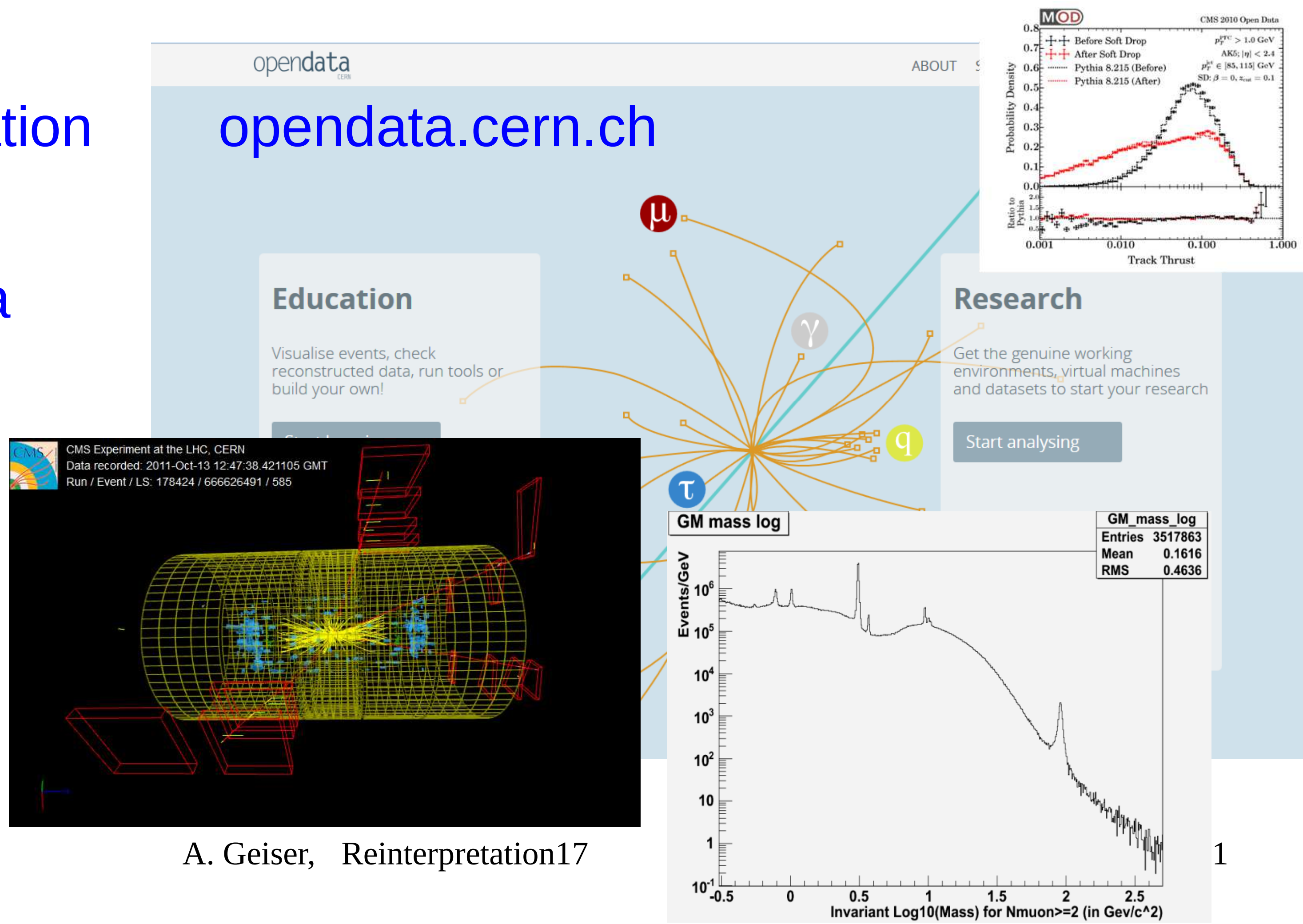
CMS Open Data in Research

Achim Geiser for the CMS collaboration (Achim.Geiser at desy.de)

Reinterpretation17 workshop, Fermilab, Batavia, IL, USA, 17.10.2017

- The vision
- The implementation
- What it is (not)
- CMS Open Data for Research
- Status, results and prospects
- (slide) Tutorial
- Conclusions



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

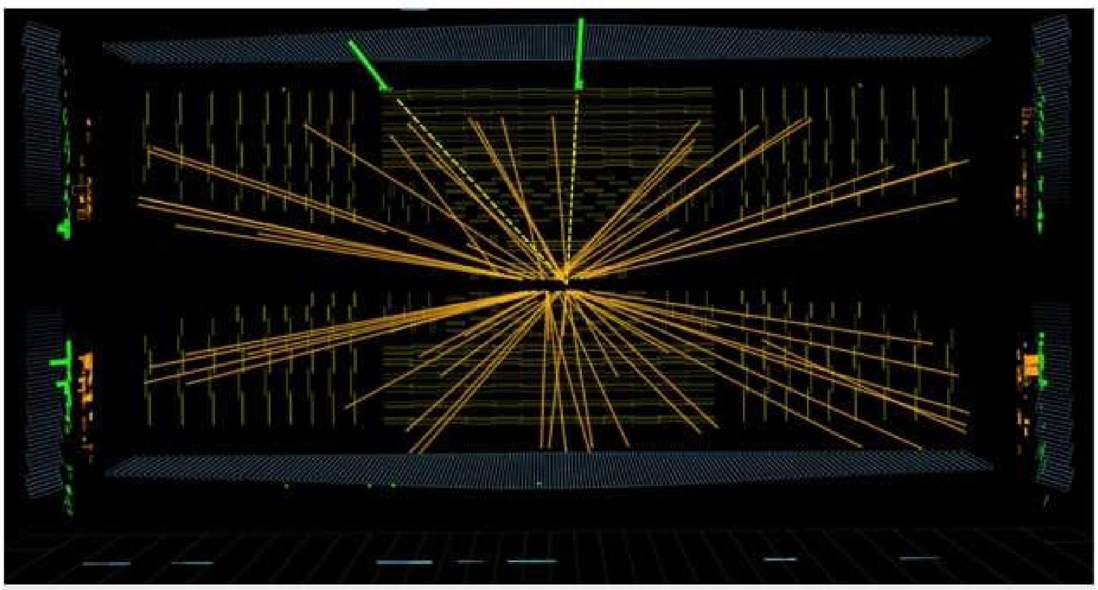
LHC plans for open data future

Researchers share results to keep them accessible.

Elizabeth Gibney

26 November 2013

PDF | Rights & Permissions



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.

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

The Vision

- **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

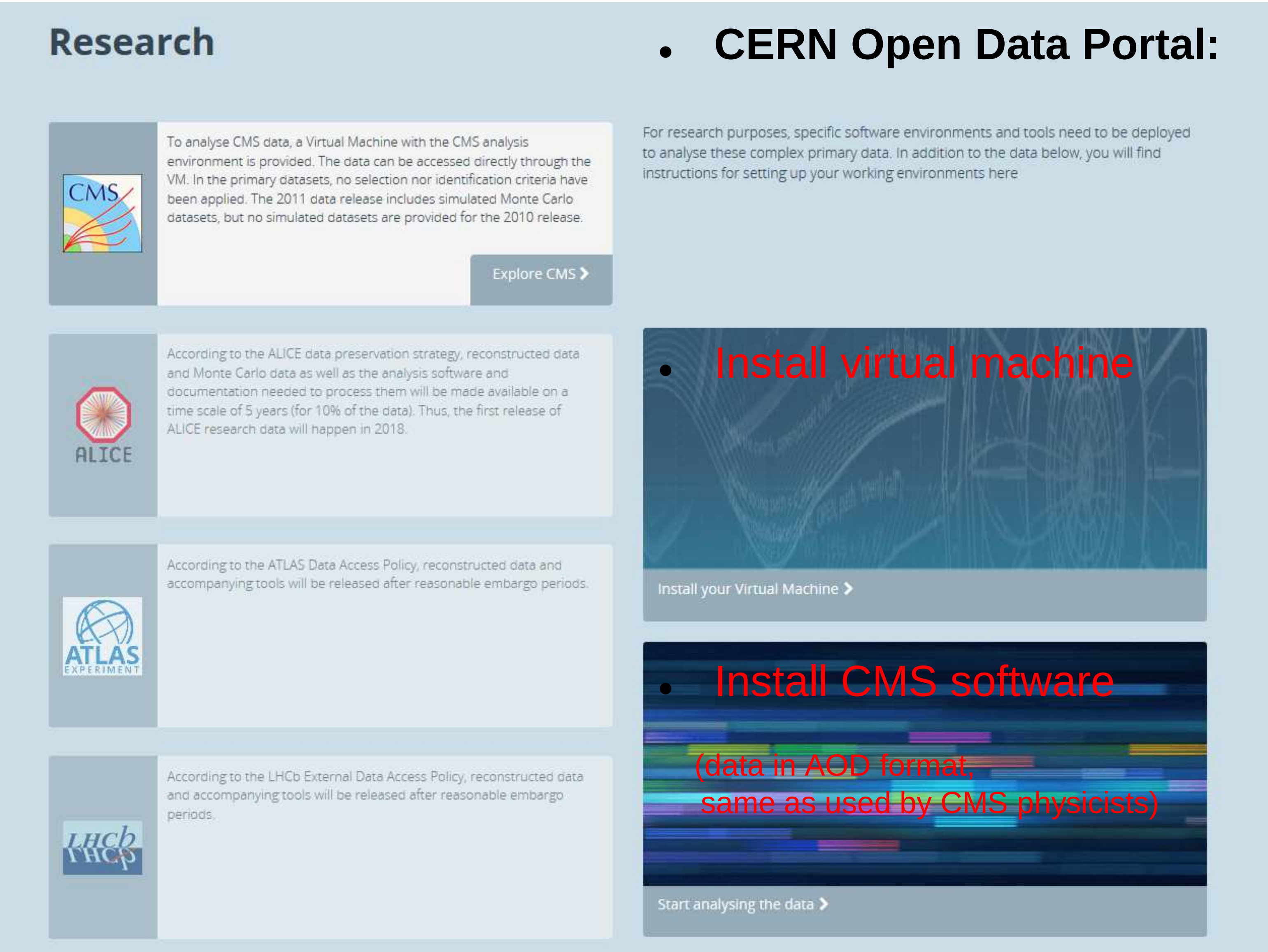
this talk:

focus on Research applications

(many educational applications available from all four experiments)

The implementation: Data and knowledge

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



- Install virtual machine
- Install CMS software

(data in AOD format, same as used by CMS physicists)

~15 min to set up

so far:

only CMS released Research level data

-> pioneer

What it is not: (in the context of this workshop)

- **not a tool to browse existing published CMS results**

  -> use e.g. INSpire, arXiv, …

- **not a tool to (re)interpret published results by comparing with theory**

  -> use e.g. HEPdata, Rivet, …

- **not a toolbox to recast published results into a different form**

  -> use recasting tools (see preceding and later contributions)

- **not dedicated to BSM applications (scope is general, so far dominated by SM applications, but BSM use possible and encouraged)**

What it is: (in the context of this workshop)

- **a setup to do whatever a CMS member did, could have done or could still do with the CMS data, without any formal constraint for non-CMS members**
- **e.g. frequent theorist complaint/request:**

paper X does not present the results in the way I need them for my purposes, recasting is not possible for reason Y, could you please change the results? (or the way they are presented)

- **alternative solution: stop complaining, use Open Data and change them yourself !**

-> (approximately) reproduce the results, or produce new ones
-> modify whatever you want to modify
-> compare to your favorite hypothesis

real published example: next talk by J. Thaler

drawback:

- **can only be done on already released datasets (embargo period 3-4 years)**
- **will probably need a similar effort as if a CMS person or group would have done it (no magic)**

Information about CMS Open Data

- **CERN Open Data Portal: http://opendata.cern.ch/about/CMS**
  (see also https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSPublicData)
- **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: **AOD format** (CMS root)

talk J. Thaler

- 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 useful and usable**
- **information is missing for an analysis to be completed ?**

  -> we are more than happy to take the feedback at opendata.support@cern.ch **and provide it** (as long as we have it available ourselves)
- e.g. luminosity values for collision data recently added; cross sections for MC being added
- **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 have plenty of good will but are 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 (**open source**!). No grid jobs, no batch jobs on farm, no CMS account.

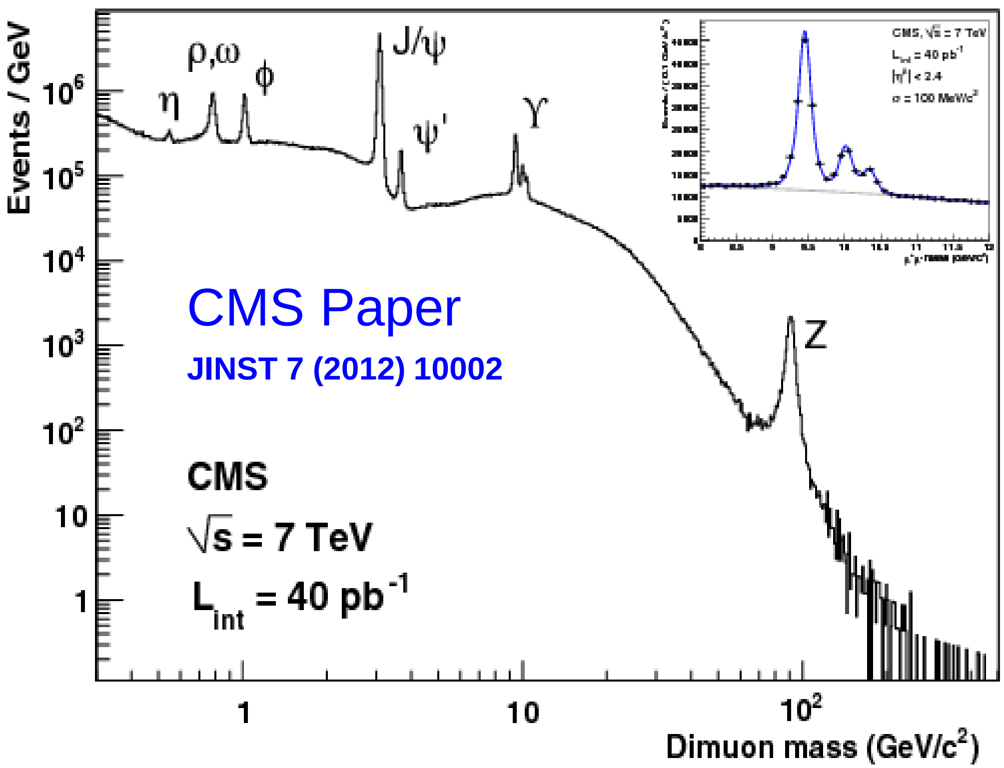
many obtained by **undergraduate students** supervised by experienced physicists -> **excellent training opportunities**!

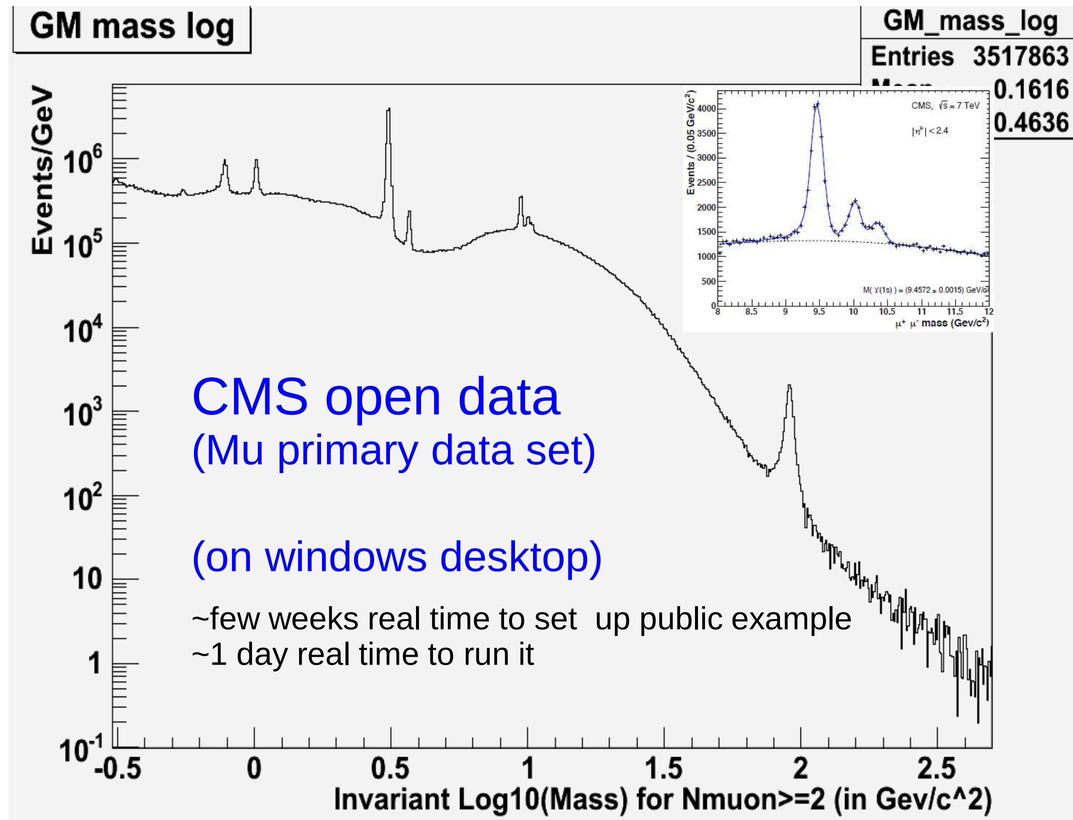
Provide references: validation/benchmarking/analysis examples

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”

on portal soon

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

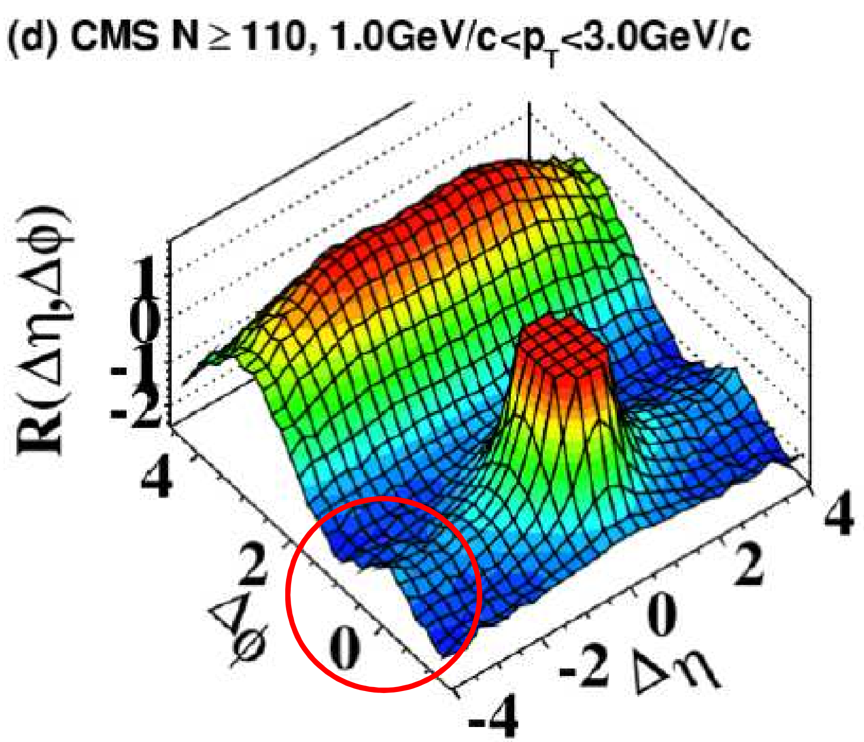
CMS Paper

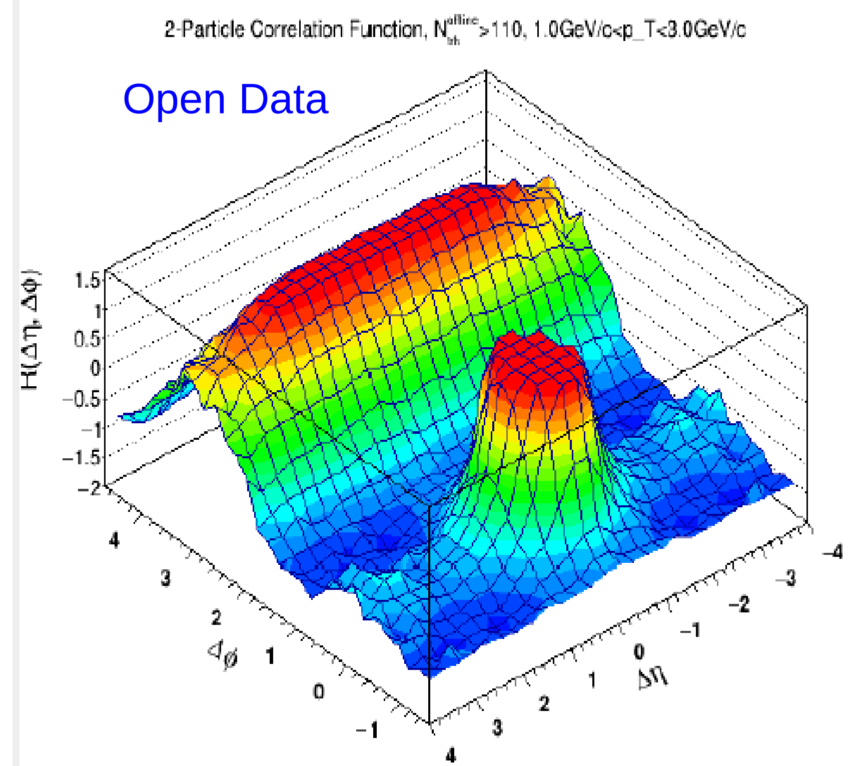
JHEP 1009 (2010) 091

CMS open data

(summer student on office desktop)

~few weeks real time





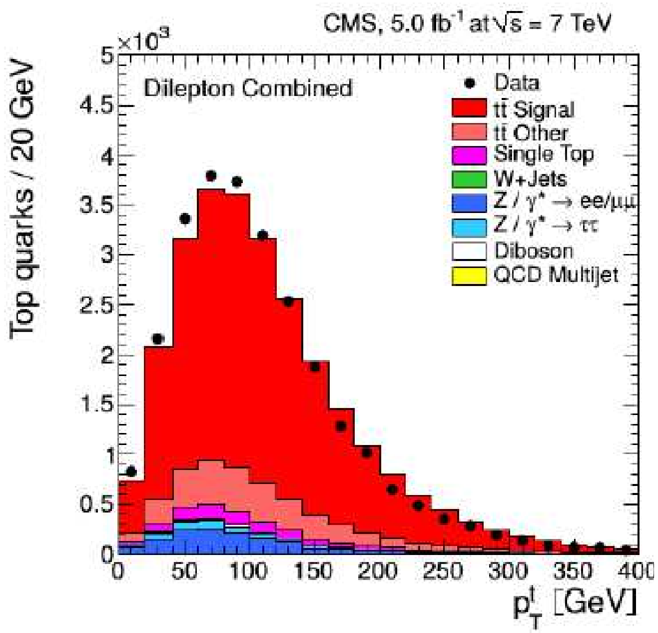
Open Data benchmark analysis: top production

on portal soon, preview on https://github.com/cms-opendata-validation

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

CMS Paper

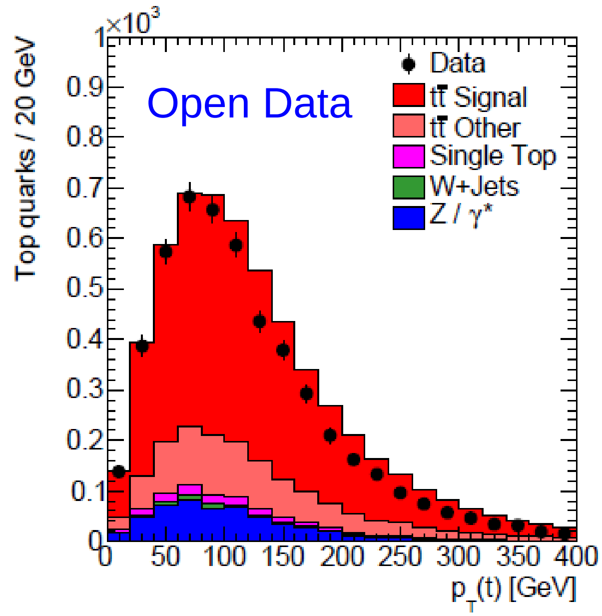
CMS-TOP-11-013,
EPJ C73 (2013) 2339



CMS Open Data

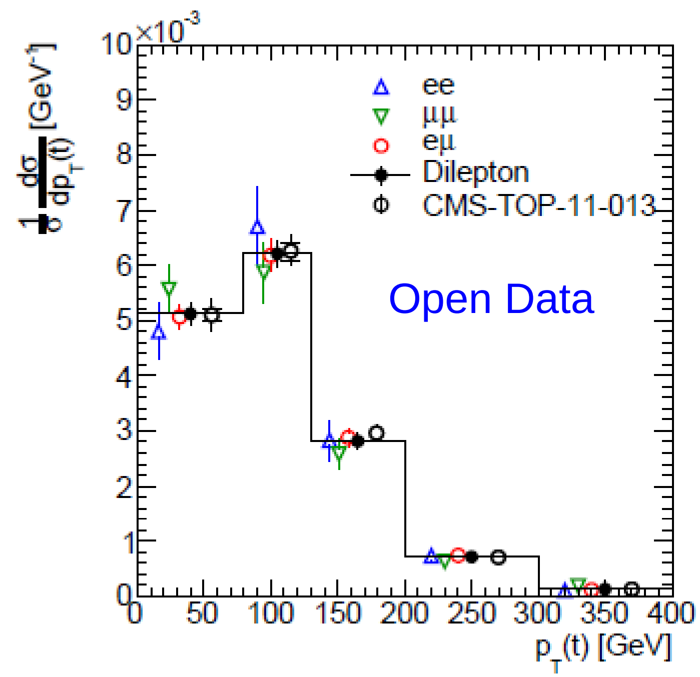
(O. Zenaiev)

~two months



comparison of norm. cross sections

(O. Zenaiev)



Machine Learning with CMS Open Data: Yandex

From seminar at DESY, 14.2.17
A. Ustyuzhanin

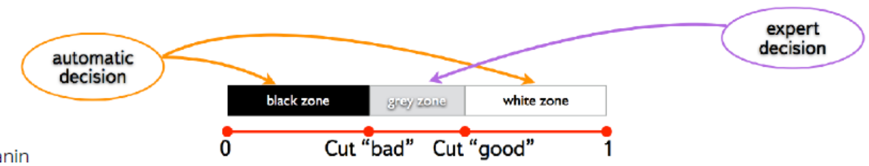
Problem 1: Data Certification (CMS)

- 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).

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

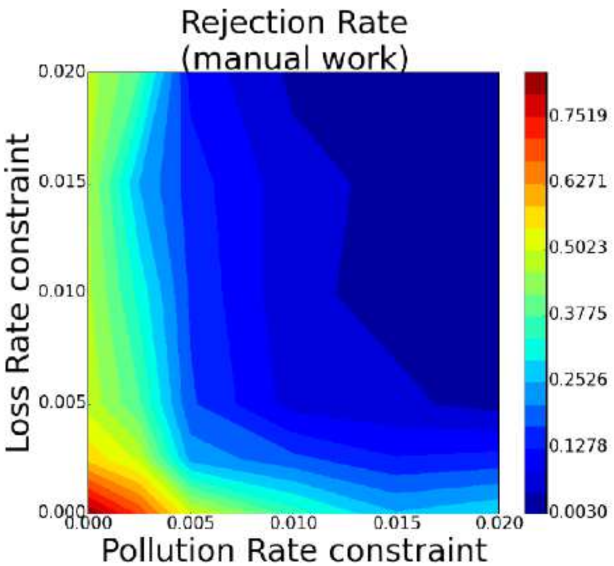
$$\text{Pollution Rate} = \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}} \leq \text{const};$$

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



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

Mine data to test new (aspects of) theories

Exposing the QCD Splitting Function with CMS Open Data

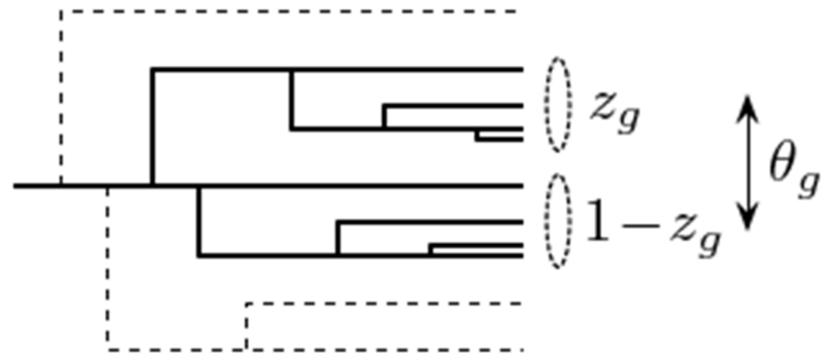
Andrew Larkoski, Simone Marzani, Jesse Thaler, Aashish Tripathee, Wei Xue

+ some CMS support (S. Rappoccio)

Phys Rev Lett 119 (2017) 132003

Apr 17, 2017 - 7 pages

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



first ever published
CMS Open Data results

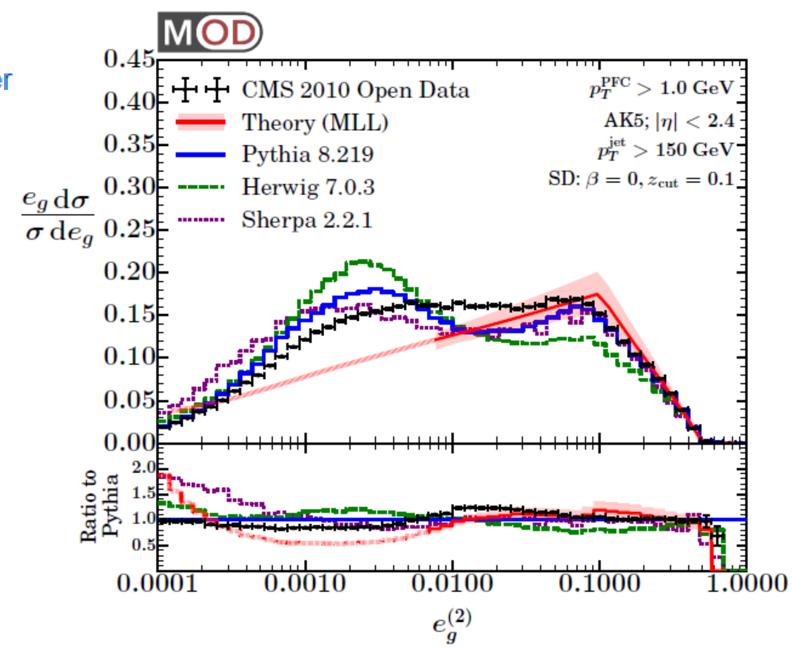
Jet Substructure Studies with CMS Open Data

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

Phys Rev D96 (2017) 074003

Apr 19, 2017 - 35 pages

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



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

-> next talk

Mine data to test new (aspects of) theories

thanks to F. Navarro, E. Carrara

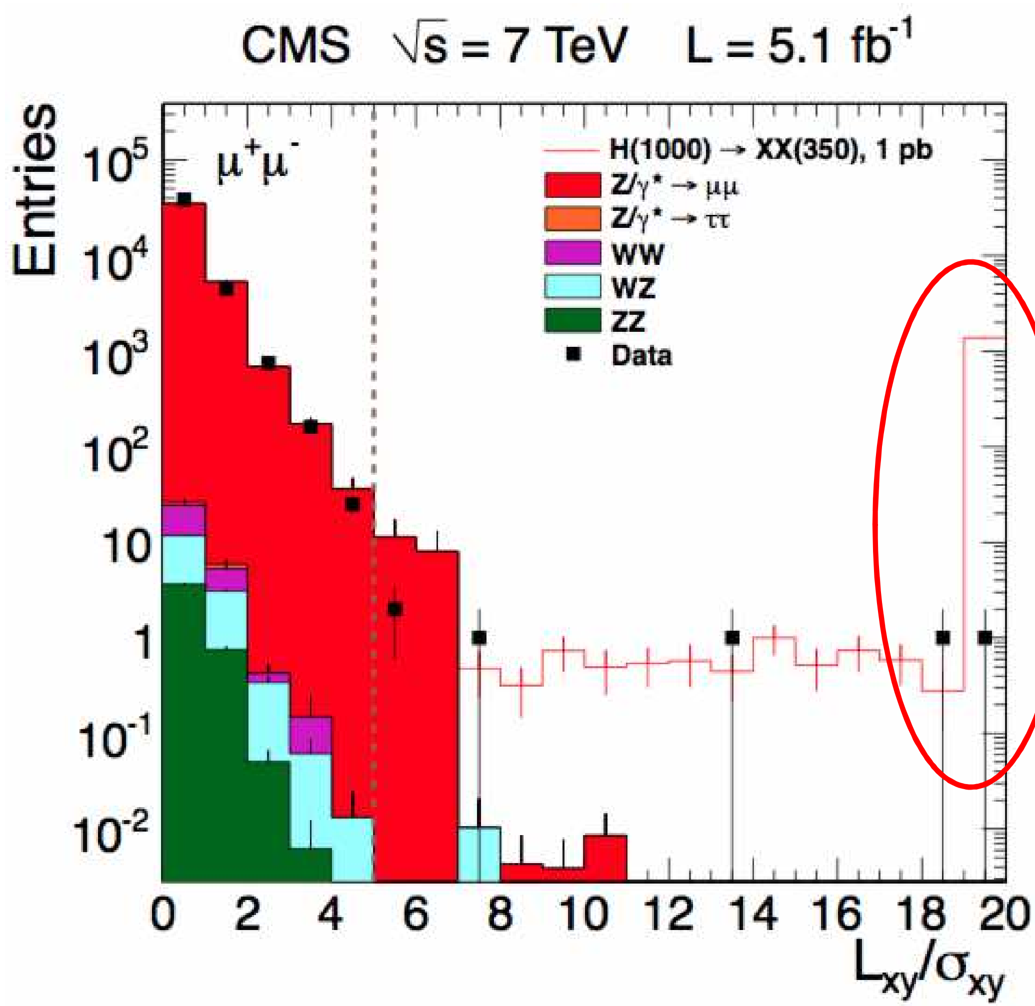
Open Data analysis example in preparation:
Search in leptonic channels for heavy resonances decaying to long-lived neutral particles

JHEP 1302 (2013) 085
CMS-EXO-11-101

Theory: “Hidden Valley” Search

in practice: Search for leptons with “long-distance” displacement from primary vertex, originating from decay of X particles

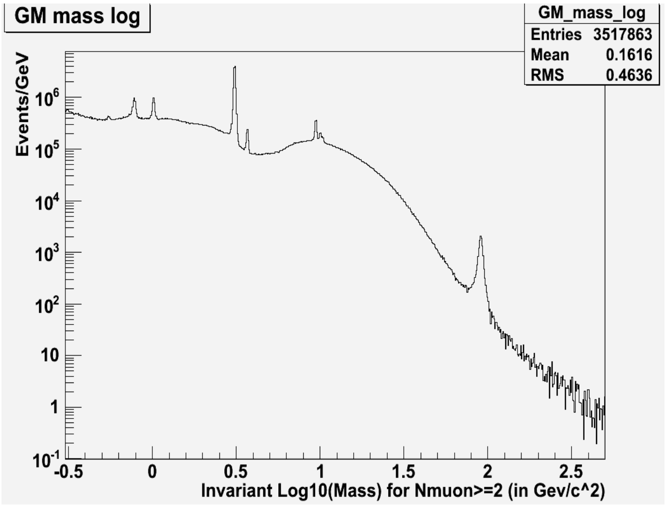
-> get limits e.g. on H->XX (details see paper)



some practical aspects below

Open Data Tutorial

Slides, not online (would take too long), but written such that it can be tried immediately

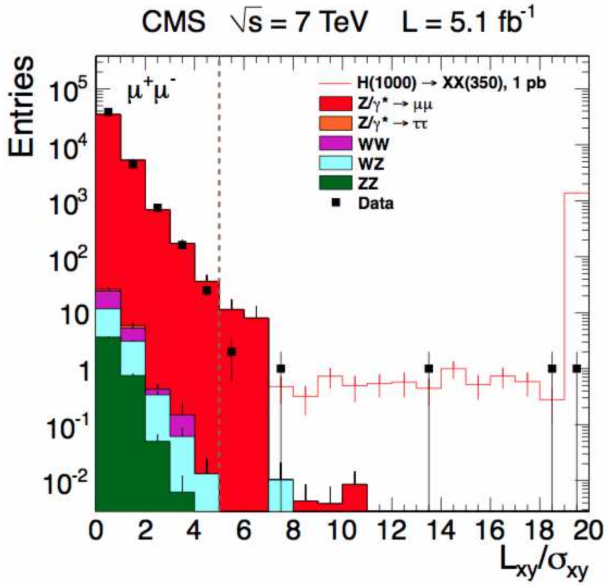


only prerequisite: know a bit of Linux and ROOT

Focus on:

Dimuon mass spectrum

(simple, exists, works)



Displaced Lepton Search

(conceptual, in preparation,

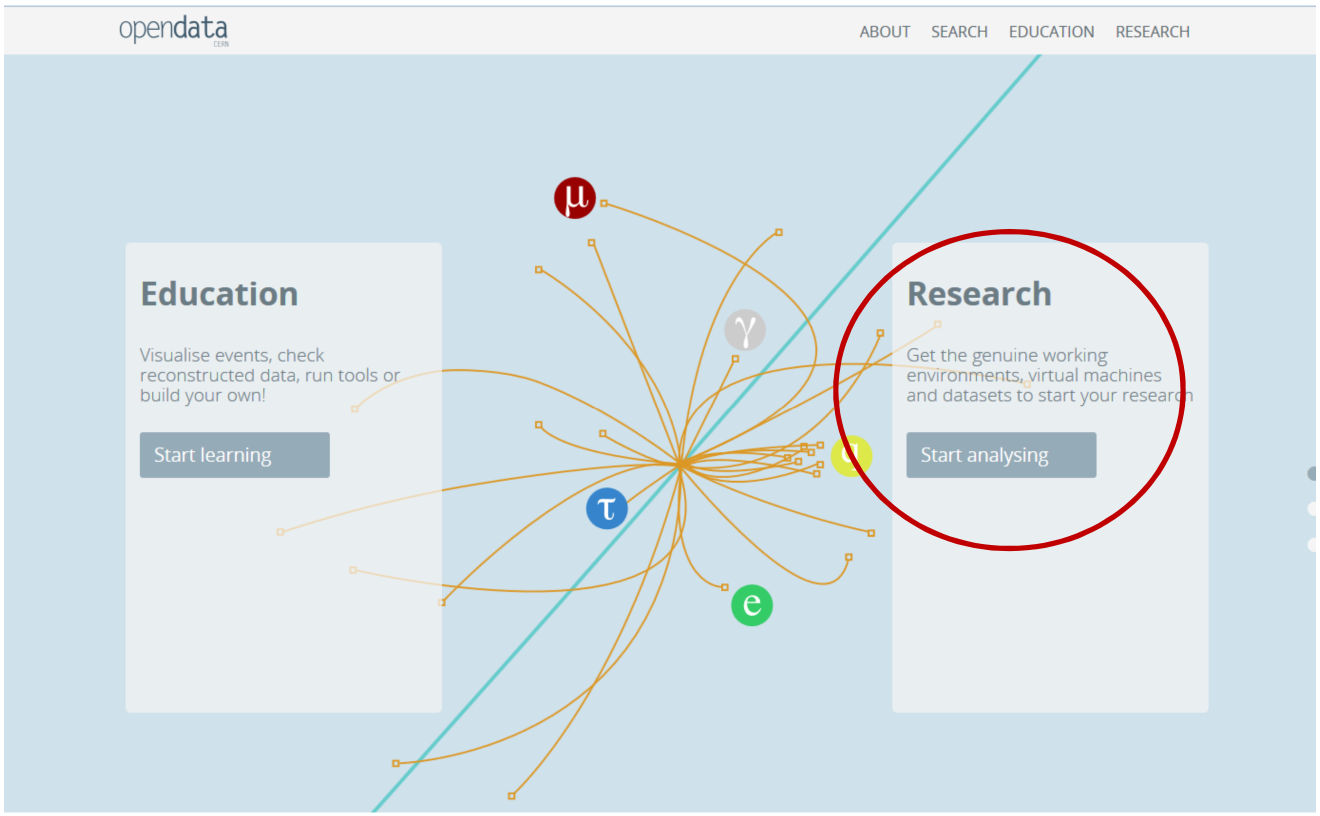
expose challenges)

Open Data Tutorial

start your favourite laptop, desktop, …   windows, linux or MacOS
(at least **2 GB memory**, administrator rights or VirtualBox preinstalled)

Portal appearance might change soon. Content will stay/be extended.

with any web browser: opendata.cern.ch
(see also https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSPublicData)



choose Research

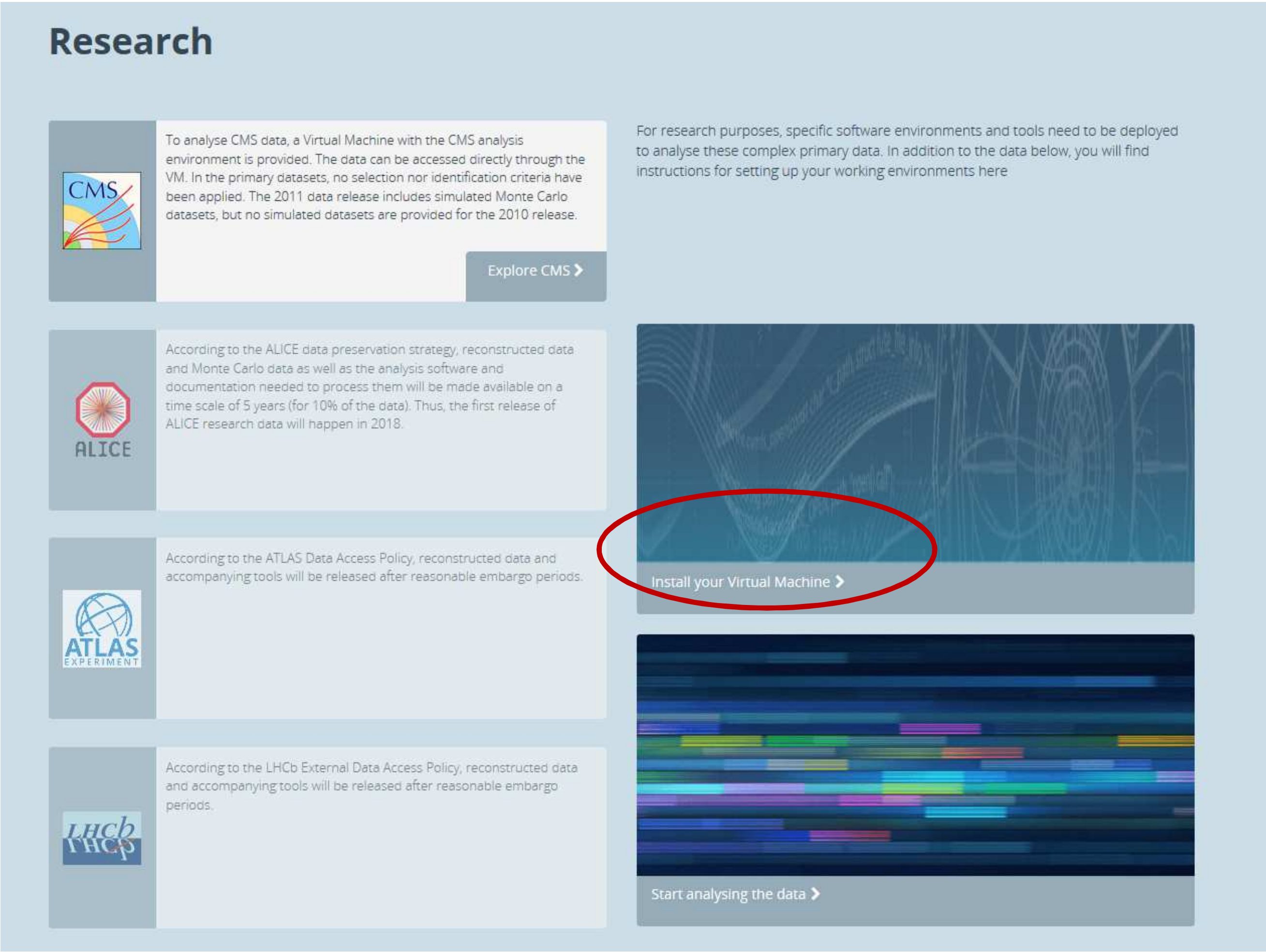
(“Start analysing”)

side remarks:

VM is faster on windows than on linux!

Tutorial will work (almost) anywhere
(except on the Fermilab guest network ..)

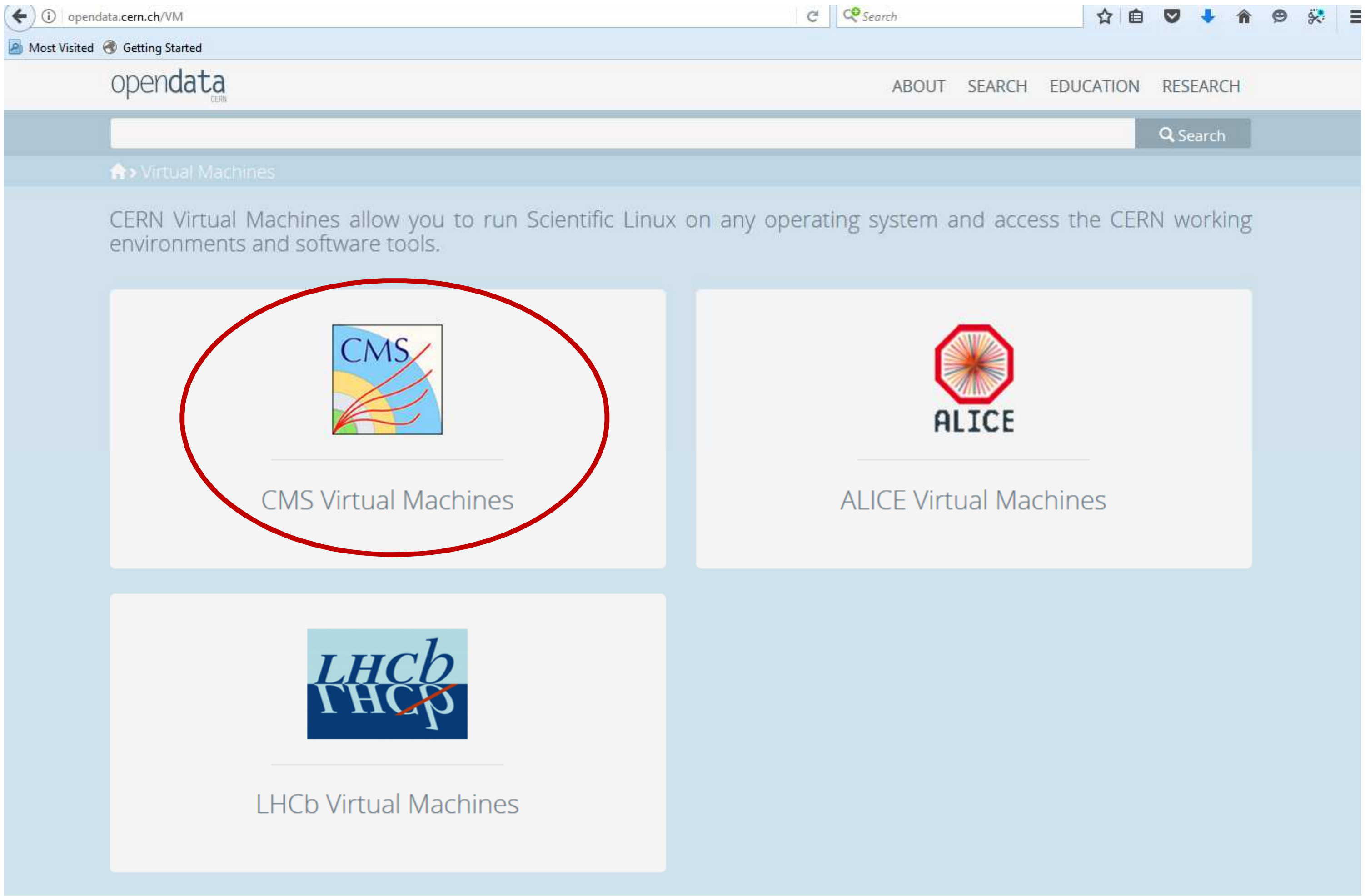
Open Data Tutorial



choose

“Install your Virtual Machine”

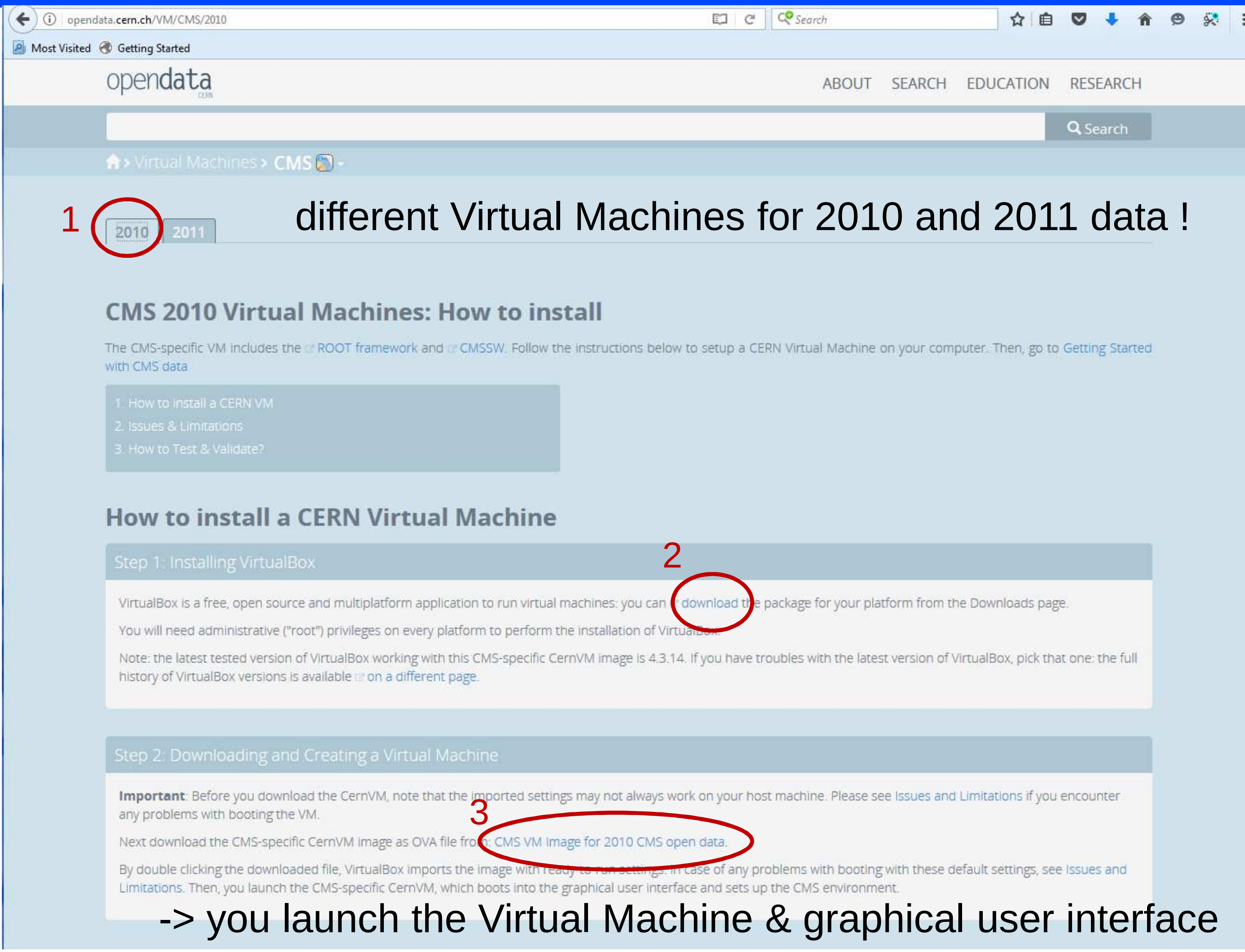
Open Data Tutorial



choose

“CMS Virtual Machines”

Open Data Tutorial



1 choose “2010”

2 if not yet done:

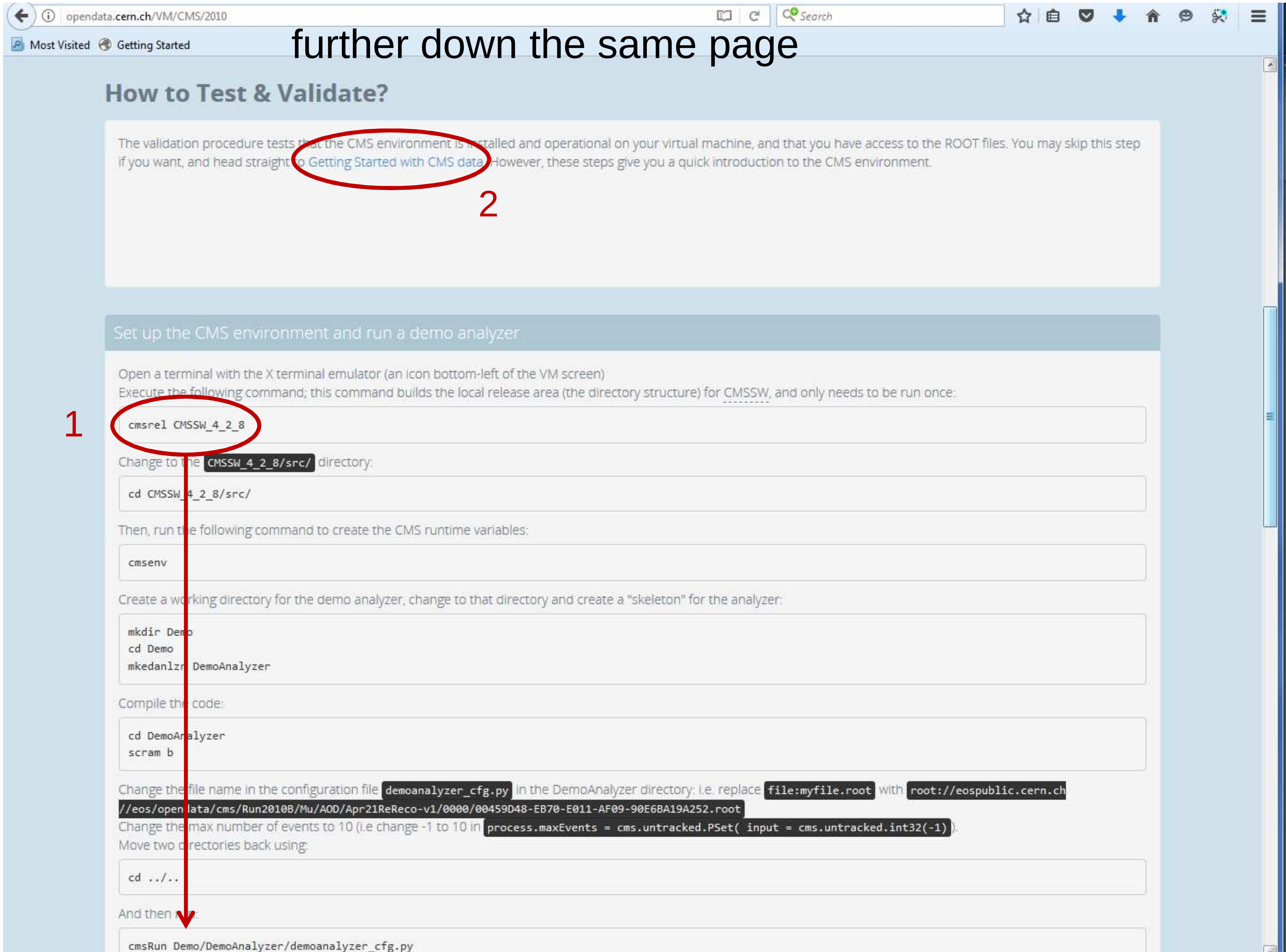
download /install VirtualBox

(see FAQ at bottom of page)

3 download CMS VM Image

& double click it

Open Data Tutorial



Test & Validate:

(do not skip)

1

install relevant CMS software

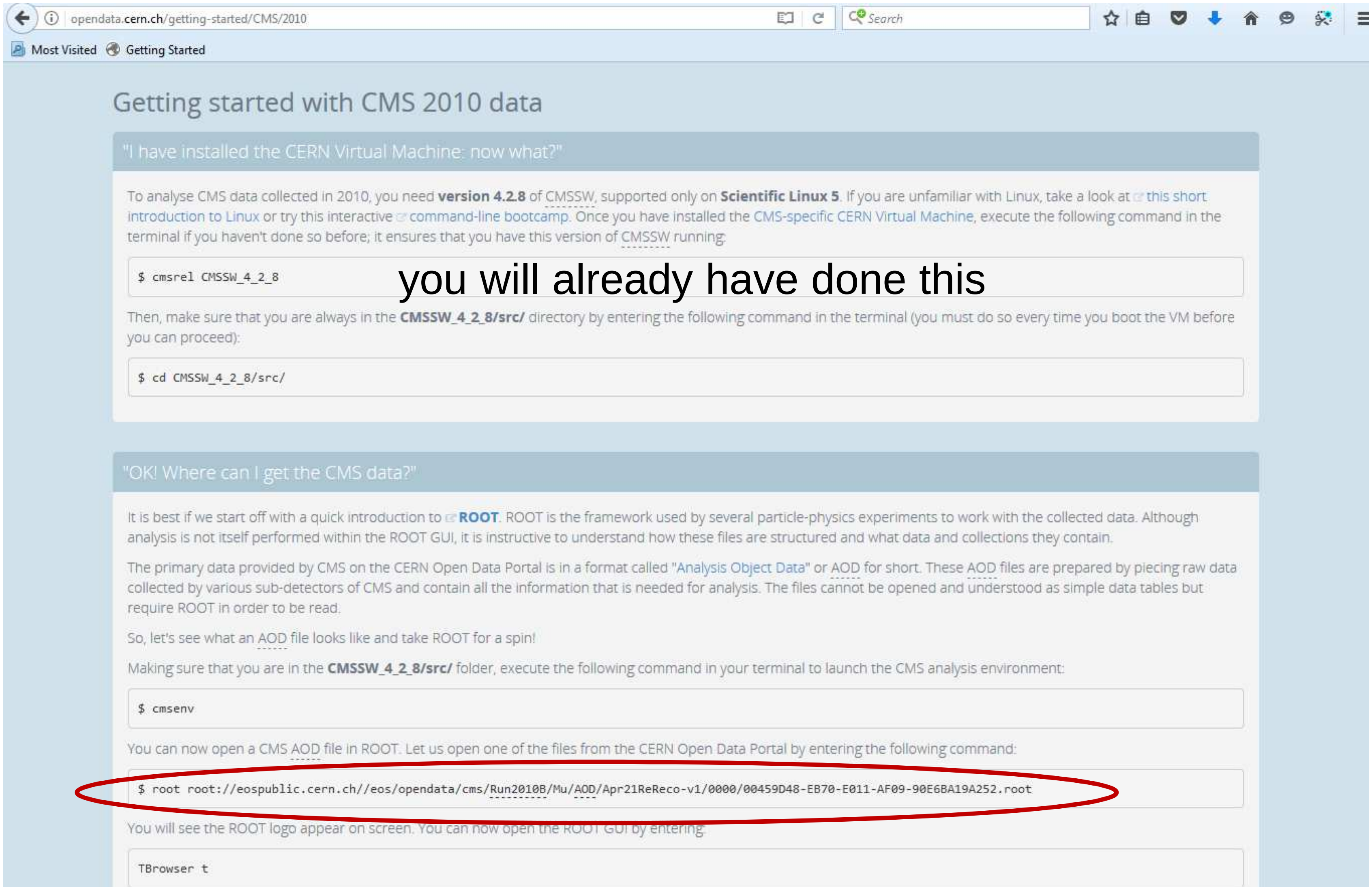
(CMSSW_4_2_8 for 2010 data)

follow instructions

-> you will have technically run your first CMS job ☺

-> continue with **2** “Getting started with CMS data”

Open Data Tutorial



opendata.cern.ch/getting-started/CMS/2010

Most Visited   Getting Started

Getting started with CMS 2010 data

"I have installed the CERN Virtual Machine: now what?"

To analyse CMS data collected in 2010, you need **version 4.2.8** of CMSSW, supported only on **Scientific Linux 5**. If you are unfamiliar with Linux, take a look at this short introduction to Linux or try this interactive command-line bootcamp. Once you have installed the CMS-specific CERN Virtual Machine, execute the following command in the terminal if you haven't done so before; it ensures that you have this version of CMSSW running:

```
$ cmsrel CMSSW_4_2_8
```

you will already have done this

Then, make sure that you are always in the **CMSSW_4_2_8/src/** directory by entering the following command in the terminal (you must do so every time you boot the VM before you can proceed):

```
$ cd CMSSW_4_2_8/src/
```

"OK! Where can I get the CMS data?"

It is best if we start off with a quick introduction to **ROOT**. ROOT is the framework used by several particle-physics experiments to work with the collected data. Although analysis is not itself performed within the ROOT GUI, it is instructive to understand how these files are structured and what data and collections they contain.

The primary data provided by CMS on the CERN Open Data Portal is in a format called "Analysis Object Data" or AOD for short. These AOD files are prepared by piecing raw data collected by various sub-detectors of CMS and contain all the information that is needed for analysis. The files cannot be opened and understood as simple data tables but require ROOT in order to be read.

So, let's see what an AOD file looks like and take ROOT for a spin!

Making sure that you are in the **CMSSW_4_2_8/src/** folder, execute the following command in your terminal to launch the CMS analysis environment:

```
$ cmsenv
```

You can now open a CMS AOD file in ROOT. Let us open one of the files from the CERN Open Data Portal by entering the following command:

```
$ root root://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00459D48-EB70-E011-AF09-90E6BA19A252.root
```

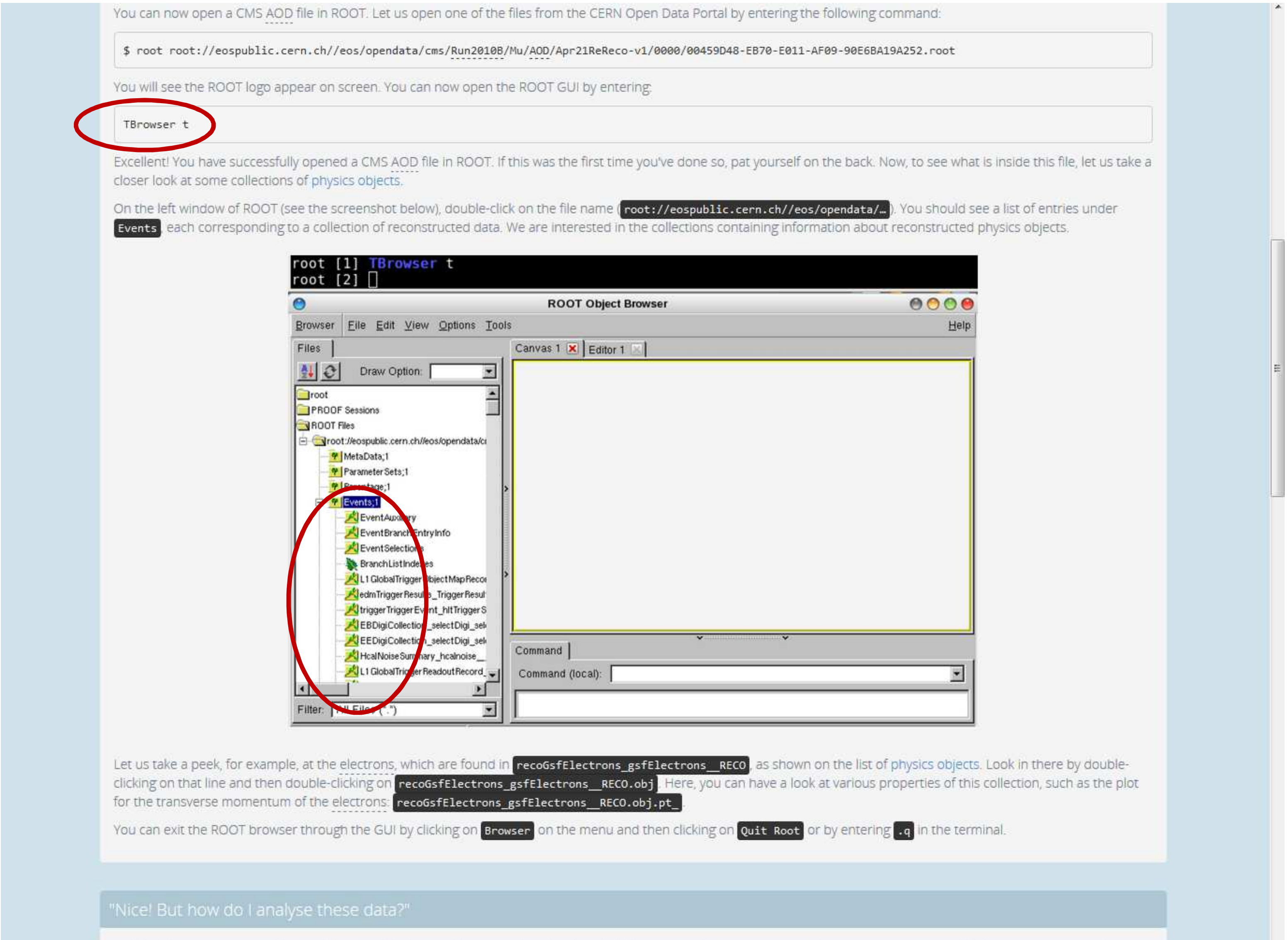
You will see the ROOT logo appear on screen. You can now open the ROOT GUI by entering:

```
TBrowser t
```

manually inspect the content of a CMS AOD ROOT file (located at CERN) in order to get a “feel”

(follow instructions)

Open Data Tutorial



You can now open a CMS AOD file in ROOT. Let us open one of the files from the CERN Open Data Portal by entering the following command:

```
$ root root://eospublic.cern.ch//eos/opendata/cms/Run2010B/Mu/AOD/Apr21ReReco-v1/0000/00459D48-EB70-E011-AF09-90E6BA19A252.root
```

You will see the ROOT logo appear on screen. You can now open the ROOT GUI by entering:

```
TBrowser t
```

Excellent! You have successfully opened a CMS AOD file in ROOT. If this was the first time you've done so, pat yourself on the back. Now, to see what is inside this file, let us take a closer look at some collections of physics objects.

On the left window of ROOT (see the screenshot below), double-click on the file name (`root://eospublic.cern.ch//eos/opendata/…`). You should see a list of entries under `Events` each corresponding to a collection of reconstructed data. We are interested in the collections containing information about reconstructed physics objects.

Let us take a peek, for example, at the electrons, which are found in `recoGsfElectrons_gsfElectrons__RECO`, as shown on the list of physics objects. Look in there by double-clicking on that line and then double-clicking on `recoGsfElectrons_gsfElectrons__RECO.obj`. Here, you can have a look at various properties of this collection, such as the plot for the transverse momentum of the electrons: `recoGsfElectrons_gsfElectrons__RECO.obj.pt_`.

You can exit the ROOT browser through the GUI by clicking on `Browser` on the menu and then clicking on `Quit Root` or by entering `.q` in the terminal.

"Nice! But how do I analyse these data?"

start the ROOT browser

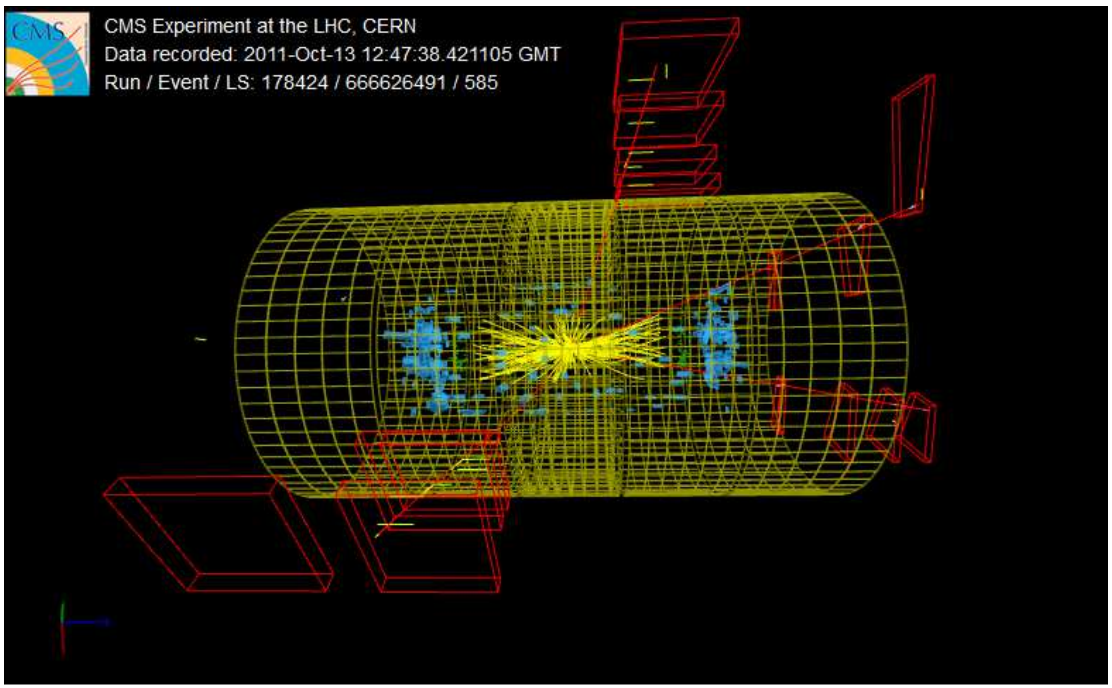
(previous experience with ROOT is helpful)

inspect the variables

Open Data Tutorial

Intermezzo: at this stage (or at any other time), might want to looki at some CMS **event displays** on your standard browser (**no VM needed**).

e.g. real Higgs -> 4 muon candidate



in browser:

go to **education** part

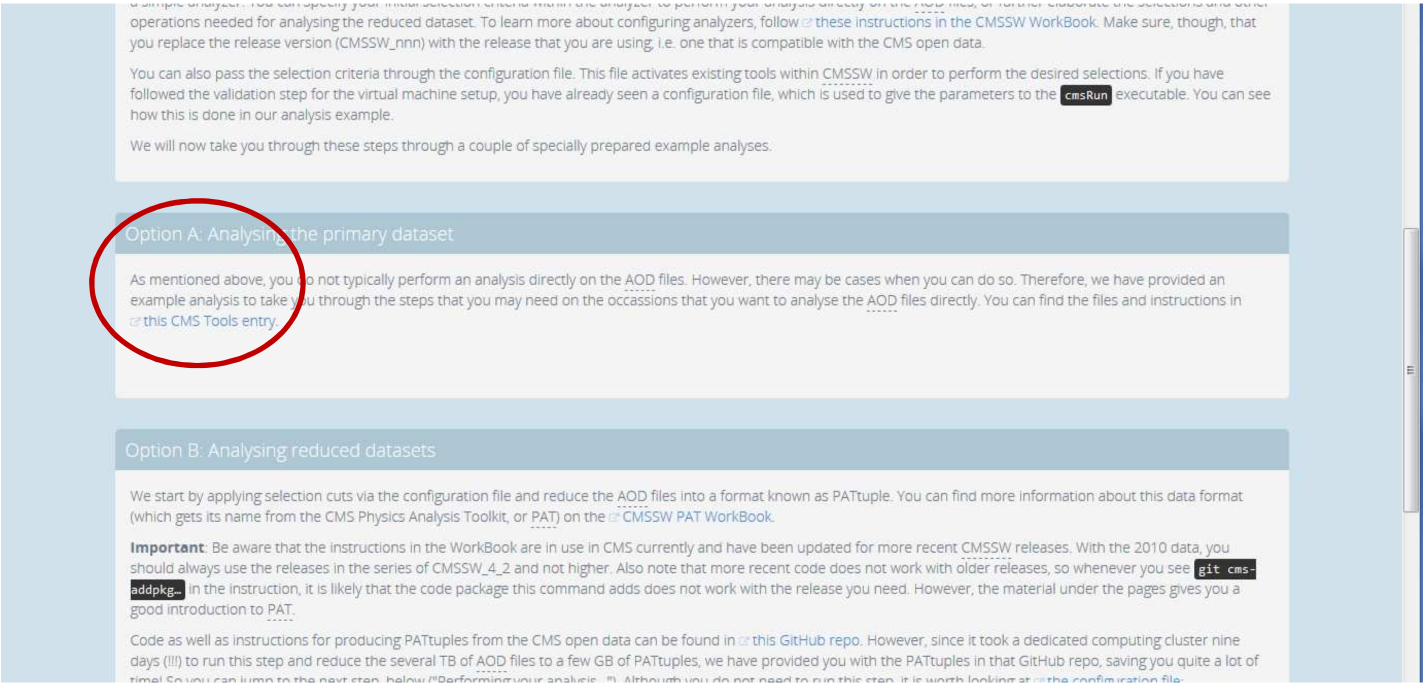
-> “visualize events”

and follow instructions

consider also https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSPublicData

Open Data Tutorial

continue further down the “Getting Started” page



operations needed for analysing the reduced dataset. To learn more about configuring analyzers, follow these instructions in the CMSSW WorkBook. Make sure, though, that you replace the release version (CMSSW_nnn) with the release that you are using, i.e. one that is compatible with the CMS open data.

You can also pass the selection criteria through the configuration file. This file activates existing tools within CMSSW in order to perform the desired selections. If you have followed the validation step for the virtual machine setup, you have already seen a configuration file, which is used to give the parameters to the `cmsRun` executable. You can see how this is done in our analysis example.

We will now take you through these steps through a couple of specially prepared example analyses.

Option A: Analysing the primary dataset

As mentioned above, you do not typically perform an analysis directly on the AOD files. However, there may be cases when you can do so. Therefore, we have provided an example analysis to take you through the steps that you may need on the occassions that you want to analyse the AOD files directly. You can find the files and instructions in this CMS Tools entry.

Option B: Analysing reduced datasets

We start by applying selection cuts via the configuration file and reduce the AOD files into a format known as PATtuple. You can find more information about this data format (which gets its name from the CMS Physics Analysis Toolkit, or PAT) on the CMSSW PAT WorkBook.

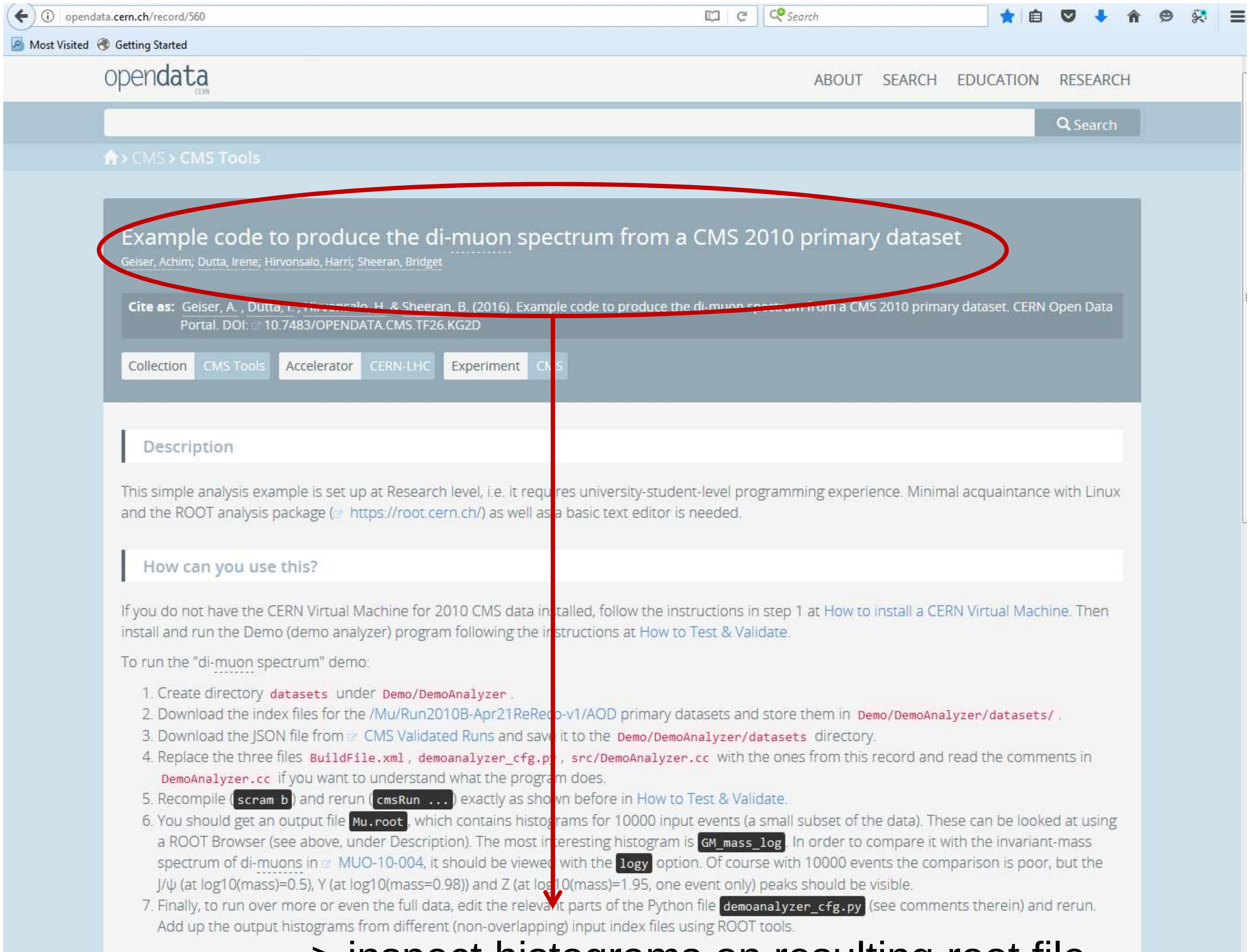
Important: Be aware that the instructions in the WorkBook are in use in CMS currently and have been updated for more recent CMSSW releases. With the 2010 data, you should always use the releases in the series of CMSSW_4_2 and not higher. Also note that more recent code does not work with older releases, so whenever you see `git cms-addpkg…` in the instruction, it is likely that the code package this command adds does not work with the release you need. However, the material under the pages gives you a good introduction to PAT.

Code as well as instructions for producing PATtuples from the CMS open data can be found in this GitHub repo. However, since it took a dedicated computing cluster nine days (!!!) to run this step and reduce the several TB of AOD files to a few GB of PATtuples, we have provided you with the PATtuples in that GitHub repo, saving you quite a lot of

for this tutorial:

choose Option A

Open Data Tutorial


opendata.cern.ch/record/560

Most Visited Getting Started

opendata

ABOUT SEARCH EDUCATION RESEARCH

Search

CMS › CMS Tools

Example code to produce the di-muon spectrum from a CMS 2010 primary dataset

Geiser, Achim; Dutta, Irene; Hirvonsalo, Harri; Sheeran, Bridget

Cite as: Geiser, A., Dutta, I., Hirvonsalo, H. & Sheeran, B. (2016). Example code to produce the di-muon spectrum from a CMS 2010 primary dataset. CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.TF26.KG2D

Collection CMS Tools | Accelerator CERN-LHC | Experiment CMS

Description

This simple analysis example is set up at Research level, i.e. it requires university-student-level programming experience. Minimal acquaintance with Linux and the ROOT analysis package (https://root.cern.ch/) as well as a basic text editor is needed.

How can you use this?

If you do not have the CERN Virtual Machine for 2010 CMS data installed, follow the instructions in step 1 at How to install a CERN Virtual Machine. Then install and run the Demo (demo analyzer) program following the instructions at How to Test & Validate.

To run the "di-muon spectrum" demo:

1. Create directory `datasets` under `Demo/DemoAnalyzer`.
2. Download the index files for the /Mu/Run2010B-Apr21ReReco-v1/AOD primary datasets and store them in `Demo/DemoAnalyzer/datasets/`.
3. Download the JSON file from CMS Validated Runs and save it to the `Demo/DemoAnalyzer/datasets` directory.
4. Replace the three files `BuildFile.xml`, `demoanalyzer_cfg.py`, `src/DemoAnalyzer.cc` with the ones from this record and read the comments in `DemoAnalyzer.cc` if you want to understand what the program does.
5. Recompile (`scram b`) and rerun (`cmsRun ...`) exactly as shown before in How to Test & Validate.
6. You should get an output file `Mu.root`, which contains histograms for 10000 input events (a small subset of the data). These can be looked at using a ROOT Browser (see above, under Description). The most interesting histogram is `GM_mass_log`. In order to compare it with the invariant-mass spectrum of di-muons in MUO-10-004, it should be viewed with the `logy` option. Of course with 10000 events the comparison is poor, but the J/ψ (at log10(mass)=0.5), Υ (at log10(mass)=0.98)) and Z (at log10(mass)=1.95, one event only) peaks should be visible.
7. Finally, to run over more or even the full data, edit the relevant parts of the Python file `demoanalyzer_cfg.py` (see comments therein) and rerun. Add up the output histograms from different (non-overlapping) input index files using ROOT tools.

you have now reached the dimuon mass spectrum example
(works technically like “Test & Validate”)

follow instructions to download and run it

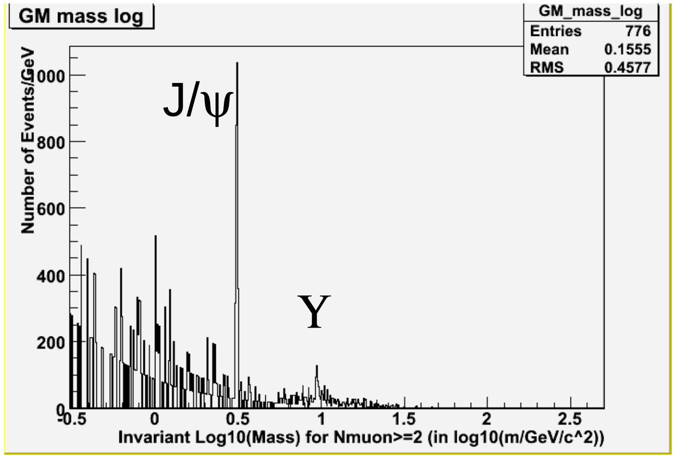
-> inspect histograms on resulting root file

Open Data Tutorial

-> inspect histograms on resulting ROOT file

with 10000
input events

(default,
runs ~few minutes)

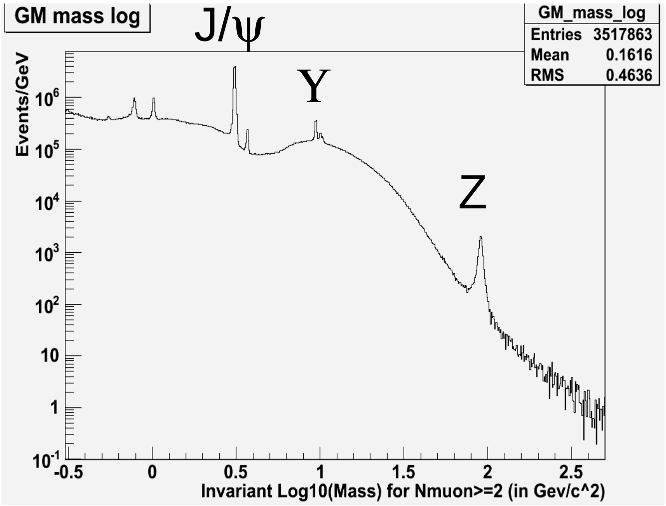


there you are!

now ready to
edit example
and
add/change
what you like

with full 2010
Mu dataset,
set logy option

(I/O to CERN
takes ~1 day
at ~10 Mbit/s)



Open Data Tutorial: conceptual extension

e.g. for **“displaced lepton” analysis**: (example in preparation)

repeat previous exercise (or other) with **2011 data** (fully documented)

learn how to **find, access and treat physics objects** you are interested in
(e.g. select non-vertex-associated leptons, possibly do revertexing (tools exist!))

learn how to **select the most relevant dataset(s)**
and how to **identify and treat the most relevant triggers**

we can (try to) help in all these steps

check availability of **suitable MC sets** (among the already released ones),
possibly **reweight for your application**;
if necessary, **try to generate your own exotic signal set** (not at all documented yet, being tried; if successful, might become possible with full detector simulation, but will remain a challenge! Or try to use external simplified simulation tools)

possibly extract your **personal reduced data set** (not necessary, but allowed)

do your analysis and
publish with your (non-CMS) name on it, cite DOIs of CMS open datasets used

Extended Vision (for discussion)

my **personal extension of initial vision:**

(for discussion, not a collaboration statement)

with ~1% of additional resources aim to achieve

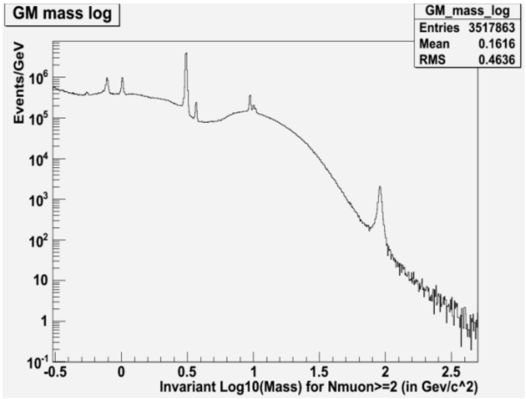
~10% additional scientific output (physics papers)

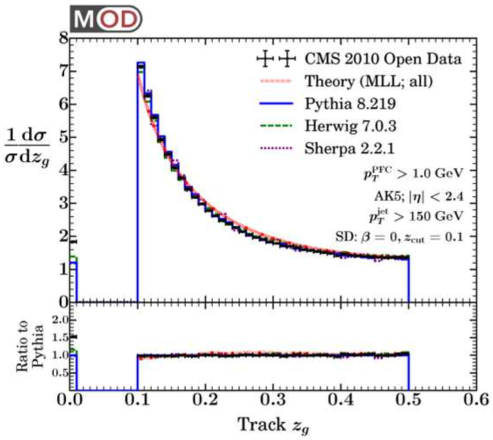
from both external and internal use of preserved/open data

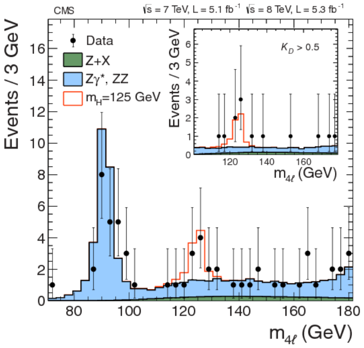
over lifetime of experiment + 10-20 years

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 (more coming)
- 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
- upcoming 2012 data release: on the way towards public reconstruction of the Higgs discovery
- **feel encouraged to use it for your purposes!**







Backup

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

see also talk J. Thaler
+ discussion

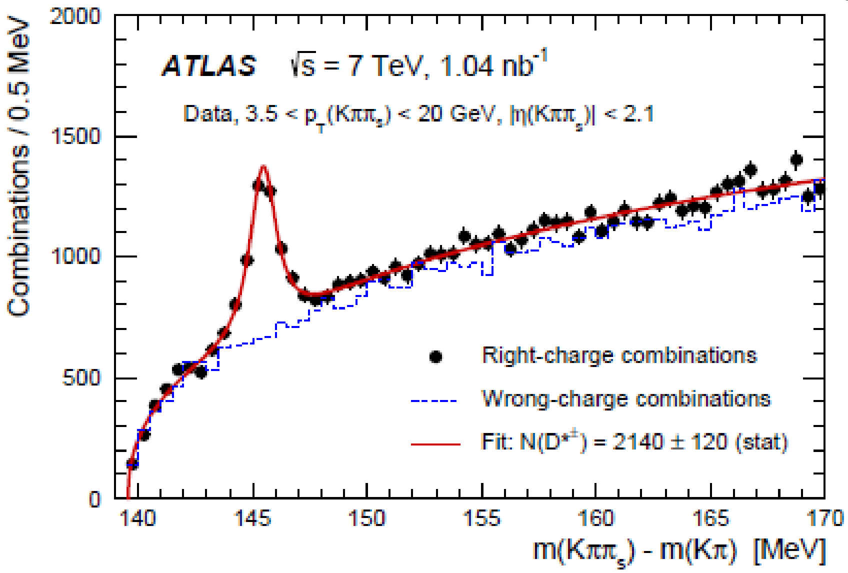
Contains section with Advice to community, Challenges, and Recommendations (see there)

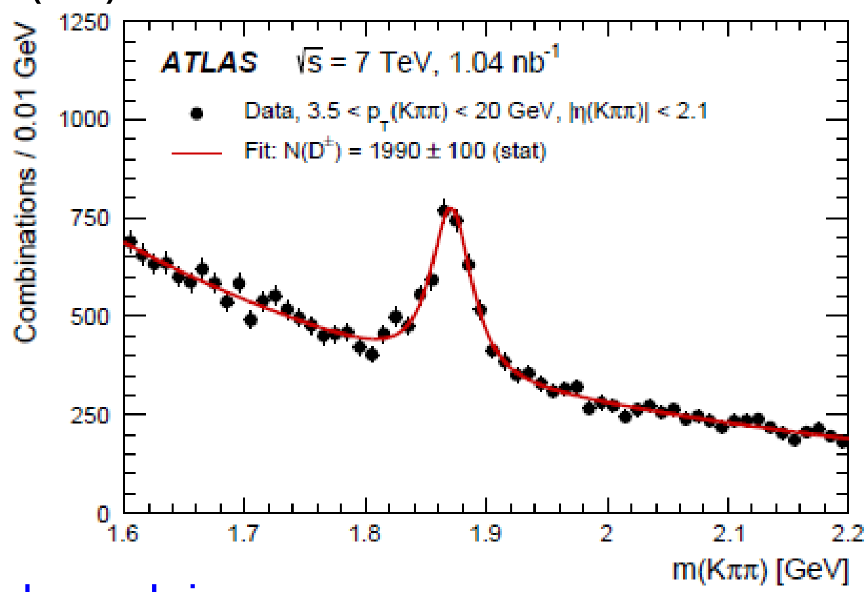
Releases of 2011 CMS data+MC "exciting"
-> properly evaluate detector systematics

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"

"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