


Beyond repositories

Enabling actionable FAIR open data reuse services in particle physics

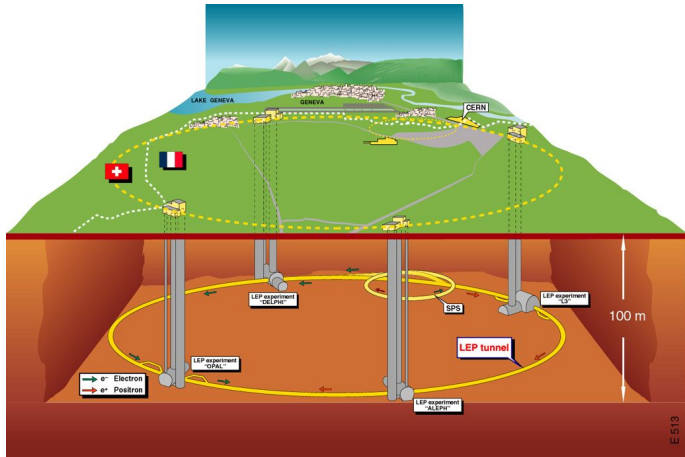
Diego Rodríguez

 @diego_delemos

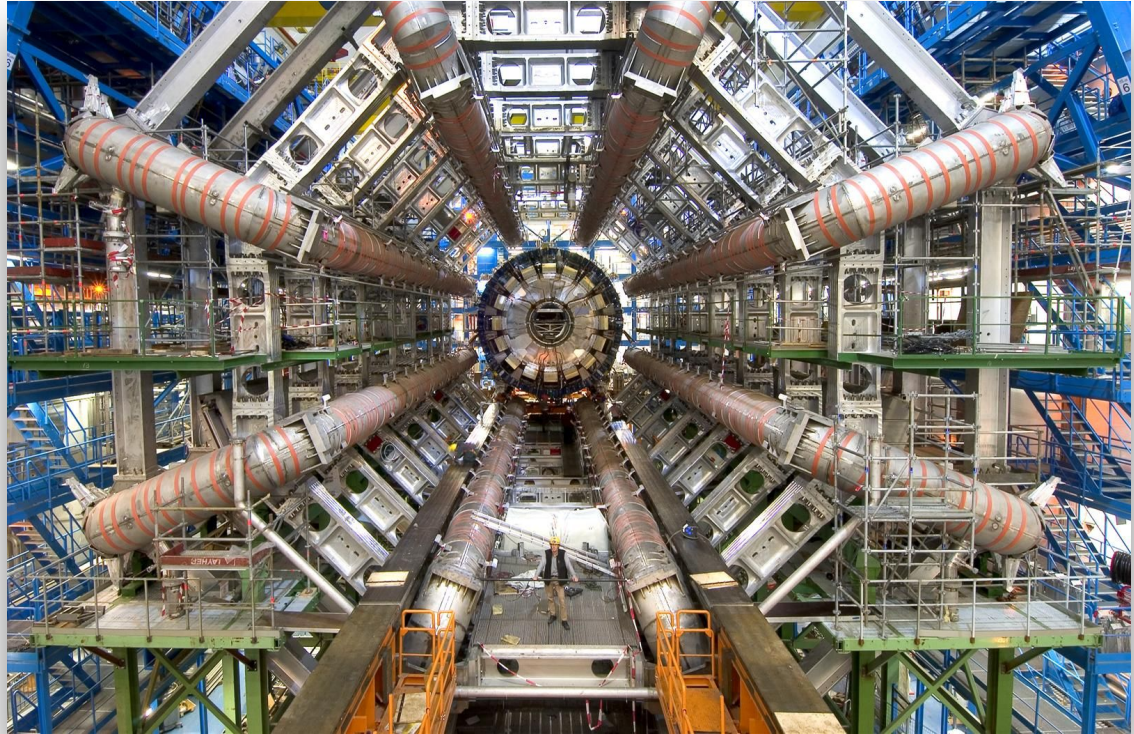
Tibor Šimko, Sünje Dallmeier-Tiessen, Sebastian Feger, Pamfilos Fokianos, Dinos Kousidis, Artemis Lavasa, Rokas Mačiulaitis, Jan Okraska, Diego Rodríguez, Anna Trzcińska, Ioannis Tsanaktsidis, Stephanie van de Sandt

CERN, Switzerland

CERN Hadron Collider

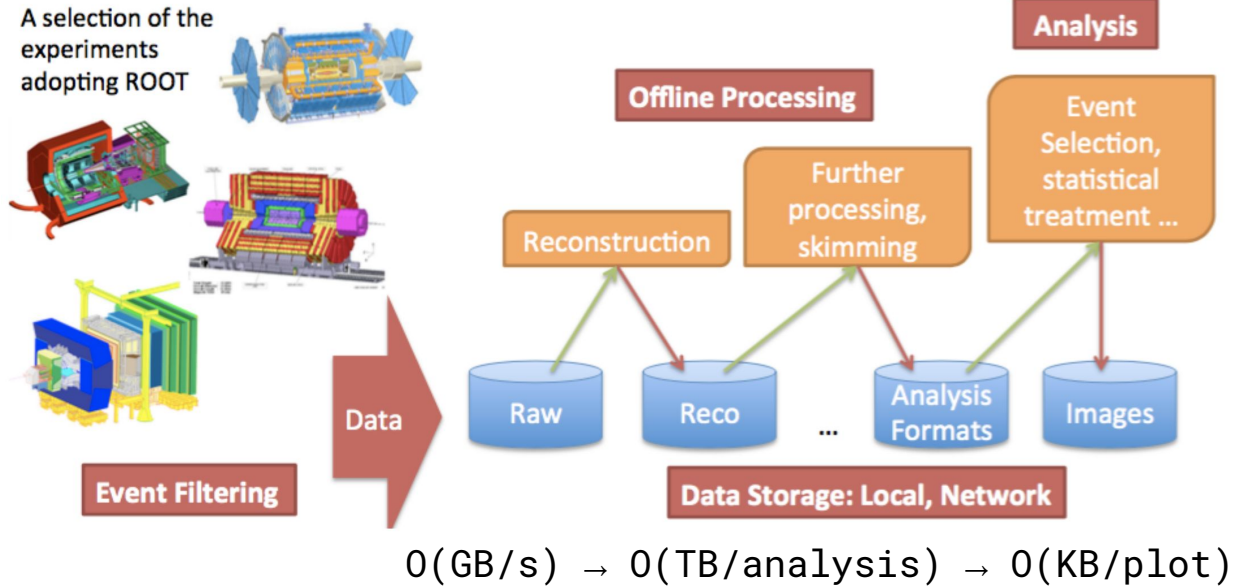


<http://cds.cern.ch/record/842153>



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

HEP data analyses



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

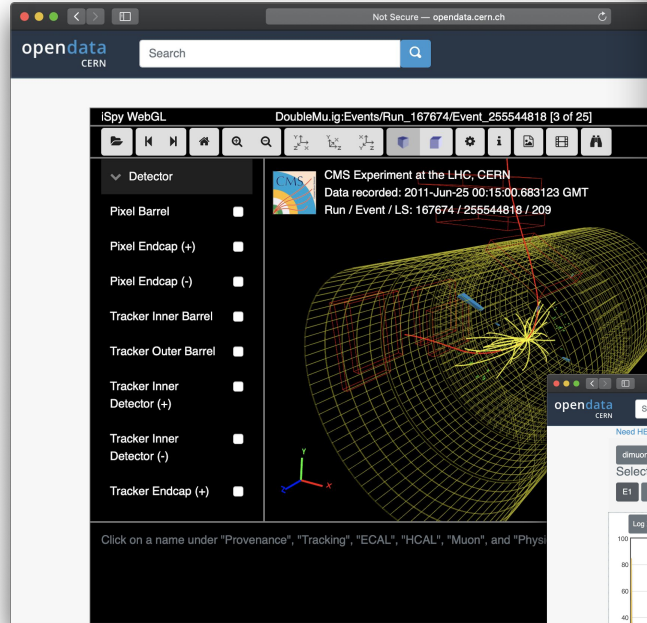
CERN Open Data portal

The screenshot shows the homepage of the CERN Open Data portal. At the top left, the logo reads "opendata CERN". The main heading says "Explore more than 1 petabyte of open data from particle physics!". Below this is a search bar with the placeholder text "Start typing..." and a blue "Search" button. Under the search bar, there are "search examples: collision datasets, keywords.education, energy.7TeV". To the left, under the heading "Explore", there are links for "datasets", "software", "environments", and "documentation". To the right, under the heading "Focus on", there are links for "ATLAS", "ALICE", "CMS", "LHCb", and "OPERA". The background features a stylized particle detector diagram.

The screenshot shows the search results page for the query "higgs". The search bar at the top contains "higgs". The page is divided into several sections. On the left, there are filter options: "Filter by type" (Dataset: 270, Derived: 22, Simulated: 248, Documentation: 8, About: 1, Activities: 3, Guide: 4, News: 3, Software: 2, Analysis: 2, Supplementaries: 37, Configuration SIM: 37), "Filter by experiment" (ATLAS: 25, CMS: 295), "Filter by year" (2011: 279, 2011-2012: 2, 2012: 24), and "Filter by file type" (aodsim: 248, cc: 1, csv: 2, gz: 2, lg: 2, json: 1). On the right, the results are sorted by "Best match" in ascending order, displaying "20 results". The first result is "Learning to discover: the Higgs boson machine learning challenge - Documentation", with a sub-header "Documentation" and a description: "A documentation for the Challenge is provided, with just the sufficient level of explanation for people without any background in physics or machine learning...". The second result is "Software examples for the ATLAS Higgs Learning Challenge 2014", with sub-headings "Software", "Analysis", and "ATLAS", and a description: "This software package contains several helper scripts that demonstrate event filtering and AMS score evaluation techniques on the ATLAS Higgs Challenge dataset as well as one script that allows to rep...". The third result is "Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014", with a description: "The dataset has been built from official ATLAS full-detector simulation, with 'Higgs to tautau' events mixed with different backgrounds. The simulator has two parts. In the first, random proton-proto...".

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

Education



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

Independent research

1904.11195v2.pdf (page 1 of 20)

MIT-CTP 4890

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*CERN/EP, University of Cambridge, UK*
²*Massachusetts Institute of Technology, Cambridge, USA*
³*Johannes Gutenberg University, Mainz, Germany*
E-mail: lester@hep.phy.cam.ac.uk, matthias.schott@cern.ch

ABSTRACT: The Standard Model violates parity, but invisible to Large Hadron Collider (LHC) experiments (a state polarisation or spin-sensitivity in the detectors). Non could potentially violate parity in ways which are detected. If such sources of new physics occur only at LHC energy searches. 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 initial is primarily sensitive to non-standard parity violating effects. Within our measurements, no significant deviation from no-observable experimental limitations have been found. We for one-standard parity violation could be performed, not to very different sets of models to those which our measurement initial studies provide a valuable starting point for full LHC datasets at 13 TeV with a careful and less precise uncertainties.

Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum

Carli Cesrotti,^{1,*} Yotam Soreq,^{2,3,†} Matthew J. Strasser,^{4,1,†} and Wei Xia⁵

¹*Department of Physics, Harvard University, Cambridge, MA 02138*
²*Theoretical Physics Department, CERN, Geneva, Switzerland*
³*Department of Physics, Technion, Haifa 32000, Israel*
⁴*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Abstract

We study dimuon events in 2.11fb^{-1} of 7 TeV pp collisions, using CMS Open Data, and search narrow dimuon resonances with moderate mass (14–60 GeV) and substantial transverse momentum. Applying dimuon p_T cuts of 25 GeV and 60 GeV, we explore two overlapping samples: one with p_T momenta, and one with prompt muons without an isolation requirement. Using the latter sample we information about detector effects and QCD backgrounds, which we obtain directly from the CMS 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 p_T by factors of as much as nine. Our “ p_T -enhanced” dimuon search strategy provides improved sensitivity models in which a new particle is produced mainly in the decay of something heavier, as could occur, for example, in the decay of the Higgs boson or of a TeV-scale top partner. An implementation of this method in the current 13 TeV data should improve the sensitivity to such signals further by roughly an order of magnitude.

* cesrotti@lig.harvard.edu
† yotam.soreq@cern.ch
mstrasser@physics.harvard.edu
jshaver@mit.edu
wxi@cern.ch

1704.05842v3.pdf (page 1 of 35)

MIT-CTP 4890

Jet Substructure Studies with CMS Open Data

Aashish Tripathi,¹ Wei Xue,¹ Andrew Larsook,² Simone Marzani,³ and Jose Thaler¹

¹*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
²*Physics Department, Reed College, Portland, OR 97082, USA*
³*University at Buffalo, The State University of New York, Buffalo, NY 14260-1506, USA*

We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on 33.9pb^{-1} of 7 TeV pp proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 706,677 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 analysis observables to validate those substructure studies.

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

CONTENTS	
I. Introduction	11
II. The CMS Open Data	21
A. The CMS Software Framework	21
B. The Jet Primary Dataset	22
C. The MIT Open Data Format	22
D. Analysis Tools	27
E. Parton Shower Generators	27
III. Hardest Jet Properties	30
A. Jet Kinematics	30
B. Back-Substructure Observables	31
C. Jet Angularities	31
IV. Two-Prong Jet Substructure	33
A. Soft Drop Declustering	33
B. ML Analytic Predictions	34
C. Open Data Results	35
V. Advice to the Community	221
A. Key Applications	221
B. Recommendations	221
VI. Conclusion	223
Acknowledgments	223
A. Additional Open Data Information	224
B. Additional Soft-Dropped Distributions	224
References	225

* aashish@mit.edu
wxi@mit.edu
larsook@reed.edu
simone.marzani@cern.ch
jthaler@mit.edu

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

1704.05066v3.pdf (page 1 of 8)

MIT-CTP 4891

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larsook,¹ Simone Marzani,² Jose Thaler,^{3,†} Aashish Tripathi,³ and Wei Xue¹

¹*Physics Department, Reed College, Portland, OR 97082, USA*
²*University at Buffalo, The State University of New York, Buffalo, NY 14260-1506, 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 most 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 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 → 2 splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

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

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 factors into a $2 \rightarrow n-1$ scattering cross section multiplied by a universal $1 \rightarrow 2$ splitting probability, with corrections suppressed by the degree of collinearity. Collinear universality is a fundamental property of QCD and appears in many applications, most famously in deriving the DGLAP 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 $1 \rightarrow 2$ splittings [16–18]. Bond-energies substructure schemes used in the $1 \rightarrow 2$ splitting function [19–21], and the k_T jet clustering metric is based on $2 \rightarrow 1$ recombination [22–24]. 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 inapplicable 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

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

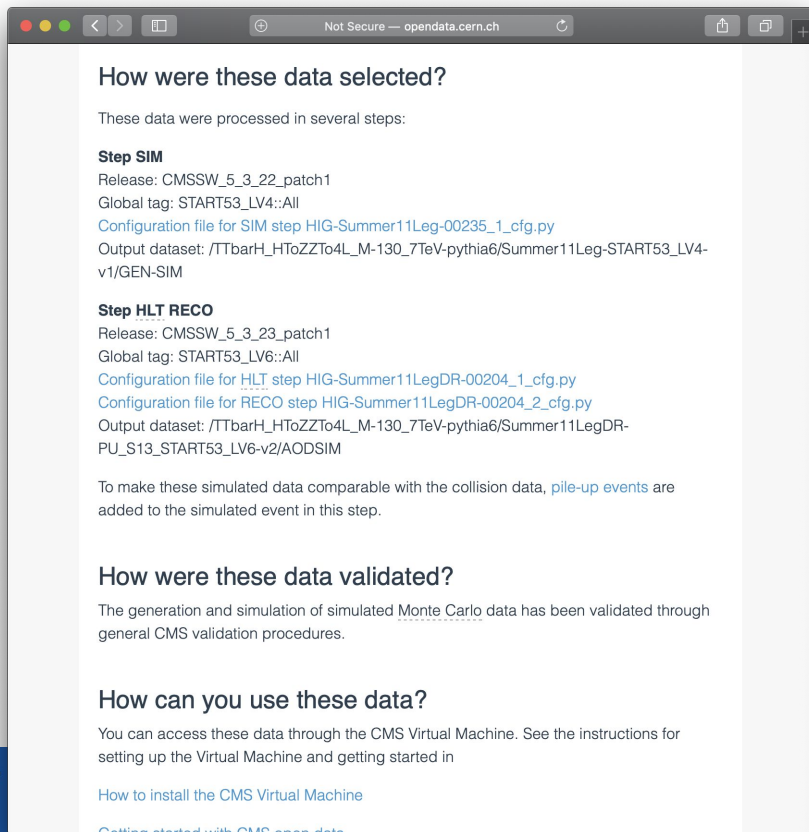
where the $P_{1 \rightarrow 2}$ are the Altarelli-Parisi QCD splitting functions [5] which depend on the momentum fraction z and the parton flavors i, j , and C . Crucially, this expression has a real emission singularity in the $\theta \rightarrow 0$ limit, so requires careful corresponding virtual singularities from loop diagrams. In this sense, there is no way to

directly measure the splitting function $P_{1 \rightarrow 2}(\theta)$ in nature, though there is of course overwhelming indirect evidence that $P_{1 \rightarrow 2}(\theta)$ is a universal function (as the 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 [8] (see also [32, 60, 70]), which recursively removes soft radiation from a jet until hard 2-prong substructure is found. When applied to ordinary quark and gluon-initiated jets with an intrinsic collinear quark, soft drop exposes the collinear core of the jet. As shown in ref. [71], the momentum sharing between the two prongs (denoted z_1) is closely related to the momentum fraction z appearing in eq. (1), and the cross section for z_1 asymptotes to the QCD splitting function in the high-energy limit. While variants of z_1 have appeared in many jet substructure studies (notably the \sqrt{z} parameter in refs. [52, 72]), to the best of our knowledge, no published z_1 distribution has ever been presented using actual collider data, though there are preliminary z_1 results from CMS [73, STAR [74], and ALICE [75]. Here, we present the first analysis of z_1 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 as 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 includes jet energy correction (JEC) factors and recommended jet quality cuts, though supporting calibration tools for jet substructure studies. The main



Tracking data provenance



How were these data selected?

These data were processed in several steps:

Step SIM
Release: CMSSW_5_3_22_patch1
Global tag: START53_LV4::All
[Configuration file for SIM step HIG-Summer11Leg-00235_1_cfg.py](#)
Output dataset: /TTbarH_HToZZTo4L_M-130_7TeV-pythia6/Summer11Leg-START53_LV4-v1/GEN-SIM

Step HLT RECO
Release: CMSSW_5_3_23_patch1
Global tag: START53_LV6::All
[Configuration file for HLT step HIG-Summer11LegDR-00204_1_cfg.py](#)
[Configuration file for RECO step HIG-Summer11LegDR-00204_2_cfg.py](#)
Output dataset: /TTbarH_HToZZTo4L_M-130_7TeV-pythia6/Summer11LegDR-PU_S13_START53_LV6-v2/AODSIM

To make these simulated data comparable with the collision data, [pile-up events](#) are added to the simulated event in this step.

How were these data validated?

The generation and simulation of simulated [Monte Carlo](#) data has been validated through general CMS validation procedures.

How can you use these data?

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

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

[Getting started with CMS open data](#)



opendata
CERN

Search

About

Software to preprocess the CMS 2011 DoubleMu and DoubleElectron datasets for the two-lepton/four-lepton analysis example of CMS open data.

Rodriguez Marrero, Ana; Lassila-Perini, Kati;

Cite as: Rodriguez Marrero, Ana; Lassila-Perini, Kati; (2016). Software to preprocess the CMS 2011 DoubleMu and DoubleElectron datasets for the two-lepton/four-lepton analysis example of CMS open data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.ETJK.JKMB

Software Tool CMS Accelerator CERN-LHC

Description

Software to produce the intermediate data files, derived from the primary datasets, for a simple analysis on Z decays to two leptons and ZZ decays to four leptons, of CMS open data.

Use with

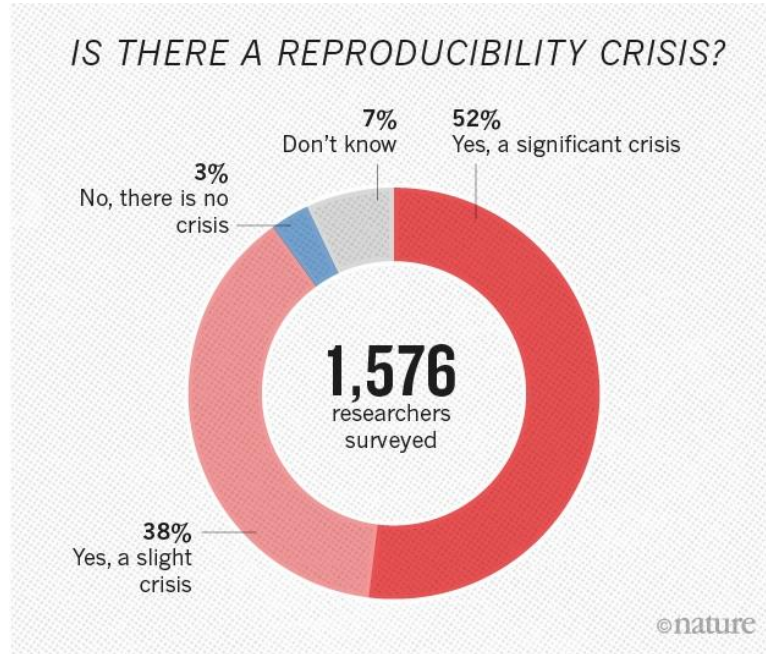
Use this with the following datasets:

[/DoubleElectron/Run2011A-12Oct2013-v1/AOD](#)

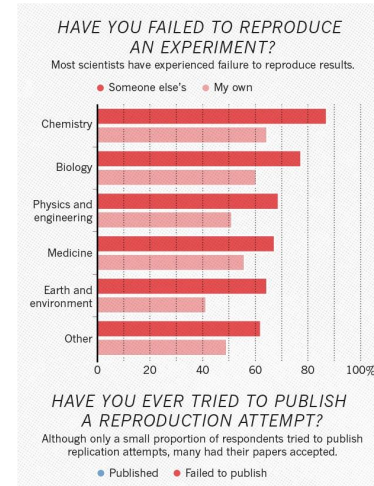
[/DoubleMu/Run2011A-12Oct2013-v1/AOD](#)

Notes

Reproducibility crisis



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



How FAIR are we?

Findable	✓	Accessible	✓
Interoperable	✓	Reusable	?

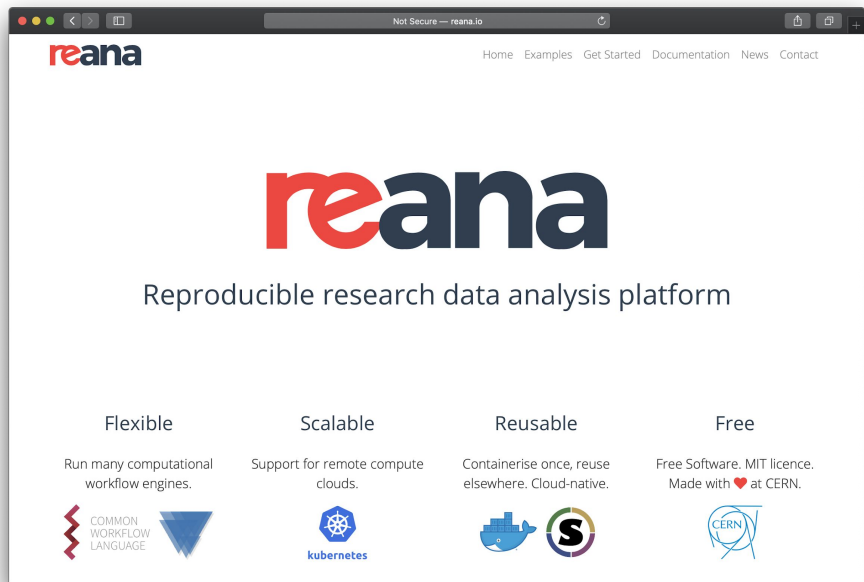
where is the **data**?

where is the **code**?

what is the **environment**?

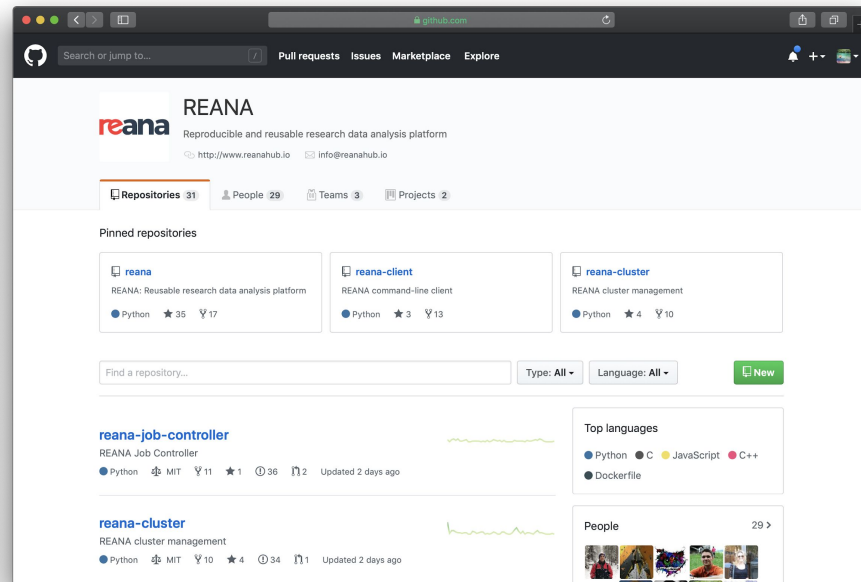
what is the **workflow**?

Reproducible analyses



The screenshot shows the homepage of the reana.io website. At the top left is the reana logo. The main heading is "reana" in a large, bold font, with "Reproducible research data analysis platform" underneath. Below this, there are four columns of features: "Flexible" (Run many computational workflow engines), "Scalable" (Support for remote compute clouds), "Reusable" (Containerise once, reuse elsewhere. Cloud-native), and "Free" (Free Software, MIT licence. Made with ❤️ at CERN.). At the bottom, there are logos for Common Workflow Language, Kubernetes, Docker, and CERN.

<http://reana.io/>



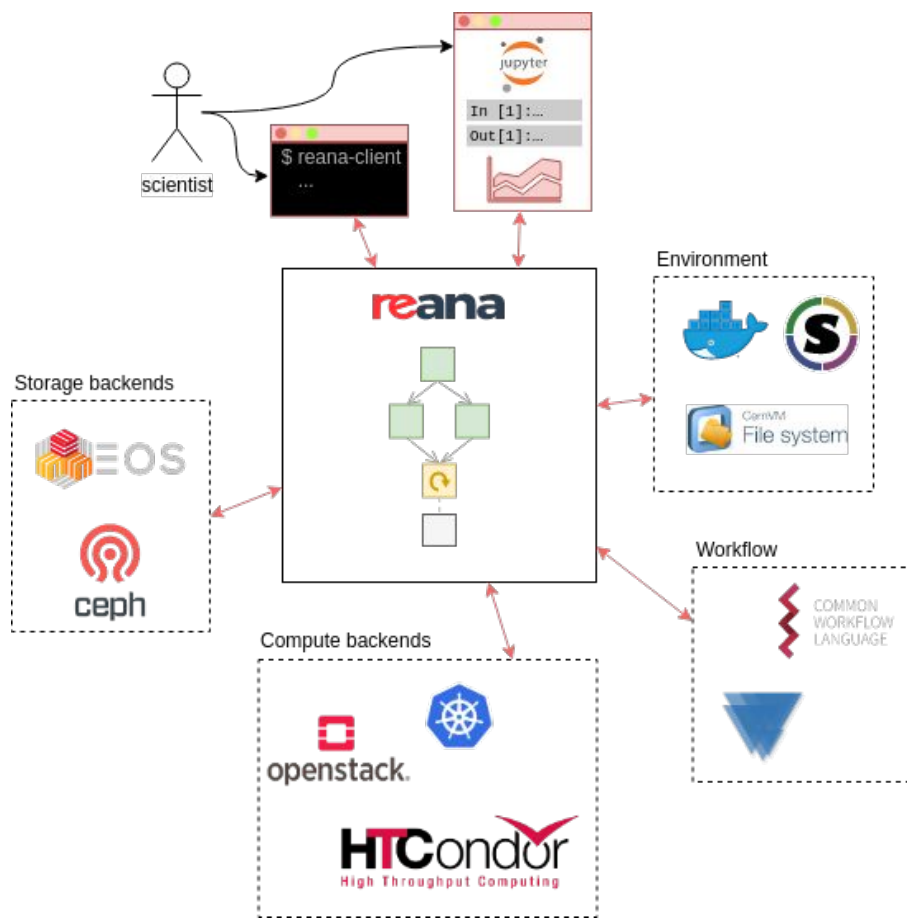
The screenshot shows the GitHub repository page for REANA. The repository name is "REANA" and the description is "Reproducible and reusable research data analysis platform". It lists 31 repositories, 29 people, 3 teams, and 2 projects. Pinned repositories include "reana", "reana-client", and "reana-cluster". A search bar and filters for "Type" and "Language" are visible. The "Top languages" section shows Python, C, JavaScript, and C++. The "People" section shows a list of contributors.

<https://github.com/reanahub>

REANA

Architecture

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





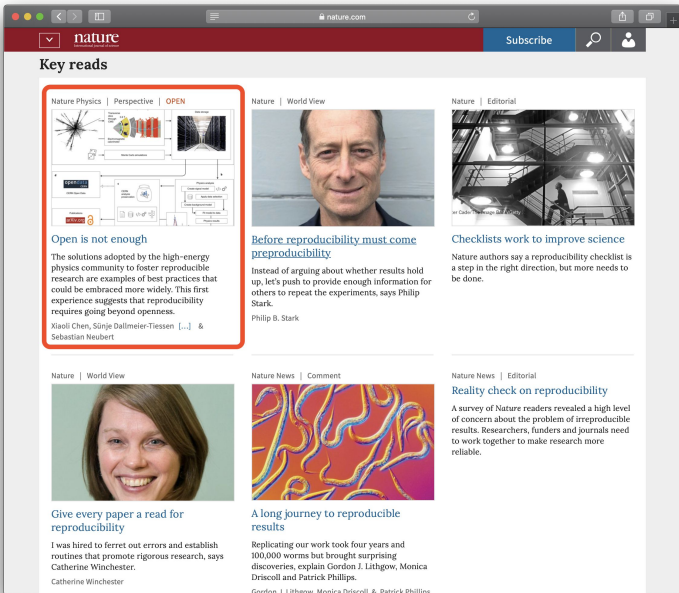
data + code + environment + workflow =



Conclusions

- successful **open data** and **reproducible research**
example in particle physics
- beyond repositories: **actionable data via runnable examples**
- **FAIR ≠ open**, FAIR should start early
- “top-down” approach: funding agencies, **best practices**
- “bottom-up” activities: building **useful tools** for scientists

Open is not enough



@cernopendata
@reanahub



@cernopendata
@reanahub

Diego Rodríguez
@diego_delemos

Danke schön!