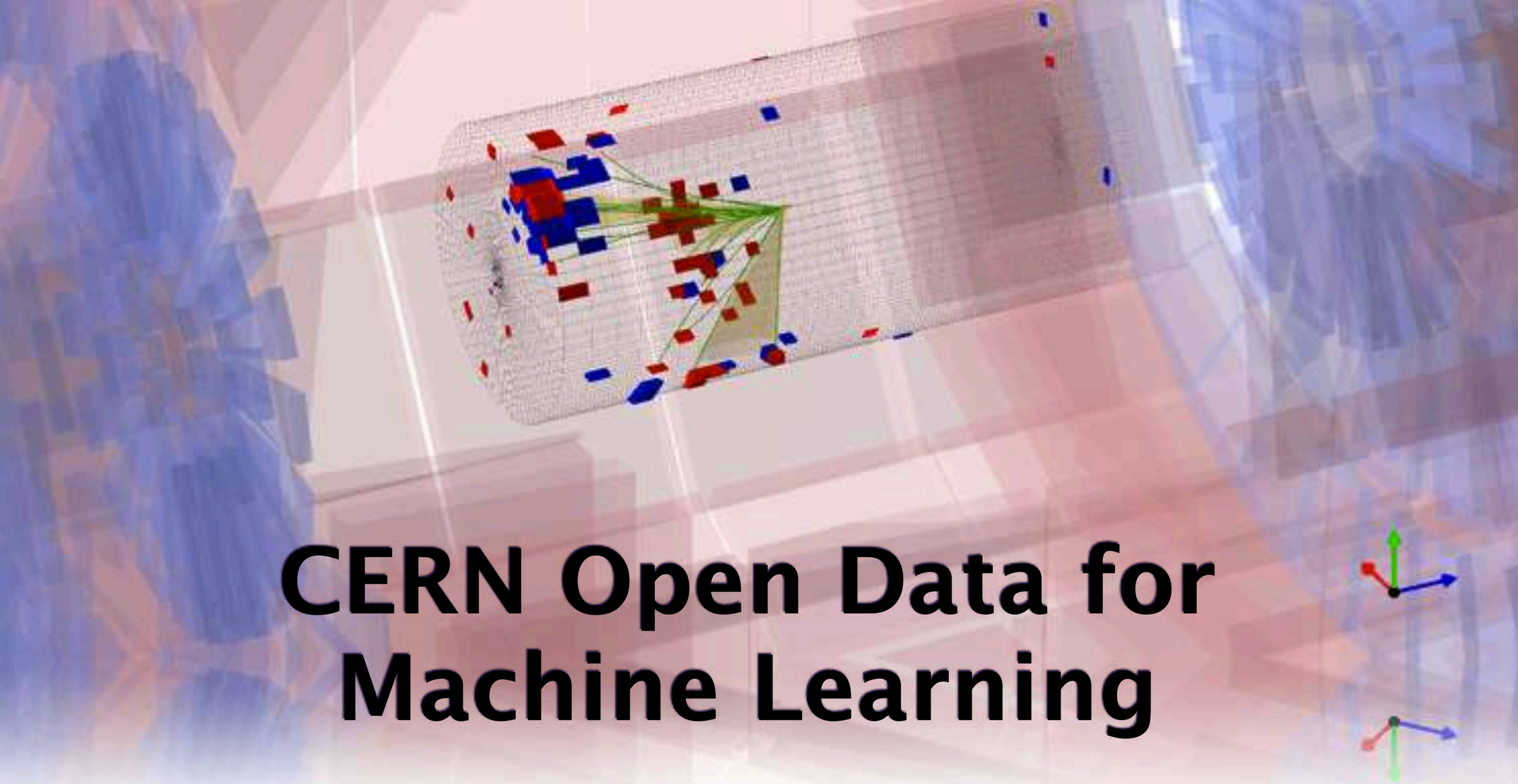


CERN Open Data for Machine Learning

Dr. Sercan Şen
(Hacettepe University)

Sercan.Sen@cern.ch

NEAR EAST UNIVERSITY
February 21, 2020

Outline

- **What is machine learning (ML)?**
    - Types of ML / classifiers / decision trees
- **LHC experiments at CERN**
    - Data format
    - High energy physics + ML
- **CERN open data**
- **ML dedicated open data release**
    - ML approach to jet identification
    - pixel tracking studies
    - Higgs to bb tagging
- **Summary and outlook**

What is Machine Learning (ML)?

- Wide field with increasing applications in research and industry
- First ideas already since 1950's
- Machine Learning not new to physics
- In HEP, many 'classic' ML methods are already in use.
  - Boosted decision trees (BDTs) in Higgs search
  - In ALICE, e.g. BDTs for signal extraction for charmed baryons

- Relatively new: Deep learning in physics
  → Huge progress done in last years
  → More and more analyses upcoming in all experiments
  Development of Machine Learning techniques is incredibly fast

- A lot of progress from tech companies and industry
- Boost from big data, most progress in deep learning
- Many advanced solutions for "human problems" (e.g. image recognition, text understanding) can also be adapted to HEP problems!

Hello World - C++ / Python

Example 1: Hello World Program

```cpp
#include <iostream>
using namespace std;

int main()
{
    cout << "Hello, World!";
    return 0;
}
```

Output

```
Hello, World!
```

Source Code

```python
# This program prints Hello, world!

print('Hello, world!')
```

Output

```
Hello, world!
```

To Run the code:

g++ example.cpp -o exp
./exp

To Run the code:

python example.py

Hello World - C++ / Python

This program checks whether user input number is positive, negative, or zero and displays the result

```
num = float(input("Enter a number: "))
if num >= 0:
    if num == 0:
        print("Zero")
    else:
        print("Positive number")
else:
    print("Negative number")
```

Output 1

```
Enter a number: 5
Positive number
```

Rules are written by hand.

Rule #1:
if the number is greater than zero, then the number is "positive".

Hello World – Machine Learning

Can you write code to tell the difference between an apple and an orange?

Tons of lines of rules are needed to do that

```
def detect_colors(image):
    # lots of code

def detect_edges(image):
    # lots of code

def analyze_shapes(image):
    # lots of code

def guess_texture(image):
    # lots of code

def define_fruit():
    # lots of code

def handle_probability():
    # lots of code
```

We need an algorithm that can figure out the rules for us, so we do not have to write them by hand.

→Traine a classifier

If you change the problem (objects), rules change

What is a Classifier?

It takes some data as input and assigns a label to it as output.

Classifier **Apple**

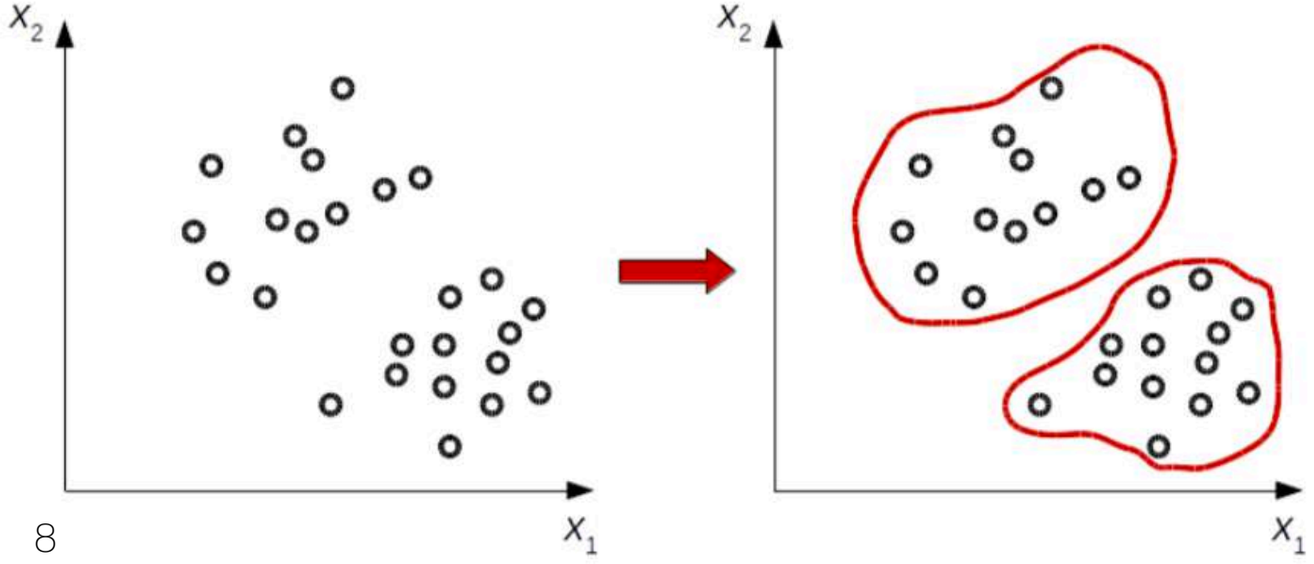
Classifier **Spam**

The technique to write the classifier automatically is called supervised learning.

Types of Machine Learning

Supervised Learning algorithms need **labeled training data** and learn the correct mapping of input data and desired output

Unsupervised Learning tries to **find a structure in the data** without a priori knowledge of desired outcome



Training Data

Features: Weight and Texture
Label: Orange and Apple

labeled training data
supervised learning

Weight	Texture	Label
150g	Bumpy	Orange
170g	Bumpy	Orange
140g	Smooth	Apple
130g	Smooth	Apple
…	…	…

The more training data you have, the better a classifier you can create

Training Data

```
from sklearn import tree
features = [[140, 1], [130, 1], [150, 0], [170, 0]]
labels = [0, 0, 1, 1]
clf = tree.DecisionTreeClassifier()
clf = clf.fit(features, labels)
print clf.predict([[150, 0]])
```

features:
smooth = 1
bumpy = 0

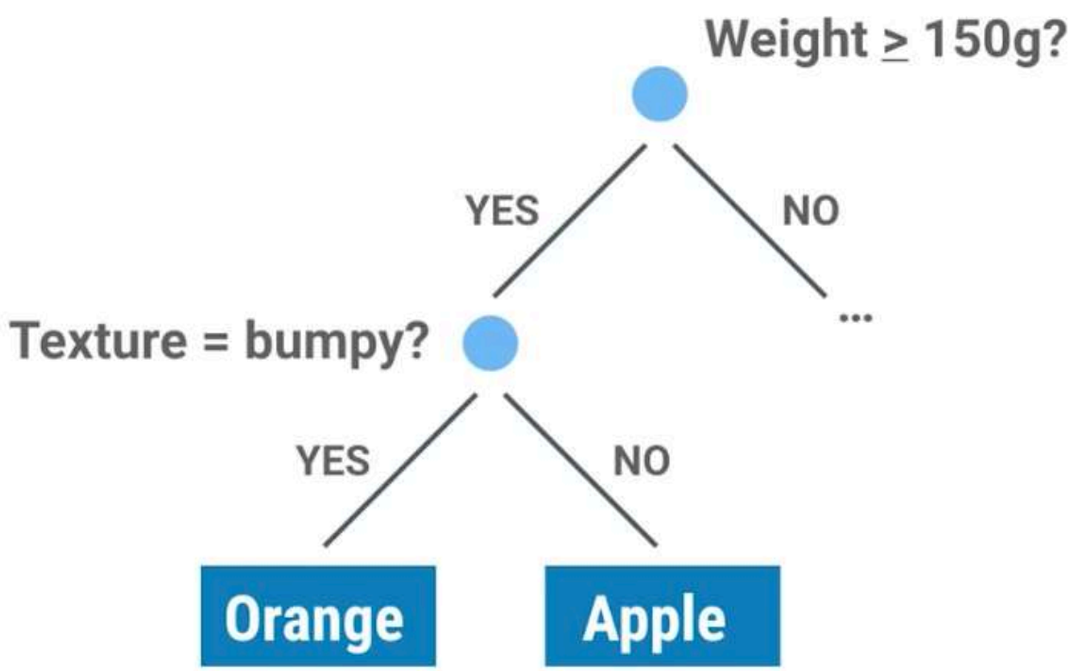
labels:
apple = 0
orange = 1

Output:

```
plus2:tutorials ssen$ python hello-world.py
[1]
```

training algorithm. can find patterns in data.

To train data, we use a type of classifier called a **decision tree.**



New classifier for a new problem can be created, just by changing the training data!

Much better than writing new rules for each problem.

Decision Trees (I)

Iris versicolor

Iris virginica

Iris setosa

Data set [edit]

The dataset contains a set of 150 records under five attributes - petal length, petal width, sepal length, sepal width and species.

Fisher's *Iris* Data [hide]

Dataset Order	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	*I. setosa*
2	4.9	3.0	1.4	0.2	*I. setosa*
3	4.7	3.2	1.3	0.2	*I. setosa*
4	4.6	3.1	1.5	0.2	*I. setosa*
5	5.0	3.6	1.4	0.3	*I. setosa*
6	5.4	3.9	1.7	0.4	*I. setosa*
7	4.6	3.4	1.4	0.3	*I. setosa*
8	5.0	3.4	1.5	0.2	*I. setosa*
9	4.4	2.9	1.4	0.2	*I. setosa*
10	4.9	3.1	1.5	0.1	*I. setosa*

→Three types of flowers

→50 examples for each type

→150 examples in total

https://en.wikipedia.org/wiki/Iris_flower_data_set

Decision Trees (II)

7.2. Toy datasets

scikit-learn comes with a few small standard datasets that do not require to download any file from some external website.

They can be loaded using the following functions:

`load_boston`([return_X_y])	Load and return the boston house-prices dataset (regression).
`load_iris`([return_X_y])	Load and return the iris dataset (classification).
`load_diabetes`([return_X_y])	Load and return the diabetes dataset (regression).
`load_digits`([n_class, return_X_y])	Load and return the digits dataset (classification).
`load_linnerud`([return_X_y])	Load and return the linnerud dataset (multivariate regression).
`load_wine`([return_X_y])	Load and return the wine dataset (classification).
`load_breast_cancer`([return_X_y])	Load and return the breast cancer wisconsin dataset (classification).

```
⌥⌘2                    ssen@lxplus700:/afs/cern.ch/wo
from sklearn.datasets import load_iris
iris = load_iris()
print iris.feature_names
print iris.target_names
print iris.data[0]
print iris.target[0]
~
```

import the iris into scikit-learn

https://scikit-learn.org/stable/datasets/index.html

```
plus2:tutorials ssen$ python iris.py
['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
['setosa' 'versicolor' 'virginica']
[5.1 3.5 1.4 0.2]
0
```

Instructions

1. Import dataset
2. Train a classifier
3. Predict label for new flower
4. Visualize the tree

Testing Data

- Examples used to "test" the classifier's accuracy.
- Not part of the training data

Testing Data

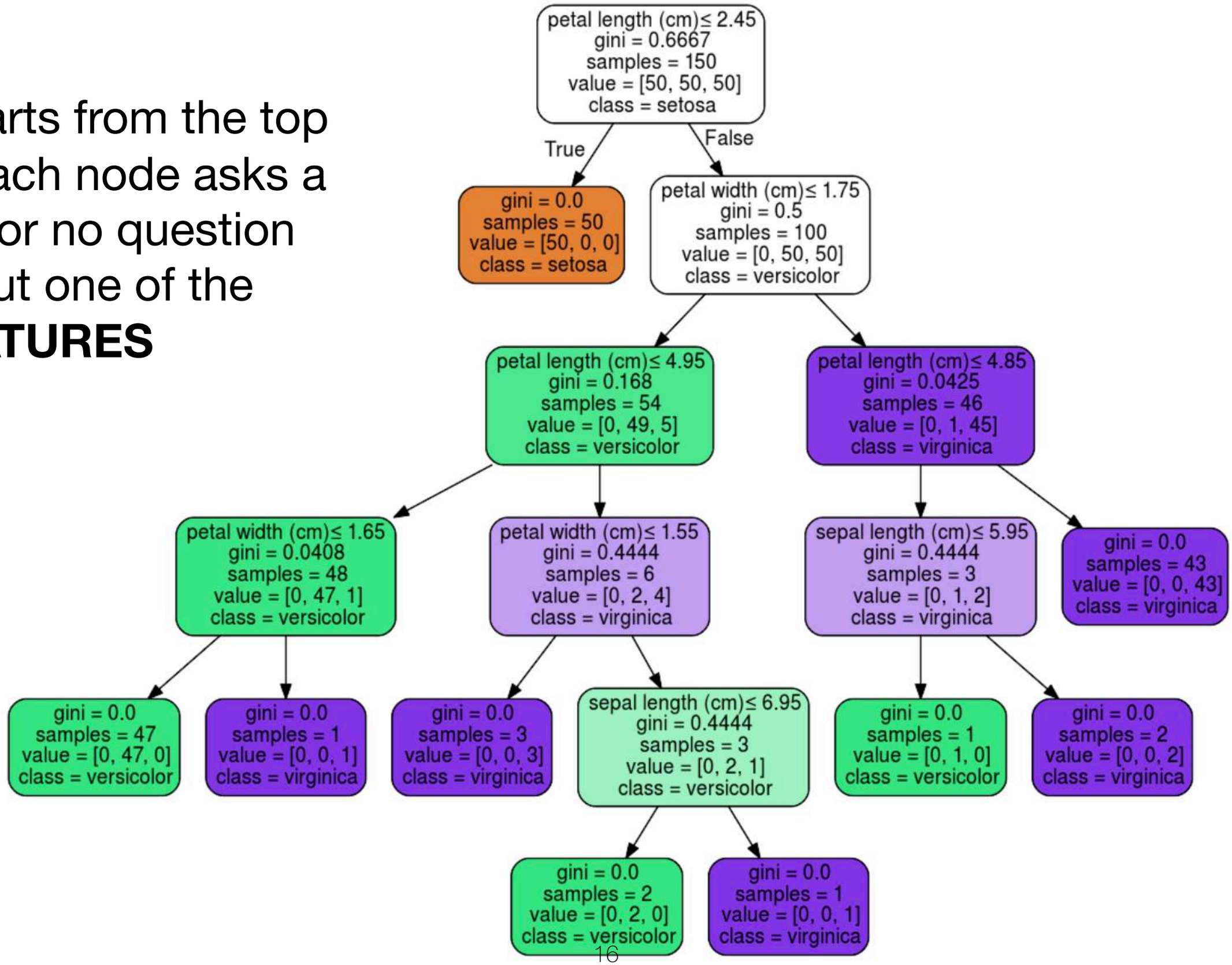
```
import numpy as np
from sklearn.datasets import load_iris
from sklearn import tree

iris = load_iris()
test_idx = [0,50,100]

#training data
train_target = np.delete(iris.target, test_idx)
train_data = np.delete(iris.data, test_idx, axis=0)

#testing data
test_target = iris.target[test_idx]
test_data = iris.data[test_idx]

clf = tree.DecisionTreeClassifier()
clf.fit(train_data, train_target)

print 'testing data:',(test_target)
print 'predicted label:', (clf.predict(test_data))
~
```

Output:

```
plus2:tutorials ssen$ python iris.py
testing data: [0 1 2]
predicted label: [0 1 2]
```

Decision Tree

- Starts from the top
→Each node asks a yes or no question about one of the **FEATURES**



More to Learn

- How are decision trees built automatically from examples?
- How well do they work in practice?
- The tree asks the questions about the features. Choosing good features is one of the important tasks in the first place.

How do we use machine learning in high energy particle physics?

A roadmap for ML in HEP

Machine Learning in High Energy Physics Community White Paper

May 17, 2019

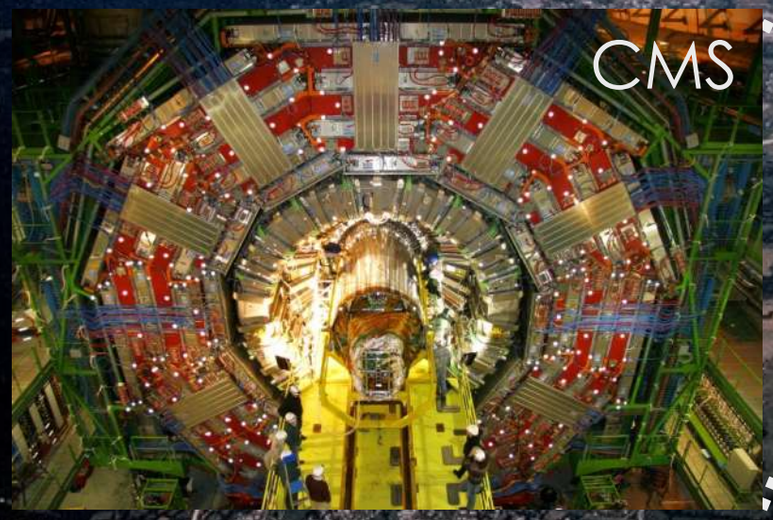
Data science and HEP communities

Abstract: Machine learning has been applied to several problems in particle physics research, beginning with applications to high-level physics analysis in the 1990s and 2000s, followed by an explosion of applications in particle and event identification and reconstruction in the 2010s. In this document we discuss promising future research and development areas for machine learning in particle physics. We detail a roadmap for their implementation, software and hardware resource requirements, collaborative initiatives with the data science community, academia and industry, and training the particle physics community in data science. The main objective of the document is to connect and motivate these areas of research and development with the physics drivers of the High-Luminosity Large Hadron Collider and future neutrino experiments and identify the resource needs for their implementation. Additionally we identify areas where collaboration with external communities will be of great benefit.

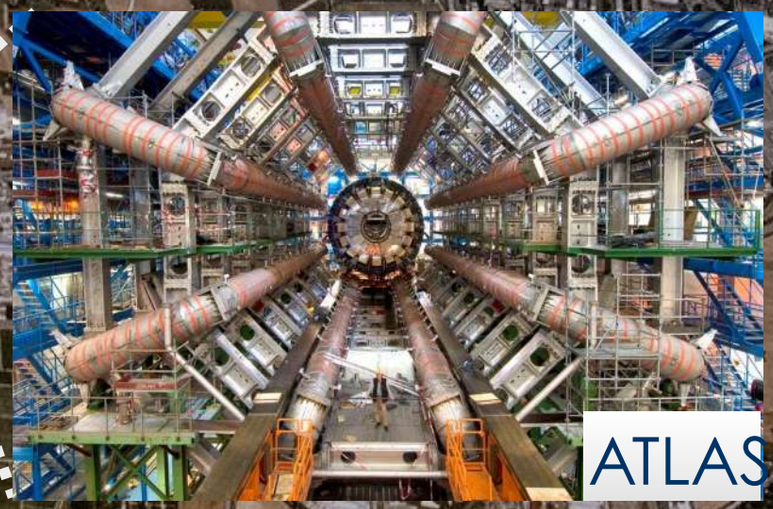
Editors: Sergei Gleyzer[30], Paul Seyfert[13], Steven Schramm[32]

Contributors: Kim Albertsson[1], Piero Altoe[2], Dustin Anderson[3], John Anderson[4], Michael Andrews[5], Juan Pedro Araque Espinosa[6], Adam Aurisano[7], Laurent Basara[8], Adrian Bevan[9], Wahid Bhimji[10], Daniele Bonacorsi[11], Bjorn Burkle[12], Paolo Calafiura[10], Mario Campanelli[9], Louis Capps[2], Federico Carminati[13], Stefano Carrazza[13], Yi-Fan Chen[4], Taylor Childers[14], Yann Coadou[15], Elias Coniavitis[16], Kyle Cranmer[17], Claire David[18], Douglas Davis[19], Andrea De Simone[20], Javier Duarte[21], Martin Erdmann[22], Jonas Eschle[23], Amir Farbin[24], Matthew Feickert[25], Nuno Filipe Castro[6], Conor Fitzpatrick[26], Michele Floris[13], Alessandra Forti[27], Jordi Garra-Tico[28],

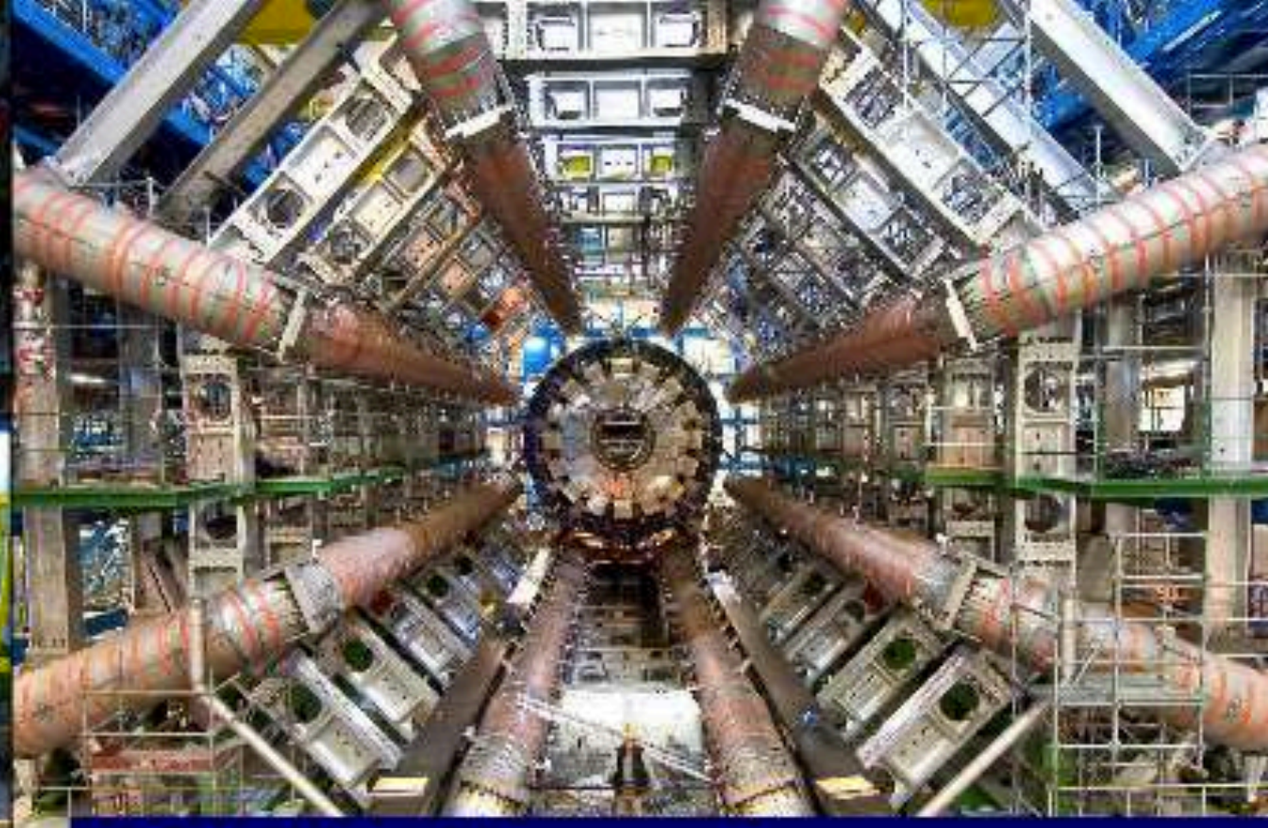
2010: a New Era in Fundamental Science



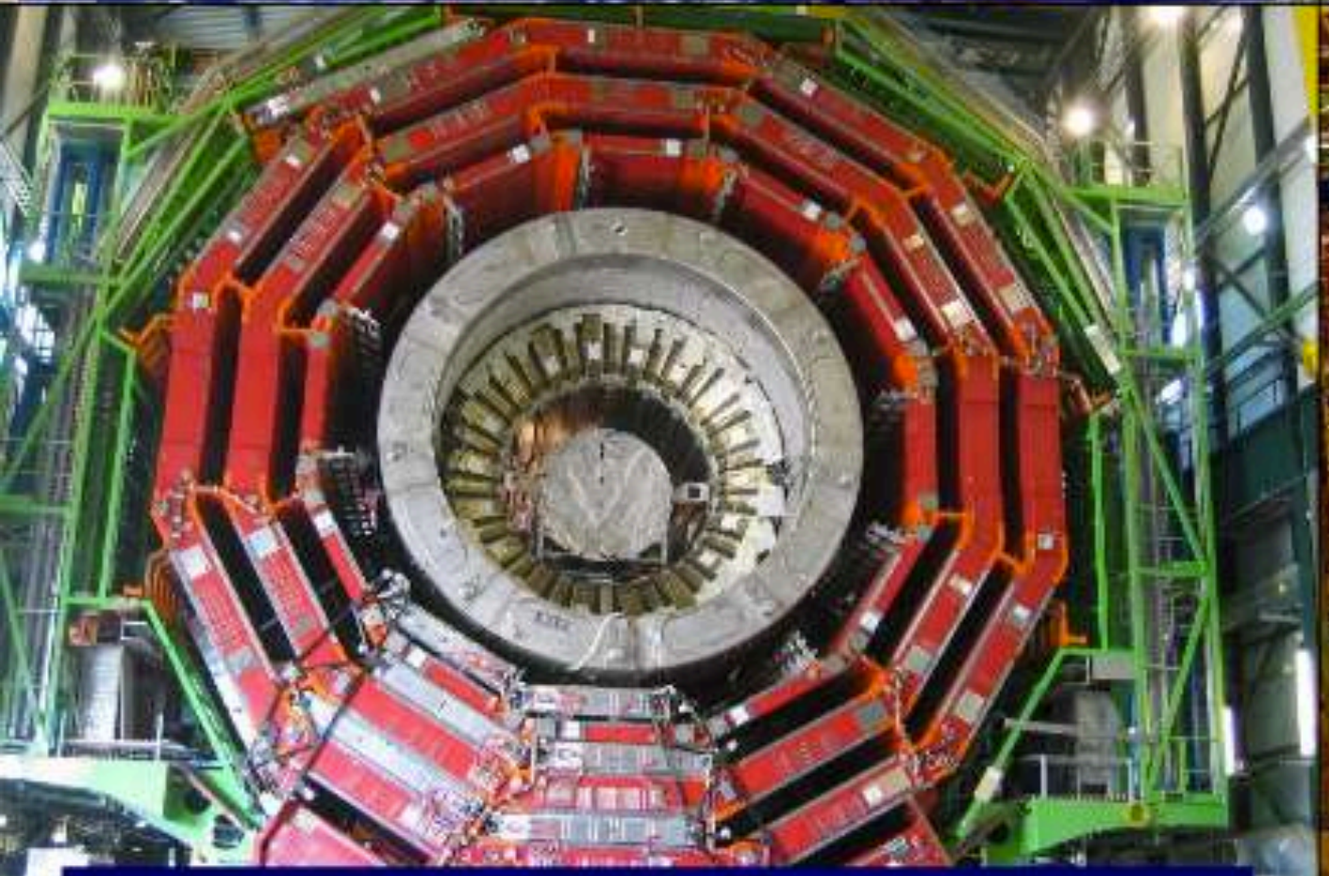
Exploration of a new energy frontier
Proton-proton and Heavy Ion collisions
at E_{CM} up to 14 TeV



ALICE: Primordial cosmic plasma



ATLAS: Higgs and supersymmetr



CMS: Higgs and supersymmetry

LHCb: Matter-antimatter differen

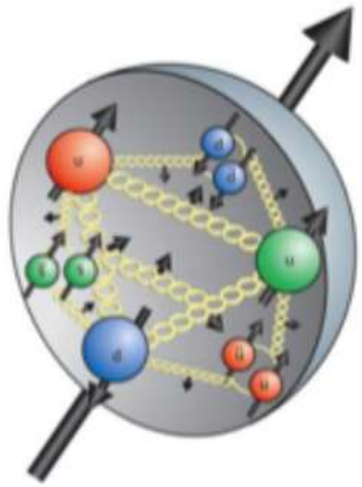
Highly heterogeneous system
Raw data is 100 M channels sampled every 25 ns.

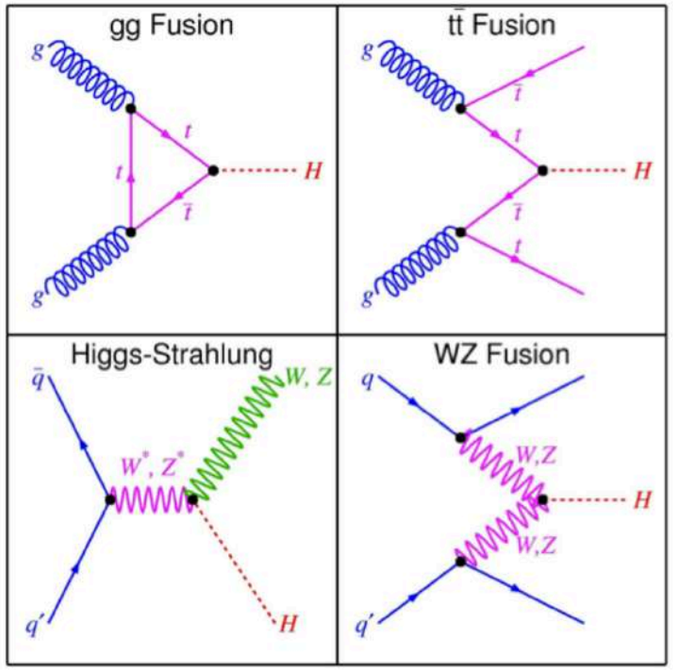
4 JULY 2012
CERN Press conference

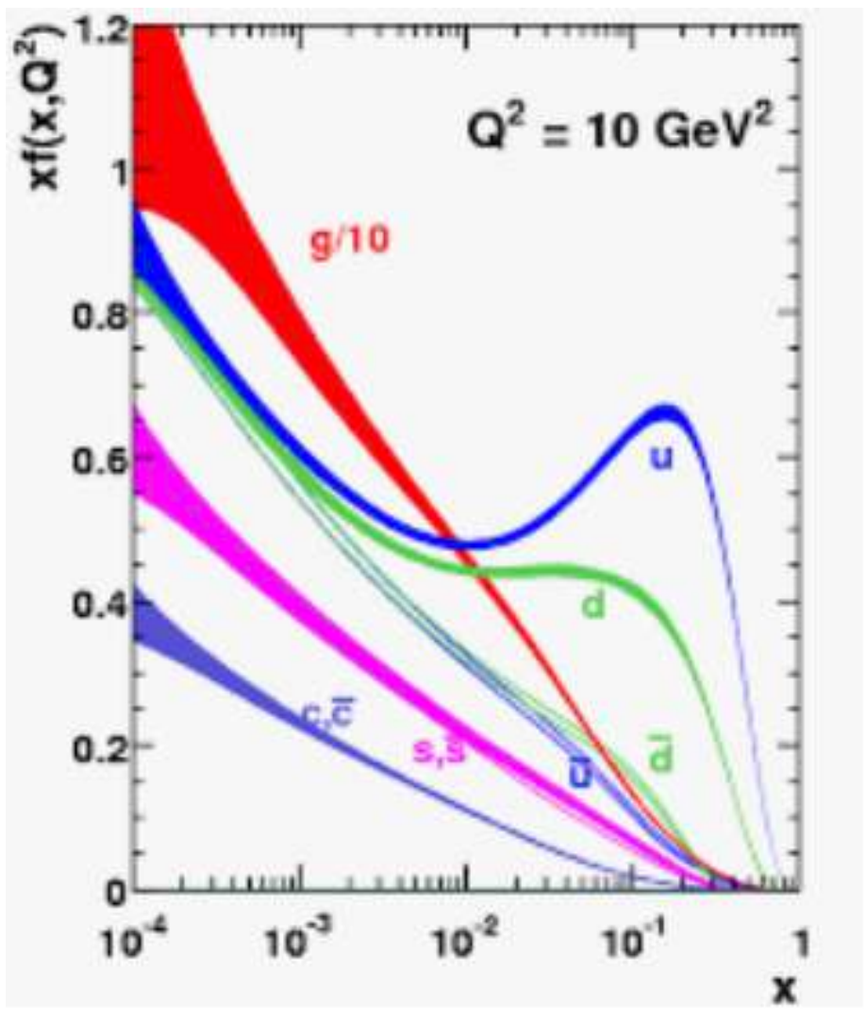
Nobel Prize in Physics 2013

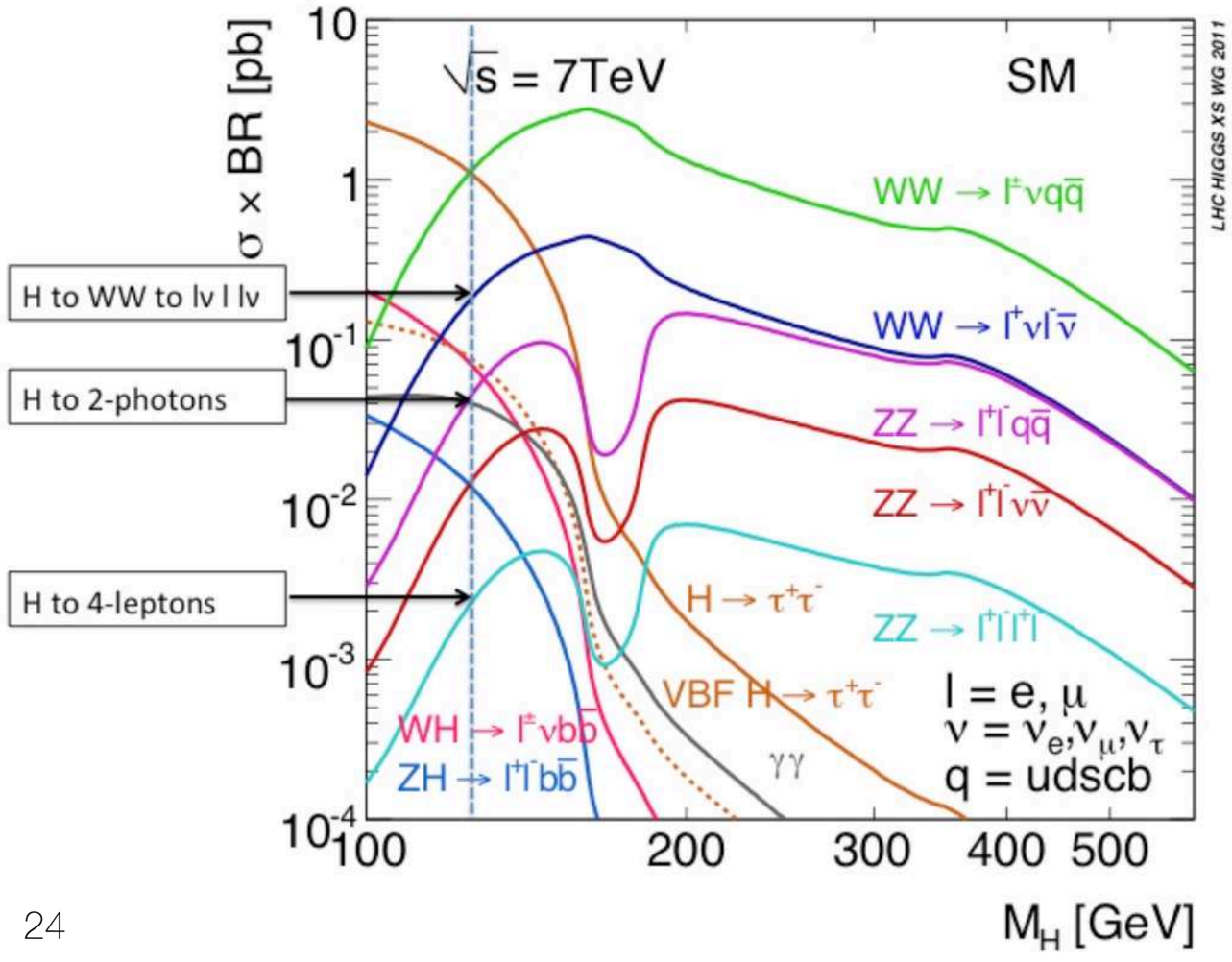
The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider".*

Higgs production and decays at the LHC









Higgs Boson Machine Learning Challenge 2014

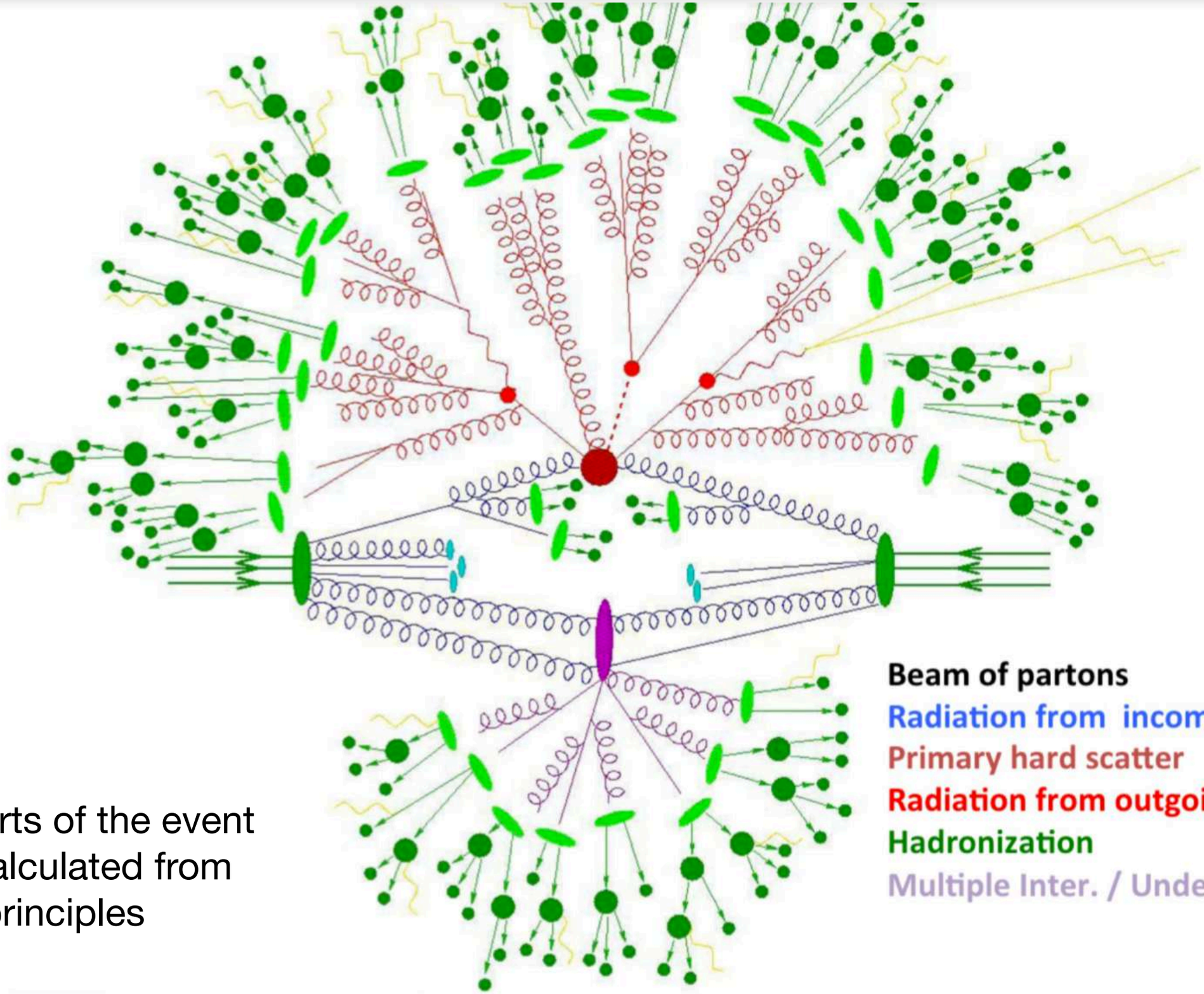
~1 million events (rows)
35 features (coloumns)

EventId	DER_mass_MI	DER_mass_tran	DER_mass_vis	DER_pt_h	DER_deltaet	DER_mass_jet_jet	DER_prod	PRI_jet_subleading_phi	PRI_jet_all_pt	Weight	Label	KaggleSet	KaggleWeight
100000	138.47	51.655	97.827	27.98	0.91	124.711		-2.475	113.497	0.00081448039868	s	t	0.00265331133733
100001	160.937	68.768	103.235	48.146	-999.0	-999.0		-999.0	46.226	0.681041906806	b	t	2.23358448717
100002	-999.0	162.172	125.953	35.635	-999.0	-999.0		-999.0	44.251	0.715742006349	b	t	2.34738894364
100003	143.905	81.417	80.943	0.414	-999.0	-999.0		-999.0	0.0	1.66065435355	b	t	5.44637821192
100004	175.864	16.915	134.805	16.405	-999.0	-999.0		-999.0	0.0	1.90426344118	b	t	6.24533268686
100005	89.744	13.55	59.149	116.344	2.636	284.584		3.106	193.66	0.0254337596084	b	t	0.0834140312717
100006	148.754	28.862	107.782	106.13	0.733	158.359		-2.767	179.877	0.00081448039868	s	t	0.00265331133733
100007	154.916	10.418	94.714	29.169	-999.0	-999.0		-999.0	30.638	0.00572068250088	s	t	0.018636116672
100008	105.594	50.559	100.989	4.288	-999.0	-999.0		-999.0	0.0	1.614803466	b	t	5.29600298518
100009	128.053	88.941	69.272	193.392	-999.0	-999.0		-999.0	167.735	0.000461025356734	s	t	0.00150187015894
100010	-999.0	86.24	79.692	27.201	-999.0	-999.0		-999.0	0.0	0.70114133537	b	t	2.29950373735
100011	114.744	10.286	75.712	30.816	2.563	252.599		-2.079	165.64	0.0936590129021	b	t	0.307169523947
100012	145.297	64.234	103.565	106.999	-999.0	-999.0		-999.0	93.117	0.512739889611	b	t	1.68161144262
100013	82.488	31.663	64.128	8.232	-999.0	-999.0		-999.0	0.0	0.665890037882	b	t	2.18389154017
100014	-999.0	109.412	14.398	17.323	-999.0	-999.0		-999.0	0.0	0.655921659265	b	t	2.15119866823
100015	111.026	32.096	75.271	23.067	-999.0	-999.0		-999.0	0.0	0.00572068250088	s	t	0.018636116672
100016	114.256	4.351	67.963	47.221	-999.0	-999.0		-999.0	36.263	0.443597626883	b	t	1.45484847268
100017	127.861	50.953	77.267	26.967	-999.0	-999.0		-999.0	0.0	0.0004612815739 49	s	t	0.00150270483101
100018	-999.0	85.186	68.827	5.042	-999.0	-999.0		-999.0	0.0	1.56163346758	b	t	5.12162357847
								-999.0	0.0	1.82316290593	b	t	5.97935067368

Training data
Events are labeled as "signal" (Higgs) or "background"
Train a classifier and predict the signal events

http://opendata.cern.ch/record/328

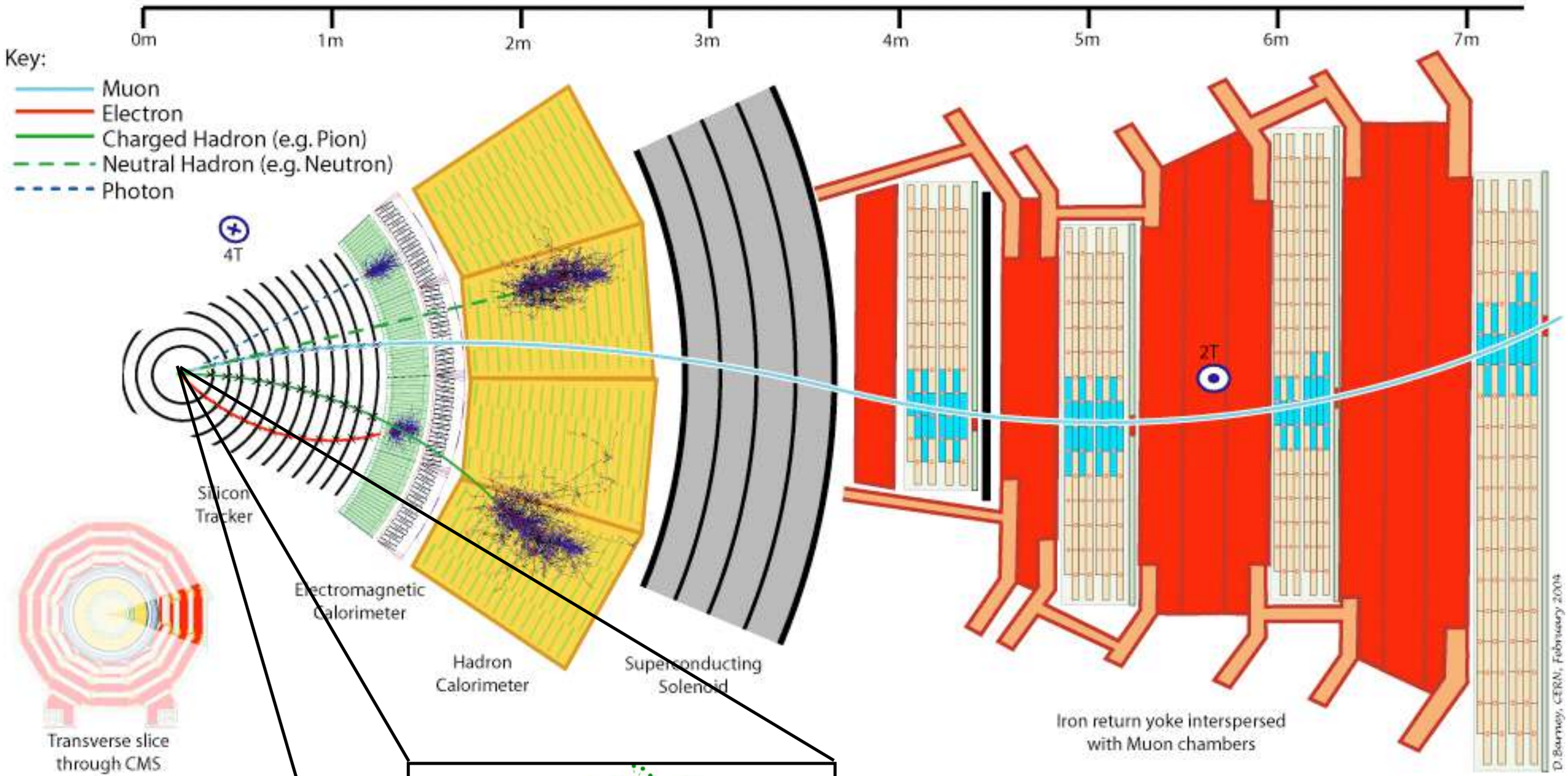
Typical proton-proton collision



not all parts of the event can be calculated from the first principles

Add 40 such on top of each other currently. Up to 200 such overlay in the horizon 2025.

Jets: any particle decaying in quark/gluons will result in a “jet” of particles in the direction of the original particle.



Particles leave hints of their passage in sub-detectors. Specific (but overlapping) pattern for each particle type.

10-100 billion events/year

6 m

2 m

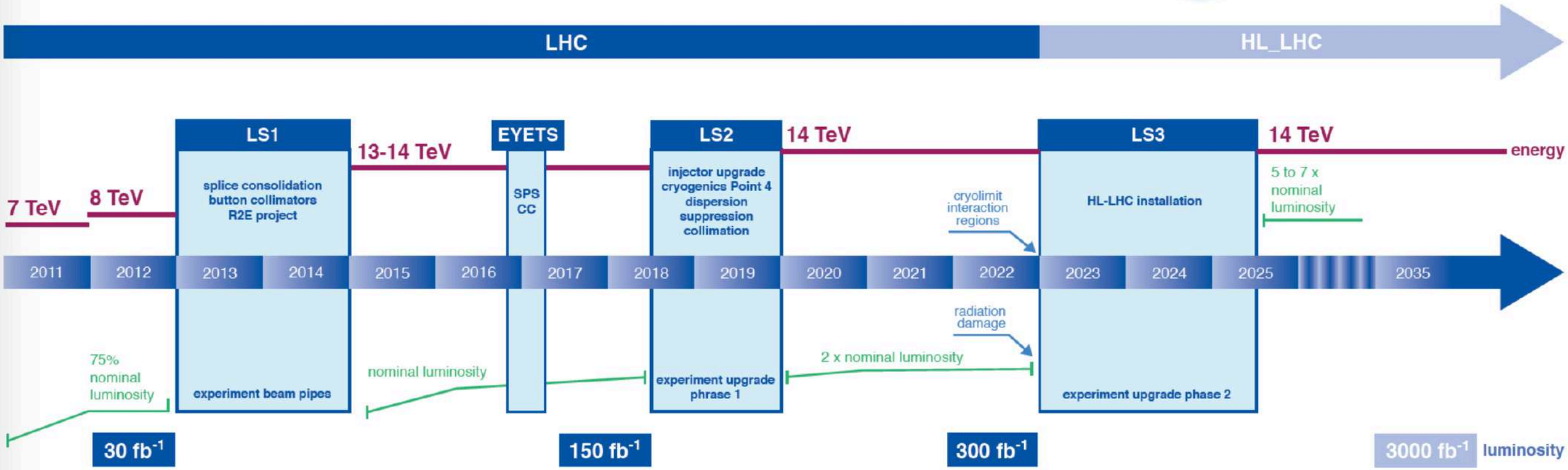
Point precision ~5 μm to 3mm

100k points 10k tracks / event

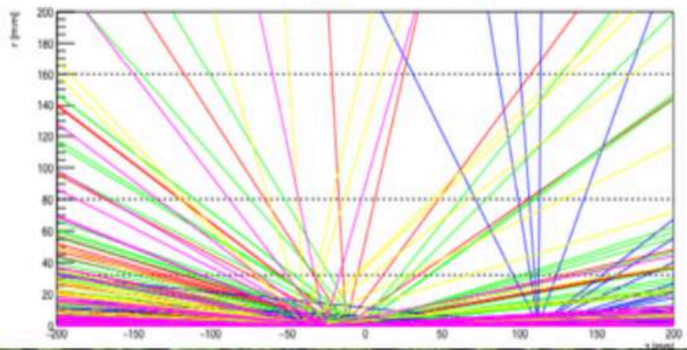
image: David Rousseau

LHC / HL-LHC Plan

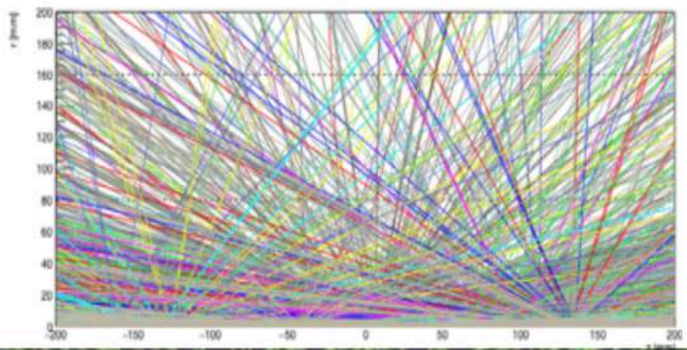
High Luminosity LHC



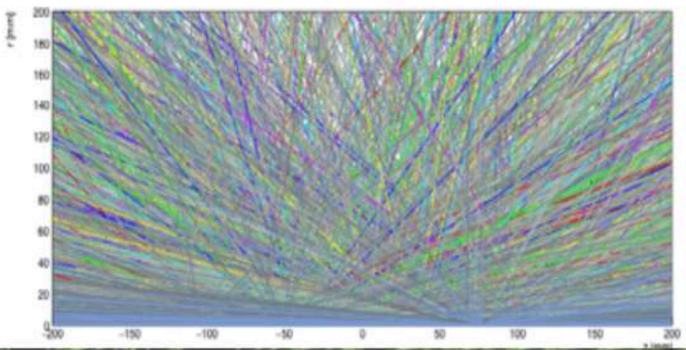
5 p-p collisions
LHC early Run-1 2010

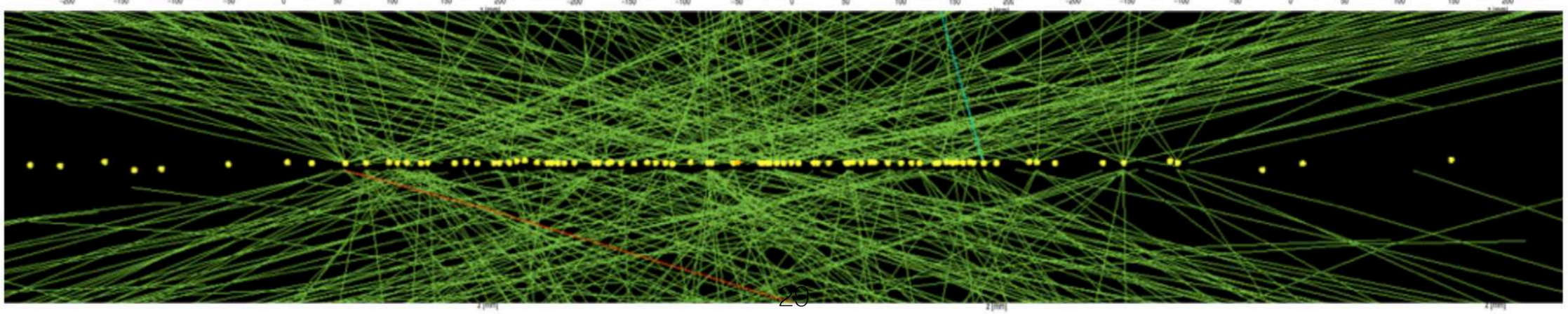


40 p-p collisions
LHC early Run-2 2015/16



200 p-p collisions
HL-LHC conditions

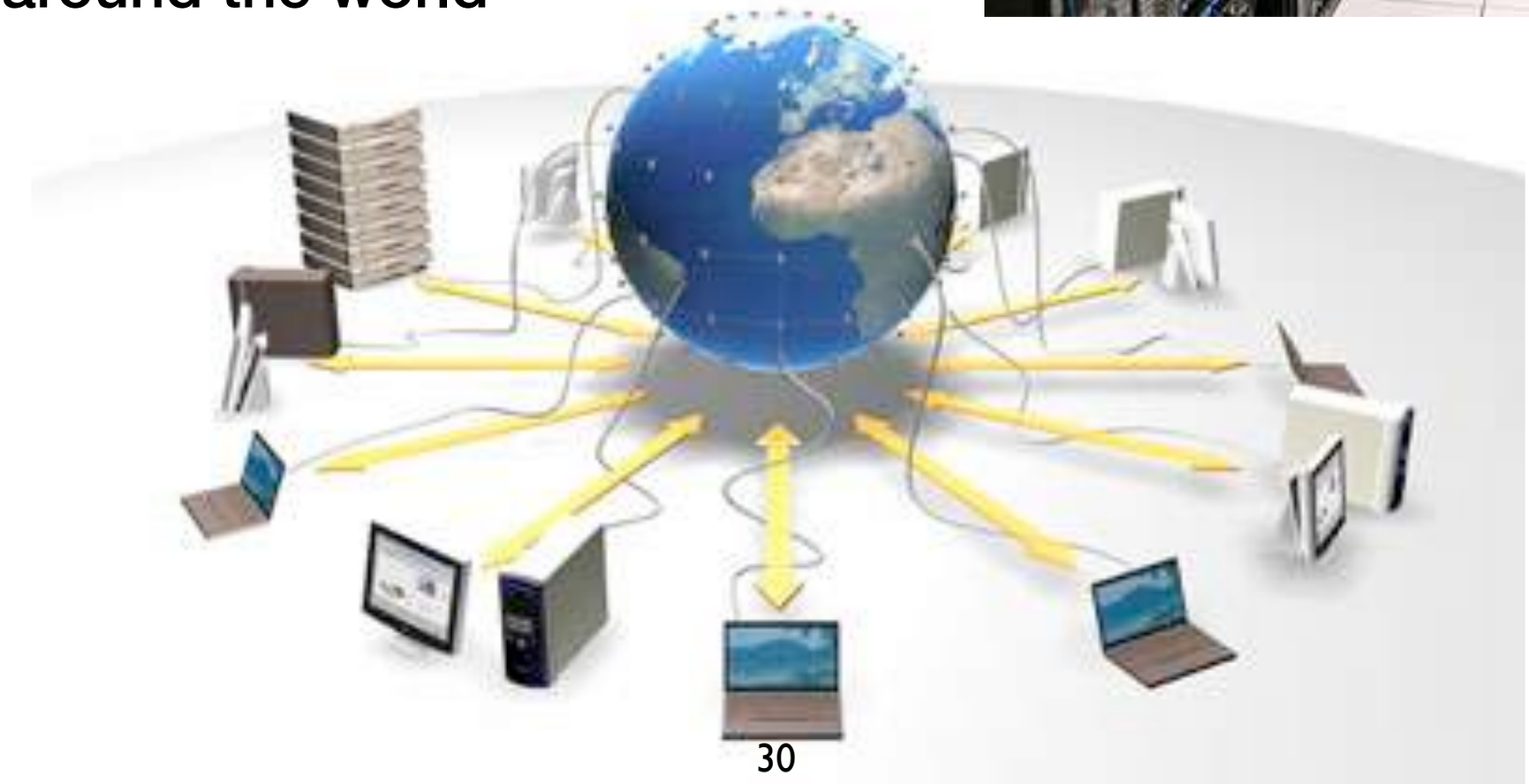




Grid CERN

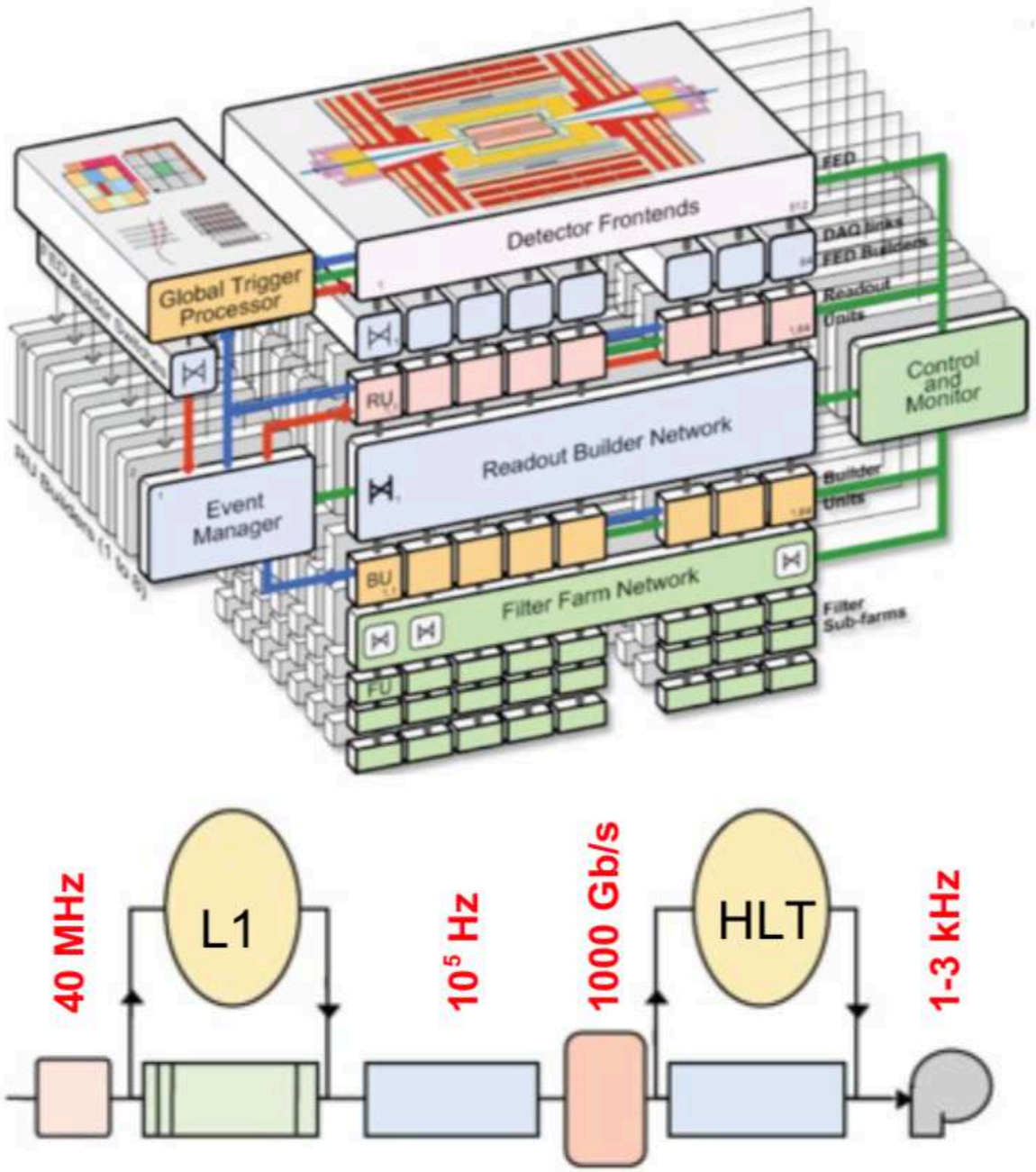
World Wide Web: an information system where documents interlinked by hypertext, and accesible over the internet.

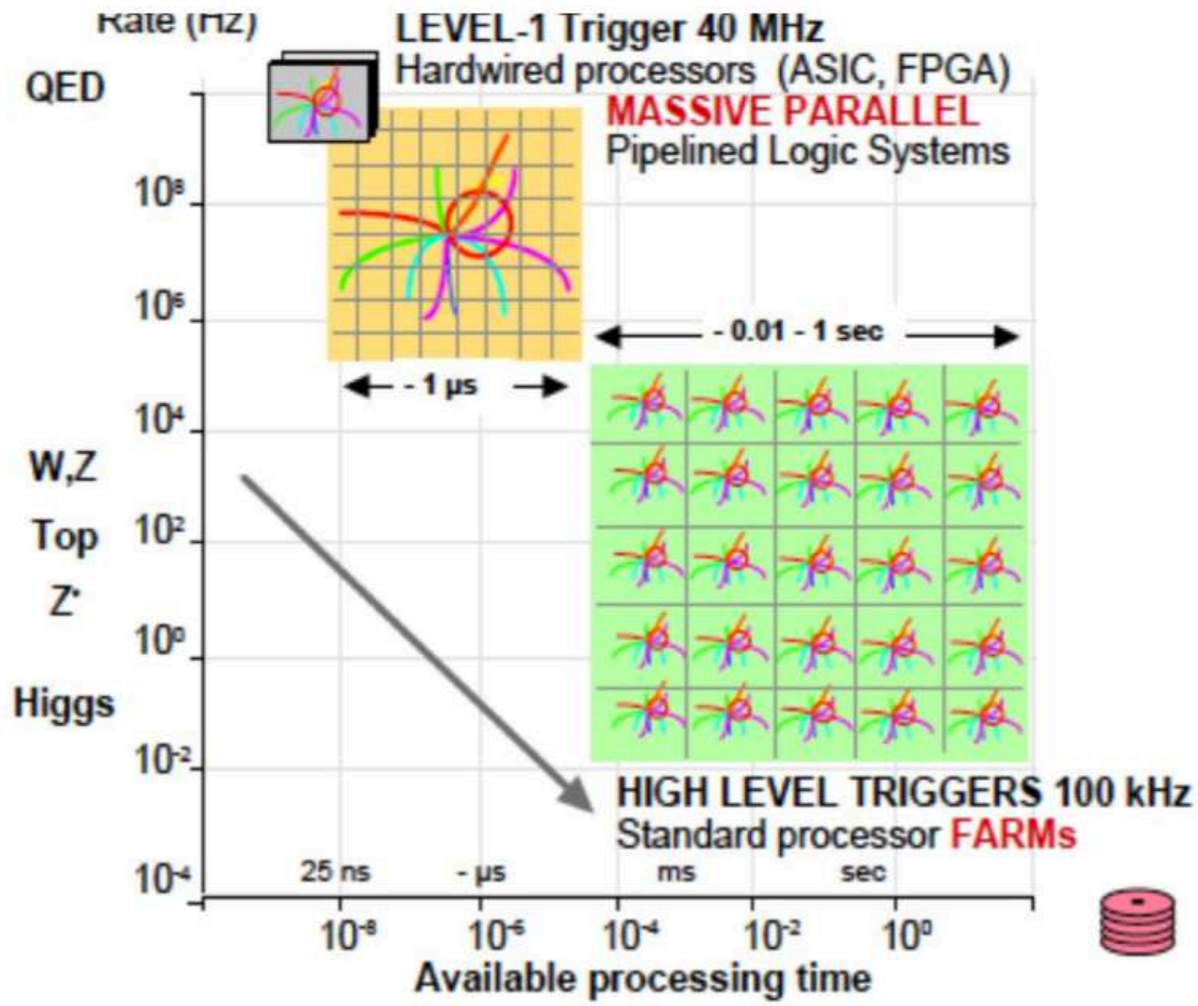
LHC Grid: A global collaboration of computer centres distributes and stores LHC data, giving real-time access to physicists around the world



Running jobs: 268149
Transfer rate: 11.38 GiB/sec

Event Filtering

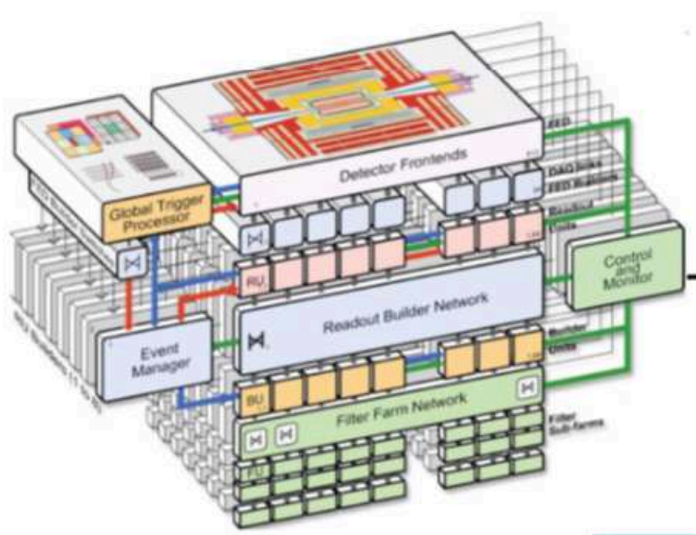




Ultra fast decision to keep the relevant data.
In hardware and software.

image: Jean-Roch Vlimant

Computing Grid



300-1500 MB/s

Tier0 CERN Computer Center

10 – 40 to 100 Gb/s

IN2P3 T1 ... FNAL T1 ... INFN T1

10 – 40 to 100 Gb/s

10 – 40 to 100 Gb/s

T2 T2 T2 T2 T2 T2 T2 T2 T2 T2 T2 T2 T2 T2 T2

10 to Nx10 Gb/s

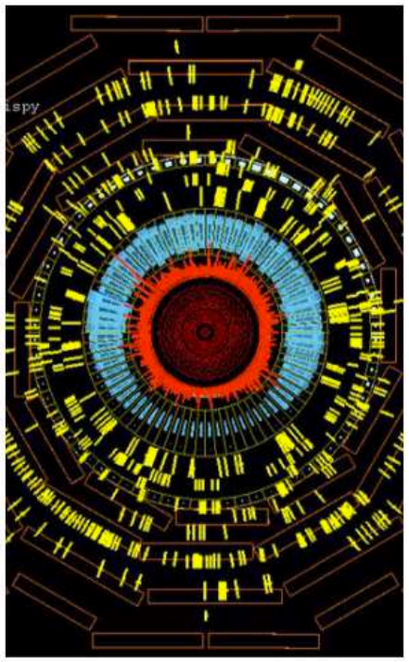
T3 T3 T3 T3

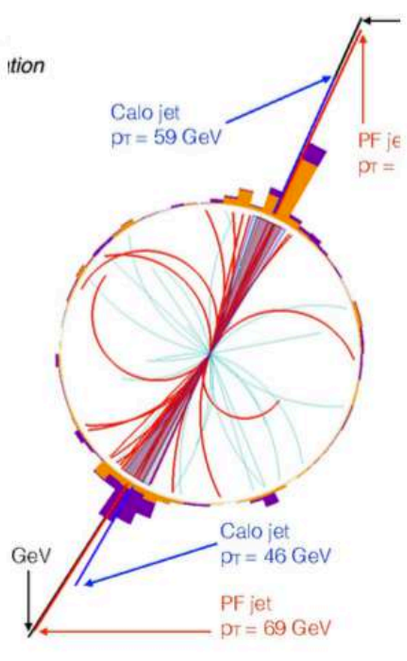
- **Hundreds of computer centers** (100-10k cores per site)
- Increased use as a cloud resources (any job anywhere)
- Increasing use of additional cloud and HPC resource
- Real time data processing at Tier0
- Data and Simulation production at Tier1 and Tier2
- High bandwidth networks between disk storage

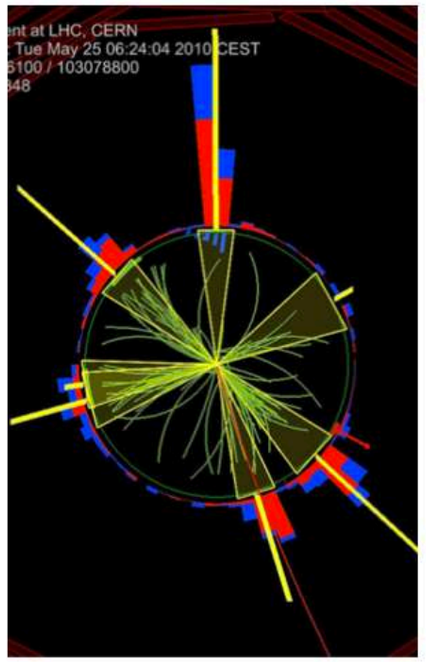
image: Jean-Roch Vlimant

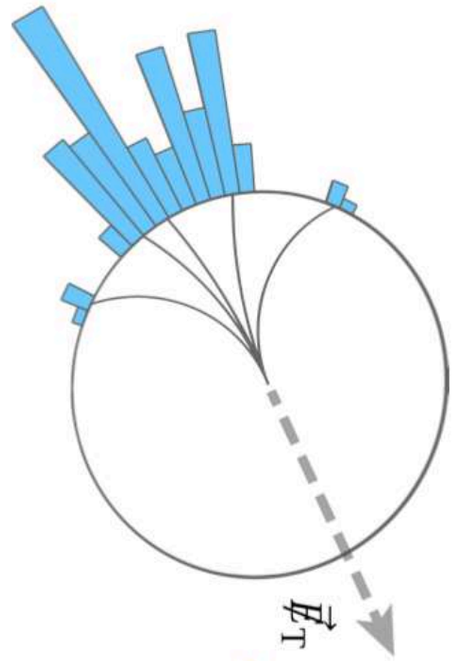
From Low to High Level Data

Detector Data | Local reconstruction | Particle representation | Jet Clustering | High level features









Event Processing

Dimensionality reduction

Globalization of information

The reconstruction of an event goes from the digital signal of the individual sub-detector to a sequence of particles, jets, and high-level features

RAW → RECO → AOD → NanoAOD

image: Jean-Roch Vlimant

Who can access the data?

CERN Open Data portal

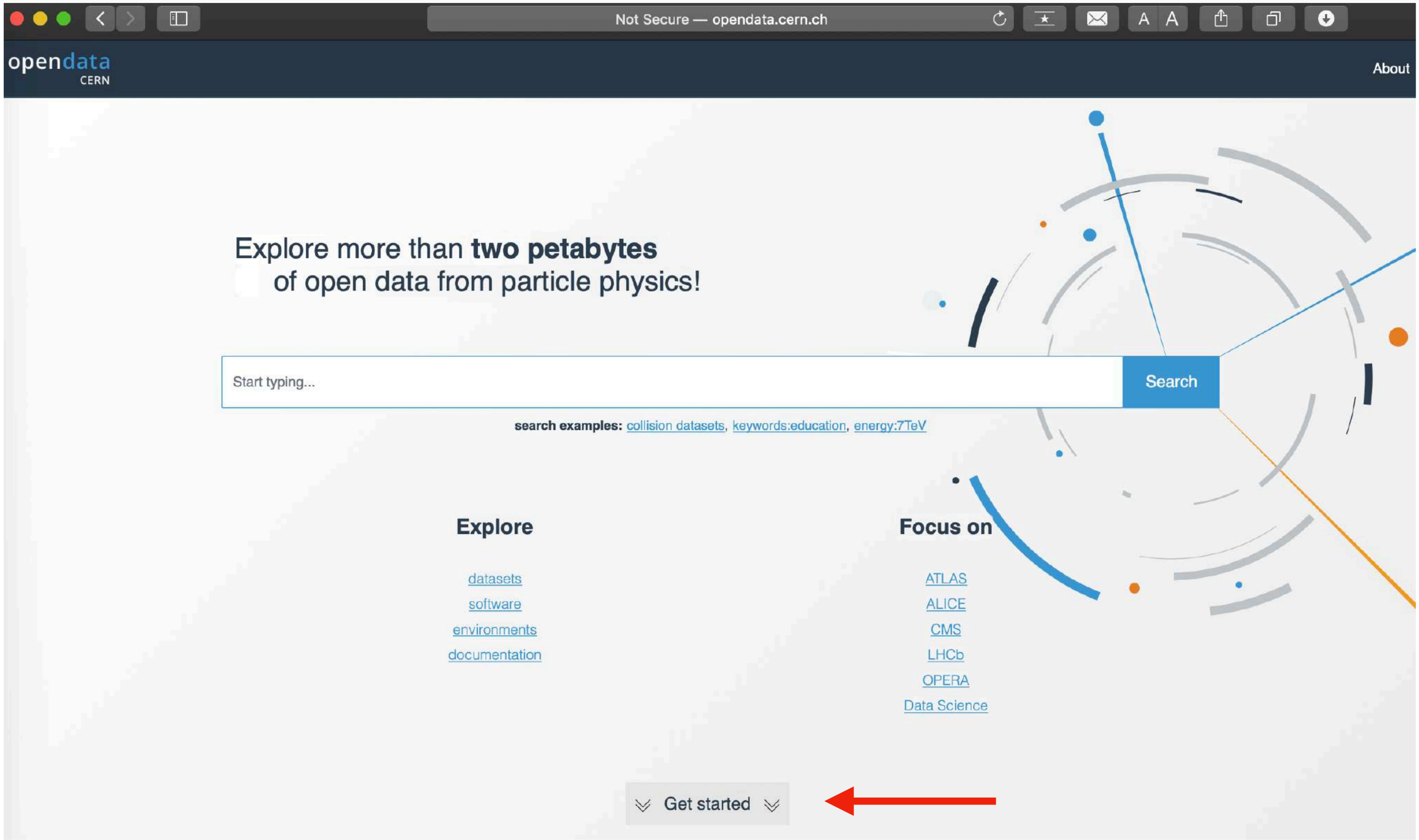
- opendata.cern.ch launched in November 2014
- LHC collaboration data policies
  - restricted → embargo period (~5 years) → open
- Over 1.5 Petabytes of open particle physics data
  - datasets, software, VMs, configuration, documentation,...
- Users
  - education: general public, high-school students, masterclasses
  - research: data scientists, physicists

☑ Openly accesible data
☑ Reproducible analysis examples
☑ Lightweight and easily readable data format
☑ Good usability for non-physics experts

Developed by CERN-IT and CERN-SIS

in collaboration with Experiments

CERN Open Data portal



http://opendata.cern.ch

http://github.com/cernopendata

CERN Open Data portal

Learn

Discover the world of open data from particle physics

Welcome to our updated portal

CMS Guide to education use of CMS Open Data

Improving educational content with high school teachers: A field report from our summer students

Glossary

more

Visualise

Explore detector events and run basic histogramming

CMS Event Display

OPERA Event Display

CMS Histograms

Analyse

Run your own physics analyses, start virtual machines

CMS Guide to research use of CMS Open Data

ATLAS Higgs Machine Learning Challenge

Getting Started with LHCb Open Data

Getting Started with ALICE Open Data

more

News

News & Update

2020-02-11 by ATLAS Collaboration

ATLAS Experiment releases 13 TeV Open Data for Science Education

News ATLAS

2019-07-18 by CMS Collaboration

CMS releases open data for Machine Learning

News CMS

2018-05-22 by OPERA Collaboration

Release of the first set of data samples by the OPERA Collaboration

News OPERA

2017-12-20 by CMS Collaboration

Observing the Higgs with over one petabyte of new CMS Open Data

News CMS

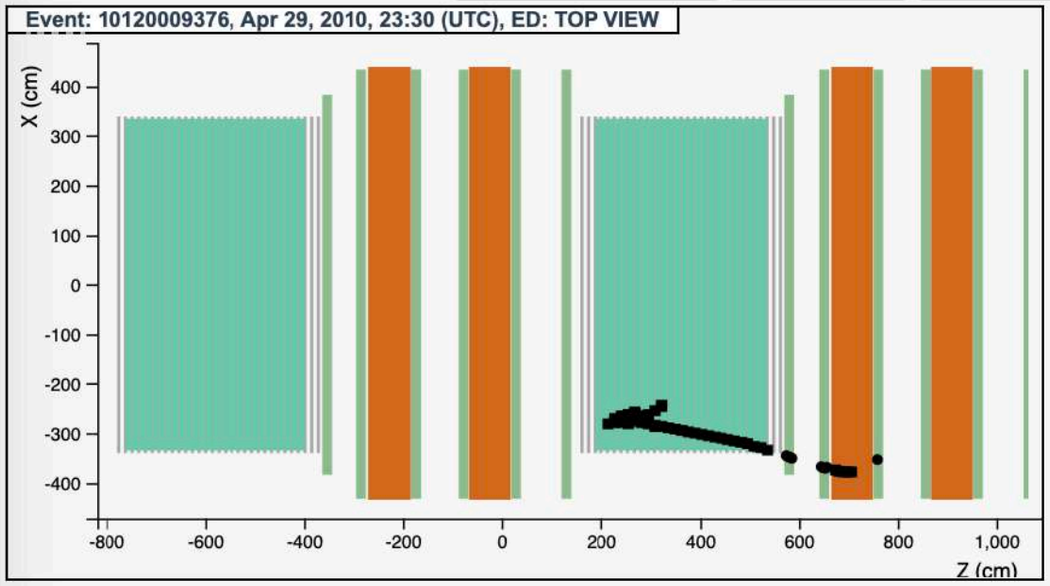
First non-LHC data releases

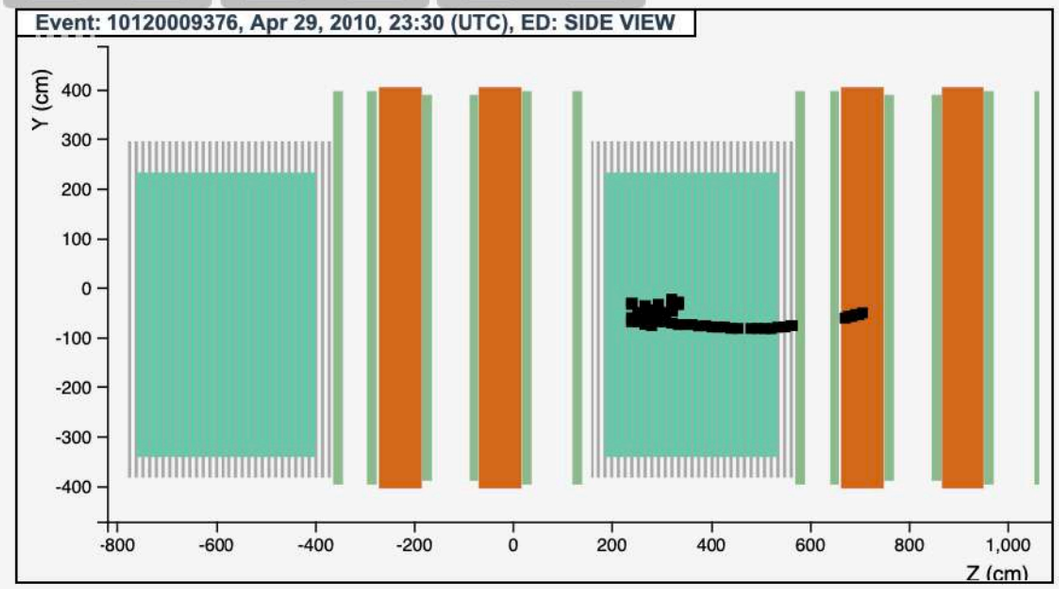
opendata CERN

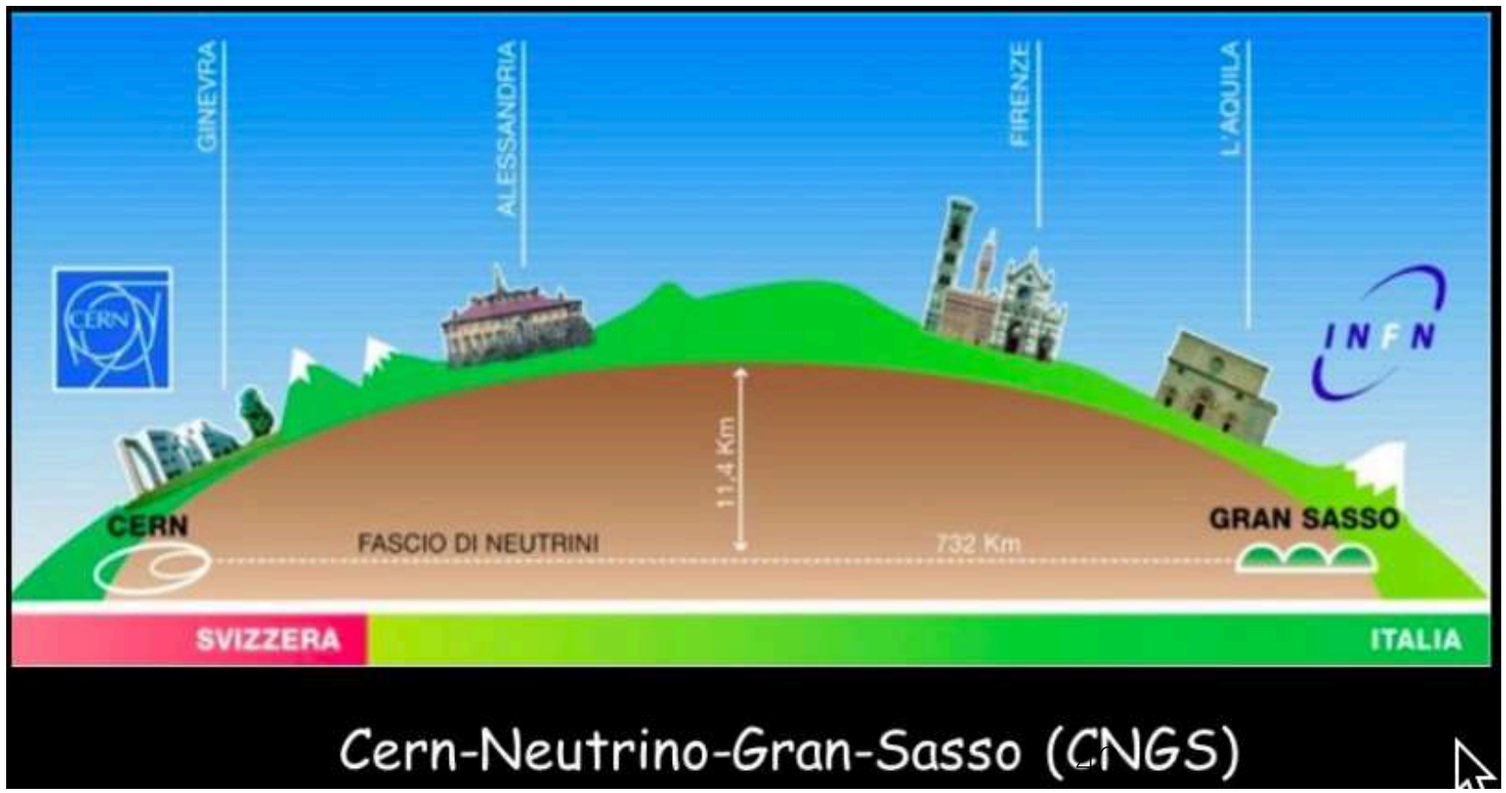
Search

OPERA electronic detectors (ED) display

nuMu Event: 10120009376







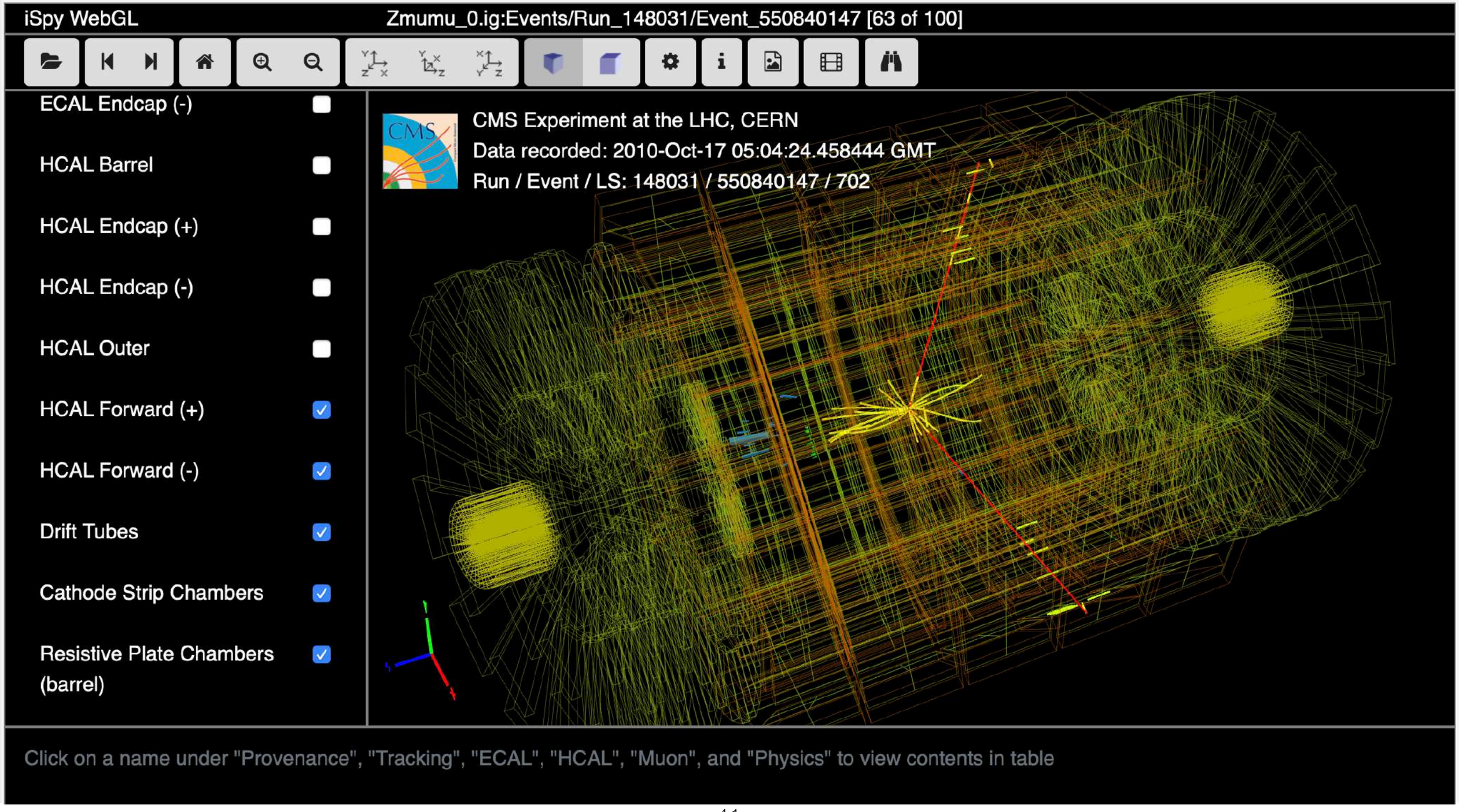
OPERA tau neutrino candidate and multiplicity studies data

Visualize collisions

opendata CERN

Search

Need HELP?



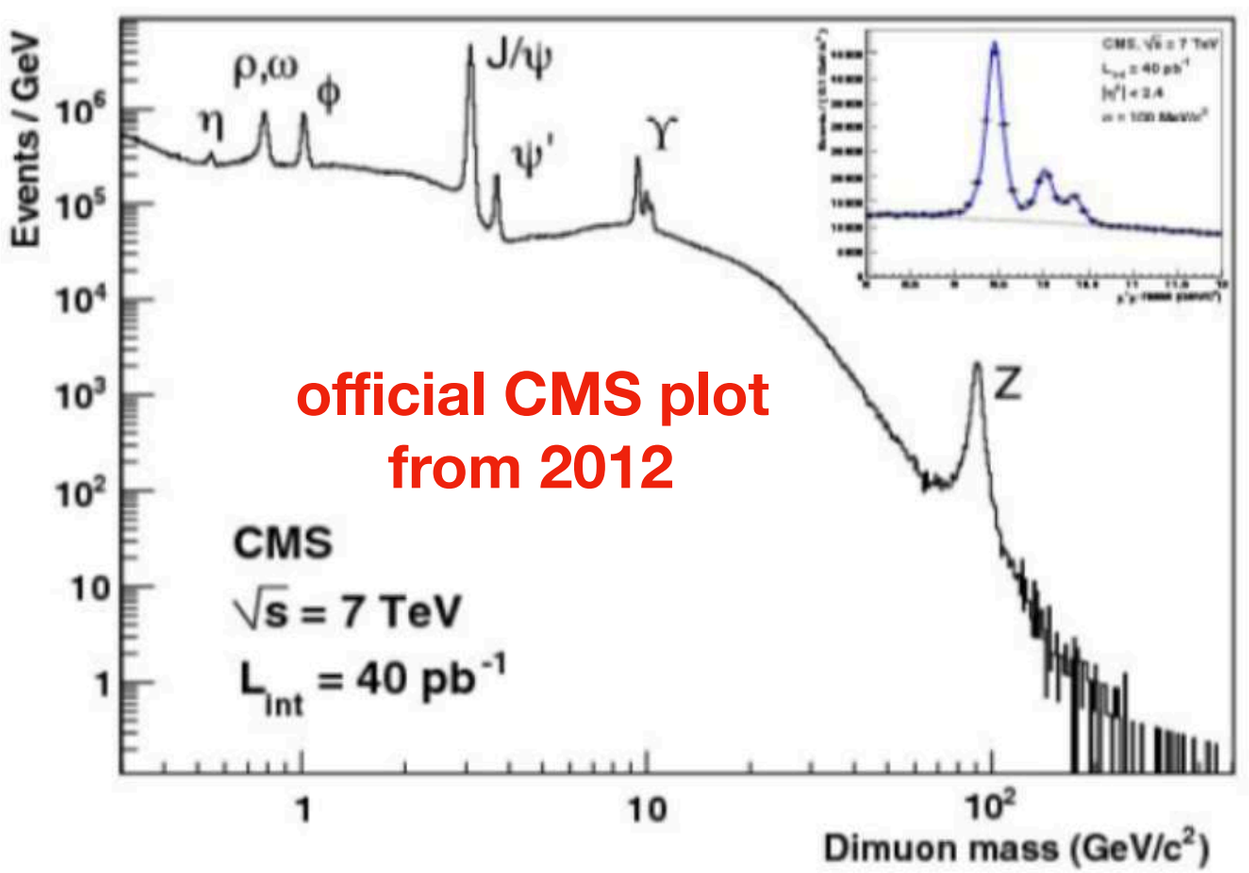
Interactive histogramming

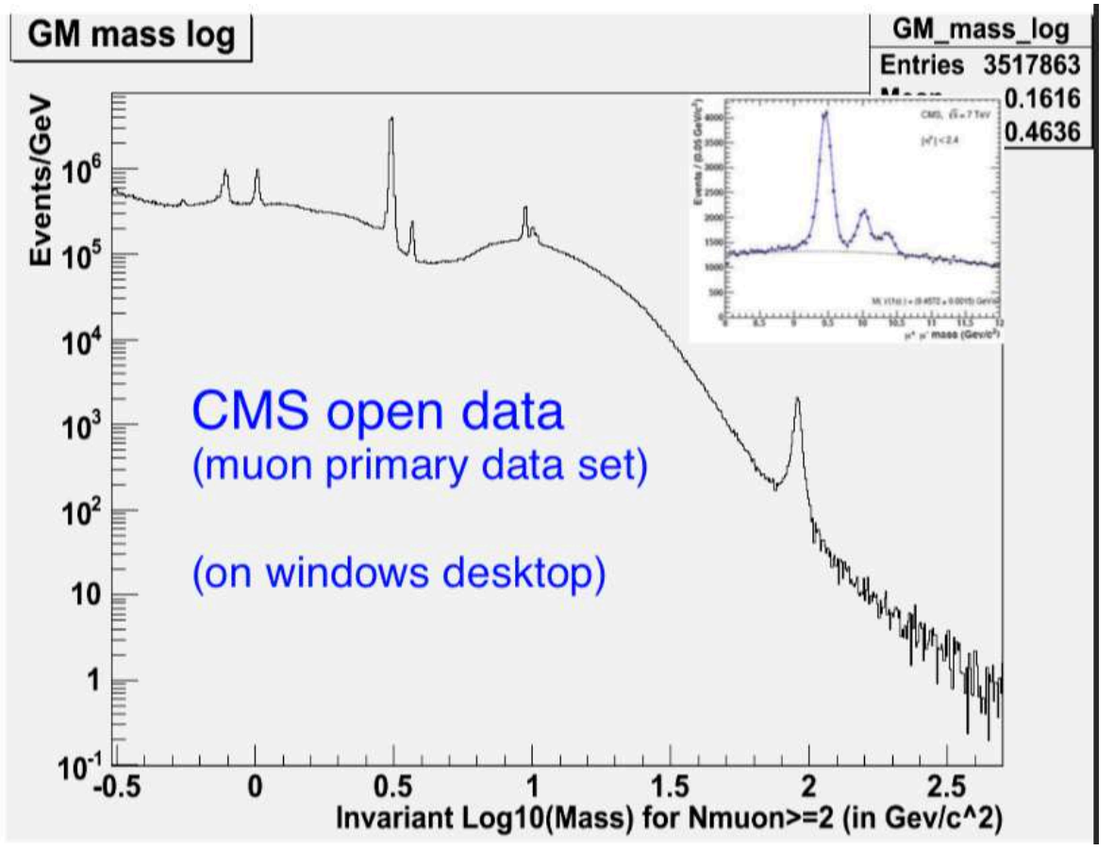
online histogramming for high-level derived datasets

Dive deeper into the data (I)

Analysis of the di-muon spectrum using data from the CMS detector taken in 2012.

- Rediscover particle resonances in a wide energy range up to the Z boson.
- About 62 M events from data taken at CMS in 2012.
- Only ROOT as dependency.
- NanoAOD format.
- Analysis code in Python, C++ or a Jupyter notebook.





Dive deeper into the data (II)

opendata CERN | Search | About

Higgs-to-four-lepton analysis example using 2011-2012 data

Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin;

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JKB8.RR42

Software | Analysis | CMS | Accelerator CERN-LHC

Description

This research level example is a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis published in Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4l_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

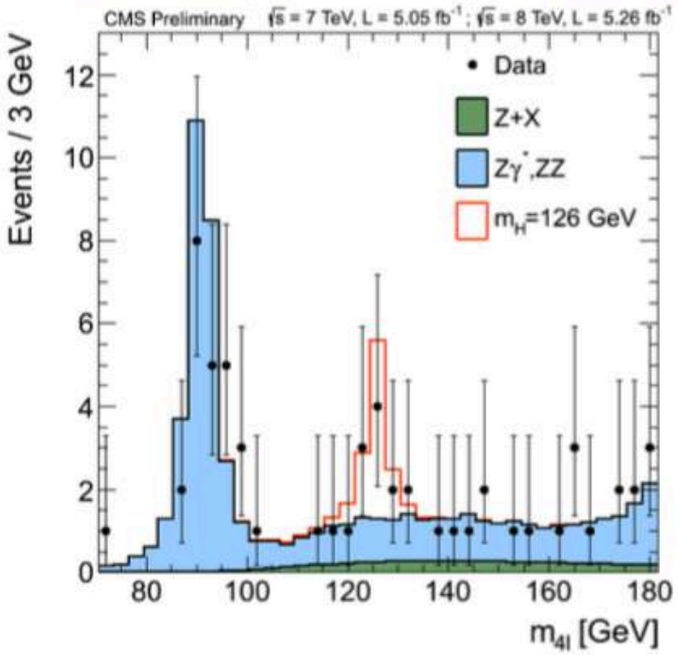
The example consists of different levels of complexity. The highest leve
minimal understanding of the content of this paper and of the meanin
educational exercises. The lower levels might also be interesting for ed
with the linux operating system and the ROOT analysis tool.

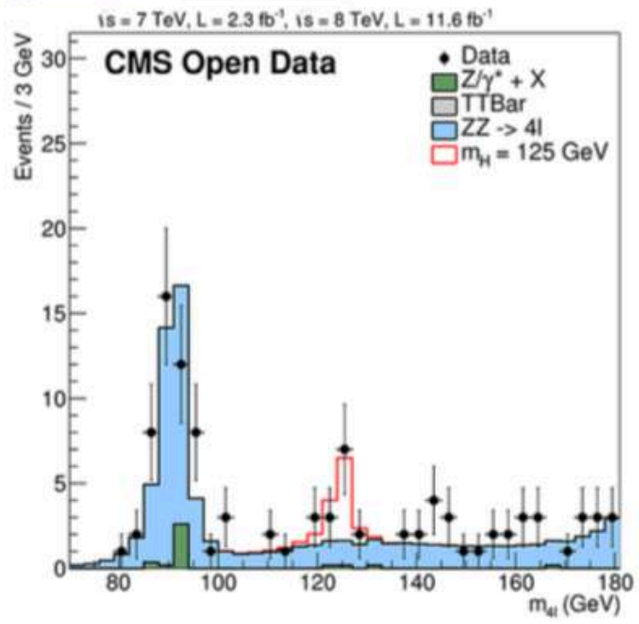
Use with

The example uses legacy versions of the original CMS datasets in the A
publication due to improved calibrations. It also uses legacy versions o
but not identical to, the ones in the original publication. These legacy d
in many later CMS publications.

/DoubleElectron/Run2011A-12Oct2013-v1/AOD

/DoubleMu/Run2011A-12Oct2013-v1/AOD

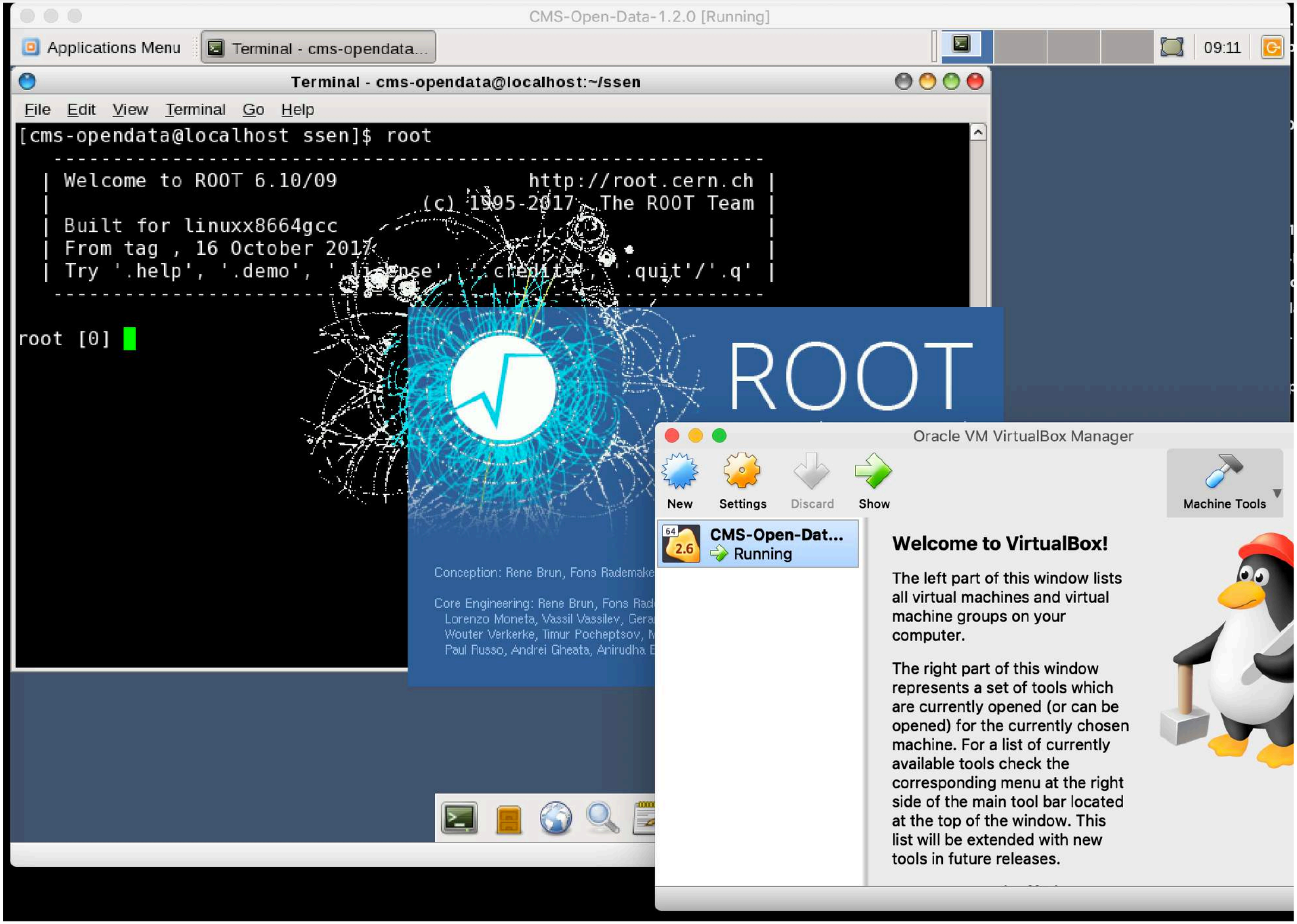




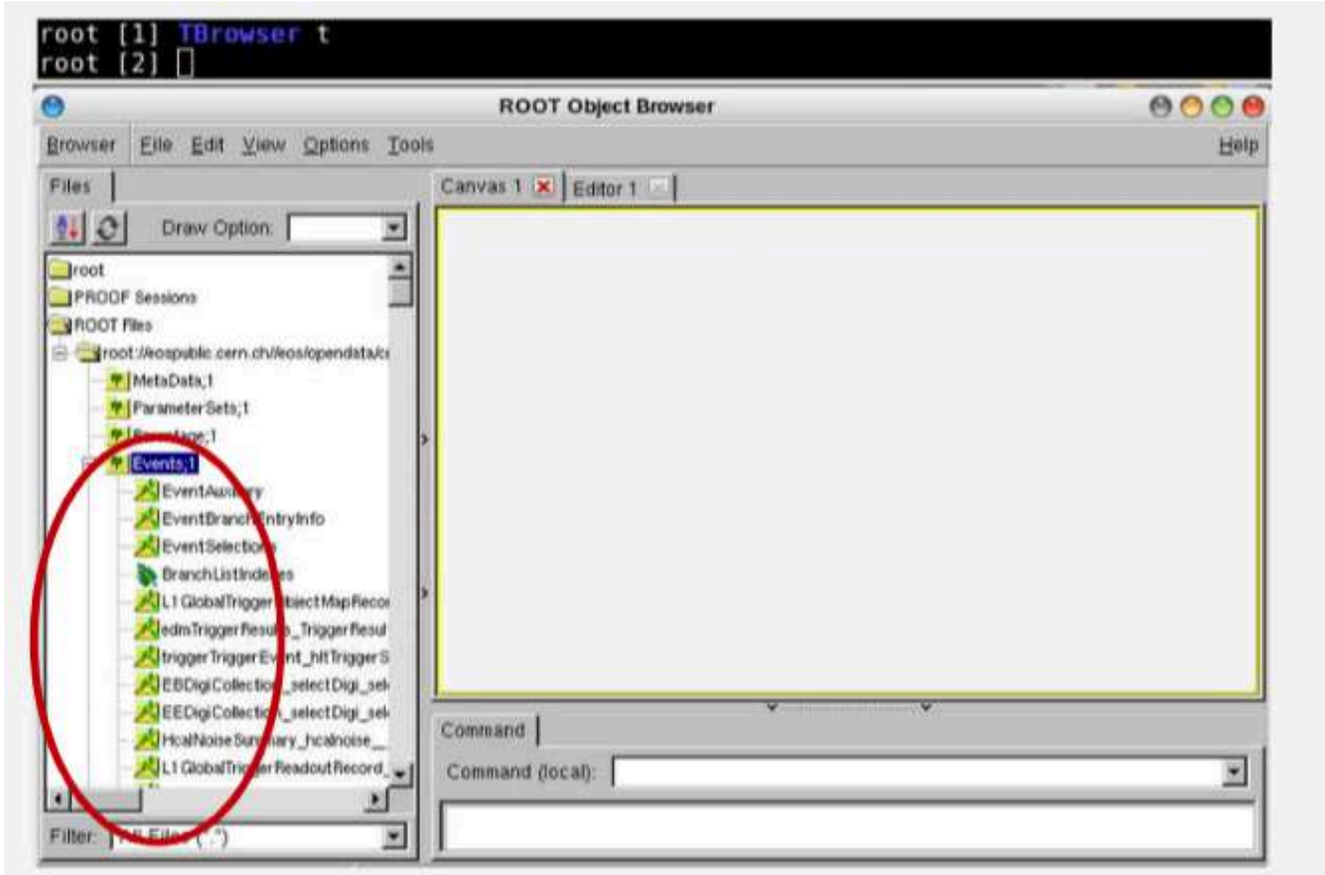
Data Format

SW dependencies?

Virtual machines



Software tutorials: ROOT



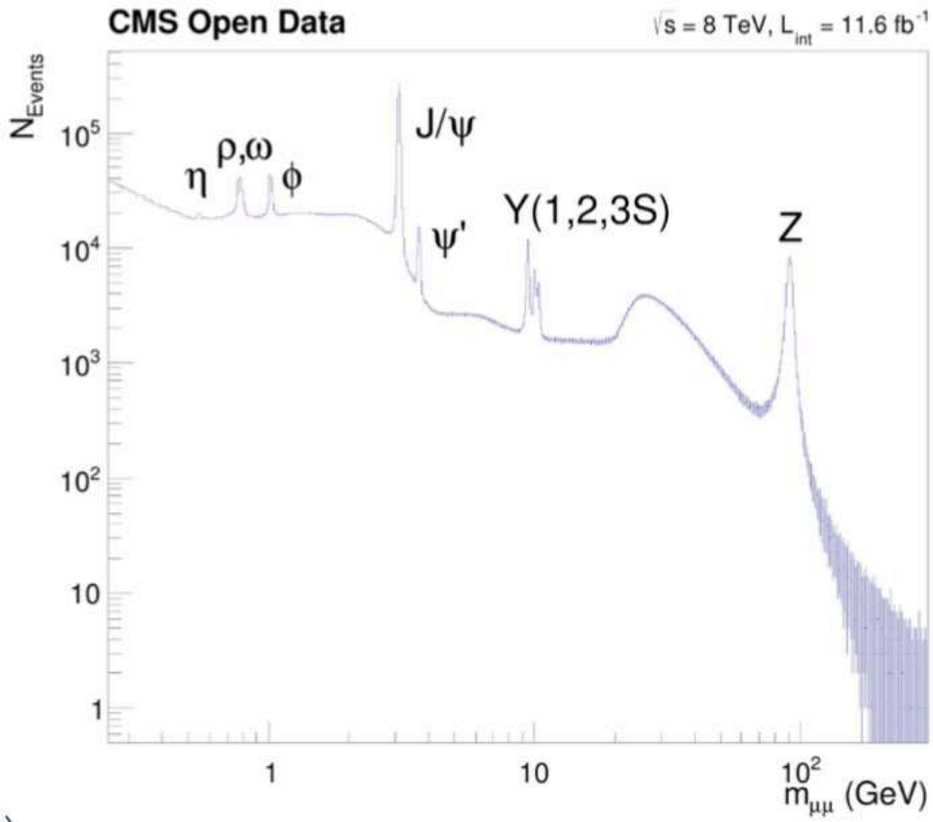
- Very positive feedback from hands-on in ROOT workshops
  - Teaching usage of software with real physics results

```
histo = ROOT.RDataFrame("Events", "Run2012BC_DoubleMuParked_Muons.root")
            .Filter("nMuon == 2")
            .Filter("Muon_charge[0] != Muon_charge[1]")
            .Define("Dimuon_mass",
                    "InvariantMass(Muon_pt, Muon_eta, Muon_phi, Muon_mass)")
            .Histo1D("Dimuon_mass")
```

CMS Open Data in use for ROOT tutorials

- Providing examples very close to the user's use-cases

ROOT
Data Analysis Framework



View Notebook Open in SWAN

ROOT tutorials: Di-muon, Higgs to four leptons, introductory course

Data Format

AOD vs NanoAOD

- **AOD:** Format of primary samples provided on the portal
  - Serialized C++ objects
  - Only fully accessible using CMS software and ROOT
  - Large files with much information (~500 kB/event)
  - Powerful but complex

Variable	Type	Description
nMuon	*unsigned int*	Number of muons in this event
Muon_pt	*float[nMuon]*	Transverse momentum of the muons (stored as an array of size nMuon)
Muon_eta	*float[nMuon]*	Pseudorapidity of the muons
Muon_phi	*float[nMuon]*	Azimuth of the muons
Muon_mass	*float[nMuon]*	Mass of the muons
Muon_charge	*int[nMuon]*	Charge of the muons (either 1 or -1)

Muon collection in NanoAOD format

- **NanoAOD:** Reduced format often used in recent CMS analyses
  - Basic types (floats, integers, ...) or arrays thereof
  - Accessible with any library capable to read ROOT files
  - Smaller files with reduced information (~1-2 kB/event)
  - **Used by actual CMS analyses and easily accessible for everyone**

- Tool to convert Run 1 AOD files to subset of NanoAOD format for education and outreach
  - Ongoing effort to provide Run 1 data in NanoAOD format for research

Jupyter notebooks

Search or jump to… / Pull requests Issues Marketplace Explore

cms-opendata-education / **cms-jupyter-materials-english** Watch 5 Star 7 Fork 7

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Branch: master Find file Copy path

cms-jupyter-materials-english / Exercises-with-open-data / Advanced / **Normfit-transversemomentum+pseudorapidity.ipynb**

mediumlager Add files via upload 7a5d2f6 on 4 Jul

1 contributor

246 lines (245 sloc) 7.54 KB Raw Blame History

Creating a normfit and transverse momentum+pseudorapidity

The point of this exercise is to learn to create a normal distribution fit for the data, and to learn what are transverse momentum and pseudorapidity (and how are they linked together). The data used is open data released by the CMS experiment.

First the fit

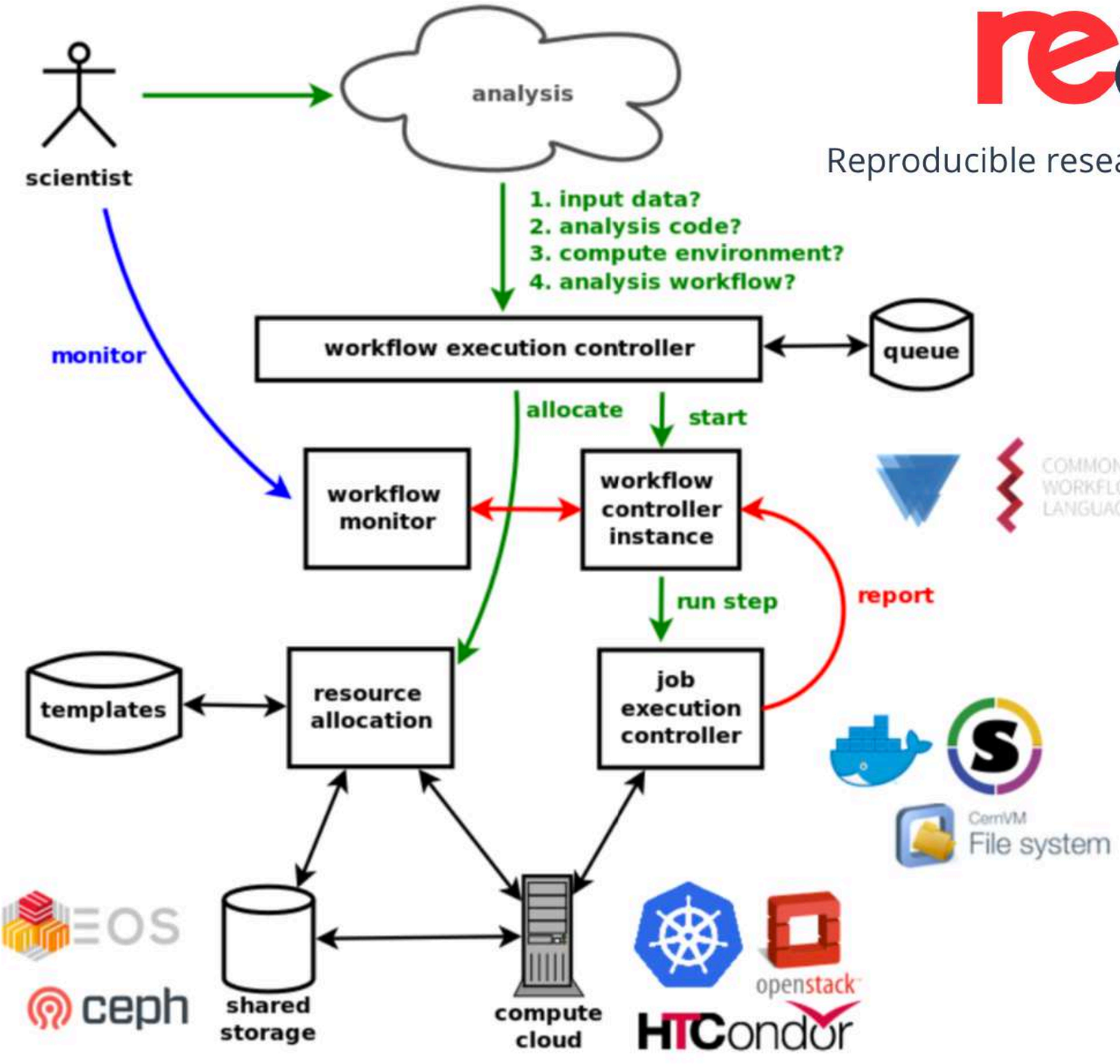
Let's begin by loading the needed modules, data and creating a histogram of the data to see the more interesting points (the area for which we want to create the fit).

In [ ]:
```
# This is needed to create the fit
from scipy.stats import norm

import pandas as pd
import numpy as np
import matplotlib.mlab as mlab
import matplotlib.pyplot as plt
```

In [ ]:
```
# Let's choose Dimuon_DoubleMu.csv
data = pd.read_csv('http://opendata.cern.ch/record/545/files/Dimuon_DoubleMu.csv')

# And save the invariant masses to iMass
iMass = data['M']

# Plus draw the histogram
n, bins, patches = plt.hist(iMass, 300, facecolor='g')
plt.xlabel('Invariant Mass (GeV)')
plt.ylabel('Amount')
```

CMS education activities using notebooks and CMS open data

Run analyses on the clouds



reana

Reproducible research data analysis platform

http://www.reana.io/

Physics publications in peer-reviewed journals

PRL 119, 132003 (2017) PHYSICAL REVIEW LETTERS week ending 29 SEPTEMBER 2017

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,[1,*] Simone Marzani,[2,†] Jesse Thaler,[3,‡] Aashish Tripathee,[3,§] and Wei Xue[3,‖]
[1]*Physics Department, Reed College, Portland, Oregon 97202, USA*
[2]*University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA*
[3]*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
(Received 9 May 2017; revised manuscript received 27 July 2017; published 26 September 2017)

The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between partons. Despite its ubiquitous appearance in many QCD calculations, the

PHYSICAL REVIEW D 96, 074003 (2017)

Jet substructure studies with CMS open data

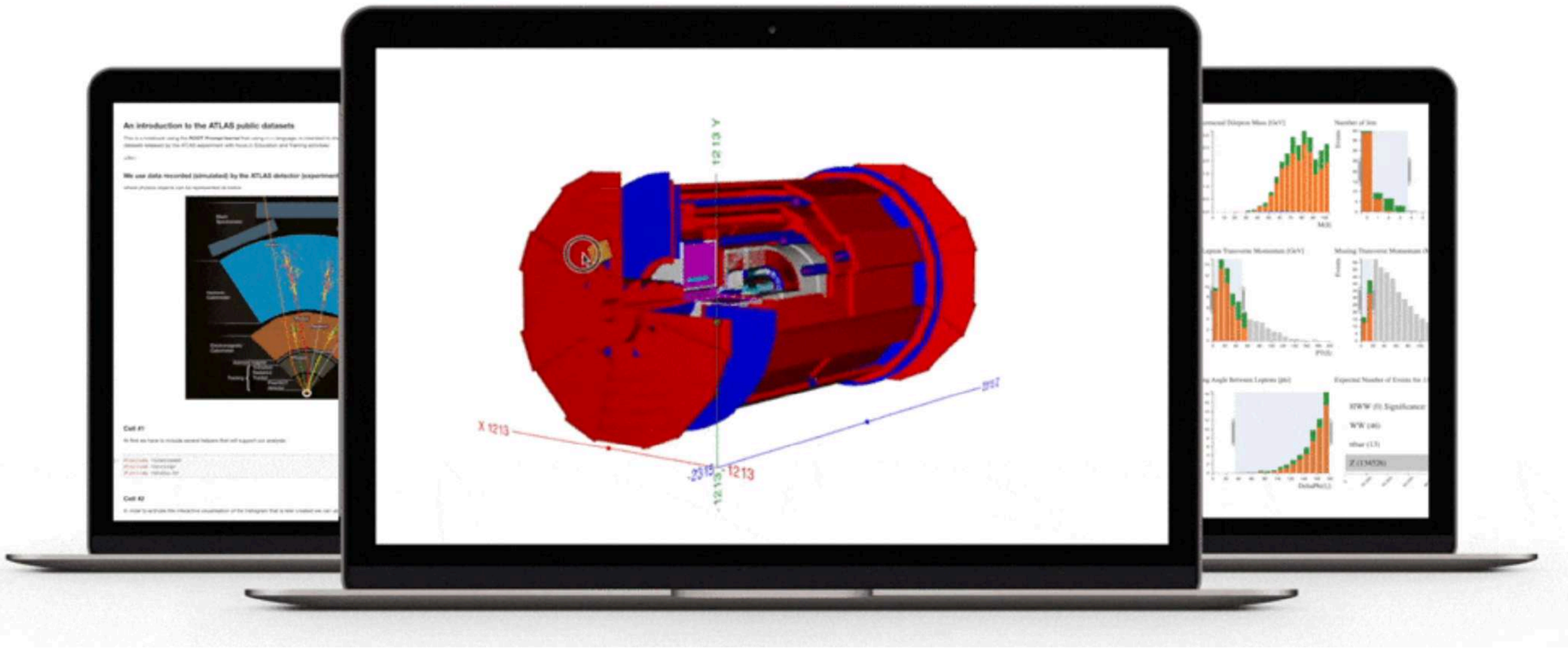
Aashish Tripathee,[1,*] Wei Xue,[1,†] Andrew Larkoski,[2,‡] Simone Marzani,[3,§] and Jesse Thaler[1,‖]
[1]*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*Physics Department, Reed College, Portland, Oregon 97202, USA*
[3]*University at Buffalo, The State University of New York, Buffalo, New York 14260-1500, USA*
(Received 9 May 2017; published 3 October 2017)

We use public data from the CMS experiment to study the two-prong substructure of jets. The CMS open data are based on 31.8 pb^{-1} of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768 687 events containing a high-quality central jet with transverse momentum

First analyses by theorists (Jesse Thaler *et al*, MIT)

CERN Open Data portal

Highest-energy particle-collision data ever released through open access.



The ATLAS Collaboration makes public 10 inverse femtobarns (fb^{-1}) of the **13 TeV** data.

Corresponds to about 1 quadrillion proton-proton collisions (that's 1 followed by 15 zeros), or **500 thousand produced Higgs bosons**. Approximately the same amount of data that the ATLAS Collaboration used to discover the Higgs boson in 2012.

Open Data ML release

CMS releases open data for Machine Learning

2019-07-18 by CMS Collaboration

News

The CMS Collaboration at CERN is happy to announce the release of its fourth batch of open data to the public. With this release, which brings the volume of its open data to more than 2 PB (or two million GB), CMS has now provided open access to 100% of its research data recorded in proton-proton collisions in 2010, in line with the collaboration's data-release policy. The release also includes several new data and simulation samples. The new release builds upon and expands the scope of the successful use of CMS open data in research and in education.

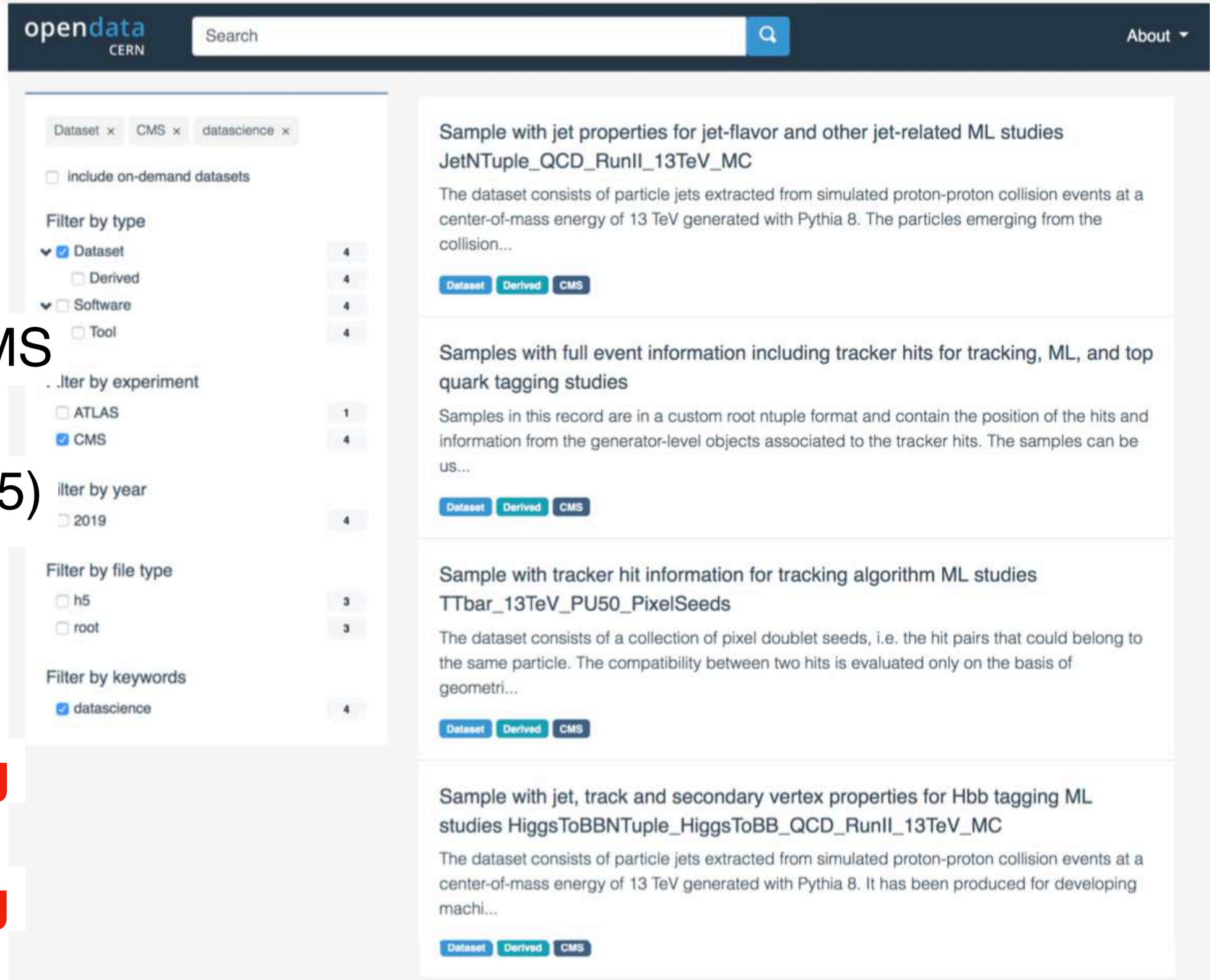
http://opendata.cern.ch/docs/cms-releases-open-data-for-machine-learning

In this release, CMS open data address the ever-growing application of machine learning (ML) to challenges in high-energy physics. According to a recent paper, collaboration with the data-science and ML community is considered a high-priority to help advance the application of state-of-the-art algorithms in particle physics.

ML studies dedicated datasets on CERN Open Data

Four *derived* datasets from official 2016 CMS simulation (ROOT & HDF5)

- **Jet flavor studies**
- **Top tagging**
- **Pixel tracking studies**
- **H(bb) tagging**



http://opendata.cern.ch/record/12102

jet-related ML studies

Sample with jet properties for jet-flavor and other jet-related ML studies
JetNTuple_QCD_RunII_13TeV_MC

Description: The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8.

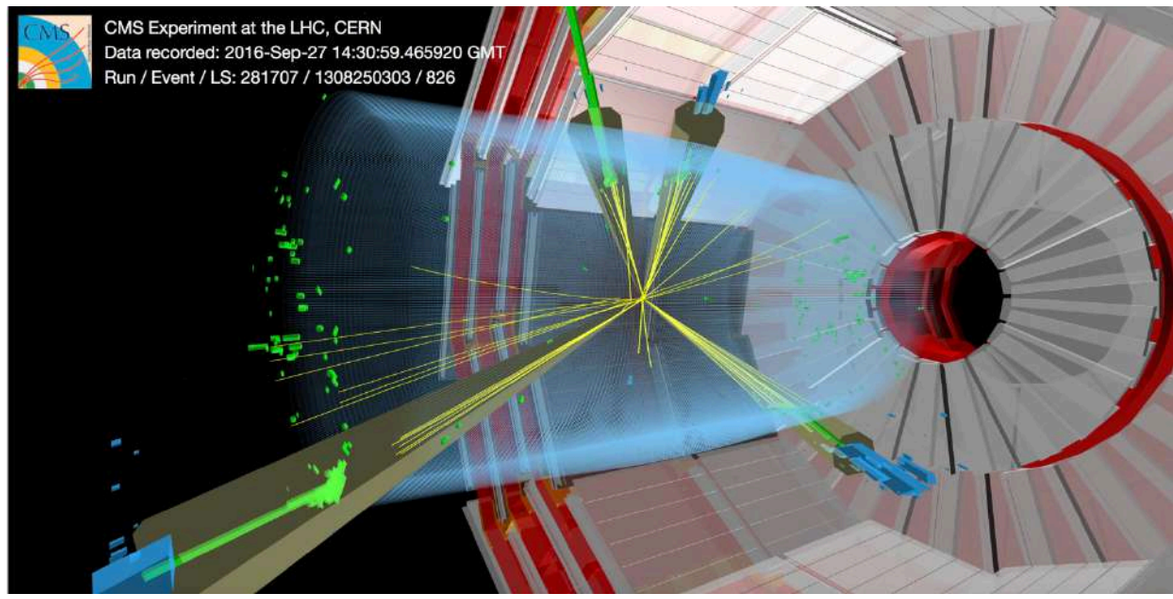
Dataset characteristics

22554294 entries. **244** files. **204.6 GB** in total.

Dataset semantics

Variable	Type	Description
jetPt	*Float_t*	Transverse momentum of the jet.
jetEta	*Float_t*	Pseudorapidity (η) of the jet.
jetPhi	*Float_t*	Azimuthal angle (ϕ) of the jet.
jetMass	*Float_t*	Mass of the jet.
jetGirth	*Float_t*	Girth of the jet (as defined in arXiv:1106.3076 [hep-ph]).
jetArea	*Float_t*	Catchment area of the jet; used for jet energy corrections.
jetRawPt	*Float_t*	Transverse momentum of the jet before the energy corrections.
jetRawMass	*Float_t*	Mass of the jet before the energy corrections.
jetLooseID	*UInt_t*	Binary variable indicating whether the jet passes 'loose' criteria f
jetTightID	*UInt_t*	Binary variable indicating whether the jet passes 'tight' criteria fc

- Training data produced with JetNtupleProducerTool



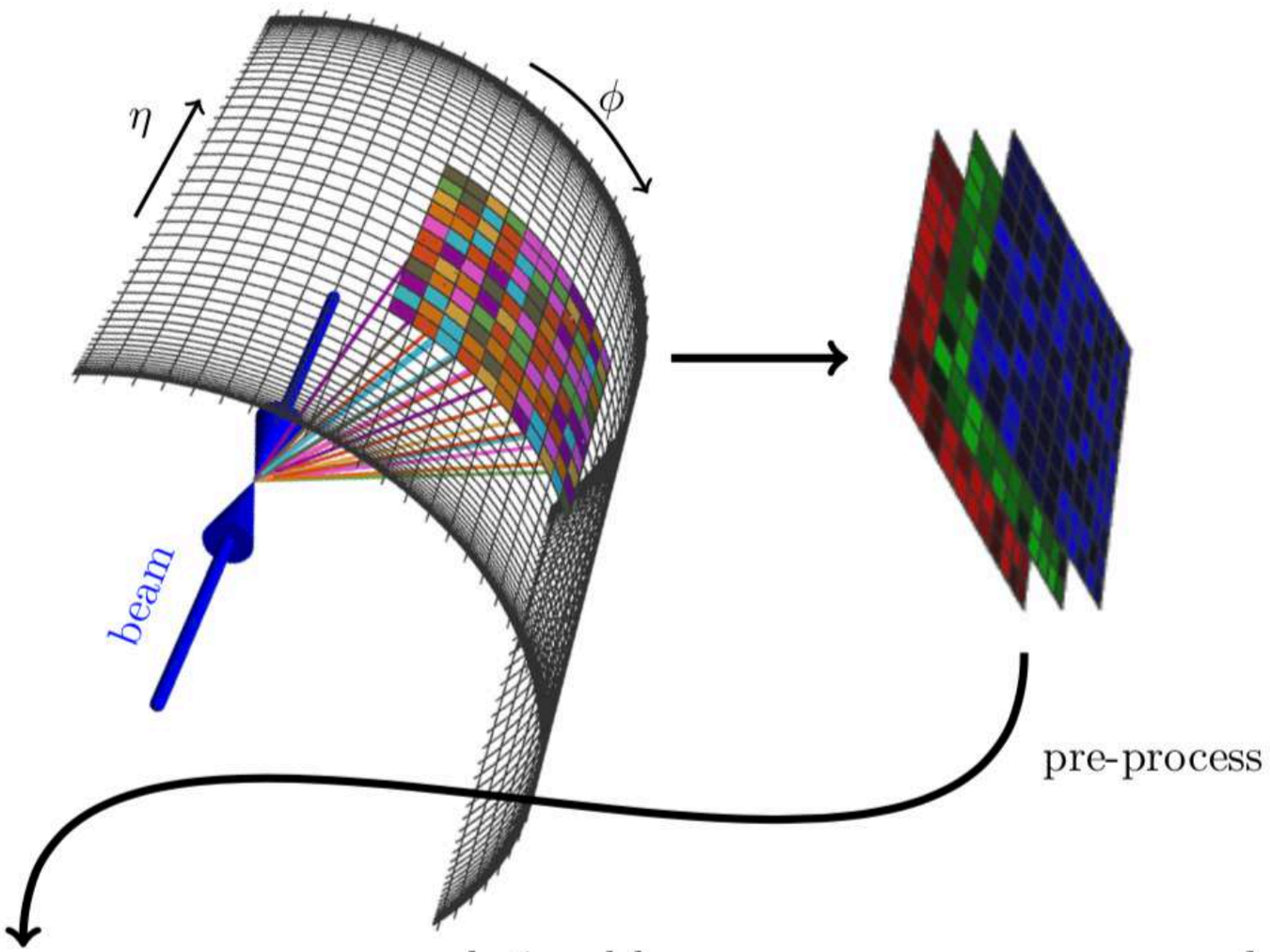
ML approach to jet identification

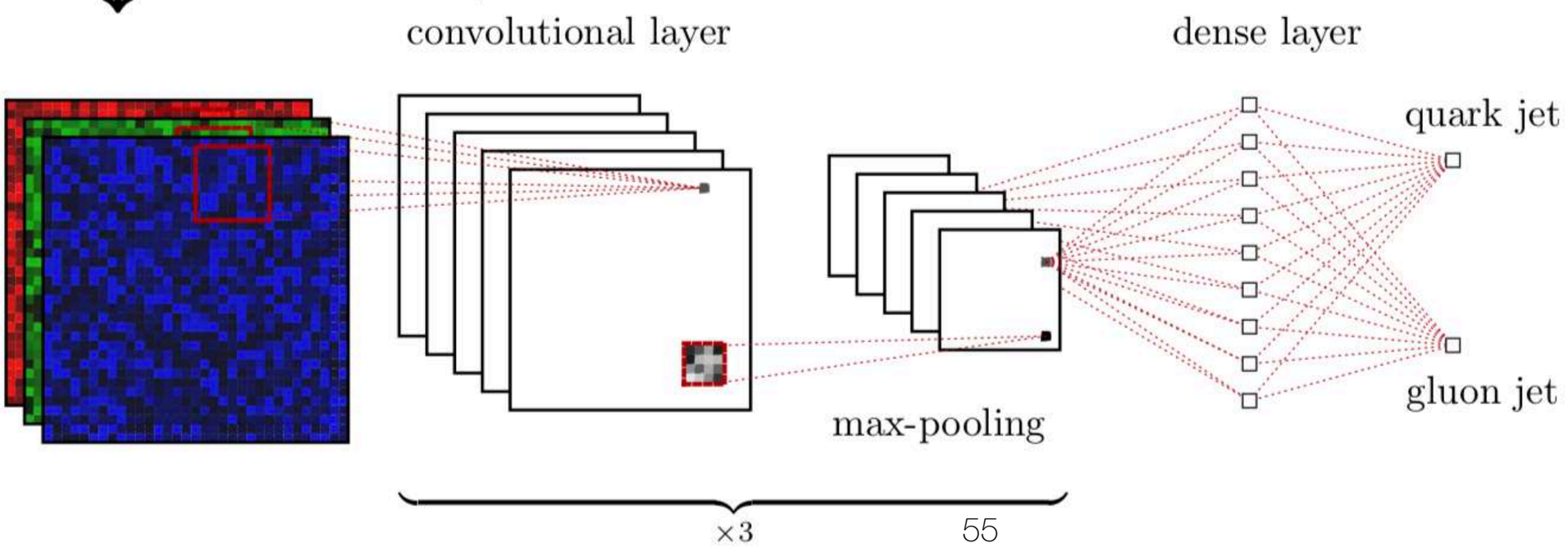
image: Komiske, Metodiev, Schwartz arxiv:1612.01551

Oliveira, et. al arXiv:1511.05190

Whiteson, et al arXiv:1603.09349

Dawe, et al arXiv:1609.00607

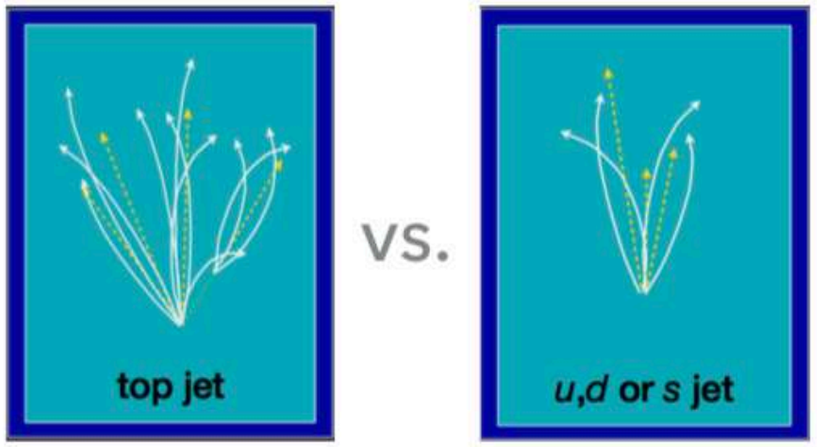




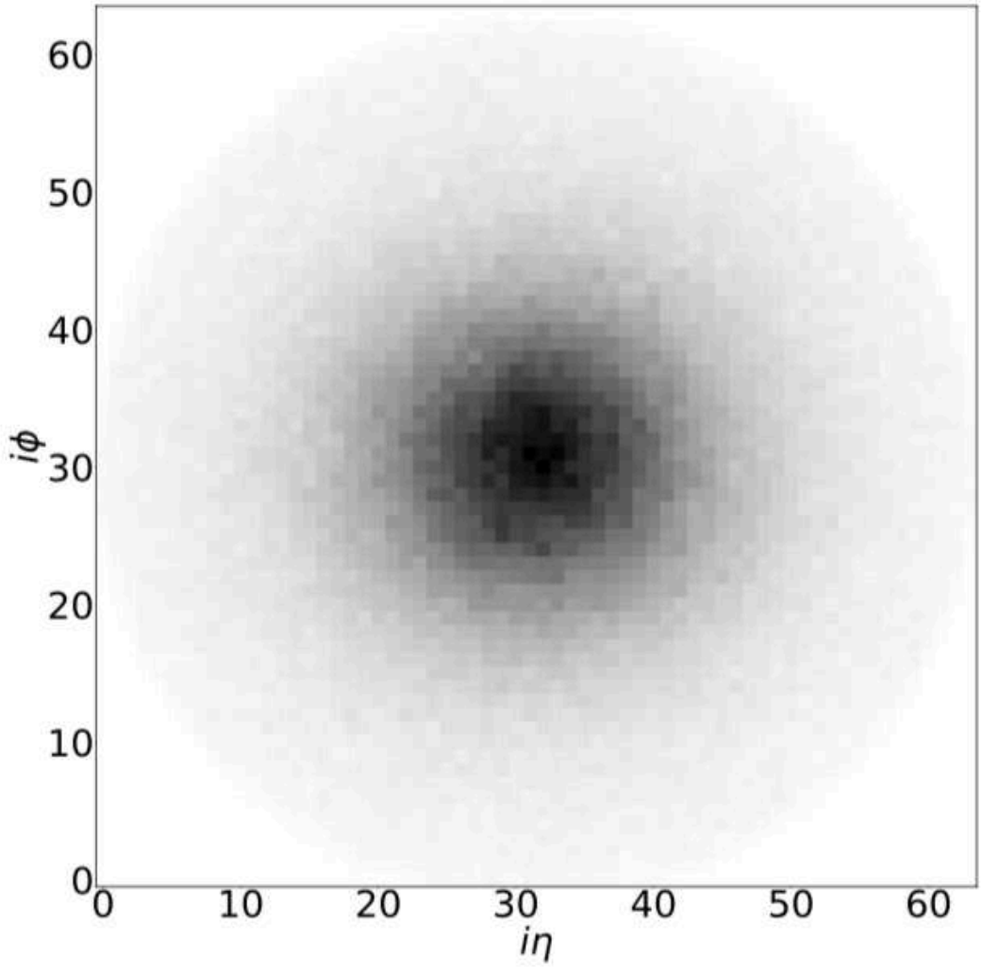
Apply deep learning algorithms to classify to “jet images”

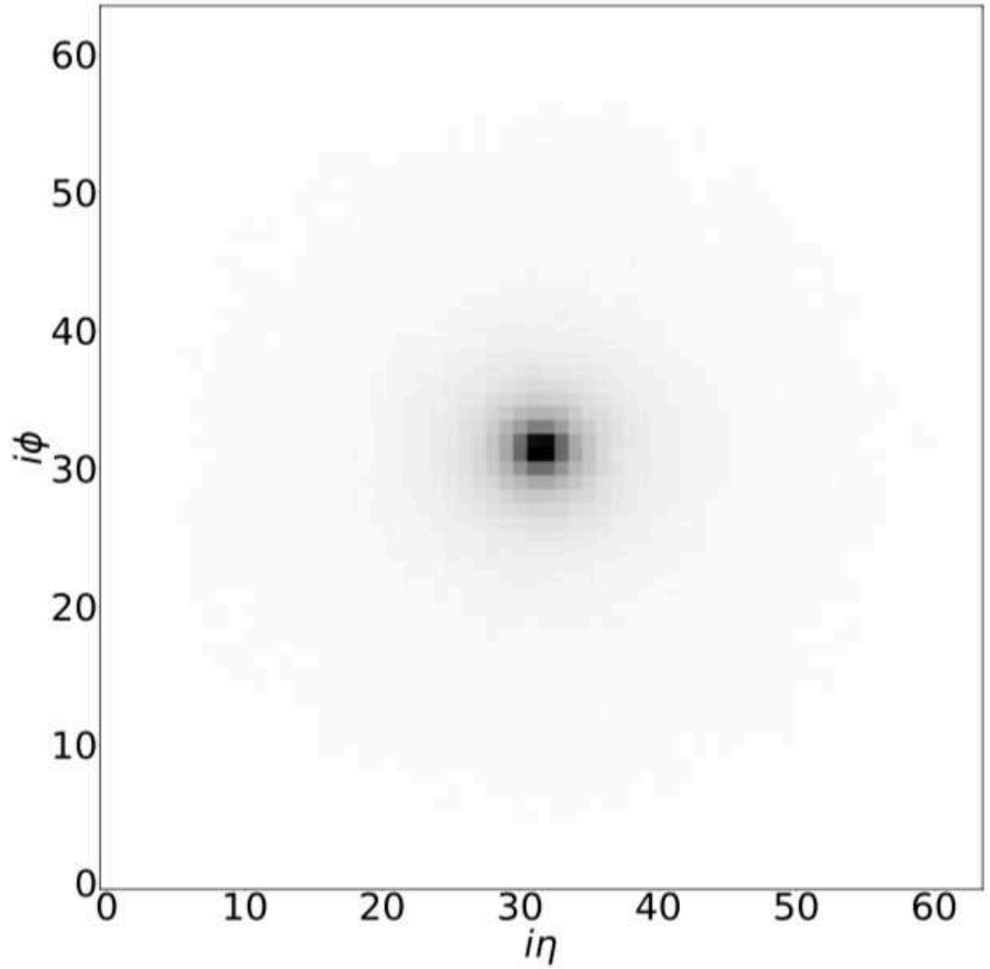
ML approach to jet identification

- Re-train ResNet-50 to identify the origin of jets
- Inputs are jet images = pixelated versions of calorimeter hits in 2D (η, Φ)



Note: averaged over 10k jets; 1 jet gives a ***sparse*** image





Higgs to bb

opendata CERN

Search

Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC

Duarte, Javier;

Dataset | Derived | Datascience | CMS | CERN-LHC

Parent Dataset: /BulkGravTohhTohbbhbb_narrow_M-600_13TeV-madgraph/RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM

Description

The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8. It has been produced for developing machine-learning algorithms to differentiate jets originating from a Higgs boson decaying to a bottom quark-antiquark pair (Hbb) from quark or gluon jets originating from quantum chromodynamic (QCD) multijet production.

The reconstructed jets are clustered using the anti-kT algorithm with R=0.8 from particle flow (PF) candidates (AK8 jets). The standard L1+L2+L3+residual jet energy corrections are applied to the jets and pileup contamination is mitigated using the charged hadron subtraction (CHS) algorithm. Features of the AK8 jets with transverse momentum pT > 200 GeV and pseudorapidity $|\eta| < 2.4$ are provided. Selected features of inclusive (both charged and neutral) PF candidates with pT > 0.95 GeV associated to the AK8 jet are provided. Additional features of charged PF candidates (formed primarily by a charged particle track) with pT > 0.95 GeV associated to the AK8 jet are also provided. Finally, additional features of reconstructed secondary vertices (SVs) associated to the AK8 jet (within $\Delta R < 0.8$) are also provided.

Derived datasets (ROOT & HDF5):
http://opendata-dev.web.cern.ch/record/12102

- 182 files, 245 GB, 18 million total entries (jets)
- event features, e.g. MET, ρ (average density)
- jet features, e.g. mass, p_T, N-subjettiness variables
- particle candidate features, e.g. p_T, η, **ϕ** (*for up to 100 particles*)
- charged particle / track features, e.g. impact parameter (*for up to 60 tracks*)
- secondary vertex features, e.g. flight distance (*for up to 5 vertices*)

Higgs to bb

Dataset semantics

Variable	Type	Description
event_no	*UInt_t*	Event number
npv	*Float_t*	Number of reconstructed primary vertices (PVs)
ntrueInt	*Float_t*	True mean number of the poisson distribution for this event from which the number of interactions in each bunch crossing has been sampled
rho	*Float_t*	Median density (in GeV/A) of pile-up contamination per event; computed from all PF candidates of the event
sample_isQCD	*Float_t*	Boolean that is 1 if the simulated sample corresponds to QCD multijet production
fj_doubleb	*Float_t*	Double-b tagging discriminant based on a boosted decision tree calculated for the AK8 jet (see CMS-BTV-16-002)
fj_eta	*Float_t*	Pseudorapidity η of the AK8 jet
fj_gen_eta	*Float_t*	Pseudorapidity η of the generator-level, matched heavy particle: H, W, Z, top, etc. (default = -999)
fj_gen_pt	*Float_t*	Transverse momentum of the generator-level, geometrically matched heavy particle: H, W, Z, t, etc. (default = -999)
fj_isBB	*Int_t*	Boolean that is 1 if two or more b hadrons are clustered within the AK8 jet (see SWGuideBTagMCTools)

links to variable definitions / methods

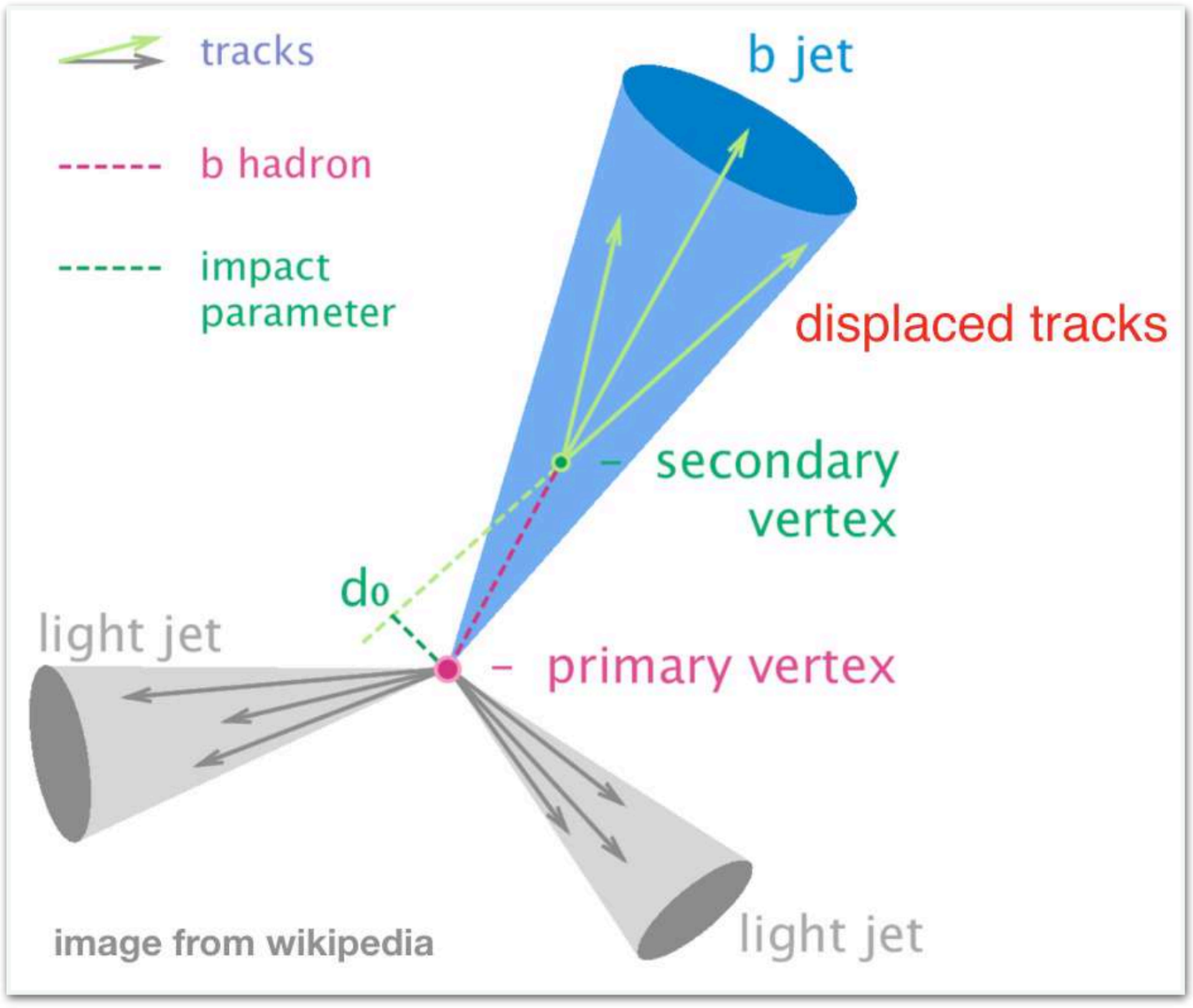
.....

b-tagging

b hadrons have long lifetimes: travel O(mm) before decay!

- displaced tracks
- secondary vertices
- soft leptons

→ **Combined taggers**



b-tagging

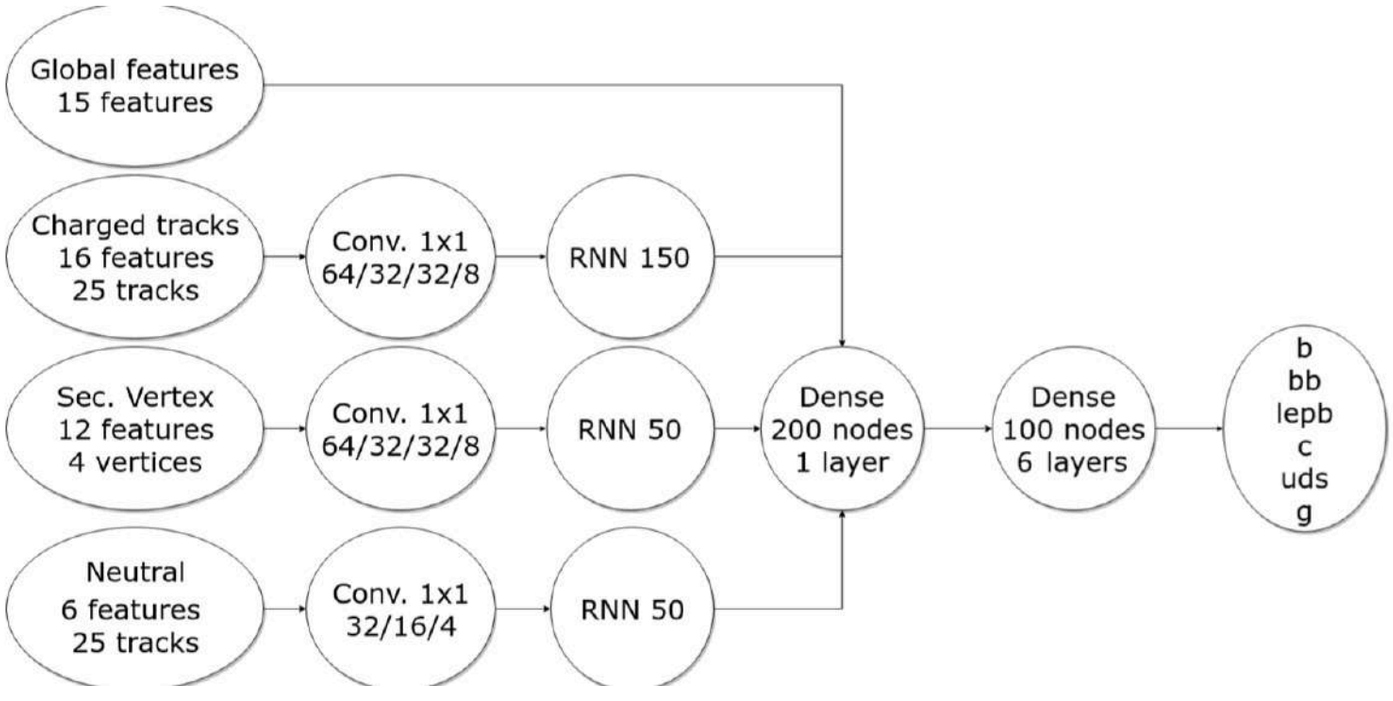
b hadrons have long lifetimes: travel O(mm) before decay!

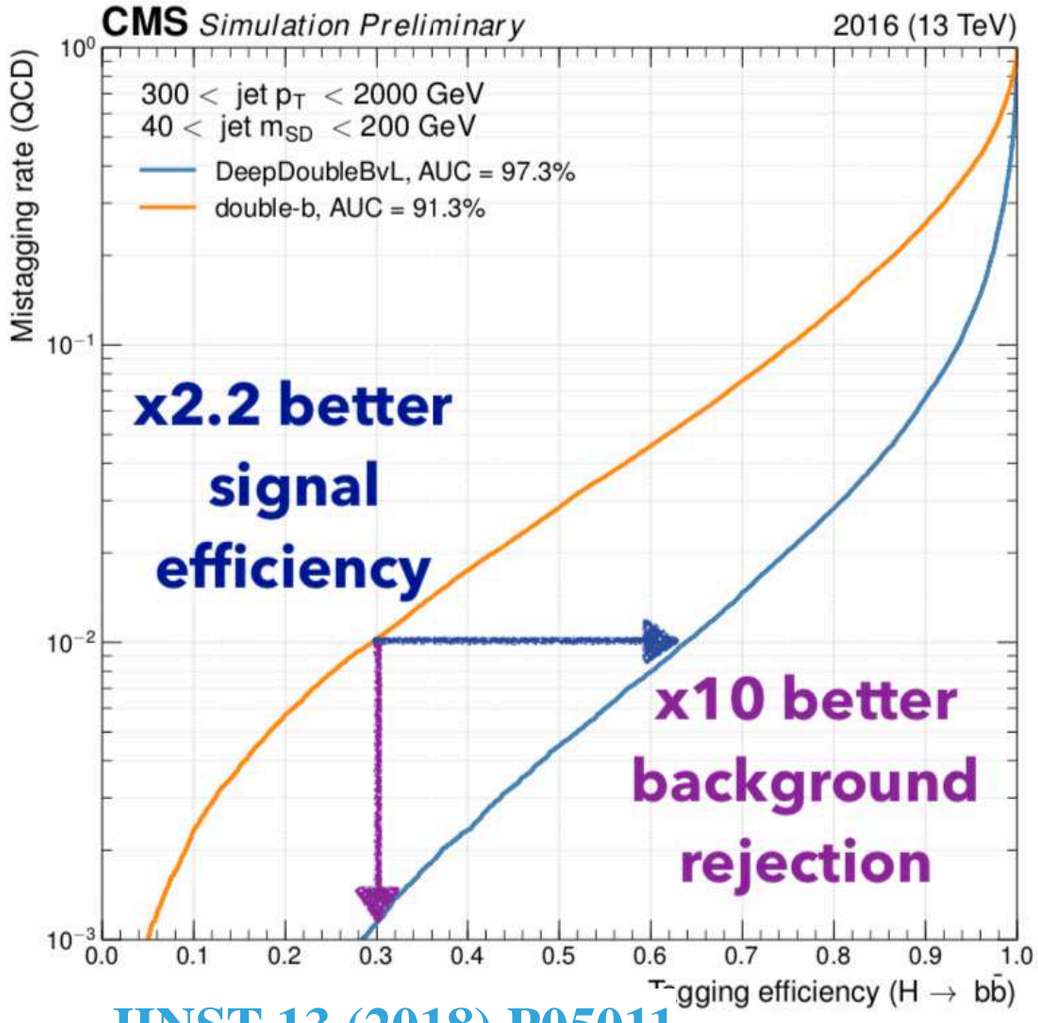
- displaced tracks
- secondary vertices
- soft leptons

→ **Super Combined taggers**

Deep bb tagging

Large performance gain over previous algorithm (BDT)





JINST 13 (2018) P05011

Summary and Outlook

- Data science and particle physics communities closely work together.

- A white paper has been published recently.

    - Future research and development areas for machine learning in particle physics.

    - Roadmap for software, hardware resource requirements.

- CERN open data released new data dedicated to Machine Learning studies.