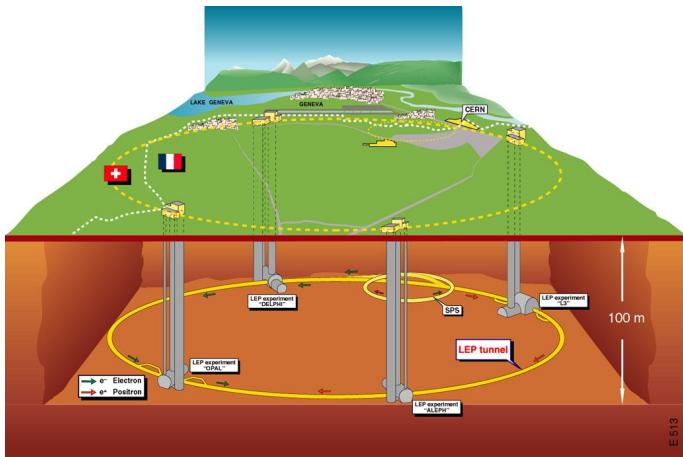


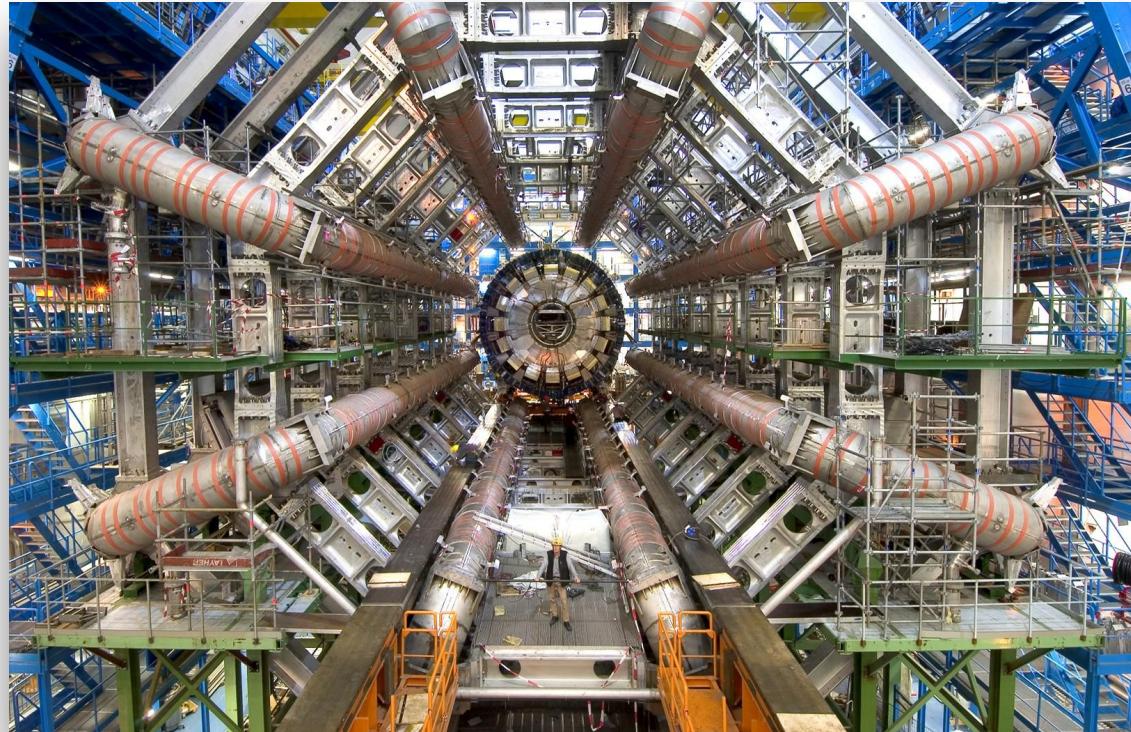
# Open and Reproducible Research Services in LHC Particle Physics

Diego Rodríguez  
CERN

# CERN Large Hadron Collider

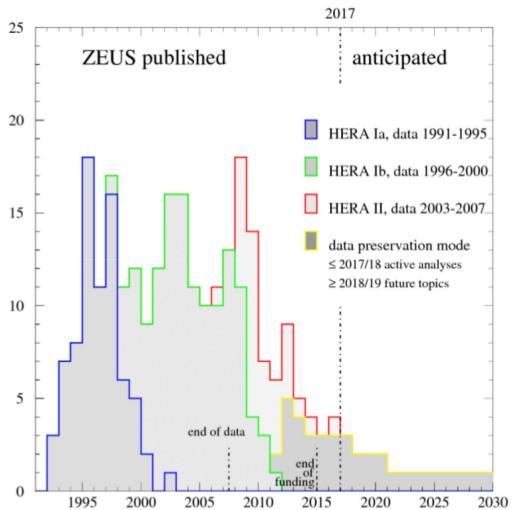


<http://cds.cern.ch/record/842153>



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

# Data and knowledge



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



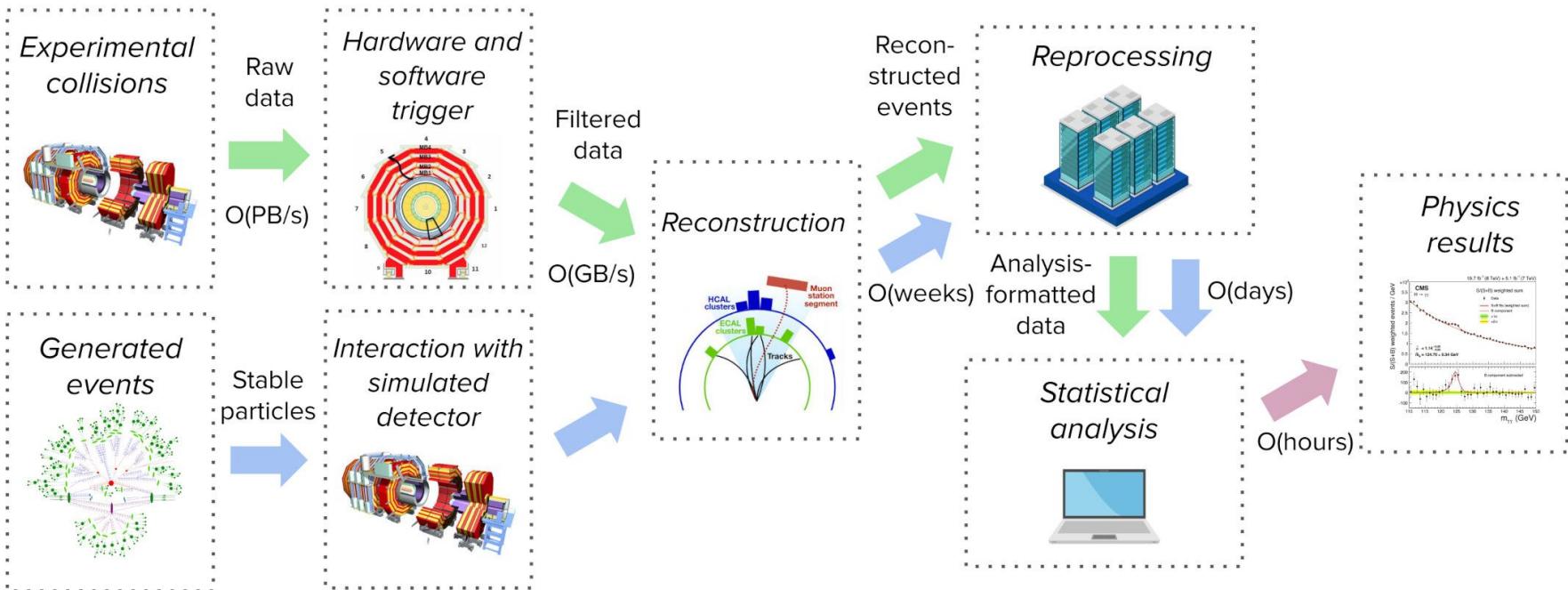
**Patrick Koppenburg** @PKoppenburg · 4 sept.

The @LHCbExperiment collaboration submitted its first physics paper 10 years ago. It had 629 authors [arxiv.org/abs/1008.3105](https://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



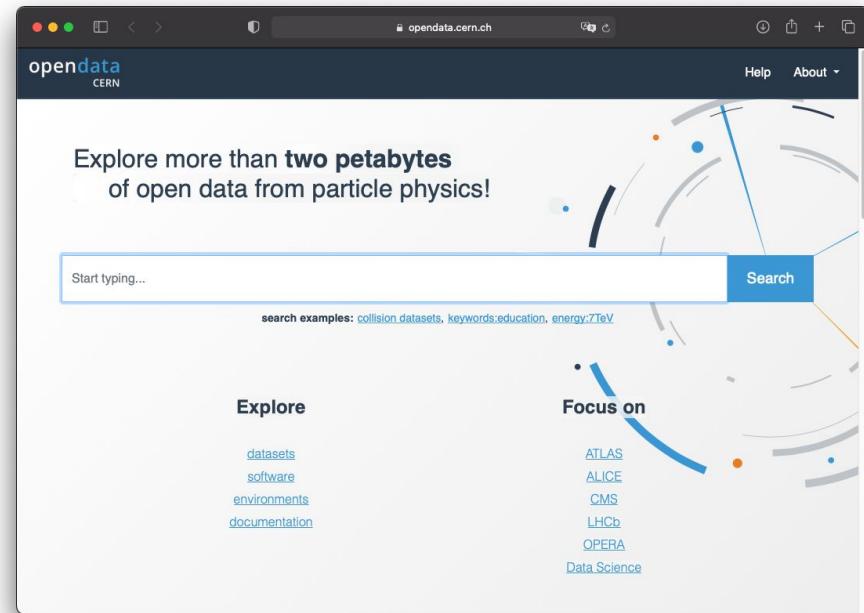
# CERN Open Data portal

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:

- Screenshot 1 (Left):** A detailed view of a dataset record. It includes sections for "Description", "Related datasets", "Dataset characteristics", and "Dataset semantics". The "Description" section contains a link to "How were these data selected?", which provides a detailed explanation of the selection criteria.
- Screenshot 2 (Middle):** A search results page showing a list of 1170 results. The results are filtered by type (e.g., Dataset, Documentation, Environment, Software) and experiment (e.g., ALICE, CMS, OPERA). The results are sorted by best match and displayed in a detailed view with 20 results per page.
- Screenshot 3 (Right):** A file download interface for a configuration file named "Razor filter and analyzer for SUSY searches". The interface shows the file size as 1.4 MB and provides download links for "Supplementary", "Configure", and "CMS".

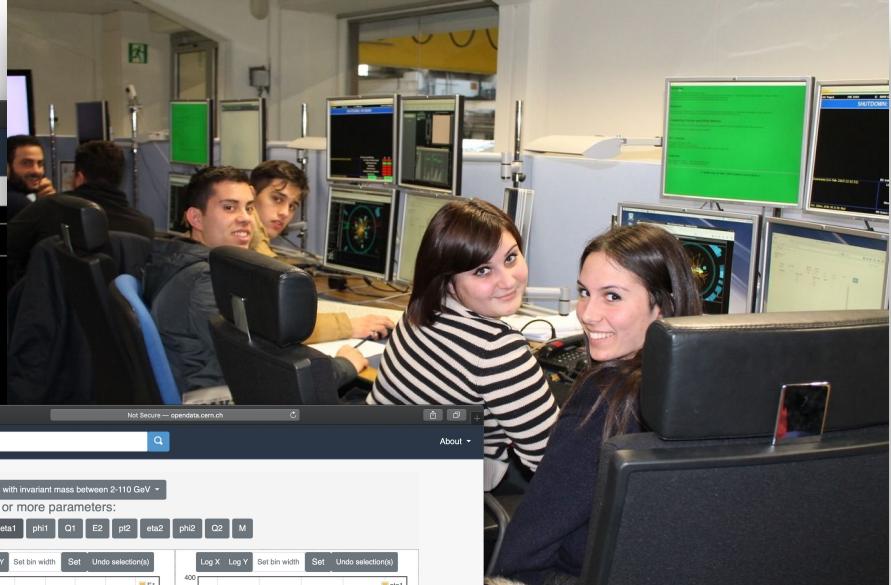
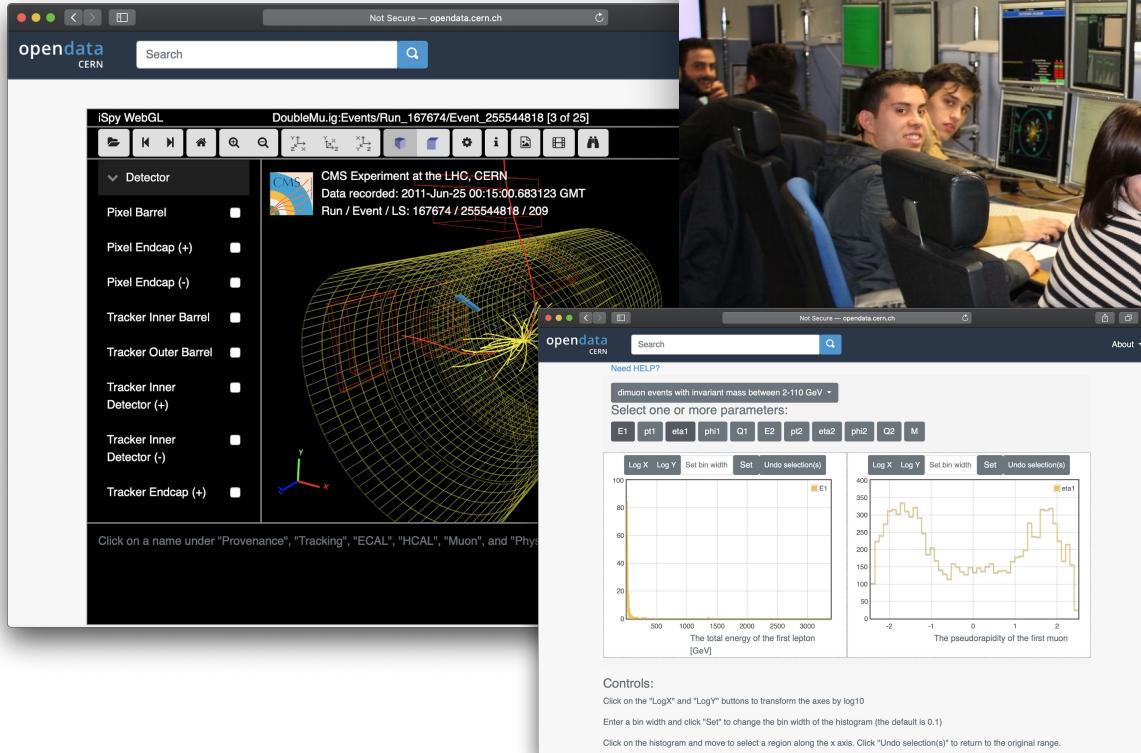
Faceted search

Large file download 0(GB)

JSON Schema

Data provenance

# Education use cases



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

# Research use cases

The figure displays four separate windows from a Mac OS X desktop environment, illustrating various research use cases:

- Top Left Window:** A PDF viewer showing arXiv:1904.11195v1 [hep-ex] 29 Apr 2019. The title is "Non-Standard Sources of Parity Violation in Jets and a First Search at  $\sqrt{s}=8$  TeV with CMS Open Data". It features a standard LaTeX-style document structure with sections like Abstract, Introduction, and References.
- Top Right Window:** A PDF viewer showing arXiv:1704.05842.pdf (page 1 of 35). The title is "Jet Substructure Studies with CMS Open Data". It includes a table of contents and several sections of text and figures related to jet substructure analysis.
- Bottom Left Window:** A PDF viewer showing arXiv:1902.04222.pdf (page 1 of 37). The title is "Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum". It discusses the search for dimuon resonances, mentioning the Harvard University and CERN collaborations.
- Bottom Right Window:** A PDF viewer showing arXiv:1704.05066.pdf (page 1 of 8). The title is "Exposing the QCD Splitting Function with CMS Open Data". It focuses on exposing the splitting function in QCD, comparing it with previous work and discussing its implications for particle physics.

Below the windows, a caption reads: "The core of our analysis is based on soft drop declustering. DELPHES fast simulation is used to generate jets with varying multiplicities and momenta. The core of our analysis is based on soft drop declustering. DELPHES fast simulation is used to generate jets with varying multiplicities and momenta."

**Footnotes (from bottom left window):**

- <sup>1</sup>cesanott@harvard.edu
- <sup>1</sup>yotan.soreq@cern.ch
- <sup>1</sup>strasser@physics.harvard.edu
- <sup>1</sup>phabre@mit.edu
- <sup>1</sup>wei.xue@cern.ch

**Footnotes (from top right window):**

- <sup>1</sup>Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
- <sup>2</sup>Physics Department, Reed College, Portland, OR 97202, USA
- <sup>3</sup>University of New York, Stony Brook, NY 11794-3630, USA

**Footnotes (from bottom right window):**

- <sup>1</sup>Physics Department, Boston University of New York, Buffalo, 1426 Delaware Avenue, Buffalo, NY 14214, USA
- <sup>2</sup>Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

**Page Number:** 8

# 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 CERN-LHC

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](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 files also contain many undocumented plots which grew as a side product from setting up this example.

Events / 3 GeV

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

• Data

Z + X

Z $_t$  ZZ

m<sub>h</sub>=126 GeV

Events / 3 GeV

CMS Open Data

• Data

Z $_t$  + X

T $t\bar{b}$

ZZ > 4l

m<sub>h</sub> = 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

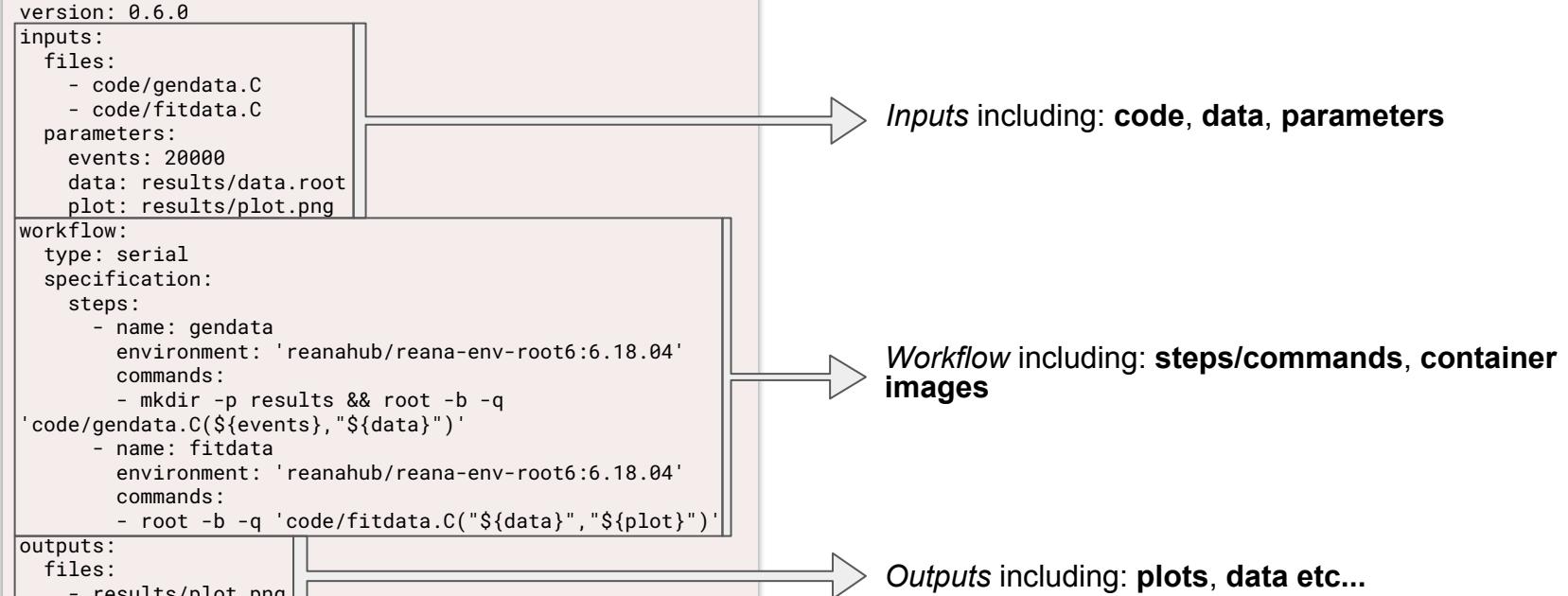
The screenshot shows the official website for REANA at <http://reana.io>. The page features a large "reana" logo with "re" in red and "ana" in dark blue. Below the logo, the tagline "Reproducible research data analysis platform" is displayed. The page is divided into four main sections: "Flexible" (Run 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/>

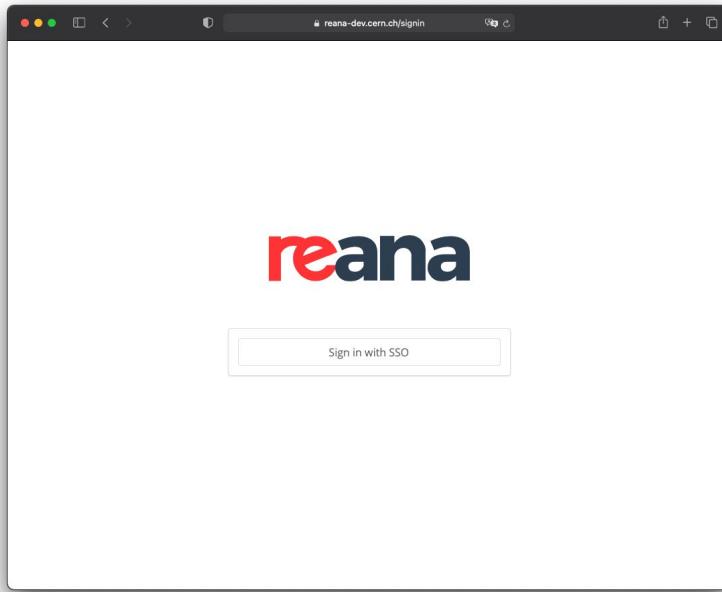
The screenshot shows the GitHub organization page for REANA at <https://github.com/reanahub>. The page title is "REANA" and the description is "Reproducible and reusable research data analysis platform". It lists several pinned repositories: "reana" (Python, 35 stars, 17 forks), "reana-client" (Python, 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 grid of user profiles.

<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**

**Job #7: roofit** (finished in 28 seconds, step 2/2)

- Logs:

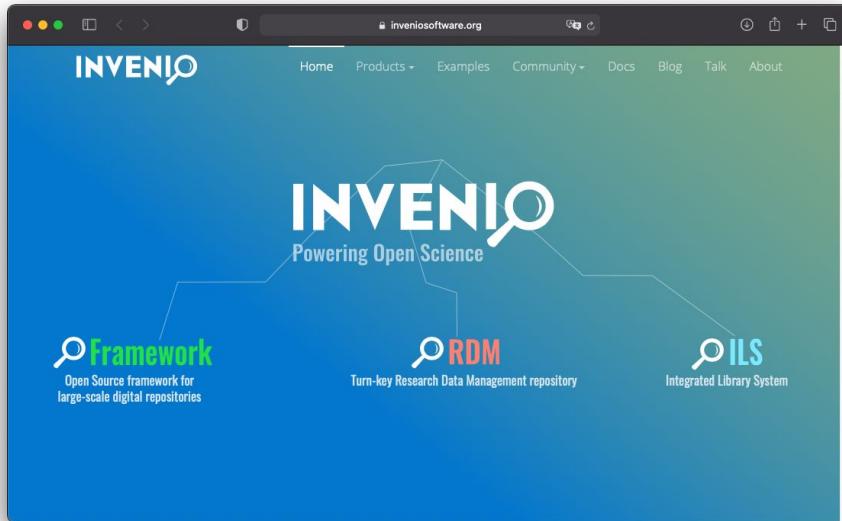
```
I Welcome to ROOT 6.18/04
|
I Built for linuxx86_64gcc on 2021-02-09T08:45:17
I From tags/v6-18-04#v6-18-04
I Try '.help', '.demo', '.ls'
```
- Workspace:
  - code/gendata.C (2002-02-09T08:45:17)
  - code/fitdata.C (2021-02-09T08:45:17)
  - results/plot.png (2021-02-09T08:45:43)
  - results/data.root (2021-02-09T08:45:35)
- Specification:

**Right Panel: Results**

**Fit example**

A histogram titled "Fit example" showing Events / (0.1 GeV) versus x. The plot displays a peak around x=5.5. The data points are black dots, and the fit is a blue line.

# Technology: repository



<https://inveniosoftware.org>

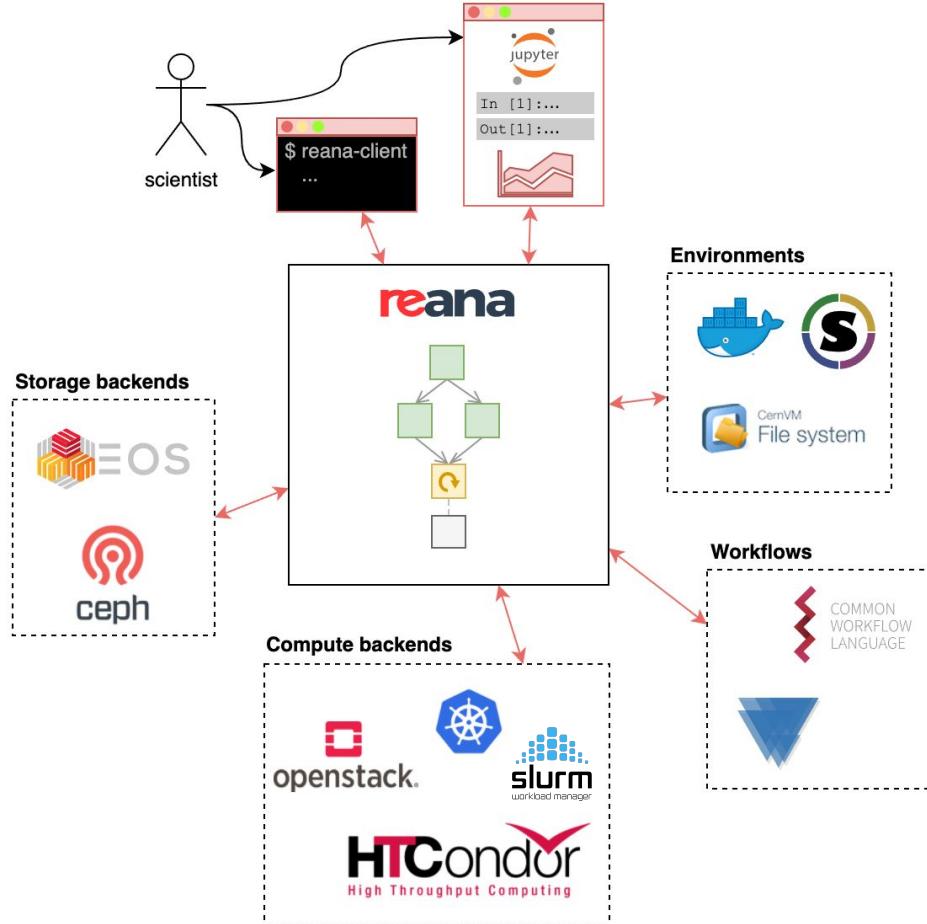
The technology behind:

<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](https://opendata.cern.ch)

A screenshot of a web browser showing the opendata.cern.ch search results for "cms virtual machine". The search bar at the top contains "opendata.cern.ch/search?page=1&size=20&q=cmssw%20virtual%20machine". The results page displays a list of 4098 items, sorted by best match. The left sidebar shows a navigation menu with categories like Dataset, Documentation, Environment, Software, and Framework, each with a list of sub-items and their counts. The main content area shows two sections: "CMS 2011 Virtual Machines: How to install" and "CMS Guide to research use of CMS Open Data". Both sections provide links to "Documentation", "Guide", "CMS", and "Getting Started".

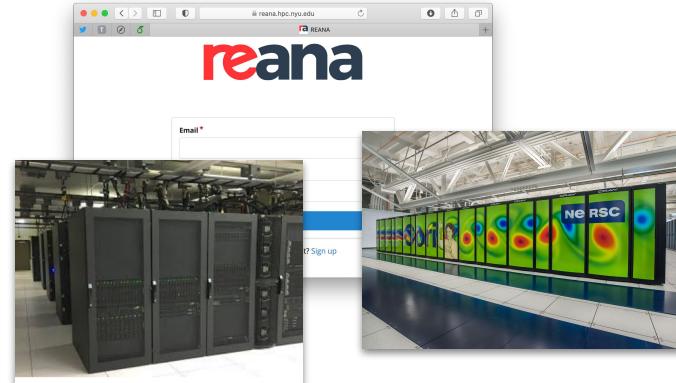
Install REANA on premises/locally

```
$ 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/>



# 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

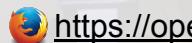
Use various remote storage backends for workflow workspace.

#### 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>