

# CMS open data for LLP Machine learning – a case study –

Freya Blekman

Vrije Universiteit Brussel

with large overlap with talks from the CMS open data group, check out also the [comprehensive talk by Xavier Duarte at LHCP 2019](#)

FNAL LHC Physics Centre  
distinguished Researcher 2020

# Outline

- Intro ML and Open data
- Short summary of what needs to be done to have a ML-accessible dataset
  - Describing my experience as someone familiar with the CMS software to get access to these samples
  - Which hopefully is at ‘tutorial’ level so is useful to outside world

# Machine Learning in HEP

- Machine Learning and HEP have a well-established connection
  - Simple NN have been used since LEP
- Machine Learning was vital for the discovery of the Higgs boson in 2012
  - Nowadays, Machine Learning is used in all aspects of analysis, reconstruction and software
    - These searches and techniques create opportunities that are created results with sensitivity well beyond what was expected from the LHC during its design phase
  - Most scenarios however are used inside the LHC collaborations as detailed information is used
- Often ML is progressing faster outside the LHC collaborations – it is relevant to engage the wider community to tackle advanced challenges



Welcome to [INSPIRE](#), the High Energy Physics information system. Please direct questions, comments or concerns to [feedback@inspirehep.net](mailto:feedback@inspirehep.net).

HEP :: [HepNames](#) :: [Institutions](#) :: [Conferences](#) :: [Jobs](#) :: [Experiments](#) :: [Journals](#) :: [Help](#)

Information [References \(10\)](#) [Citations \(59\)](#) [Files](#) [Plots](#)

## Track Finding With Neural Networks

[Carsten Peterson](#) (Lund U.)

Apr 1988 - 16 pages

**Nucl.Instrum.Meth. A279 (1989) 537**  
DOI: [10.1016/0168-9002\(89\)91300-4](https://doi.org/10.1016/0168-9002(89)91300-4)  
LU-TP-88-8

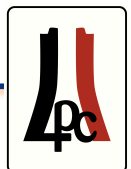
**Abstract** (Elsevier)  
A neural network algorithm for finding tracks in high energy physics experiments is presented. The performance of the algorithm is explored on modest size samples with encouraging results. It is inherently parallel and thus suitable for execution on a conventional SIMD architecture. More important, it naturally lends itself to direct implementations in custom made hardware, which would permit real time operations and hence facilitate fast triggers. Both VLSI and optical technology implementations are briefly discussed.

**Keyword(s):** [INSPIRE: track data analysis](#) | [mathematical methods: neural network](#) | [computer: multiprocessor](#) | [numerical calculations](#)

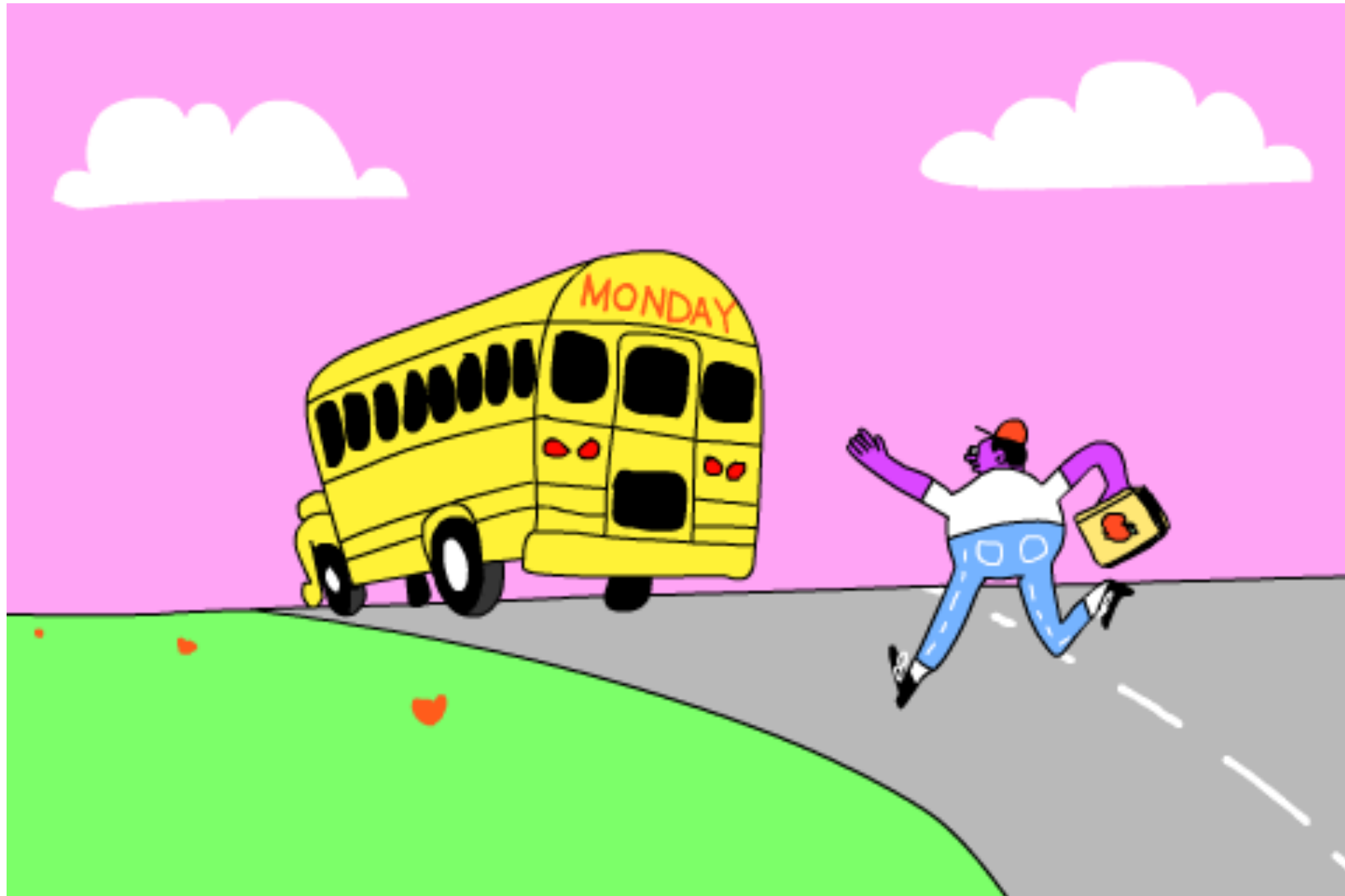
Record added 1988-12-14, last modified 2016-03-27

Export  
[BibTeX](#), [EndNote](#), [LaTeX\(US\)](#),  
[LaTeX\(EU\)](#), [Harvmac](#), [MARC](#),  
[MARCXML](#), [NLM](#), [DC](#)

One of the first ANN papers in HEP



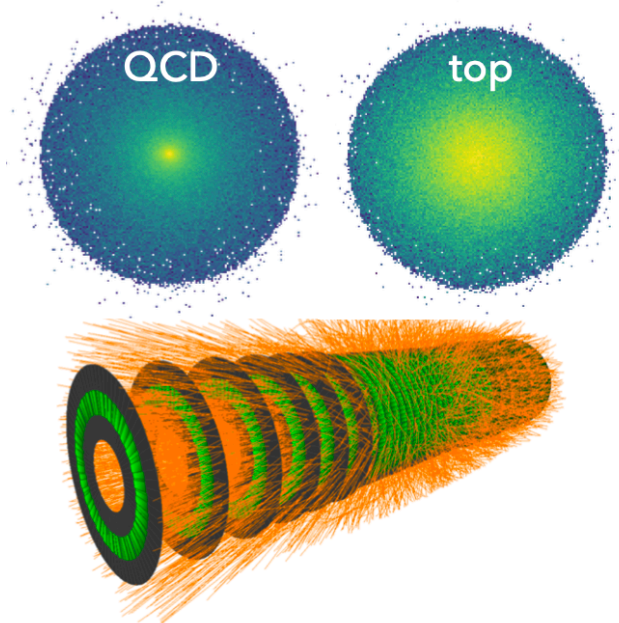
# Machine learning has evolved



- ▶ Engage ML community for interesting, realistic tasks in experimental HEP
- ▶ Calls at [ML4Jets](#) and [Connecting the Dots](#) workshops for more public HEP data sets with real detector simulation for ML applications

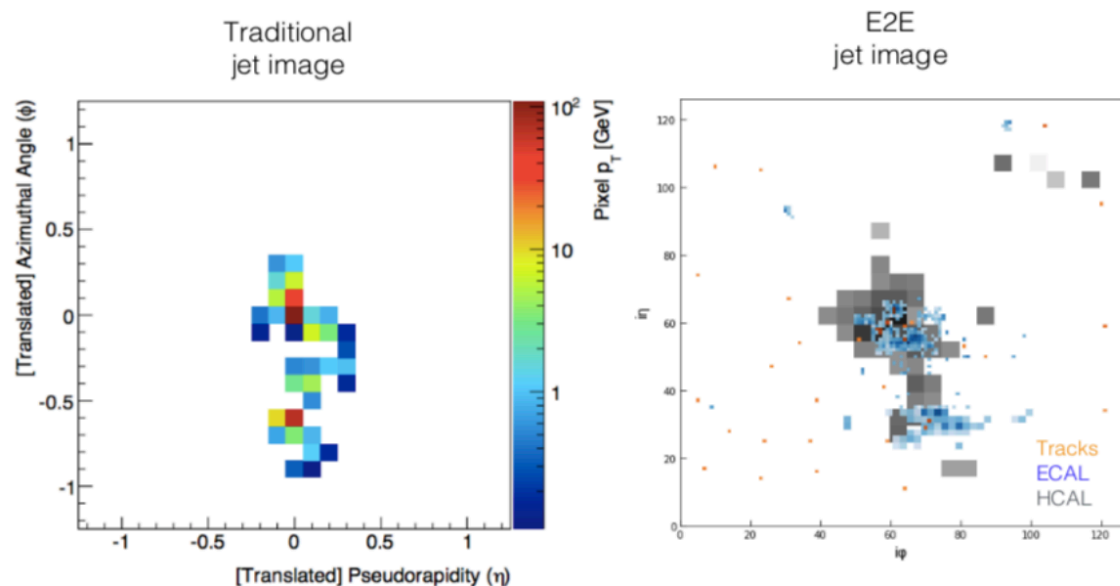
- ▶ Example: [data set](#) for top tagging based on Pythia+Delphes

- ▶ Example: [data set](#) for tracking based on ACTS (kaggle TrackML challenge)



- ▶ Can **CMS open data** fill this role for many ML applications?

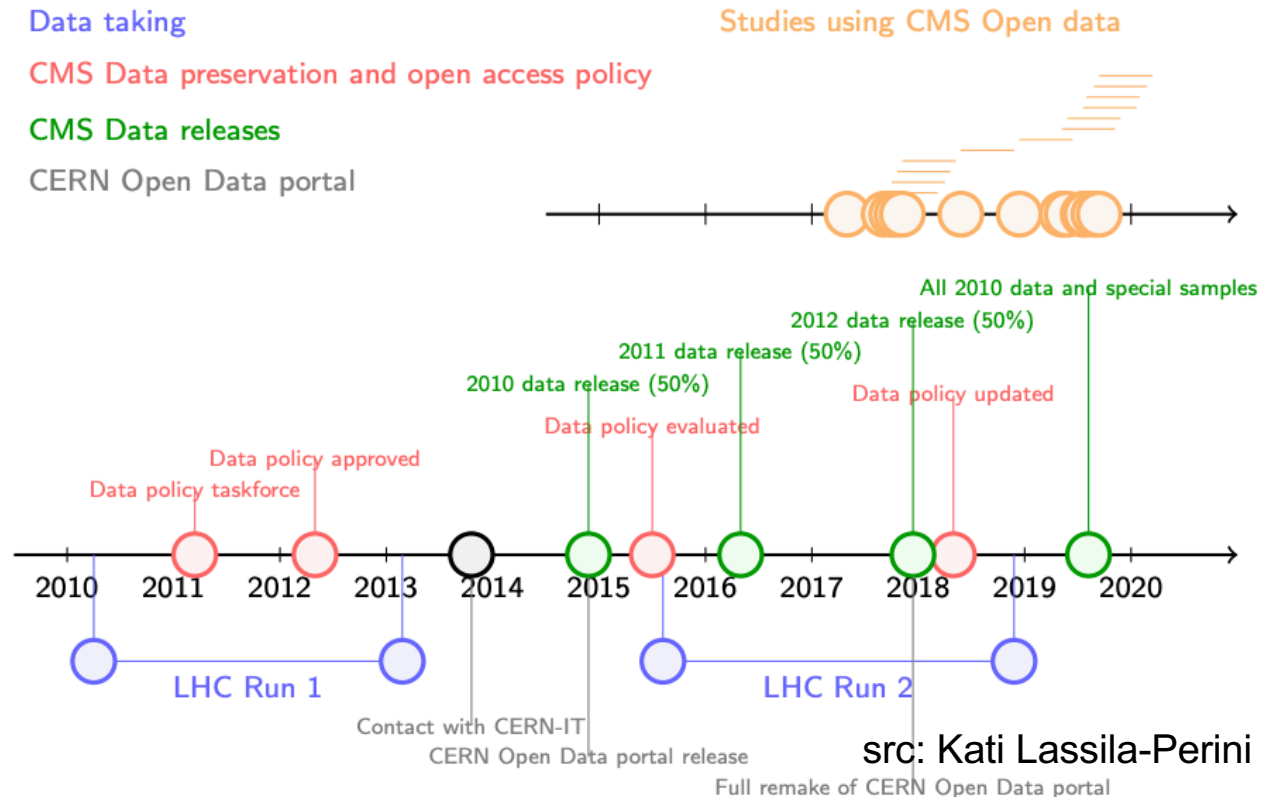
- ▶ Open data & simulation is also useful for ML-focused studies, e.g. [kaggle ATLAS  \$H \rightarrow \tau\tau\$  challenge](http://opendata.cern.ch/record/328): <http://opendata.cern.ch/record/328>
- ▶ Most existing efforts based on reducing AOD/MINIAOD samples
  - ▶ Requires CMS domain knowledge, CMS software, ...
- ▶ Convolutional neural networks image-based event classification [[arXiv:1708.07034](https://arxiv.org/abs/1708.07034)]
- ▶ End-to-end physics event classification [[arXiv:1807.11916](https://arxiv.org/abs/1807.11916)]
- ▶ End-to-end jet classification of quarks and gluons [[arXiv:1902.08276](https://arxiv.org/abs/1902.08276)]



Many more details: [Javier Duarte's talk at LHCP19](#)

# The LHC open data

- LHC open data is used mostly for science communication and approach depends on the collaboration



CMS Collaboration has released almost all of its LHC Run 1 data



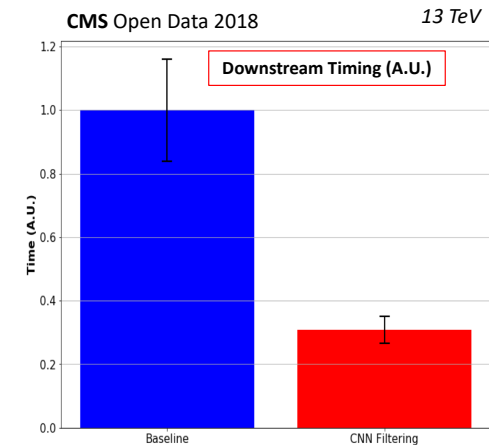
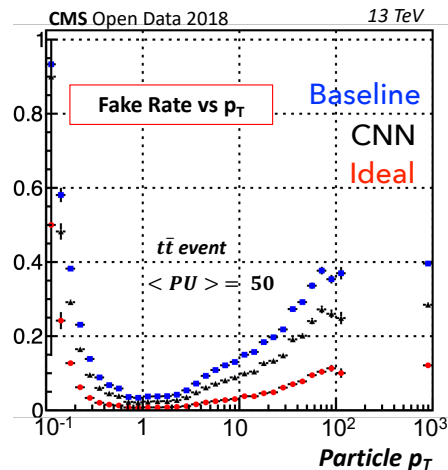
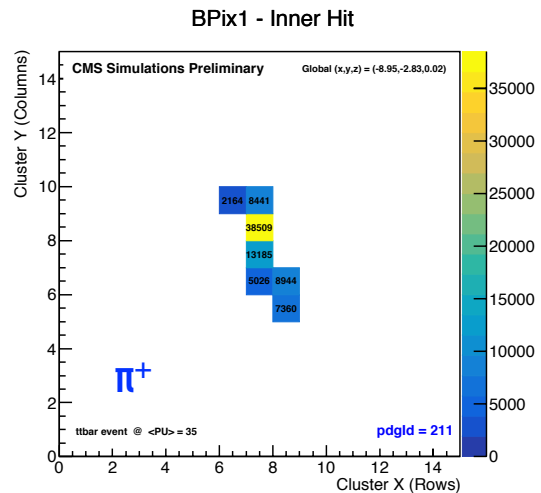
CMS RELEASES OPEN  
DATA FOR MACHINE  
LEARNING

- Derived datasets in ML-friendly formats
- Using 2016 (so 13 TeV!) CMS simulation instead of the 2009-2010 samples
- Provided in ROOT and HDF5
  - Jet flavor studies “b-tagging” - <http://opendata.cern.ch/record/12100>
  - Substructure studies “top-tagging” - <http://opendata.cern.ch/record/12220>
  - Pixel tracking studies - <http://opendata.cern.ch/record/12320>
  - H->bb tagging - <http://opendata.cern.ch/record/12102>



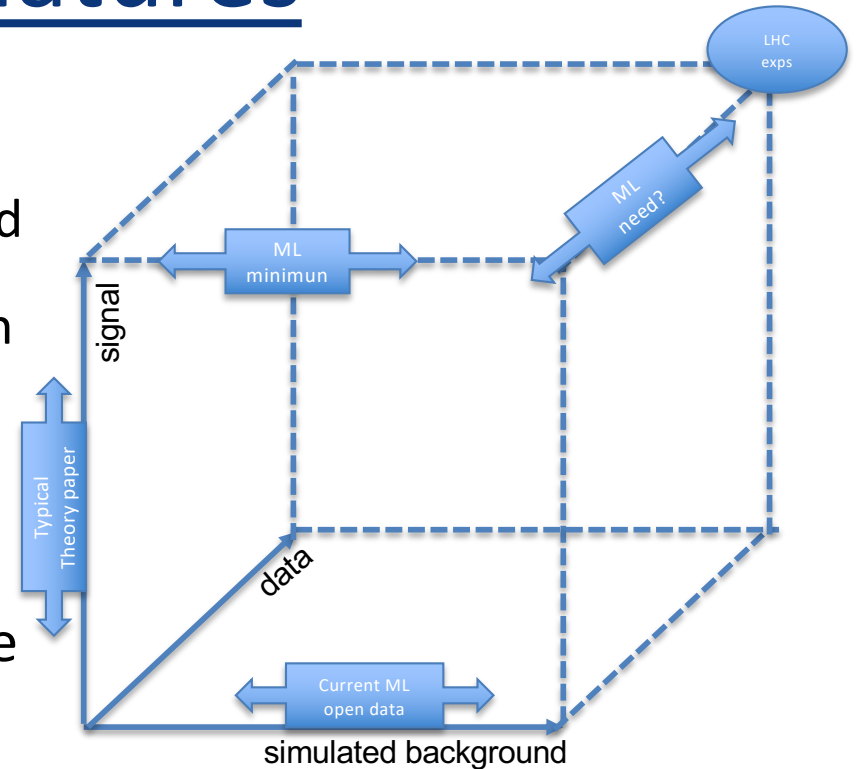
# Some highlights:

- CMS Pixel dataset: (200 GB after compression)
- high fake rate problem (S/B for good pixel tracks very low)
- Convolutional NN used to identify good pixel tracks
  - More details in this [talk](#) and [talk](#)

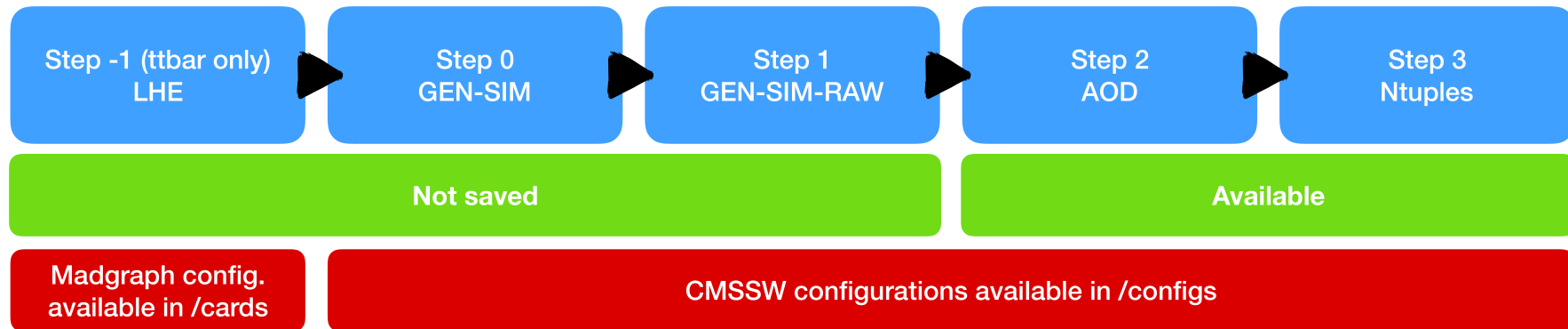


# Goal: adapt one of these for ML for displaced signatures

- Conundrum:
  - No signal samples available at all
  - Only MC available for background – and only the specific samples and preselection cuts that the collaboration chose
  - Data available is old(ish): LHC Run I
  - All three together necessary to be relevant for LLP studies
- My first goal: get sample of jets and the tracks therein, including displacement (impact parameter) information that can be used as background for ML studies
  - Signal? Challenging – ideas particularly from theory welcome



# Short intro: CMS software



ML examples on CMS open data portal give background data at level of Step3

- To get realistic signal samples the same chain need to be followed for signal as well
- This is highly CPU intensive and needs expert knowledge (as it includes detector settings etc)
  - From experience I guesstimate that >90% of CPU time is spent in steps 0, 1 and 2
  - Not clear how this can be done without CMS providing samples or involving CMS collaborators

# <http://opendata.cern.ch/>

opendata  
CERN

About ▾

Explore more than **two petabytes** of open data from particle physics!

Start typing... Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

**Explore**

- [datasets](#)
- [software](#)
- [environments](#)
- [documentation](#)

**Focus on**

- [ATLAS](#)
- [ALICE](#)
- [CMS](#)
- [LHCb](#)
- [OPERA](#)
- [Data Science](#)

Get started ▾

# First try: b-tagging, as very close in signature to long-lived jets

- So picked b-tag tutorial  
<http://opendata.cern.ch/record/12100>

(I tried some of the other tutorials as well, they are at various difficulty levels)

- Goal is to provide step-by-step instructions here so people can use it also at home

# Instructions

I'm a root user, so used root (there is a jupyter workbook available in a git repo linked from <http://opendata.cern.ch/record/12100>)

On my (osx with conda & root installed): the following command worked straight out of the box:

```
fblekman$ root
root://eospublic.cern.ch//eos/opendata/cms/datascience/JetNtupleProducerTool/JetNTuple_Q
CD_RunII_13TeV_MC/JetNtuple_RunIISummer16_13TeV_MC_1.root
```

So the files are readable straight from the open data portal

# Instructions

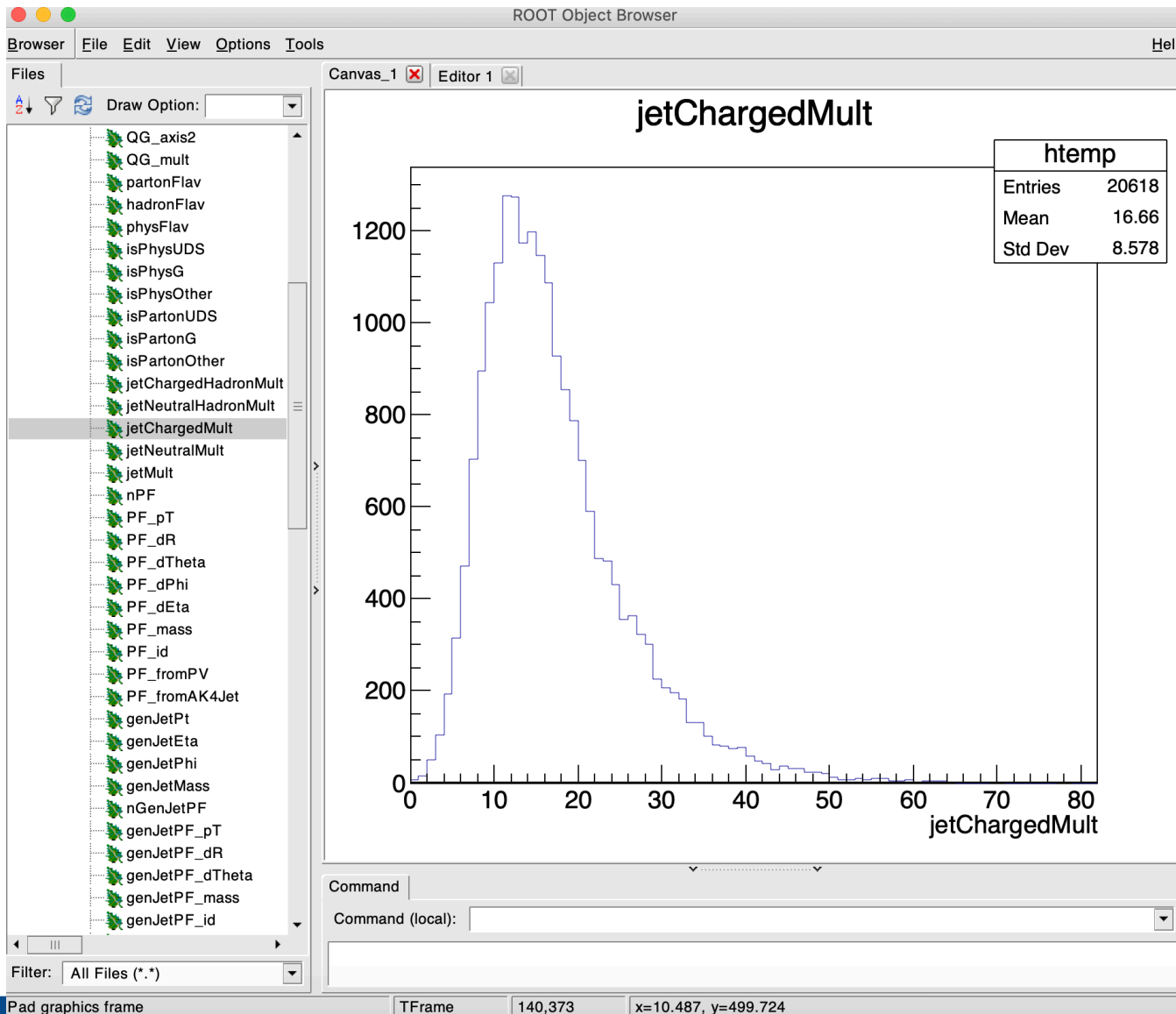
```
fblekman$ root
```

```
root://eospublic.cern.ch//eos/opendata/cms/datascience/JetNtupleProducerTool/JetNTuple_QCD_RunII_13TeV_MC/JetNtuple_RunIISummer16_13TeV_MC_1.root
```

More explanation:

- This is an Ntuple hosted on the CERN open data portal
- Produced by running the JetNtupleProducerTool ([documented here](#)) on CMS reconstructed simulation with all corrections applied
- The CMS open data team has run on the simulated [sample](#) documented on the open data page, which is a flat (in pT, 15 to 7000 GeV) spectrum QCD multijet production sample produced with Pythia 8 in the CUETP8M1 (“CMS tune”) tune at sqrt(s)=13 TeV
- It opens straight on my laptop in root!
- And is just one of [122 files](#) (TChain is your friend)

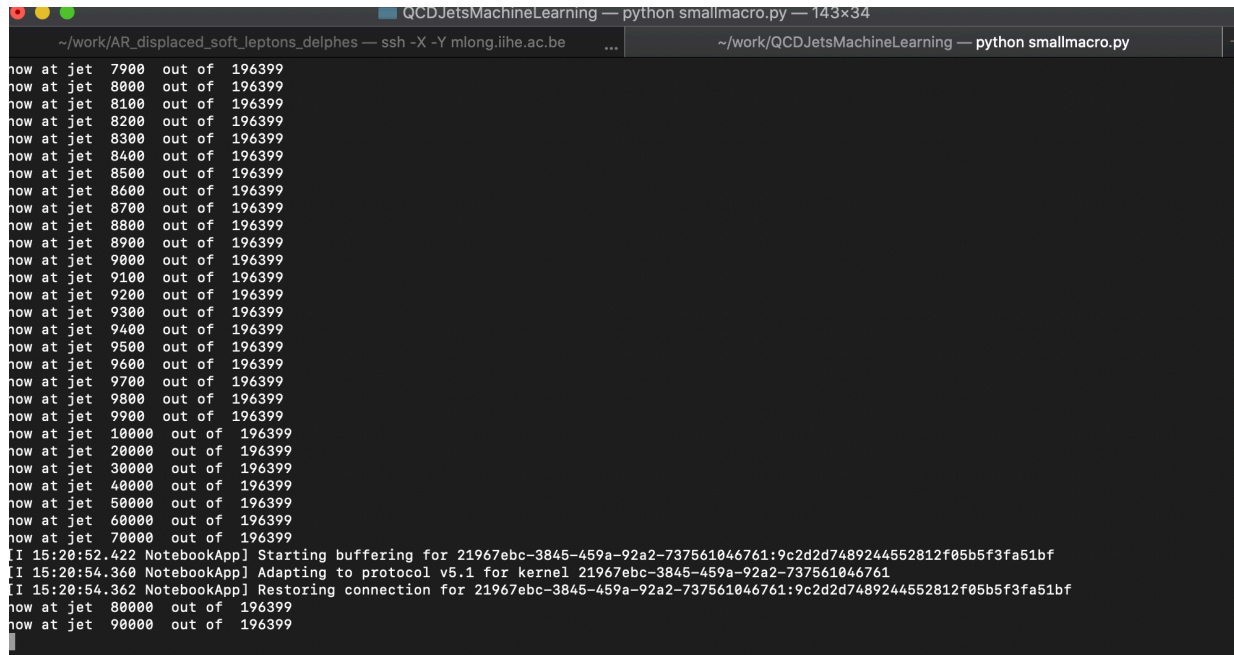
# It really works!





# It really works!

- I looped over one file filling one histogram during a coffee break at this workshop
- I've attached very small macro to the indico page of this talk at the LLP workshop – you should only need pyroot installed as is installed on any HEP-oriented computing cluster



```
~/work/AR_displaced_soft_leptons_delphes — ssh -X -Y mliong.ihe.ac.be ...  ~/work/QCDJetsMachineLearning — python smallmacro.py — 143x34
now at jet 7900 out of 196399
now at jet 8000 out of 196399
now at jet 8100 out of 196399
now at jet 8200 out of 196399
now at jet 8300 out of 196399
now at jet 8400 out of 196399
now at jet 8500 out of 196399
now at jet 8600 out of 196399
now at jet 8700 out of 196399
now at jet 8800 out of 196399
now at jet 8900 out of 196399
now at jet 9000 out of 196399
now at jet 9100 out of 196399
now at jet 9200 out of 196399
now at jet 9300 out of 196399
now at jet 9400 out of 196399
now at jet 9500 out of 196399
now at jet 9600 out of 196399
now at jet 9700 out of 196399
now at jet 9800 out of 196399
now at jet 9900 out of 196399
now at jet 10000 out of 196399
now at jet 20000 out of 196399
now at jet 30000 out of 196399
now at jet 40000 out of 196399
now at jet 50000 out of 196399
now at jet 60000 out of 196399
now at jet 70000 out of 196399
[I 15:20:52.422 NotebookApp] Starting buffering for 21967ebc-3845-459a-92a2-737561046761:9c2d2d7489244552812f05b5f3fa51bf
[I 15:20:54.360 NotebookApp] Adapting to protocol v5.1 for kernel 21967ebc-3845-459a-92a2-737561046761
[I 15:20:54.362 NotebookApp] Restoring connection for 21967ebc-3845-459a-92a2-737561046761:9c2d2d7489244552812f05b5f3fa51bf
now at jet 80000 out of 196399
now at jet 90000 out of 196399
```

- (accessing large samples remotely is slow – downloading the 100 GB will be substantially faster)

# Is this sample useful for LLP background studies?

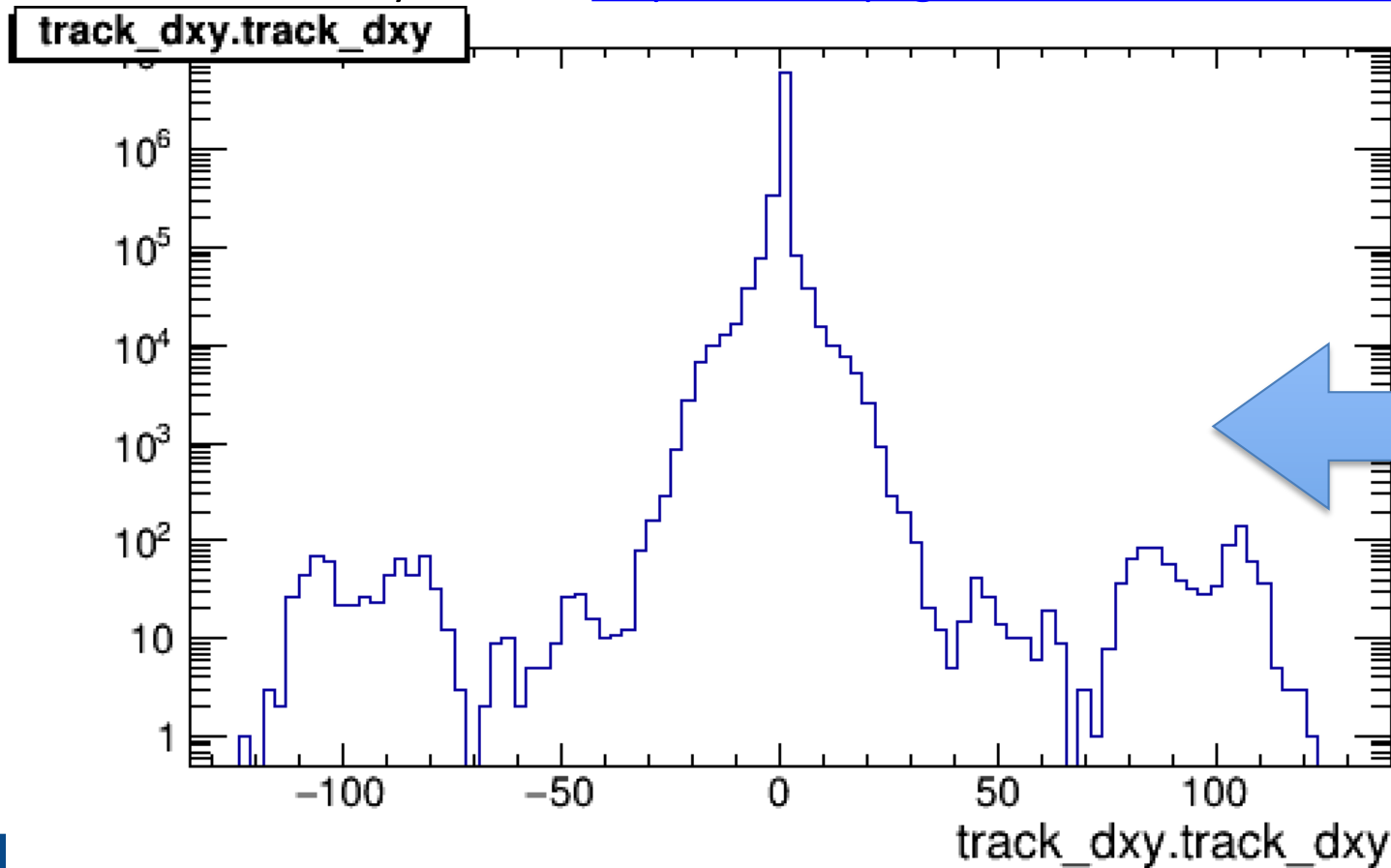
- Some essentials are missing, specifically: decay distances, Distance to PV and things like impact parameter and their uncertainty
  - Requesting that these are added can be done, but needs a CMS member to work with the CMS open data team
- Is this sample still useful?
  - Possibly for example if one is interested in signatures like more EM activity, this sample would be perfectly suitable as a realistic background model
  - or wants to look/train for signatures without using lifetime info
    - Possibly – for practically any obvious LLP search (displaced jets of various sorts, displaced leptons, etc) that uses short lifetimes, jets are a main background

# I tried to do this

- Any CMS author with substantial CMS software experience create new samples that add information to open data samples relatively easily (it took me few days and that was while multitasking – it can definitely be done)
  - But for Run I data if you want **data**!
  - For example I modified the [top-tag example](#) to contain jets and not include hit information and then produced small(ish) root tuples of minimum bias **data** and **QCD MC** containing jets and tracks from Run I (8 TeV) samples
    - They're here:
      - <https://homepage.iihe.ac.be/~fblekman/opendata-test/>
      - if you are interested in using them let me know
      - But no signal!

# I tried to do this

- small(ish) root tuples of minimum bias **data** and **QCD MC** containing jets and tracks from Run I (8 TeV) samples
  - They're here: <https://homepage.iihe.ac.be/~fblekman/opendata-test/>



Minimum bias **data** from 2011, showing  $d_{xy}$  of any track inside a 0.5 cone of a anti-kT 0,5 cone jet (no cuts applied beyond pT)

Note: Unit of x-axis I am not so sure, so don't count on new physics at  $|d_{xy}|=100$  <unit?>

# Some final thoughts

- In my opinion, the lack of signal models is the main limitation to make this an interesting avenue to pursue
  - Useful CMS 13 TeV samples are all simulation of QCD processes
  - Getting samples with lifetime info can be done
  - I produced small(ish) root tuples of minimum bias data and QCD MC containing jets and tracks from the 8 TeV samples
    - They're here:
      - <https://homepage.iihe.ac.be/~fblekman/opendata-test/>
      - if you are interested in using them let me know
      - But no signal!
    - No long-lived signal samples available (also not in 2010-2012 Run 1 Open data)
- Other options:
  - Go fully DELPHES – fixes the signal problem
  - DELPHES also for background, but there are challenges to get realistic background then

# Thoughts? Suggestions?



- CMS Open data community should be able to produce other samples
  - With specific suggestions this can be done  
(but takes someone's time so specific requests need to be motivated/organized)