CERN openlab

Data Popularity and Data Certification

CERN openlab Open Day

Andrey Ustyuzhanin

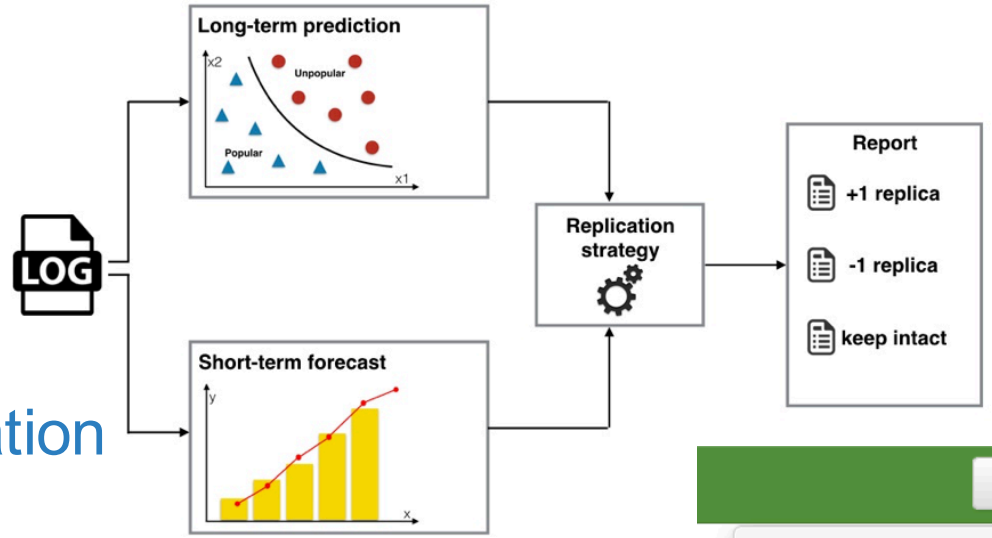
2017-09-21

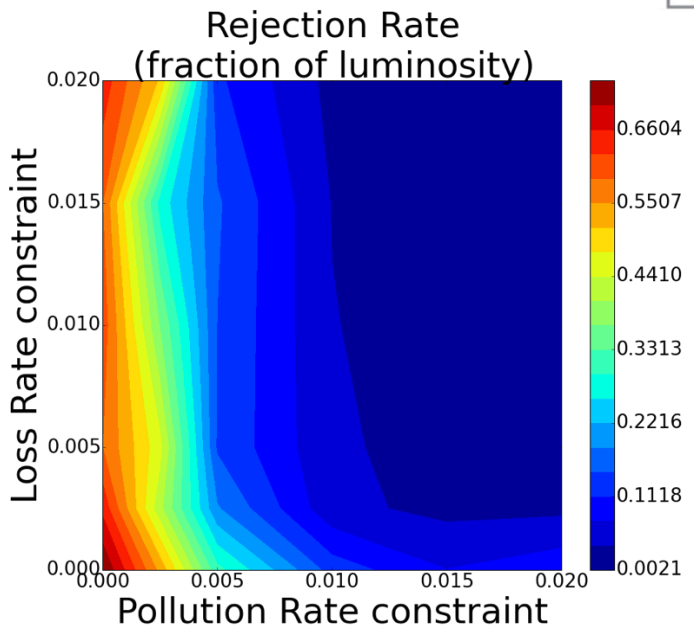
Yandex

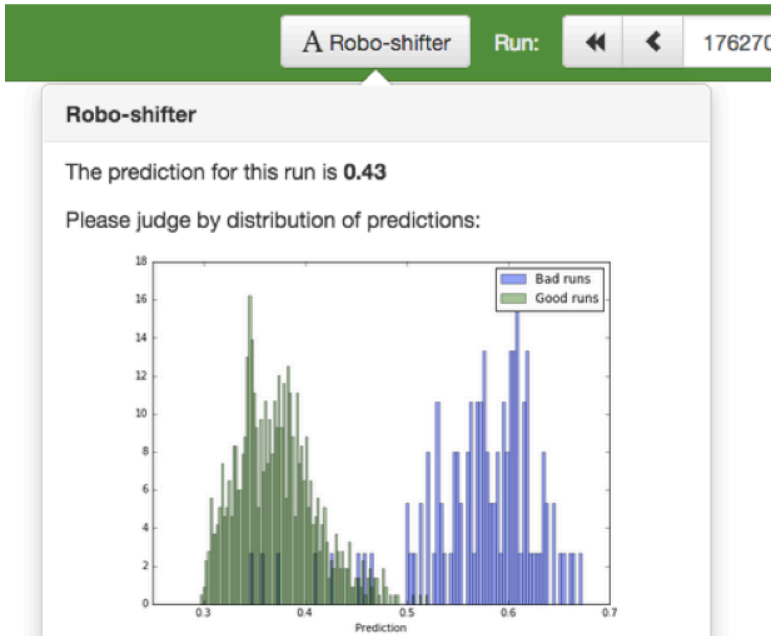
Projects

overview

1. LHCb Data Tiering
2. LHCb Data Certification
3. CMS Data Certification



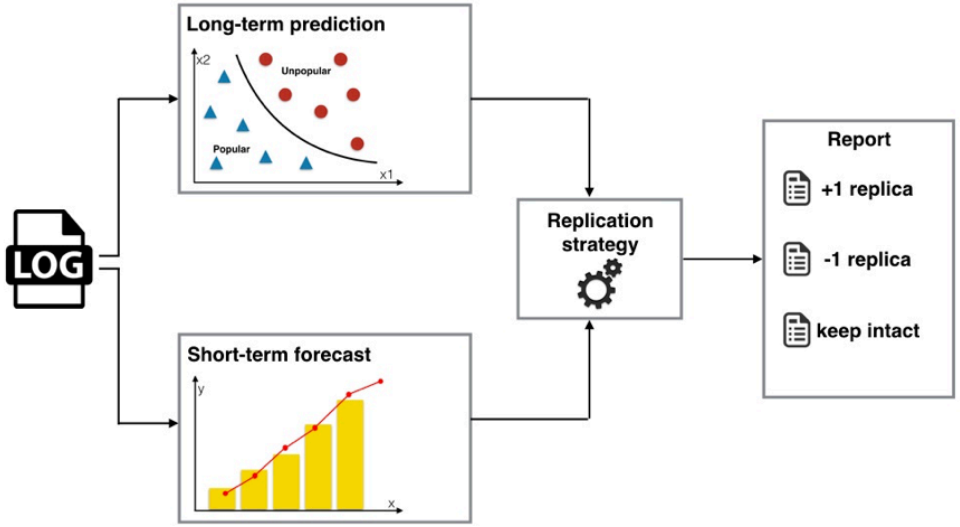




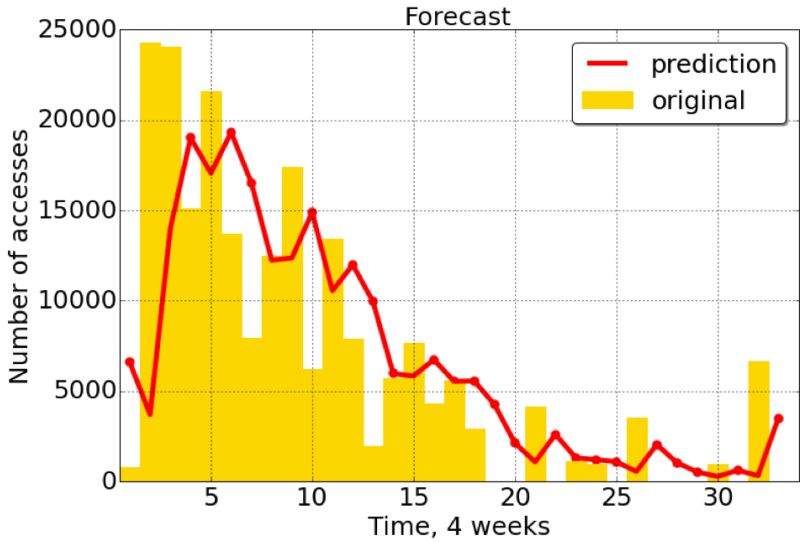
LHCb Data Popularity. Context

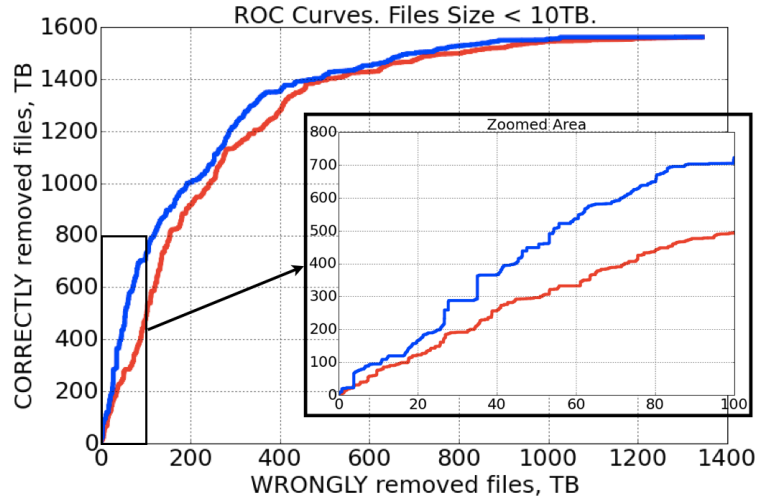
Time series prediction

1. Analyze data usage
2. Predict future usage



#time-series
#storage
#predictive
#assistive intelligence



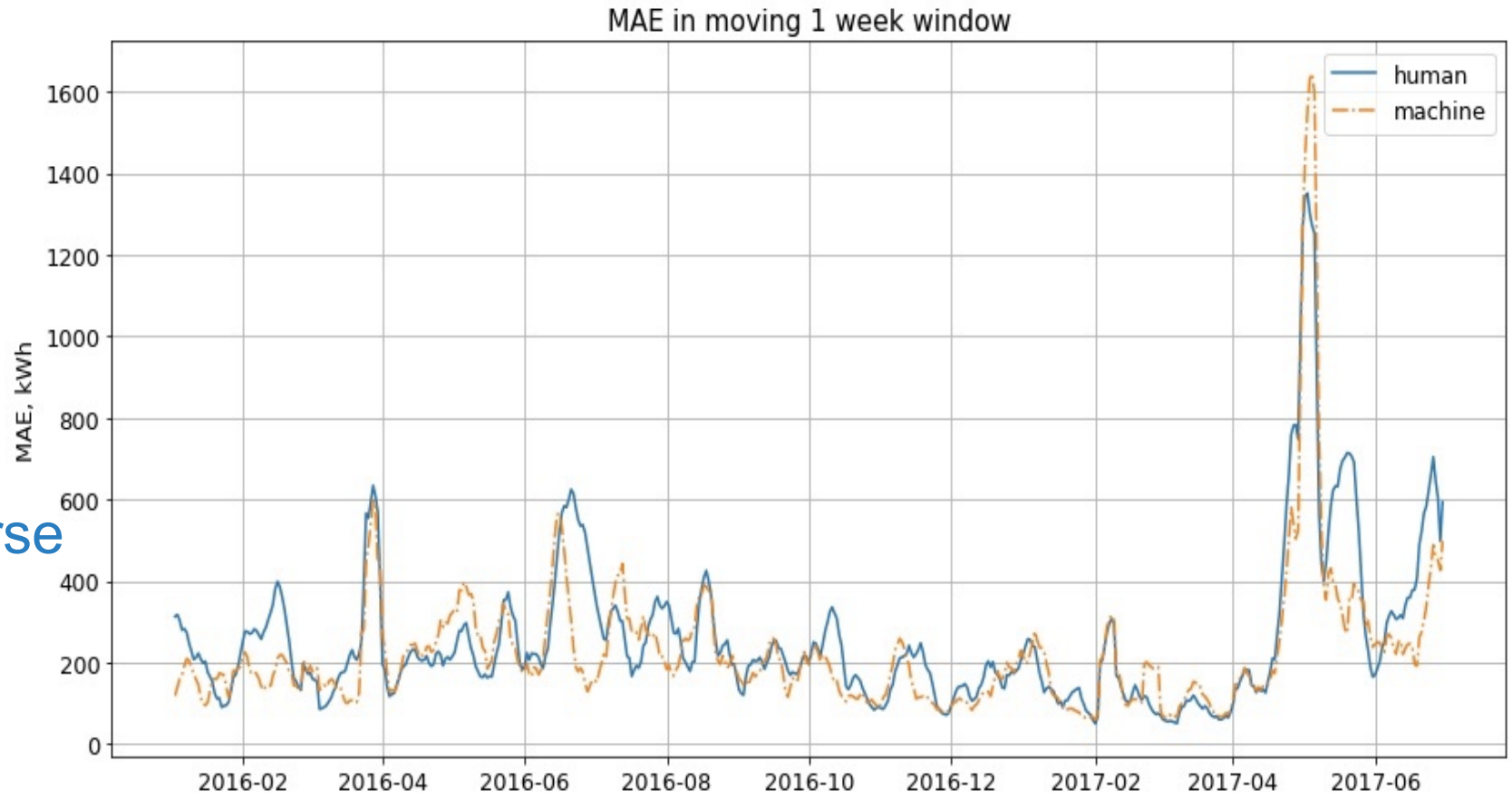


2017

Result

Apply very similar approach to predict Yandex Data Center energy consumption to reduce electricity cost overhead.

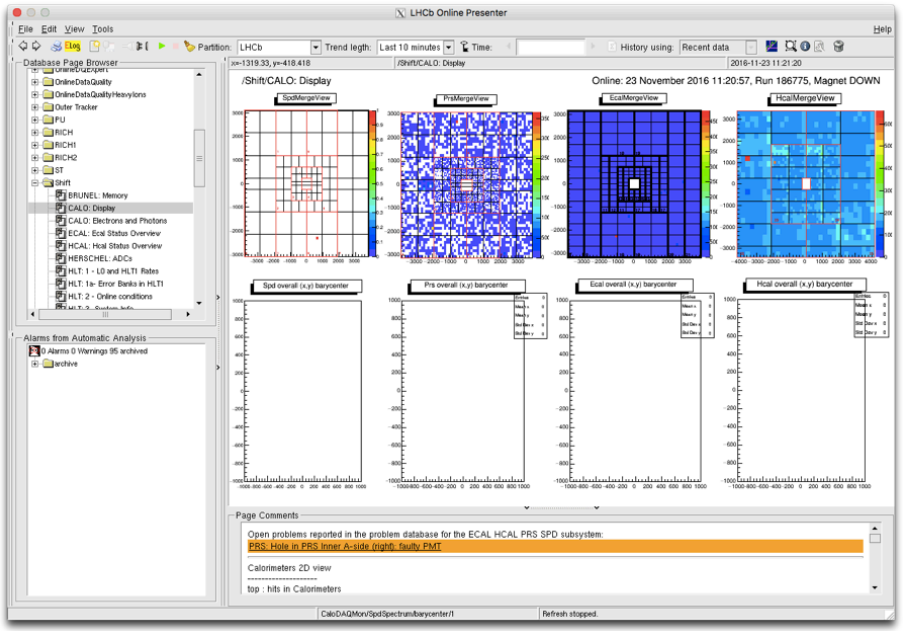
Relative **error** is not worse than human, and **is halved** during certain periods.

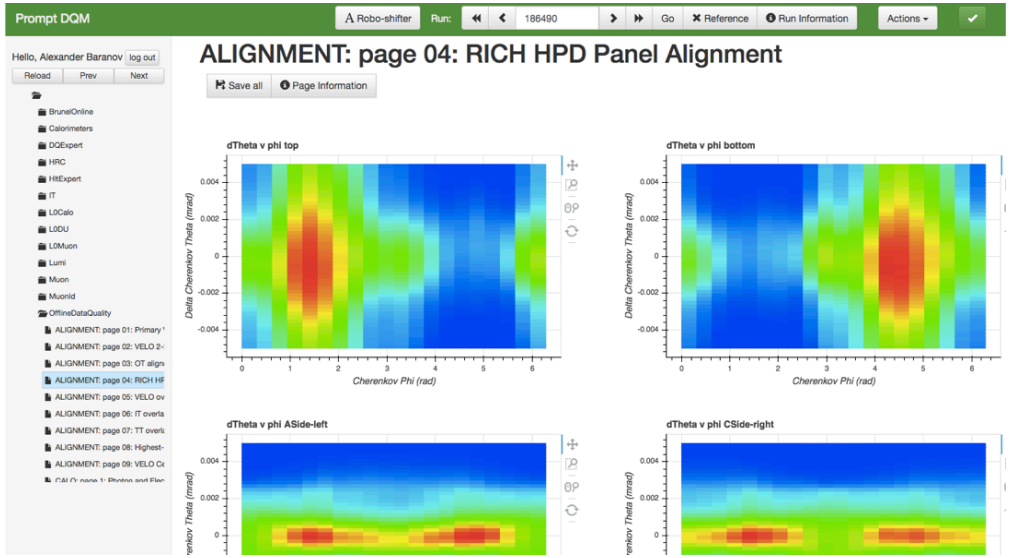


LHCb data quality

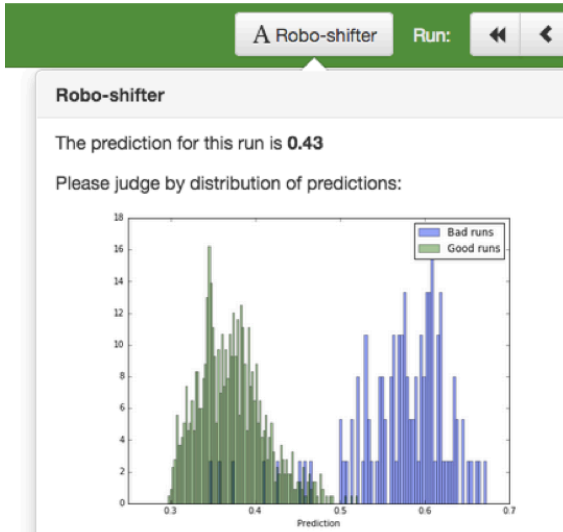
Online & Offline Data Quality Monitoring

- Upgrade LHCb monitoring software
- Add predictive capabilities





#monitoring
#predictive
#anomaly management
#assistive intelligence



CERN openlab

Andrey Ustyuzhanin @ Yandex

2017

Result

- Compare different algorithms
- Decision interpretability
- Class dis-balance mitigation

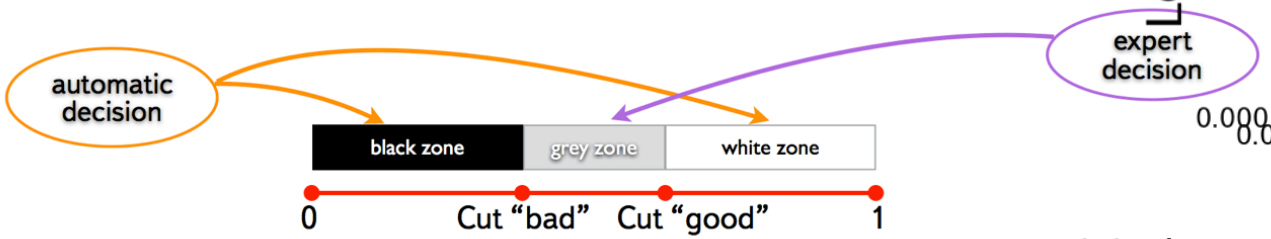
Algorithm	ROC-AUC	Precision	Recall
SVM	0.894	0.895	0.78
Random Forest (simple)	**0.899**	**0.962**	**0.80**
AdaBoost	0.898	0.949	**0.80**
RandomForest (bins)	0.878	0.913	0.77
RandomForest (fft)	0.881	0.924	0.75

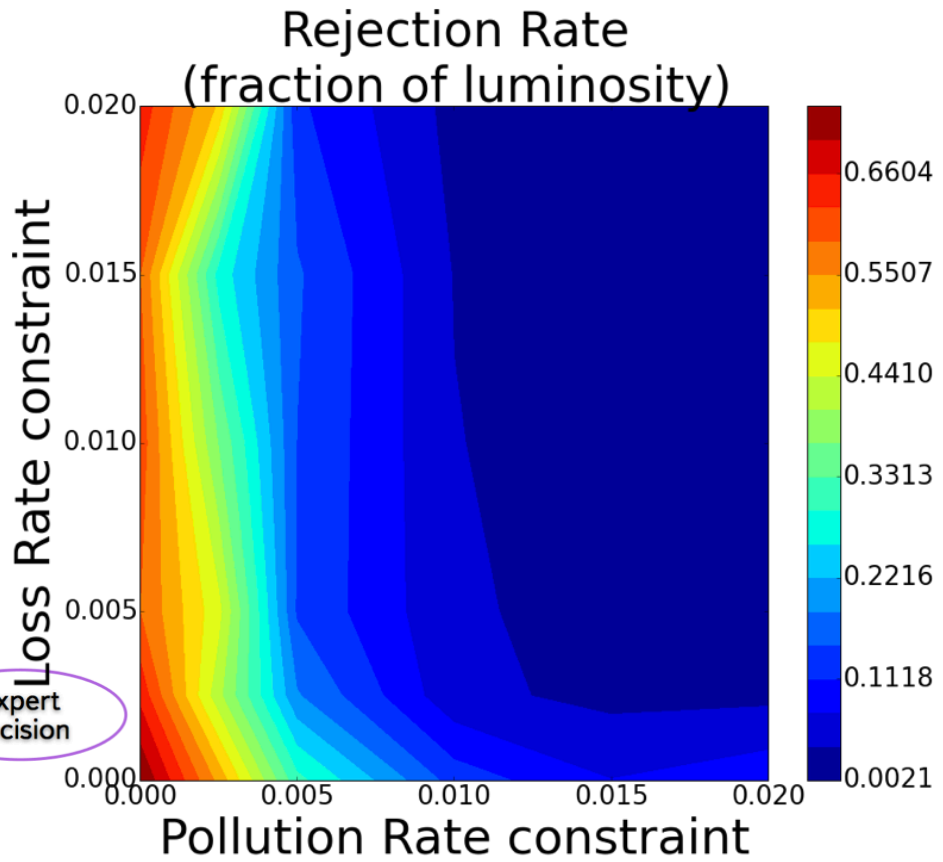
Algorithm	MRR	MAP@1	MAP@5	MAP@10
TreeInterpreter	0.47	0.35	**0.22**	**0.17**
Decision Rule Gradient	**0.56**	0.26	0.12	0.07
AdaBoost	0.54	**0.43**	0.21	0.12

CMS Data Certification

Automatically highlight anomalous lumisections

- Analyze 2010 CERN open data
- Develop algorithm for bad lumisection identification, based on event characteristics





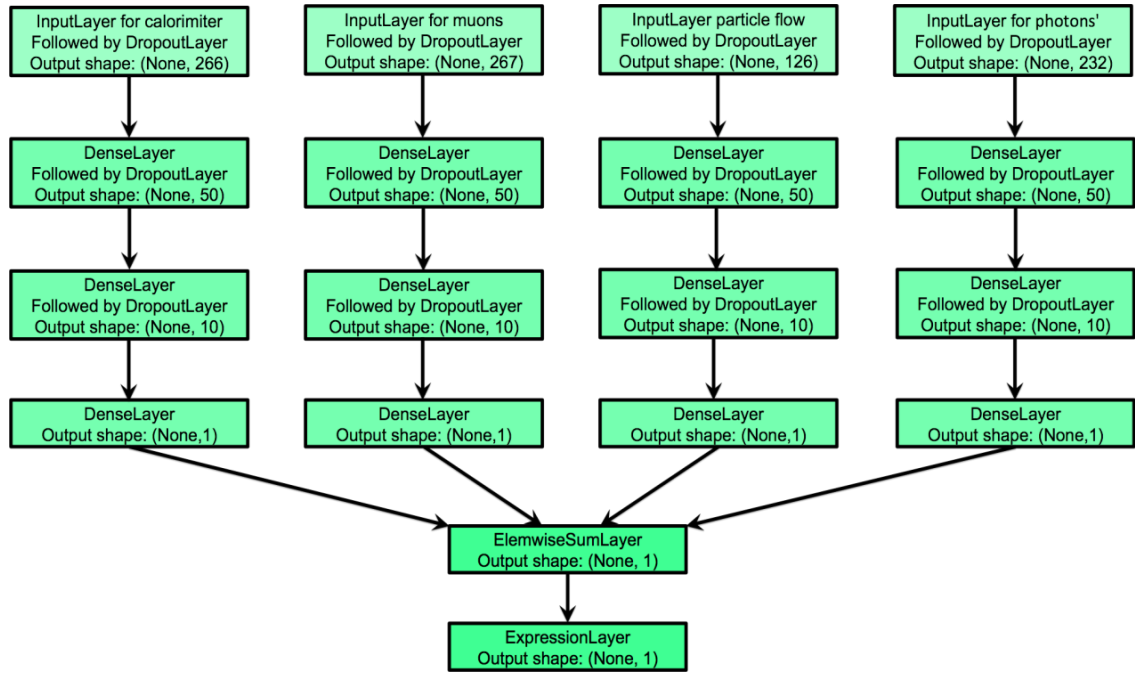
#monitoring
#predictive
#anomaly mgmt
#assistive int

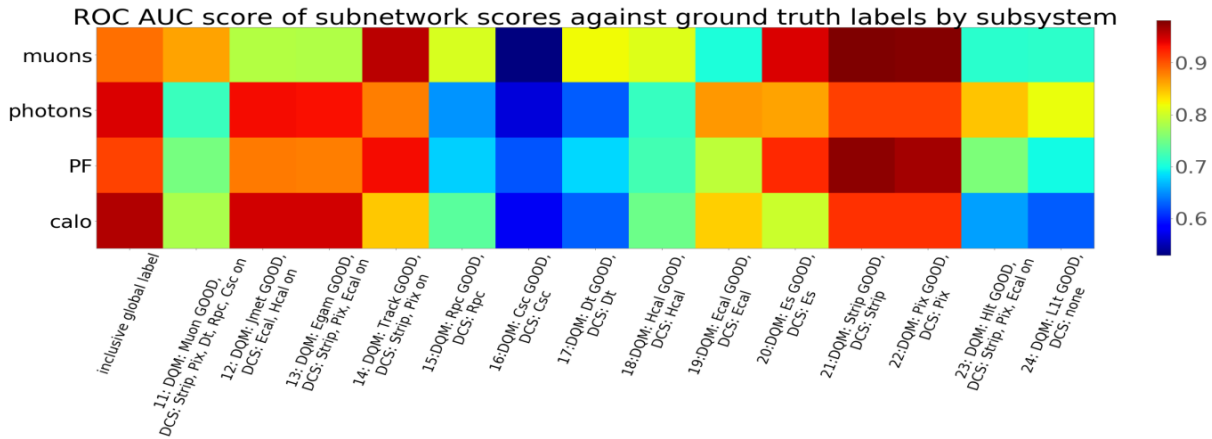
~80% saving on manual work is feasible for Pollution Rate at 5‰ and zero Loss Rate

2017

results

- Developed semi-supervised algorithm for identifying channel anomalies using just general labels.
- Started working on 2016 (much more realistic dataset)





ML HEP, summer school, invitation

2015 – Saint Petersburg, Russia

2016 – Lund, Sweden

2017 – Reading, UK

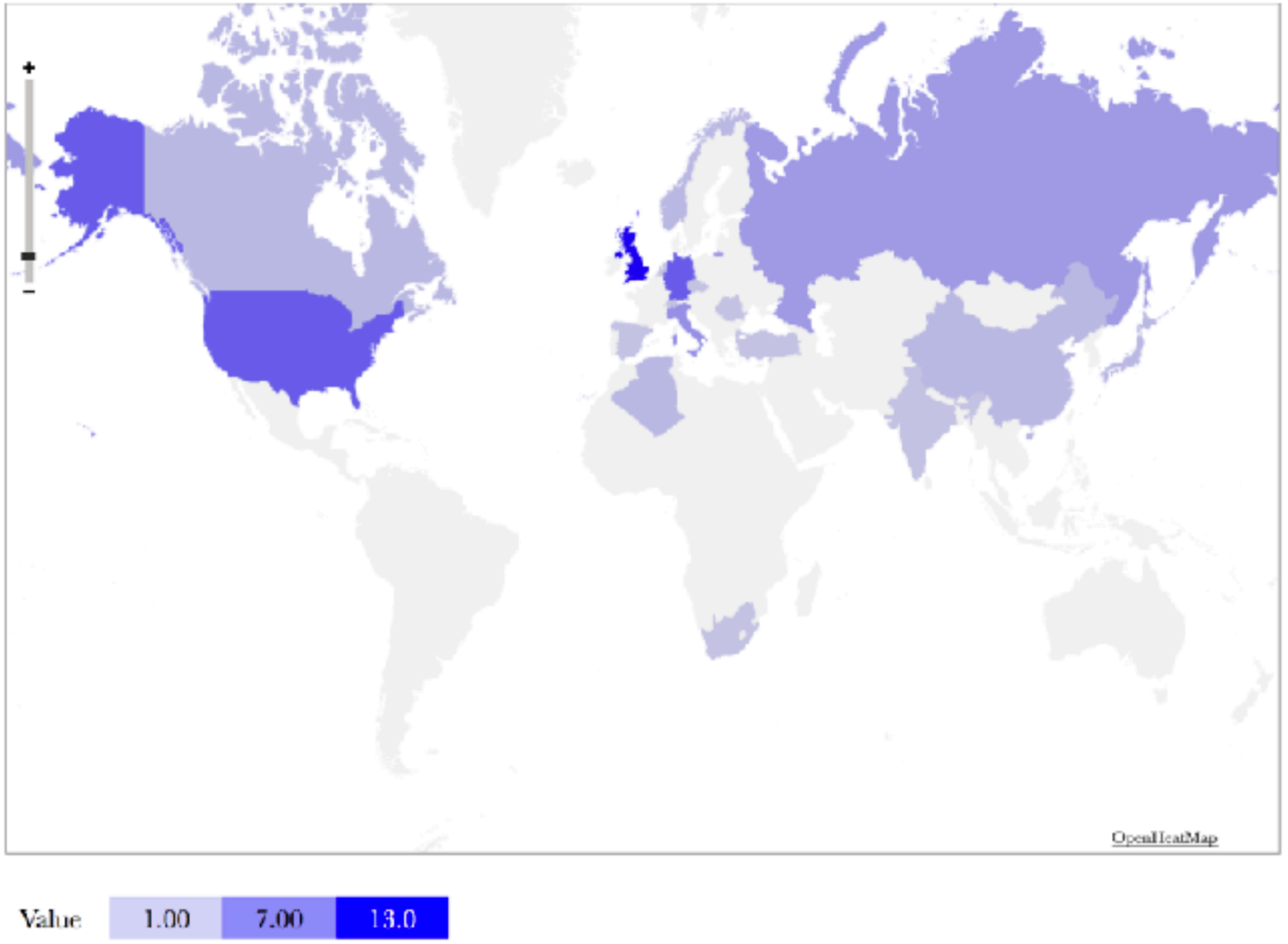
- 60 participants from 18 countries, 47 universities
- ML basics, Deep Learning, GANs, GP
- Invited speakers from academia and industry

Data Challenge

https://bit.ly/mlhep2017

2018 – Oxford, UK (TBC)

- **speakers?**
- **sponsors?**



QUESTIONS?

Andrey Ustyuzhanin
@anaderiRu