Automation of Data Quality & Certification w/ Deep Learning

EP Software R&D First Lightning Talks Session

G. Cerminara, *G. Franzoni*, M. Pierini, F. Pantaleo, A. Pol, F.Siroky

CERN EP-CMG

Automated anomaly detection: physics motivation 1/2

- Historically, artificial intelligence/ machine learning (ML) tools employed in HEP for particle identification and regression

- Recently, new areas of R&D involving ML for **automated** anomaly detection
  - **Real time feedback** on data quality to the shift crew
  - **Offline certification** of reconstructed data for physics analysis

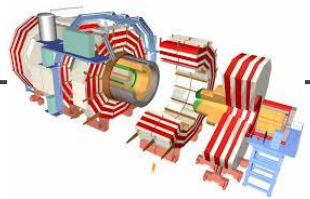
Automated anomaly detection: physics motivation 2/2

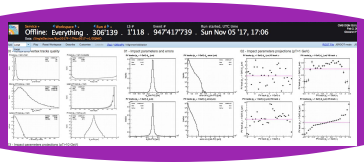
- **Real time feedback** on data quality to the shift crew
  - Currently based on human inspection of plots summarising $O(10^3)$ histos
  - automate the identification of detector anomalies presented to the shifter → increase sensitivity and promptness of reaction to problems
  - Longer term perspective: trained automated agent to assist (→ replace?) most of the shift crew

- **Offline certification** of reconstructed data for analysis
  - Drastically reduce person power needed to inspect for anomalies
  - Smallest time granularity over which the problems are identified → maximise the efficiency of truly-bad data

online histograms → data quality in Real time

- **Real time feedback** on data quality to the shift crew





Online Histograms

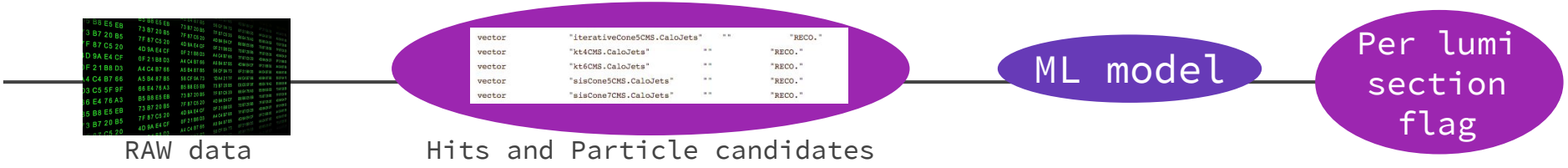
ML model

Alarm ?

- Semi-supervised deep learning models are trained on past histograms to encapsulate the nature of “good data”
- Histograms from live monitoring are evaluated to identify problems with the hardware or the data taking conditions

Reconstructed data → detailed offline certification flags

- **Offline certification** of reconstructed data



- Semi-supervised deep learning models trained on initial runs to encapsulate the nature of “good data”
- Exploit the entire statistical power and full event reconstruction to single out time intervals affected by failures or bad performanve
- Replacing labor intensive run level selection

The technology we need: model training

- Generic **interfaces to aggregate:**
  - **Histograms from recorded events (**RAW data no longer available**)**
  - **Non event data (**DCS, voltages, status of accelerator**)**
  - **reconstructed quantities**

as python accessible objects → use ML frameworks to train

- Framework for **historical/browsability of ground truth**
  - at the moment logging and further reuse mostly kept in logbooks by detector experts
- ML training facility available to CERN project to streamline reoutine re-training

The technology we need: inference w/ ML models on data

- ML growing in data processing (signals reco, particle ID) → **memory & cpu efficient containers for ML model weights**
  - in c++
  - Tools for systematic weights pruning and compression
- ML models need re-training as the detectors evolve/age → **tools to bookkeep versioning of ML models**,
  and efficiently serve multiple instances of the model in the same job