

October 19 - 25, 2024

# CHEP 2024



## Collaboration, Reinterpretation, Outreach and Education

### Track 8 highlights

Lene Kristian Bryngemark (Lund), James Catmore (Oslo), **Giovanni Guerrieri (CERN)**, Jake Bennet (Mississippi) on behalf of the track 8 speakers

October 25<sup>th</sup> 2024

# Our speakers

[Lauren Mowberry \(STFC/UKRI\)](#)

[Thomas James \(CERN\)](#)

[Piet Nogga \(Bonn\)](#)

[Greg Corbett \(STFC/UKRI\)](#)

[Rodrigo Sierra \(CERN\)](#)

[Joni Pham \(CDM\)](#)

[Roger Jones \(Lancaster\)](#)

[Muhammad Imran \(PK\)](#)

[Kevin Pedro \(Fermilab\)](#)

[Kyle Knoepfel \(Fermilab\)](#)

[Maxim Potekhin \(BNL\)](#)

[Eoin J. Clerkin \(FAIR\)](#)

[Mindaugas Šarpis \(Vilnius\)](#)

[Axel Naumann \(CERN\)](#)

[Giovanni Guerrieri \(CERN\)](#)

[Pablo Saiz \(CERN\)](#)

[Gerardo Ganis \(CERN\)](#)

[Gordon Watts \(UW/Seattle\)](#)

[Jim Pivarski \(Princeton\)](#)

[Kati Lassila-Perini \(Helsinki\)](#)

large software challenges research code format cms  
energy physics collaboration analysis students public access  
facilitate web different development source particle  
open storage scientific atlas experiment  
lhcb science atlas system various  
project part engagement future discuss  
vr engagement cern program  
data new users

# A tentative outline

## Stealing from Axel's slides

Some of [CERN's Open Science elements](#), from its Policy:

Open Data

Open Source

Research Integrity

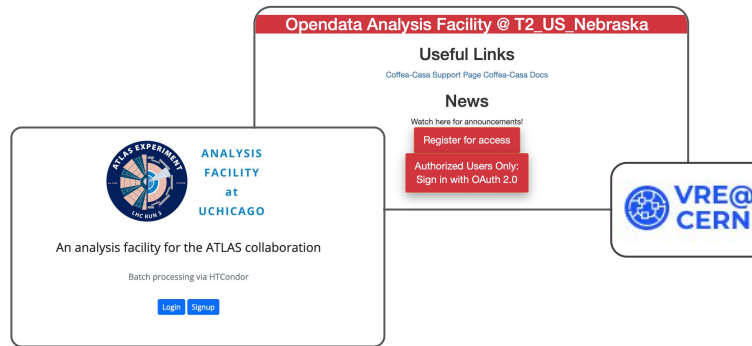
Training and Outreach

(?)

# Open Data

## The dawn of analysis facilities

Sometimes our laptop is not enough



Available at the click of a button ✓

Data persistence ✓

No account restrictions ✗

No timeout time for sessions ✓

Spawn time <1min ✓

Change easily the software stack ✓

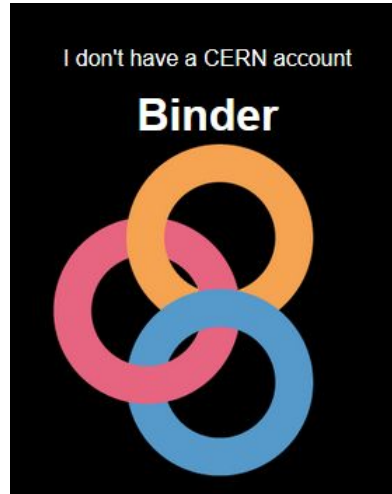
I have a CERN account

### SWAN



I don't have a CERN account

### Binder



✓ Available at the click of a button

✗ Data persistence

✓ No account restrictions

✗ No timeout time for sessions (1 CPU-h max)

✓ Spawn time ~O(min)

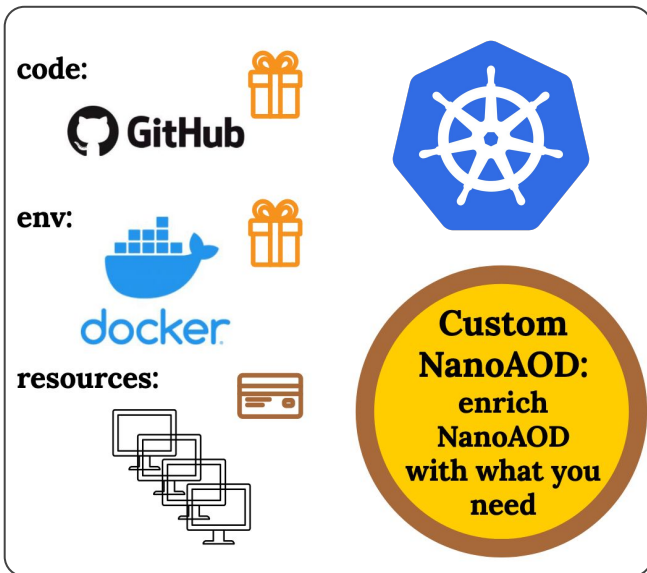
✗ Change easily the software stack  
Need to re-build the underlying image

Giovanni Guerrieri (CERN)

# Open Data

## Data processing as a service

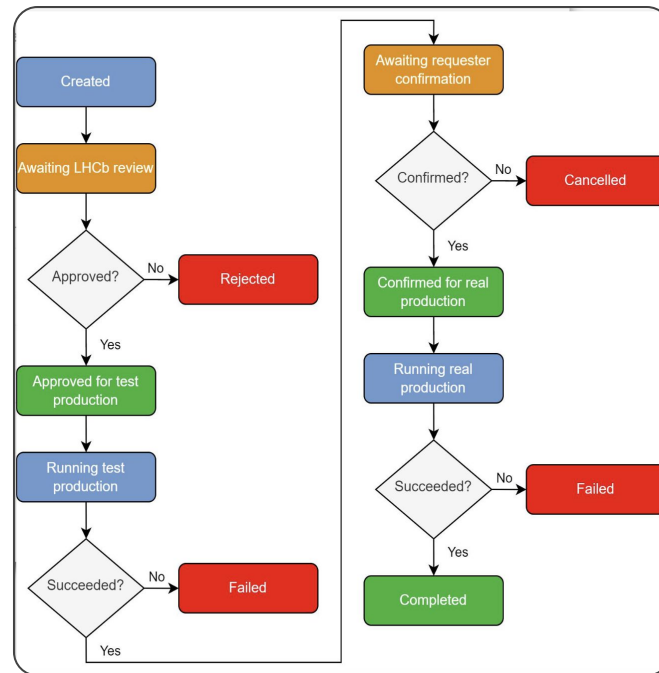
Customizing CMS NanoAOD on clouds is possible, and (reasonably) affordable



Kati Lassila-Perini (Helsinki)



The LHCb Ntupling service (coming soon - 2025)



Piet Nogga (Bonn)

# Open Data

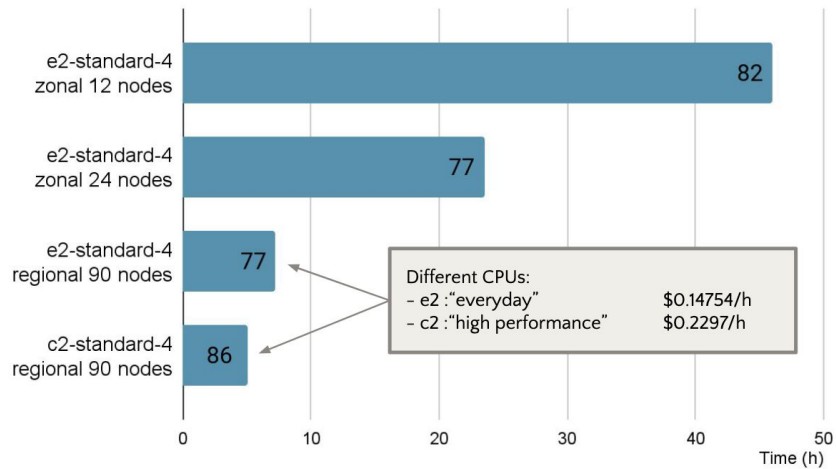
## Data processing as a service

Customizing CMS NanoAOD on clouds is possible, and (reasonably) affordable



The LHCb Ntupling service (coming soon - 2025)

Processing time (hours) and price (CHF) - autoscaling clusters



1TB of data

[Kati Lassila-Perini \(Helsinki\)](#)

[Piet Nogga \(Bonn\)](#)

### Your ntuples

#### For demonstration purposes only

This page has been created solely for demonstration purposes and cannot be used to view and download your ntuples as of yet.

All ntuples are removed from the system after 90 days. If you want to keep your ntuples, please download them in time.

Filename	Expires in	Size	Produced by			
B2HHH_MagnetDown.root	10 days	636 MiB	B Meson Decay Distributions			
PhaseSpaceSimulation.root	10 days	2 MiB	B Meson Decay Distributions			
B2psiKsPiPiTuple_2011MagDown.root	10 days	5 MiB	B Meson Decay Distributions			
00012345_00006789_1.dvntuple.root	10 days	88 KiB	B Meson Decay Distributions			

# Open Data

Store and preserve our data

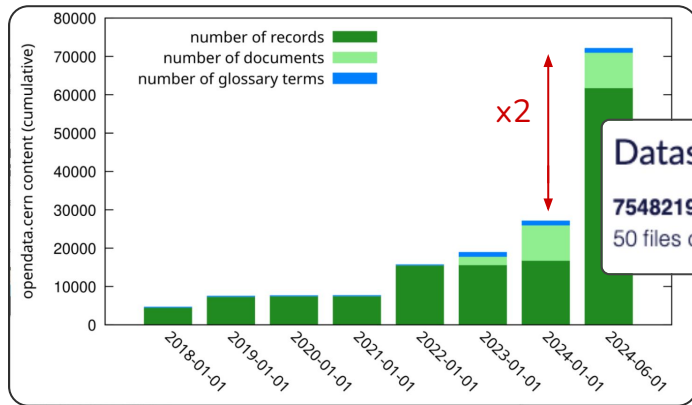


Bringing cold storage to the CERN Open Data portal

Files might be stored on cheaper media, **not immediately accessible**.

Interface with FTS and CTA to **archive/stage files**

First round with manual operations triggered by **user/curators requests**



Pablo Saiz (CERN)

## Data Preservation in High Energy Physics

2023 10-years report (DPHEP Collaboration)

DPO = Data Preservation Organisation

0: DPO during experiment proposal.

1: DPO during data taking.

2: DPO after data taking and during analysis-only mode.

3: DPO after the collaboration funding scheme.

4: DP Rescue organisational scheme

Taking **no action** == decommissioning (deleting) the data.

Gerardo Ganis (CERN)



# Open Source

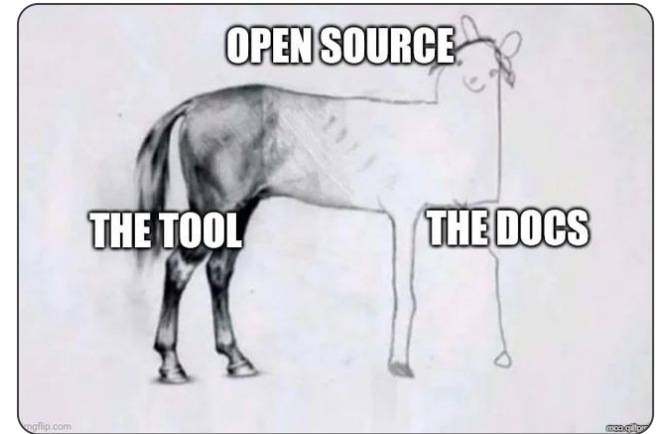
Changing the (HENP) software world, pragmatically



CERN Open Source Program Office (OSPO)

OSPO  $\neq$  open-source evangelists (but they are, really)

- CERN-wide official recommendations about procedures and best practices
- Guidance on how to run open-source projects, hardware-specific aspects
- Providing open-source hardware and software catalogs
- Center of expertise, providing KPIs and story telling



[Axel Naumann \(CERN\)](#)

# Open Science

## The CMB collaboration successfully tackling Open Science best practices

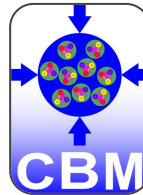
### The Compressed Baryonic Matter experiment

Ensuring that Open Science policy aids advancement to the collaboration's physics goals

Validation and verification tests on an open data policy

Providing an open and licensed software suite

[Eoin J. Clerkin \(FAIR\)](#)



• General Public License v3 (GPL-3)

Can	Cannot	Must
Commercial Use	Sublicense	Include Original
Modify	Hold Liable	State Changes
Distribute		Disclose Source
Place Warranty		Include Licence and ICopyright
Use Patent Claims		Include Install Instructions

 **GitLab** 

Our source code as well as installation instructions are publicly available online via <https://git.cbm.gsi.de/computing/cbmroot>

A next step: Obtain permanent DOI.

 **ESCAPE**  **OSSR** | Open-source Scientific Software and Service Repository 

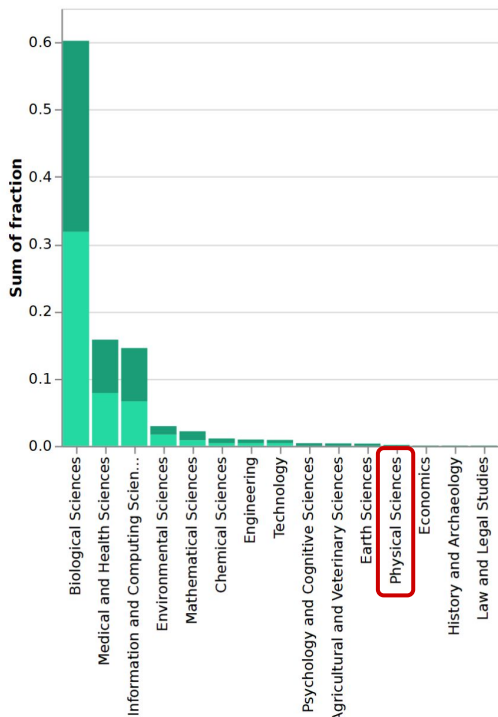
F<sub>indable</sub> A<sub>ccessible</sub> I<sub>nteroperable</sub> R<sub>eusable</sub>

F<sub>acility for</sub> A<sub>ntiProton</sub> and I<sub>on</sub> R<sub>esearch</sub> goes F<sub>indable</sub> A<sub>ccessible</sub> I<sub>nteroperable</sub> R<sub>eusable</sub>

# Research integrity

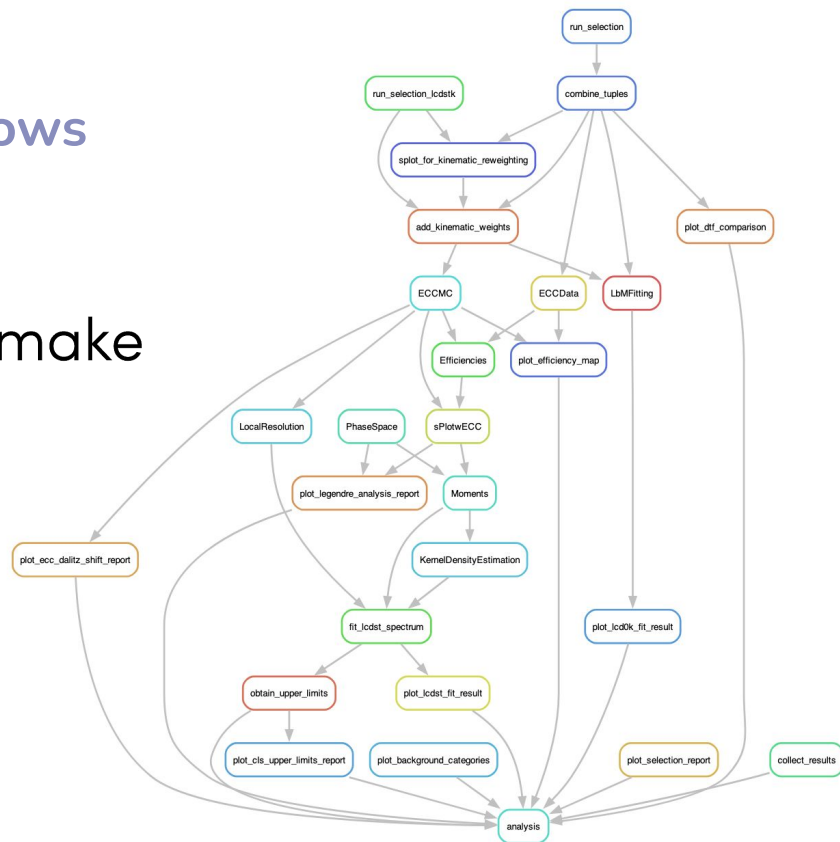
Or how to manage our physics workflows



publication  
■ original  
■ rolling



snakemake

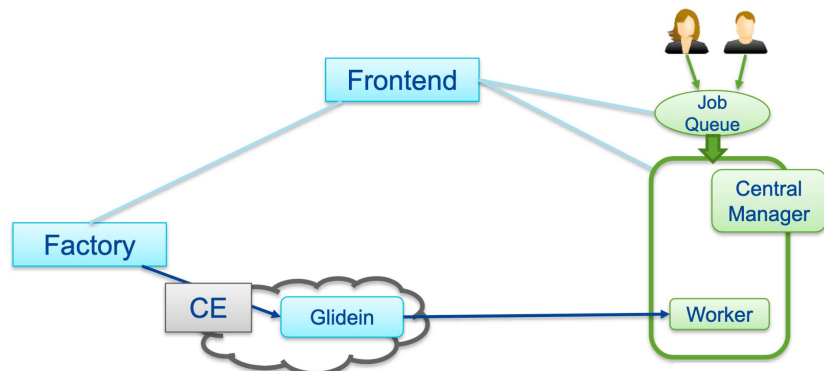


Mindaugas Šarpis (Vilnius)

# Training and Outreach

## Building the next generation of scientists

Learn how to use GlideinWMS with containers [here](#)



```
mkdir ws-test; cd ws-test
TEST_DIR=$(pwd)
git clone https://github.com/glideinWMS/containers.git
cd containers/workspaces
mkdir "$TEST_DIR"/gwms #Optional
GWMS_PATH="$TEST_DIR"/gwms/ podman-compose up -d
```

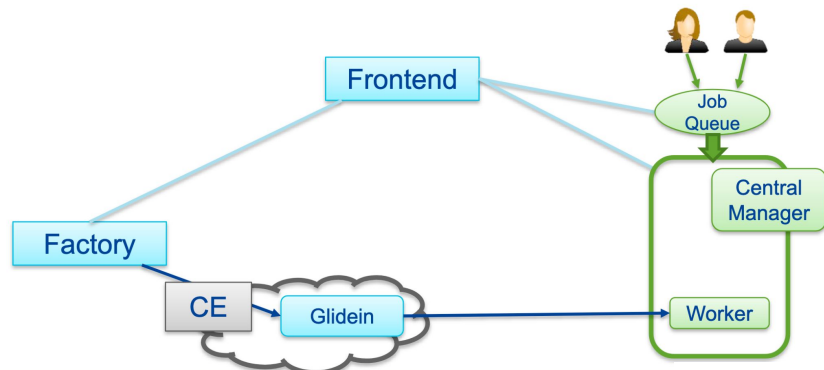
[Kevin Pedro \(Fermilab\)](#)

# Training and Outreach

## Building the next generation of scientists



Learn how to use GlideinWMS with containers [here](#)

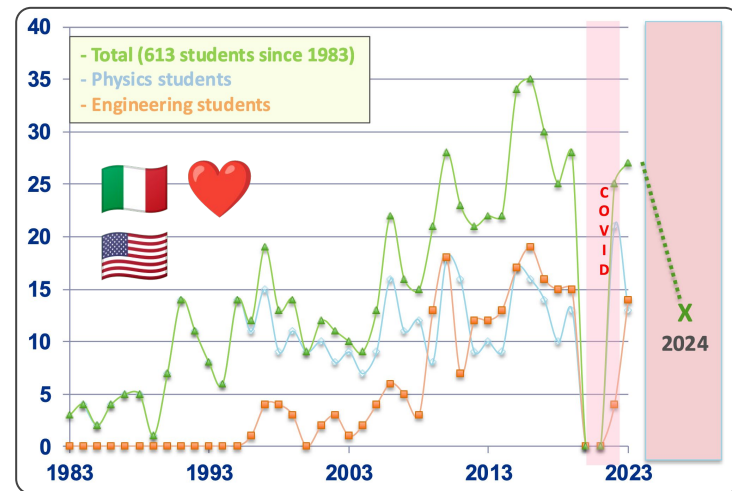


```
mkdir ws-test; cd ws-test
TEST_DIR=$(pwd)
git clone https://github.com/glideinWMS/containers.git
cd containers/workspaces
mkdir "$TEST_DIR"/gwms #Optional
GWMS_PATH="$TEST_DIR"/gwms/ podman-compose up -d
```

[Kevin Pedro \(Fermilab\)](#)

Are you Italian? Are your students Italian?

Try the [Italian summer student program](#) (not only) at Fermilab!



[Kyle Knoepfel \(Fermilab\)](#)

# Training and Outreach

Discovering experiments



The ATLAS Virtual visit program



Joni Pham (CDM)



# Training and Outreach

Discovering experiments



The ATLAS Virtual visit program



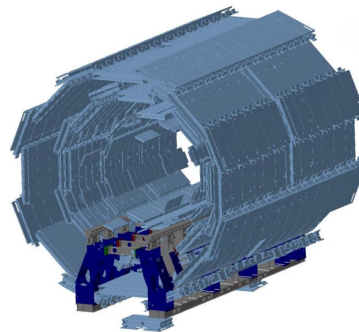
Joni Pham (CDM)



The ATLAS VR Application

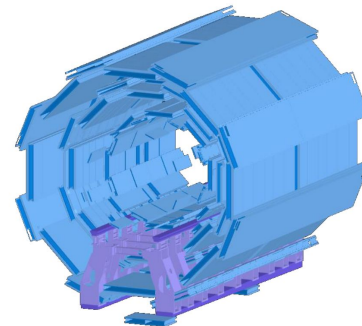


Source Geometry



Frame per Second : 7  
Triangles : 61 million  
Load time : 17.210 sec  
Draw calls : 340

Simplified Geometry

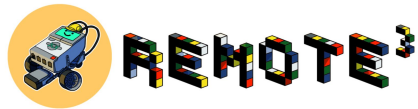


Frame per Second : 59  
Triangles : 0.015 million  
Load time : 1.210 sec  
Draw calls : 87

Roger Jones (Lancaster)

# Training and Outreach

## Discovering STEM in the UK



Students work to program LEGO Mars Rovers that would operate 1.1km underground.

[Lauren Mowberry \(STFC/UKRI\)](#)

The STFC Open Weeks project

-Engage and involve under-served communities with STFC's work

-Enable participants to feel that "science and technology are for people like me"

-STFC Open Weeks run every 4 years. *Open Week 2027?*

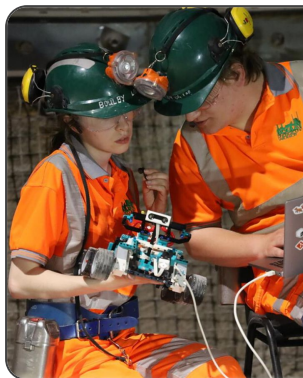


[Greg Corbett \(STFC/UKRI\)](#)



# Training

## Discovering



*"I loved working with all your colleagues but in my opinion, Will was my favourite but you're all fantastic"*  
[A Will's fan]

Lauren Mowberry (STFC/UKRI)

Greg Corbett (STFC/UKRI)

8 Weeks project

to involve under-served  
in STFC's work

to ensure all students  
want to feel that  
technology are for

8 weeks run every 4 years.  
7?



# Fueling innovation through collaboration

## The CERN Openlab industry-science journey



### Summer Student Programme

Provides undergraduate and master's level students with an opportunity to work on one of the R&D projects for nine weeks under experts' supervision

This year there was a record of more than **6600** applicants!

#### CURRENT INDUSTRY AND RESEARCH MEMBERS

SIEMENS

micron.

E4  
COMPUTER  
ENGINEERING

ORACLE

cerabyte

UNIVERSITÀ  
DEGLI STUDI DI TRIESTE

SIM NS  
FOUNDATION

INFN

intel.

#### PRE-AGREEMENT STAGE

PURESTORAGE

PASQAL

NVIDIA.

Johnson & Johnson

In discussions and negotiations with ~10 more

#### High-level: Accelerating Computing for Science

Pioneering sustainable and emerging computing and storage solutions

Harnessing heterogeneous computing and AI for a greener future

Fostering synergies and technology transfers between industry and sciences

OBJECTIVES

#### Sustainable Infrastructures

- Heterogeneous computing platforms and infrastructures
- Computer architectures and software engineering
- Storage and data management
- Artificial intelligence algorithms, platforms and applications
- Applications for society and environment

#### Emerging Technologies

- New materials for long term digital storage
- Digital twins
- Quantum computing and networks

Thomas James (CERN)

AI

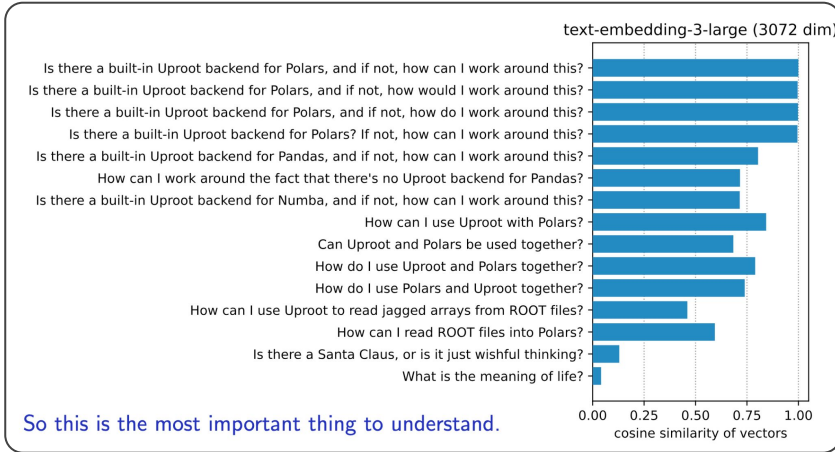
# Harness LLMs to help us

A careful approach

<https://hep-help.org>

GitHub discussion format, with LLM-chatbot to help.

Without a good embedding, we won't find the right documents



Jim Pivarski (Princeton)

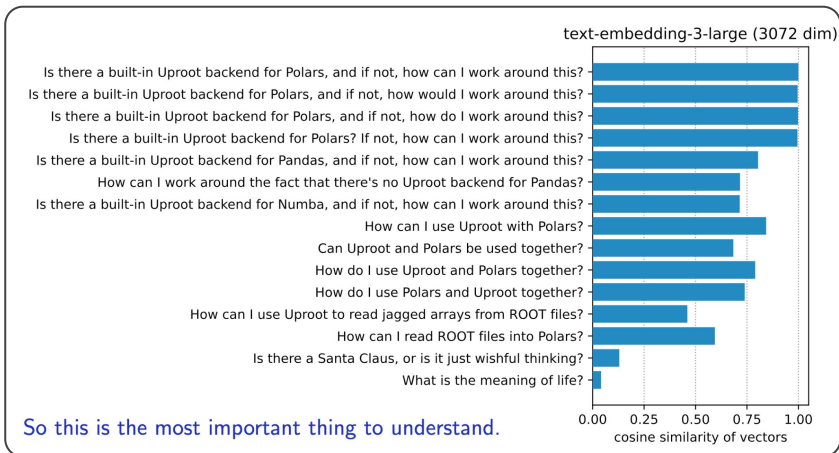
# Harness LLMs to help us

A careful approach

<https://hep-help.org>

GitHub discussion format, with LLM-chatbot to help.

Without a good embedding, we won't find the right documents



[Jim Pivarski \(Princeton\)](#)

Navigating conference abstracts

Open-Source configuration and running is painful

The large models really do make a big difference for this sort of task

```
6 # Raw prompt for the LLM
7 abstract_ranking_prompt = """Help me judge the following conference presentation as interesting or
8 not by summarizing the abstract and ranking it according to topics I'm interested in or not.
9 """
10
11 interested_topics = [
12     "Hidden Sector Physics",
13     "Long Lived Particles (Exotics or RPV SUSY)",
14     "Analysis techniques and methods and frameworks, columnar analysis, particularly those based around python or "
15     "ROOT's DataFrame (RDF)",
16     "Machine Learning and AI for particle physics",
17     "The ServiceX tool",
18     "Distributed computing for analysis (e.g. Dask, Spark, etc)",
19     "Data Preservation and FAIR principles",
20     "Differentiable Programming",
21 ]
22
23 not_interested_topics = [
24     "Quantum Computing",
25     "Lattice Gauge Theory",
26     "Neutrino Physics",
27 ]
```

[Gordon Watts \(UW/Seattle\)](#)

# Harness LLMs to help us

A careful approach

<https://hep-help.org>

GitHub discussion format, with LLM-chatbot to help.

Without a good embedding, we won't find the right documents

Navigating conference abstracts

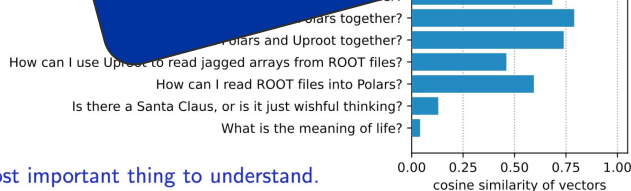
Open-Source configuration is painful

The difference

**Unity makes strength**  
Get together and tackle common LLM challenges!

text-embedder

Is there a built-in Uproot backend for Polars, and if not, how can I work around it?  
Is there a built-in Uproot backend for Polars, and if not, how can I work around it?  
Is there a built-in Uproot backend for Polars, and if not, how can I work around it?  
Is there a built-in Uproot backend for Polars, and if not, how can I work around it?  
How can I work around the issue of reading jagged arrays from ROOT files into Polars?  
Is there a built-in Uproot backend for Polars, and if not, how can I work around it?



So this is the most important thing to understand.

```
selected_topics = [
    "Hidden Sector Physics",
    "Long Lived Particles (Exotics or RPV SUSY)",
    "Analysis techniques and methods and frameworks, columnar analysis, particularly those based around python or "
    "ROOT's DataFrame (RDF)",
    "Machine Learning and AI for particle physics",
    "The ServiceX tool",
    "Distributed computing for analysis (e.g. Dask, Spark, etc)",
    "Data Preservation and FAIR principles",
    "Differentiable Programming",
]

not_interested_topics = [
    "Quantum Computing",
    "Lattice Gauge Theory",
    "Neutrino Physics",
]
```

[Jim Pivarski \(Princeton\)](#)

[Gordon Watts \(UW/Seattle\)](#)

# Summary

- **Open Data** is looking good and is gaining momentum
  - Education, training, research
- Enabling **analysis and data preservation** is critical (always has been)
  - Sustainable workflows, data preservation best practices
- **Collaborating to accelerate** physics research!
  - Enable researchers to do better science with less computing challenges
- Investing in the **next gen of HEP** is important
  - Analysis facilities, LLM helpers
- Building the **next gen of scientists**
  - Outreach, training, and public engagement



# Posters

## A twenty first century telephony solution for

**Background**

**Obsolete hardware and arduous service evolution:**

- Annual Private Automatic Branch Exchange (PABX) Analog cabling installed since the 50s
- Strategy reviewed to modernise the fixed telephony ser

**TONE (2016 - )**

**Telephony Open-source Network Evolution (TONE):**

- Provide the fixed-telephony features required by the central offices, switchboards, interconnections, or any call using standard VoIP (Voice over IP) protocols
- Avoiding vendor lock-in by using open-source components
- Significant reduction in operational costs
- Built on top of the IT department's agile infrastructure

**CERNphone (2020 - )**

**CERNphone is CERN's softphone-based telephony solution:**

- Fully standard-compliant and open source based
- Sharing the infrastructure and building blocks with TONE
- Offering all telephony features required by the users
- Desktop (Windows, macOS, Linux) and mobile (Android, iOS)
- CERNphone Users Portal for numbers, devices and advanced features: team calls, delegation, standstill, enriched features: presence, calendar, phonebook, etc.

**Call Centres (2023 - )**

**Call Centres developed in-house using TONE's building blocks:**

- Waiting queues with customised behaviour
- Automatic call distribution according to selected routes
- IP phones and CERNphone clients as agents
- Realtime information (queue, agents, calls, recording)
- Automatic statistics per call centre

**Service migration and adoption**

**All telephony services migrated to TONE/CERNphone:**

- More than 10000 CERNphone users and 100000 call
- ~11500 legacy phones/faxes/modems decommissioned
- Call centres became a standard for services, control
- Flexible voice service fully based on standards and open source
- User requirements quickly integrated into the development
- Based on IT standard components and IT best practices
- Users can use and manage their softphone anywhere
- Enriched and advanced features possible via softphone

Rodrigo SIERRA, Ihor OLKHOVSKIY, AGUST OLSZEWICZ

## Migration of CADI to Fence

Muhammad Imran<sup>1</sup>, Andrea Pflüger<sup>2</sup>, Haq Muhammad Ali Shah<sup>1</sup>, Joel Chai<sup>1</sup>, Maha (Croc) Khan<sup>1</sup> for CMS Collaboration,  
<sup>1</sup>National Centre for Physics, Islamabad, Pakistan  
<sup>2</sup>CERN, Geneva, Switzerland

**Abstract**

**CMS Analysis Database Interface (CADI):** Central tool for managing publications in the CMS experiment at CERN's Large Hadron Collider (LHC).

- Frontend engine for Glance (FENCE): Developed by EPFL-CERN collaboration, FENCE enables systems by creating interfaces through JSON configuration files.
- **FENCE Redesign:** LHC's modernized FENCE, splitting it into a PHP REST API backend and a Vue.js frontend for improved flexibility.
- **Migration to FENCE:** CMS decided to migrate CADI to the modular FENCE system, following the new ATLAS LHC and CMS strategy.
- **Subsystems:** Migration initially focused on two FENCE subsystems:
  - Membership: Manages members, institutes, authorship, and reports.
  - ALCM: Manages workflows such as CADI, facilitating the publication process.
- **Challenges:** Several issues were encountered during the migration process.

**Introduction**

- CADI is a management tool for physics publications in the CMS experiment. Also:
- Acts as a central database, tracking analysis papers conducted by CMS researchers.
- The journey from early analysis to publication involves multiple stakeholders providing feedback and approval.

**Motivation**

- FENCE Originally used by the ATLAS experiment, now has been adopted by LHC and also, utilizing a modular architecture with PHP-based REST API backend and VueJS frontend.
- CMS plans to migrate to the FENCE system, starting with the membership and ALCM subsystems.
- NCP took institutional responsibility for leading the migration of CADI to Fence. The plan involved NCP personnel rotating to CERN for onsite collaboration, working closely with the CMS and LHC teams, while the remaining team members provided remote support.

**Comparison of Architectures**

- A comparison of the architectures of the old CADI and new FENCE systems.



Figure 1. CADI vs FENCE architecture comparison

CHEP BOB4

**Membership**

- A prerequisite for all other applications to function.
- It holds essential information such as members, institutes, authorship, and reports.

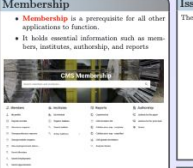


Figure 2. Membership

**ALCM**

- The ALCM module manages publications like CADI and includes the following sub-modules:
  - **Figures:** Manage the figure types associated with publications.
  - **Analysis:** Handle the analysis aspects of the publication life-cycle.
  - **Notes:** Provide functionality for managing notes related to the publications.



Figure 3. ALCM

CHEP BOB4

## Collaborative Tools for the ePIC Experiment



The ePIC Collaboration is working on the design of a future particle detector for the Electron-Ion Collider (EIC) at the Brookhaven National Laboratory. This experiment will collide a high-energy electron beam with protons and ions, and will use streaming readout to seamlessly deliver the data obtained by the EIC systems to distributed computing facilities for processing. This will make possible analyses that will address some of the most fundamental questions in science regarding the visible world, including the origin of the nuclear matter, the nuclear spin, and the emergent properties of a dense system of quarks.

Currently the Collaboration is working on the Technical Design Report (TDR) for the detector, which requires a plethora of physics and detector studies based on sophisticated and well-coordinated simulations. Work is underway on the ePIC streaming readout, and the software and computing framework.

These activities are conducted by a large international Collaboration with over 850 members from 25 countries, and require an effective set of Collaborative Tools and an open, collaborative environment to be able to succeed.

**ePIC Collaborative Tools by Category**

- Web Presence
- Digital Repositories
- Collaborative Document Development
- Collaborative Software Development
- File Sharing and Data Delivery
- Collaborative Communication

**The New Website (in Development)**

- Static website generation (Jekyll) - same as on the GitHub pages
- Simplification use of structured data is possible by parsing files formatted in YAML, with the help of the Liquid templating language
- The text content format is Markdown
- Once generated, the site becomes a static collection of HTML files that can be deployed on any web server, and offers the following advantages:
  - High performance (due to the static nature of the content)
  - Security (due to absence of a live backend database)

**The Digital Repository**

- ePIC uses a state-of-the-art digital repository - Zenodo at CERN
- Recent advances in this technology platform include powerful management tools for the content created by the organization: the Zenodo Communities
- ePIC keeps a list of managed keywords, with automatic links on the website for user and content queries
- Given the convergence of Zenodo and Invenio RDM, techniques for migration to a possible future document development workflow system based on open source are being investigated

**File Sharing and Data Delivery**

- Google Drive remains a popular tool for document sharing and collaborative development in ePIC, complementing the repository function of Zenodo
- Regarding the needs of data handling at scale, ePIC is forming the software experience of many HEPWFN experiments, and is utilizing XRootD to share and store the data needed for its Monte Carlo campaigns in excess for the TDR, and in software training programs
- ePIC is operating an XRootD cluster deployed at Jefferson Laboratory as its central data hub, with keen hardware acceleration for access control
- There are plans to enhance the ePIC distributed storage capability by deploying another XRootD cluster at Brookhaven National Laboratory
- Please see the talk: **The ePIC Simulation Campaign Workflow on the Open Science Grid** given at this conference. <https://indico.cern.ch/event/1318668/contributions/6101555>

**Collaborative Software Development**

- Please see the talk: **Collaborative software and experiment at ePIC** given at this conference. <https://indico.cern.ch/event/1318668/contributions/6101570>

**Scheduling and Communication**

- Indico
- Zoom
- Mattermost

**Summary**

The ePIC Collaboration is actively engaged in the development of the most recent software for its future detector to open Collaborator, and its software and computing capability remain in support of this work, the Collaboration has undertaken an area of Collaborative Tools, utilizing best practices and patterns.

## Evolution of Regional, Age and Gender Demographics in the ATLAS Collaboration

Gordon Watts, for the ATLAS Collaboration

**Why is this work important?**

"UNITE PEOPLE FROM ALL OVER THE WORLD TO PUSH THE FRONTIERS OF SCIENCE AND TECHNOLOGY, FOR THE BENEFIT OF ALL"  
 - CERN Mission Statement

**ATLAS is at its best when everyone is fully participating**

This work is one component tracking how well the policies & procedures in place are working. It is impossible to trace a single policy to a single effect.

In many cases control over policy is indirect - institutional hiring policies and procedures, etc. But those can be over time. Some effects can be seen in the data...  
**Evolution of Gender in Leadership Roles**

Statistics are very low, so time evolution is hard to track!

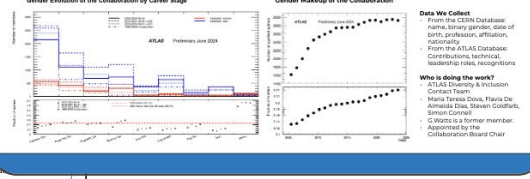


**Why do we track statistics by age?**

Many positions are long-term, so the policies of the day are baked in.

**Definitions**

- **Member:** someone who belongs to the ATLAS Collaboration
- **Author:** someone whose name is included in ATLAS publications
- **Gender:** binary options (male/female), from government ID
- **Home Institution:** research or educational organization to which a person belongs (not nationality)



Rodrigo Sierra (CERN)

Muhammad Imran (PK)

Maxim Potekhin (BNL)

Gordon Watts (UW/Seattle)



# Thanks

To our speakers, to our organisers, and to everyone!