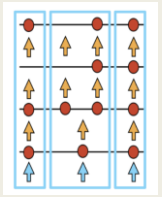
Project Run-Thru: Innovative Algorithms

Tracking

ACTS



Line-Segment Tracking

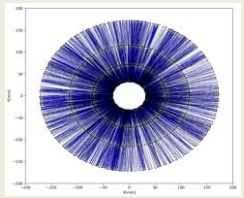


PV-Finder

Trigger

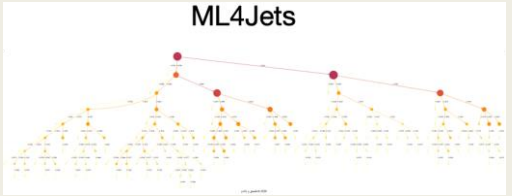
Allen GPU Trigger - Monitoring

Exploratory



Accelerated GNN Tracking

exploratory-ml



ML4Jets

We are not ready for the HL-LHC Data Rates

Like the LHC's Run 1, the HL-LHC will push the networking technology available at startup.

- We can take simple expected physics parameters (duration of data, event rates, event sizes) and computing models to generate minimum data movement scenarios.
- Unreasonable to expect we can take today's system and "throw more money" at the problem.

A coordinated approach with the whole ecosystem is needed!

Today: ~500 Gbps globally

WLCG estimate of required data rates:

4.8 Tbps

Minimal scenario: running flat-out, no margin for error, no unnecessary reprocessing.

9.6 Tbps

Flexible scenario: capacity for mistake, bursting, leveraging new opportunities. **How we run today.**

Project Run-Thru: DOMA

Intra-Facilities

XCache

XCache

tpc

Third Party Copy

Facilities

IDDS

IDDS

Coffea-Casa

Coffea-casa

Data Distribution

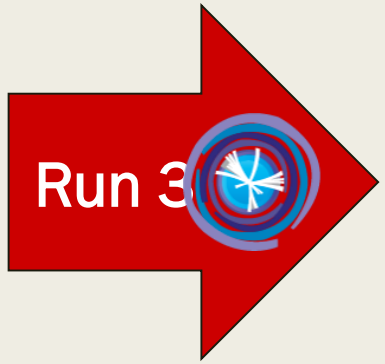
ServiceX

ServiceX

SkyhookDM

Transforming analysis for the HL-LHC Era

	LHC (Run 1&2)	HL-LHC (Run 4+)
Analysis Dataset size	10 TB	1,000 TB
Target Scan Turnaround time	Weeks	Hours
Analysis team size (physicists)	5-10	< 5
Primary analysis resource	Laptop	Analysis Facility



LHC analysis:

- Search & Precision Physics
- Simple ML techniques (BDT)
- Reproducibility in its infancy

HL-LHC analysis:

- Very High Precision Physics
- Modern ML (Deep Learning)
- Reproducible and Open Data

Project Run-Thru: Analysis Systems

Data

Awkward Array

DASK + Awkward Array

Awkward Array

uproot

uproot

VECTOR

vector

func_adl

Statistical Models & Fitting

cabinetry

cabinetry

pyhf

differentiable Likelihoods

pyhf

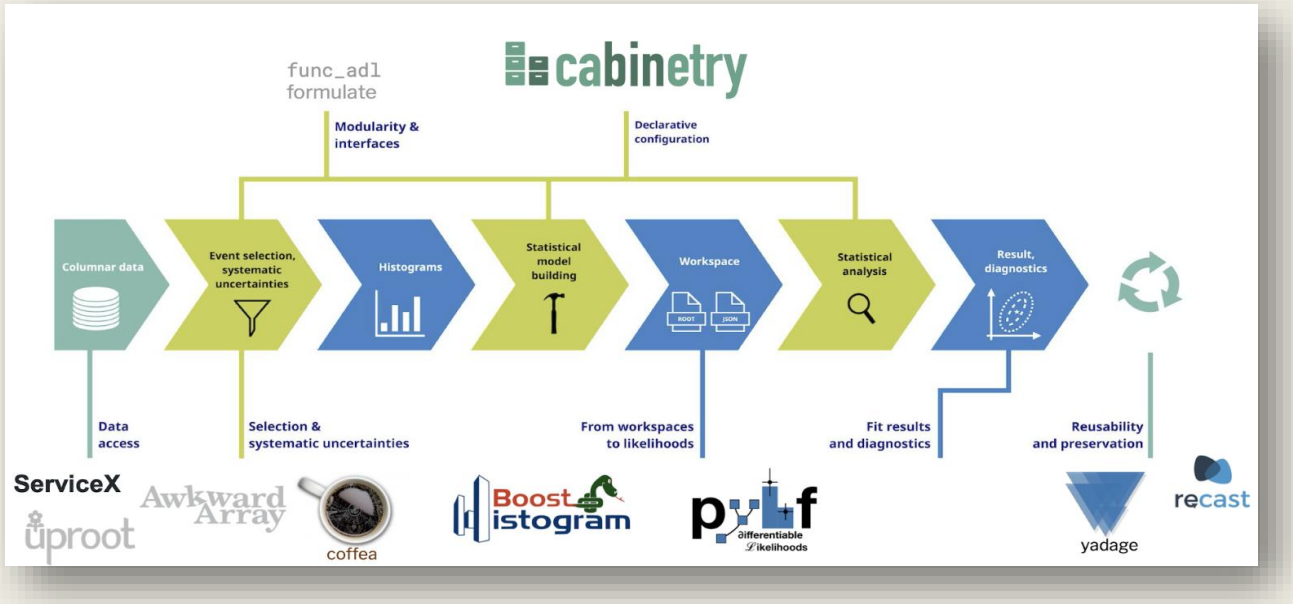
abcd-pyhf

Boost Histogram

Boost Histogram

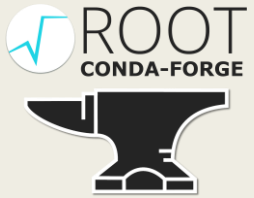
Project Run-Thru: Analysis Systems

Analysis Grand Challenge



Project Run-Thru: Analysis Systems

Community Packaging



ROOT Conda Forge

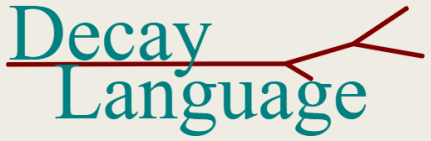
Scikit-HEP

Intellectual Hub

ADL Benchmarks

Awesome-hep

Generic Particle Support Packages



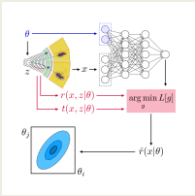
Decay Language

Particle

Project Run-Thru: Analysis Systems

AmpGen (and goofit)

Monte Carlo & Simulation

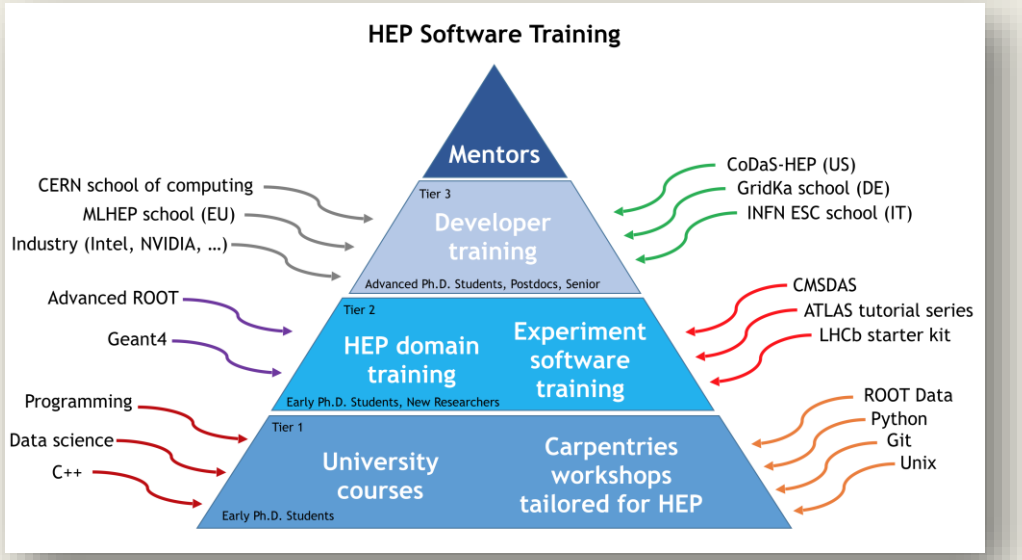


MadMiner

PPX

recast

Project Run-Through: SSC



See Killian's talk later today

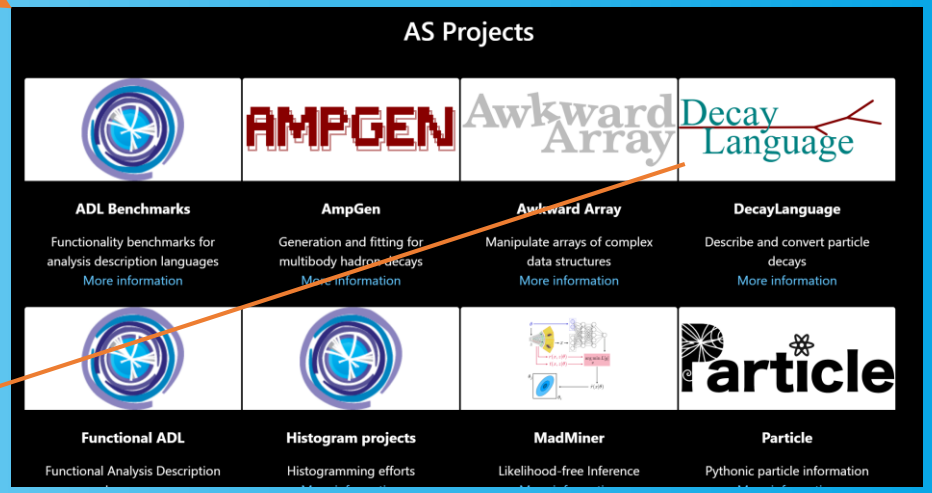
Project Run-Through: OSG-LHC

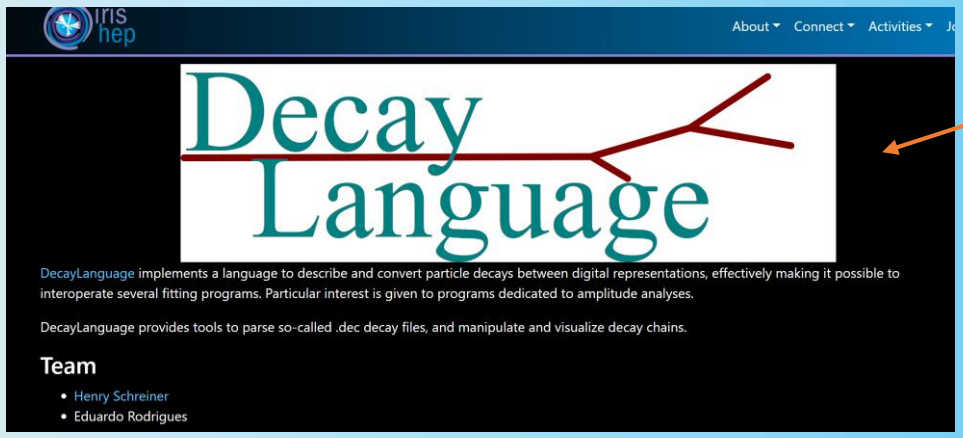
- IRIS-HEP's OSG-LHC area is where the LHC community intersects with the production cyberinfrastructure.
  - *Evolve & operate the integrated technologies and services.*
  - *Manage the software lifecycle necessary to sustain production.*
  - *Translate the work done for LHC to national community in conjunction with PATh (NSF OAC #2030508).*
- Provides ⅓ of the effort for the OSG Consortium:
  - *Integration of the OSG Software Stack.*
  - *Led Globus toolkit fork & retirement. Used by wider community, incl. XSEDE.*
  - *Drove the Containerization of OSG Services.*
  - *Via OSG Consortium, Operational Cybersecurity for the US LHC (jointly funded with PATh).*
  - *WLCG Accounting & Runtime environment.*
  - *Network performance Monitoring*
- CI Coordination with stakeholders (WLCG, ATLAS, CMS, in US and globally)

Where to look for more details

http://iris-hep.org

Almost all projects are detailed on our website





CWP Topics

- Physics Generators
- Machine Learning
- Visualization
- Detector Simulation
- DOMA
- Software Development, Deployment, Validation, Verification
- Software Trigger & Event Reconstruction
- Data-Flow Processing Framework
- Data and Software Preservation
- Data Analysis & Interpretation
- Conditions Data
- Security
- Training

CWP Topics – IRIS-HEP

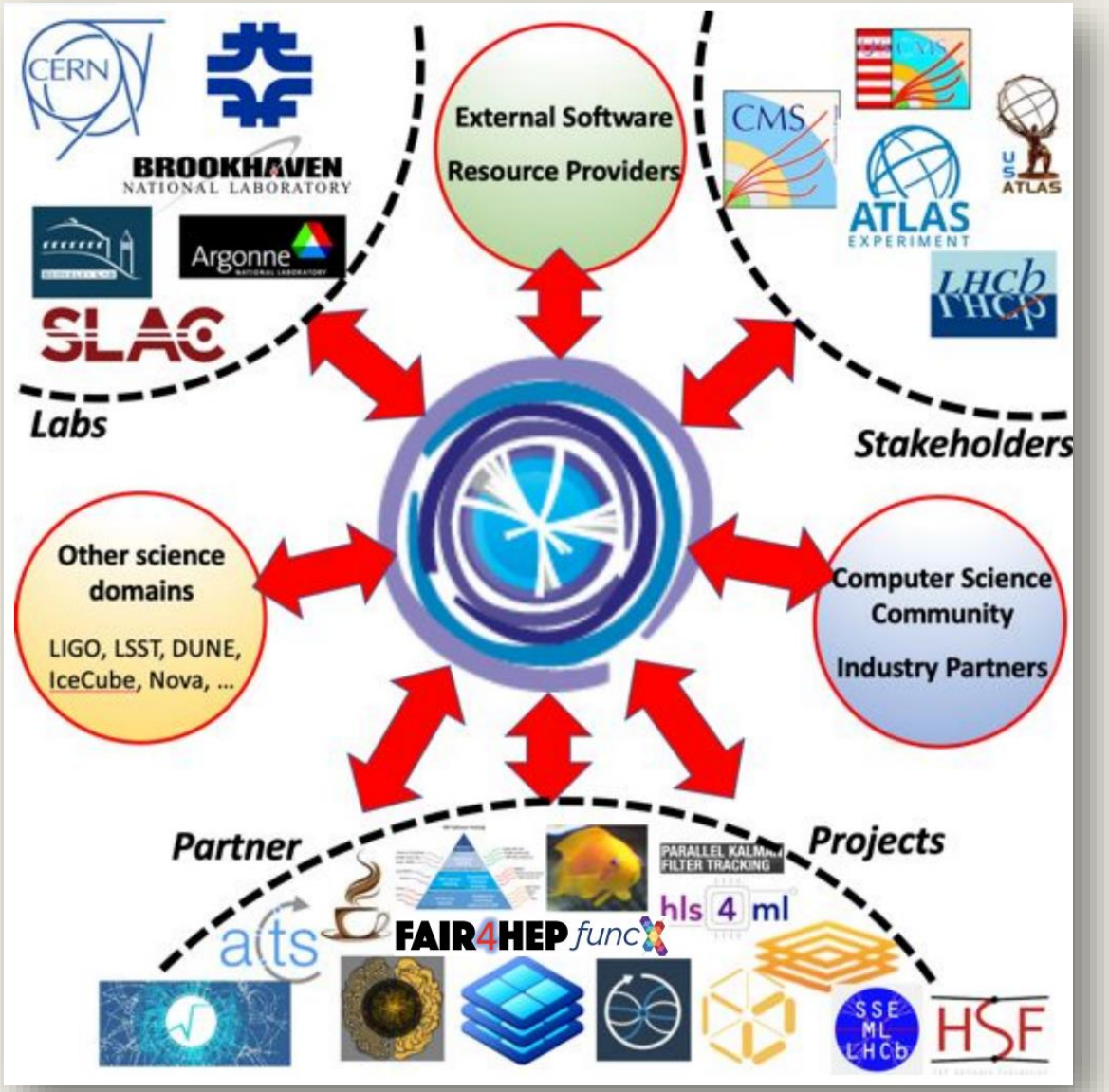
- Physics Generators
- Machine Learning
- Visualization
- Detector Simulation
- DOMA
- Software Development, Deployment, Validation Verification
- Software Trigger & Event Reconstruction
- Data-Flow Processing Framework
- Data and Software Preservation
- Data Analysis & Interpretation
- Conditions Data
- Security
- Training

Areas of large investment

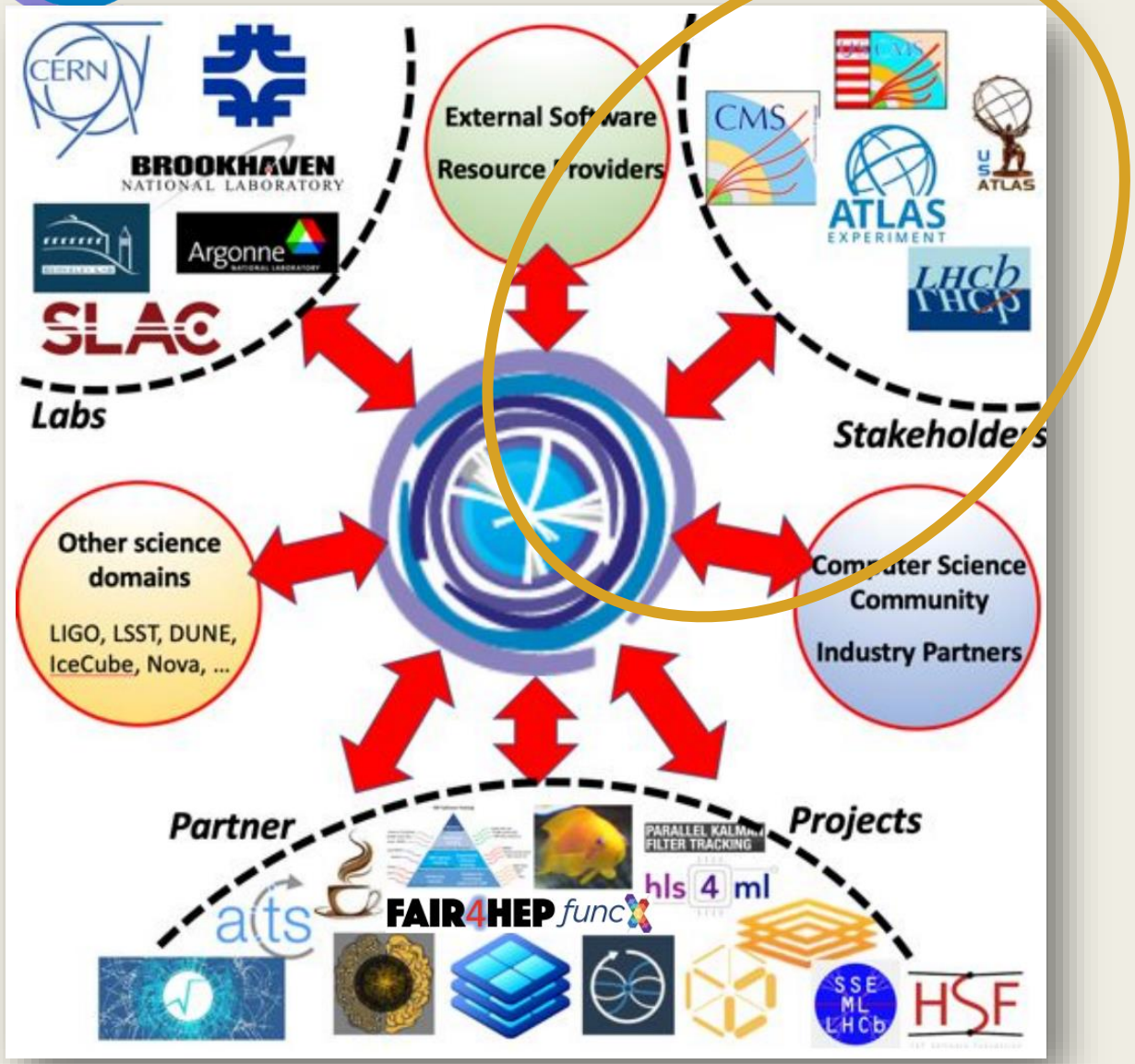
Intellectual Hub

We are building a common cyberinfrastructure vision and catalyze collaborations across the larger communities of users and developers.

- Community planning/“blueprint” workshops
  - Coordination with CERN, DOE.
  - User and Developer Communities (e.g. PyHEP)
  - Analysis, Data, Training Grand Challenges
- **Extensive Training & Fellows Program**
  - ~1600 people trained
  - >100 Fellows
- Topical Meeting Series
- Online Presence
  - **Website**: https://iris-hep.org
  - Twitter, Slack, YouTube
  - **Slack**: 450+ members and growing



US Operations & Experiments



IRIS-HEP Steering Board (formal)
- Meets once a quarter
- Regular explicit feedback from US Operations Programs
- Computing Coordinators from ATLAS, CMS, and LHCb are members
- Head of S&C of US Ops programs are members

IRIS-HEP Executive Board (formal)
- Ex-officio members from US Ops

US Ops S&C Meetings (informal)
- Members of IRIS-HEP attend

Collaboration (informal)
* IRIS-HEP people are members of the experiments

How are we delivering the software?

IRIS-HEP's goal is to help develop and improve software for the HL-LHC. *Not* maintain it for the full length of the HL-LHC.

- Technique varies by software and people
  - Some projects we do not lead
- IRIS-HEP has SSL and OSG
  - SSL can be used to test projects at scale before integration into outside
  - OSG's packaging experience

Basic Process

1. Develop idea & software
2. Build sustainable community of users and developers
3. Build appropriate structures for release
   1. Connections to AF, Ops programs, etc.
   2. Connections with user community…

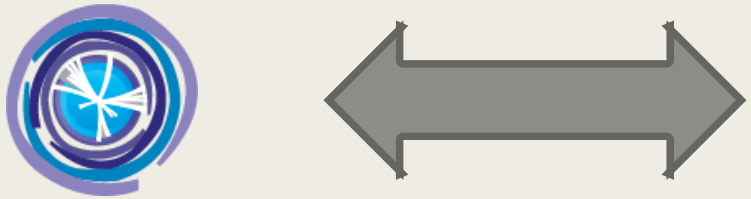
<u>Collaboration</u>:

- e.g for pyhf, awkward, etc. we work with end users to use it in their analyses

<u>Integration</u>:

- For TPC, work with OSG to integrate into software stack. Similar for tokens.
- Analysis Facility (coffee-casa) is prototyped on our SSL clusters before getting integrated with facilities in US ATLAS and US CMS

Communication

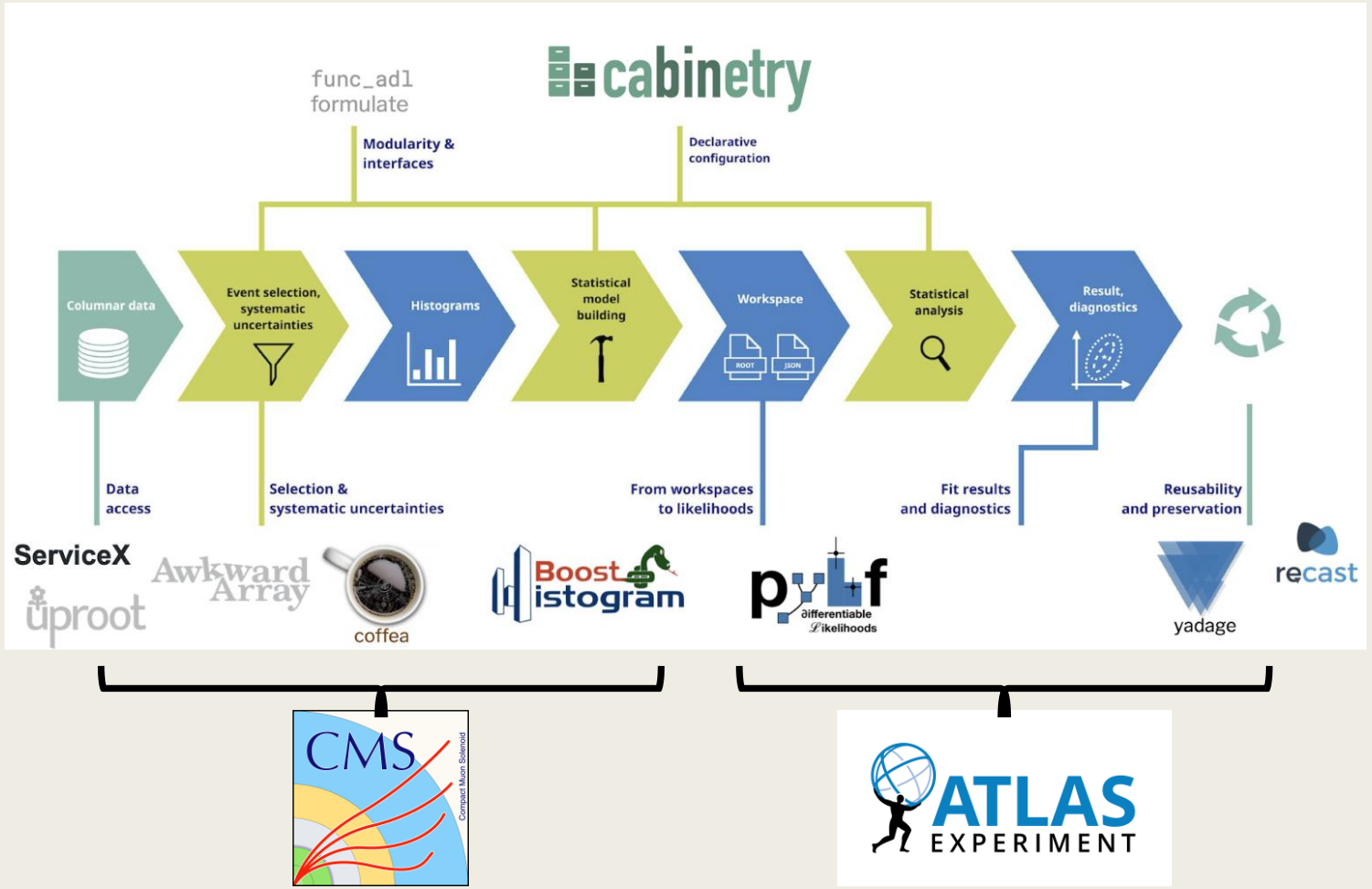


Awkward Array CSSI – NSF
C++ as a Service CSSI – NSF
Scikit-build CSSI – NSF
SkyHook POSE – NSF
funcX – NSF
ServiceX Dark Matter CSSI - NSF

Arrow Open Source Community
HEP Software Foundation

Experiments (LHC)
Nuclear and other communities – working on connections

Introducing the Experiments to each other



Opportunities

Engaging with the Community on Translational ML

Through the **strategic planning process**, we are identifying areas with the most opportunity for impact. Preliminary examples:

Algorithms:
- Expand work in ML for fast simulation and reco so it can be integrated with production workflows.

Analysis:
- Innovate new **fully differentiable analysis** pipelines
- **Provide a "landing pad" for new techniques** from AI research by integrating pytorch, TensorFlow, and JAX and other tools into our analysis workflows for ML training.
- Include systematic errors in ML training in analysis pipelines.

Facilities:
- Integrate workloads with ML **training and inference services** (either inside LHC facilities or from national facilities).
- Using composable facilities as a means for deploying new services from the data science community (MLFlow).

The IRIS-HEP's institute's broad view of all parts of HEP production, data, and analysis means a **unique and global approach to solving these problems in a cross-experiment way**

Strategic Opportunities

HL-LHC Cyberinfrastructure

Analysis Facilities

- Bring new, interactive paradigms from R&D into production.
- Demonstrate at HL-LHC-scale data rates.
- Start transforming existing facilities into **composable systems**, opening new capabilities for end-users

Analysis Tools and Pipeline

- Expand the AGC pipeline to cover a wider breadth of HL-LHC-style analyses.
- Incorporate more complex ML techniques into the AGC pipeline, bringing in modern tools from the data science community.

HL-LHC Analysis Techniques

Fully Differentiable Analysis

- **Vision**: make entire analysis differentiable so it can be optimized with ML tools instead of human-guided selections.
- End-to-end optimization of analysis taking into account systematic errors
- Ideally improves the interpretability of the ML we use.
- Techniques require support in entire toolchain – need an "institute-wide" vantage point.
  - Applicability outside HEP.

Systematic Errors

- The field is changing the approach of how systematic errors are calculated and stored at the LHC:
  - General Solution applicable across domains

DOMA Opportunities

Two activities for ‘23 and beyond (using the data challenge as a test bed):

- **Demonstrating software at scale:** Using the PRP (CC*DIBBs, #1541349) technologies and NRP resources (Cat II, #2112167), show our reference software stack can reach 500Gbps between sites.
- **Integration with engineered network paths:** With ESNet’s SENSE project, adapt the production stack to work with engineered network paths & apply them at production sites during the data challenge.

Leverages the national ecosystem of CI projects to do translational CS research for the broader community.

OSG Going Forward

While the **OSG-LHC area is the engine** keeping together the US LHC's production cyberinfrastructure today, there are several opportunities looking forward:

- Finish transformation of authorization technologies to utilize capability tokens.
  - *Enables a more trustworthy & more secure infrastructure.*
- Help deploy improved network monitoring, allowing differentiation of experiment-level flows, and "managed" networks.
- Expand the reach of our data federations:
  - *Grow the global network of caches for data delivery with hardware partner projects.*
  - *Increasingly contribute to the core of the XRootD to enable more rapid deployment of new features.*
  - *Unify the various LHC & OSDF data federations, reducing duplication of efforts.*
- Make full range of NSF compute facilities available to US LHC community.
  - *Deploy hosted CE services for the LHC - usable by others!*

Conclusions

- IRIS-HEP has ~31 FTE's working on a large number of CWP topics
  - *Far from covering every topic in the CWP!*
- Focused Research
  - *Analysis Systems, Innovative Algorithms, DOMA*
  - *Infrastructure to move ideas and projects from the lab to the community and production environments.*
- Designed from the ground up to collaborate
  - *With LHC operations programs*
  - *With other funded programs working towards the same goals*
  - *With individuals working towards similar goals*
  - *Working on connecting with other communities, like Nuclear Physics*
- For details use our website: http://iris-hep.org.