Analysis preservations efforts at the CMS Collaboration

Alejandro Gomez Espinosa
On behalf of the CMS Experiment

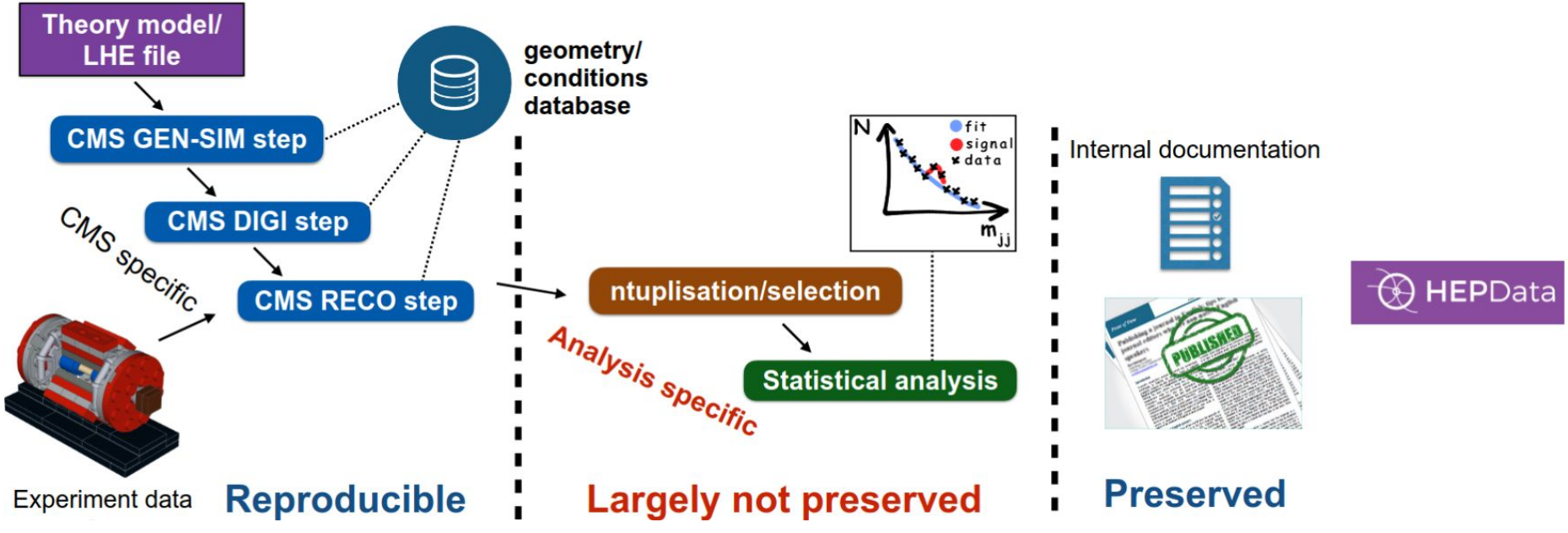
Feb 26th, 2025

Carnegie Mellon University

CMS analysis preservation efforts

CMS has two sets of well-established workflows related to physics analyses:

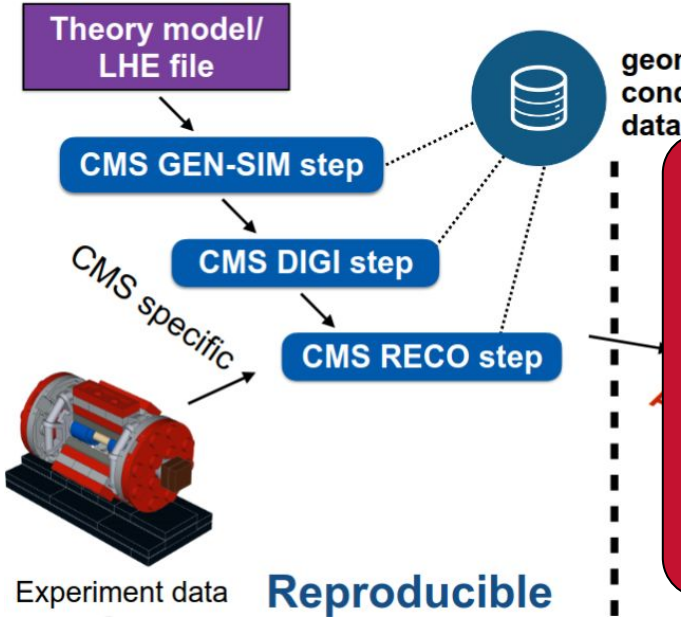
- Production system for simulation and collision data sets
- Analysis review and publication. The individual physics analysis workflows show significant preservation gaps



CMS analysis preservation efforts

CMS has two sets of well-established workflows related to physics analyses:

- **Production system for simulation and collision data sets**
- Analysis review and publication. The individual physics analysis workflows show significant preservation gaps

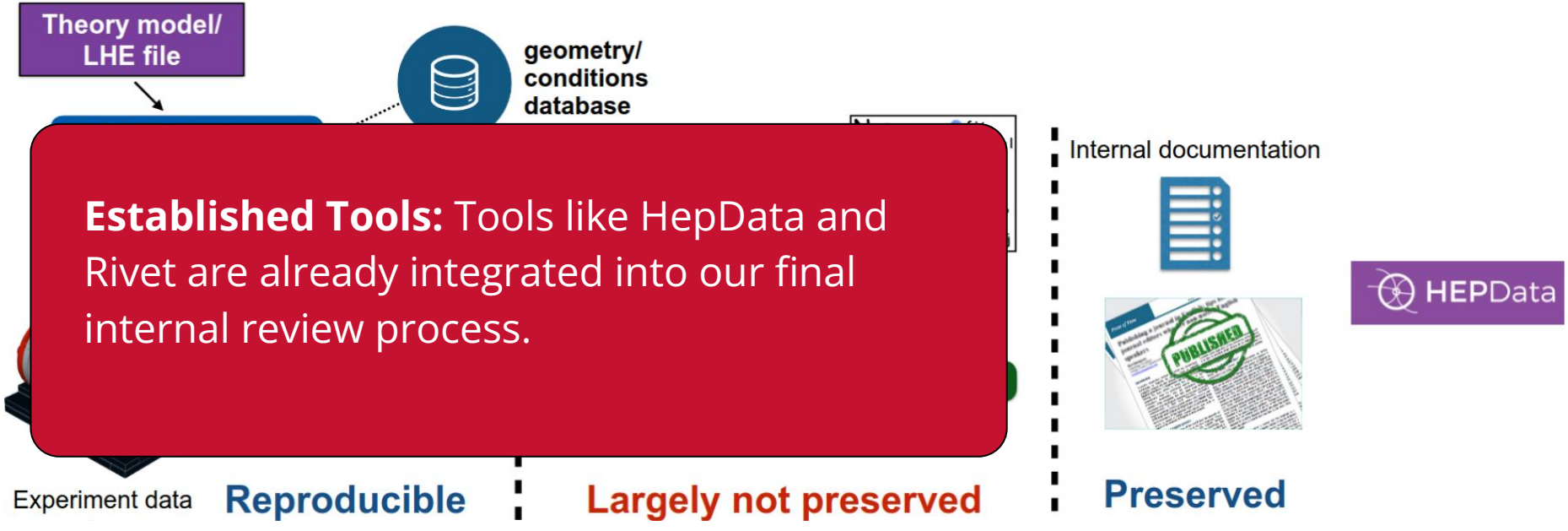


This section is well-preserved and thoroughly documented to ensure internal reproducibility. CMS has made significant efforts to enhance its usefulness for OpenData releases. (Refer to Kati's talk for more details.)

Introduction

CMS has two sets of well-established workflows related to physics analyses:

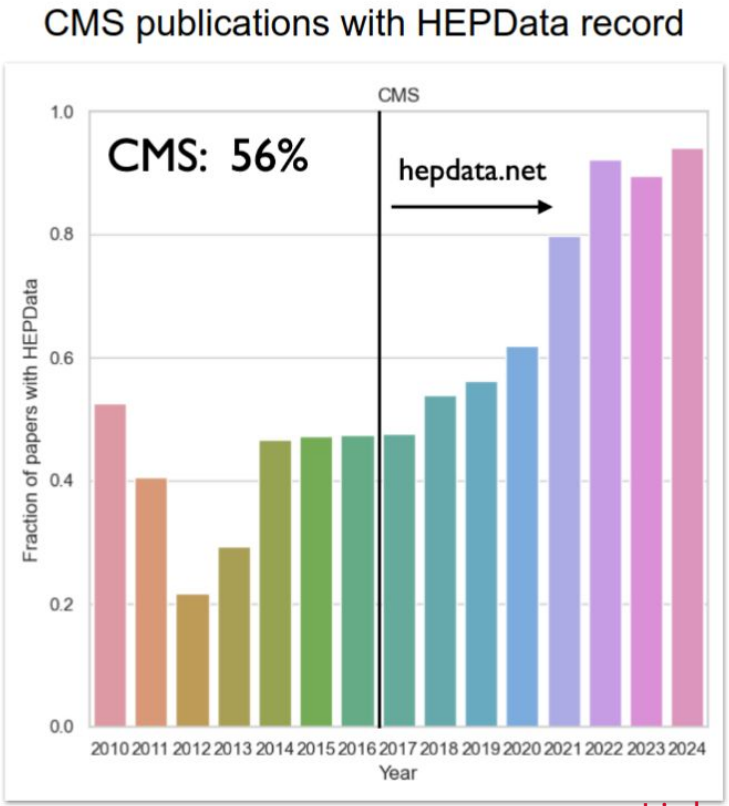
- Production system for simulation and collision data sets
- Analysis review and **publication**. The individual physics analysis workflows show significant preservation gaps



Established Tools: Tools like HepData and Rivet are already integrated into our final internal review process.

HEPData

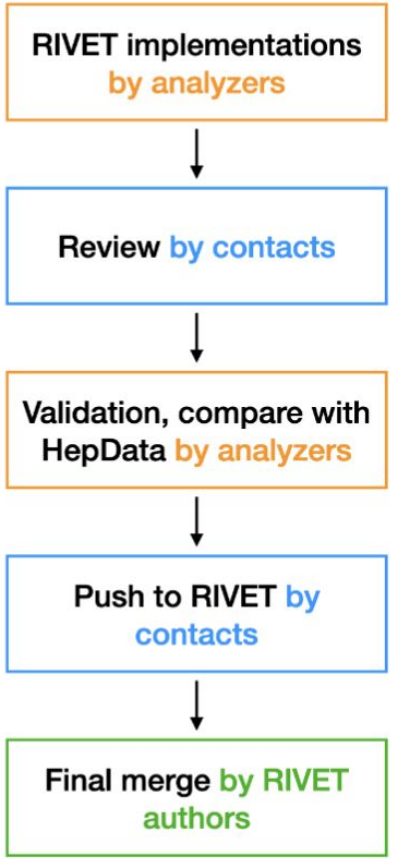
- Since 2021, CMS policy mandates that each analysis creates a HEPData record at the end of the internal review.
- The hepdata_lib tool and dedicated HEPData contacts have facilitated this task.
- This effort is evident in the increasing number of CMS publications with HEPData records in recent years.
- Recently, CMS has begun releasing full statistical models, which are now linked from HEPData.



Link

Alejandro Gomez Espinosa @ Carnegie Mellon University

Rivet plugins



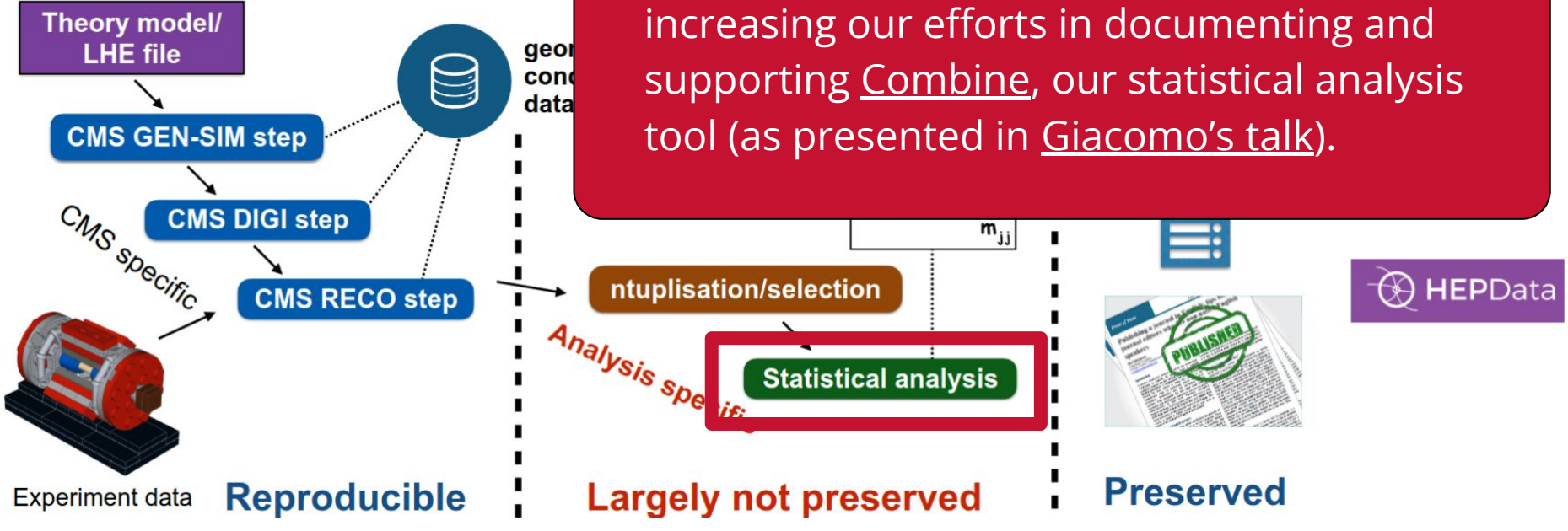
- Rivet implementation is not strictly enforced for CMS analysts, as it is mainly useful for Standard Model measurements.
- Rivet implementation often occurs only after publication.
- Fortunately, based on analysis notes and the paper, a third party can often still implement the Rivet analysis.

CMS analysis preservation efforts

CMS has two sets of well-established workflows related to physics analyses:

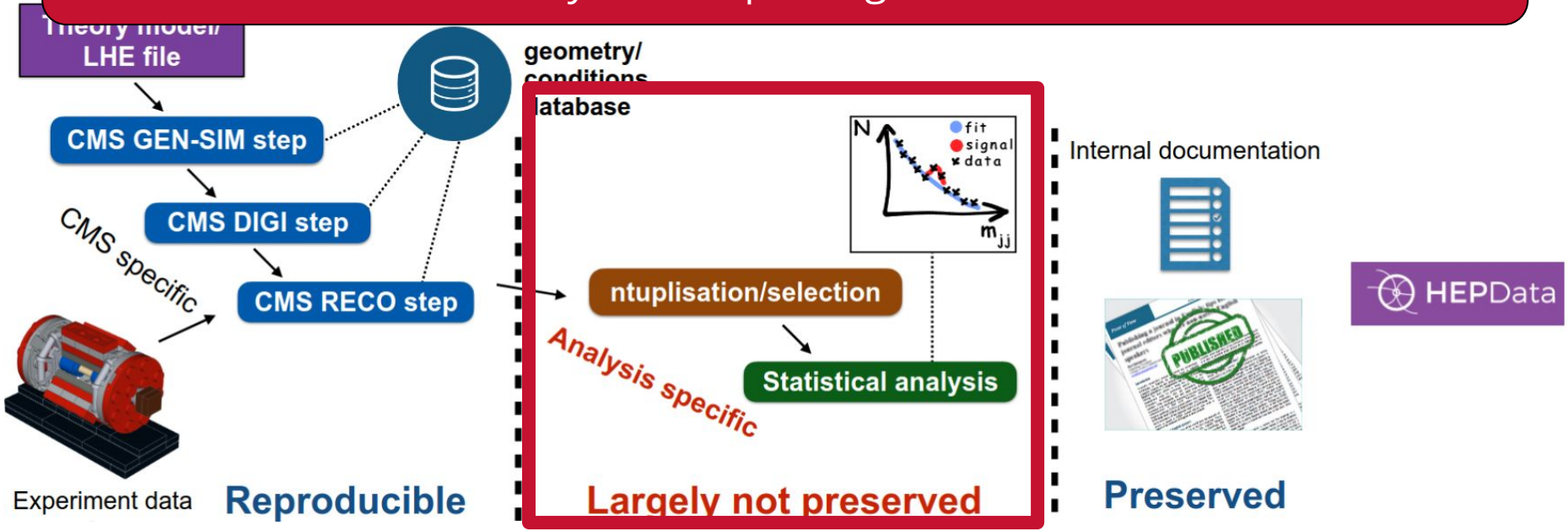
- Production system for simulation and collision data sets
- **Analysis review** and publication. The individual physics analysis workflows show signific

Improved Support for Combine: We're increasing our efforts in documenting and supporting Combine, our statistical analysis tool (as presented in Giacomo's talk).



CMS analysis preservation efforts

Our collaboration is diligently working to close the current preservation gaps in individual physics analyses. The CMS Common Analysis and Tools group (CAT) is dedicated to assisting CMS analysts in improving this effort.



Capturing the actual analysis

1. Capture software

How to capture the code and its dependencies?

2. Capture commands

How to run the captured software?

3. Capture workflow

How to connect the individual analysis steps?

Capturing the actual analysis

1. Capture software

How to capture the code and its dependencies?

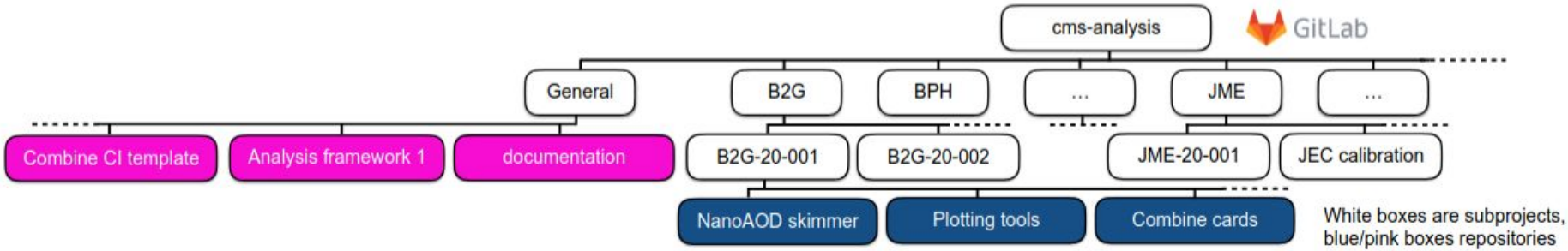
2. Capture commands

How to run the captured software?

3. Capture workflow

How to connect the individual analysis steps?

- We provide a common area for CMS analyses: https://gitlab.cern.ch/cms-analysis
- We also provide tools and services to enable analysis testing and preservation of the code.



Capturing the actual analysis

1. Capture software	2. Capture commands	3. Capture workflow
How to capture the code and its dependencies?	How to run the captured software?	How to connect the individual analysis steps?

- The overall analysis workflow is often *only known to the analysts.*
- We are gradually implementing workflow management tools to capture analysis workflows:
  - LAW (based on Spotify's Luigi) has gained the most traction.
  - Snakemake is becoming more visible as an alternative for managing local workflows and preserving analysis in REANA.
    - CAT has providing support through mattermost channels and tutorials.
- CMS would like to use REANA as our tool to preserve analysis workflows and for reinterpretations. **Progress is being made steadily**.

Further work

CMS is committed to prioritize the following topics:

- **Automated publication of generator information for signal samples**: Ensuring that external researchers can correctly reproduce our signals requires publishing all generator-related model files and cards.
- **Publishing analysis physics algorithms independently of frameworks:** Simply preserving analysis code is not sufficient for colleagues outside CMS to reproduce our analyses. Even within CMS, deciphering code from other analysis teams can be challenging. To address this, we aim to move toward publishing analysis physics algorithms in a framework-independent manner.
- **Systematic preservation of analysis validation information:** This includes cutflows, event selection efficiency maps, object efficiencies, and other key validation details that would enhance the reusability of our results.

Alejandro Gomez Espinosa @ Carnegie Mellon University

Conclusions

- The CMS collaboration is committed to preserving all components of our analyses.
- While there are extensive collaborative efforts to preserve the processing and calibration of our data and simulations, and established mechanisms to preserve final results, some gaps remain that the CMS CAT group is diligently working to address.
- We have tools to preserve individual analysis software and code, and we are actively working on preserving the entire analysis workflow.

Alejandro Gomez Espinosa @ Carnegie Mellon University

Thank you

Alejandro Gomez Espinosa @ Carnegie Mellon University