The IRIS-HEP Analysis Grand Challenge

Alexander Held[1], Oksana Shadura[2]

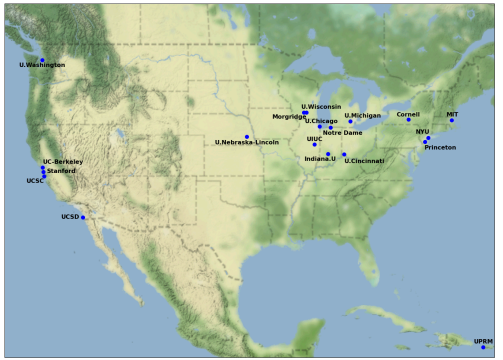
[1] University of Wisconsin–Madison
[2] University of Nebraska–Lincoln

ICHEP 2022 (Bologna)
https://agenda.infn.it/event/28874/contributions/169204/
July 09, 2022

This work was supported by the U.S. National Science Foundation (NSF) Cooperative Agreement OAC-1836650 (IRIS-HEP).

IRIS-HEP and the Analysis Grand Challenge

- **IRIS-HEP**: *"Institute for Research and Innovation in Software for High Energy Physics"*
  - Software institute funded by the US National Science Foundation
  - Research & development for the HL-LHC
    - innovative algorithms for data reconstruction & triggering
    - analysis systems to reduce time-to-insight and maximize physics potential
    - data organization, management and access systems
  - more information: https://iris-hep.org/

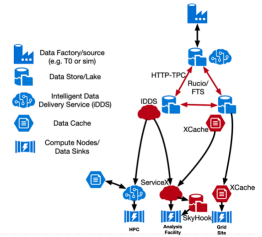


institutes participating in IRIS-HEP

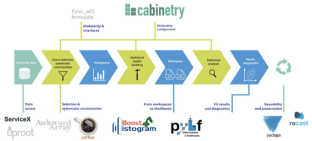
IRIS-HEP and the Analysis Grand Challenge

- **AGC**: *"Analysis Grand Challenge"*
  - historically, an integration exercise
    - test realistic end-to-end analysis pipelines aimed at HL-LHC use
    - combine technologies being developed in various ares of IRIS-HEP & adjacent ecosystem
    - identify & address performance bottlenecks and usability issues
  - organized jointly with the US ATLAS & US CMS operations programs

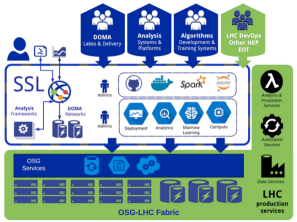
Data Organization, Management and Access (DOMA):
data delivery

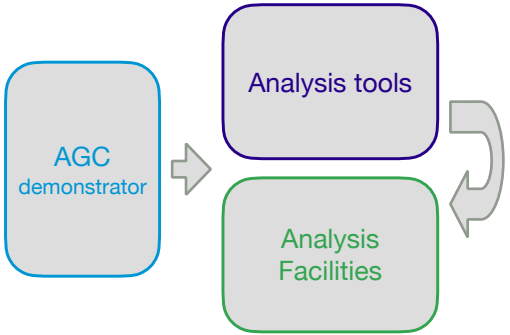


Analysis Systems (AS):
tools



Scalable Systems Laboratory (SSL):
deployment techniques and resources

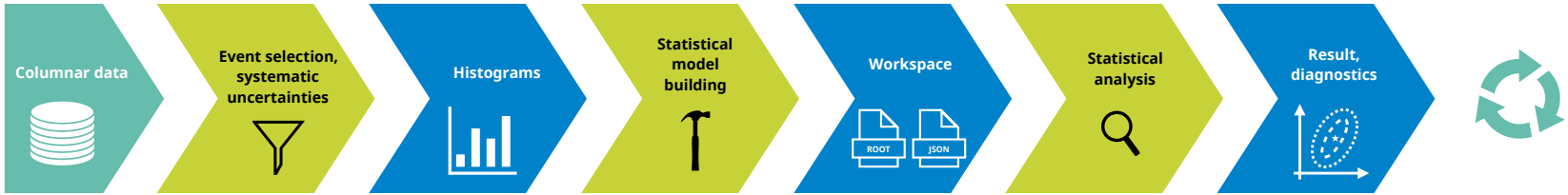




AGC combining IRIS-HEP focus areas

“Analysis” in the AGC context

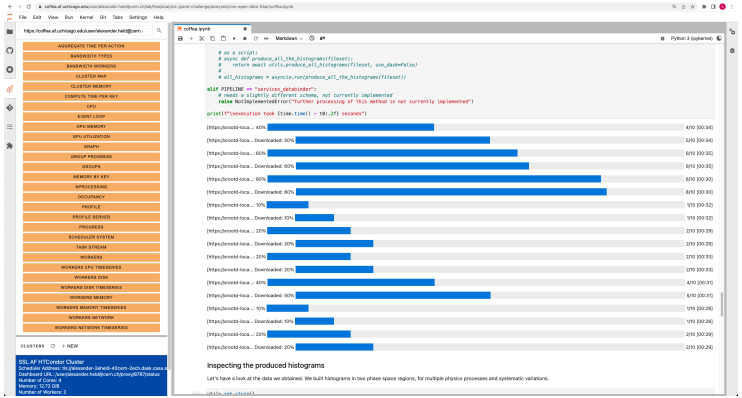
- In view of the HL-LHC: “analysis” **starts** from centrally produced **common data samples**
- Includes all **subsequent steps** to produce results needed for publication
  - Extract relevant data
  - (Re-) calibrate objects & calculate systematic variations
  - Filter events & calculate observables
  - Histogramming (for binned analyses)
  - Construct statistical model + perform statistical inference
  - Visualize results & provide all relevant information to study analysis details
- Do all these steps in a **reproducible** way



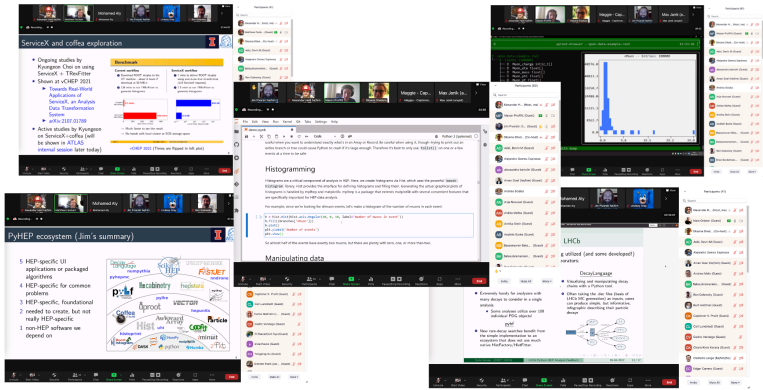
Moving beyond an integration exercise

- Investigating the possibility of **“interactive analysis”**: turnaround time of minutes or less
  - made possible by highly parallel execution in short bursts, low latency & heavy use of caching
- We hope that the AGC can be **useful to the broader community**!
  - testbed for software library development
  - environment to prototype analysis workflows
  - functionality & integration test for analysis facility development

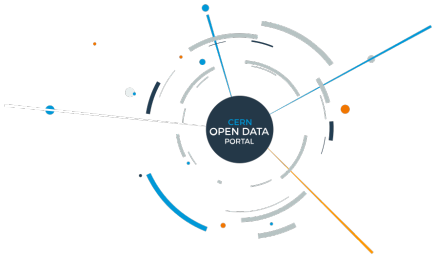
interactive analysis in a notebook



AGC tools 2022 workshop

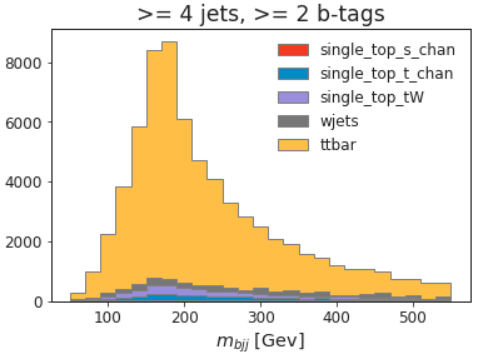


The AGC analysis setup



- Main AGC analysis task: **ttbar cross-section measurement** in single lepton channel
  - includes simple top reconstruction
  - setup chosen as it captures relevant workflow aspects and can easily be extended
    - e.g. conversion into a BSM search
  - analysis task prominently features handling of systematic uncertainties

- Analysis is based on **Run-2 CMS Open Data** (~400 TB of MiniAOD available)
  - Open Data is crucial: everyone can participate
  - currently using 4 TB of ntuple inputs (pre-converted, ~1B events before cuts)



reconstructed top mass

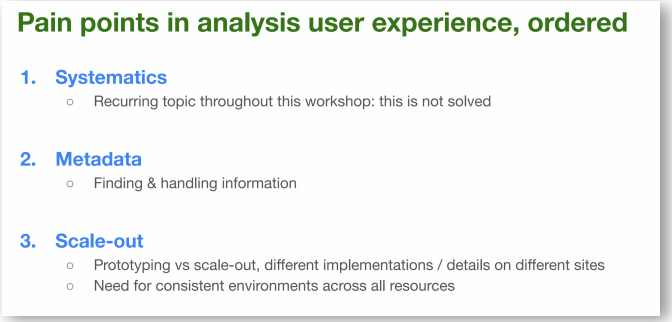
- Goal of setup is showing **functionality**, not discovering new physics
  - want to capture *workflow*, but can use made-up tools for evaluating calibrations & systematic uncertainties

Systematics and other analyzer user experience aspects

- Handling **systematic uncertainties** is a **key challenge** in analysis workflows
  - AGC analysis task includes different types of systematic uncertainties to mirror practical requirements
    - weight-based uncertainties
    - object-based systematic variations affecting kinematics (+ thereby event selection / observables)
    - non-histogram-based uncertainties (e.g. cross-section uncertainties)

- **Metadata** handling
  - capturing various bookkeeping aspects in analysis task

- **Scale-out**: from laptop to analysis facility
  - challenge: write analysis implementation that can run anywhere



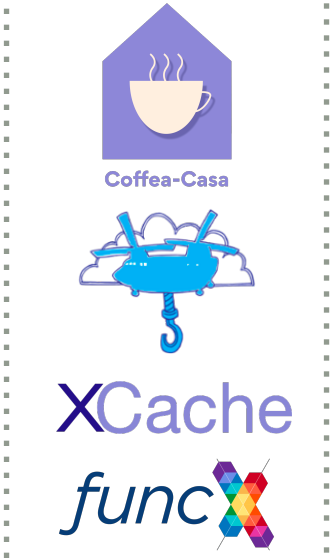
Analysis Ecosystem Workwshop II
User experience & Declarative Languages summary

Tools and services in our implementation

- Employing stack of **Python HEP libraries** for analysis tasks
- **ServiceX** used as data delivery service
- Execution on a **coffea-casa analysis facility**

HEP-specific libraries used for data analysis

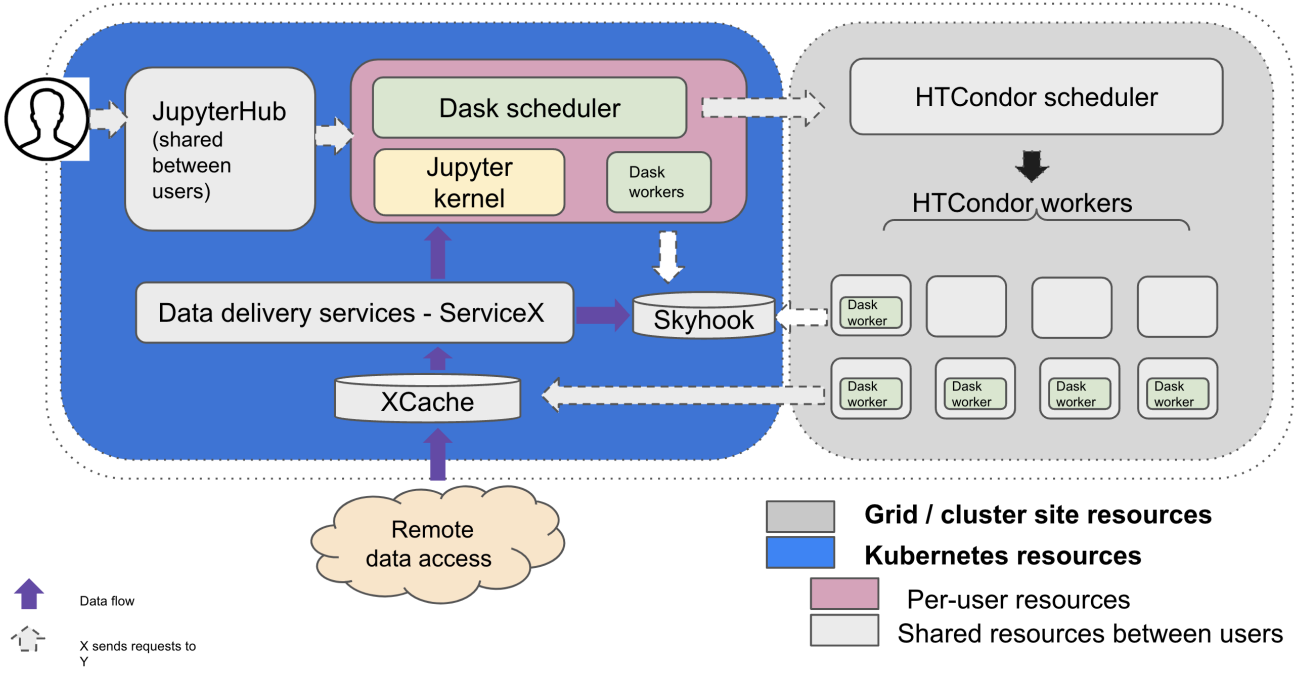
data delivery services



optional services

Analysis Facilities for execution

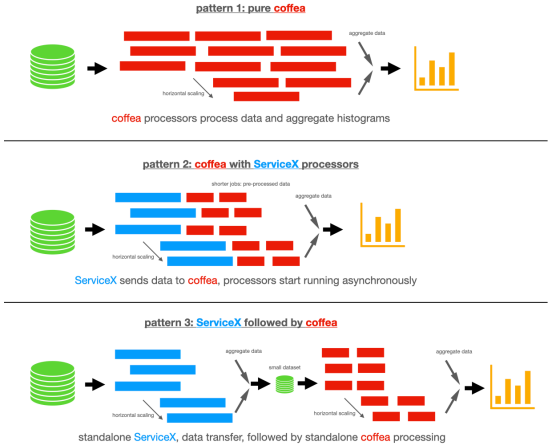
- **coffea-casa** is a **prototype analysis facility** for the HL-LHC
  - interactive facility for columnar analysis providing analysis tools & scaling to computing resources
  - more information: https://iris-hep.org/projects/coffea-casa.html



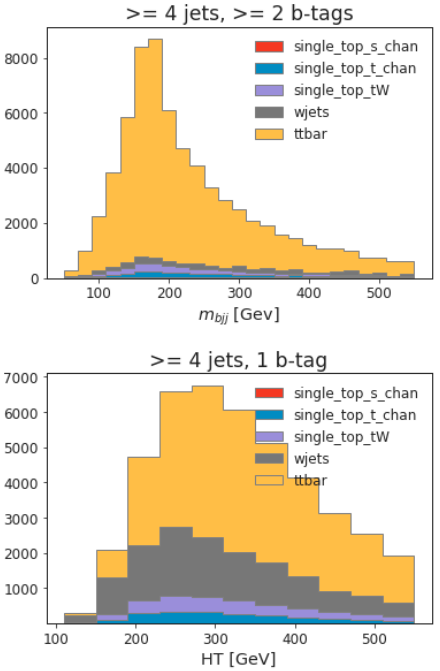
Implementation: ttbar analysis in a notebook

- **From data delivery to statistical inference** in a notebook

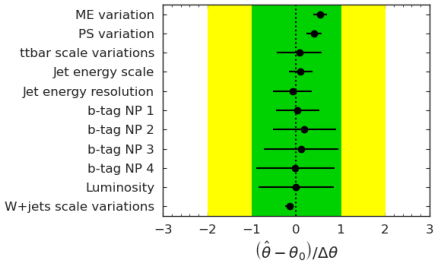
multiple supported processing schemes



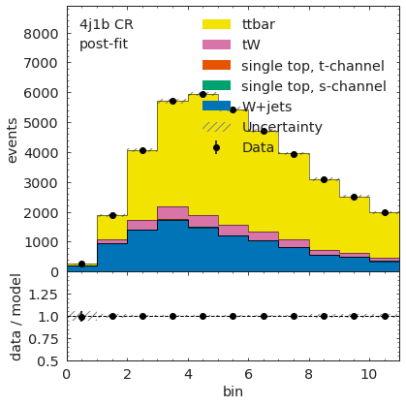
reconstructed observables



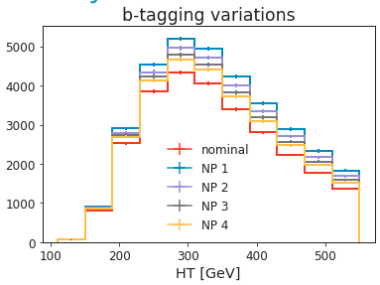
nuisance parameter pulls



post-fit distributions



systematic variations

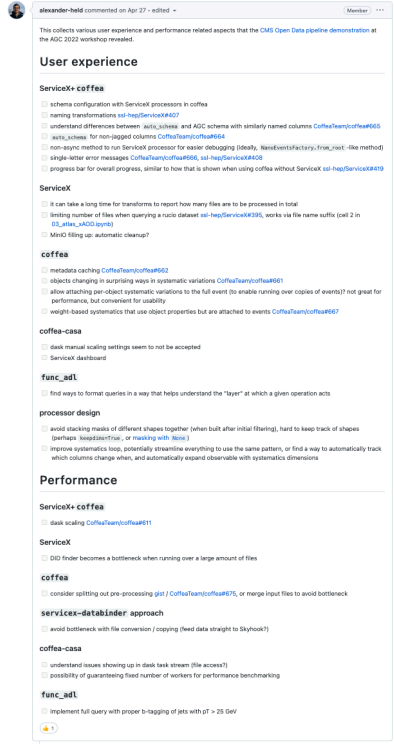


coffea processor

Our status

- **Implemented & demonstrated ttbar analysis pipeline**
  - showcased at dedicated AGC tools workshop: https://indico.cern.ch/e/agc-tools-2
  - implementation available on GitHub: iris-hep/analysis-grand-challenge
    - everything developed and tracked openly
  - pipeline generally works well, with well-defined interfaces between components

- Implementation was a useful exercise revealing **areas for further improvement**
  - several performance bottlenecks
    - being addressed now, expect significant improvements!
  - sharp edges related to user experience
    - following up with developers & experts to smoothen the user experience



follow-up items tracked in agc#64

Next steps and plans for the future

- **Expanding** the **ttbar analysis task**
  - machine learning component, extended set of systematic uncertainties, larger amount of data to process
- **Develop & compare** different **implementations**
  - ongoing project to implement analysis using `ROOT` `RDataFrame`
- **Documentation**
  - provide complete analysis task description as text
- **Benchmarking**
  - investigate performance, identify potential additional bottlenecks & implement solutions
- Longer term plan: **differentiable analysis pipeline** (related meeting: https://indico.cern.ch/event/1096431/)
  - investigate end-to-end analysis optimization, evaluate usefulness vs cost of gradient information

Summary

- The **Analysis Grand Challenge** is an **integration exercise** to study **HL-LHC analysis workflows**

- Developed **ttbar analysis task & implementation** based on **CMS Open Data**
  - all data & our implementation are publicly available

- We hope that the **Analysis Grand Challenge** can be **useful to the broader community**
  - test analysis tools, compare different workflows, test analysis facilities, ...

- Stay in touch via our **mailing list**
  - analysis-grand-challenge@iris-hep.org (sign up at this Google grop)

Thank you!

- The **AGC is made possible** thanks to the **help of a large number of people** working on many different projects.

- **Thank you** in particular to the teams behind:
  - coffea-casa
  - Scikit-HEP, coffea, IRIS-HEP Analysis Systems
  - ServiceX, IRIS-HEP DOMA
  - IRIS-HEP SSL
  - CMS Open Data

Backup

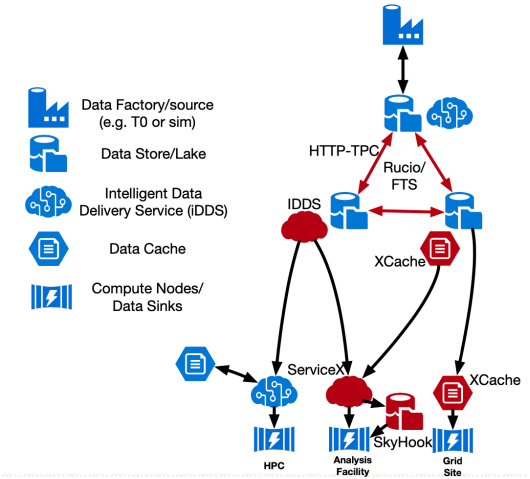
Abstract

Analysis workflows commonly used at the LHC experiments do not scale to the requirements of the HL-LHC. To address this challenge, a rich research and development program is ongoing, proposing new tools, techniques, and approaches. The IRIS-HEP software institute and its partners are bringing together many of these developments and putting them to the test in a project called the “Analysis Grand Challenge” (AGC).

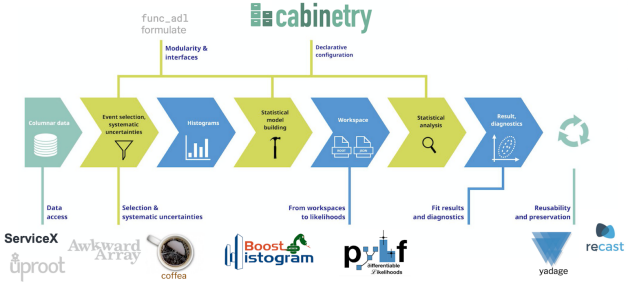
The AGC aims to demonstrate how novel workflows can scale to analysis needs at the HL-LHC. It is based around a physics analysis using publicly available Open Data and includes the relevant technical requirements and features that analyzers at the LHC need. The analysis demonstration developed in this context is heavily based on tools from the HEP Python ecosystem and makes use of modern analysis facilities.
This talk will review the state of the ecosystem, describe the status of the AGC, and showcase how the envisioned workflows look in practice.

Integration: connecting IRIS-HEP focus areas

Data Organization, Management and Access (DOMA):
data delivery



Analysis Systems (AS):
tools



Scalable Systems Laboratory (SSL):
deployment techniques and resources