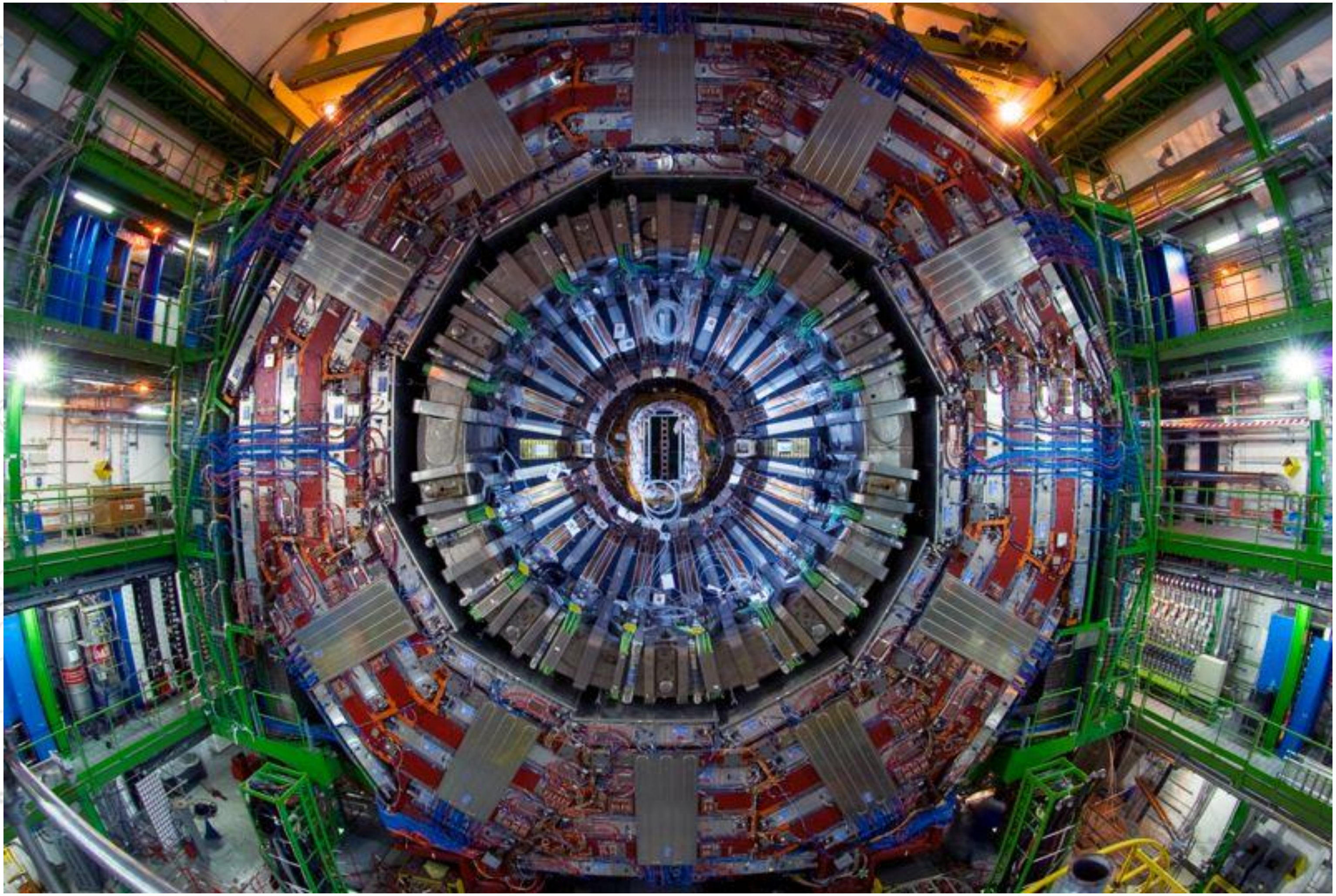
Data Quality Monitoring with Machine Learning at CMS

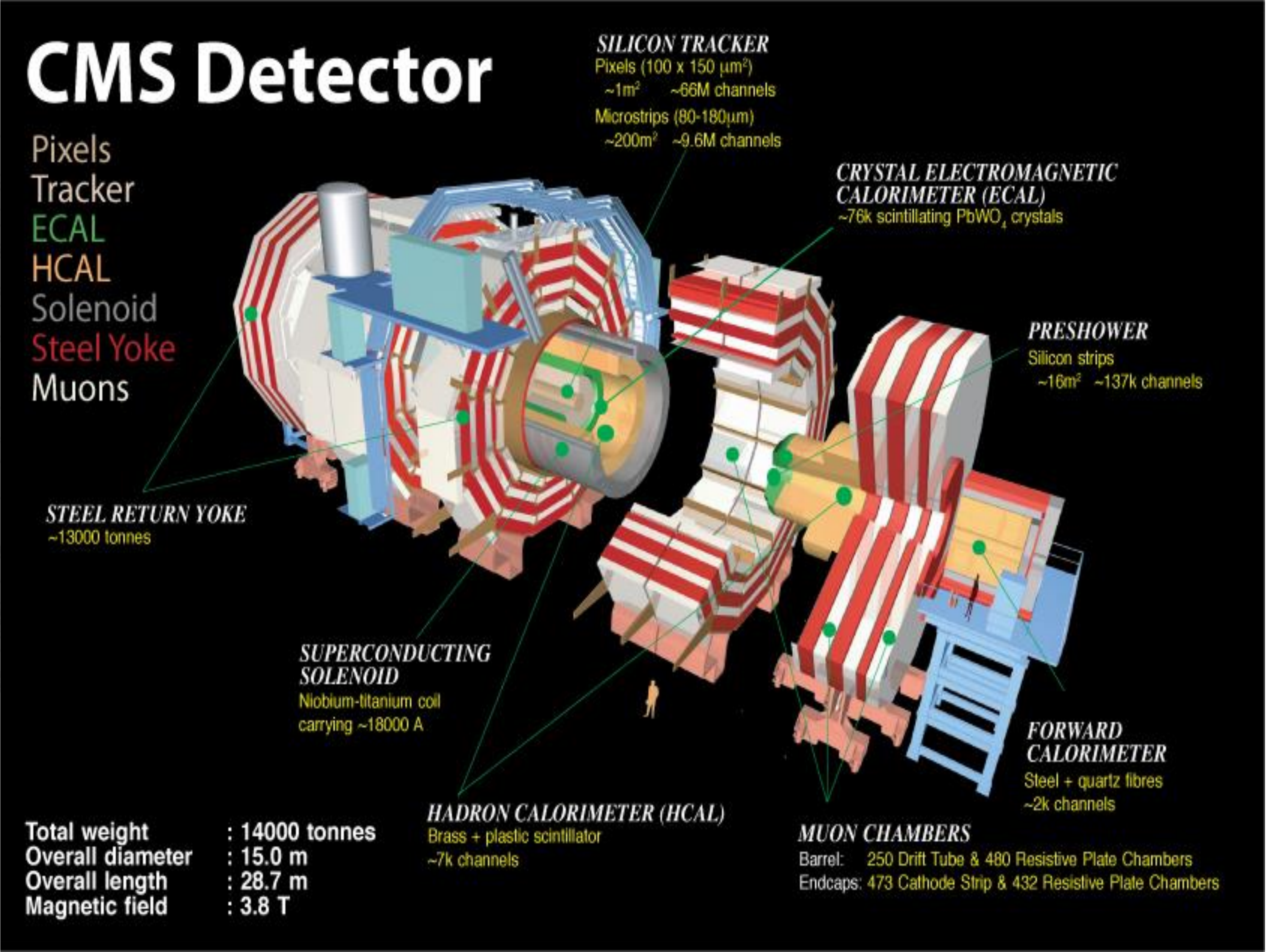
- 12/8/2016

Mentor: Professor Nural Akchurin
Postdoc Mentor: Dr. Federico de Guio
Hector F. Lacera Otalora

The Compact Muon Solenoid Detector



Background image: Shutterstock

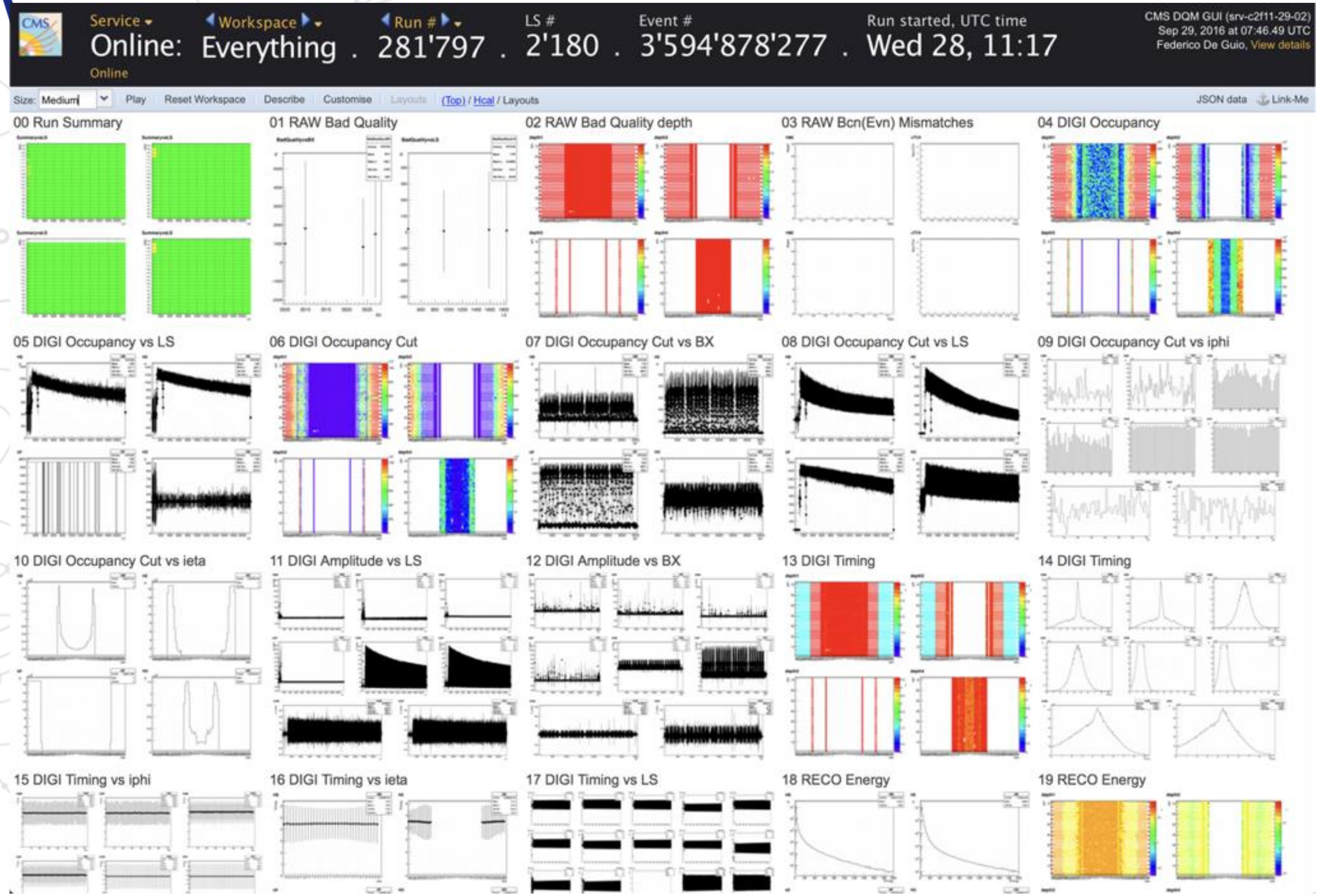


CMS Detector
Pixels
Tracker
ECAL
HCAL
Solenoid
Steel Yoke
Muons
SILICON TRACKER
Pixels (100 x 150 μm²)
~1m² ~66M channels
Microstrips (80-180μm)
~200m² ~9.6M channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76k scintillating PbWO₄ crystals
PRESHOWER
Silicon strips
~16m² ~137k channels
STEEL RETURN YOKE
~13000 tonnes
SUPERCONDUCTING SOLENOID
Niobium-titanium coil carrying ~18000 A
FORWARD CALORIMETER
Steel + quartz fibres
~2k channels
HADRON CALORIMETER (HCAL)
Brass + plastic scintillator
~7k channels
MUON CHAMBERS
Barrel: 250 Drift Tube & 480 Resistive Plate Chambers
Endcaps: 473 Cathode Strip & 432 Resistive Plate Chambers
Total weight : 14000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T

Data Quality Monitoring for CMS

- The CMS detector is complex and produces large amounts of data.
- A crucial part of CMS is to identify errors and problems in the detector.
- Central component: DQM GUI
  - High quality histogram viewing
  - Web server accessible worldwide

DQM GUI website



Machine Learning for better DQM

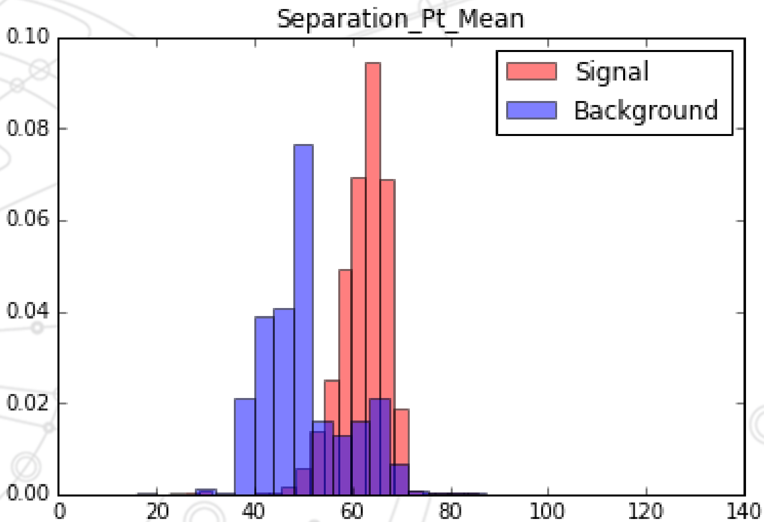
- Current method: Physicists on shift check incoming data from the detector and test it against trusted references.
- Better way: Implement Machine Learning techniques that recognize patterns and are able to accept or reject data in the trivial cases.
- Leave only non-trivial cases to experts.

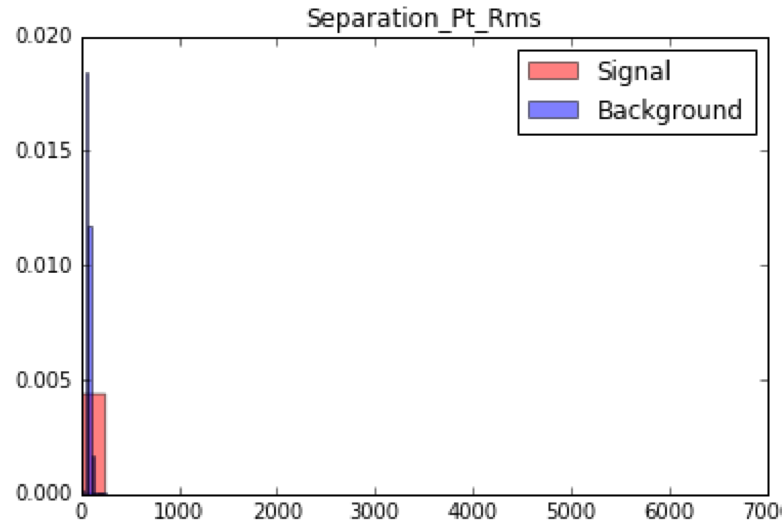
Objects of study: Jets

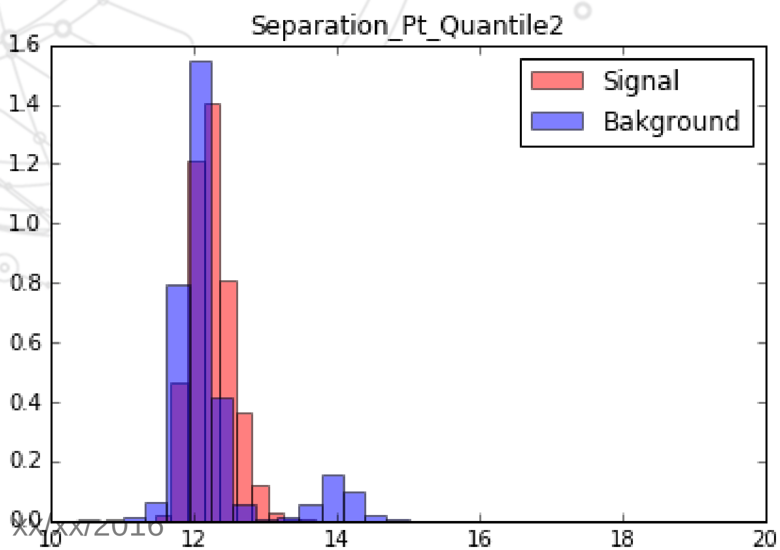
- Variables used:
- Pt = Momentum of the jet measured in transverse plane
(7 quantiles)
- Eta= Jet pseudorapidity
(7 quantiles)
- Phi = azimuthal angle
(7 quantiles)
- Vtx = Number of primary vertices in the event
(7 quantiles)
- Cross Section = Probability that two particles will collide and react in a certain way

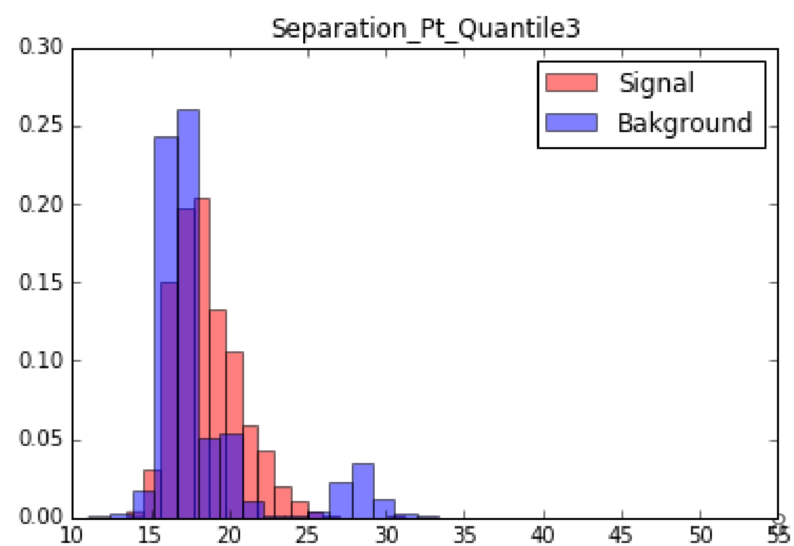
Total = 29 Features

Discrimination Power by feature

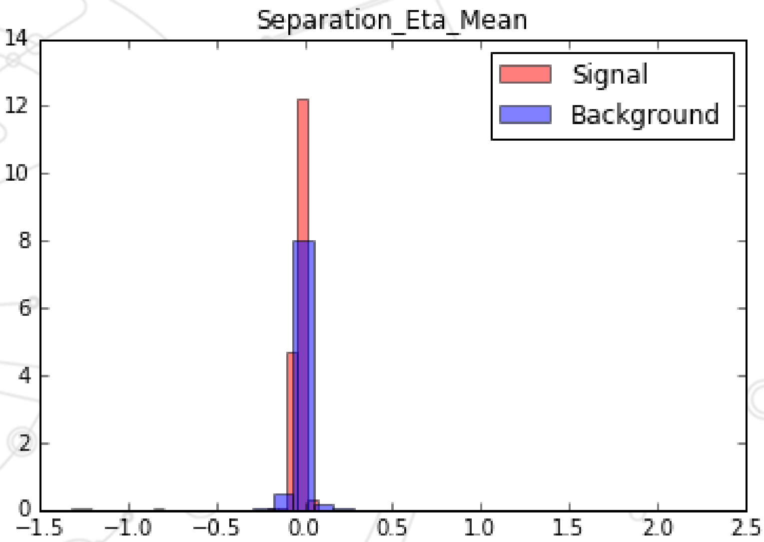


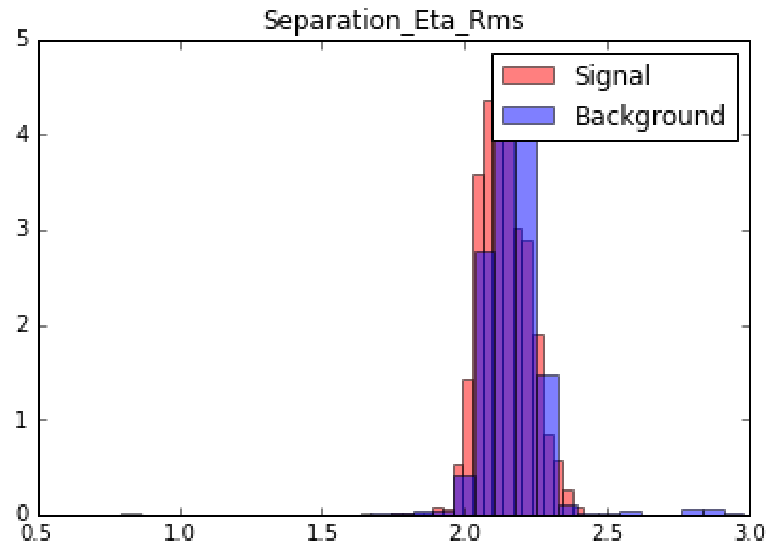


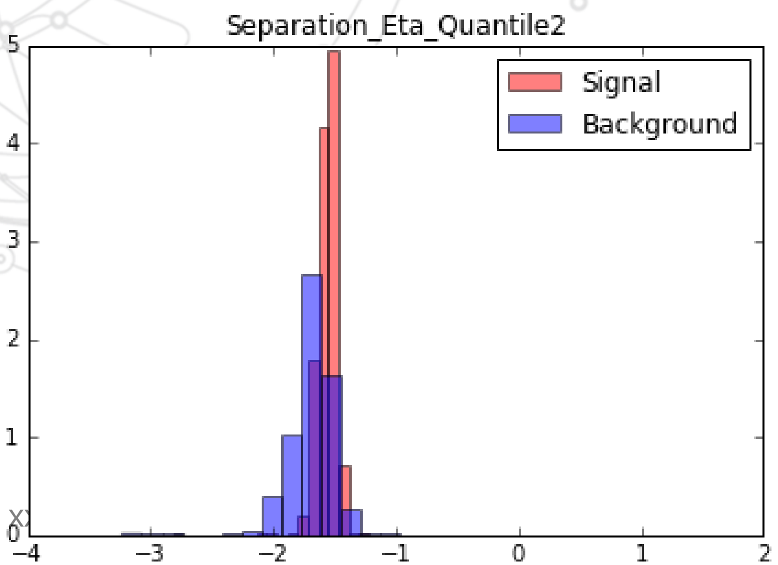


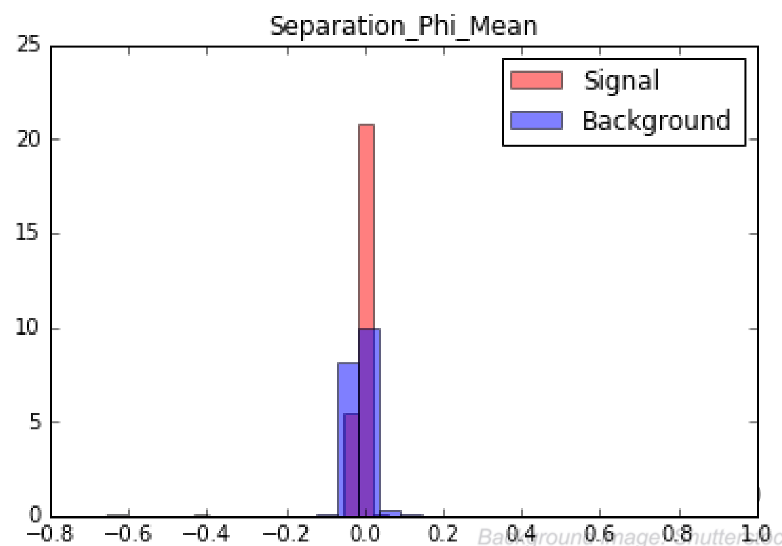


Discrimination Power continued









Summary of Data

- Entire Database from 2016 runs .
- ~ 250,000 events (lumisections).
- Data has “flags” telling us if it was signal or background.
- Data has “flags” telling us which subsystem failed during that run.
- Only 10% of bad data for 2016.
- Used 50% of data for training and 50% of data for testing.

Definitions

$$- \; Precision = \frac{tp}{tp+fp}$$ (Background rejection rate)

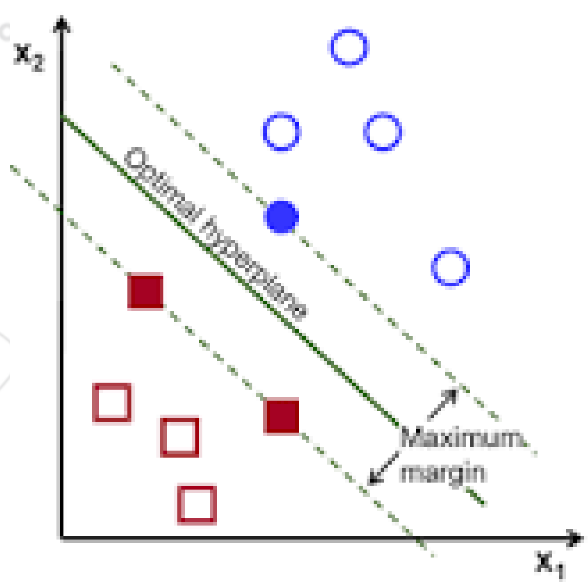
where tp is the number of true positives and fp the number of false positives.

$$- \; Recall = \frac{tp}{tp+fn}$$ (Signal Efficiency)

where tp is the number of true positives and fn the number of false negatives.

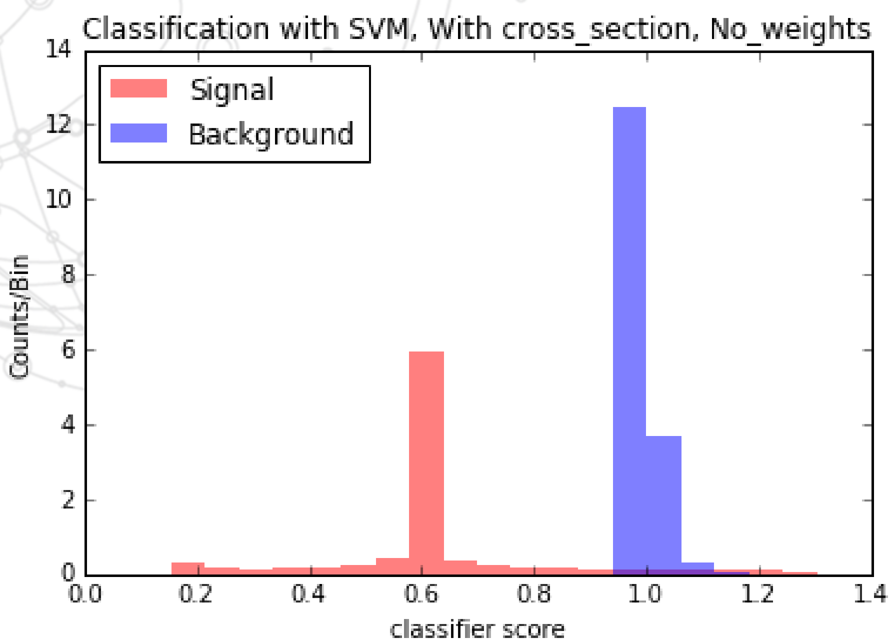
Support Vector Machines

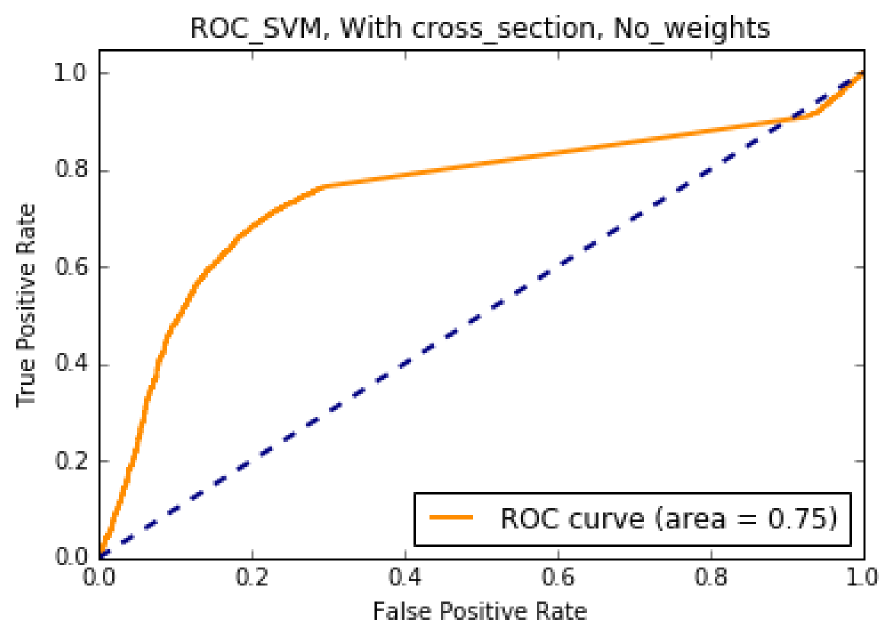
- A discriminating classifier that is described by a hyperplane.
- Given some training data, the classifier generates a hyperplane to classify new examples.



Training SVM with entire dataset

- - Training completed in ~ 1 hour and 10 min.
- - Precision = 0.97
- - Recall = 0.97

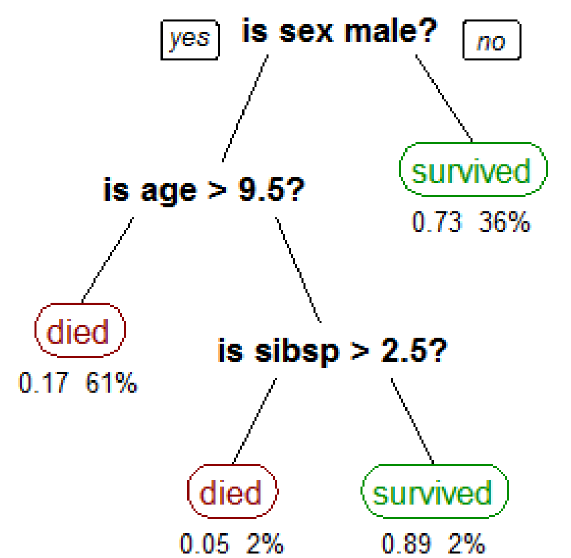




Background image: Shutterstock

Boosted Decision Trees

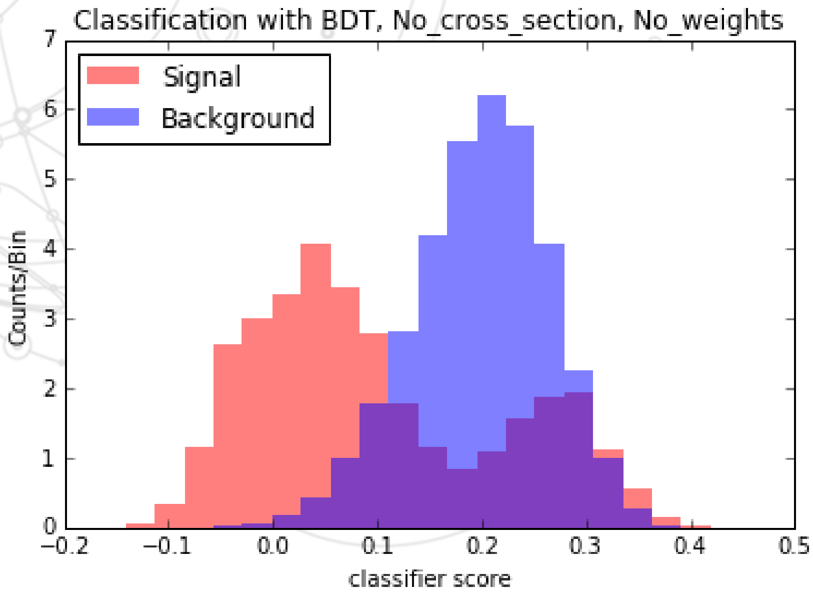
- **Decision tree learning** uses a decision tree as a predictive model.
- This model maps observations about an item to conclusions about the item's target value.
- Gradient boosting: produces a prediction model in the form of an ensemble

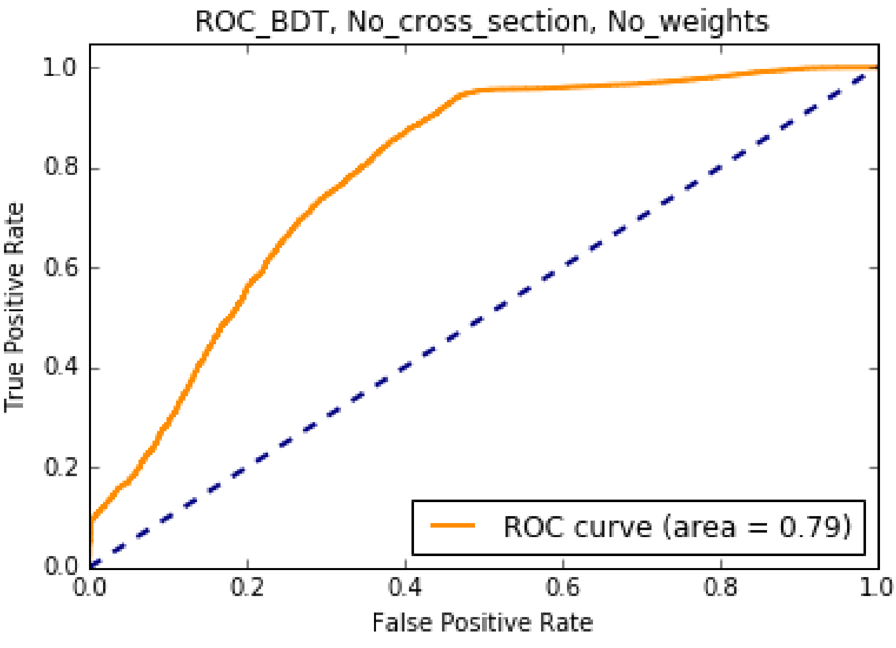


Background image: Shutterstock

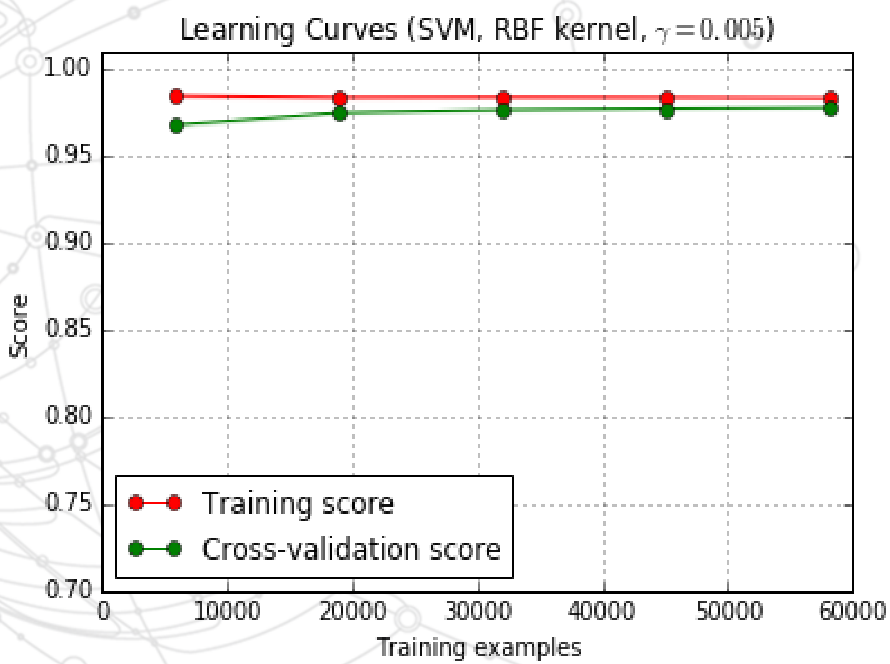
Training BDT with entire data

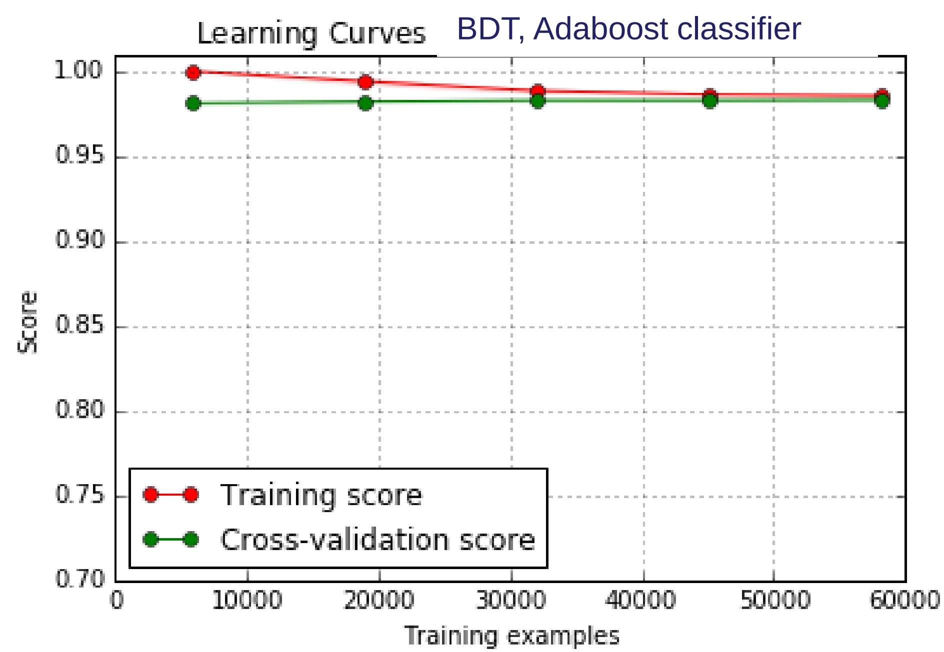
- - Training completed in ~3 minutes.
- - Precision = 0.92
- - Recall = 0.93





Did the classifiers learn?





Post mortem Analysis

Top ten bad runs	Classified Correctly (Reason for bad Run)	Misclassified
Run (275831)	Hcal	
Run (273017)	Hcal	
Run (273318)	L1tmu	
Run (274161)	ECal	
Run (280099)		Misclassified
Run (277218)		Misclassified
Run (274157)	Pixels	
Run (277220)	Hcal	
Run (277202)		Misclassified
Run (273301)	Hcal	

Conclusions and Work to be done

- After using several algorithms to perform training, SVM proved to have the best performance.
- The classifier was able to recognize more often failures in the Hcal.
- Proceed to a more realistic workflow in 2017.

My experience at CERN!

Questions?

Thank you!

Background image: Shutterstock