Data preservation and Open access in CMS

Kati Lassila-Perini

Helsinki Institute of Physics

DPHEP Collaboration meeting
CERN
June 9, 2015

Outline

Defining the legacy data

- 2011-2012 data have been reprocessed with CMSSW 5.3
  - collision data in AOD format ≈200 Tb for 2011, ≈800 Tb for 2012
  - similar amount of MC (or larger upto $\times 2$)
- The current plan is to keep one complete AOD reprocessing (in addition to 2×RAW)
  - no reconstructed collision data have yet been deleted, but deletion campaigns are starting.
- Most Run 2 analyses will use miniAOD's which are significantly smaller in size.

Evaluation of the open data release at the CB in June

- Evaluation of the data release, foreseen in the CMS data preservation policy
  - CMS data preservation policy: *"...This release will be followed by a full analysis of the procedure and the experience will be evaluated by the Collaboration Board and in absence of unexpected overhead to the Collaboration the public data release will be accepted as a standard practice."*
- The policy itself is not to be questioned, but the practical aspects of the implementation could be.
- Contents:
  - Brief summary of the released data (data= data, instructions, sw), usage, feedback, impact
- Start preparing the next release of 2011-2012 collision data and MC later this year.
  - Now as the CERN Open Data Portal is in place, the preparation is much easier.

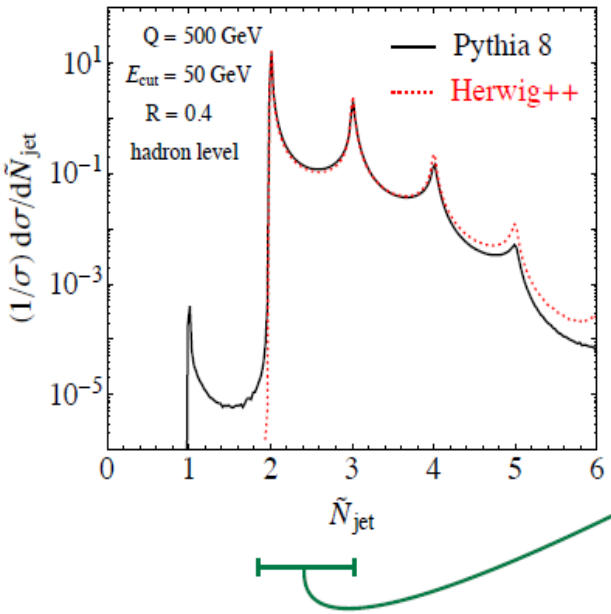
Examples of open data usage

- 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 in Finland
  - Ideas from physics teachers on further education course at CERN
  - Based on the existing tools online tools (event display...)
  - Two summer students started, continue development with Lapland Univ. of Applied Sciences (collaborated already for CODP testing)
  - A consortium of 20 high-schools offering online courses as test-bed
- IFCA provides computing resources https://cmsopendata.ifca.es/

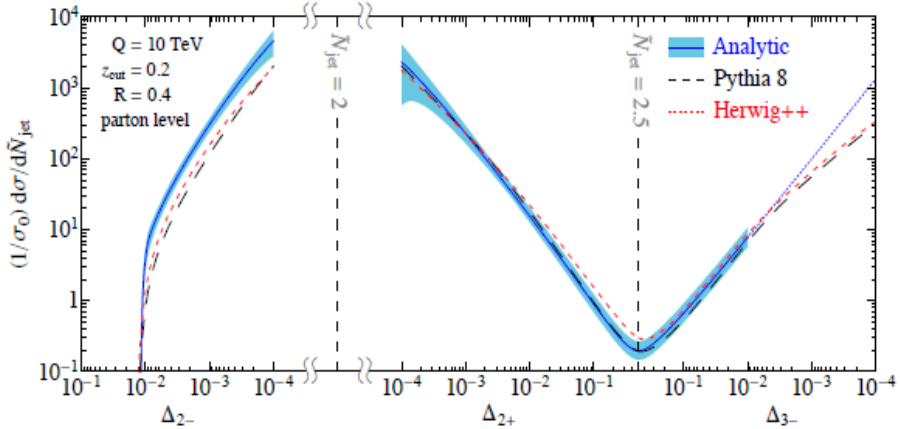
Planning for Archival Data Access?

"Jets Without Jets" $\widetilde{N}_{\text{jet}}(p_{T\text{cut}}, R) = \sum_{i \in \text{event}} \frac{p_{Ti}}{p_{Ti,R}} \Theta(p_{Ti,R} - p_{T\text{cut}})$

In Monte Carlo…



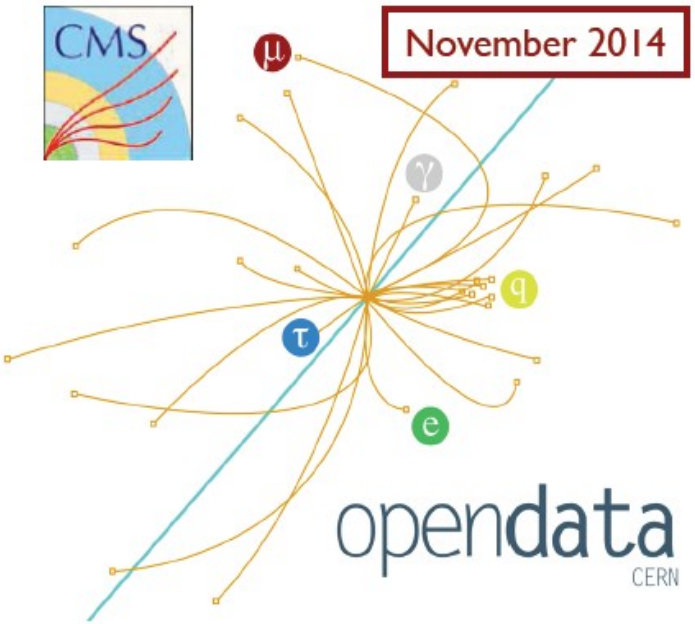
…in QCD at $O(\alpha_s^2)$…

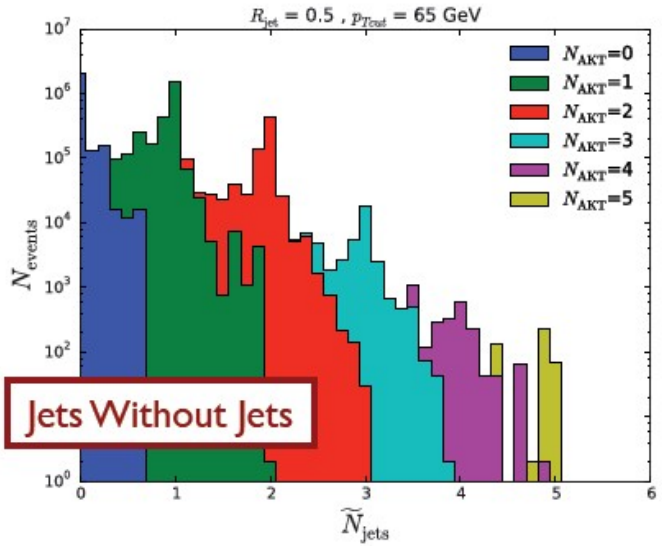


…in data?

[Bertolini, Chan, JDT, 2013; Bertolini, JDT, Walsh, 2015]

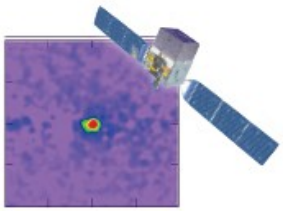
Planning for Archival Data Access?





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

c.f. Fermi



Can FCC accelerate scientific progress through judicious open data releases?

Open data benchmark analyses

- CMS plans to offer analysis examples, and tools and benchmarks to validate open primary datasets
  - These two goals can coincide.
  - Analysis example chain for two leptons already available on the portal.
- Work on more benchmark analyses on AOD for external users and for validation has started:
  - feasible on 7 TeV 2010 AOD data, with small luminosity
  - possibility for comparison (later) with data at other center-of-mass energies
  - not too complicated but nevertheless interesting physics objects
  - basic studies feasable without MC
  - published reference available.

Data format exercise - a step to the future?

- CMS data format is difficult for external users.
- Software being developed for the future HEP experiments aims to offer easier usability, see. e.g.
  - ‣ Benedikt Hegner: Software for FCC Physics and Experiments
- Start an exercise based on the di-lepton example available on the portal, trying to match our objects to the future format (or vice-versa...)
  - leptons "easy" and well defined
  - get feedback from the open data users
  - not a complete data format definition, but a demonstator/play-ground.
- Summer student from CMS Open Data project with the help of Benedikt Hegner (PH-SFT)
  - More news by the end of the summer.

CERN Analysis Preservation framework

- CMS has provided input for the data model and user interface design, and defining pipelines for automated ingestion from CMS services.
- The CAP use-cases are well acknowledged by CMS.
  - ▸ See Tim's slides from Monday
- CAP will be valuable tool to start data preservation while the analysis is active.
- Several physics analyses have volunteered as test users for the interface.
- Some of the analyses (cfr validation benchmarks for open data) could eventually be directed to the Open Data Portal.
- Use of miniAODs in Run2 ($\approx$10% in volume of the AODs) will ease recording of data of the intermediate processings.

Outlook

- Impact of the open data release has been very positive
  - 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 services and with DASPOS
  - Common projects are essential for long-term preservation
  - Benefit from expertise in digital archiving and library services
  - Fruitful discussion with other experiments.
- Long-term vision and planning is difficult for ongoing experiments
  - DPHEP offers a unique viewpoint.
- CMS is looking forward to
  - stress-testing CERN Open Data Portal with the new data release
  - stress-testing CERN Analysis Preservation framework with our analyses!