The CMS Open Data workshop: Introduction

September 30, 2020

Kati Lassila-Perini
CMS Data preservation and open access coordinator
Helsinki Institute of Physics (Finland)

Welcome!

On behalf of the CMS Open data team

Involved in the preparation of this workshop:

Matt

Clemens

Edgar

Kati

The tutorial team 💙

Speakers

Edgar Carrera Jarrin - USF Quito

Matthew Bellis - Siena College

Kati Lassila-Perini - HIP Finland

Jesse Thaler - MIT

Clemens Lange - CERN

Julie Hogan - Bethel U/Brown U

Santeri Laurila - CERN

Tom McCauley - Notre Dame

Facilitators

Speakers +

Kevin Pedro - FNAL

Stefan Wunsch - CERN

Gabriele Benelli - Brown U

Farrah Simpson

Nikolas Pervan

Adelina Lintuluoto - CERN

Marguerite Belt Tonjes - Illinois U

Emanuele Usai - Brown U

Steve Mrenna - FNAL

Andrew Malone Melo - Vanderbild U

Theodor Herwig - FNAL

David Yu - Brown U

Nada Mohamed - Siena College

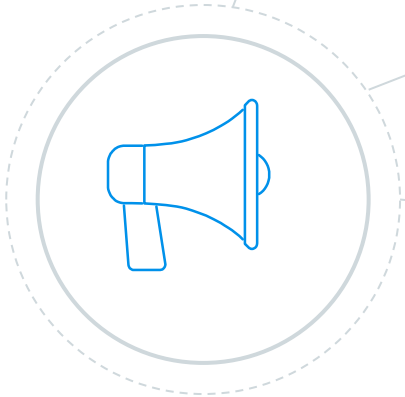
CMS participants helping

Nick Manganelli

Clemencia Mora

Special thanks to LPC events committee and coordinators!

Get to know each other!



Please, tell something about you in this slide set!

1.

Goals?

What do you expect?
What do we expect?

We made some assumptions

We think you want to understand:

- the basic object usage (object access, id, corrections, how to write them out)
- the event selection and triggers
- the luminosity evaluation
- the possibilities for large-scale data processing.

In addition, we think you will be interested in

- examples how to estimate systematic uncertainties
- examples of analyses that have already been carried out, along with lessons learned.

But that's not all - we get something as well

We want to:

- build a community of users
- introduce https://opendata-forum.cern.ch/
- get understanding of the usage patterns and needs
- get feedback of what is missing in the documentation and tutorial material
- build a proper CMS open data user guide.

Ambitious goals ➟
Do we reach them?

Bear with us, this is the first workshop of this kind.

2.

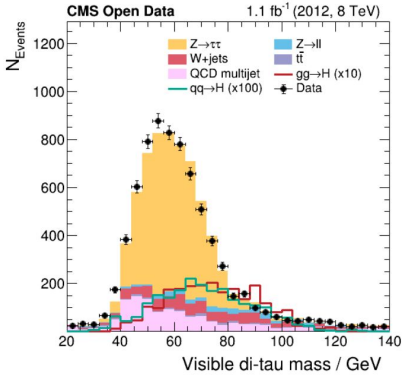
How to get there?

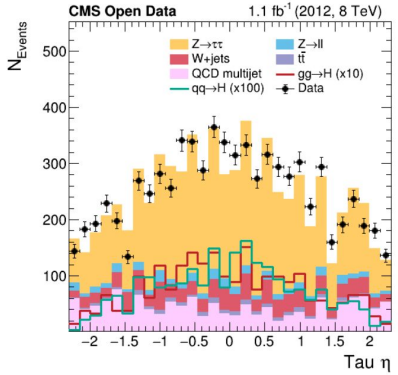
Workshop structure
Working methods

Walk-through of an analysis

- The tutorial is built thinking of you writing your own analysis
- You will follow the steps of an existing simplified example: H→$\tau\tau$

- A typical two-step process
  - Write smaller format
  - Analyse that format
- You will expand it in some areas to get an idea what it needs to take it to the research level.





A set of mandatory pre-exercises

Pre-exercises

(*Mandatory exercises must be completed before the start of the workshop*)

Mandatory 5 min	Orientation
Mandatory 2 h	Virtual Machine **or** Docker container setup for CMS Open Data
Mandatory 2 h	CMSSW fundamentals
Mandatory 2.4 h	ROOT basics
Optional (external lesson)	The Unix Shell
Optional (external lesson)	Version Control with Git
Optional (external lesson)	Programming with Python

Very important:

reply to the poll!

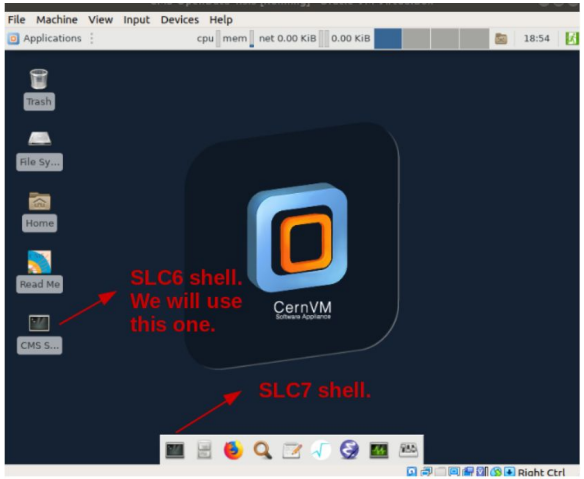
Pre-exercises

- To set and test your working environment before the workshop
- To give some background information on the tools in use at this workshop
  - Many thanks for your questions!

- If neither of these two is familiar, you're in trouble!

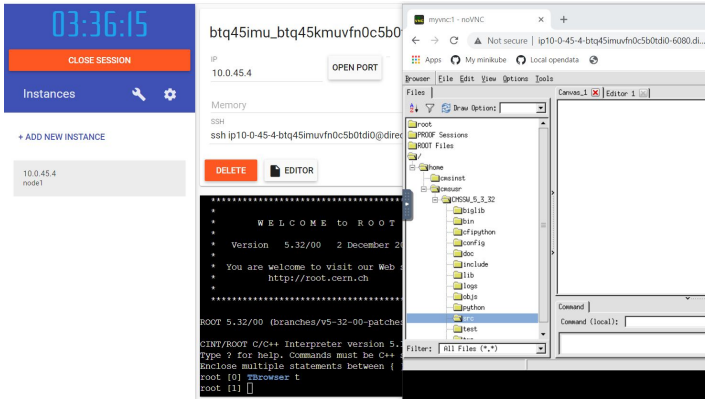
```
Setting up CMSSW_5_3_32
CMSSW should now be available.
```

- For this time, we have a temporary solution:
  - You can use a browser-based docker in
  - http://learn.cms-cloud.ml/
    - NB: nothing is saved after the session.

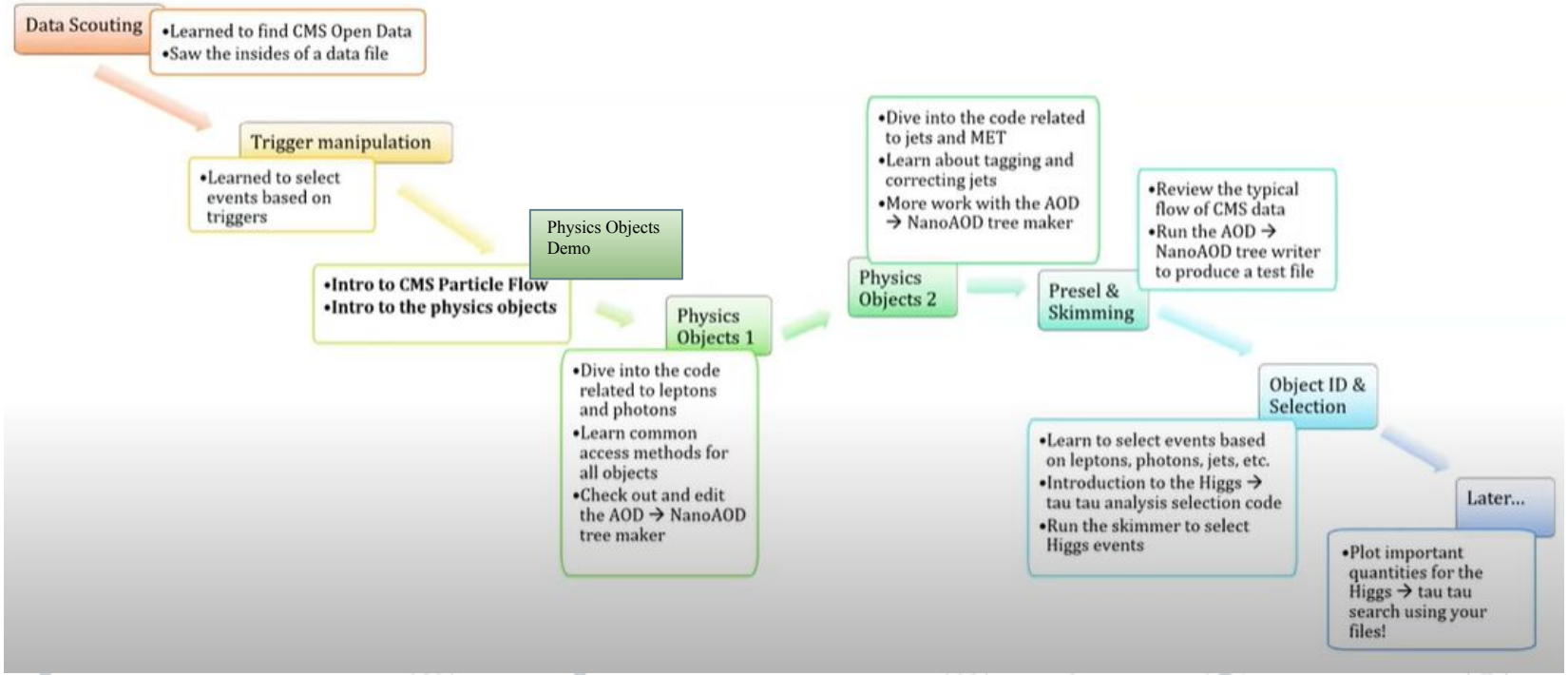


Play with docker in browser - backup environment

- In addition to the “learn” environment, we provide
  - http://docker.cms-cloud.ml/ with /cvmfs and GUI in browser
- To use it:
  - Click on “Add new instance”
  - docker run -it -P -p 5901:5901 -p 6080:6080 -v "/cvmfs:/cvmfs" cmsopendata/cmssw_5_3_32_vnc:latest /bin/bash
  - Type start_vnc (choose a password)
  - Click on “Open port” and type 6080
  - Click on vnc.html In the browser tab which opens
  - Connect (give the password).
- One session is max 4 hours
- Be aware: nothing is saved after the session.



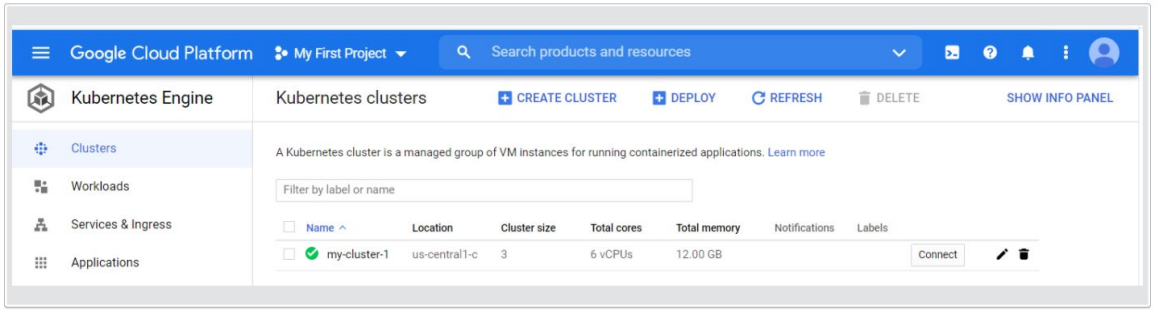
Tutorial workflow (from the Physics Objects Demo video)



Friday morning: CMS analysis on a cloud environment

- You will have the opportunity to learn how to run a CMS open data processing job in real scale on commercial cloud environment
- It will be hands-on and you will get a temporary account
  - Details will be sent to you later
- We've got resources for it through Archiver project in which CERN Open Data is a use-case
- We'll be using Kubernetes engine on Google Cloud Platform

- Don't miss it!



Friday afternoon - a featured demo on ADL

- Our colleagues will offer a short demo on use of Analysis Description Language
  - See further information
- This is not part of the CMS open data distribution but you may find it interesting
  - It will be built around the H→$\tau\tau$ example.

Schedule

Wednesday

08:30-09:00	Welcome and orientation	Organizers
09:00-09:45	Live Presentation: Workshop Introduction	Kati Lassila-Perini
09:45-10:00	Break	
10:00-10:45	Live Hands-on lesson: Dataset scouting	Matt Bellis
10:45-12:00	Live Hands-on lesson: Trigger manipulation	Edgar Carrera
12:00-12:15	Break	
12:15-13:00	Async Demo: Physics objects	(Julie Hogan)
13:00-14:30	Lunch	
14:30-16:00	Async Hands-on lesson: Physics objects I	(Julie Hogan)
16:00-17:00	Live Fermilab Colloquium: The Future is Open	Jesse Thaler

Thursday

08:30-10:00	Live Hands-on lesson: Physics objects II	Julie Hogan
10:00-10:15	Break	
10:15-11:15	Live Hands-on lesson: Pre-selection and skimming	Julie Hogan
11:15-11:45	Live Hands-on lesson: Object ID and selection.	Julie Hogan
11:45-12:00	Break	
12:00-13:00	Live Hands-on lesson: Plotting and interpretation	Matt Bellis
13:00-14:30	Lunch	
14:30-15:15	Async Demo: Luminosity and data quality.	(Tom McCauley)
15:15-17:00	Async Hands-on lesson: TBA	(TBC)

Friday

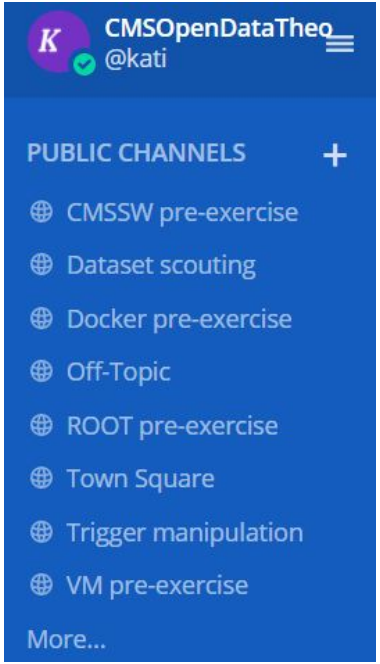
08:30-10:30	Live Demo: CMS analysis in the cloud	Clemens Lange Adelina Lintuluoto
10:30-10:45	Break	
10:45-11:45	Live Hands-on lesson: CMS analysis in the cloud	Clemens Lange Adelina Lintuluoto
11:45-12:00	Break	
12:00-13:00	Live Hands-on lesson: Pile-up corrections	Santeri Laurila
13:00-14:30	Lunch	
14:30-15:00	Live Featured Demo: ADL	Sezen Sekmen
15:00-17:00	Wrap up and time for discussion	Everybody

Mornings: Live Hands-on lessons and demos - through zoom, with dedicated mattermost channel
Afternoons: Async lessons - dedicated mattermost channel and zoom as “open office” for everyone

Material available from the schedule and from the indico agenda

Getting help - live

- In mattermost, choose the channel corresponding to the lesson.
- Do not hesitate to ask!
  - But check if the same question has already been asked.
- Cut and paste the command and the error message
  - If needed, use `some code in line`
  - or ```block of code or output```
  - shift-return for a line break in a message
- Reload the tutorial page every now and then for updates.
- During live lessons
  - Type in quick questions in zoom chat
  - Use “Raise hand” for voice questions



Ask! Ask! Ask!

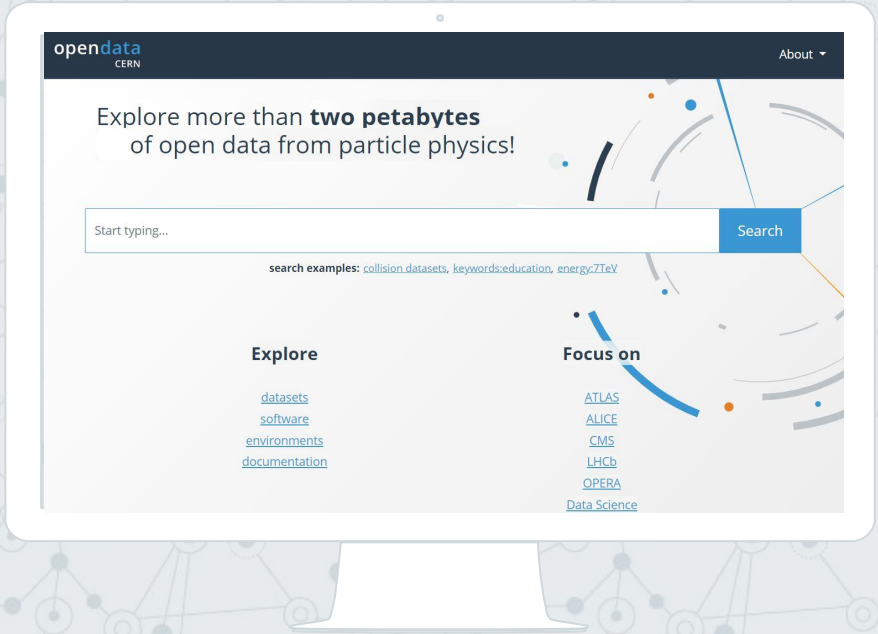
The excellence in teaching is not pushing forward those who know the most,

but taking care that no one is left behind.

3.

How to get help after?

Information sources
Communication



CERN Open data portal

Serves the data, associated analysis artefacts, usage examples

CMS products on the CERN Open data portal

Data + metadata

Collision data

Simulated data

Derived data

Associated artefacts

VMs, containers

List of certified runs

Luminosity tables

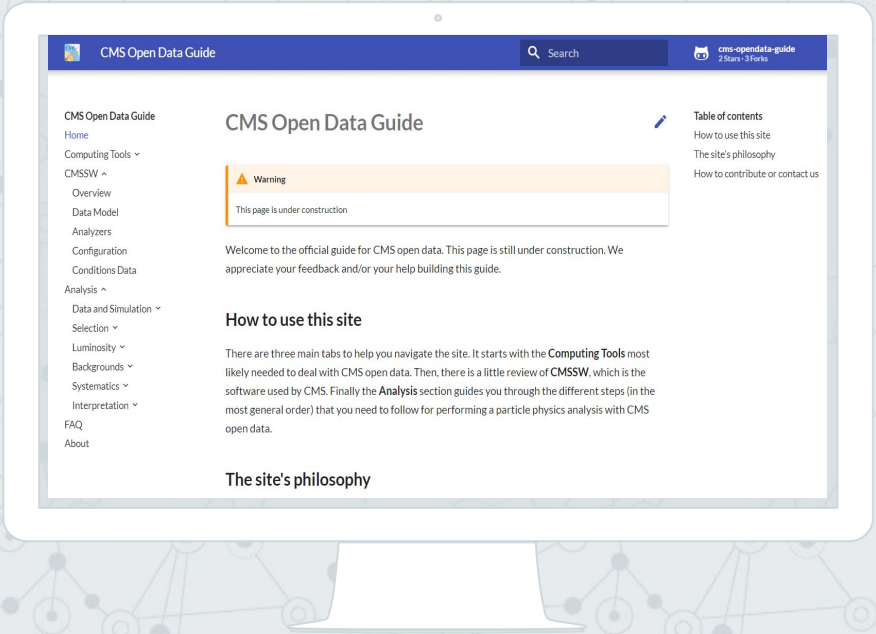
Condition data...

Example and guides

Topical guides: condition data usage, trigger systems, Monte Carlo generation, luminosity estimation, docker containers

Analysis examples...

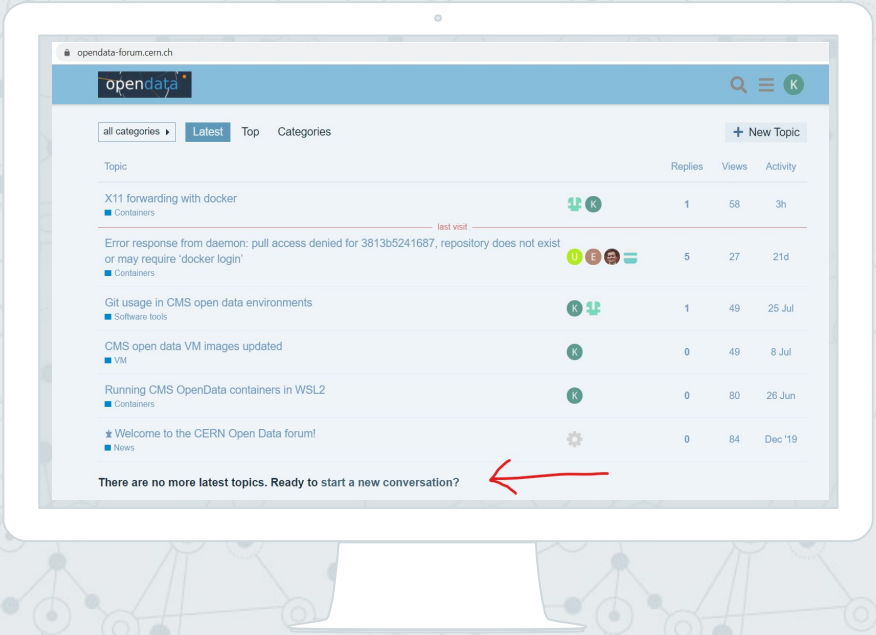
CERN Open data portal is primarily a data catalogue.



CMS Open data guide

Work in progress, will be completed with the material in this tutorial.

Do you want to help?



CERN Open data forum

Feel free to post questions! Feel free to reply as well!

Most frequently asked questions at this workshop will be added.

Other sources of information

- Open data portal support mail: opendata-support@cern.ch
  - Technical issues
  - Questions to limited audience
- CMS WorkBook and SWGuide
  - Careful: instructions might not correspond to the CMSSW version needed for open data
- CMSSW source code
  - Keep in mind the versioning, for 2011-2012 open data use CMSSW_5_3_X as tag.

4.

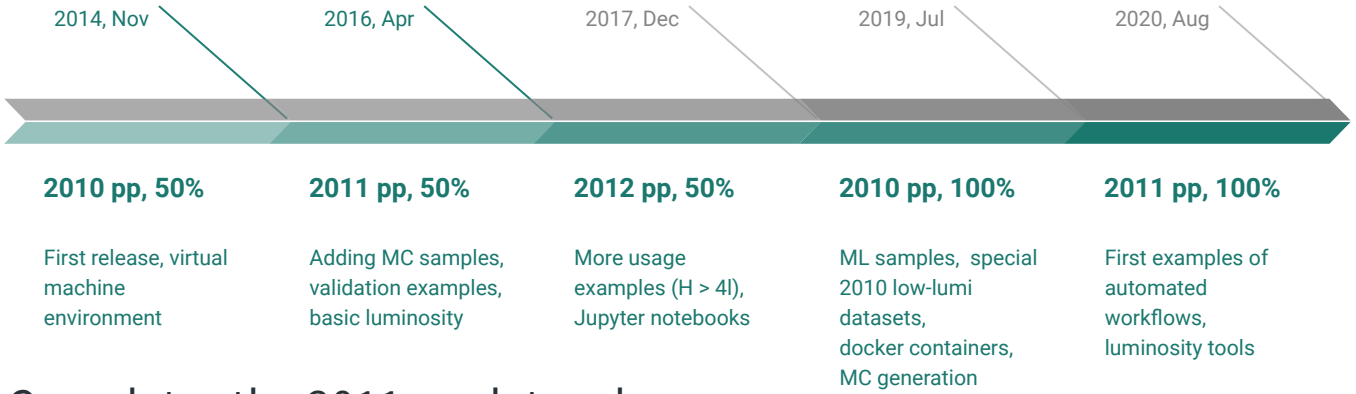
News and outlook

Data release in August 2020

CMS open data plans

CMS data release - August 2020

- The 5th CMS open data release since 2014



- Completes the 2011 pp data release
  - Now all 2010-2011 and 50% of the 2012 pp data in public domain.

Highlights

- Now full 5 fb^{-1} of 2011 pp data available
  - In addition to all 35 pb^{-1} of 2010 pp and 11 fb^{-1} of 2012 pp data released earlier.
- Improved usability with examples of automated workflows with:
  - software containers which can be used on users' local resources or on public cloud platforms enabling large-scale processing
  - state-of-the-art software tools with workflow definitions preserving the exact analysis procedure.
  - Used e.g in a new validation and usage example of early low-luminosity special data with the CASTOR calorimeter.
- Regular release schedule and follow-up with open data users helps improving usability, in this release:
  - Luminosity tools and improved information now available, very important for research use
  - Improved documentation for software container usage
  - Improved search functionalities for simplified data for educators.

Future releases

- CMS data policy was recently updated
  - to reflect the current delay:
    - 2015 data release foreseen in 2021, 6 years after data taking
  - and to take into account the slow down of luminosity increase
    - Open data will be limited to 20% of the total luminosity of similar data while data taking is still ongoing.
- CMS continues regular data releases, the next releases will include
  - 2010-2011 heavy ion data
  - first Run2 data from 2015.
- Working towards a common CERN open data policy.

5.

Now, let’s get to work!

Enjoy the workshop!
We’ll love to hear feedback from you on Friday.

Thanks!

Any questions?

Find us in mattermost

Credits

Thanks to my colleagues

- in the DPOA group in CMS
  - Edgar Carrera, Clemens Lange, Matt Bellis, Lara Lloret, Achim Geiser and many others
- in the CERN Data preservation services
  - CERN Open data portal team, and many other services that we rely on
- in the CERN Open Data policy working group

And great thanks to all CMS open data users!

And thanks to SlidesCarnival for this free presentation template