

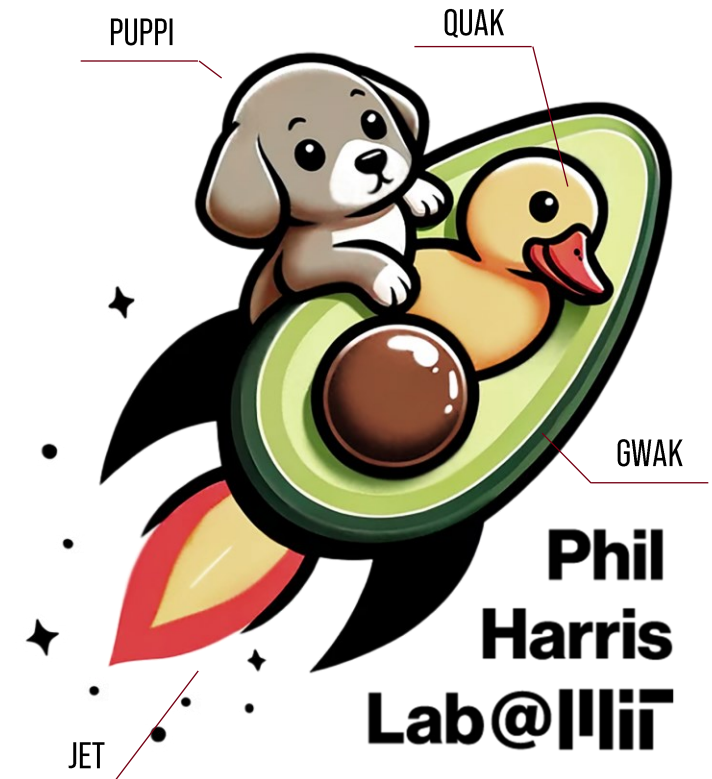
Common Analysis Tools in CMS

... bridging the gap from datasets to publications

Andrzej Novak for the CMS Collaboration

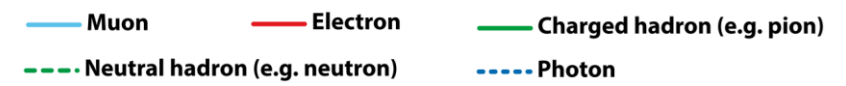
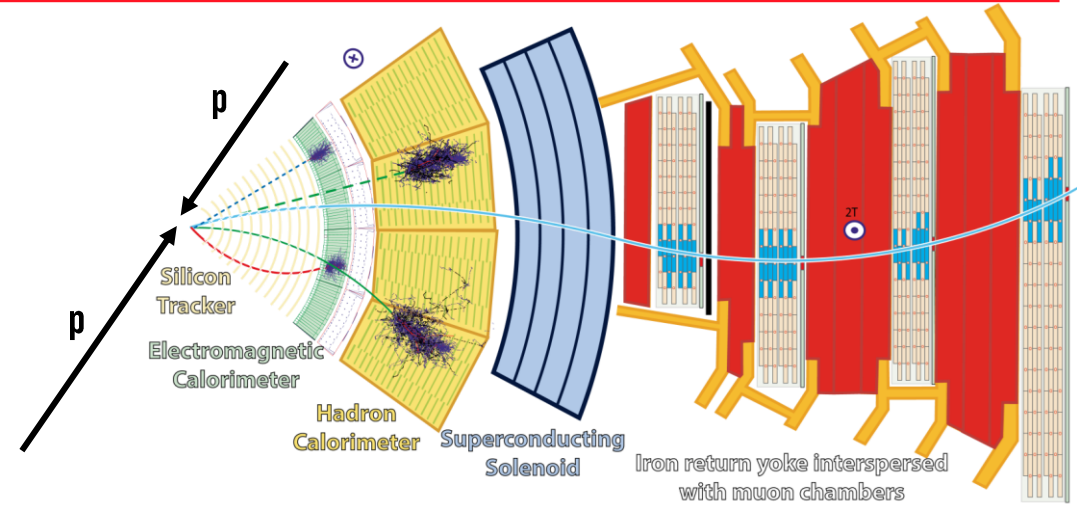
07/19/2024

ICHEP 2024 (Prague)

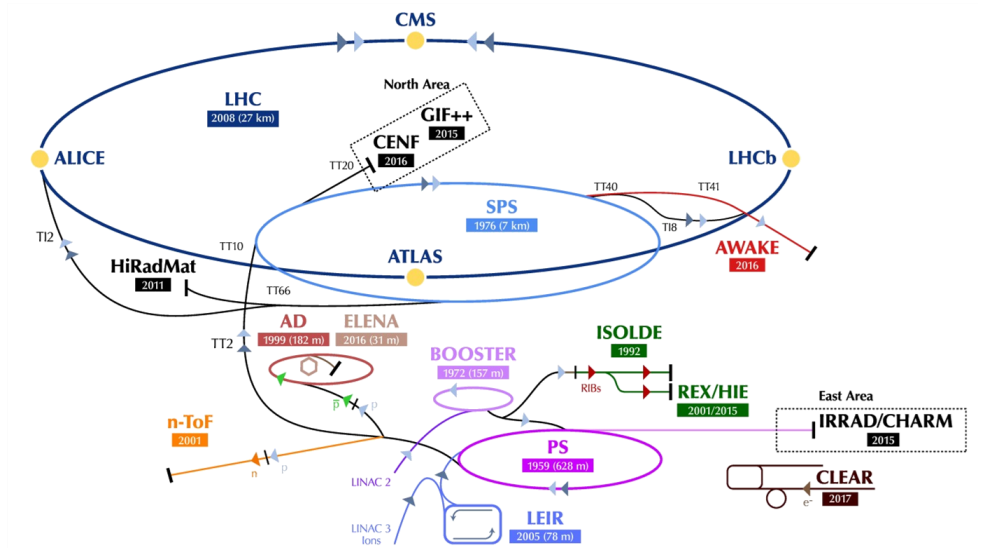
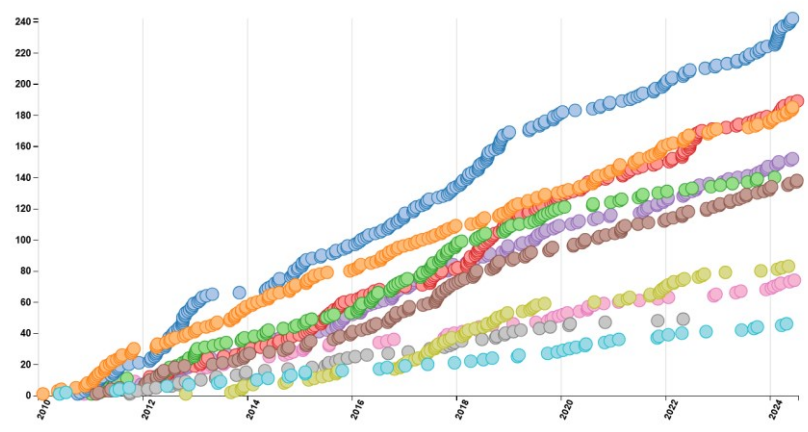


The CMS Experiment

- Compact Muons Solenoid (CMS) Collaboration
 - 4000+ physicists, engineers, computer scientists, technicians, students from ~240 institutes and 50+ countries
 - LHC collision rate at 40 MHz, full detector readout ~ 1MB/evt
 - 100 MB/s of interesting collision events passing trigger (~5 years of data)
 - At any given point nearly 200 analyses are worked on...



Show all Total Exotica Standard Model Supersymmetry Higgs Top Heavy Ions
 B and Quarkonia Forward and Soft QCD Beyond 2 Generations Detector Performance
 1300 collider data papers submitted as of 2024-07-12

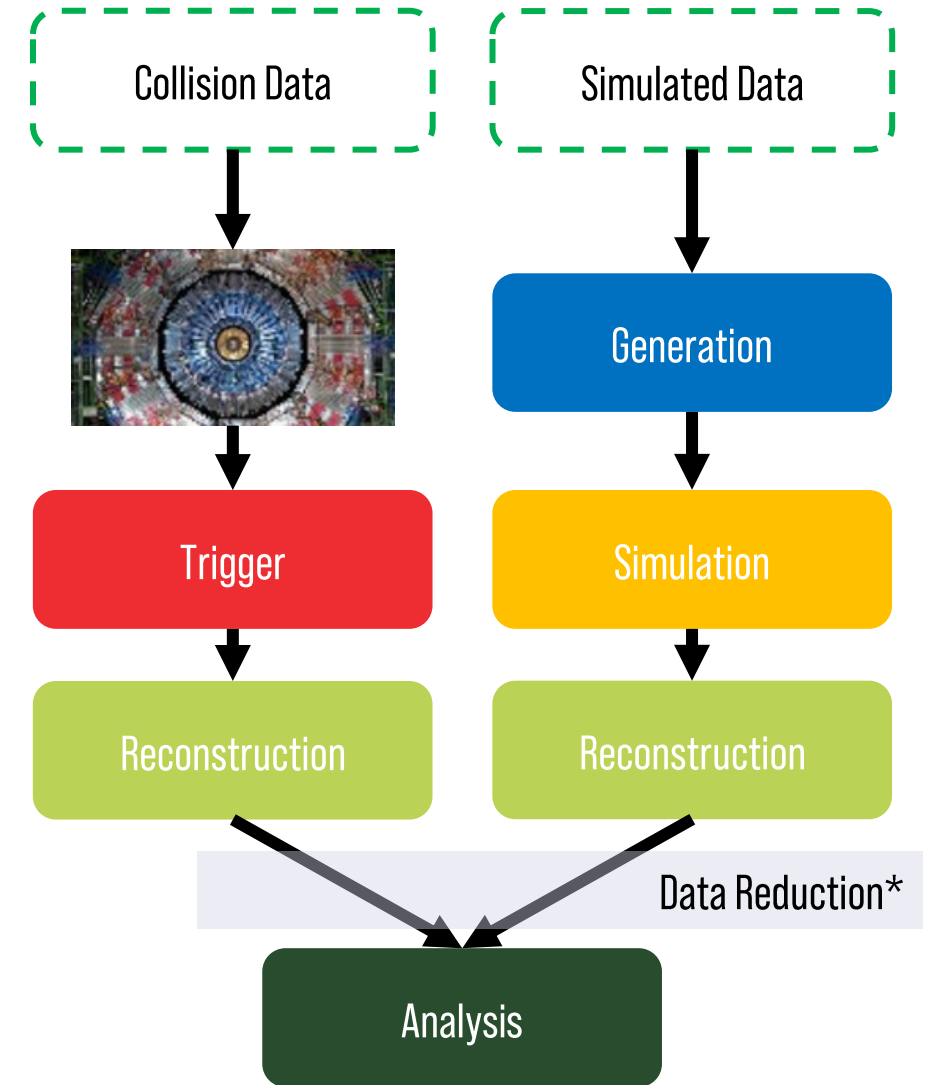


- How do we manage?

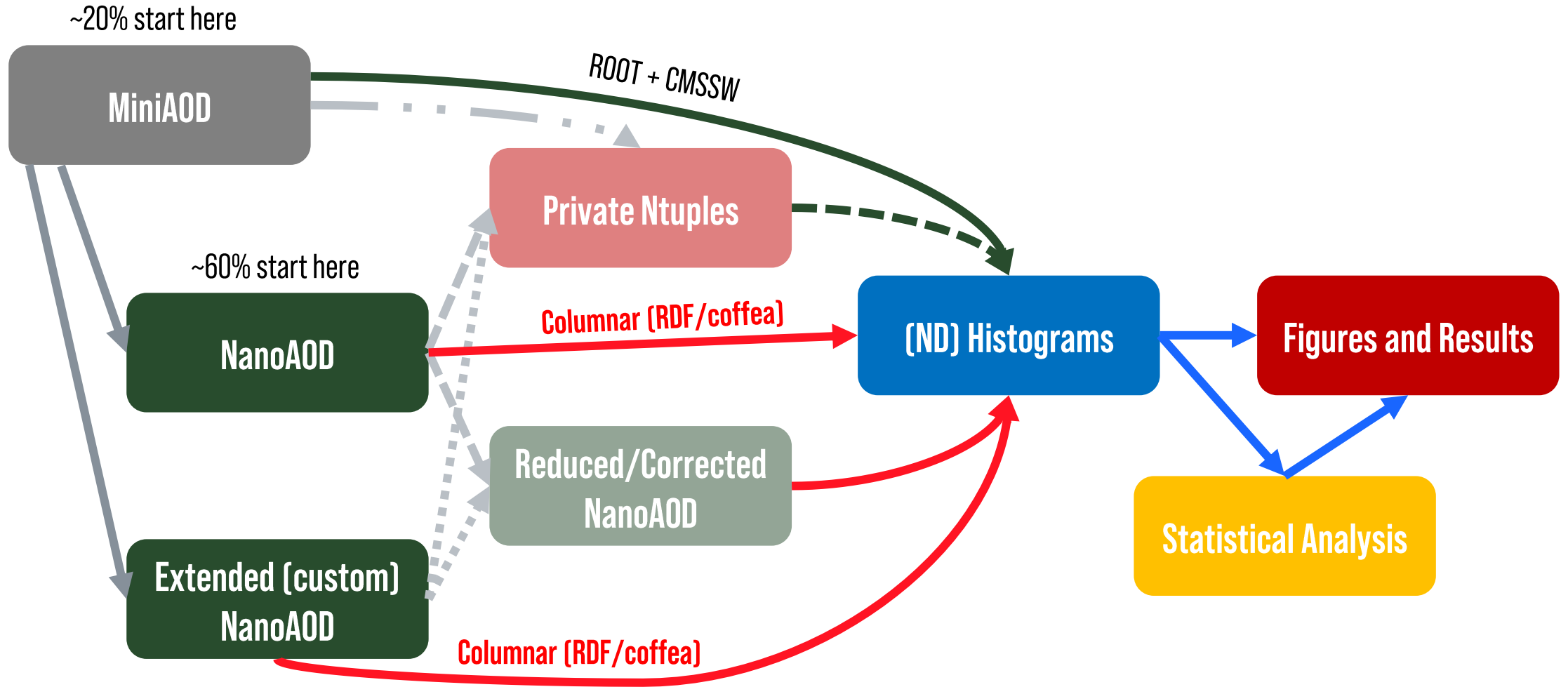
Data Processing & Tiers (Central Processing)

- RAW – detector readout
- AOD - Analysis Object Data (Introduced in 2011, ~2x smaller than the RAW)
- MiniAOD (Introduced in 2013, ~10x smaller than AOD)
- **NanoAOD** (~2018, ~5x smaller than MiniAOD)
 - Primary analysis format, sufficient in ~60% cases
 - “Flat” structure based on simple ROOT TTrees
 - Only basic data types (e.g. float, int, arrays)
 - Only variables related to high-level physical objects like m/p4/id of jet/e/mu

Tier	Event size [MB]	
	200 PU	140 PU
RAW	5.9	4.3
AOD	2	1.4
MiniAOD	0.25	0.18
NanoAOD	0.004	0.004



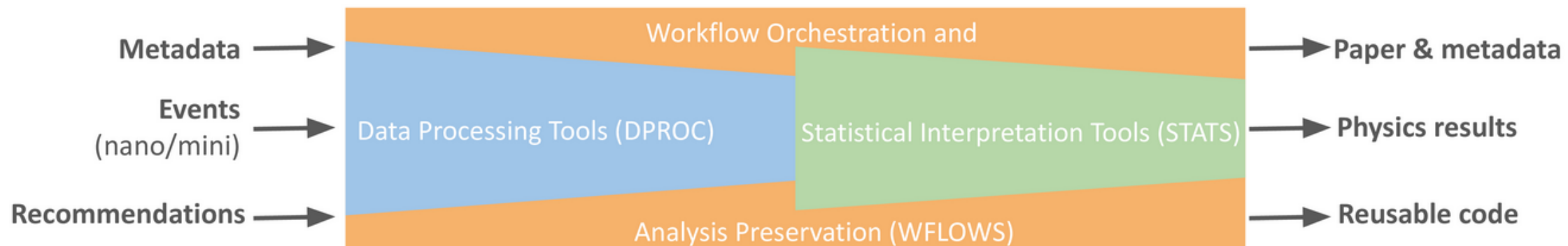
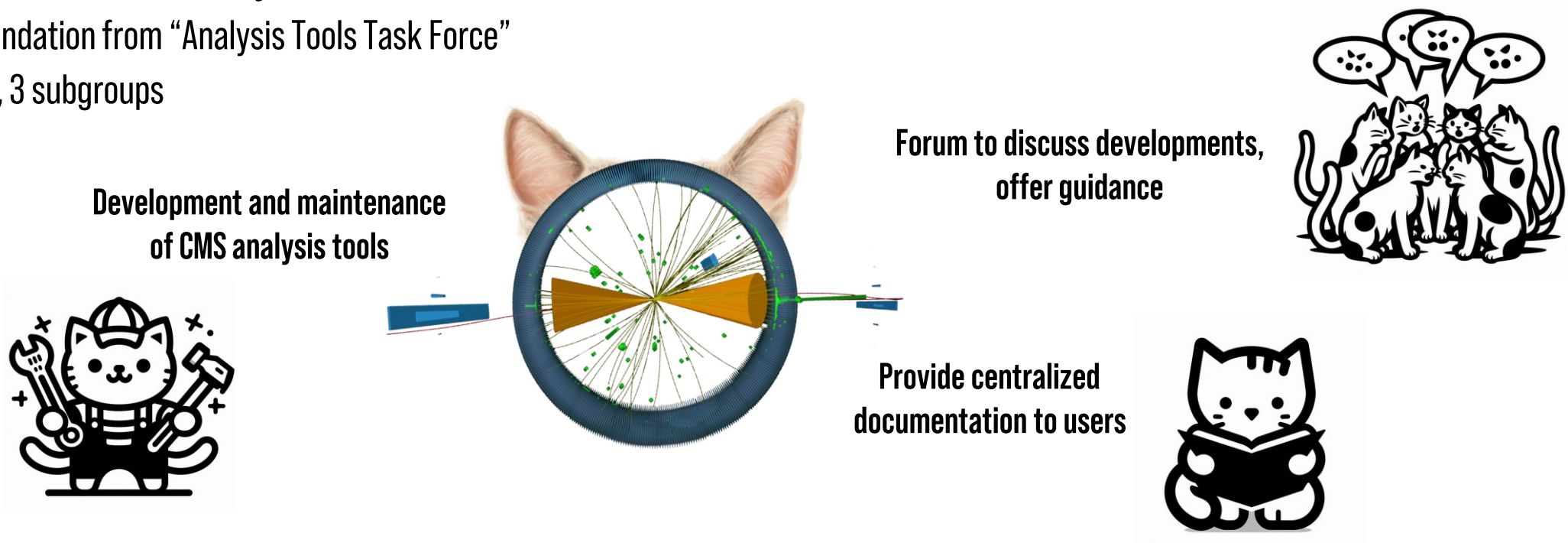
Typical Analysis Schematic



Common Analysis Tools (CAT) Group

Established in 2022 with the Physics Coordination area

- On recommendation from “Analysis Tools Task Force”
- 2 conveners, 3 subgroups



CAT Activities

General meetings every two weeks

- News about recent developments and contributions

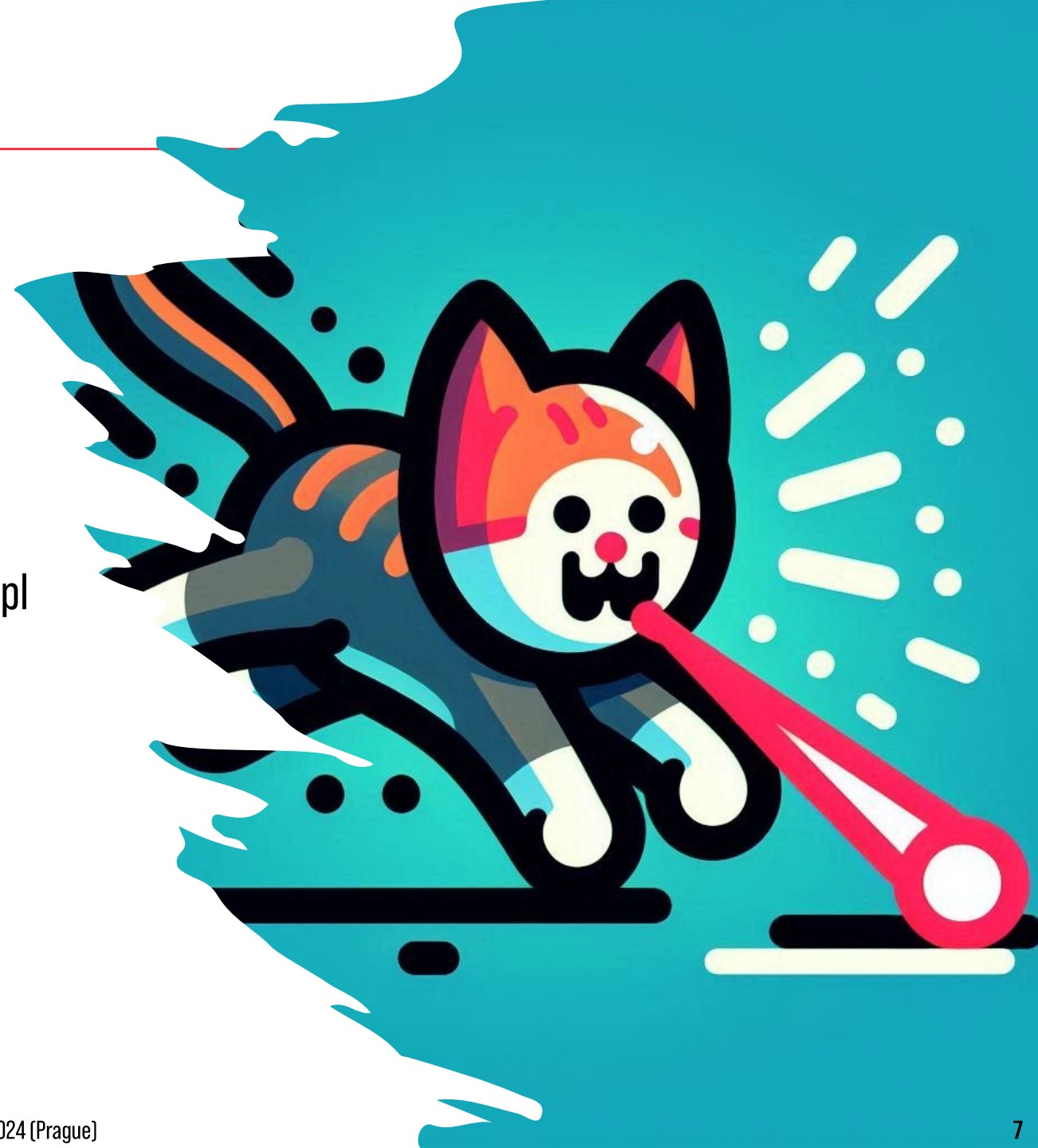
CMS-talk (a customized version of [Discourse](#))

- Forum for users to ask questions and raise issues

CAT documentation website (for now internal only)

CAT HaCATHons (both hacking and training events) ~30/40 ppl

- 1st HaCATHon - Apr 3-6 2023 (CERN)
- 2nd HaCATHon - Sep 25-29 2023 (CERN)
- 3rd HaCATHon - Feb 19-23 2024 (GGI - Florence)
- 4th HaCATHon - Jun 17-28 2023 (CERN/Remote/Async)



CAT Activities

General meetings every two weeks

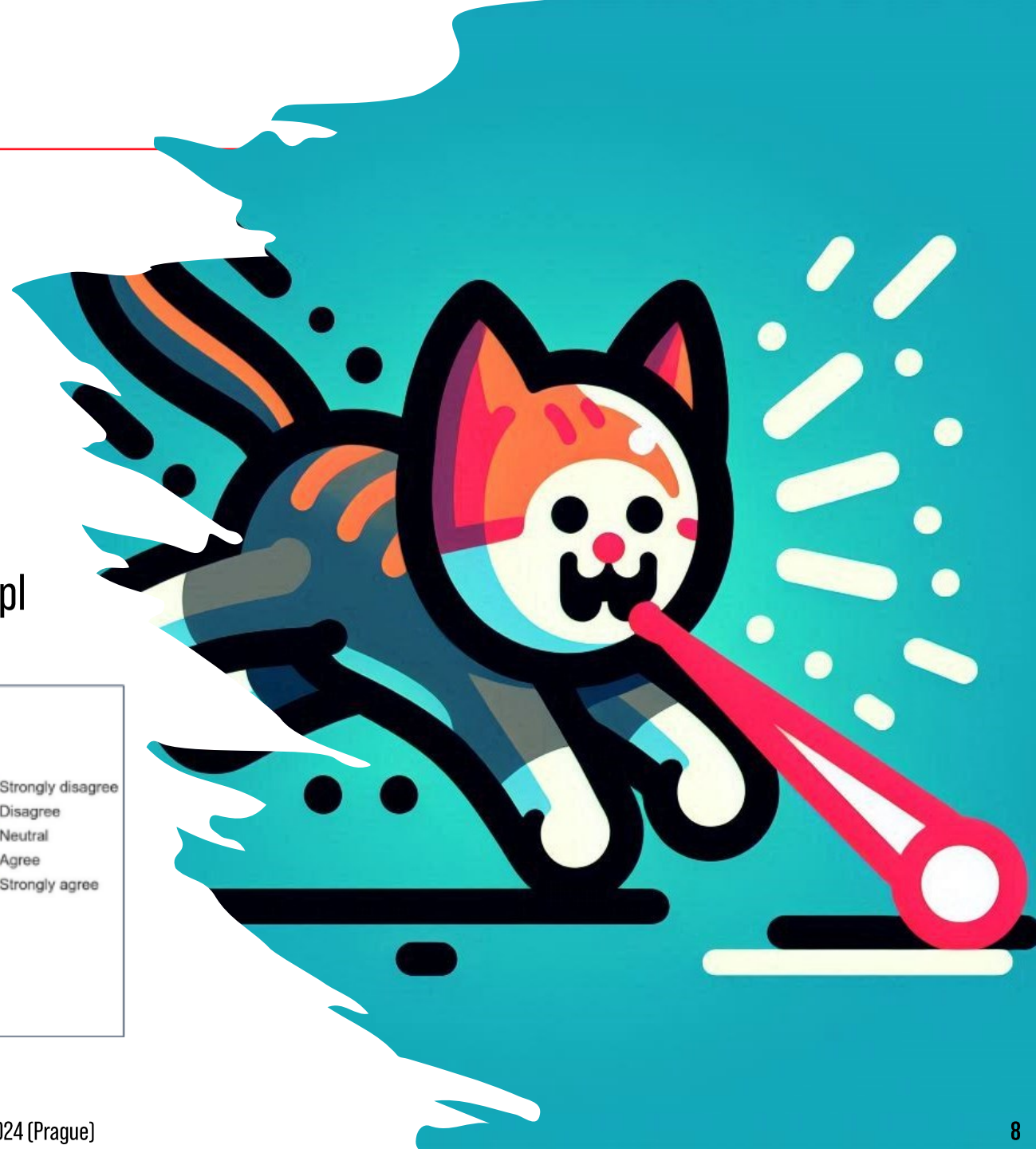
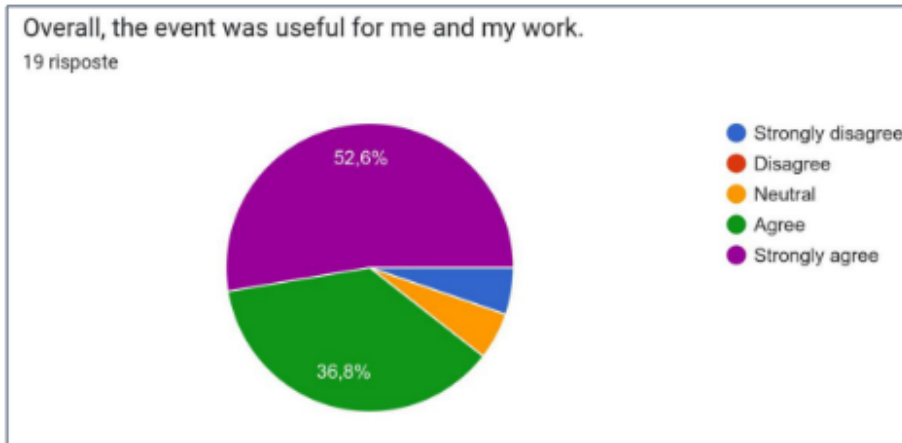
- News about recent developments and contributions

CMS-talk (a customized version of [Discourse](#))

- Forum for users to ask questions and raise issues

CAT documentation website (for now internal only)

CAT HaCAThons (both hacking and training events) ~30/40 ppl



CAT Documentation

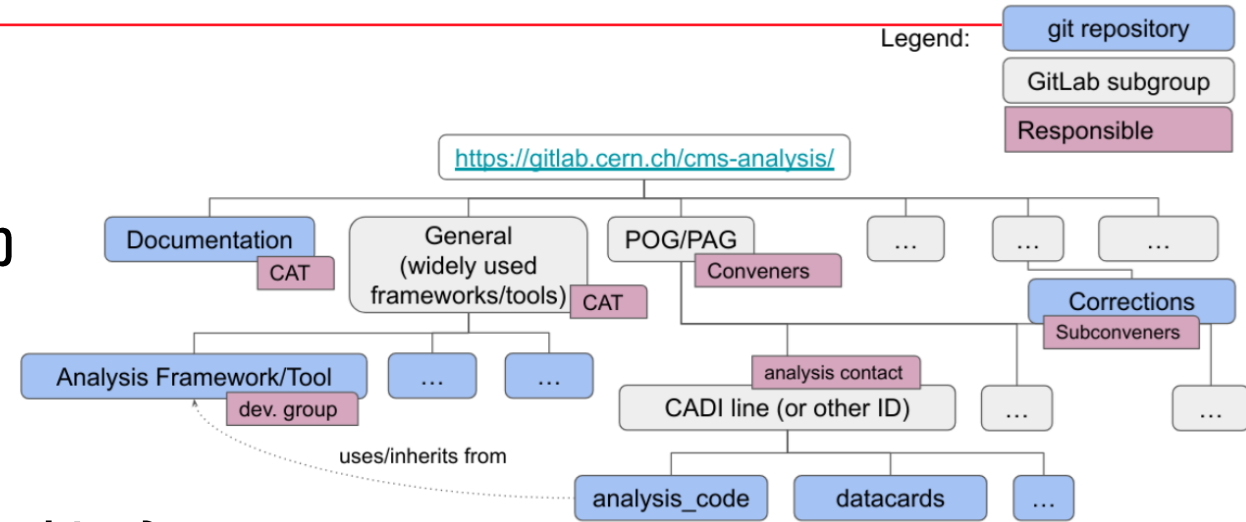
Maintaining a documentation page that should take a junior analyst from NanoAOD to limits...

- Or at least link to all other relevant resources...
- Recommendations for **CMS NanoAOD analysis**
 - Instructions on how to setup code areas
- Overview of supported tools
 - **Data processing**
 - **Workflow management**
 - **Statistical analysis**
 - Miscellaneous snippets
- **Plotting guidelines**
- Collection of links available **Analysis Facilities**
- Links to useful **tutorials** and communication channels

The screenshot shows the homepage of the CMS Common Analysis Tools Documentation. The page has a purple header with the title 'CMS Common Analysis Tools Documentation' and a search bar. Below the header is a navigation menu with links for 'Home', 'Analysis 101', 'General Recommendations', 'Analysis Code Area', and 'CAT structure and contributing'. The main content area features a large heading 'CMS Common Analysis Tools Documentation' and a welcome message: 'Welcome to the CMS Common Analysis Tools (CAT) group documentation pages!'. It then states 'The documentation is split as follows:' and lists four categories: 'General information (this page)', 'General Recommendations', 'Analysis Code Area', and 'Group structure and how to contribute'. Below this, it says 'Use the links at the top to navigate between the groups.' There is also an 'Important Links' section with a sub-heading 'CAT on CMS Talk' and three links: 'Main category - Announcements, getting help', 'Analysis support - Analysis specific questions', and 'Statistical interpretation tools - Statistical inference, combine'. On the right side of the page, there are icons for 'Table of contents' and 'Important Links'.

Analysis Code Areas

- Version control is crucial to **reproducibility**
- CAT hosts unified code areas for analyses on GitLab
 - Full analysis code to be accessible in the repo
 - Either developed directly there or **mirrored** from private GitHubs
- Already well established for statistical analysis (combine)
- Ideally, analysis code just a configuration layer on top of existing framework code
- CI use is encouraged (tutorials/demos)
 - Some analyses fully ran in CI already



CMS analysis repository

Recent activity Last 30 days: Merge requests created 26, Issues created 6, Members added 126

Subgroups and projects	Shared projects	Archived projects	Q Search	Name	1s
> A AnalysisExamples					0 3 1
> B B2G					25 0 3
B BPH					0 0 2
B BTV					0 0 2
E EGM					0 0 2
E EXO					87 0 3

CAT Supported Tools

CAT arranges support/maintenance for CMS-specific tools

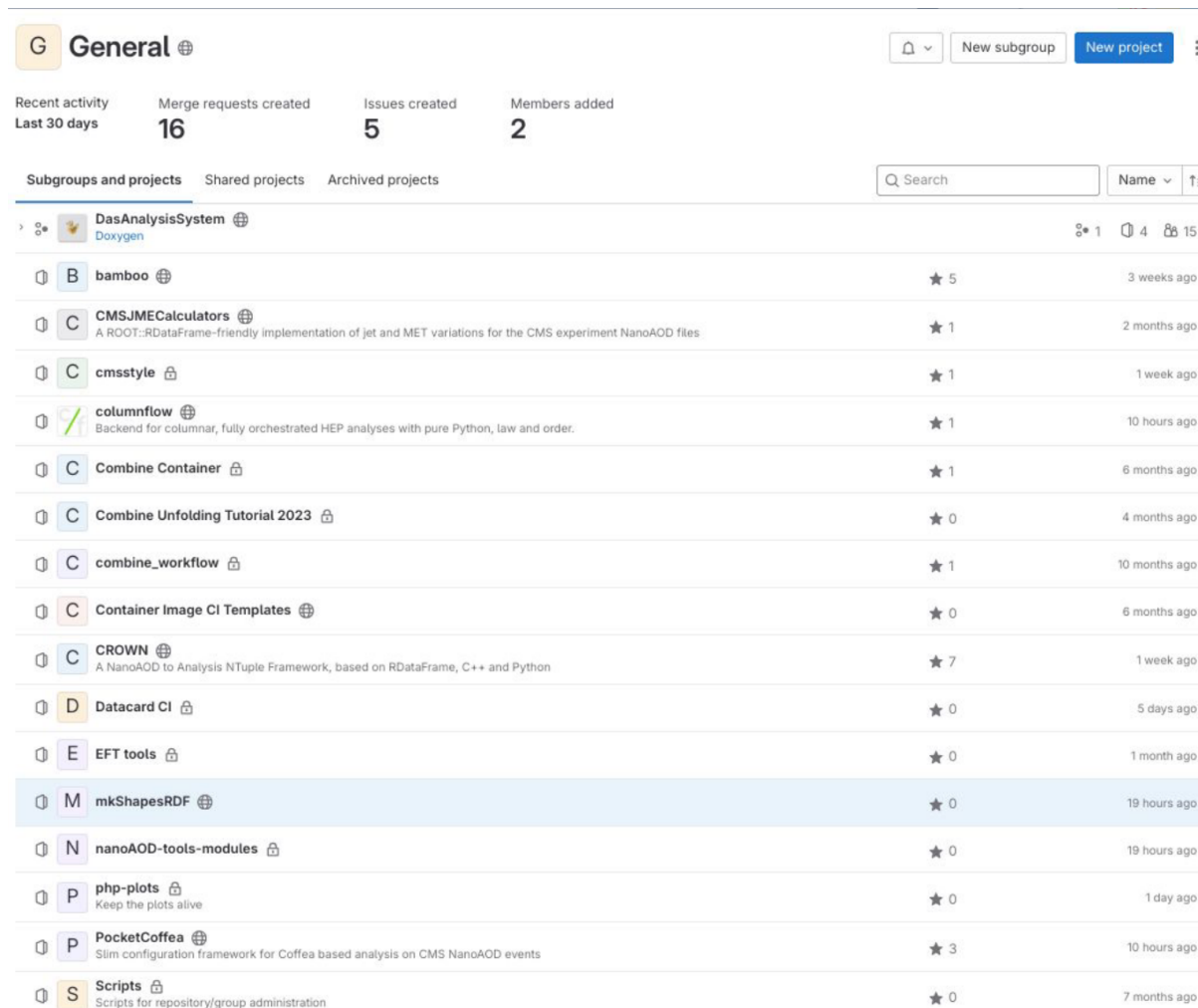
- Mostly coffea or RDF based
- Targeting NanoAOD format

nanoAOD-tools: legacy pyROOT-based sequential framework to skim/extend nanoAODs, and produce plots (modules [here](#))

bamboo: RDF-based python framework that allows to express analysis in a functional style

CMSJMECalculators: RDF-friendly implementation of the recipes for jet and MET variations for CMS

CROWN: RDF-based (C++ and python) framework to generate analysis ntuples (and friends)



The screenshot shows the GitHub organization page for 'General'. It features a header with 'G General' and buttons for 'New subgroup' and 'New project'. Below the header, there are statistics for 'Recent activity Last 30 days': 16 Merge requests created, 5 Issues created, and 2 Members added. A search bar and a 'Name' dropdown are also present. The main content is a list of subgroups and projects, each with a letter icon, name, description, star count, and last activity date.

Letter	Name	Description	Stars	Last Activity
	DasAnalysisSystem	Doxygen	1	4
B	bamboo		5	3 weeks ago
C	CMSJMECalculators	A ROOT::RDataFrame-friendly implementation of jet and MET variations for the CMS experiment NanoAOD files.	1	2 months ago
C	cmsstyle		1	1 week ago
	columnflow	Backend for columnar, fully orchestrated HEP analyses with pure Python, law and order.	1	10 hours ago
C	Combine Container		1	6 months ago
C	Combine Unfolding Tutorial 2023		0	4 months ago
C	combine_workflow		1	10 months ago
C	Container Image CI Templates		0	6 months ago
C	CROWN	A NanoAOD to Analysis NTuple Framework, based on RDataFrame, C++ and Python	7	1 week ago
D	Datacard CI		0	5 days ago
E	EFT tools		0	1 month ago
M	mkShapesRDF		0	19 hours ago
N	nanoAOD-tools-modules		0	19 hours ago
P	php-plots	Keep the plots alive	0	1 day ago
P	PocketCoffea	Slim configuration framework for Coffea based analysis on CMS NanoAOD events	3	10 hours ago
S	Scripts	Scripts for repository/group administration	0	7 months ago

CAT Supported Tools

CAT arranges support/maintenance for CMS-specific tools

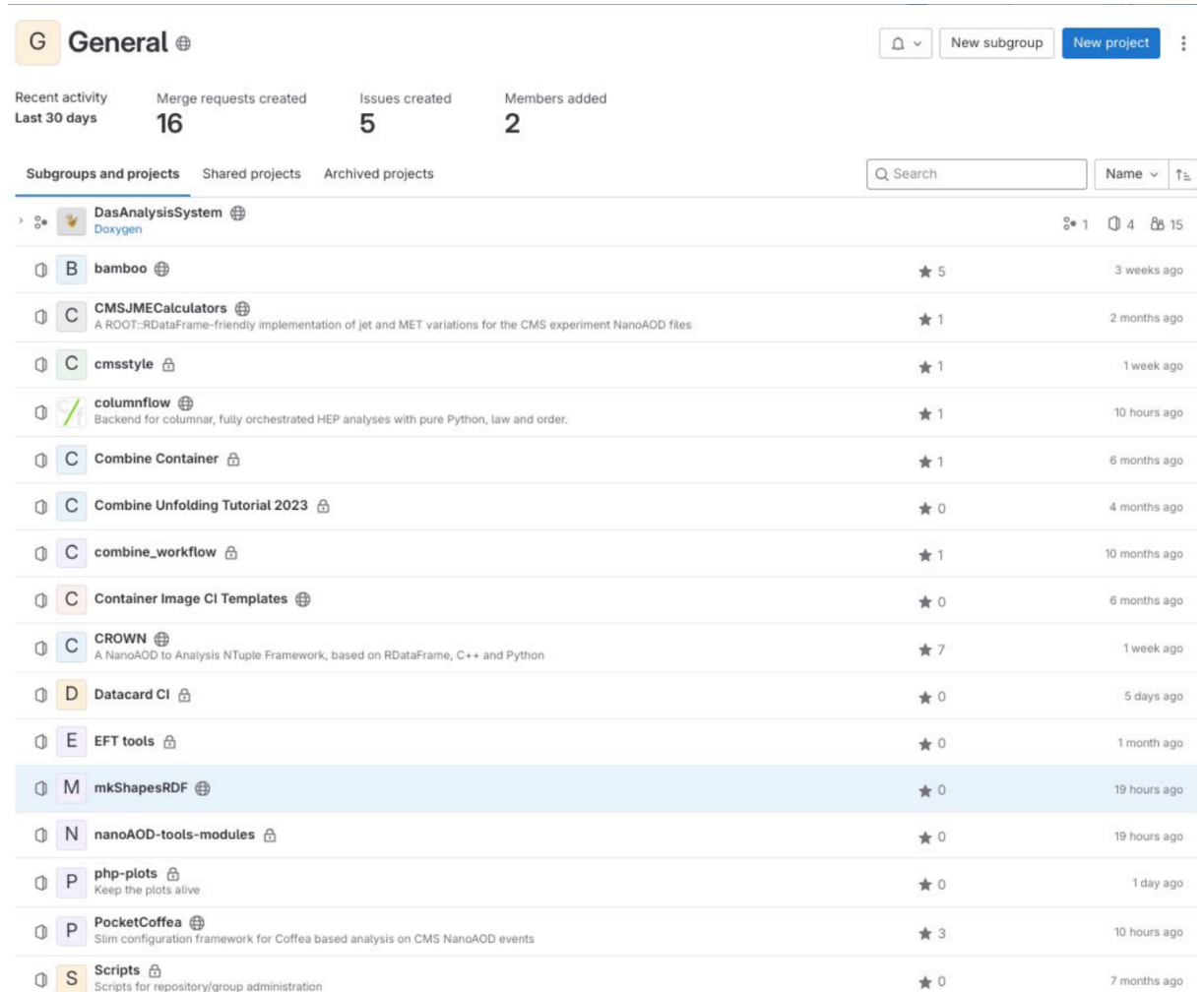
- Mostly coffea or RDF based
- Targeting NanoAOD format

[columnflow](#): python (Awkward Arrays)-based backend for columnar, fully-orchestrated HEP analyses

[DasAnalysisSystem](#): ROOT-based tools for analysis with high-level objects

[PocketCoffea](#): configuration framework for Coffea-based analyses on NanoAODs

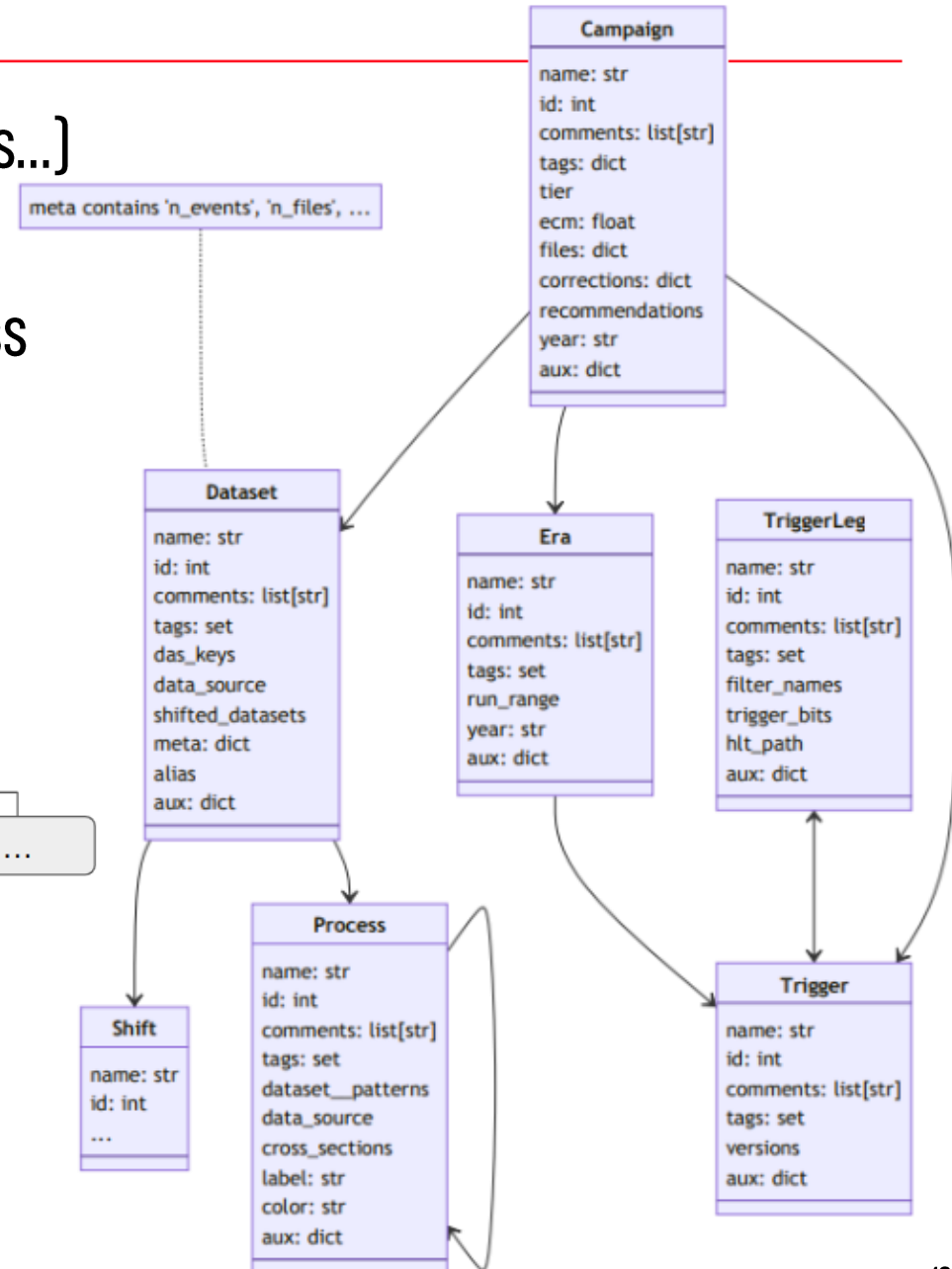
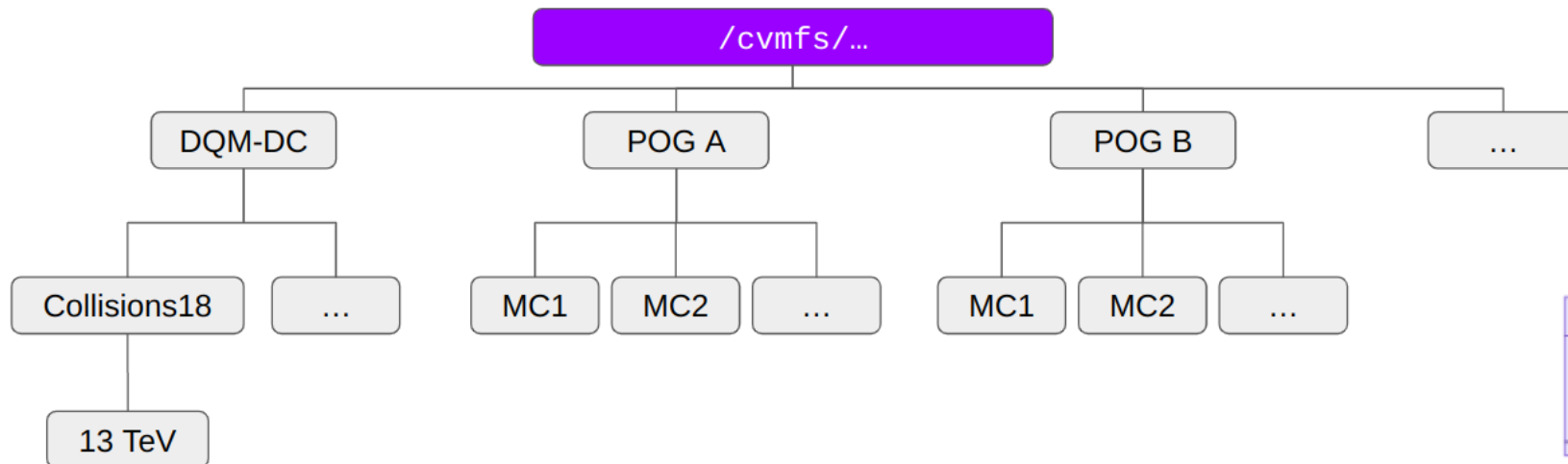
[mkShapesRDF](#): RDF-based framework for analyses on NanoAODs, which are implemented through config files



Subgroups and projects	Shared projects	Archived projects	Search	Name	Sort
G General 🌐 🔔 New subgroup New project ⋮					
Recent activity	Merge requests created	Issues created	Members added		
Last 30 days	16	5	2		
Subgroups and projects Shared projects Archived projects <input type="text" value="Search"/> Name Sort					
DasAnalysisSystem 🌐 Doxygen 🔔 🌐 1 4 15					
B	bamboo 🌐	★ 5	3 weeks ago		
C	CMSJMECalculators 🌐 A ROOT::RDataFrame-friendly implementation of jet and MET variations for the CMS experiment NanoAOD files.	★ 1	2 months ago		
C	cmsstyle 🔒	★ 1	1 week ago		
C	columnflow 🌐 Backend for columnar, fully orchestrated HEP analyses with pure Python, law and order.	★ 1	10 hours ago		
C	Combine Container 🔒	★ 1	6 months ago		
C	Combine Unfolding Tutorial 2023 🔒	★ 0	4 months ago		
C	combine_workflow 🔒	★ 1	10 months ago		
C	Container Image CI Templates 🌐	★ 0	6 months ago		
C	CROWN 🌐 A NanoAOD to Analysis NTuple Framework, based on RDataFrame, C++ and Python	★ 7	1 week ago		
D	Datacard CI 🔒	★ 0	5 days ago		
E	EFT tools 🔒	★ 0	1 month ago		
M	mkShapesRDF 🌐	★ 0	19 hours ago		
N	nanoAOD-tools-modules 🔒	★ 0	19 hours ago		
P	php-plots 🔒 Keep the plots alive	★ 0	1 day ago		
P	PocketCoffea 🌐 Slim configuration framework for Coffea based analysis on CMS NanoAOD events	★ 3	10 hours ago		
S	Scripts 🔒 Scripts for repository/group administration	★ 0	7 months ago		

Metadata Management

- Analyses require non-trivial amount of metadata (random TWikis...)
 - Cross-sections, DQM-validations, calibrations, corrections, systematics...
- Ongoing work to design a metadata schema and a **tool** for access
 - Majority of corrections already aggregated in a **single repo** in **json** format
 - Corrections applied via **correctionlib** (C++ with pybind11 bindings)
- Target – easy to understand versions, distribution via /cvmfs



Workflow Management and Analysis Preservation

- **Workflow management tools** are a way of helping analysis reusability and reproducibility

- Organizing, managing, and scaling job submission
- Publishing results into reusable formats e.g. HEPData, Rivet

- **Orchestration & workflow tools**

- **luigi**: Package for building complex pipelines with dependency resolution, workflow management, and visualization.
- **law**: Extension of luigi with full decoupling of resources on HEP infrastructure
- **airflow**: Platform to programmatically author, schedule and monitor workflows
- **snakemake**: Workflow management system to create reproducible and scalable data analyses

- **Preservation**

- **HEPData portal**: Repository for publication-related High-Energy Physics data
- **Reana**: Reproducible research data analysis platform
- **Rivet**: Toolkit for robust independent validation of experiment and theory
- **MadAnalysis**: Framework for phenomenological investigations at particle colliders
- **CheckMate**: Toolkit for checking models at terascale energies
- **SModelS**: A tool for interpreting simplified-model results from the LHC

- **Complimentary to efforts for establishing CI pipelines for analysis**

- Best practices recommended in CAT docs
- Ongoing work to offload CI jobs to analysis facilities

Plotting Guidelines and Tools

Established uniform style guide for the experiment

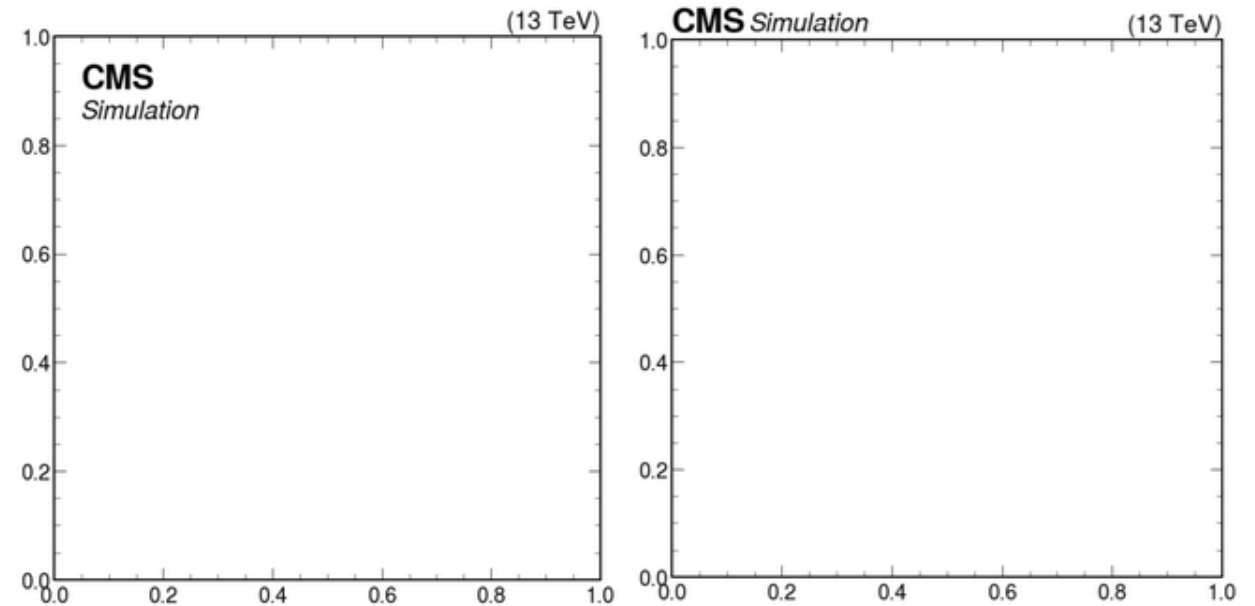
- Maintenance of packages and canned recipes for application both in python ([mplhep](#)) and ROOT ([cmsstyle](#))
- **Plots in both python and ROOT should look the same** (also regardless of OS)

```
In-frame
Python  ROOT
hep.cms.label("Preliminary", loc=2, ax=ax)

Out-of-frame
Python  ROOT
hep.cms.label("Preliminary", loc=0, ax=ax)

In-frame
Python  ROOT
CMS.SetExtraText("Simulation Preliminary")
CMS.SetLumi("")
canv = CMS.cmsCanvas('', 0, 1, 0, 1, '', '', square = CMS.kSquare, extraSpace=0.01,

Out-of-frame
Python  ROOT
CMS.SetExtraText("Simulation Preliminary")
CMS.SetLumi("")
canv = CMS.cmsCanvas('', 0, 1, 0, 1, '', '', square = CMS.kSquare, extraSpace=0.01,
```



- Tex Gyre Heros - open license Helvetica clone
- Now also available in [ROOT](#)

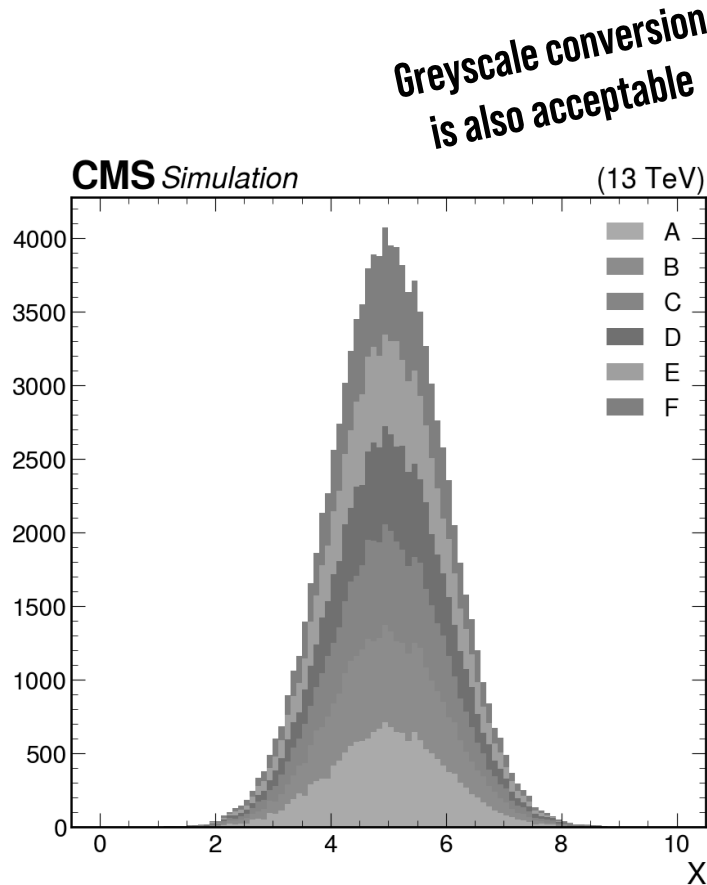
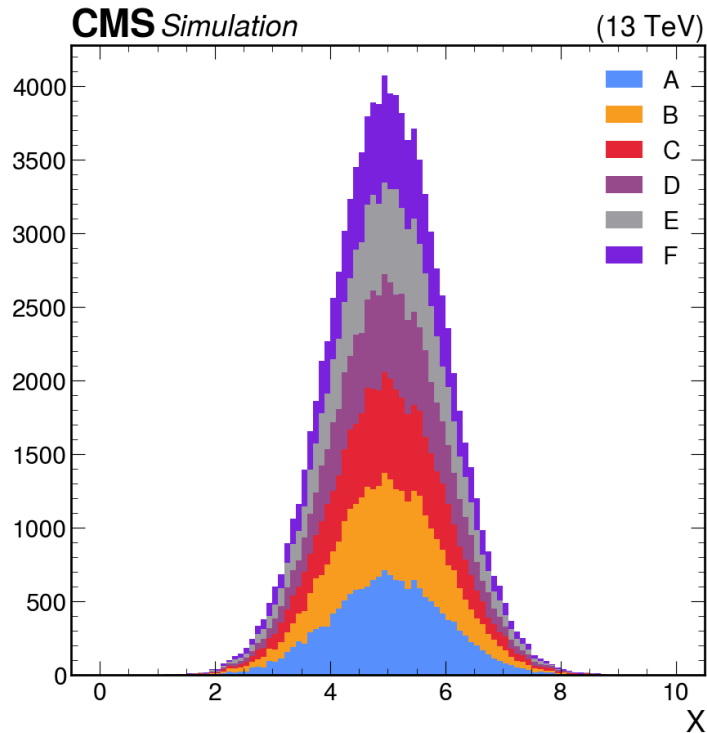
Accessible Color Schemes

Recently standardized CMS *recommended* color scheme

Chosen for **colorblind accessibility** (joint effort with **CMS Diversity Office**)

- From M. Petroff [arxiv:2107.02270](https://arxiv.org/abs/2107.02270) – petroff6 or petroff10

Now also used by ATLAS



Hex Code	Color Preview
#3f90da	
#ffa90e	
#bd1f01	
#94a4a2	
#832db6	
#a96b59	
#e76300	
#b9ac70	
#717581	
#92dadd	

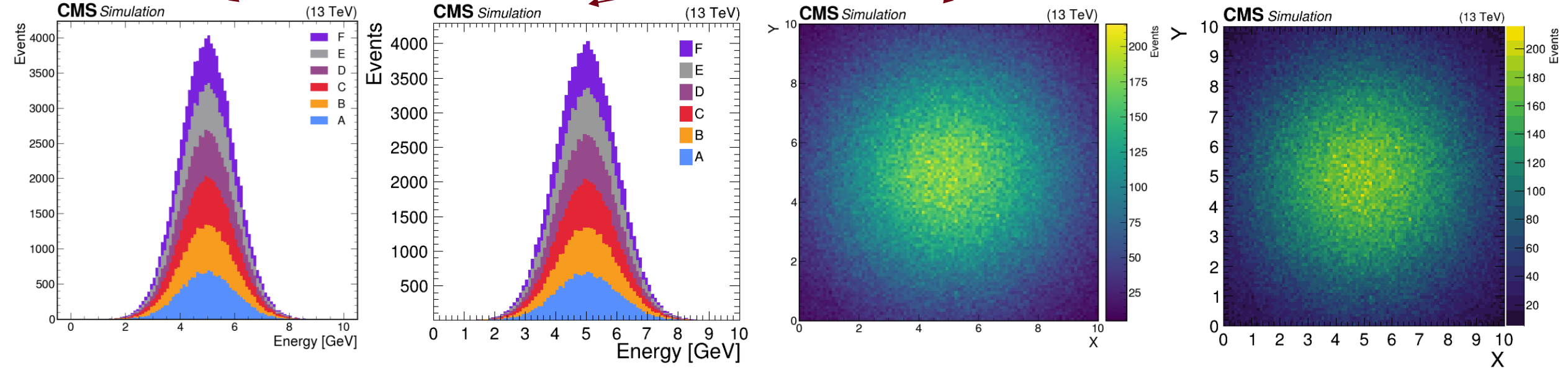
Hex Code	Color Preview
#5790fc	
#f89c20	
#e42536	
#964a8b	
#9c9ca1	
#7a21dd	

Accessible Color Schemes

For sequential colormaps a number of CVD-accessible options exist - using **viridis**

python

ROOT



Consistent look regardless of choice of programming language or OS

Code snippets for users provided in docs

Python ROOT

Combine

Combine is a high-level tool based on RooFit/RooStats

- Provides CLI to various statistical techniques used in CMS
- Statistical models encoded in **human-readable *datacards***

```
imax 1 number of bins
jmax 4 number of processes minus 1
kmax * number of nuisance parameters
-----
bin          signal_region
observation  10.0
-----
bin          signal_region  signal_region  signal_region  signal_region  signal_reg:
process      ttbar          diboson      Ztautau        jetFakes        bbHtautau
process      1              2              3              4              0
rate         4.43803       3.18309       3.7804         1.63396         0.711064
-----
CMS_eff_b    lnN  1.02         1.02         1.02         -              1.02
CMS_eff_t    lnN  1.12         1.12         1.12         -              1.12
CMS_eff_t_highpt lnN  1.1         1.1         1.1         -              1.1
acceptance_Ztautau lnN  -           -           1.08         -              -
acceptance_bbH lnN  -           -           -           -              1.05
acceptance_ttbar lnN  1.005       -           -           -              -
norm_jetFakes lnN  -           -           -           1.2           -
xsec_diboson lnN  -           1.05        -           -              -
```

De facto official CMS tool for statistical analysis

- Support, maintenance, and development organized in CAT
- Recently submitted to **CSBS**



CMS-CAT-23-001



CERN-EP-2024-078
2024/04/11

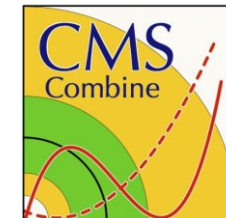
The CMS statistical analysis and combination tool:
COMBINE

The CMS Collaboration*

Abstract

This paper describes the COMBINE software package used for statistical analyses by the CMS Collaboration. The package, originally designed to perform searches for a Higgs boson and the combined analysis of those searches, has evolved to become the statistical analysis tool presently used in the majority of measurements and searches performed by the CMS Collaboration. It is not specific to the CMS experiment, and this paper is intended to serve as a reference for users outside of the CMS Collaboration, providing an outline of the most salient features and capabilities. Readers are provided with the possibility to run COMBINE and reproduce examples provided in this paper using a publicly available container image. Since the package is constantly evolving to meet the demands of ever-increasing data sets and analysis sophistication, this paper cannot cover all details of COMBINE. However, the online documentation referenced within this paper provides an up-to-date and complete user guide.

Submitted to Computing and Software for Big Science



arXiv:2404.06614v1 [physics.data-an] 9 Apr 2024

Likelihood Publishing and HS3

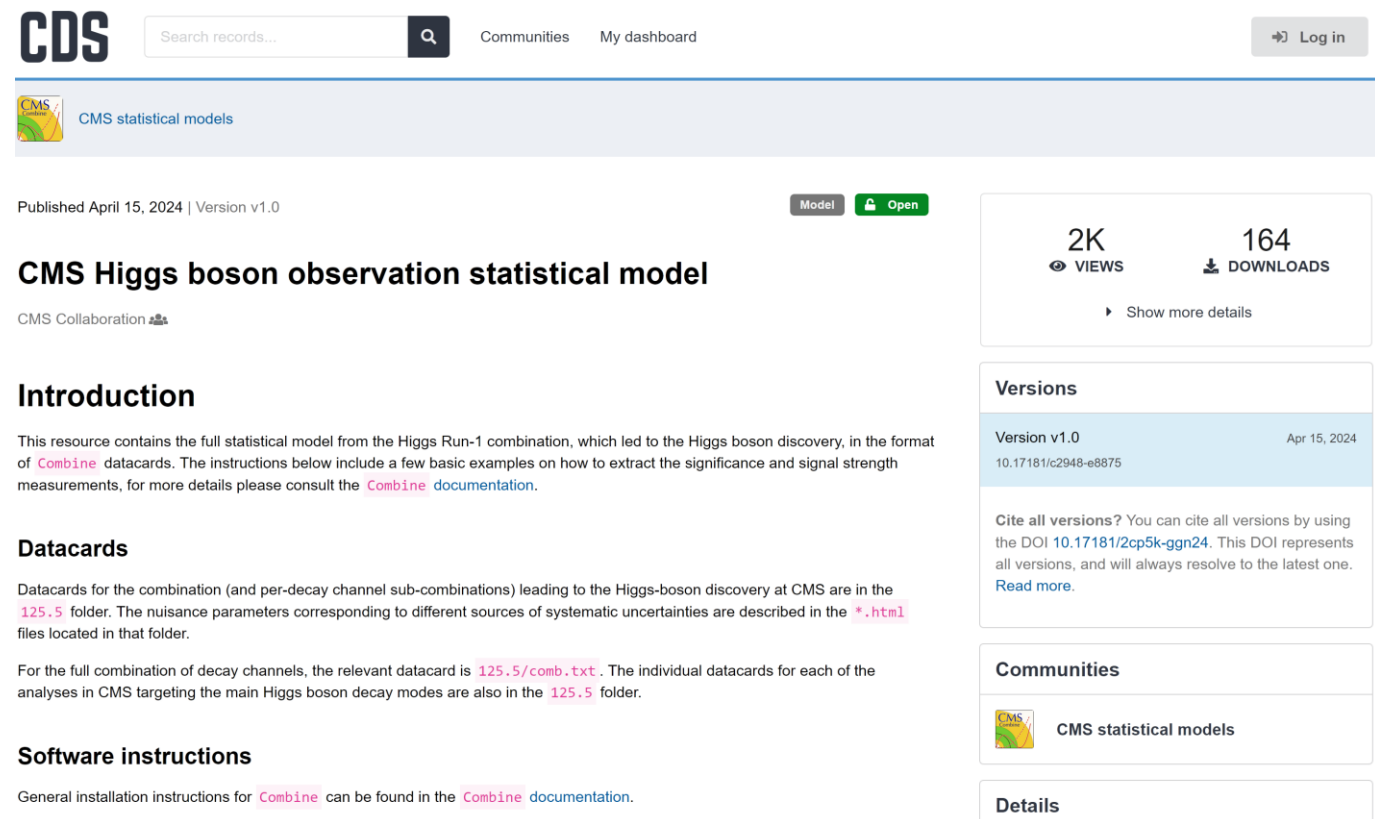
CMS will now publish full likelihood models (docker + combine + cards -> reproducible stats analysis)

- Linked to **HEPData** records and released under CC4 license
- First example already available for [CMS Higgs boson observation](#)

More details in [Sezen's talk on Saturday 14:30](#)

HEP Statistics Serialization Standard (HS3)

- Ongoing discussion on a new standard for likelihood publishing in HEP independent of particular software across experiments
- <https://indico.cern.ch/event/1348309/>



The screenshot shows the CMS statistical models page for the "CMS Higgs boson observation statistical model". The page is published on April 15, 2024, and is version v1.0. It features a search bar, a "Communities" link, and a "My dashboard" link. The main content area includes an introduction, datacards, and software instructions. The right sidebar shows 2K views and 164 downloads, along with a "Versions" section listing the current version v1.0 and a "Communities" section listing the "CMS statistical models" community.

CDS Search records... Communities My dashboard Log in

CMS statistical models

Published April 15, 2024 | Version v1.0 Model Open

CMS Higgs boson observation statistical model

CMS Collaboration

Introduction

This resource contains the full statistical model from the Higgs Run-1 combination, which led to the Higgs boson discovery, in the format of **Combine** datacards. The instructions below include a few basic examples on how to extract the significance and signal strength measurements, for more details please consult the **Combine** documentation.

Datacards

Datacards for the combination (and per-decay channel sub-combinations) leading to the Higgs-boson discovery at CMS are in the **125.5** folder. The nuisance parameters corresponding to different sources of systematic uncertainties are described in the ***.html** files located in that folder.

For the full combination of decay channels, the relevant datacard is **125.5/comb.txt**. The individual datacards for each of the analyses in CMS targeting the main Higgs boson decay modes are also in the **125.5** folder.

Software instructions

General installation instructions for **Combine** can be found in the **Combine** documentation.

2K VIEWS 164 DOWNLOADS Show more details

Versions

Version v1.0	Apr 15, 2024
10.17181/c2948-e8875	

Cite all versions? You can cite all versions by using the DOI [10.17181/2cp5k-ggn24](https://doi.org/10.17181/2cp5k-ggn24). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Communities

CMS statistical models

Details

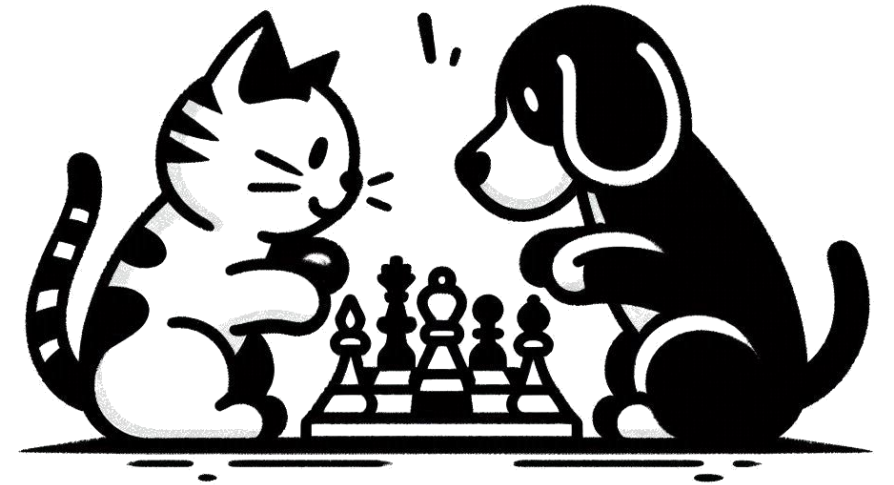
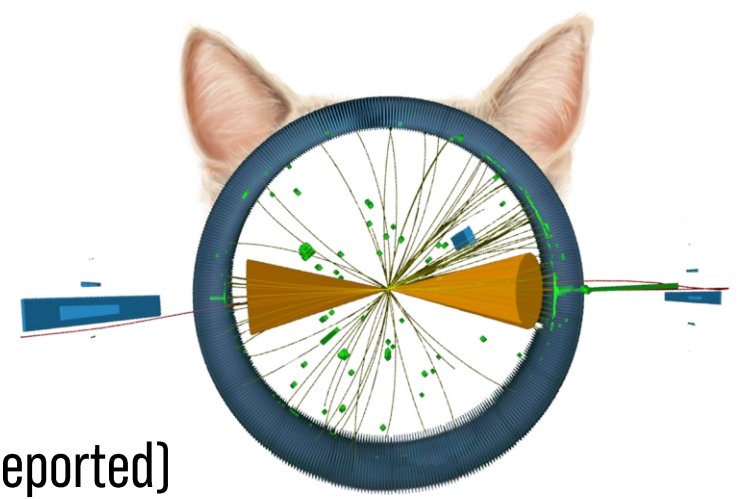
Summary of Progress and Results

- Substantial progress since CAT group was established 2 years ago
- Established **docs** that should take a junior analyst **from NanoAOD to statistical analysis**
- Coordinate **maintenance** for key internal tools
 - Serve as a channel for feedback for outside packages such as **scikit-hep/coffea/ROOT**
- **Standardized figure style** and provided code to achieve it
- Introduced a recommended **colorblind accessible color scheme**
- Publishing a reference for **Combine** and established process to **release full likelihoods**

Ongoing Efforts and How to Engage

- Metadata scheme standardization and distribution
- Further improvements of automation
 - Ongoing efforts to include CI in analysis development (first positive experiences reported)
- Standardize a likelihood serialization format across HEP experiments
- Better integration with various analysis facilities (like coffea.casa)

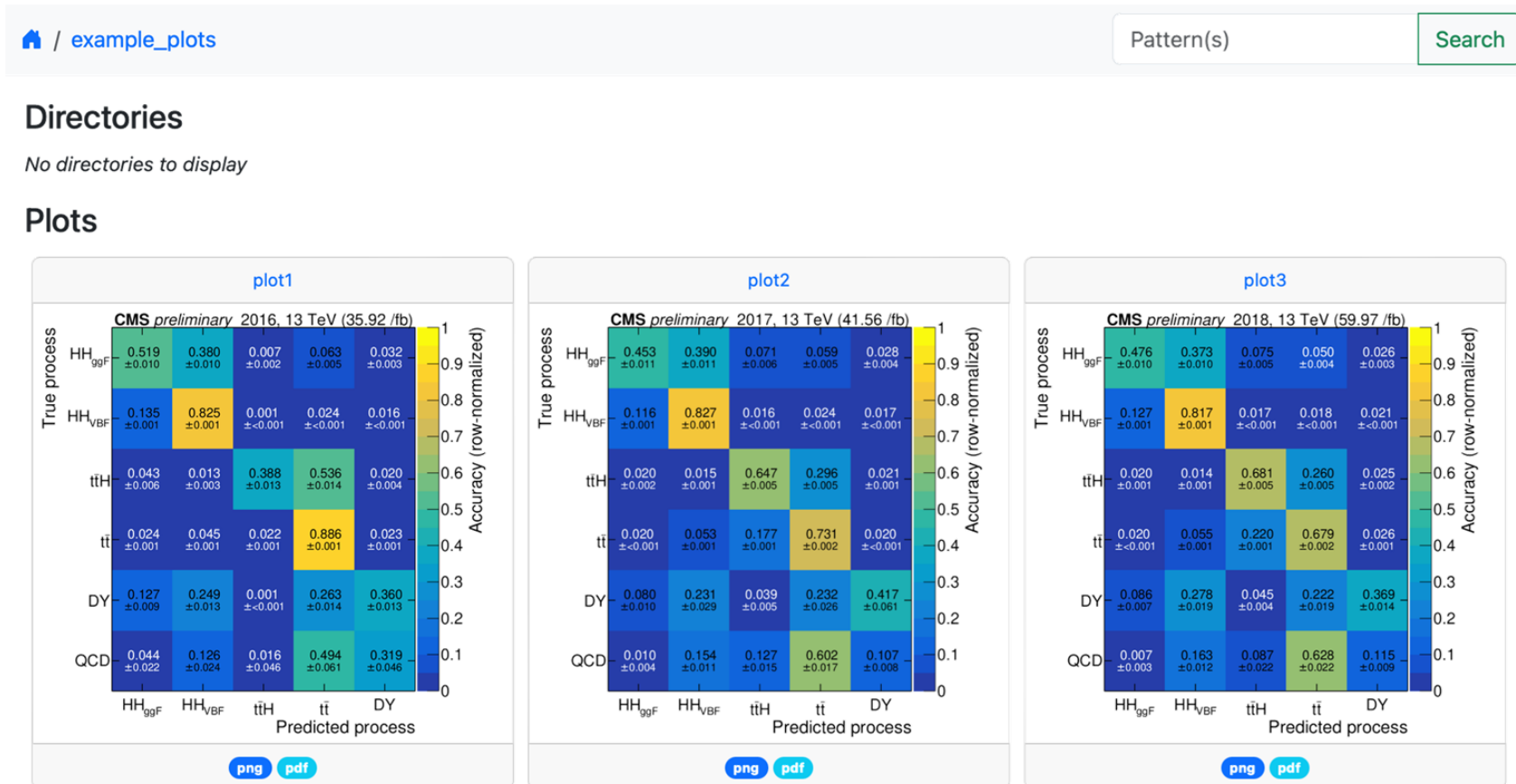
- Reach us at cms-cat-coordination@cern.ch to discuss
- [HSF Data Analysis Working Group](#) (DAWG)
 - Inter-experiment forum to discuss needs and developments



Thank You

Other Useful Tools: Updated php-plots

- Easy to set up plot browser for a personal webpages hosted on CERN eos



CAT Supported Tools

nanoAOD-tools: legacy pyROOT-based sequential framework to skim/extend nanoAODs, and produce plots (modules here)

bamboo: RDF-based python framework that allows to express analysis in a functional style

CMSJMECalculators: RDF-friendly implementation of the recipes for jet and MET variations for CMS

CROWN: RDF-based (C++ and python) framework to generate analysis ntuples (and friends)

columnflow: python (Awkward Arrays)-based backend for columnar, fully-orchestrated HEP analyses

DasAnalysisSystem: ROOT-based tools for analysis with high-level objects

PocketCoffea: configuration framework for Coffea-based analyses on NanoAODs

mkShapesRDF: RDF-based framework for analyses on NanoAODs, which are implemented through config files

1. Significant amount of metadata needed to perform analyses or to extract central recommendations (calibrations, SFs, ...)
2. Complex relations and dependencies render book-keeping highly non-trivial
3. Twiki-as-a-database does not work
 - Information is scattered across various pages with unclear responsibilities and practically no proper versioning
 - High degree of duplication, oftentimes without references to actual sources
4. Work to cope with this is repeated by every member or group
 - This is a very ambitious project & first versions might not be 100% complete, but we intend to release early & often
 - ▷ We see high potential in having a single, common, centralized effort with community-driven input!

-
- Status update given in [previous CAT meetings](#)

The word "order" is written in a large, blue, sans-serif font. The letter 'o' is replaced by a circular image of a diamond ring with many facets, set against a dark background.

- Access a dataset by name

```
In [8]: c.datasets.names()
Out[8]: dict_keys(['wjets', 'ttbar'])
```

list all datasets

```
In [9]: %od.show c.datasets.n.ttbar
```

load and show a dataset

```
Dataset(
  id: 1
  name: 'ttbar'
  variations: {
    nominal: DatasetVariation(
      keys: [
        '/TTbb_4f_TTToSemiLeptonic_TuneCP5-Powheg-OpenLoops-Pythia8/RunIISummer20UL18NanoAODv9-106X_upgrade2018_realistic_v16_L1v1-v1/NANOADSIM'
      ]
      gen_order: 'nnlo'
      n_files: lazy:das_dataset.n_files
      n_events: lazy:das_dataset.n_events
      lfns: lazy:das_lfns.lfns
    )
    scale_up: DatasetVariation(
      keys: [
        '/TTbb_4f_TTToSemiLeptonic_TuneCP5-Powheg-OpenLoops-Pythia8/RunIISummer20UL18NanoAODv9-106X_upgrade2018_realistic_v16_L1v1-v1/NANOADSIM'
      ]
      gen_order: 'nnlo'
    )
  }
)
```

nominal variation

lazy fields

scale_up variation

- When accessing a dataset (`c.datasets.n.ttbar`), its contents are loaded lazily fake values

- In this case, the order-db is accessed and a yaml file is loaded

- The *materialized* dataset still has lazy fields, e.g. `n_files`, `n_events`, `lfns`

- So called "adapters" are defined for them that control how values are obtained when fields are accessed

- Fields can share the same adapter if their values can result from the same request

- Already a handful adapters implemented in order: DAS dataset and LFNS, DBS, order-internal adapters

```
n_files: AdapterModel(
  adapter: 'das_dataset'
  key: 'n_files'
  arguments: {
    keys: [
      '/TTbb_4f_TTToSemiLeptonic_TuneCP5-Powheg-OpenLoops-Pythia8/RunIISummer20UL18NanoAODv9-106X_upgrade2018_realistic_v16_L1v1-v1/NANOADSIM'
    ]
  }
)
```

