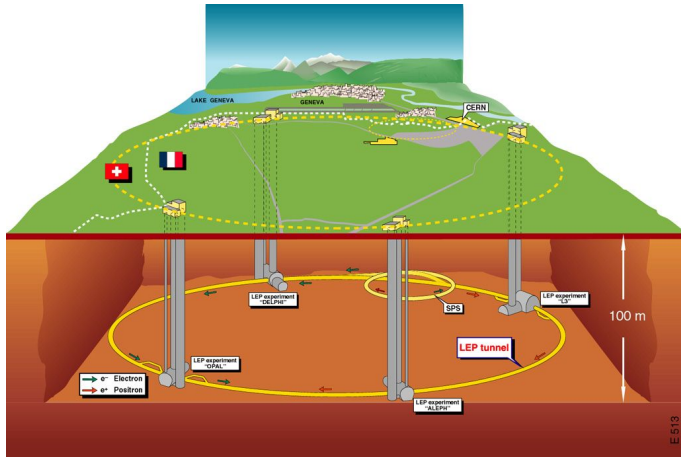


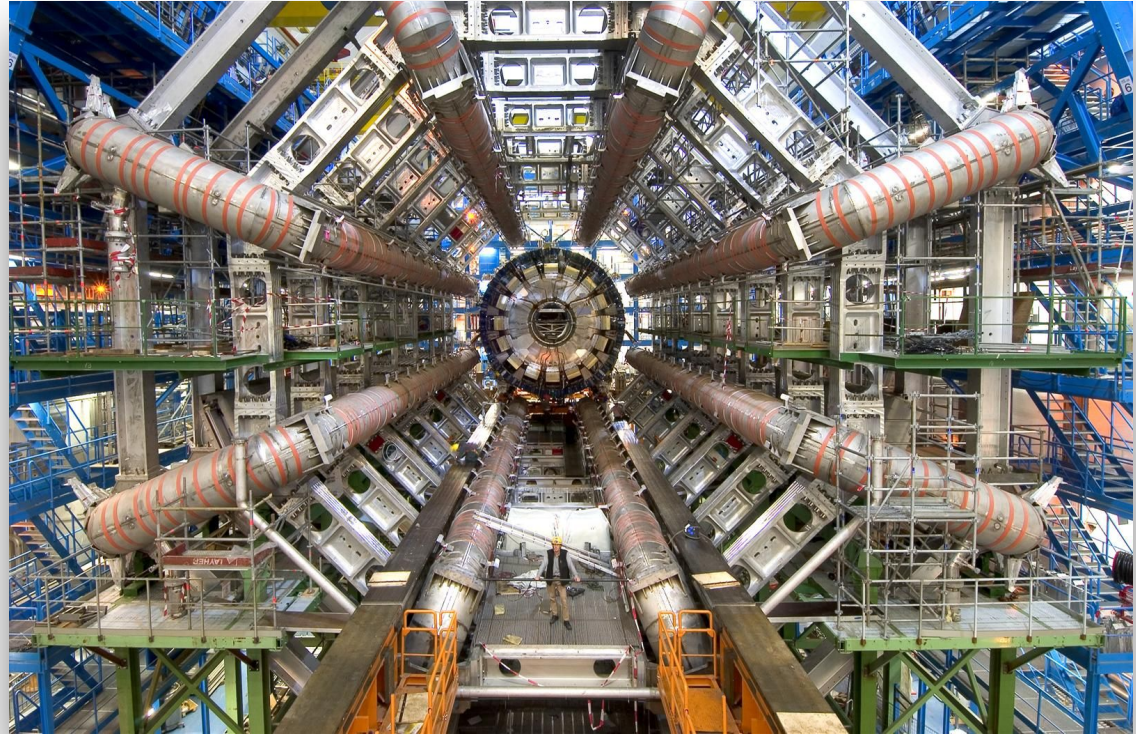
Open and Reproducible Research Services in LHC Particle Physics

Diego Rodríguez
CERN

CERN Large Hadron Collider

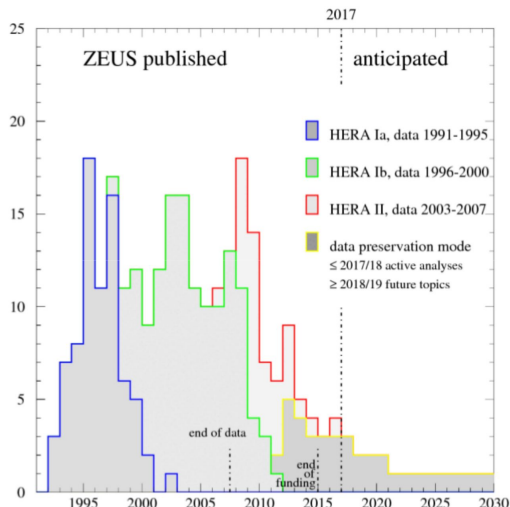


<http://cds.cern.ch/record/842153>



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

Data and knowledge



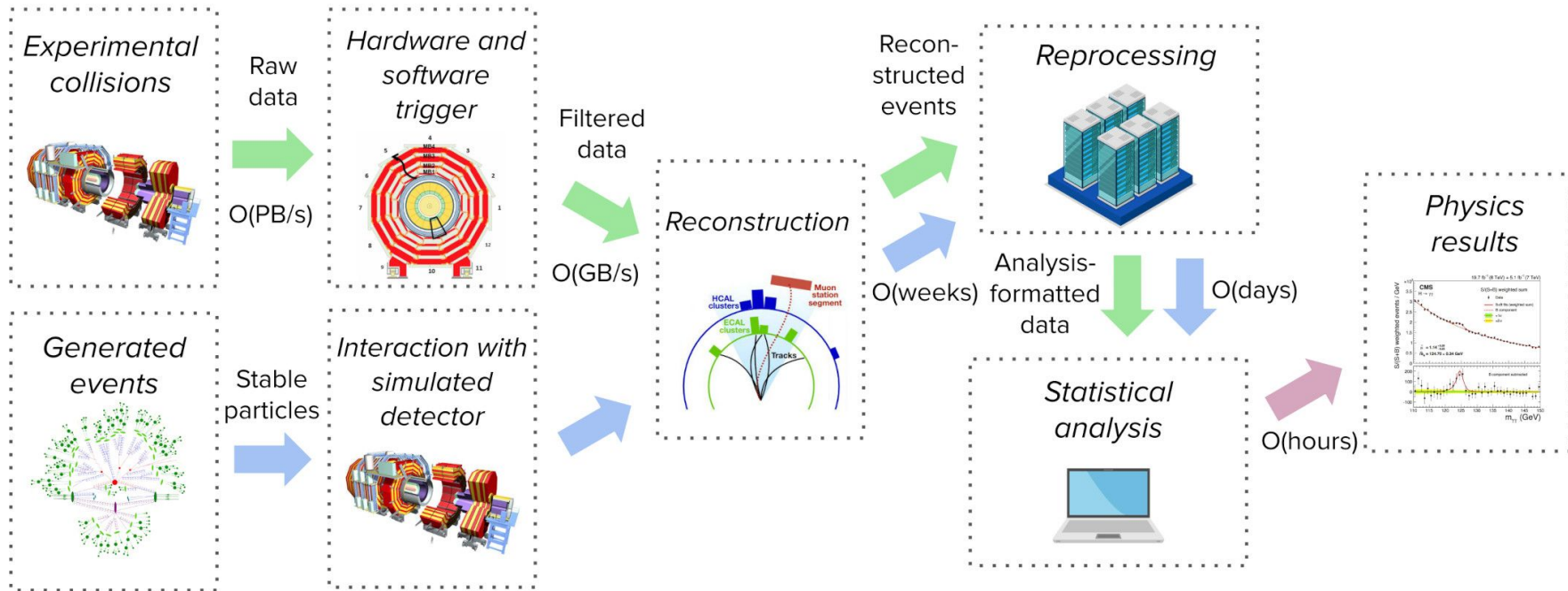
Achim Geiser <https://indico.cern.ch/event/588219>

Patrick Koppenburg @PKoppenburg · 4 retweets · 19 replies

The @LHCbExperiment collaboration submitted its first paper 10 years ago. It had 629 authors [arxiv.org/abs/hep-ex/08.3135](#). Now we are 972, and only half the number of authors are still with us. They are in violet in the list below.

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

HEP data analyses



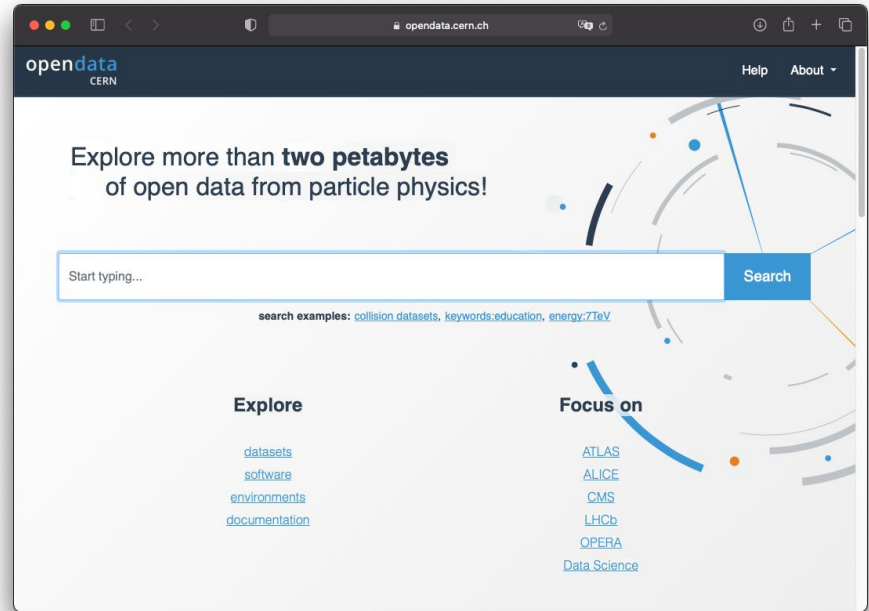
CERN Open Data portal

Launched in 2014

Disseminating over 2.4 PB of data

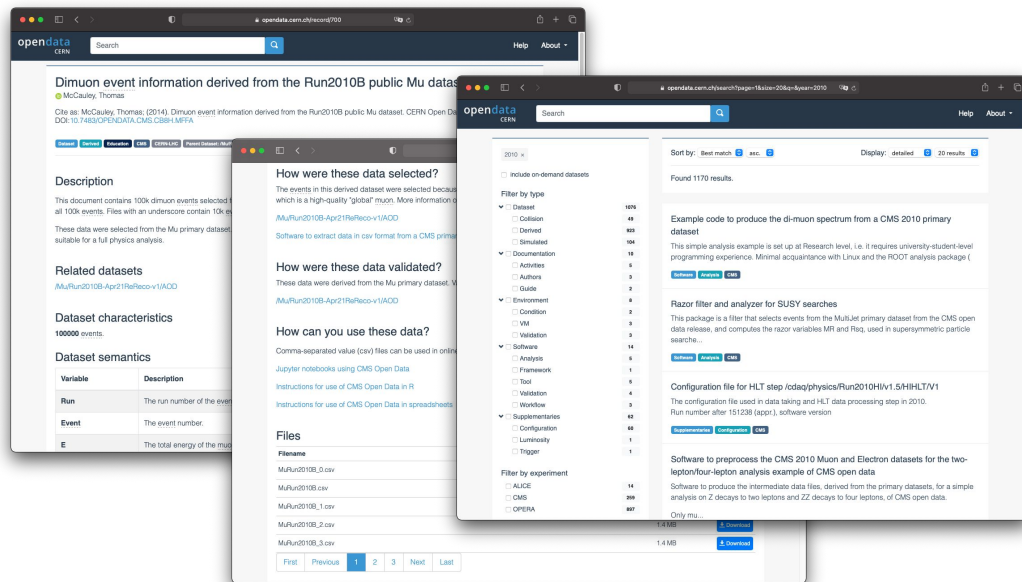
7.500 records

900.000 files



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

Information organisation



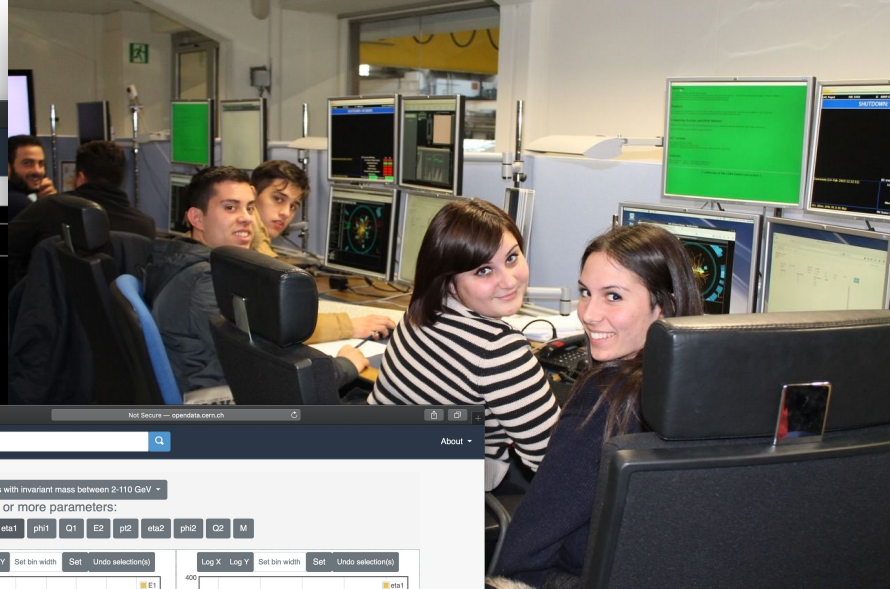
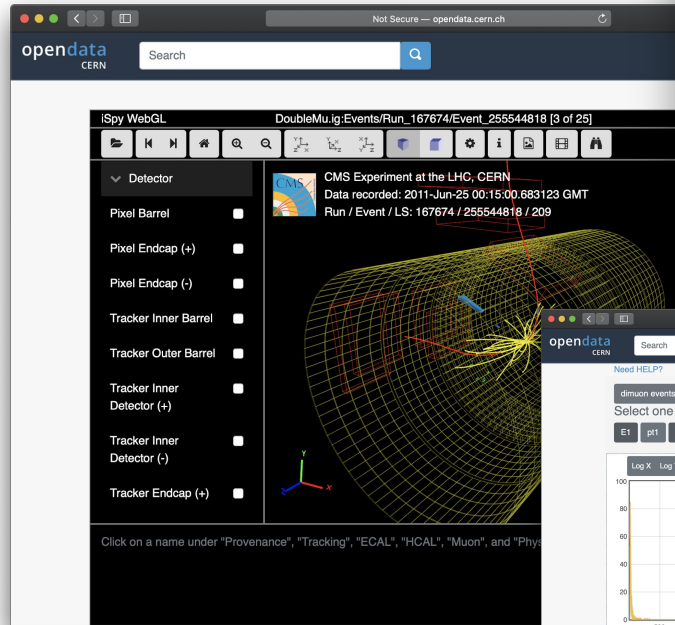
Faceted search

Large file download $O(\text{GB})$

JSON Schema

Data provenance

Education use cases



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

Research use cases



Non-Standard Sources of Parity Violation in Jets and a First Search at $\sqrt{s}=8$ TeV with CMS Open Data

Christopher G. Lester¹ Matthias Schott^{2,*}

¹*Cambridge Laboratory, University of Cambridge, UK*

²*Massachusetts Institute of Technology, Cambridge, USA*

³*Johannes Gutenberg-University, Mainz, Germany*

E-mail: lesterthep.phy.cam.ac.uk, matthias.schott@mit-cp-ctpp.org

ABSTRACT: The Standard Model violates parity, but invisible to Large Hadron Collider (LHC) experiments (e.g. state polarisation or spin-sensitivity in the detectors). New could potentially violate parity in ways which are detectable. If those sources of new physics occur only at LHC energy scales. We probe the feasibility of such measurements data which was revealed 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 for non-standard parity violation could be performed, not to very different sets of models to those which our measurements. Our initial studies provide a valuable starting point for full LHC datasets at 13 TeV with a careful and less conservative analysis.



Jet Substructure Studies with CMS Open Data

Aashish Tripathi^{1,2} Wei Xue^{1,3} Andrew Larkoski^{2,3} Simone Mariani^{2,3} and Jesse Thaler^{1,3}

¹*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

²*Physics Department, Reed College, Portland, OR 97202, USA*

³*University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA*

We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on 33.9 pb⁻¹ of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,677 events containing a high-quality central jet with transverse momentum larger than 60 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 declustering. We find good agreement between results obtained from the CMS Open Data and those obtained from particle shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS Open Data does not include simulated data to help estimate systematic uncertainties, our results are observable to validate these substructure studies.

CONTENTS

I. Introduction	11
II. The CMS Open Data	12
A. The CMS Software Framework	12
B. The Jet Primary Dataset	12
C. The MIT Open Data Format	12
D. Analysis Tools	12
E. Parton Shower Generators	12
III. Hardjet Jet Properties	13
A. Jet Kinematics	13
B. Back Substructure Observables	13
C. Jet Angularities	13
IV. Two-Prong Jet Substructure	13
A. Soft Drop Declustering	13
B. ML Analytic Predictions	13
C. Open Data Results	13
V. Advice to the Community	21
A. Key applications	21
B. Recommendations	21
VI. Conclusion	23
Acknowledgments	23
A. Additional Open Data Information	24
B. Additional Soft-Dropped Distributions	24
References	25

* cschott@hig.harvard.edu
matthias.schott@mit-cp-ctpp.org
larkoski@red.cmu.edu
mariani@buffalo.edu
jthaler@mit.edu
wei.xue@reed.edu

The core of our analysis is based on soft drop declustering.

DELPHES fast simulation.

MIT-CTP 4890



Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski^{1,2} Simone Mariani^{2,3} Jesse Thaler^{2,3} Aashish Tripathi^{2,3} and Wei Xue^{1,3}

¹*Physics Department, Reed College, Portland, OR 97202, USA*

²*University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA*

³*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

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. However, however, a new jet substructure observable was introduced which is sensitive to the splitting function for collinearly high jet energy. 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 2-prong substructure of jets and test the 1 \rightarrow 2 splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

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 \rightarrow n scattering process factorizes into a 2 \rightarrow n-1 scattering cross section multiplied by a universal 1 \rightarrow 2 splitting probability, with corrections suppressed by the degree of collinearity. Collinear universality is a fundamental property of QCD and appears in many applications, most famously in the factorization theorem [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 \rightarrow 2 splittings [16–18]. Fast-order substructure observables like the 1 \rightarrow 2 splitting function [19–21], and the k_T jet clustering metric is based on 2 \rightarrow 1 recombination [22, 23]. Collinear universality can be extended to multi-parton 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 \rightarrow n decays of heavy particles from 1 \rightarrow n splittings in QCD in order to enhance the search for new physics at the Large Hadron Collider (LHC) [53–56].

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

$$P_{1 \rightarrow 2}(\theta) = \frac{d\sigma}{d\theta} P_{1 \rightarrow 2}(\theta), \quad (1)$$

where the $P_{1 \rightarrow 2}$ are the Altarelli-Parisi QCD splitting functions [5] 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 $\theta \rightarrow 0$ limit, so requires to cancel corresponding virtual singularities from loop diagrams. In this sense, there is no way to

directly measure the splitting function $P_{1 \rightarrow 2}(\theta)$ in data, though there is evidence of overwhelming indirect evidence that $P_{1 \rightarrow 2}(\theta)$ is a universal function in the many success of QCD in describing high-energy scattering (see e.g. [57–62]).

In this letter, we present a semi-direct method to test the 1 \rightarrow 2 splitting function in QCD by studying the 2-prong substructure of jets. Our method is based on soft drop declustering [63] (see also [62, 60, 70]), which recursively removes soft radiation from a jet until hard 2-prong substructure is found. When applied to collinear 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) is closely related to the momentum fraction z appearing in eq. (1), and the cross section for z splittings to the QCD splitting function in the high-energy limit. While variants of z have appeared in many jet substructure studies (notably the $\sqrt{p_T}$ parameter in refs. [62, 72]), to the best of our knowledge, no published z distribution has ever been presented using actual collider data, though there are preliminary z results from CMS [73, STAR [74], and ALICE [75]. Here, we present the first analysis of z using LHC data, taking advantage for the first time of public data released by the CMS experiment [76].

The CMS Open Data is derived from 7 TeV center-of-mass proton-proton collisions recorded in 2010 and released to the public on the CMS Open Data Portal in November 2014 [77]. The data is provided in AOD (Analysis Object Data) format, which is a CMS-specific data scheme based on the ROOT framework [78]. Crucially for the purposes of studying jet substructure, the AOD format contains all of the particle flow candidates (PFCs) [79, 80] used for jet finding within CMS [81], and we can apply jet substructure techniques directly on the PFCs themselves. The AOD files have no associated condition database which include jet energy correction (JEC) factors and recommended jet quality cuts, though no specific calibration tools for jet substructure studies. The main

MIT-CTP 4891

arXiv:1904.11195v2 [hep-ex] 29 Apr 2019

arXiv:1902.04222v1 [hep-ph] 12 Feb 2019

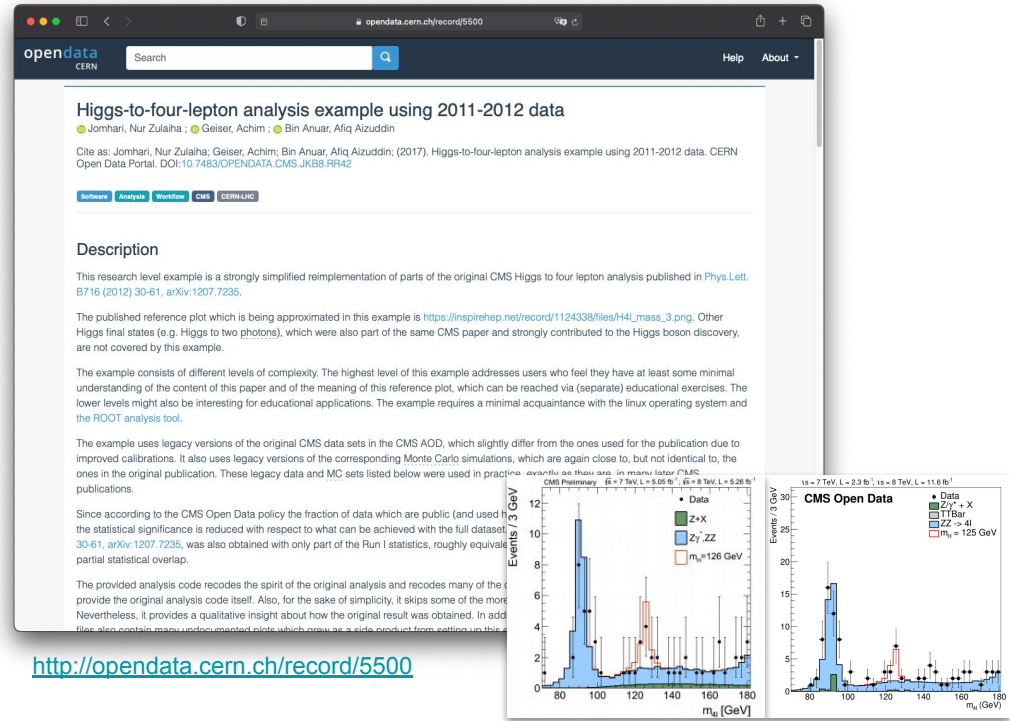
arXiv:1704.05842v3 [hep-ph] 28 Sep 2017

arXiv:1704.05066v3 [hep-ph] 25 Sep 2017

* cschott@hig.harvard.edu
yotam.soreq@cern.ch
jthaler@mit.edu
wei.xue@reed.edu

Analysis examples

code
+
data
+
environment



The four questions

where is **data**?

hard drive, distributed storage

where is the **code**?

GitLab, local copy, email

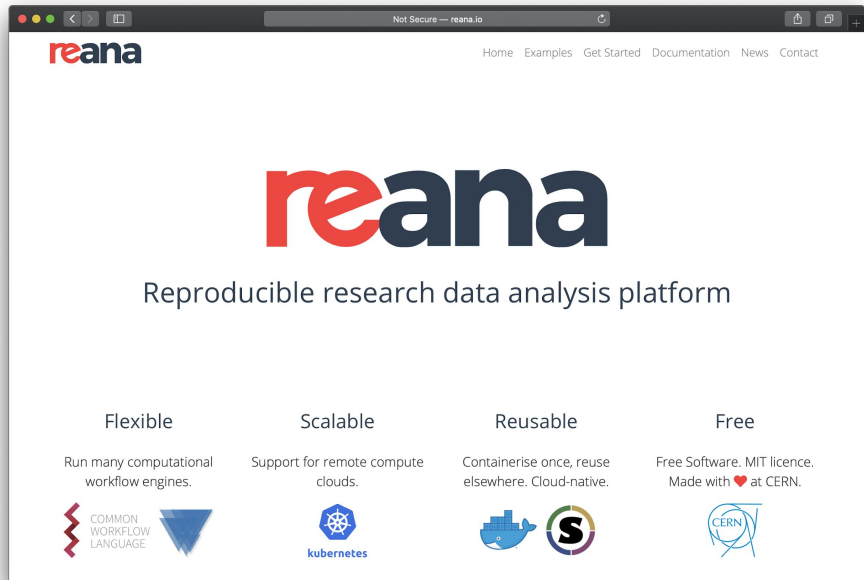
what **environment** do you use?

my own laptop, remote server

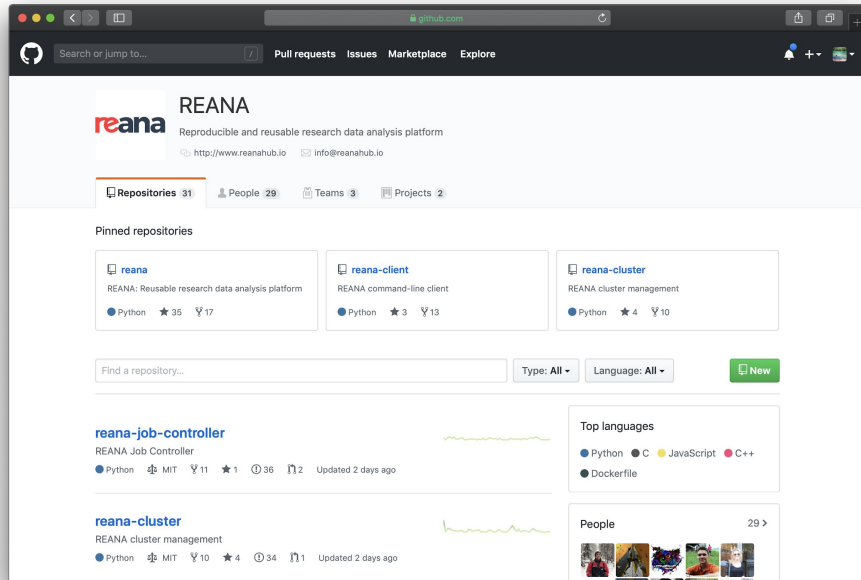
what **workflow** do you use?

Interactive commands, bash script, README file

Reproducible analyses



<http://reana.io/>



<https://github.com/reanahub>

Example

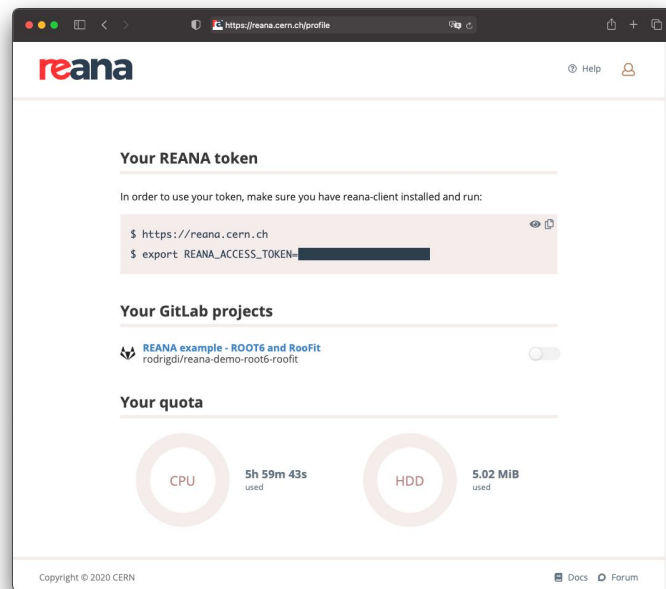
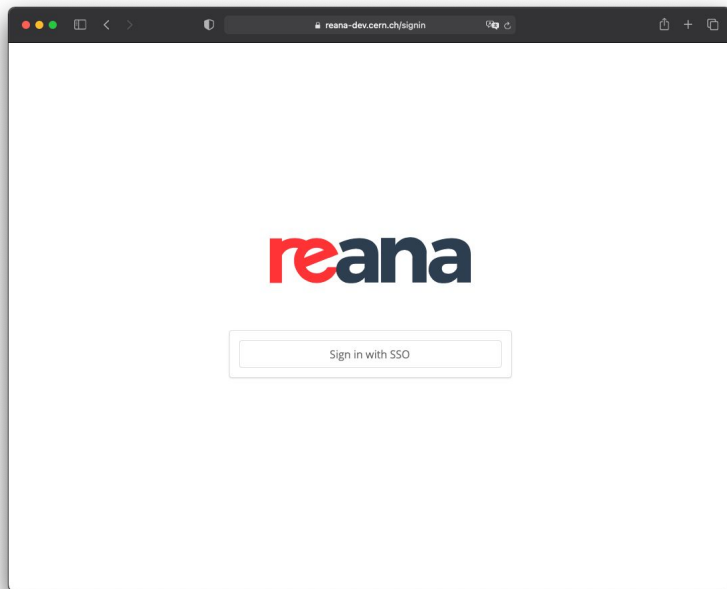
```
version: 0.6.0
inputs:
  files:
    - code/gendata.C
    - code/fitdata.C
  parameters:
    events: 20000
    data: results/data.root
    plot: results/plot.png
workflow:
  type: serial
  specification:
    steps:
      - name: gendata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
          - mkdir -p results && root -b -q
            'code/gendata.C(${events},"${data}")'
      - name: fitdata
        environment: 'reanahub/reana-env-root6:6.18.04'
        commands:
          - root -b -q 'code/fitdata.C("${data}","${plot}")'
outputs:
  files:
    - results/plot.png
```

Inputs including: **code, data, parameters**

Workflow including: **steps/commands, container images**

Outputs including: **plots, data etc...**

Example



Example

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

```
Terminal
$ reana-client logs -w roofit
==> 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.

==> 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 displays four overlapping screenshots of the reana web interface, illustrating the workflow execution and results.

Top Left Screenshot: Shows the "Your workflows" section. It lists three workflows:

- roofit #7**: Finished 8 hours ago, finished in 28 seconds step 2/2.
- roofit #6**: Finished 8 hours ago, finished in 32 seconds step 2/2.
- roofit #5**: Finished 14 days ago, finished in 28 seconds step 2/2.

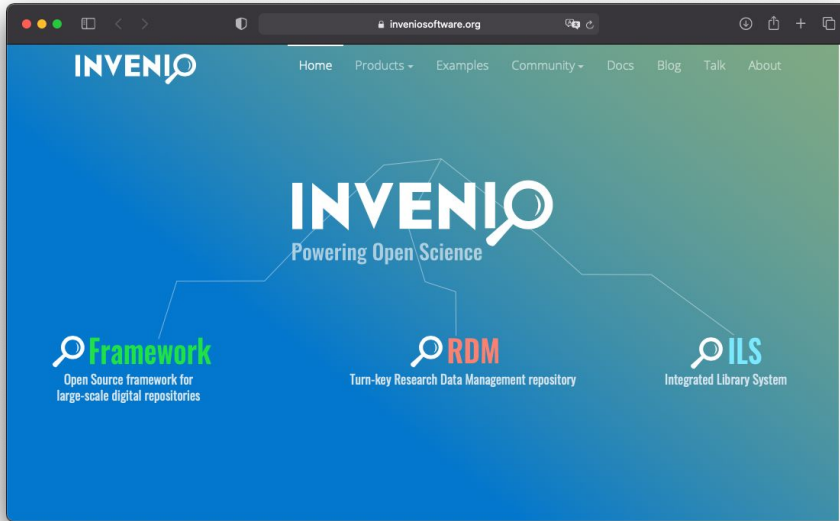
Top Right Screenshot: Shows the details of workflow **roofit #7**. It indicates the workflow is **finished** in 28 seconds, step 2/2.

Bottom Left Screenshot: Shows the "Workspace" tab for workflow **roofit #7**. It displays a table of files:

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

Bottom Right Screenshot: Shows a plot titled "Fit example". The y-axis is labeled "Events / (0.1)" and ranges from 0 to 1000. The x-axis ranges from 0 to 10. The plot shows a sharp peak at x=5. A "Download" button is visible at the bottom right of the plot area.

Technology: repository



<https://inveniosoftware.org>

The technology behind:

<https://www.hepdata.net/>

<https://inspirehep.net/>

<https://zenodo.org>

 Python

 Flask

 React

 JSONSchema

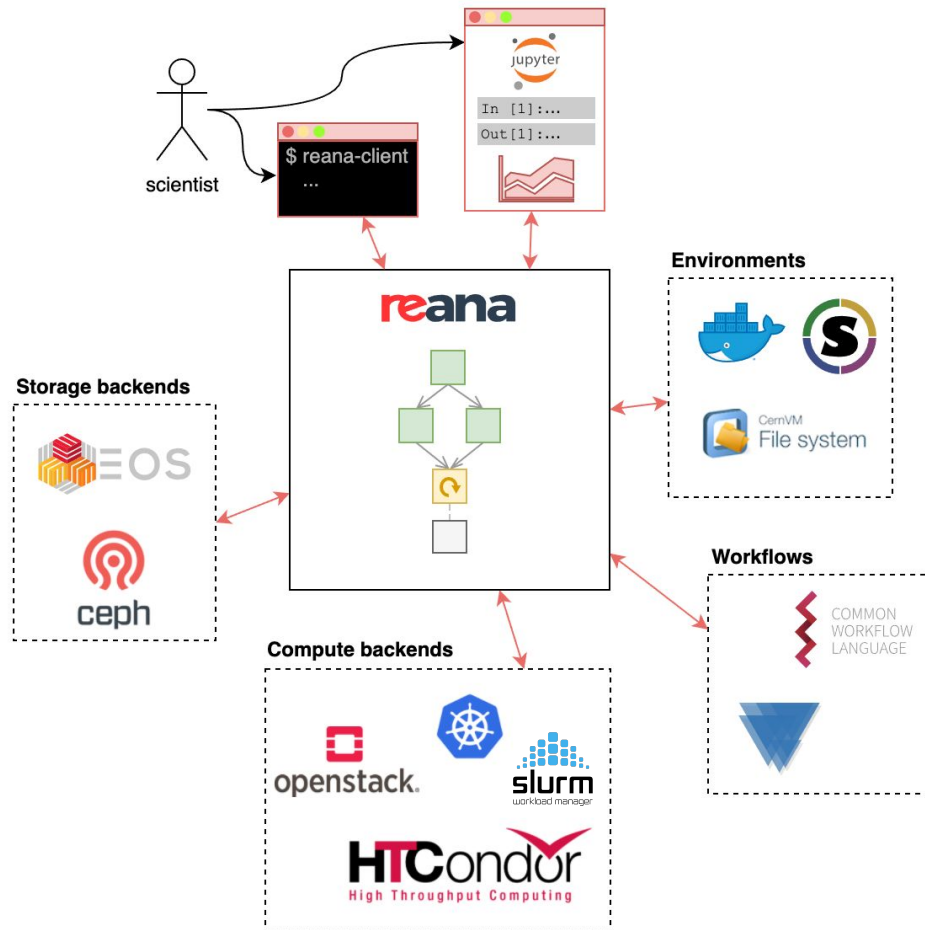
 Redis

 Elasticsearch

 PostgreSQL

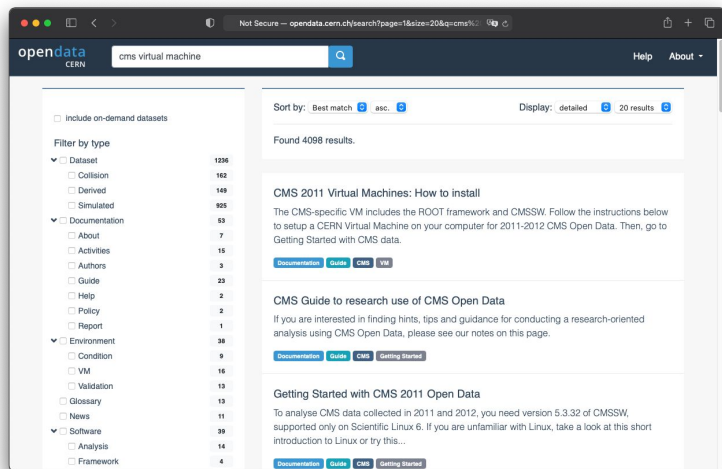
Technology: REANA

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




Try them out!

Use opendata.cern.ch



Install REANA on premises/locally

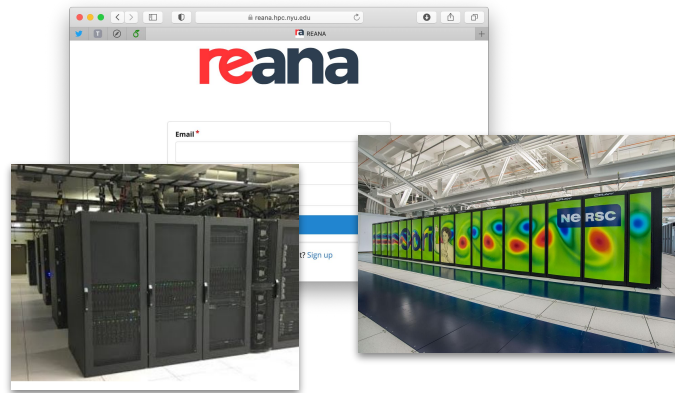
```
$ helm repo add reanahub \
https://reanahub.github.io/reana
```



```
$ 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>