PSSOH

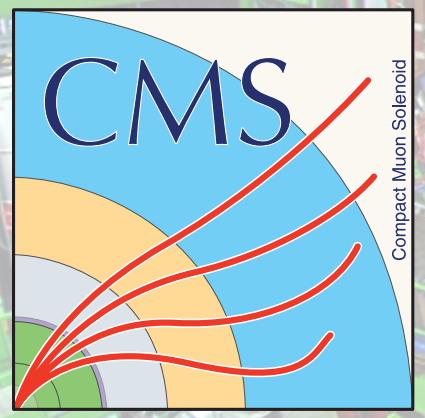
Open Data from CMS at CERN: Status and Plans

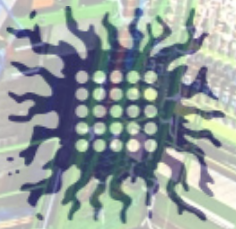
Milos Dordevic

Vinca Institute of Nuclear Sciences, National Institute

of the Republic of Serbia, University of Belgrade

on behalf of the CMS Collaboration





October 15, 2022, School of Electrical Engineering, University of Belgrade

Outline

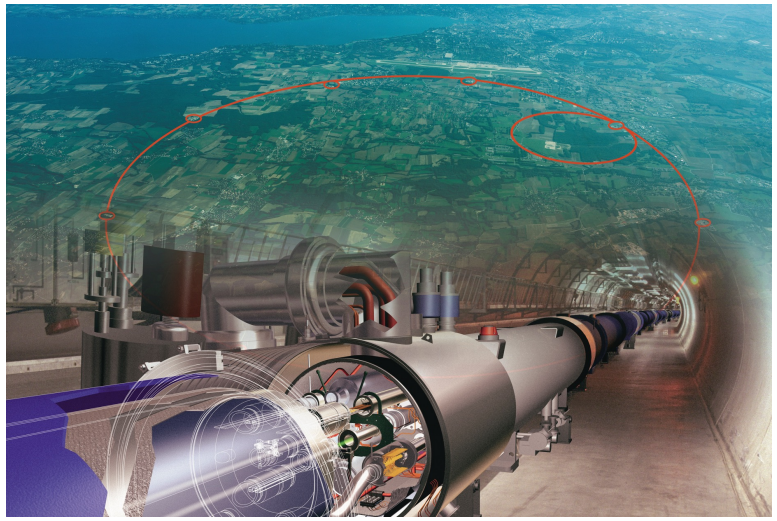
- LHC and CMS experiment at CERN
- **Recording the particle collisions**

https://home.cern/science/accelerators/large-hadron-collider

opendata CERN

- Motivation for releasing CMS Open Data
- Data format, accessibility and an example
- Feedback and experiences from the users

The Large Hadron Collider and CMS Experiment at CERN



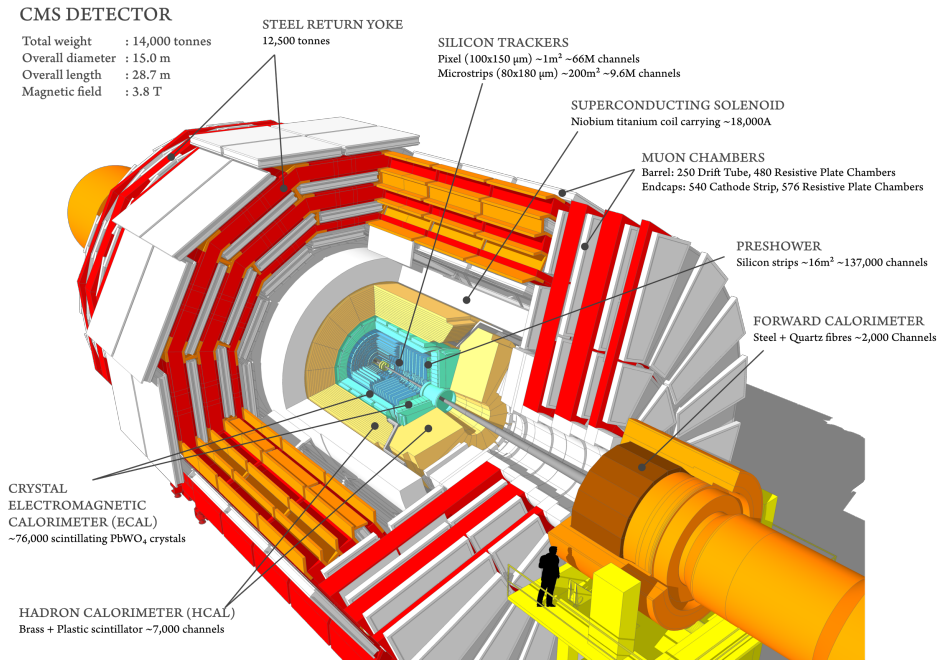
- **Large Hadron Collider** is operating since year of 2010
- 2022: colliding protons at record energies of 13.6 TeV
- Four major experiments: ATLAS, CMS, ALICE and LHCb

- **Compact Muon Solenoid (CMS)** experiment

  https://home.cern/science/experiments/cms

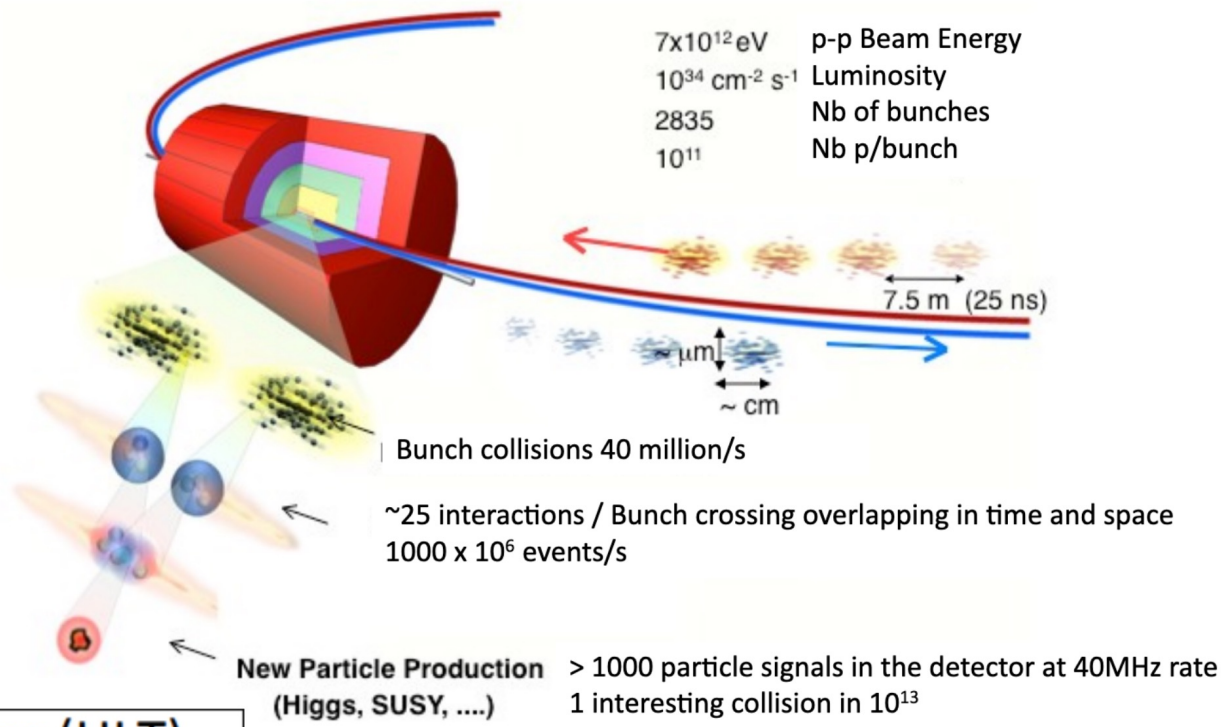
- Multi-layered, multipurpose experiment for

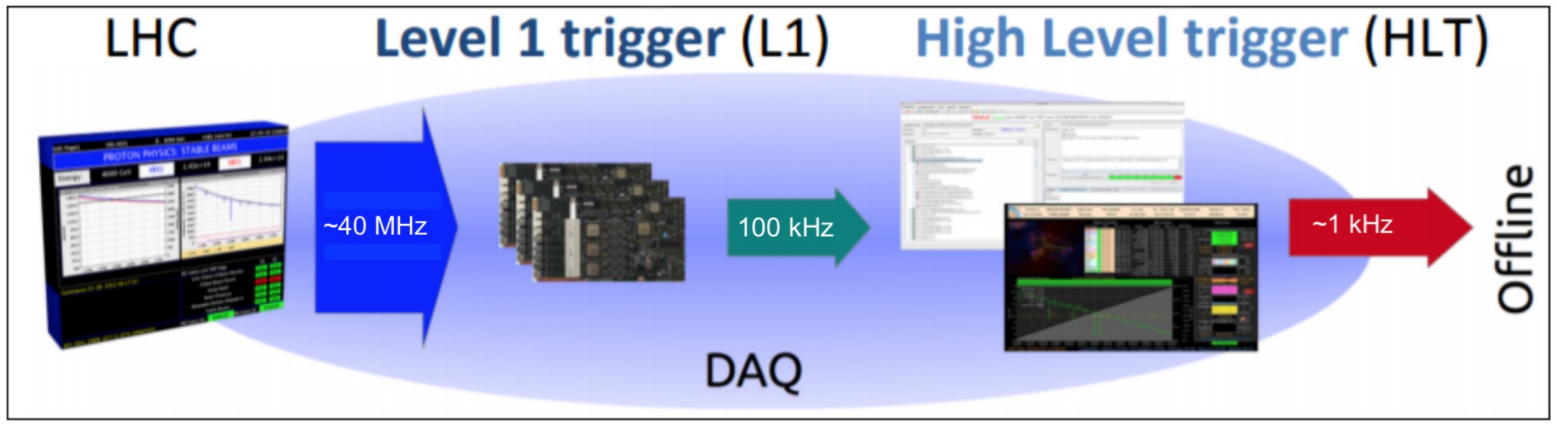
  Higgs studies, precision SM & BSM searches



Recording the particle collisions at CMS experiment

- **Cannot take all data** (storage and processing)
  - may generate 50 terabytes at 40MHz rate
- Selective read out of data by **Trigger system**
- Two-tiered Trigger system: **Level-1** based on fast (custom) electronics (**40 MHz->100 kHz**)





CMS Collaboration, *The CMS trigger system*, JINST 12 (2017) no.01, P01020

- **High Level Trigger** is streamlined offline-> runs on computing farm
- **100 kHz to 1.5 kHz** with full event content & prompt reconstruction

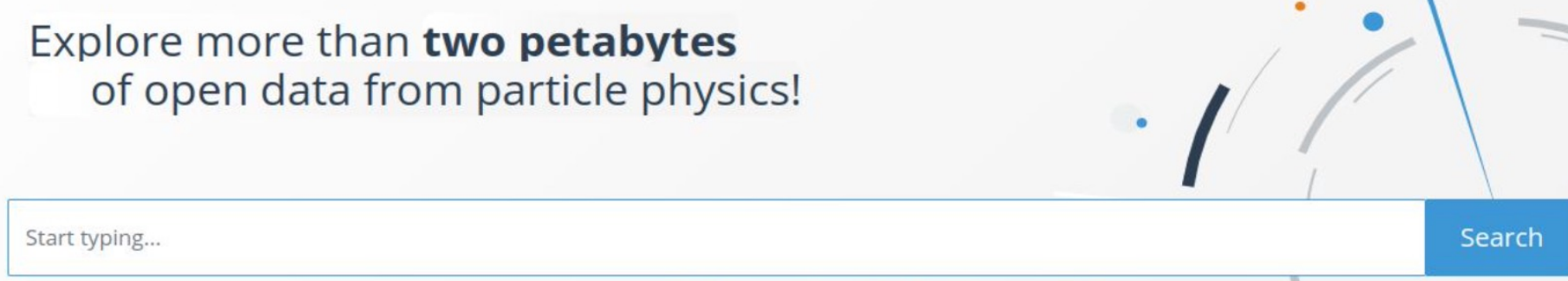
Motivation for releasing the CMS data to the public

- **Inclusiveness:**
  - science should be inclusive and knowledge open to everyone
- **Engagement:**
  - availability of data is key to engage people with research
- **Education:**
  - provide teaching and attract students to particle physics

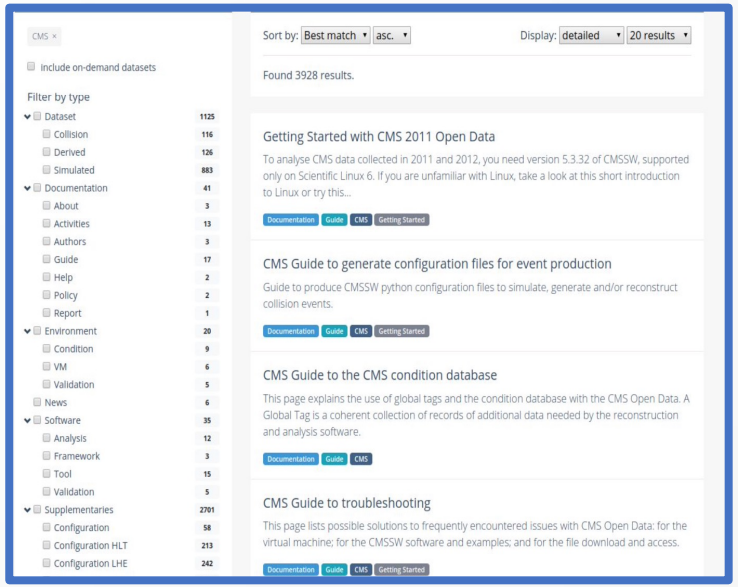
- **Societal impact:**
  - another way to return something back to the society
- **Scientific research:**
  - improving the exchange of knowledge with the non-collaboration members in the same or different field

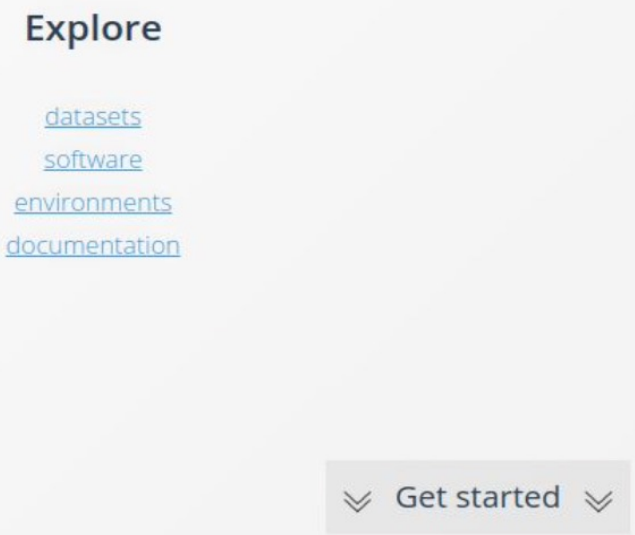
Where to start from with using the CMS Open Data?

http://opendata.cern.ch/



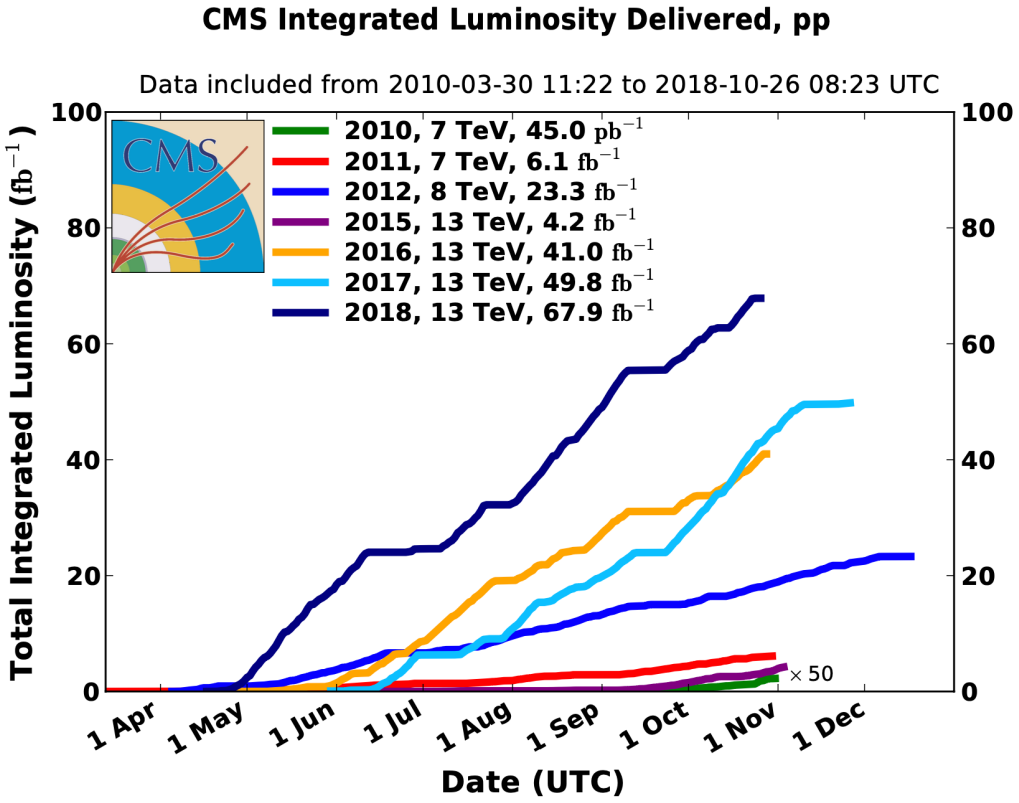
How to browse the CMS Open Data:

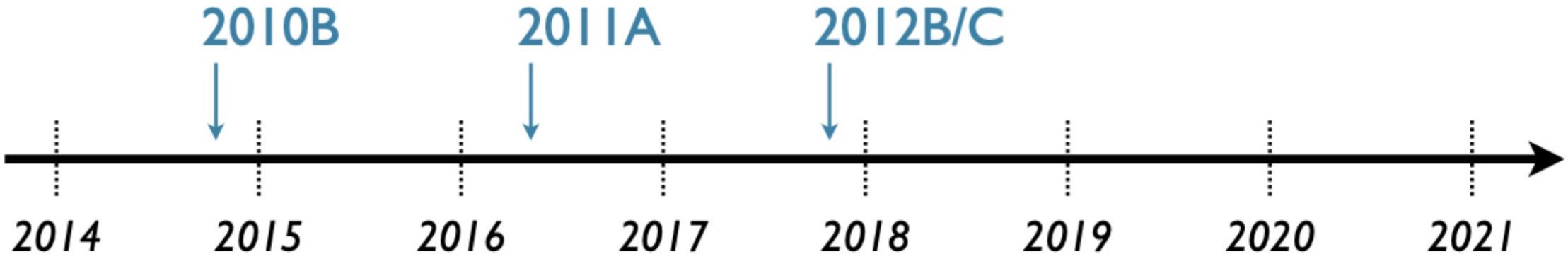




Basic information about releasing the CMS Open Data

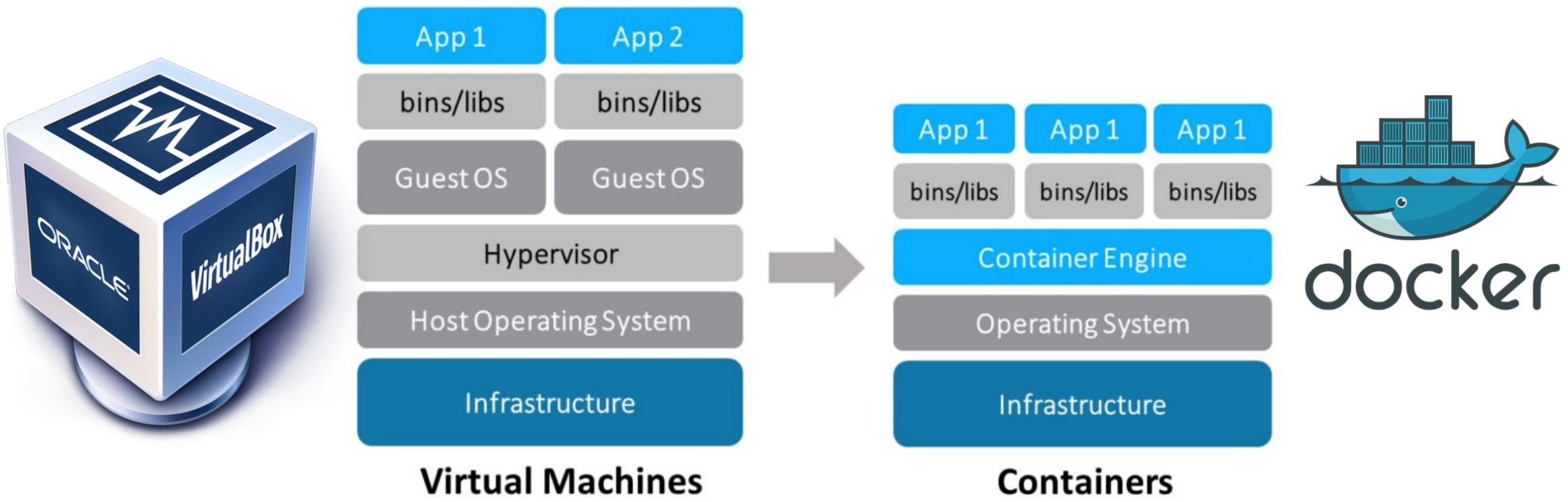
- CMS data preservation, re-use and open access policy
  DOI:10.7483/OPENDATA.CMS.7347.JDWH
- Publish 50(100)% of the collision data after 3(10) years
- Data with open license (Creative Commons CC0 waiver)
- Currently available:
  - 2010: 32 1/pb p-p collision data at 7 TeV
  - 2011: 2.3 1/fb p-p collision data at 7 TeV + MC
  - 2012: 11.6 1/fb p-p collision data at 8 TeV + MC





How to access CMS Open Data: Virtual Machines & Docker

- CERN Virtual Machine to access CMS data, but also Docker containers (allows to preserve full analysis)



http://opendata.cern.ch/docs/cms-virtual-machine-2011

http://opendata.cern.ch/docs/cms-guide-docker

Data format of CMS Open Data

- Most of the CMS Open Data is published in the Analysis Object Data (AOD) format
  - serialized C++ objects requiring specific (CMSSW) environment and also ROOT
  - each of the events holds about 500 kB of information, resulting in large files

- Reduced information content formats:
  - miniAOD: reduced version of the AOD stores serialized C++ objects
  - nanoAOD: storage of basic types (floats, integers, arrays), 2kB/evt

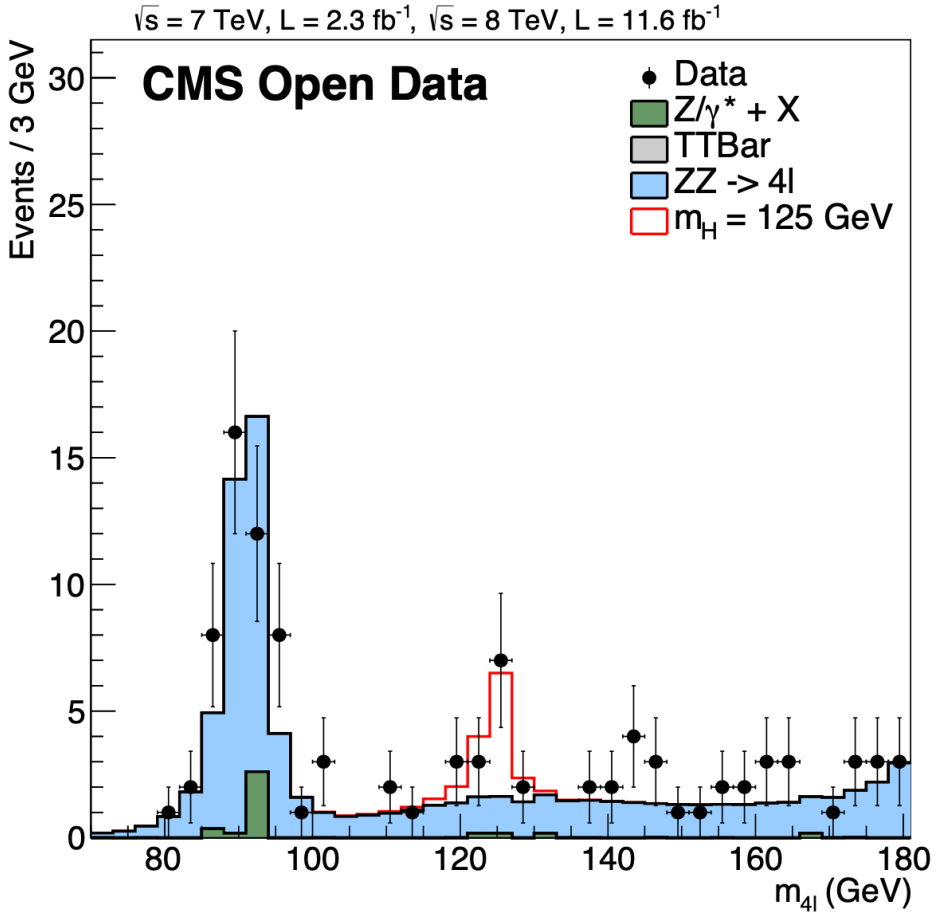
Variable	Type	Description
nMuon	unsigned int	Number of muons in this event
Muon_pt	float[nMuon]	Transverse momentum of the muons
Muon_eta	float[nMuon]	Pseudorapidity of the muons
Muon_phi	float[nMuon]	Azimuth of the muons
Muon_mass	float[nMuon]	Mass of the muons
Muons_charge	int[nMuon]	Charge of the muons (either 1 or -1)

- NanoAOD readable independent from CMSSW, with any library reading ROOT files

- Tool for converting AOD to nanoAOD developed, to ease access of CMS open data

“Re-descovering” the Higgs boson with CMS Open Data

- An example of the Higgs to four leptons analysis provided in http://opendata.cern.ch/record/5500
- Different levels of computational complexity available
  - from reproducing the plot from pre-processed files
  - to processing ~80 TB of CMS AOD files in CMSSW
- Possibility to perform full-fledged CMS physics analysis



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

Simplified reimplementation of the original CMS H -> 4 lepton analysis

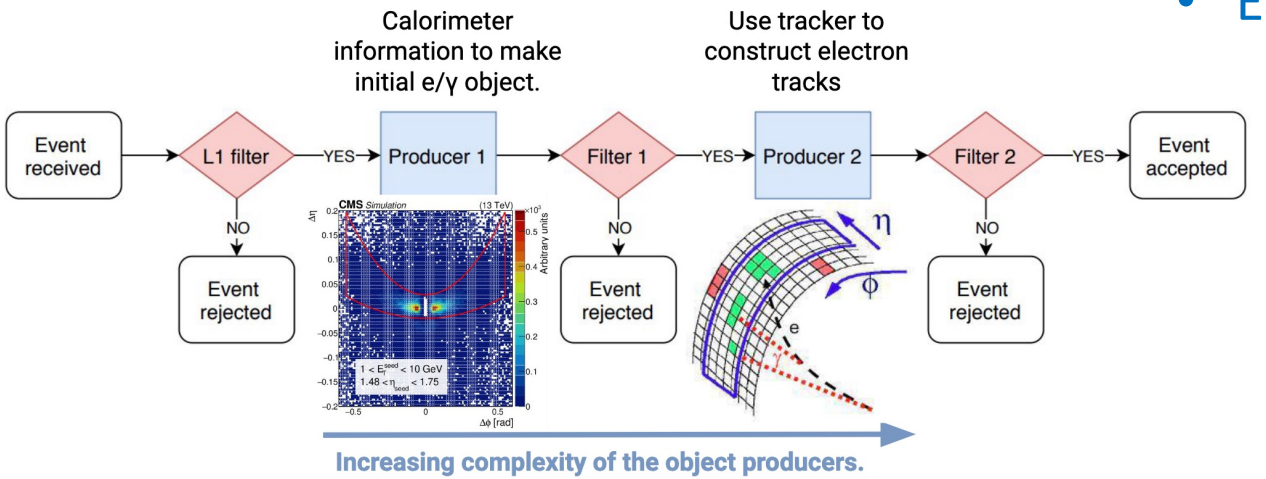
- Full Higgs to four-leptons analysis: EPJC 81 (2021) 488 (w/ Run2 data)

CMS Trigger Analysis with CMS Open Data: Introduction

- DAQ & HLT systems create RAW data events containing:
  - the detector data
  - the Level-1 and HLT trigger results (trigger bits)
  - higher-level objects made during HLT processing

Guide to the CMS Trigger System

http://opendata.cern.ch/docs/cms-guide-trigger-system



- A path consists of several steps (software **modules**)
- Each module performing a well-defined task such as
  - performing the unpacking (**raw->digi conversion**)
  - **reconstruction of physics objects** (e, μ, jet, MET)
  - making some **intermediate triggering decisions**
  - calculating the **final decision** for HLT trigger path

Analysis code for trigger information from CMS 2011 data

- Example of the C++/Python code to extract the trigger information from CMS Open/Legacy data http://opendata.cern.ch/record/5004, source code: https://github.com/cms-opendata-analyses/TriggerInfoTool/tree/2011
- **GeneralInfoAnalyzer:**
  - C++ snippets on how to access trigger information such as metadata, prescales, module info etc.
- **ModuleInTriggerAnalyzer:**
  - how to dump all the modules for a specific trigger and obtain last active trigger module (filter)
- **TriggerMatchingAnalyzer**:
  - how to match a reconstructed tracks to objects that fired a trigger containing a specific module
- **TriggerSimplePrescalesAnalyzer**:
  - check the trigger L1 and HLT prescales, and whether the trigger has accepted the event or not

```
cmsrel CMSSW_5_3_32
cd CMSSW_5_3_32/src/
cmsenv
git clone -b 2011 git://github.com/cms-legacydata-analyses/TriggerInfoTool.git cd TriggerInfoTool
cd {packagename}
scram b
ln -s python/{configname} .
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA FT_53_LV5_AN1
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db FT_53_LV5_AN1_RUNA.db
ls -l
ls -l /cvmfs/
cmsRun {configname} > full.log 2>&1 &      (checking w/ “tail -f full.log”)
```

ModuleInTriggerAnalyzer: Output

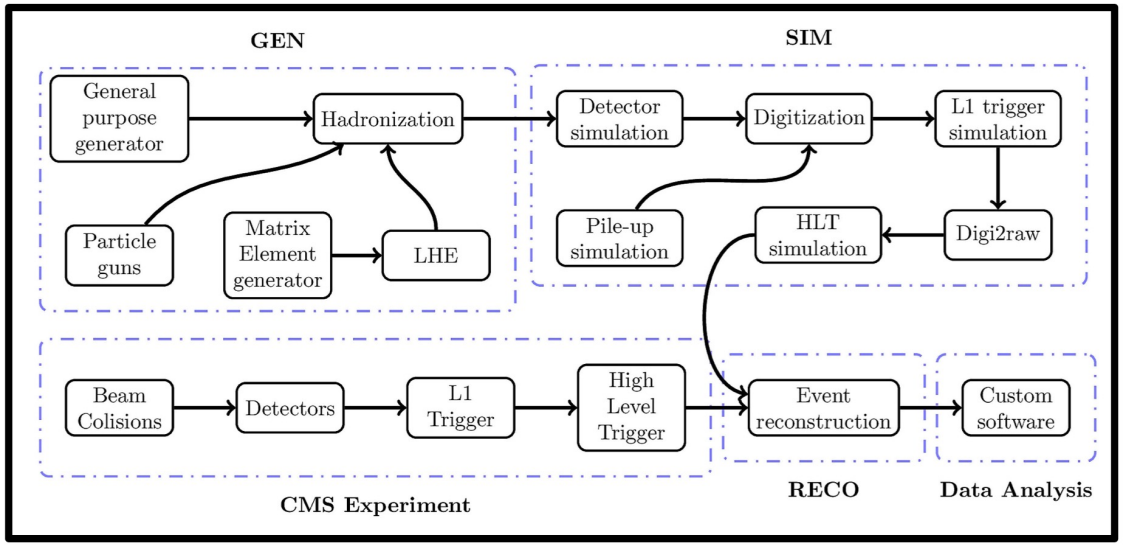
The modules in trigger HLT_Jet190_v6 are:
hltTriggerType
hltGtDigis
hltGctDigis
hltL1GtObjectMap
hltL1extraParticles
hltScalersRawToDigi
hltOnlineBeamSpot
hltOfflineBeamSpot
hltL1sL1SingleJet92
hltPreJet190
hltEcalRawToRecHitFacility
hltEcalRegionalJetsFEDs
hltEcalRegionalJetsRecHit
hltHcalDigis
hltHbhereco
hltHfreco
hltHoreco
hltTowerMakerForJets
hltAntiKT5CaloJetsRegional
hltCaloJetL1MatchedRegional
hltCaloJetIDPassedRegional
hltCaloJetCorrectedRegional
hltSingleJet190Regional
hltBoolEnd

(partial) content of the full.log of ModuleInTriggerAnalyzer:

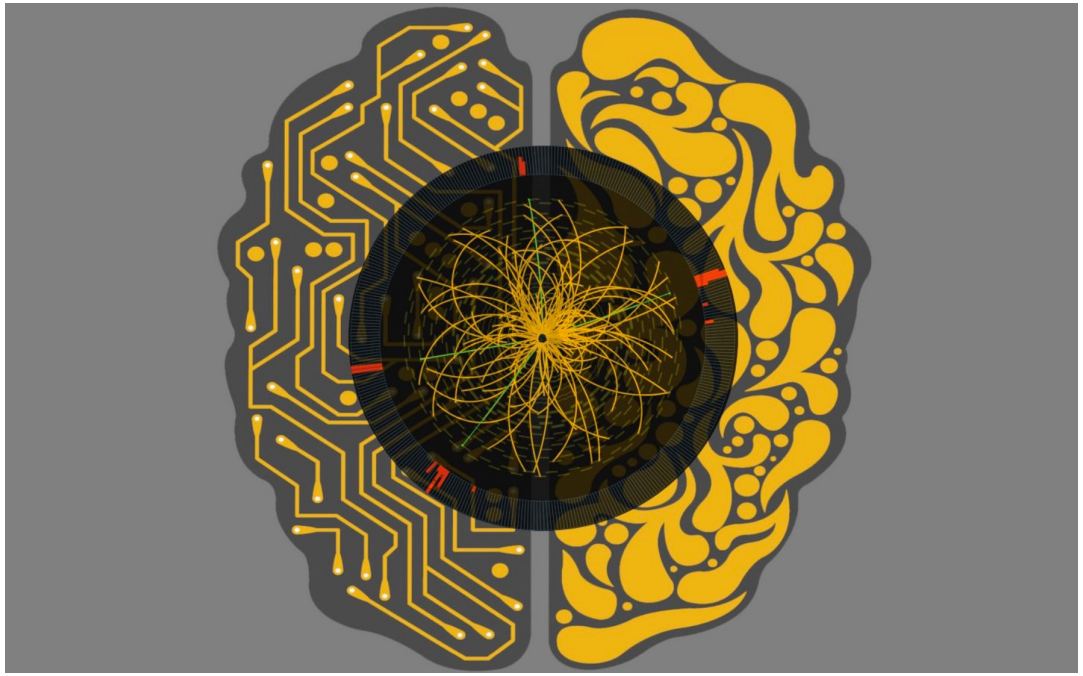
…
Begin processing the 41st record. Run 171897, Event 489806429, LumiSection 452 at 19-Sep-2022 06:52:41.279 CEST
Currently analyzing trigger HLT_Jet190_v6
Last active module - label/type: hltPreJet190/HLTPrescaler [9 out of 0-23 on this path]
Begin processing the 42nd record. Run 171897, Event 489992533, LumiSection 452 at 19-Sep-2022 06:52:41.279 CEST
Currently analyzing trigger HLT_Jet190_v6
Last active module - label/type: hltL1sL1SingleJet92/HLTLevel1GTSeed [8 out of 0-23 on this path]
Begin processing the 43rd record. Run 171897, Event 489970773, LumiSection 452 at 19-Sep-2022 06:52:41.280 CEST
Currently analyzing trigger HLT_Jet190_v6
Last active module - label/type: hltPreJet190/HLTPrescaler [9 out of 0-23 on this path]
Begin processing the 44th record. Run 171897, Event 488919432, LumiSection 452 at 19-Sep-2022 06:52:41.280 CEST
Currently analyzing trigger HLT_Jet190_v6
Last active module - label/type: hltSingleJet190Regional/HLT1CaloJet [22 out of 0-23 on this path]…

CMS Open Data for Machine Learning

- The open data address application of Machine Learning to challenges in high-energy physics
- Reconstructed data and simulations from the CASTOR calorimeter (in a very forward region)



CMS Collaboration, *The very forward CASTOR calorimeter of the CMS experiment*, JINST 16 (2021) P02010



https://home.cern/news/news/knowledge-sharing/cms-releases-open-data-machine-learning

- Instructions and examples provided for you on how to generate your own events

Jet Substructure Studies with CMS Open Data

- Jesse Thaler (MIT theorist): two papers with CMS Open Data

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski, Simone Marzani, Jesse Thaler, Aashish Tripathee, Wei Xue

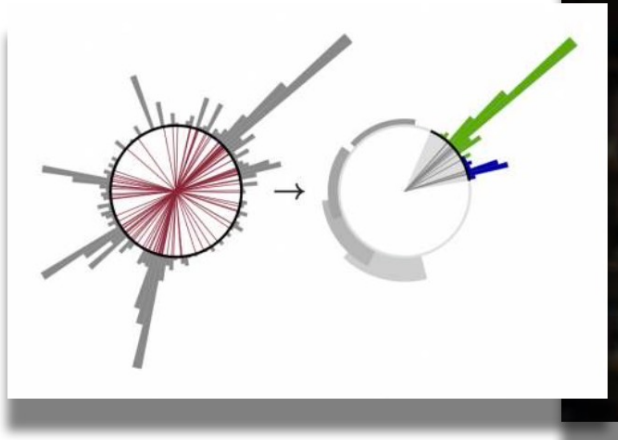
(Submitted on 17 Apr 2017 (v1), last revised 25 Sep 2017 (this version, v3))

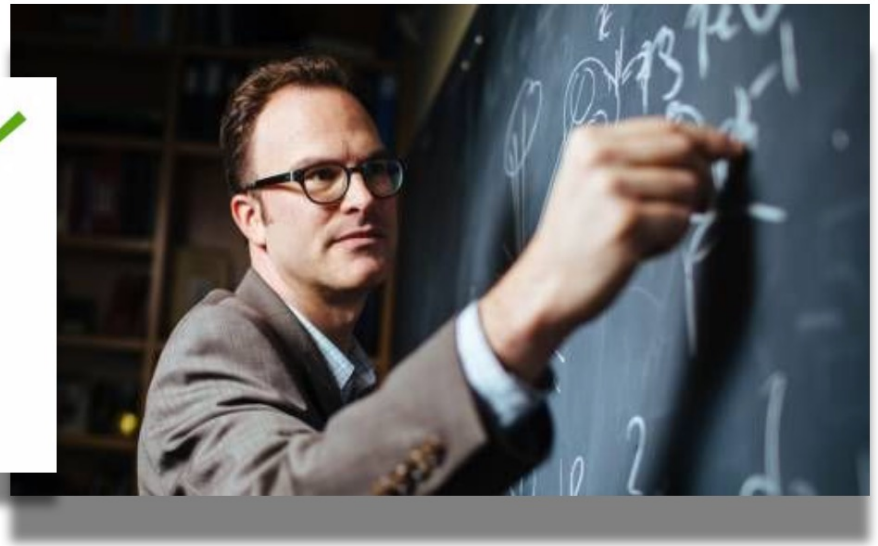
10.1103/PhysRevLett.119.132003

Jet substructure studies with CMS open data

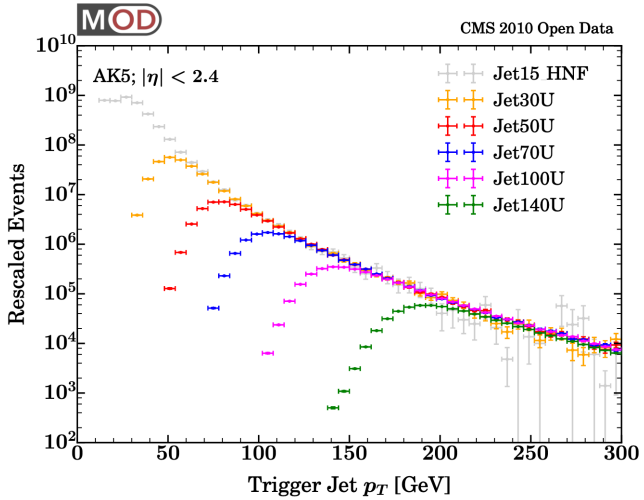
Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, and Jesse Thaler
Phys. Rev. D **96**, 074003 – Published 3 October 2017

https://doi.org/10.1103/PhysRevD.96.074003





- Interesting lessons from the paper:



- *"We then converted AOD files into a text-based MIT Open Data (MOD) format to facilitate the use of external analysis tools."*
- *"From the physics perspective, our experience with the CMS Open Data was fantastic"*
- *"From a technical perspective, though, we have encountered a number of challenges"*

An opinion from Nature Physics

Slow and steady

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[1–4], 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') discards all but the most interesting collisions, and accounting for its effects and those of the heterogeneous LHC detectors is challenging. The resources required to make such a complex dataset public and usable are substantial, but in short supply.

However, data from the LHC — whose successor is decades away — are priceless for future scientists and must be carefully archived, along with all necessary associated knowledge. As it is archived, the data should be made public, though not immediately. A delay of several years, enough for the experimenters who collected the data to perform thorough analyses, is appropriate; only those who spent years building the experiments have earned quick access. Furthermore, making LHC data ready for public use, with documentation and example code, requires significant funding and time.

But steady publication of LHC data has multiple benefits. First, it encourages prompt archiving, before collective memory fades and knowledge is lost. Second, other scientists can analyse the data while the LHC is still running, testing unconventional strategies and potentially leading to unexpected discoveries, new approaches and fruitful discussions. And third, as a by-product, these scientists can stress test the archiving methods; any deficiencies found are easier to fix now than later.

In this way, public collider data can complement the overall LHC research effort. We, therefore, favour a slow but steady approach to full publication of the LHC experiments' data; it is in the best interest of particle physics. ❐

Matthew Strassler[1] and Jesse Thaler[2]
[1]*Department of Physics, Harvard University, Cambridge, MA, USA.* [2]*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA, USA.*
e-mail: strassler@physics.harvard.edu; jthaler@mit.edu

Published online: 1 August 2019
https://doi.org/10.1038/s41567-019-0628-z

References

1. Larkoski, A., Marzani, S., Thaler, J., Tripathee, A. & Xue, W. *Phys. Rev. Lett.* **119**, 132003 (2017).
2. Madrazo, C. F., Cacha, I. H., Iglesias, L. L. & de Lucas, J. M. Preprint at https://arxiv.org/abs/1708.07034 (2017).
3. Andrews, M., Paulini, M., Gleyzer, S. & Poczos, B. Preprint at https://arxiv.org/abs/1807.11916 (2018).
4. Lester, C. G. & Schott, M. Preprint at https://arxiv.org/abs/1904.11195 (2019).
5. Cesarotti, C., Soreq, Y., Strassler, M. J., Thaler, J. & Xue, W. *Phys. Rev. D* **100**, 015021 (2019).

Nature Physics opinion

"only those who spent years building the experiment have earned quick access"

"other scientists can analyse the data while LHC is still running, testing unconventional strategies"

"public data can complement the overall research effort"

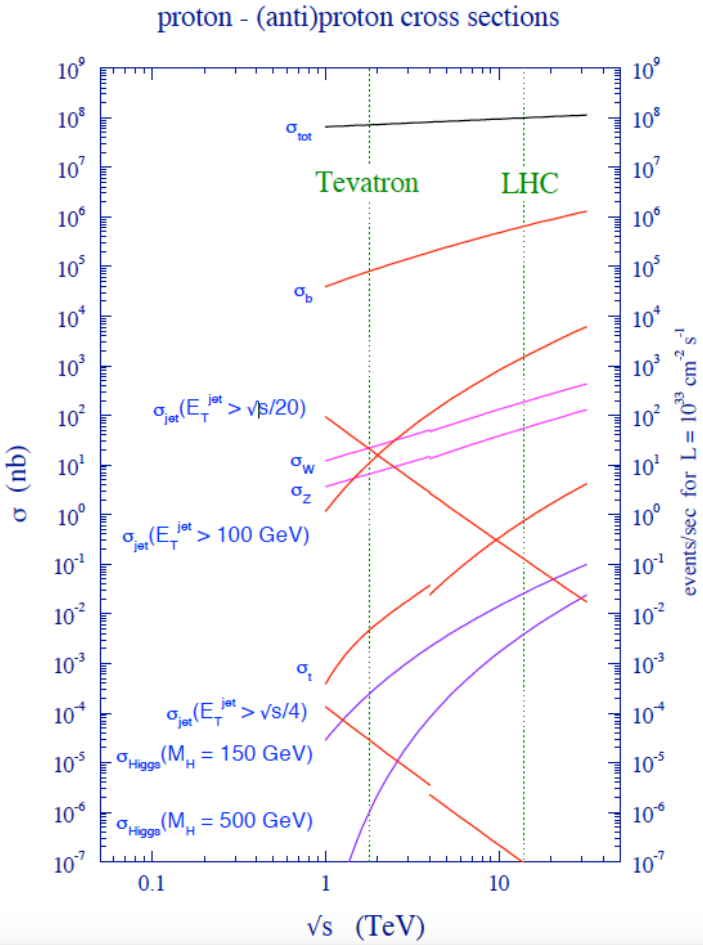
Summary

- CMS makes **a very strong Open Data effort** within the LHC
- We are trying to facilitate the usage of the CMS Open Data
  - improved documentation and software tools + containers
  - working towards simplified and easy to use data formats
- Release of the new data is becoming imminent
- **Please let us know if you have any feedback!**

BACKUP SLIDES

The CMS Detector and Rates of Physics Processes

- CMS is general purpose detector at the CERN LHC
- Sub - detectors to identify particles & Particle Flow
- Real time decision to store interesting events (Trigger)



Lumi: 2 x 10^{34} cm^2 s^{-1} in the Run2

2556 bunches, 2.5x10^{11} p/bunch

Total collision rate around 2 GHz

b-quark production rate 10 MHz

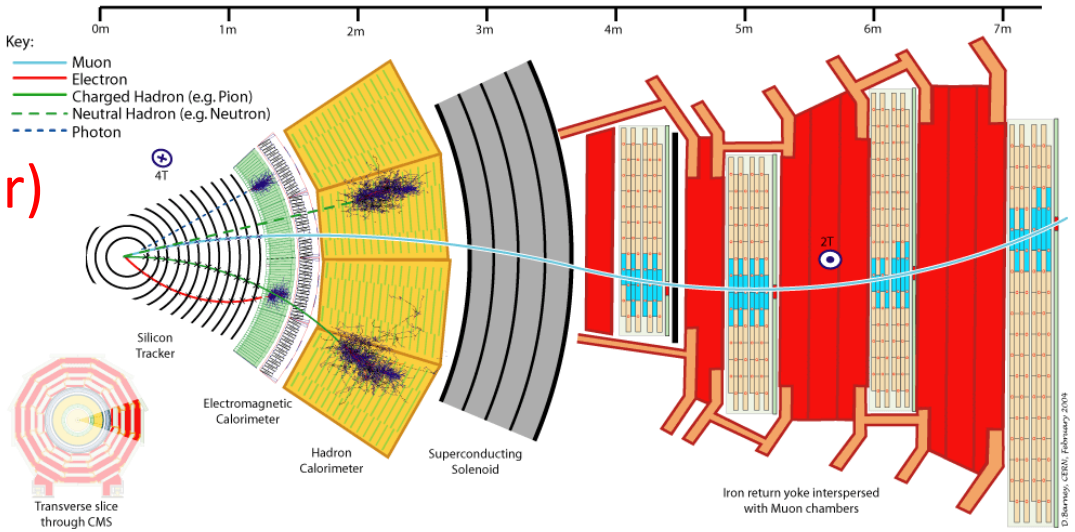
W boson production rate 4 kHz

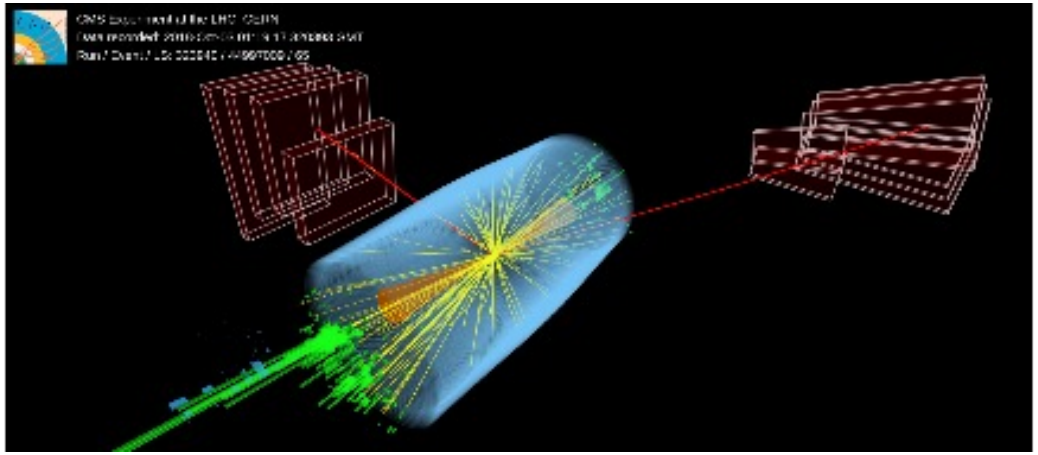
Top quark production rate 20 Hz

Higgs boson prod. rate only 1 Hz

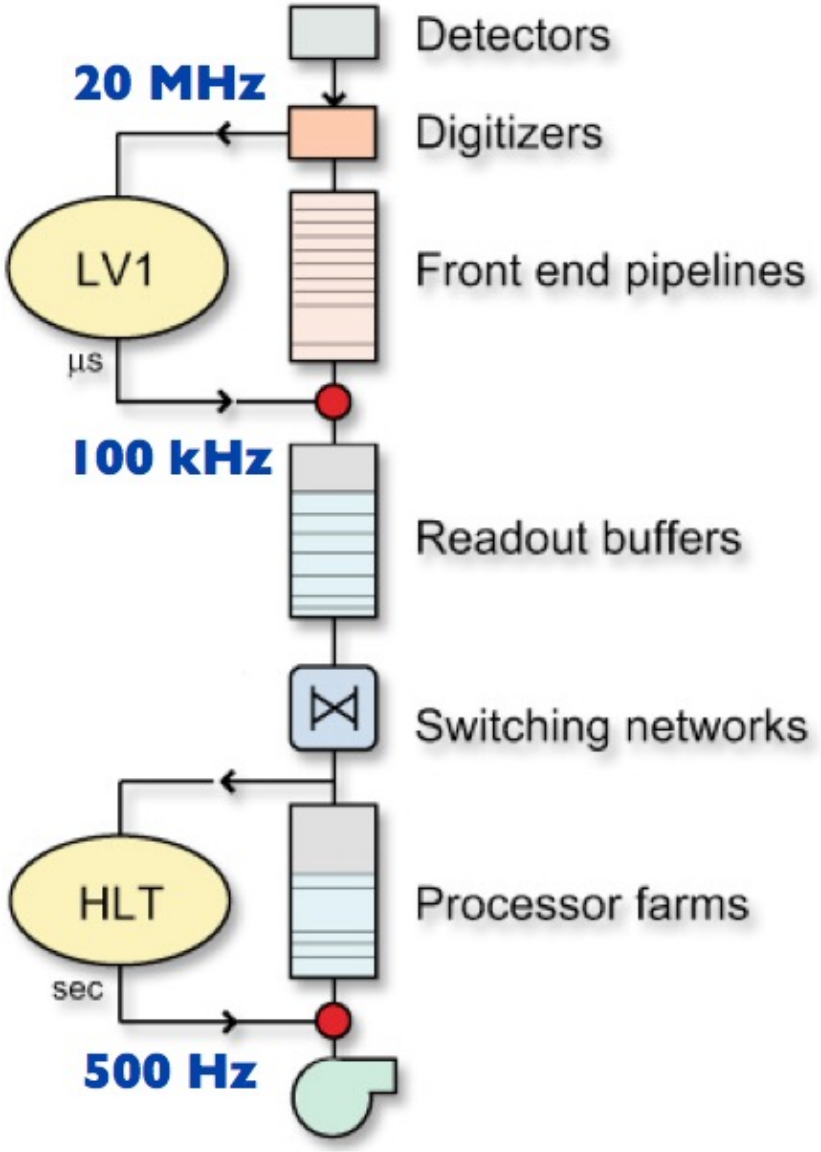
SUSY rate(m@TeV) below 0.1 Hz

Interesting events at low rates!



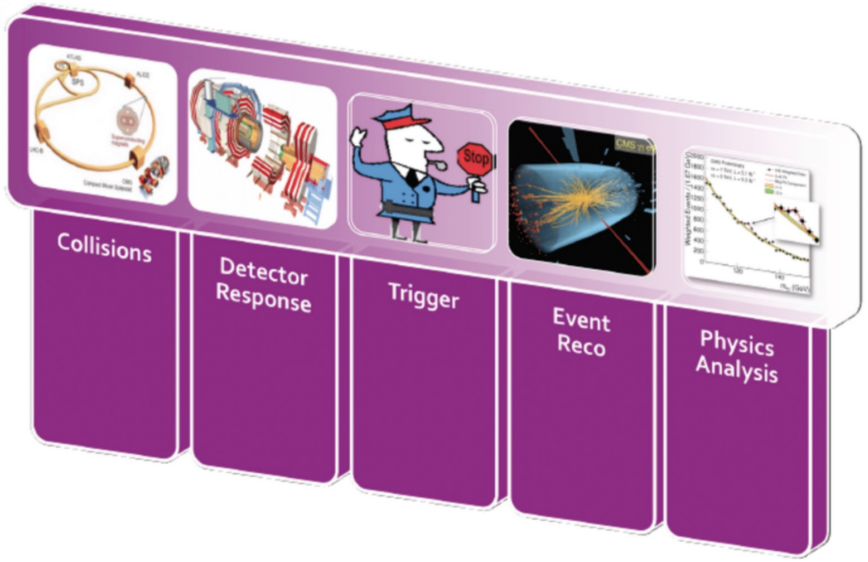


The CMS Trigger System: Design

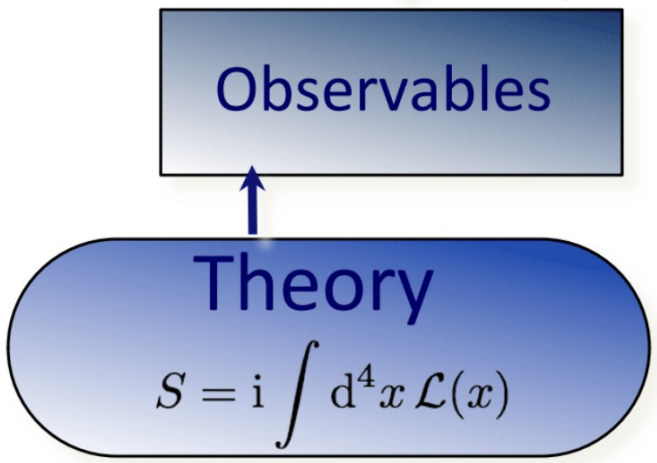


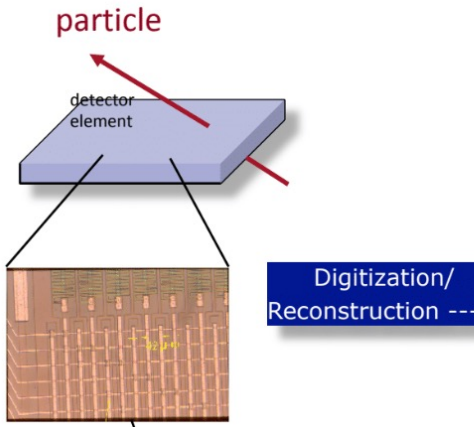
- The CMS Trigger System is organized in two tiers/levels:
  - **Level-1 Trigger** based on custom-made electronics to reduce the data/event rate from the crossing rate of 40 MHz to no more than 100 kHz, with 4µs latency
  - **High Level Trigger (HLT)** filtering events with software running on computing farm based on commercial CPU and now also GPUs, to further reduce the event rate for storage to 1 kHz (in the Run2), now around 1.5 kHz

From raw data to physics results

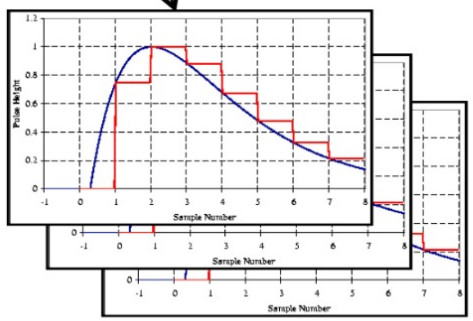


(same steps for the MC)

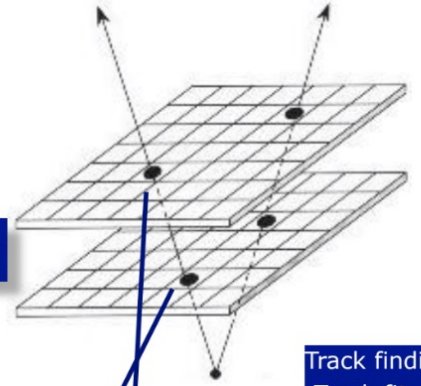




Digitization/ Reconstruction --->

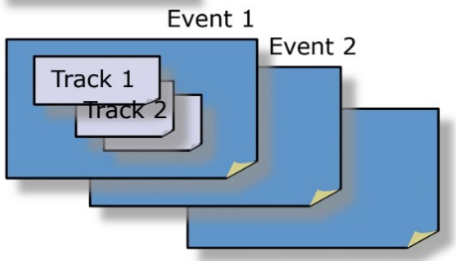


Analog signals

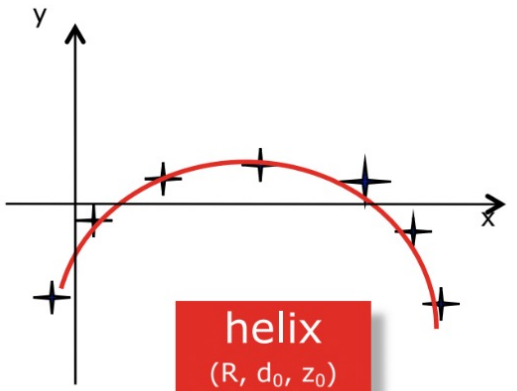


Track finding + Track fit --->

hits
(x_1, y_1, z_1, t_1)
(x_2, y_2, z_2, t_2)
...



Store the info for every event and every track <-------



helix
(R, d_0, z_0)

Magnetic field B: reconstruct

$\mathbf{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix}$

Track momentum

Open Data levels as defined in Data Access Policy

- 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 analyse them
    - Open Data - Research
  - **Level 4**: Raw data, and the software to reconstruct and analyse it

The CMS Data Tiers