Analysis user experience with Python HEP data science tools in ATLAS

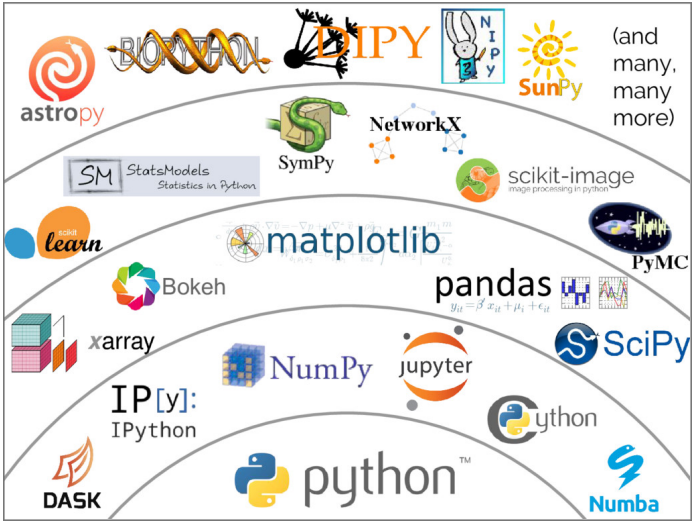
Matthew Feickert

University of Illinois at Urbana-Champaign

IRIS-HEP Analysis Grand Challenge Tools 2022 Workshop

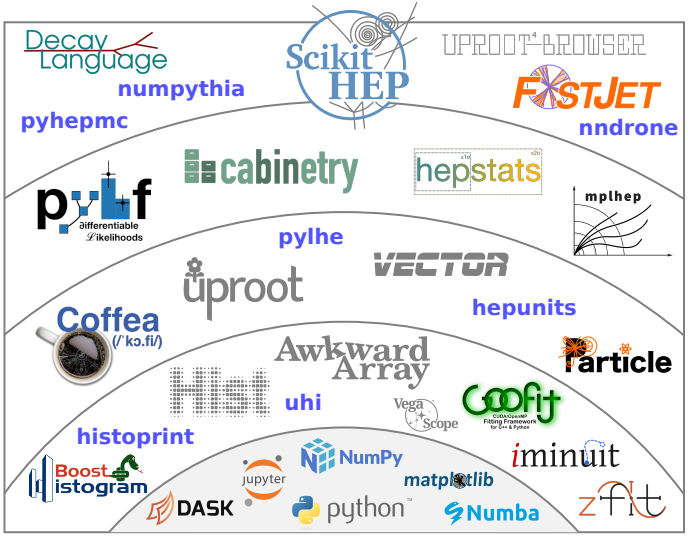
April 26th, 2022

PyHEP ecosystem



In his PyCon 2017 keynote, Jake VanderPlas gave us the iconic "PyData ecosystem" image

PyHEP ecosystem



In his 2022 PyHEP topical meeting update, Jim Pivarski gave us a view for the Scikit-HEP ecosystem

PyHEP ecosystem (Jim's summary)

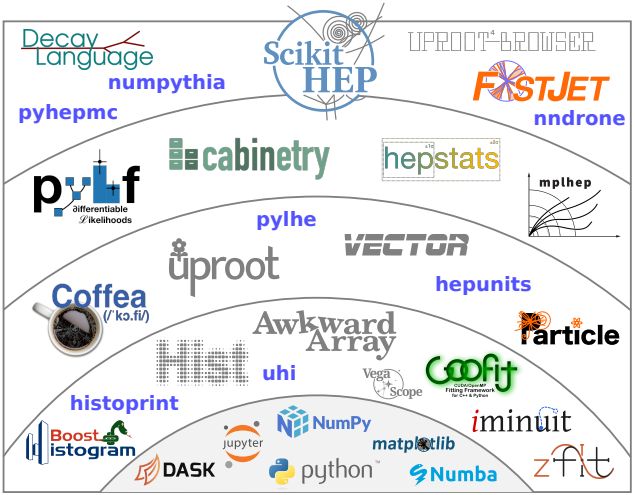
5 HEP-specific UI applications or packaged algorithms

4 HEP-specific for common problems

3 HEP-specific, foundational

2 needed to create, but not really HEP-specific

1 non-HEP software we depend on



PyHEP ecosystem

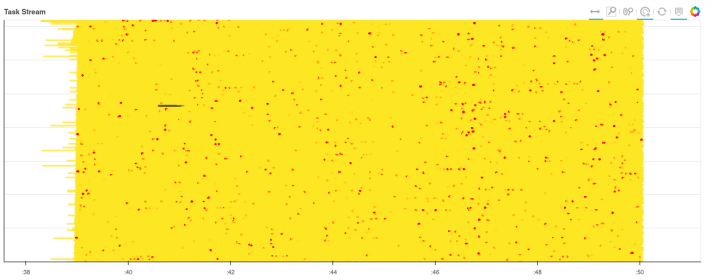
These ecosystems are still growing and evolving, so these visual representations of stacks are not supposed to be static or fully inclusive (plenty of IRIS-HEP and other tools not shown).

They are useful exercises for consideration of the ecosystem state, growth, and needs.

This talk is meant to try to highlight how ATLAS is exploring these ecosystems now and give pause for discussion.

Scaling workflows with Dask for columnar analysis

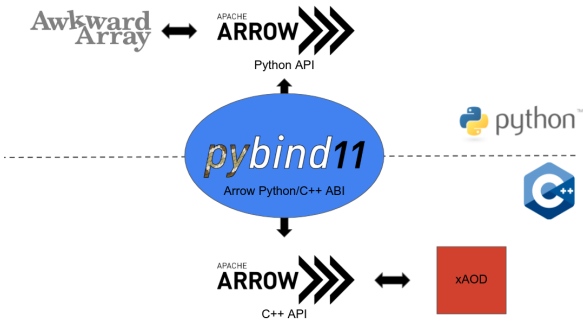
- Nikolai Hartmann has shown promising tests on PHYSLITE (data format) workflows
  - Nikolai's talk from vCHEP 2021
  - Nikolai's talk from IRIS-HEP AGC Tools 2021 Workshop ATLAS interal session
- Shown can also scale out via Dask (or Ray) to O(1000) CPUs
  - Think good integration with coffea based workflows as well



Plot of 100k Dask tasks (ATLAS Software & Computing Week October 2021)

Ongoing studies on providing columnar analysis

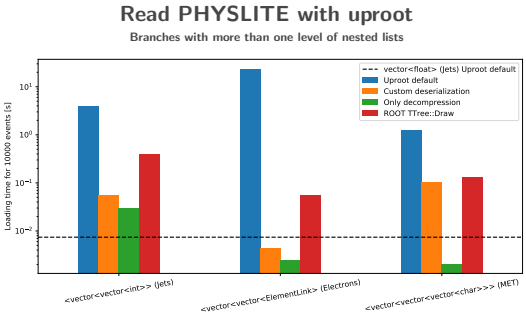
- Growing group of ATLAS people looking at providing columnar analysis for ATLAS tools
- Have ongoing studies of `awkward`-xAOD APIs with a focus on finding zero-copy methods
- Jim has suggested alternative approaches that might simplify things



Example possible approach with PyArrow and pybind11 (AGC Tools 2021 Workshop ATLAS internal session)

uproot used galore

- ATLAS ❤ uproot
- Safe to say that any analysis using Python is using uproot at some level
- Nikolai Hartmann's (internal) studies show reading PHYSLITE with uproot feasible

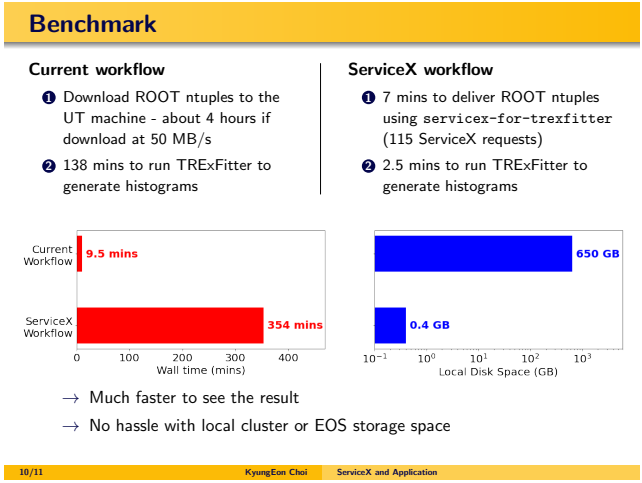


→ efficient columnar reading of PHYSLITE ROOT files possible!

5 / 12

ServiceX and coffea exploration

- Ongoing studies by Kyungeon Choi on using ServiceX + TRexFitter
- Shown at vCHEP 2021
  - Towards Real-World Applications of ServiceX, an Analysis Data Transformation System
  - arXiv:2107.01789
- Active studies by Kyungeon on ServiceX+coffea (will be shown in ATLAS internal session later today)



vCHEP 2021 (Times are flipped in left plot)

Pythonic visualization

- `matplotlib` visualization has been a growing cultural movement in ATLAS (and broader HEP) in the past years
  - Scikit-HEP: `mplhep`
  - `atlas-mpl-style` (uses UHI)
  - `uhepp` (used for an ATLAS analysis)
  - `atlasify`
  - I'm sure there are more
- ATLAS PUB notes with `mplhep` plots
  - Simultaneous Jet Energy and Mass Calibrations with Neural Networks
- Despite `mplhep`'s popularity and use in the Scikit-HEP stack, there hasn't been a clear rallying around a single library
  - Pro: Bazaar of ideas
  - Con: Loss of contributions and upstreaming work to `matplotlib`

Slow uptick in use of pyhf for analysis

- pyhf has been cited as being used by ATLAS in 4 papers and 2 PUB notes
  - arxiv:2203.01009 (2022)
  - arxiv:2203.00587 (2022)
  - ATL-PHYS-PUB-2021-038 (2021)
  - arxiv:2106.01676 (2021)
  - arxiv:2102.01444 (2021)
  - ATL-PHYS-PUB-2019-029 (2019)
- pyhf has received well designed contributions from ATLAS contributors
- cabinetry is still being socialized in the experiment
  - Alexander Held has been invited to give presentations on cabinetry to different subgroups. (ATLAS internal link)
  - Feedback from talks indicates that there is some mental shifting from fitting "frameworks" (e.g. HistFitter) to libraries (i.e. pyhf, cabinetry)

Summary and Story

- ATLAS is mostly using the PyHEP / Scikit-HEP stack at the beginning (`uproot`) and end (`pyhf`, `cabinetry`)
- Plenty of ongoing R&D efforts that are ongoing that deeply utilize the PyHEP and IRIS-HEP tools
  - Looking at Dask, `uproot`, and `coffea` for working with PHYSLITE
  - ServiceX+`coffea` for analysis workflows
- ATLAS and IRIS-HEP would benefit from ATLAS performing an analysis user survey like CMS did (will follow up on this)
- ❗ Personal opinion injection: I think most people are interested in using PyHEP tooling in some point in their analysis. Uptake hesitancy shouldn't be viewed as lack of interest, but probably more uncertainty on how to integrate tools into current workflow. So workshops like this that socialize tools and show technical examples are critical for adoption.

Additional examples on columnar analysis in HEP

- Columnar Analysis with PHYSLITE, Nikolai Hartmann, IRIS-HEP AGC ATLAS-Internal Meeting 2021
- The Case for Columnar Analysis , Nick Smith, HOW 2019
- The NanoEvents object, Nick Smith, PyHEP 2020
- Columnar Analysis at Scale with Coffea, Mat Adamec, PyHEP 2020