ML in the online data acquisition

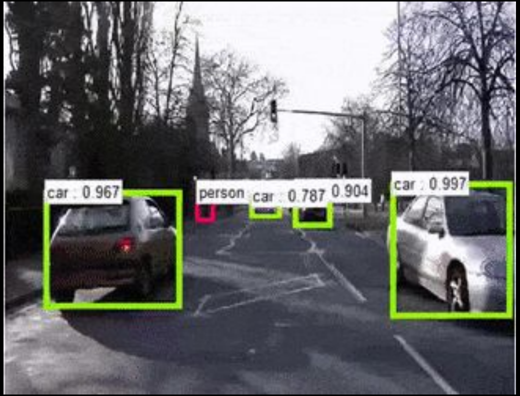
Cristiano Fanelli

EIC^2
EIC Center at Jefferson Lab

2019

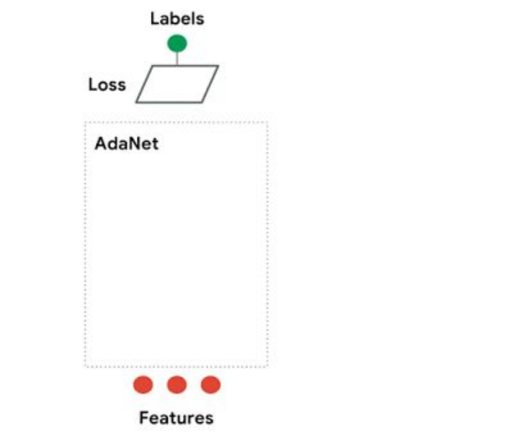
we stand at the height of some of the greatest accomplishments that happened in DL

Autopilot [2]

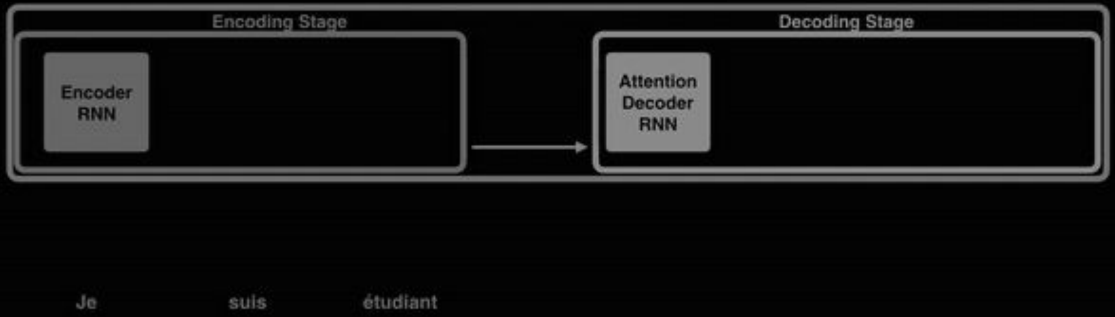


Ref [1] [2] [3] [4]

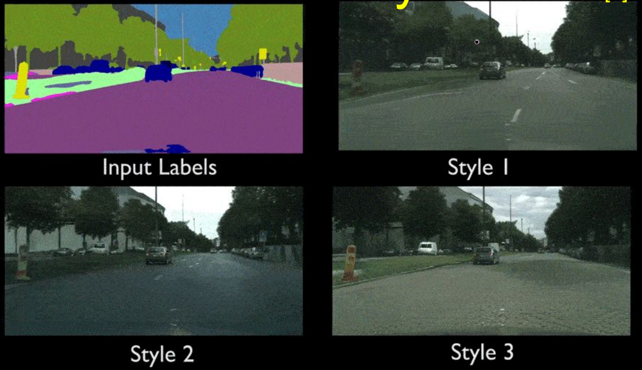
Meta-learning [3]



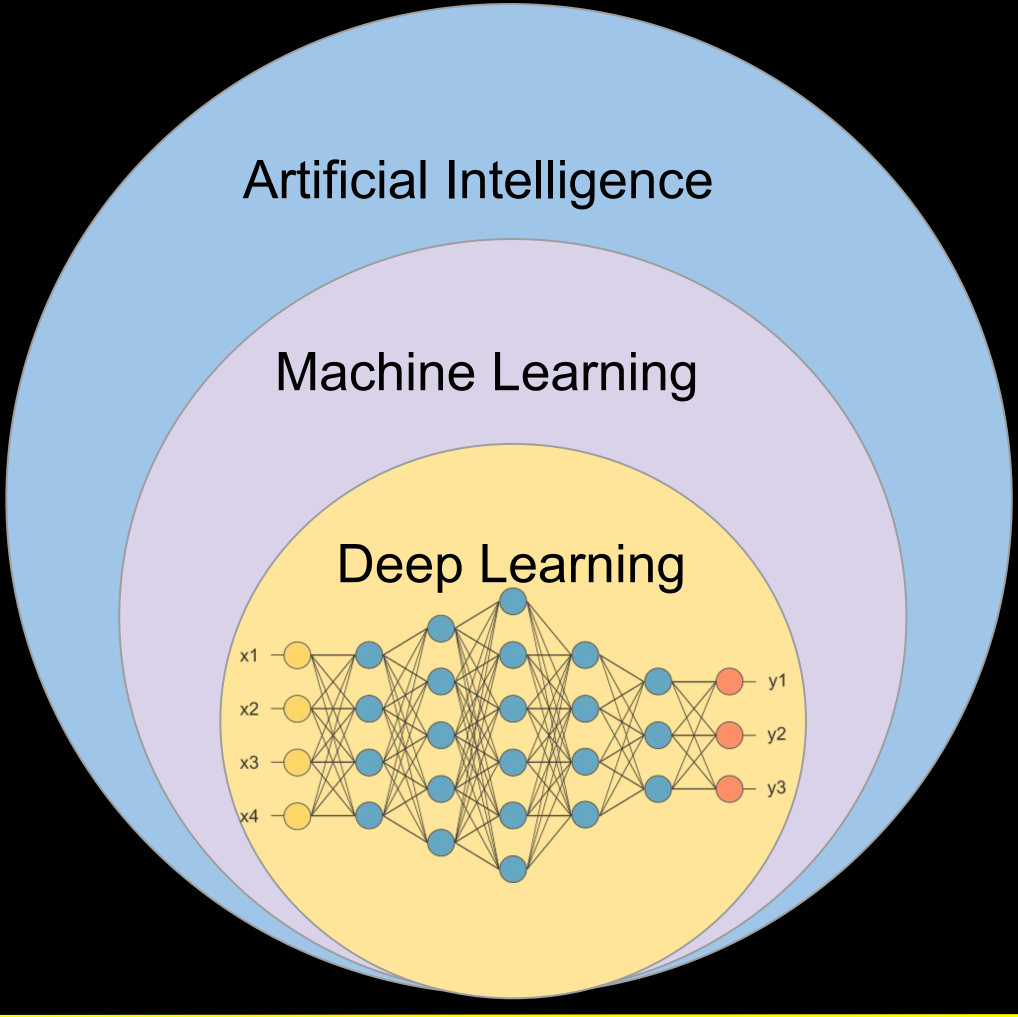
Natural Language Processing [1]



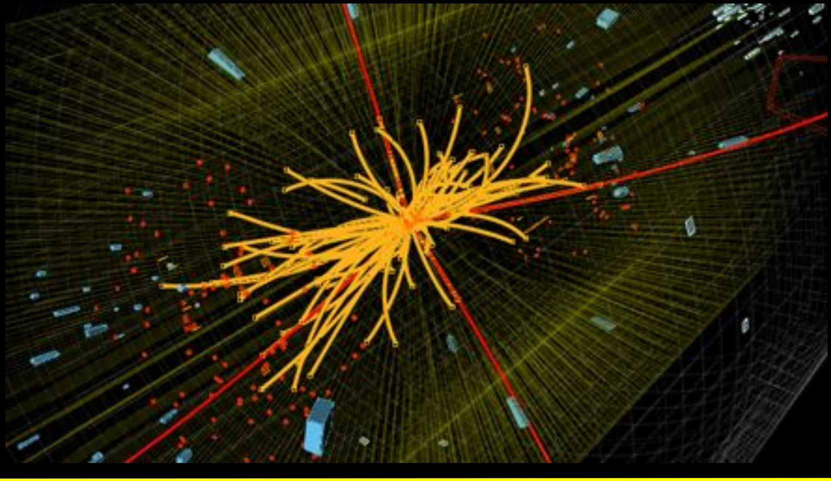
Video to video synthesis [4]

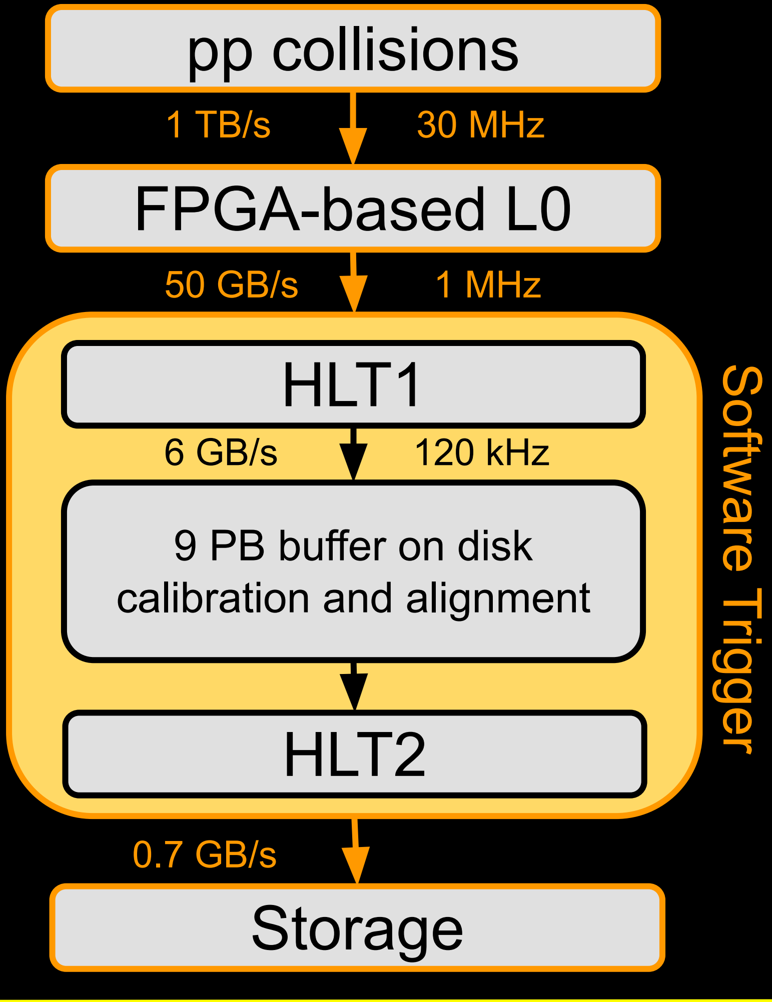


...but this is also the beginning of this incredible data-driven technology, in particular in our field



- DL is a subset of ML which makes the computation of multi-layer NN feasible. When applied to massive datasets and giving massive computer power it outperforms all other models most of the time.
- ML is becoming ubiquitous in nuclear and particle physics.
- DL just started having an impact in nuclear/particle physics





pp collisions

1 TB/s 30 MHz

FPGA-based L0

50 GB/s 1 MHz

HLT1

6 GB/s 120 kHz

9 PB buffer on disk
calibration and alignment

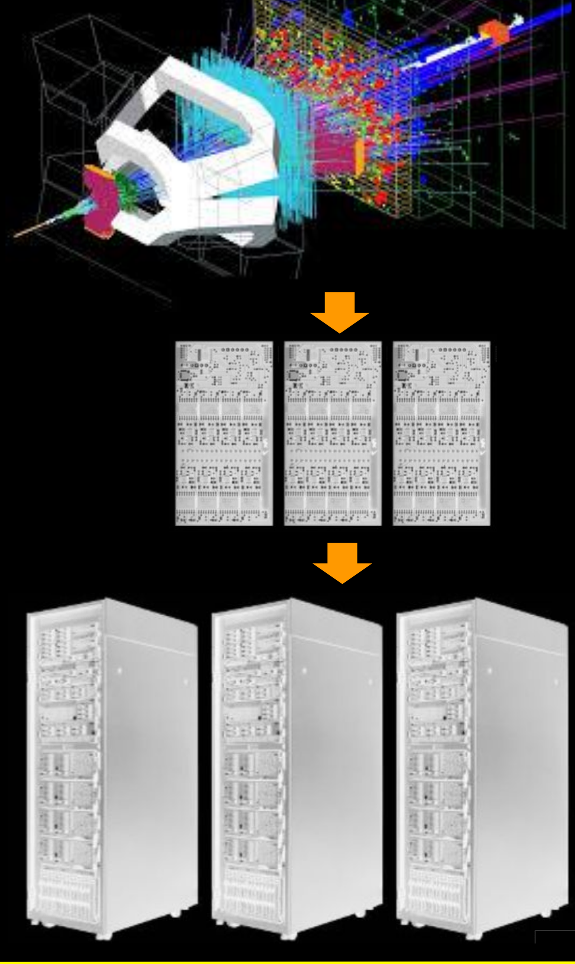
HLT2

Software Trigger

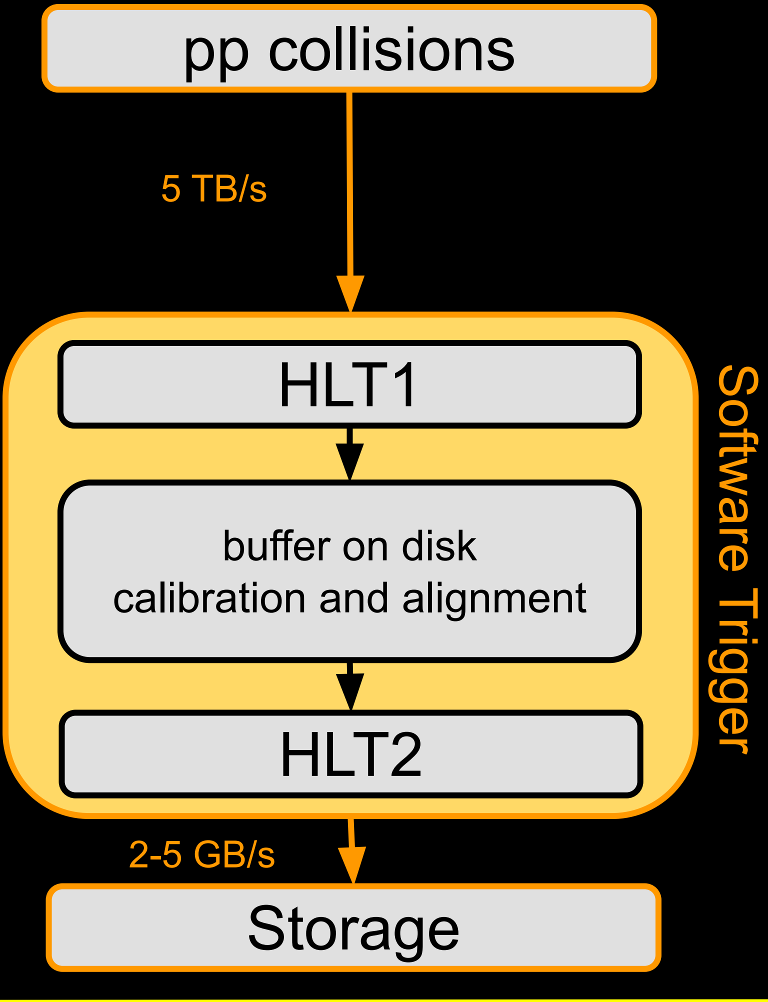
0.7 GB/s

Storage

in the example: LHCb (Run 2)



M. Williams,
ML in the LHCb Trigger and Beyond



pp collisions

5 TB/s

HLT1

buffer on disk
calibration and alignment

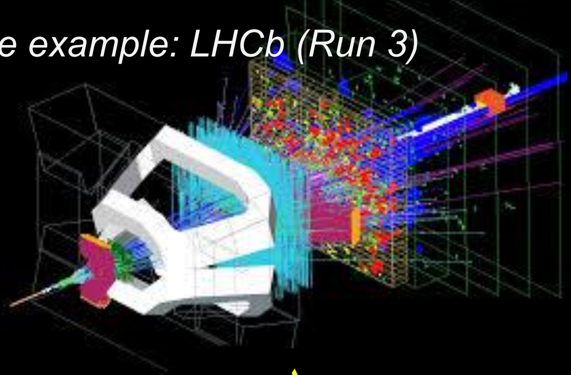
HLT2

Software Trigger

2-5 GB/s

Storage

in the example: LHCb (Run 3)



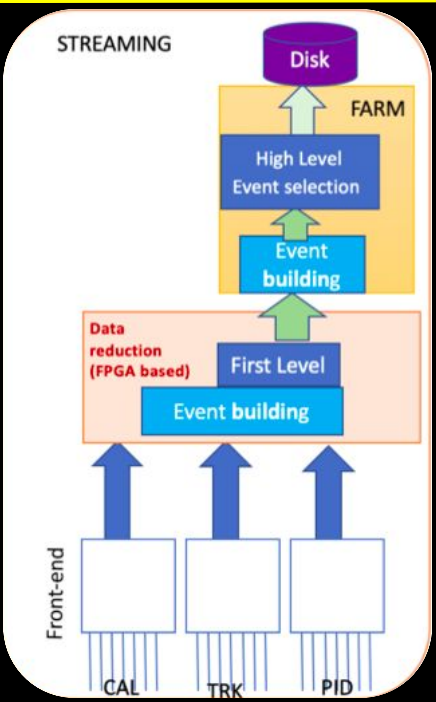
“Triggerless” Readout

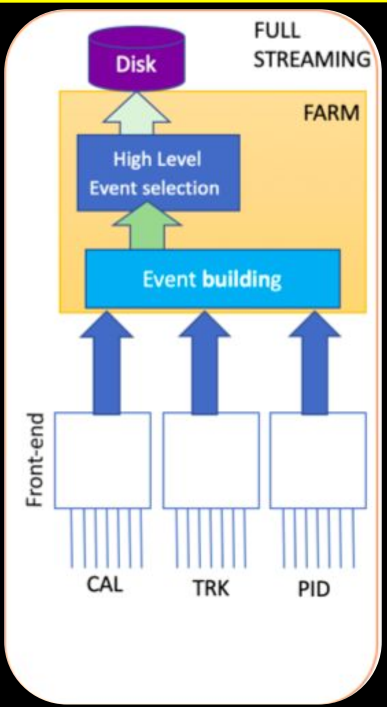
Towards Streaming Readout

G. Heyes, JLab12 streaming readout and future EIC
Y. Furletova, ML for particle identification

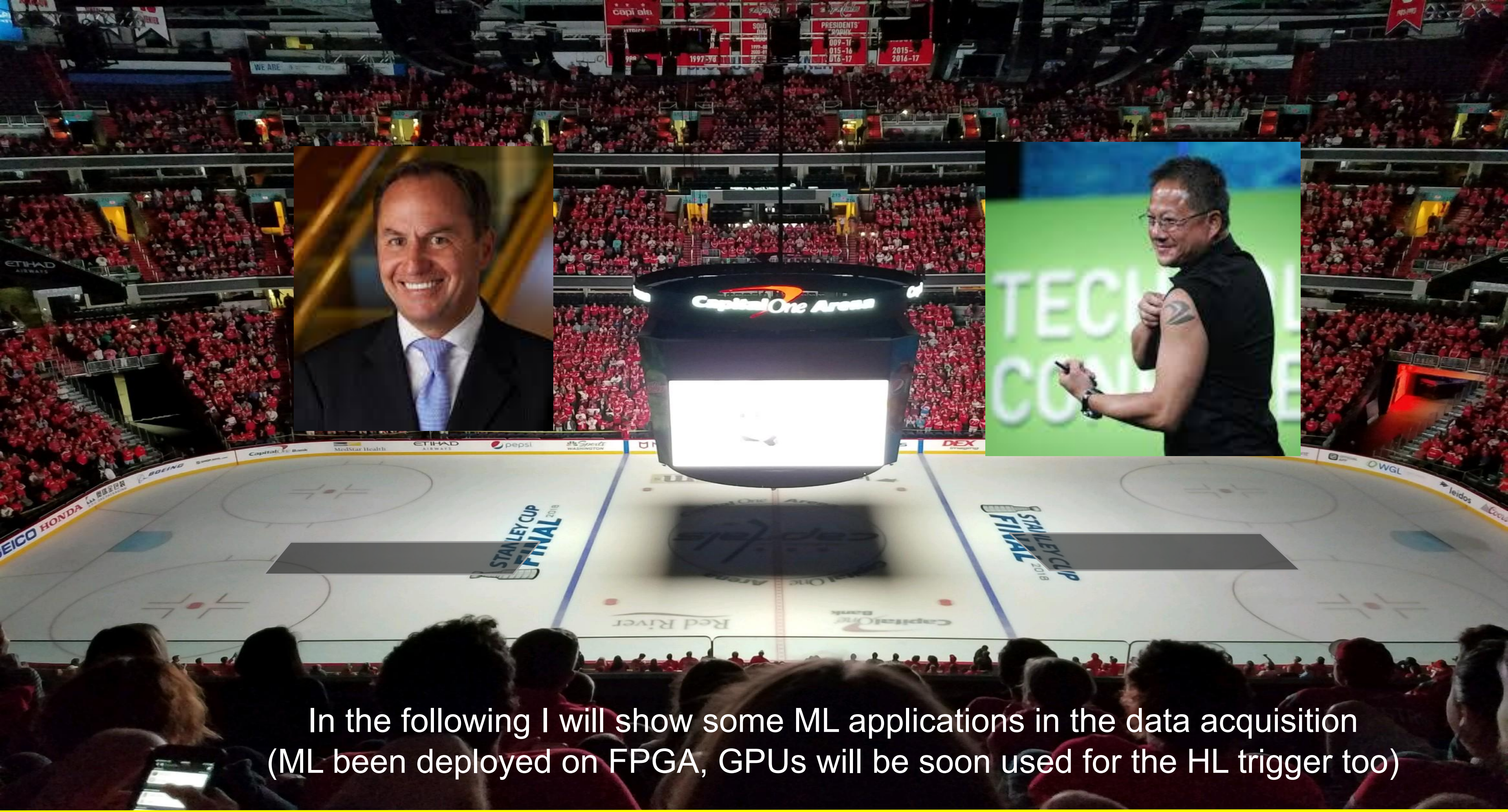
data read continuously from all channels

- Validation checks at source reject noise and suppress empty channels.
- Data then flows unimpeded in parallel channels to storage or a local compute resource.
- Data organized in multi-dimensions by channel and time.





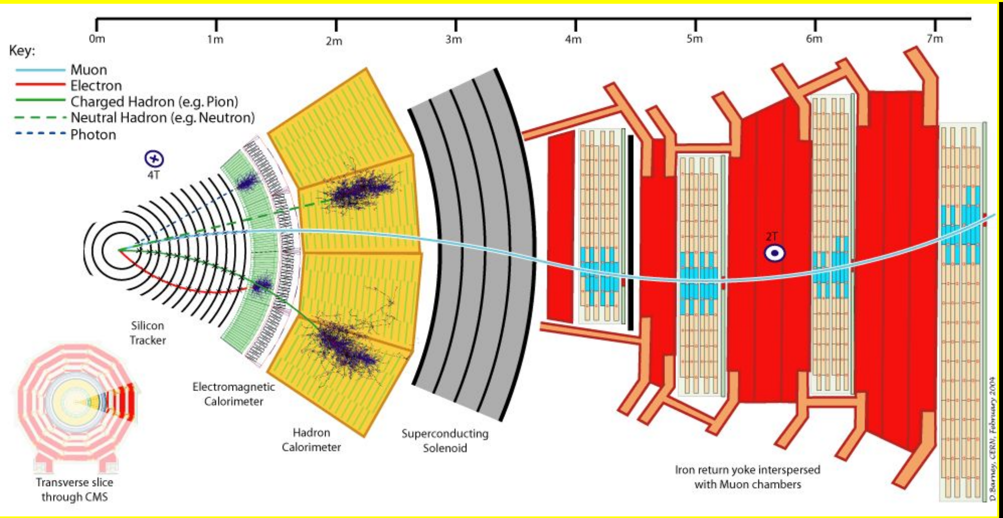
Different streaming pipelines:
FPGA based (w/ data reduction) and full streaming.
ML naturally suited for online data reduction or high level physics event selection/trigger



In the following I will show some ML applications in the data acquisition (ML been deployed on FPGA, GPUs will be soon used for the HL trigger too)

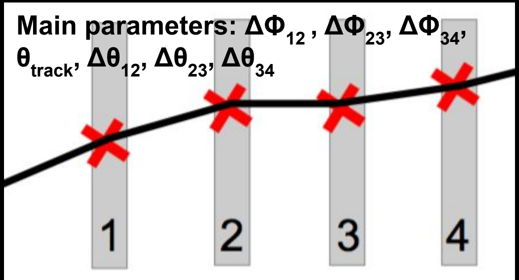
Hardware Trigger: Muon ID CMS

(CMS CR -2017/357)



0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D Barney, CERN, February 2004

- L1 trigger responsible for selecting 100k/s interesting events out of the 40M/s
- Endcap Muon Track Finder (EMTF)
  - Needs to operate fast (~ 500 ns)
  - No tracker info available, only muon chambers
- Want Machine Learning to do the p_T assignment (implemented on FPGA)



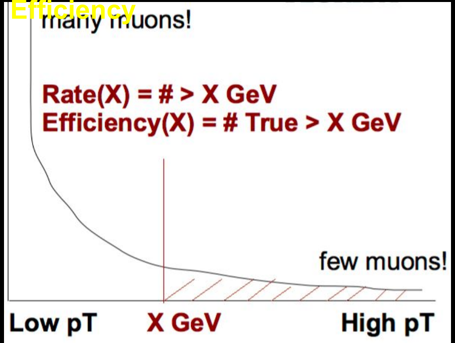
Main parameters: $\Delta\Phi_{12}$, $\Delta\Phi_{23}$, $\Delta\Phi_{34}$, θ_{track}, $\Delta\theta_{12}$, $\Delta\theta_{23}$, $\Delta\theta_{34}$

- transverse momentum (p_T) is assigned based on curvature
- The Endcap Muon Track Finder (EMTF) needs to process hits and assign a momentum
- Interesting muons have large p_T

Muon ID CMS

(CMS CR -2017/357)

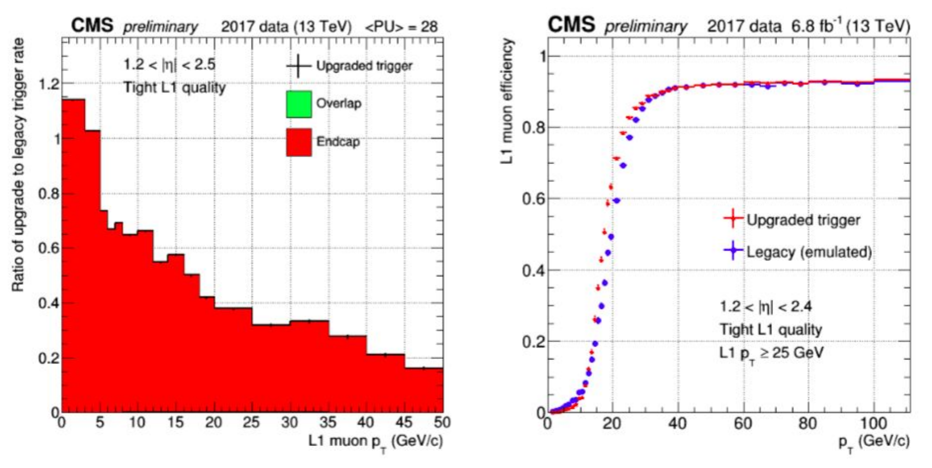
Goal:
Minimize Rate Maximize Efficiency



- 2500 operations to assign the p_T for a single track.
- A BDT would take ~2500 ns with standard settings.
- Create a look-up table* to reduce the 2500 operations into 1 operation

*LHCb first employed the discretized LUT BDT approach in 2011

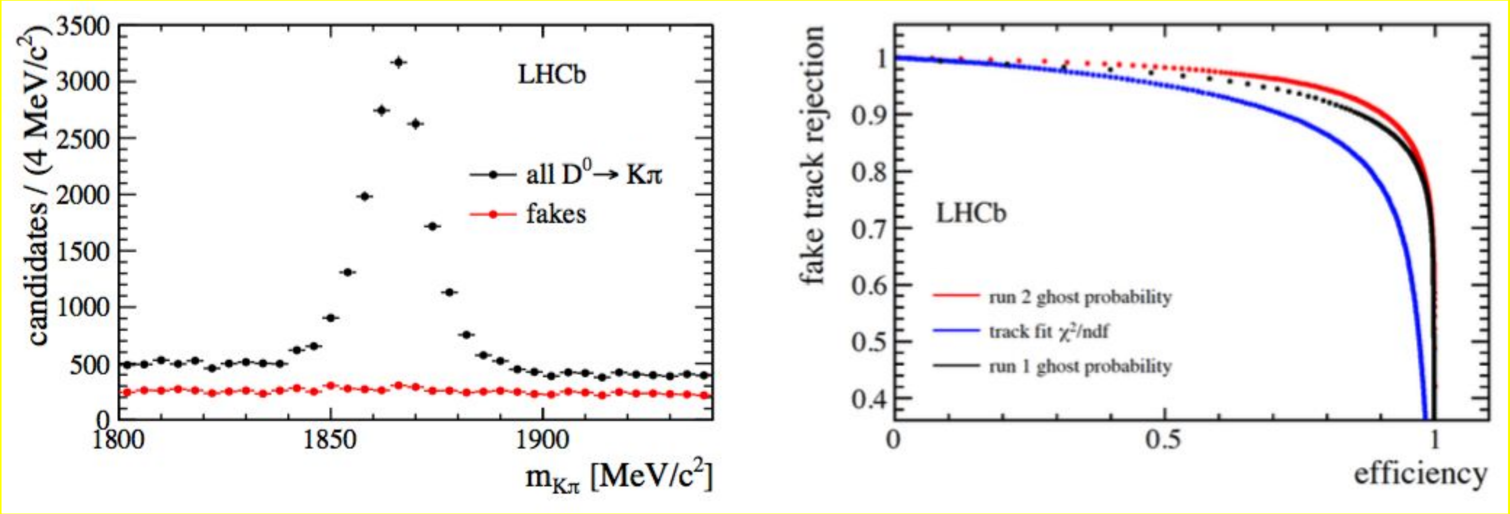
- Trade time for memory by discretizing features and fit into 30 bits:
  - e.g. var 1 = 10 bits, var2 = 5 bits, var3 = 5 bits, var4 =5 bits, var5=5 bits => Input = 30 bits
  - Map each input to the ML output and save map 2^30 possibilities w/ 9 bit output = 1.2 GB LUT



- LUT into the FPGA Implemented this design in 2016/2017 data taking
- Improved 3x rate reduction (for p_T > 25 GeV) with small loss of efficiency

“Ghost Tracks Killer” (LHCb-PUB-2017-011)

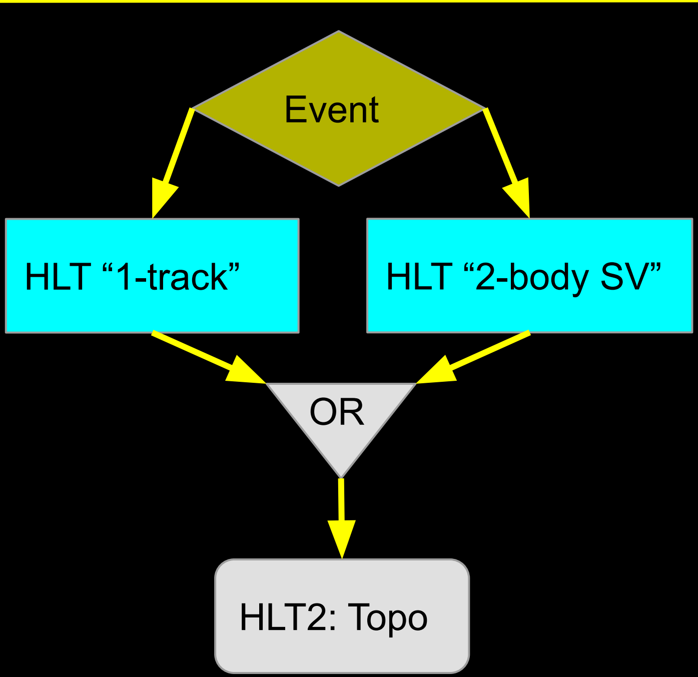
Fake-track (ghost) killing DNN based on 22 features, most important are hit multiplicities and track-segment chi2 values from tracking subsystems. Significantly reduces the rate of events selected in the HLT1.



Run in the trigger on all tracks (it must be very fast). Use of custom activation function and highly-optimized C++ implementation.

HLT2 Topological Trigger

T.Likhomanenko et al [1510.00572]
V.Gligorov, M. Williams, JINST 8 (2012) P02013



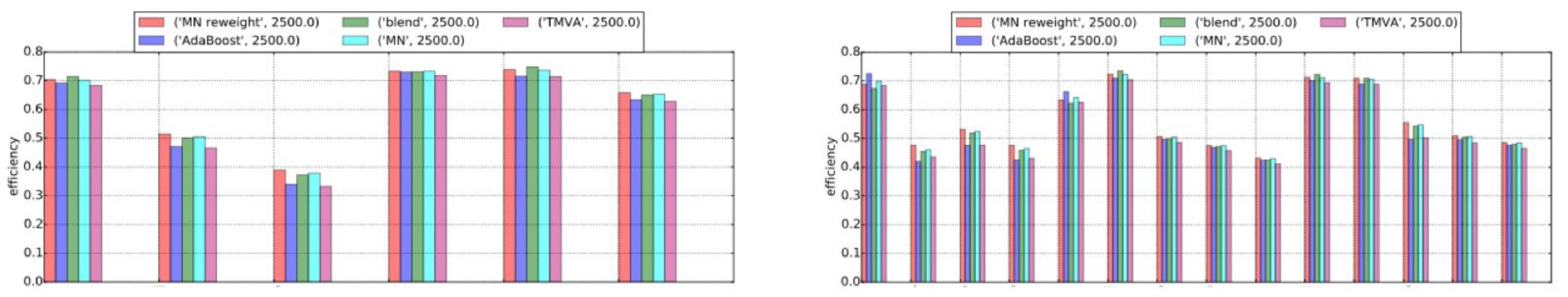
- The main b-physics trigger used by LHCb. Selects vertices which are:
  - ➢ Detached from the primary pp
  - ➢ Compatible with coming from a b-hadron decay

Consists of:

- An SV algorithm that considers 2, 3, and 4-track vertices (seeded by HLT1 ML selections).
- The ML uses a list of features: n(tracks), corrected mass, vertex χ^2, scalar track p_T sum, flight distance χ^2, pseudorapidity (PV-SV), min(track p_T), n(small IP tracks), IP χ^2, n(very b-like tracks).
- All features are discretized in the ML for stability, robustness, etc. This allows to control growth of DT.

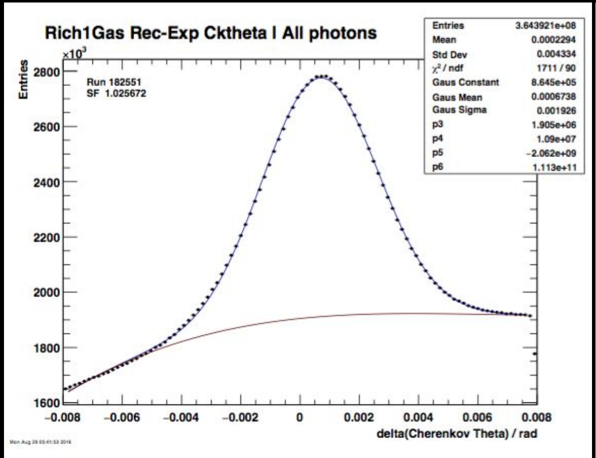
HLT2 Topological Trigger

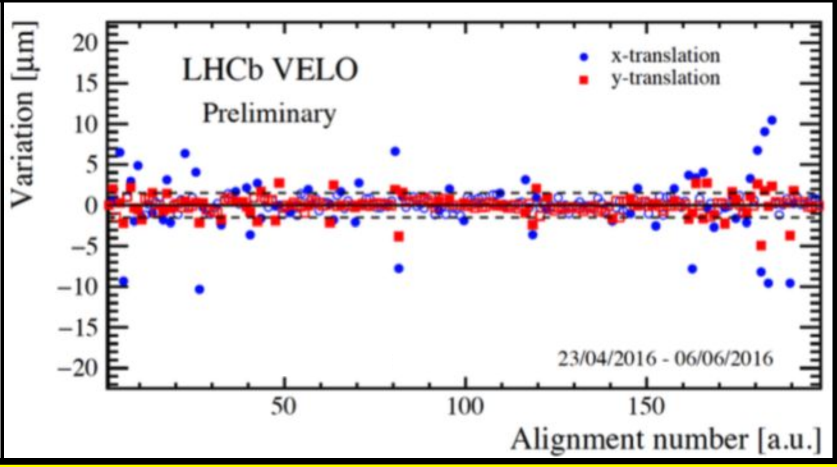
- This BDT algorithm has run since 2011, collecting the data used by ~200 papers.
- It was re-tuned for Run 2 by Yandex and it is now based on MatrixNet.
- In the LHC Run 1, this trigger, which utilized a custom boosted decision tree algorithm, selected a nearly 100% pure sample of b-hadrons with a typical efficiency of 60-70%.



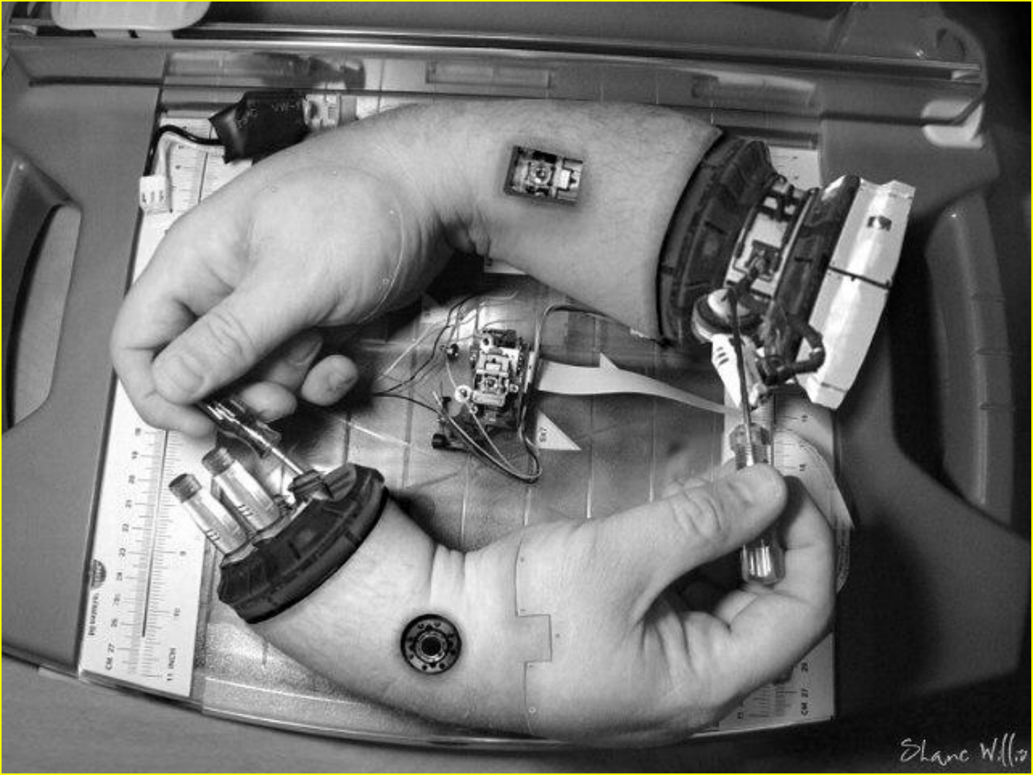
Real-Time Calibration in LHCb

- VELO opens/closes every fill, expect updates every few fills. Rest of tracking stations only need updated every few weeks.
- RICH gases indices of refraction must be calibrated in real time; requires ~1 min to run, and new calibrations are required for each run.
- Calibration data is sent to a separate “stream” from the physics data after the first software-trigger stage. This permits running the calibrations on the online farm simultaneously with running the trigger.

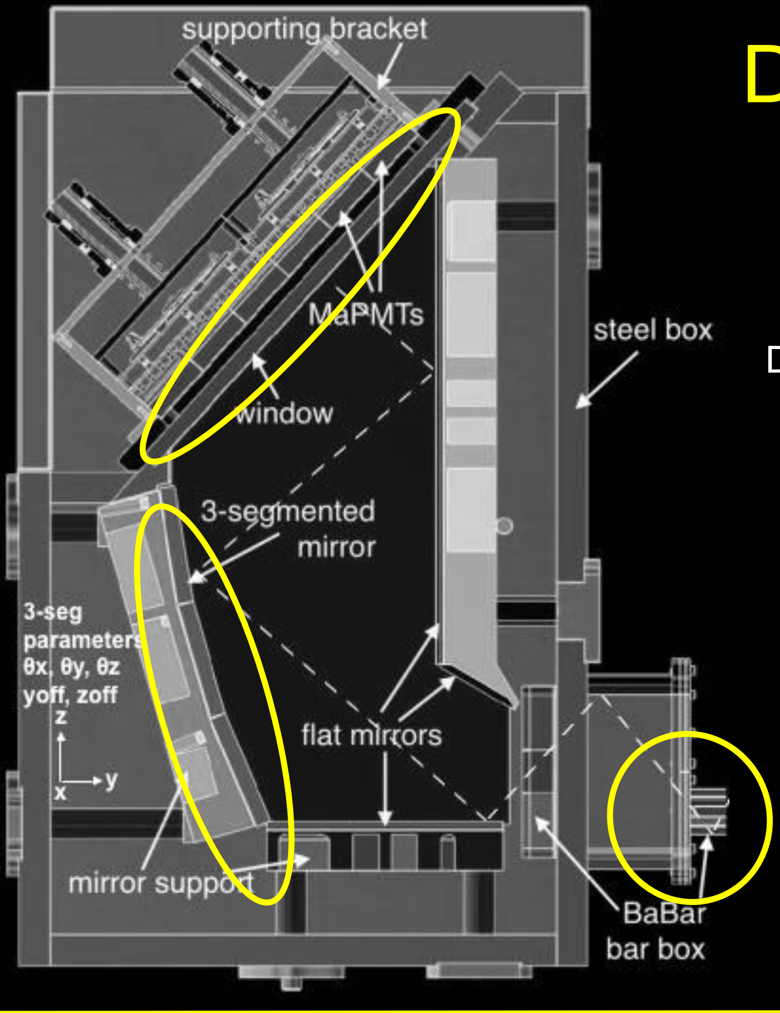




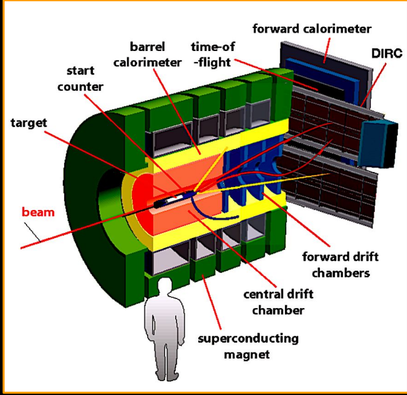
Further Improvements



New Directions

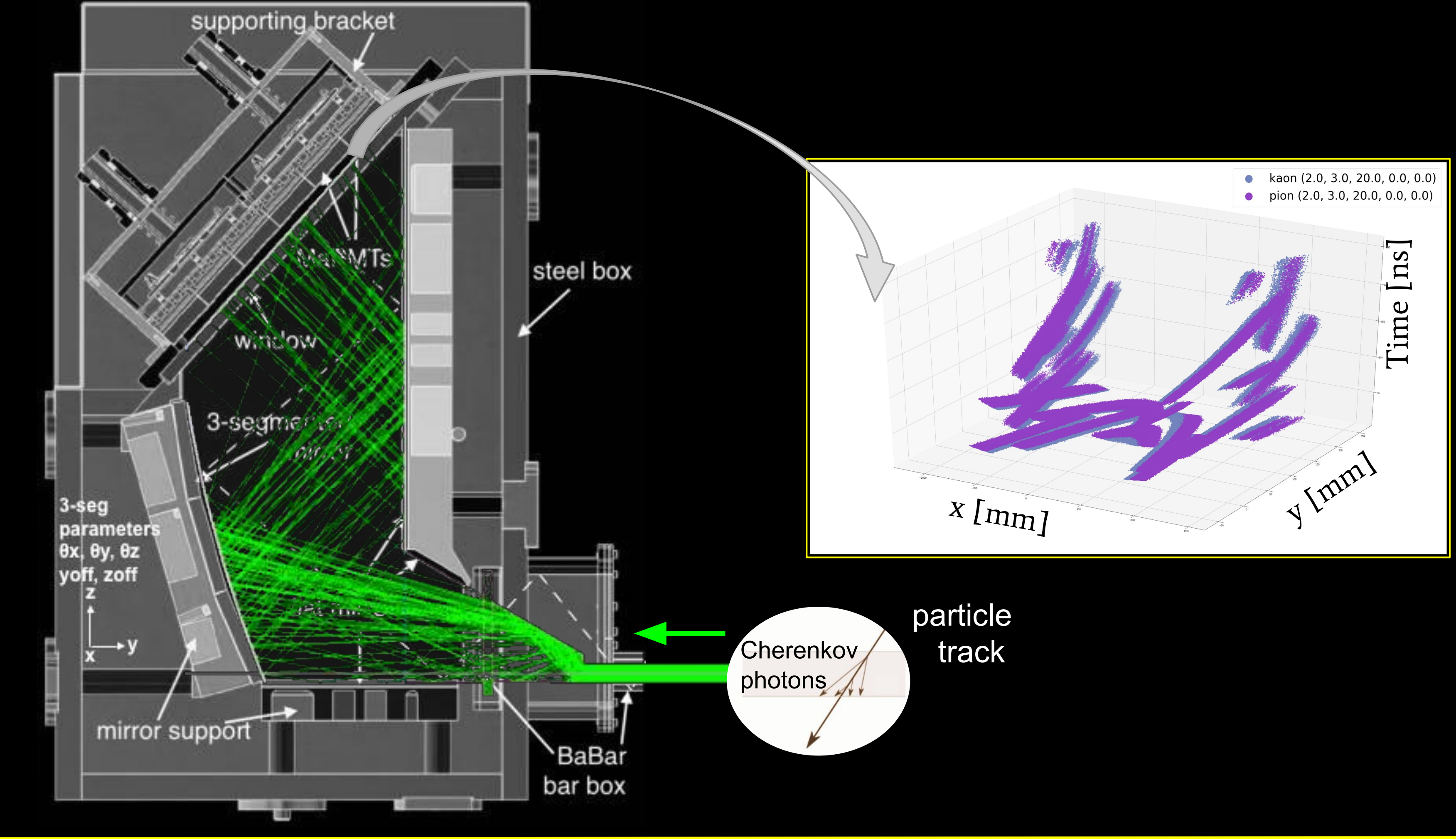


Detector Alignment

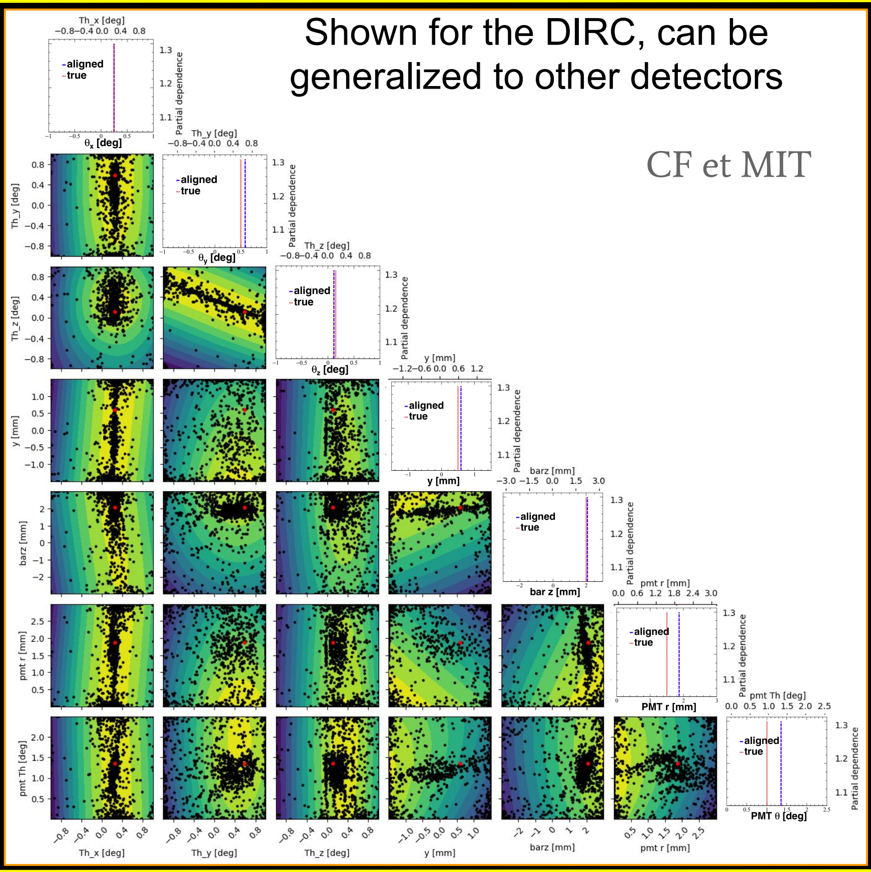


DIRC @ GlueX/JLab

- Optical box made by several components and filled by water.
- During data-taking this becomes a noisy black-box problem with many non-differentiable terms.
  - relative alignment of the tracking system with the location and angle of the bars
  - mirrors shifts cause parts of the image change
  - other offsets
- These aspects make seemingly impossible to analytically understand the change in PMT pattern
- Requires dedicated system for calibration.

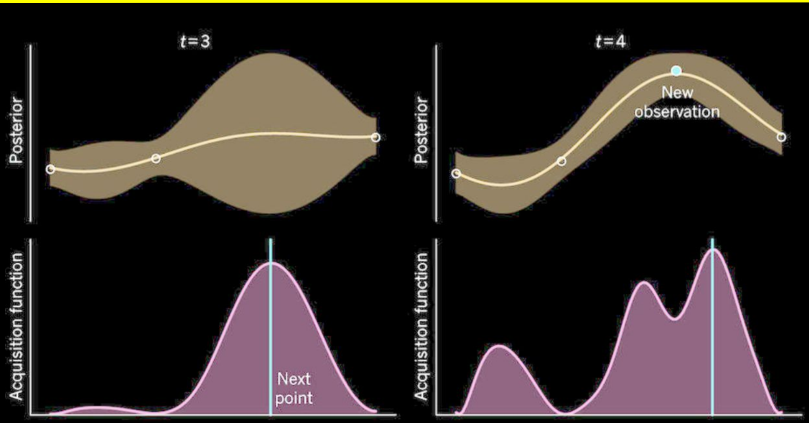

supporting bracket
MaPMTs
steel box
window
3-segmented mirror
3-seg parameters θx, θy, θz yoff, zoff
z
y
x
mirror support
BaBar bar box
Cherenkov photons
particle track
kaon (2.0, 3.0, 20.0, 0.0, 0.0)
pion (2.0, 3.0, 20.0, 0.0, 0.0)
Time [ns]
x [mm]
y [mm]

Self-learning alignment parameters



Shown for the DIRC, can be generalized to other detectors

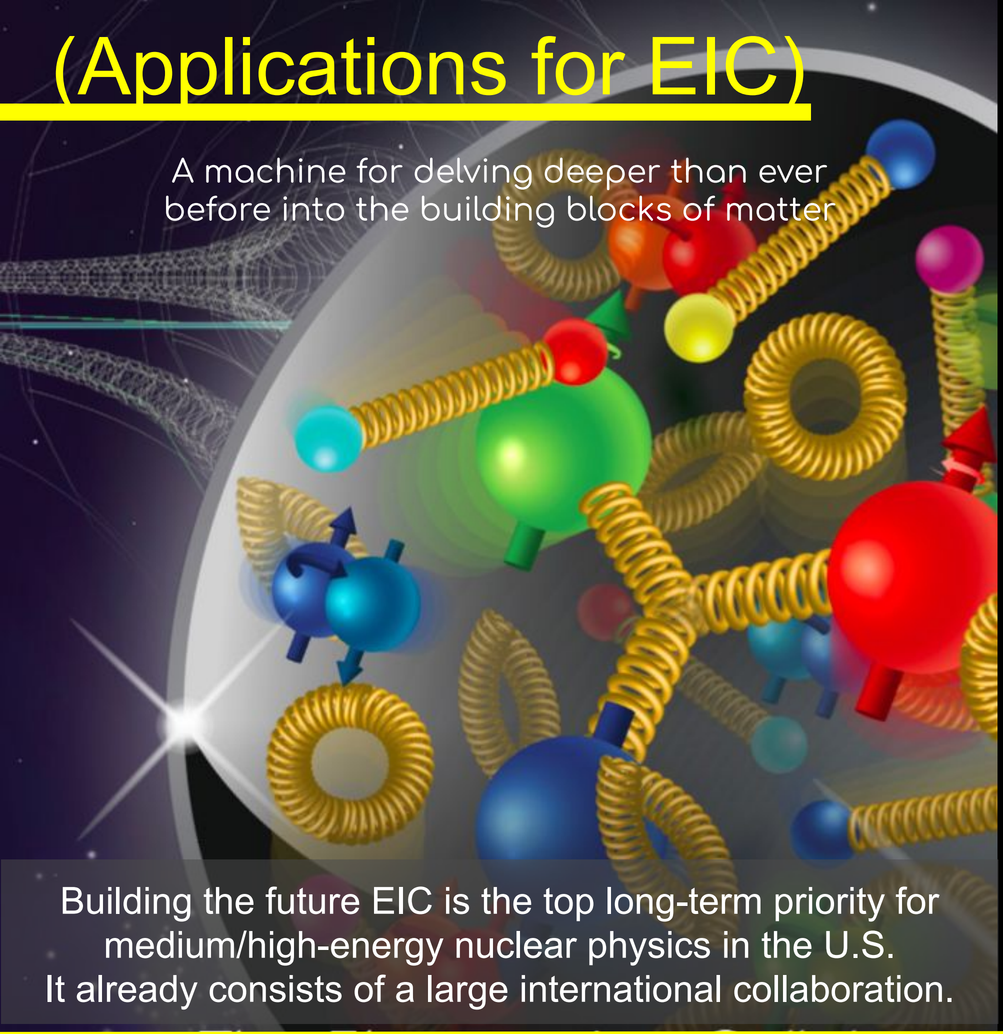
CF et MIT



Bayesian Optimization

(left) correlations among the main misalignment parameters

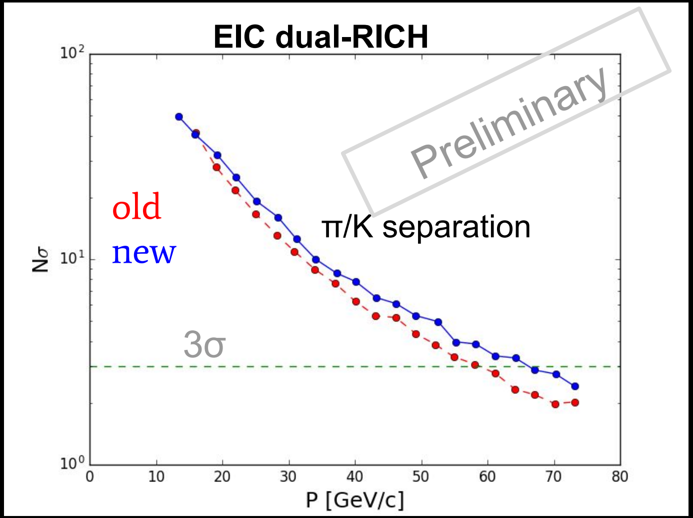
(Applications for EIC)



A machine for delving deeper than ever before into the building blocks of matter

Building the future EIC is the top long-term priority for medium/high-energy nuclear physics in the U.S.
It already consists of a large international collaboration.

E. Cisbani, A. Del Dotto, CF



This approach finds a lot of useful applications:

- Optimal Design (hardware, ...)
- Tuning Simulations (cf. Ilten, Williams, Yang [1610.08328])
- Hyperparameters (e.g. DNN)
- Calibration (cf. GlueX DIRC)
- etc…

Anomaly Detection

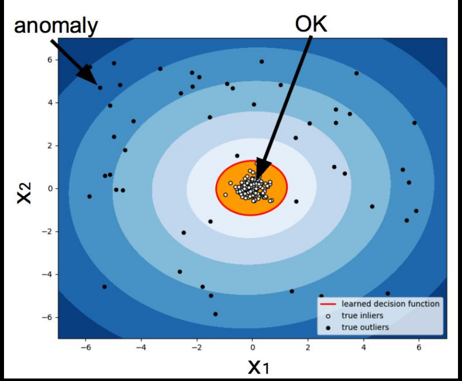
M Borisyak et al 2017 J. Phys.: Conf. Ser. **898** 092041

Towards Automated Data Quality System

It’s a continuous supervised learning approach:

- Historical data processed by experts (classifying data good/bad)
- System learns patterns

Establish procedure to split data into “definitely bad”, “definitely good”, and “expert needed”



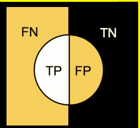
- As new data is coming the supervisor continue making complicated labelling
- Can think of approaching the problem in minimal chunks of data to label

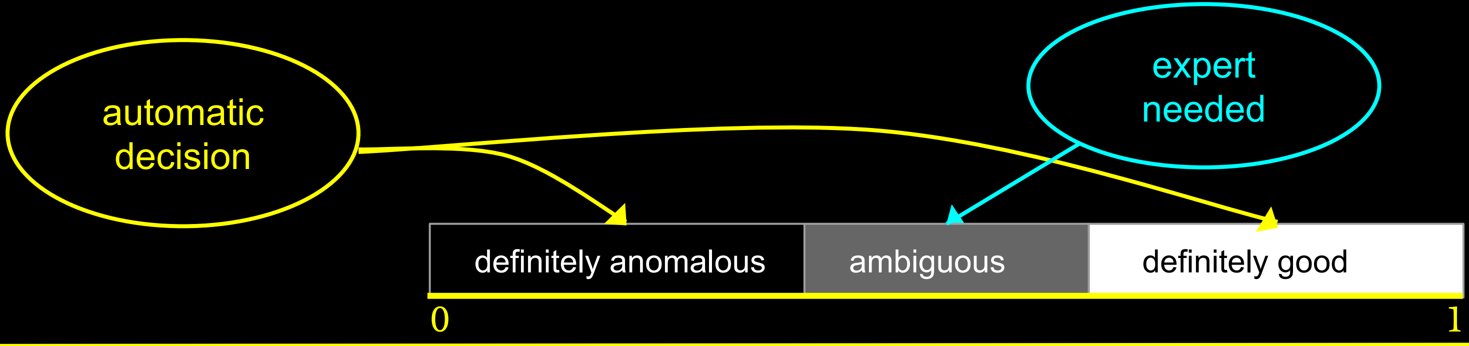
scores

Loss rate
L = FN/(TP+FN)

Pollution rate
P = FP/(TP+FP)

Rejection rate
R = grey / (black + grey + white)





reduce Rejection minimizes human evaluation

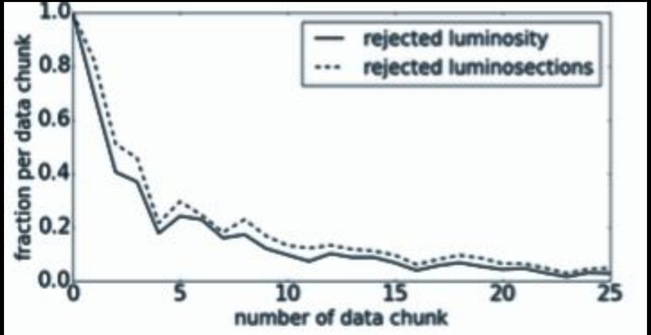
↻

retrain

Anomaly Detection

Different approaches under study for the identification of anomaly (N.B.: specific failure modes can be labeled). Used “lumi-sections” (CMS open data).

- Combining different “channels”. Used a BDT.

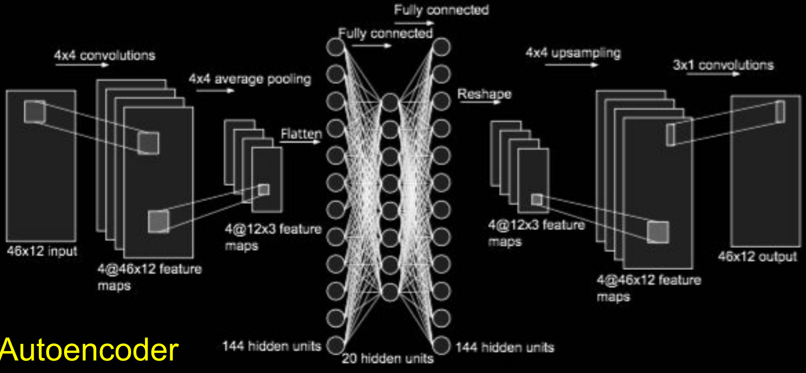


M Borisyak et al (YANDEX)
J. Phys.: Conf. Ser. **898** 092041

fractions of rejected luminosity and lumisections gradually decrease as classifier gets more labeled data

- Explored autoencoders for rare anomalies. When trained on the inliers, testing on unseen faul sample tend to yield sub-optimal representations, thus providing a metric for quantifying the anomaly in occupancy plots.

1808.00911 A.A. Pol et al



Autoencoder

All models instructed to minimize the mean squared error between original and reconstructed samples

$$\epsilon^k = \frac{1}{ij}\sum_{i,j}(\dot{x}^k_{i,j} - \ddot{x}^k_{i,j})^2$$

severity of a potential anomaly ~ p-value

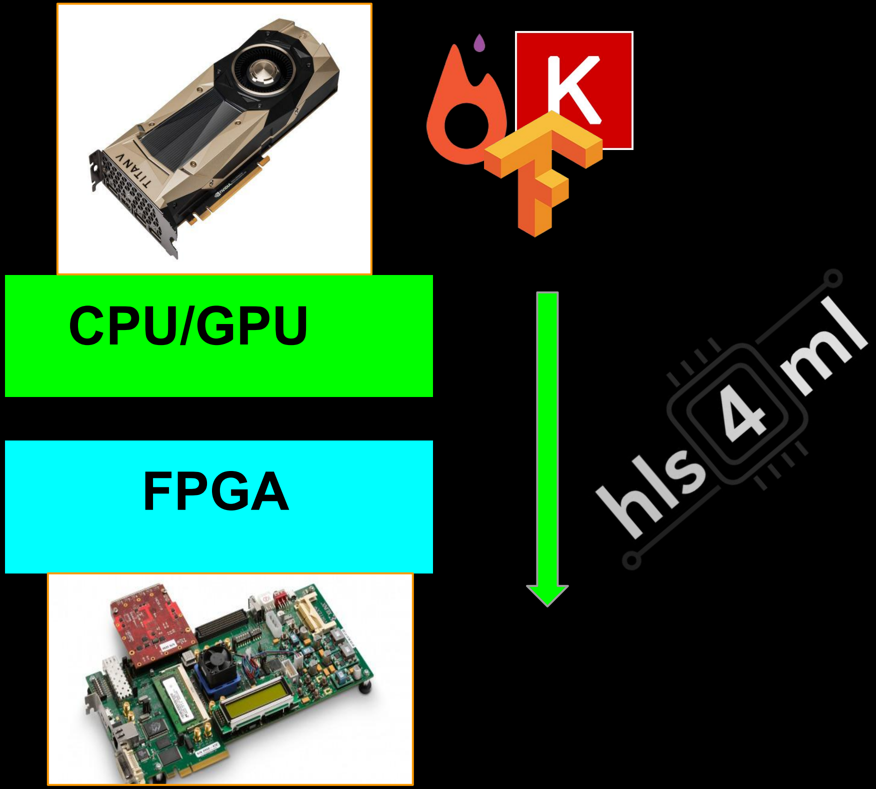
FPGA: A “Top/Down” Approach

arXiv:1804.06913v3

- hls4ml is a package for creating HLS implementations of neural networks

  https://hls-fpga-machine-learning.github.io/hls4ml/

- Supports common layer architectures and model software
- Highly customizable output to map different latency and size needs
- Simple workflow to allow quick translation to HLS



Summary

- With high luminosity and high data rate environment we have to be able to make FAST decisions along data transfer.
- Heavy use of ML @LHC during Run 1 and Run 2.
- ML on FPGA allows online data reduction.
- Run3 LHCb will upgrade to a triggerless readout.
- Both data science and detector expertise needed to implement advanced approaches, e.g. detection of anomalies in detector data.

Deep Net Painting of Portofino