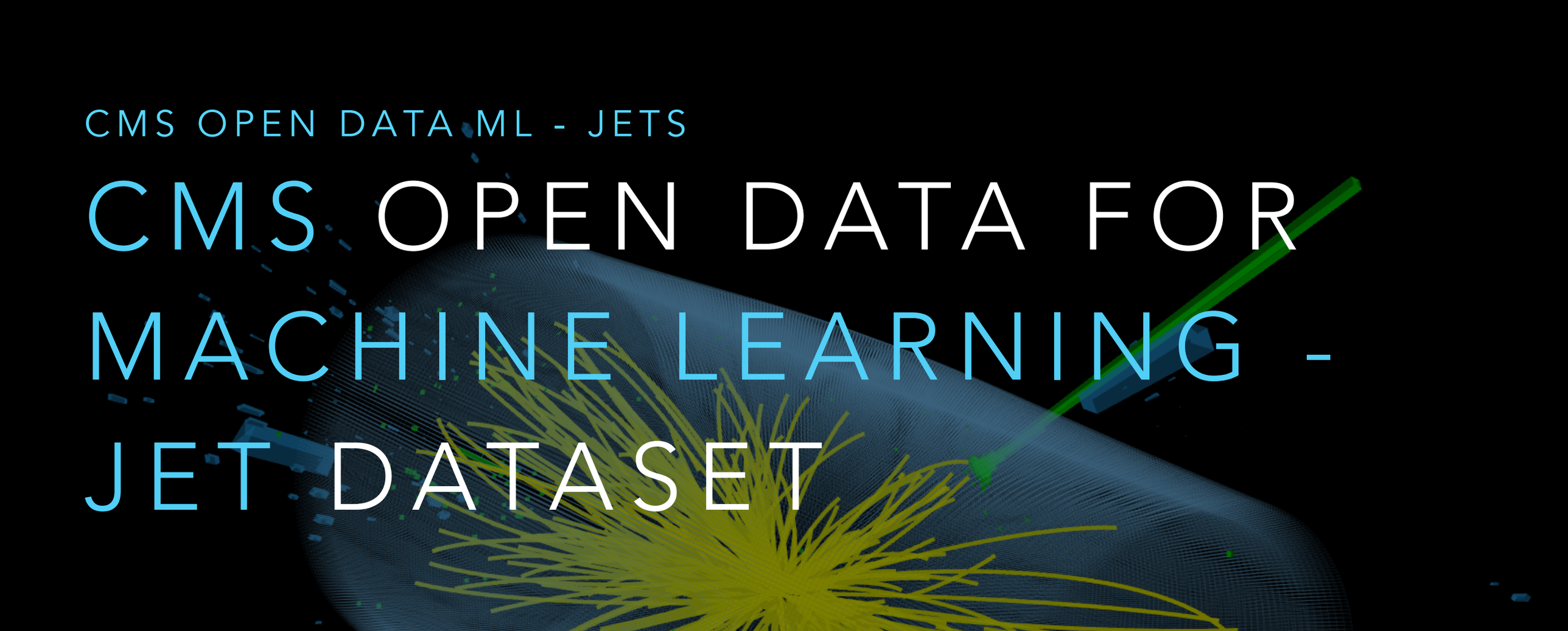
CMS OPEN DATA ML - JETS

CMS OPEN DATA FOR MACHINE LEARNING - JET DATASET



DATA SCIENCE @ HIGH ENERGY PHYSICS 2017

FERMILAB

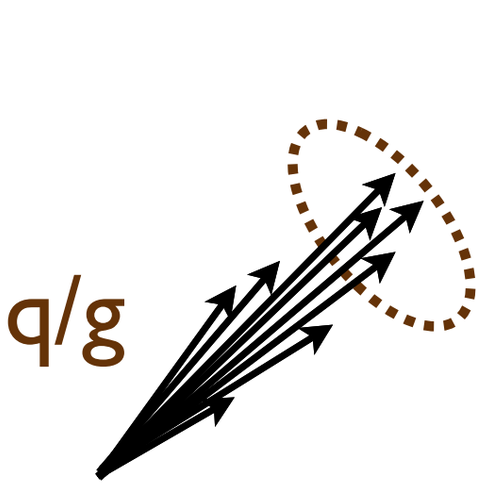
BATAVIA, IL, USA

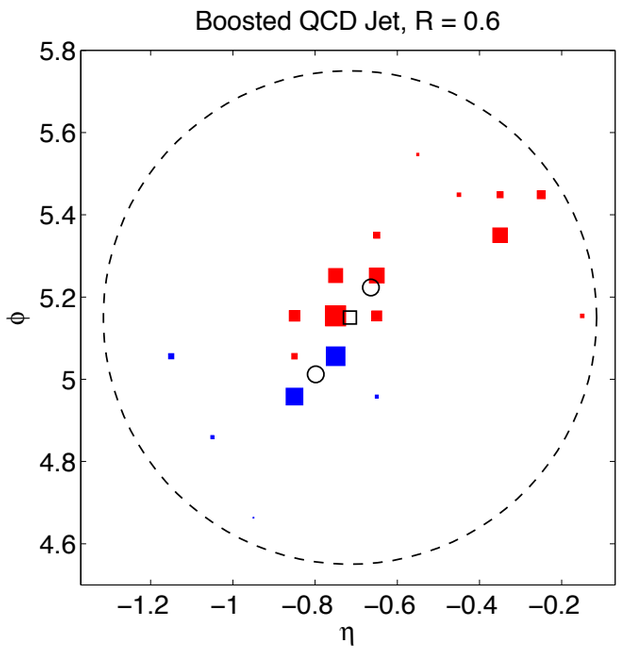
MAY 8, 2017

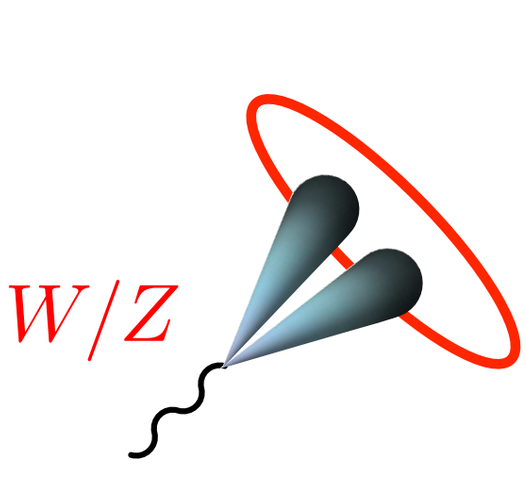
Javier Duarte
Fermilab

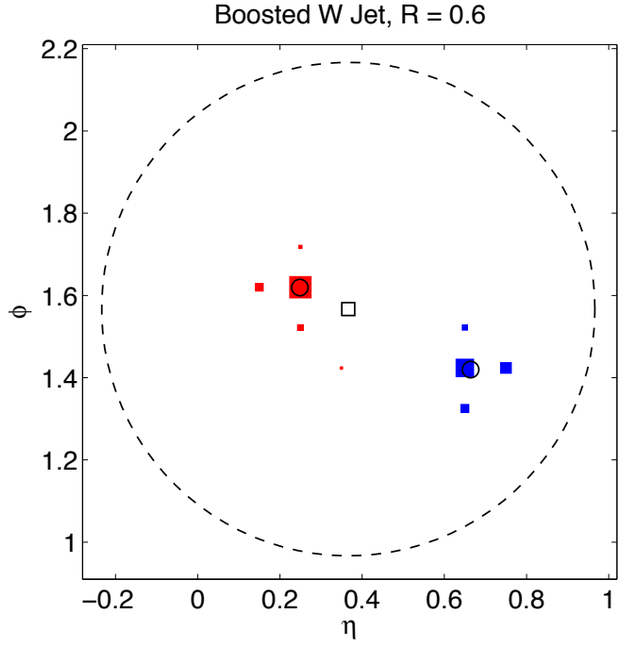
CLASSIC JET PROBLEM

- A jet is a collimated spray of energetic particles originating from the fragmentation of scattered partons (quarks or gluons)
- One classic problem is identifying whether the jet originates from the decay of a boosted particle W/Z/H/t or simply from a quark/gluon (QCD)







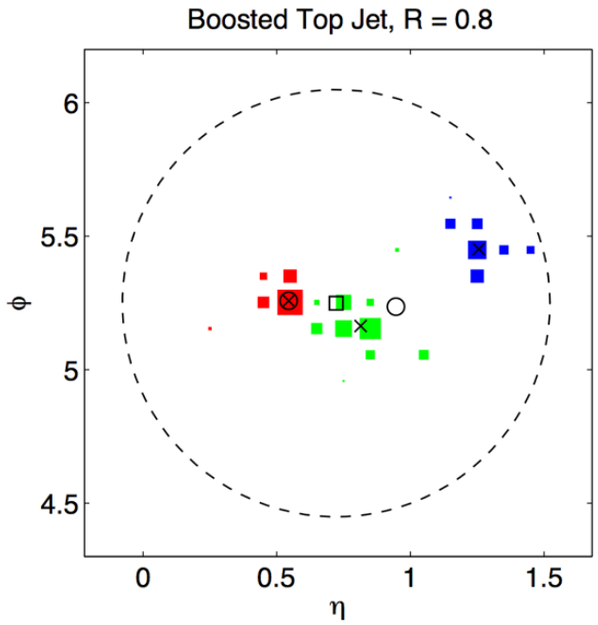


[J. Thaler, et al. arXiv:1011.2268]

Javier Duarte
Fermilab

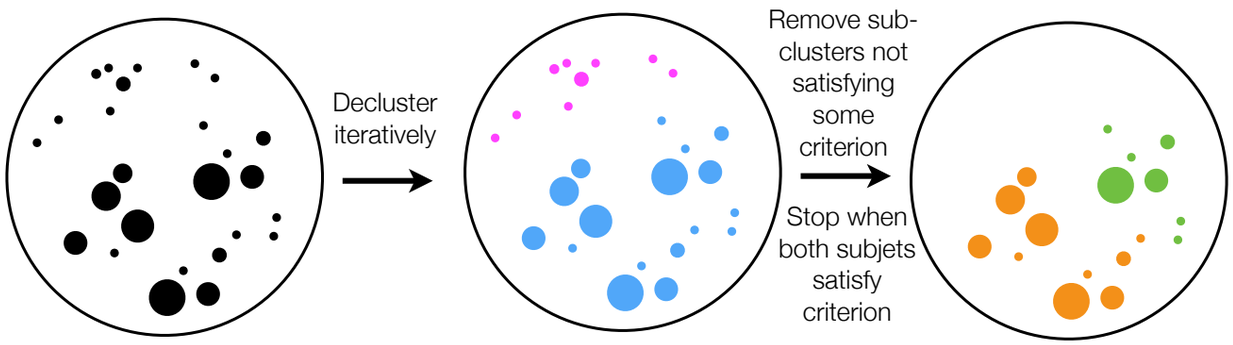
STATE OF THE ART

- Many techniques to aid in the identification of heavy particles have been developed such as
  - N-subjettiness [arXiv:1011.2268] and energy correlation functions [arXiv:1305.0007] which provide discrimination by looking at the shape of the jet
  - Jet trimming [arXiv:0912.1342], pruning [arXiv:0912.0033], and soft drop [arXiv:1402.2657] algorithms which remove "soft" radiation to better identify the "hard" part of the jet



$\tau_N \to 0 \Longrightarrow$ energy spread is close to the subjet axes

$$\tau_N = \frac{\sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}\}}{\sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} R}$$



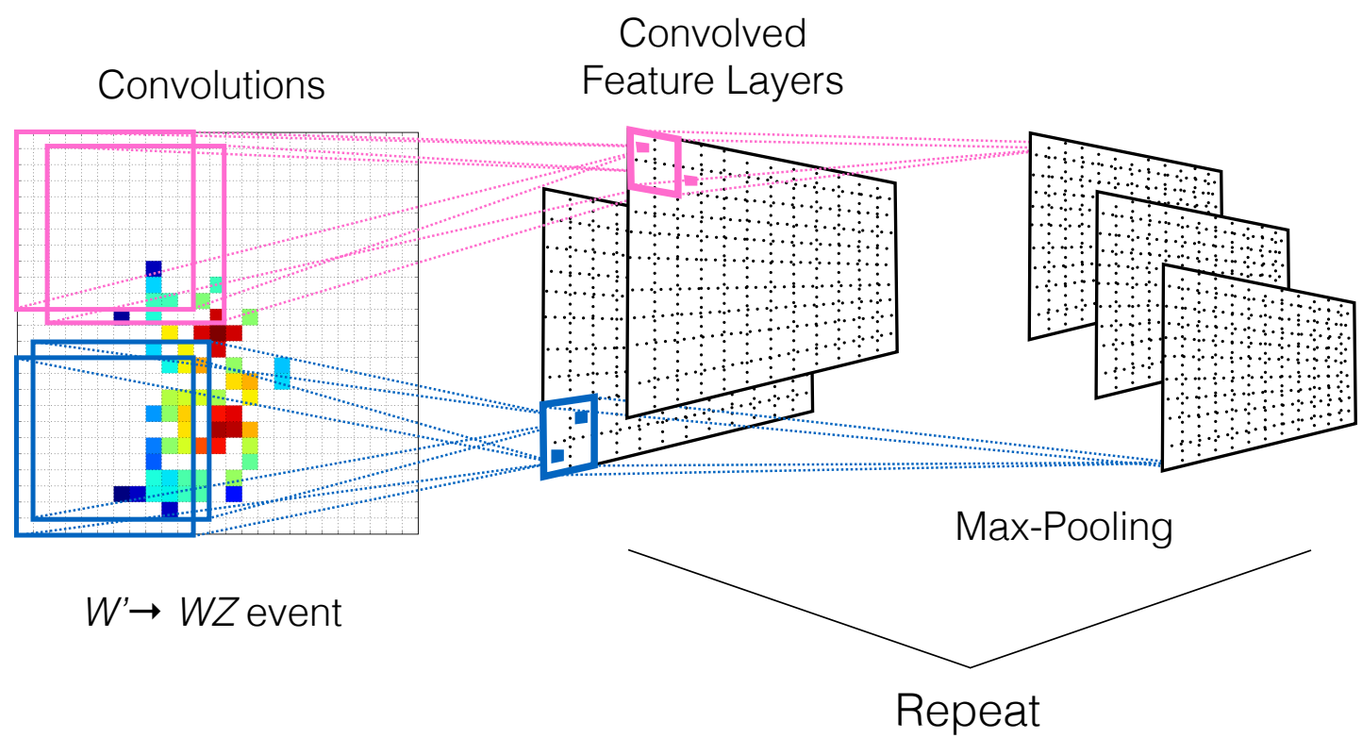
$$\text{Soft Drop Condition: } \frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0}\right)^{\beta}$$

[A. J. Larkoski, et al. arXiv:1402.2657]

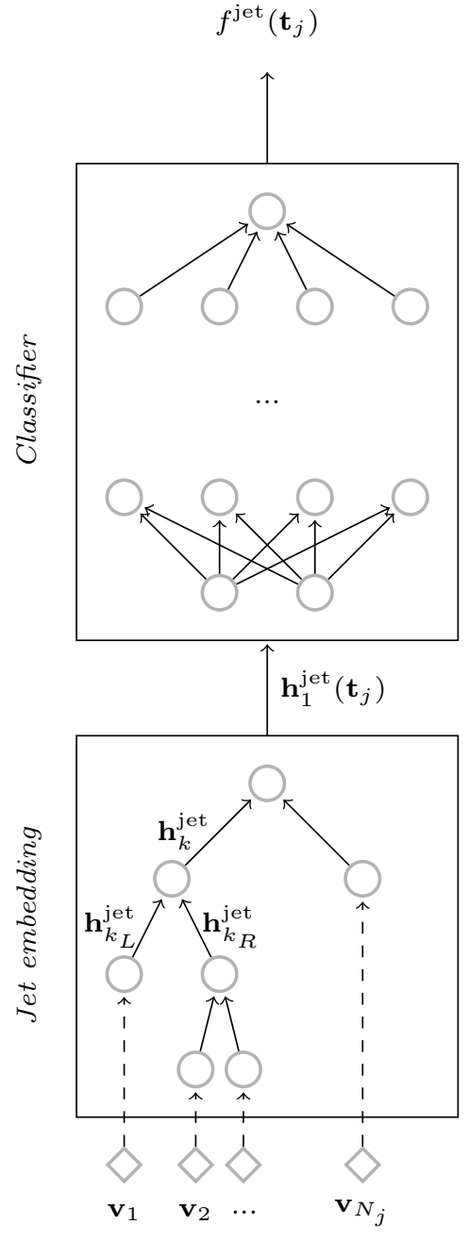
Javier Duarte
Fermilab

MACHINE LEARNING APPS

- Many groups have also tried to apply machine learning to aid in the solution of this problem, such as
  - Convolutional neural networks using an analogy between calorimeters and images [arXiv:1407.5675, arXiv:1511.05190, arXiv:1704.02124]
  - Recursive neural networks built upon an analogy between QCD and natural languages [arXiv:1702.00748]



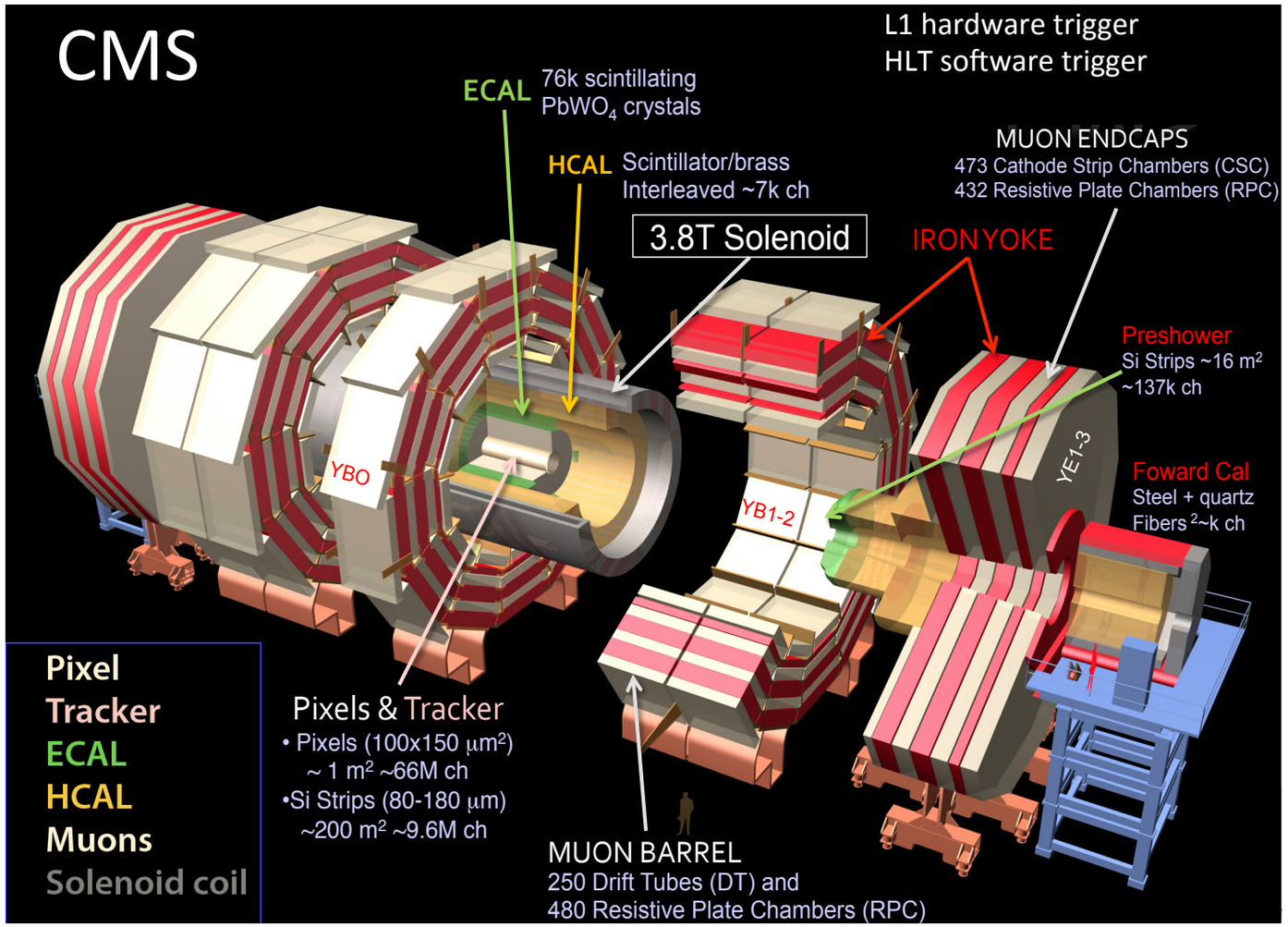
[L. de Oliveira, et al. arXiv:1511.05190]



[G. Louppe, et al. arXiv:1702.00748]

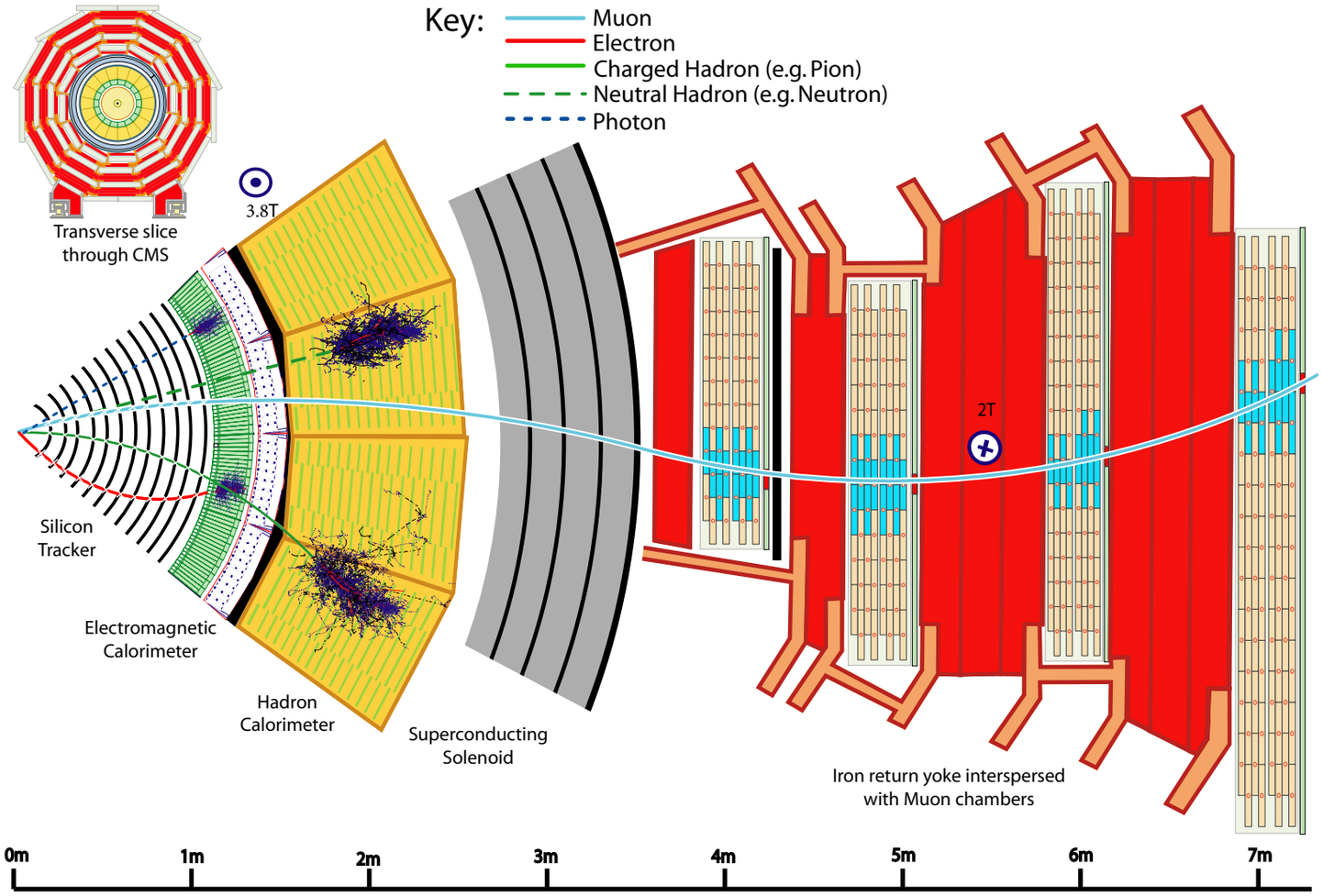
COMPACT MUON SOLENOID

- CMS is one of the two general-purpose detectors at the LHC
- 3.8 T magnetic field bends particle trajectories allowing for excellent tracking
- ECAL: $PbWO_4$ crystals (high density, short radiation length and Molière radius)
- HCAL: plastic scintillator and brass absorber interleaved
- Muon system: drift tubes (DT), resistive plate chambers (RPC), and cathode strip chambers (CSC)



PARTICLE FLOW RECONSTRUCTION

- "Particle flow" (PF) reconstruction: holistic approach to particle reconstruction, combining measurements in the tracker, calorimeters, and muon system to provide an improved determination of the energy and direction of each class of particle
- Five main classes: Muon, Electron, Charged Hadron, Neutral Hadron, and Photon



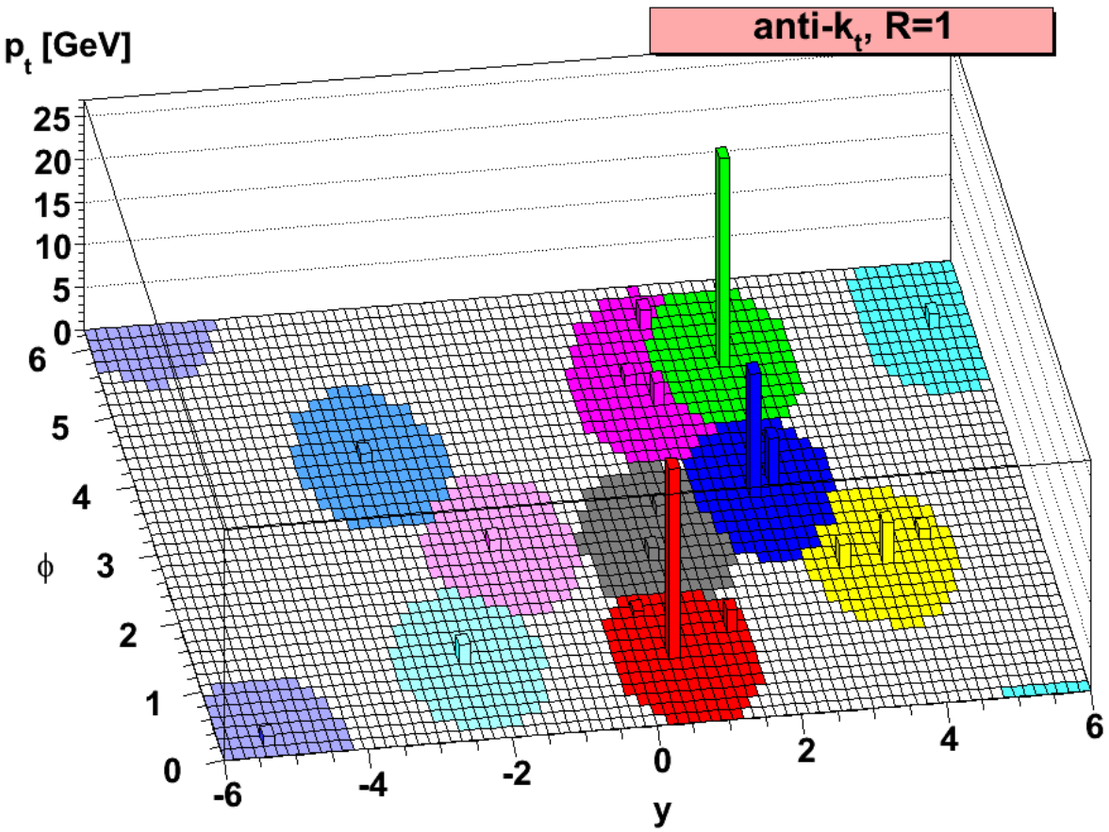
PF JET RECONSTRUCTION

- PF jets are clustered from PF candidates (belonging to the 5 classes) using anti-kT algorithm [arXiv:0802.1189] and FastJet [arXiv:1111.6097] with jet radius parameter R=0.7

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} ,$$

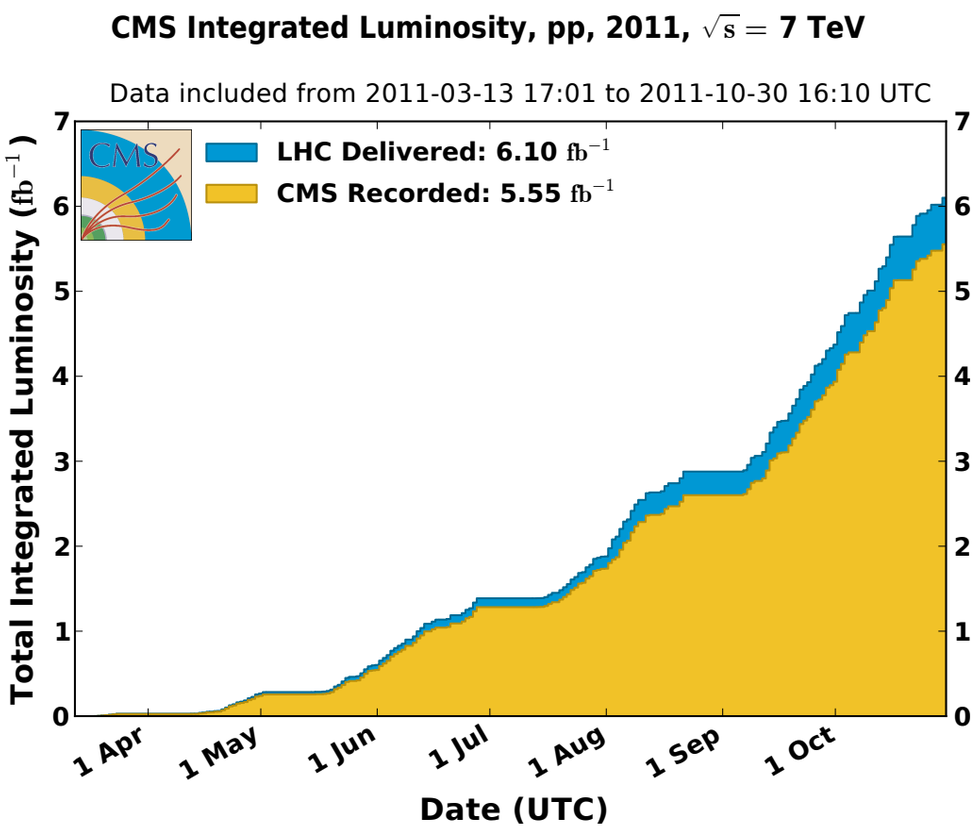
$$d_{iB} = k_{ti}^{2p} ,$$

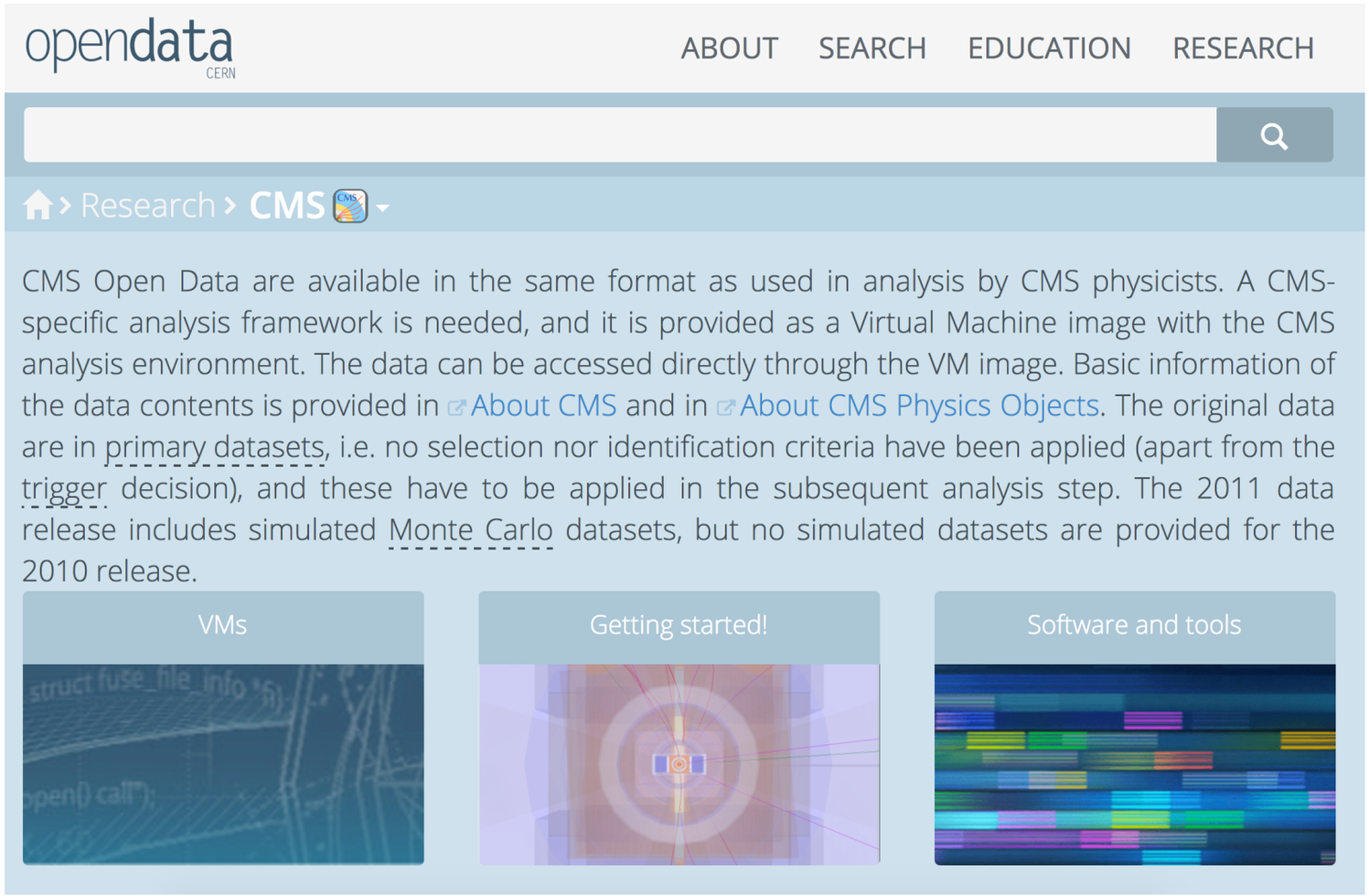
$$p = -1$$



CMS OPEN DATA

- 2011 data is second public release of CMS data (this time with simulation) [portal]
- Data format is AOD (same used by CMS analysts)





PREPARATION OF THE DATA

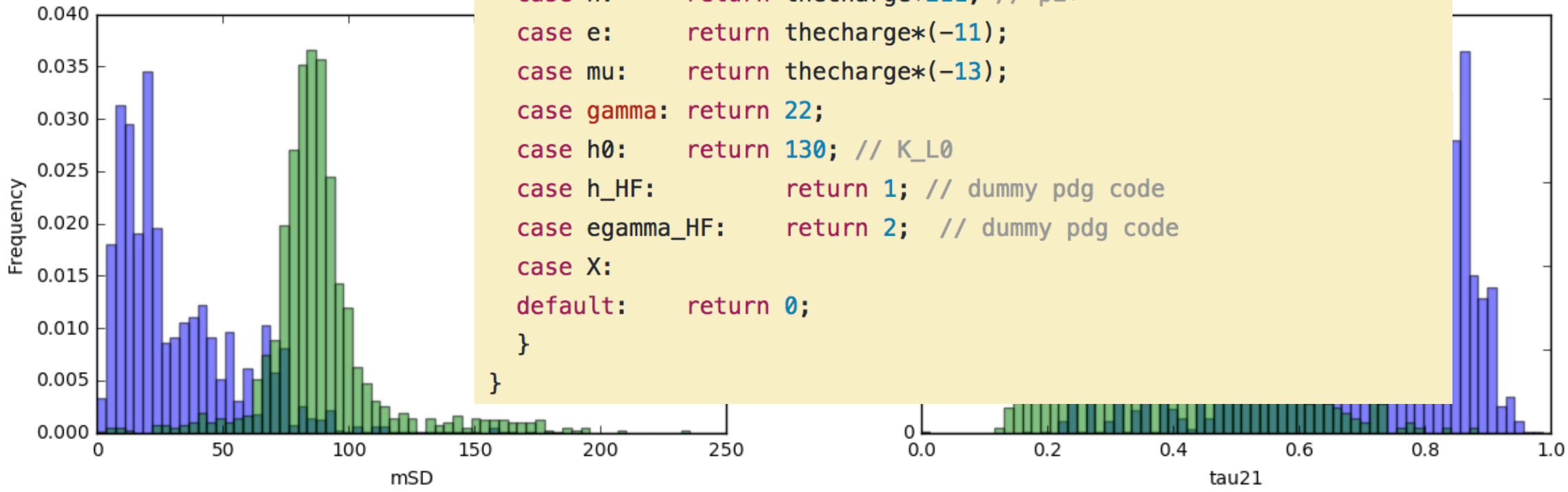
- Using modified publicly available CMSSW code [github] running over different samples of CMS open simulation to produce flat numpy arrays
- Simulation datasets:
  - ttbar (for boosted W sample): TT_weights_CT10_TuneZ2_7TeV-powheg-pythia-tauola
  - QCD (for background QCD jets): QCD_Pt-80to120_TuneZ2_7TeV_pythia6 QCD_Pt-120to170_TuneZ2_7TeV_pythia6 QCD_Pt-170to300_TuneZ2_7TeV_pythia6 QCD_Pt-300to470_TuneZ2_7TeV_pythia6 QCD_Pt-470to600_TuneZ2_7TeV_pythia6

DATASET FEATURES

- Event level features:
  ('run', 'lumi', 'event', 'met', 'sumet', 'rho', 'pthat', 'mcweight', 'njet_ak7')
- Jet-level features:
  ('jet_pt_ak7', 'jet_eta_ak7', 'jet_phi_ak7', 'jet_E_ak7', 'jet_msd_ak7', 'jet_area_ak7', 'jet_jes_ak7', 'jet_tau21_ak7', 'jet_isW_ak7',
- PF-candidate-level feature
  ('ak7pfcand_pt', 'ak7pfcan
  'ak7pfcand_id', 'ak7pfcanc

- Boolean 'jet_isW_ak7` is 1 if generator-level W boson is matched within dR<0.7 of the jet **and** both quark daughters have dR<0.7

```
int PFCandidate::translateTypeToPdgId( ParticleType type ) const {

  int thecharge = charge();

  switch( type ) {
  case h:      return thecharge*211; // pi+
  case e:      return thecharge*(-11);
  case mu:     return thecharge*(-13);
  case gamma:  return 22;
  case h0:     return 130; // K_L0
  case h_HF:           return 1; // dummy pdg code
  case egamma_HF:      return 2;  // dummy pdg code
  case X:
  default:     return 0;
  }
}
```



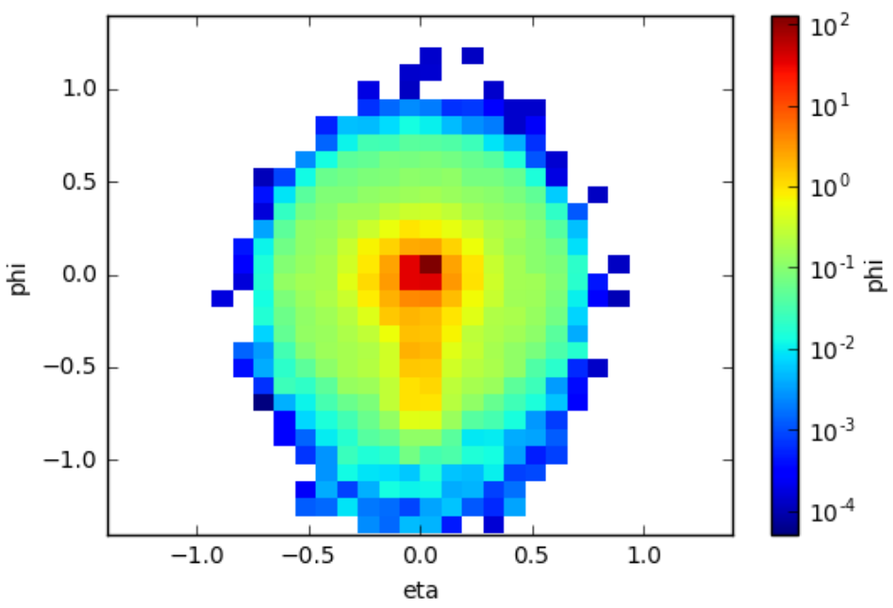
DATASET LOCATION

- Available on CMS LPC: root://cmseos.fnal.gov//eos/uscms/store/user/woodson/DSHEP2017/
- and Fermilab public dCache: root://fndca4a.fnal.gov//pnfs/fnal.gov/usr/hlml/persistent/DSHEP2017/
- Also be available on Amazon S3 storage (special thanks to B. Holzman): s3://ds-hep/
- Small subset available on Dropbox: https://www.dropbox.com/sh/zgrsduzuaclmzs2/AADvCY1i6uz3A5UhGrPrY30da?dl=0

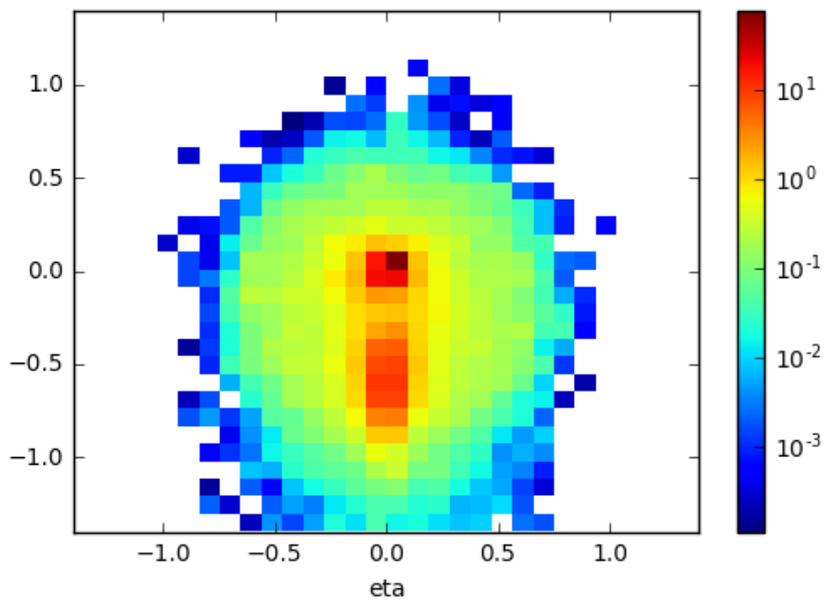
HANDS-ON SESSION

- Example notebook [github] shows how to access the data and build classifiers based on a fully connected NN and a convolutional NN
- Possible extensions: tuning metaparameters, testing different pre-processing steps, separating image representation into layers based PF candidate classes, training a recursive NN or a completely new network architecture we haven't thought of!
- Any feedback: missing features? different data structure? let us know!
- Sign up at google doc [doc] and join the slack channel [slack]

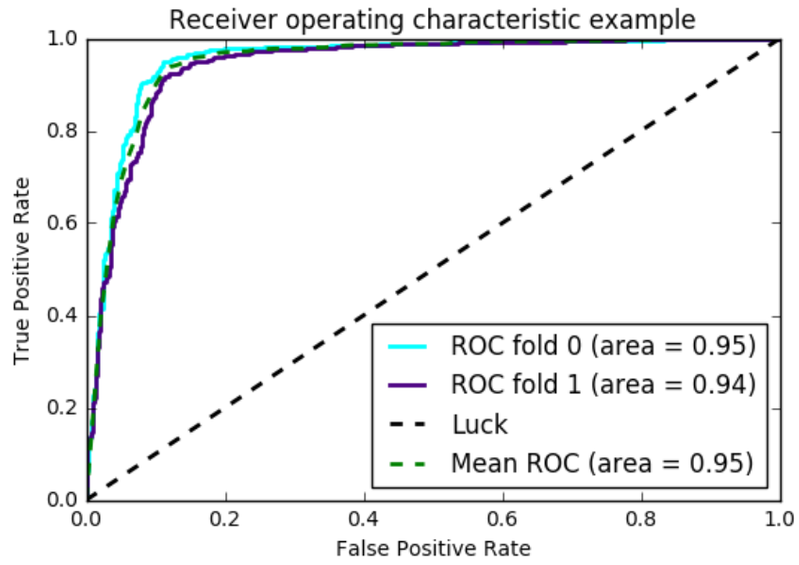
QCD jet image



W jet image



ROC curve



CMS OPEN DATA ML - JETS

BACKUP

Javier Duarte
Fermilab

DATASET FEATURES

- Event level features:
  ('run', 'lumi', 'event', 'met', 'sumet', 'rho', 'pthat',
  'mcweight', 'njet_ak7')
- Jet-level features:
  ('jet_pt_ak7', 'jet_eta_ak7', 'jet_phi_ak7', 'jet_E_ak7',
  'jet_msd_ak7', 'jet_area_ak7', 'jet_jes_ak7',
  'jet_tau21_ak7', 'jet_isW_ak7', 'jet_ncand_ak7')
- PF-candidate-level features:
  ('ak7pfcand_pt', 'ak7pfcand_eta', 'ak7pfcand_phi',
  'ak7pfcand_id', 'ak7pfcand_charge', 'ak7pfcand_ijet')

https://github.com/cms-sw/cmssw/blob/CMSSW_5_3_32/DataFormats/ParticleFlowCandidate/src/PFCandidate.cc#L148-L163

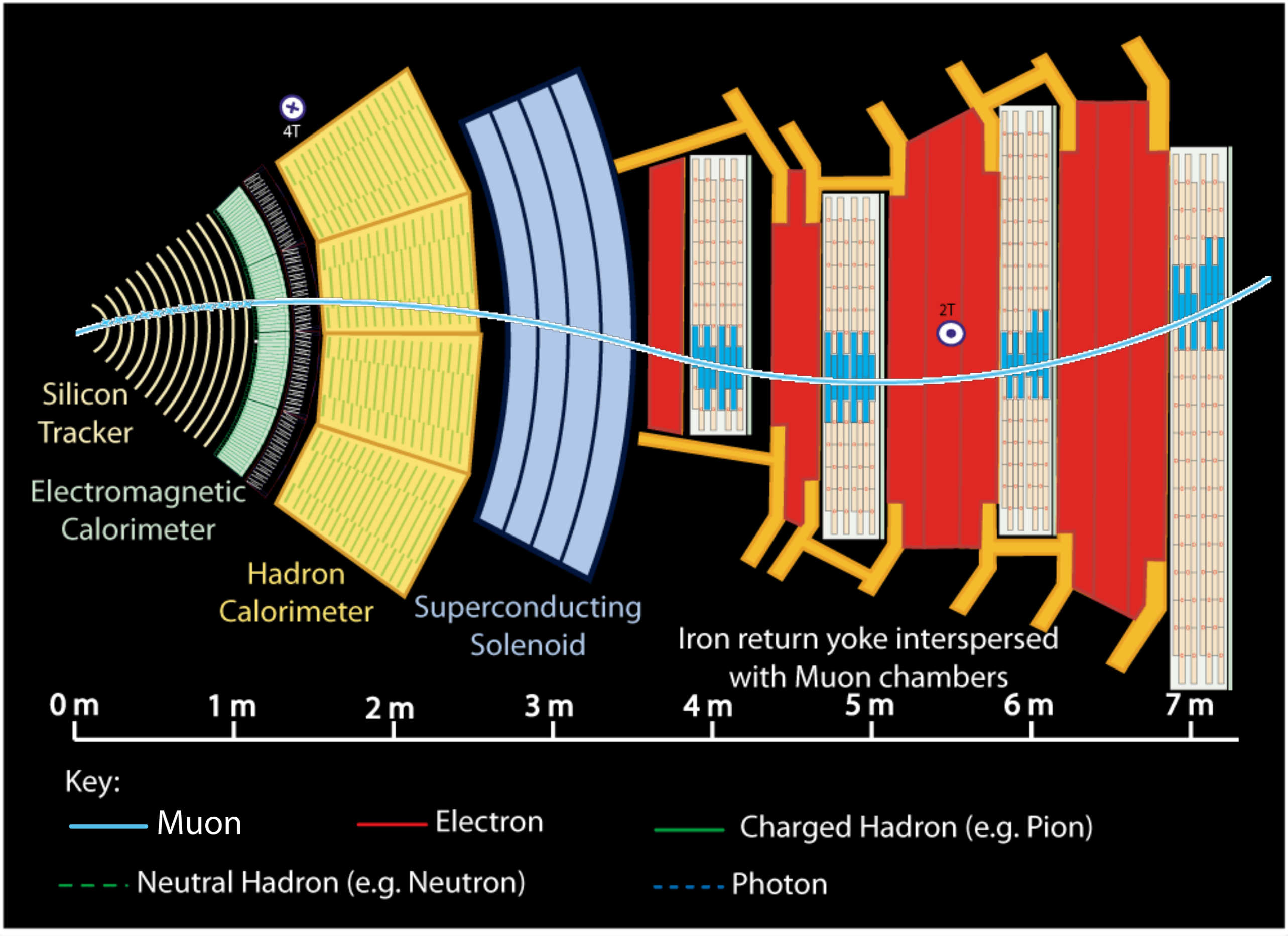
```
int PFCandidate::translateTypeToPdgId( ParticleType type ) const {

  int thecharge = charge();

  switch( type ) {
  case h:      return thecharge*211; // pi+
  case e:      return thecharge*(-11);
  case mu:     return thecharge*(-13);
  case gamma: return 22;
  case h0:     return 130; // K_L0
  case h_HF:          return 1; // dummy pdg code
  case egamma_HF:     return 2;  // dummy pdg code
  case X:
  default:     return 0;
  }
}
```

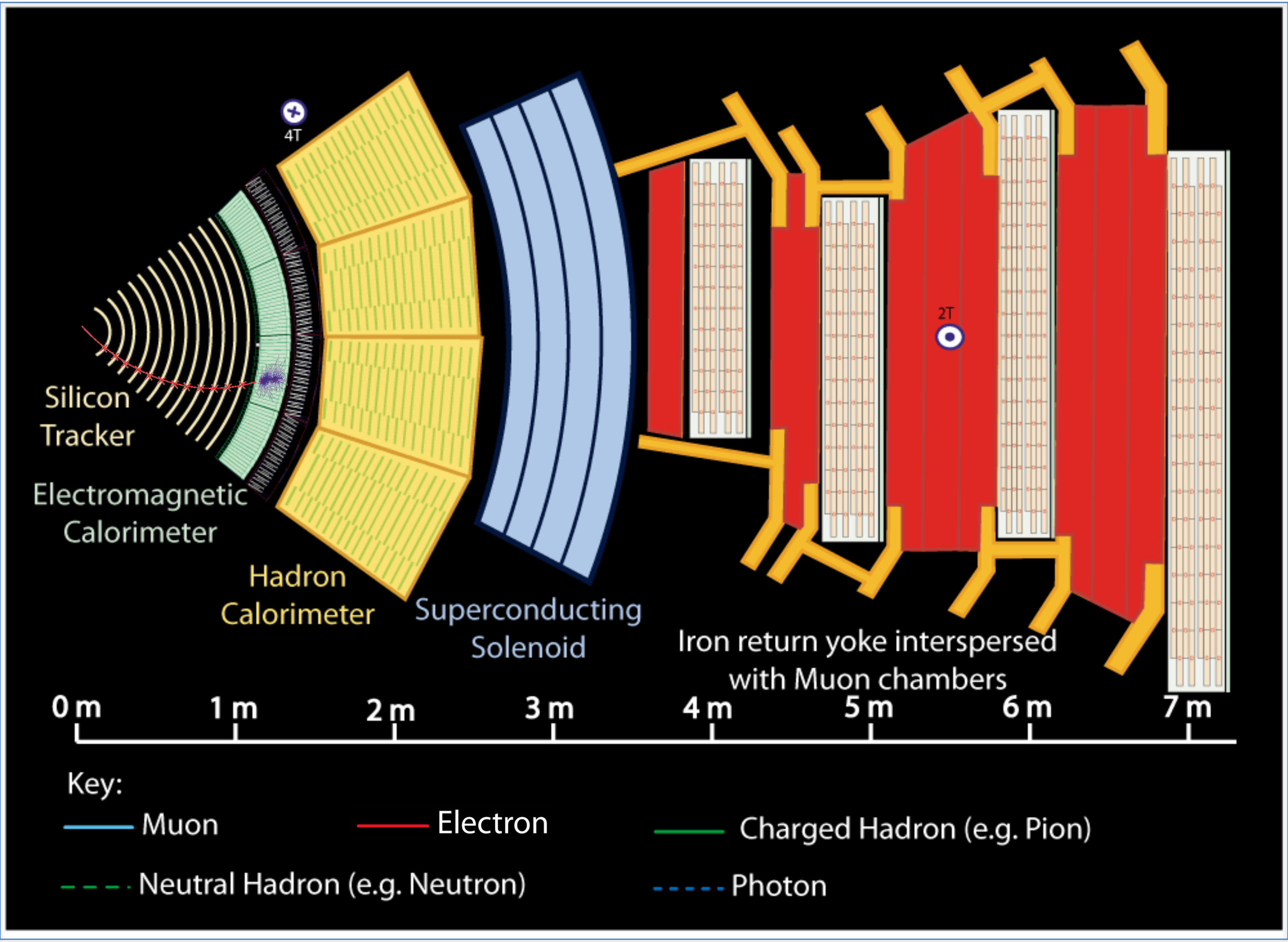
LARGE HADRON COLLIDER

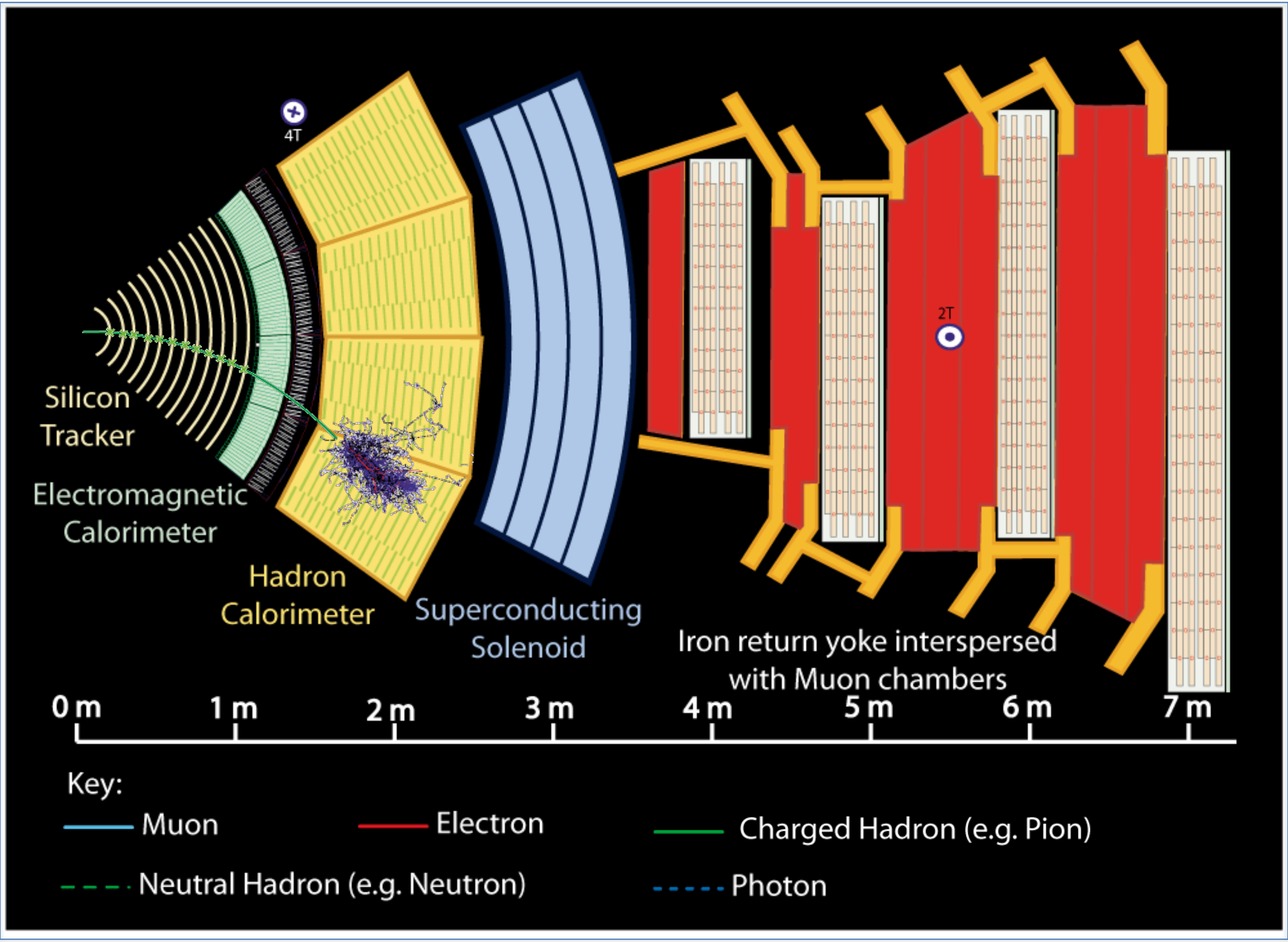
Lake Geneva

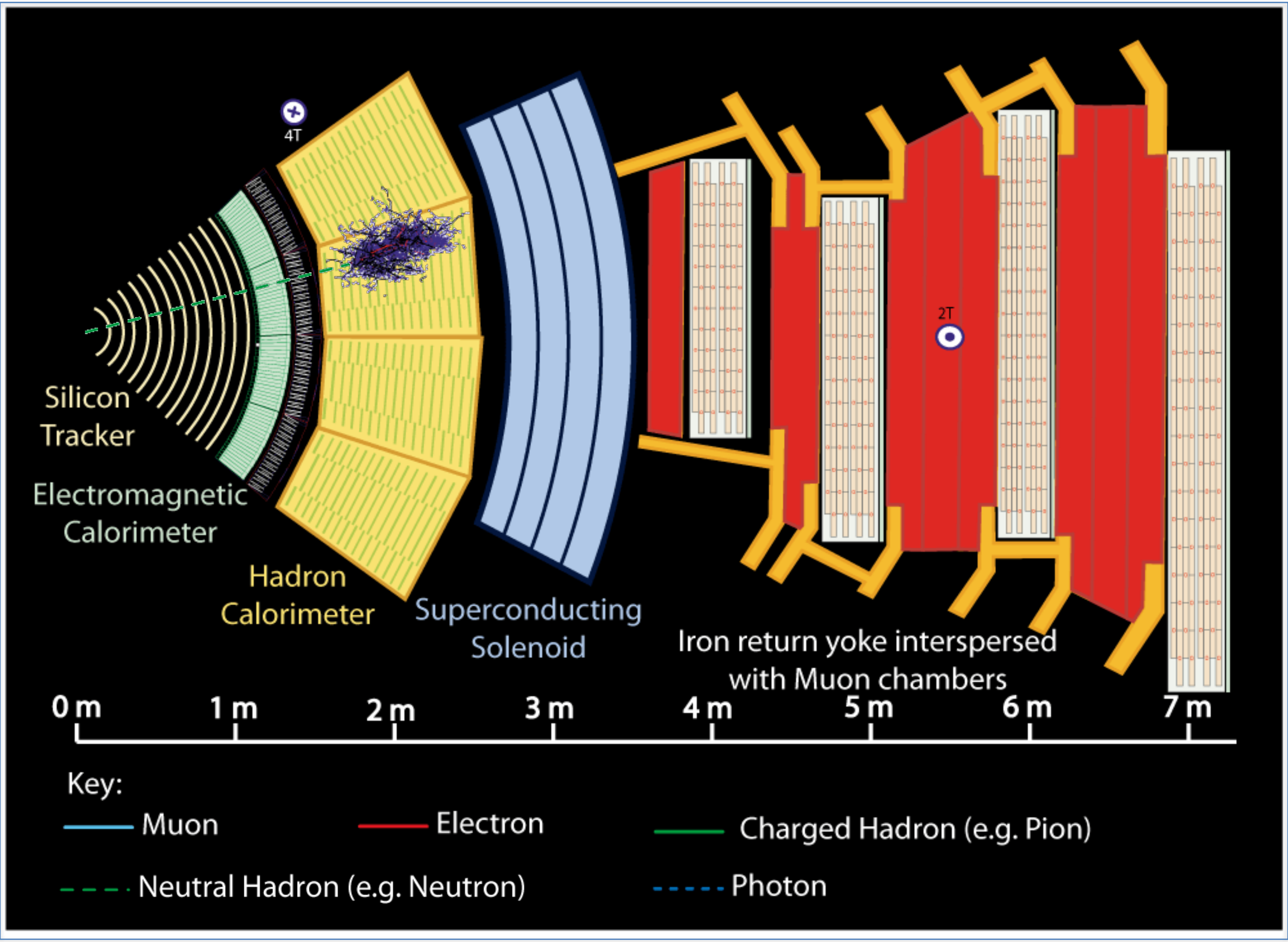
CMS

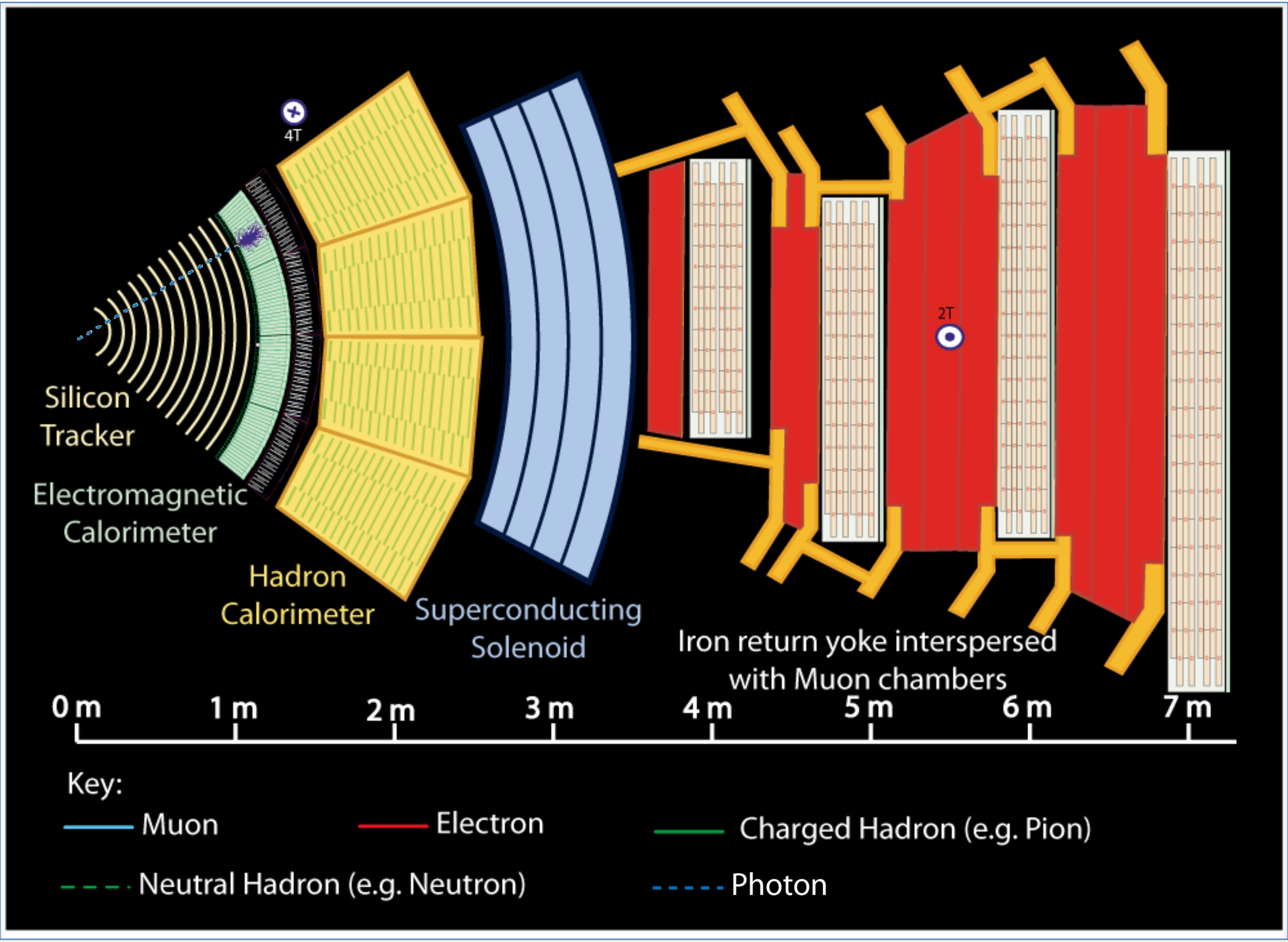
CERN


4T
2T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
Superconducting
Solenoid
Iron return yoke interspersed
with Muon chambers
0 m
1 m
2 m
3 m
4 m
5 m
6 m
7 m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon


4T
2T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
Superconducting
Solenoid
Iron return yoke interspersed
with Muon chambers
0 m
1 m
2 m
3 m
4 m
5 m
6 m
7 m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon


4T
2T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
Superconducting
Solenoid
Iron return yoke interspersed
with Muon chambers
0 m
1 m
2 m
3 m
4 m
5 m
6 m
7 m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon


4T
2T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
Superconducting
Solenoid
Iron return yoke interspersed
with Muon chambers
0 m
1 m
2 m
3 m
4 m
5 m
6 m
7 m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon


4T
2T
Silicon
Tracker
Electromagnetic
Calorimeter
Hadron
Calorimeter
Superconducting
Solenoid
Iron return yoke interspersed
with Muon chambers
0 m
1 m
2 m
3 m
4 m
5 m
6 m
7 m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon