

Preservation/reinterpretation efforts in CMS

Louis Moureaux for the CMS Collaboration

Contact: cms-physics-coordinators@cern.ch
LHC BSM WG General meeting – 13.11.2025



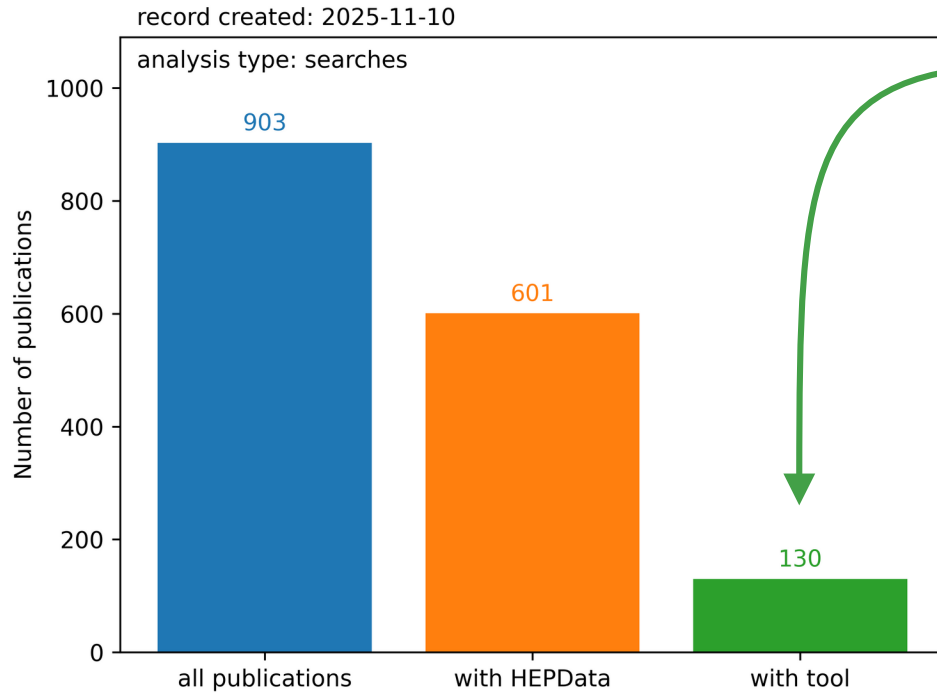
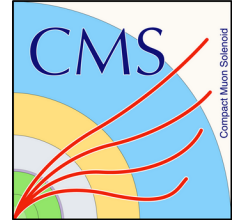
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Sponsored by the:



Federal Ministry
of Research, Technology
and Space

Reinterpretation material

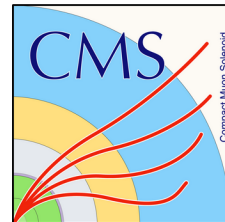


General status: not perfect...

- Looks like a structural issue
- Structural solutions needed

This talk: what is CMS doing?

What we do publish



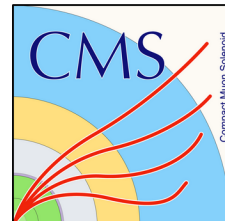
HepData

- Requirement for every analysis (few exceptions)
- Quality of records varies
- Created by analysts

Statistical models on CDS **new** ([link](#))

- COMBINE datacards used to get the results
- Built upon internal review pipelines
- Largely automated

What we do publish



Open data ([link](#))

- All (most) samples become available after some time
- Not easy to match to papers
- Not always easy to inspect

51,955 result(s) found

Sort by Most recent



[/ZZZ_TuneCP5_13TeV-amcatnlo-pythia8/RunII Summer20UL16NanoAODv9-106X_mcRun2_asymptotic_v17-v1/NANO AODSIM](#)

Simulated dataset ZZZ_TuneCP5_13TeV-amcatnlo-pythia8 in NANO AODSIM format for 2016 collision data. See the description of the simulated dataset names in: About CMS simulated dataset names. These simu...

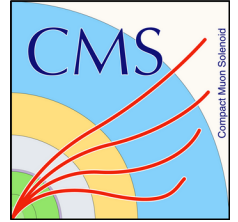
Dataset Simulated Standard Model Physics ElectroWeak CMS

[/ZZZ_TuneCP5_13TeV-amcatnlo-pythia8/RunII Summer20UL16NanoAODv9-106X_mcRun2_asymptotic_v17_ext1-v1/NANO AODSIM](#)

Simulated dataset ZZZ_TuneCP5_13TeV-amcatnlo-pythia8 in NANO AODSIM format for 2016 collision data. See the description of the simulated dataset names in: About CMS simulated dataset names. These simu...

Dataset Simulated Standard Model Physics ElectroWeak CMS

A story



I once had a student...

- Excellent physicist
- Well aware of CMS sample production mechanisms
- Writing their PhD thesis
- Could not figure out precisely what was inside one of our own datasets

How can we prepare good reinterpretation material in this situation?

Lifetime of a simulated sample

- 1) Idea of a topology/model to test
- 2) Find a way to generate it, submit jobs

...some months...

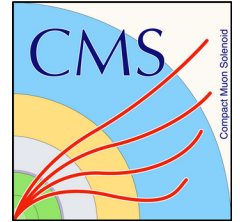
- 3) Samples are ready

...2 years...

- 4) Analysis is done

...some more years...

- 5) Released as open data

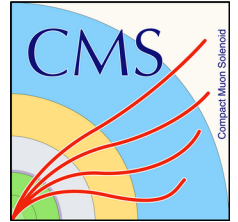


Finding & preserving knowledge

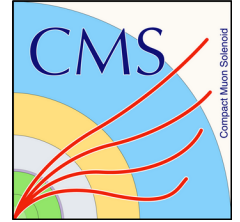
Writing good documentation is *hard*

- How to answer questions that will come up 5 years later?
- Guidelines/tutorials help a lot but not sufficient on their own

We also need to make *finding* the answers easier

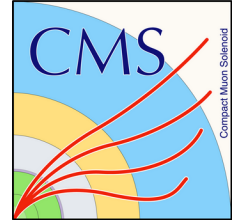


Some really basic questions



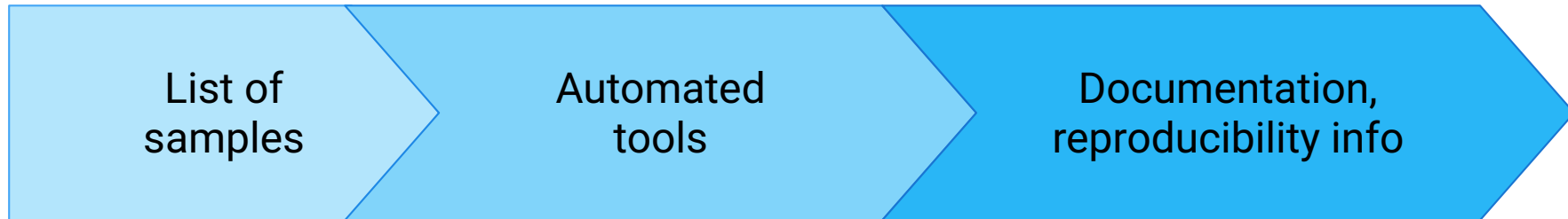
- Which samples? → Starting to record this centrally
- Cross sections? → To be recorded centrally
- Specific decays? → Generator-specific knowledge required
- Physics processes? → Hints in the sample name
- Which generators? → Hints in the sample name
- Which versions? → Not trivial to find
- Specific parameters? → Generator-specific knowledge required
- Postprocessing? → Not easy to find/understand
- How to generate my own copy? → Requires answering all these questions

Answering questions

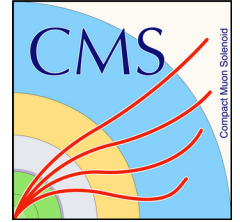


To answer the questions, we need:

- **Organizational changes:** map datasets to analyses
- **New tooling:** answer the basic questions automatically
- **Goal:** enable full reproduction of the samples outside CMS



Internal organization



Every CMS analysis has a GitLab area:

- Used to store COMBINE datacards
- Pipelines to check & publish them

Extending the idea with more info:

- Samples, cross sections
- Cut flows, ...?
- Minimize overhead, maximize impact

A screenshot of a GitLab repository structure. It shows three main repository entries, each with a dropdown arrow, a group icon, a letter in a colored box, and a lock icon. The first entry is 'SUS-24-015' with a grey 'S' box. Below it is a sub-repository 'datacards' with a red 'D' box and the description 'SUS-24-015 datacard repository'. The second entry is 'SUS-24-006' with an orange 'S' box. Below it is a sub-repository 'datacards' with a grey 'D' box and the description 'SUS-24-006 datacard repository'. The third entry is 'SUS-24-014' with an orange 'S' box, and this entire row is highlighted with a light blue background. Below it is a sub-repository 'datacards' with an orange 'D' box. The text 'Compact Muon Solenoid' is visible vertically on the right side of the image.

New tooling

Collect information about a sample

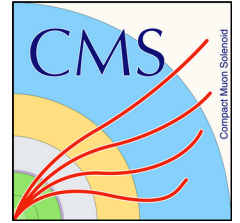
From sample configuration, gridpacks, cards, ...

Present it in computer- and human-readable formats

Could be released together with every paper

Welcoming feedback on the concept:

- What should we include?
- How to present information?
- Interest in code sharing? Shared output format?



Summary

CMS aims to provide reinterpretation material more systematically:

- Statistical models for COMBINE **new** ([link](#))
- Datasets used by analyses
- Cut flows, other metadata

Organization and automation are key:

- Minimize overhead to maximize adhesion
- Tool to describe MC samples automatically, provide descriptions alongside papers

