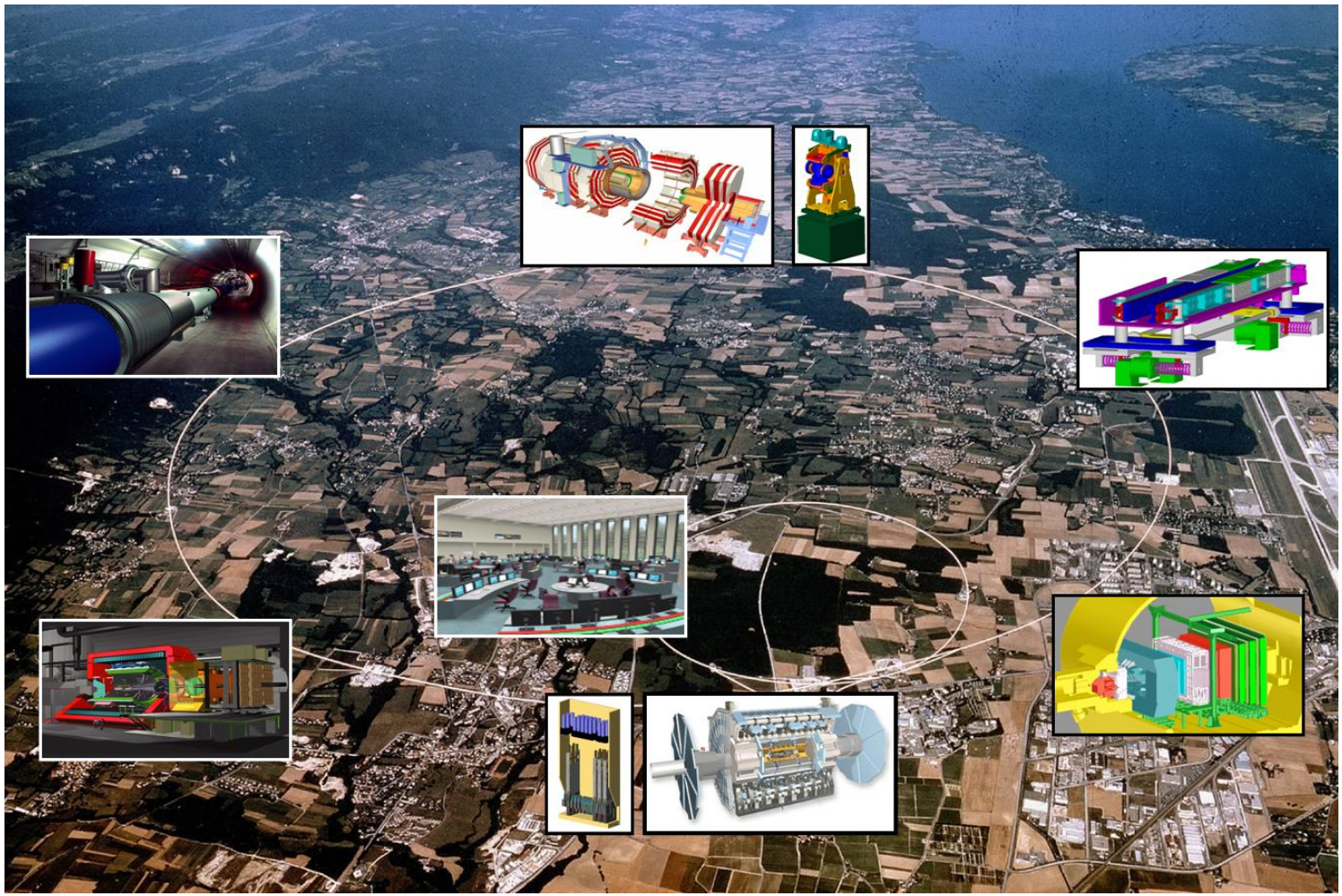

# Open is not enough: fostering reproducible and reusable particle physics research

Tibor Šimko

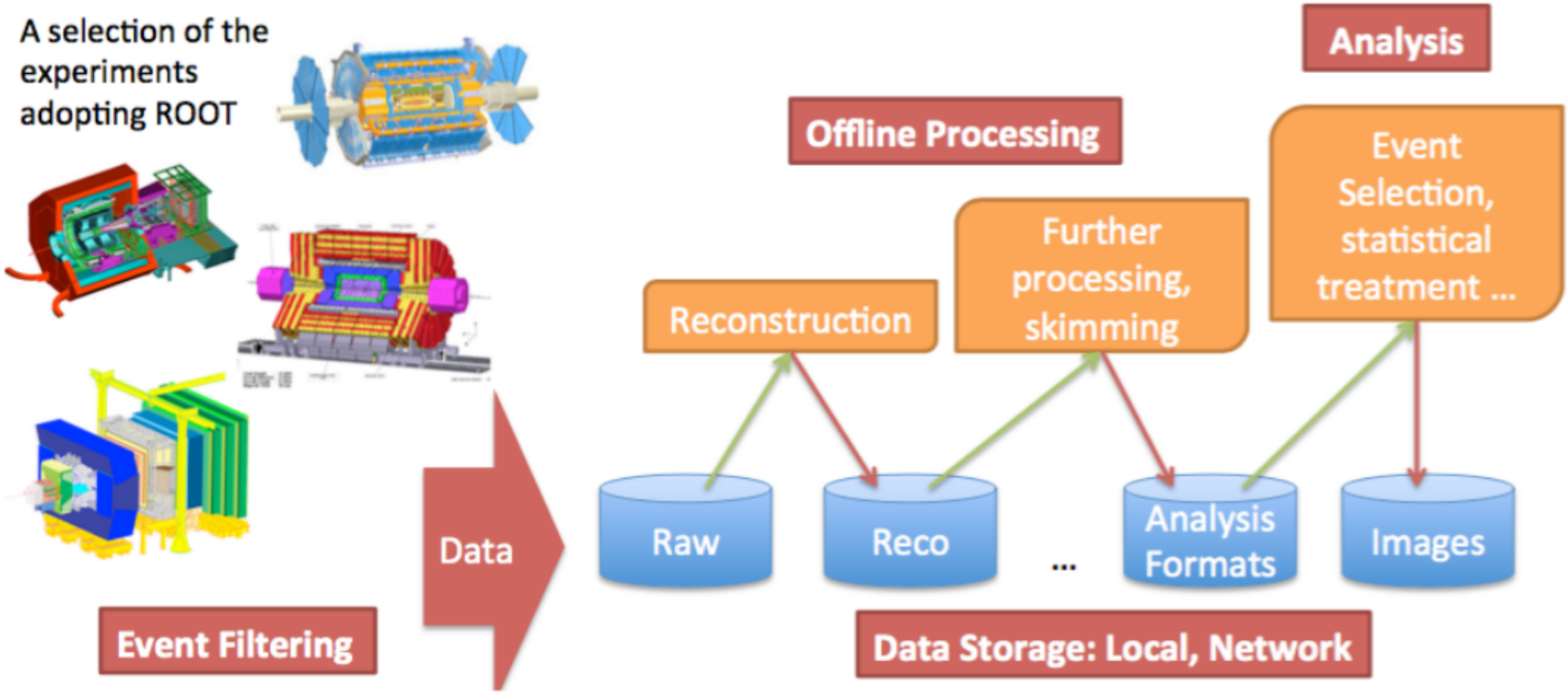
CERN

# CERN Large Hadron Collider



Large Hadron Collider with ATLAS, ALICE, CMS, LHCb detectors

# HEP data analyses




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

Typical data chain from acquisition through analysis to plots

# LHC data pyramid

~KB/paper

~GB/analysis

~TB/analysis

~PB/year

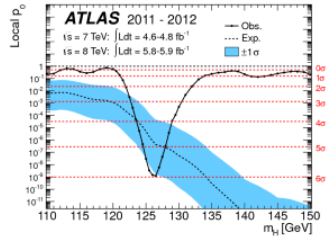
~GB/sec

1. Provision of additional documentation for the published results

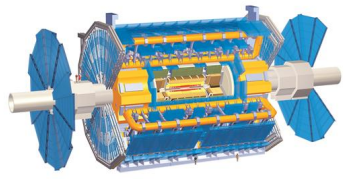
2. Simplified data formats for analysis in outreach and training exercises

3. Reconstructed data and simulations as well as the analysis level software to allow a full scientific analysis

4. Basic raw level data (if not yet covered as level 3 data) and their associated software which allows access to the full potential of the experimental data

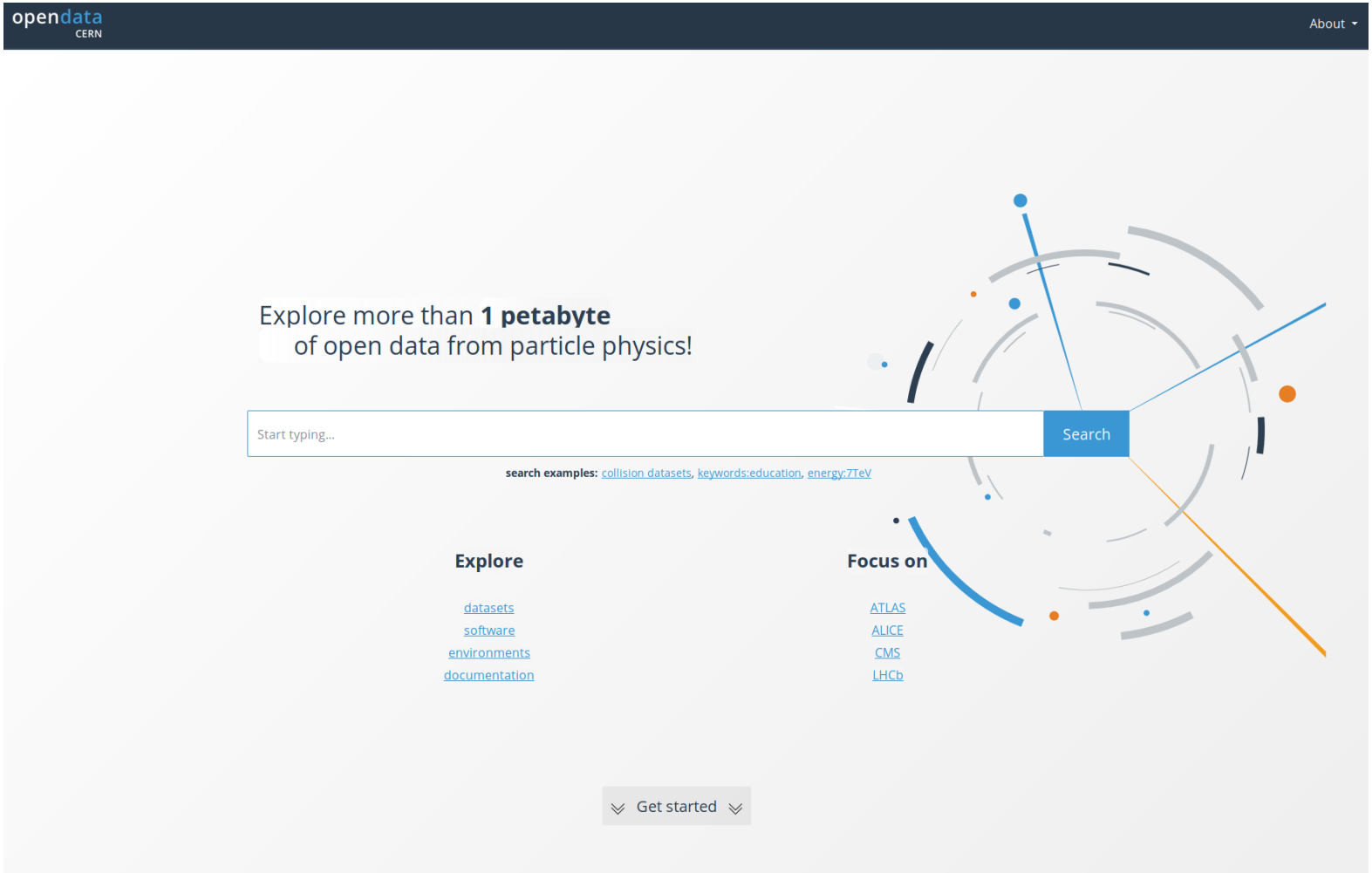



analysis



Four data levels for capture, preservation and opening

# CERN Open Data portal




http://opendata.cern.ch/

# LHC collaboration data policies

ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community. These data embed unique scientific information for the in depth understanding of the profound nature and origin of matter. Because of their uniqueness, long term preservation must be an essential objective of the data processing framework and will lay the foundations of the ALICE Collaboration legacy to the scientific community as well as to the general public. These considerations call for a detailed assessment of the ALICE data preservation strategy and policy. Documentation, long term preservation at various levels of abstraction, data access and analysis policy and software availability constitute the key elements of such a data preservation strategy allowing future collaborators, the wider scientific community and the general public to analyze data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the redaction addressed by the ALICE data preservation policy.

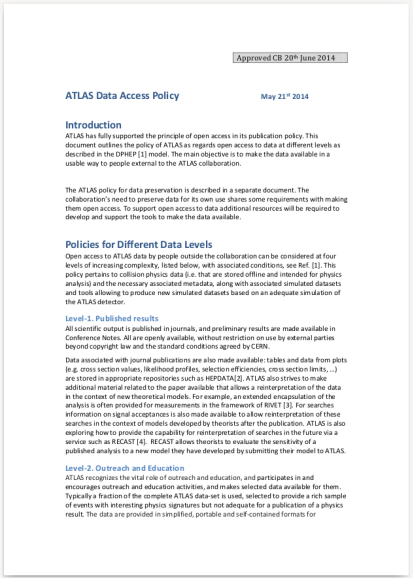
ALICE data formats

The level of abstraction of ALICE data increases at every step of the data processing chain starting from basic raw data delivered by the detectors of the experiment evolving into physics-analysis-ready data and ending with physics data suitable for publication. At each stage of the data processing ancillary meta-data, such as calibration, alignment and running condition parameters are invoked to transform raw detector information into physics information, free of detector biases. The various ALICE data formats are classified as follows:

a) *Raw data* embedding the signals delivered by the detectors along with the associated status data containing various information on the running conditions constitutes the primary information collected by the ALICE experiment. They provide the input of the reconstruction algorithm, together with the calibration data stored in a dedicated database;
b) *Monte-Carlo data*, including data at the event generator level (MC truth) and data mimicking the raw data format (digits), anchored to real data reproducing the running conditions;
c) *Event Summary Data* (ESD) produced by the reconstruction algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific information allowing a full physics analysis;
d) General purpose *Analysis Object Data* (AOD), derived from ESD data. The AOD data format contains a simplified event model with few additional high level detector specific parameters;
e) *Custom analysis object data*, used standalone or together with the general purpose AOD for specific analysis;
f) *Published physics results* and highly abstracted data resulting from the analysis.

These different formats of the ALICE data lead to a specific schema for data preservation. While formats can change with time, the collaboration provides software releases suitable to read and process any format, or alternatively to migrate data from one format to another. Since processed data can exist in several versions, only the version used for the final publication of the results is considered as a candidate for data preservation.

The ALICE Computing Model includes the provision for permanent storage of two copies of the raw data. They are not presently being considered for open access, but they can be reprocessed at any time by members of the ALICE collaboration upon approval by the ALICE Physics Board. The original datasets used to produce published results, together with the adequate software version (framework and macros) are subject to long-term preservation.



Approved CB 20th June 2014

ATLAS Data Access Policy May 21st 2014

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS as regards open access to data at different levels as described in the DPHEP [1] model. The main objective is to make the data available in a usable way to people external to the ATLAS collaboration.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them open access. To support open access to data additional resources will be required to develop and support the tools to make the data available.

Policies for Different Data Levels

Open access to ATLAS data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions, see Ref. [1]. This policy pertains to collision physics data (i.e. that are stored offline and intended for physics analysis) and the necessary associated metadata, along with associated simulated datasets and tools allowing to produce new simulated datasets based on an adequate simulation of the ATLAS detector.

Level-1. Published results

All scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, ...) are stored in appropriate repositories such as HEPDATA[2]. ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data in the context of new theoretical models. For example, an extended encapsulation of the analysis is often provided for measurements in the framework of RIVET [3]. For searches information on signal acceptances is also made available to allow reinterpretation of these searches in the context of models developed by theorists after the publication. ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as RECAST [4]. RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

Level-2. Outreach and Education

ATLAS recognizes the vital role of outreach and education, and participates in and encourages outreach and education activities, and makes selected data available for them. Typically a fraction of the complete ATLAS data-set is used, selected to provide a rich sample of events with interesting physics signatures but not adequate for a publication of a physics result. The data are provided in simplified, portable and self-contained formats for

Version 1.2
Approved by the CMS Collaboration Board 20th April 2018

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 level 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.

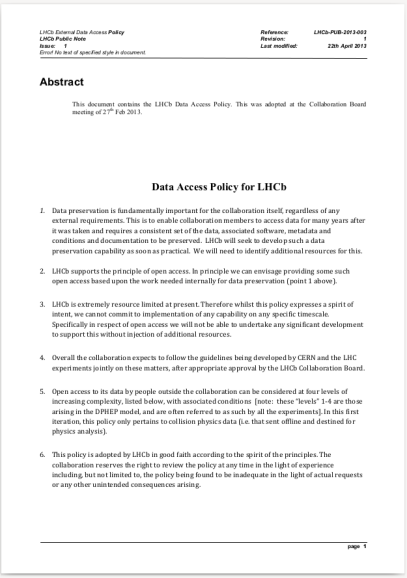
CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their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

This policy describes the CMS principles of data preservation, re-use and open access, as well as the relevant actors in all these tasks and their roles and responsibilities. CMS understands that in order to fully exploit all these re-use opportunities, immediate and continued resources are needed. The level of support that CMS will be able to provide to external users depends on the available funding. This policy addresses the moral responsibility of CMS for its data, as well as the increasing concern of funding agencies worldwide and the civil society for the preservation and re-use of scientific data.

Notwithstanding the long-term perspective of the LHC programme, the time for action is now: lower-energy and lower-luminosity LHC runs at centre-of-mass energies of 0.9, 2.36, 2.76, 7 and 8 TeV may never be repeated, and their preservation and preparation for later re-use, has to be addressed urgently. Meeting this challenge is a unique way to stress-test and evaluate the entire preservation, re-use and open access concept for the CMS data.

CMS data take many forms. Starting from either raw experimental or simulated data through to reconstructed data and the datasets of higher abstraction generated by analysis workflows, and finally all the way to data represented in scientific publications. Each of these layers has the potential to afford different opportunities for long-term re-use and poses different challenges for preservation. Data represented in publications can already be preserved by building on the existing practices of the Collaboration (e.g. open access publishing) and existing third-party platforms (e.g. INSPIRE[2]), simply expanding the concept of publication to include additional data sets of a high level of abstraction. At the other extreme of the spectrum, closer to the raw data, different challenges appear which imply a paradigm shift from in-depth documenting and archiving of analyses during the publication process, to a preservation of reconstruction and simulation software packages with all of their dependencies.

[1] A. Holzner, P. Igo-Kemenes, S. Mele, "First results from the RAISE:insight project: HEP survey on data preservation, re-use and (open) access" http://arxiv.org/abs/0906.0485
[2] http://inspirehep.net



LHCb External Data Access Policy
LHCb Public Note
Issue: 1
Error! No text of specified style in document.

Reference: LHCb-PUB-2013-003
Revision: 1
Last modified: 22th April 2013

Abstract

This document contains the LHCb Data Access Policy. This was adopted at the Collaboration Board meeting of 27th Feb 2013.

Data Access Policy for LHCb

1. Data preservation is fundamentally important for the collaboration itself, regardless of any external requirements. This is to enable collaboration members to access data for many years after it was taken and requires a consistent set of the data, associated software, metadata and conditions and documentation to be preserved. LHCb will seek to develop such a data preservation capability as soon as practical. We will need to identify additional resources for this.
2. LHCb supports the principle of open access. In principle we can envisage providing some such open access based upon the work needed internally for data preservation (point 1 above).
3. LHCb is extremely resource limited at present. Therefore whilst this policy expresses a spirit of intent, we cannot commit to implementation of any capability on any specific timescale. Specifically in respect of open access we will not be able to undertake any significant development to support this without injection of additional resources.
4. Overall the collaboration expects to follow the guidelines being developed by CERN and the LHC experiments jointly on these matters, after appropriate approval by the LHCb Collaboration Board.
5. Open access to its data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions [note: these "levels" 1-4 are those arising in the DPHEP model, and are often referred to as such by all the experiments]. In this first iteration, this policy only pertains to collision physics data (i.e. that sent offline and destined for physics analysis).
6. This policy is adopted by LHCb in good faith according to the spirit of the principles. The collaboration reserves the right to review the policy at any time in the light of experience including, but not limited to, the policy being found to be inadequate in the light of actual requests or any other unintended consequences arising.

page 1

ALICE

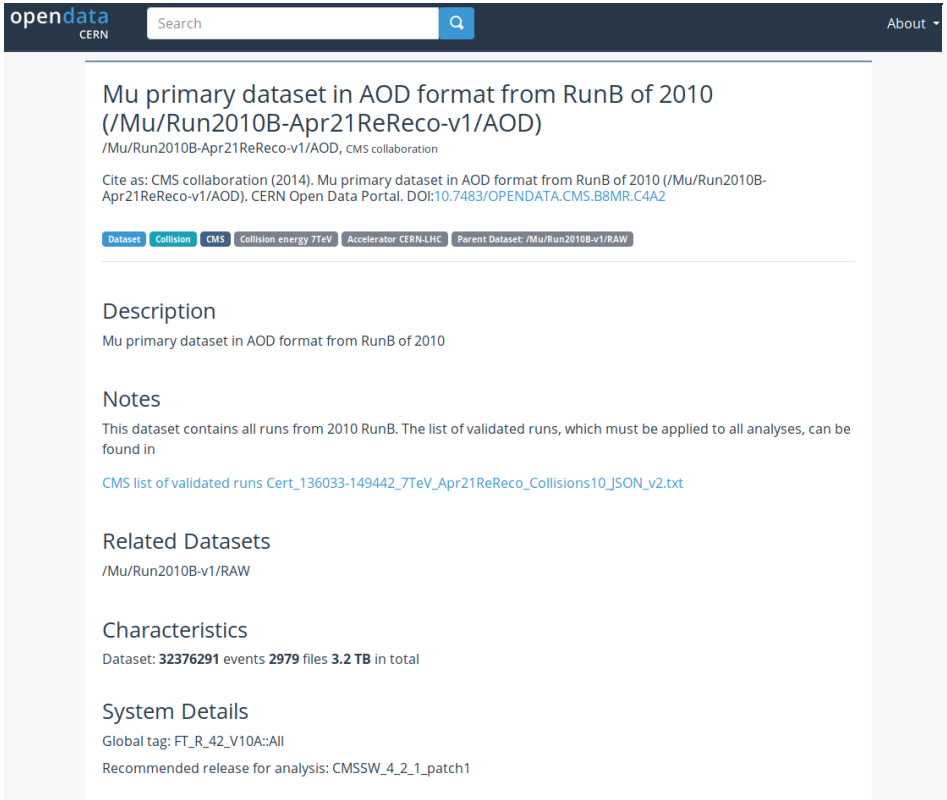
ATLAS EXPERIMENT

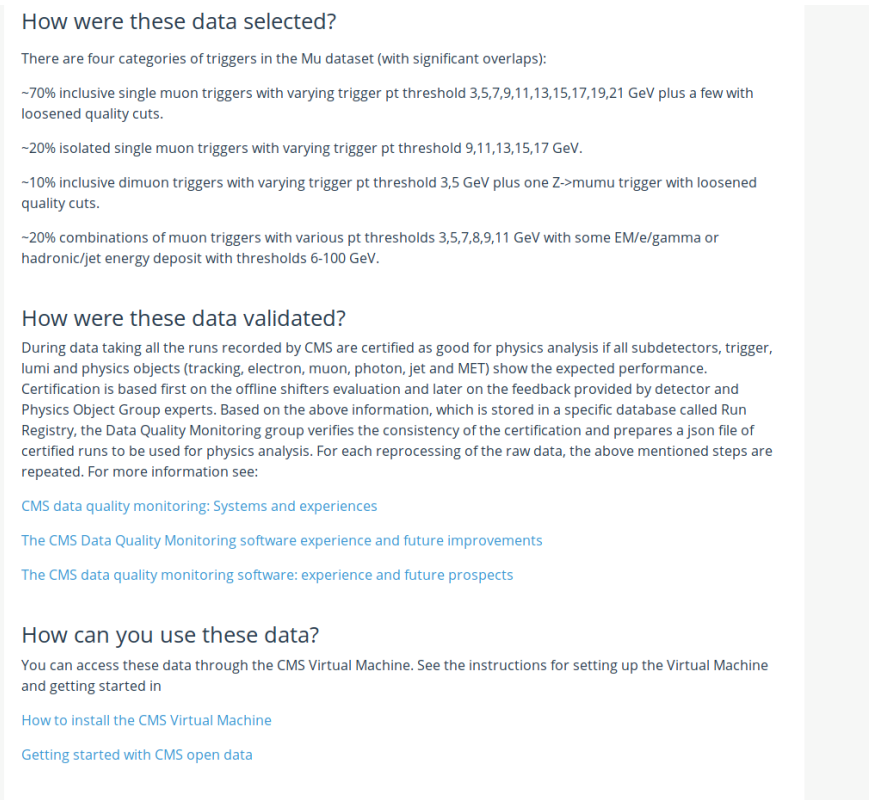
CMS

LHCb

Restricted data → embargo period (∼5 years) → open data

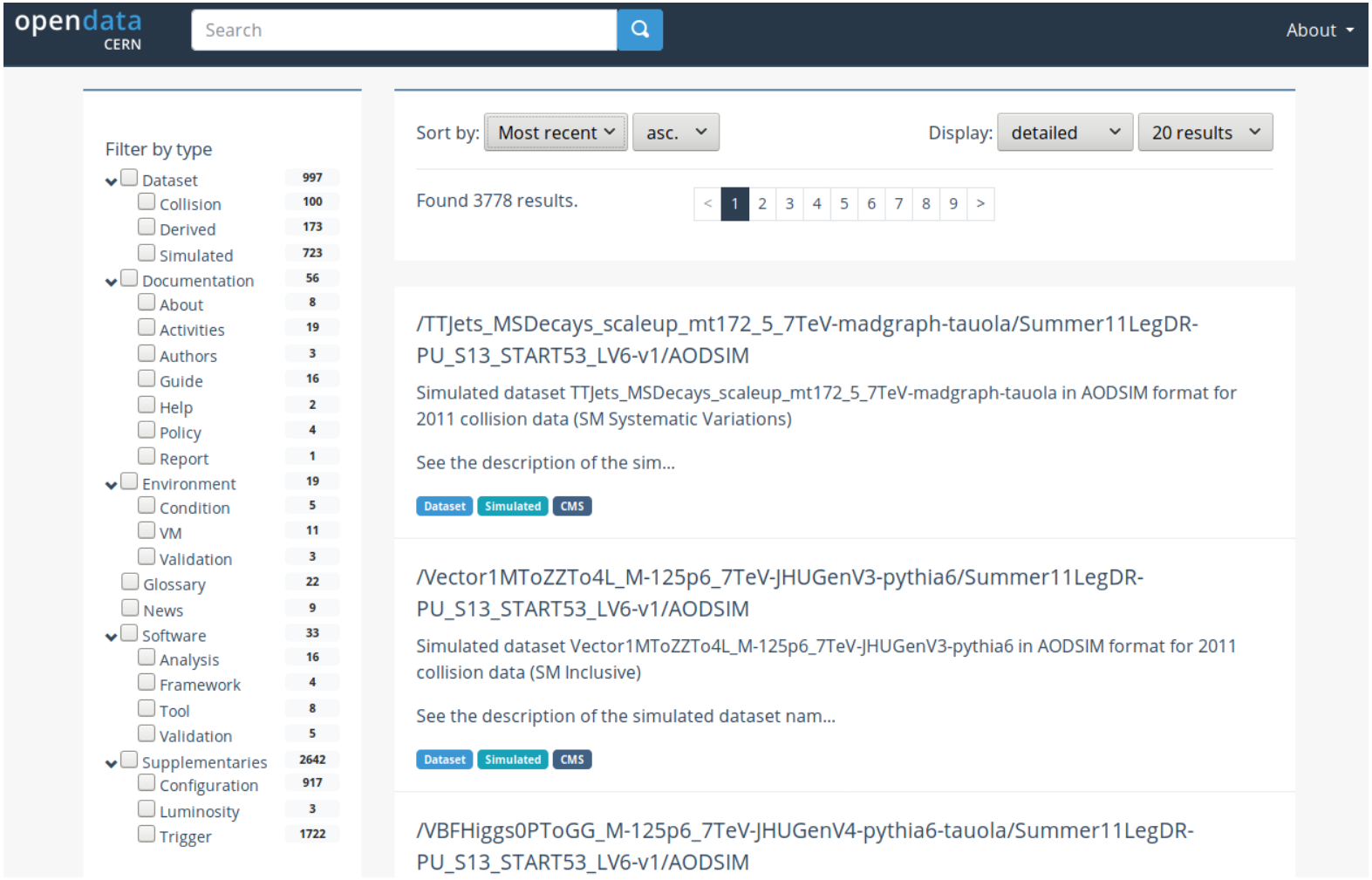
# Information organisation



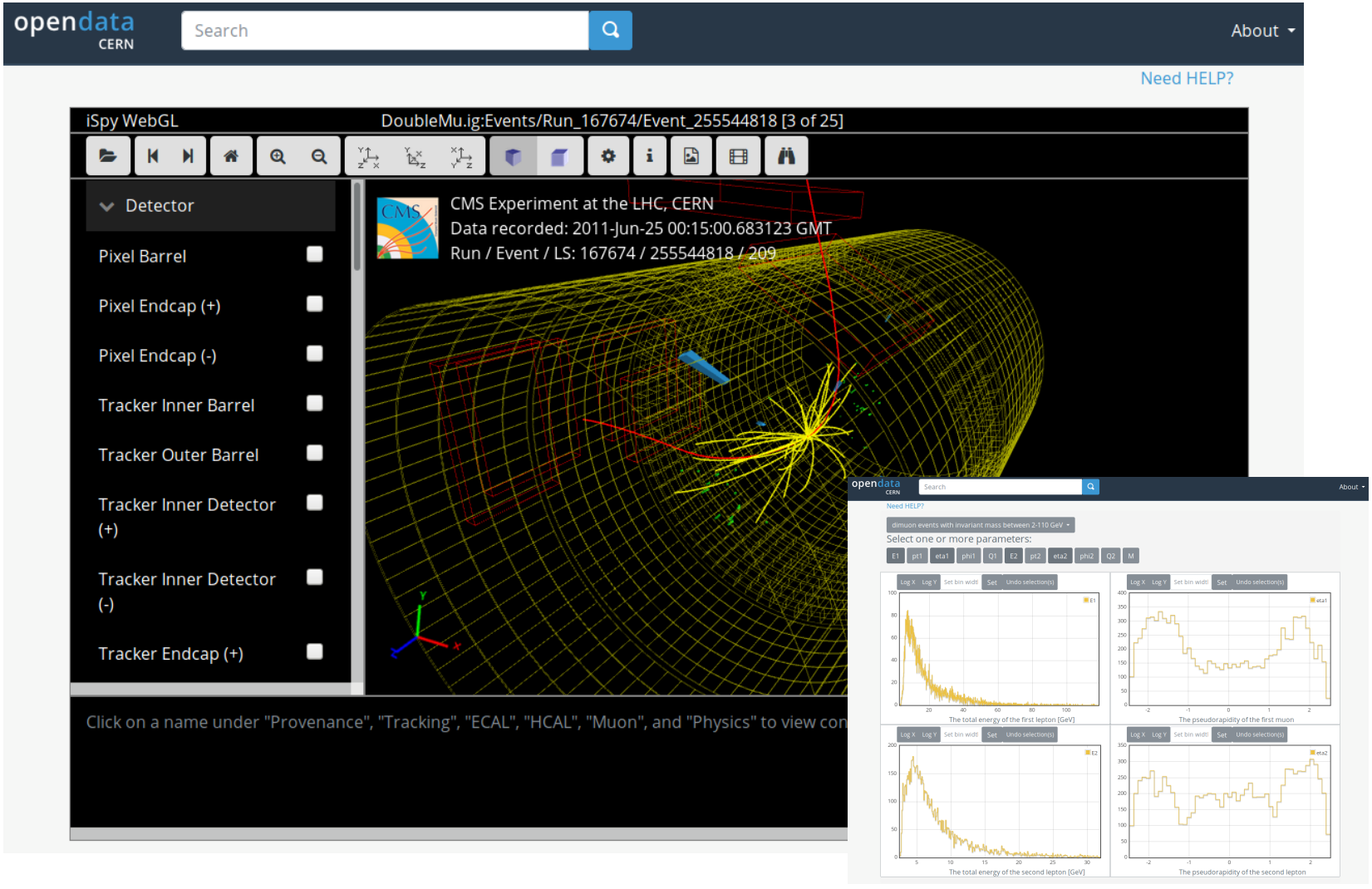

Curated context information about data selection, validation, use

# Information discovery




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

# Education use cases




Use interactive event display and basic histogramming

# Analysis examples

opendata CERN | Search | About

## 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 reimplementation 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. 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.

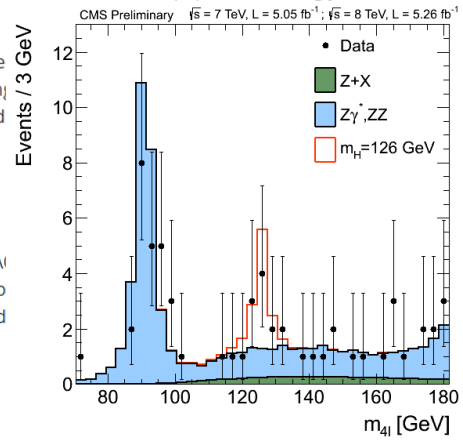
The example consists of different levels of complexity. The highest leve
minimal understanding of the content of this paper and of the meanin
educational exercises. The lower levels might also be interesting for ed
with the linux operating system and the ROOT analysis tool.

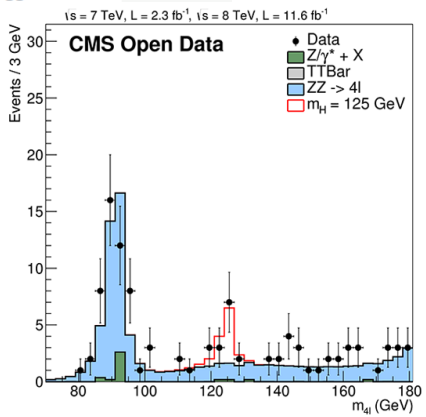
### Use with

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




Install virtual machines and run realistic physics analysis examples

# Research use cases

PRL 119, 132003 (2017) PHYSICAL REVIEW LETTERS week ending 29 SEPTEMBER 2017

## Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,‖]
[1]Physics Department, Reed College, Portland, Oregon 97202, USA
[2]University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA
[3]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 9 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 partons. 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 asymptotes to the splitting function for sufficiently high jet energies. This provides a way to expose the splitting function through jet substructure measurements 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 test the $1 \to 2$ 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 partons become collinear in QCD, the cross section for a $2 \to n$ scattering process factorizes into a $2 \to n-1$ scattering cross section multiplied by a universal $1 \to 2$ 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-Parisi evolution equations [1–3] (see also [4–13]), and it is at the heart of the factorization theorem in hadron-hadron collisions [14,15]. In addition, parton shower generators are based on recursively applying $1 \to 2$ splittings [16–18], fixed-order subtraction schemes utilize the $1 \to 2$ splitting function [19–21], and the $k_t$ jet clustering metric is based on $2 \to 1$ recombination [22–24]. Collinear universality can be extended to multiparton splittings at tree level and beyond [25–41]; however, its all-orders validity [42,43] is spoiled in the presence of Glauber modes [44–47]. More recently, jet substructure techniques [48–52] have been introduced to distinguish $1 \to n$ decays of heavy particles from $1 \to n$ splittings in QCD in order to enhance the search for new physics at the Large Hadron Collider (LHC) [53–56].

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

$$dP_{i\to jk} = \frac{d\theta}{\theta} dz P_{i\to jk}(z), \tag{1}$$

where the $P_{i\to jk}$ are the Altarelli-Parisi QCD splitting functions [3] which depend on the momentum fraction $z$ and the parton flavors $i$, $j$, and $k$. Crucially, this expression has a real emission singularity in the $\theta \to 0$ limit, as required to cancel corresponding virtual singularities from loop diagrams. In this sense, there is no way to directly measure the splitting function $P_{i\to jk}(z)$ in data, though there is of course overwhelming indirect evidence that $P_{i\to jk}(z)$ is a universal function from the many successes of QCD in describing high-energy scattering (see, e.g., [57–67]).

In this Letter, we present a semidirect method to test the $1 \to 2$ splitting function in QCD by studying the two-prong substructure of jets. Our method is based on soft drop declustering [68] (see also [52,69,70]), which recursively removes soft radiation from a jet until hard two-prong substructure is found. When applied to ordinary quark- and gluon-initiated jets with no intrinsic substructure, soft drop exposes the collinear core of the jet. As shown in Ref. [71], the momentum sharing between the two prongs (denoted $z_g$) is closely related to the momentum fraction $z$ appearing in Eq. (1), and the cross section for $z_g$ asymptotes to the QCD splitting function in the high-energy limit. While variants of $z_g$ have appeared in many jet substructure studies (notably the $\sqrt{y}$ parameter in Refs. [52,72]), to the best of our knowledge, no published $z_g$ distribution has ever been presented using actual collider data, though there are preliminary $z_g$ results from CMS [73], STAR [74], and ALICE [75] Collaborations. Here, we present the first analysis of $z_g$ using LHC data, taking advantage for 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 data (AOD) format, which is a CMS-specific data scheme based

PHYSICAL REVIEW D 96, 074003 (2017)

## Jet substructure studies with CMS open data

Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,‖]
[1]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Physics Department, Reed College, Portland, Oregon 97202, USA
[3]University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA
(Received 9 May 2017; published 3 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 31.8 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 substructure studies.

DOI: 10.1103/PhysRevD.96.074003

### I. 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 collision 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 unique low-luminosity running environment where pileup contamination was minimal and trigger thresholds were relatively low. The CMS open data present an enormous opportunity 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 copiously produced in 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 boosted heavy objects like top quarks [11–31] and electroweak bosons [3,4,6,14,22,30–59]. To successfully tag such objects, though, one first has to understand the radiation patterns of ordinary quark and gluon jets [26,60–75], which are the main backgrounds to boosted objects. The CMS open data are a fantastic resource for performing these baseline quark/gluon studies. Using the Jet Primary Dataset [76], we perform initial investigations of the two-prong substructure of jets as well as present a general analysis framework to facilitate future studies. This effort is complementary to the growing catalog of jet substructure measurements performed within the ATLAS and CMS collaborations [77–199].[1]

The core of our analysis is based on soft-drop declustering [46], which is a jet grooming technique [6,200–202] that mitigates jet contamination from initial state radiation (ISR), underlying event (UE), and pileup. For the studies in this paper, we set the soft-drop parameter $\beta$ equal to zero, such that soft drop behaves like the modified mass drop tagger (mMDT) [203,204].[2] After soft drop, a jet is composed of two well-defined subjets, which can then be used to derive various two-prong substructure observables. In addition to comparing the CMS open data to parton shower generators, we perform first-principles calculations of soft-dropped observables using recently developed analytic techniques [46,205,206]. In a companion paper, we use soft drop to expose the QCD splitting function using the CMS open data [207]; a similar strategy was used in preliminary CMS [167], STAR [208], and ALICE [209] heavy ion studies to test for possible modifications to the splitting function from the dense QCD medium [210,211].

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

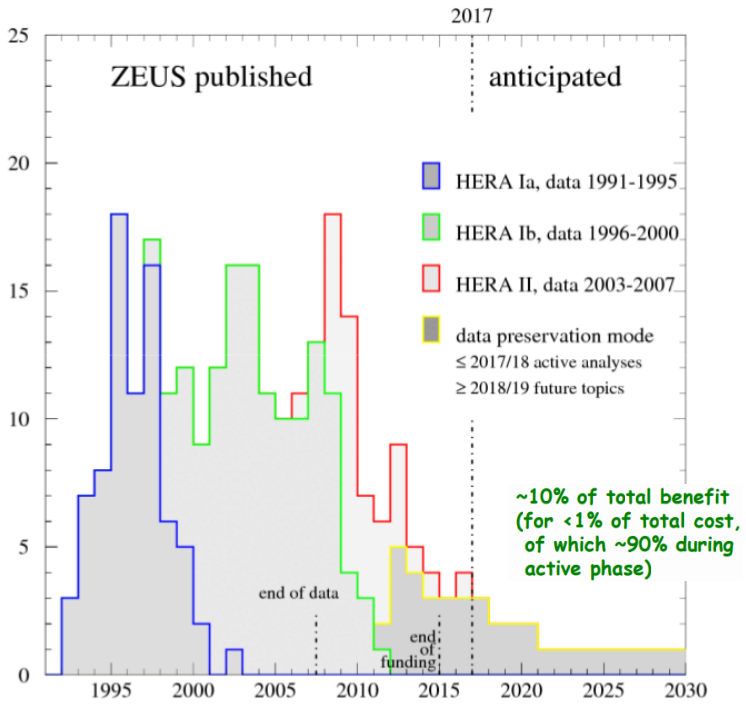
*aashisht@mit.edu
†weixue@mit.edu
‡larkoski@reed.edu
§smarzani@buffalo.edu
‖jthaler@mit.edu

[1]To highlight the vibrancy 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.
[2]The original mass drop tagger [6] was a pioneering technique in jet substructure; see also precursor work in Refs. [2–5].

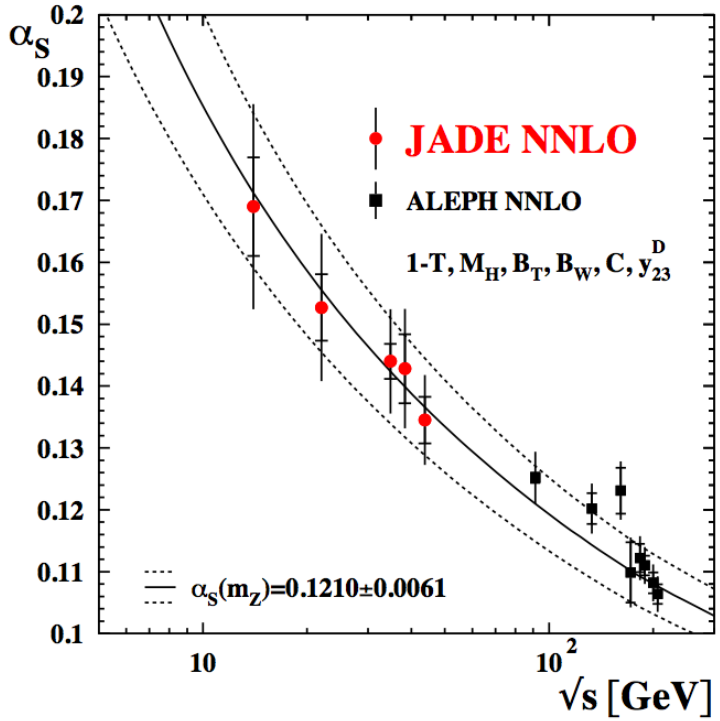
 

Independent analyses by theorists (Jesse Thaler *et al*, MIT)

# Long-term value of data!




Achim Geiser `https://indico.cern.ch/event/588219`



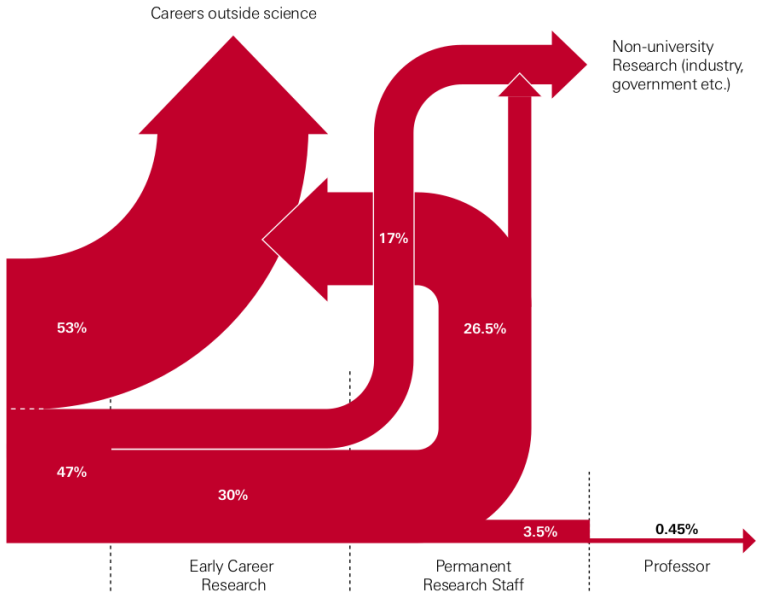

DPHEP `https://arxiv.org/abs/1205.4667`

Collaborations publish papers even ∼15 years after data taking ends

JADE data (1979–1986) still unique even ∼35 years later

# Long-term value of knowledge?

CMS collaboration



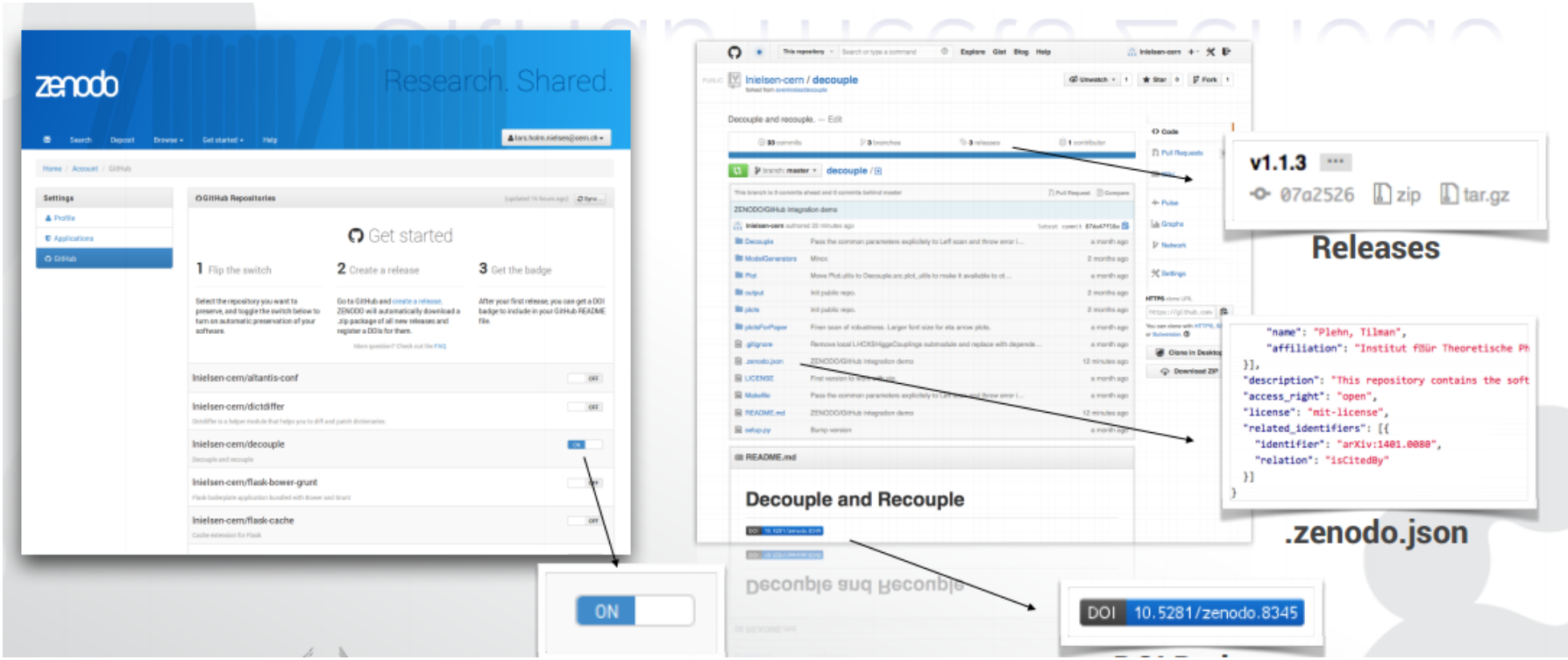

Career after PhD

THE ROYAL SOCIETY

Experimental physics done by groups of ∼3000 physicists

High turnover of young researchers

# Preserving scientific code




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

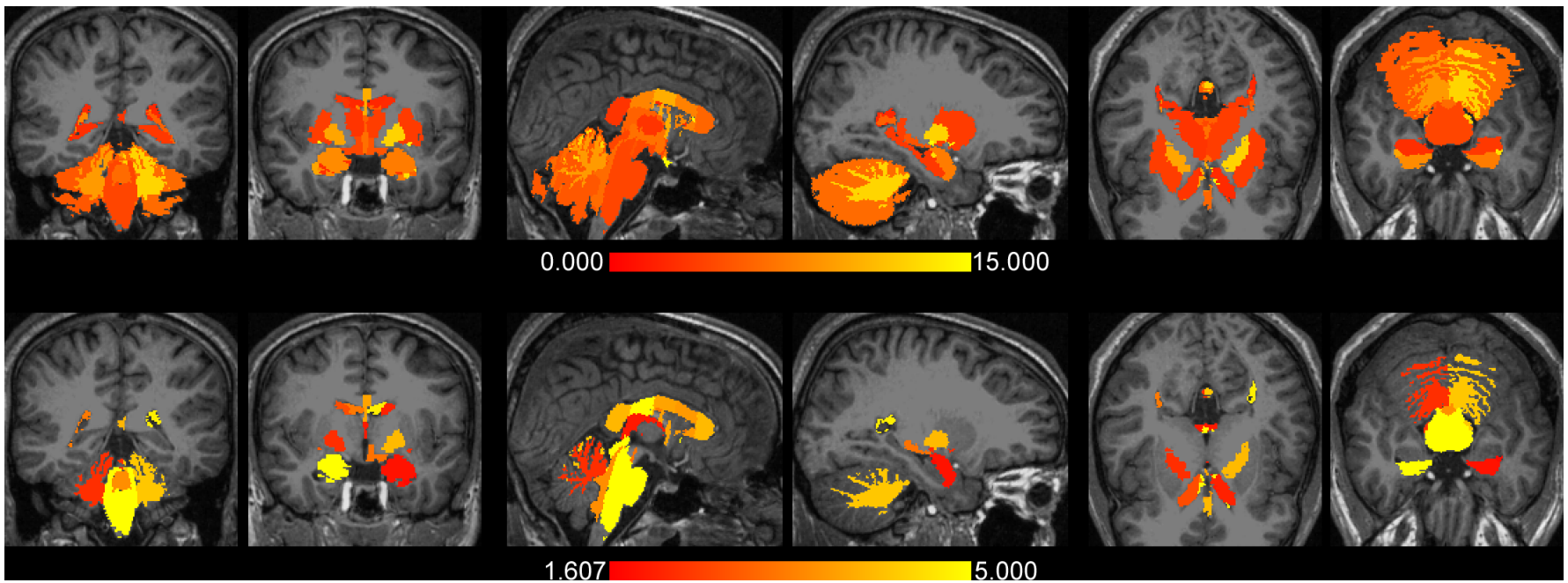
GitHub ↔ Zenodo bridge to automatically preserve releases

# Code is not enough

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

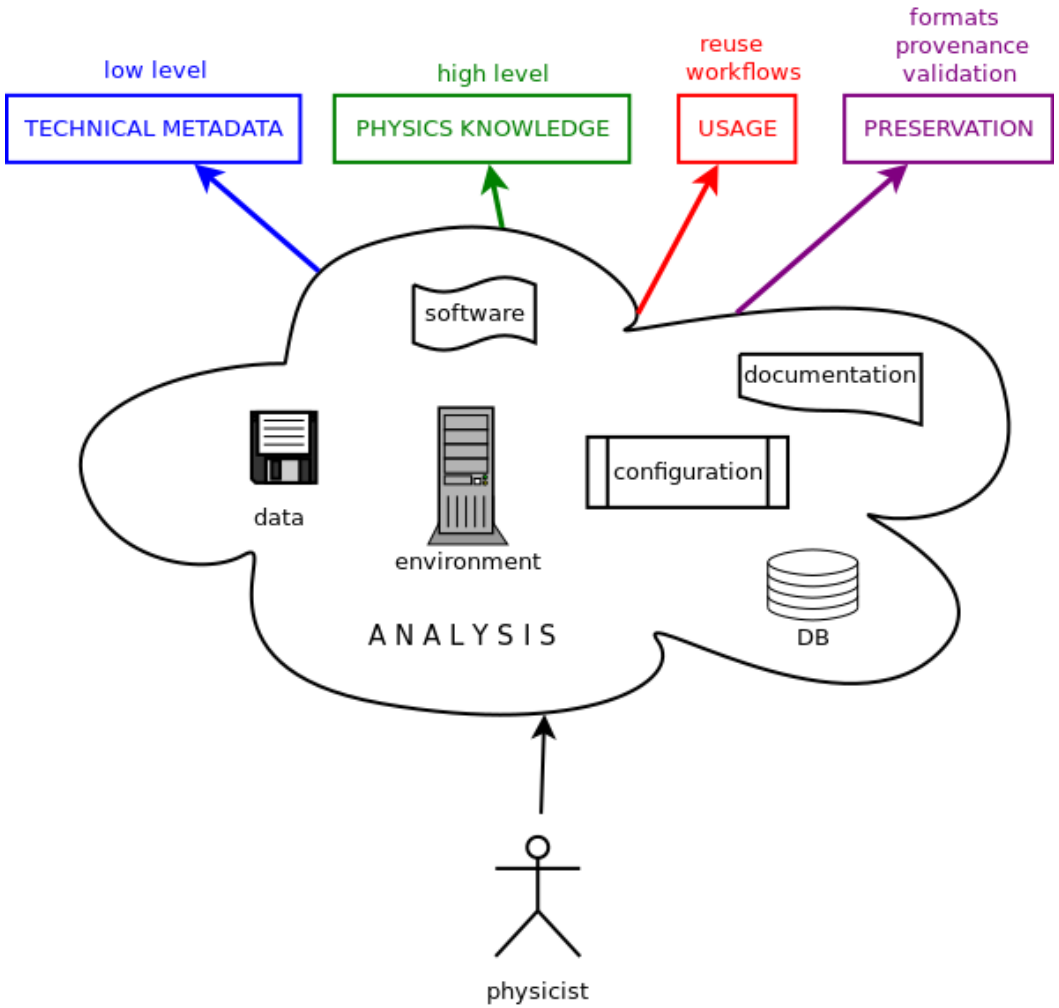
Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

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




8.8±6.6% (volume) and 2.8±1.3% (cortical thickness)

# Preserving analyses 1/2

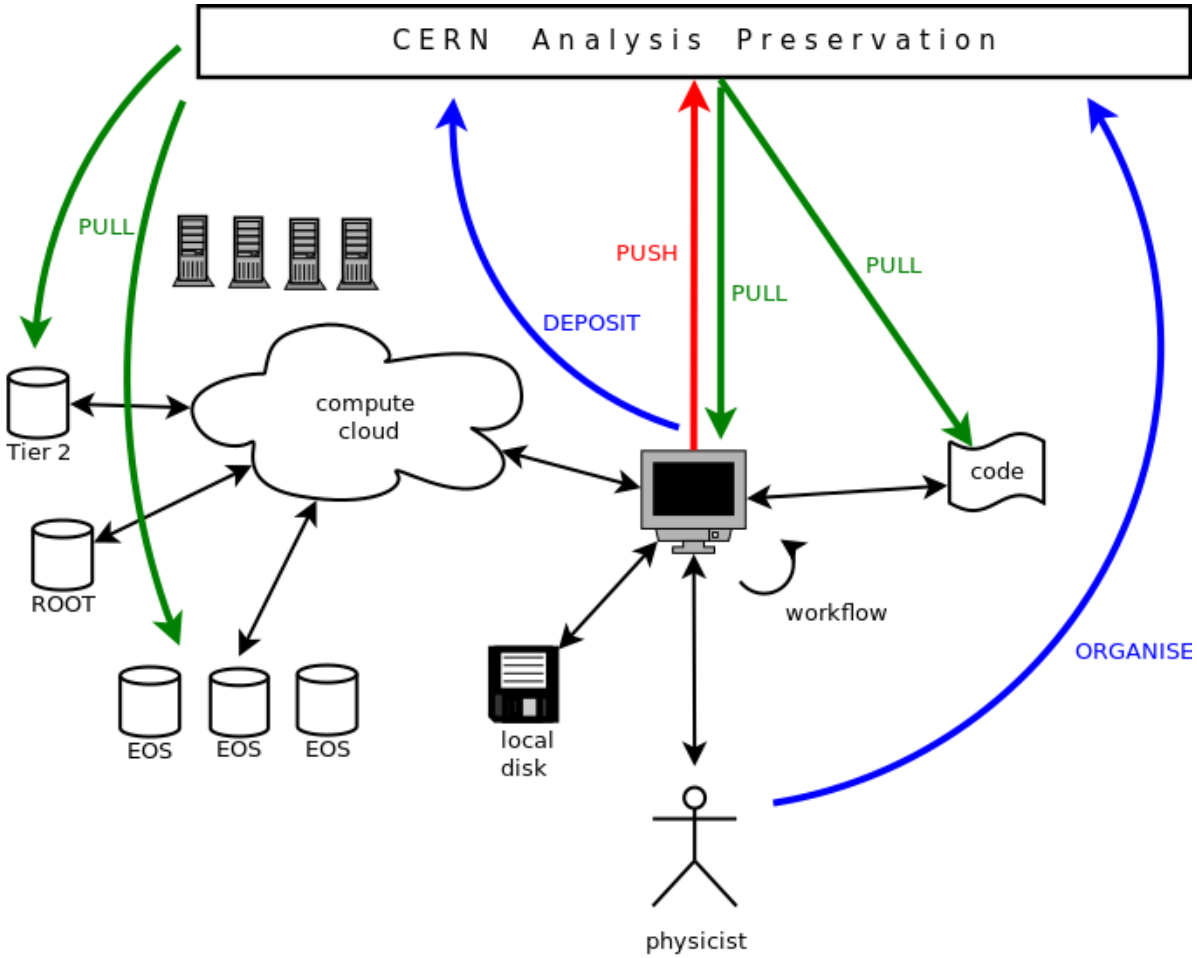



INVENIO

- JSON Schema
- W3C DCAT
- domain-specific fields

Structuring knowledge behind research data analysis

# Preserving analyses 2/2

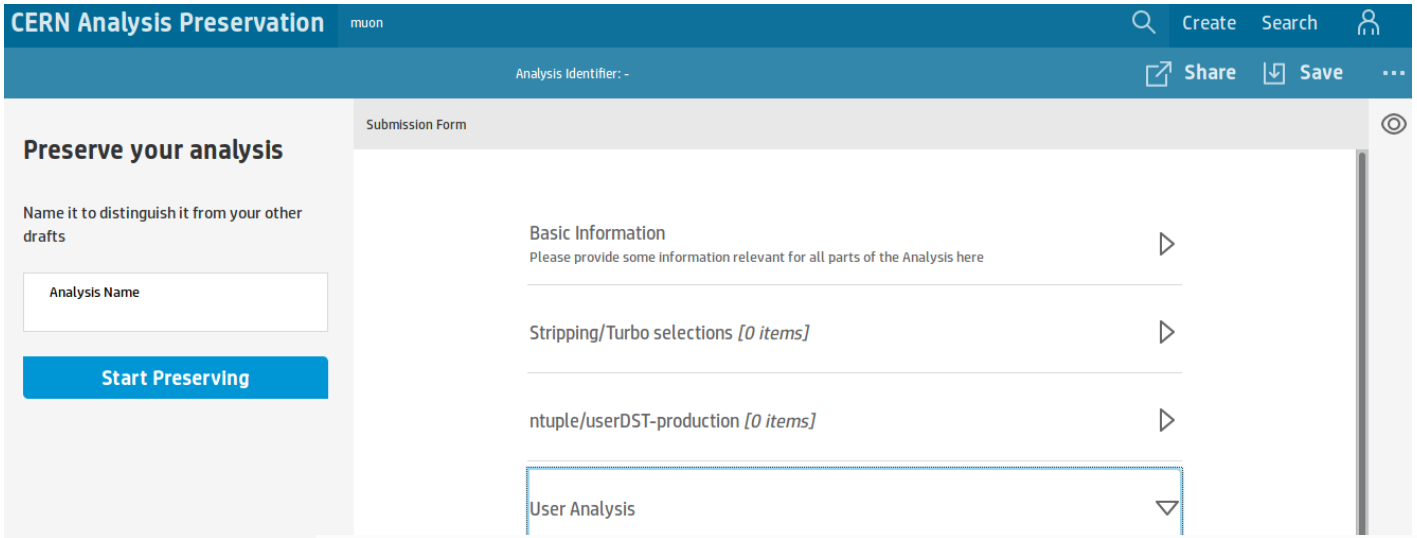



INVENIO

- 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




```
$ 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.
```

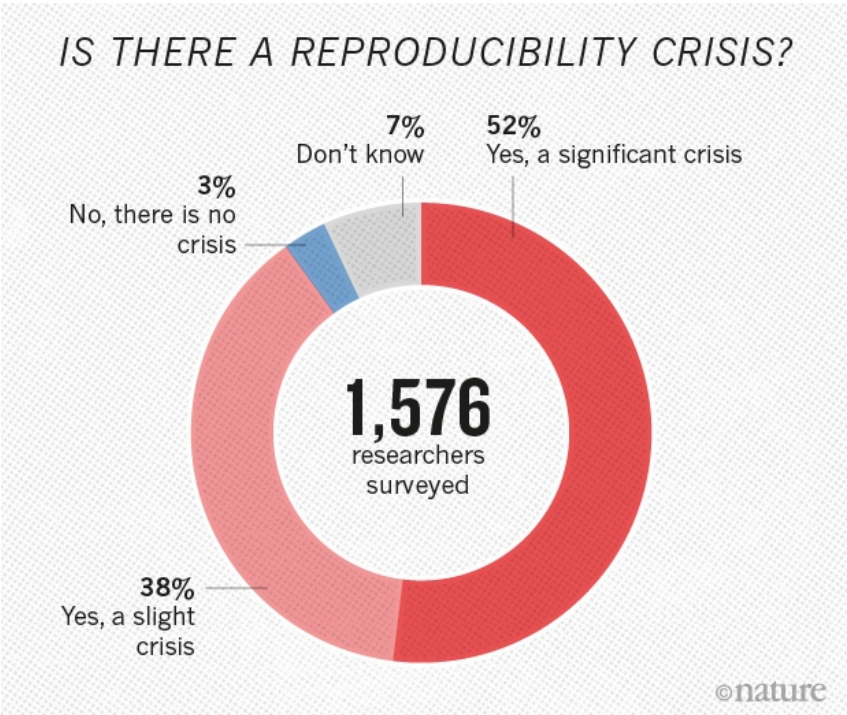
Web based and command-line based deposit workflows

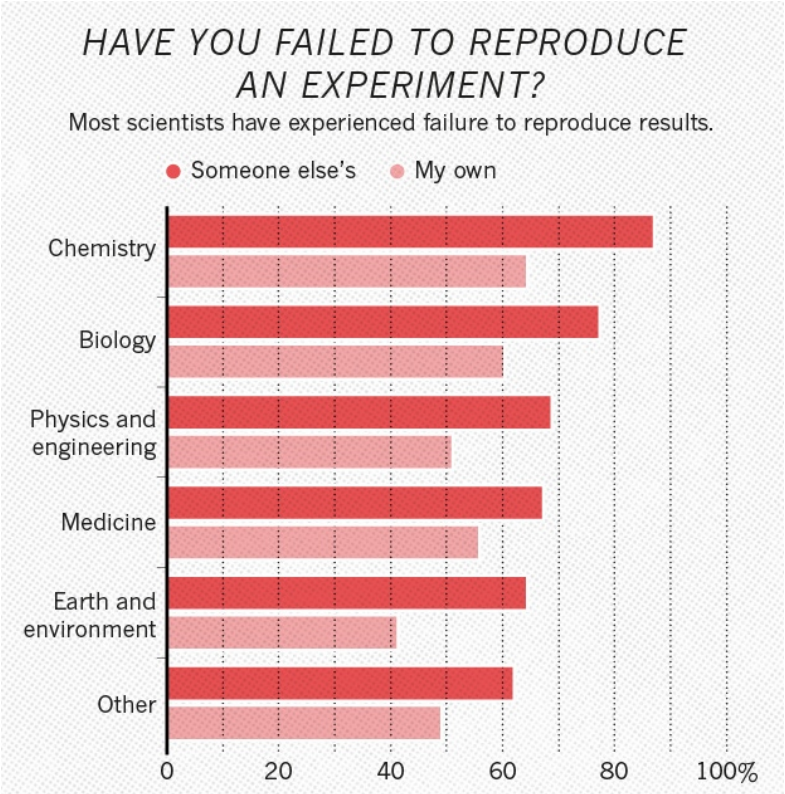
# How FAIR is scientific data?

- **F**indable
  - persistent identifiers
  - rich metadata
  - indexed and searchable
- **A**ccessible
  - retrievable by identifiers
  - standard protocols
  - metadata vs data accessibility
- **I**nteroperable
  - knowledge representation language
  - common vocabularies
  - references to other metadata and data
- **R**eusable
  - domain-relevant attributes and community standards
  - clear licensing
  - provenance tracking

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

# Reusable and reproducible?

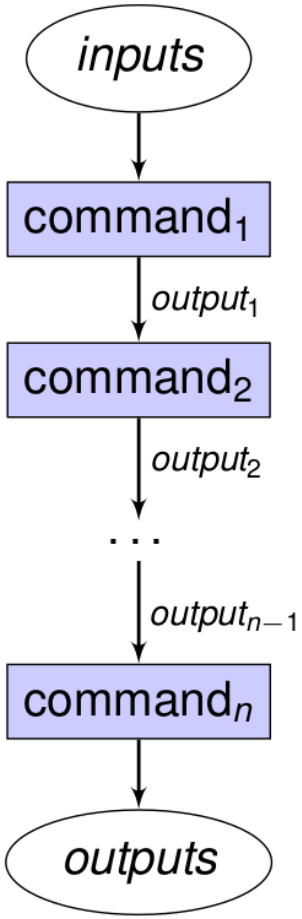


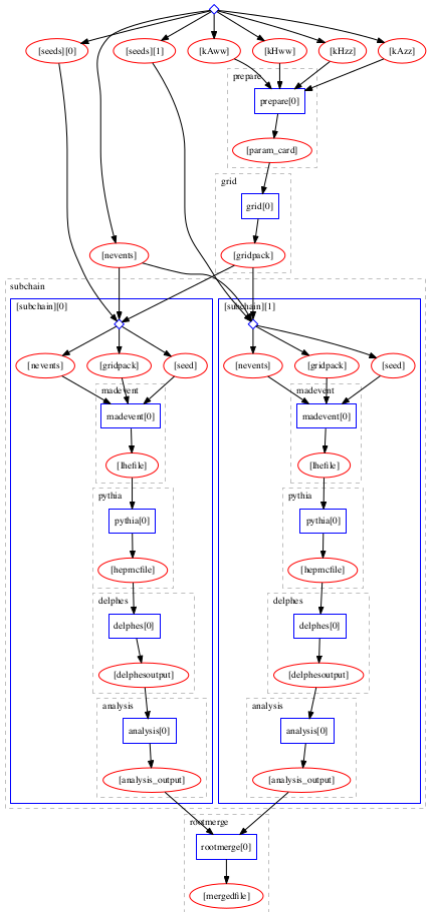



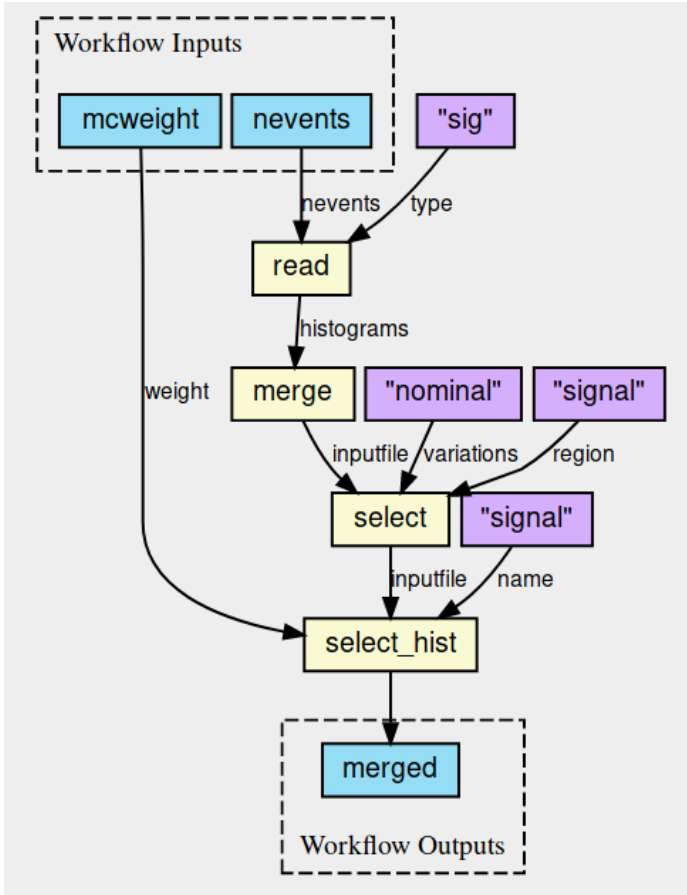

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Half of researchers cannot reproduce their own experimental results

# Computational workflows






Serial Yadage CWL

# REANA Reusable Analyses


Reproducible research data analysis platform

**Flexible**

Run many computational workflow engines.


**Scalable**

Support for remote compute clouds.


**Reusable**

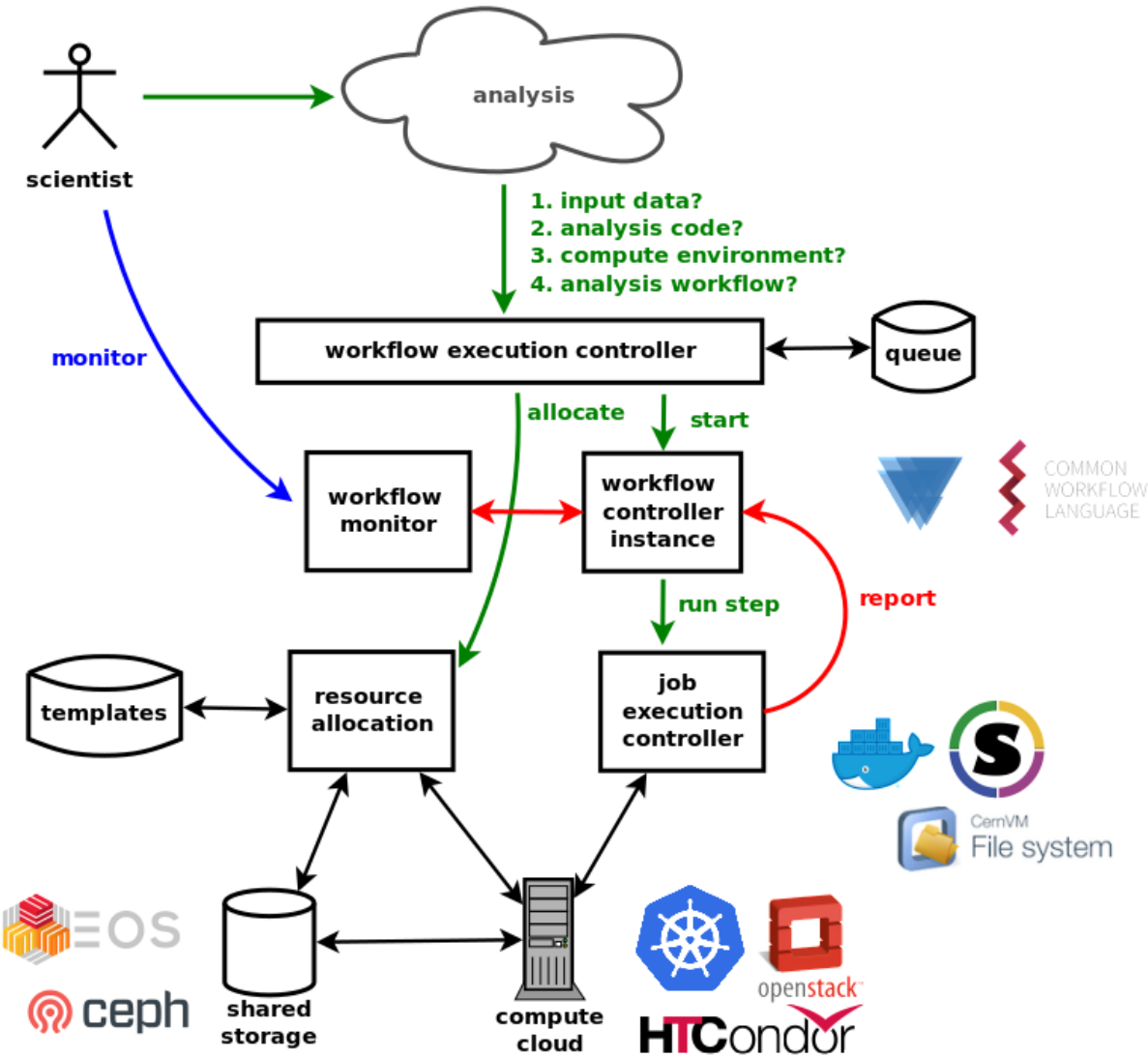
Containerise once, reuse elsewhere. Cloud-native.

**Free**

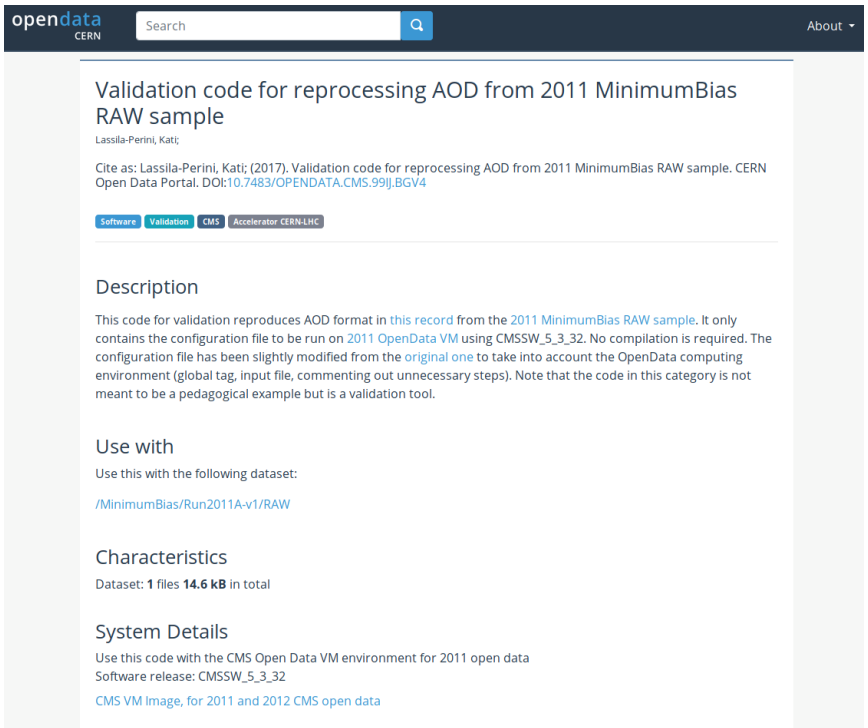
Free Software. MIT licence. Made with ❤ at CERN.

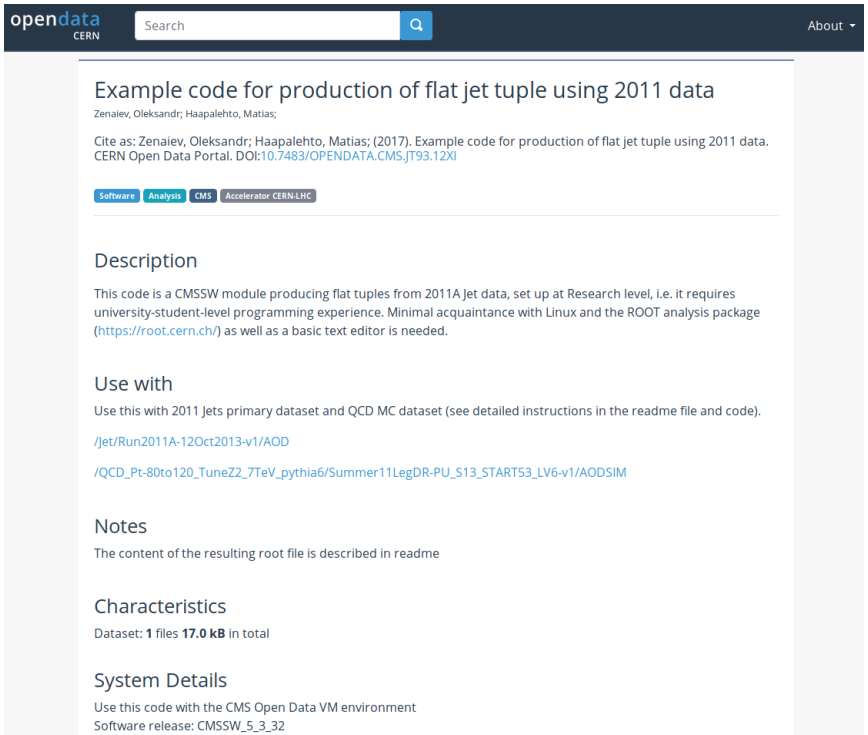

`http://www.reana.io/`

# REANA architecture



# Data production examples



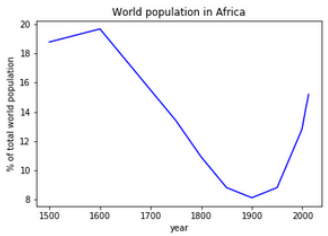




Reconstruction and flat jet tuple production from CMS open data

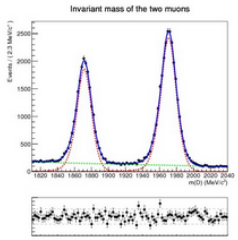
# Data analysis examples

```
$ cat inputs/names.txt
  Jane Doe
  Joe Bloggs
$ reana-client start
$ reana-client download
$ cat outputs/greetings.txt
  Hello Jane Doe!
  Hello Joe Bloggs!
```

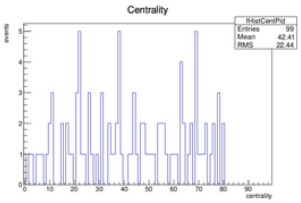
"Hello world"



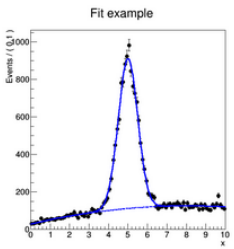

Parametrised Jupyter notebook



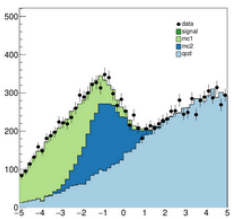

LHCb rare charm decay search



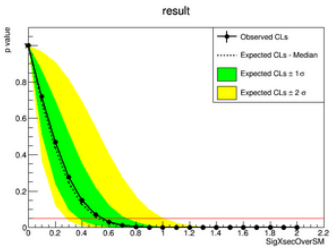

ALICE LEGO train test run



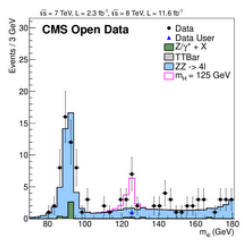

ROOT/RooFit physics analysis



ATLAS BSM search



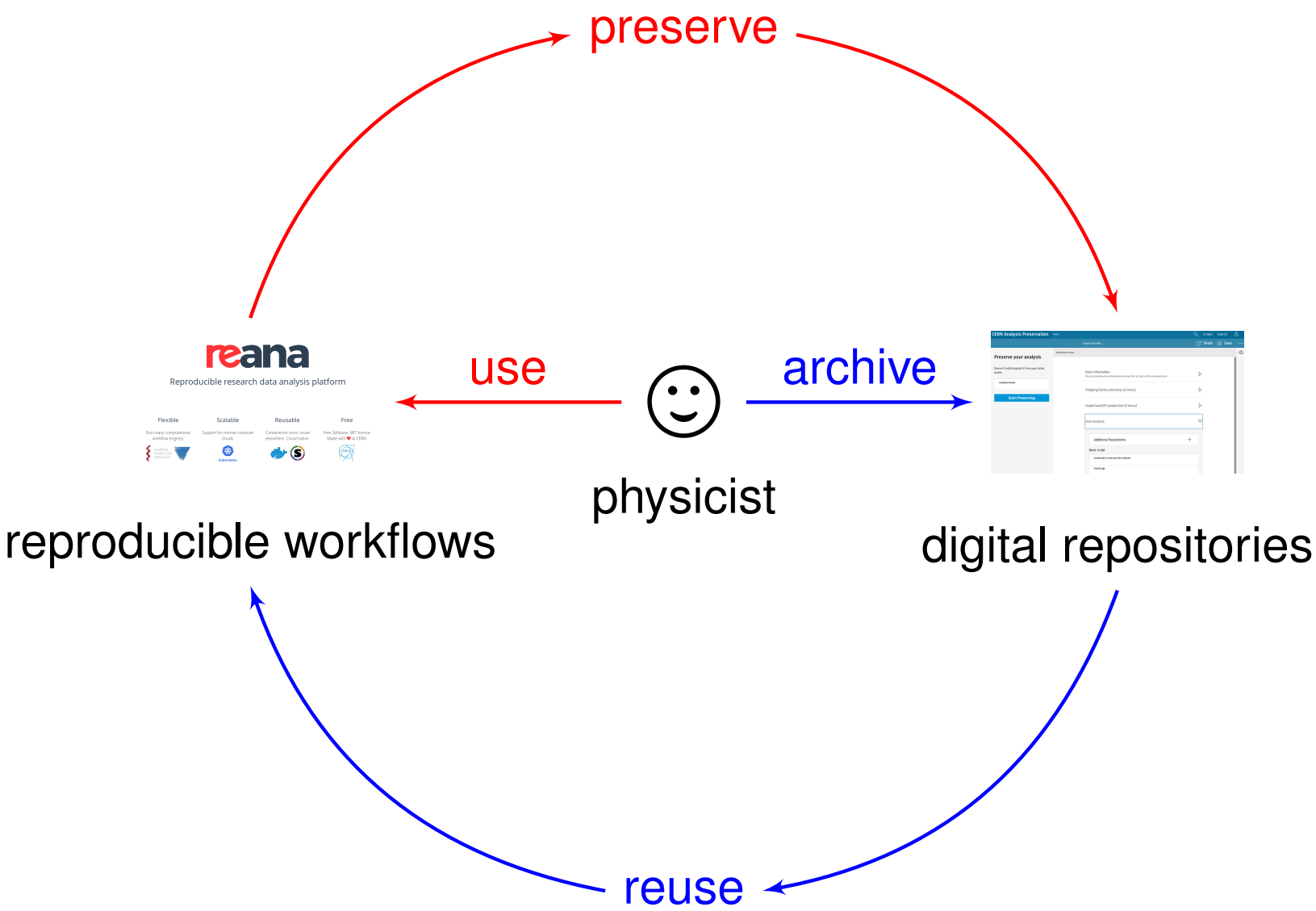

ATLAS RECAST



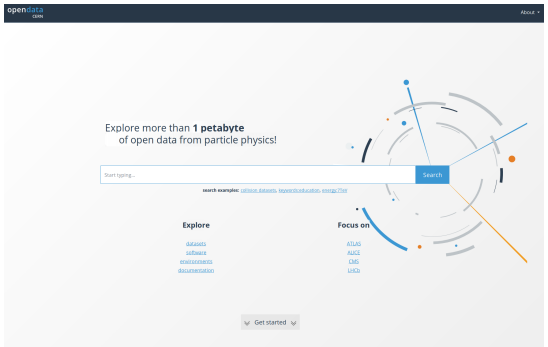

CMS Higgs-to-four-leptons

Several physics analysis examples

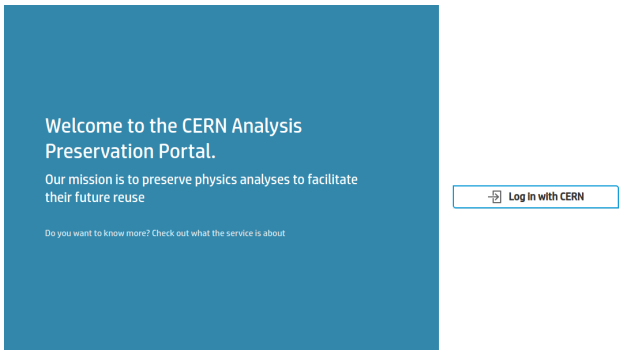
# Reproducibility ⇌ Preservation



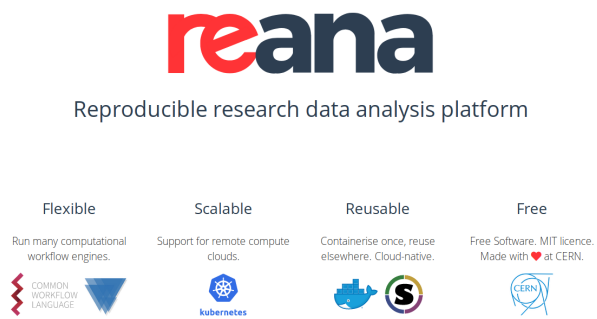
# Fostering FAIR science




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** R. Maciulaitis, J. Okraska, D. Rodriguez, T. Šimko · **CERN SIS** S. Feger, P. Fokianos, A. Lavasa, S. van de Sandt, 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

# Conclusions

- opening complex LHC physics data beyond education use cases
  - enables independent physics research
  - validates and enriches knowledge preservation techniques
- open data is not enough
  - code, environments, runnable recipes and analysis workflows
  - science is born naturally ‘closed’; FAIR principles apply early
- “top-down” approach
  - funding agency requirements for preservation and openness
  - collaboration best practices for better reproducibility
- “bottom-up” activities
  - building tools to fit scientists’ daily workflow
  - pre-producibility rather then re-producibility
- time is ripe for driving the change
  - technological progress meets sociological challenges

# Further reading 1: CERN Courier


CERN Courier March/April 2019

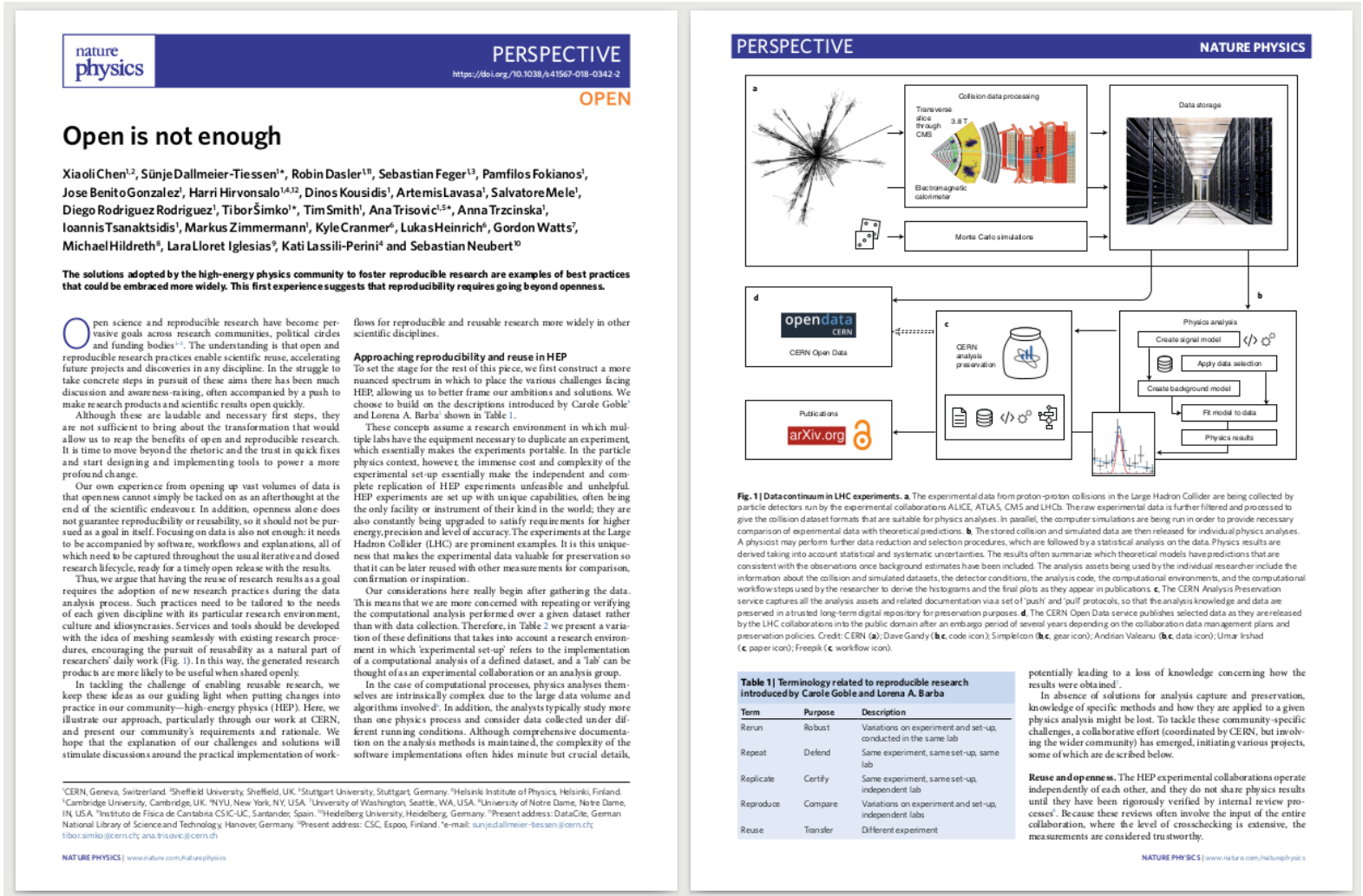
“The rise of Open Science”

Features:

- open access
- open data
- open source
- open science

`https://cerncourier.com/`

# Further reading 2: Nature Physics



nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

OPEN

## Open is not enough

Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1]*, Robin Dasler[1,11], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1]*, Tim Smith[1], Ana Trisovic[1,5]*, Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassila-Perini[4] and Sebastian Neubert[10]

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

Open science and reproducible research have become pervasive goals across research communities, political circles and funding bodies[1–3]. The understanding is that open and reproducible research practices enable scientific reuse, accelerating future projects and discoveries in any discipline. In the struggle to take concrete steps in pursuit of these aims there has been much discussion and awareness-raising, often accompanied by a push to make research products and scientific results open quickly.

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

Our own experience from opening up vast volumes of data is that openness cannot simply be tacked on as an afterthought at the end of the scientific endeavour. In addition, openness alone does not guarantee reproducibility or reusability, so it should not be pursued as a goal in itself. Focusing on data is also not enough: it needs to be accompanied by software, workflows and explanations, all of which need to be captured throughout the usual iterative and closed research lifecycle, ready for a timely open release with the results.

Thus, we argue that having the reuse of research results as a goal requires the adoption of new research practices during the data analysis process. Such practices need to be tailored to the needs of each given discipline with its particular research environment, culture and idiosyncrasies. Services and tools should be developed with the idea of meshing seamlessly with existing research procedures, encouraging the pursuit of reusability as a natural part of researchers' daily work (Fig. 1). In this way, the generated research products are more likely to be useful when shared openly.

In tackling the challenge of enabling reusable research, we keep these ideas as our guiding light when putting changes into practice in our community—high-energy physics (HEP). Here, we illustrate our approach, particularly through our work at CERN, and present our community's requirements and rationale. We hope that the explanation of our challenges and solutions will stimulate discussions around the practical implementation of workflows 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 construct a more nuanced 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 Carole Goble[4] and Lorena A. Barba[5] 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 immense 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 constantly being upgraded 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 really begin after gathering the data. This means that we are more concerned with repeating or verifying the computational analysis performed over a given dataset rather than with data collection. Therefore, in Table 2 we present a variation of these definitions that takes into account a research environment in which 'experimental set-up' refers to 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, physics analyses themselves are intrinsically complex due to the large data volume and algorithms involved[6]. In addition, the analysts typically study more than one physics process and consider data collected under different running conditions. Although comprehensive documentation on the analysis methods is maintained, the complexity of the software implementations often hides minute but crucial details, potentially leading to a loss of knowledge concerning how the results were obtained[7].

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. To tackle these community-specific challenges, a collaborative effort (coordinated by CERN, but involving the wider community) has emerged, initiating various projects, some of which are described below.

**Reuse and openness.** The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by internal review processes[8]. Because these reviews often involve the input of the entire collaboration, where the level of crosschecking is extensive, the measurements are considered trustworthy.

[1]CERN, Geneva, Switzerland. [2]Sheffield University, Sheffield, UK. [3]Stuttgart University, Stuttgart, Germany. [4]Helsinki Institute of Physics, Helsinki, Finland. [5]Cambridge University, Cambridge, UK. [6]NYU, New York, NY, USA. [7]University of Washington, Seattle, WA, USA. [8]University of Notre Dame, Notre Dame, IN, USA. [9]Instituto de Física de Cantabria CSIC-UC, Santander, Spain. [10]Heidelberg University, Heidelberg, Germany. [11]Present address: DataCite, German National Library of Science and Technology, Hanover, Germany. [12]Present address: CSC, Espoo, Finland. *e-mail: sunje.dallmeier-tiessen@cern.ch; tibor.simko@cern.ch; ana.trisovic@cern.ch







**Fig. 1 | Data continuum in LHC experiments. a**, The experimental data from proton–proton collisions in the Large Hadron Collider are being collected by particle detectors run by the experimental collaborations ALICE, ATLAS, CMS and LHCb. The raw experimental data is further filtered and processed to give the collision dataset formats that are suitable for physics analyses. In parallel, the computer simulations are being run in order to provide necessary comparison of experimental data with theoretical predictions. **b**, The stored collision and simulated data are then released for individual physics analyses. A physicist may perform further data reduction and selection 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 summarize which theoretical models have predictions that are consistent with the observations once background estimates have been included. The analysis assets being used by the individual researcher include the information about the collision and simulated datasets, the detector conditions, the analysis code, the computational environments, and the computational workflow steps used by the researcher to derive the histograms and the final plots as they appear in publications. **c**, The CERN Analysis Preservation service captures all the analysis assets and related documentation via a set of 'push' and 'pull' protocols, so that the analysis knowledge and data are preserved in a trusted long-term digital repository for preservation purposes. **d**, The CERN Open Data service publishes selected data as they are released 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 (**a**); Dave Gandy (**b**,**c**, code icon); Simpleicon (**b**,**c**, gear icon); Andrian Valeanu (**b**,**c**, data icon); Umar Irshad (**c**, paper icon); Freepik (**c**, workflow icon).

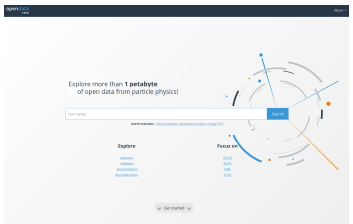
**Table 1 | Terminology related to reproducible research introduced by Carole Goble and Lorena A. Barba**

| Term | Purpose | Description |
|---|---|---|
| Rerun | Robust | Variations on experiment and set-up, conducted in the same lab |
| Repeat | Defend | Same experiment, same set-up, same lab |
| Replicate | Certify | Same experiment, same set-up, independent lab |
| Reproduce | Compare | Variations on experiment and set-up, independent labs |
| Reuse | Transfer | Different experiment |



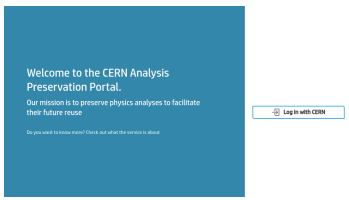
https://www.nature.com/articles/s41567-018-0342-2.pdf

# References



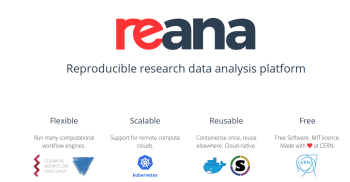

## CERN Open Data

http://opendata.cern.ch
http://github.com/cernopendata
cernopendata



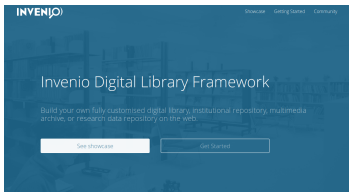

## CERN Analysis Preservation

http://analysispreservation.cern.ch
http://github.com/cernanalysispreservation
analysispreserv



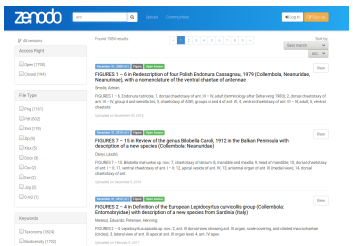

## REANA

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