Open Data in CMS

Kati Lassila-Perini

Helsinki Institute of Physics

HEP Software Foundation
LAL - Orsay
May 3, 2016

CMS data levels and open data

- CMS experiment has approved a data preservation, re-use and open access policy, which underlines the will to preserve the data and defines the approach to access to them at various levels:
  - Level 1 - Open access publication and additional numerical data
  - Level 2 - Simplified data for outreach and education
  - Level 3 - Reconstructed data and the software to analyze them
  - Level 4 - Raw data, and the software to reconstruct and analyze them.

CMS Open Data

- CMS continues publishing and promoting levels 1 & 2.
- CMS made the first release of reconstructed data in November 2014.
  - 28 TB of 2010 collision data in AOD format.
- The latest CMS data release in April 2016
  - $>$ 100 TB of 2011 collision data in AOD format
  - $>$ 200 TB of corresponding MC data

CMS Open Data release

- Data
  - CMS collision data in format used in analysis by CMS physicists (AOD)
  - In the latest release, a partial set of simulated MC included (for the first release no corresponding MC available)
  - For future releases, include "miniAOD" (less complete, but more compact and easier to use)
- Tools
  - VM image of the computing environment
  - Access to the corresponding software and condition data
  - Access to data through xrootd or direct download
- Instructions
  - Basic instructions to get started ($\approx$ 15 mins to setup) with examples
  - Basic description of the physics objects
- Examples of derived datasets to be used in different education and outreach contexts
  - Event display, online histogramming
  - Code to produce the derived datasets

The challenge: knowledge preservation

- In HEP, we are doing well with the "immediate" metadata, such as
  - beam conditions, event and run numbers, provenance information (raw data from which data have been reconstructed, the software version used in the reconstruction)...

  recorded together with the data records at the time of creation.
- We are doing poorly with the "context" metadata, such as
  - how to pick up the right objects in the data
  - how to know if there are additional selections, corrections...

  in general, the practical information needed to put the data in context and analyze them: information, which is readily available and even obvious at the time of the data analysis, but easily forgotten.

Open Data helps/forces us to meet this challenge

- Information must be collected and released together with the data.

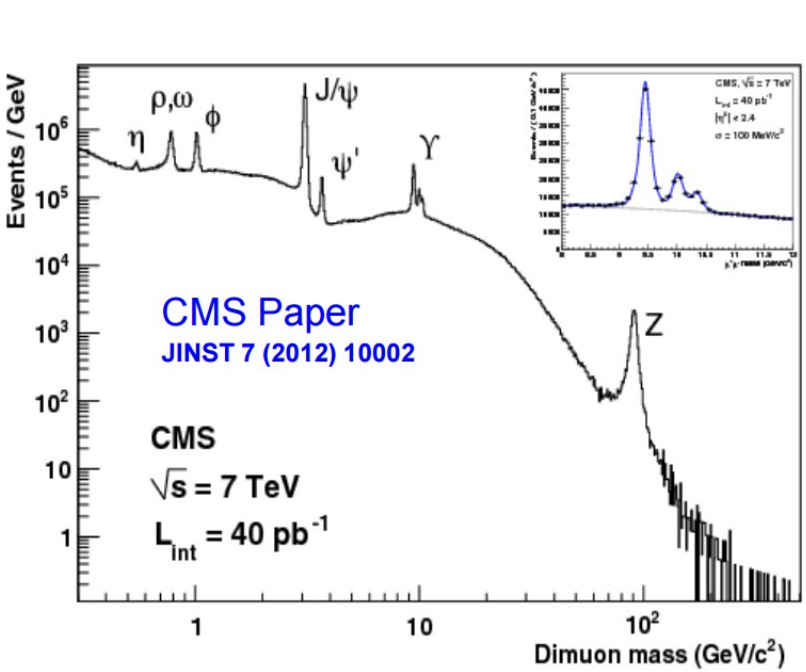
What is new on the portal: a walk-through

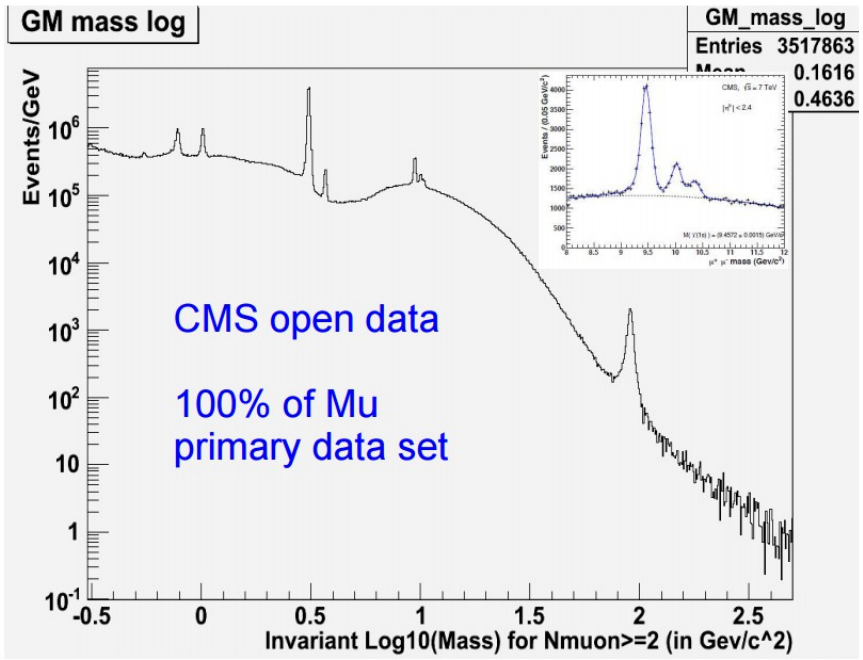
- About CMS data - updated to mention simulated data
- New event display with enhanced functionalities and new example events from 2011
- Updated instructions for 2011 data for installing CMS Open Data VM
- Updated instructions for 2011 data for Getting started
- New example to reproduce di-muon spectrum with 2010 data
- CMS research collections - several new collections
- Enhanced trigger explanation with trigger paths and config files for primary datasets
- Run-Release-Trigger configuration information summarized in a table
- New MC records with detailed generator information through config files
- Config files stored in a collection - not for direct browsing but useful as linked to the datasets
- New validation utilities collection to host also the validation code for legacy datasets
- The condition data is stored on the portal for the sake of consistency (no need to download)
- New records in the Tools collections for updated examples, CMSSW, new VM and VM contextualisation scripts...

Open data benchmark/validation analyses

- Several benchmark analyses on AOD for validation and for external users soon on the portal:
  - high-level validation for each released primary dataset
  - feasible with the available data
  - possibility for comparison (later) with data at other beam energies
  - not too complicated but nevertheless interesting physics objects
  - published reference available.

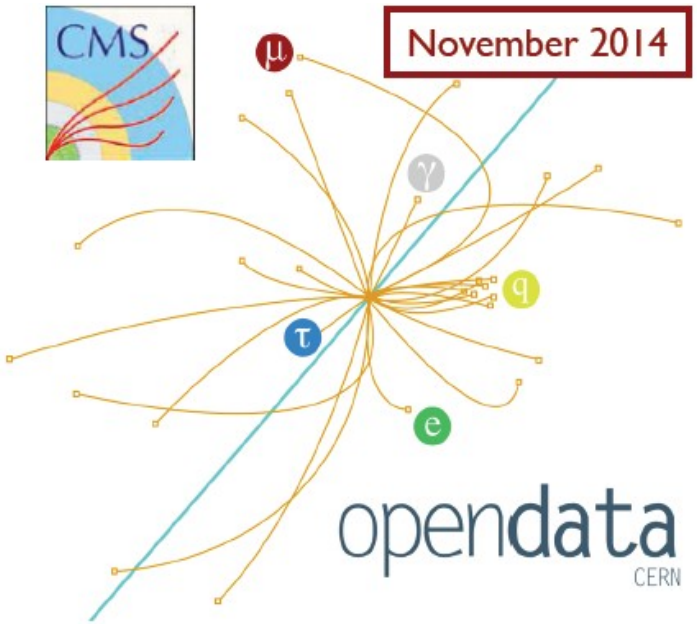
Achim Geiser et al (DESY)

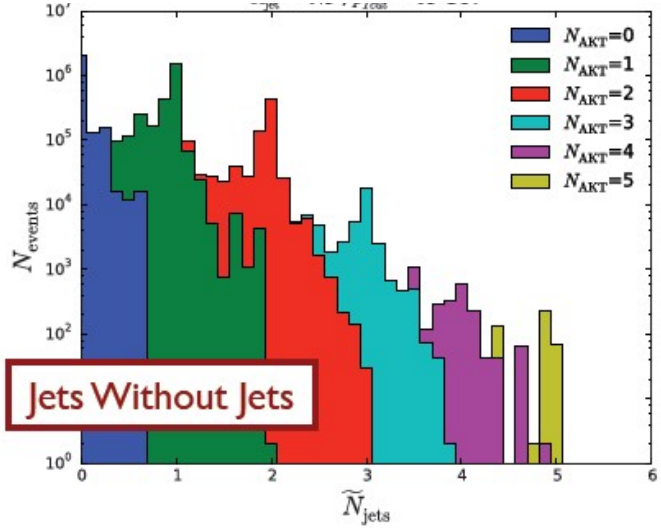




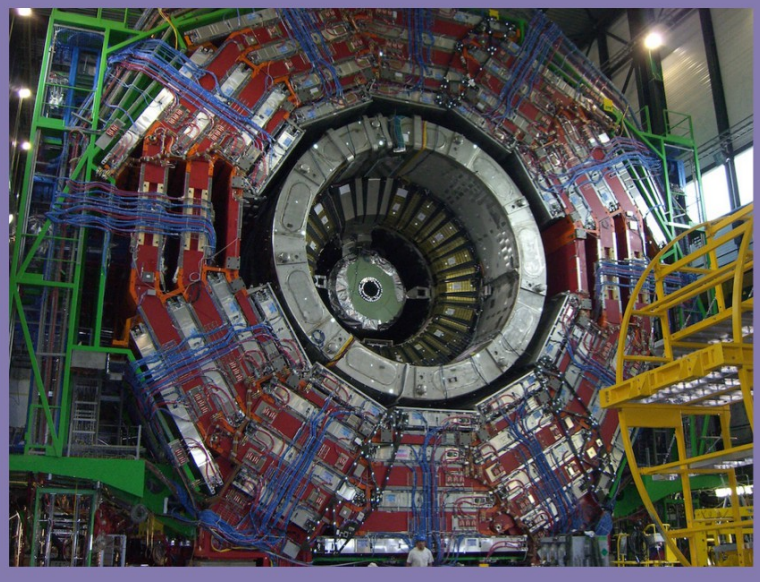
Examples of open data usage

- Ongoing analysis at MIT on jet substructure
  - a small group with a theorist, a post-doc and undergraduate
  - got started with the instructions on portal, and got help on volunteering basis from MIT and US CMS colleagues
  - aiming for a publication
  - willing to contribute to the documentation to help other users
- Research into cloud computing security
  - testing data deletions and operations by the local file system
  - the nature of the data itself is not relevant, but LHC data ideal.
- Pilot project on teaching applicatios for high-schools
  - ideas from physics teachers on further education course at CERN
  - based on the existing tools online tools (event display...)
- External resources have been generated
  - IFCA provides computing resources https://cmsopendata.ifca.es/





Extremely preliminary from Wei Xue
(limited sample size, missing MinBias, no JEC factors)



ABOUT

Look to the LHC CMS detector from inside, start analyzing its data.

Instituto de Física de Cantabria provides you with a virtual environment for CMS Open Data analysis for educational use, developed in collaboration with aeonium.

Outlook

- Impact of the Open Data release has been very positive
  - a modest start, but well received by the public and the funding agencies
  - little unexpected additional workload to the collaboration
  - the data are in use!
- Excellent collaboration with CERN services developing data preservation and open access tools
  - common solutions essential for long-term preservation
  - benefit from expertise in digital archiving and library services.
- Issues
  - data preservation must start when data analysis in ongoing, but we compete for resources for data taking and operations.
- CMS is looking forward to
  - seeing the newly released CMS open data widely in use.
  - your feedback: **are these data potentially useful to you?**