Physics analysis for the HL-LHC: concepts and pipelines in practice with the Analysis Grand Challenge

Alexander Held[1], Elliott Kauffman[2], Oksana Shadura[3], Andrew Wightman[3]

[1] University of Wisconsin–Madison

[2] Princeton University

[3] University of Nebraska–Lincoln

CHEP 2023

https://indico.jlab.org/event/459/

May 9, 2023

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

The Analysis Grand Challenge (AGC) project

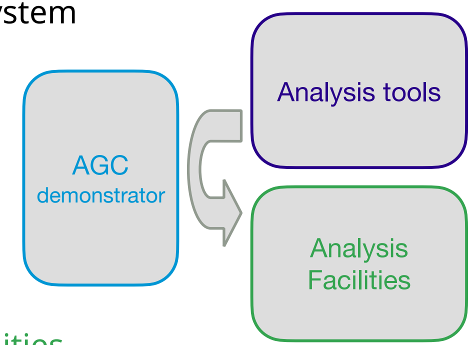
- The **"Analysis Grand Challenge" (AGC)** aims to help **address the computing challenges** of the HL-LHC
  - coordinated by IRIS-HEP: research and development for HL-LHC (https://iris-hep.org/)
  - organized jointly with the US ATLAS & US CMS operations programs

- The AGC has **two aspects**

1. define a physics analysis task of realistic scope & scale
2. develop analysis pipelines that implements the task
   - find & address performance bottlenecks & usability concerns

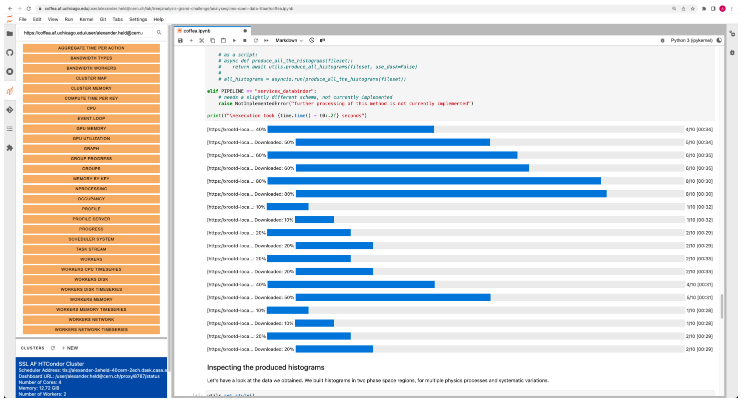
Goals of the AGC project

- Started out as an **integration exercise** combining efforts within IRIS-HEP + broader ecosystem
  - test realistic end-to-end analysis pipelines aimed at HL-LHC use
  - employ modern analysis facilities & new services, evaluate usability & performance
  - investigate possibility of interactive analysis (done in a ☕ break)

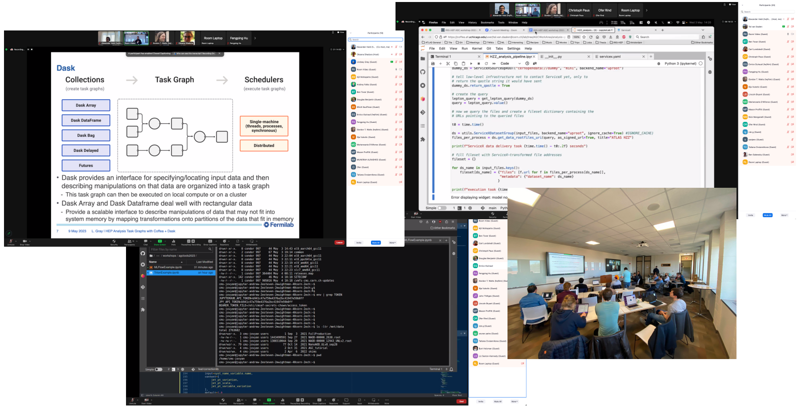


- **Build & engage community**: central gathering point to test new libraries, workflows, facilities

interactive analysis in a notebook

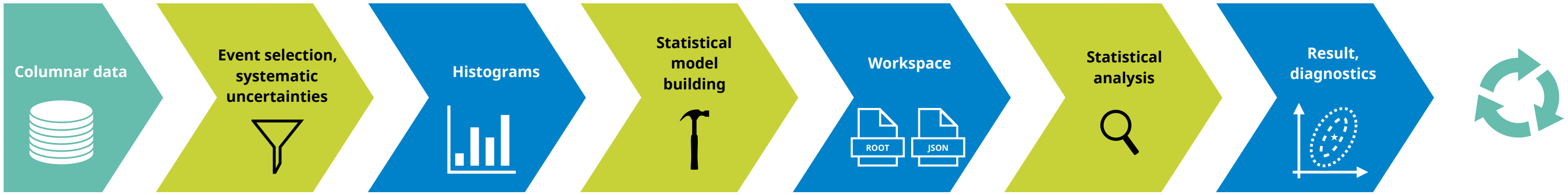


AGC workshop 2023

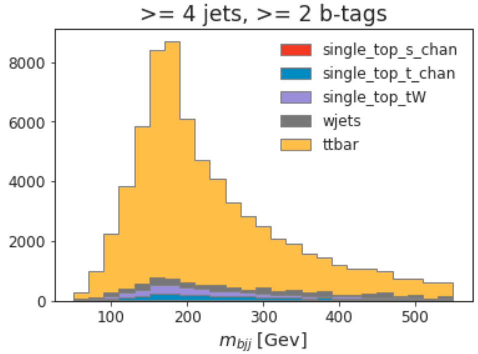


The AGC analysis setup

- **“Analysis”** in the AGC context: starting from centrally produced **common data samples**
  - extract & filter data, calibrate objects & evaluate systematic variations, fill histograms
  - perform statistical inference, visualizations, ensure reproducibility



- Main AGC analysis task: **ttbar cross-section measurement**
  - using CMS Open Data (reformatted to 2 TB of NanoAODs): anyone can participate
  - key feature: different kinds of systematic uncertainties & metadata handling
  - sufficient complexity to demonstrate distributed scale-out performance

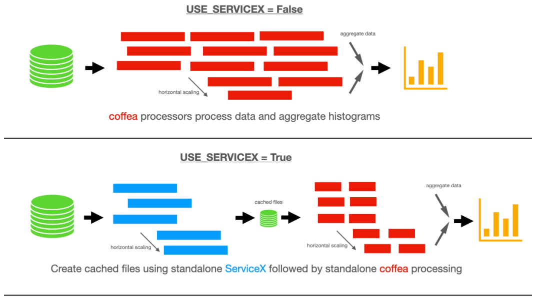


reconstructed top mass

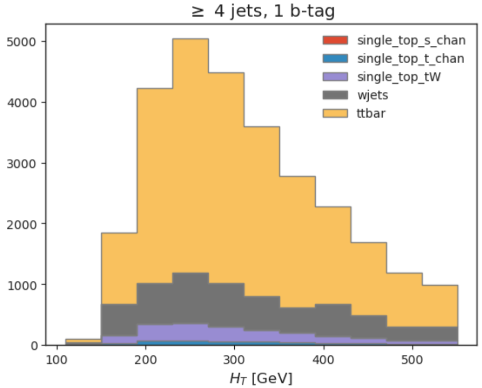
Implementation: ttbar analysis in a notebook

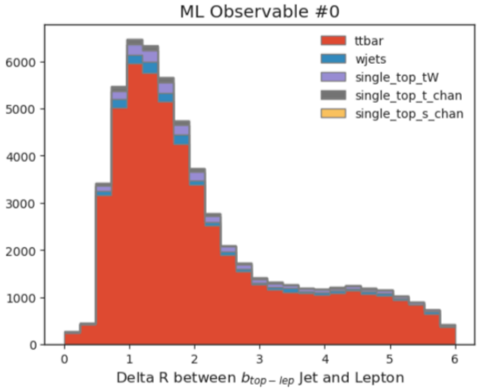
- **From data delivery to statistical inference** in a notebook

multiple supported processing schemes

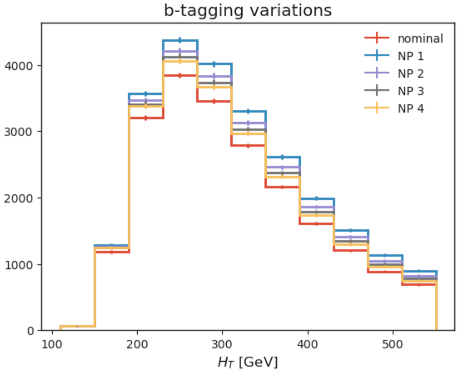


reconstructed observables

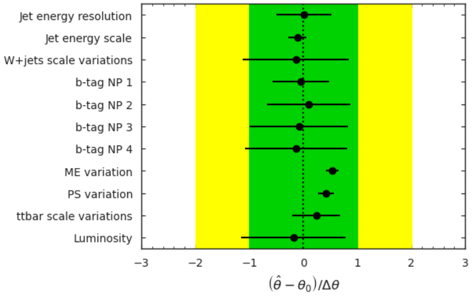




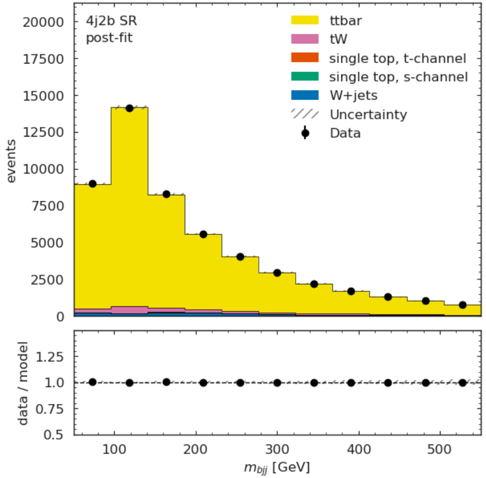
systematic variations



nuisance parameter pulls

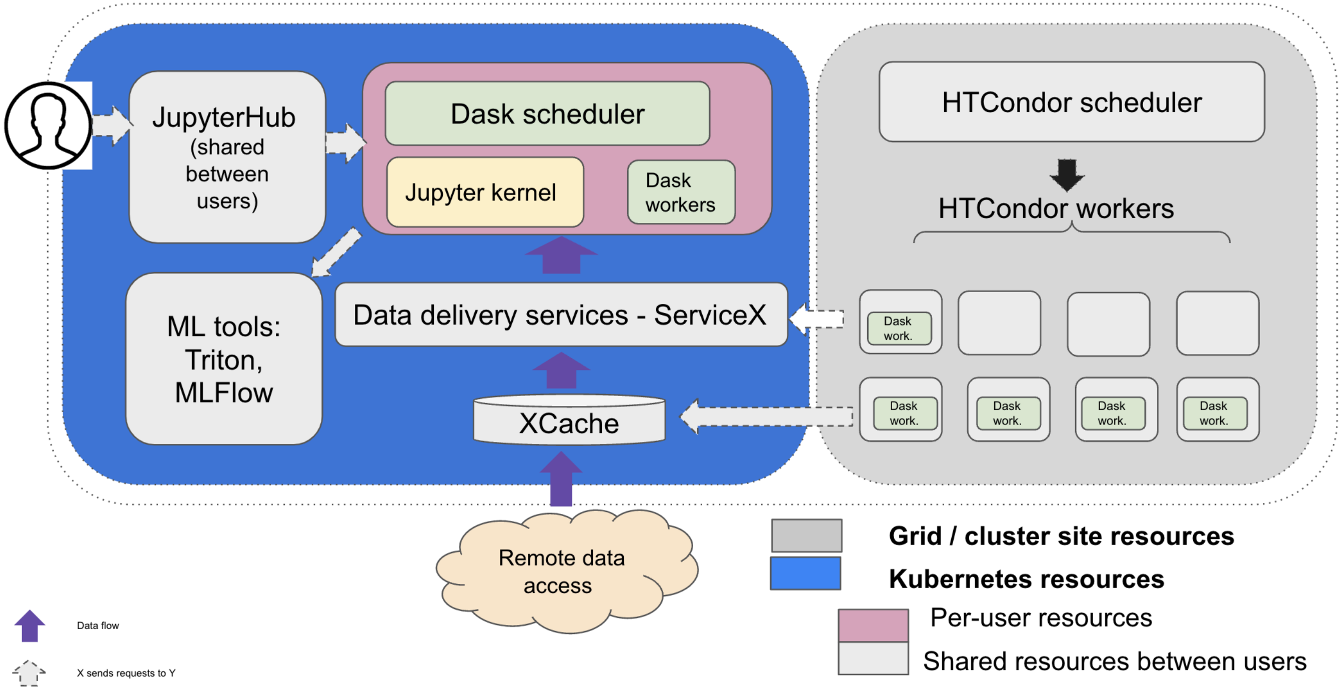


post-fit distributions



Preparing the next generation of analysis facilities

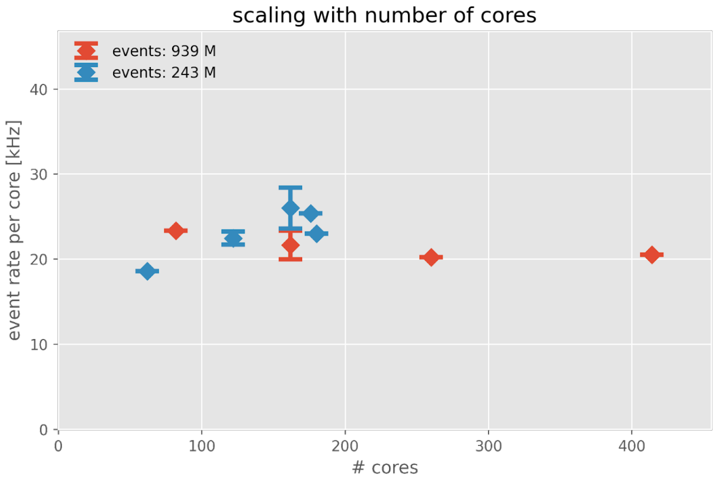
- **coffea-casa** is a **prototype analysis facility** for the HL-LHC providing an **AGC execution environment**
  - interactive facility for columnar analysis providing analysis tools & scaling to computing resources
  - more information: see Oksana Shadura's talk



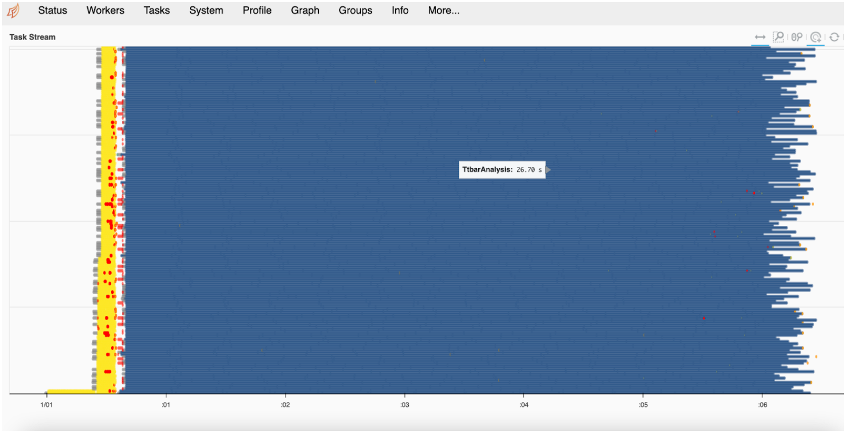
AGC for benchmarking

- **Benchmarking AGC v0.1 implementation performance** at the University of Nebraska–Lincoln CMS Tier-2
  - tested various configurations of hardware, data pipeline and analysis task (see ACAT 2022 contribution)
  - new results show at CHEP! see David Koch's talk and Andrea Sciabà's talk

ACAT 2022 results



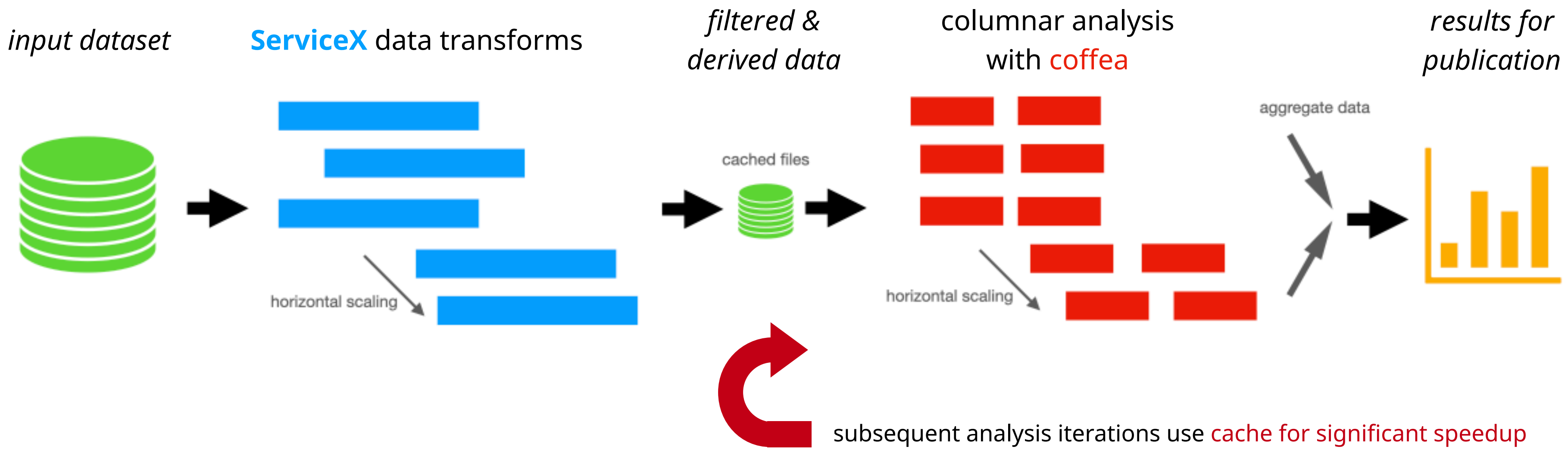
good scaling to hundreds of cores



efficient resource usage via Dask

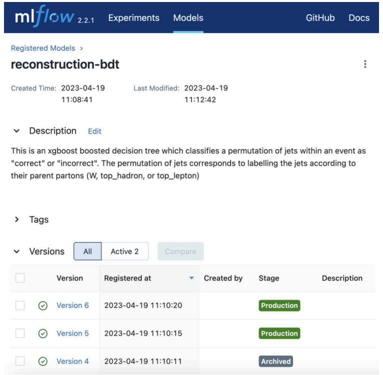
On-demand columnar data delivery: ServiceX

- **ServiceX** is a **data extraction and delivery** service
  - users provide list of datasets to process + instructions for how to extract data (e.g. declarative)
  - ServiceX can be co-located with input datasets for fast execution
  - columnar data is returned and cached -> subsequent executions hit the cache
  - see Ben Galewsky's talk for more information!

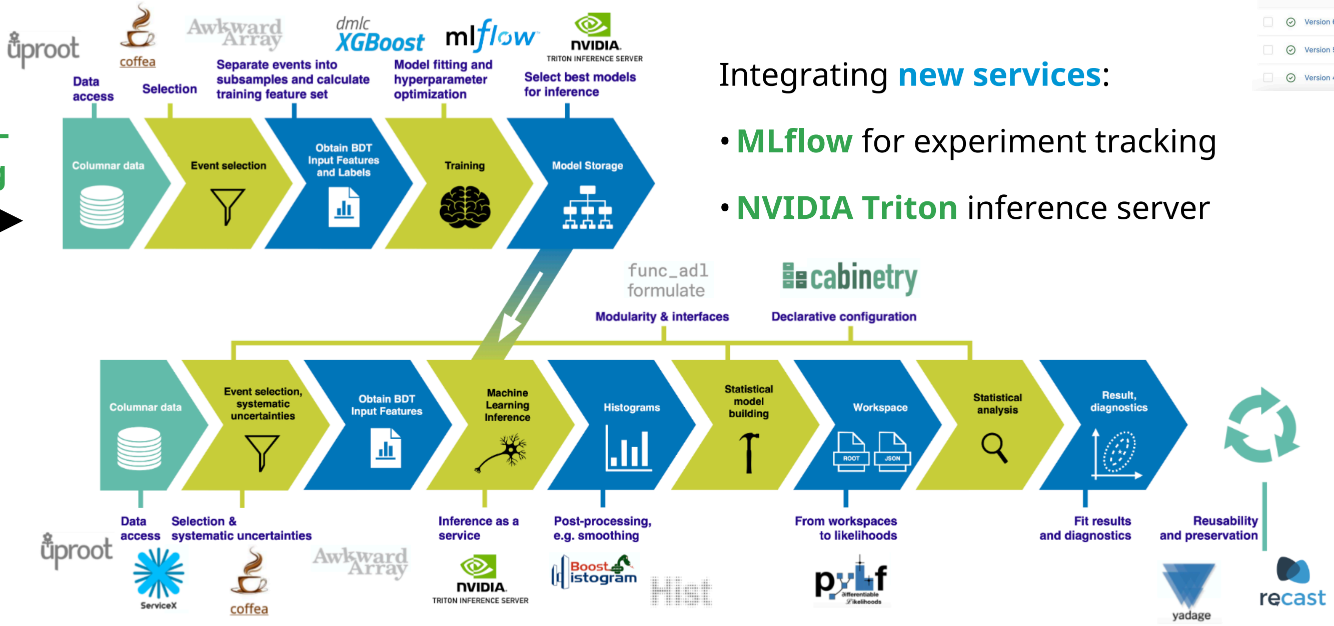


Extending AGC: ML integration

- Development of a **"version 2" of the AGC analysis task** is ongoing
  - expanded task: more complexity and data to process
  - inclusion of machine learning aspects (training & inference): frequently requested!
  - see Elliott Kauffman's talk for more information



new: ML training



Integrating **new services**:

- **MLflow** for experiment tracking
- **NVIDIA Triton** inference server

AGC versions: the project is evolving

- The CMS Open Data ttbar analysis task was first defined in 2022 and has since **evolved** based on **community feedback**

  - **v0.1: ACAT 2022** setup (related talk), using ntuple inputs[1]
    - current RDF implementation[2] (Vincenzo Padulano's talk), summer fellow project this year to update
  - **v0.2:** same analysis as v0.1, improved **ServiceX pipeline** (coffea streaming files from object store)
  - **v1.0:** switch to **NanoAOD inputs** (replaces v0), minimal analysis changes (new column names)
  - **today:** towards v2: **machine learning training + inference** (w/ MLflow + Triton), **correctionlib** adoption
  - **v2.0:** (~ mid June target): machine learning + further expanded systematics (increased I/O and CPU)

- **AGC showcase event** in September featuring demonstrations based on v2

- See also the website for version information

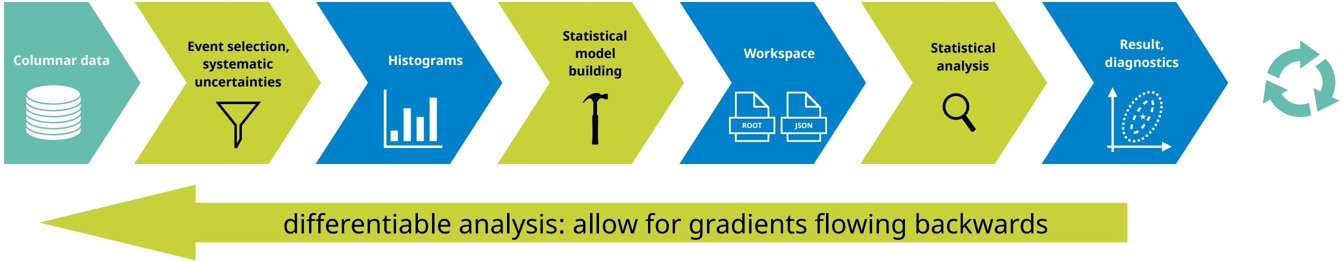
[1] with `ntuples_merged.json`, no point in using the older `ntuples.json`
[2] currently misses statistical inference part of pipeline

Future plans for the AGC

- Short term: wrap up development of **AGC v2** and perform **showcase event** in September
  - including benchmarking: continuously identify & address bottlenecks

Year	Target
2024	• Define analysis tasks for the top quark mass and di-Higgs measurement. • High-volume analysis done on dataset 20% the scale needed for HL-LHC and completed within 1 hour. • Integrate ML inference service with AGC.
2025	• High-volume analysis done on dataset 40% the scale needed for HL-LHC and completed within 1 hour. • Demonstrate AOD column extraction workflow
2026	• High-volume analysis done on dataset 60% the scale needed for HL-LHC and completed within 1 hour. • Demonstrate fully differentiable analysis
2027	• High-volume analysis done on dataset 80% the scale needed for HL-LHC and completed within 1 hour.
2028	• High-volume analysis done on dataset 100% the scale needed for HL-LHC and completed within 1 hour.

- Longer term: **IRIS-HEP strategic plan** (arXiv:2302.01317 [hep-ex])
  - two new flagship analyses: complexity of methodology & scale of data, closer ATLAS & CMS connections
  - column joining: enhance analyzer-level data with missing information (e.g. NanoAOD with MiniAOD enhancement)
  - differentiable analysis: investigate end-to-end analysis optimization

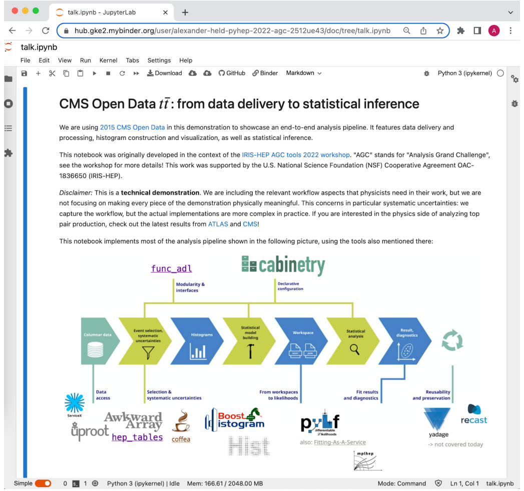


Summary

- The **Analysis Grand Challenge** project develops and studies **HL-LHC analysis workflows**
  - provides gathering point for community & context for discussions
- Developed **ttbar analysis task & implementation** based on **CMS Open Data**
  - all data & implementations are publicly available
  - used for benchmarking & improving performance and user experience
- **More information**
  - AGC workshop last week, AGC documentation, GitHub repository
  - mailing list: analysis-grand-challenge@iris-hep.org (sign-up link)
- **Give it a try!**
  - run AGC in your browser: Binder link, context: PyHEP 2022 contribution

AGC workshop last week

AGC example via Binder



Thank you!

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

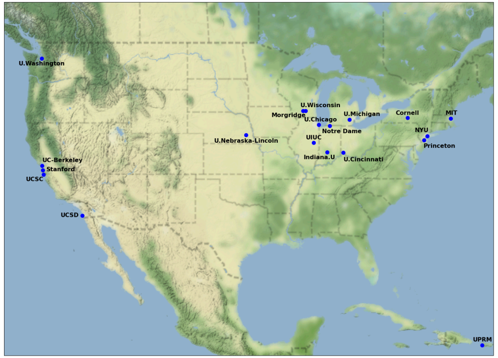
- **Thank you** in particular to the teams behind:
  - coffea-casa
  - Scikit-HEP, coffea, IRIS-HEP Analysis Systems
  - ServiceX, IRIS-HEP DOMA
  - IRIS-HEP SSL
  - CMS Open Data

- Lots of (directly) **related CHEP contributions**
  - David Koch: Monday 12:15, track 4
  - Elliott Kauffman: Monday 14:00, track 8
  - Andrea Sciabà: Monday 15:15, track 7
  - Oksana Shadura: Tuesday 10:00, plenary
  - Vincenzo Padulano: Tuesday 17:15, track 6

Backup

IRIS-HEP and the Analysis Grand Challenge

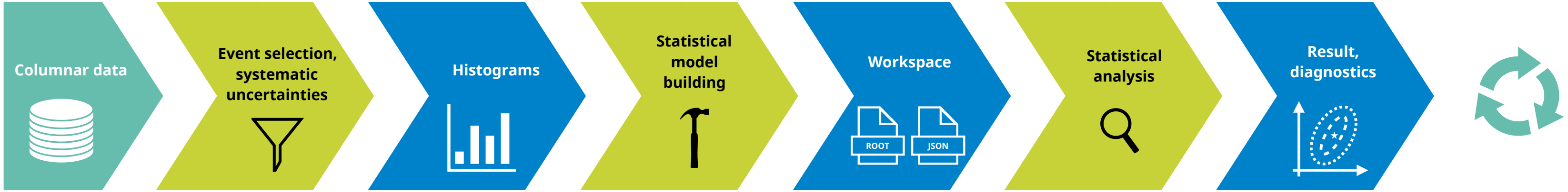
- **IRIS-HEP**: *"Institute for Research and Innovation in Software for High Energy Physics"*
  - software institute funded by the US National Science Foundation
  - research & development for the HL-LHC
    - innovative algorithms for data reconstruction & triggering
    - analysis systems to reduce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



institutes participating in IRIS-HEP

“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



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

Pain points in analysis user experience, ordered

1. **Systematics**
    - Recurring topic throughout this workshop: this is not solved
2. **Metadata**
    - Finding & handling information
3. **Scale-out**
    - Prototyping vs scale-out, different implementations / details on different sites
    - Need for consistent environments across all resources

Analysis Ecosystem Workwshop II
User experience & Declarative Languages summary

AGC showcase event

- Working towards an **AGC showcase event**
  - date to be confirmed, likely September 14
  - short, half-day event
  - inviting interested community to share setup and present results obtained with the AGC
    - opportunity to test variety of AGC implementations and hardware configurations at different sites
  - may also include performance measurements
  - opportunity to showcase computing resources & services to physics analysis community

- **Targets** for contributions
  - baseline: demonstrate distributed scaling with Dask
  - advanced: ServiceX, ML training & inference, MLflow / Triton integration
  - performance studies: variations in I/O requirements, CPU variations: skip columnar processing / ML inference

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

Abstract

Realistic environments for prototyping, studying and improving analysis workflows are a crucial element on the way towards user-friendly physics analysis at HL-LHC scale. The IRIS-HEP Analysis Grand Challenge (AGC) provides such an environment. It defines a scalable and modular analysis task that captures relevant workflow aspects, ranging from large-scale data processing and handling of systematic uncertainties to statistical inference and analysis preservation. By being based on publicly available Open Data, the AGC provides a point of contact for the broader community. Multiple different implementations of the analysis task that make use of various pipelines and software stacks already exist.

This contribution presents an updated AGC analysis task. It features a machine learning component and expanded analysis complexity, including the handling of an extended and more realistic set of systematic uncertainties. These changes both align the AGC further with analysis needs at the HL-LHC and allow for probing an increased set of functionality.

Another focus is the showcase of a reference AGC implementation, which is heavily based on the HEP Python ecosystem and uses modern analysis facilities. The integration of various data delivery strategies is described, resulting in multiple analysis pipelines that are compared to each other.