Using CMS Open Data in research — challenges and directions

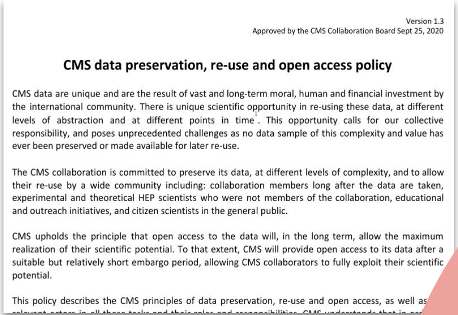
CHEP 2021

Kati Lassila-Perini[1], Clemens Lange[2], Edgar Carrera Jarrin[3], and Mathew Bellis[4].

[1]Helsinki Institute of Physics, Finland
[2]CERN, Switzerland
[3]Universidad San Francisco de Quito, Ecuador
[4]Siena College, USA

May 20, 2021

Introduction



Version 1.3
Approved by the CMS Collaboration Board Sept 25, 2020

CMS data preservation, re-use and open access policy

CMS data are unique and are the result of vast and long-term moral, human and financial investment by the international community. There is unique scientific opportunity in re-using these data, at different levels of abstraction and at different points in time[1]. This opportunity calls for our collective responsibility, and poses unprecedented challenges as no data sample of this complexity and value has ever been preserved or made available for later re-use.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their re-use by a wide community including: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outreach initiatives, and citizen scientists in the general public.

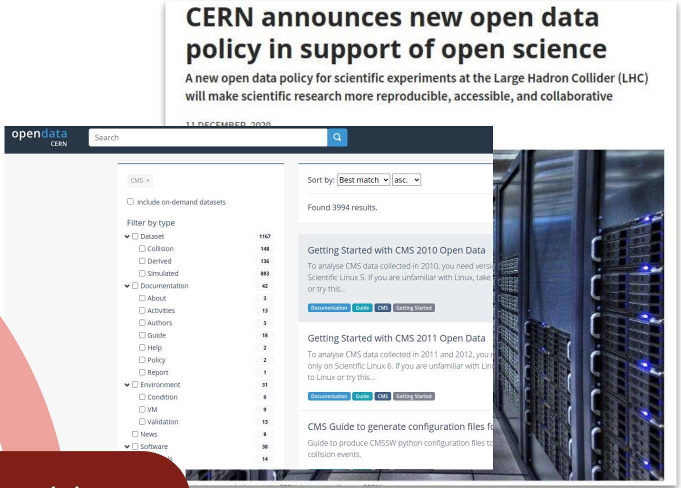
CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their scientific potential.

This policy describes the CMS principles of data preservation, re-use and open access, as well as

CMS Open Data releases

- Leading CERN's and LHC's efforts in open science

- > 2PB of data released since 2014 using the CERN Open Data Portal (CODP)



CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER 2020

CMS ×

□ Include on-demand datasets

Filter by type

□ Dataset	1167
□ Collision	148
□ Derived	136
□ Simulated	883
□ Documentation	42
□ About	3
□ Activities	13
□ Authors	3
□ Guide	18
□ Help	2
□ Policy	2
□ Report	1
□ Environment	31
□ Condition	9
□ VM	9
□ Validation	13
□ News	8
□ Software	38
	14

Sort by: Best match asc.

Found 3994 results.

Getting Started with CMS 2010 Open Data
To analyse CMS data collected in 2010, you need versi
Scientific Linux 5. If you are unfamiliar with Linux, take
or try this...
Documentation Guide CMS Getting Started

Getting Started with CMS 2011 Open Data
To analyse CMS data collected in 2011 and 2012, you
only on Scientific Linux 6. If you are unfamiliar with Lin
to Linux or try this...
Documentation Guide CMS Getting Started

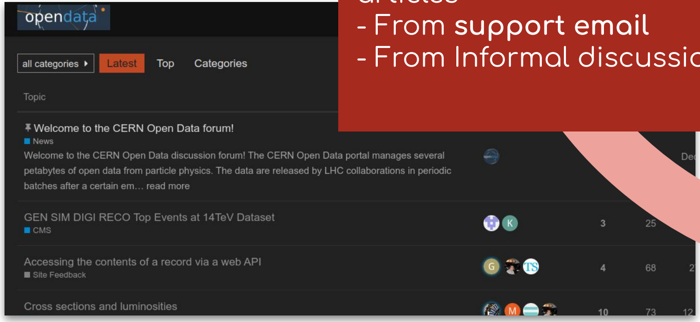
CMS Guide to generate configuration files f
Guide to produce CMSSW python configuration files to
collision events.

Feedback

- CERN Open Data Forum
- From comments in published articles
- From **support email**
- From Informal discussions

Usage in Research*

Topics include: standard model (SM) novel studies, re-measurements, searches beyond the SM, new methods and techniques (e.g., machine learning algos)

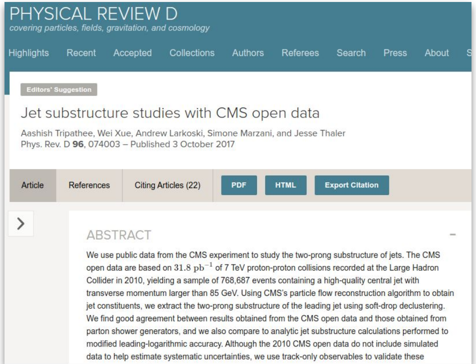


Welcome to the CERN Open Data forum!
News
Welcome to the CERN Open Data discussion forum! The CERN Open Data portal manages several petabytes of open data from particle physics. The data are released by LHC collaborations in periodic batches after a certain em… read more
De

GEN SIM DIGI RECO Top Events at 14TeV Dataset
CMS
3 25

Accessing the contents of a record via a web API
Site Feedback
4 68 2

Cross sections and luminosities
10 73 12



Jet substructure studies with CMS open data

Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, and Jesse Thaler
Phys. Rev. D **96**, 074003 – Published 3 October 2017

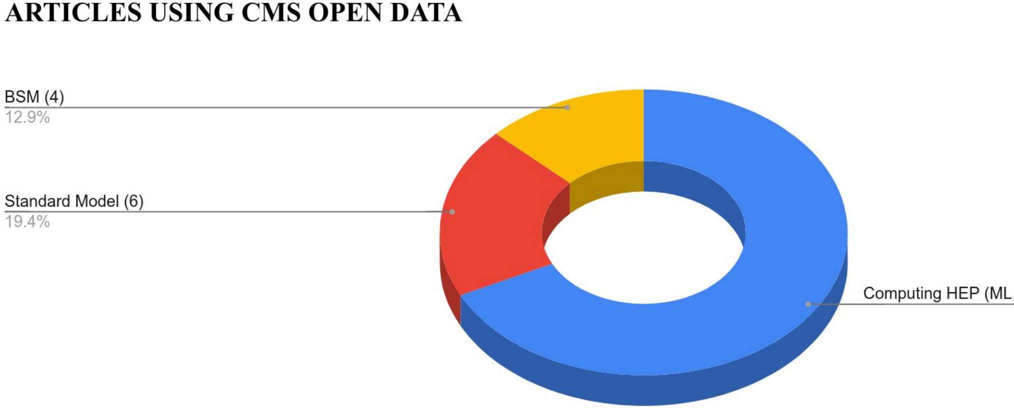
ABSTRACT

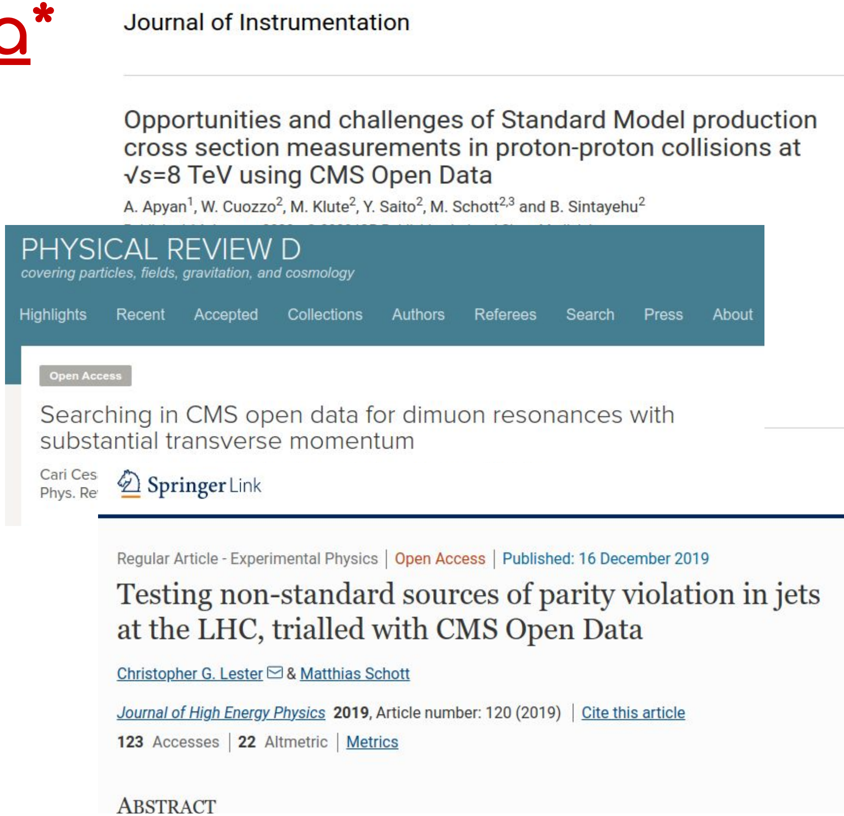
We use public data from the CMS experiment to study the two-prong substructure of jets. The CMS open data are based on $31.8\ \mathrm{pb}^{-1}$ of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,687 events containing a high-quality central jet with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the two-prong substructure of the leading jet using soft-drop declustering. We find good agreement between results obtained from the CMS open data and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS open data do not include simulated data to help estimate systematic uncertainties, we use track-only observables to validate these

*Not an exact search link but a reference

Articles using CMS open data*

AREA	NUMBER OF ARTICLES (as of may 13, 2021)			
	Published in a journal	Arxiv only or submitted to a journal	Contribution to conferences	TOTAL
Computing HEP (ML, new algos and techniques, etc.)	8	5	8	21
Standard Model	5		1	6
BSM	2	2		4
TOTAL	15	7	9	31





Journal of Instrumentation

Opportunities and challenges of Standard Model production cross section measurements in proton-proton collisions at $\sqrt{s}$=8 TeV using CMS Open Data

A. Apyan[1], W. Cuozzo[2], M. Klute[2], Y. Saito[2], M. Schott[2,3] and B. Sintayehu[2]

PHYSICAL REVIEW D

covering particles, fields, gravitation, and cosmology

Highlights Recent Accepted Collections Authors Referees Search Press About

Open Access

Searching in CMS open data for dimuon resonances with substantial transverse momentum

Cari Ces

Phys. Re

SpringerLink

Regular Article - Experimental Physics | Open Access | Published: 16 December 2019

Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS Open Data

Christopher G. Lester & Matthias Schott

Journal of High Energy Physics **2019**, Article number: 120 (2019) | Cite this article

123 Accesses | **22** Altmetric | Metrics

ABSTRACT

* Link is not exact but just a reference for easier search.

CMS open data

Release Policy

- Some embargo time and restrictions apply.
- Start of release after a few years of the end of data taking
- Most of Run 1 data released

Data Format

- Analysis Object Data (AOD) format
- Based on ROOT and CMSSW
- Research quality
- Slimmer miniAOD and nanoAOD used and foreseen in Run 2.

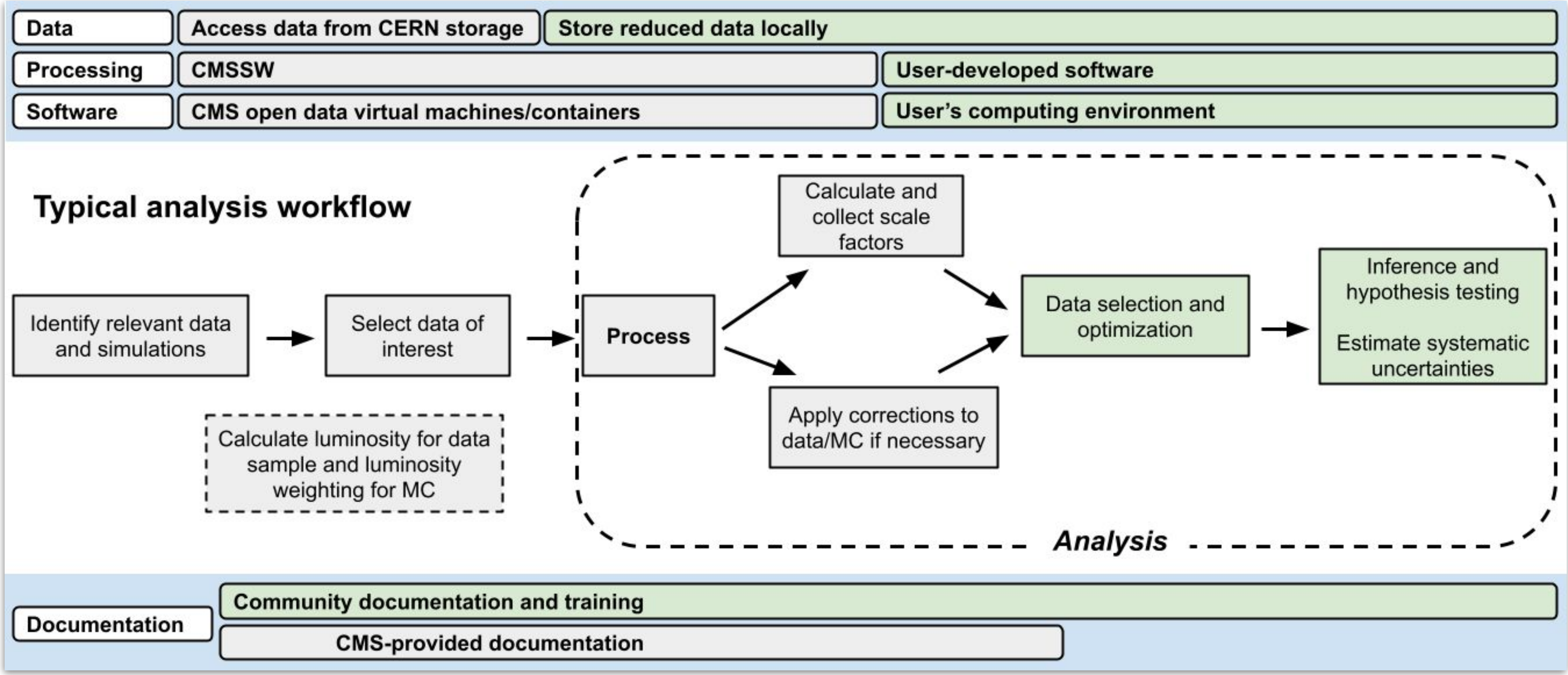
Data products

- Collision Data
- Simulated Data (MC)

Software and Associated products

- CMSSW
- Data Quality
- Conditions database (alignment, calibration, etc)
- Luminosity information
- Examples (some with automated workflows) and topical guide pages

Using CMS open data



Data | Access data from CERN storage | Store reduced data locally

Processing | CMSSW | User-developed software

Software | CMS open data virtual machines/containers | User's computing environment

Typical analysis workflow

Identify relevant data and simulations → Select data of interest → Process

Calculate luminosity for data sample and luminosity weighting for MC

Process → Calculate and collect scale factors → Data selection and optimization

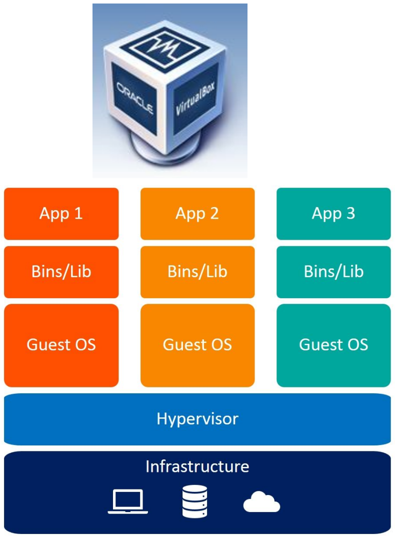
Process → Apply corrections to data/MC if necessary → Data selection and optimization

Data selection and optimization → Inference and hypothesis testing; Estimate systematic uncertainties

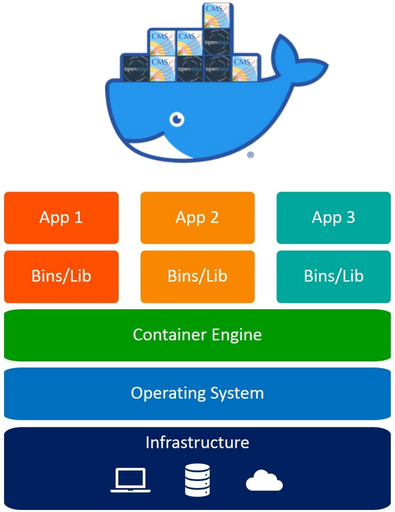
Analysis

Documentation | Community documentation and training | CMS-provided documentation

Using CMS open data



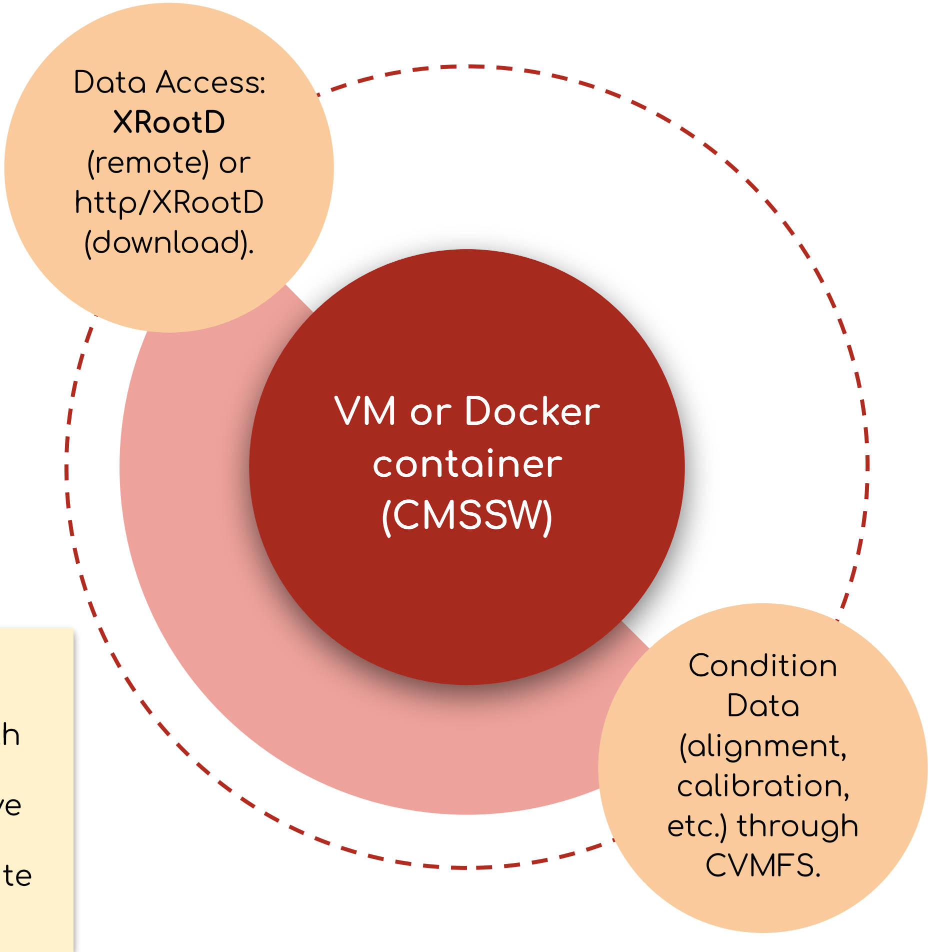
Virtual Machines



Containers

- Based on CernVM
- Quick first access
- CMSSW dependencies accessed through CVMFS
- Challenging to scale up for full analysis

- Different kind of images provided
- CVMFS may be dispensed with
- Layered file system makes it easier to maintain and preserve (Gitlab CI).
- Easier to use in batch compute systems.

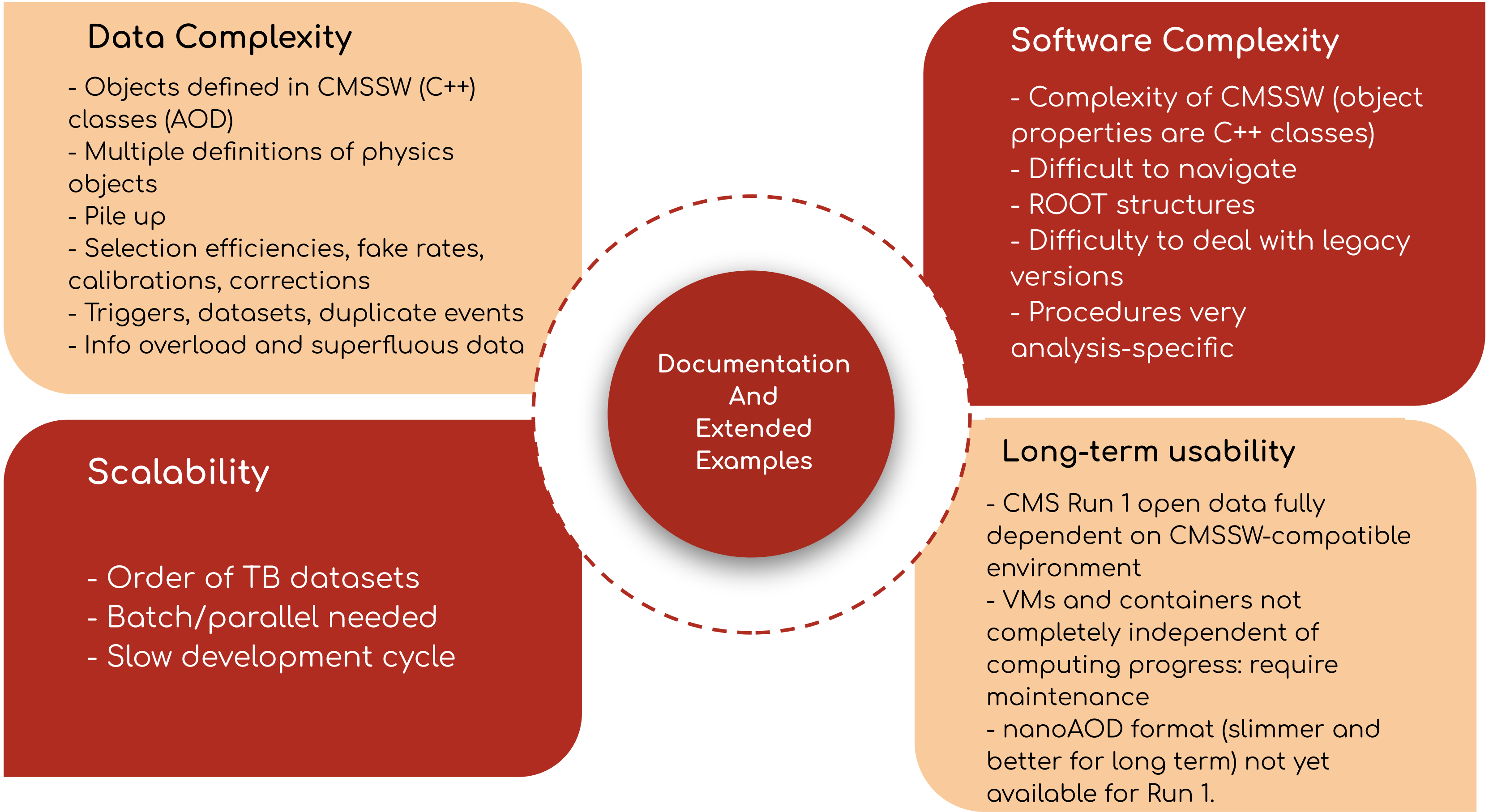


Data Access: **XRootD** (remote) or http/XRootD (download).

VM or Docker container (CMSSW)

Condition Data (alignment, calibration, etc.) through CVMFS.

User feedback and challenges



Data Complexity

- Objects defined in CMSSW (C++) classes (AOD)
- Multiple definitions of physics objects
- Pile up
- Selection efficiencies, fake rates, calibrations, corrections
- Triggers, datasets, duplicate events
- Info overload and superfluous data

Software Complexity

- Complexity of CMSSW (object properties are C++ classes)
- Difficult to navigate
- ROOT structures
- Difficulty to deal with legacy versions
- Procedures very analysis-specific

Documentation And Extended Examples

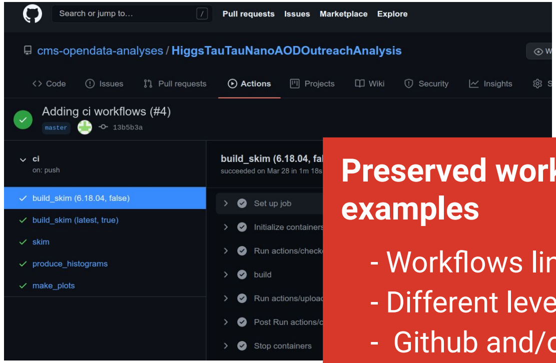
Scalability

- Order of TB datasets
- Batch/parallel needed
- Slow development cycle

Long-term usability

- CMS Run 1 open data fully dependent on CMSSW-compatible environment
- VMs and containers not completely independent of computing progress: require maintenance
- nanoAOD format (slimmer and better for long term) not yet available for Run 1.

Measures to improve usability



cms-opendata-analyses / HiggsTauTauNanoAODOutreachAnalysis

Adding ci workflows (#4)

master · 13b5b3a

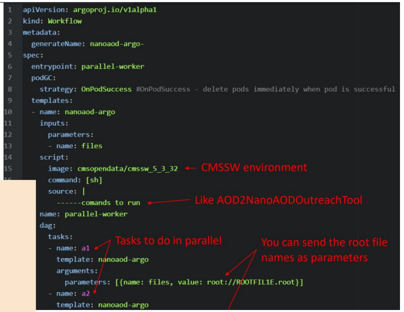
ci
on: push

- ✓ build_skim (6.18.04, false)
- ✓ build_skim (latest, true)
- ✓ skim
- ✓ produce_histograms
- ✓ make_plots

build_skim (6.18.04, fal
succeeded on Mar 28 in 1m 18s

- Set up job
- Initialize containers
- Run actions/checko
- build
- Run actions/uploac
- Post Run actions/c
- Stop containers

- Easy way to manage a workflow inside Kubernetes clusters



```yaml
apiVersion: argoproj.io/v1alpha1
kind: Workflow
metadata:
  generateName: nanoaod-argo-
spec:
  entrypoint: parallel-worker
  podGC:
    strategy: OnPodSuccess #OnPodSuccess - delete pods immediately when pod is successful
  templates:
  - name: nanoaod-argo
    inputs:
      parameters:
      - name: files
    script:
      image: cmsopendata/cmssw_5_3_32
      command: [sh]
      source: |
       ------comands to run
    name: parallel-worker
    dag:
      tasks:
      - name: a1
        template: nanoaod-argo
        arguments:
          parameters: [{name: files, value: root://ROOTFIL1E.root}]
      - name: a2
        template: nanoaod-argo
```

CMSSW environment

Like AOD2NanoAODOutreachTool

Tasks to do in parallel

You can send the root file names as parameters

Preserved workflows as examples

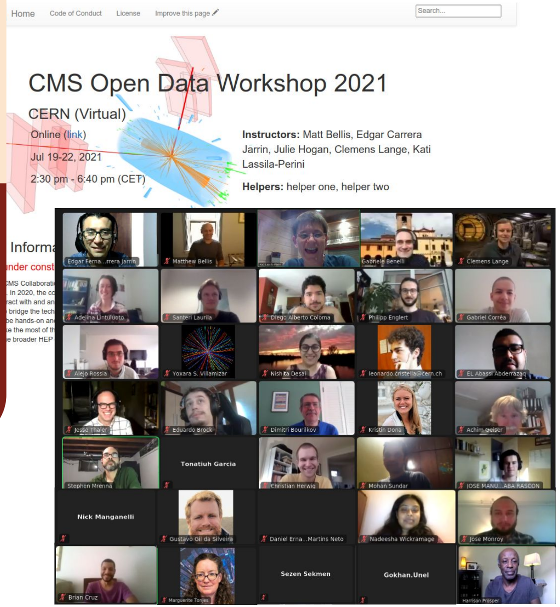
- Workflows linked from CODP
- Different levels of complexity
- Github and/or GitLab CI/CD
- Usage of Kubernetes/minikube
- Some workflows executed in REANA

Data analysis in the cloud

- Usage of scalable commercial clusters
- Simplify and test workflows on Kubernetes cloud clusters
- Test cloud local data download (via XRootD) and local cached CVMFS server

Documentation and Training

- Twiki pages, CODP records, Github code available but difficult
- Analysis examples in CODP hard to navigate
- CMS Open Data Guide
- Since 2020, regular hands-on workshops and training events



CMS Open Data Workshop 2021

CERN (Virtual)

Online (link)

Jul 19-22, 2021

2:30 pm - 6:40 pm (CET)

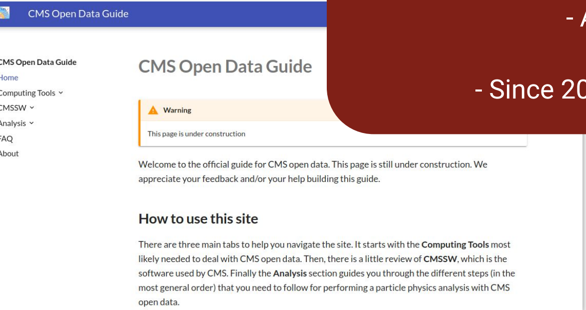
Instructors: Matt Bellis, Edgar Carrera Jarrin, Julie Hogan, Clemens Lange, Kati Lassila-Perini

Helpers: helper one, helper two

Informa

nder const

MS Collaborati
In 2020, the co
act with and an
bridge the tech
be hands-on an
e the most of th
e broader HEP



CMS Open Data Guide

⚠ Warning

This page is under construction

Welcome to the official guide for CMS open data. This page is still under construction. We appreciate your feedback and/or your help building this guide.

How to use this site

There are three main tabs to help you navigate the site. It starts with the **Computing Tools** most likely needed to deal with CMS open data. Then, there is a little review of **CMSSW**, which is the software used by CMS. Finally the **Analysis** section guides you through the different steps (in the most general order) that you need to follow for performing a particle physics analysis with CMS open data.

Summary and Outlook

- CMS has spearheaded CERN's open data efforts
- CMS open data have been successfully used in original scientific research (jet physics, sm measurements, new methods and algorithms)
- Usage of these data has opened opportunities for deeper collaboration between theorists and experimentalists
- Limitations and challenges have been identified from user feedback and self assessment:
    - Information extraction from data, procedures, documentation and examples.
- Measures (within limited person-power capabilities) are being taken in order to improve usability:
    - Better (automated) implementation of workflows, cloud computing, CMS Open Data Guide and training events.
- CMS will maintain its commitment to open data and open science
- Next CMS Open Data Workshop, July 19-22, 2021. Registrations open:https://indico.cern.ch/e/CmsODW2021