

Machine Learning and Quality Control in ALICE

3-4 Dec 2018, CERN

<https://indico.cern.ch/event/766450>

56 registered participants!

Goals

- Bring together ML and HEP communities to discuss applications of ML techniques for data quality control (QC)
- Discuss data sets and tools for application of ML techniques in ALICE
- Discuss problems to be solved with ML techniques for online and offline QC in ALICE

Outline

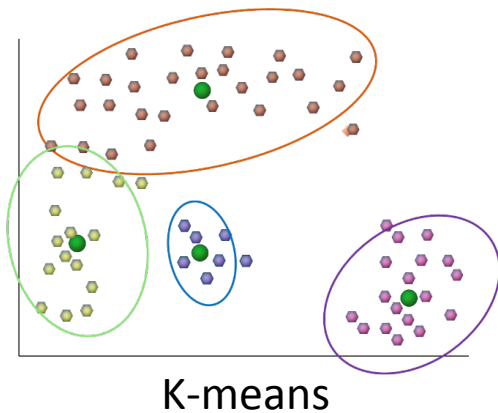
- ML techniques for anomalies detection
- CMS and ATLAS experience with ML for QC
- QC data sets for ML in ALICE
- Examples of ML usage for QC in ALICE
- ML tools in ALICE
- ML for ALICE QC in Run3

Anomalies detection with ML techniques

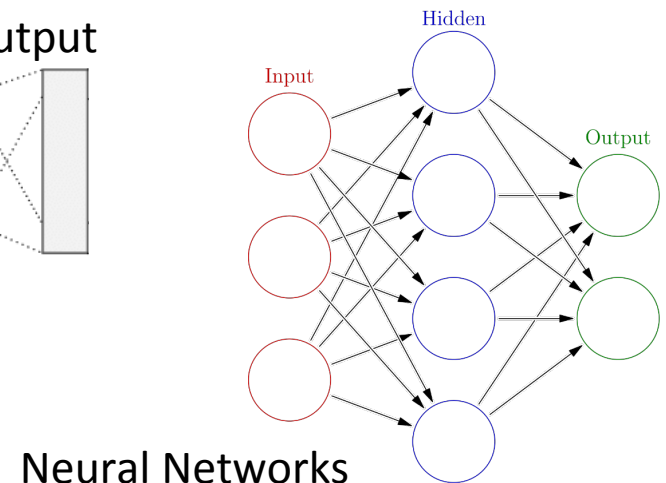
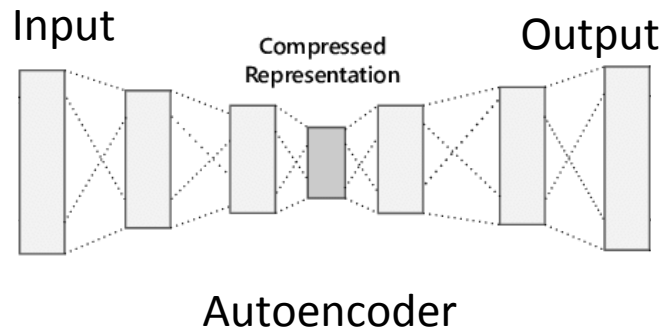
Kamil Deja

- Classification of anomalies (needed: labelled dataset)
- Regression of one value which may indicate anomalies (needed: dataset with known values)
- Clustering of unknown data and searching for outliers (needed: noisy data)
- Dimensionality reduction for sparse data representation and searching of outliers (needed: high dimensional data)

Clusterization



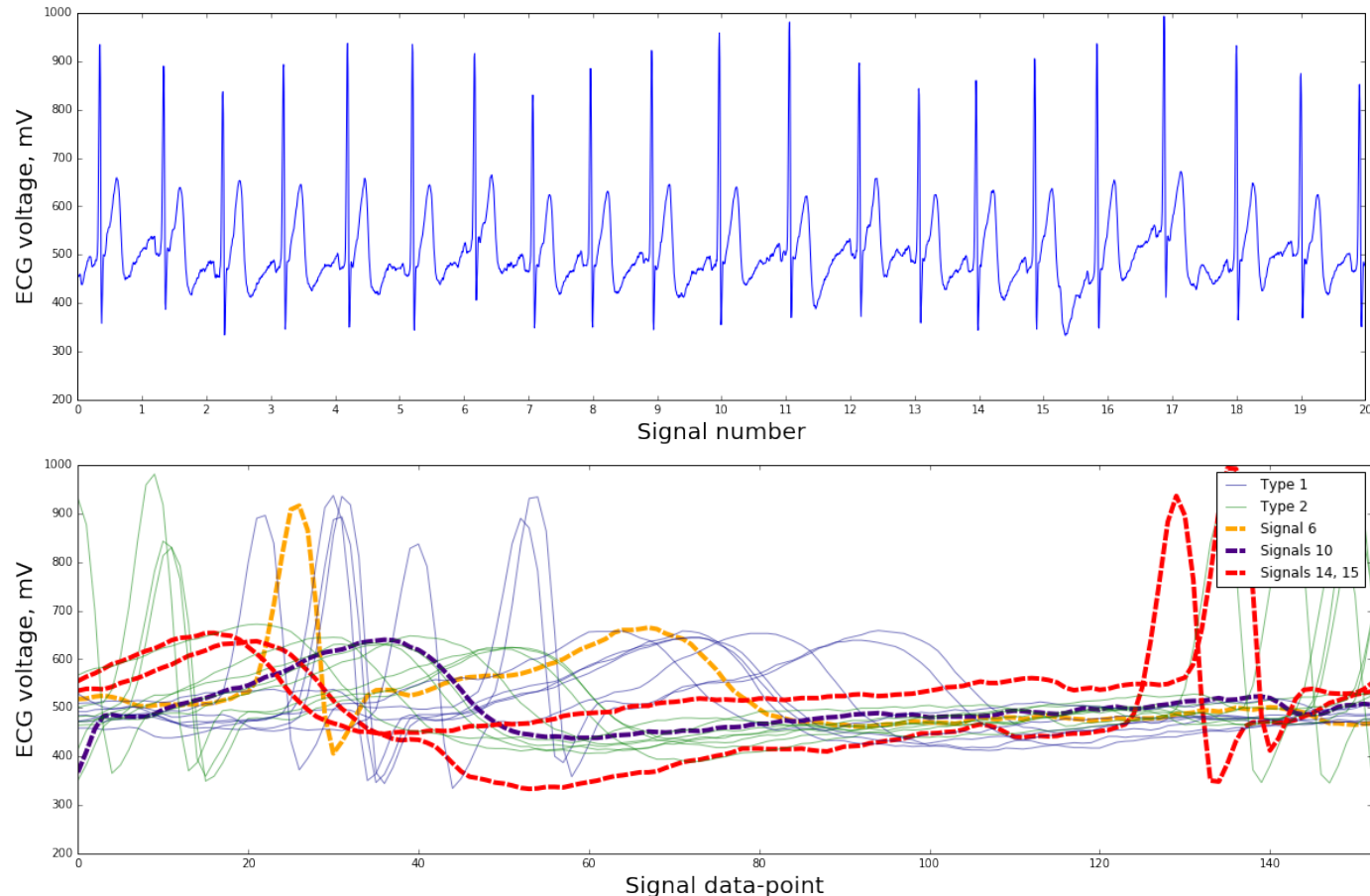
Classification and Regression



Online deep learning for pulsed-signal forecasting

Example: ECG forecasting

Creig Bower

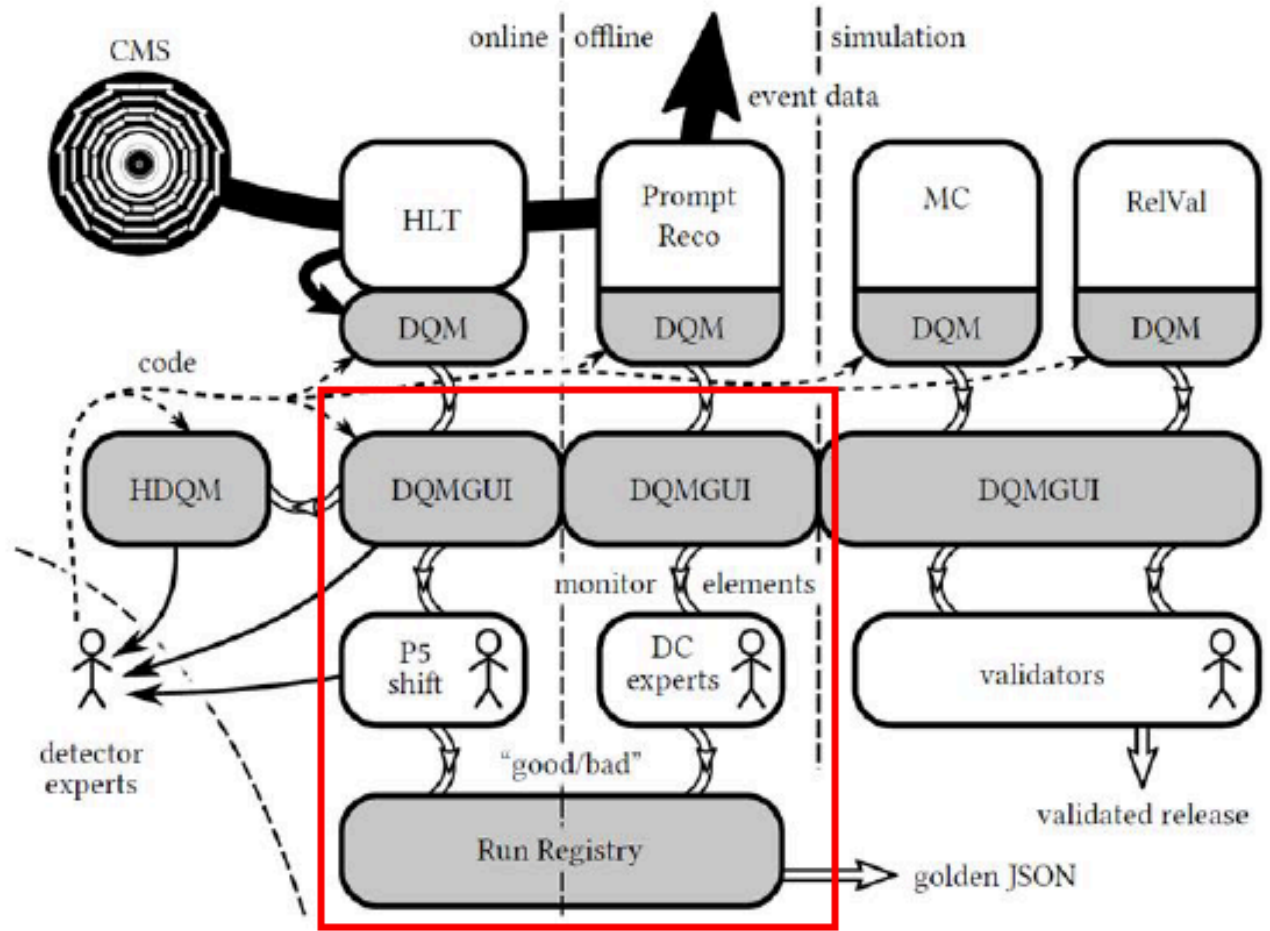


Combines evolutionary *survival of the fittest* strategy with parallel stochastic average gradient descent → can be consider for DCS data

CMS – Data Quality Monitoring (DQM)

Mantas Stankevicius

- ML techniques are used for online anomaly detection and offline data certification
- **Do not replace experts but minimize human errors**



CMS Drift Tubes – online anomalies detection

Mantas Stankevicius

Local:

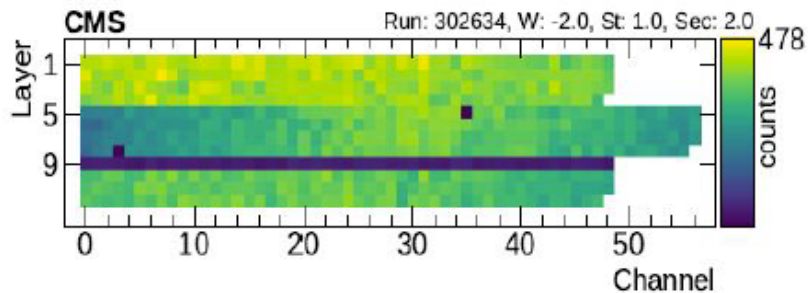
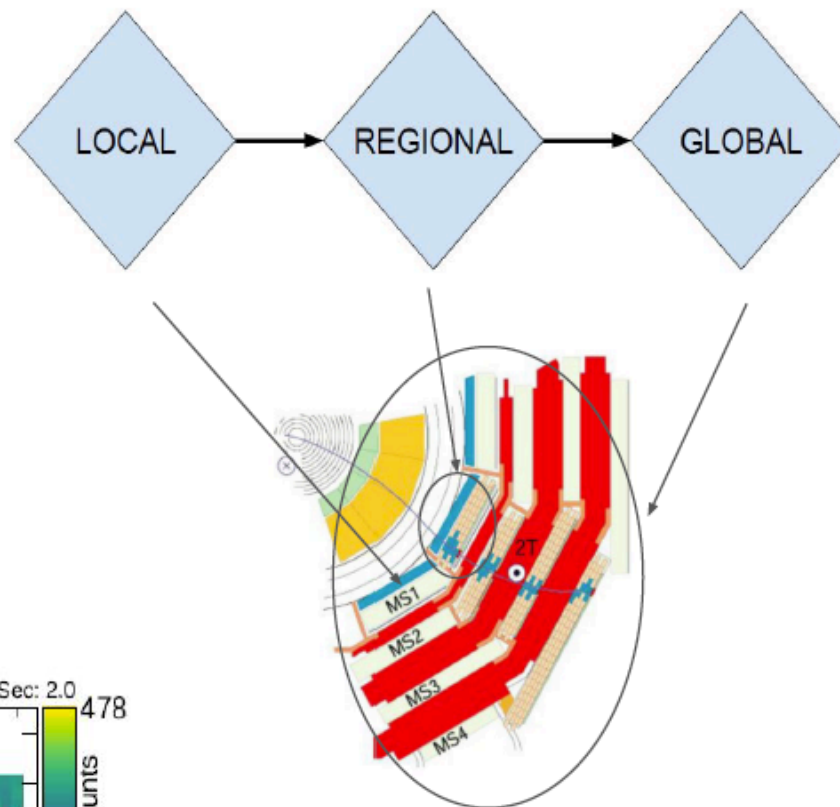
Each layer is treated independently

Regional:

Use information from all layers from individual chambers

Global:

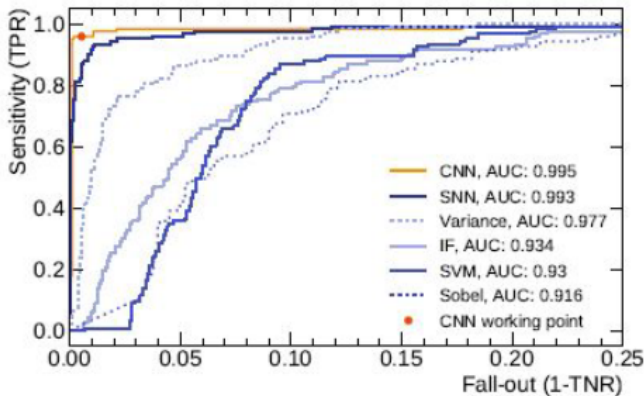
Information from all chambers for a given run



CMS Drift Tubes – online anomalies detection

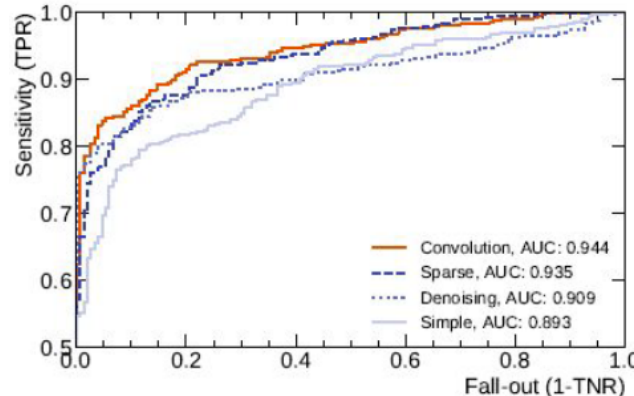
Mantas Stankevicius

Local



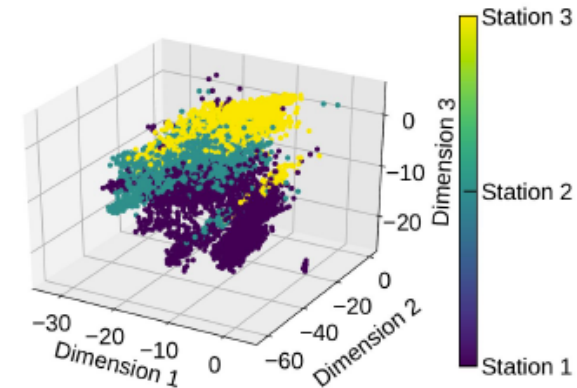
- Supervised Convolutional Neural Networks (CNN) outperforms other methods (ROC AUC: 0.995)
- Successfully applied in production

Regional



- Semi-supervised autoencoder variation

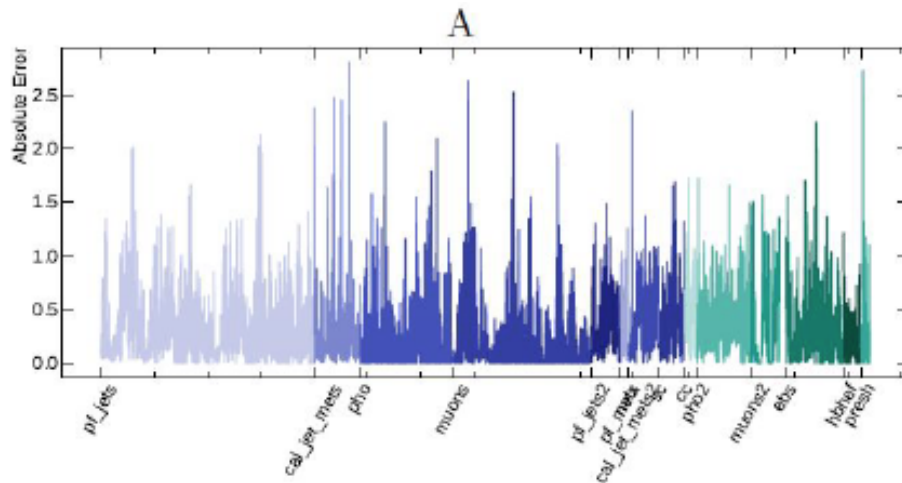
Global



- The position impact occupancy patterns
- Autoencoders learn a compressed representation of chamber data

CMS – offline anomalies detection with Autoencoders

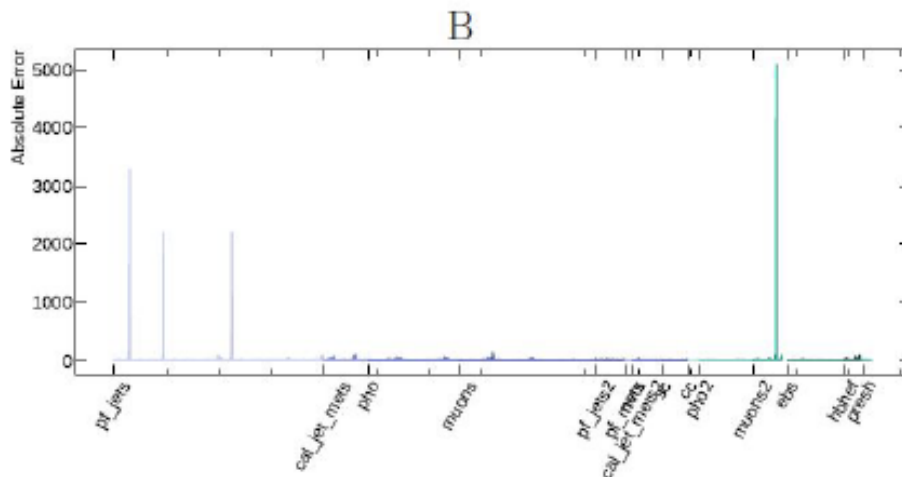
Mantas Stankevicius



Features are grouped by physics objects:
A: Reconstruction errors is small

B: Reconstruction error is high

- Anomalies seen for muons and jets



ROC AUC = 0.978

ATLAS – Data Quality Monitoring (DQM)

Peter Onyisi

Run 363033, 1/physics_Main

Select Run= 363033 Stream= physics_Main Pass= 1 LB Change= Entire Run Click to Go Superimpose
 or jump to run Jump to Run Display side-by-side with Run(-processing)= 362354 Show side-by-side

- Overall Status: **Red**
- CaloMonitoring: **Yellow**
 - CaloMonExpert: Undefined
 - CaloMonShift: **Yellow**
 - CaloMonBAR: **Yellow**
 - CaloTopoClustersBAR: **Yellow**
 - EMTopoClustersBAR: **Yellow**
 - CaloMonECA: **Yellow**
 - CaloMonECC: **Yellow**
- CentralTrigger: **Red**
- Global: **Red**
- HLT: **Red**
- InnerDetector: **Red**
- JetTagging: **Green**
- Jets: Undefined
- L1Calo: **Green**
- L1Interfaces: **Green**
- LAr: **Red**
- MissingEt: **Red**
- MuonDetectors: **Red**
- MuonTracking: **Yellow**
- Tau: **Red**
- TileCal: **Red**
- egamma: **Red**

CaloMonitoring/CaloMonShift/CaloMonBAR/EMTopoClustersBAR/m_clus_etaphi_Et_thresh1@BAR

Location in ROOT file: CaloMonitoring/ClusterMon/LArClusterEMNoTrigSel/2d_Rates/m_clus_etaphi_Et_thresh1

Occupancy of clusters with Et_clus > 4.0 GeV

Assessment Details:

Name: m_clus_etaphi_Et_thresh1@BAR
Algorithm: BinsDiffFromStripMedian
Num. of Entries: 220969535.0

Configuration Parameters:

MaxPublish: 200.0
SuppressFactor: 0.0
SuppressRedFactor: 0.0
xmax: 1.5
xmin: -1.5

MaxDeviation
 XXXXXXI XXXXXXI XXXXXXI
 25.0 200.0

Results:

NRedBins: 0.0
NYellowBins: 85.0
Y0-(eta,phi)
[OSRatio] 98300.0
(-0.250,1.526)
[4.08e+01]:
Y1-(eta,phi)
[OSRatio] 97430.0
(-0.250,1.427)

ATLAS – DQ Defect Entry System



Peter Onyisi

ATLAS DQ Defect Entry System

You are logged in as *ponyisi*. [Log out of CERN applications](#) 

Database:  Tag: 

Show defects in a run

Filter: 
Show defects marked absent: 
Show defects in run

*Hover mouse pointer over LB ranges to see comments, over defect names to see descriptions.
Bold defects are considered intolerable by someone.*

Defect	Present in LBs
EGAMMA_ETAPHI_SPIKES	1-719
GLOBAL_BUSY	1-7, 12-16, 49, 456, 501, 547, 608, 613, 658
GLOBAL_NOTREADY	1-64, 718-719
ID_BS_RUNAVERAGE	1-64, 718-719
ID_IBL_TRACKCOVERAGE	1-719
JET_LOWOCUPANCY	1-719
LAR_DATACORRUPT	117, 118-132, 133, 199, 717
LAR_EMBA_DATACORRUPT	118-132
LAR_EMECA_NOISEBURST	69, 75, 78, 118, 123, 142, 150, 154, 161, 165-166, 169-171, 176, 206, 224, 227, 250, 263, 265, 280, 283, 298, 300, 328, 352, 357, 366, 371, 375, 376, 380-381, 383, 390, 394, 400-401, 404, 426, 440-442, 445, 460-461, 468, 469, 481, 516, 534, 549, 592, 599, 611, 634, 639, 643, 646, 651, 661, 670, 712-713
LAR_EMECC_NOISEBURST	125, 140, 165, 167, 173, 180, 190, 203, 212, 215, 218, 221, 228, 235, 257, 261, 277, 296-297, 299, 324, 344, 352, 379, 403, 405, 408, 416, 431, 441, 445, 468, 488, 518, 551, 599, 607, 633-634, 659, 710

ATLAS – ML for DQM

Peter Onyisi

- No active deployment of ML for DQM during Run2. Developments under consideration for Run3.
- Investigated so far
 - Prediction of L1 trigger rates from luminosity, learning from time series in a given run
 - Anomaly detection: flag luminosity blocks which look “different” from others using Autoencoders, Boosted Decision Trees, ...
- Conceptual ideas
 - Automated predictions of reference histograms for a given e.g. luminosity, run length
 - Discover correlations of detector „defects” and characteristic of predicted histograms
- General
 - very easy to have false positives (keep discovering that luminosity / prescales changes during Run)
 - **need value-added over human checks**

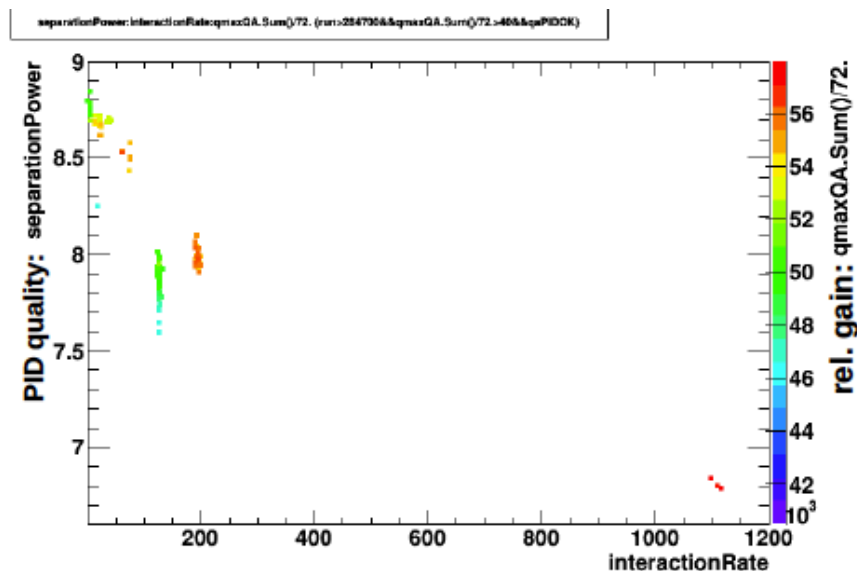
ALICE data sets for ML

Jens Wiechula

- Online/offline detector QA (root files)
- Calibration (OCDB – root files, partially contains DCS info)
- Logbook (SQL)
- MonALISA
- DCS (ORACLE – DARMA interface) – no automatized access by user

Online/offline detector QA to be extracted in smaller time intervals (~5 min.) – better for ML

Simple interface to access this data



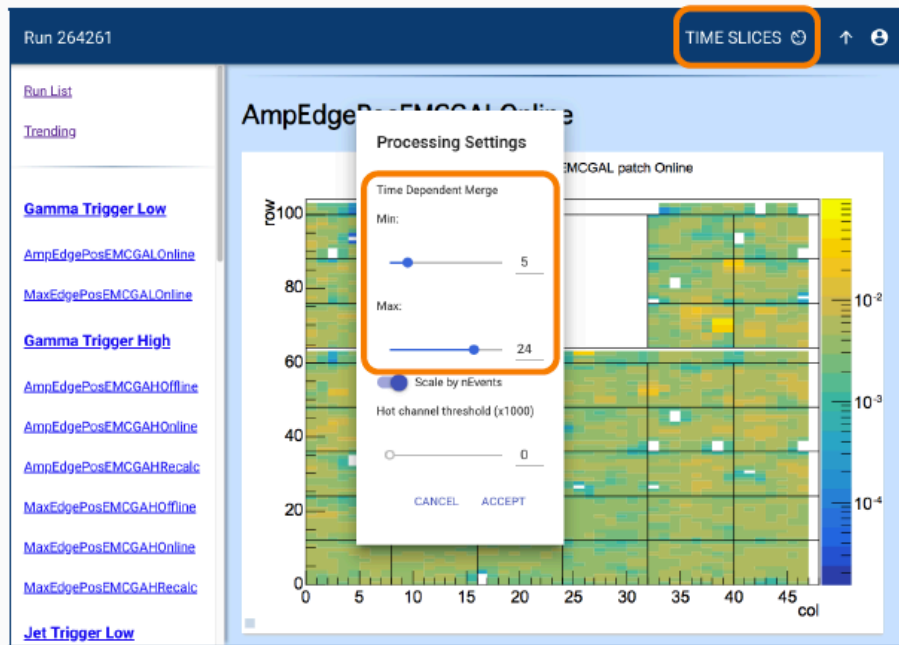
Materialized view - usage example:

```
AliExternalInfo info;  
treeCalib = info.GetChain("QA.rawTPC", "LHC18*",  
                          "cpass1_pass1", "QA.TPC;QA.EVS;Logbook");  
treeCalib->SetAlias("separationPower", "2* (meanMIPeLe-meanMIP) /  
              (resolutionMIP*meanMIP+resolutionMIPeLe*meanMIPeLe)");  
treeCalib->SetAlias("qaPIDOK", "resolutionMIP>0&&resolutionMIPeLe>0");  
treeCalib->Draw("separationPower:interactionRate:qmaxQA.Sum()/72.",  
              "run>284700&&qmaxQA.Sum()/72.>30&&qPIDOK", "colz")
```

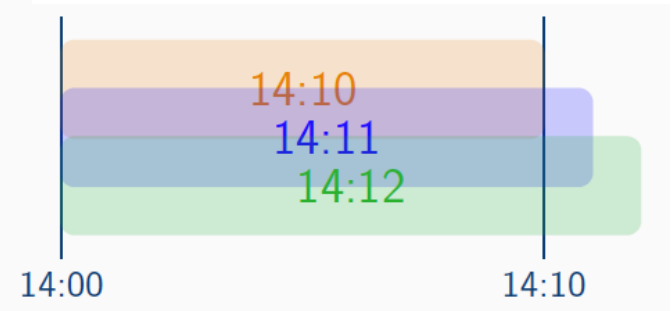
ALICE data sets from HLT/Overwatch for ML

- Available at <https://aliceoverwatch.physics.yale.edu>.

Raymond Ehlers



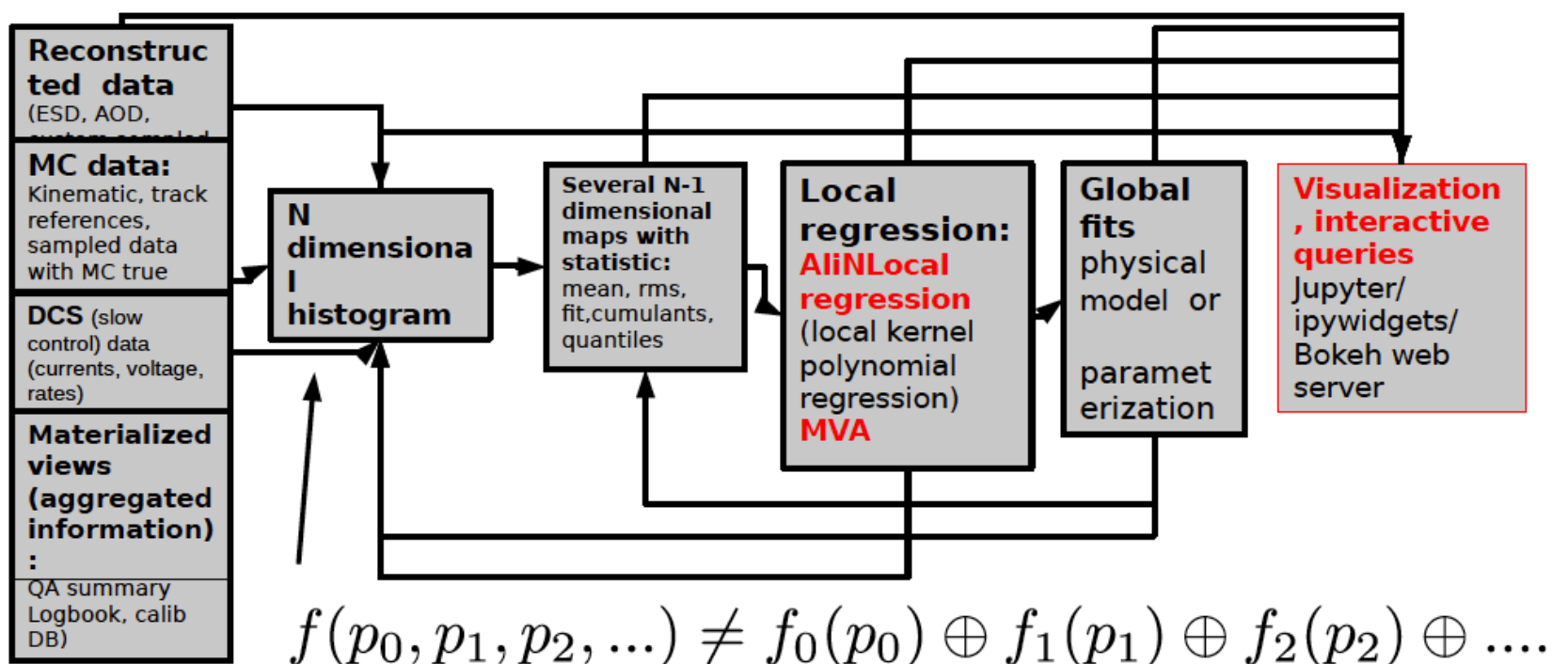
- Data stored in short time intervals (~ 5 min)
- Cumulative statistics



- Accessible from EOS
- Trending information is also extracted
- Very good data sample for ML

Pipeline data analysis

Marian Ivanov



- Differential QA
- Study detector performance – parametrization maps, physical models, ...
- Feasibility studies
- MC vs data comparison and MC tuning on data
- **Enable ML techniques (MVA)**

Interface to MVA methods in ALICE

Marian Ivanov

AliNDFunctionInterface : TMVA wrapper in ALICE analysis framework AliRoot (C++ implementation usable also in Python)

- Simple and compact user interface
 - similar to TTree::Draw and Histogram::Fit queries
- Store all the data as ROOT objects in ROOT files (instead of weight files, no xml files)
 - possibility to store data in Alice calibration DB
- Easy usage providing TFormula/TTreeformula interface
 - possibility to combine/normalize/operate with other formulas (other TMVA, global fits, NDimensional local tables (e.g AliNDLocalRegression object))

New wrapper (written in Python - to be interfaced also to C++)

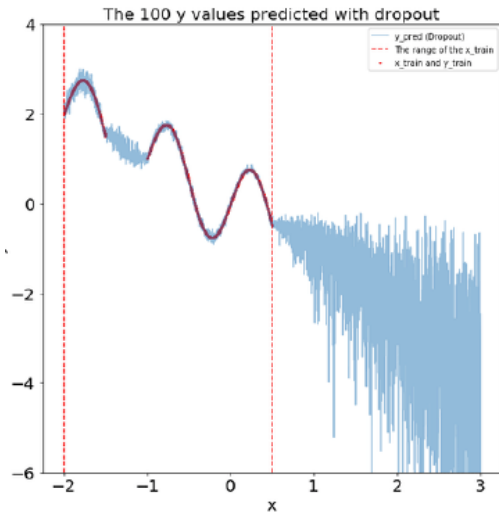
- **Local error estimates** (reducible and irreducible errors) and local robust estimators
- Combined/weighted evaluation, caching and model compression (**WORK IN progress**)

Goal - make the usage of the MVA almost as easy as standard fits in root

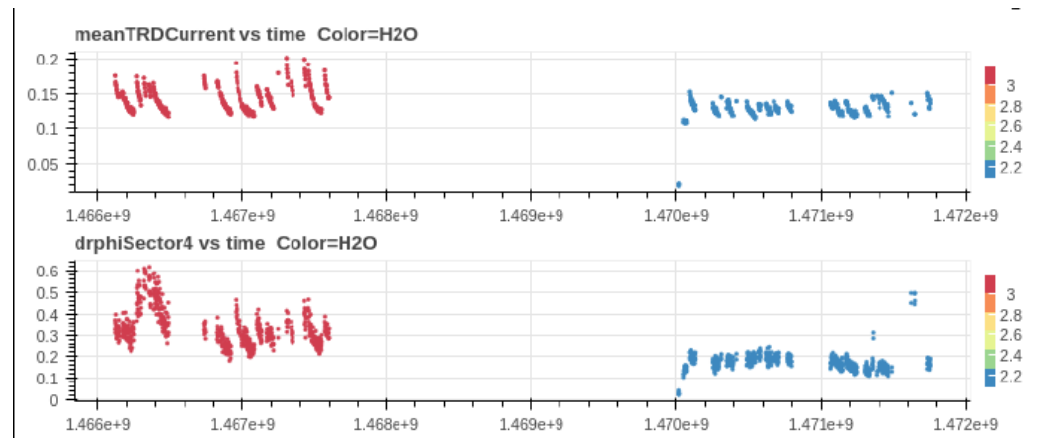
ML and error estimate with ML methods

Marian Ivanov

<https://fairyonice.github.io/Measure-the-uncertainty-in-deep-learning-models-using-dropout.htm>



Currently no standard methods in ML to estimate errors in the regions with sparse data



Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (<https://arxiv.org/abs/1506.02142> - 2015)

- *test-time dropout can be seen as Bayesian approximation to a Gaussian process related to the original network*

Bootstrap approach

- provides “prediction” intervals for all methods

TPC QC data classes example with ML

Kamil Deja

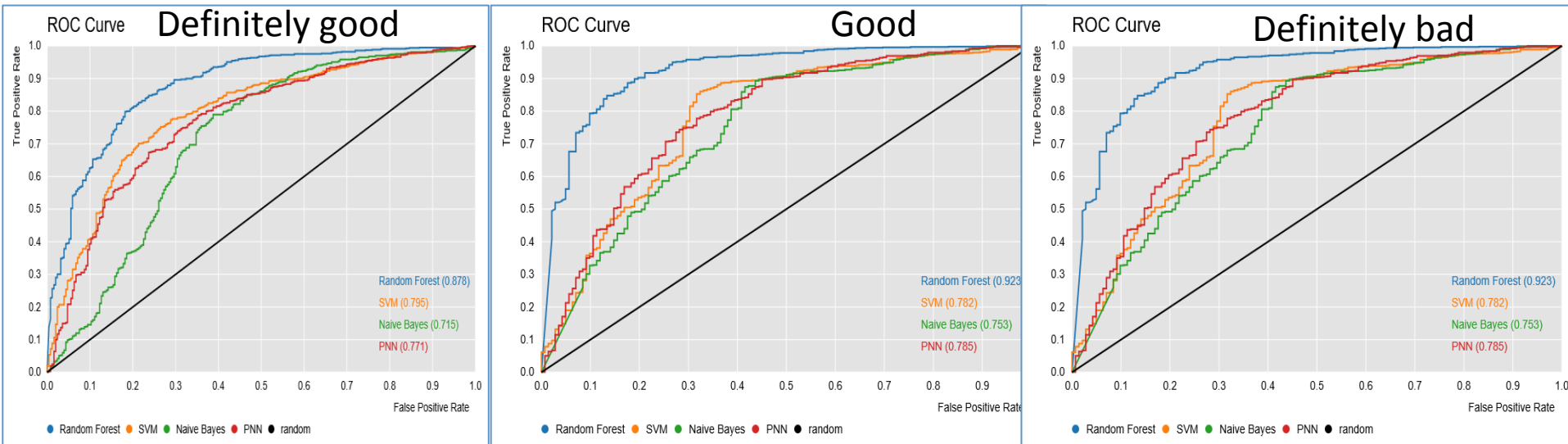
Example (TPC Jan 2016 – Dec 2017) ~ 1000 runs

Class\Approach	Definitely good	Good	Definitely bad
Detector was ON and running according to nominal specifications	ok	ok	To check
Not set	To check	To check	To check
Good data but some not full TPC acceptance	To check	ok	To check
detector was ON; but output can not be trusted / is known to be not usable	To check	To check	bad

- 217 numerical physical parameters mapped with **PCA to 26 dimensions** with 100% information preserved
- New dimensions showed **no significant correlation** with run number

TPC QC data classification example with ML

Kamil Deja

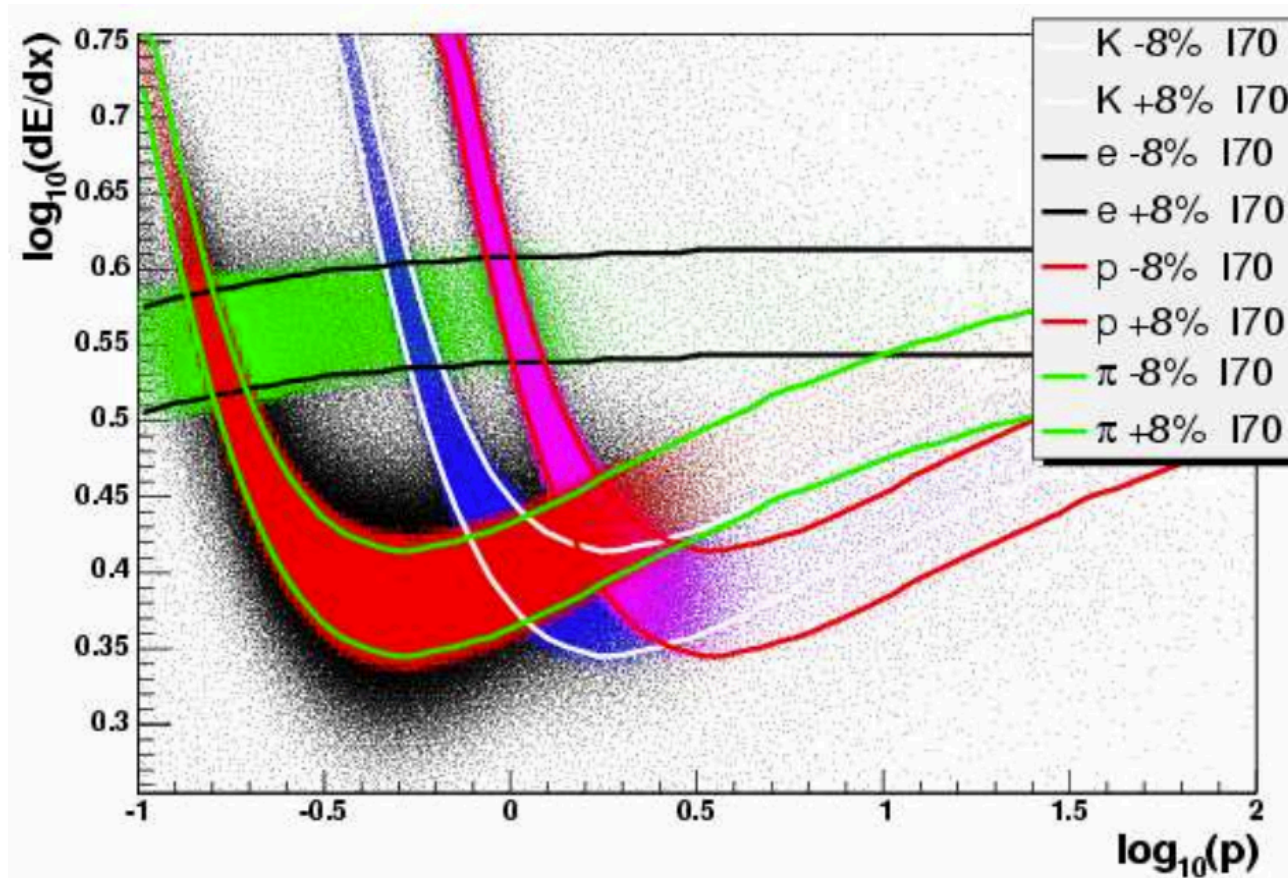


- Best performance of Random Forest
- Assignment of the quality label in **75% of the cases** with over **95% accuracy**

Example: PID with ML in ALICE

Focus on Kaons

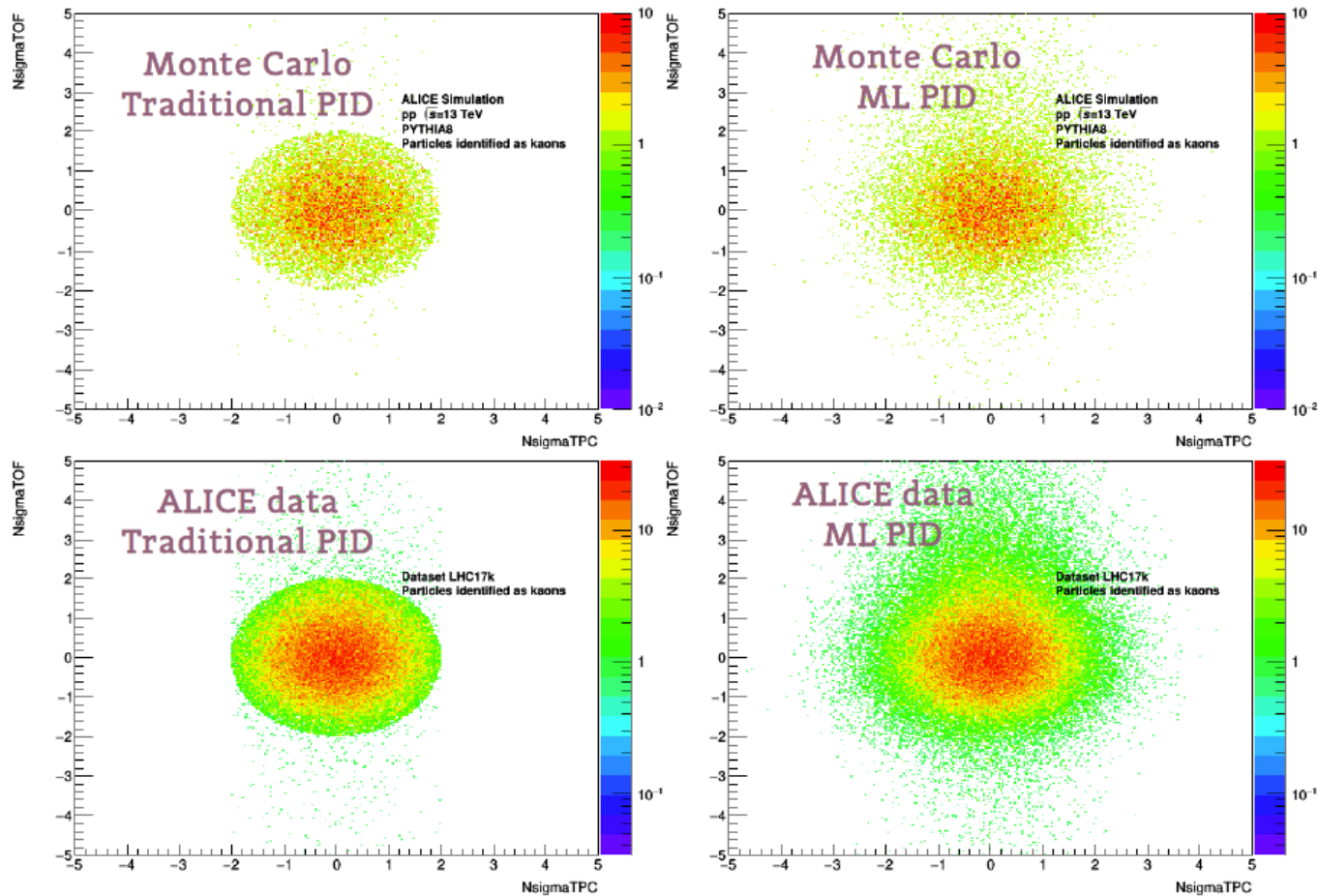
Łukasz Graczykowski et al.



PID with ML example in ALICE

Kaon in TPC and TOF

Łukasz Graczykowski et al.

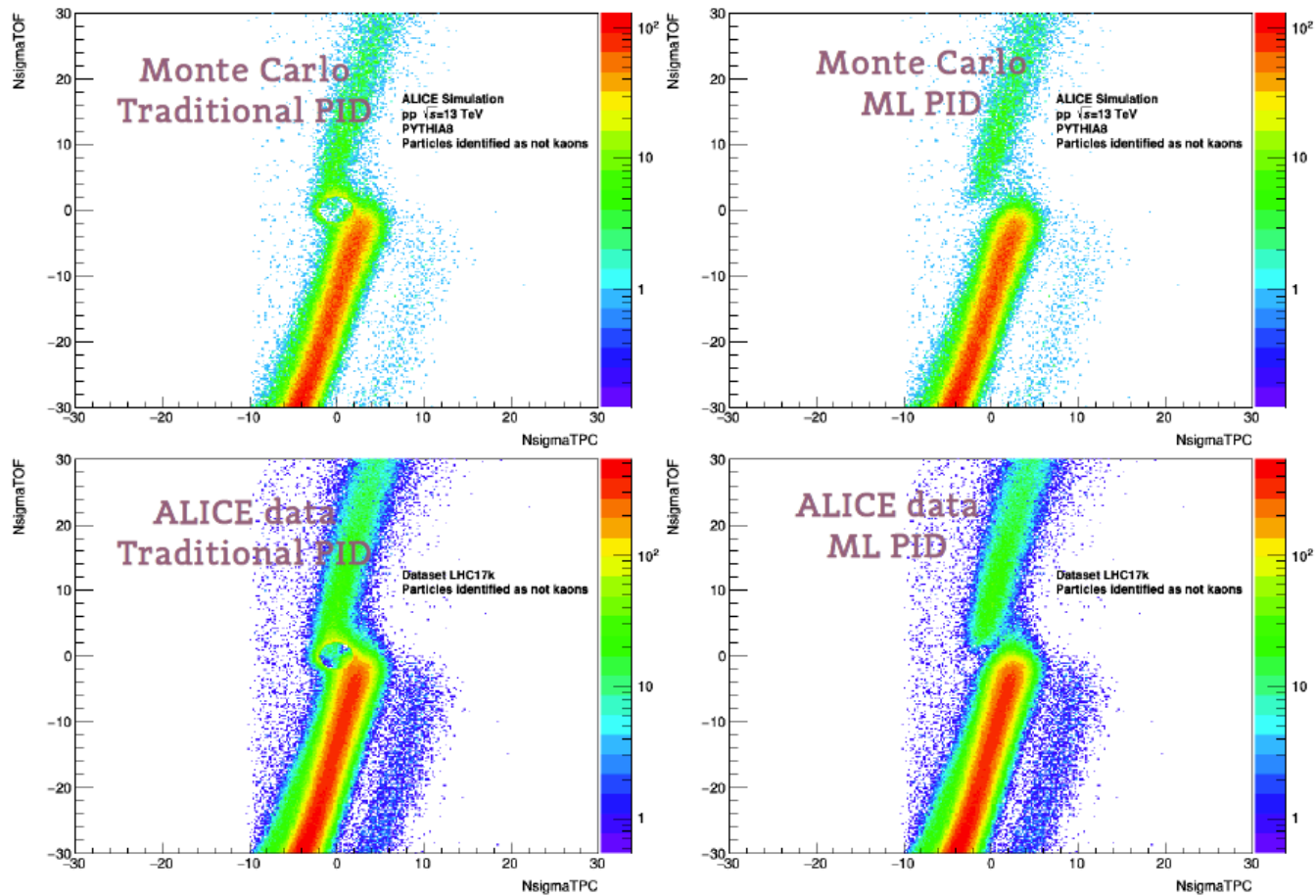


More efficient kaon identification with by ML methods (Random Forest)

PID with ML example in ALICE

Background (not Kaons) in TPC and TOF

Łukasz Graczykowski et al.

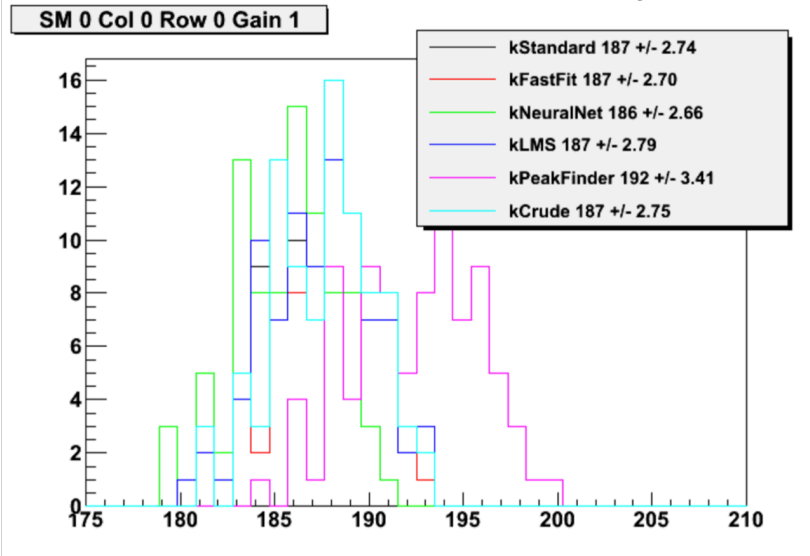
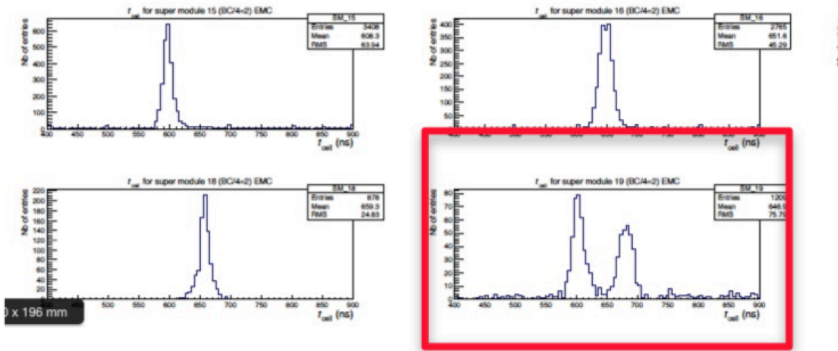


Much better background (not Kaon) rejection by ML methods (Random Forest)

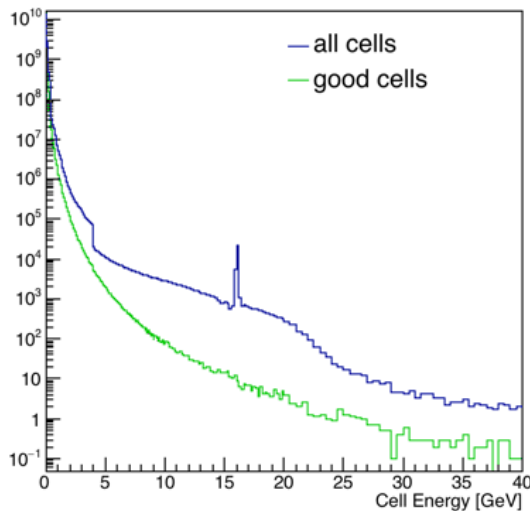
EMCAL QC with ML examples

Markus Fasel

PAR/L1 phase problem



Bad channel calibration



Neural-Network based signal extraction implemented for EMCAL

- Training data: LED run
 - Fast
 - Good time resolution (similar to standard method)
 - Amplitude underpredicted
- Not yet production ready

Common framework for ML analysis in ALICE

Gian Michele Innocenti

Proving flexible tool to perform Machine Learning analysis. It includes:

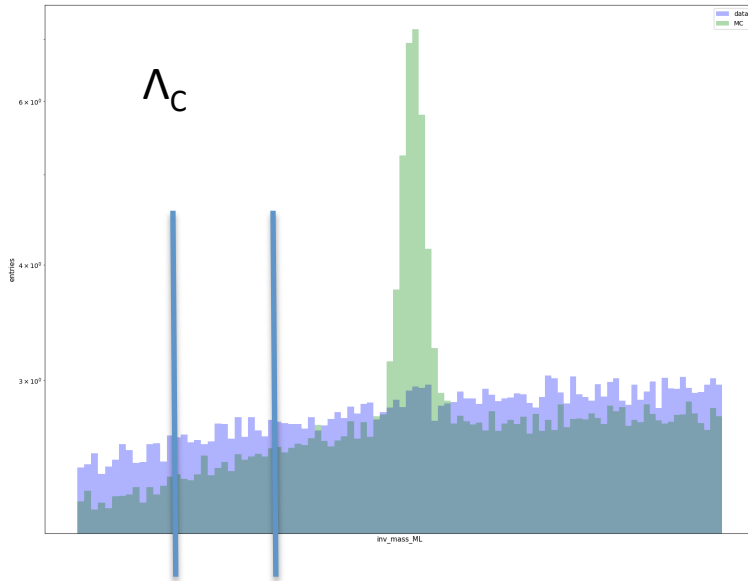
- Common Ntuplizer for TTree creation that can run on the Grid using LEGO trains (effort led by Andrea Festanti, important help from Markus/Jan for the LEGO part)
- ROOT to Pandas DataFrame conversion:
 - convert the root TTree of MC and Data into Pandas data frames
 - create training samples mixing MC and data
 - create testing and training samples
- Training/Testing and common validation routines with Scikit/TensorFlow:
 - implement most common ML algorithms and Deep Neural network for classification using SciKit and TensorFlow
 - Automatic validation with cross score validation, confusion matrix, learning curves, ROC, etc.
- Testing on large samples for analysis and new TTree creation:
 - new decision flag is added to the data frame
 - a new TTree is created including flags and probabilities of all the ML algorithm
 - Possibility of exporting the model in C++ for running testing on the Grid

It can be used for any type of ML analysis including QC

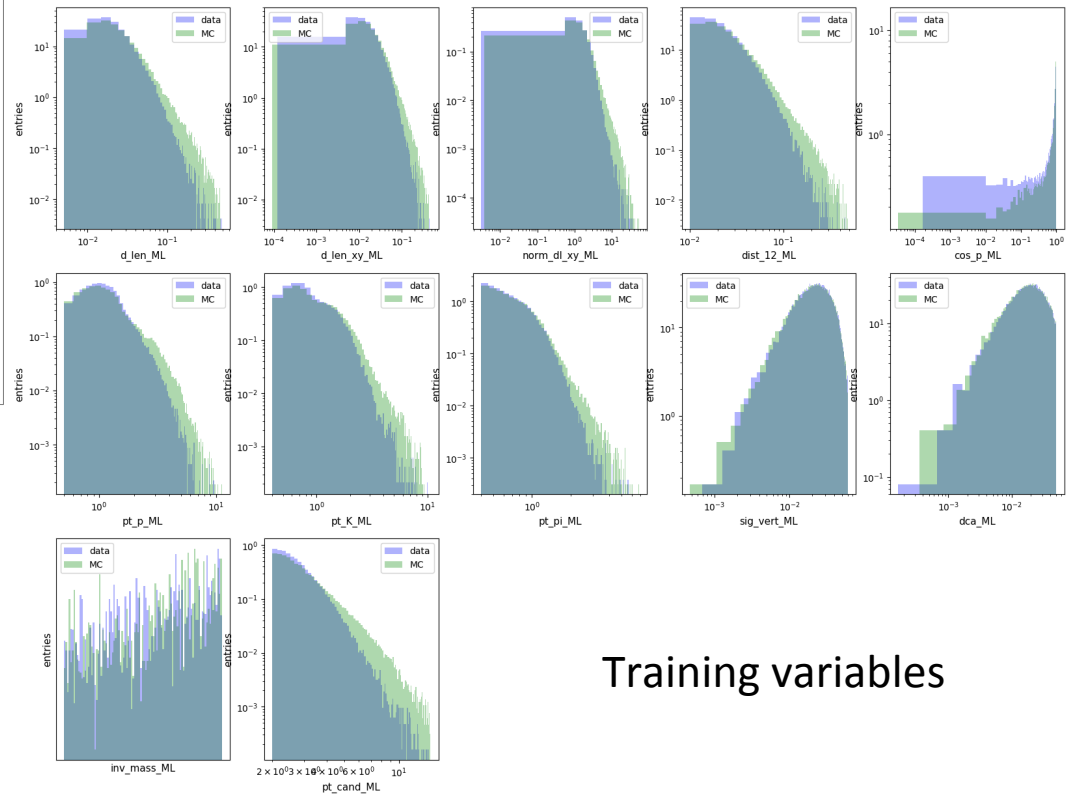
MC tuning on data with ML in ALICE

ML algorithm learns to discriminate between data and MC

Gian Michele Innocenti



Cut observables **before** MC tuning



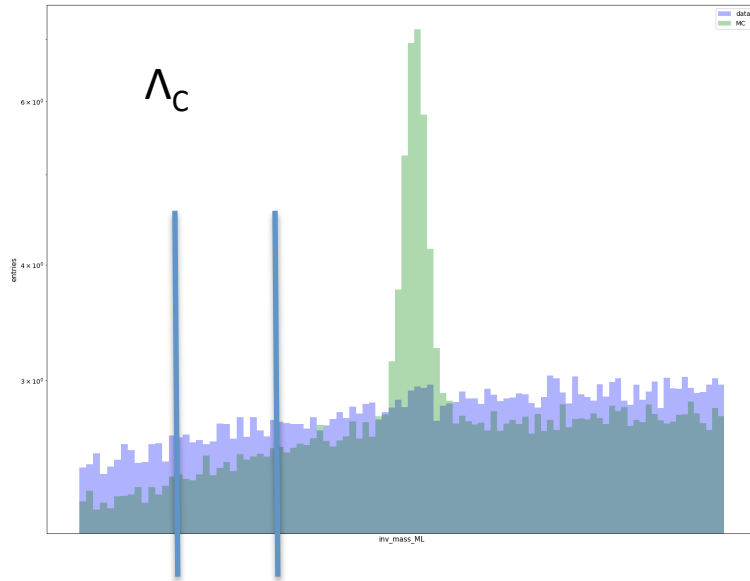
- select background events from side bands for MC and data
- tag the **data as signal** and **MC as background**

Training variables

MC tuning on data with ML in ALICE

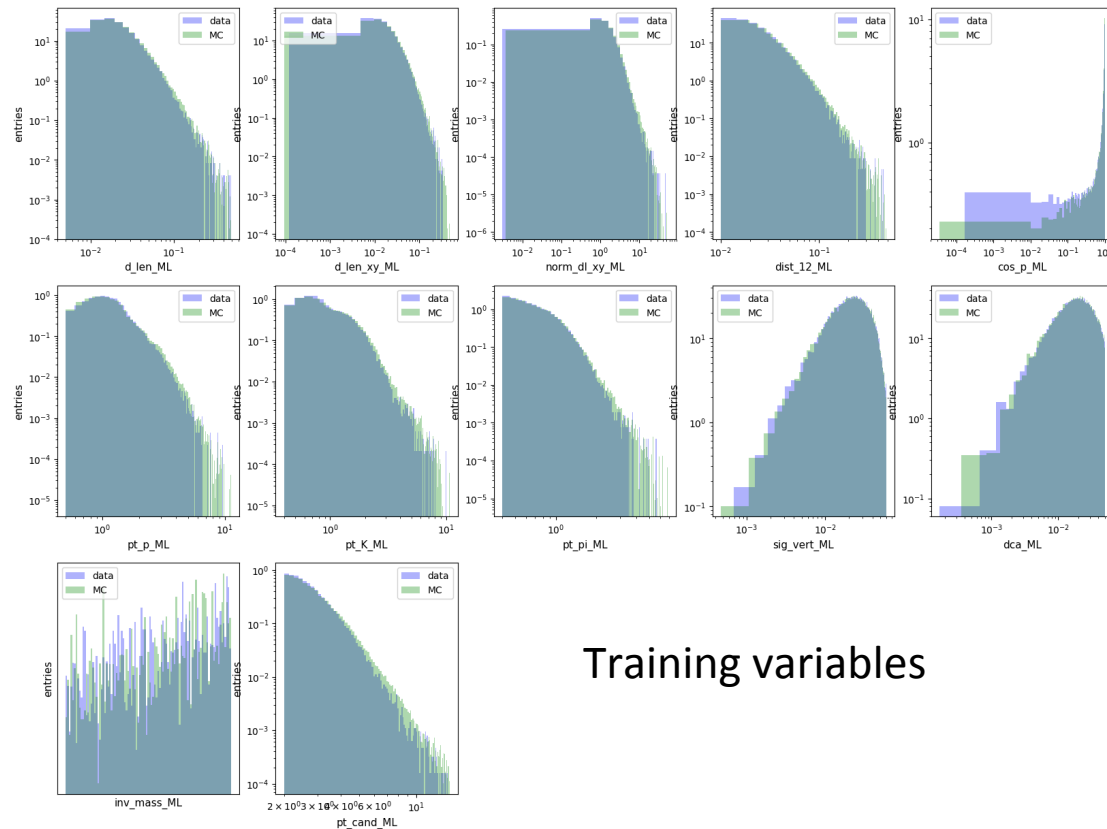
ML algorithm learns to discriminate between data and MC

Gian Michele Innocenti



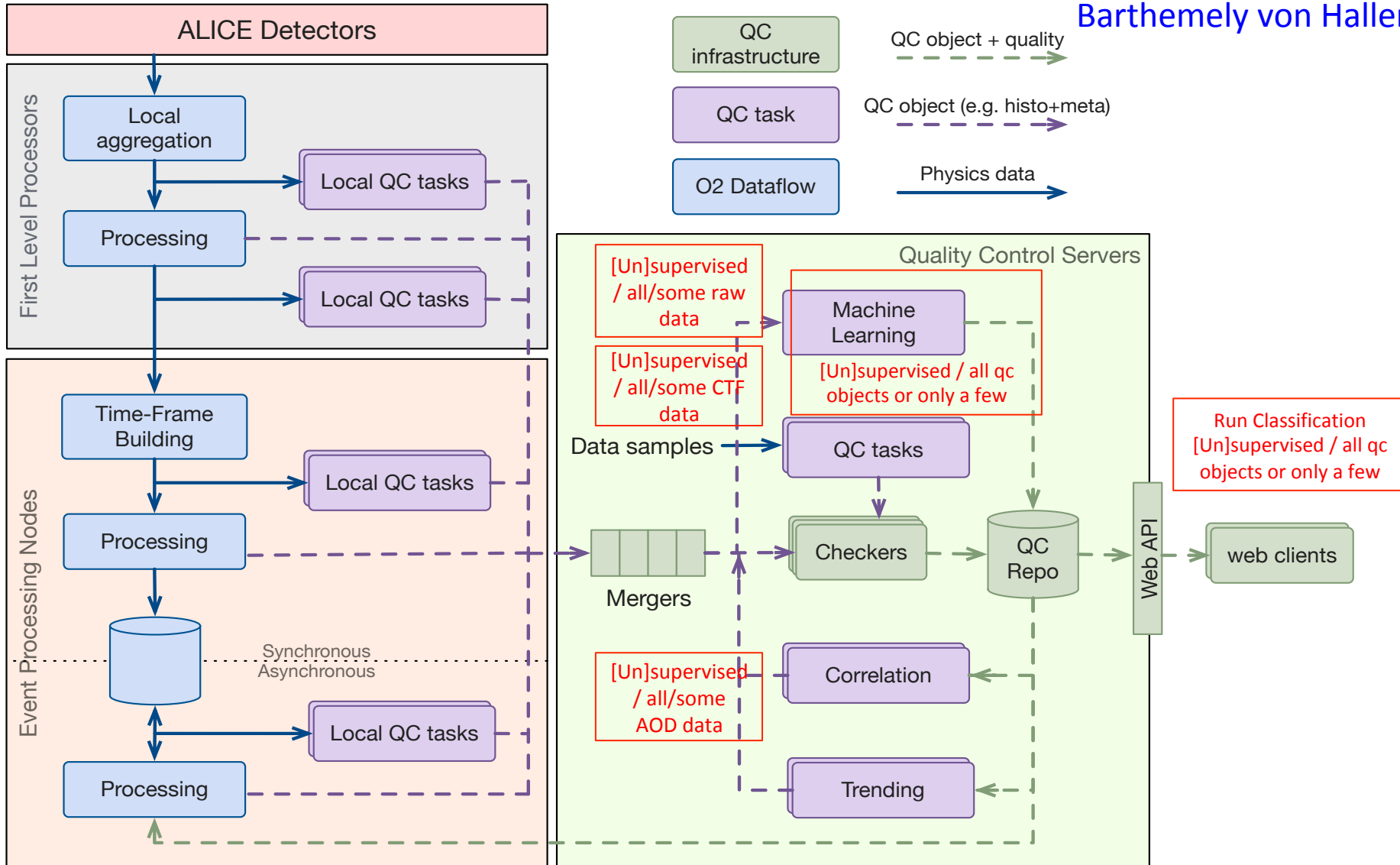
- select background events from side bands for MC and data
- tag the data as signal and MC as background

Cut observables **after** MC tuning



ML usage foreseen in Run3

Barthemely von Haller



Outlook

- ML techniques started to be used in HEP experiments for Data Quality Monitoring and data certification
- ALICE QC data to be prepared for ML applications
 - Smaller time intervals for offline data
 - Trending parameter extraction for online data
- ML techniques successfully applied for offline TPC QC data classification (but only 1000 chunks/runs)
- EMCAL QC with ML started
- Good performance of ML for PID identification (example for kaons)
- MC tuning on data with ML tested. Alternative solution to reweighting based on parameterization maps
- ML tools development ongoing
 - Interface to MVA from pipeline analysis
 - ML framework for all purpose analysis in ALICE
 - Possibility to work out one solution
- ML application for online/offline QC foreseen in Run3