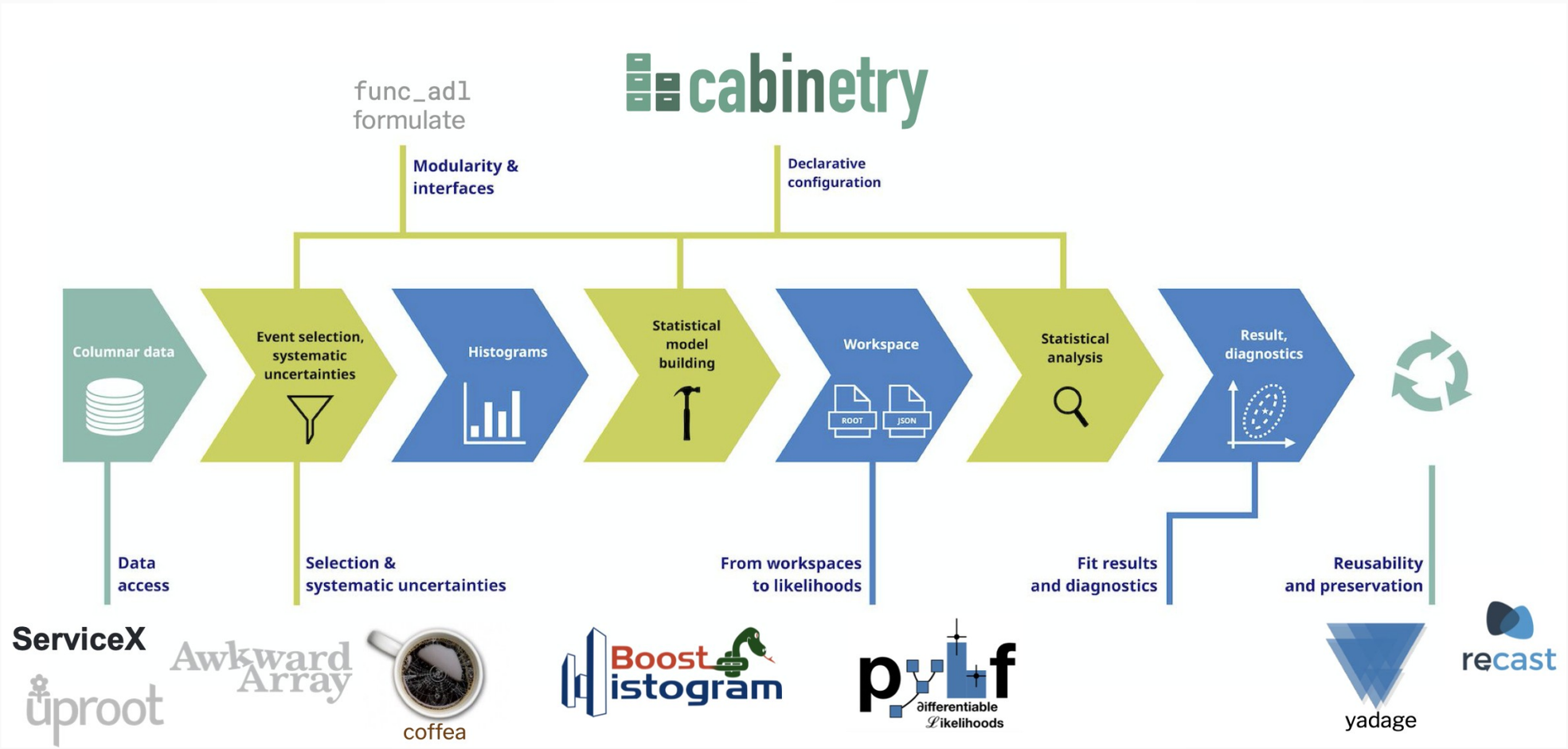
Benchmarking Prototype Analysis Codes on the Coffea-Casa Analysis Facility

Storm Lin
IRIS-HEP Summer 2021 Fellowship Project
Mentors: Alexander Held, Oksana Shadura

Motivations

- Works towards IRIS-HEP's Analysis Grand Challenge
- Investigating new software tools to effectively analyze the large volume of data from the HL-LHC
- Several analysis tool prototypes are in development, but need to be further tested and improved

Analysis Grand Challenge Tools



Project Objectives

- Recreate the ATLAS Open Data analysis of the discovery of the Higgs Boson via the H→ZZ→llll channel using prototype tools
- Demonstrate Coffea, Cabinetry, and ServiceX being used in a realistic data analysis and interacting with each other
- Run benchmark test comparisons with existing analysis methods
- Test these analysis codes on the Coffea-Casa analysis facility

ServiceX

- Efficient HEP data delivery service
- Can pre-select specific events and columns
- Works with multiple data formats

Coffea

- Columnar Object Framework For Effective Analysis
- Tools for easily and efficiently analyzing and plotting HEP experiment event data
- Facilitates horizontal scaling for servers or clusters with tools like Dask

Cabinetry

- Library for constructing HistFactory models and binned template fits
- Uses declarative instructions to build models
- Includes utilities to visualize models

Coffea-Casa Analysis Facility

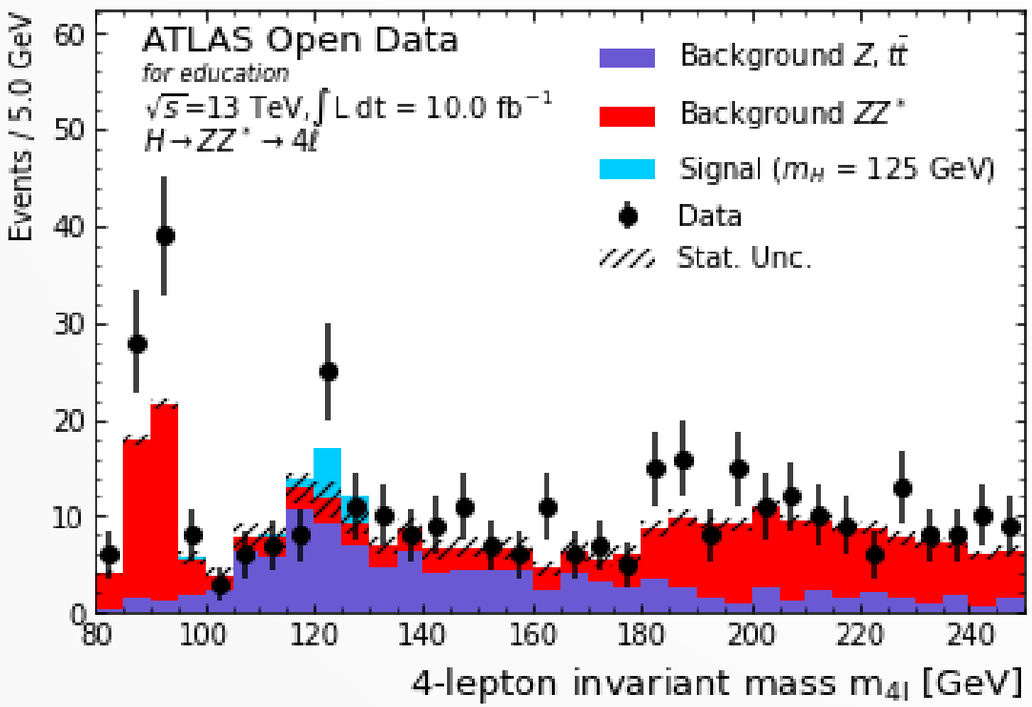
- Prototype analysis facility for Coffea-based analyses
- Speed up analysis times
- Uses Dask and Jupyter notebooks on a Kubernetes cluster

Recreated Analysis Steps

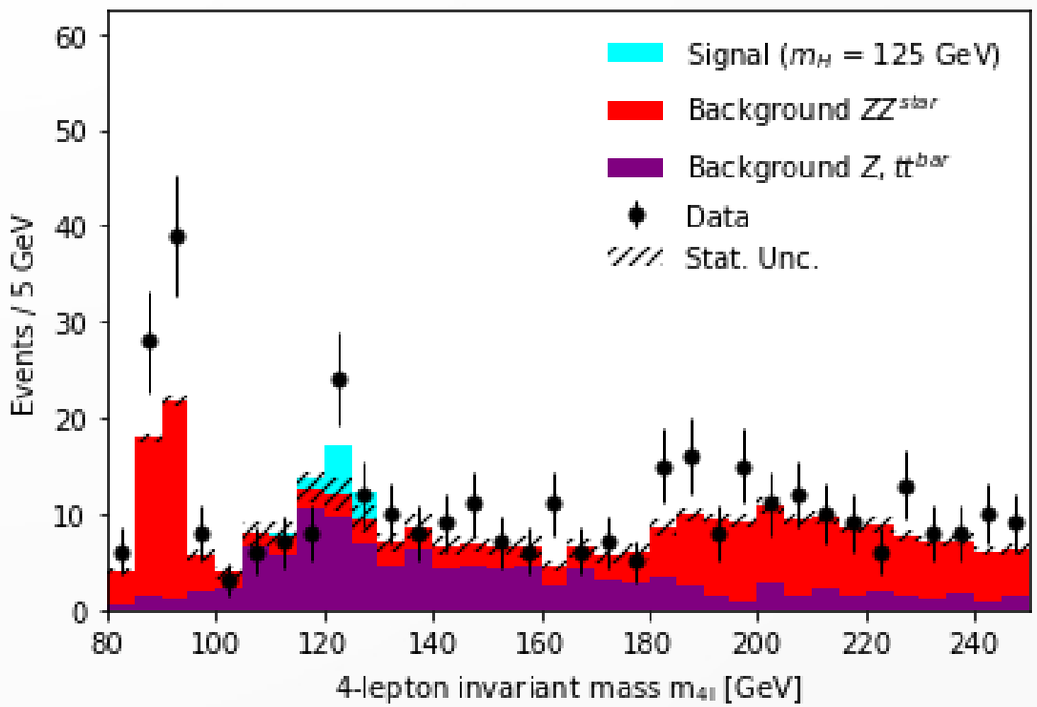
- Process Monte Carlo simulation sample and experimental event data from .root files
- Cut to only events where the sum of lepton charges is 0 and the lepton flavors are eeee, eeμμ, or μμμμ
- Compute the invariant mass of the 4 leptons in each event
- Create a histogram of these invariant masses

Output Histograms

Original Plot



Plot Created With Coffea



Performance Benchmarks

- Run times locally (w/ data from internet)
  - Original Notebook: 33 seconds (16 minutes)
  - Coffea: 6 seconds (46 seconds)
  - Coffea w/ ServiceX: 2 seconds (33 seconds)
  - Cabinetry w/ Uproot: 4 seconds
  - Cabinetry w/ Coffea: 10 seconds
- All run times were the same on Coffea-Casa as on a personal computer

Difficulties with Coffea and Cabinetry

- Coffea's processor class functions have input and output types that were difficult to understand initially
- Cabinetry cannot easily assign different histogram weights to different data files within the same MC sample
- Had to create my own patch to make the Coffea backend of Cabinetry accept the same built-in functions as Uproot
- Cabinetry cannot readily use https or ServiceX data sources like Coffea can

Difficulties with Coffea and Cabinetry, cont.

- Cabinetry specifies histogram variables and cuts as strings which cannot contain custom functions
  - “sqrt((lep_E[:,0] + lep_E[:,1] + lep_E[:,2] + lep_E[:,3])**2 - (lep_pt[:,0]*cos(lep_phi[:,0]) + lep_pt[:,1]*cos(lep_phi[:,1]) + lep_pt[:,2]*cos(lep_phi[:,2]) + lep_pt[:,3]*cos(lep_phi[:,3]))**2 - (lep_pt[:,0]*sin(lep_phi[:,0]) + lep_pt[:,1]*sin(lep_phi[:,1]) + lep_pt[:,2]*sin(lep_phi[:,2]) + lep_pt[:,3]*sin(lep_phi[:,3]))**2 - (lep_pt[:,0]*sinh(lep_eta[:,0]) + lep_pt[:,1]*sinh(lep_eta[:,1]) + lep_pt[:,2]*sinh(lep_eta[:,2]) + lep_pt[:,3]*sinh(lep_eta[:,3]))**2)/1000”

Difficulties with ServiceX and Coffea-Casa

- Running an analysis with ServiceX data sources can be quite confusing at first
  - Have to create an ObjectStream with a dummy data source to set up multiple data sources
- ServiceX always turns ROOT data sources into .parquet files
  - Cannot use Coffea's processor.run_uproot_job even though the original files are .root files
- Programs using patches or draft features may require tuning to run on Coffea-Casa

Outcomes

- Successfully created, ran, and studied the performance of analyses using prototype software tools
- Contributed new example notebook to the ATLAS Open Data collection
- Led to the creation of a new milestone task aimed at streamlining the interactions found in this analysis

Links/References

- All code created for this project with full notes and citations:
  https://github.com/stormsomething/CoffeaHZZAnalysis
- Original Open Data H→ZZ analysis:
  https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/blob/master/13-TeV-examples/uproot_python/HZZAnalysis.ipynb
- More about Coffea: https://github.com/CoffeaTeam/coffea
- More about Cabinetry:
  https://github.com/scikit-hep/cabinetry

Links/References, cont.

- More about ServiceX: https://github.com/ssl-hep/ServiceX
- More about Coffea-Casa: https://github.com/CoffeaTeam/coffea-casa
- Analysis Grand Challenge milestone: https://github.com/iris-hep/analysis-grand-challenge/issues/1
- PyHEP 2021 (talks here helped this project a lot): https://indico.cern.ch/event/1019958/