Towards automation of data quality system for CERN CMS experiment

Authors: **Maxim Borisyak**[1,2], **Jean-Roch Vlimant**[3], **Andrey Ustyuzhanin**[1,2], **Maurizio Pierini**[4], **Maria Stenina**[5] and **Dmitry Smolyakov**[2]

[1] National Research University Higher School of Economics;
[2] Yandex School of Data Analysis; [3] California Institute of Technology;
[4] CERN; [5] Yandex

22nd International Conference on Computing in High Energy and Nuclear Physics, San Francisco, October 10-14, 2016

Introduction

Traditionally, quality of the data at CERN CMS experiment is determined manually which requires tremendous amount of person power. In this work, we describes an approach for automated Data Quality system.

Data and feature extraction

CERN open portal data, 2010.
Over 2500 features were extracted. Each feature is defined by:

- **stream**: minimal bias, muon or photon enriched;
- **channel**: muons, photons, flows or calorimiter particles;
- **quantile by particle momentum**: $\frac{5}{5}, \frac{4}{5}, \dots, \frac{1}{5}$;
- **physical property of particle**: η, ϕ, p_T, f_x, f_y, f_z or m;
- **statistic within lumisection**: one of 5 percentiles, mean or standard deviation.

Additionally:

- total momentum of event;
- instant luminosity;
- number of particles in event.

Decision making

Possible labels:

- almost surely good ('white zone');
- almost surely contains an anomaly ('black zone');
- ambiguous ('grey zone').

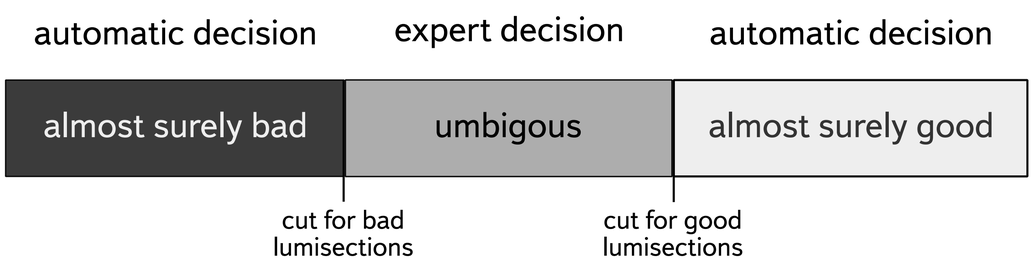

Figure 1: Decision making schematics. Horizontal axis corresponds to classifier's score.

Quality metrics

$$\text{Rejection Rate} = \frac{\text{Rejected}}{\text{Total quantity of samples}} \to \min;$$

$$\text{Pollution Rate} = \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}} \leq \text{const};$$

$$\text{Loss Rate} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}} \leq \text{const}.$$

Stream Learning algorithm

1. set constraints on Pollution and Loss rates;
2. initialize training set as empty;
3. train classifier using k-fold;
4. evaluate classifier's scores;
5. estimate cuts on classifier's scores;
6. receive new chunk of data and classify it;
7. receive labels (expert decisions) for rejected samples;
8. extend training set with rejected samples.
9. repeat from step 3 until out of chunks.

Results

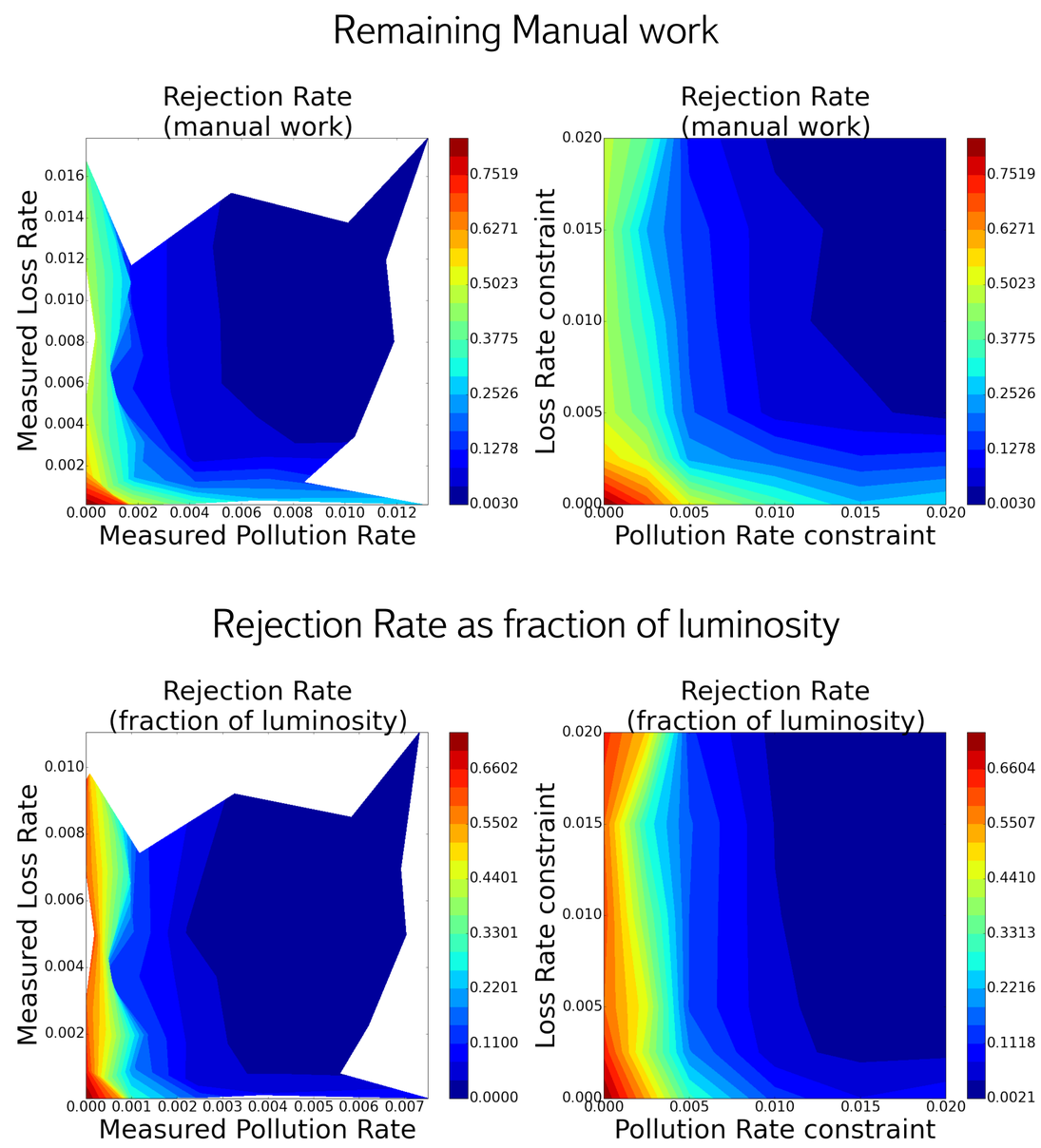

Figure 2: Fractions of manual work (top row) depending on measured Pollution and Loss rates and rejected luminosity (bottom row) depending on constraints on Pollution and Loss rates. Constrains violated only for Pollution Rate $= 0$ or Loss Rate $= 0$, but measured Pollution and Loss rates are both under 0.05%.

Conclusions

- **20% saved person power** for Pollution and Loss rates 0.05%;
- **80% saved person power** for Pollution and Loss rates 0.5%.

In addition, for data not labeled automatically system provides its estimates and hints for a possible source of anomalies which leads to overall improvement of data quality estimations speed and higher purity of collected data.