Open Data in CMS

Kati Lassila-Perini

Helsinki Institute of Physics

Data Science
CERN
November 13, 2015

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.
- Next release in preparation
  - 130 Tb of 2011 collision data in AOD format
  - 300 Tb of corresponding MC data

CMS Open Data release

- Data
  - CMS collision data in format used in analysis by CMS physicists (AOD)
  - For the next 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 cleaner)
- 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)
  - 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.

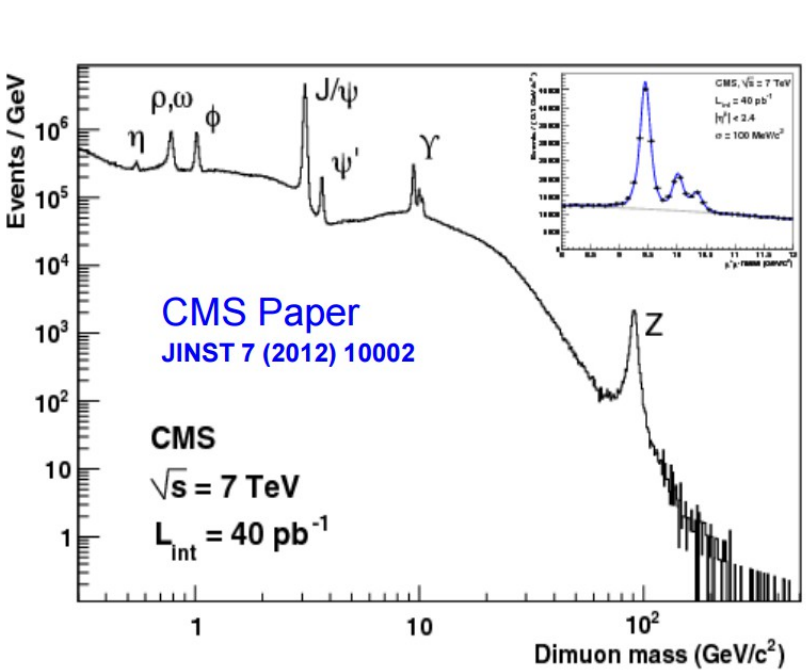
Open Data helps/forces us to meet this challenge

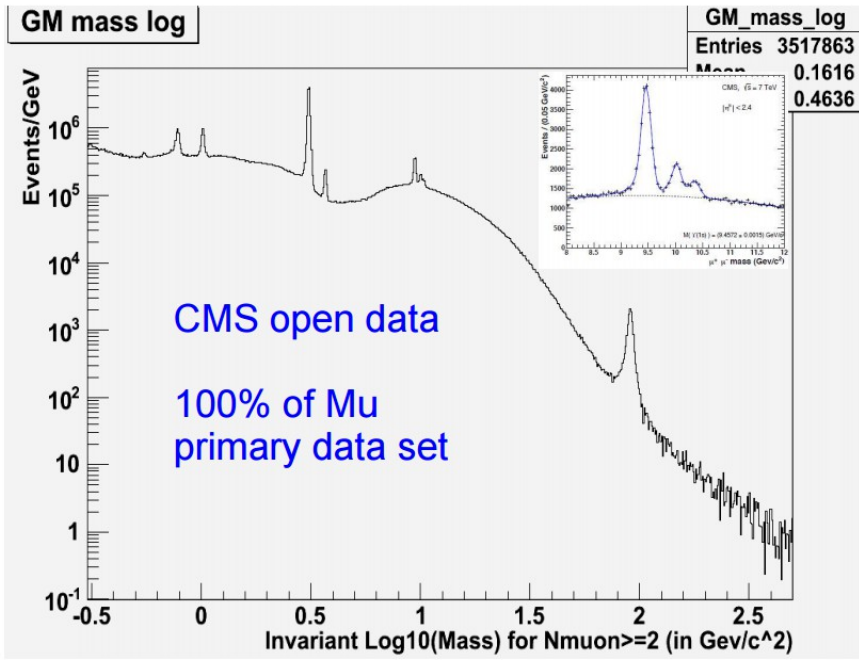
- Information must be collected and released together with the data.

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.

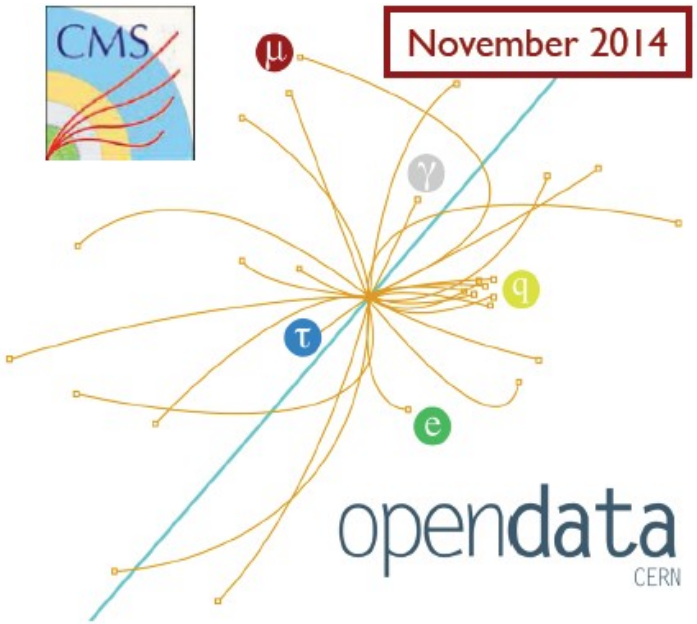
Achim Geiser et al (DESY)

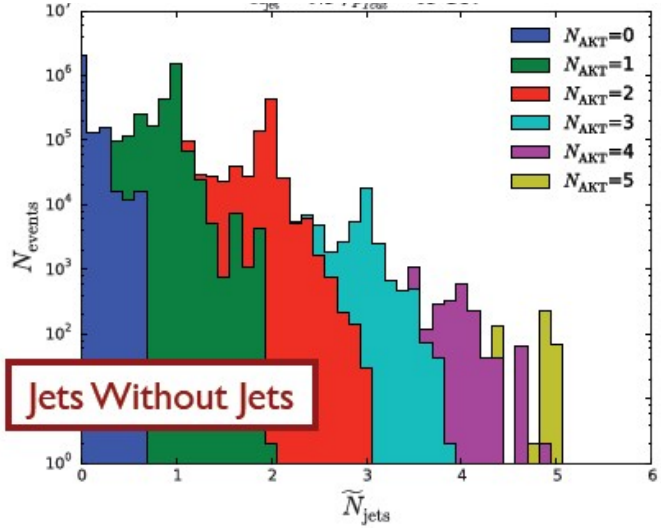




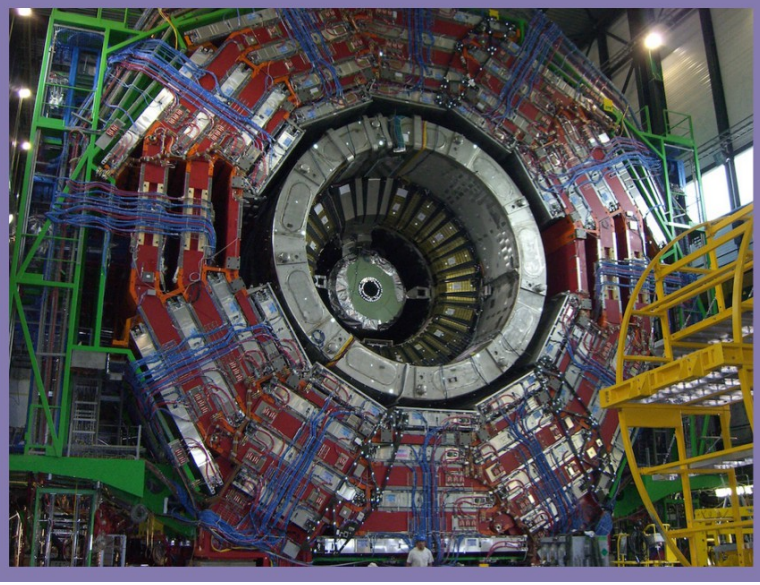
Examples of open data usage

- 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
  - no 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
  - long-term vision is difficult for experiments in data-taking phase.
  - data preservation must start when data analysis in ongoing, but we compete for resources for data taking and operations.
- CMS is looking forward to
  - releasing a new batch of open data through CERN Open Data Portal
  - using the tools and services set up for Open Data to achieve a true long-term data and knowledge preservation in HEP!