Virginia Azzolini

Machine Learning in Data Quality Monitoring

… a CMS point of view

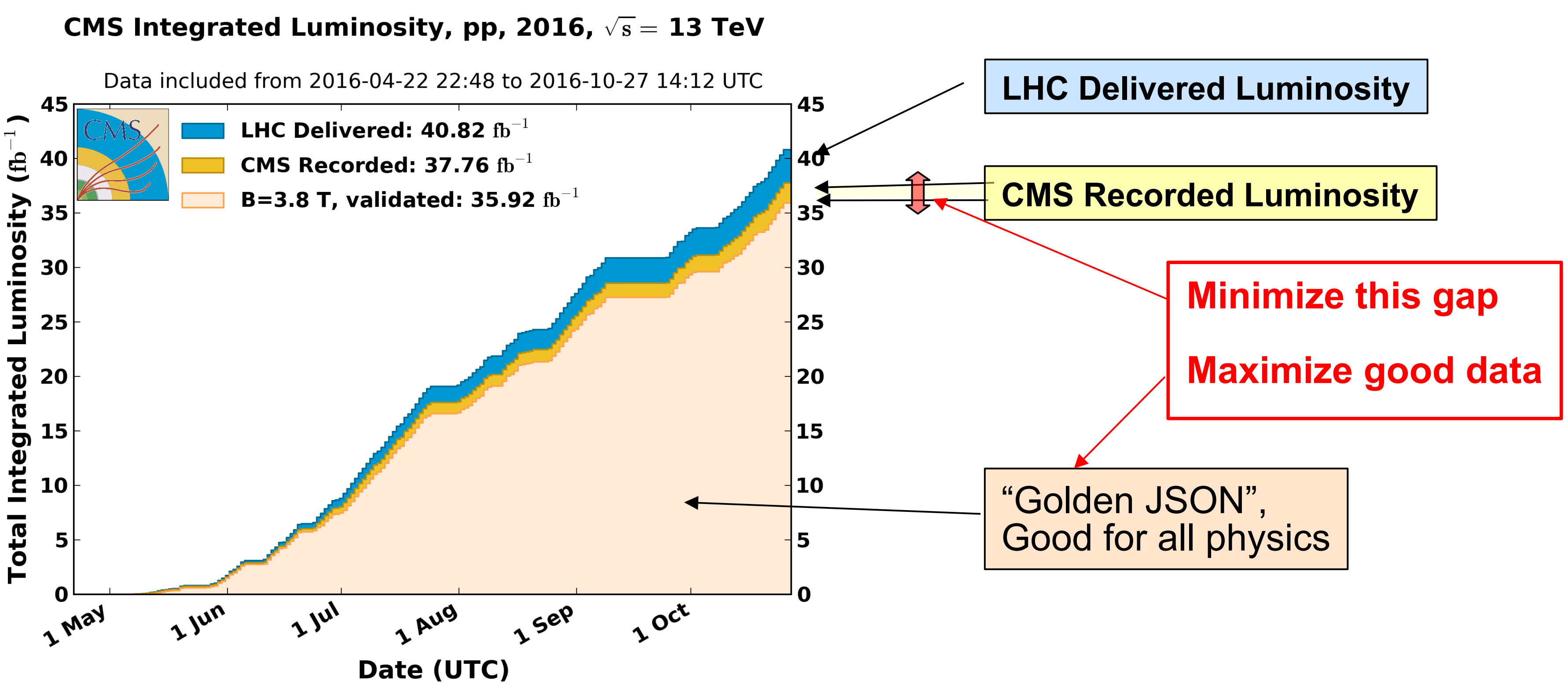
CERN openlab workshop on
Machine Learning and Data Analytics

April 27th, 2017

Goal

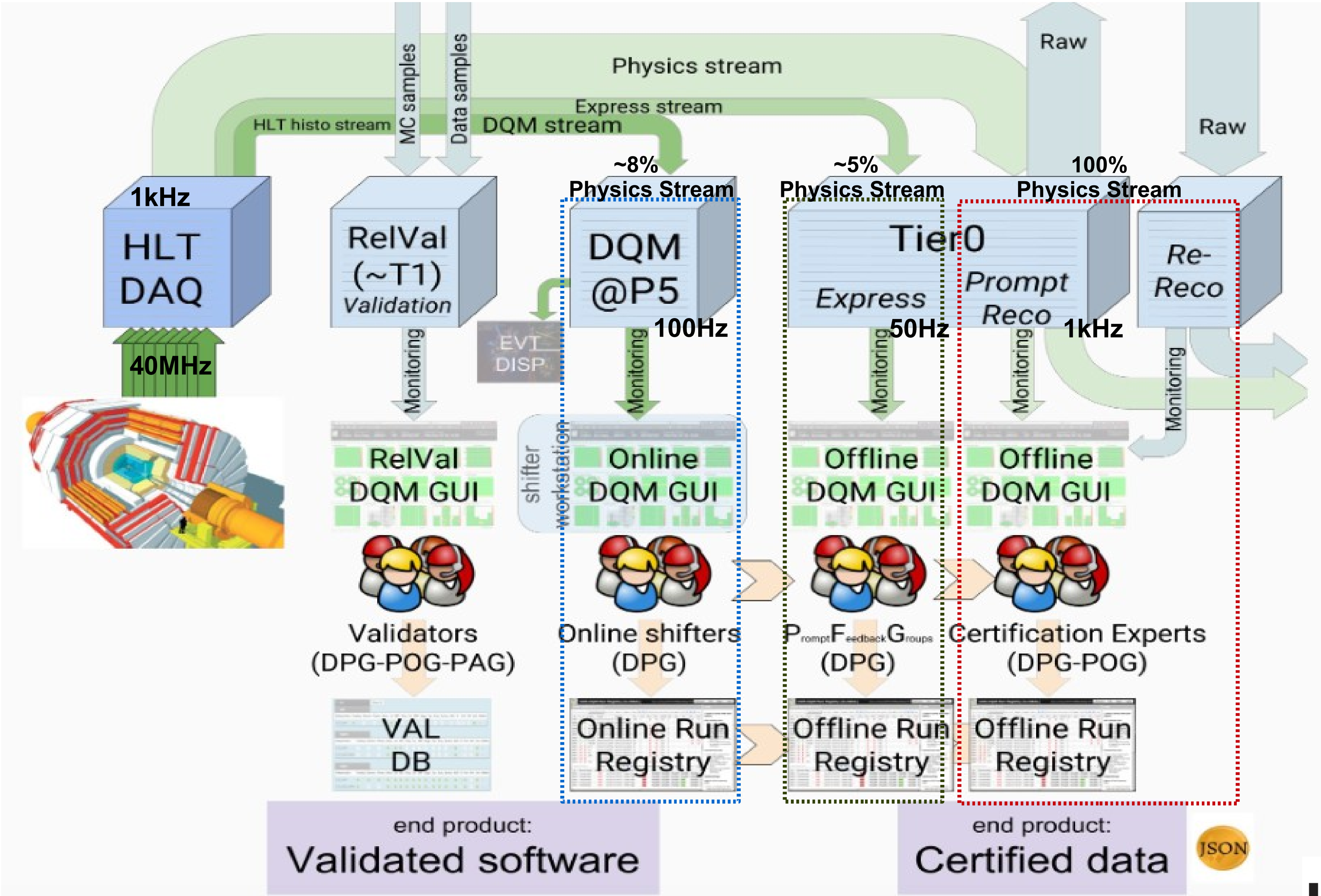
Maximize the best Quality Data for physics analysis

Data Quality Monitoring (DQM) and Data Certification:
- Monitors and ensures data quality of each data
- Measuring data properties
- Anomaly detection
- Certification



Virginia Azzolini

DQM system used in CMS



Data Quality Assessment

AUTOMATIC QA

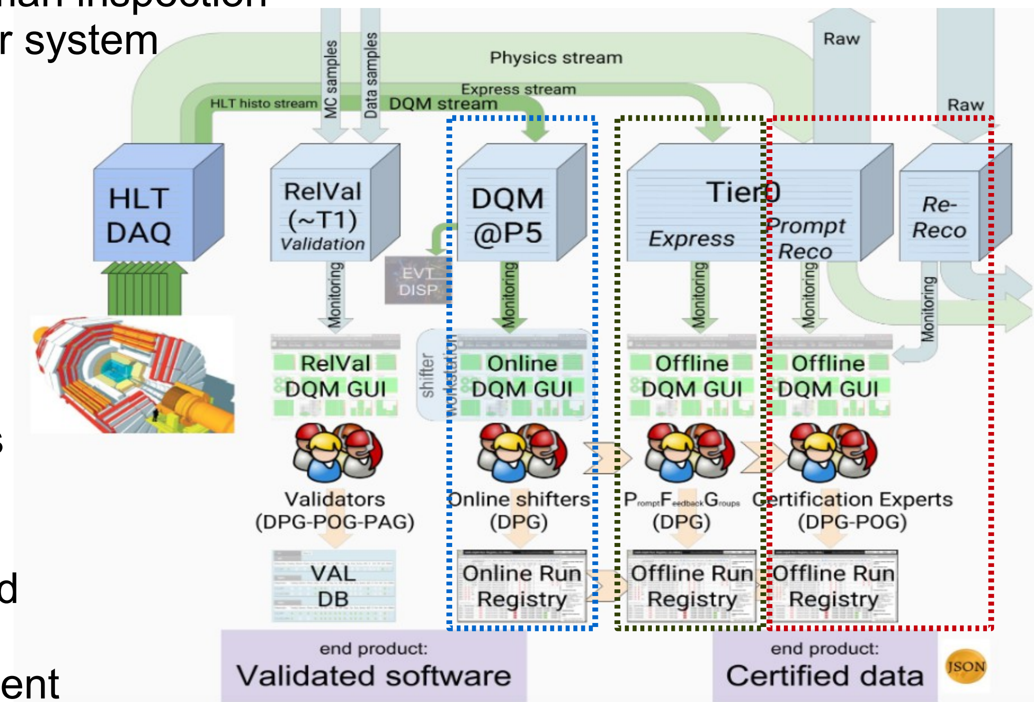
1) near-real-time applications

. fraction of the events with a rate of about 100 Hz
. automatic tests are validated via visual human inspection
. identify problems in the detector and trigger system

OPERATIONAL QA

2) fast reconstruction on a part of data

. subset of the data promptly reconstructed and monitored with ~1h
. goodness of the data regarding also the reconstruction software and the alignment and calibration constants



SCIENCE QA

3) full reconstructed data

. full set of data taken promptly reconstructed and monitored with ~48h latency
. same aim as 2), but typically better alignment and calibration constants are available

3-bis) reprocessed data once per year or at need

. data are again monitored and certified
. same aim as 2) and 3), but typically better reconstruction software and better alignment and calibration constants are available

On the side: release validation on Monte Carlo production,

. validate functionalities and performance of the reconstruction software

DQM GUI: Summary Workspace

One plot per sub system

What we monitor for Quality Assessment

Online DQM: mostly focused on Hardware level checks

. integrity of the data-format, errors from the read-out electronics

count errors, classify errors, monitor # of errors vs LS

. occupancy of signals (hits) in the various channels

maps and distributions in the detector

presence of noisy/dead read-out channels

. distribution of energy/momentum/time of the signals

. resolution plots, pulls

Offline Data Certification: principally focus on Physics

. detector subsystem:

..Certify the correctness of detector calibration and alignment application,

these conditions are recalculated una tantum, because statistics dependent

Almost same distributions as online

. physics objects (muon, electron, photons, tracks, jets)

.. Monitoring quantities product of the reconstruction, ingredient of future analysis

(# vertices, 3 tracks, energy, typology, topology of the particles, key quantities

Summary and occupancy maps

Distribution of quantities used to characterize the candidate particles

Limits of a Human-based QA

. Volume budget

Limited amount of quantities that a human can process in a finite time interval

. Time delay

Online: automatic test+ human intervention = ~ minutes → trigger stop/continue data taking

Offline: reconstruction data time +human intervention = ~ 1 week → Need intermediate step

. Expensive, in terms of human resources

Online: shifter 24/7 + the effort to train her, maintain instructions, etc

Offline: duplication of effort (many detector and physics object experts) on weekly basis

. Makes assumptions on potential failure scenarios

QA paradigm: scrutiny of a large,but chosen, # of histograms in comparison with a reference

Conservative strategy that could prevent unforeseen anomaly detection

. missed time dimension

Granularity of certification is lumi-section* to reduce data lost due to short pbl condition, but evaluation relies often on integrated quantities, punctual anomaly may not surface

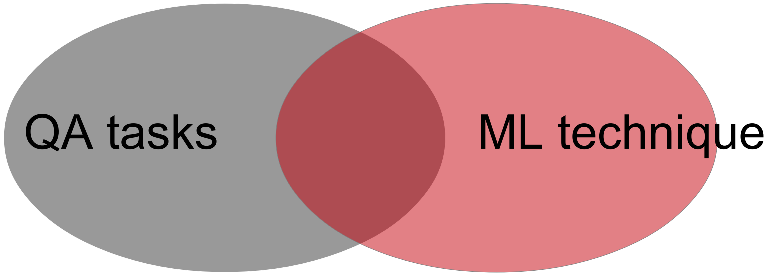
Good news is the current system works

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

*minimum time interval within a run

Technical characteristics required for QA



WE NEED:

- operate effectively with minimal human guidance
- anticipate important events
- adapt behavior in response to changes in data content, user needs, or available resources

WE HOPE:

Developed machine learning and analytics-based solutions will:

- improve the accuracy of data quality
- establish new data quality rules to sharpen data error detection and correction
- increase the speed at which this is achieved

Technical characteristics required for QA

REQUIREMENTS:
To do this the "intelligent data archive" will need to

- learn from its own experience
  - recognize data quality problems solely from its experience with past data, rather than having to be told explicitly the rules for recognizing such problems.
  - E.g. flag suspect data by recognizing departures from past norms
  - E.g. categorize data based on the type or severity of data quality problem.

- recognize hidden patterns in incoming data streams, could learn to recognize problems
  either from explicit examples or simply its own observation of different types of data

- data access requests, ability of an archive to respond automatically to data quality problems.
  E.g. significant increases, in the amount of data flagged as bad or missing,
  might indicate that the data are exceeding the bounds expected by science QA algorithms
  The intelligent archive could:
  - notify DQ experts so that the issue can be further examined and resolved.
  - archive could retrieve old data to confirm a data quality problem,
    obtain data from an alternate source, or
    request that the data be recollected or reprocessed in response to confirmed DQ pbl

(from automatic QA → to autonomous data QA)

Technical characteristics required for QA

REQUIREMENTS:

- Fast and efficient operation

  Some current DQ applications require data to be delivered < 1h hour
  so the need to perform the DQ of a relatively large amount of data within a few minutes.

  Machine Learning algos are compute intensive,but we could still meet this requirement
  the computational effort is associated to the deriving rules or training the system;
  the rules themselves are generally computational easy to apply in an operational mode.
  If QA is a function of applying the rules, rather than deriving the rules,
  then the computational complexity associated with deriving the rules is not an issue,
  because this can be done on a subset of the data in an offline process.

Which approach?

Numerous ML methods, techniques, and algorithms can be applied to data QA

Probably a tool-only approach would not suffice as a complete and manageable solution. We aim, certainly at the beginning, to a mixed situation where the intelligent automatic QA is integrated to human expertise

The QA problem is primarily one of classification
prevalence of good /bad “flag” types rather than numerical quality indicator

We focus on classifiers, though numerical predictors may be useful in intermediate steps in the QA process

Which approach?

Supervised classifiers:
Require a training set of data previously classified (human interpretation or direct observation)

Direct application: train the classifier on data with known quality signatures.
Pros: . incorporate information not available in the data to be classified, metadata
(e.g., determine a set of data “good” or “bad” based on derived data products or human eye)
. additional flexibility + opportunity to refine directly QA process vs time
(adding new examples to the training data set)
Cons: . accumulate sufficient training set presents a challenge

Indirect application: E.g.classify all of the data according to a set of positive categories
A behavior that cannot be easily classified into a category may then be signs of quality problems such as random events or mixed contributions.
Important : is it a bug or is it a feature?

Unsupervised classifiers:
Generate classes directly from the observed data, “clustering”.

Pros: . identify new classes that may not have been defined a priori.
. identify anomalous datasets without explicit training,
either by directly identifying separate clusters for “typical” and “unusual” data, or
by identifying normal clusters to used as references to identify outliers in a data stream.
. less human effort, no need to identify different classes of DQ problems and good training examples of each

CONCLUSIONS

. Goal: Maximize the best Quality Data for physics analysis

.. Introduction to CMS multi-steps human based system for data monitoring and certification

… Satisfied for the goodness of it, we are able to the limits of it:
Human effort, delay time, volume increasing, protective umbrella (too blind?)

…. Looking toward automatic QA:
Needs: minimal human guidance, predictivity of events, adjustable behavior

….. Requirements: be intelligent
experience docet, be fast but efficient, recognize the unknown, be proactive

….. which approach?
Bring together computing, data scientists and physicists is the winning approach

Outgoing projects:
IBM – CMS-DQM project for near-real-time anomaly detection
Yandex – CMS DC project for automatic certification for science QA

Thank you
Virginia

Machine Learning in Data Quality Monitoring

… a CMS point of view