Machine Learning applications for Data Quality Monitoring and Data Certification within CMS

Vichayanun Wachirapusitanand on behalf of CMS collaboration
Chulalongkorn University, Bangkok, Thailand and European Organization for Nuclear Research (CERN), Geneva, Switzerland
Email: vichayanun.wachirapusitanand@cern.ch

What to expect from future data taking

In the next data taking campaign, Run 3 starting in early 2022, the CMS detector [1] is expected to collect twice the amount of proton-proton collision data collected during Run 2 (2015-2018).

300 fb^{-1} Recorded luminosity goal for Run 3 (2022 and beyond)

500 trillion Collision events expected to occur on CMS

98% Golden/recorded data ratio during CMS Run 2 (2015-2018) operations

Data Quality Monitoring (DQM) and Data Certification (DC) in a nutshell

DQM and DC teams oversee the data quality monitoring during data-taking (online) and data certification after data has been obtained from the detector (offline).

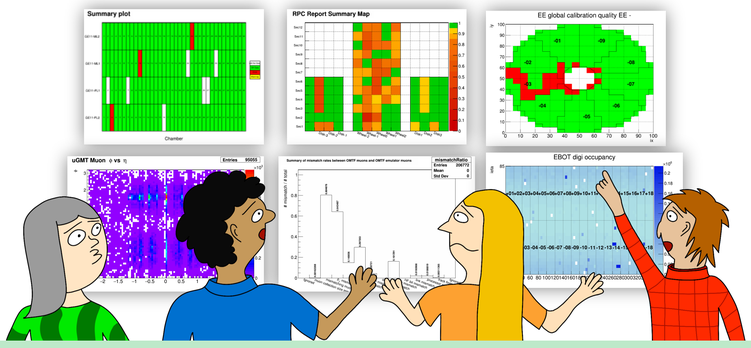
sends a small portion of data stream to

Online DQM backend

processes and sends histograms to

DQM GUI

DQM shifters then look at the histograms and consult with subsystem experts. They must react quickly when problems affecting data quality arise.



What is the problem with the current workflow?

Data from CMS is stored as a **run** (a period of time where CMS had stable detector conditions), which contains several **lumi sections** (the minimum bit of data-taking in CMS, roughly 23 seconds).

Human DQM shifters and subsystem experts certify the data on a run-by-run basis, both during online and offline.

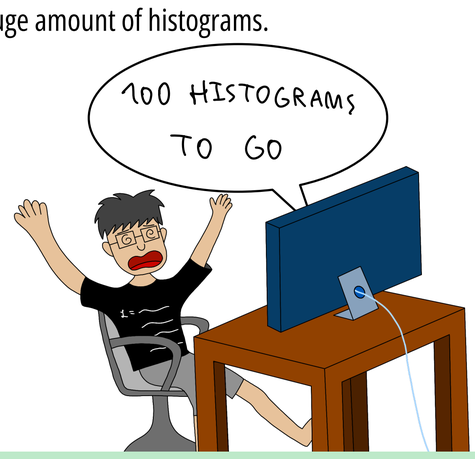
run 123456

lumi section	lumi section
lumi section	**BAD lumi section**

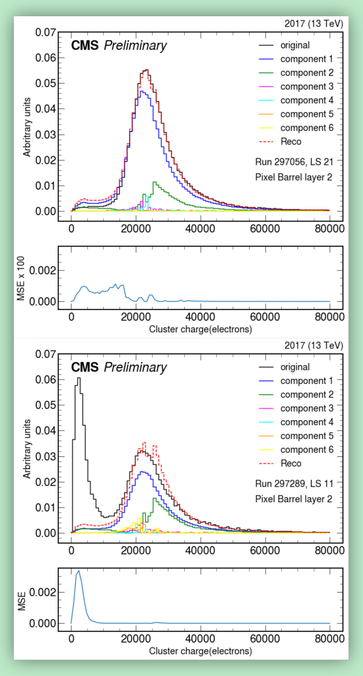
We now have two problems:

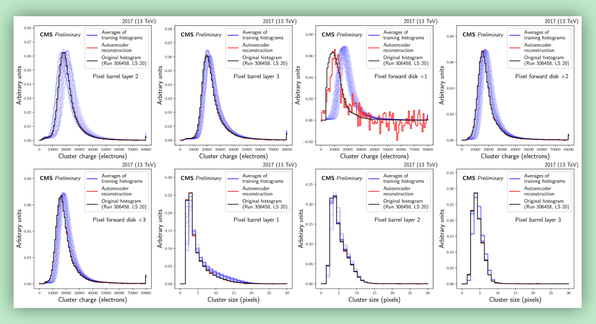
Not enough granularity: Shifters may miss anomalous lumi sections in a run

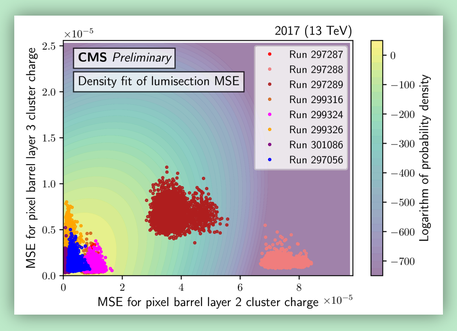
Human error and fatigue: Humans need to inspect a huge amount of histograms.

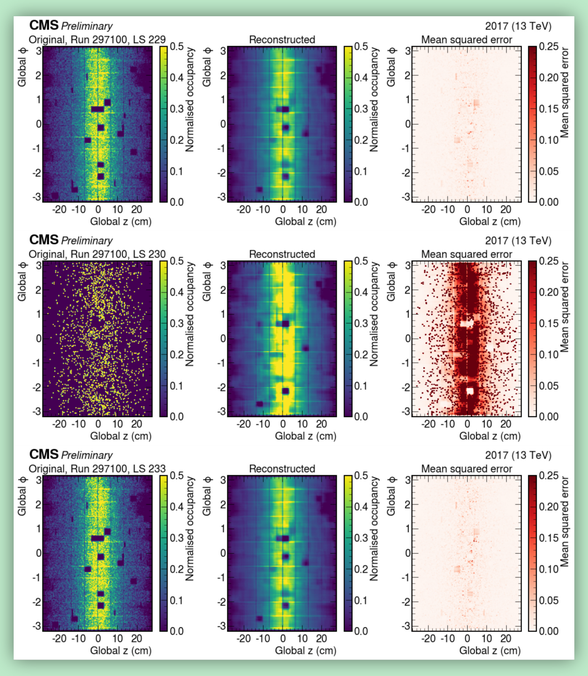


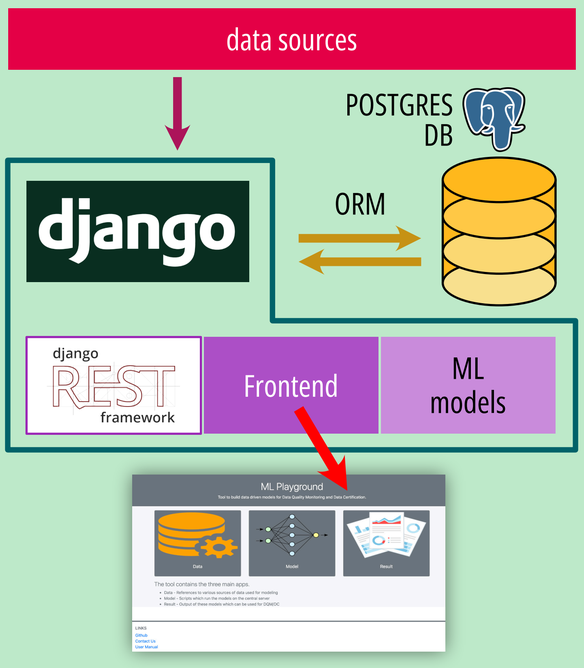
Current Machine Learning techniques being explored [2]











Non-negative Matrix Factorization (NMF) [3]

Unsupervised models for dimensionality reduction. We apply NMF in order to distinguish standard and anomalous data by the value of weights for the different components.

1-dimensional autoencoders

Train a dense feedforward autoencoder on a collection of 1D histograms. Keep the autoencoder simple and the training restricted, so it learns to reconstruct the good histograms (majority) but not the anomalous ones (minority).

2-dimensional residual networks [4]

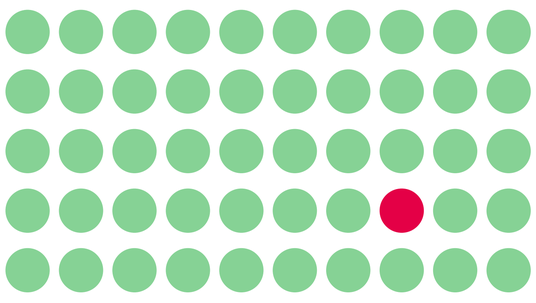
Train residual networks (ResNets) to encode and reconstruct 2D histograms, and classify anomalous histograms based on metrics, such as MSE loss. We can also apply PCA on MSE loss information to further classify and detect anomalous histograms.

Machine Learning playground

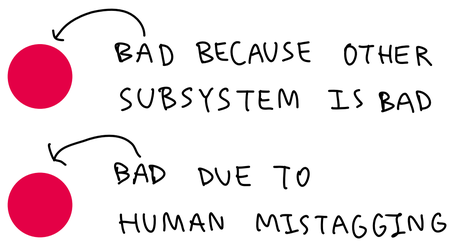
A framework based on Django intended to consolidate training dataset information, automate training of ML models, and generate training reports based on the ML model performance.

Challenges to developing ML-based anomaly detectors and integration to human-centric certification process

Anomaly detection **has a class imbalance problem**, since the number of bad histograms are small compared to good histograms.



Human-based labels may be incorrect, both due to run-by-run certification in the past and human error and fatigue. The final decision for such labels is also derived by subdetectors under the condition that **all subsystems must be good in order for the data to be labelled as good**.



This ML system **does not intend to replace human shifters and experts completely**, but rather assist humans to focus on problematic histograms and skip histograms that look good.

With High Luminosity LHC upgrade, expected to start operations at 2025, **the data taking rate is expected to increase significantly**, making the use of ML even more compelling. ML techniques will also have to adapt to ever-evolving detector conditions (calibration, alignments, and changing detectors). **Work is ongoing to assessing the amount of data needed for optimal training of ML algorithms.**

References

1. CMS Collaboration, J. Instrum. 3 (2008) S08004
2. https://twiki.cern.ch/twiki/bin/view/CMSPublic/TrackerMLDQMStudiesForDCNov2021
3. D. D. Lee and H. S. Seung, Adv. Neural Inf. Process. Syst. 13 (2000) 556-562
4. K. He, X. Zhang, S. Ren, and J. Sun, arXiv:1512.03385

Acknowledgements

Vichayanun would like to acknowledge Chulalongkorn University and the Chulalongkorn Academic into Its 2nd Century Project Advancement Project (Thailand) for their financial support.