Yandex Data Factory

SCHOOL OF DATA ANALYSIS

Welcome to the school «Machine Learning in High Energy Physics»

Andrey Ustyuzhanin

27th August 2015

Organizers, a short intro

- **Yandex (est. 1994)**
  - World-wide search engine, leading position (~60%) in Russia
  - CERN openlab partner since 2013
- **Yandex School of Data Analysis (est. 2007)**
  - member of CERN LHCb & SHiP collaborations since 2014
- **Yandex Data Factory (est. 2014)**
- **Higher School of Economics**
  - Computer Science faculty (founded by Yandex)
  - LAMBDA (**LA**boratory of **M**ethods for **B**ig **D**ata **A**nalysis)

HEP Challenges

- Online event selection (10s TB/sec), data storage optimization
- Automatic event reconstruction
  - Reconstruction of tracks from hits, or higher-level properties
  - Semi-supervised algorithms
  - Parallelized execution (GPU, Xeon Phi, etc)
- Anomaly detection & prediction
- Machine Learning for hypothesis testing
- Systematic error estimation for parametrized prediction models
- Fast event simulation
- High luminosity LHC «era»: 100x increase of data flow in 2025

Machine Learning recent advances

- Algorithm Ensembling
- Deep learning
- Feature extraction
- Representation Learning
- Transfer learning
- Clustering
- Outlier detection
- Collaborative filtering
- …

Why MLHEP?

- Machine Learning: powerful approaches for complex tasks
- HEP: lots of challenges which could be solved with ML
- Increase people expertise in ML → advances in the field

Open Science

ML tools

- TMVA citations: ~750
- scikit-learn citations: ~1760
- ...

Open data

- CERN CMS, EMBL, ESA, ...

Research Reproducibility

- gitxiv.com, codalab.com, openml.org, binder, everware, github.com/yandex/rep, ...

MLHEP focus

- Bring variety of ML methods & tools to HEP projects in systematic way
  - Offline data analysis
  - Trigger optimization
  - Also: Data popularity, Anomaly detection & prediction
- Foster communication between HEP & ML communities
- Find possibilities for new joint projects
- Improve science (both HEP & ML)
- Research Reproducibility

Infrastructure

- MLHEP cloud
  - 8 machine x 16 CPU cores x 16 RAM
  - Docker containers for participants
  - authenticated by github.com
  - Ubuntu with all necessary libraries (yandex/rep-mlhep2015)
  - https://rep-mlhep2015.mlhep.yandex.net/
  - mlhep2015@yandex.ru

MLHEP highlights

Experimental to certain degree

- isolated campus, dogs after midnight
- 2 tracks: Separation is tentative
- Timetable may change! stay tuned
  https://indico.cern.ch/export/event/439520.ics

School-wide competition

Get help:

- mlhep2015@, https://gitter.im/yandexdataschool/mlhep2015

All feedback is welcomed

- twitter (#mlhep2015), email, whatever!

http://fc04.deviantart.net/fs39/f/2008/328/9/3/Beware_of_Dog_by_BrianGriffinFan.jpg

Upcoming Events

- YSDA conference on ML applications (Oct'15)
  https://yandexdataschool.com/conference/
- Data Science at LHC (Nov'15)
- ALEPH workshop at NIPS (Dec'15)
  http://yandexdataschool.github.io/aleph2015/
- Data Science at LHC (Mar'16)

Instead of Conclusion

- Machine Learning: powerful approaches for complex tasks
- HEP: lots of challenges which could be solved with ML
- Increase people expertise in ML → advances in the field
- Interdisciplinary research - points of growth
- We are open to joint projects (YSDA, YDF, LAMBDA)

Dr. Andrey Ustyuzhanin

head of CERN-Yandex
R&D group

twitter @AnaderiRu

anaderi@yandex-team.ru

Thank you

Yandex
Data Factory

SCHOOL OF DATA ANALYSIS