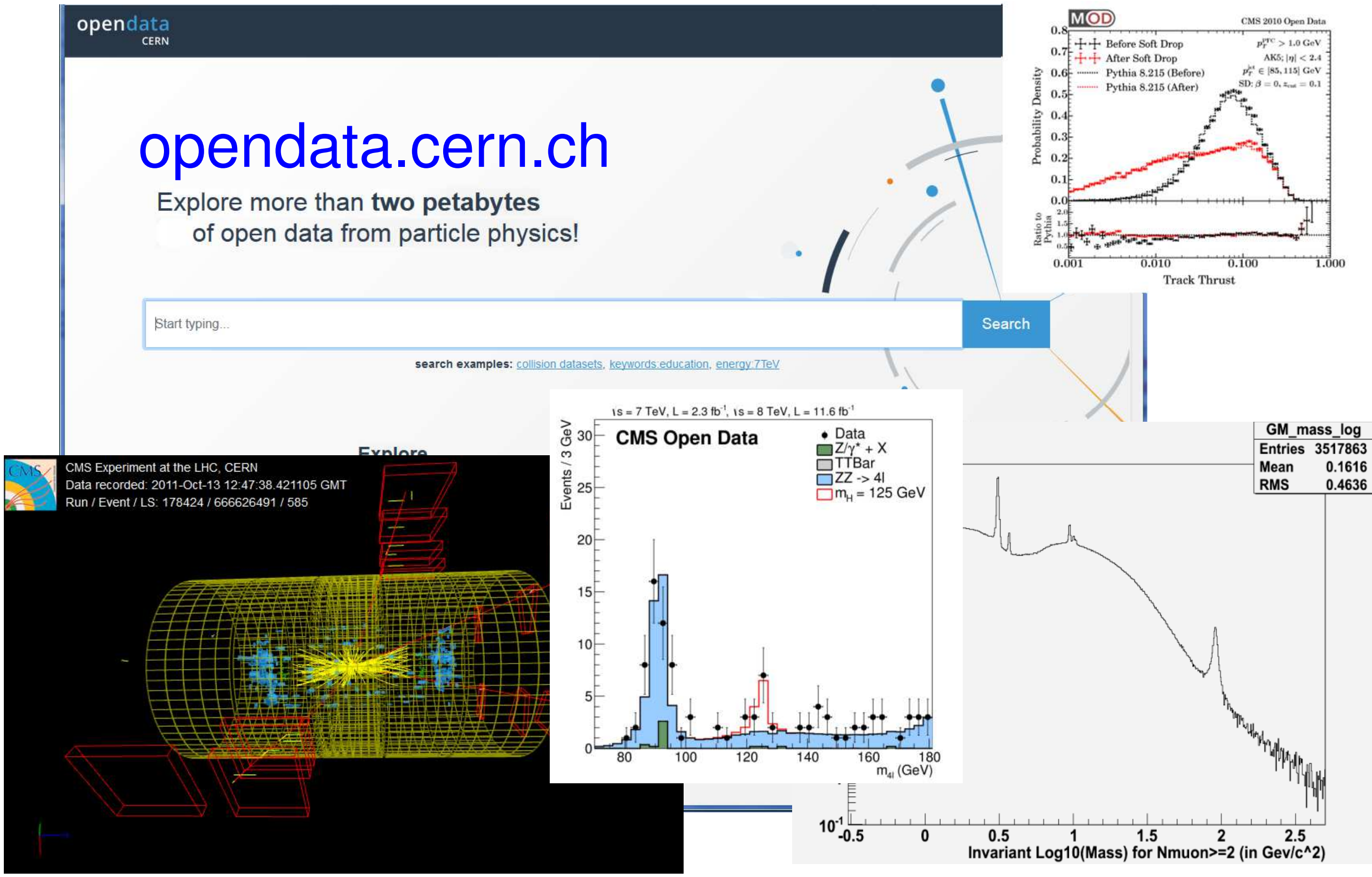
CERN Open Data, and open/preserved data in HEP

Achim Geiser, DESY Hamburg

member of the CMS collaboration (and of ZEUS/HERA, PROSA)

PUNCH4NFDI Open Data workshop, DESY remote, 11.02.2021



opendata CERN

opendata.cern.ch

Explore more than **two petabytes** of open data from particle physics!

Start typing...

Search

search examples: collision datasets, keywords:education, energy:7TeV

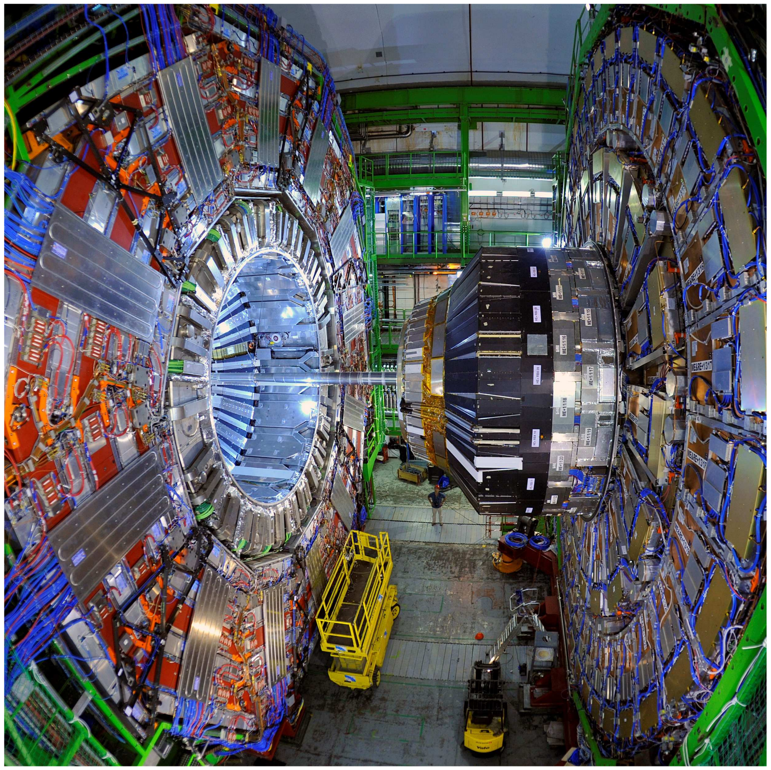
Explore

CMS Experiment at the LHC, CERN
Data recorded: 2011-Oct-13 12:47:38.421105 GMT
Run / Event / LS: 178424 / 666626491 / 585

- HEP in general
- The vision for LHC
- The implementation
- Conclusions

HEP and data preservation/Open Data

HEP (High Energy Physics) experiments:



- **years/decades of preparation**
- **very complex detectors**, complicated reconstruction and calibration of thousands or millions of channels of very different kinds, complex software environments, very large data volumes, tradition of “indefinite” embargo times
- **-> big challenge, both technical and sociological, for data preservation and Open Data**

(data preservation is prerequisite for Open Data)

(semi)open/preserved data from e^+e^- colliders: LEP (CERN, 1989-2000) and PETRA (DESY, 1979-1986)

https://dphep.web.cern.ch/experiment/aleph

partial data available to former members and their collaborators, at CERN

https://dphep.web.cern.ch/experiment/delphi

partial data available to members, at CERN

https://dphep.web.cern.ch/experiment/opal

partial data available to members, e.g. via MPI Munich

LEP overall: ~2% of publications in “archive” mode (after 2007/2008)
arXiv:0912.1803

https://wwwjade.mpp.mpg.de/ **JADE**

partial data available to interested parties at MPI (PETRA)

~9% of JADE publications in “archive“ mode (after 1994, „private“ initiative), S. Bethke, arXiv:1009.3763 contacts at MPI: Stefan Kluth, Andrii Verbytskyi

preserved/accessible data from ep colliders: HERA (DESY, 1992-2007)

for recent status and experiment contacts, see e.g.

https://indico.bnl.gov/event/9287/contributions/41457/attachments/30600/48033/EIC_2020.pdf

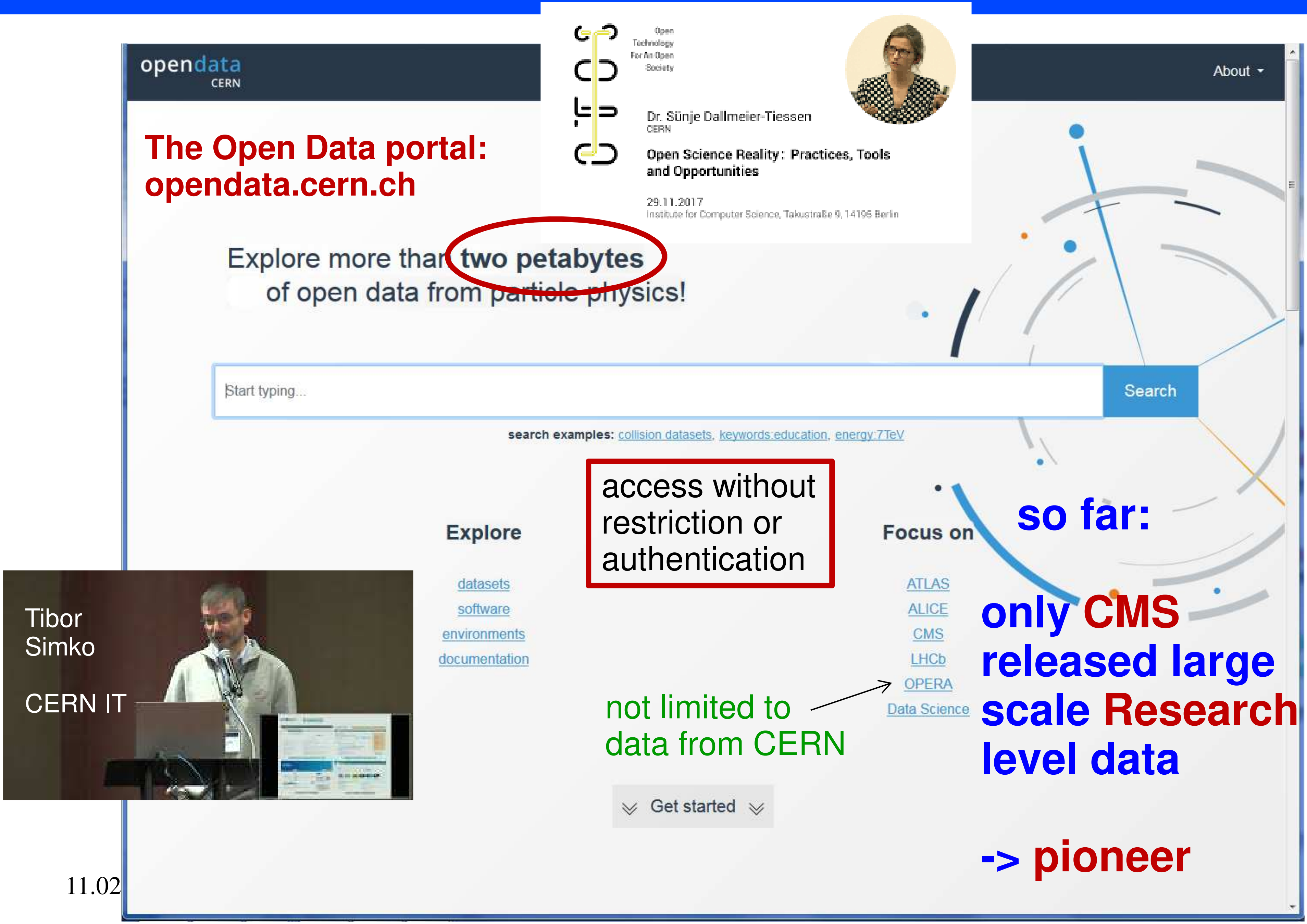
collaborations still alive (very small remaining person power, scientific review, but no explicit funding)

full original H1, Hermes, and ZEUS research data (internally) available in "**archive" mode** (since 2015) at DESY, parts at MPI; low threshold for **participation of external communities** (e.g. EIC, heavy ions, theory, …) -> many of recent publications from these

so far 28 papers in "archive" mode (out of 821, **3.5%**), more coming

Open Data possible in principle, but no resources/person power

CERN Open Data



Open Technology For An Open Society

Dr. Sünje Dallmeier-Tiessen
CERN

Open Science Reality: Practices, Tools and Opportunities

29.11.2017
Institute for Computer Science, Takustraße 9, 14195 Berlin

The Open Data portal:
opendata.cern.ch

Explore more than **two petabytes** of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

access without restriction or authentication

Explore

datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA
Data Science

not limited to data from CERN

so far:

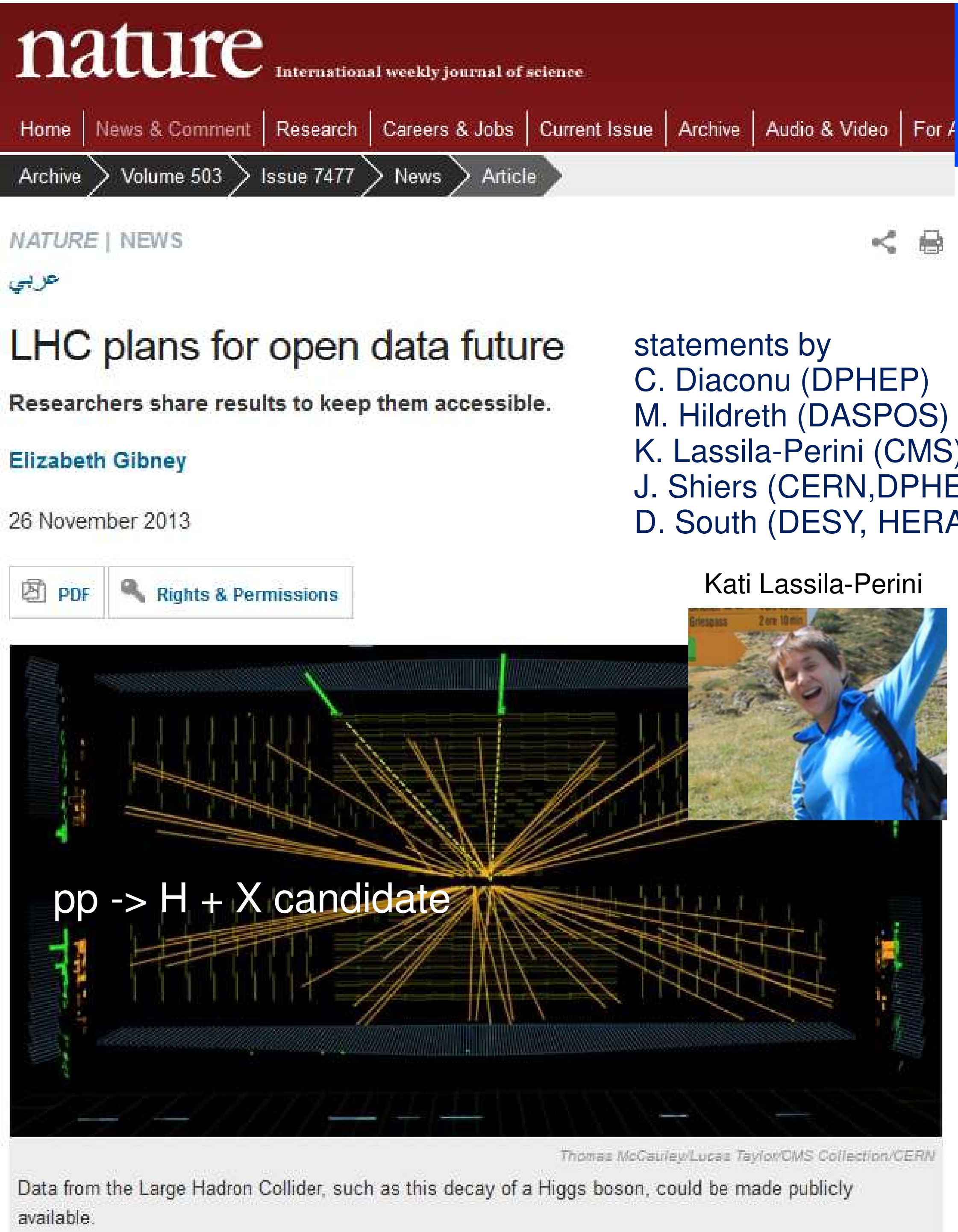
only CMS released large scale Research level data

-> pioneer

Tibor Simko

CERN IT

The Vision, 2013



LHC plans for open data future

Researchers share results to keep them accessible.

Elizabeth Gibney

26 November 2013

statements by
C. Diaconu (DPHEP)
M. Hildreth (DASPOS)
K. Lassila-Perini (CMS)
J. Shiers (CERN,DPHEP)
D. South (DESY, HERA)

Kati Lassila-Perini

pp -> H + X candidate

Thomas McCauley/Lucas Taylor/CMS Collection/CERN

Data from the Large Hadron Collider, such as this decay of a Higgs boson, could be made publicly available.

- **Preserve data and knowledge (metadata)**
- **Open sharing – data and knowledge more likely to survive if constantly used -> enlightened self-interest**
- **Make data available to school pupils and researchers alike - allow them e.g. to reconstruct the Higgs discovery**
- **(Allow CMS physicists to recreate results from ATLAS and vice versa -> backup)**
- **Mine data to test new theories and provide crucial references**
- **Contain cost to ~1% of operating costs -> worth the effort**

Extended Vision

my **personal extension of initial vision:**

(formulated 2015, **not** a collaboration statement)

with ~1% of additional resources aim to achieve

~10% additional scientific output (e.g. physics papers)

from both external and internal use of **preserved/open data**

over lifetime of experiment + 10-20 years

recent addition in view of PUNCH4NFDI:

enable common analyses of
HEP, Hadron, Astroparticle and Astrophysical data

CERN Open Data policy for LHC experiments

http://opendata.cern.ch/docs/cern-open-data-policy-for-lhc-experiments

made public on Dec. 11, 2020

policy for the release of Open Data at the various “DPHEP” data levels:

Level 1: Publications (open since a long time, e.g. arXiv, Inspire, open access journals), plus related information in machine readable form (e.g. HEPDATA), as well as **binned and unbinned likelihoods** (**main focus of ATLAS**, see **contribution Lukas Heinrich,** see also initiative S. Neubert, LHCb)

Level 2: Education and Outreach, simplified derived data sets (not the topic of this contribution, all collaborations contribute)

Level 3: Fully calibrated reconstructed data sets as used internally by the collaborations for their analysis (mainly **pioneered by CMS** since 2014, ~**20 papers** published so far from CMS Open Data, **~2%**, fraction rising) see also CMS Open data workshop, https://indico.cern.ch/event/882586/

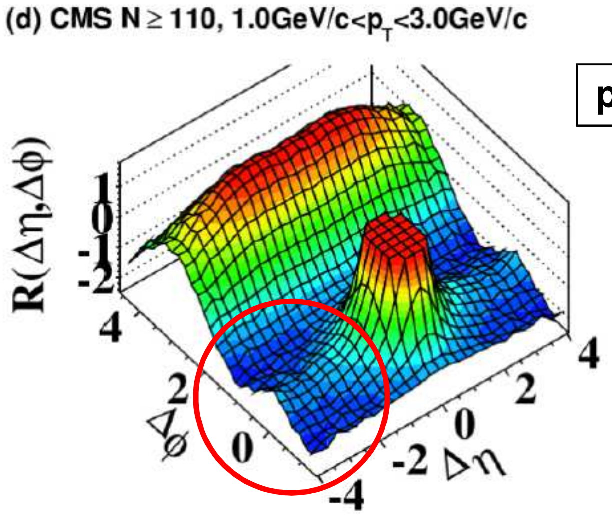
data to be **partially released** by the collaborations with a typical **embargo** time of about **5 years**, with **full release** after **10 years** or at the close of the collaborations (CMS and ALICE: since 2014, LHCb: ~now/soon, ATLAS: from ~2023).

Level 4: Raw data: usually not useful to outsiders, mostly not being released

example candidate for cross-experiment Open Data analysis: “Ridge” in long range particle correlations

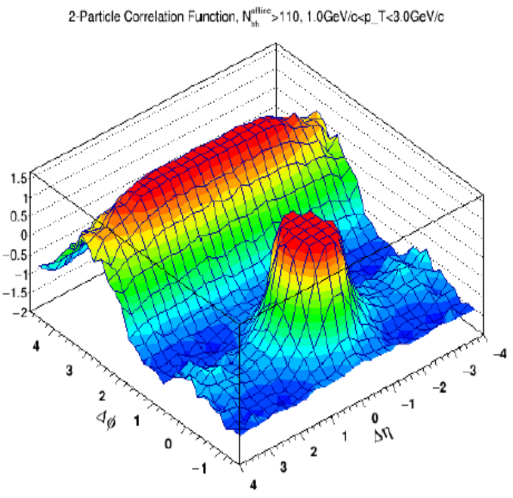
unexpected „Ridge“ observed in 2010 pp data, **JHEP 1009 (2010) 091** **(most-cited non-Higgs LHC result)**
can be ~reproduced on 2010 CMS Open Data

CMS Paper
JHEP 1009 (2010) 091

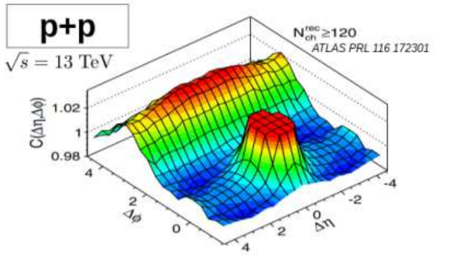


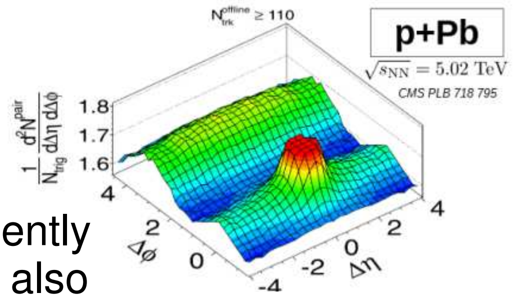
pp

CMS Open Data
(summer student on office desktop)

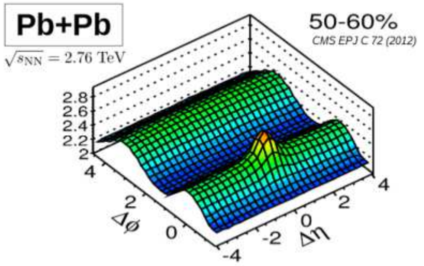


ALICE pp
Open Data
not yet
analyzed



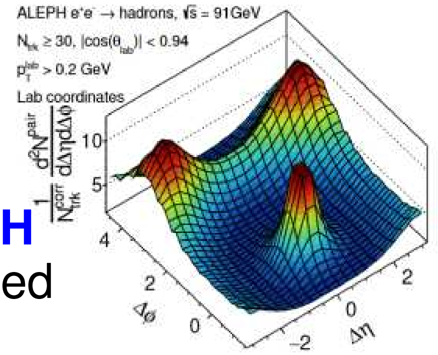


CMS recently
released also
Heavy Ion
Open Data



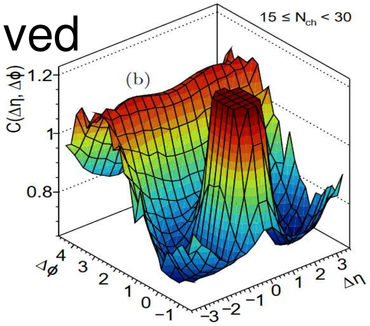
e^+e^-

ALEPH
archived
data



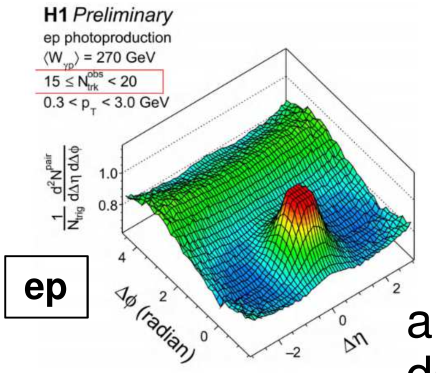
Phys.Rev.Lett. 123 (2019) 212002

ZEUS
archived
data



JHEP 04 (2020) 070

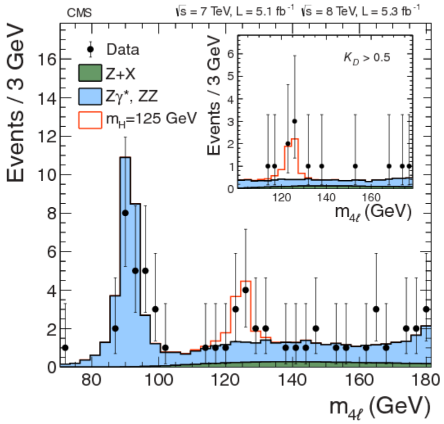
ep



H1
archived
data

Conclusions: Open Data in HEP

- Open Data releases for educational purposes (not topic of today) pursued successfully by essentially all modern HEP experiments since a long time
- Many **research level archived data sets** available from e^+e^- and ep collisions
- **CERN Open Data policy** released in December 2020
- **LHC Open Data releases** for research purposes are **pioneered successfully by CMS**. Presumably the highest complexity data ever made public (2 Pb). Iterative process: started small, expanding fast. Meanwhile, commitment to release all original data within 10 years. All other LHC experiments are in the process of joining in.



- LHC Open Data have **nontrivial but solvable challenges**
- **it works !** so far ~20 scientific results published, more coming (well, cannot compete with astrophysics … yet)
- **overwhelmingly positive feedback** from all sides
- also a lot of great PR

see backup

- **PUNCH4NFDI will hopefully further accelerate this process and link HEP data with those from other communities**

Backup

The challenge: knowledge preservation

HEP doing well with "immediate" metadata, such as

- beam conditions, event and run numbers, provenance information (processing and reconstruction chain, software versions) recorded together with data at time of data set creation

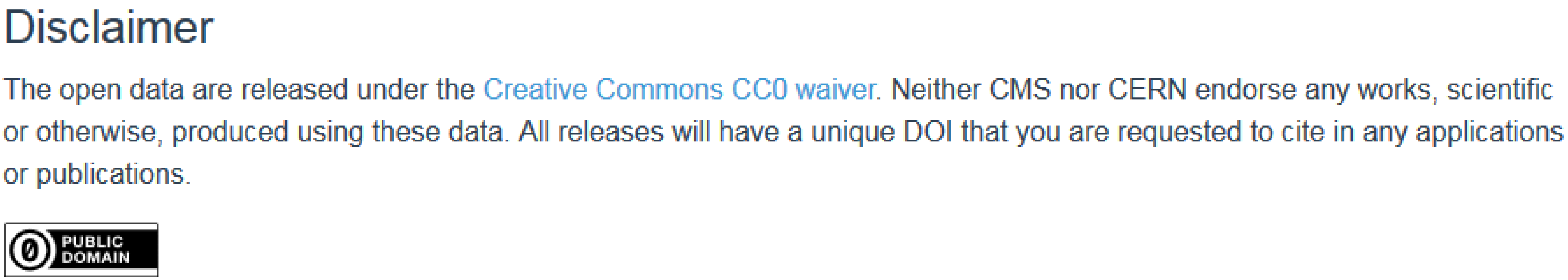
doing poorly with "context" metadata, such as

- how to pick up the right objects in the data and their documentation
- how to know if there are additional selections, corrections, …
- in general, practical information needed to put data in context and analyze them: information readily available and even obvious at time of immediate data analysis, but then easily forgotten
- **Open Data helps/forces us to meet this challenge**

Information must be collected and released together with the data

The implementation: Disclaimer

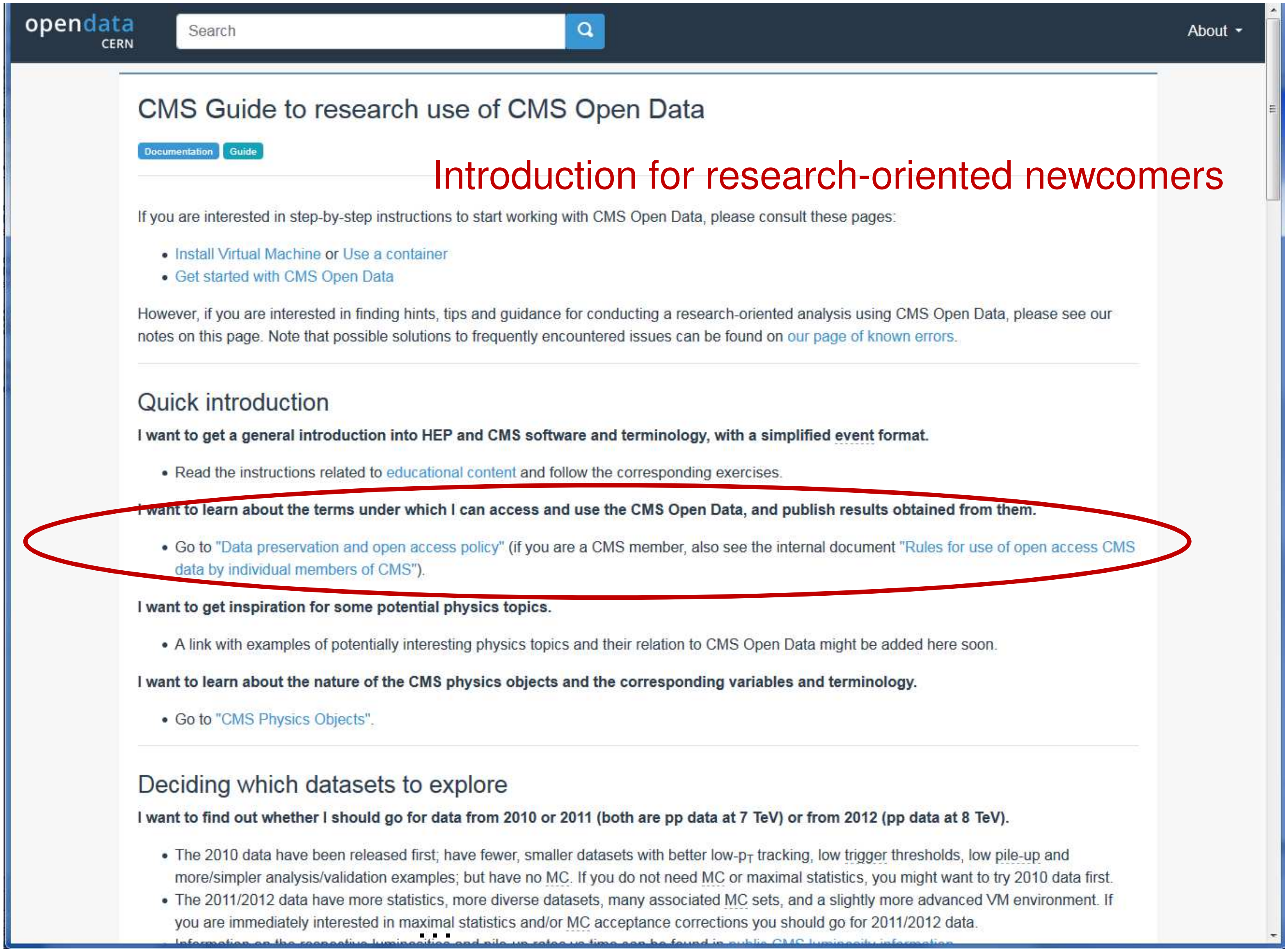
- **On every relevant Open Data portal record (not only CMS)**



Disclaimer

The open data are released under the Creative Commons CC0 waiver. Neither CMS nor CERN endorse any works, scientific or otherwise, produced using these data. All releases will have a unique DOI that you are requested to cite in any applications or publications.

The implementation: Data and knowledge



CMS Guide to research use of CMS Open Data

Introduction for research-oriented newcomers

If you are interested in step-by-step instructions to start working with CMS Open Data, please consult these pages:

- Install Virtual Machine or Use a container
- Get started with CMS Open Data

However, if you are interested in finding hints, tips and guidance for conducting a research-oriented analysis using CMS Open Data, please see our notes on this page. Note that possible solutions to frequently encountered issues can be found on our page of known errors.

Quick introduction

I want to get a general introduction into HEP and CMS software and terminology, with a simplified event format.

- Read the instructions related to educational content and follow the corresponding exercises.

I want to learn about the terms under which I can access and use the CMS Open Data, and publish results obtained from them.

- Go to "Data preservation and open access policy" (if you are a CMS member, also see the internal document "Rules for use of open access CMS data by individual members of CMS").

I want to get inspiration for some potential physics topics.

- A link with examples of potentially interesting physics topics and their relation to CMS Open Data might be added here soon.

I want to learn about the nature of the CMS physics objects and the corresponding variables and terminology.

- Go to "CMS Physics Objects".

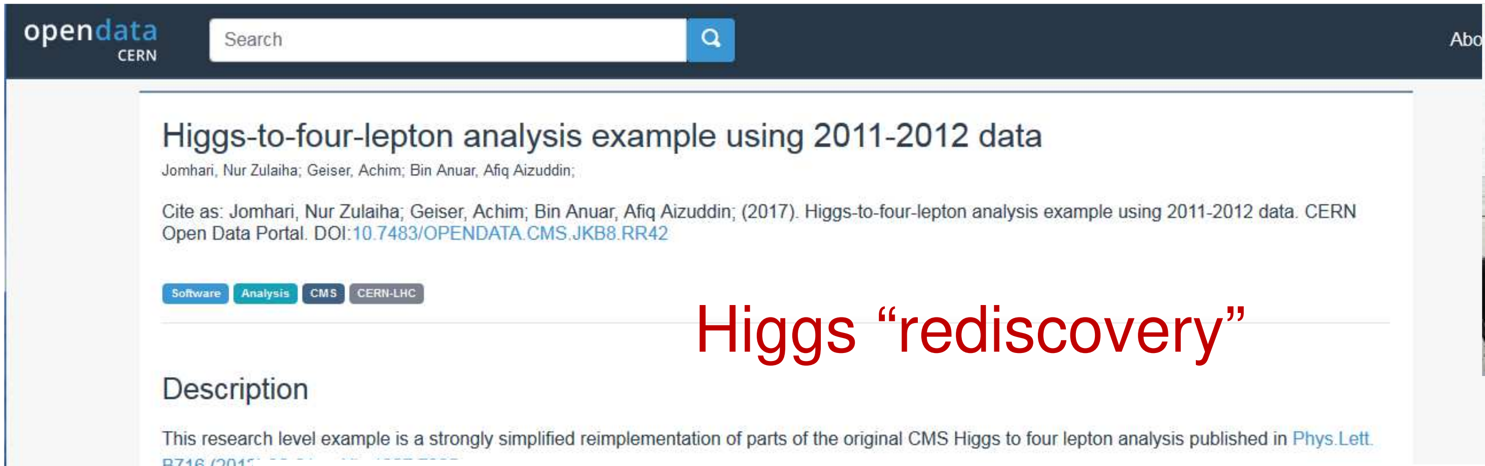
Deciding which datasets to explore

I want to find out whether I should go for data from 2010 or 2011 (both are pp data at 7 TeV) or from 2012 (pp data at 8 TeV).

- The 2010 data have been released first; have fewer, smaller datasets with better low-p_T tracking, low trigger thresholds, low pile-up and more/simpler analysis/validation examples; but have no MC. If you do not need MC or maximal statistics, you might want to try 2010 data first.
- The 2011/2012 data have more statistics, more diverse datasets, many associated MC sets, and a slightly more advanced VM environment. If you are immediately interested in maximal statistics and/or MC acceptance corrections you should go for 2011/2012 data.
- Information on the respective luminosities and pile-up rates vs time can be found in public CMS luminosity information.

...

The implementation: Make data available to school pupils and researchers alike



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 CERN-LHC

Higgs “rediscovery”

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

The publish final states covered by … t/record … y contri

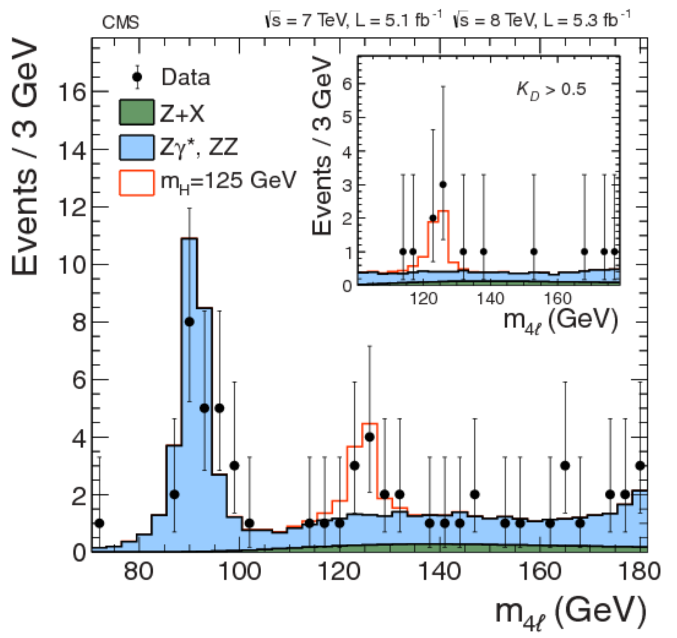
The examp understand lower levels the ROOT … es users … e reache … al acqu

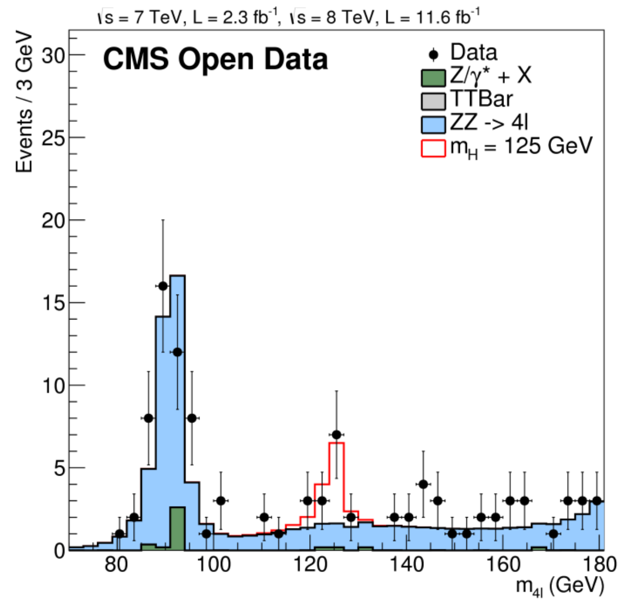
The examp improved c ones in the … differ fr … which … e, exact

Since acco the statistic 30-61, arXi partial stati … e) is on … lowever … t to the

The provide provide the Neverthele files also co Higgs 'exce difference i … iginal cu … advanc … on to th … ample a … s in this … public) CMS data sets with higher statistics the signal is observed in a preliminary analysis with more than 5 standard deviations in this channel alone CMS-PAS-HIG-16-041.

Nur Zulaiha Jomhari demonstrating the Higgs example at DESY

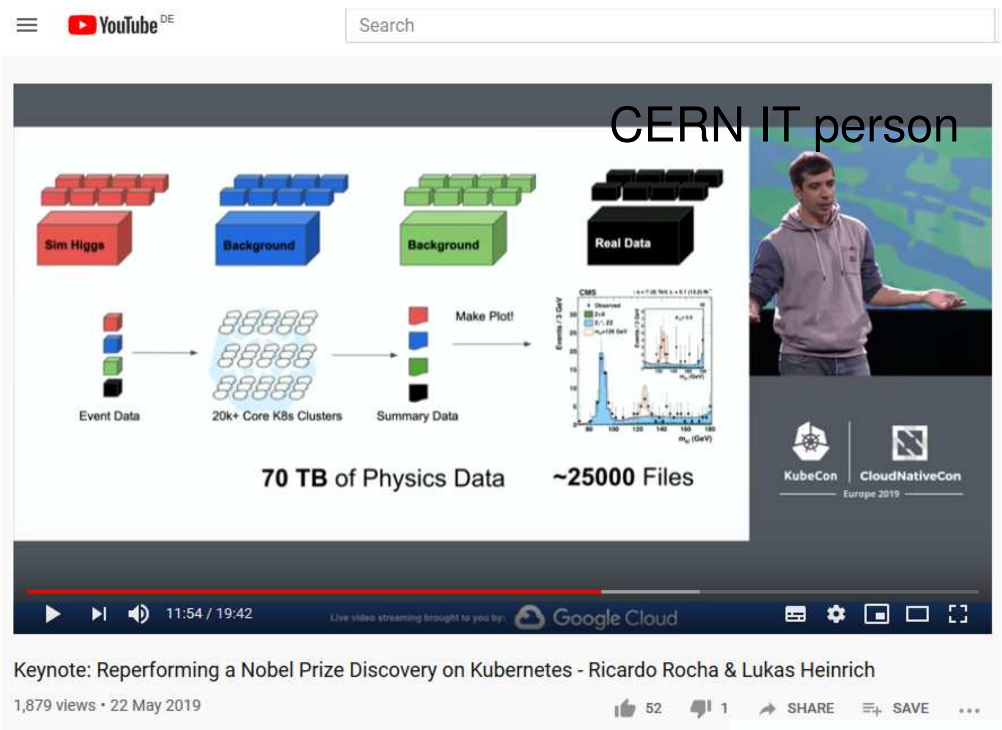




four stages of complexity (seconds, minutes, hours, months)

The implementation: Make data available to school pupils and researchers alike

https://www.youtube.com/watch?v=CTfp2woVEkA

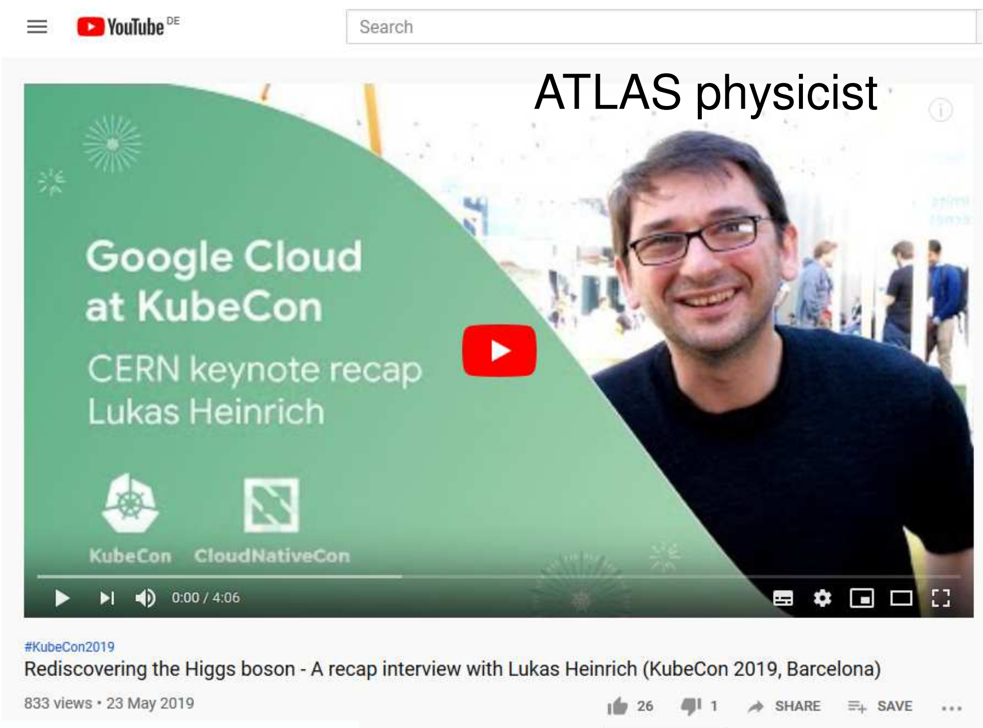


CERN IT person

Keynote: Reperforming a Nobel Prize Discovery on Kubernetes - Ricardo Rocha & Lukas Heinrich

1,879 views • 22 May 2019 · 52 · 1 · SHARE · SAVE

https://www.youtube.com/watch?v=9EczBbF5OLU



ATLAS physicist

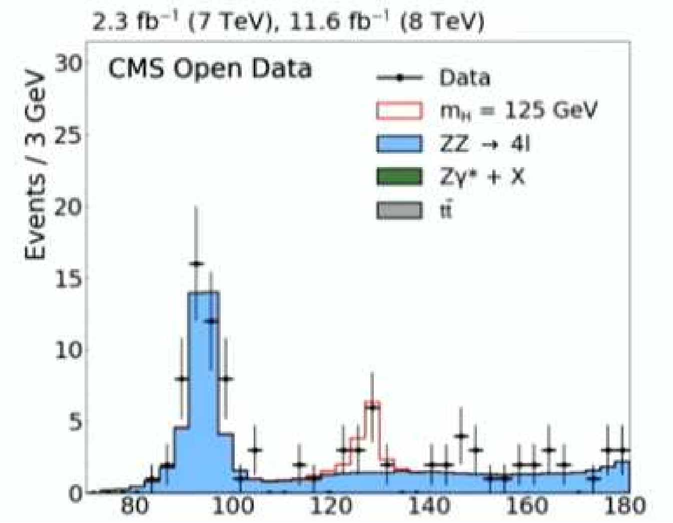
#KubeCon2019

Rediscovering the Higgs boson - A recap interview with Lukas Heinrich (KubeCon 2019, Barcelona)

833 views • 23 May 2019 · 26 · 1 · SHARE · SAVE

level 4:
run on 25000 cores
of Google cloud

+ 3 minutes



Clemens Lange, CMS

Information about CMS Open Data

- **CERN Open Data Portal: http://opendata.cern.ch/about/CMS**
- **CMS (DPHEP) Open Data levels:**
  - **Level 1 – Open access publication and additional numerical data** **INSPIRE**
  - **Level 2 – Simplified data for Outreach and Education** **Open Data - Education**
  - **Level 3 – Reconstructed data and the software to analyze them** **Open Data - Research**
  - **Level 4 – Raw data, and the software to reconstruct and analyze them**

CMS Open Data for Research: **AOD format** (CMS ROOT)

- 1st release of 28 TB of reconstructed 2010 **7 TeV pp collision data** in Nov. 2014
- 2nd release of 130 TB of 2011 **7 TeV pp collision data** and >200 TB of corresponding **MC data** in April 2016

  ~ half the respective full datasets
- 3rd release of 2012 **8 TeV pp data + MC** (~1 PB) in December 2017
- 4th release of remainder of 2010 data+MC + dedicated **special and machine learning datasets** in June 2019

  ↘ including very forward data
- 5th release: 2010 and 2011 **heavy ion** data, dec. 2020

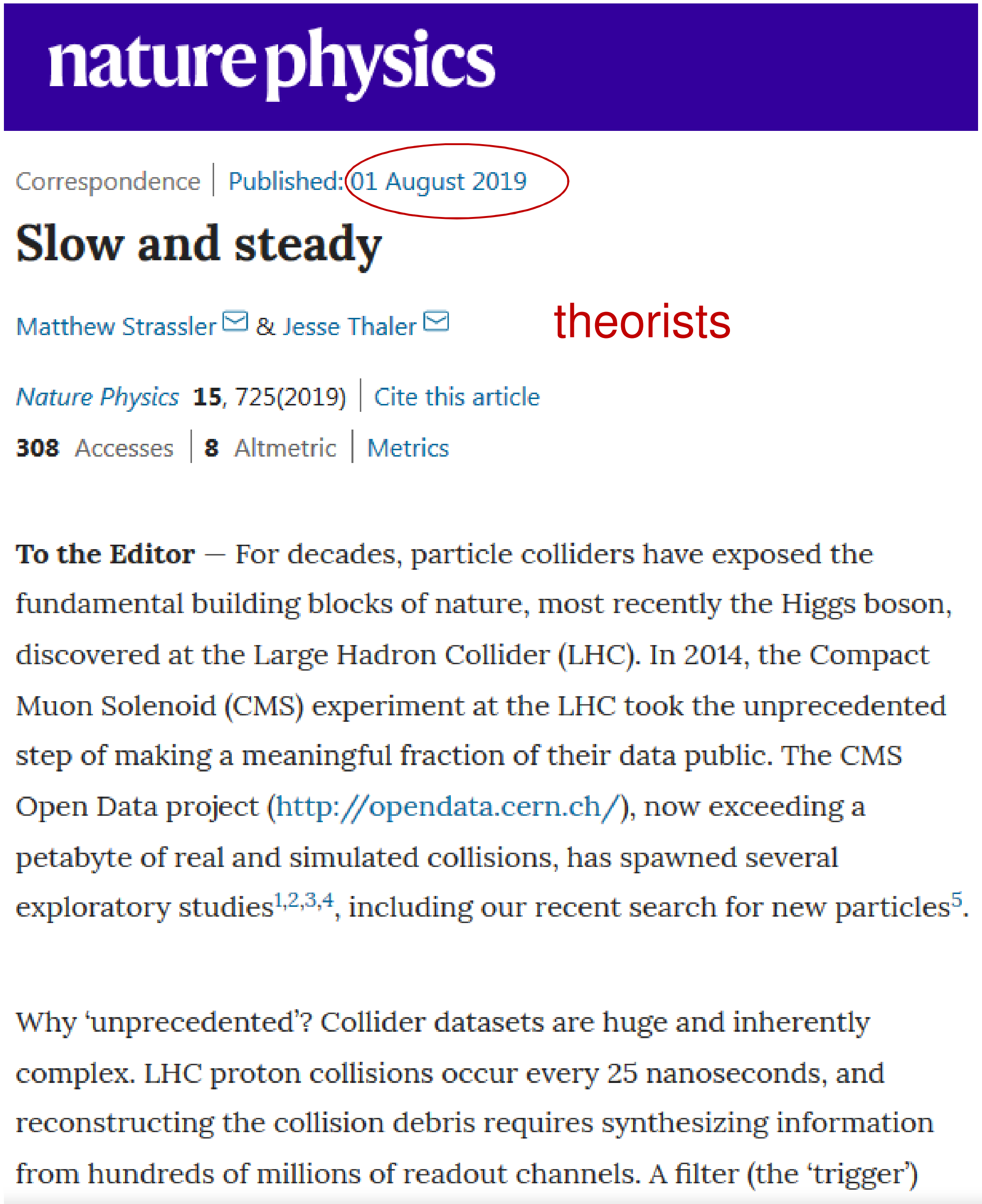
How we = CMS (try to) meet the challenge

- release data exactly as they were last used by CMS members before being released -> fully validated, latest calibrations, no additional action for content
- core CMS software has been public from the very beginning (on github)

  -> no extra action to release the software

  analysis programs for the last step(s) of an analysis to be written by the users themselves (analysis examples provided by individuals)
- part of documentation is public (Google), but not fully sufficient
  -> add special instructions and examples on best effort basis;
  add crucial metadata, + other metadata on best effort basis
- release also MC for signal efficiency corrections and background studies (typically last round of MC used for our own studies, no extra efforts)
- infrastructure for public data storage (currently ~2 PB) and management, as well as Open Data portal infrastructure, provided by CERN
- data can either be accessed at CERN from remote through VMs or containers, (from concerted CERN/experiment effort) or users can download (smaller)datasets

Closing the loop



nature physics

Correspondence | Published: 01 August 2019

Slow and steady

Matthew Strassler & Jesse Thaler

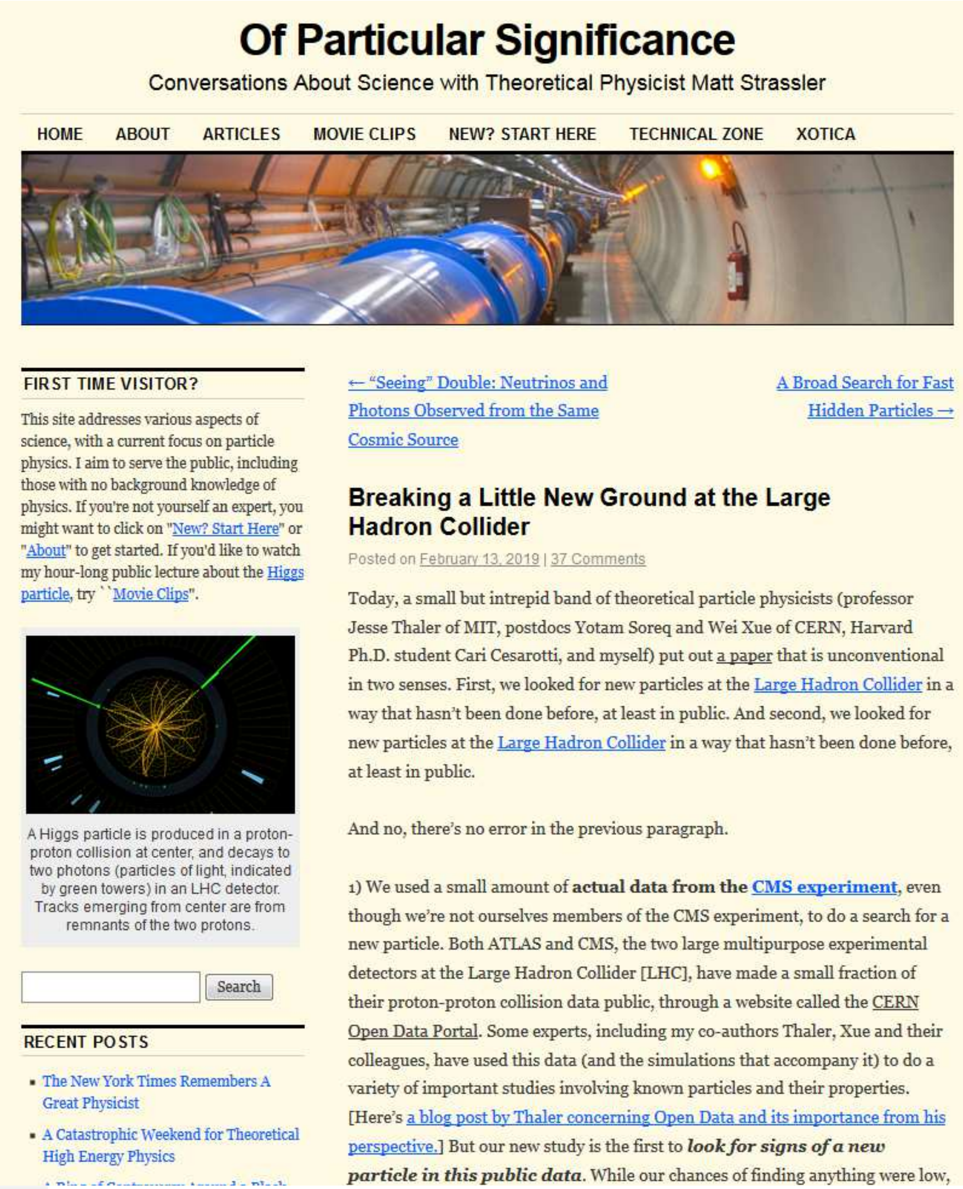
theorists

Nature Physics **15**, 725(2019) | Cite this article

308 Accesses | **8** Altmetric | Metrics

To the Editor — For decades, particle colliders have exposed the fundamental building blocks of nature, most recently the Higgs boson, discovered at the Large Hadron Collider (LHC). In 2014, the Compact Muon Solenoid (CMS) experiment at the LHC took the unprecedented step of making a meaningful fraction of their data public. The CMS Open Data project (http://opendata.cern.ch/), now exceeding a petabyte of real and simulated collisions, has spawned several exploratory studies[1,2,3,4], including our recent search for new particles[5].

Why 'unprecedented'? Collider datasets are huge and inherently complex. LHC proton collisions occur every 25 nanoseconds, and reconstructing the collision debris requires synthesizing information from hundreds of millions of readout channels. A filter (the 'trigger')



Of Particular Significance

Conversations About Science with Theoretical Physicist Matt Strassler

FIRST TIME VISITOR?

This site addresses various aspects of science, with a current focus on particle physics. I aim to serve the public, including those with no background knowledge of physics. If you're not yourself an expert, you might want to click on "New? Start Here" or "About" to get started. If you'd like to watch my hour-long public lecture about the Higgs particle, try ``Movie Clips".

A Higgs particle is produced in a proton-proton collision at center, and decays to two photons (particles of light, indicated by green towers) in an LHC detector. Tracks emerging from center are from remnants of the two protons.

RECENT POSTS

- The New York Times Remembers A Great Physicist
- A Catastrophic Weekend for Theoretical High Energy Physics

← "Seeing" Double: Neutrinos and Photons Observed from the Same Cosmic Source

A Broad Search for Fast Hidden Particles →

Breaking a Little New Ground at the Large Hadron Collider

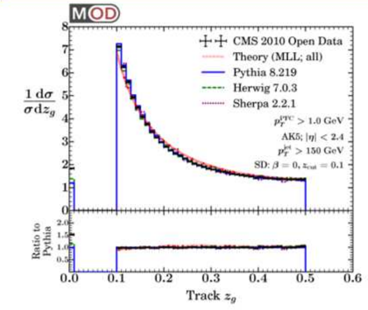
Posted on February 13, 2019 | 37 Comments

Today, a small but intrepid band of theoretical particle physicists (professor Jesse Thaler of MIT, postdocs Yotam Soreq and Wei Xue of CERN, Harvard Ph.D. student Cari Cesarotti, and myself) put out a paper that is unconventional in two senses. First, we looked for new particles at the Large Hadron Collider in a way that hasn't been done before, at least in public. And second, we looked for new particles at the Large Hadron Collider in a way that hasn't been done before, at least in public.

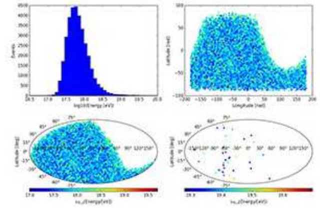
And no, there's no error in the previous paragraph.

1) We used a small amount of **actual data from the CMS experiment**, even though we're not ourselves members of the CMS experiment, to do a search for a new particle. Both ATLAS and CMS, the two large multipurpose experimental detectors at the Large Hadron Collider [LHC], have made a small fraction of their proton-proton collision data public, through a website called the CERN Open Data Portal. Some experts, including my co-authors Thaler, Xue and their colleagues, have used this data (and the simulations that accompany it) to do a variety of important studies involving known particles and their properties. [Here's a blog post by Thaler concerning Open Data and its importance from his perspective.] But our new study is the first to ***look for signs of a new particle in this public data***. While our chances of finding anything were low,

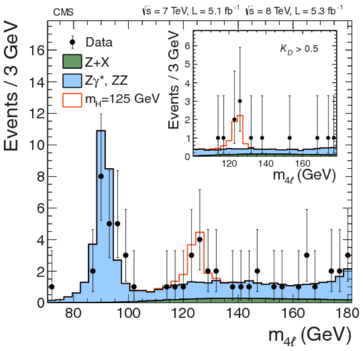
Open Data: research and education

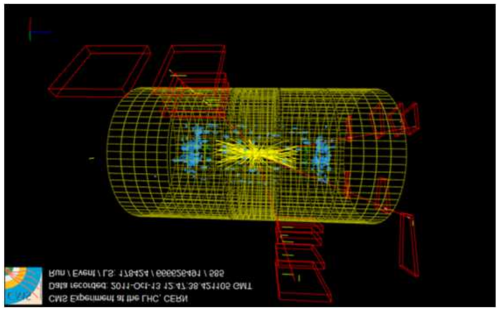


- Use for scientific research papers (in refereed journals)
- Use for machine learning purposes (scientific or procedural)
- Use for preparation of students to join or get acquainted with experiment without need for internal access (e.g. summer students, bachelor projects)



- Use for general education of HEP students (e.g. master classes, at the level of the Auger VISPA project (?))
- Use for school-level teaching, or students/researchers from other fields -> motivate them that physics in general/HEP in particular is great
- Use for PR towards general public and funding agencies -> motivate them that physics in general/HEP in particular is worth while funding





H->4μ event display

What opendata.cern.ch is not:

- not a tool to browse existing published CMS results

  -> use e.g. INSpire, arXiv, …

- not a tool to (re)interpret published results by comparing with theory

  -> use e.g. HEPdata, Rivet, …

- not a toolbox to recast published results into a different form

  -> use recasting tools   (also see separate ReAna effort)

What it is: (for research applications)

- a setup to do whatever a CMS member did, could have done or could still do with the CMS data, without any formal constraint for non-CMS members

- e.g. frequent theorist complaint/request:

  paper X does not present the results in the way I need them for my purposes, recasting is not possible for reason Y,
  could you please change the results? (or the way they are presented)

- alternative solution: stop complaining,
  use Open Data and change them yourself !

  -> (approximately) reproduce the results, or produce new ones
  -> modify whatever you want to modify
  -> compare to your favorite hypothesis

real published examples: see above/below

drawback:

- can only be done on already released datasets (embargo period 3-5 years)

- will probably need a similar effort as if a CMS person or group would have done it (no magic)