Virginia Azzolini

# Machine Learning for CMS Online Data Quality Monitoring and Data Certification

# Limits of a Human-based Data Certification

- Volume budget
  - Limited amount of quantities that a human can process in a finite time interval
- Time delay
  - Online: # plots, can cause delay in spotting a problem or cause a transient problem to be overlooked
  - Offline: reconstruction data time +human intervention = ~ 1 week → Need PFG intermediate step
- Expensive, in terms of human resources
  - Duplication of effort (many detector and physics object experts) on weekly basis
  - There is a possibility that the monitoring decisions can vary from shifter to shifter.
- Makes assumptions on our level of understanding
  - Scrutiny of a large # of histograms in comparison with a reference visually or via automatic threshold checking. Static threshold, led by actual conditions understanding, do not scale
- Strategy tailored to certain failure modes,
  - the certain set of quantities monitored might not have enough discriminatory power against all the possible problems

Good news is the current system works
but volume of data has grown so large it is becoming increasingly difficult to QA all data

We aim to incorporate modern ML techniques to perform quality in future intelligent archives

# A way forward: automatisation

CMS started in 2017 long-term programs to automatise the system

. started from the use cases
pbl to address, definition working roles, solution driven users activities, business process

. continued with establishing models

. developing dedicated applications as single blocks of ML-based DQM

. will commissioning the new system by running it in the shadow of the shifters/experts for 2018

. find similarities across problems and solutions and design a common framework of operations

CMS projects aim to improve the operations of the two domains of CMS quality assessments:
#ml4dqm : online data quality monitoring
#ml4dc : offline data certification

courtesy of M.Pierini, PPD workshop 2017 - updated to 2018 status

# ml4dqm: the big picture

## ml4dc: the big picture

# Formalizing the Problem:
# Continuously Supervised learning approach,
# Semi-(Un)supervised learning … both?

GOAL: Offline we aim to reduce the human burning load and to model good target Physics Object

## Supervised learning

Aim: Assist Data Quality managers by filtering most obvious cases, 3-class classification
model: Gradient Tree Boosting classifier + 10-fold cross validation scheme to estimate cuts
Pro: Data reduction & Limit the need of human intervention and save experts burnout
Cons: Indicate the uncertain LS ranges, unfortunately the certification happen mostly by run,
so even if we would give the experts a limited LS range to certify this would not solve their life

## Supervised learning

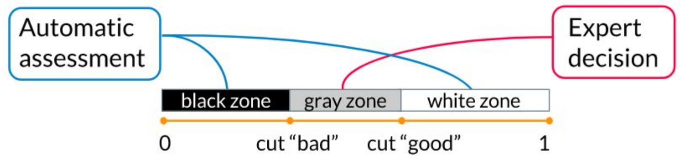
Aim: Identify channels in which anomalies occurred. If photon is bad, may I still use for muon analysis?
model: Multi-head NN
Pro: Given CMS global tag, model restores quality of data for each channel separately
Use: Identification good channels in anomalous data samples

## Semi(Un)-supervised learning

Aim: Find a model that not only can mimic data quality prescription on today's data,
but will make good predictions based on future data which will be different
model: Auto-Encoders
Pro: Possibility to learn which feature would be more responsible of the failure of data,
point the experts on the right direction, perform a solid QA per LS , results intuitive interpretability

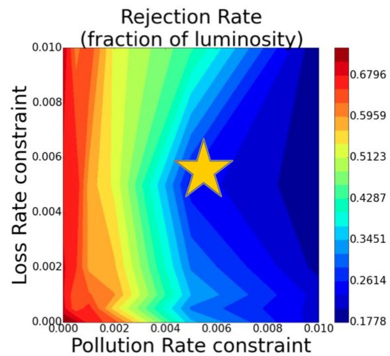
# Status of art of ml4dc application

## Data Reduction & human effort ( with Yandex)



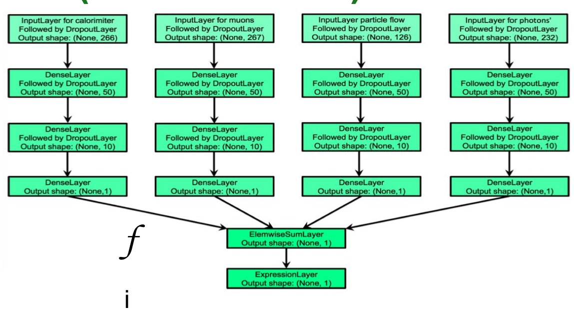

<u>Performance:</u>
20% saved person power for
Pollution and Loss rates 0.05%
80% saved person power for
Pollution and Loss rates 0.5 %

published on 2010B
CMS opendata



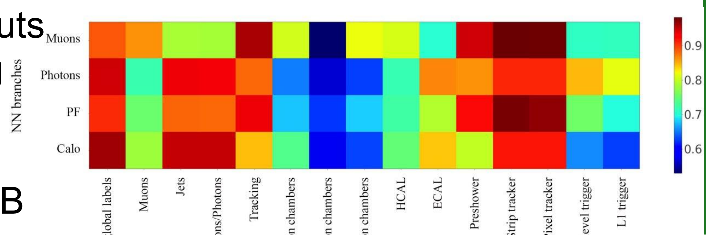

## Channel decomposition ( with Yandex)

input feature
related to the
channel(phys obj)

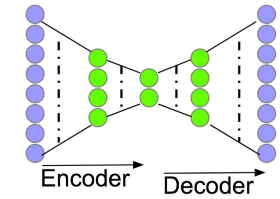



<u>Performance:</u>
correlation between
subnetworks' outputs
and corresponding
subsystem labels

published on 2010B
CMS opendata

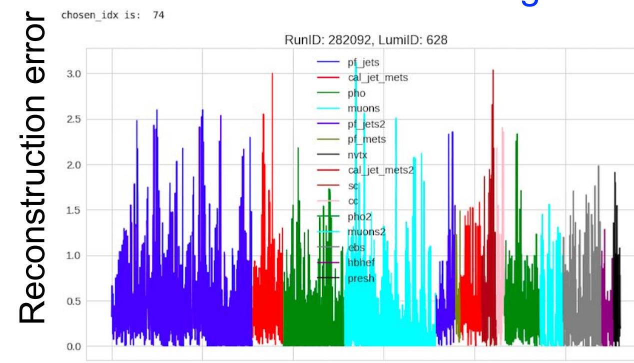



## Model the output, $E_X(X - g(f(X)))^2 \rightarrow \min$

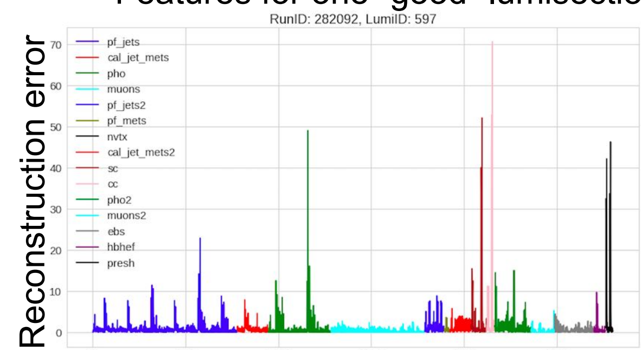



<u>Performance</u>: scrore 0.97

<u>Easy Interpretability</u>
peaks of higher reconstructed errors for anomalous LS
uniform distribution of reconstruction errors for good LS




Features for one "good" lumisection




Features for one "bad" lumisection

# #ml4dc project:
# Deliverables and Timeline

## R & D

. preprocessed data & storage location

. labeling of data

. application
  . data reduction
  . channel decomposition
  . modeling & feature indicator

. ML documentation & expertise support

. project documentation

. results web interface for commissioning in the shadow of DC

. local WS in RR

. hand filling of LS in RR for commissioning

## Consolidation

. involvement of sub-syst as testers

. operation workflow commissioning

. DQM area and layout filled by ML outputs only

. injection of QA decision in RR via API

## Application

. wrapping 2018 workshop

. use on DC operations support

# ml4dc project: Milestones

. Jan- Feb: ( on 2016 data)
Yandex/HSE team concentrating on sophisticated method for anomaly detection
CMS team resolve dilemma about Lumi as feature
data preparation, model and notebook documentation
ml4dc project supporting documentation for presentations and publication

. March: given good data list, preprocess 2017 data and make them available

. April: CMS team welcome new member
counter validation of 2017 Rereco certification via ML

. May : use training on 2017 to predict 2018 quality

. June : collect success/ unsuccess statistic in view of CHEP presentations (july )
( Yandex/HSE (ML section) and CMS (ML section) abstract submitted)

. Aug: definition of workflow in real data taking, how to split samples and train and re-train frequency
work on general framework
learn lessons from simulation of “normal workflow”

. Sept/Oct: wrapping 2018 workshop: lessons learnt, results, promises and vision

After long shutdown
full implementation of ML algo in the normal workflow as support to the normal operations

# ml4dc project:
# Person Power

. interest: very high

core group : . about 11 people ( listed at the end of this talk)

. well eclectic group of expertises, particle physicists,
data scientists, computer engineers evolving physicists

. opportunity for young members

2017: summer student developed a standalone muon algo

2017: technical student responsible of the 2016 CMS-only AE, finishing now

2018: technical student follow up the CMS-only AE effort

2018: summer students

ml4dc fan and curious about: many more

scouting for possible efforts ongoing the shadow

Please start to attend our meeting and bring in your ideas, we’re open to alternatives

# ml4dqm and ml4dc:
# Working Meetings

working meeting is the key formula desired

Rhythm: 1 h in alternate weeks
intermediate chats on need or demand

2017 meeting day and time: adjusted to involved people availability

2018 meeting day and time:
Looking forward for a consistent time lot for easing life of our active members and
as reference for the irregulars and newcomers
Doodle to be circulated scouting for new day and time slot: Please fill it to get involved!
Wish we could accommodate everybody needs, unfortunately being a working meeting,
key people availability weight more in the final decision

Working e-group: ( get involved! please subscribe directly )
cms-ml4dqm@cern.ch
cms-ml4dc@cern.ch

Indico pages:
(under PPD DQM-DC meeting indico page): https://indico.cern.ch/category/3904/
Machine Learning for Data Quality Monitoring
Machine Learning for Data Certification

*Thank you !*

*questions ?*

# Machine Learning for CMS
# Online Data Quality Monitoring ( with* IBM)

V. Azzolini,
M. Andrews
G. Cerminara,
N. Dev
E. Eskandari
R. A. Gerosa
C. Jessop
N. Marinelli
T. Mudholkar
M. Pierini,
A. Pol,
A. Vartak
J-R. Vlimant

N. Twebti,
U. Walter,
N. Altaf,
M. Lucrezia

# and
# Data Certification ( with* Yandex / HSE)

V. Azzolini,
G. Cerminara,
F. De Guio,
G. Franzoni,
M. Pierini,
A. Pol,
F. Siroky,
J-R. Vlimant

M. Borisyak,
D. Derkach,
O. Koval,
F. Ratnikov,
A. Ustyuzhanin

PPD projects via CERN openlab as facilitator

BACKUP

# Automated Data Quality Assessment

2 possible approaches:

- robots like anomaly detection

  Routinely monitor same measured or reconstructed properties – “DPG”
  Rely on set of statistics and rules, and automatically state normal vs abnormal behaviors

  Pro: Immediate benefit of save human intervention
  Pro: 1 to 1 verification
  Con: Constructing these statistics requires an exhaustive knowledge of the detector and all possible anomalies
  Con: no easily scalability with data volumes and detectors configurations changement

- Machine Learning-based Automated Anomaly Detection

  Online: system aim to monitor XYZ detector occupancy data, streaming in and updating plots every LS
  Offline: system is based on the measured or reconstructed physical properties – “POG/PAG”

  Pro: statistics can be learned directly from data →possibility of automated detection of anomaly
  Pro: reduce the human burning load
  Pro: adaptable to different experimental setups (including changes in the detector)
  Pro: not orthogonal to expert statistics approaches, expert statistics injection into the feature set is a starting point for improvement of the system
  Con: commissioning time

# Data Access Policies

This is not straightforward for cooperation beyond collaboration:
◊ to be useful, the system needs access to data in real time.
◊ collaboration restricts access to physics data during grace period
◊ Yandex is not a member of CMS collaboration

Practical solution:
◊ CMS members of the team provide data processing and collecting statistics over periods of data taking
  ◊ collected data contain only integrated information, no information from individual events
◊ Yandex members of the team develop classification algorithms based on these integrated features