# Analysis Grand Challenge

**Alexander Held (University of Wisconsin–Madison)**
**Oksana Shadura (University Nebraska–Lincoln)**

**March 14, 2023**

**IRIS-HEP / Ops Program Analysis Grand Challenge Planning**
**https://indico.cern.ch/event/1263386/**

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

# AGC: two components

**The IRIS-HEP Analysis Grand Challenge (AGC) has two components:**

- Defining a **physics analysis task** of realistic HL-LHC scope & scale

- Developing an **analysis pipeline** that implements this task
  - Finding & addressing performance bottlenecks & usability concerns

You can (for example) take take an analysis task and develop a different implementation, take a pipeline and try it with a new analysis task, or adopt task & implementation and run it on your favorite facility.

# An AGC implementation: software stack

## Involves large number of packages from IRIS-HEP and partners

uproot

Awkward Array

Func ADL

Coffea

VECTOR

Boost Histogram

mplhep

cabinetry

pyhf differentiable Likelihoods

iminuit

ServiceX

Coffea-Casa

XCache

funcX

Analysis specific frameworks and packages (available in Docker container)

Data delivery service (k8s)

Optional services (k8s)

# AGC Demo Day #2

## Feb 24, 2023

- "Demo Day" format: see agenda & GitHub issue
  - Short, technical talks (~ 15 min)
  - Target date for project convergence
  - Recording on YouTube

**17:00** → 17:07 **ServiceX allowing to use multiple code generators (xAOD + uproot)**
**Speaker**: Benjamin Galewsky (Univ. Illinois at Urbana Champaign (US))

**17:07** → 17:14 **Demo with ServiceX using self-signed JWT tokens**
**Speaker**: Benjamin Galewsky (Univ. Illinois at Urbana Champaign (US))

**17:15** → 17:30 **Awkward-dask integration into coffea framework**
**Speaker**: Lindsey Gray (Fermi National Accelerator Lab. (US))
demo_day.ipynb | nbviewer

**17:30** → 17:45 **Integrating MLflow in AGC workflow**
**Speaker**: Elliott Kauffman (Princeton University (US))
emk_agcdemoday_... | GitHub Repository | nbviewer

**17:45** → 18:00 **Demo with Coffea-casa facility working with integrated CephFS**
**Speaker**: Sam Albin (UNL)

**18:00** → 18:15 **Dependency management for complex analysis at Coffea-casa facility**
**Speaker**: Oksana Shadura (University of Nebraska Lincoln (US))

**18:15** → 18:45 **Discussion**

- Variety of topics covered (AS/DOMA/SSL)
  - Opportunity to showcase latest developments -> open to contributions!
- Will repeat "Demo Day" format every 2 months
  - Next IRIS-HEP Demo day will be integrated into AGC workshop

# AGC workshop 2023

- AGC workshop taking place at **UW-Madison, May 3–5** (just before CHEP)
  - Please sign up! https://indico.cern.ch/e/agc-workshop-2023
  - Due to room limitations, only 32 spots for in-person attendance

- **Workshop format:**
  - Sessions focused on AS, ServiceX, DOMA, facilities, caching
  - Mix of demos and discussions, plus planning session on Friday
  - Focused on identifying next steps towards AGC showcase event

# Workshop schedule

## This is preliminary! We expect it will still evolve a bit.

### Wednesday: AS + ServiceX

**9:00 AM** → 12:30 PM **Morning session: Analysis Systems and demos**

**Convener**: Matthew Feickert (University of Wisconsin Madison (US))

| | | |
|---|---|---|
| 9:00 AM | **Workshop introduction**<br>**Speakers**: Alexander Held (University of Wisconsin Madison (US)), Oksana Shadura (University of Nebraska Lincoln (US)) | 10m |
| 9:10 AM | **coffea + awkward + dask**<br>**Speaker**: Lindsey Gray (Fermi National Accelerator Lab. (US)) | 50m |
| 10:00 AM | **AGC with RDataFrame (via Zoom)**<br>**Speaker**: Vincenzo Eduardo Padulano (Valencia Polytechnic University (ES)) | 30m |
| 10:30 AM | **Coffee break** | 30m |
| 11:00 AM | **User experience for ML**<br>**Speaker**: Elliott Kauffman (Princeton University (US)) | 30m |
| 11:30 AM | **Systematic uncertainties and correctionlib**<br>**Speakers**: Alexander Held (University of Wisconsin Madison (US)), Andrew Wightman, Andrew Wightman (University of Nebraska Lincoln (US)) | 30m |
| 12:00 PM | **tbd** | 30m |

**2:00 PM** → 5:30 PM **Afternoon session: ServiceX, AGC for ATLAS + CMS**

**Conveners**: Benjamin Galewsky (Univ. Illinois at Urbana Champaign (US)), Gordon Watts (University of Washington (US))

| | | |
|---|---|---|
| 2:00 PM | **User experience**<br>including demo, func-adl + ServiceX client + interacting with servers<br>**Speaker**: Tal van Daalen (University of Washington (US)) | 30m |
| 2:30 PM | **tbd** | 1h |
| 3:30 PM | **Coffee break** | 30m |
| 4:00 PM | **ServiceX for ATLAS**<br>including PHYSLITE / PHYS demo | 30m |
| 4:30 PM | **ServiceX for CMS**<br>miniAOD transformer, column joining, use in CMS version of AGC | 30m |
| 5:00 PM | **tbd**<br>performance / latency? | 30m |

# Workshop schedule

## This is preliminary! We expect it will still evolve a bit.

**Thursday:** DOMA, facilities, caching

**9:00 AM** → 12:30 PM **Morning session: DOMA, facilities and performance**

**Convener**: Brian Paul Bockelman (University of Wisconsin Madison (US))

- **9:00 AM** — **DOMA R&D and connections to AGC** — 30m
  **Speakers**: Jayjeet Chakraborty, Jayjeet Chakraborty (University of California, Santa Cruz)
- **9:30 AM** — **End-to-end AGC walkthrough with facility focus** — 1h
  including ServiceX demo: ServiceX + JWT at coffea-casa, writing output to shared FS, dashboard / JupyterLab plugin, ML components at coffea-casa (Triton, MLFlow)
- **10:30 AM** — **Coffee break** — 30m
- **11:00 AM** — **Benchmarking discussion** — 1h 30m
  local vs remote files, XCache (including https status), scaling
  **Speakers**: Fengping Hu (University of Chicago (US)), Ilija Vukotic (University of Chicago (US))

**1:00 PM** → 6:10 PM **Afternoon session: AF reports and caching discussion**

- **2:00 PM** — **Experience with AGC at German facilities** — 20m
  **Speakers**: David Koch, David Martin Koch (Ludwig Maximilians Universitat (DE))
- **2:20 PM** — **AGC at US AFs** — 30m
- **2:50 PM** — **tbd** — 40m
- **3:30 PM** — **Coffee break** — 30m
- **4:00 PM** — **Caching strategies discussion** — 1h
  **Speaker**: Lindsey Gray (Fermi National Accelerator Lab. (US))
- **5:00 PM** — **tbd** — 30m

# Workshop schedule

## This is preliminary! We expect it will still evolve a bit.

**Friday:** planning

- ending around lunchtime to allow for travel to pre-CHEP workshop

**9:00 AM** → 12:30 PM **Planning session: towards an AGC showcase event**

**Conveners**: Alexander Held (University of Wisconsin Madison (US)), Oksana Shadura (University of Nebraska Lincoln (US))

| | | |
|---|---|---|
| 9:00 AM | **AS: workshop outcomes and action items**<br>**Speaker**: Matthew Feickert (University of Wisconsin Madison (US)) | 30m |
| 9:30 AM | **DOMA: workshop outcomes and action items**<br>**Speakers**: Brian Paul Bockelman (University of Wisconsin Madison (US)), Gordon Watts (University of Washington (US)) | 30m |
| 10:00 AM | **SSL: workshop outcomes and action items**<br>**Speakers**: Brian Paul Bockelman (University of Wisconsin Madison (US)), Robert William Gardner Jr (University of Chicago (US)) | 30m |
| 10:30 AM | **Coffee break** | 30m |
| 11:00 AM | **Discussion: towards an AGC showcase event** | 1h |
| 12:00 PM | **Closing**<br>**Speakers**: Alexander Held (University of Wisconsin Madison (US)), Oksana Shadura (University of Nebraska Lincoln (US)) | 30m |

# AGC showcase event

## Timeline: around September

- Possibly co-located with **IRIS-HEP all hands meeting** (maybe second week of September)
- Fairly short (one afternoon) event
- Inviting **everyone** who is interested to **share their setup and to present the results**
  - Interesting combinations of hardware, network site configurations
  - Any type of “combinatorics” of AGC analysis implementation / components setup
  - Can include performance measurements
  - The chance to publicize your computing resources to physics analysis community :-)
- Not meant as the end of the AGC project, but a **big milestone**!

# Evolution of the AGC analysis task

## Towards AGC v2

- AGC analysis task thus far will become "**AGC v0**"
- We are now adopting **NanoAOD inputs** in our implementation
  - This will become "**AGC v1**": same analysis, different input file format
- We are working on defining "**AGC v2**" for **CHEP**
  - Same **NanoAOD inputs**
  - Include **ML training + inference**
  - Increased analysis **complexity**: larger set of **systematic uncertainties**
  - Will present setup at CHEP (+ related talks focused on ML + coffea-casa AF)

# Evolution of the AGC analysis task

## Towards AGC v2 for AFs

- We are working on AF related components "**AGC v2**" @ **CHEP**
  - Include **ML training**
  - Include **inference: Triton**
    - **Will require preferably to have GPUs available for users @ AFs**
  - Additional useful ML services: **MLFlow**

# Summary

- **AGC workshop May 3–5:** please sign up and join us!
  - https://indico.cern.ch/e/agc-workshop-2023
- Work ongoing towards **AGC v2** and **CHEP** contributions
- Planning AGC showcase event around September

- Stay in touch: analysis-grand-challenge@iris-hep.org (sign up: google group link), and please also feel free to contact us if you'd like to get involved or have any questions!

# Backup

# Strategic plan for a 2nd phase of IRIS-HEP

- **Strategic plan: https://arxiv.org/abs/2302.01317**

- Includes section with **AGC plans**
  - Expand to two flagship analyses (high volume, high complexity)
    - Further increase scale & complexity (+ ML)
  - Continue annual workshops
  - Demonstrate AOD column joining, differentiable analysis pipeline
  - Many connections to IRIS-HEP focus areas

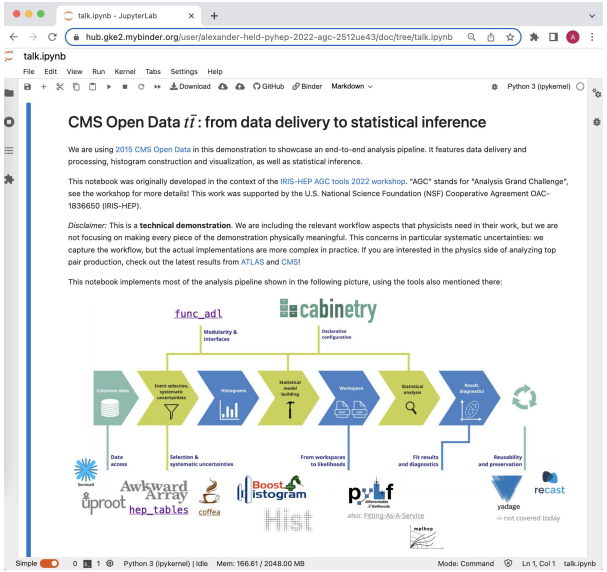
- **Experiment-specific (ATLAS/CMS) implementations**

# AGC: give it a try!

## We are making it easy for you to try out our setup

- **One click** to get PyHEP notebook in Binder environment
  - **Try it out today!**

- You can also use the **UNL Open Data coffea-casa**
  - Or **SSL** (ATLAS members), or your favorite facility
  - This allows you to scale up (limited on Binder)
  - Everything is available in the **AGC repository**

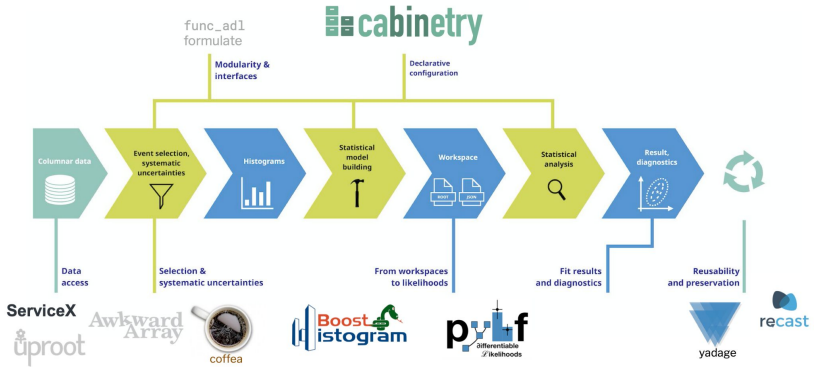


# Analysis pipeline

- **Pipeline setup**
  - **ServiceX** delivers columns following declarative **func_adl** request
  - **coffea** orchestrates distributed event processing & histogram production
    - Using **uproot, awkward-array, hist**
  - Visualization with **hist & mplhep**
  - Statistical model construction with **cabinetry** & inference with **pyhf**
- **Everything is openly developed** (IRIS-HEP AGC repository)
  - Including categorization of datasets in terms of role in AGC demonstrator
- Will be executed on various partner facilities: *University Nebraska-Lincoln, UChicago, FNAL, BNL, others*

***Other (partial) AGC implementations:***

- *ROOT RDF* (Andrii Falko, Enrico Guiraud): andriiknu/RDF/
- *Julia* (Jerry Ling): Moelf/LHC_AGC.jl



# AGC: how we envisioned it initially

## An “integration exercise” for IRIS-HEP

- Demonstrate method for **handling HL-LHC data pipeline requirements**
  - Large data volumes + bookkeeping
  - Handling of different types of systematic uncertainties
  - Use of reduced data formats (PHYSLITE / NanoAOD), aligned with LHC experiments

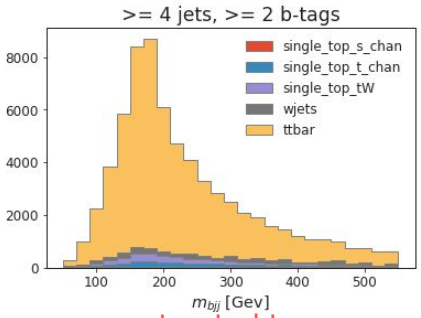
- Aiming for **“interactive analysis”**: turnaround time of ~minutes or less
  - Made possible by highly parallel execution in short bursts, low latency & heavy use of caching

- **Specify all analysis details** to allow for **re-implementations** and re-use for benchmarking

- Execution on **Analysis Facilities**

# AGC: analysis task

## Community benchmark



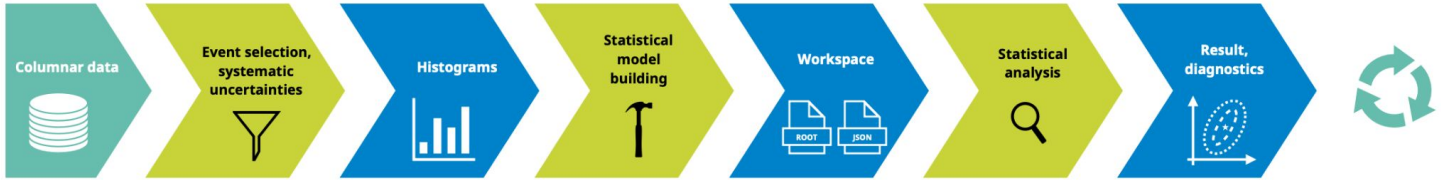

reconstructed top mass

- Analysis task: **ttbar cross-section measurement** in single lepton channel
  - Includes simple top reconstruction
  - 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)
- Goal of setup is showing **functionality**, not discovering new physics
  - Want to capture workflow; use made-up tools for calibrations & systematic uncertainties

# AGC: what we mean by “analysis”

## Typical steps in an analysis workflow

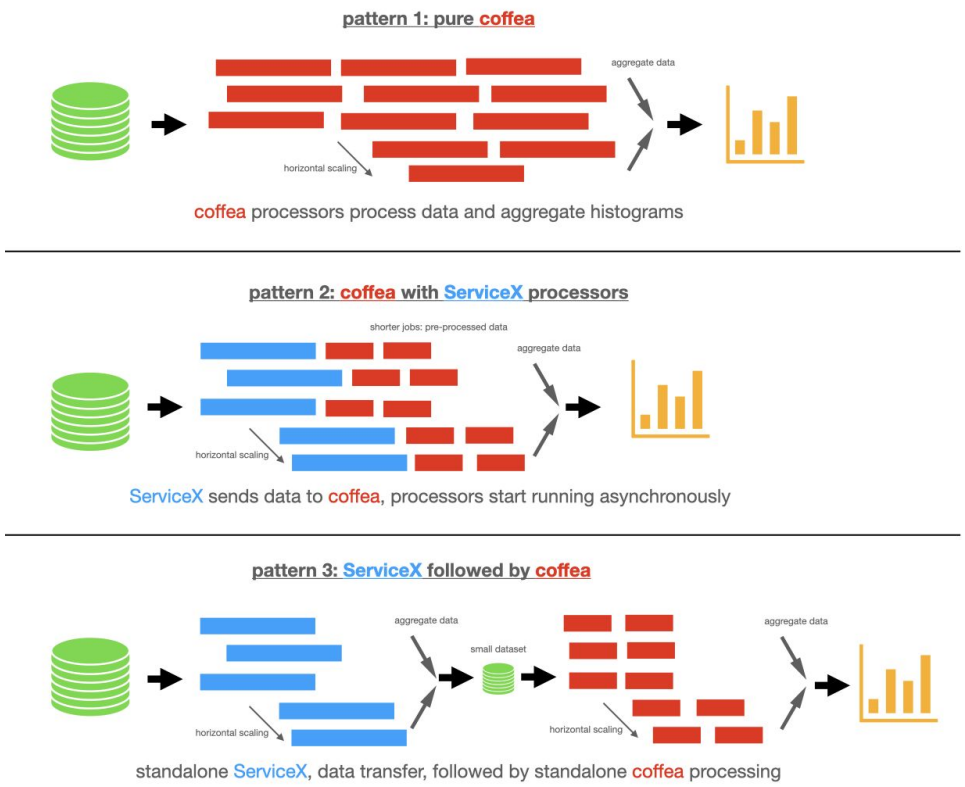
- Start from centrally produced **common data samples**
- Perform all subsequent steps (in a reproducible way)
  - Extract relevant data
  - (Re-) calibrate objects & calculate systematic variations
  - Filter events & calculate observables
  - Histogramming (for binned analyses)
  - Construct statistical model & perform statistical inference
  - Visualize results & provide all relevant information to study analysis details



# Adding ServiceX to the mix

## Benefits of caching

- Investigating different **data pipelines**
- Data delivered by ServiceX can be **filtered** and is **cached locally**
  - Subsequent runs can hit **(filtered) cache for significant speedup**



# What currently runs where?

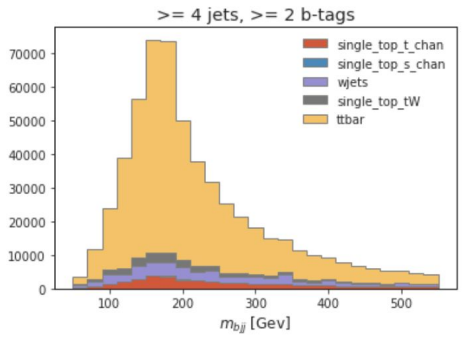
## (please help us update the gaps)

| | BNL | FNAL | SLAC | UNL | UChicago |
|---|---|---|---|---|---|
| basic `coffea` (e.g. `IterativeExecutor`)<br>-> notebook with `USE_DASK = False` | ✓ | ✓ | ✓ | ✓ | ✓ |
| `coffea` scaling (e.g. with Dask)<br>-> notebook with default settings* | | ✓ | ✓ | ✓<br>(using HTCondor @ Tier2, planning to switch to k8s) | ✓ |
| standalone `ServiceX`<br>-> notebook (no configuration) | ✓ | ✓ | | ✓ | ✓ |
| `ServiceX`+`coffea`+scaling<br>-> notebook with `PIPELINE = "servicex_processor"` | | | | ✓ | ✓ |
| `XCache` support | ✓ | ✓<br>(some performance caveats, to be understood) | ✓ | ✓ | ✓ |

* may need site-dependent Dask cluster configuration, see implementation, please get in touch in case of questions

# AGC implementations

## Community effort




- *coffea*: iris-hep/analysis-grand-challenge/
- *ROOT RDF* (Andrii Falko, Enrico Guiraud): andriiknu/RDF/
- *Julia* (Jerry Ling): Moelf/LHC_AGC.jl