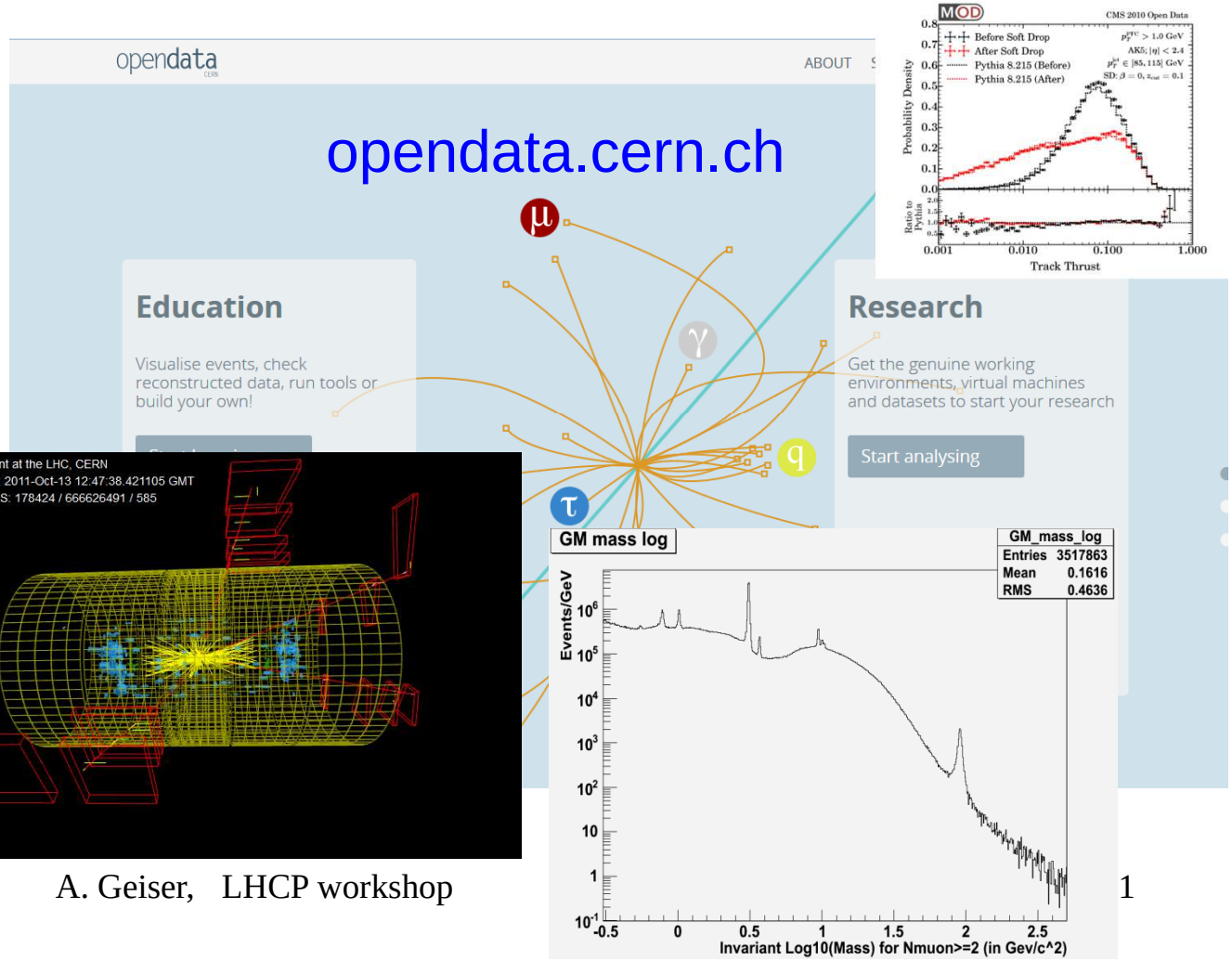
CMS Open Data in Research

Achim Geiser for the CMS collaboration

Outreach session of LHCP workshop, Shanghai, China, 15.05.2017

- The vision
- The implementation
- CMS Open Data for Research
- Status, results and prospects
- Conclusions



The Vision

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 503 > Issue 7477 > News > Article

NATURE | NEWS

عربي

LHC plans for open data future

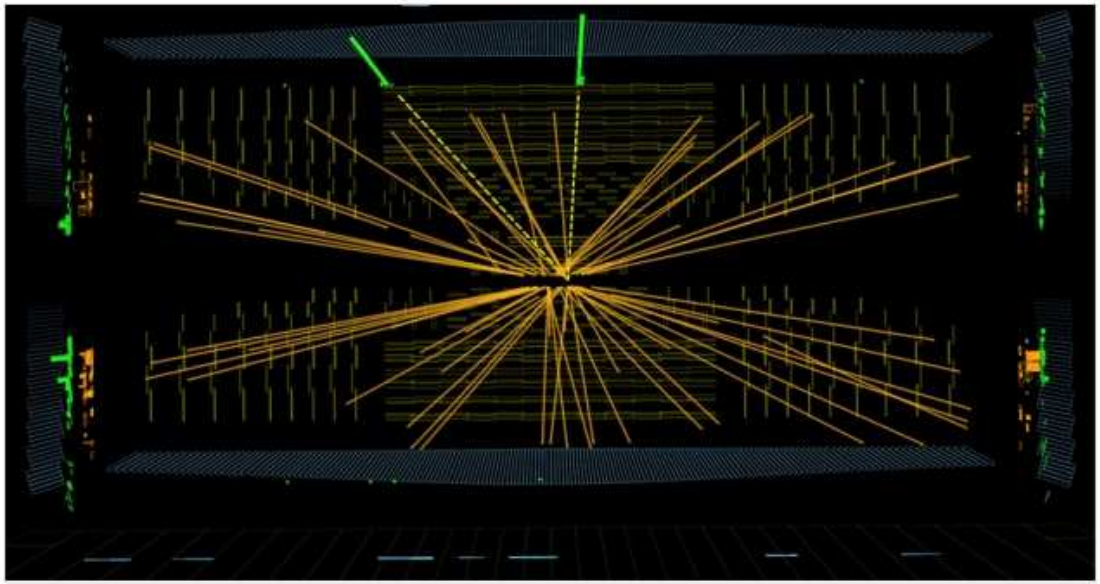
Researchers share results to keep them accessible.

Elizabeth Gibney

26 November 2013

PDF | Rights & Permissions

statements by
C. Diaconu (DPHEP)
M. Hildreth (DASPOS)
K. Lassila-Perini (CMS)
J. Shiers (CERN,DPHEP)
D. South (DESY, HERA)



Thomas McCauley/Lucas Taylor/CMS Collection/CERN

Data from the Large Hadron Collider, such as this decay of a Higgs boson, could be made publicly available.

- **Preserve data and knowledge (metadata)**
- **Open sharing – data and knowledge more likely to survive if constantly used -> enlightened self-interest**
- **Make data available to school pupils and researchers alike - allow them e.g. to reconstruct the Higgs discovery**
- **(Allow CMS physicists to recreate results from ATLAS and vice versa -> backup)**
- **Mine data to test new theories and provide crucial references**
- **Contain cost to ~1% of operating costs -> worth the effort**

The implementation: Open sharing

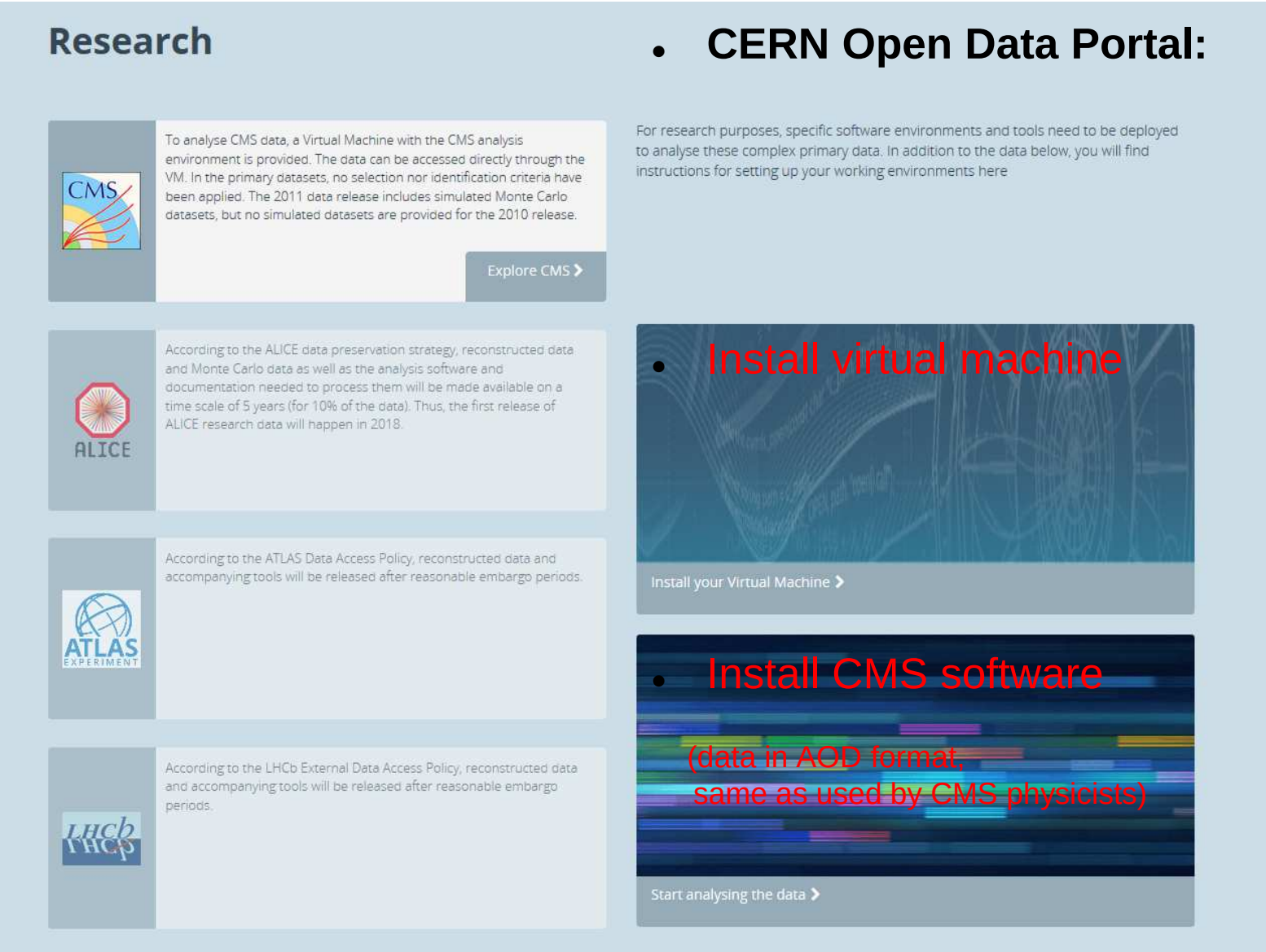
- **CERN Open Data Portal: opendata.cern.ch**
- Access point to growing range of data produced through research at CERN. Disseminates preserved output from various research activities, including accompanying software and documentation needed to understand and analyze the data being shared.
- Adheres to established global standards in data preservation and Open Science: the products are shared under open licenses; issued with a digital object identifier (DOI) to make them citable objects in the scientific discourse.
- Close collaboration between experiments, CERN IT and scientific information services

this talk:

focus on Research applications

(many educational applications available from all four experiments)

The implementation: Data and knowledge



- **CERN Open Data Portal:** **opendata.cern.ch**

~15 min to set up

so far:

only CMS released Research level data

-> pioneer

Open Data in CMS

- **CERN Open Data Portal: http://opendata.cern.ch/about/CMS**
- **CMS data preservation, re-use and open access policy,** **http://opendata.cern.ch/record/411,**
  defines approach to data access at various levels:
- **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 analyze them** **Open Data - Research**
  - **Level 4 – Raw data, and the software to reconstruct and analyze them**

CMS Open Data for Research:

- 1st release of 28 TB of reconstructed 2010 7 TeV pp collision data in Nov. 2014
- 2nd release of 130 TB of 2011 7 TeV pp collision data and >200 TB of corresponding MC data in April 2016 **~ half the respective full datasets**
- 3rd release of 8 TeV pp data + MC (~2 PB) approved for later this year

The challenge: knowledge preservation

HEP doing well with "immediate" metadata, such as

- beam conditions, event and run numbers, provenance information (processing and reconstruction chain, software versions) recorded together with data at time of data set creation

doing poorly with "context" metadata, such as

- how to pick up the right objects in the data and their documentation
- how to know if there are additional selections, corrections, …
- in general, practical information needed to put data in context and analyze them: information readily available and even obvious at time of immediate data analysis, but then easily forgotten
- **Open Data helps/forces us to meet this challenge**

Information must be collected and released together with the data

How we (try to) meet the challenge

- **information provided is not perfect** (and will not be) **but still useful and usable**
- **information is missing for an analysis to be completed ?**
  (e.g. currently luminosity values for collision data and cross sections for MC)
  -> we are more than happy to take the feedback at opendata.support@cern.ch
  and provide them (as long as we have them available ourselves)
- **being done for the first time** (in HEP) **-> learning process for everyone,**
  for users to learn to use these data, for us to gather and provide the necessary information from internal sources
- **we are full of good will but very low on resources** -> be patient

Most results presented in next slides obtained starting from scratch on CERNVM virtual machines, using windows or linux office desktop or laptop computers (can do it “from your kitchen”!), using publicly available documentation of CMS software. No grid jobs, no batch jobs on farm, no CMS account needed.

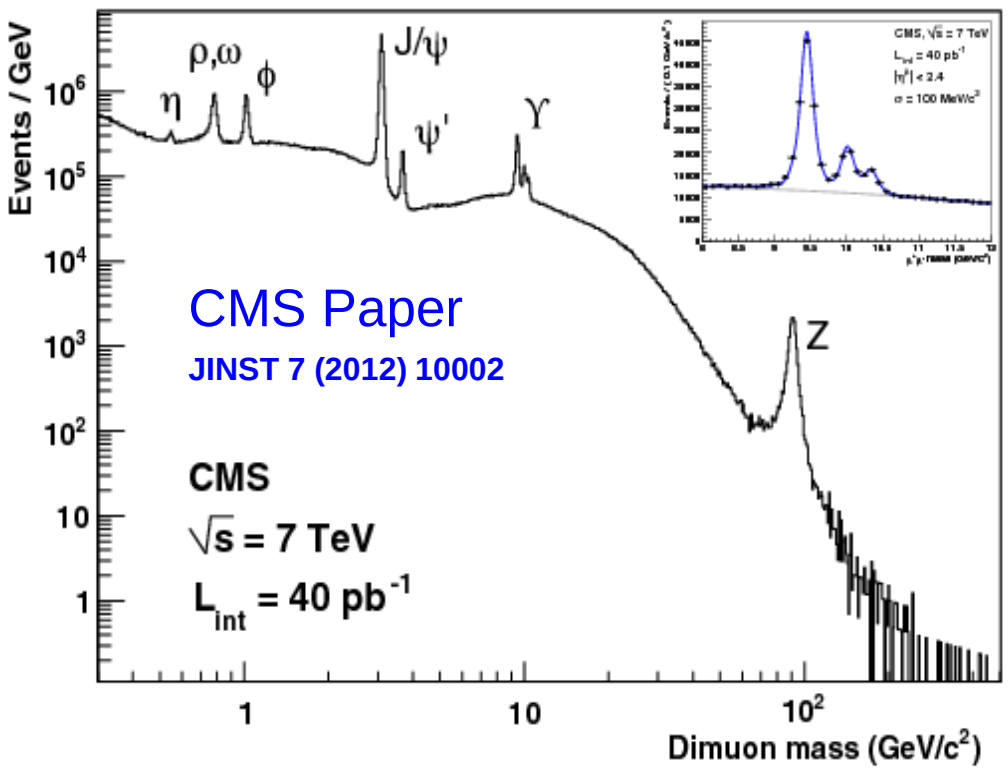
Many obtained by undergraduate students supervised by experienced physicists -> excellent training opportunities!

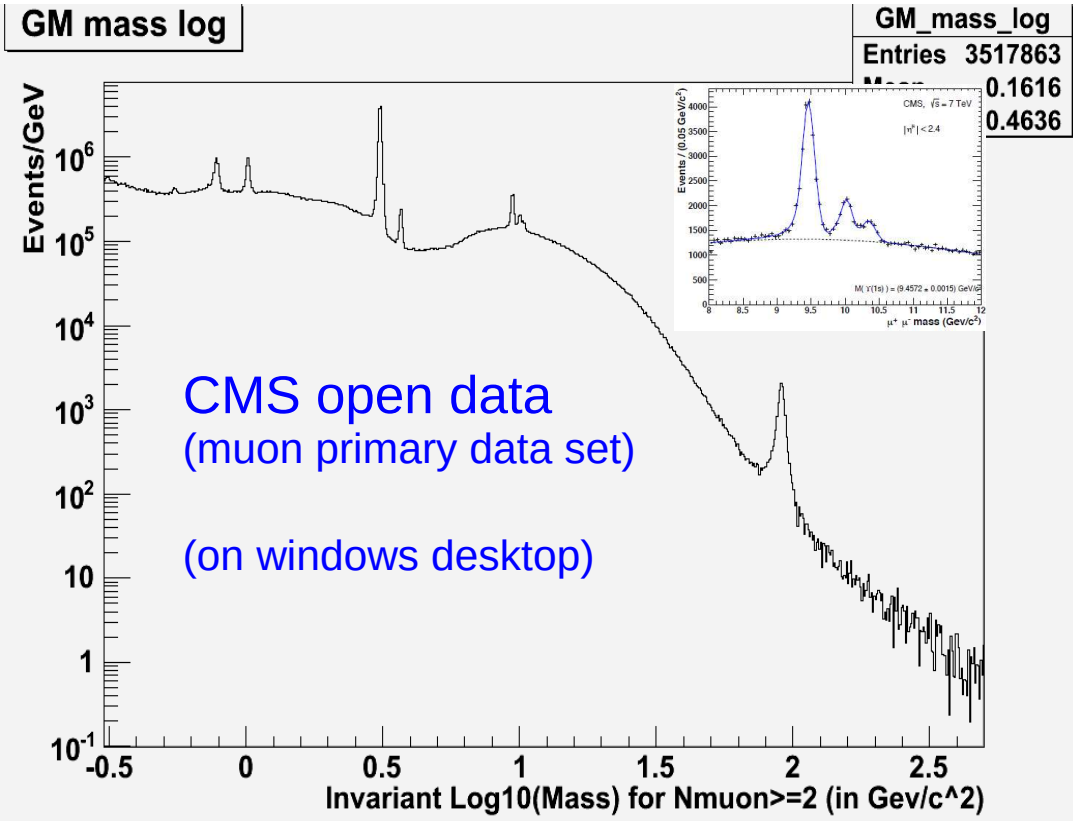
Provide references: validation/benchmarking/analysis examples

Open release of 2010 data in fall 2014

Using open data portal: http://opendata.cern.ch/about/CMS

Dimuon invariant mass distribution





Open Data benchmark analysis: “Ridge”

A. Nassirpour, summer student 2016

https://indico.desy.de/getFile.py/access?contribId=4&resId=0&materialId=slides&confId=15932

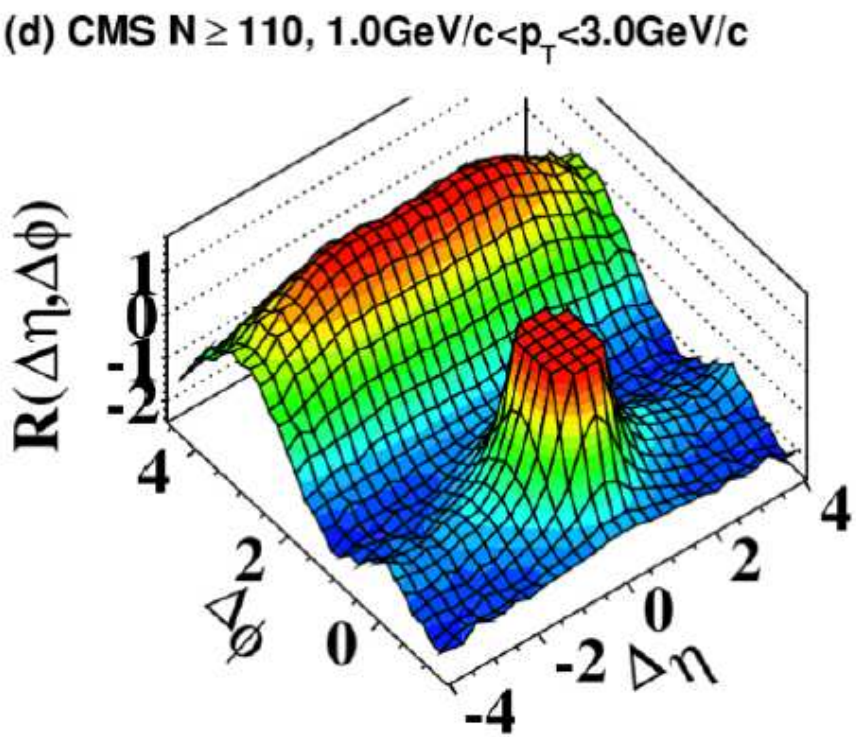
Unexpected „Ridge“ was observed in 2010 pp data,

JHEP 1009 (2010) 091 (topcite 500+)

Can be ~reproduced by selecting high multiplicity triggers in Minimum Bias dataset of 2010 open data

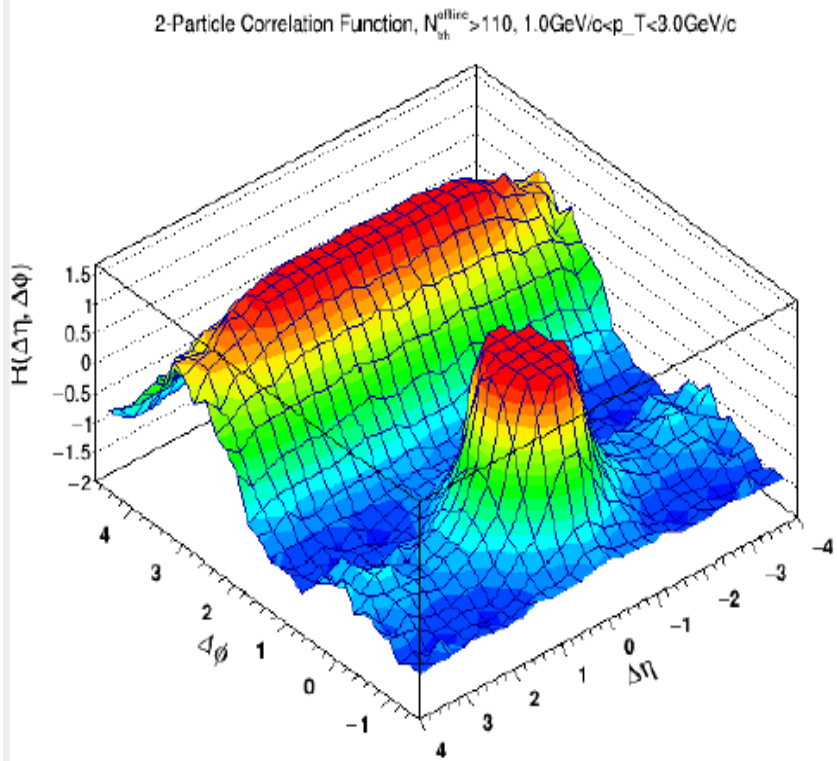
CMS Paper

JHEP 1009 (2010) 091



CMS open data

(summer student on office desktop)



Mine data to test new (aspects of) theories

Exposing the QCD Splitting Function with CMS Open Data

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

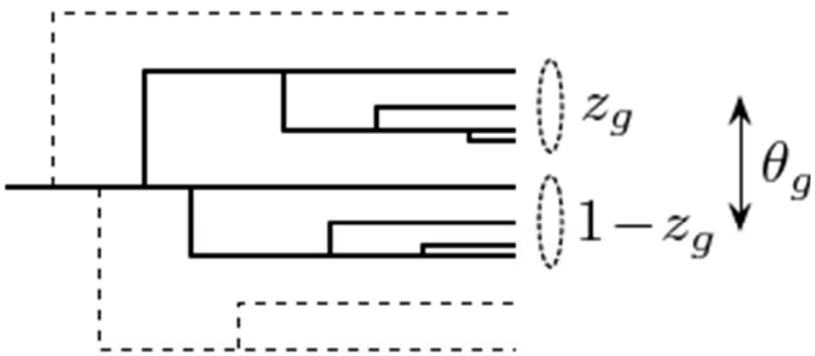
+ some CMS support (S. Rappoccio)

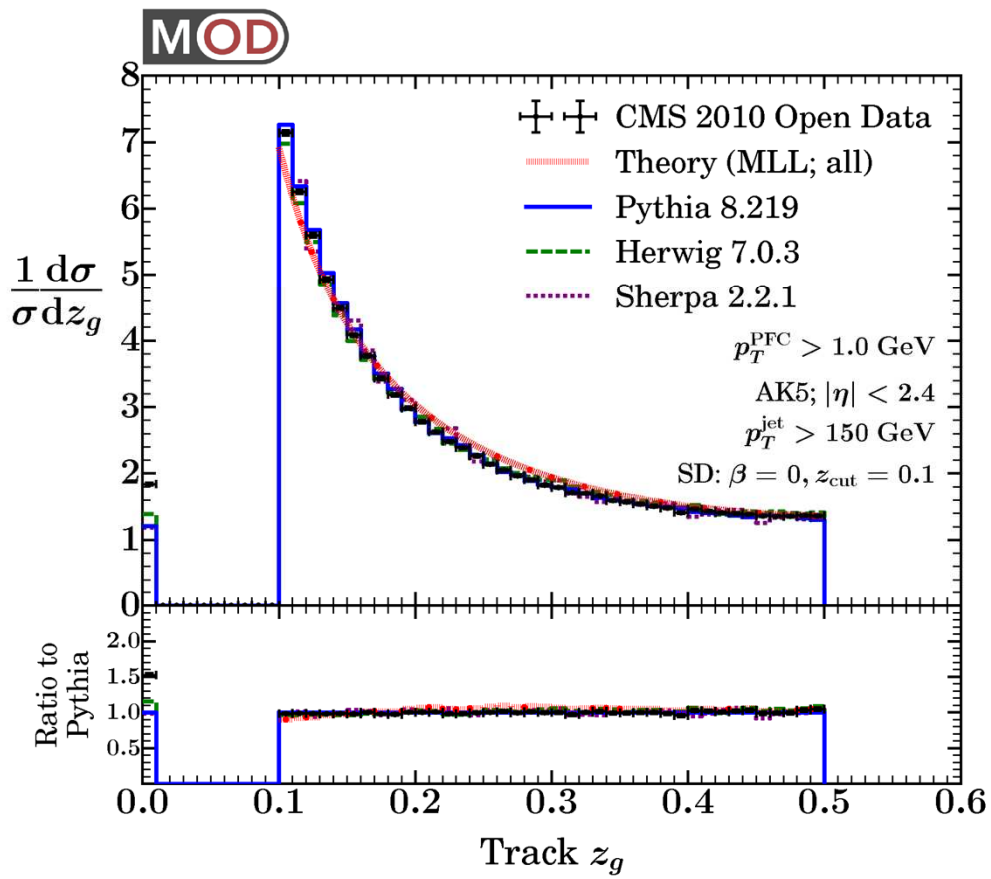
Apr 17, 2017 - 7 pages

MIT-CTP-4891
e-Print: arXiv:1704.05066 [hep-ph] | PDF

(soon) **first ever published CMS Open Data result**

using track (and PF) candidates from CMS 2010 pp jet data





observed jet substructure agrees with predictions from first principles using QCD splitting functions

Mine data to test new (aspects of) theories

Jet Substructure Studies with CMS Open Data

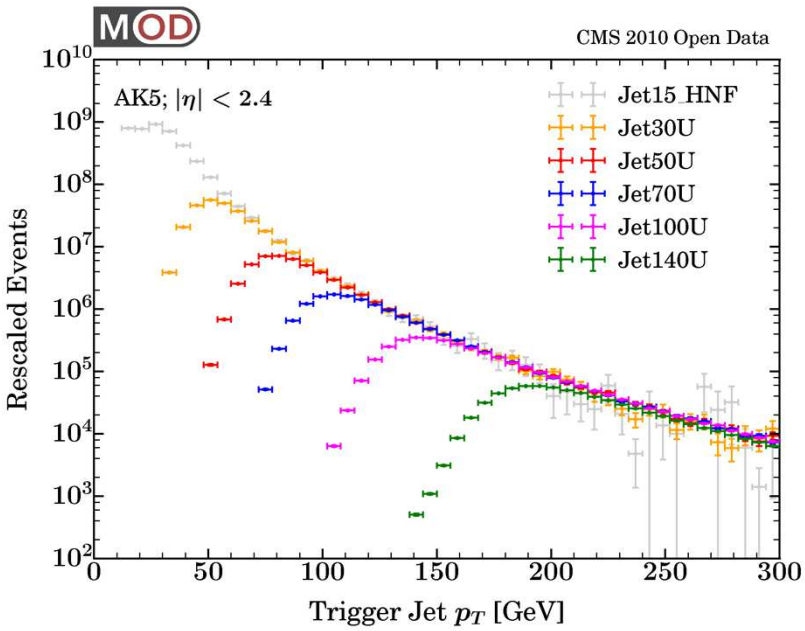
Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, Jesse Thaler + some CMS support (S. Rappoccio)

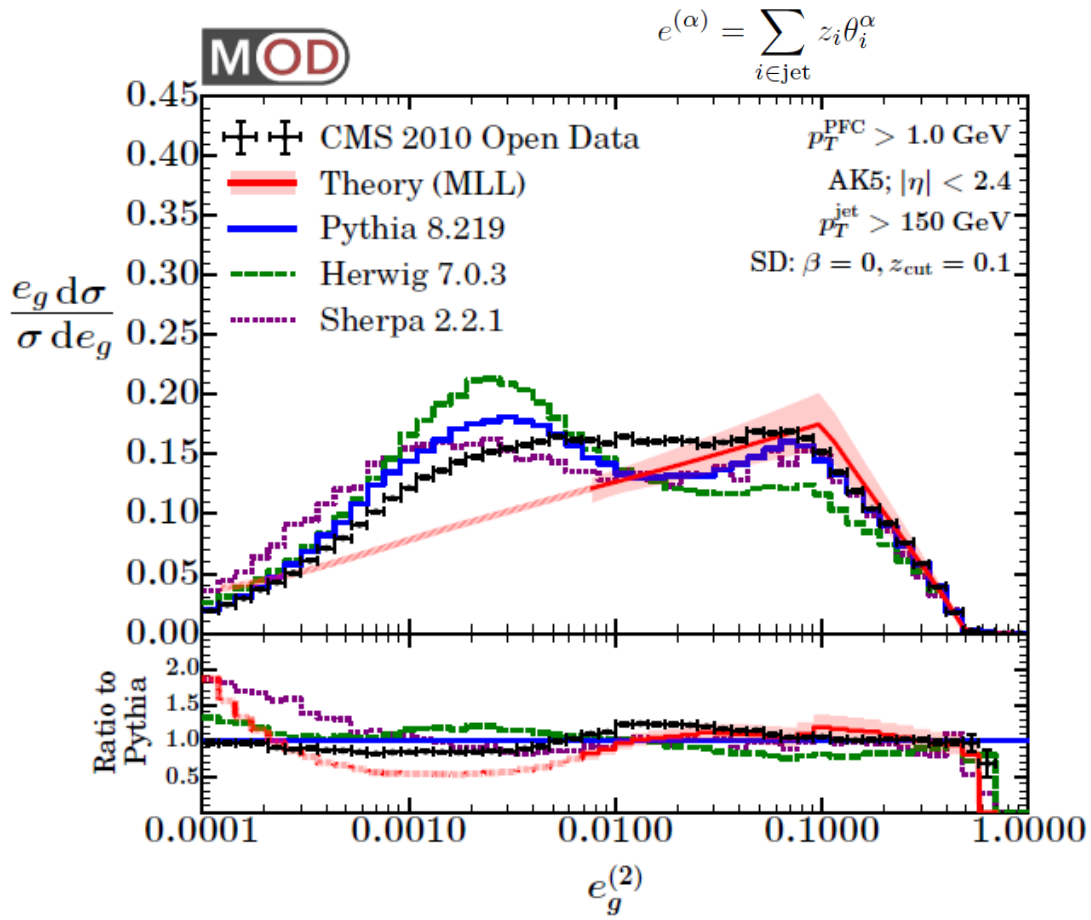
Apr 19, 2017 - 35 pages

MIT-CTP-4890
e-Print: arXiv:1704.05842 [hep-ph] | PDF

combine jet data from different trigger thresholds -> challenge

using particle flow (PF) candidates from CMS 2010 pp jet data





expected kink structure observed,

higher p_T jet measurements needed to disentangle perturbative from nonperturbative effects

Feedback to community

Jet Substructure Studies with CMS Open Data

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

Apr 19, 2017 - 35 pages

MIT-CTP-4890
e-Print: arXiv:1704.05842 [hep-ph] | PDF

Contains section with Advice to community, Challenges, and Recommendations (see there)

Releases of 2011 CMS data+MC "exciting"
-> properly evaluate detector systematics

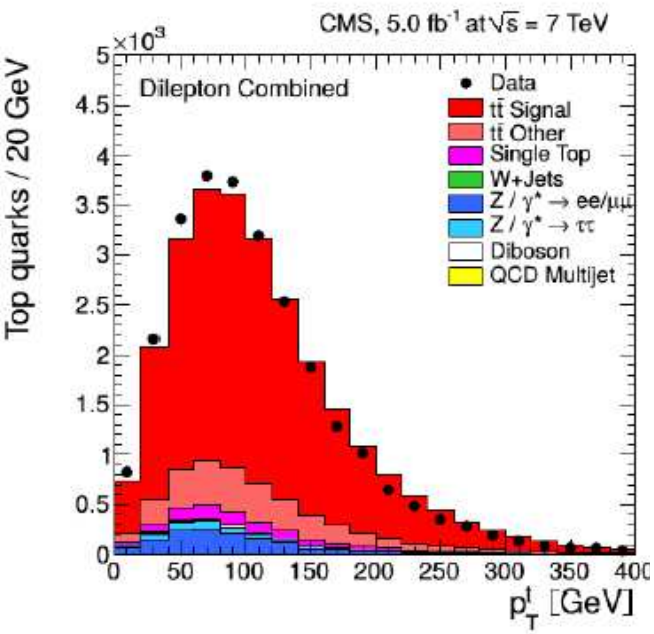
Conclusions: "We hope our experience motivates the LHC collaborations to further their investment in public data release and encourages the particle physics community to exploit the scientific potential of open datasets"

Open data benchmark analysis: top production

use 2011 pp Open Data (2.5 pb^{-1}) + MC,
no usage of advanced CMS tools, simplified acceptance correction

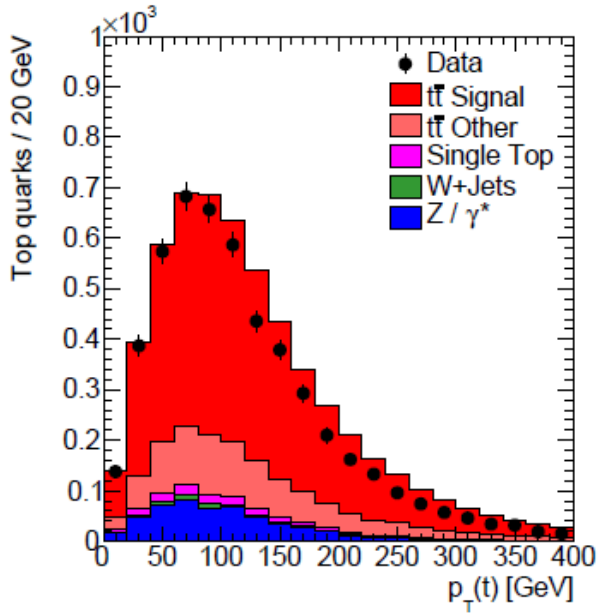
CMS Paper

CMS-TOP-11-013,
EPJ C73 (2013) 2339



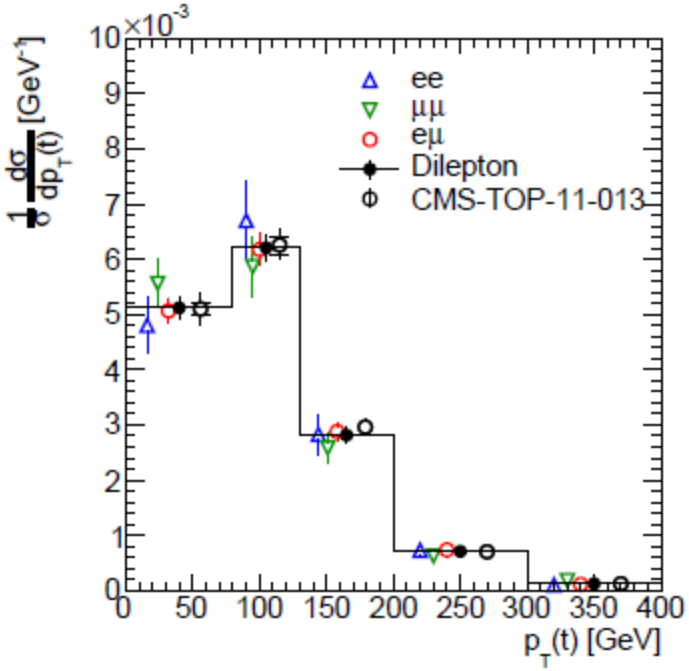
CMS Open Data

(O. Zenaiev, work in progress)



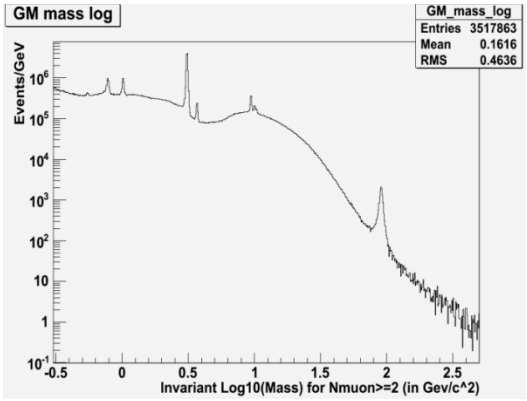
comparison of norm. cross sections

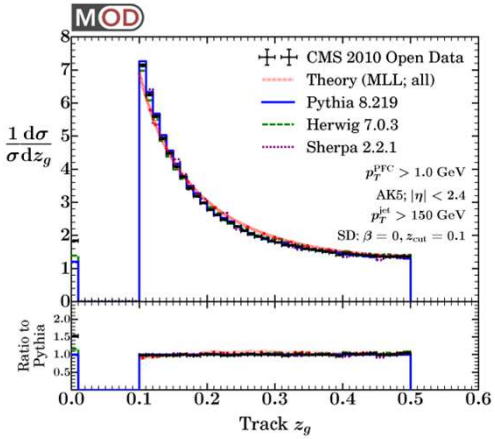
(O. Zenaiev, work in progress)

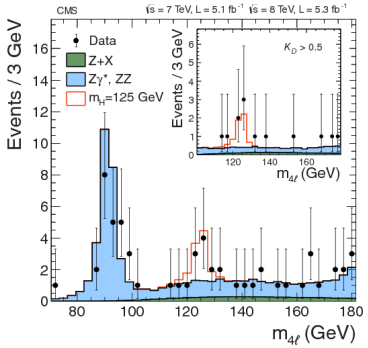


Conclusions

- Open Data release of full CMS 2010 run B data and 2011 run A data + MC available on http://opendata.cern.ch
- well prepared by CERN and CMS IT and open data teams: anybody can use it and it works
- contains benchmark physics analysis and validation examples
- involves nontrivial challenges being worked on
- first physics results from 2010 open data just published by group of theorists from MIT
  -> hopefully start of long and fruitful future of full exploitation of High Energy Physics data beyond actual collaborations
- also used for machine learning (Yandex, see backup)
- upcoming 2012 data release: on the way towards public reconstruction of the Higgs discovery







Backup

Extended Vision (for discussion)

My personal extension of initial vision:

(for discussion, not a collaboration statement)

With ~1% of additional resources, aim to achieve

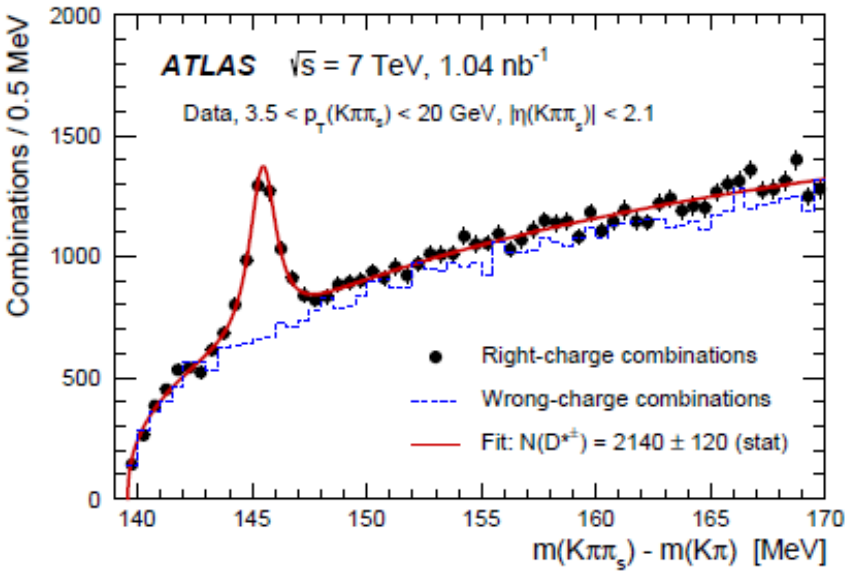
~10% additional scientific output (physics papers)

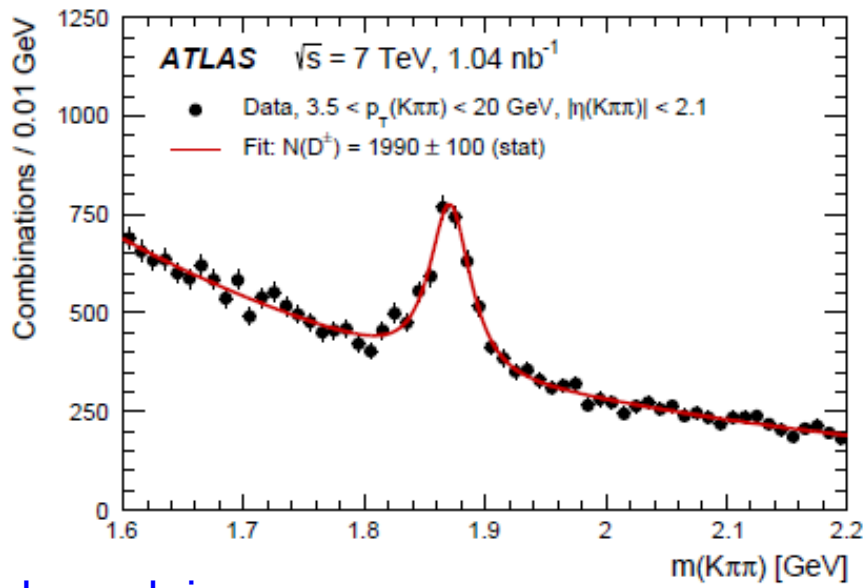
from both external and internal use of preserved data

over lifetime of experiment + 10-20 years

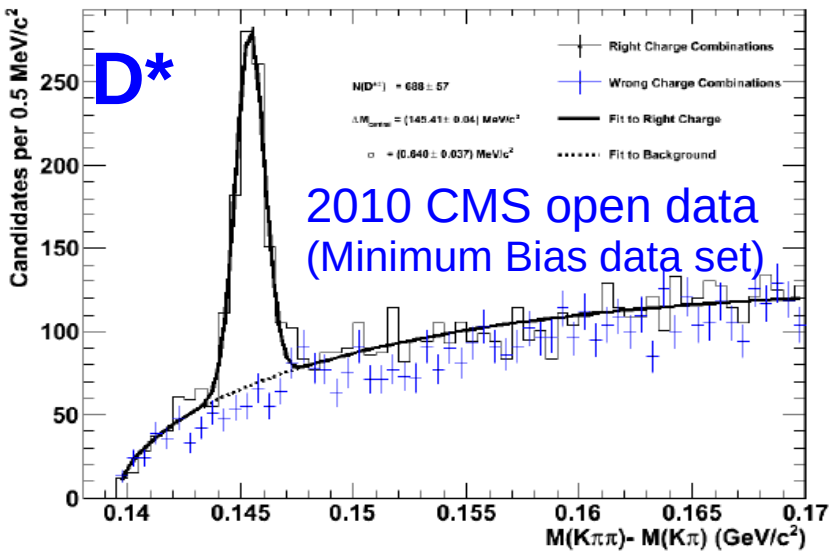
“Recreate” ATLAS result from CMS data: Low p_T D* production (new for CMS)

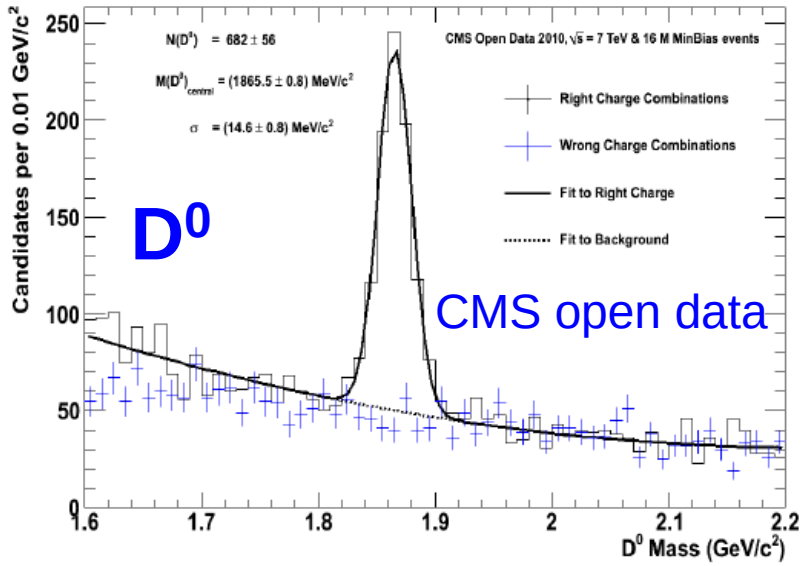
Nucl. Phys. B907 (2016) 717





B. Sheeran, N. Stefaniuk, work in progress





Machine Learning with CMS Open Data: Yandex

From seminar at DESY, 14.2.17

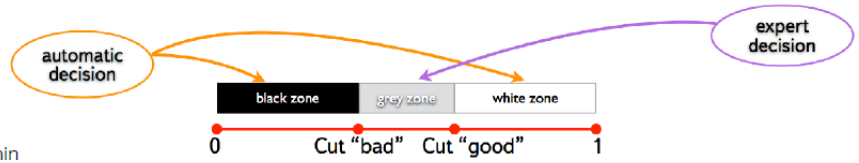
Problem 1: Data Certification (CMS)

- Traditionally, quality of the data at CERN CMS experiment is determined manually which requires considerable amount of human efforts;
- ML can save some of those efforts;
- Data: CMS 2010B run open data;
- Aim: automated classification of Lumisections as "good" or "bad";
- Features: particle flow jets, Calorimeter Jets, Photons, Muons;
- The dataset was flagged by experts (3 FTE).

$$\text{Rejection Rate} = \frac{\text{Rejected}}{\text{Total quantity of samples}} \to \min;$$

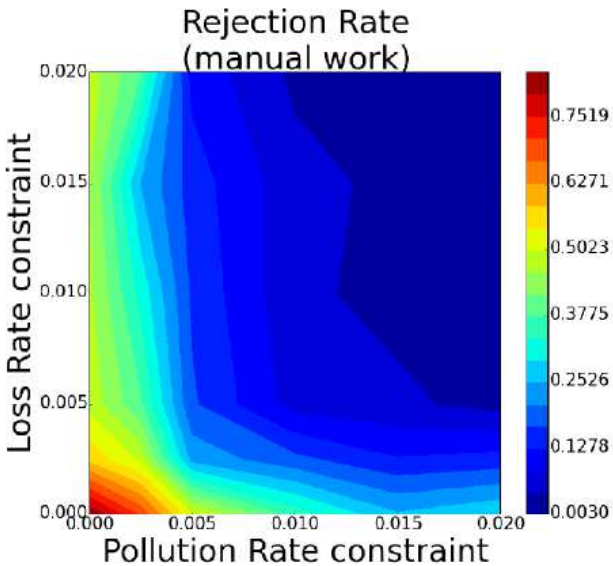
$$\text{Pollution Rate} = \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}} \leq \text{const};$$

$$\text{Loss Rate} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}} \leq \text{const.}$$



Andrey Ustyuzhanin 12

Results



- The aim is to minimise the Manual work with low Loss Rate ("good" classified as "bad") and Pollution Rate ("bad" classified as "good");
- ~80% saving on manual work is feasible for Pollution & Loss rate of 0.5%.
- Next steps: adopt technique for 2016 data & run in production

http://bit.ly/2l0MLiN

Andrey Ustyuzhanin 13