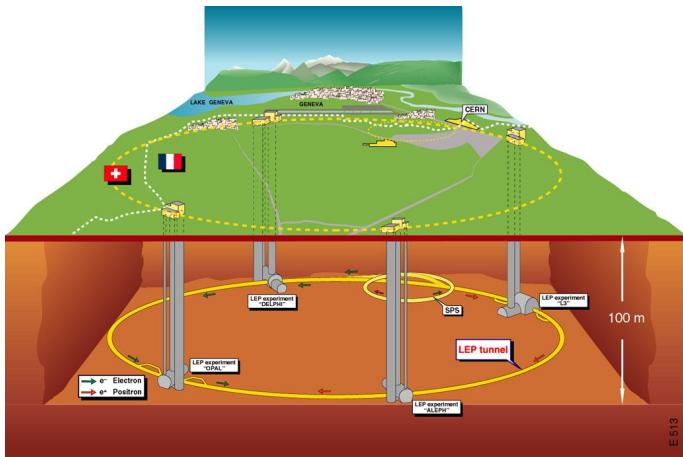


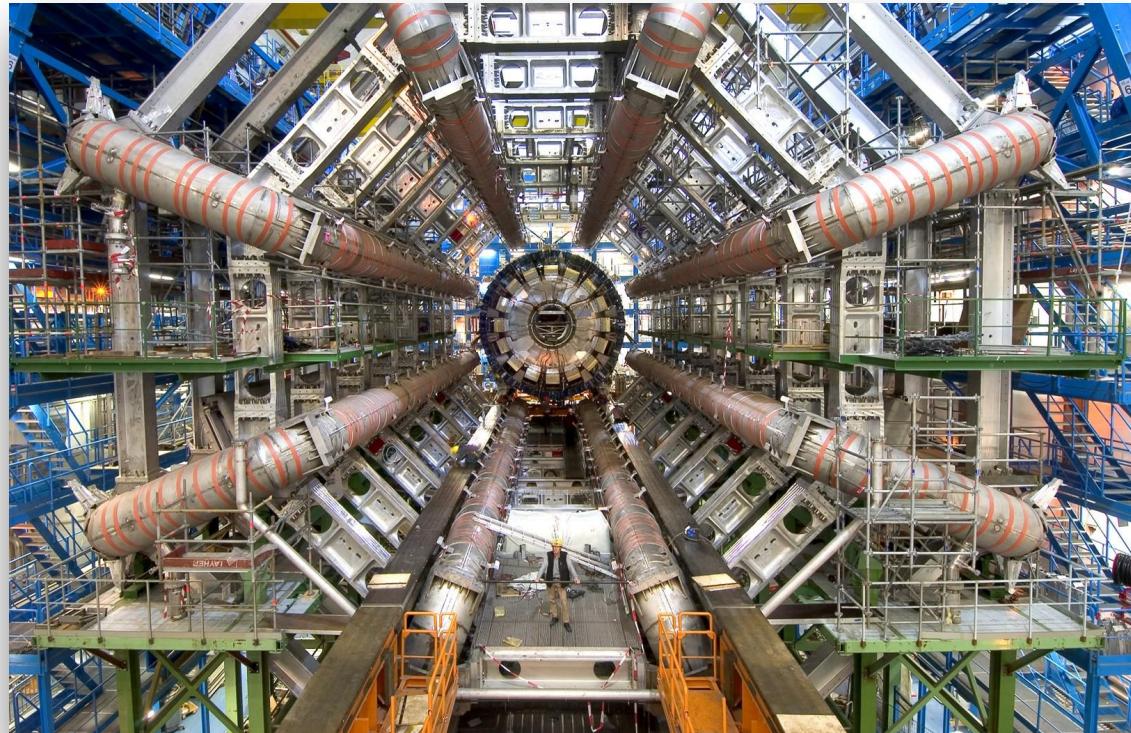
Open and Reproducible Research Services in LHC Particle Physics

Diego Rodríguez
CERN

CERN Large Hadron Collider

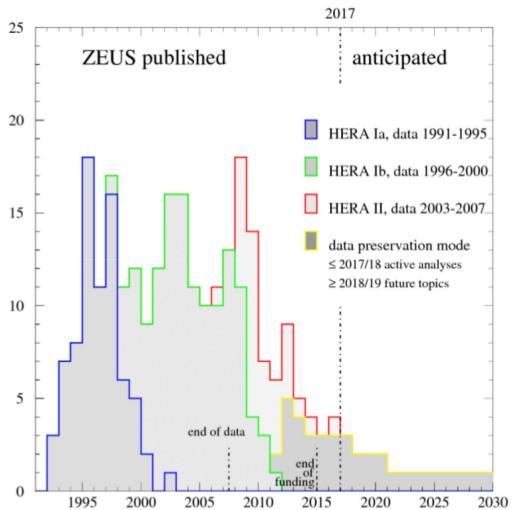


<http://cds.cern.ch/record/842153>



<https://cds.cern.ch/record/910381>

Data and knowledge



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



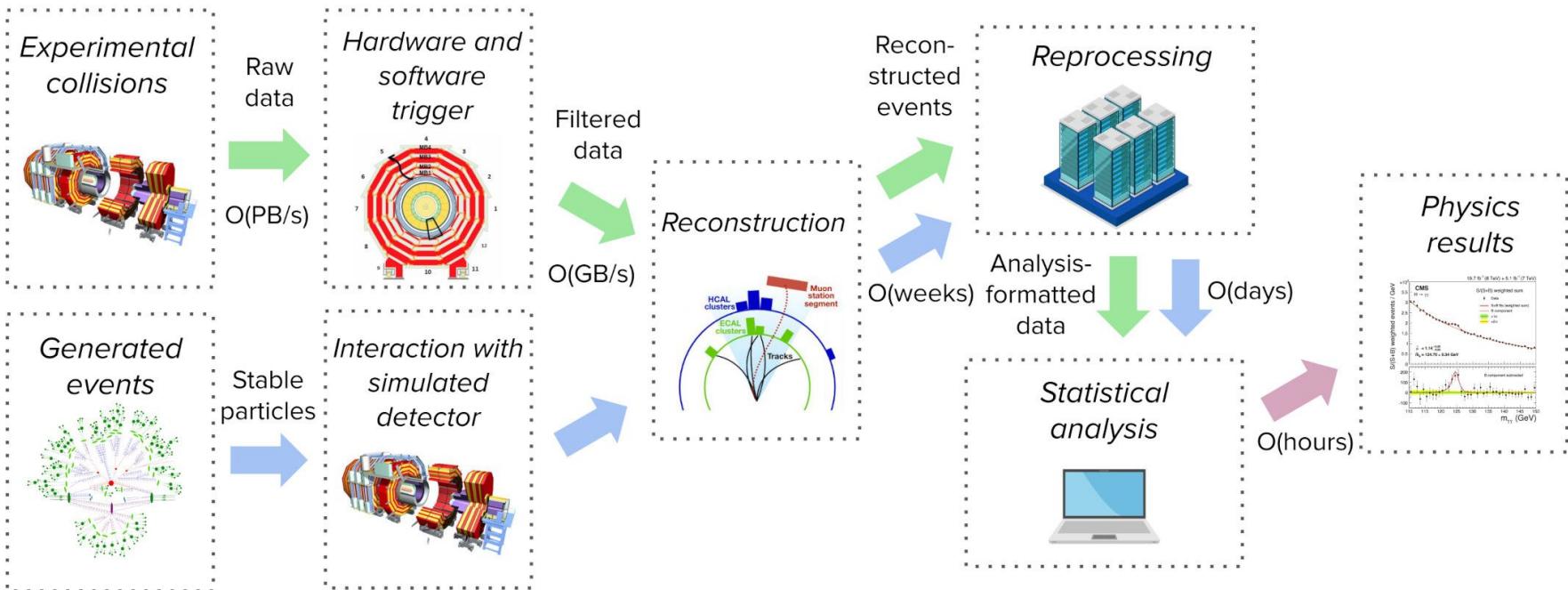
Patrick Koppenburg @PKoppenburg · 4 sept.

The @LHCbExperiment collaboration submitted its first physics paper 10 years ago. It had 629 authors arxiv.org/abs/1008.3105. Now we are 972, but only half of the authors of the first paper are still with us. They are in violet in the list below.

Prompt K_short production in pp collisions at sqrt(s)=0.9 TeV

<https://twitter.com/PKoppenburg/status/1301813341460066304>

HEP data analyses



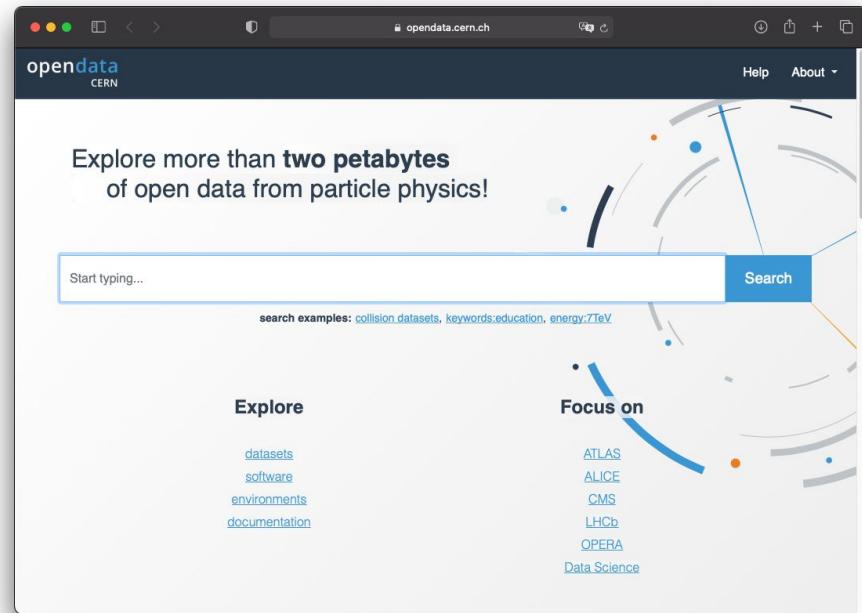
CERN Open Data portal

Launched in 2014

Disseminating over 2.4 PB of data

7.500 records

900.000 files



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

Information organisation

The figure displays three screenshots of the CERN Open Data portal (opendata.cern.ch) illustrating various information organization features:

- Screenshot 1: Detailed Dataset Page**

This screenshot shows a detailed view of a dataset. It includes:
 - Description:** A summary of the Dimuon event information derived from the Run2010B public Mu dataset.
 - Related datasets:** Links to MuRun2010B-Apr21Reco-v1AOD and MuRun2010B-Apr21Reco-v1AOD.
 - Dataset characteristics:** Information about 100000 events.
 - Dataset semantics:** Definitions for Variable (Run, Event, E), Description (The run number of the event, The event number, The total energy of the muon), and Files (MuRun2010B_0.csv, MuRun2010B.csv, MuRun2010B_1.csv, MuRun2010B_2.csv, MuRun2010B_3.csv).
- Screenshot 2: Faceted Search Results**

This screenshot shows a search results page for "dimuon". The results are filtered by type (Dataset, Document, Environment, Software) and experiment (ALICE, CMS, OPERA). The results list includes:
 - MuRun2010B-Apr21Reco-v1AOD
 - Razor filter and analyzer for SUSY searches
 - Configuration file for HLT step /cdaq/physics/Run2010Hi/v1.5/MILHTV1
- Screenshot 3: Large File Download**

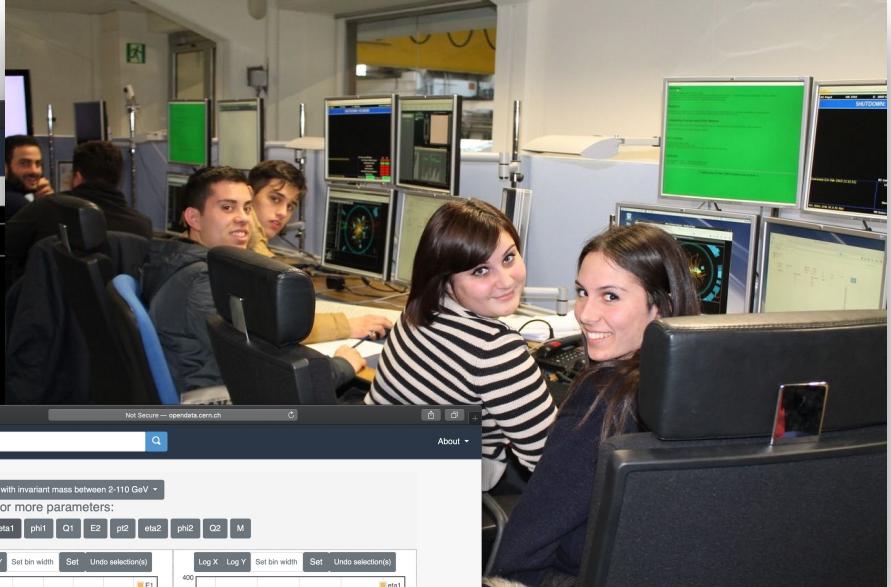
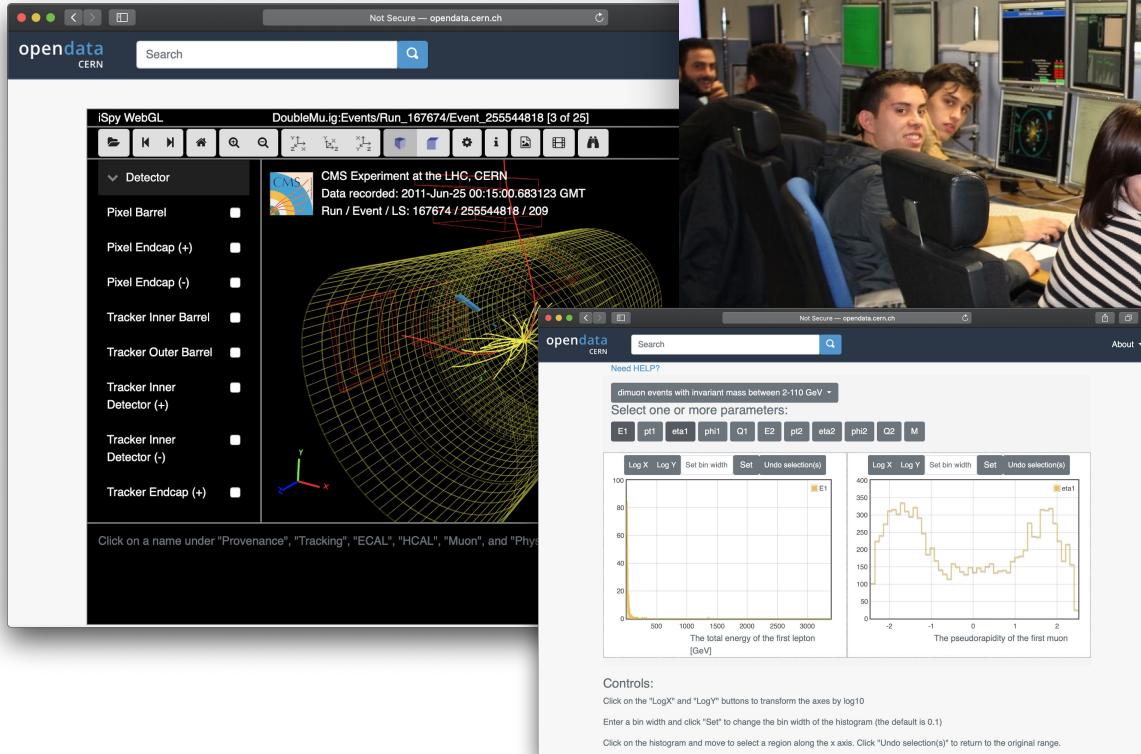
This screenshot shows a file download page for the configuration file "Configuration file for HLT step /cdaq/physics/Run2010Hi/v1.5/MILHTV1". The file is 1.4 MB and can be downloaded in compressed (.tar.gz) or raw (.root) formats.

Faceted search

Large file download $O(GB)$

JSON Schema Data provenance

Education use cases



<https://cds.cern.ch/record/1994217>

Research use cases

The figure displays four research papers arranged in a grid:

- Top Left:** "Non-Standard Sources of Parity Violation in Jets and a First Search at $\sqrt{s}=8$ TeV with CMS Open Data" by Christopher G. Lester^a, Matthias Schott^{b,c}.
Abstract: The Standard Model violates parity, but invisible to Large Hadron Collider (LHC) experiments (e.g. state polarization or spin-sensitivity in the detectors). New could potentially violate parity in ways which are detected. If there are sources of parity violation other than LHC center-of-mass scattering. We probe the feasibility of such measurements data which was recorded in 2012 by the CMS collaboration CMS Open Data initiative. In particular, we test an inclusion is primarily sensitive to non-standard parity violating effects. Within our measurements, no significant deviation from no obvious experimental limitations have been found. We find that non-standard parity violation could be performed, not very different sorts of models to those which we measure our initial studies provide a valuable starting point for full LHC datasets at 13 TeV with a careful and less conservative uncertainties.
- Top Right:** "Jet Substructure Studies with CMS Open Data" by Aashish Tripathie^{1,2}, Wei Xie^{1,3}, Andrew Larkoski^{2,4}, Simone Marzani^{2,5}, and Jesse Thaler^{1,6}.
Abstract: We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on 31.8 pb^{-1} of 7 TeV proton-proton collisions recorded at the Large Hadron Collider. We use a soft drop jet clustering algorithm to identify jets with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the 2-prong substructure of the leading jet using soft drop deconvolution. We compare the distributions of the two-prong mass and the two-prong radius and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations informed by modified leading-logarithmic accuracy. Although the 2010 CMS Open Data does not include a jet substructure analysis, we use it to validate our analysis and to obtain observables to validate these substructure studies.
- Bottom Left:** "Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum" by Cari Cesario^{1,2}, Yotam Soreq^{2,3,4}, Matthew J. Strassler^{1,3}, Jesse Thaler^{4,1,3}, and Wei Xu^{1,3}.
Abstract: We study dimuon events in 2.11 fb^{-1} of 7 TeV $p\bar{p}$ collisions, using CMS Open Data, and search for narrow dimuon resonance with moderate mass (14–66 GeV) and substantial transverse momentum. Applying dimuon ρ cuts of 25 GeV and 60 GeV, we explore two overlapping samples: one with low masses, and one with prompt muons without an isolation requirement. Using the latter sample to obtain information about detector effects and QCD backgrounds, which we obtain directly from the CMS Open Data. We present model-independent limits on the product of cross section, branching fraction, acceptance, and efficiencies. These limits are stronger, relative to a corresponding inclusive search without a ρ cut, by factors of as much as nine. Our ρ_3 -enhanced dimuon search strategy provides improved sensitivity in models in which a new particle is produced mainly in the decay of something heavier, as could occur, for example, in decays of the Higgs boson or of a TeV-scale top squark. An implementation of this method with the current 13 TeV data should improve the sensitivity to such signals further by roughly an order of magnitude.
- Bottom Right:** "Exposing the QCD Splitting Function with CMS Open Data" by Andrew Larkoski^{1,2}, Simone Marzani^{2,3}, Jesse Thaler^{1,3}, Aashish Tripathie^{1,4}, and Wei Xie^{1,5}.
Abstract: 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 is often suppressed by a small parameter, the anomalous 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 in QCD. In this letter, we use the 2-prong substructure of jets recorded in the CMS Open Data to study the $1 \rightarrow 2$ splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

Footnotes and references are visible in the bottom right corner of each paper's page.

Analysis examples

code
+
data
+
environment

<http://opendata.cern.ch/record/5500>

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 Workflow CMS CERNLHC

Description

This research level example is a strongly simplified reimplementations 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 level of this example addresses users who feel they have at least some minimal understanding of the content of this paper and of the meaning of this reference plot, which can be reached via (separate) educational exercises. The lower levels might also be interesting for educational applications. The example requires a minimal acquaintance with the linux operating system and the ROOT analysis tool.

The example uses legacy versions of the original CMS data sets in the CMS AOD, which slightly differ from the ones used for the publication due to improved calibrations. It also uses legacy versions of the corresponding Monte Carlo simulations, which are again close to, but not identical to, the ones in the original publication. These legacy data and MC sets listed below were used in practice, as they are in many later CMS publications.

Since according to the CMS Open Data policy the fraction of data which are public (and used here) is reduced with respect to what can be achieved with the full dataset 30-61, arXiv:1207.7235, was also obtained with only part of the Run 1 statistics, roughly equivalent partial statistical overlap.

The provided analysis code recodes the spirit of the original analysis and recodes many of the details of the original analysis code itself. Also, for the sake of simplicity, it skips some of the more complex parts of the original analysis. Nevertheless, it provides a qualitative insight about how the original result was obtained. In addition, the code also contains many undocumented plots which grew as a side product from setting up this example.

Events / 3 GeV

CMS Preliminary $\mathcal{S} = 7 \text{ TeV}, L = 5.0 \text{ fb}^{-1}, \sqrt{s} = 8 \text{ TeV}, L = 5.26 \text{ fb}^{-1}$

• Data

Z γ + X

Z γ ZZ

m $_h$ =126 GeV

Events / 3 GeV

CMS Open Data

• Data

Z γ + X

T \bar{T} Bar

ZZ > 4l

m $_h$ = 125 GeV

9

The four questions

where is data?

hard drive, distributed storage

where is the code?

GitLab, local copy, email

what environment do you use?

my own laptop, remote server

what workflow do you use?

Interactive commands, bash script, README file

Reproducible analyses

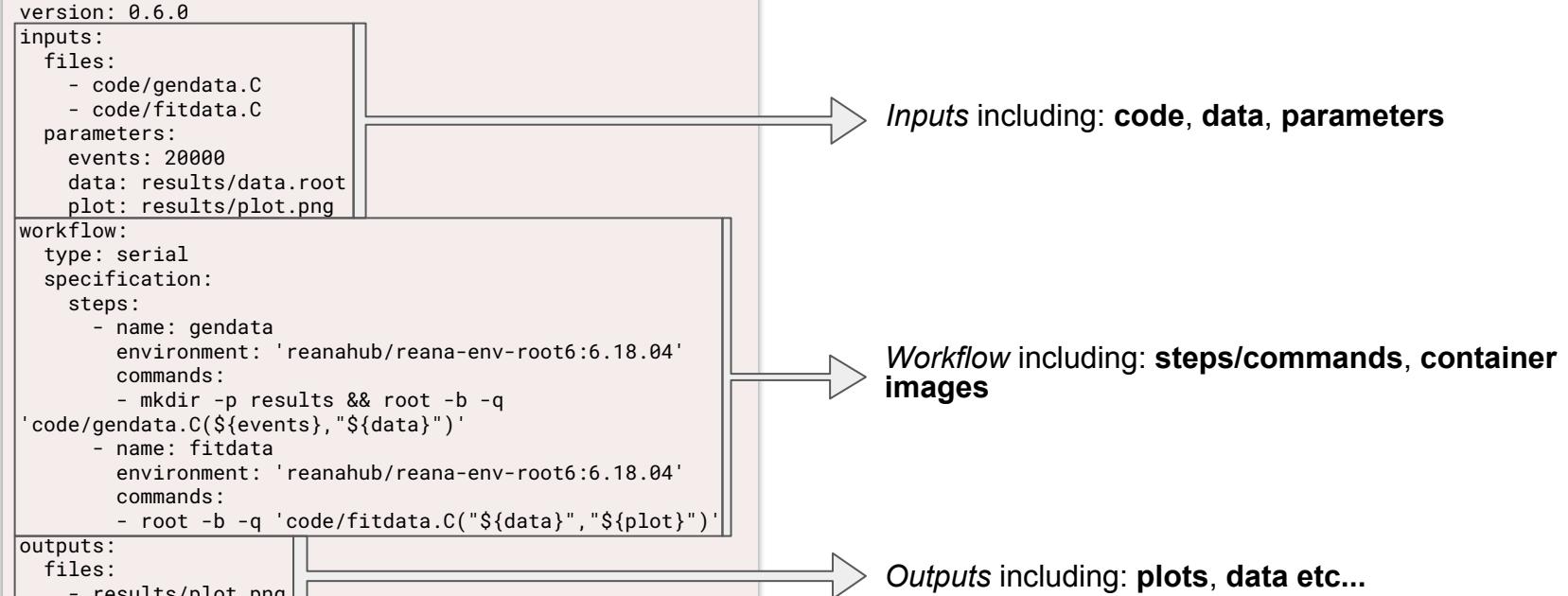
The screenshot shows the official website for REANA at <http://reana.io>. The page features a large "reana" logo with "re" in red and "ana" in dark blue. Below the logo, the tagline "Reproducible research data analysis platform" is displayed. The page is divided into four main sections: "Flexible" (Run many computational workflow engines, supported by Common Workflow Language and Apache Airflow icons), "Scalable" (Support for remote compute clouds, supported by Kubernetes icon), "Reusable" (Containerise once, reuse elsewhere. Cloud-native, supported by Docker and Kubernetes icons), and "Free" (Free Software. MIT licence. Made with ❤️ at CERN, supported by CERN icon).

<http://reana.io/>

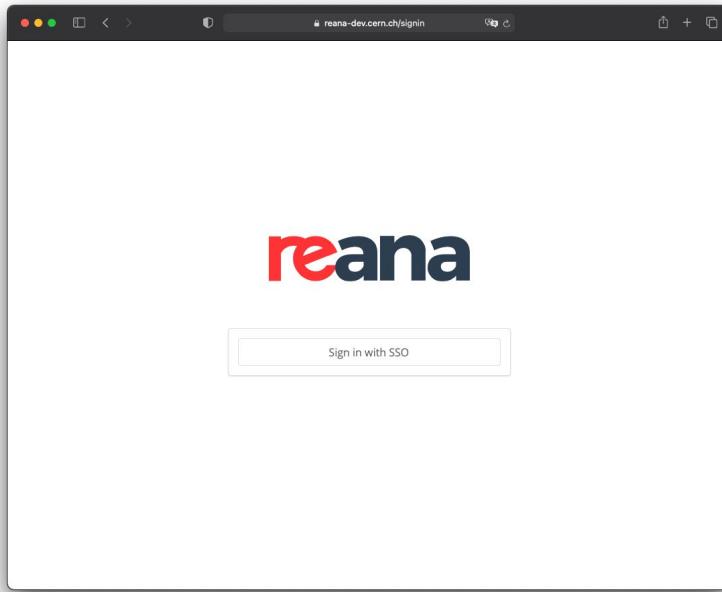
The screenshot shows the GitHub organization page for REANA at <https://github.com/reanahub>. The page title is "REANA" and the description is "Reproducible and reusable research data analysis platform". It lists several pinned repositories: "reana" (Python, 35 stars, 17 forks), "reana-client" (Python, 3 stars, 13 forks), and "reana-cluster" (Python, 4 stars, 10 forks). Below these are other repositories like "reana-job-controller" and "reana-cluster". A "Top languages" section shows Python, C, JavaScript, and C++ as the most used languages. A "People" section shows a list of contributors.

<https://github.com/reanahub>

Example



Example



A screenshot of a web browser showing the REANA profile page at <https://reana.cern.ch/profile>. The page has a header with the REANA logo and navigation links for "Help" and "Profile". The main content includes sections for "Your REANA token" (with instructions and a code snippet), "Your GitLab projects" (listing "REANA example - ROOT6 and RooFit" by "rodrigdl/reana-demo-root6-roofit"), and "Your quota" (showing CPU usage of 5h 59m 43s and HDD usage of 5.02 MiB). At the bottom, there are links for "Docs" and "Forum".

Example

```
Terminal
$ reana-client create -w rootfit
rootfit.7
$ reana-client upload -w rootfit
file /code/gendata.C was successfully uploaded.
file /code/fitdata.C was successfully uploaded.
$ reana-client start -w rootfit
rootfit has been queued
$ reana-client status -w rootfit
NAME      RUN_NUMBER    CREATED        STARTED        STATUS
rootfit 7   2021-02-09T08:45:04  2021-02-09T08:46:20  running
$ reana-client ls -w rootfit
NAME      SIZE  LAST-MODIFIED
code/gendata.C  1937  2021-02-09T08:45:17
code/fitdata.C  1648  2021-02-09T08:45:17
$ reana-client status -w rootfit
NAME      RUN_NUMBER    CREATED        STARTED        ENDED          STATUS      PROGRESS
rootfit 7   2021-02-09T08:45:04  2021-02-09T08:45:20  2021-02-09T08:45:48  finished  2/2
$ reana-client ls -w rootfit | grep plot
results/plot.png  15450  2021-02-09T08:45:43
$
```

```
Terminal
$ reana-client logs -w rootfit
=> Workflow engine logs
2021-02-09 08:45:33.723 | root | MainThread | INFO | Publishing step:0, cmd: mkdir -p results && root -b -q 'code/gendata.C(20000,"results/data.root")', total steps 2 to MQ
2021-02-09 08:45:39.827 | root | MainThread | INFO | Publishing step:1, cmd: root -b -q 'code/fitdata.C("results/data.root","results/plot.png")', total steps 2 to MQ
2021-02-09 08:45:48.865 | root | MainThread | INFO | Workflow 5958a639-32b5-45d3-b6d0-215896b26692 finished. File s available at /var/reana/users/444eb8dc-968c-454c-a3ca-4faec439fc82/workflows/5958a639-32b5-45d3-b6d0-215896b26692
92.

=> Job logs
=> Step: gendata
=> Workflow ID: 5958a639-32b5-45d3-b6d0-215896b26692
=> Compute backend: Kubernetes
=> Job ID: reana-run-job-b4046e72-c5f0-4db0-89ca-c1a5c38b1e95
=> Docker image: reanahub/reana-env-root6:6.18.04
=> Command: mkdir -p results && root -b -q 'code/gendata.C(20000,"results/data.root")'
=> Status: finished
=> Logs:
job: :

| Welcome to ROOT 6.18/04 https://root.cern |
| (c) 1995-2019, The ROOT Team |
| Built for linuxx86_64gcc on Jan 08 2020, 14:10:00 |
| From tags/v6-18-04@v6-18-04 |
```

Example

The image shows a screenshot of the reana.cern.ch web interface, demonstrating a workflow management system.

Left Panel: Your workflows

- roofit #7** (finished 8 hours ago)
- roofit #6** (finished 8 hours ago)
- roofit #5** (finished 14 days ago)

Middle Panel: Workflow Details

Workflow ID: 5958a639-32b5-45d3-b6d0-215890b26692

Step 1: gendata (finished in 28 seconds, step 2/2)

Logs:

```
job: :  
-----  
| Welcome to ROOT 6.18/04  
|  
| Built for linuxx86_64gcc  
| From tags/v6-18-04#v6-18-04  
| Try '.help', '.demo', '.lis
```

Step 2: fitdata (finished in 28 seconds, step 2/2)

Logs:

```
Processing code/gendata.C(2000)  
[1]RooFit v3.60 -- Developed by Wouter Verkerke --
```

Workspace

Name	Modified
code/gendata.C	2021-02-09T08:45:17
code/fitdata.C	2021-02-09T08:45:17
results/plot.png	2021-02-09T08:45:43
results/data.root	2021-02-09T08:45:35

Specification

Right Panel: Results

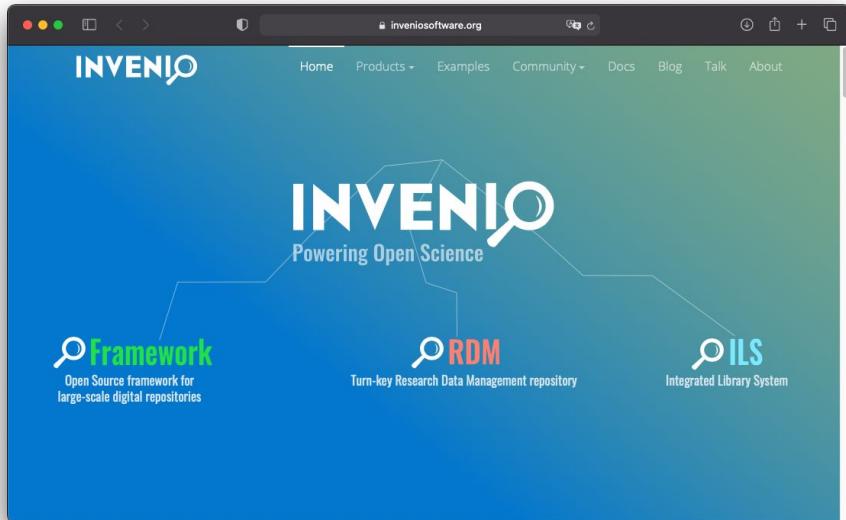
Plot: results/plot.png

Fit example

A histogram showing Events / (0.1) vs x. The x-axis ranges from 0 to 10, and the y-axis ranges from 0 to 1000. The plot shows a sharp peak at x ≈ 5.5, with a fitted curve overlaid.

Download

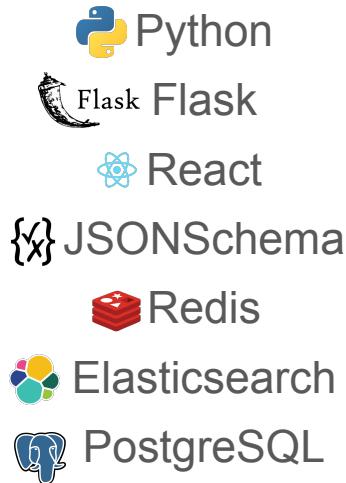
Technology: repository



<https://inveniosoftware.org>

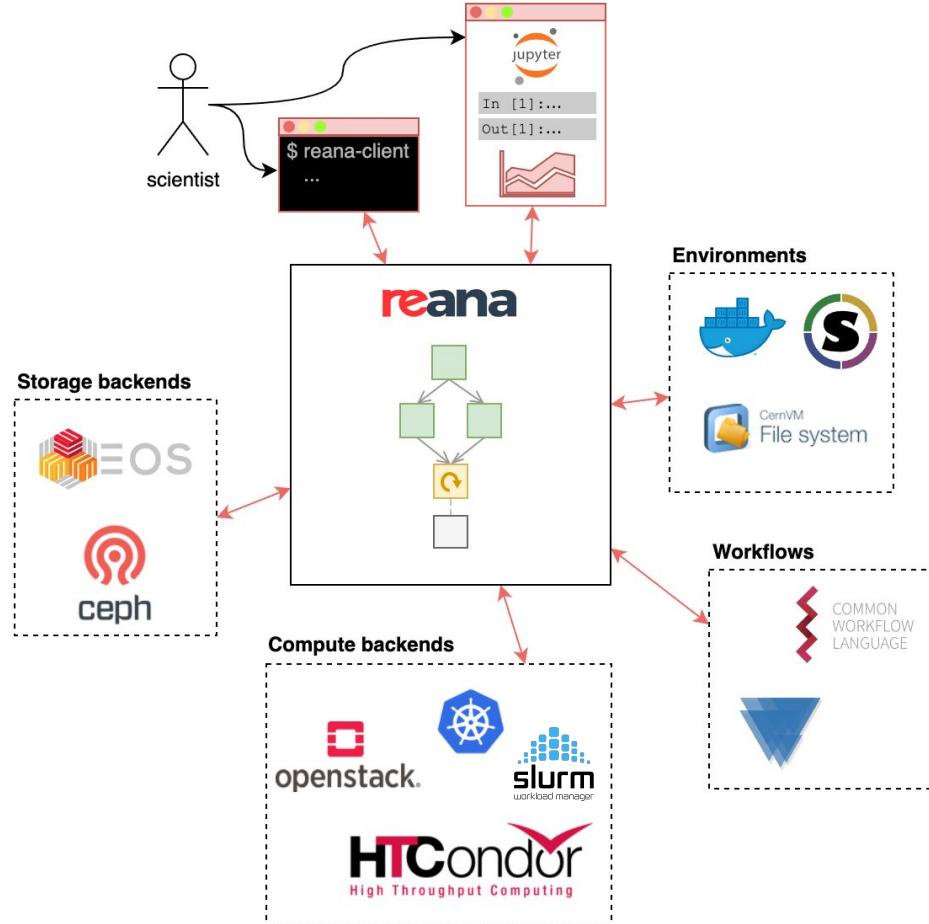
The technology behind:

<https://www.hepdata.net/>
<https://inspirehep.net/>
<https://zenodo.org>



Technology: REANA

- Cloud-native application
- Extensible
 - Storage backends
 - Compute backends
 - Container technologies
 - Workflow engines



Try them out!

Use opendata.cern.ch

opendata CERN

Sort by: Best match asc 20 results

Found 4098 results.

CMS 2011 Virtual Machines: How to install

The CMS-specific VM includes the ROOT framework and CMSSW. Follow the instructions below to setup a CERN Virtual Machine on your computer for 2011-2012 CMS Open Data. Then, go to Getting Started with CMS data.

CMS Guide to research use of CMS Open Data

If you are interested in finding hints, tips and guidance for conducting a research-oriented analysis using CMS Open Data, please see our notes on this page.

Getting Started with CMS 2011 Open Data

To analyse CMS data collected in 2011 and 2012, you need version 5.3.32 of CMSSW, supported only on Scientific Linux 6. If you are unfamiliar with Linux, take a look at this short introduction to Linux or try this...

Documentation Guide CMS Getting Started

Dataset Collision Derived Simulated Documentation About Activities Authors Guide Help Policy Report Environment Condition Validation Glossary News Software Analysis Framework

Install REANA on premises/locally

```
$ helm repo add reanahub \
  https://reanahub.github.io/reana
```



```
$ helm repo update
```

```
$ helm install reana reanahub/reana
```

Helm repository at <http://reanahub.github.io/reana>, more documentation at <http://docs.reana.io/development/deploying-at-scale/>

reana

RIVER now ostensibly running REANA

- As a start, need to test things

Name	READY	Status	RESTARTS	Age
reana-cache-88b76b854-2b9c1	1/1	Running	0	267h
reana-db-5f45666d85-bm49v	1/1	Running	0	267h
reana-worker-6855778564-2sg7r	1/2	Running	0	267h
reana-server-6855778564-7sqn2	2/2	Running	0	267h
reana-worker-6855778564-2sg7r	2/2	Running	0	267h

Sign up

What's next



Roadmap

Near-term

What we plan to work on next

Live logs

Introduce live job log streaming for CLI and Web UI.

LHC community

Introduce abstract dataset concept to handle a set of related files.
Use various remote storage backends for workflow workspace.

Future

What is coming later

Abstraction of data storage

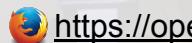
Use various remote storage backends for workflow workspace.

User groups and authorisations

Introduce OpenID Connect to support more authentication mechanisms.
Introduce user groups and role-based authorisation control models.

Get in touch

CERN Open Data



<https://opendata.cern.ch>



<https://github.com/cernopendata>



<https://forum.opendata.ch/>



<https://gitter.im/cernopendata/opendata.cern.ch>



<https://twitter.com/cernopendata>

REANA



<https://www.reana.io>



<https://github.com/reanahub/reana>



<https://forum.reana.io/>



<https://gitter.im/reanahub/reana>



<https://twitter.com/reanahub>



<https://docs.reana.io>