IRIS/HEP Retreat: Analysis Systems
Year 3 Plans, Year 4 & 5 Goals

Kyle Cranmer (NYU)
Overview

It was a productive “retreat”!

- many thanks for all those that provided input to prepare for the discussions
- Good progress in improving the interface between DOMA and AS
  - Eg. a nice talk in Parallel II by Carl Lundstedt
- Analysis Grand Challenge has already proved useful as a mechanism to understand interoperability and integration of our projects into a products and systems
- Analysis Systems parallels focused largely on making progress on our shared understanding of integration points.
  - We spent less time going through each individual product listing milestones for Y3 and goals for Y4 & Y5.
  - IMO a good use of our time for the retreat, but more to be done in preparation for PEP.
Highlights

- talk in Parallel II by Carl Lundstedt
- skyhook / ServiceX plugin draft

Making this a reality

- Finally, we use a slightly-patched version of the HTCondorCluster integration from desk to allow auto-scaling out to the local HTCondor pool.
- Jobs run in the container on the HTCondor worker node. HTCondor exposes an incoming port to provide the necessary connectivity.
- All of this is being incorporated into a Helm chart — many rough edges, but can eventually be portable to other sites.

Can Run Using Built-in Dask Clustering on Host CPUs

- SkyhookDM Client
  - Client maps tables to sets of objects
  - Map is also stored in objects
  - Client API designed for plugins
  - Allows pushdown to solve out tabular data operations
  - Reduces data movement (CPU cycles)

- BIRD-HEP
  - Plugins for ServiceX readers and writers

- CROSS
  - Plugins for Postgres, Spark, Pandas, HDF5
Highlights

- First serious previews of hep_tables and cabinetry
- More great stuff from pyhf

Reducing time to insight: Fitting as a service

1. pyhf installed on different clusters with GPUs around the world
2. User hits a REST API with JSON
3. pyhf workspace as a request
4. pyhf fits the workspace on the cluster on demand
5. Returns fit results over REST API to user

```python
pyhf HistFactory model spec is pure JSON: Very natural to use a REST web API for remote fitting!
```
A coherent ecosystem

- ServiceX
- awkward array
- coffea
- hep_tables
- pyhf
- yadage
- recast
- Data access
- Selection & systematic uncertainties
- Event selection, systematic uncertainties
- Columnar data
- Histograms
- Statistical model building
- Modularity & interfaces
- Declarative configuration
- Workspace
- Statistical analysis
- Result, diagnostics
- From workspaces to likelihoods
- Fit results and diagnostics
- Reusability and preservation
- Logical flow of processes

func_adl
formulate
cabinetry
Y3 Milestones

Integration:

- Many of the individual tools are at beta stage or better.
- Increase our efforts towards integrating tools into systems (vertical slices)
  - this is expected to expose areas where tools can be improved, modified, etc.
- More explicit coordination and planning with Analysis Facility / DOMA / SSL
- An important step towards almost any Grand Challenge involving Analysis Systems

Adoption:

- Some of our tools and projects are at “tipping point,” rapidly gaining traction within experiments and in user communities.
  - Example: pyhf adoption is rapid (papers, likelihood publishing, etc.)
  - Example: ATLAS is ramping up RECAST efforts (papers to come near end Y3)
  - Example: Scikit-hep as an example of community-driven software effort
- Good to invest effort in these areas for results and to build IRIS-HEP reputation
  - Development, Training, Documentation
  - Misc. experiment specific contributions also valuable for ”delivery to experiments”
Analysis Grand Challenge

- Generally a positive reaction to the proposed Grand Challenge
- Need to start working backwards into milestones (some will be in Y3)
  - Challenge motivates one or more blueprint meetings
    - DOMA/AS focused on named scenarios for functionality of “Analysis Facilities” and baseline what is needed for grand challenge
    - One focusing on other aspects of the challenge (autodiff issues, firming up other details)
  - Identify and settle on the actual facilities to carry out the challenge (Y3/Y4?)
- To do:
  - Quick prototype / exploratory work to understand better what is involved in making libraries with autodiff (eg. Jax, Numpy, Torch) work with awkward
    - Milestone: a demo of something like neos/INFERNO using awkward & func_adl
  - Clarify how SkyHook fits in to analysis systems
Y4 & Y5 Goals

Individual projects have Y4 and Y5 goals

- Some of these were discussed in parallel, but not uniformly
- Not clear it’s worth while to repeat them here anyways
- More work is needed to aggregate them for the PEP
  - aiming at ~1 milestone (+ 1 stretch) per significant project

Many of our Y4 and Y5 goals can be tied to Analysis Grand Challenge

- many of these are fairly obvious, but need to be written down.
- started prior to and during the session, but not done
- not all our goals are connected to Grand Challenge

We did not spend any significant time talking about metrics.
Backup
What would be potential Year 3 milestones for each of the projects? (First ideas, to be iterated with PIs and the whole team as this process moves forward.)

- Integration of `func_adl` specification for variable definition and selections with the emerging cabinetry specification for high-level template fit analysis
- Demonstration of differentiable analysis pipeline (e.g., cabinetry) ending with `pyhf` limit back-propagating through selection implemented with awkward, `func_adl`, etc.
  - connect to `pyhf/nees` demo. Need autodiff-able analysis over awkward arrays
  - connect to histogramming projects
  - Discussion in Slack to connect this with the Sally algorithm in MadMiner
- Documentation and training event using new tools
- Use of new IRIS-HEP tools (MadMiner, awkward, ...) for analysis in LHC experiment (may not be published by end of Y3)
- Snowmass (tools & REANA workflows for sensitivity studies)
Questions from Area Lead Chat:

Are there new opportunities where effort from IRIS-HEP can make an impact? Is the alignment of the focus areas in IRIS-HEP appropriate?

- Visualization tools (e.g., altair like declarative visualization)
  - yes for AS, but expansion of scope
- excursion alg. to streamline MC production for ATLAS or CMS reinterpretation campaign
  - yes for AS
- Improve efficiency of event gen. with ML-inspired tools & techniques
  - yes for AS, but an expansion into "theory" tools
- pyhf and astrophysics (HEALPix for boost histogram)
  - yes for AS, but secondary aim of IRIS-HEP
- MadMiner like tools for EIC
  - yes for AS, but secondary aim of IRIS-HEP. Brought up at 18 mo review
- python library for fastjet that plays well with columnar analysis
- Documentation efforts
  - yes, aligns with "lowering barriers" goal of AS
Analysis Systems Team

Bulk Data Processing

Reconstruction Algorithms

Analysis Code

Analysis code in HEP is often more free-form with less organized development:
- one-off approach limits functionality
- slow iteration cycle
- slow on-boarding and lack of interoperability
- difficult to reproduce and reuse

- primarily ROOT & C++
- lack of developer community
- overlapping solutions
- data redundancy
Analysis Systems

ad hoc analysis code

Analysis Systems strategies:
- improve functionality & interoperability
- more modular, less dependence on ROOT
- declarative: focus on what to do not how to do it
- align with modern data science practices
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.*
Value of IRIS-HEP as an Institute

IRIS-HEP as a tugboat:

- direct and navigate large efforts in the collaborations with significant inertia
- take advantage of consistent presence and messaging within the large collaborations
- Examples:
  - *pythonic analysis tools*
  - *software practices*
  - *industry-standards*
Value of IRIS-HEP as an Institute

IRIS-HEP as a lighthouse:

- provide cohesive, long-term vision for how software should evolve to meet needs of HL-LHC
- take advantage of holistic perspective of the institute
- Examples:
  - columnar analysis
  - declarative programming
  - differentiable programming
  - preservation & reuse
A coherent ecosystem

One of our analysis use cases involves a vertical slice from ServiceX to final limits for a real-world ATLAS Higgs analysis. See Alex Held’s poster.
The field is at a tipping point, DIANA/DASPOS/IRIS-HEP contributions have been transformational.

First results using the RECAST reinterpretation framework and publishing full statistical likelihoods (using pyhf)
Highlight

CERN Council appoints Fabiola Gianotti for second term of office as CERN Director General

Press release | 6 November, 2019

LATEST NEWS

- New open release allows theorists to explore...
- Relive 2019 at CERN
- Dive into the world of accelerators

Featured on CERN homepage
Scikit-HEP

A broad community project with heavy IRIS-HEP involvement.

Scikit-HEP project - welcome!

The Scikit-HEP project is a community-driven and community-oriented project with the aim of providing Particle Physics at large with an ecosystem for data analysis in Python. The project started in Autumn 2016 and is in full swing.

It is not just about providing core and common tools for the community. It is also about improving the interoperability between HEP tools and the scientific ecosystem in Python, and about improving on discoverability of utility packages and projects.

For what concerns the project grand structure, it should be seen as a toolset rather than a toolkit. The project defines a set of five pillars, which are seen to embrace all major topics involved in a physicist's work. These are:

- **Datasets**: data in various sources, such as ROOT, Numpy/Pandas, databases, wrapped in a common interface.
- **Aggregations**: e.g. histograms that summarize or project a dataset.
- **Modeling**: data models and fitting utilities.
- **Simulation**: wrappers for Monte Carlo engines and other generators of simulated data.
- **Visualization**: interface to graphics engines, from ROOT and Matplotlib to even beyond.

Toolset packages

To get started, have a look at our GitHub repository. The list of presently available packages follows, together with a very short description of their goals:
The Future

Tight integration of

- Simulation
- Machine Learning
- Statistical Inference

Repository and tutorials:
github.com/johannbrehmer/madminer

Documentation:
madminer.readthedocs.io

Installation:
pip install madminer

Deployed with Docker, yadage, REANA:
github.com/irinaespejo/workflow-madminer

Thanks to Kyle, Gilles, Felix, Irina, and Sam for material and inspiration for slides!

Slides from Johann Brehmer’s Keynote talk at ACAT on Constraining Effective Field Theories with Machine Learning
Major Activities

- Development of declarative specifications for different stages of analysis
- Identification and benchmarking of traditional implementations for benchmark example use-cases that span the scope of AS
- Implementation of prototype components & integration
  - connection with DOMA (particularly ServiceX)
- Benchmarking and assessment of prototype implementations and declarative specifications for the same example use cases
  - connection with SSL (dedicated Blueprint Activity)
- Exploratory research in machine learning that may impact how analysis is performed
- Engagement with community of early adopters and developers
Are there internal or external collaborations associated with each project or activity? For external collaborations, is IRIS-HEP leading, contributing or simply “connecting/liaising”?

Internal:
- **SSL**: benchmarking and scaling, REANA testbeds, etc.
- **SSL & DOMA**: ServiceX

External:
- **DIANA/HEP**: last bits of funding on NCE supporting various items very aligned
- **SCAILFIN**: developing products, good synergy w/ IRIS-HEP. **REANA** dev team
- **INSPIRE-HEP, HEPData, CAP, Invenio**: Advisory boards, join in development
- **ATLAS stats effort**: [docker containers for RooFit-based statistical analysis & combinations](https://example.com) and development of pyhf tools. IRIS-HEP (Matthew, Kyle, Alex) & Lukas & Giordon are leading
- **HEP Statistics Serialization Standard (HS3)** similar cast of characters
- **scikit-hep**: useful umbrella (not seen as US, ATLAS/CMS, or HSF) IRIS-HEP leading by example  
  - **Awkward:**
    - formal collaboration with Amy Roberts at UC Denver on **Kaitai Structs**
    - frequent collaboration with **LPC/Coffea** (Lindsey Gray)
    - close liaisons with **Anaconda.com**: Numba and Dask developers
    - intermittent contact with **Oxford Big Data Institute** (genetics, developers of **Zarr**)


Projects

- Analysis systems are connected to analysis use cases
- Systems are composed of components
- Most of these projects refer to those components
  - many projects include people beyond IRIS-HEP
- Milestones and activities mainly oriented towards integration, evaluation, with a global overview of the vertical slice