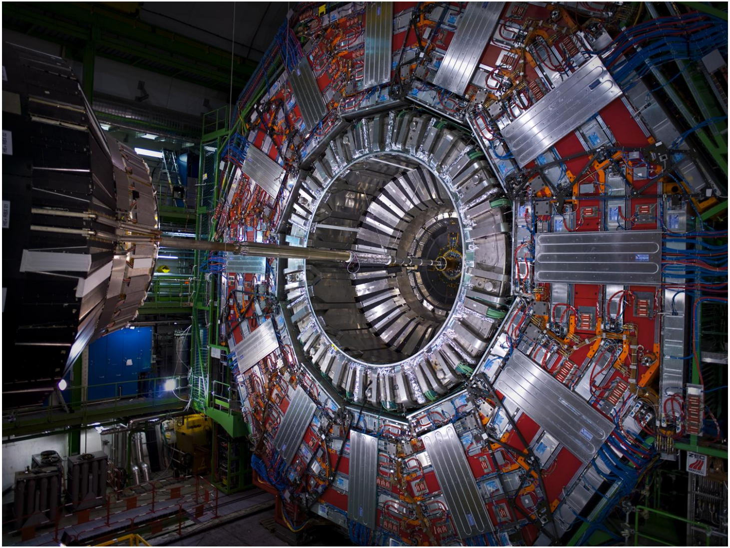
Data Quality Monitoring with Machine Learning in High Energy Physics

*Adrian Alan Pol[†,‡]   Cecile Germain[†]   Gianluca Cerminara[‡]   Maurizio Pierini[‡]

[†]University of Paris-Sud

[‡]CERN CMS



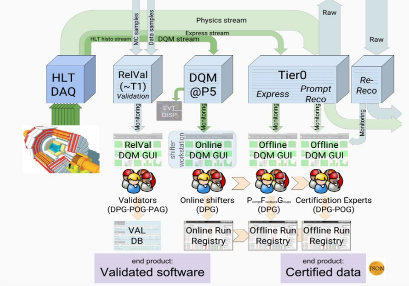
Compact Muon Solenoid (CMS) Experiment at LHC, CERN

Data Quality Monitoring (DQM) system

- a critical asset to **guarantee a high-quality data** for physics analyses
  - **live** during data taking (**online DQM**)
  - during offline data processing (offline DQM)
- online DQM assess data goodness and identifies emerging problems in the detector
- data with poor quality is flagged by **eyeballing dashboards** and comparing a set of histograms to a reference good sample

Identifying problems in real-time: summary of the current strategy

- identify problems in the detector & trigger system, e.g. read-out electronics errors
- fraction of the events with a rate of ~100Hz
- monitor 15 subsystems, each with unique parameters
- currently implemented **static thresholds** perform data reduction tasks
- based on results of those threshold tests and set of instructions, operator spots problems by visual inspection



DQM system used in CMS

Problems with current strategy

- **delay**: human intervention and thresholds require collecting sufficient statistics;
- **volume budget**: amount of quantities a human can process in a finite time period;
- **static thresholds don't scale**: assumptions we included all failure scenarios;
- **human driven decision process**: alarms based on shifter judgment;
- **changing running conditions**: reference samples change over time;
- **manpower**: the effort to train a shifter and maintain instructions

Can we solve or reduce some/all of the above problems?

Let’s have a

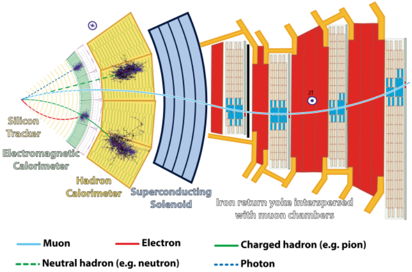
self-sustaining and **autonomous**,

reliable and **fast**

data quality monitoring framework at CMS

Muon detector occupancy data reduction

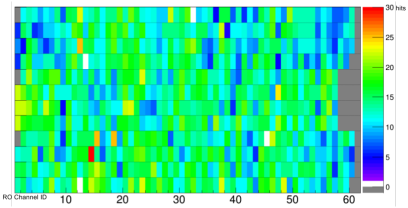
- the number of muons crossing a certain region (single electronic read-out channel) of the muon chambers
- expected behavior: occupancy of the hits with small variance between neighbouring read-out channels



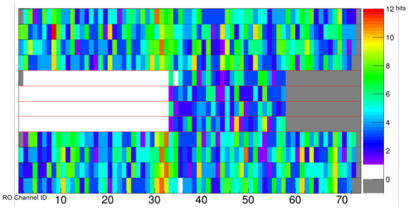
CMS Detector Slice

Muon detector occupancy data reduction cont.

- noisy or under-performing area is reported as a problem
- each layer (row) is a separate sample to detect problems with a thinner granularity $X = \begin{pmatrix} x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix}$, m - layer, n - cell number



Occupancy plot: expected behavior



Occupancy plot: fault

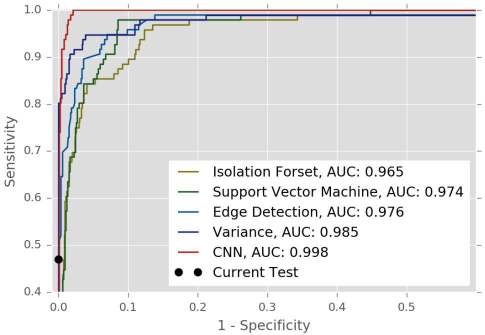
Results

- **supervised learning**: binary classification
- labels, by experts: 4300/476, ~.1 positives
- 80:20 train/test split
- stratified folds for cross-validation
- trained **SVM**, **Isolation Forest**, and **CNN** with 1D convolutions

```
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_7 (Conv1D)            (None, 47, 5)             30
_________________________________________________________________
max_pooling1d_7 (MaxPooling1 (None, 10, 5)             0
_________________________________________________________________
flatten_7 (Flatten)          (None, 50)                0
_________________________________________________________________
dense_13 (Dense)             (None, 128)               6528
_________________________________________________________________
dense_14 (Dense)             (None, 2)                 258
=================================================================
Total params: 6,816
Trainable params: 6,816
Non-trainable params: 0
_________________________________________________________________
None
```

Results cont.

- CNN outperforms other approaches, 0.95 hit rate for 0.01 fall-out rate.



Bottom line

- the current paradigm of the quality assessment in the CMS collaboration is based on the scrutiny of a large number of histograms by detector experts comparing them with a reference
- the project aims at applying machine learning techniques to the automation of this process allowing the check of large volumes of data in real-time and improving the ability to detect unexpected failures
- the muon detector exercise is rather about learning if/how to plug machine learning technology in the CMS DQM, than solving real problems
- other projects happening within CERN OpenLab:
  - certification of data for physics analysis with Yandex
  - online DQM with IBM