



# Open Data at CERN

Tibor Šimko  
CERN IT

UNIGE Open Science Day · 21 November 2018

# CERN Open Data portal

- `opendata.cern.ch` launched in November 2014
- LHC collaboration data policies
  - restricted → embargo period (~5 years) → open
- over 1.5 Petabytes of open particle physics data
  - datasets, software, VMs, configuration, documentation, ...
- users
  - education: general public, high-school students, masterclasses
  - research: data scientists, physicists

*Developed by CERN-IT and CERN-SIS  
in close collaboration with Experiments*



# CERN Open Data portal

opendata  
CERN

About

Explore more than **1 petabyte**  
of open data from particle physics!

Start typing...

Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

**Explore**

- [datasets](#)
- [software](#)
- [environments](#)
- [documentation](#)

**Focus on**

- [ATLAS](#)
- [ALICE](#)
- [CMS](#)
- [LHCb](#)

Get started

<http://opendata.cern.ch/>

# Information organisation

open data CERN Search About

## Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)

/Mu/Run2010B-Apr21ReReco-v1/AOD, CMS collaboration

Cite as: CMS collaboration (2014). Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.B8MR.C4A2

Home Contact CMS Software energy 1TeV Accelerator CERN LHC Fermilab Dataset /Mu/Run2010B-Apr21ReReco-v1/AOD

### Description

Mu primary dataset in AOD format from RunB of 2010

### Notes

This dataset contains all runs from 2010 RunB. The list of validated runs, which must be applied to all analyses, can be found in

[CMS list of validated runs Cert\\_136033-149442\\_7TeV\\_Apr21ReReco\\_Collisions10\\_JSON\\_v2.txt](#)

### Related Datasets

[/Mu/Run2010B-v1/RAW](#)

### Characteristics

Dataset: **32376291** events **2979** files **3.2 TB** in total

### System Details

Global tag: FT\_R\_42\_V10A:AB  
Recommended release for analysis: CMSSW\_4\_2\_1\_patch1

## How were these data selected?

There are four categories of triggers in the Mu dataset (with significant overlaps):

- 70% inclusive single muon triggers with varying trigger pt threshold 3.5,7.9,11,13,15,17,19,21 GeV plus a few with loosened quality cuts.
- 20% isolated single muon triggers with varying trigger pt threshold 9,11,13,15,17 GeV.
- 10% inclusive dimuon triggers with varying trigger pt threshold 3.5 GeV plus one Z-muon trigger with loosened quality cuts.
- 20% combinations of muon triggers with various pt thresholds 3.5,7.8,9,11 GeV with some EM/e/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

## How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

[CMS data quality monitoring: Systems and experiences](#)

[The CMS Data Quality Monitoring software experience and future improvements](#)

[The CMS data quality monitoring software: experience and future prospects](#)

## How can you use these data?

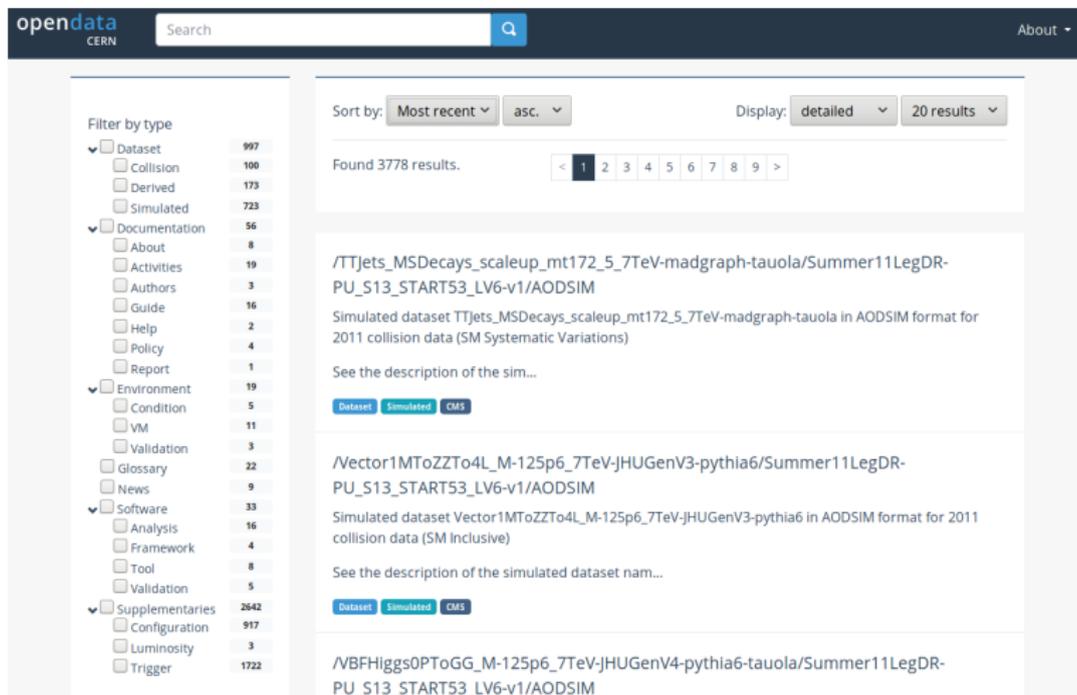
You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

[How to install the CMS Virtual Machine](#)

[Getting started with CMS open data](#)

Context information about data selection, validation, use

# Information discovery



The screenshot shows the OpenData CERN website interface. At the top left is the 'opendata CERN' logo. A search bar is located at the top center. On the right, there is an 'About' link. Below the search bar, there is a 'Filter by type' sidebar on the left and search controls on the right. The sidebar lists various categories with their respective counts, such as 'Dataset' (997), 'Documentation' (56), 'Environment' (19), 'Software' (33), and 'Supplementaries' (2642). The search controls include 'Sort by: Most recent', 'asc.', 'Display: detailed', and '20 results'. Below these, it says 'Found 3778 results.' and shows a pagination bar with page numbers 1 through 9. The main content area displays two search results. The first result is for a simulated dataset: '/TTjets\_MSDecays\_scaleup\_mt172\_5\_7TeV-madgraph-tauola/Summer11LegDR-PU\_S13\_START53\_LV6-v1/AODSIM'. The second result is for a simulated dataset: '/Vector1MToZZto4L\_M-125p6\_7TeV-JHUGenV3-pythia6/Summer11LegDR-PU\_S13\_START53\_LV6-v1/AODSIM'. Each result includes a description and a link to see the description. There are also buttons for 'Dataset', 'Simulated', and 'CMS' for each result.

opendata CERN

Search

About

Filter by type

- Dataset 997
  - Collision 100
  - Derived 173
  - Simulated 723
- Documentation 56
  - About 8
  - Activities 19
  - Authors 3
  - Guide 16
  - Help 2
  - Policy 4
  - Report 1
- Environment 19
  - Condition 5
  - VM 11
  - Validation 3
  - Glossary 22
  - News 9
- Software 33
  - Analysis 16
  - Framework 4
  - Tool 8
  - Validation 5
- Supplementaries 2642
  - Configuration 917
  - Luminosity 3
  - Trigger 1722

Sort by: Most recent asc. Display: detailed 20 results

Found 3778 results.

< 1 2 3 4 5 6 7 8 9 >

/TTjets\_MSDecays\_scaleup\_mt172\_5\_7TeV-madgraph-tauola/Summer11LegDR-PU\_S13\_START53\_LV6-v1/AODSIM

Simulated dataset TTjets\_MSDecays\_scaleup\_mt172\_5\_7TeV-madgraph-tauola in AODSIM format for 2011 collision data (SM Systematic Variations)

See the description of the sim...

Dataset Simulated CMS

/Vector1MToZZto4L\_M-125p6\_7TeV-JHUGenV3-pythia6/Summer11LegDR-PU\_S13\_START53\_LV6-v1/AODSIM

Simulated dataset Vector1MToZZto4L\_M-125p6\_7TeV-JHUGenV3-pythia6 in AODSIM format for 2011 collision data (SM Inclusive)

See the description of the simulated dataset nam...

Dataset Simulated CMS

/VBFHiggs0PToGG\_M-125p6\_7TeV-JHUGenV4-pythia6-tauola/Summer11LegDR-PU\_S13\_START53\_LV6-v1/AODSIM

Explore a variety of data, software, VMs, supplementary material. . .

# Visualise detector events

opendata CERN Search About

Need HELP?

iSpy WebGL DoubleMuon:Events/Run\_167674/Event\_255544818 [3 of 25]

Detector

- Pixel Barrel
- Pixel Endcap (+)
- Pixel Endcap (-)
- Tracker Inner Barrel
- Tracker Outer Barrel
- Tracker Inner Detector (+)
- Tracker Inner Detector (-)
- Tracker Endcap (+)

CMS Experiment at the LHC, CERN  
Data recorded: 2011-Jun-25 00:15:00.683123 GMT  
Run / Event / LS: 167674 / 255544818 / 209

Click on a name under "Provenance", "Tracking", "ECAL", "HCAL", "Muon", and "Physics" to view contents in table

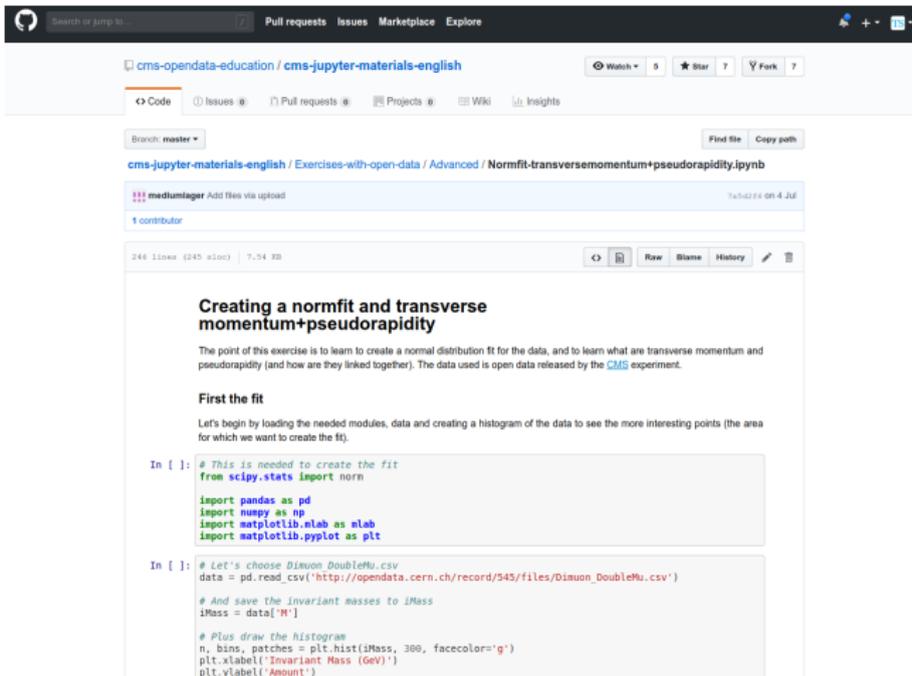
Interactive event display for high-level derived datasets

# Visualise histograms



Interactive histogramming for high-level derived datasets

# Jupyter notebooks



The screenshot shows a GitHub repository page for 'cms-opensource / cms-jupyter-materials-english'. The repository is on the 'master' branch. The selected file is 'Normfit-transversemomentum+pseudorapidity.ipynb'. The notebook content is displayed in a code editor with a light blue background. It includes a title, a description of the exercise, and two code blocks.

## Creating a normfit and transverse momentum+pseudorapidity

The point of this exercise is to learn to create a normal distribution fit for the data, and to learn what are transverse momentum and pseudorapidity (and how are they linked together). The data used is open data released by the [CMS](#) experiment.

### First the fit

Let's begin by loading the needed modules, data and creating a histogram of the data to see the more interesting points (the area for which we want to create the fit).

```
In [ ]: # This is needed to create the fit
from scipy.stats import norm

import pandas as pd
import numpy as np
import matplotlib.mlab as mlab
import matplotlib.pyplot as plt
```

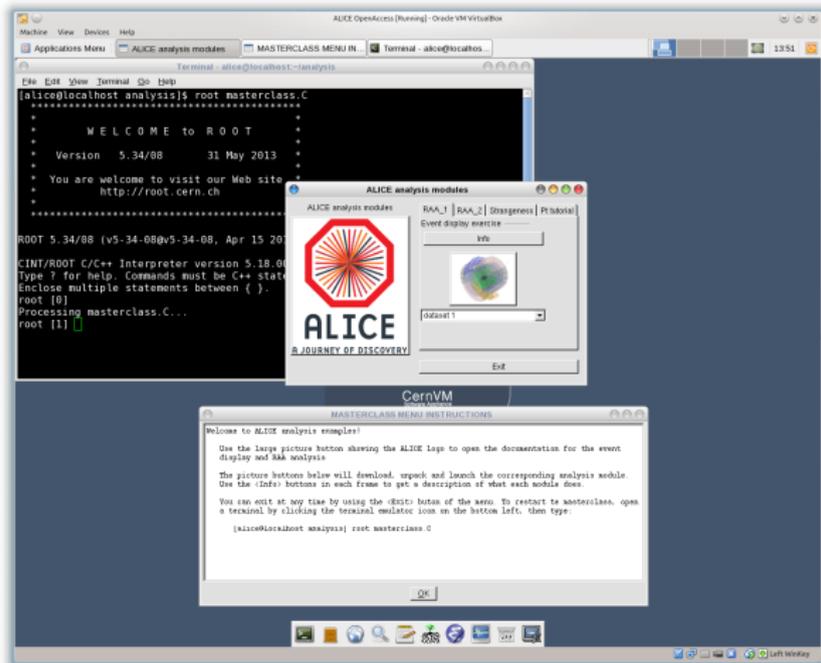
```
In [ ]: # Let's choose Dimaon_DoubleMu.csv
data = pd.read_csv('http://opendata.cern.ch/record/545/files/Dimaon_DoubleMu.csv')

# And save the invariant masses to iMass
iMass = data['M']

# Plus draw the histogram
n, bins, patches = plt.hist(iMass, 300, facecolor='g')
plt.xlabel('Invariant Mass (GeV)')
plt.ylabel('Amount')
```

CMS education activities using notebooks and CMS open data

# Virtual machines



Install CernVM virtual machines to explore primary datasets

# Analysis examples

## Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin;

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS\_JKB8\_RR42

Software Analysis CMS Accelerator CERN/LHC

### Description

This research level example is a strongly simplified reimplemention of parts of the original CMS Higgs to four lepton analysis published in [Phys.Lett. B716 \(2012\) 30-61](#), [arXiv:1207.7235](#).

The published reference plot which is being approximated in this example is [https://inspirehep.net/record/1124338/files/H4l\\_mass\\_3.png](https://inspirehep.net/record/1124338/files/H4l_mass_3.png). Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

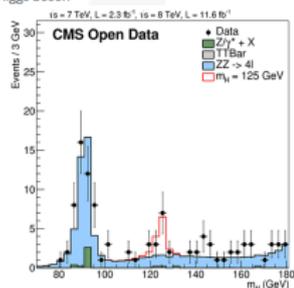
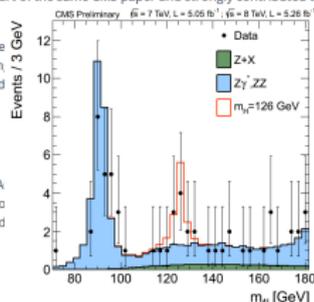
The example consists of different levels of complexity. The highest level minimal understanding of the content of this paper and of the meanin educational exercises. The lower levels might also be interesting for ed with the linux operating system and [the ROOT analysis tool](#).

### Use with

The example uses legacy versions of the original CMS datasets in the A publication due to improved calibrations. It also uses legacy versions o but not identical to, the ones in the original publication. These legacy d in many later CMS publications.

`/DoubleElectron/Run2011A-12Oct2013-v1/AOD`

`/DoubleMu/Run2011A-12Oct2013-v1/AOD`



Run realistic physics analysis examples

# Release-driven usage patterns

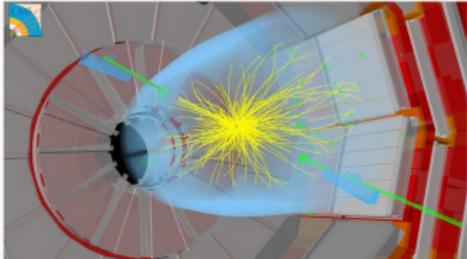
TC News Video Events Crunchbase

10TH ANNUAL CRUNCHIES AWARDS See who Will Take home Top Honors At Our Tech Awards Event This February. Get Your Tickets Today >

Education Large Hadron Collider LHC Physics CERN

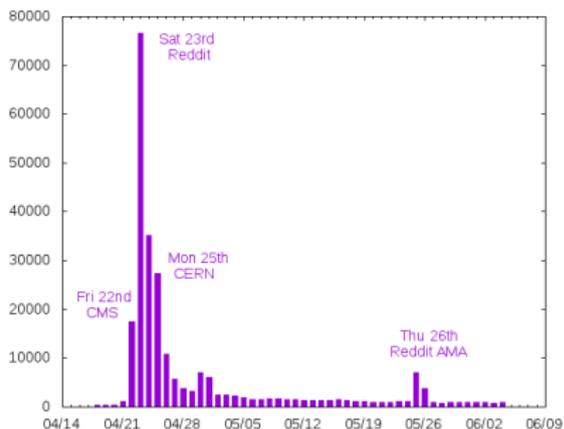
## CERN releases 300TB of Large Hadron Collider data into open access

Posted Apr 20, 2016 by Devin Coldewey, Contributor



Cancel your plans for this weekend! CERN just dropped 300 terabytes of hot collider data on the world and you know you want to take a look.

Open data releases are widely covered by general media



Six weeks in 2016: 200K users, 40K viewed records, 70K used event display, 3K used histogramming

# Research made open at large

PHYSICAL REVIEW LETTERS

PHYSICAL REVIEW D

## Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,<sup>1,2</sup> Srinivas Muzaffar,<sup>1,2</sup> Jesse Thaler,<sup>1,2</sup> Anshu Tripathi,<sup>1,2</sup> and Wei Xue<sup>1,2</sup>

<sup>1</sup>Princeton University, 807 GFDL, Princeton, New Jersey 08542, USA  
<sup>2</sup>University of Cambridge, The Isaac Newton Institute for Mathematical Sciences, 90 Williamstown Road, Port Melbourne, Victoria 3207, Australia  
<sup>3</sup>Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA  
(Received 1 May 2017; revised manuscript received 27 July 2017; published 26 September 2017)

The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between particles. Despite its ubiquitous appearance in many QCD calculations, the splitting function cannot be measured directly, since it always appears multiplied by a collinear singularity factor. Recently, however, a new jet substructure observable was introduced which compensates for the splitting function for sufficiently light jet emission. This provides a way to expose the splitting function through an observable measurement at the Large Hadron Collider. In this Letter, we use public data released by the CMS experiment to study the two-prong substructure of jets and use the  $1 \rightarrow 2$  splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

DOI: 10.1103/PhysRevLett.119.132003

Quantum chromodynamics (QCD), like any weakly coupled gauge theory, exhibits universal behavior in the small angle limit. When two particles become collinear in QCD, the cross section for a  $2 \rightarrow n$  scattering process factorizes into a  $2 \rightarrow n-1$  scattering cross section multiplied by a universal  $1 \rightarrow 2$  splitting probability, with corrections suppressed by the degree of collinearity. Collinear universality is a fundamental property of QCD and appears in many applications, most famously in deriving the Dokshitzer-Gribov-Lipatov-Altarelli-Feynman evolution equations [1–3] (for text [4–13]), and in the heart of the factorization theorem in hadron-hadron collisions [14,15]. In addition, parton shower generators are based on recursively applying  $1 \rightarrow 2$  splitting [16–18], fixed-order resummation schemes utilize the  $1 \rightarrow 2$  splitting function [19–21], and the  $1 \rightarrow 2$  splitting [22–24] is spoiled in the presence of Glauber modes [25–27]. More recently, jet substructure techniques [28–32] have been introduced to distinguish  $1 \rightarrow n$  decays of heavy particles from  $1 \rightarrow n$  splittings in QCD in order to enhance the search for new physics at the Large Hadron Collider (LHC) [33–36].

Despite its ubiquity, however, the  $1 \rightarrow 2$  splitting function cannot be directly measured at a collider, since collinear universality is inseparable from the existence of collinear singularities and only related nonambiguously to fragmentation functions. Specifically, when two particles are separated by an angle  $\theta$ , the  $1 \rightarrow 2$  splitting probability takes the form

$$dP_{1 \rightarrow 2} = \frac{d\sigma}{d\sigma_0} dP_{1 \rightarrow 2}(\theta). \quad (1)$$

where the  $P_{1 \rightarrow 2}$  are the Altarelli-Parisi QCD splitting functions [1] which depend on the momentum fraction  $z$  and the parent flavor,  $f$ , and  $d\sigma_0$  is the cross section for a hard scattering process. In order to cancel corresponding virtual singularities from loop diagrams, in this sense, there is no way to directly measure the splitting function,  $P_{1 \rightarrow 2}(\theta)$ , in data, though there is a coarse overbalancing indirect evidence that  $P_{1 \rightarrow 2}(\theta)$  is a universal function from the many successes of QCD in describing high-energy scattering (see, e.g., [37–39]).

In this Letter, we present a novel method to test the  $1 \rightarrow 2$  splitting function in QCD by studying the two-prong substructure of jets. Our method is based on soft drop substructure [40] (also known as [41,42]), which recursively removes soft radiation from a jet until hard two-prong substructure is found. When applied to collinear quark and gluon related jets with intrinsic substructure, soft drop exposes the collinear core of the jet. As shown in Ref. [7], the momentum sharing between the two prongs (denoted  $z_{12}$ ) is closely related to the momentum fraction  $z$  appearing in Fig. 1(a), and the cross section for  $z_{12}$  corresponds to the QCD splitting function in the high-energy limit. While values of  $z_{12}$  have appeared in many jet substructure studies, notably the  $\beta$  parameter in Refs. [43–72], to the best of our knowledge, no published distribution has ever been presented using actual collider data, though there are preliminary  $z_{12}$  results from CMS [73], STAR [74], and ALICE [75] Collaborations. Here, we present the first analysis of  $z_{12}$  using LHC data, taking advantage of the first time of public data released by the CMS experiment [76].

The CMS Open Data are derived from 7 TeV center-of-mass proton-proton collisions recorded in 2010 and released to the public on the CERN Open Data Portal in November 2014 [77]. The data are provided in analysis object (AOD) format, which is a CMS-specific data scheme based

PHYSICAL REVIEW D

119

## Jet substructure studies with CMS open data

Anshu Tripathi,<sup>1,2</sup> Wei Xue,<sup>1,2</sup> Andrew Larkoski,<sup>1,2</sup> Srinivas Muzaffar,<sup>1,2</sup> and Jesse Thaler<sup>1,2</sup>

<sup>1</sup>Center for Theoretical Physics, Massachusetts Institute of Technology,

Cambridge, Massachusetts 02139, USA

<sup>2</sup>Physics Department, Reed College, Portland, Oregon 97202, USA

<sup>3</sup>University of Angles, The State University of New York, Stony Brook 11794-3500, USA

(Received 1 May 2017; published 1 October 2017)

We use public data from the CMS experiment to study the two-prong substructure of jets. The CMS open data are based on 34.3 fb<sup>-1</sup> of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, holding a sample of 706,607 events containing a high quality control jet with numerous 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 proton shower generators, and we also attempt to analyze jet substructure calculations performed on modified leading-jet-like momenta. Although the 2010 CMS open data do not include simulated data to help estimate systematic uncertainties, we use track-level observables to validate these substructure studies.

DOI: 10.1103/PhysRevD.96.104003

## 1. INTRODUCTION

In November 2014, the CMS experiment at the Large Hadron Collider (LHC) announced the CMS Open Data project [1]. To our knowledge, this is the first time in the history of particle physics that research-grade collider data has been made publicly available for use outside of an official experimental collaboration. The CMS open data were reconstructed from 7 TeV proton-proton collisions in 2010, corresponding to a single low-luminosity running environment where pile-up contributions were minimal and trigger thresholds were relatively low. The CMS open data present an occasion especially to the particle physics community, both for performing physics studies that would be more difficult at higher luminosities and for demonstrating the scientific value of open data releases.

In this paper, we use the CMS open data to analyze the substructure of jets. Jets are collimated sprays of particles that are typically produced by LHC collisions, and by studying the substructure of jets, one can gain valuable information about their parentage [2–10]. A key application of jet substructure is tagging heavy objects like top quarks [11–13] and electroweak bosons [14,15,16,22–26,50]. To successfully tag such objects, though, one first has to understand the radiation patterns of ordinary quark and gluon jets [28–40,73], which are the main backgrounds to boosted objects. The CMS open data are a fantastic

resource for performing these baseline quark-hadron studies. Using the Jet Primary Dataset [78], we perform initial investigations of the two-prong substructure of jets as well as present a general analysis framework to facilitate future studies. The effort is complementary to the growing catalog of jet substructure measurements performed within the ATLAS and CMS collaborations [77–199].

The core of our analysis is based on soft-drop declustering [40], which is a jet grooming technique [20–202] that mitigates jet contamination from initial state radiation (ISR), underlying event (UE), and pileup. For the studies in this paper, we use the soft-drop parameter  $\beta$  equal to zero, such that the soft drop behaves like the modified mass drop trigger (MDT) [20,203]. After soft dropping, a jet is composed of two well-defined subjets, which can then be used to derive various two-prong substructure observables. In addition to computing the CMS open data to probe shower generators, we perform first-principles calculations of soft-dropped observables using recently developed analytic techniques [40,203,206]. In a companion paper, we use soft drop to expose the QCD splitting function in the CMS open data [207], another strategy we used to probe shower generators. We perform first-principles calculations of soft-dropped observables using recently developed analytic techniques [40,203,206]. In a companion paper, we use soft drop to expose the QCD splitting function in the CMS open data [207], another strategy we used to probe shower generators. We perform first-principles calculations of soft-dropped observables using recently developed analytic techniques [40,203,206].

For studying jet substructure, the key features of the CMS open data is that they contain full information about particle

published for the  
copyright for this  
article by the  
author(s) under  
the terms of the  
Creative Commons  
Attribution  
License (CC BY)  
4.0  
International  
License  
http://creativecommons.org/licenses/by/4.0/

To highlight the openness of the field, we have attempted to list all published jet substructure measurements from ATLAS and CMS. Please contact us if we missed a reference.  
The original manuscript (pre-proof) was submitted to the journal on 1 May 2017. The final version (proof) was submitted on 27 July 2017. The article was published on 26 September 2017.  
The copyright term for this article will expire on 10/10/2037.

0031-9007/17/119(13):132003

132003-1

© 2017 American Physical Society

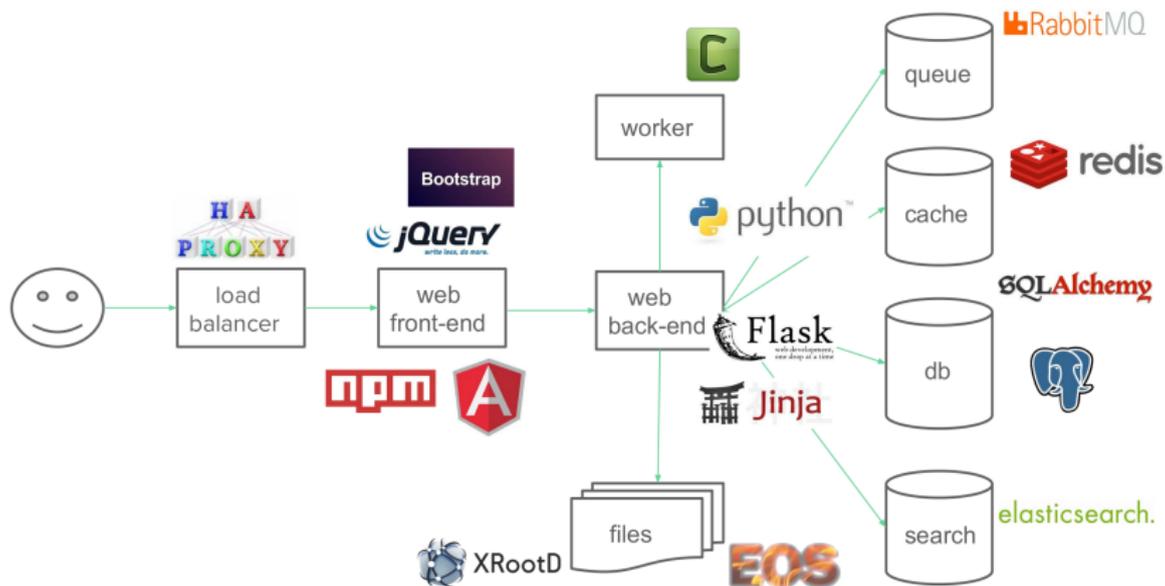
2478-0033/17/119(13):132003

04000-1

© 2017 American Physical Society

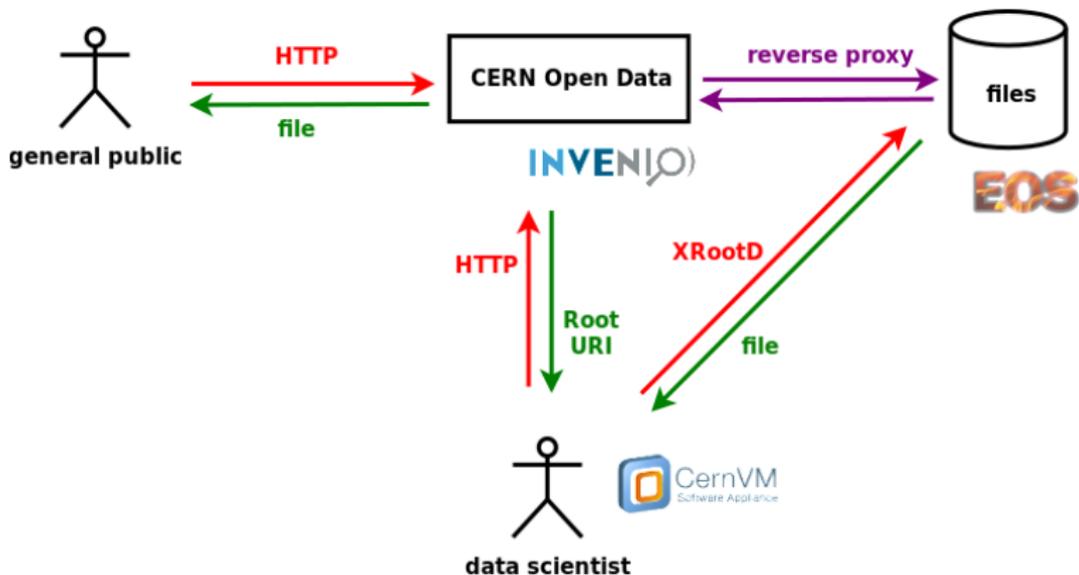
# First interesting analyses by theorists Jesse Thaler et al, MIT)

# Technology



Technology stack using **INVENIO** digital repository

# Data exposure



HTTP and XRootD access scenarios

# Open development

The screenshot shows a GitHub repository page for 'opendata CERN Open Data'. The navigation bar at the top includes 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. The repository name 'opendata CERN Open Data' is displayed prominently. Below the name, there are filters for 'Repositories', 'People', 'Teams', 'Projects', and 'Settings'. A section titled 'Pinned repositories' contains a card for 'opendata.cern.ch', which is described as 'Source code for the CERN Open Data portal' and has 238 stars and 82 forks. Below this, there is a search bar for repositories and filters for 'Type: All' and 'Language: All'. The main content area lists several repositories: 'opendata.cern.ch' (Source code for the CERN Open Data portal, 238 stars, 82 forks, updated 11 hours ago), 'data-curation' (Data ingestion and curation tools, 1 star, 3 forks, updated 7 days ago), and 'demobbed-viewer'. On the right side, there are two sidebars: 'Top languages' showing HTML and JavaScript, and 'People' showing a grid of 35 contributors.

<https://github.com/cernopendata>

# FAIR data principles

## ■ Findable

- persistent identifiers
- rich metadata
- indexed and searchable

## ■ Accessible

- retrievable by identifiers
- standard protocols
- metadata vs data accessibility

## ■ Interoperable

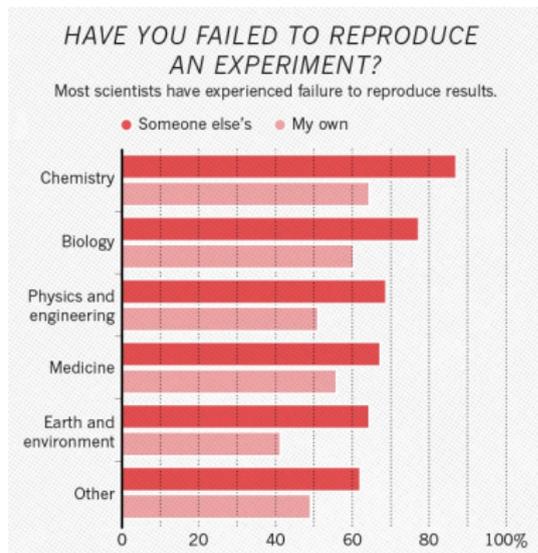
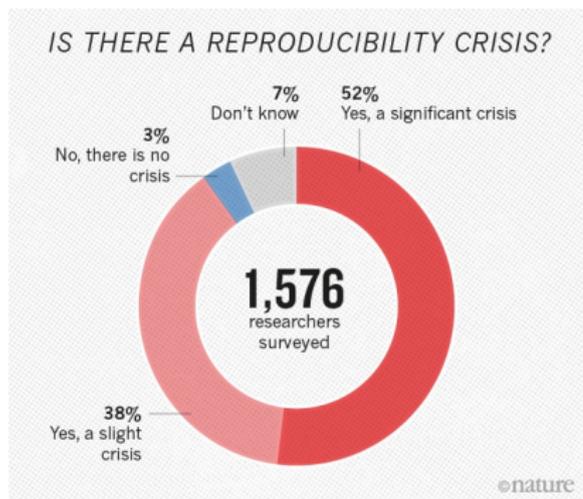
- knowledge representation language
- common vocabularies
- references to other metadata and data

## ■ Reusable

- domain-relevant attributes and community standards
- clear licensing
- provenance tracking

<https://www.nature.com/articles/sdata201618>

# Reusable and reproducible?



<https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>

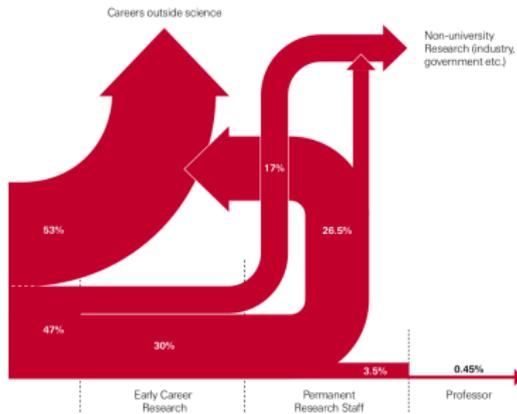
Half of researchers cannot reproduce their own experimental results

# Long-term value of knowledge?



## CMS collaboration

Experimental physics done by groups of  $\sim 3000$  physicists

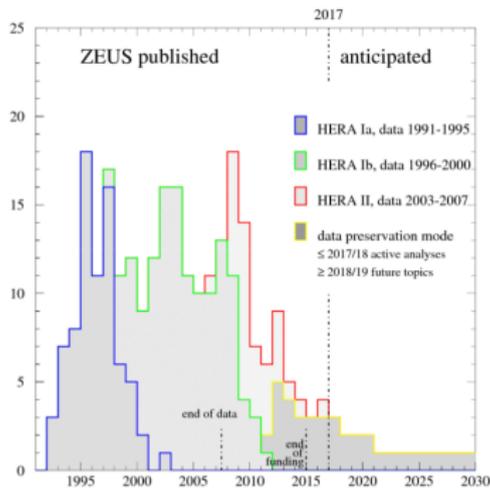


## Career after PhD

THE ROYAL SOCIETY

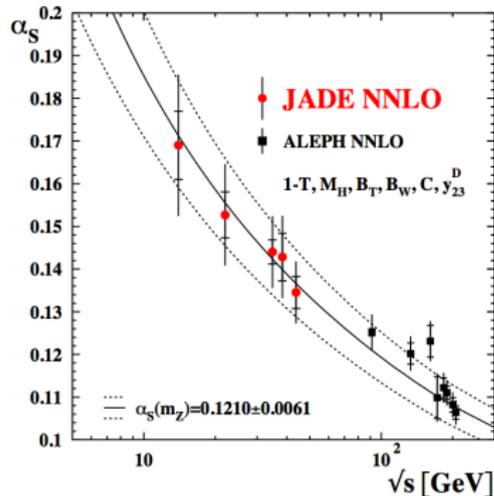
High turnover of young researchers

# Long-term value of data!



Achim Geiser <https://indico.cern.ch/event/588219>

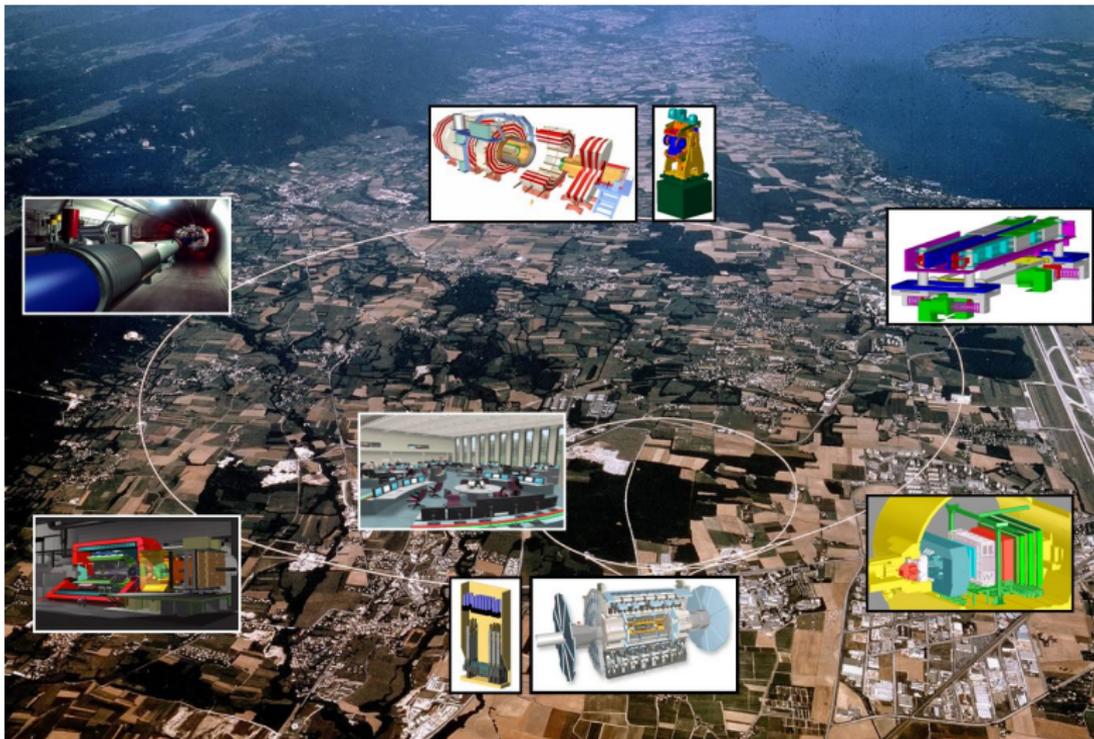
Collaborations publish papers even  $\sim 15$  years after data taking ends



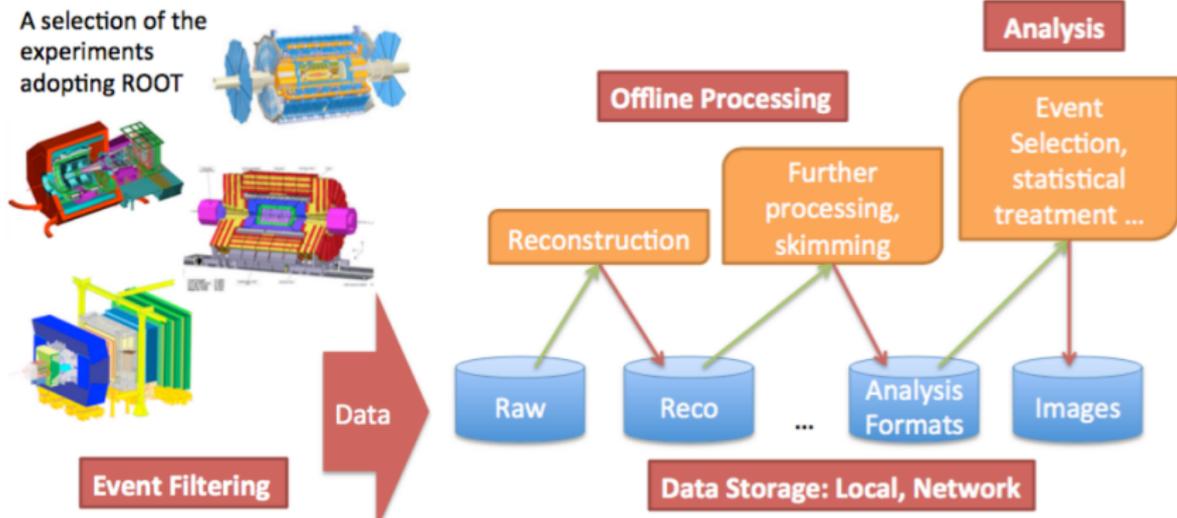
DPHEP <https://arxiv.org/abs/1205.4667>

JADE data (1979–1986) still unique even  $\sim 35$  years later

# CERN LHC Experiments



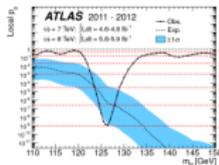
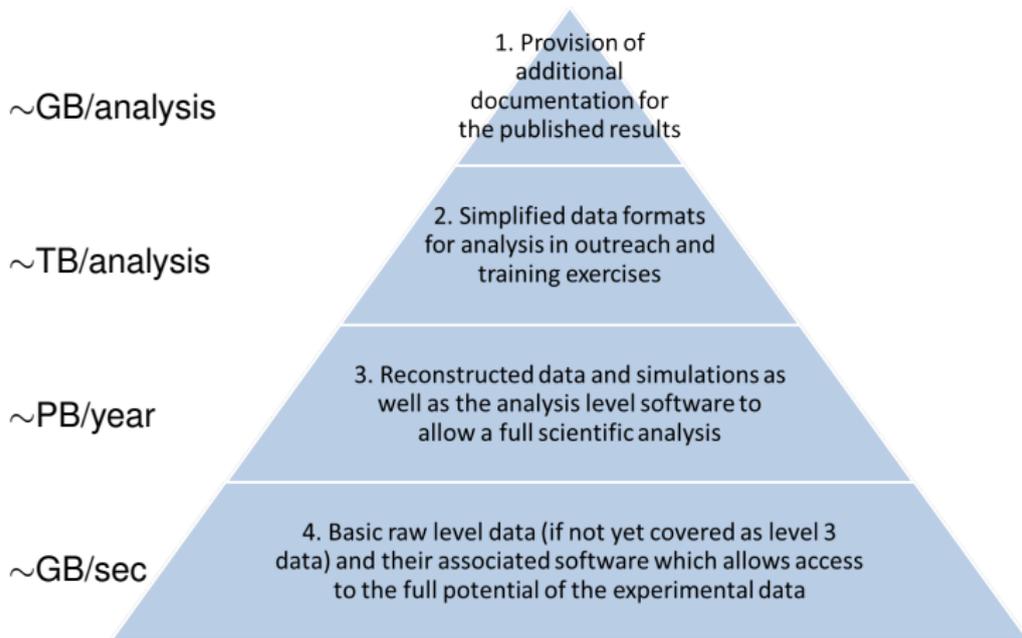
# HEP data analyses



D. Krücker *et al* <https://indico.desy.de/indico/event/18343>

Targeting both data production and data analysis stages

# Preserving data



analysis



Slicing through LHC data pyramid

# Preserving code

The image shows a composite of three screenshots illustrating the Zenodo workflow. On the left is the 'Settings' page for the repository 'instan-cern / decouple', where the 'Publish releases' toggle is turned ON. In the center is the 'Releases' page for the same repository, showing a list of releases with the latest one being v1.1.3. On the right is a 'Release' detail view for v1.1.3, displaying the release name, DOI (10.5281/zenodo.8345), and a .zenodo.json file with the following content:

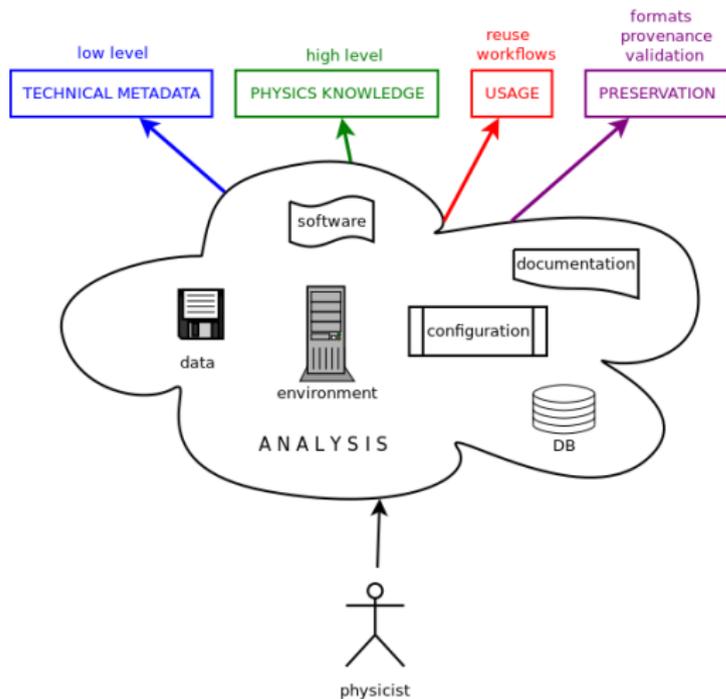
```
{
  "name": "Plein, T1lean",
  "affiliation": "Institut für Theoretische Ph",
  "description": "This repository contains the soft",
  "access_right": "open",
  "license": "mit-license",
  "related_identifiers": [
    {
      "identifier": "arXiv:1401.0000",
      "relation": "arXiv:1401.0000"
    }
  ]
}
```

Arrows indicate the flow from the 'Publish releases' toggle to the 'Releases' page, and from a specific release to its details, including the DOI and the .zenodo.json file.

<https://guides.github.com/activities/citable-code>

GitHub ↔ Zenodo bridge to automatically preserve releases

# Structuring analyses

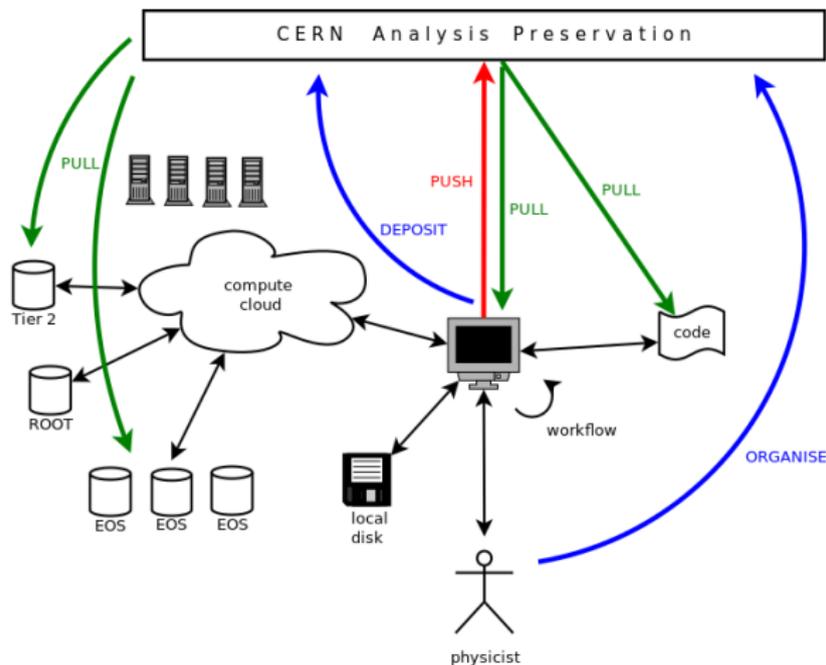


INVENIO

- JSON Schema
- W3C DCAT
- domain-specific fields

Structuring knowledge behind research data analysis

# Capturing analyses



INVENIO

- datasets:  
local storage,  
cloud storage
- software:  
Git, SVN
- information:  
DBs, TWiki,  
SharePoint
- protocols:  
HTTP, XRootD

Taking consistent snapshot of analysis assets at a certain time

# CERN Analysis Preservation

CERN Analysis Preservation main

Analysis Identifier: -

Share Save ...

### Preserve your analysis

Name it to distinguish it from your other drafts

Start Preserving

Submission Form

Basic Information  
Please provide some information relevant for all parts of the Analysis here

Stripping/Turbo selections [0 items]

ntuple/userDST-production [0 items]

User Analysis

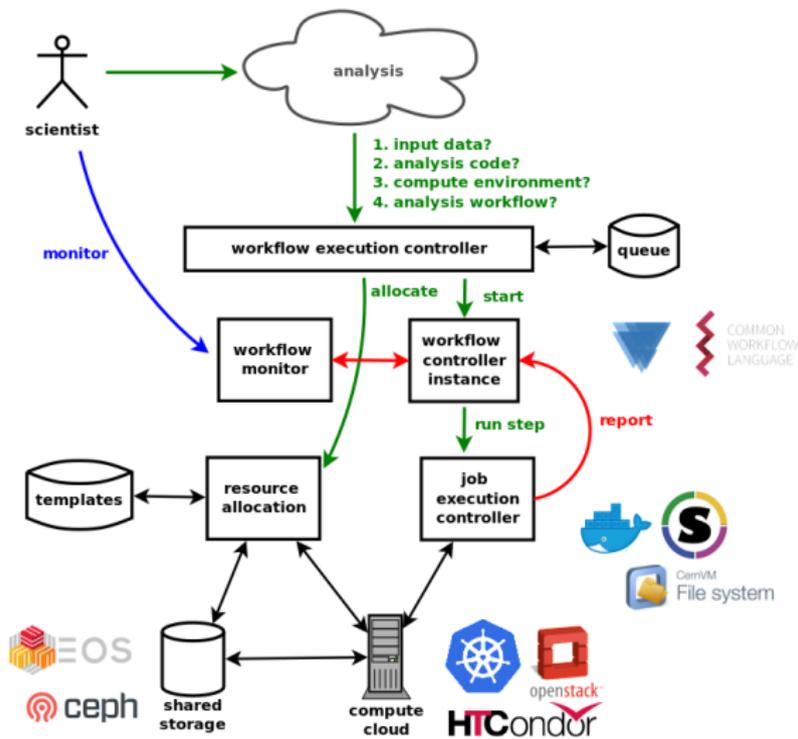
```
$ pip install cap-client
```

```
$ export CAP_SERVER_URL=https://analysispreservation.cern.ch/  
$ export CAP_ACCESS_TOKEN=<your generated access token from server>
```

```
$ cap-client files upload <file path> --pid/-p <existing pid>  
$ cap-client files upload file.json -p 89b593c498874ec8bcafc88944c458a7  
File uploaded successfully.
```

Web based and command-line based deposit workflows

# Reusing analyses



Reproduce analysis computations on containerised compute clouds

# Four questions

## 1 Input data

What is your input data?

- input files
- input parameters

## 3 Compute environment

What is your environment?

- operating system
- database calls

## 2 Analysis code

Which code analyses it?

- software frameworks
- user code

## 4 Analysis workflow

Which steps did you take?

- single command
- complex workflows

# Simple example

```
Region,1500,1600,1700,1750,1800,1850,1900,1950,1999,2008,2010,2012,2050,2150
World,100,100,100,100,100,100,100,100,100,100,100,100,100,100
Africa,18.8,19.7,15.5,13.4,10.9,8.8,8.1,8.8,12.8,14.5,14.8,15.2,19.8,23.7
Asia,53.1,58.4,63.9,63.5,64.9,64.1,57.4,55.6,60.8,60.4,60.4,60.3,59.1,57.1
Europe,18.3,19.1,18.3,20.6,20.8,21.9,24.7,21.7,12.2,10.9,10.7,10.5,7,5.3
Latin America and the Caribbean,8.5,1.7,1.5,2,2.5,3,4.5,6.6,8.5,8.6,8.6,8.6,9.1,9.4
Northern America,0.7,0.5,0.3,0.3,0.7,2.1,5.6,8.5,1.5,5,5.4,4.4,1
Oceania,0.7,0.5,0.4,0.3,0.2,0.2,0.4,0.5,0.5,0.5,0.5,0.5,0.5,0.5
```

## 1 input: CSV file

```
FROM centos:7
RUN yum install -y epel-release
RUN yum install -y \
    gcc \
    python-devel \
    python-pip
RUN pip install ipython==5.0.0 jupyter==1.0.0
ADD world_population_analysis.ipynb /code/
ADD World_historical_and_predicted_populations_in_percentage.csv /code/
WORKDIR /code
CMD ["jupyter", "nbconvert", "world_population_analysis.ipynb"]
```

## 3 environment: CentOS7, IP5

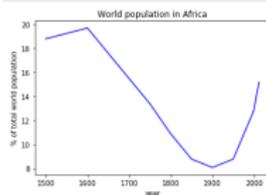
```
In [2]: # define input parameters
input_file = "../data/World_historical_and_predicted_populations_in_percentage.csv"
output_file = "../results/plot.png"
region = 'Africa'
year_min = 1500
year_max = 2012

In [3]: # read input data file
df = pd.read_csv(input_file)

In [4]: # add index
df = df.set_index("Region", drop=False)

In [5]: # select region and years based on input parameters
dfs = df.loc[region, str(year_min):str(year_max)]
dft = pd.DataFrame({'year': dfs.index.astype(int), 'percentage': dfs.values}, columns=['year', 'percentage'])

In [6]: # create output plot and save it to a file
plot = plt.plot(dft['year'], dft['percentage'], color='blue')
plt.title('World population in {}'.format(region))
plt.xlabel('year')
plt.ylabel('% of total world population')
plt.savefig(output_file)
```



## 2 code: Jupyter notebook

## 4 workflow: papermill ...

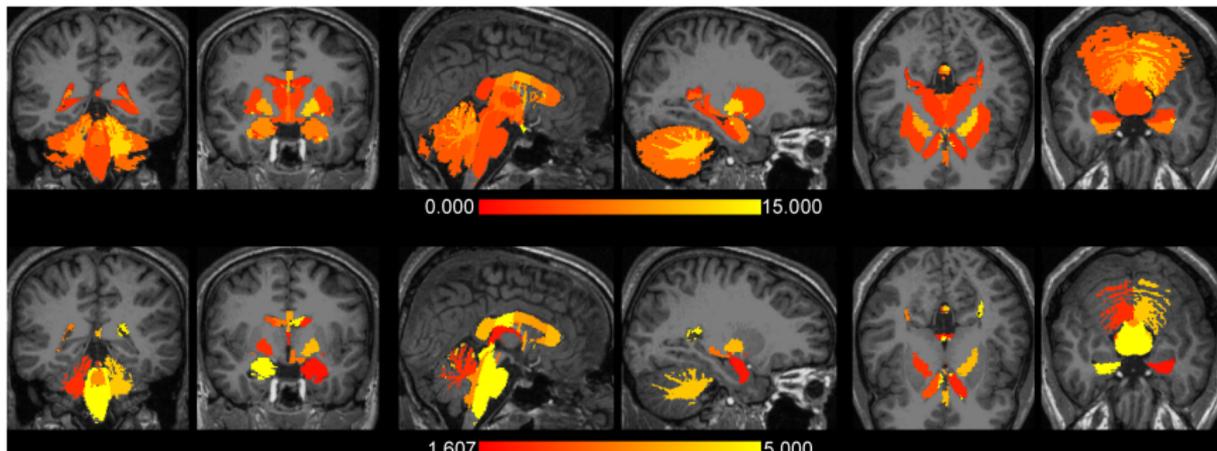
 <https://github.com/reanahub/reana-demo-worldpopulation>

# Example from medicine

## The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

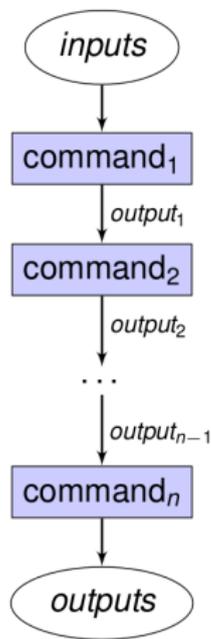
Ed H. B. M. Gronenschild , Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: June 1, 2012 • DOI: 10.1371/journal.pone.0038234

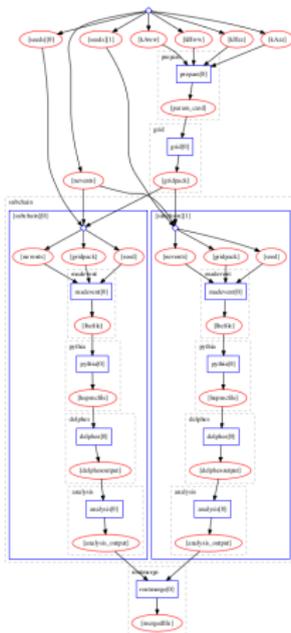


$8.8 \pm 6.6\%$  (volume) and  $2.8 \pm 1.3\%$  (cortical thickness)

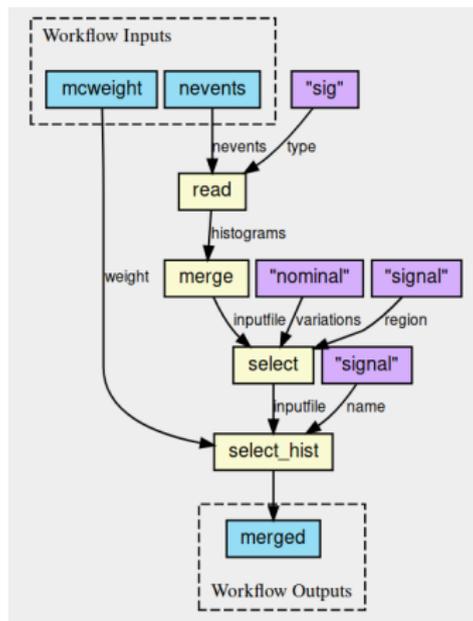
# Computational workflows



Serial

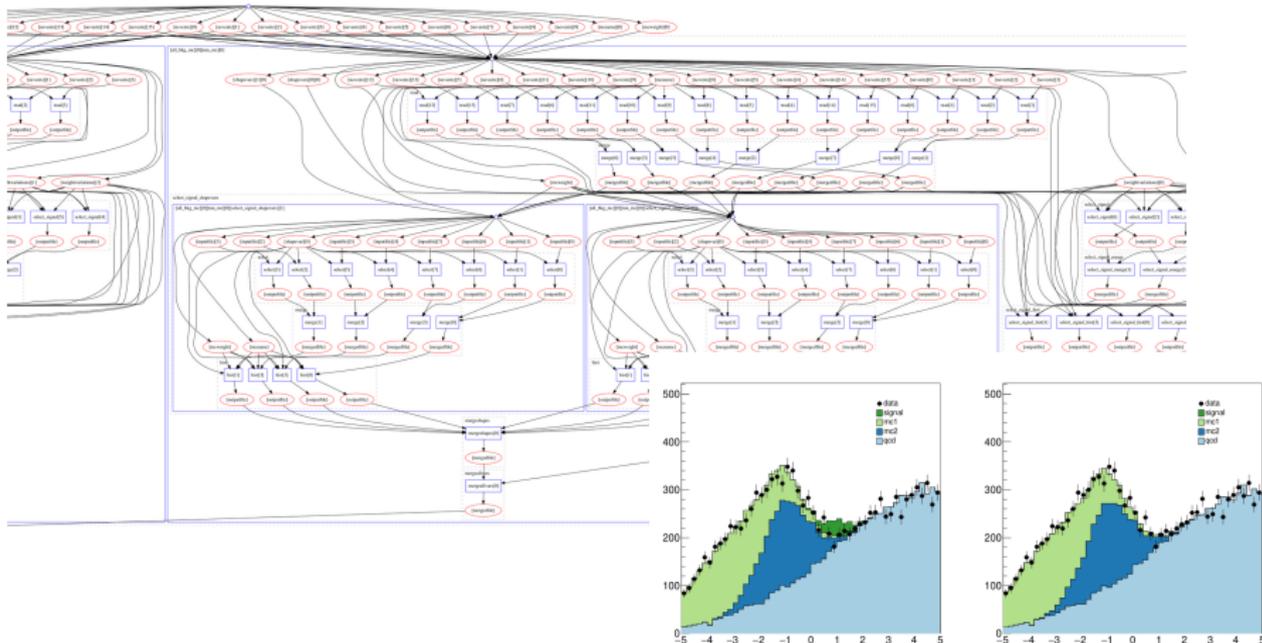


Yadage



CWL

# Example: Beyond Standard Model



<https://github.com/reanahub/reana-demo-bsm-search/>

Complex computational workflows typical in particle physics analyses

# REANA Reusable Analyses

# reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



Reusable

Containerise once, reuse elsewhere. Cloud-native.



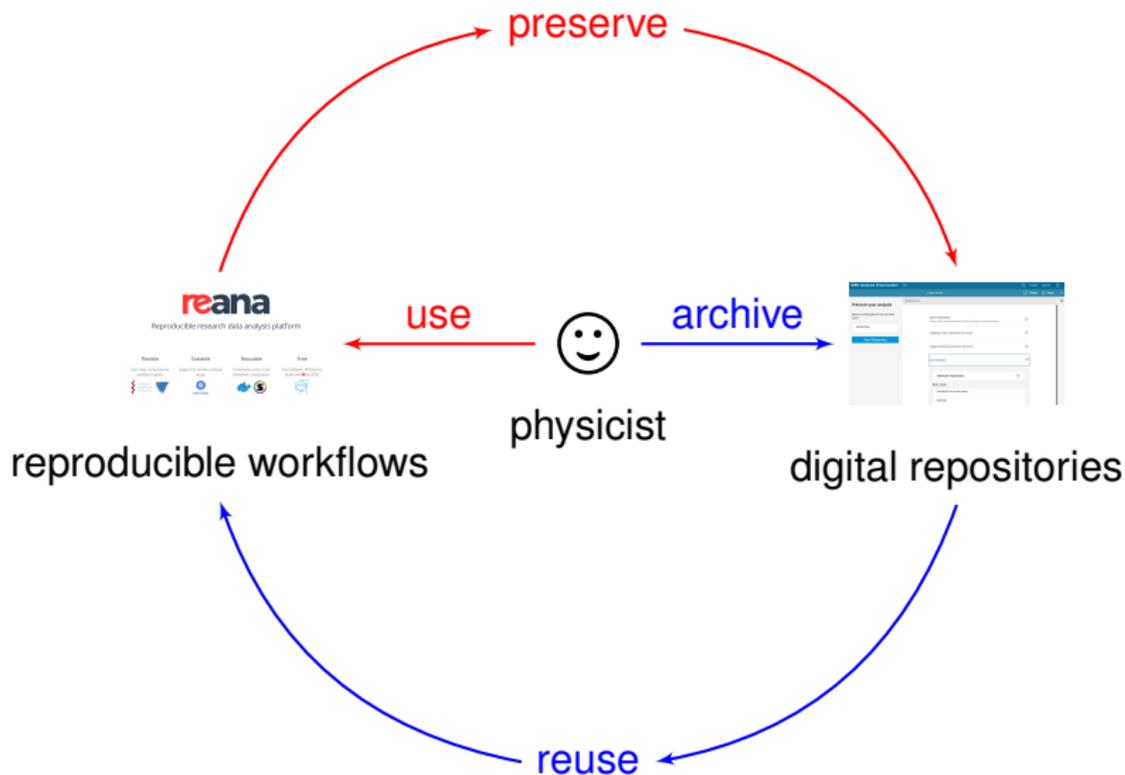
Free

Free Software. MIT licence. Made with ❤️ at CERN.



Reproducible and reusable research data analysis platform

# Reproducibility $\rightleftharpoons$ Preservation



# “Open is not enough”

nature  
physics

PERSPECTIVE

<https://doi.org/10.1038/nph157-018-0342-2>

OPEN

## Open is not enough

Xiaoli Chen<sup>1</sup>, Sünje Dalmeier-Tiessen<sup>2</sup>, Robin Dasler<sup>1</sup>, Sebastian Feger<sup>1</sup>, Pamfilos Fokianos<sup>3</sup>, Jose Benit González<sup>4</sup>, Heri Hirvonen<sup>5</sup>, Dinos Kourdis<sup>6</sup>, Artemis Lavasa<sup>7</sup>, Salvatore Mele<sup>8</sup>, Diego Rodríguez Rodríguez<sup>9</sup>, Viktor Simik<sup>10</sup>, Tim Smith<sup>11</sup>, Ana Trisovic<sup>12</sup>, Anna Trzcinska<sup>13</sup>, Ioannis Tsanaktsidis<sup>14</sup>, Markus Zimmermann<sup>15</sup>, Kyle Cramer<sup>16</sup>, Lukas Heinrich<sup>17</sup>, Gordon Watts<sup>18</sup>, Michael Hildner<sup>19</sup>, Laraloret Iglesias<sup>19</sup>, Kati Lassitz-Perini<sup>19</sup> and Sebastian Neubert<sup>20</sup>

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be followed more widely. This first perspective suggests that reproducibility requires going beyond open access.

Open science and reproducible research have become popular goals across research communities, political circles and funding bodies<sup>1</sup>. The understanding is that open and reproducible research practices enable scientific reuse, accelerating future projects and discoveries to our discipline. In the struggle to take concrete steps in practice these aims have been much discussed and more rarely achieved, often accompanied by a push to make research products and scientific results open quickly.

Although there are laudable and necessary first steps, they are not sufficient to bring about the transformations that would allow us to reap the benefits of open and reproducible research. It is time to move beyond the rhetoric and the quest for quick fixes and start developing and implementing tools to power a more profound change.

Our own experience from opening up vast volumes of data at CERN opens some thoughts to be followed at the end of the scientific endeavour. In addition, open access does not guarantee reproducibility or reliability, so it should not be pursued as a goal in itself. Focusing data to do so enough it needs to be accompanied by software, workflows and explanations, all of which need to be captured throughout the entire life cycle of research. We argue for a steady open science as the result.

Our own experience from opening up vast volumes of data at CERN opens some thoughts to be followed at the end of the scientific endeavour. In addition, open access does not guarantee reproducibility or reliability, so it should not be pursued as a goal in itself. Focusing data to do so enough it needs to be accompanied by software, workflows and explanations, all of which need to be captured throughout the entire life cycle of research. We argue for a steady open science as the result.

steps for reproducible and reusable research more widely in other scientific disciplines.

### Approaching reproducibility and reuse in HEP

To set the stage for the rest of this piece, we first contrast a more recent spectrum in which to place the various challenges facing HEP, allowing us to better frame our ambitions and solutions. We choose to build on the descriptions introduced by Gerde Gebke<sup>2</sup> and Lorenz A. Barba shown in Table 1.

These concepts assume a research environment in which multiple labs have the equipment necessary to duplicate an experiment, which essentially makes the experiments portable. In the particle physics context, however the intrinsic cost and complexity of the experimental set-up essentially make the independent and complete replication of HEP experiments unfeasible and unhelpful. HEP experiments are set up with unique capabilities, often being the only facility or instrument of their kind in the world; they are also extremely high upgrated to satisfy requirements for higher energy precision and level of accuracy. The experiments at the Large Hadron Collider (LHC) are prominent examples. It is this uniqueness that makes the experimental data valuable for preservation so that it can be later reused with other measurements for comparison, confirmation or inspiration.

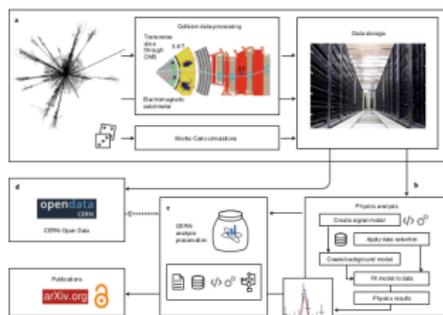
Our considerations here mainly begin after gathering the data. This means that we are more concerned with regarding verifying the computational analysis performed over a given dataset rather than with data collection. Therefore, in Table 2 we present a variation of the idea of reaching goals that have been assessed a great extent to which implemented set-up actions in the implementation of a computational analysis of a defined dataset, and a ‘lab’ can be thought of as an experimental collaboration or an analysis group. In the case of computational processes, analysts analyse themselves an increasingly complex flow to the large data volume and algorithms involved<sup>3</sup>. In practice, the physicist typically study more than one physics problem and compare data collected under different running conditions through comprehensive documentation of the analysis methods to maximize the discovery of the software implementations often leads to more critical details.

<sup>1</sup>CERN, Geneva, Switzerland; <sup>2</sup>Sheffield University, Sheffield, UK; <sup>3</sup>Stuttgart University, Stuttgart, Germany; <sup>4</sup>National Institute of Physics, Frascati, Italy; <sup>5</sup>Finland Centre for Physics, Centre for IT, NUTU, New York, NY, USA; <sup>6</sup>University of Washington, Seattle, WA, USA; <sup>7</sup>University of Bonn, Bonn, Germany; <sup>8</sup>IN2P3-CNRS, Université de Clermont, Clermont-Ferrand, France; <sup>9</sup>University of Valencia, Valencia, Spain; <sup>10</sup>University of Cambridge, Cambridge, UK; <sup>11</sup>University of Edinburgh, Edinburgh, UK; <sup>12</sup>University of Bonn, Bonn, Germany; <sup>13</sup>University of Bonn, Bonn, Germany; <sup>14</sup>University of Bonn, Bonn, Germany; <sup>15</sup>University of Bonn, Bonn, Germany; <sup>16</sup>University of Bonn, Bonn, Germany; <sup>17</sup>University of Bonn, Bonn, Germany; <sup>18</sup>University of Bonn, Bonn, Germany; <sup>19</sup>University of Bonn, Bonn, Germany; <sup>20</sup>University of Bonn, Bonn, Germany

NATURE PHYSICS | [www.nature.com/naturephysics](https://doi.org/10.1038/nph157-018-0342-2)

PERSPECTIVE

NATURE PHYSICS



**Fig. 1** | Datacentricity in LHC experiments. a, The experimental data from proton-proton collisions in the large Hadron Collider are being collected by particle detectors such as the experimental collaborations ALICE, ATLAS, CMS and LHCb. These experimental data are further filtered and processed to give the collision dataset formally that are suitable for physics analysis. In parallel, the computer simulations are being run in order to provide necessary comparison of experimental data with theoretical predictions. b, The standard software and simulated data are then released for individual physics analyses. A physicist may perform further data reduction and selective procedures, which are followed by a statistical analysis on the data. Physics results are derived taking into account statistical and systematic uncertainties. The results often summarise which theoretical models best describe the data that are consistent with the observations once background uncertainties have been included. The analysis results being used by the individual researchers include the information about the software and simulated datasets, the detector conditions, the analysis code, the computational environment, and the experimental workflow used by the researcher to derive the histogram and the final plots as they appear in publications. c, The CERN Analysis Preservation service captures all the analysis events and related documentation (such as input and output products), so that the analysis knowledge and data are preserved as a shared long-term digital repository for re-use and research purposes. d, The CERN Open Data service publishes selected data sets generated by the LHC collaborations into the public domain after an embargo period of several years depending on the collaboration data management plans and preservation policies. Credit: CERN. <https://doi.org/10.1038/nph157-018-0342-2>; <https://arxiv.org>; <https://www.cern.ch>; <https://open.data.cern.ch>

### Table 1 | Terminology related to reproducible research

introduced by Gerde Gebke and Lorenz A. Barba

Term	Purpose	Description
Reprint	Robot	Iterations on experiment and set-up, produced as a new lab
Repeat	Defined	Same experiment, same set-up, same lab
Replicate	Copy	Same experiment, same set-up, independent lab
Reproduce	Compare	Iterations on experiment and set-up, independent lab
Reuse	Transfer	Different experiment

potentially leading to a loss of knowledge concerning how the results were obtained.<sup>4</sup> In absence of solutions for analysis capture and preservation, knowledge of specific methods and how they are applied to a given physics analysis might be lost. In fact, there currently are specific challenges, a culture effort coordinated by CERN, but including the wider community, has emerged, initiating a series of projects, some of which are described below.

**Reuse and preservation.** The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by statistical error processors. Because these reviews often involve the input of the entire collaboration, where the level of cross-checking is extensive, the re-experiments are considered to be trustworthy.

NATURE PHYSICS | [www.nature.com/naturephysics](https://doi.org/10.1038/nph157-018-0342-2)

<https://www.nature.com/articles/s41567-018-0342-2.pdf>

# Conclusions



CERN Open Data



CERN Analysis Preservation



REANA Reusable Analyses

*“Capturing, preserving and sharing FAIR data and actionable knowledge behind particle physics data analyses in order to facilitate future data reuse”*

**CERN IT** D. Kousidis, R. Maciulaitis, J. Okraska, D. Rodriguez, T. Šimko · **CERN SIS** S. Dallmeier-Tiessen, S. Feger, P. Fokianos, A. Lavasa, S. van de Sandt, I. Tsanaksidis, A. Trzcinska · **ALICE** Y. Foka, M. Gheata, C. Grigoras, M. Zimmermann · **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** H. Bittencourt, A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, S. Roiser, A. Trisovic · **OPERA** G. De Lellis, S. Dmitrievsky · **CERN CernVM** J. Blomer · **CERN EOS** L. Mascetti, H. Rousseau · **CERN Kubernetes** R. Rocha · **CERN OpenShift** A. Lossent, A. Peon

# References



## CERN Open Data

-  <http://opendata.cern.ch>
-  <http://github.com/cernopendata>
-  [cernopendata](#)



## CERN Analysis Preservation

-  <http://analysispreservation.cern.ch>
-  <http://github.com/cernanalysispreservation>
-  [analysispreserv](#)



## REANA

-  <http://www.reanahub.io>
-  <http://github.com/reanahub>
-  [reanahub](#)



## Invenio

-  <http://inveniosoftware.org>
-  <http://github.com/inveniosoftware>
-  [inveniosoftware](#)



## Zenodo

-  <https://zenodo.org>
-  <http://github.com/zenodo>
-  [zenodo\\_org](#)