PAUL SCHERRER INSTITUT

PSI

CERN Open Science

Open source software projects in high-energy physics

Some lessons learned

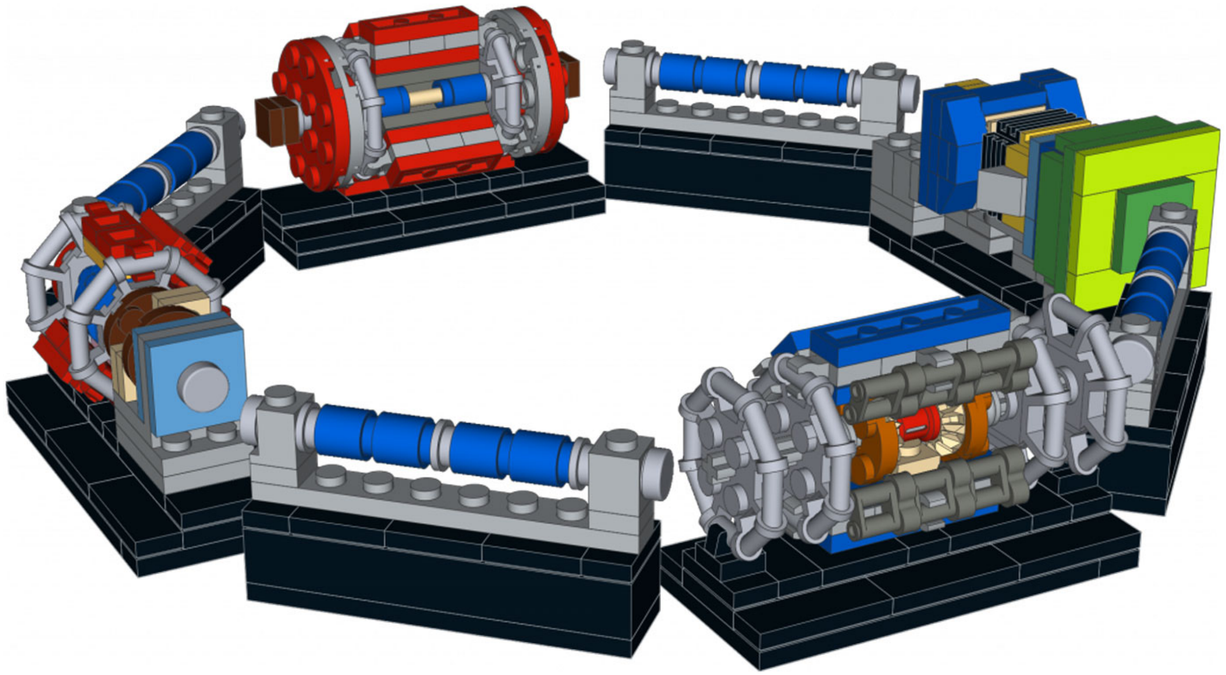
Clemens Lange (Paul Scherrer Institute PSI)

BOINC Workshop 2024

29th May 2024

High-energy physics at the LHC

> The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator



> Four large experiments:

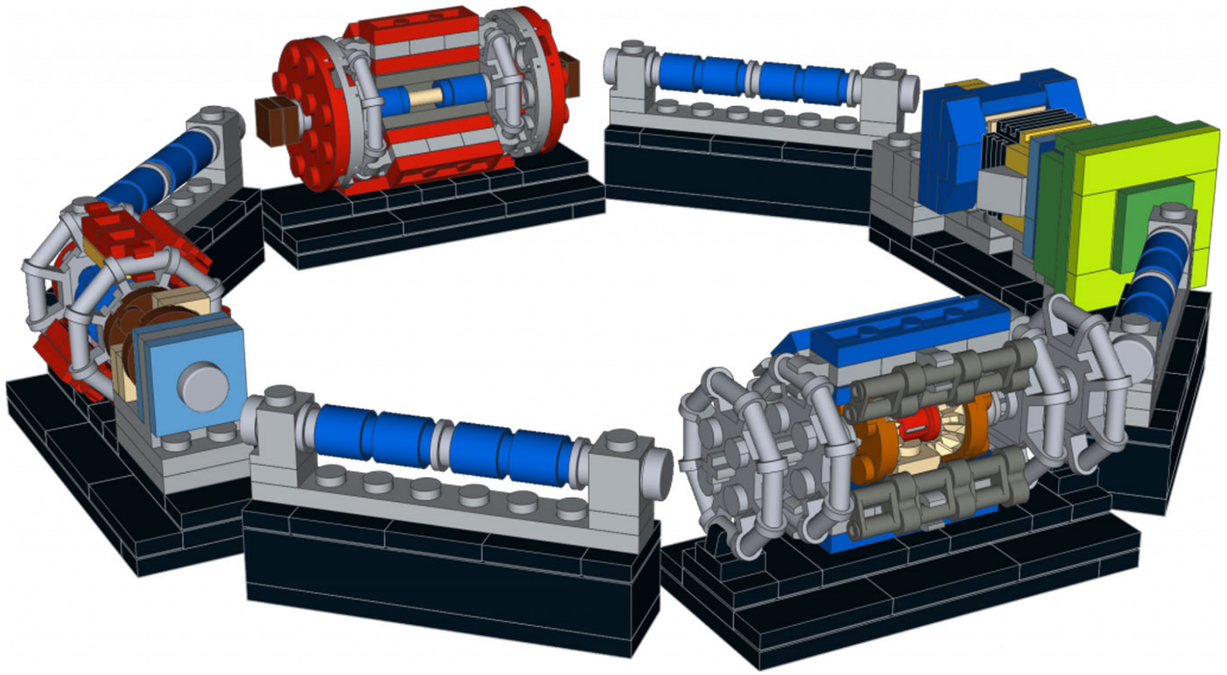
- ATLAS (5500 members of which almost 3000 scientific authors)
- ALICE (almost 2000 members)
- CMS (4000 particle physicists, engineers, computer scientists, technicians and students)
- LHCb (about 1700 scientists, engineers and technicians)

> … plus several smaller ones

Image: Nathan Readioff

High-energy physics at the LHC

> The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator



> Four large experiments:

- ATLAS (5500 members of which almost 3000 scientific authors)
- ALICE (almost 2000 members)
- CMS (4000 particle physicists, engineers, computer scientists, technicians and students)
- LHCb (about 1700 scientists, engineers and technicians)

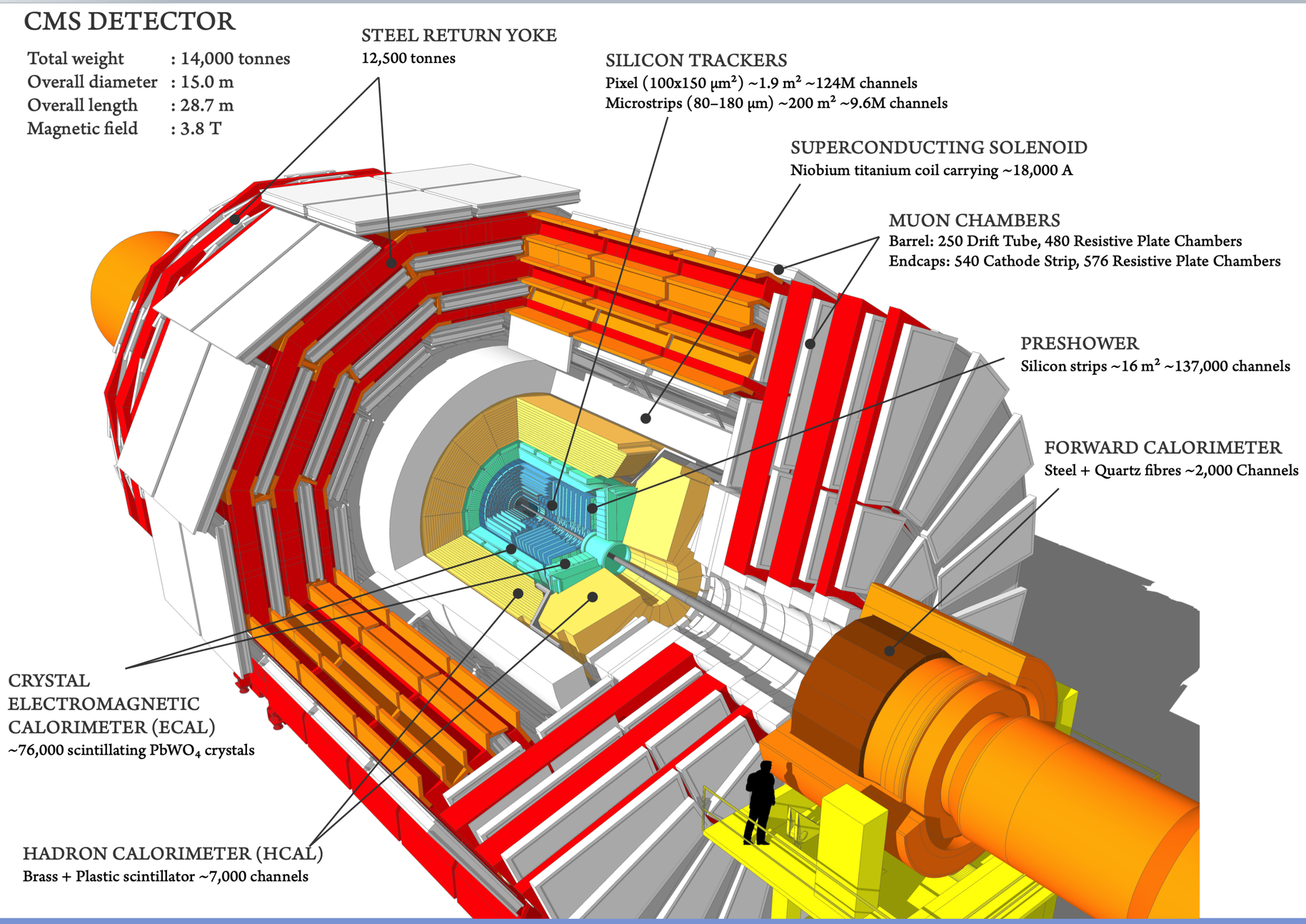
> … plus several smaller ones

Today: more than 13,000 people involved in the experiments

Image: Nathan Readioff

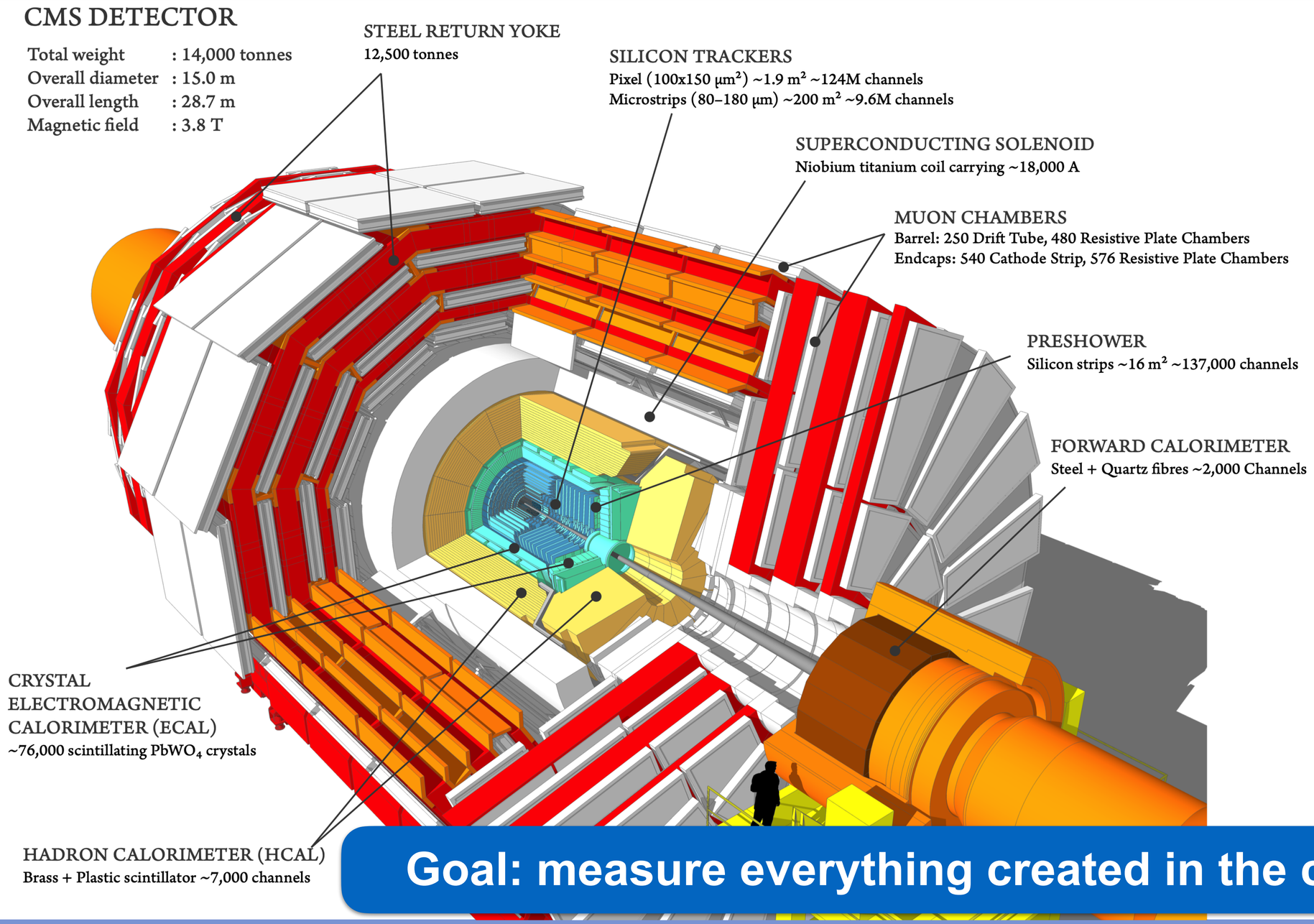
Large collaborations with huge detectors

arXiv:2309.05466



Large collaborations with huge detectors

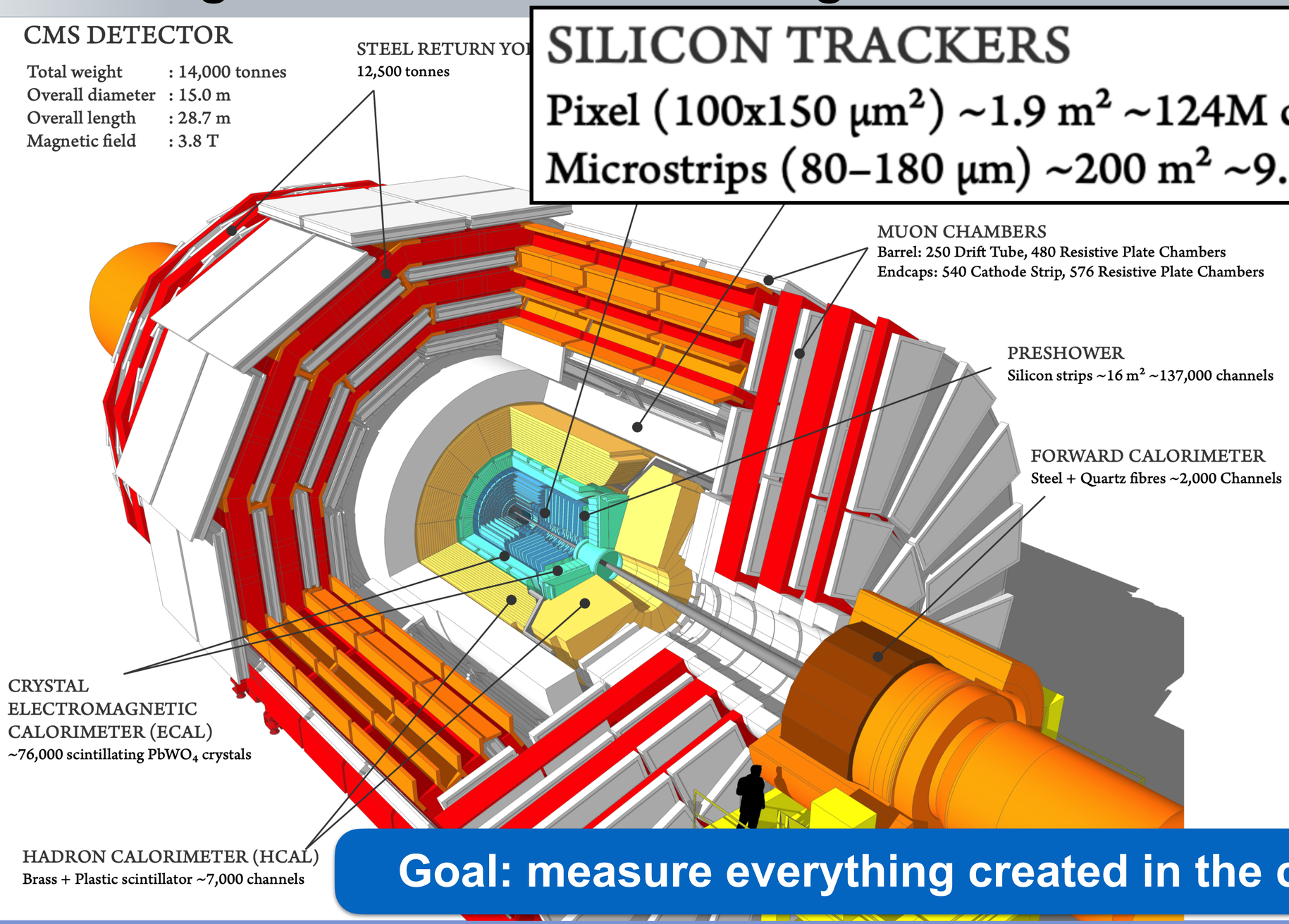
arXiv:2309.05466



Goal: measure everything created in the collisions

Large collaborations with huge detectors

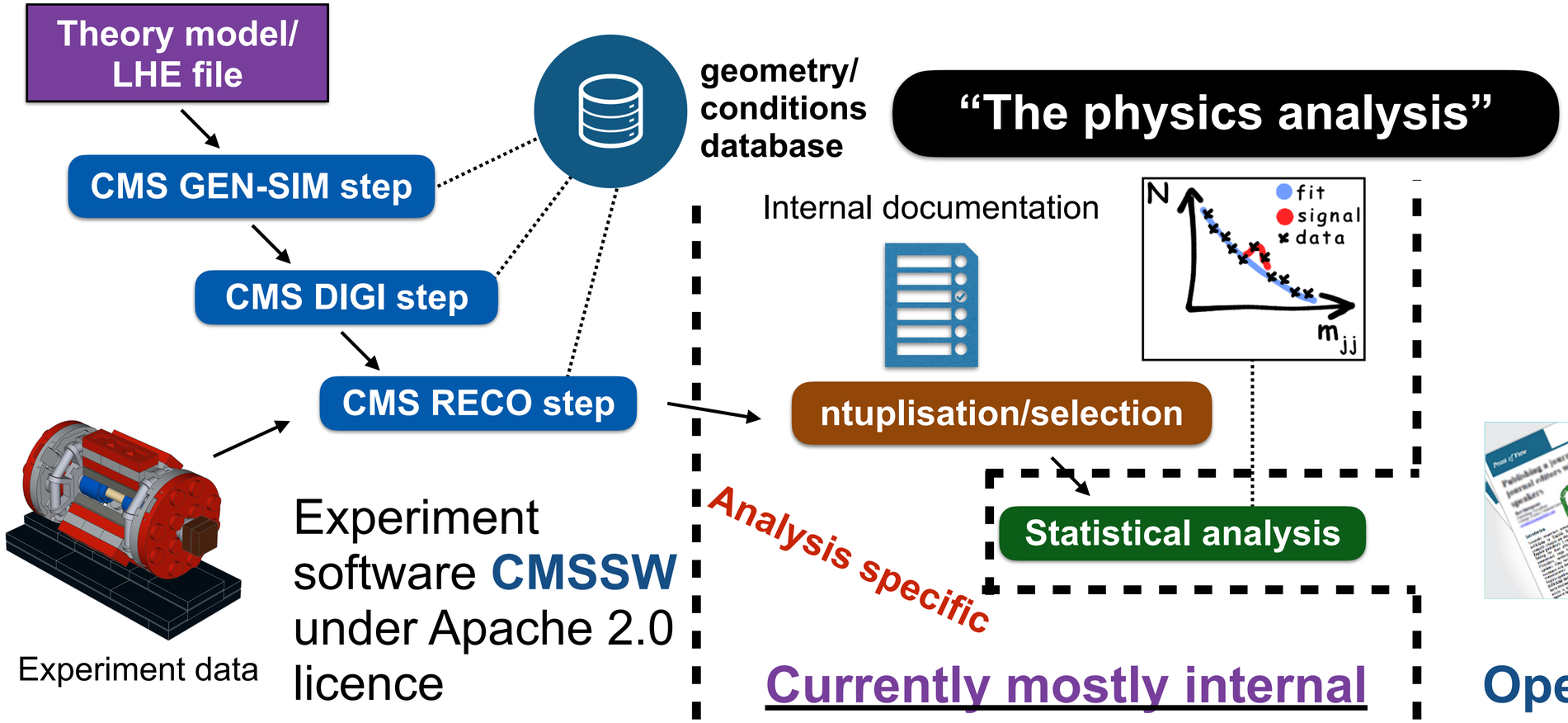
arXiv:2309.05466



Goal: measure everything created in the collisions

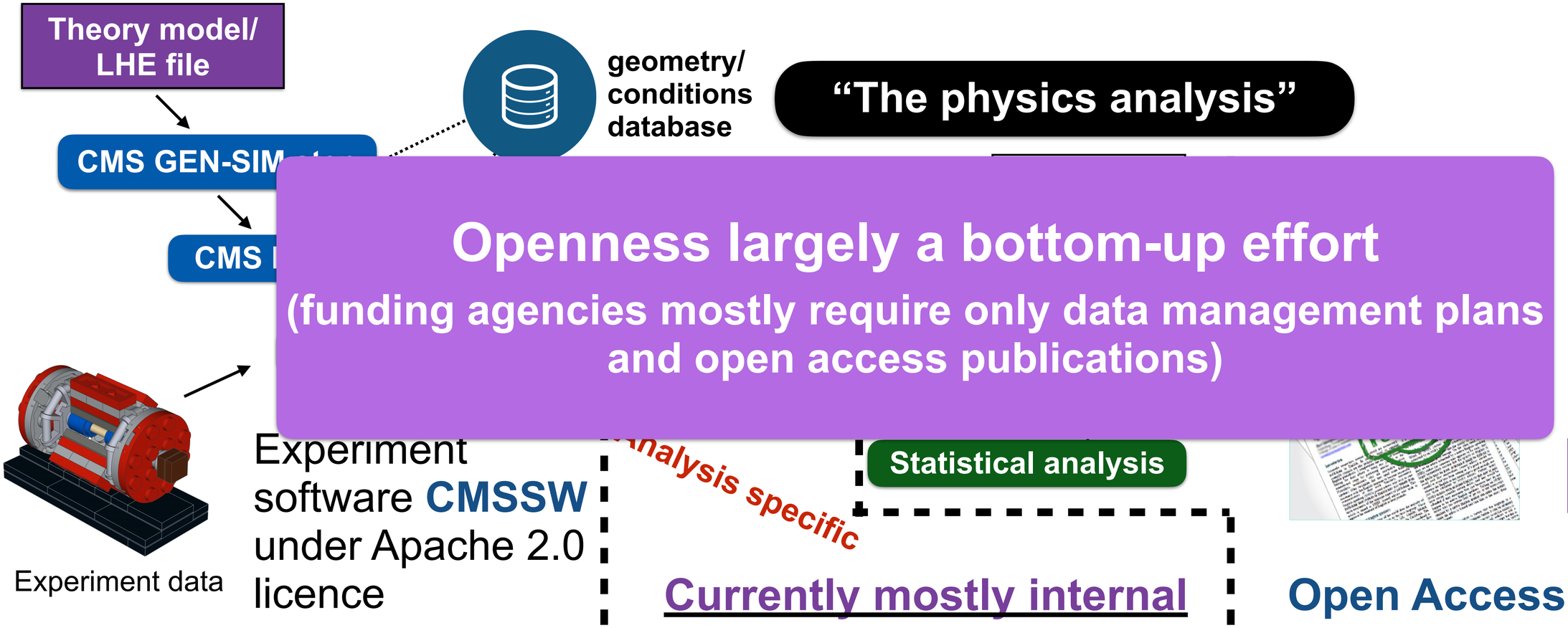
Open Science in CMS

- CMS first LHC experiment to release research-level open data, by now several petabytes available (as well as data for education and outreach)
  - Available after embargo period
- Large parts of the software open source; publications openly accessible



Open Science in CMS

- CMS first LHC experiment to release research-level open data, by now several petabytes available (as well as data for education and outreach)
  - Available after embargo period
- Large parts of the software open source; publications openly accessible



CERN Open Science Policy

Captures current practice and states vision across multiple Open Science domains:

- Open Access to Publications
- Open Research Data
- **Open Software**
- Open Hardware
- **Citizen Science**

v1.0 released Oct 2022: https://cds.cern.ch/record/2835057

- For more information, see https://openscience.cern/
  - Have a look at the implementation plan!

CERN Open Science

CERN Open Science Policy

Captures current practice and states vision across multiple Open Science domains:

- Open Access to Publications
- Open Research Data
- **Open Software**
- Open Hardware
- **Citizen Science**

Policy goes beyond what is currently already being done

→ Call to action

v1.0 released Oct 2022: https://cds.cern.ch/record/2835057

- For more information, see https://openscience.cern/
  - Have a look at the implementation plan!

CERN Open Science

PAUL SCHERRER INSTITUT

PSI

CERN Open Science

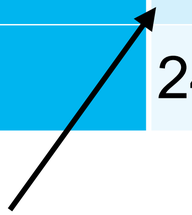
Open Source software projects in HEP

Selected examples

CMSSW vs. BOINC — an arbitrary comparison

- Both projects are hosted on GitHub: https://github.com/cms-sw/cmssw/ and https://github.com/BOINC/boinc
  - (Mind: the reason CMSSW is hosted on GitHub is that CERN only provided a simple Git server at the time the move to Git was decided — other experiment software largely on CERN GitLab instance)

numbers as of 28th May 2024	CMSSW	BOINC
Forks	~4200	438
Stars	1.1k	1.9k
Contributors	1,149	128
Commits	244,708	36,389



of a potential 18k

CMSSW vs. BOINC — an arbitrary comparison

- Both projects are hosted on GitHub: https://github.com/cms-sw/cmssw/ and https://github.com/BOINC/boinc
  - (Mind: the reason CMSSW is hosted on GitHub is that CERN only provided a simple Git server at the time the move to Git was decided — other experiment software largely on CERN GitLab instance)

numbers as of 28th May 2024	**CMSSW**	**BOINC**
Forks	~4200	438
Stars	1.1k	1.9k
Contributors	1,149	128
Commits	244,708	36,389

of a potential 18k

~70% of all forks never contribute any changes upstream

GitHub insights over the past month

CMSSW

April 28, 2024 – May 28, 2024

Period: 1 month

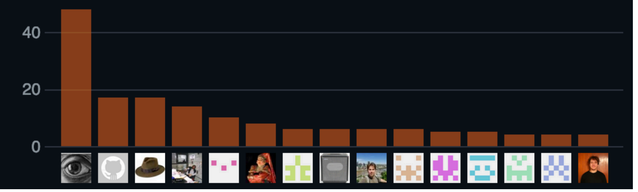
Overview

222 Active pull requests

71 Active issues

162	60	39	32
Merged pull requests	Open pull requests	Closed issues	New issues

Excluding merges, **45 authors** have pushed **138 commits** to master and **213 commits** to all branches. On master, **1,119 files** have changed and there have been **62,640 additions** and **31,516 deletions**.



BOINC

April 28, 2024 – May 28, 2024

Period: 1 month

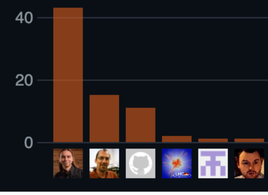
Overview

20 Active pull requests

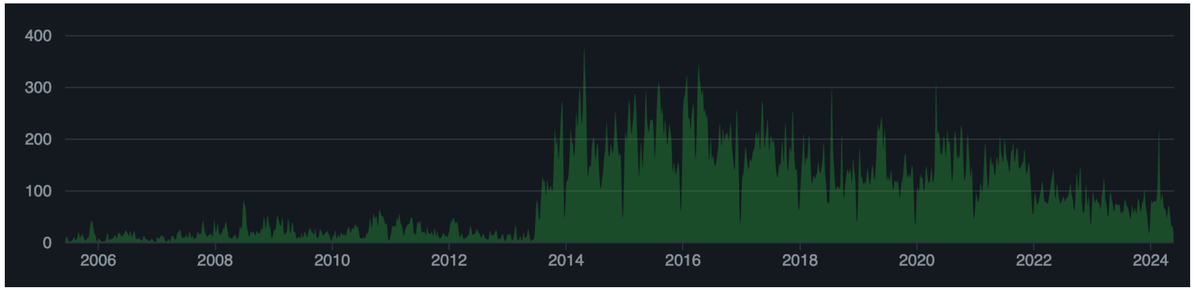
22 Active issues

19	1	16	6
Merged pull requests	Open pull request	Closed issues	New issues

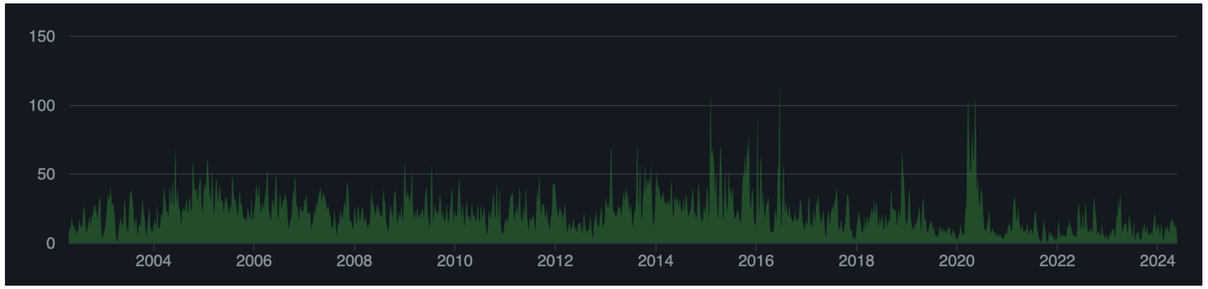
Excluding merges, **6 authors** have pushed **39 commits** to master and **73 commits** to all branches. On master, **83 files** have changed and there have been **2,673 additions** and **394 deletions**.



Contributions to master (limited usefulness in case of CMSSW) :



2006 2024



2004 2024

Managing experiment software in HEP

- In CMS, **every researcher** signing the scientific publications **needs to contribute to the operation, maintenance, and upgrade of the experiment**
  - Qualify as an author (6 months of service work)
  - Service work and detector shifts (4 months/year)

→ Continuous influx of possible contributors

- Subgroups in the experiment responsible for “their” part of the software
- Additionally, release coordinators and reconstruction group conveners
- Discussion beyond GitHub during **weekly meetings**
- **Rely heavily on automation**

Challenges and issues maintaining CMSSW

- CMSSW is ~18 years old (and will need to continue to work for about the same time)
  - Continuous modernisation efforts required
  - Removing unused packages difficult
- Software largely written in C++ (Python for configuration/steering)
  - **(Physics) students do not learn C++ anymore these days** → risk losing developer base
- High complexity
  - Size tends to increase more and more
  - Software needs to be written with focus on efficient computing
- Complicated sign-off process
  - **Medium-size changes often take months to get merged**
- **Lack of documentation**

Mind: these are personal observations

Typical open source projects in HEP

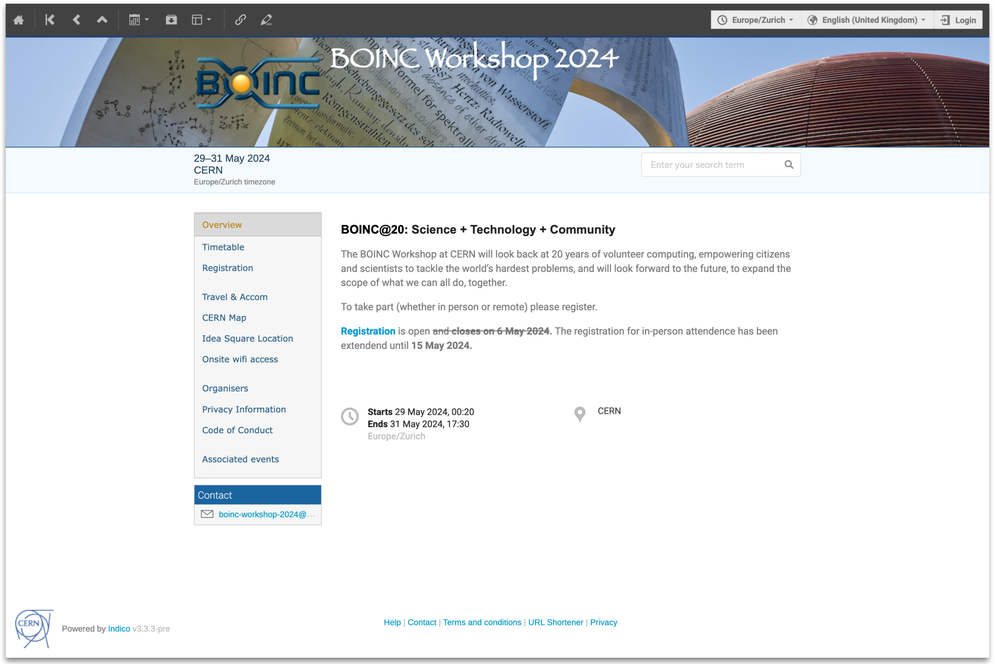
Differentiate between institutional and grant-based funding

- Institutional funding example: ROOT (https://root.cern/, since 1995)
  - Guaranteed longevity due to staff with indefinite contract duration
  - Project age/complexity makes contributing more difficult
  - Vast majority of contributions from internal contributors
- Grant-based funding example: Scikit-HEP ecosystem (https://scikit-hep.org/, since 2016)
  - Smaller projects with small number of core contributors
  - Making significant effort to attract and guide new contributors to enable possible longevity (including documentation)
  - User risk of projects getting abandoned
- Commonalities:
  - Contributions reflect contract durations

Mind: these are personal observations

Beyond research-level software

- Software projects on previous slides mostly only with applications in research
- However, several other projects from the CERN community also used outside
- Example: Indico (the tool used for the organisation of this event)
  - Largely developed by CERN IT (i.e. non-scientific personnel)
  - CERN as a lab is an infrastructure provider
- Efforts to attract contributors:
  - CONTRIBUTING.md file in repository
  - Discussion forum
  - Regular meetings/office hours(?)
  - CERN-internal Mattermost channel
- Deployed by technical personnel
  - Might be more inclined to contribute back(?)



Personal (closing) observations

- Researchers and technical personnel (and citizen scientists) need to **learn how to contribute** to open source software projects and make their project open source
  - Involves steep learning curve and personal effort
  - Requires **training** opportunities (e.g. HEP Software Foundation Training Initiative)
  - **New**: CERN Open Source Project Office
- Contributions typically only to solve personal problems/feature needs
- **Heavy reliance on individual maintainers** with long-term contracts
  - These are also often busy with other projects… → need **automation**
  - External contributions from part-time developers rare, **hackathons can help**
  - Slow project velocity might indicate a stable and feature-complete product
- Being “too technical” often seen as negative for research careers
  - Need **new research assessment practices** (see e.g. https://coara.eu/)


PAUL SCHERRER INSTITUT
PSI


CERN
Open Science