Latin American Symposium on Nuclear Physics and Applications (LASNPA) June 17-21, 2024





1

Fostering educational efforts in Latin America through the use of CERN Open Data

Dr. Leonid Serkin

Instituto de Ciencias Nucleares, UNAM, Mexico



Open data is a crucial part of science and brings many benefits to society and the public:

- Increases trust in scientific research.
- Allows the public to understand the results obtained.
- Promotes citizen science.
- Increases understanding of innovation and economic growth.
- Fosters educational efforts, allowing students and teachers to obtain and reproduce our results.

Open Data is available

for all of them!



<image>

EHT Data Products



tps://eventhorizontelescope.org/for-astronomers/data

2012: Discovery of the Higgs boson at the LHC at CERN.

2015: Detection of gravitational waves by LIGO.

2019: Direct detection of the event horizon of a black hole by EHT.

Open data is not enough



PERSPECTIVE https://doi.org/10.1038/s41567-018-0342-2

To set the stage for the rest of this piece, we first construct a more

choose to build on the descriptions introduced by Carole Goble'

tiple labs have the equipment necessary to duplicate an experiment

physics context, however, the immense cost and complexity of the

energy, precision and level of accuracy. The experiments at the Large

that it can be later reused with other measurements for comparison.

the computational analysis performed over a given dataset rather

thought of as an experimental collaboration or an analysis group.

In the case of computational processes, physics analyses ther

algorithms involved". In addition, the analysts typically study mor

Our considerations here really begin after gathering the data.

rimental set-up essentially make the independent and com

These concepts assume a research environment in which mul-

and Lorena A. Barbas shown in Table 1

confirmation or inspiration.

OPEN

Corrected: Publisher Correction

Open is not enough

Xiaoli Chen^{1,2}, Sünje Dallmeier-Tiessen^{1*}, Robin Dasler^{1,11}, Sebastian Feger^{1,3}, Pamfilos Fokianos¹, Jose Benito Gonzalez¹, Harri Hirvonsalo^{1,412}, Dinos Kousidis¹, Artemis Lavasa¹, Salvatore Mele¹, Diego Rodriguez Rodriguez', Tibor Šimko1*, Tim Smith', Ana Trisovic15*, Anna Trzcinska1, Ioannis Tsanaktsidis¹, Markus Zimmermann¹, Kyle Cranmer⁶, Lukas Heinrich⁶, Gordon Watts⁷, Michael Hildreth⁶, Lara Lloret Iglesias⁹, Kati Lassila-Perini⁴ and Sebastian Neubert¹⁰

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness

pen science and reproducible research have become per-flows for reproducible and reusable research more widely in other vasive goals across research communities, political circles scientific disciplines. nd funding bodies1-1. The understanding is that open and

reproducible research practices enable scientific reuse, accelerating Approaching reproducibility and reuse in HEP future projects and discoveries in any discipline. In the struggle to take concrete steps in pursuit of these aims there has been much nuanced spectrum in which to place the various challenges facing discussion and awareness-raising, often accompanied by a push to HEP, allowing us to better frame our ambitions and solutions. We make research products and scientific results open quickly. Although these are laudable and necessary first steps, they

are not sufficient to bring about the transformation that would allow us to reap the benefits of open and reproducible research. It is time to move beyond the rhetoric and the trust in quick fixes which essentially makes the experiments portable. In the particle and start designing and implementing tools to power a more profound change.

Our own experience from opening up vast volumes of data is plete replication of HEP experiments unfeasible and unhelpful. that openness cannot simply be tacked on as an afterthought at the HEP experiments are set up with unique capabilities, often being end of the scientific endeavour. In addition, openness alone does the only facility or instrument of their kind in the world; they are not guarantee reproducibility or reusability, so it should not be pursued as a goal in itself. Focusing on data is also not enough: it needs to be accompanied by software, workflows and explanations, all of Hadron Collider (LHC) are prominent examples. It is this uniquewhich need to be captured throughout the usual iterative and closed ness that makes the experimental data valuable for preservation so research lifecycle, ready for a timely open release with the results. Thus, we argue that having the reuse of research results as a goal requires the adoption of new research practices during the data

analysis process. Such practices need to be tailored to the needs This means that we are more concerned with repeating or verifying of each given discipline with its particular research environment, culture and idiosyncrasies. Services and tools should be developed than with data collection. Therefore, in Table 2 we present a variawith the idea of meshing seamlessly with existing research proce- tion of these definitions that takes into account a research environ dures, encouraging the pursuit of reusability as a natural part of ment in which 'experimental set-up' refers to the implementation researchers' daily work (Fig. 1). In this way, the generated research of a computational analysis of a defined dataset, and a 'lab' can be products are more likely to be useful when shared openly. In tackling the challenge of enabling reusable research, we

keep these ideas as our guiding light when putting changes into selves are intrinsically complex due to the large data volume and practice in our community-high-energy physics (HEP). Here, we illustrate our approach, particularly through our work at CERN, than one physics process and consider data collected under dif and present our community's requirements and rationale. We ferent running conditions. Although comprehensive documenta hope that the explanation of our challenges and solutions will tion on the analysis methods is maintained, the complexity of the stimulate discussions around the practical implementation of work- software implementations often hides minute but crucial details.

CERN, Geneva, Switzerland. "Sheffield University, Sheffield, UK. "Stuttgart University, Stuttgart, Germany. "Heisinki Institute of Physics, Heisinki, Finland. Cambridge Liniversity Cambridge LIK (NYL) New York NY USA "Liniversity of Washington Seattle WA USA "Liniversity of Notes Dame Notes Dame N. USA "Instituto de Fisica de Cantabria CSIC-UC, Santander, Spain "Heidelberg University, Heidelberg, Germany, "Present address: DataCite, German National Library of Science and Technology, Hanover, Germany. ^oPresent address: CSC, Espoo, Finland. 'e-mail: sunje.dallmeier-tie

NATURE PHYSICS | VOL 15 | FEBRUARY 2019 | 113-119 | www.nature.com/haturephysics



particle detectors run by the experimental collaborations ALICE, ATLAS, CMS and LHCb. The raw experimental data is further filtered and processed to give the collision dataset formats that are suitable for physics analyses. In parallel, the computer simulations are being run in order to provide necessary comparison of experimental data with theoretical predictions. b. The stored collision and simulated data are then released for individual physics analyses A physicist may perform further data reduction and selection procedures, which are followed by a statistical analysis on the data. Physics results are derived taking into account statistical and systematic uncertainties. The results often summarize which theoretical models have predictions that are consistent with the observations once background estimates have been included. The analysis assets being used by the individual researcher include the Information about the collision and simulated datasets, the detector conditions, the analysis code, the computational environments, and the computational workflow steps used by the researcher to derive the histograms and the final plots as they appear in publications, c. The CERN Analysis Preservation service captures all the analysis assets and related documentation via a set of 'push' and 'pull' protocols, so that the analysis knowledge and data are preserved in a trusted long-term digital repository for preservation purposes. d. The CERN Open Data service publishes selected data as they are released by the LHC collaborations into the public domain after an embargo period of several years depending on the collaboration data management plans and preservation policies. Credit: CERN (a): Dave Gandy (b.c. code icon): SimpleIcon (b.c. even icon): Andrian Valeanu (b.c. data icon): Umar Irshad (c. paper icon): Freepik (c. workflow icon).

Table 1 | Terminology related to reproducible research ntroduced by Carole Goble and Lorena A. Barba Descriptio

erun	Robust	Variations on experiment and set-up conducted in the same lab Same experiment, same set-up, sam lab Same experiment, same set-up, independent lab			
epeat	Defend				
eplicate	Certify				
eproduce	Compare	Variations on experiment and set-up independent labs			
		and the second se			

potentially leading to a loss of knowledge concerning how the results were obtained In absence of solutions for analysis capture and preservation

knowledge of specific methods and how they are applied to a given physics analysis might be lost. To tackle these community-specific challenges, a collaborative effort (coordinated by CERN, but involving the wider community) has emerged, initiating various projects some of which are described below.

Reuse and openness. The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by internal review processes". Because these reviews often involve the input of the entire collaboration, where the level of crosschecking is extensive, the measurements are considered trustworthy.

NATURE REPORTS 1 VOL 15 LEERRUARY 2019 LTIS, 119 Leave nature com/habitenities

Open is not enough:

- define your reproducibility.
- runnable usage examples.
- preserve computational workflow.

• enable FAIR reuse

Nature Physics 15, 113–119 (2019)

FAIR principles

scientific data

Explore content Y About the journal Y Publish with us Y

nature > scientific data > comment > article

Comment Open access Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... <u>Barend Mons</u> + Show authors

<u>Scientific Data</u> 3, Article number: 160018 (2016) Cite this article

732k Accesses | 5469 Citations | 2223 Altmetric | Metrics

An <u>Addendum</u> to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A

Scientific Data 3, 160018 (2016)

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- 11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- 13. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

Latin America involvement at the LHC

Brazil Colombia Ecuador Mexico



Brazil Colombia

Argentina Brazil Chile Colombia

M. Mulders, II LASF4RI 6

Brazil Cuba Mexico Peru

LHC data collection:

• 90 petabytes (PB = 1024 TB) per year from LHC + 25 PB from other experiments

LHC data storage:

- CERN data center has ~400 PB on tapes + ~350 PB on disks
- CERN storage system served 2.5 exabyte (EB = 1024 PB) of data (just in 2020!)

"...and the results of its experimental and theoretical work shall be published or otherwise made generally available"

CERN Founding Convention (1953)

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CONVENTION

FOR THE ESTABLISHMENT OF A EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

PARIS, 1st JULY, 1953

ALICE data preservation strategy

Sunday, October 6, 2013

The data humsdall by the ALCL Department as to now and to be threated in the future constructs the neural information in the simular instructs by the intermental community, then data or and using assemble information to the in-information structure by the intermental community, then data or and using a information to the in-information with the second address or of origin and instruct. Unservice and the intermentation with the second address or of origin and the intermental and with the dataset of the ALCL Collideration larges to the size restrict ammunities will as the the grant place. These could be address of the ALCL Collideration interpret systems and an exclusions will be a structure and the address of the ALCL Collideration interpret systems and an exclusions will be a structure and the large extension of the address of the ALCL Collideration in the size intermediation of the address of the address of the ALCL Collideration in the size is structure. The size is structure is the greaner address areas data for exclusional process or for the extent assessment of the public the size the greaner address areas data for exclusional process or for the extent assessment of the public the size the greaner address areas data for exclusional process or for the extent assessment of the public the size the greaner address areas the for exclusional process or for the extent assessment of the public the size the greaner address areas the for exclusional process or for the extent assessment of the public the size the greaner address areas the for exclusional process or for the extent assessment of the public the size the greaner address areas the for exclusional process or for the extent addresses of the exclusional process of the exclusion

ALICE data formats

The level of sharestion of AUCE data accesses as every step of the data processing dubin storing from basic new data delivered by the detectors of the experiment owhing into stypics analysis ready data and ending with physics data studielies from planestic and page of the data processing and page rated and a salitared to data studielies from planesters are involved to the data processing activity and a salitared to detector observations and ending the data processing to the data activity data and ending with a data detector observation. The ends with the data processing activity and ending the data activity of the data detector observation. The processing data activity of the data processing activity of the data activity of the data processing activity of

- a) Rear data embedding the signal delivered by the detectors along with the associated status data constaining various information on the number constitutes the primary information collected by the AUCL agariment. They provide the input of the reconstruction algorithm, together with the calibration data stored in a decisted database:
- Monte-Carlo data, including data at the event generator level (MC truth) and data mimicking the raw data format (data), and/ored to real data reproducing the running conditions;
- c) Event Summary Data (ESD) produced by the reconstruction algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific information of the event of th
- real care break point careful to be the series of the seri
- e) Custom analysis object data, used stands brie or together with the general purpose ADD for specific analysis
- f) Published physics results and highly abstracted data resulting from the analysis.

This offerent formatis of the AUCL data is at to a poofic scheme for data presentation. While formatic can charge with time, the collaboration provides confruence relates untable to read indepression systematic, or altern twice migrate data freen one format to another, since processed data can exist in several versions, only the version used for the final guildation of the resentation concludes the model and an exist in several versions, only the version used for the final guildation of the resentation concludes the data measurement.

The AUCE Computing Model includes the provision for permanent storage of two copies of the new data. They are not presently being considered for span access, but they can be reprocessed at any time by members of the AUCE collaboration upon approval by the AUCE Physics Board. The original datasets used to produce published results, target her with head quarks of there we show in fifthmerowitz and marcoid are being estimated to long corm presentation.

Approved CB 20th June 2014

ATLAS Data Access Policy May 21# 2014

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS as regards open access to data at different levels as described in the DPHEP [3] model. The main objective is to make the data available in a usable way to people external to the ATLAS collaboration.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them openaccess. To support openaccess to data additional resources will be required to develop and support the tools to make the data available.

Policies for Different Data Levels

Open access to AILAS data by people outside the collaboration can be envisited at form level of increasing on dony, lined below, with associated contrains, see field, [1]. This policy parts in to collision physics data ji as, that are stored off instand intended for physics analysis and the encessary associated metadata, along with the acclused with added dutants and tools allowing to produce new simulated distancts based on an adequate simulation of the ATLAS-feetcor.

Level-1. Published results

All scientific output is published in journals, and preliminary results are made available in Conference Notou. All are openly available, without restriction on use by external parties beyond copyinght law and the standard conditions agreed by CERN.

Data anostan with journal pairtings on we also made a validate tables and data from figure (e) provisions and an isolahoop and pairtings in the strength of t

level-2. Outreach and Education

ATUA's recognizes the situal role of outreach and education, and participates in and encourage outreach and education activities, and makes selected data available for them. Typically a fraction of the complete ATUA So date sets to acts, selected to provide a roll sample of events with interesting physics signatures but nat adequate for a publication of a physics result. The data are provided in simplified, portable and self-contain of comma for mask for the self-contains for an anti-

Uses an L. Approved by the CMS Collaboration Board 20th April 2021

CMS data preservation, re-use and open access policy

ONS data are unique and are the result of wast and brag-term moral, human and financial investment by the international community. There is unique scientific opportunity in revealing base data, at different level of absorbation and at different points in the . This opportunity calls for our collective repensibility, and poses unprecedented challenges as no data sample of this completity and wake has ever been preserved or made available for litter revease.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their reuse by a wide community including: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outmachinitiative, and cities n clientists in the general public.

CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to faily exploit their scientific potential.

This policy describes the CMS principles of disa presentation, reuse and open access, as well as the research access in a three tasks and three roles and reasonables. CAS understands that in order to follywapibul all three reuse opportunities, immediate and continued resources are needed. The level of support that CMS will be able to power to extrant users depends on the available transfer. This power, address the model angle control states and the advection of the available transfer. This power denotes worksholds and the foll power for the more advection of notating advections the transfer and the foll access the more of section field.

Notwithstanding the long-term perspective of the UHC programme, the time for action is now: lowerenergy and lower-luminosity UHC runs at entired-mass arrangies of 0, 0, 3, 3, 2%, 7 and 8 TWC may never be repeated, and their preservations and proparation for latter reuse, has to be addressed outgently. Maketing this challings is a unique way to stress test and evaluate the entire preservation, re-use and open access concerning for the CMS data.

Old data bala many from: Starting from either raw expenses of e-multised data through to construct total and a start of tight activation granted baraying workflow, and Hully all the way to also any entry activation. Unit of even symm has the potential autout manual activation of the start of the star

¹A. Hidnes: P. Ba-Hennesen, S. Mele, "First results from the PARSE insight project: HEP survey on data preservation, re-use and lopest access? http://oriviergi.doi.0006.0085 *#Bac/Anapiere.pn.ort LHCh Datested Data Access Palicy LHCh Public Note Nation 1 Data No bast of specified style in decament. Reference: LHCb-PCB-2013-003 Redular: 7 Lest modifiet: 22th April 2013

none 1

Abstract

This document contains the LHCb Data Access Policy. This was adopted at the Collaboration Boost meeting of 27^9 Feb 2013.

Data Access Policy for LHCb

- I. Data preservative is fundamentally important for the onlinkous length regardless of any external requirements. This is to mishe colluboration members to access data for many years after it was then an frequires a consistent set of the data, associated software, metadata and conditions and documentation to be preserved. LHOs will seek to develop such a data preservation capability associate particle. We will have to develop action a data
- LHCb supports the principle of open access. In principle we can envisage providing some such open access based upon the work needed internally for data preservation (point 1 above).
- LHCh is extremely resource limited at present. Therefore whilst this policy expresses a split of
 intent, we cannot exemute to implementation of any capability on any specific time scala.
 Specifically is respect of open access we will not be able to undertake any significant development
 to support this without high-time of add from it resources.
- 4. Overall the collaboration expects to follow the guidelines being developed by CERN and the LHC experiments jointly on these matters, after appropriate approval by the LHCs Collaboration Board.
- 5. Open access to its data by people studiet the collaboration can be considered at four level of increasing ecosyloxity. Birth below, with associated ecositions. [note: these "levels" 1-4-are those relating in the DPHDE model, and are a data of an elevent on a study and the experiments.] In this first iteration, this policy only pertains to collision physics data (i.e. that sent offline and destined for physics analysis).
- 6. This policy is adopted by LHCb in good faith according to the spirit of the principles. The collaboration reserves the right to review the policy at any time in the light of experience including, but not limited to, the policy being found to be inadequate in the light of actual requests or any other unitstended consumences arising.









Restricted data \rightarrow embargo period \rightarrow open data

Launched in November 2014:

- Collision and simulated datasets for research
- Derived datasets for education
- Configuration files and documentation
- Virtual machines and container images
- Software tools and analysis examples

As of today:

- over 40.000 datasets
- over 50 different software
- over 5 petabytes of data



Each layer of the ATLAS detector is sensitive to different types of particles and/or radiation produced in the collision: photons, electrons, muons, charged and neutral hadrons...



Software

Data analyses require hundreds of petabytes of data storage and hundreds of thousands of CPUs to execute millions of lines of code.

ATLAS has collected ~4 PB of data so far.

Data reconstruction and preprocessing take ~100 PB of space.

The software for reconstruction, simulation, and analysis consists of ~4 million lines of C++ code.

None of this is trivial to communicate and share with the public...





April 2024: CMS publishes proton collision data at 13 TeV from 2016. More than 70 TB of 13 TeV data and 830 TB of MC.

February 2024: The ATLAS Experiment releases 25% of the main proton-proton datasets for research purposes.

December 2023: LHCb releases the complete Run I dataset. 800 TB of data and algorithms for research and education.

December 2022: CMS completes the release of all its Run I proton-proton collision data. 491 TB of data and code.

February 2020: The ATLAS Experiment releases open data at 13 TeV for education.

November 2014: ALICE releases educational datasets. ALICE's datasets customized for demonstration and education.

Open data available from LHC experiments

CMS releases 1.3 TeV proton collision data from 2016
2014/2 Ty 040 Cateware
T
T
T
Cate Cateware
T
T
Cate Cateware
T
Catew



For the first these, the vicentity community has across to addecard databactel of 13 WV collisions. This release agreems the V2215 data set unulation that were made paths in 2021. Over 20,000 simulations of different adjusces processes have been released adorphot the collision data, as well as new software containers and a new inflam technic to the adjusces. The collision data data data databacter additional data

ATLAS releases 13-TeV open data for science education

and a subset of the collision data is evaluable in an exampled NancAOD format that includes information about particle candidates from the CMS "particle flow" algorithm

13-02-2020



The ATLAS collaboration has just released the <u>first open dataset</u> from the <u>Large Hadron Collider</u>'s (LHC) highest-energy run at 13 teraelectronvolts (TeV). The new release is specially developed for science education, underlining the collaboration's long-standing commitment to students and teachers using open-access ATLAS data and released tools.

ATLAS has made public 10 inverse femebarns (th⁻¹) of the 13-RV data, which corresponds to about 1 quadrillion proton-proton collisions, or 500 thousand produced <u>Higgs horons</u>. It is also approximately the same amount of data that the ATLAS collaboration used to discover the Higgs boron in 2012. The datasets, software and tools are available on the <u>ATLAS public website</u> and on the <u>CERN Open Data Portal</u>.

"Our high-energy collision open data, recorded during the second run of the LHC, provides insight into the real world of particle-physics analysis. Students, scholars and interested members of the public will be able to reproduce ATLAS physics results in a fully realistic manner, understanding for themselves the fascinating study of nature at its deepest level", spays kard LASs physics preson.

ATUS has also released new simulated data sets and web-based and <u>effine analysis coltrains</u> as well as extensive documentation and tutorital. "These are the tools of a particle physicits t trade, allowing us to go from data-taking to physics measurements and eventually discovery", says Atruo Sainchez Printead, co-leader of the ATUS Spine tatus mit (Nineviry) of Udine, ICTP and NIFN, Italyi. "Simulated datasets allow physicits to compare theory with real data. They are based on theoretical models of the expected physics processes taking place in the collisions, together with a detailed description of the ATLAS Genet. By providing such resources, we hope to empower students, professors and dedicated self-learners worldwide to learn and teach experimental particle physics, as well as the comparer science binth the field."



Interactive event display and histogram creation using small datasets

Research-oriented use cases





CERN Virtual Machines

Simplified research-level analysis

Independent theoretical research



Welcome to <u>INSPIRE</u>, the High Energy Physics information system. Please direct questions, comments or concerns to <u>teedback@inspirehep.ret</u>.

Search took 0.12 seconds.

HEP :: HEPNAMES :: INSTITUTIONS ::	CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS
reference:10.7483/OPENDATA.CMS	Brief format Search
find i "Phys.Rev.Lett. 105" :: more	Search the new INSPIRE

 Sort by:
 Display results:

 latest first
 v
 desc. v
 times cited
 v
 25 results
 v
 single list

No exact match found for 10.7483/OPENDATA.CMS, using 10 7483 OPENDATA CMS instead...

HEP 35 records found 1 - 25 jump to record: 1

1. Exposing the QCD Splitting Function with CMS Open Data

(3) Andrew Larkoski (Reed Coli,), Simone Marzani (SUNY, Buffalo), Jesse Thater, Aashish Tripathee, Wei Xue (MIT, Cambridge, CTP). Apr 17, 2017. 7 pp. Published In Phys.RevLett 10 (2017) no.13, 132003 MIT-CTP-4891 DOI: 10.1103/PhysRevLett.110.132003

Doi: <u>10.1103/P1/9104.05066 [hep-ph] [DDF</u> <u>References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote</u> <u>ADS Abstract Service</u> Detailed record - Cited by 35 records

2. Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks

(32) Pasquale Musella (ETH. Zurich (main)), Francesco Pandolf (INEN, Rome). May 2, 2018. 8 pp. Published in Comput.Softw.Big Sci. 2 (2018) no.1, 8 DOI: 10.1007/e11781-018-0015-y e-Print: arXiv:1305.00850 [hep-sc] [PDF References | BitTe3, LIERA(US) | LaTEX(EU) | Harvmac | EndNote ADS Abstract Service Detailed record - Cited by 33 records

3. Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2

 (19) LHC Reinterpretation Forum Collaboration (Waleed Abdallah (Harish-Chandra Res. Inst. & Cairo U.) *et al.*). Mar 19, 2020. 58 pp. Published in SciPost Phys. 9 (2020) no.2, 022
 CERN-LPCC-2020-00.1; FENILAE-RN-L089-CMS-T, Impenal/HEP/2020/RIF/01
 DOI: 10.214693-SeiPostPhys.9.2.022
 e-Print: arXiv:2003.07868 [hep-ph] [PDF
 References [BirtsX] LaTeX(LS) [LaTeX(EJ) [Harvmac [EndNote CERN Document Server; ADS Abstract Service: OSTLoov Server; Link to Fermiab Library Server (fulltext available); Link to Fulltext from Publisher; Link to Fulltext

Detailed record - Cited by 19 records



arXiv:1704.05066

arXiv:1807.11916

arXiv:1902.04222

Searches, QCD jet studies, Machine Learning...

A measurement of the z_g distribution in pp collisions, using CMS open data, was recently reported [33, [34]]. In PbPb collisions, this measurement reflects how the two color-charged par-

- [33] A. Larkoski et al., "Exposing the QCD splitting function with CMS Open Data", Phys. Rev. Lett. 119 (2017) 132003, doi:10.1103/PhysRevLett.119.132003, arXiv:1704.05066
- [34] A. Tripathee et al., "Jet Substructure Studies with CMS Open Data", Phys. Rev. D 96 (2017) 074003, doi:10.1103/PhysRevD.96.074003, [arXiv:1704.05842]

New theoretical ideas... being already cited by LHC experiments

arXiv:1708.09429v

Test-bed for new ideas

		MIT-CTP 4801		MIT-CTP 1128		
	Exposing the QCD Splitting Function	n with CMS Open Data	Explorin	ig the Space of Jets with CMS Open Data		
	Andrew Larkoski, ^{1,1} Simone Marrani ^{1,1} Jesse Thaler. ³	Aschish Tripathee, ^{3,8} and Wei Xue ^{3,4}	Patrick T. Kominke, ^{1,2,*} Radha	Mastandren, ^{1,3} Eric M. Metodiev, ^{1,2,3} Preksha Naik, ^{1,3} and Jesse Thaler ^{1,2,¶}		MIT.CTP 514
Biomyin: Mag 17, 2019	¹ Physics Department, Bool College, Per ² /Integration of Reficies The State Department of New	Hend, OR 97092, USA	¹ Center for Theoretical P	Syntes, Massachusetts Institute of Technology, Corderidge, MA 02129, USA of Physics, Revenue University, Combridge, MA 09198, USA		SHT-011, 218
Bavinus: November 10, 1019 Accession December 5, 1019	¹ Center sig al Baffalo, The Stole University of New ¹ Center for Theoretical Physics, Massachusetts Institute of	Technology, Cambridge, MA 62378, USA	We explore the metric	space of jets using public collider data from the CMS experiment. Starting		
PUBLICITY Decoder 16, 2019	The splitting function is a universal property of quantum how experts is shared between matters. Togethe its abiration	chromodynamics (QCD) which describes	from 2.3 fb ⁻¹ of proton- 2011, we bedate a same	proton collisions at $\sqrt{s} = 7$ TeV collected at the Large Hadron Collider in is of 1.690.394 control acts with transverse momentum above 375 GeV. To		
	the splitting function cannot be measured directly, since it	always appears multiplied by a collision	validate the performance	r of the CMS detector in reconstructing the energy flow of jets, we compare		
Testing non-standard sources of parity violation in jets	singularity factor. Recently, however, a new jet substra seymptotes to the splitting function for sufficiently high jet	sture observable was antroduced which energies. This provides a way to expose	substructure observables	eccremponding annualed data samples for a variety of jet kinematic and a fiven without detector unfolding, we find very good agreement for track-		
the LUC A '- H- d - 'th CMC O D-t-	the splitting function through jet substructure measureme letter, we use public data released by the CMS experime	sta at the Large Hadron Collider. In this at to study the 2-prong substructure of	perform a range of nove	using charged hadron subtraction to miligate the impact of pilcup. We el analyses, using the "energy mover's distance" (EMD) to measure the	The Hidden Geometry of Particle Collisions	
t the LHC, trialled with CMS Open Data	ipts and test the 1 → 2 splitting function of QCD. To our analysis based on the CMS Open Date.	knowledge, this is the first over physics	pairwise difference betwee officete, visualize the met.	on jet energy flows. The EMD allows us to quantify the impact of detector on space of lets, extract correlation dimensions, and identify the most and	0	
	H IO		Ci least typical jet configure detects and analysis co	ations. To facilitate future jet studies with CMS Open Data, we make our de analable, amounting to around two stealastes of distilled data and one	0.2	
Indexember C. Louised and Matthias Related	[4] Quantum chromodynamics (QCD), like any weakly direct	y measure the splitting function $P_{t-ijk}(z)$ in data,	Ci hundred gigabytes of sim	milation film.	6	
Precipiter G. Dester- and machinas schedu	small angle limit. When two partons because collinear that I	there is of course overwheiming indirect evidence (z) is a universal function from the many suc-	eb		Patrick T. Komiske, Eric M. Metodiev, and Jesse Thater	
33 Thomson Assense, Cambridge, United Kingdom	in QCD, the cross section for a $2 \rightarrow n$ scattering pro- esses factorizes into a $2 \rightarrow n - 1$ scattering resulting and N	of QCD in describing high-energy scattering (see 7-67).	CONTENTS	1. INTRODUCTION	Constraint for resonance reprice, Managements Institute of Technology, Cambridge, C	repr. re.a. 60139, 635
Manachaetta Institute of Technology, Cambridge, U.S.A.	C multiplied by a universal 1 → 2 splitting probability, In (his letter, we present a semi-direct method to test	L Introduction	I Ever since the first evidence for jet structure [1], the	A-mes: pRomiskeGmit.edu, metodievOmit.edu, jthalerOmit.edu	
Institute of Physics, Johanness Catenberg University,	with corrections suppressed by the degree of collinear- ity. Collinear universality is a fundamental property of home	→ 2 splitting function in QCD by studying the substructure of lets. Our method is based on	II. Processing the CMS Open Data	2 Imgeneration of mort-distance quarks and gluons into 2 long-distance hadrons has been a rich area for experi-	ABSTRACT: We establish that many fundamental concepts and technique	www.iw.quantum.fiel
Raudingerung, Mainz, Germany	QCD and appears in many applications, most farmously work d	up declustering [68] (see also [51, 69, 70]), which	1. J. Jet Primary Dataset	mental and theoretical investigations into quantum chro-	Gir three and pollider physics can be naturally understood and unified the	rough a simple no
S-mail: lester@hep.phy.cam.ac.uk, matthias.schott@cers.ch		from a jet until hard	D. Monte Carls Event	BIYSICAL REVIEW D 100 015021 (2019)	agaage. The idea is to equip the space of collider events v	with a metric, fror
ervery. The Standard Model sideres parity but only be mechanisms which are	Connection and Suffman for No. Science (2020) 5.11	no intrinsic substruc-	E. Jet and Trigger Sele	THE OCAL REVIEW D 100, 015051 (2017)	geometric objects can be rigorously defined. Our analysis is i	based on the energy
de to Large Hadron Collider (LHC) experiments (on account of the lack of initial	https://doi.org/10.1007/v41781-621-00060-4	e core of the jet. As sharing between the	III. Analyzing Jot Substruct		which operates purely at the level of observable energy flow	information, allow
farisation or spin-sensitivity in the detectors). Nonetheless, new physical processes	ORIGINAL ARTICLE	dated to the momen-	B. Jet Constituents	Searching in CMS open data for dimuon resonance	d definition of infrared and collinear safety and related co	oncepts. A numbe
containty violate parity in ways which are detectable by those same experiments. If array of new releases ecour only at LHC gravitation, they are unitated by first our		thing function in the	C. Jet Substructure Ob	with substantial transverse momentum	n collider observables can be exactly cast as the minimum	m distance betwee
is probe the feasibility of such measurements using approximately 0.215-1 of data	Analysis Specific Fact Simulation at the LHC with Door Looping	a, have appeared in	O IV. Exploring the Space of A. Baylow of the Energy	Carl Countril,17 Yourn Seren 237 Mathew J. Strander 18 Jacob Wales 518 or	d Wri Xur ¹² Jet definitions, such as	mobulys cone and
as recorded in 2012 by the CMS collaboration and made public within the CMS	Analysis-specific rast simulation at the LHC with Deep Learning	ably the \sqrt{y} param- if our knowledge, no	00 B. Quantifying Detecto	¹ Department of Physics, Barrord University, Cambridge, Massachusattr 0213	6. USA potentiation algorithms, can be directly derived by finding the to the event. Several area, and constituent, based allows a	a causest new-particle mitigation strategie
ata initiative. In particular, we test an inclusive three-jet event selection which is pa	C. Osen ¹ - O. Cerri ² - T. O. Nguyen ¹ - J. R. Wimant ¹ - M. Pierini ³	born presented using	D. Carrelation Dimensi	² Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland ² Department of Physics, Technics, Haile (2000, Jurad)	expressed in this formalism as well. Finally, we lift our re-	easoning to develo
y senserve to non-southern parity violating effects in quark-gravit inferactions. W		of ALICE [75]. Here,	E. The Most Represent 1. F. Towards Amounts II	⁶ Center for Theoretical Physics, Massachusetts Institute of Technology	tance between theories, which are treated as collections of	events weighted b
our experimental limitations have been found. We discuss other ways that search	Received: 1.2 October 2020 / Accepted: 1.2 May 2021 / Published online: 9 June 2021 (2) The Authority 2021	ing LHC data, taking	S V. Ondoriza	cambridge, Manachaoette (92159, 658	 In all of these various cases, a better understanding of exist 	ting methods in ov
on-standard parity violation could be performed, noting that these would be sensit or different anti- of module to these which our method would concern. We have		, this research by the	H Advandedaments	(Received 8 March 2019; published 16 July 2019)	iguage suggests interesting new ideas and generalizations.	
ir initial studies provide a valuable starting point for rigorous future analyses usin	Abstract	rem 7 TeV center-of-	A Mining and Zenard Law	We study dimuon events in 2.11 fb ⁻¹ of 7 TeV pp collisions, using CMS Open Data, nervos dimuon resonance with moderate muss (14-66 GeV) and substantial transverse r	and search for a committee (n-)	
A LHC datasets at 13 TeV with a careful and less conservative estimate of experim	we present a sam-semilation application based on a deep reseat network, designed to create large analysi Taking as an example the generation of W + jot events produced in √s = 13 TeV proton-proton cellision	s, we train a neural Open Data Portal in	The second and second Lab	Applying dismon p_{γ} cuts of 25 and 60 GeV, we explore two overlapping samples: o	se with isolated	
acctalaties.	network to model detector resolution effects as a transfer function acting on an analysis-specific set of	f relevant features, wided in AOD (Anal- a CMS suscille data	D. Aspera st Passp	meens, and one with pumpt muons without an isolation requirement. Using the latter information about detector effects and QCD backgrounds, which we obtain detects from	sample requires the CMS Open	
EVWORDS: Exotics, Hadron-Hadron scattering (experiments), proton-proton scatt	compared at generation seven, i.e., in ansence or defector effects. Based on this model, we propose a no workflow that starts from a large amount of generator-level events to deliver large analysis-unocific sam	ples. The adoption work [78]. Crucially	C. Authloral Plots	Data. We present model-independent limits on the product of cross section, branching frac-	ian, acceptance,	
P violation, Jet physics	of this approach would result in about an order-of-magnitude reduction in computing and storage requir	ements for the col- w candidates (PFCs)	References	and errorences, ruese times are spreager, relative to a corresponding inclusive search will factors of an much an 9. Our "pr-enhanced" demote search strategy provides improve	out a py cut, ny. ni amnihivity to	
a Yas a Danser, 16th 11105	origination weeknow. This strategy could help the high energy physics commanity to face the co- of the future High-Laminosity LHC.	[MS [81], and we can		models in which a new particle is produced mainly in the decay of something heavier, as meaneds, in decays of the linear parts of a TeV scale tra-	could occur, for	
And a more second the	Rest Hill GRI B. J. D. R. L. B. L. J. H. S. S. S.	associated conditions	plossisle/fmit.edu	with the current 13 TeV than should improve the sensitivity to such signals forther by nu	ghty an order of	
	neywords matron Constant mysics - rost tematation - Deep Learning - High Energy Physics computing	torrection (JEC) fac-	* renetand@mit.edu 8 metcdke/Omit.edu	magnitude.		
un Accust, @ The Ashers	Interdenting and a second s	e studies. The main	* predoit an finit.ofe * jihake-fire it.osh	DOI: 10.1105/PhysRevD.100.015021		
Jele funded by SCOAP*. https://doi.org/10.1007/JH02P12(20	introduction rely extensively on an accurate simula processes under study, including a deta	are of the physics fied description of				
	At the CERN Large Hadron Collider (LHC), high-energy the response of their detector to a given a	et of incoming par-		L INTRODUCTION detector effects. The value of	the CMS Open Data for	
	proton-proton (pp) cottistors are strained to consolidate our ticles. These large sets of synthetic data understanding of physics at the energy frontier and possi- ated with experiment-specific simulation	we typically gener- software, based on		The CERN Open Data portal [1] aims to make data from Refs. [4, 5] ser Refs. [6, 8] for	has been demonstrated in machine-demonstrated on	
	bby to search for new phonomena. While these studies are the CELINE 4 [1] library. Through Mont	Carlo techniques,		the Large Hadron Collider (LHC) publicly available as a long-term archive, with the first research-arade data from	les, Refs. [9-11] for QCD	
	typically conducted according to a data driven methodology, GEANT4 provides the state of the art i synthetic data from simulated pp collisions are a key ingre- tion accuracy. The first two runs of the	n terms of simula- LHC highlighted		the CMS experiment released in 2014 [2]. In order to studies on archival ALEPH dipheton analysis with rable	data, and Ref. [12] for a LHCb data.	
	dient to a robust analysis development. Particle physicists the remarkable agreement between da	ta and simulation,		identify any issues that might interfere with their use by physicists of the future, it is important that open data	first utilization of the CMS	
	with discrepancies observed at the leve On the other hand, running UTAPT 4 in a	i of a tew perceit. Ionanding in terms		frameworks be tested today. There are pood scientific (BSM) obsources We sack	eyona me nandara monel new particle V that decays	
	50 M. Plovini of ressurces. As a consequence of this, a	elivering synthetic		motivations to make use of this resource [3]. Open data make it possible for scientists outside of the LHC collab-) and is typically produced	
	C. Chen the most challenning tasks for the common	s reat data is one of ting infrastructures		erations to study standard model (SM) processes that are based on 2.11 th ⁻¹ of 7 YeV	entan (p)). Our analysis is enter-of-mass an collision	
	e.chm@cerr.ch of the LHC experiments. It is then more	and more common		not wen moduled by bullet Carlo (b(c) generators, such as rare QCD backgrounds. Together with detector-simulated	experiment during the first	
	O Corm for LHC physics analyses to be affected shroll-taken data success analyses to be affected success ainties due to the limited amount	by large systematic of simulated data.		samples, opon data also make it possible to test event. Data portal [13]. We perform a	mrough the CERCY Open narrow resonance search in	
	T.Q. Ngayen This is particularly true for precise mean	arements of Stand-		analysis strategies the roly on a detailed understanding of the dimuon mass range $m_V \in$	[14, 66] GeV and study the	
	and Model processes for which large d 1.8. Virtual	stasets are already th luminosity LHC		effect of modest cuts on p ₂ , 'consurvati@g.harvant.edu 60 GeV; this approach (which	samety, $p_{\gamma} > 25$ GeV and we refer to as p_{γ} enhanced)	
	j-diman@cakech.edu upgrade, this will become a serious prob	iem for most of the		ystan serue @cert.ch staniler@physics.havaal.edu could be applied to larger py vi	hoes as well, or alternatively	
	* State Key Laboratory of Nuclear Physics and Technology. ILHC data analyses [2]. Our community the computing applied for community	is called to reduce		Sphaler@mit.edu to a cut on the V boost factor	p_{T}^{*}/m_{T} . This type of search time are [14], as one of	
	School of Physics, Peterg University Hastan, flows by at least one order of magnitude	, not to jeopardize		several unconventional approx	ches for finding low-mass	
	² Califients Instrume of Technology, Pasadena, CA 91125, 1983. the accuracy gain expected when open high luminous.	sing the LHC at a		the Orazine Commission Attribution 4/0 International Internet and diphoton resconar	ces [15], but to our knowl- tan a mublic analysis by the	
	² European Organization for Nuclear Research (CERN).			rurmer distribution of this work must maintain estribution to the mathor(s) and the published article's title, journal classion, LHC collaborations. For this	reason, the mass and p ₇	
	1211 Geneva 23, Switzerland			and DOT. Funded by SCOAP'. regime we cover is relatively	anespicerol. Moreover, our	
		6 mm				
		F 4.4.		2470-0010/2019/100(1)/015021(20) 015021-1 Published by	the American Physical Society	
					ender-salmas espaded bro	

and much more

How can we overcome distances and enable anyone interested in particle physics and interaction of radiation with matter to learn remotely?

- How can students replicate the procedures used at CERN?
- How can teachers improve their own lectures and science communication?

The ATLAS Collaboration at CERN launched an educational platform (<u>opendata.atlas.cern</u>) designed to help students and teachers foster educational efforts at both undergraduate and graduate levels.











Open data stored at the <u>CERN portal</u> and on the <u>online webpage</u>

Comes with codes written in C++ and Python, uproot, pandas/numpy, pyROOT and RDataFrame, available on <u>GitHub</u>.

Interactive visual <u>data analysis</u> and <u>notebooks</u> <u>Jupyter</u> that allow one to get results online

<u>Virtual machines</u> that allow contains the operating system, software and data, and <u>containers Docker</u> for a local interface.





By atlaasopendata • Updated 10 months ago Jupyter notebook containing a fresh installation of ROOT@CERN, in both python and C++ kernel

Overview Tags

ATLAS Open Data ROOT notebook

Fostering educational efforts: the case of ATLAS



Recreate the main LHC discoveries over the past 10 years of data-taking: find the heaviest top quark, or the Higgs boson in various final states



CERN Open Data incorporated into the curriculum of several universities.



Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.



Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.



Educational workshops and schools using ATLAS Open Data worldwide, including Argentina, Colombia, Ecuador, Mexico, Peru, Uruguay, and Venezuela.



Educational efforts using CERN Open Data



Open Data Challenge

CERN International Teachers Programme





PHYSICS WITHOUT FRONTIERS CALL FOR PROPOSALS Train and inspire physics and mathematics students in the Global South

CIP's Physics Without Frontiers works to teach, train, and respire physics and Instituments aniversity shadest surdivaide, with focus on countries in the Globe down, its help dual the next generations of accentation. We arguing the end generation of a constraint, which are productored researchers, or lectrans from allow of an end mathematics. Each research survival and in developed with the countries area for marked in mark and an end mathematics.

...

Science education programs in Latin America:

LA-CoNGA ICTP Physics Without Frontiers CEVALE2VE (Centro de Altos Estudios de Altas Energías) PPGCosmo

25

BSc and MSc theses using CERN Open Data







Making ATLAS Data from CERN Accessible to the General Public (2017)



Reconstrucción de masas invariantes de bosones del Modelo Estándar (2017)



<u>A Contribution to ATLAS Open Data</u> <u>Collaboration at CERN (2019)</u>

Perspectivas y Evaluación de producción de Materia Oscura (2017)



Measurements of top-quark pair production using ATLAS open data (2021)

CMS Open Data Workshop & Hackathon 2024

- III 29, 2024, 2:00 PM → Aug 1, 2024, 6:00 PM Europe/Zurich
- IdeaSquare (CERN)

👤 Julie Hogan (Brown University, Bethel University (US)) , Kati Lassila-Perini (Helsinki Institute of Physics (FI))

Description Since 2014, the CMS Collaboration has pioneered the release of LHC research quality data for public use by making a significant amount of these data accessible through the CERN Open Data portal. Recently in 2024, the CMS Collaboration has released a sizeable new set of 13 TeV data collected in 2016.

This workshop is the fifth in a series that started in 2020 and it aims to bridge the technical gap that usually exists between the scientific creativity of an external analyst and the nuts-and-bolts details of a full analysis with CMS open data.

All exercises will be hands-on and participants should be prepared to dive into the data right away. A set of pre-exercises are provided and required for participants so that they can make the most of the workshop.

New for this workshop are morning **hackathon segments** where users with Open Data projects in mind can take advantage of the facilitators' knowledge to jump-start their



CMS Open Data Workshop & Hackathon July 29th - Aug 1st, 2024 *CERN IdeaSquare*

work. Tutorial sessions in the afternoons will give users at any stage a pedagogical grounding in the CMS experiment and relevant open data analysis techniques.

The tutorial segments of the workshop will be offered in hybrid mode. The hackathon working segments will be optimized for in-person participation.

Access the tutorial site here (coming in July)

Open data from CERN (each experiment having its own open data) has proven to be highly successful in outreach and education.

The tools are designed to assist students and teachers and foster educational efforts at both undergraduate and graduate levels.

I invite you to use them in your courses or mention their existence and potential use to your professors and teachers in their classes and science outreach efforts!

To start using CERN's open data, check out the tools and resources provided on the CERN Open Data portal, where access guides, analysis examples, and necessary software to process the data are offered.



https://opendata.cern.ch/