PROCEEDINGS OF SCIENCE

# TrackML : a tracking Machine Learning challenge

**Tobias Golling***
*Département de physique nucléaire et corpusculaire, Université de Genève, Genève, Switzerland*
*E-mail:* Tobias.Golling@unige.ch

**Sabrina Amrouche, Moritz Kiehn**
*Département de physique nucléaire et corpusculaire, Université de Genève, Genève, Switzerland*
*E-mail:* moritz.kiehn@unige.ch, sabrina.amrouche@cern.ch

**Paolo Calafiura, Steven Farrell, Heather Gray**
*Physics Division, Lawrence Berkeley National Laboratory and University of California, Berkeley CA, USA*
*E-mail:* pcalafiura@lbl.gov, sfarrell@lbl.gov, heather.gray@cern.ch

**Victor Estrade, Cécile Germain**
*LRI/TAU, Université Paris-Sud/INRIA/CNRS, Université Paris-Saclay, Gif-sur-Yvette, France*
*E-mail:* victor.antoine.estrade@gmail.com, cecile.germain91@gmail.com

**Vava Gligorov**
*LPNHE, Sorbonne Université, Paris Diderot Sorbonne Paris Cité, CNRS/IN2P3, Paris, France*
*E-mail:* Vladimir.Gligorov@cern.ch

**Isabelle Guyon**
*UPSud/INRIA Université Paris-Saclay, Orsay, France*
*ChaLearn, California, USA*
*E-mail:* guyon@chalearn.org

**Mikhail Hushchyn, Andrey Ustyuzhanin**
*National Research University Higher School of Economics, Moscow, Russia*
*Yandex School of Data Analysis, Moscow, Russia*
*E-mail:* mikhail91@yandex-team.ru, andrey.u@gmail.com

**Vincenzo Innocente, Andreas Salzburger**
*CERN, Geneva, Switzerland*
*E-mail:* vincenzo.innocente@gmail.com, Andreas.Salzburger@cern.ch

**Edward Moyse**
*Department of Physics, University of Massachusetts, Amherst MA, U.S.A.*
*E-mail:* edward.moyse@gmail.com

**David Rousseau, Yetkin Yilmaz**
*LAL, Université Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France*
*E-mail:* rousseau@lal.in2p3.fr, yilmaz@lal.in2p3.fr

**Jean-Roch Vlimant**





PoS(ICHEP2018)159

*California Institute of Technology, Pasadena CA, USA*
*E-mail:* vlimant@cern.ch

The High-Luminosity LHC will see pileup levels reaching 200, which will greatly increase the complexity of the tracking component of the event reconstruction. To reach out to Computer Science specialists, a Tracking Machine Learning challenge (TrackML) was set up on Kaggle in 2018 by a team of ATLAS, CMS and LHCb physicists, tracking experts and Computer Scientists, building on the experience of the successful Higgs Machine Learning challenge in 2014. A dataset consisting of an accurate simulation of a LHC experiment tracker has been created, listing for each event the measured 3D points, and the list of 3D points associated to a true track. The data set is large to allow for appropriate training of Machine Learning methods: about 100.000 events, 1 billion tracks, 100 GigaByte. The participants of the challenge are asked to find the tracks, which means to build the list of 3D points belonging to each track (deriving the track parameters is not the topic of the challenge). Here the first lessons from the challenge are discussed, including the initial analysis of submitted results.



*The 39th International Conference on High Energy Physics (ICHEP2018)*
*4-11 July, 2018*
*Seoul, Korea*

*Speaker.

## 1. The TrackML setup

To attain its ultimate discovery goals, the Large Hadron Collider (LHC) at CERN will increase its luminosity in the High-Luminosity LHC phase, so that the amount of additional collisions will reach a level of 200 interactions per bunch crossing, a factor 7 w.r.t. the current (2018) luminosity. This will be a challenge for the ATLAS and CMS experiments, in particular for track reconstruction algorithms. The goal is to deal with the increased combinatorial complexity without increasing the required computing budget. To engage the Computer Science community to contribute new ideas, a Tracking Machine Learning challenge (TrackML) [1] was organized running on the Kaggle platform from March to June 2018, building on the experience of the successful Higgs Machine Learning challenge in 2014. The data were generated using ACTS [2], an open source accurate tracking simulator, featuring a typical all-silicon LHC tracking detector, with 10 layers of cylinders and disks.Simulated physics events (Pythia $t\bar{t}$) overlaid with 200 additional collisions yield typically 10'000 tracks (100'000 hits) per event. The detector material is approximated by uniform cylinders and disks and a realistic non-homogenous solenoid field is simulated. The luminous region has a gaussian width of 5.5 mm in $z$ and 15 $\mu$m in the transverse direction, and 15% random hits are simulated in addition. The task to be performed by participants in the challenge is the pattern recognition step, associating the hits to tracks corresponding to the original charged particles. The participants are given 100'000 events (including truth information) to train their algorithm, while the evaluation by Kaggle is run on 100 other events yielding per-mille precision of the score. This score is used to rank the candidates, and it is based on the fraction of hits correctly assigned, with a weighting mechanism to favour higher momentum tracks and hits in the innermost and outermost detector layers.In this challenge, there is no CPU constraint, however a second phase of the challenge to be run later in 2018 will have strong computational constraints. Current track reconstruction algorithms use a combinatorial Kalman filter to connect 3D points into tracks. The trajectories of charged particles are deterministic, except for multiple scattering, energy loss and hadronic interaction, i.e. they are described as non-perfect helices pointing approximately to the origin. The emphasis of the challenge is to explore innovative Machine Learning approaches, rather than hyper-optimising known combinatorial approaches. In preliminary discussions with the ML community, Convolutional Neural Network, LSTM, Deep Neural Nets, Monte Carlo Tree Search, geometric Deep Learning have been mentioned. The particular challenges of this tracking challenge is an unusual dataset with a variable number of inputs and an unknown number of outputs. At the same time the available simulation and associated truth information allows for detailed analysis and verification of the proposed solutions.

## 2. Preliminary results

Simple reference solutions were provided by the starting kit of the challenge, in particular a reference notebook called DBSCAN, which consists of a few lines of preprocessing and calling sklearn DBSCAN clusterer yielding the non-trivial score of 20%. The score was quickly brought to 50%. Discussions on the forum helped to bring the score further up yielding best submission between about 70 and 90%. The results show no dependence on the charge of charged particles nor on its $\phi$ direction. Typically the efficiency drops for low $p_T$. As expected the track efficiency

for secondary displaced particles is low. Figure 1 compares the track reconstruction efficiency as a function of the $\Delta R$ distance to the next track for various submissions. The best solutions are characterised by showing no dependence on $\Delta R$. However, variations in track efficiency as a function of $\eta$ are observed.

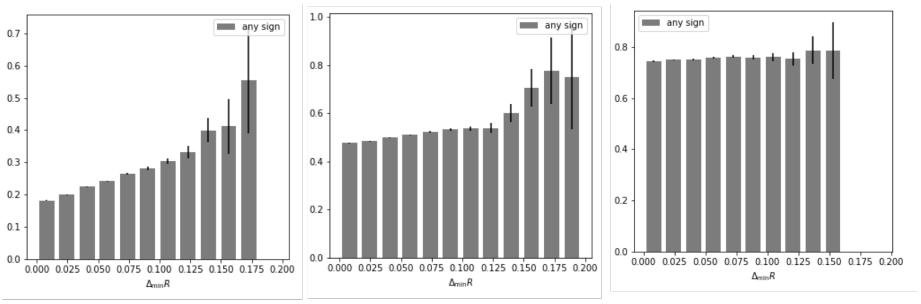


**Figure 1:** Track reconstruction efficiency as a function of the $\Delta R$ distance to the next track for DBSCAN (left), an intermediate solution of 50% (middle) and for one of the best submissions of close to 80% (right).

## 3. Conclusion

Overall the TrackML challenge was very well received by the Kaggle community with high levels of participation. The high complexity of TrackML and its associated physics domain knowledge as well as the large size of the dataset posed problems for some participants. Discussions on the forum were crucial to overcome these initial thresholds. DBSCAN was provided in the starting kit with a starting score of 20%, which quickly improved to 50% and then to about 70 to 90% for the best solutions. No clear deficiencies were identified with the best solutions. As expected towards the end of the challenge the pace of progression slowed down and less submissions were made as well as less information was shared.

## Acknowledgement

The team would like to thank CERN for allowing the use of the dataset, and Kaggle for hosting it. We are very grateful to our generous sponsors without which the challenges would not have been possible. Platinum sponsors: Kaggle, Nvidia and Université de Genève. Gold sponsors: Chalearn and DataIA. Silver sponsors: CERN Openlab, Paris-Saclay CDS, INRIA, ERC mPP, ERC RECEPT, Common Ground, Université Paris-Sud, INQNET, Fermilab and pyTorch. TG acknowledges the support of the Swiss National Science Foundation under the grant 200020_181984. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement No 724777 “RECEPT”, No 772369 “mPP” and No 654168 “AIDA-2020”.

## References

[1] *TrackML*, https://www.kaggle.com/c/trackml-particle-identification

[2] *A Common Tracking Software (ACTS)*, http://acts.web.cern.ch/ACTS/latest/doc/index.html