



BETHEL  
UNIVERSITY

# Open Data Scouting

OPEN DATA WORKSHOP 2024

WHEPP CONFERENCE

JULIE HOGAN

1/06/2024



Julie Hogan



Matt Bellis

## ► What you've done:

- Set up docker – *thanks trailblazers considering this on a cluster!!*
- “Seen” the CMS detector
- Dabbled with analyzing ROOT files
- Learned about the different physics objects

## ► What's coming up:

### Jan 6

17:45-17:55	Welcome & Introductions	Julie Hogan
17:55-18:25	Finding CMS Open Data (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:15	Inspecting CMS data files (Activity)	Julie Hogan

### Jan 7


17:45-18:25	Event selection (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:15	Event selection (Activity)	Julie Hogan
Bonus Material	Accessing trigger information	
Bonus Material	Advanced tools	

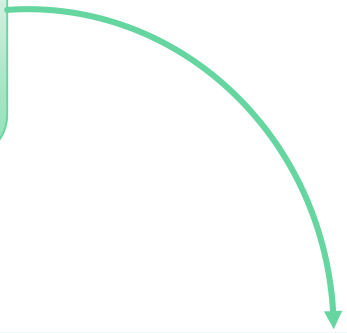
### Jan 9

17:45-18:25	Analysis example (Lesson)	Matt Bellis
18:25-18:35	Break	
18:35-19:15	Analysis example (Activity)	Matt Bellis
Bonus Material	Create a "stack plot" histogram	
Bonus Material	Systemics & Statistical interpretation	

### Jan 10

17:45-18:25	Analysis scale-up (Lesson)	Julie Hogan
18:25-18:35	Break	
18:35-19:05	Analysis scale-up (Activity)	Julie Hogan
19:05-19:15	Closing	Julie Hogan
Bonus Material	Reinterpreting CMS searches	

- Data & Simulation
- CMS Formats
  - I go find them online
- 

- My chosen events
- Format that I like
  - Apply basic choices to reduce size
  - Keep enough info to stay flexible
- 

- Do physics!! The fun part!!**
- Process my formatted files
  - Drop events, divide them up, etc
  - Analyze whatever I want

- Data & Simulation
- CMS Formats
  - I go find them online

**Today:** learn where to find the data online!  
Review the CMS data formats

**Earlier:** [you saw POET](#), our “likeable”  
file format creator

- My chosen events
- Format that I like
  - Apply basic choices to reduce size
  - Keep enough info to stay flexible

**Tomorrow:** discuss how to make basic choices

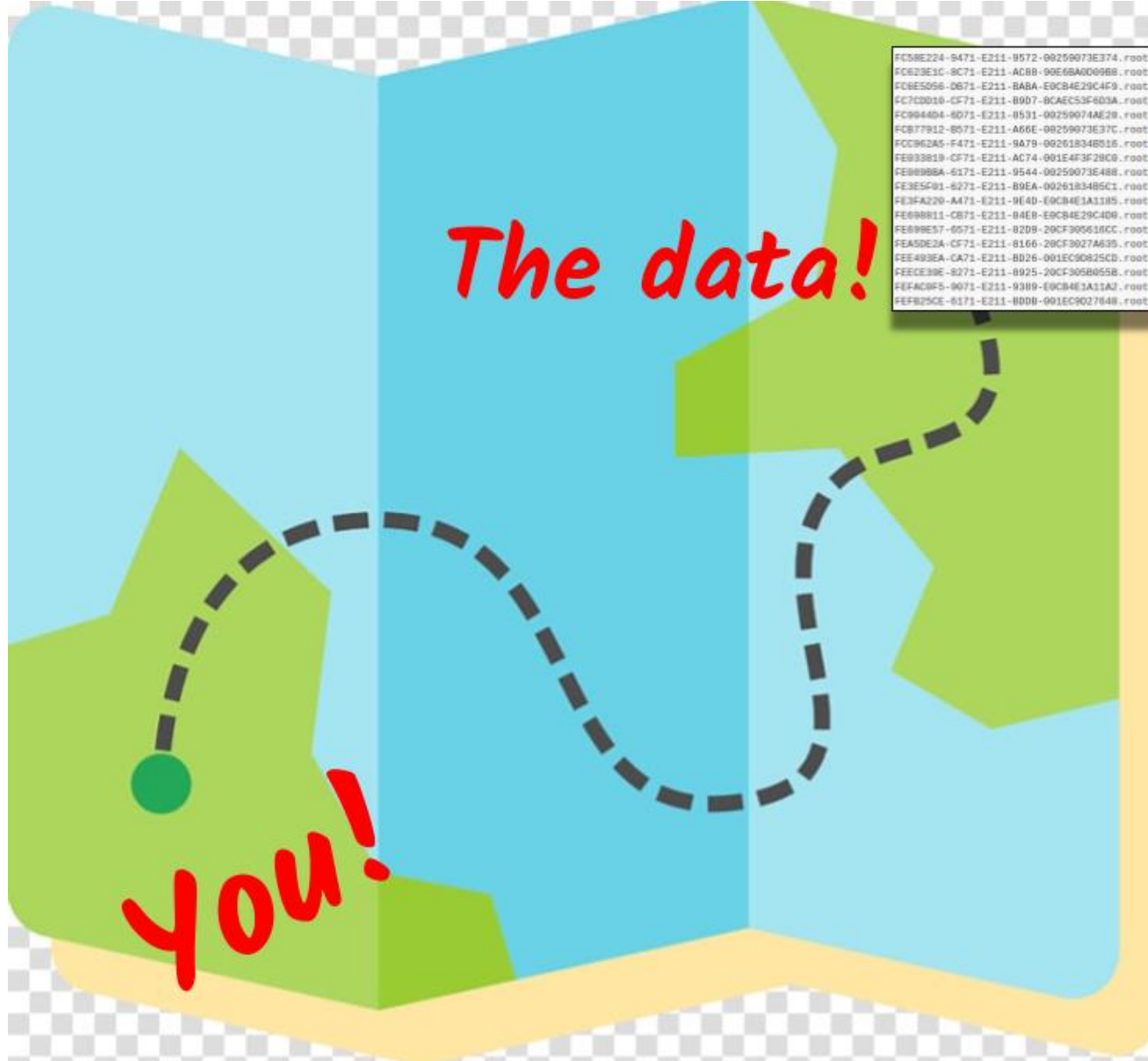
**Wednesday:** learn to process data at scale

**Earlier:** [you saw ROOT](#) analysis tools

**Tuesday:** we show python tools that can  
analyze data in the POET format easily

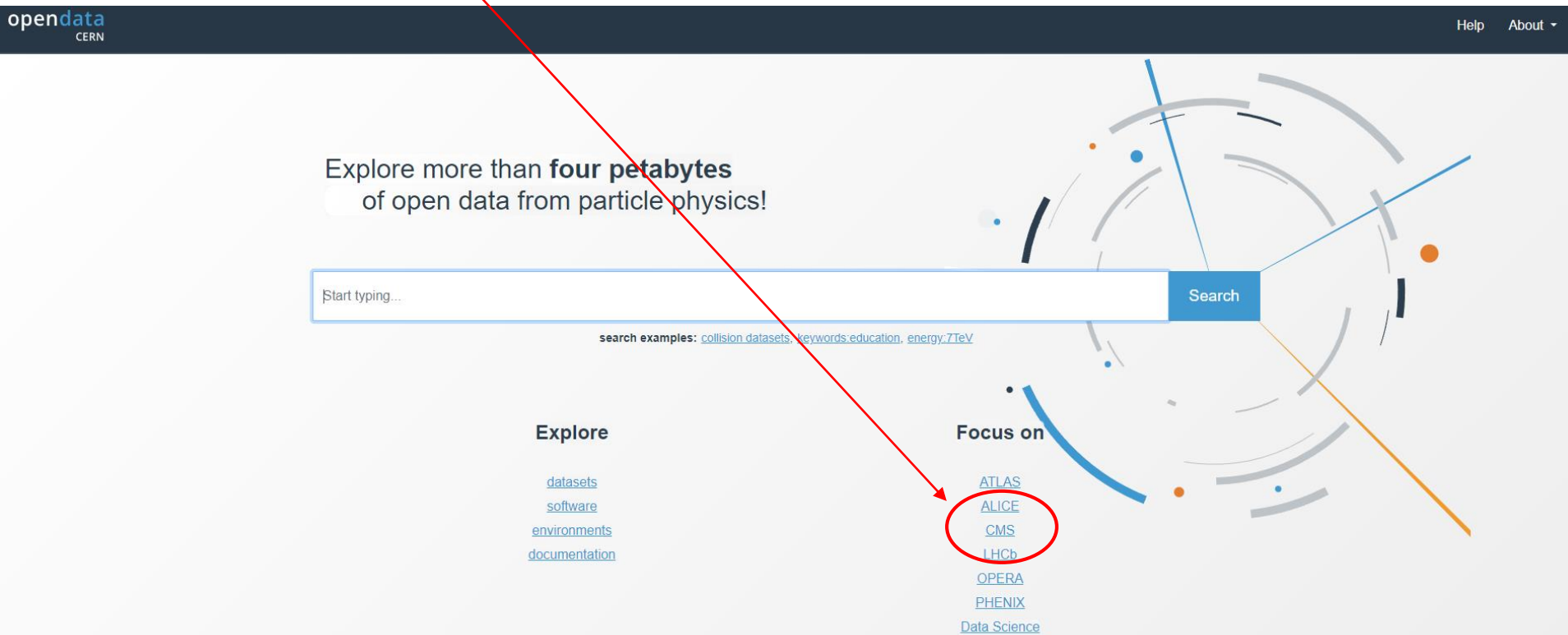
- Do physics!! The fun part!!**
- Process my formatted files
  - Drop events, divide them up, etc
  - Analyze whatever I want

# Finding the Open Data



► Go to [opendata.cern.ch](https://opendata.cern.ch)!

► Click “CMS”



The screenshot shows the CERN Open Data Portal homepage. At the top left is the 'opendata CERN' logo. At the top right are 'Help' and 'About' links. The main heading reads 'Explore more than **four petabytes** of open data from particle physics!'. Below this is a search bar with the placeholder text 'Start typing...' and a 'Search' button. Under the search bar are search examples: 'collision datasets', 'keywords:education', and 'energy:7TeV'. On the left, under the heading 'Explore', there are links for 'datasets', 'software', 'environments', and 'documentation'. On the right, under the heading 'Focus on', there is a list of particle physics experiments: 'ATLAS', 'ALICE', 'CMS', 'LHCb', 'OPERA', 'PHENIX', and 'Data Science'. A red arrow from the text 'Click “CMS”' points to the 'CMS' link, which is circled in red.

Dataset × Collision × Derived ×  
Simulated × CMS ×

include on-demand datasets

### Filter by type

- Dataset 8633
- Documentation 48
- Environment 49
  - News 11
- Software 42
- Supplementaries 4372

### Filter by experiment

- ALICE 15
- ATLAS 113
- CMS 8633
- LHCb 103
- OPERA 904
- PHENIX 1

### Filter by year

- 2010 167
- 2011 484
- 2012 523
- 2013 238
- 2015 7177
- 2016 22

Sort by:

Display:

Found 8633 results.

### Event display file derived from /DoubleMu/Run2011A-12Oct2013-v1/AOD

Sample event set from /DoubleMu/Run2011A-12Oct2013-v1/AOD readable from the browser-based 3d event display

No selection or quality criteria have been applied o...

### Event display file derived from /Jet/Run2011A-12Oct2013-v1/AOD

Sample event set from /Jet/Run2011A-12Oct2013-v1/AOD primary dataset readable from the browser-based 3d event display

No selection or quality criteria have been applied on the...

### Event display file derived from /METBTag/Run2011A-12Oct2013-v1/AOD

Sample event set from /METBTag/Run2011A-12Oct2013-v1/AOD primary dataset in json format readable from the browser-based 3d event display

No selection or quality criteria have been applied on...

### Filter by collision type

- Interfill 1
- PbPb 43
- pPb 139
- pp 8389

### Filter by collision energy

- 0.9TeV 4
- 0TeV 1
- 13TeV 7170
- 2.76TeV 111
- 5.02TeV 172
- 7TeV 595
- 8TeV 519

**Select "2012"**



Dataset × Collision × Derived ×  
Simulated × CMS × 2012 ×

include on-demand datasets

Filter by type

- Dataset 523
  - Collision 100
  - Derived 54
  - Simulated 369
- Documentation 6
  - Activities 5
  - Authors 1

Sort by: Best match ↓ asc. ↓

Display: detailed ↓ 20 results ↓

Found 523 results.

Event display file derived from /HcalNZS/Run2012C-22Jan2013-v1/AOD

Sample event set from /HcalNZS/Run2012C-22Jan2013-v1/AOD primary dataset in json format readable from the browser-based 3d event display

No selection or quality criteria have been applied on...

Dataset Derived CMS

**Select “Collision”**

/MuEG Run2012A 22Jan2013-v1 AOD

MuEG primary dataset in AOD format from RunA of 2012. Run period from run number 190456 to 193621.

Name, pointing toward the type of data in this “stream”

Year + “era”

Data format name

Processing campaign

## SingleMu primary dataset in AOD format from Run of 2012 (/SingleMu/Run2012B-22Jan2013-v1/AOD)

/SingleMu/Run2012B-22Jan2013-v1/AOD, CMS collaboration

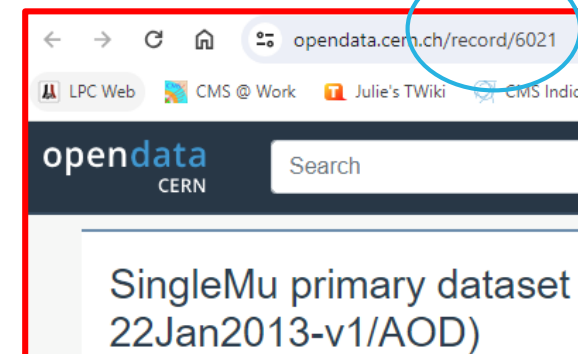
Cite as: CMS collaboration (2017). SingleMu primary dataset in AOD format from Run of 2012 (/SingleMu/Run2012B-22Jan2013-v1/AOD). CERN Open Data Portal. DOI:[10.7483/OPENDATA.CMS.IYVQ.1J0W](https://doi.org/10.7483/OPENDATA.CMS.IYVQ.1J0W)

Dataset Collision CMS 8TeV pp CERN-LHC

### *Citation info!*

- ▶ Description of the dataset (name, year, format, run numbers, validation)
- ▶ Number of events, files, total size
- ▶ Software settings and containers to process this data
- ▶ Creation path for the dataset – trigger information!
- ▶ Data quality monitoring information
- ▶ Links to instructions for analysis tools
- ▶ File lists and download links

*“Record ID”*



## Simulated dataset DY3JetsToLL\_M-50\_TuneZ2Star\_8TeV-madgraph in AODSIM format for 2012 collision data

`/DY3JetsToLL_M-50_TuneZ2Star_8TeV-madgraph` `Summer12_DR53X-PU_RD1_START53_V7N-v1` `AODSIM` CMS collaboration

Cite as: CMS collaboration (2017). Simulated dataset DY3JetsToLL\_M-50\_TuneZ2Star\_8TeV-madgraph in AODSIM format for 2012 collision data. CERN Open Data Portal. DOI:[10.7483/OPENDATA.CMS.RYNC.1VIB](https://doi.org/10.7483/OPENDATA.CMS.RYNC.1VIB)

**Dataset** **Simulated** **Standard Model Physics** **Drell-Yan** **CMS** **8TeV** **pp** **CERN-LHC**

Name, suggesting the physics process & generator

Processing campaign, with a year label

Data format name, ending in “SIM”

- ▶ Description of the dataset (name, year, format)
- ▶ Number of events, files, total size
- ▶ Software settings and containers to process this data
- ▶ Creation path for the dataset – generator information!
- ▶ Links to instructions for analysis tools
- ▶ File lists and download links

- ▶ Simulated datasets are sorted (very manually...) into categories!
  - ▶ Example for 2015 simulation:
- ▶ [There's a guide](#) to figuring out the names
- ▶ [Ask on the forum](#) if you can't find a process
  - ▶ It might exist! We might be able to find it
  - ▶ We can suggest how to get your LHE into CMS AOD

## Filter by category

<input type="checkbox"/> B physics and Quarkonia	55
<input type="checkbox"/> Beyond 2 Generations	239
▼ <input type="checkbox"/> Exotica	4450
<input type="checkbox"/> Dark Matter	600
<input type="checkbox"/> Excited Fermions	207
<input type="checkbox"/> Extra Dimensions	585
<input type="checkbox"/> Gravitons	862
<input type="checkbox"/> Heavy Fermions, Heavy Righ-Han...	493
<input type="checkbox"/> Heavy Gauge Bosons	1017
<input type="checkbox"/> Leptoquarks	443
<input type="checkbox"/> Miscellaneous	243
<input type="checkbox"/> Heavy-Ion Physics	1
▼ <input type="checkbox"/> Higgs Physics	1280
<input type="checkbox"/> Beyond Standard Model	491
<input type="checkbox"/> Standard Model	789
<input type="checkbox"/> Physics Modelling	51
▼ <input type="checkbox"/> Standard Model Physics	546
<input type="checkbox"/> Drell-Yan	73
<input type="checkbox"/> ElectroWeak	244
<input type="checkbox"/> Minimum Bias	8
<input type="checkbox"/> QCD	108
<input type="checkbox"/> Top physics	113
<input type="checkbox"/> Supersymmetry	488

- ▶ The information on the “record” webpages is stored as “metadata” that can be accessed on the command line with [cernopendata-client](#).



## cernopendata-client

python 2.7 | 3.6 | 3.7 | 3.8 | 3.9 | 3.10 CI passing docs passing codecov 80% gitter join chat  
license GPL-3.0

`cernopendata-client` is a command-line tool to facilitate downloading files from the [CERN Open Data portal](#). The tool enables to query datasets hosted on the CERN Open Data portal and to download and verify the individual data set files.

Let's follow on the webpage:

<https://cms-opendata-workshop.github.io/workshopwhepp-lesson-dataset-scouting/04-cli-through-cernopendata-client/index.html>

- ▶ Follow the webpage:

<https://cms-opendata-workshop.github.io/workshopwhepp-lesson-selection/05-solutions/index.html>

- ▶ Extra file to test: AOD format
  - ▶ `root://eospublic.cern.ch//eos/opendata/cms/Run2012A/MuEG/AOD/22Jan2013-v1/20000/00F0AA8F-D566-E211-9A55-BCAEC50971F9.root`

## Let's review!

Can you get to the starting page for the CMS Open Data Portal? Can you select the CMS data from that page?

## Let's review!

The bulk of the CMS released data covers what years? What was the collision energy for those years?

## Let's review!

What's the difference between *Collision* data and *Simulated* data?

## Let's review!

Select the CMS *collision* data for 2015. Select only the MINIAOD data. Do you remember what the MINIAOD data are?

Choose one dataset and identify different triggers that were used to direct events in this dataset. What do you think they are triggering on?

## Let's review!

Select the CMS *simulated* data for 2015. Select the **Heavy Gauge Bosons** (under *Filter by category*) and find a *Wprime* sample. Click on it. These are [hypothetical gauge bosons](#).

How can you learn *about CMS simulated dataset names*? What do you think the *W'* is decaying to? What is the assumed mass of the *W'* in this particular sample?

How many events are in this sample? How much hard drive space does this sample occupy? Can you find the generator parameters that were used to generate the collisions?