Workshop on workflow languages for HEP analysis

Matthew Feickert, Jamie Gooding, Lukas Heinrich, Clemens Lange

April 3rd, 2024

Workshop Overview

- Bring together expertise of developer and HEP user communities
- Discussions/demos are encouraged and significant time has been allocated
- Focus for each day:
  - **Today (3rd April)**: showcasing workflow languages and workflow management tools
  - **Tomorrow (4th April)**: workflow languages in HEP analyses
  - **Friday (5th April)**: workflow languages for reproducibility and workflow adoption
- Aim to establish:
  - State-of-the-art for workflow languages and direction of developments
  - Current HEP best-practices on workflow languages
  - Needs of HEP community in future analyses (e.g., HL-LHC)
- Will write brief whitepaper (to go on arXiv) to summarise outcomes

Workshop Sponsors

FAIROS-HEP

This event is sponsored in part by the National Science Foundation through grants OAC-2226378, OAC-2226379 and OAC-2226380 (FAIROS-HEP).

SMARTHEP
REAL-TIME ANALYSIS FOR
SCIENCE AND INDUSTRY

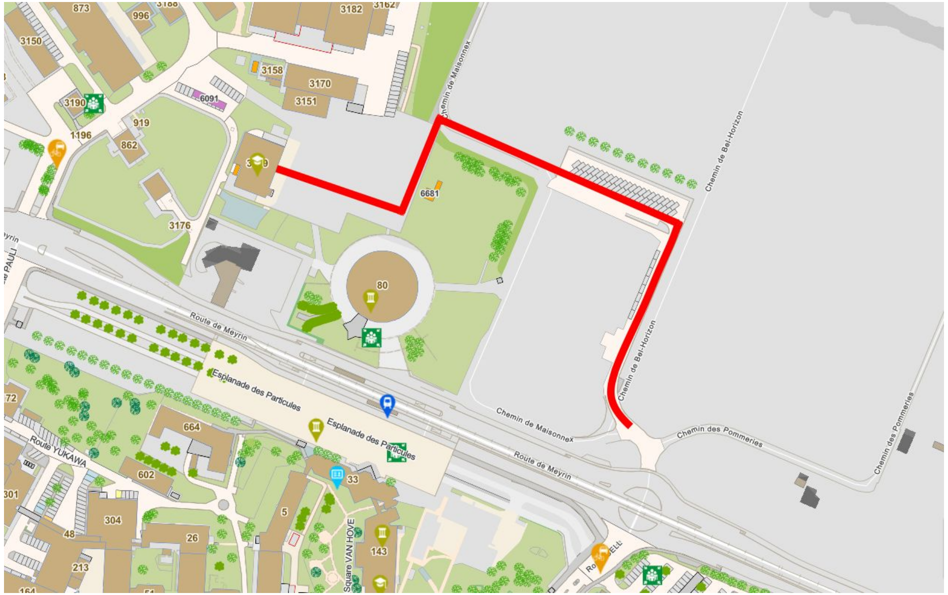
We acknowledge funding from the European Union Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086 (SMARTHEP).

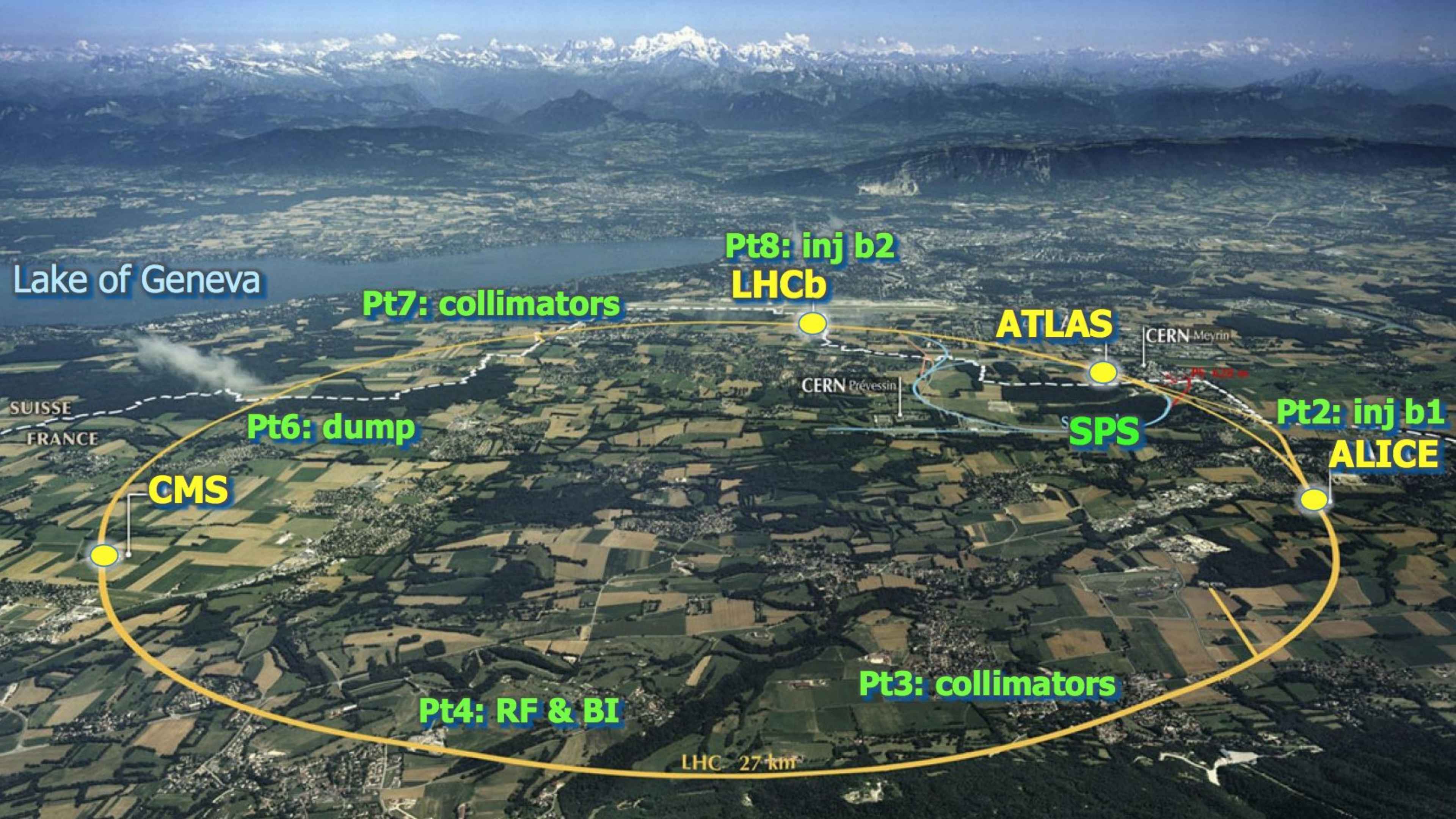
iris
hep

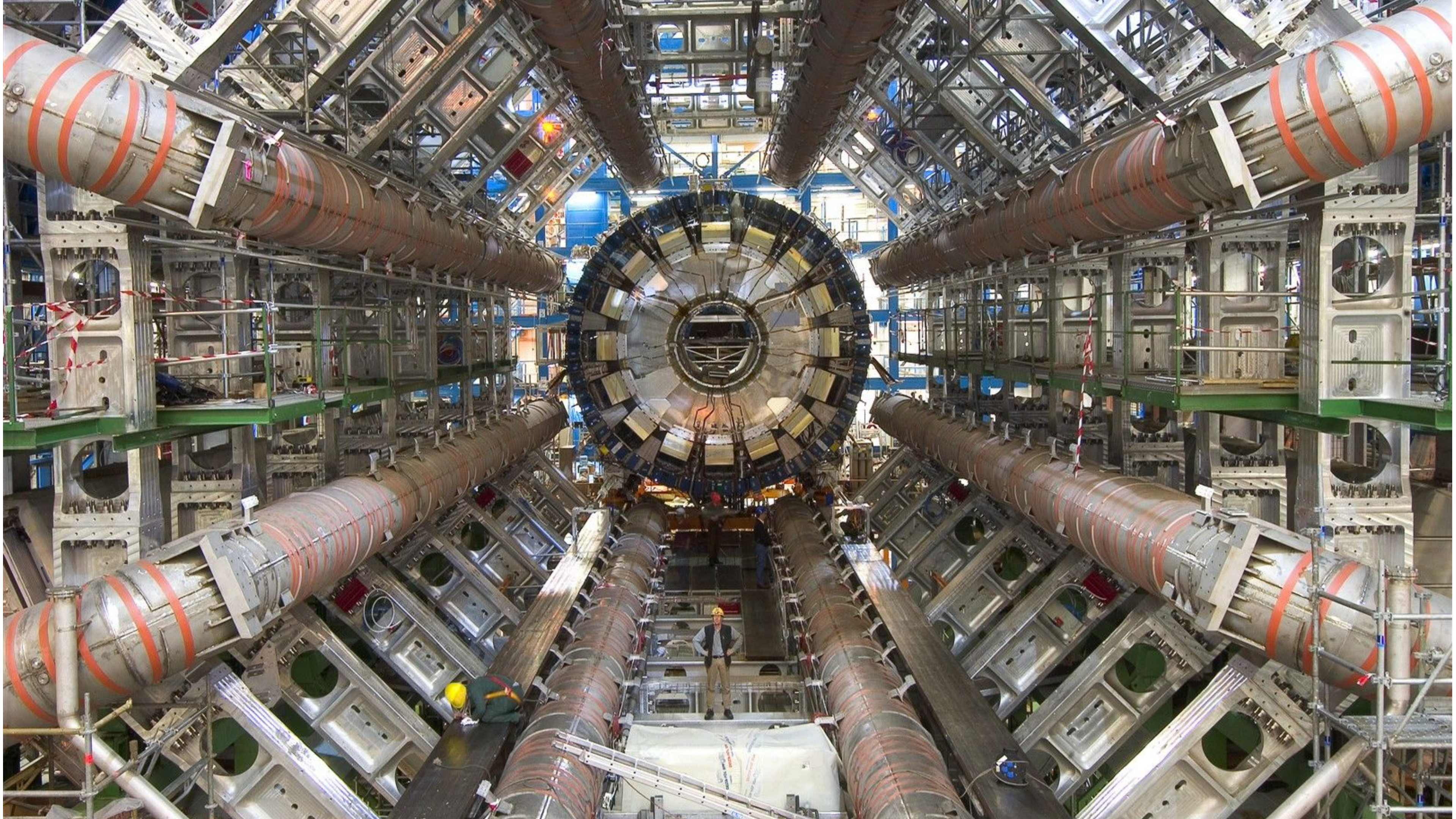
This event's computing resources were supported by the National Science Foundation under Cooperative Agreements OAC-1836650 and PHY-2323298 (IRIS-HEP).

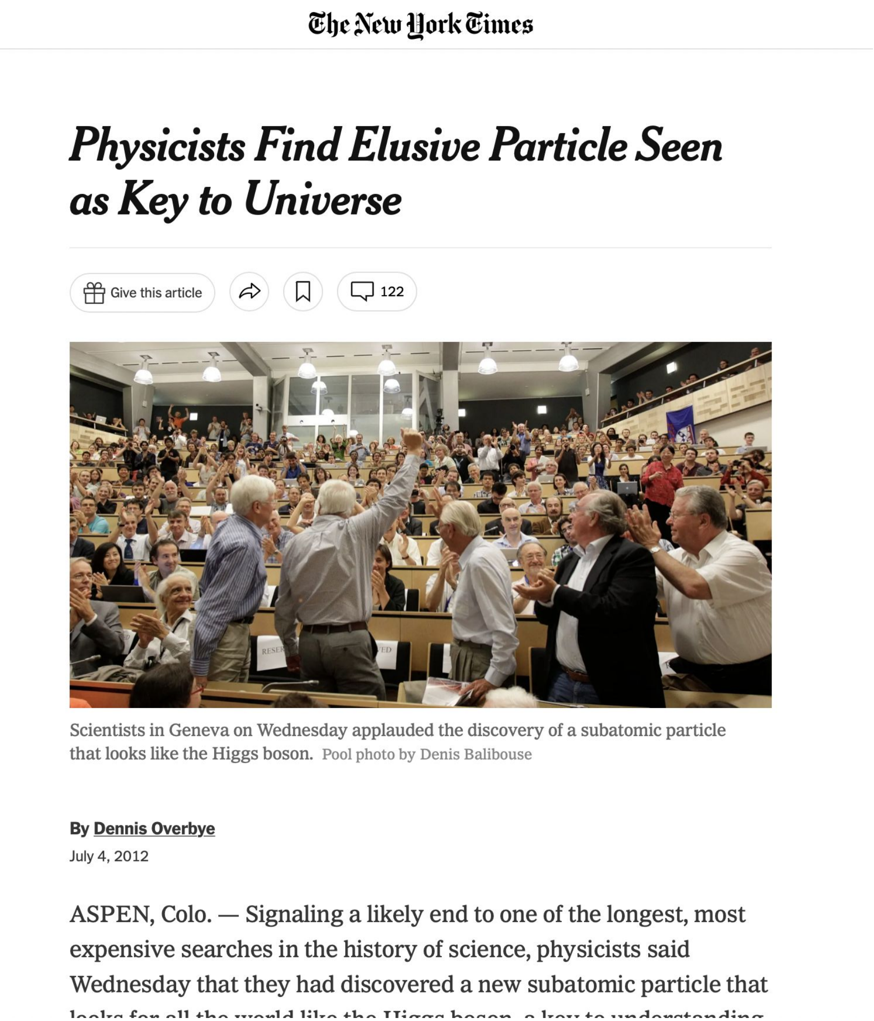
CERN IdeaSquare

- Access by walking around the globe passing to the left of the Science Gateway (not pictured in map)
- Can use coffee machines in the kitchen – please put used mugs into the dishwasher
- Can use pods for breakout discussions – please create ad hoc Zoom rooms for remote participation




Lake of Geneva
Pt8: inj b2
LHCb
Pt7: collimators
ATLAS
CERN Meyrin
PS
CERN Prévessin
SUISSE
FRANCE
Pt6: dump
SPS
Pt2: inj b1
ALICE
CMS
Pt3: collimators
Pt4: RF & BI
LHC 27 km





The New York Times

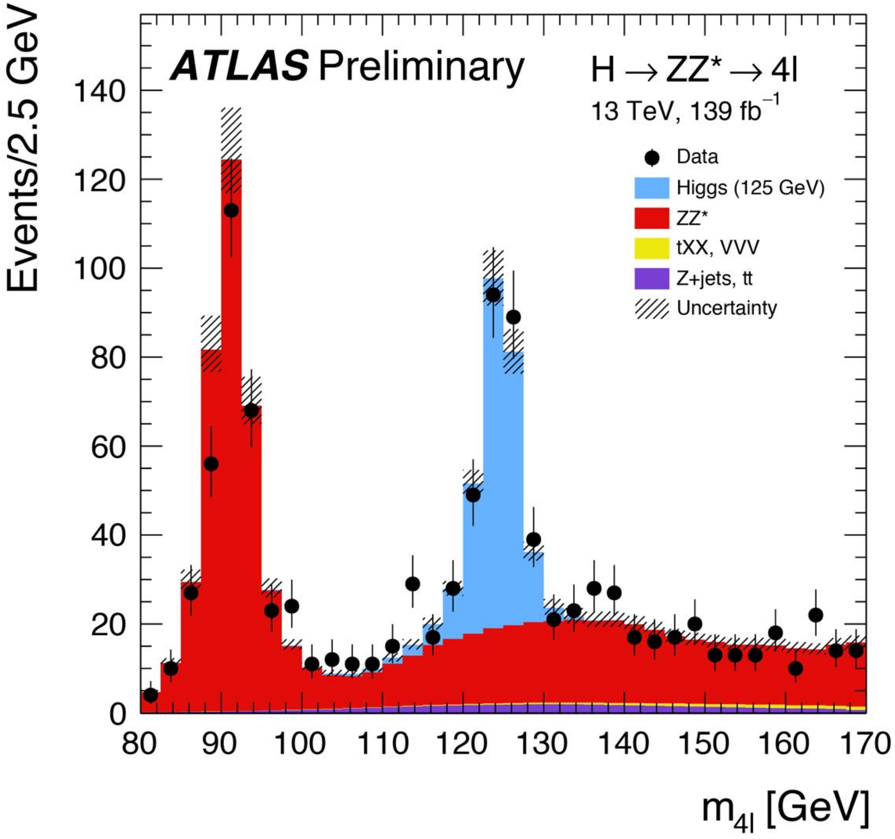
Physicists Find Elusive Particle Seen as Key to Universe

Scientists in Geneva on Wednesday applauded the discovery of a subatomic particle that looks like the Higgs boson. Pool photo by Denis Balibouse

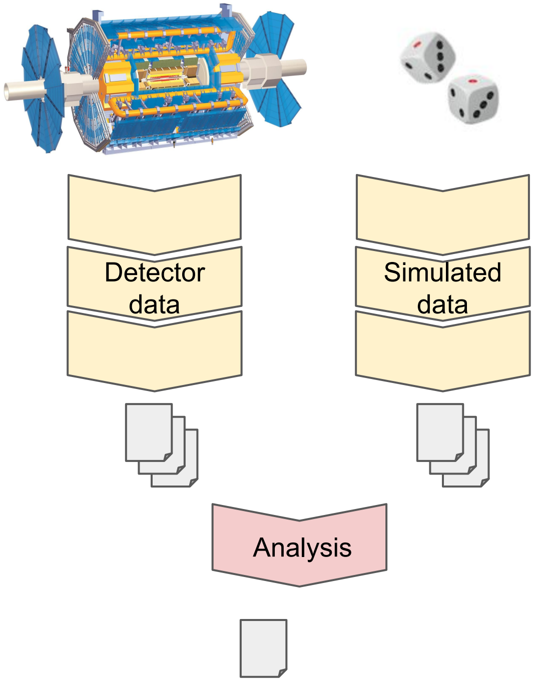
By Dennis Overbye

July 4, 2012

ASPEN, Colo. — Signaling a likely end to one of the longest, most expensive searches in the history of science, physicists said Wednesday that they had discovered a new subatomic particle that looks for all the world like the Higgs boson, a key to understanding



ATLAS computing flow



“Production”

- Generic Data Preparation for everyone within the collaboration.
- Highly Structured and centrally organized

“Analysis”

- Analysis of preprocessed Data with a specific Analysis Goal in mind
- More heterogeneous
- The “Workflow” we’re talking about in this workshop

Overview of HEP Analysis

Analyses aim to extract insight from particle collisions:

- Search for particles/processes
- Measure quantities

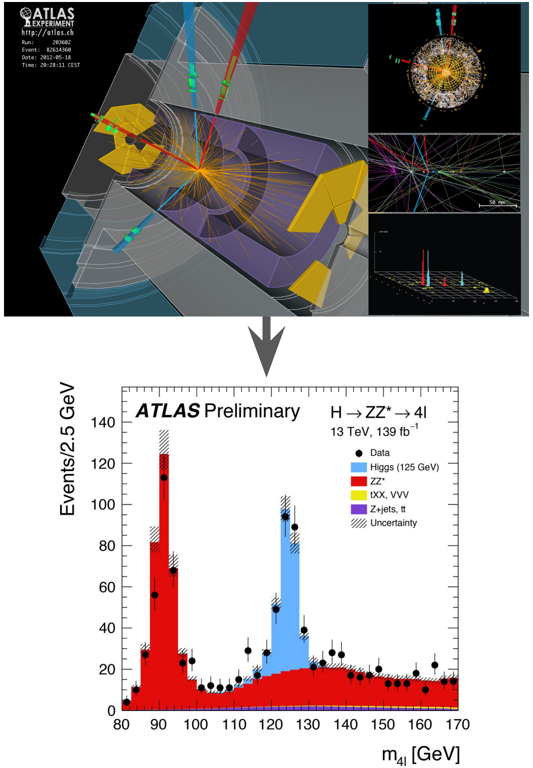
Datasets are typically:

- Large (*O*(mn) events/*O*(TB))
- Distributed

Typically many steps transforming/reducing data:

- Reduction of background
- Filling histograms
- Log-likelihood fitting

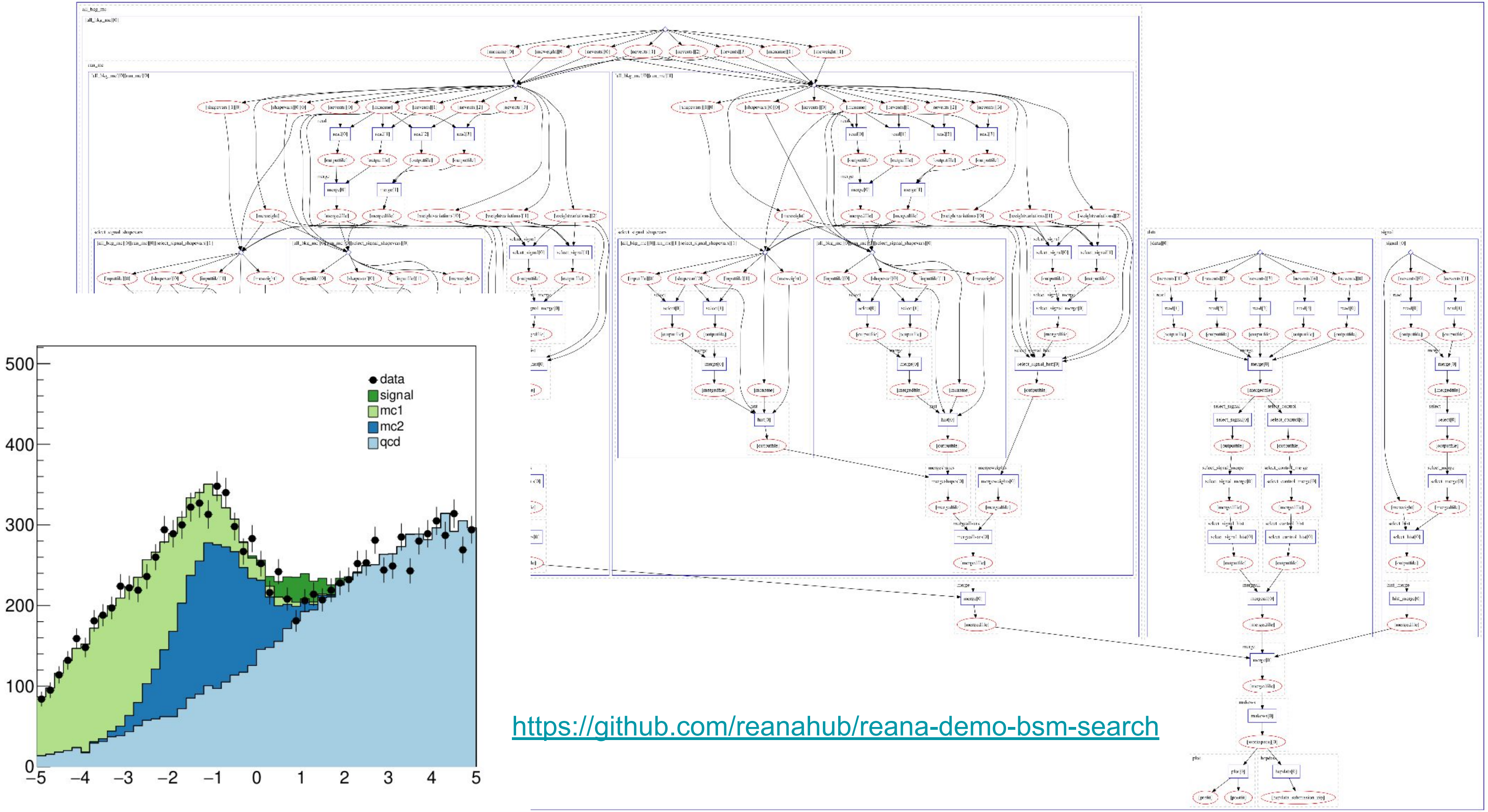
Analyses often have many branches for studies (e.g. systematics)



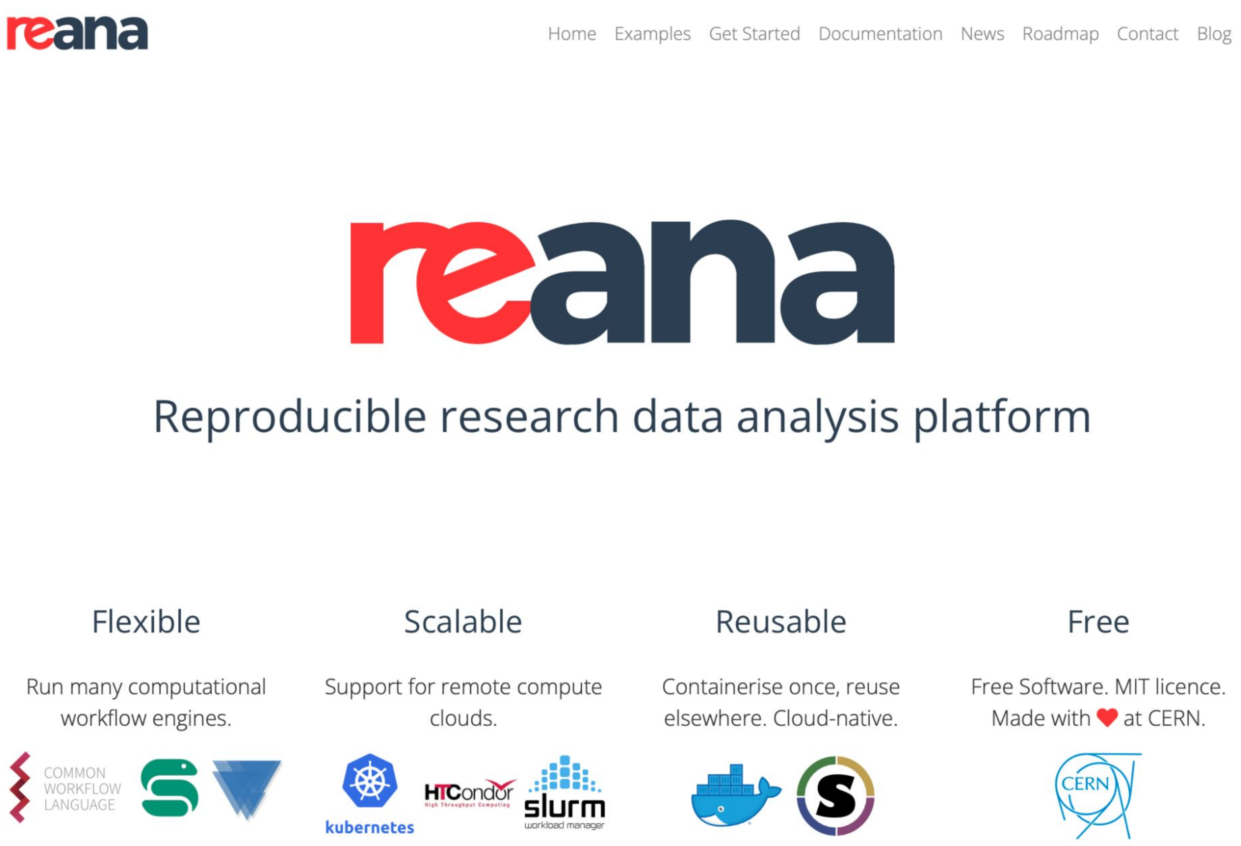
HEP-orientated questions to consider for discussion

- Need each step of a workflow to run in bespoke software environment (Linux container support is required. What runtimes are supported? E.g. Docker, Podman, Apptainer/Singularity)
- Workflow engine needs to be isolated from analysis code – how can we best separate the two while still making use of workflow commands natural during analysis development process?
  - e.g. avoid including workflow tooling in analysis software
  - Anything that needs to be changed in analysis software?
- Workflow scheduling: where can workflows be executed using typical HEP resources (HTCondor, SLURM, WLCG, Kubernetes…)
  - Can there be some generic solutions to this that don't need implementations for each engine?
- Dynamics graphs
  - Number of files could be unknown in advance of runtime
  - Want to be able to control processes that call task graph builds (e.g. Dask). How is balance created?

Example Workflow



https://github.com/reanahub/reana-demo-bsm-search



reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.

Scalable

Support for remote compute clouds.

Reusable

Containerise once, reuse elsewhere. Cloud-native.

Free

Free Software. MIT licence.
Made with ❤ at CERN.

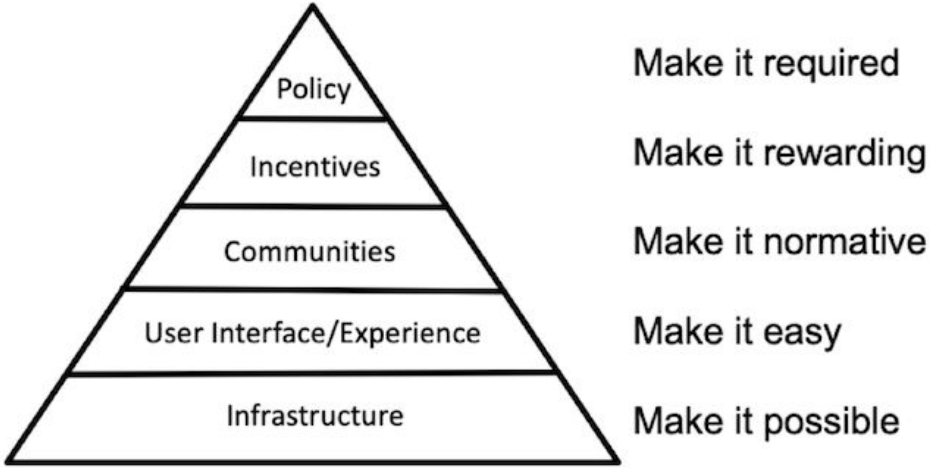
Challenges

Main observation:

We can do this technically and we've run workflows at scale at CERN

But use is still more "top-down" mandated These tools are not yet organically used in HEP

Trying to learn from other communities and identify what's missing



Open Science Pyramid

Workshop dinner

The workshop dinner will take place at Luigia Academy Meyrin on Thursday (tomorrow) at 19:00.

It’s a 20-minute walk from IdeaSquare. If you want to walk there together, we’ll **leave from CERN hostel building 39 at 18:35**.

Food is covered by our sponsors – drinks will need to be paid for individually

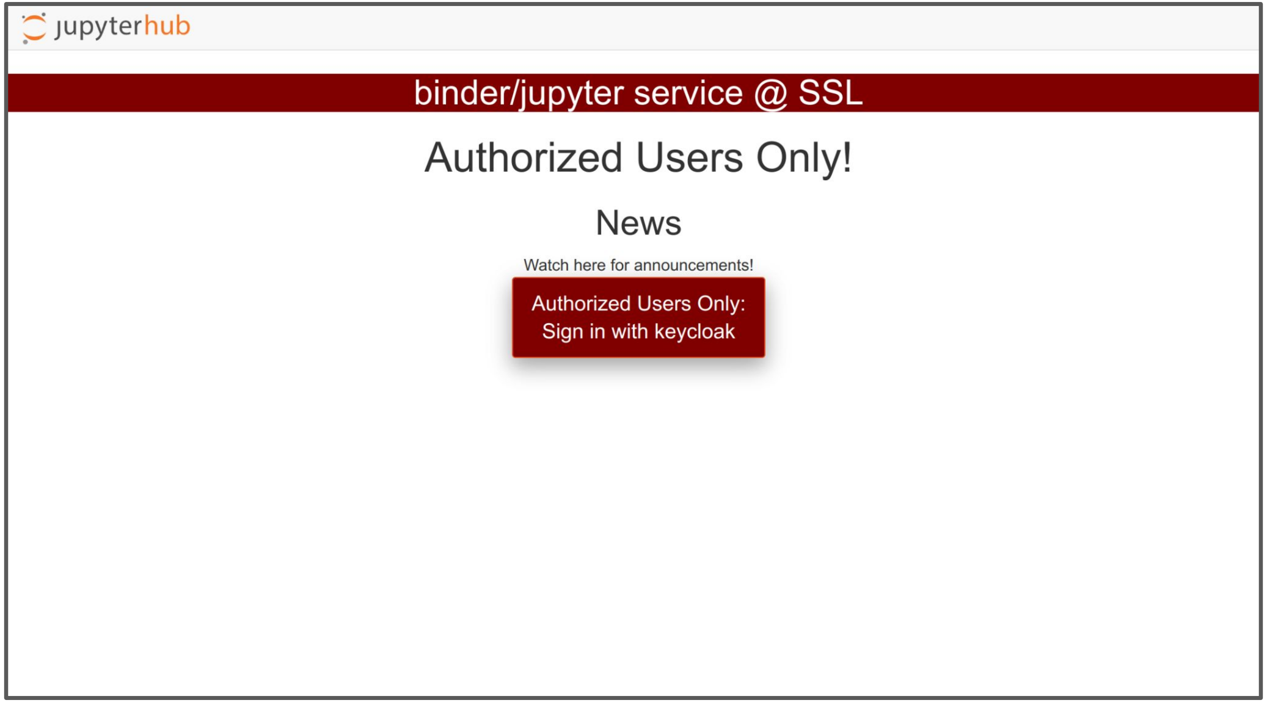
If you would like to join and haven’t filled the survey yet, please do so **by 17:00 today**

Live Notes

- Have a CodiMD setup for community live notes https://codimd.web.cern.ch/bknH2bfqS26ORazJ-eRnOA?both
- Please contribute notes, questions, and discussion there
- Will be used when writing a workshop summary white paper

IRIS-HEP SSL BinderHub

- https://binderhub.ssl-hep.org/



binder/jupyter service @ SSL

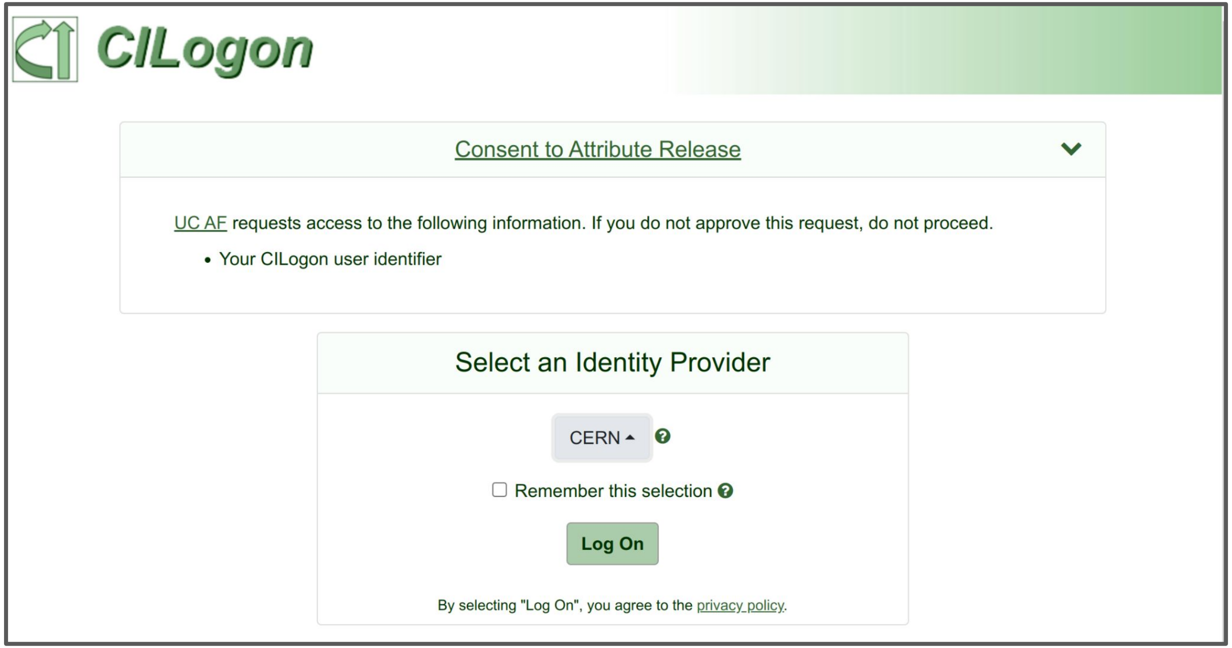
Authorized Users Only!

News

Watch here for announcements!

IRIS-HEP SSL BinderHub

- https://binderhub.ssl-hep.org/



Consent to Attribute Release

UC AF requests access to the following information. If you do not approve this request, do not proceed.

- Your CILogon user identifier

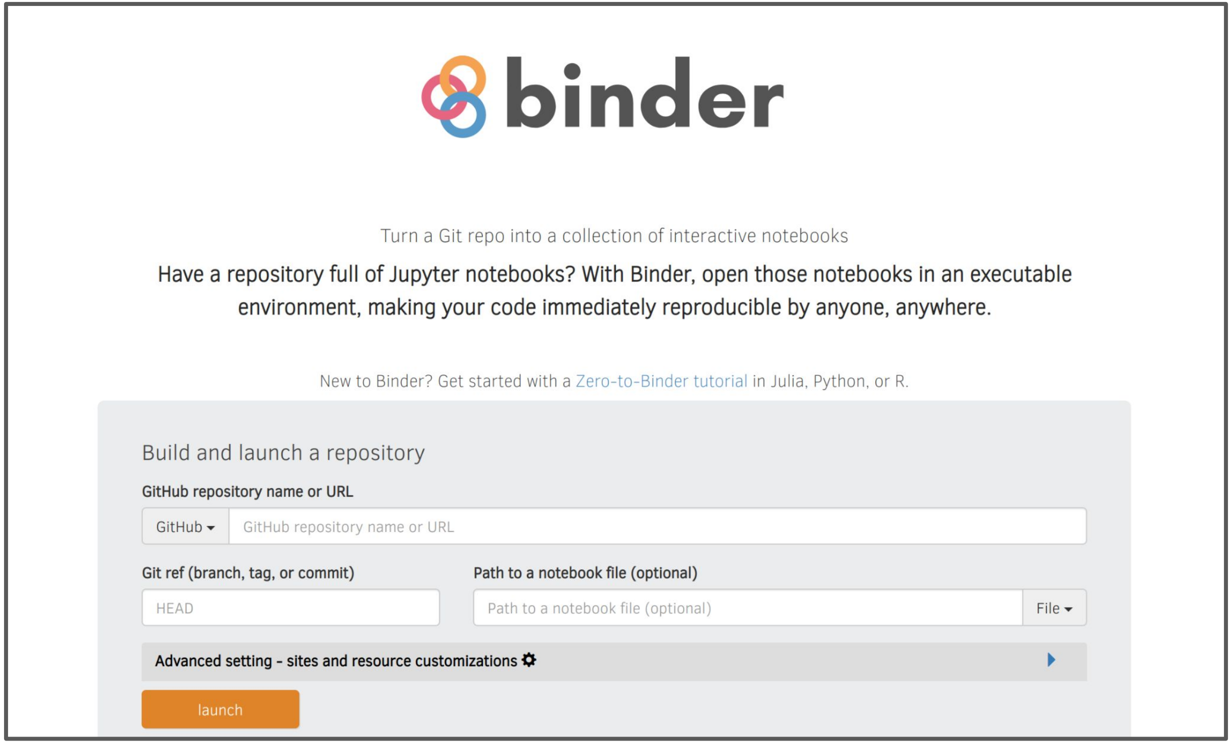
Select an Identity Provider

□ Remember this selection

By selecting "Log On", you agree to the privacy policy.

IRIS-HEP SSL BinderHub

- https://binderhub.ssl-hep.org/



Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in Julia, Python, or R.

Build and launch a repository

GitHub repository name or URL

GitHub ▾ | GitHub repository name or URL

Git ref (branch, tag, or commit): HEAD

Path to a notebook file (optional): Path to a notebook file (optional) | File ▾

Advanced setting - sites and resource customizations ⚙

Questions? Ask the organisers

Matthew Feickert
(UW-Madison)

Jamie Gooding
(Technische Universität Dortmund)

Lukas Heinrich
(Technische Universität München)

Clemens Lange
(Paul Scherrer Institut PSI)

Discussions / Questions to ask Related to HEP Analysis

- Expectations
  - Each step can run in its own unique software environment
    - Container support
    - Container runtime support (e.g. Snakemake supports Apptainer but not yet podman)
  - How do you pass state between jobs?
    - object store, filesystem, …
  - The graph needs to be dynamic
    - we need to run a step that downloads all the files, but you don’t know in advance how many files you will have to run on
    - Need the language to also not become an overhead (issue of target based languages)
  - Control flow:
    - If else constructions, dynamic, …
  - What does the user experience look like
    - We know that these tools exist, but how to they work in the typical workflow?
    - What does the scientist’s day to day look like
      - Do they work on one workflow?
      - Do you work on a partial execution for a long time?
      - How do you work at the petabyte scale?