CERN Open Data Portal

Tibor Šimko
CERN IT

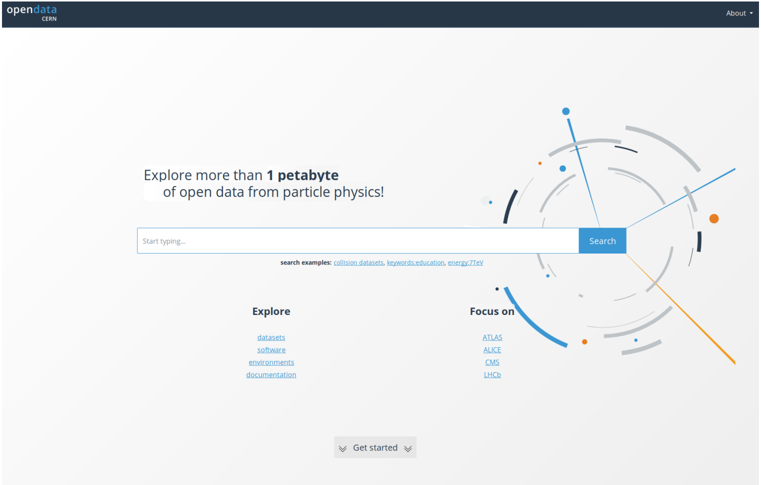
IPPOG · 4 October 2018

CERN Open Data portal

- `opendata.cern.ch` launched in November 2014
- LHC collaboration data policies
  - restricted → embargo period (∼5 years) → open
- over 1.5 Petabytes of open particle physics data
  - datasets, software, VMs, configuration, documentation, ...
- users
  - education: general public, high-school students, masterclasses
  - research: data scientists, physicists

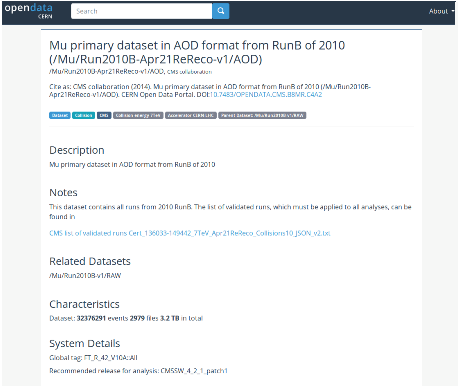
Developed by CERN-IT and CERN-SIS
in close collaboration with Experiments

CERN Open Data portal



http://opendata.cern.ch/

Information organisation



opendata CERN

Search

About

Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)

/Mu/Run2010B-Apr21ReReco-v1/AOD, CMS collaboration

Cite as: CMS collaboration (2014). Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.B8MR.C4A2

Dataset | Collision | CMS | Collision energy 7TeV | Accelerator CERN-LHC | Parent Dataset: /Mu/Run2010B-v1/RAW

Description

Mu primary dataset in AOD format from RunB of 2010

Notes

This dataset contains all runs from 2010 RunB. The list of validated runs, which must be applied to all analyses, can be found in

CMS list of validated runs Cert_136033-149442_7TeV_Apr21ReReco_Collisions10_JSON_v2.txt

Related Datasets

/Mu/Run2010B-v1/RAW

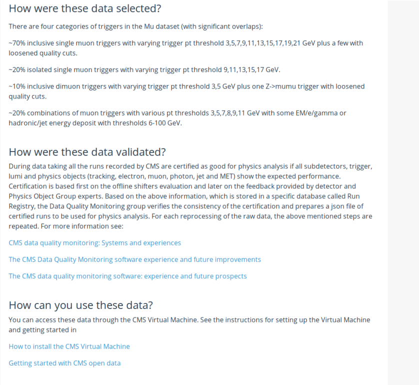
Characteristics

Dataset: **32376291** events **2979** files **3.2 TB** in total

System Details

Global tag: FT_R_42_V10A::All

Recommended release for analysis: CMSSW_4_2_1_patch1



How were these data selected?

There are four categories of triggers in the Mu dataset (with significant overlaps):

~70% inclusive single muon triggers with varying trigger pt threshold 3,5,7,9,11,13,15,17,19,21 GeV plus a few with loosened quality cuts.

~20% isolated single muon triggers with varying trigger pt threshold 9,11,13,15,17 GeV.

~10% inclusive dimuon triggers with varying trigger pt threshold 3,5 GeV plus one Z->mumu trigger with loosened quality cuts.

~20% combinations of muon triggers with various pt thresholds 3,5,7,8,9,11 GeV with some EM/e/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

CMS data quality monitoring: Systems and experiences

The CMS Data Quality Monitoring software experience and future improvements

The CMS data quality monitoring software: experience and future prospects

How can you use these data?

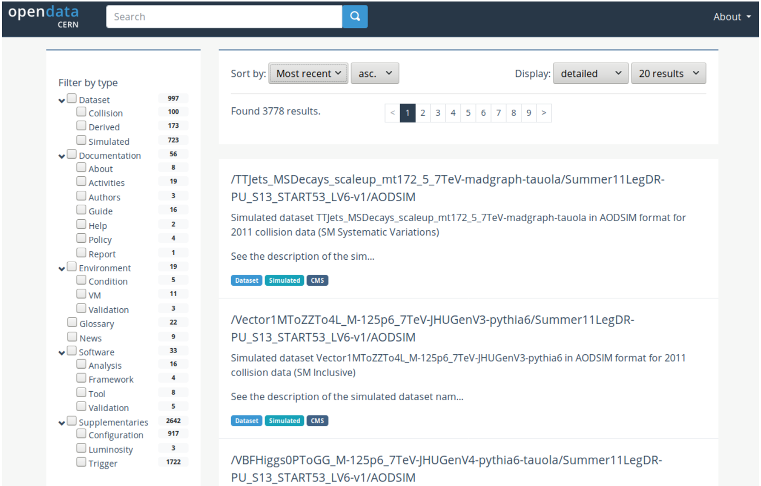
You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

How to install the CMS Virtual Machine

Getting started with CMS open data

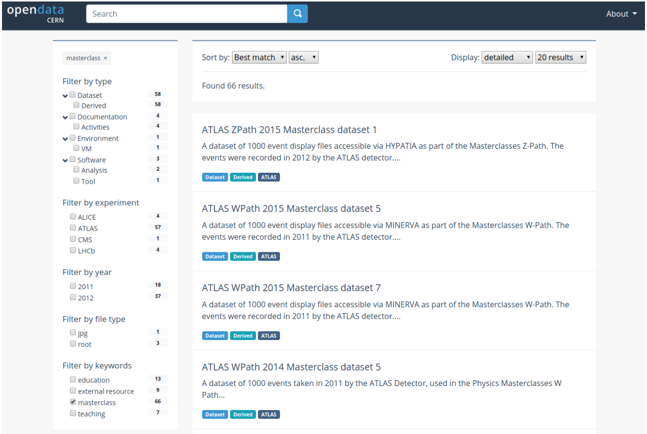
Context information about data selection, validation, use

Information discovery



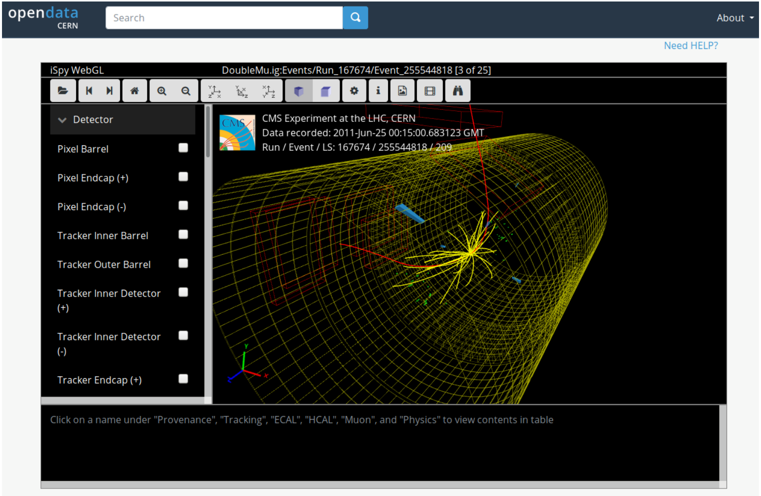
Explore a variety of data, software, VMs, supplementary material. . .

Masterclasses



Masterclass resources held on the CERN Open Data portal

Visualise detector events

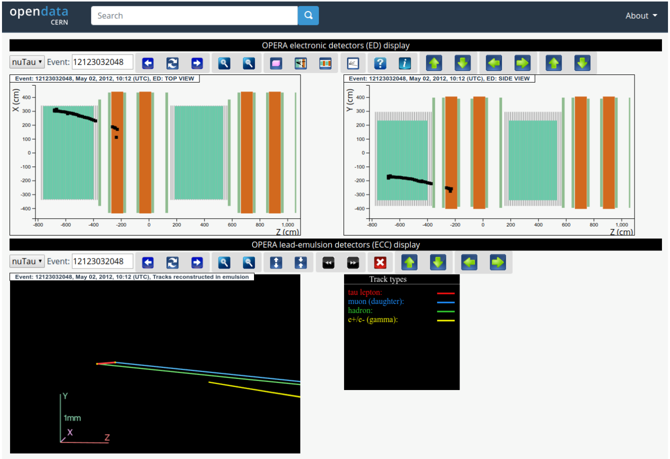


Interactive event display for high-level derived datasets

Visualise histograms

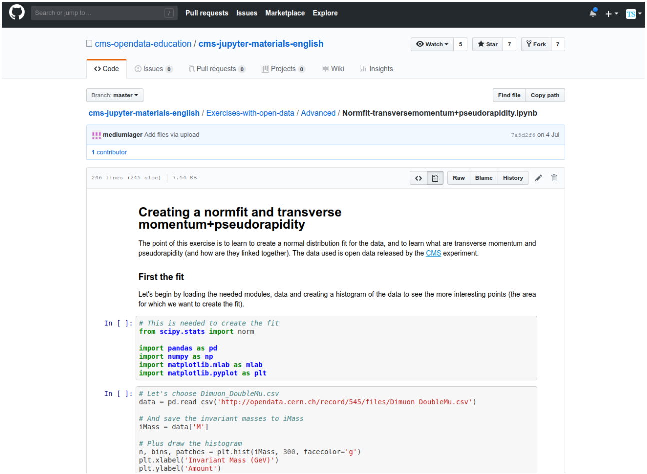
Interactive histogramming for high-level derived datasets

First non-LHC data releases



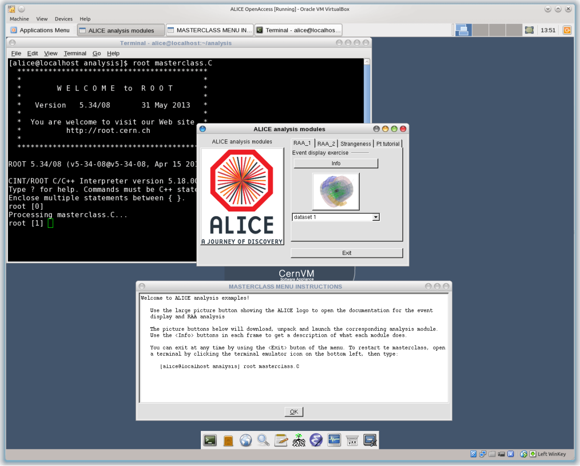
OPERA tau neutrino candidate and multiplicity studies data

Jupyter notebooks



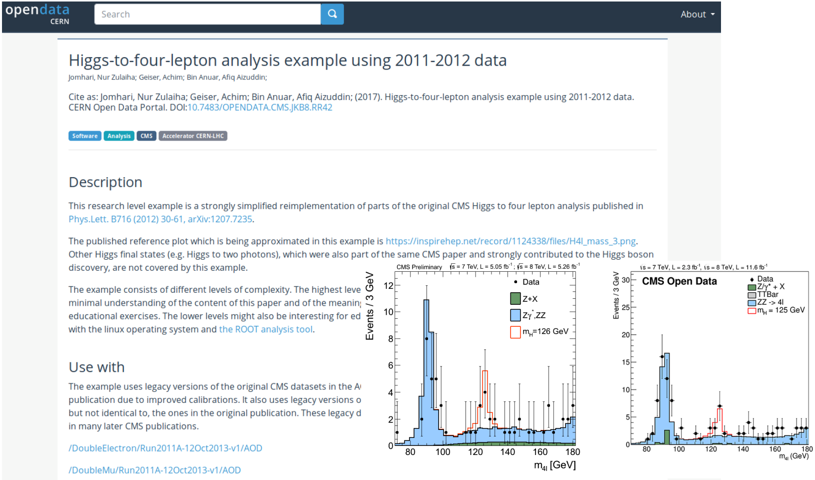
CMS education activities using notebooks and CMS open data

Virtual machines



Install CernVM virtual machines to explore primary datasets

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

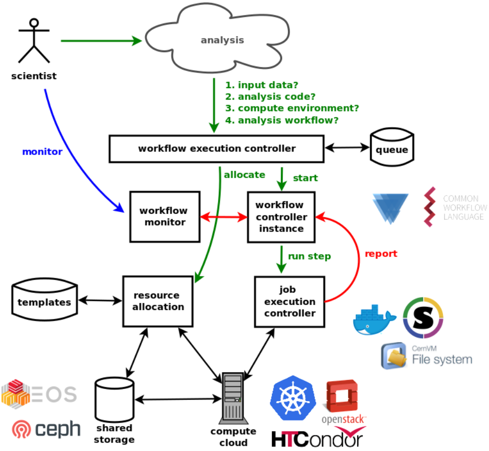
The example uses legacy versions of the original CMS datasets in the A publication due to improved calibrations. It also uses legacy versions o but not identical to, the ones in the original publication. These legacy d in many later CMS publications.

/DoubleElectron/Run2011A-12Oct2013-v1/AOD

/DoubleMu/Run2011A-12Oct2013-v1/AOD

Run realistic physics analysis examples

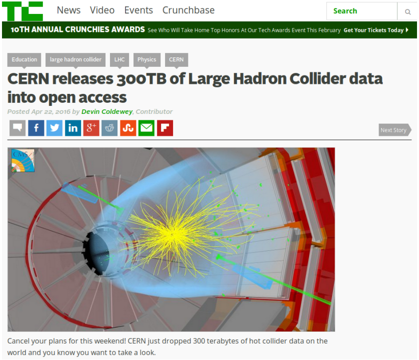
Run analyses on the cloud



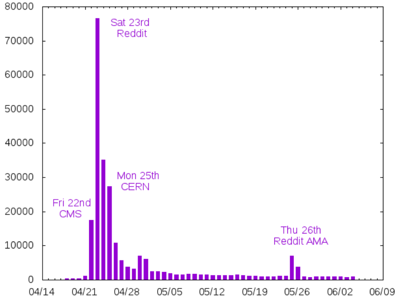
http://www.reana.io/

REANA reproducible research data analysis platform

Release-driven usage patterns



Open data releases are widely covered by general media



Six weeks in 2016: 200K users, 40K viewed records, 70K used event display, 3K used histogramming

Research-grade open data use

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 θ, the $1 \to 2$ splitting probability takes the form

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

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

0031-9007/17/119(13)/132003(7) 132003-1

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 β 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

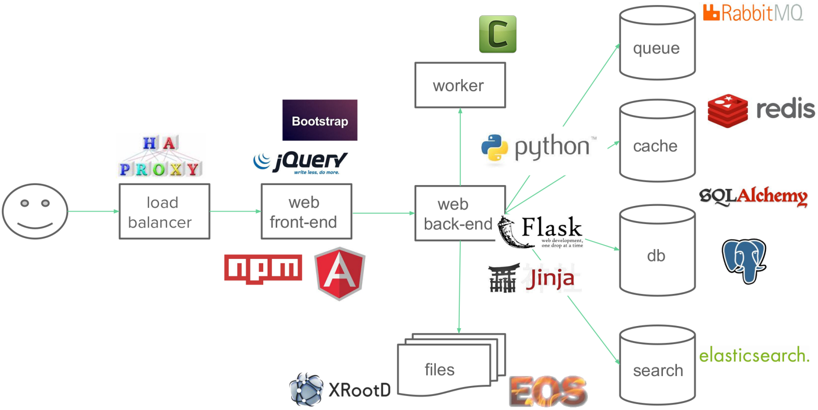
*aashisht@mit.edu
†weixue@mit.edu
‡larkoski@reed.edu
§smarzani@buffalo.edu
∥jthaler@mit.edu

[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].

2470-0010/2017/96(7)/074003(33) 074003-1

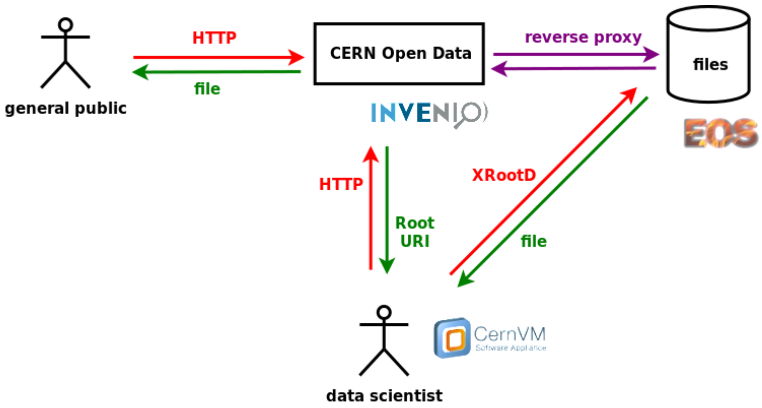
First independent analyses by theorists (Jesse Thaler *et al*, MIT)

Technology



Technology stack using INVENIO digital repository

Data exposure

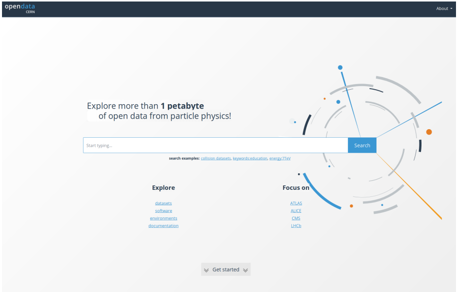


HTTP and XRootD access scenarios

Synergies for outreach

- digital repository use cases?
  - upload/download data, VM, documentation
  - version control, mint and cite persistent identifiers (DOI)
  - organise and search through data
- reaching wider scope?
  - traditionally hosting LHC experiments
  - first non-LHC data by OPERA
  - forthcoming Machine Learning reference datasets
- reaching wider audience?
  - O(200K) “visitors” in six weeks; O(40K) deeper, O(3K) heavy use
- run realistic examples?
  - in-browser visualisations and histogramming
  - analysis examples via Virtual Machines (ROOT, non-ROOT)
  - analysis examples via containers (ROOT, non-ROOT)
  - run locally or on cloud?

CERN Open Data



`http://opendata.cern.ch`

`http://github.com/cernopendata`

CERN IT D. Rodriguez, T. Šimko · **CERN SIS** S. Dallmeier-Tiessen, S. Feger, P. Fokianos, A. Lavasa, I. Tsanaktsidis, 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, K. Lassila-Perini, 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 OpenShift** A. Lossent, A. Peon