WCCI 2018 TrackML Particle Tracking Challenge

David Rousseau, Sabrina Amrouche, Paolo Calafiura, Victor Estrade, Steven Farrell, Cecile Germain, Vladimir Gligorov, Tobias Golling, Heather Gray, Isabelle Guyon, et al.

WCCI 2018 *TrackML Particle Tracking* Challenge

David Rousseau*[1], Sabrina Amrouche[2], Paolo Calafiura[3], Steven Farrell[3], Cécile Germain[4], Vladimir Vava Gligorov[5], Tobias Golling[2], Heather Gray[3], Isabelle Guyon[4], Mikhail Hushchyn[6], Vincenzo Innocente[7], Moritz Kiehn[2], Andreas Salzburger[7], Andrey Ustyuzhanin[8], Jean-Roch Vlimant[9], and Yetkin Yilmaz[1]

[1]LAL, Orsay
[2]University of Geneva
[3]Lawrence Berkeley National Laboratory
[4]UPSud, INRIA, University Paris-Saclay
[5]LPNHE-Paris
[6]Yandex, MIPT
[7]CERN
[8]Yandex, HSE
[9]CalTech

1 Competition Outline

1.1 Goal and contributions to computational intelligence and society

1.1.1 Motivation

The proposed challenge refers to recognizing trajectories in the 3D images of proton collisions at the Large Hadron Collider (LHC) at CERN. Think of this as the picture of a fireworks: the time information is lost, but all particle trajectories have roughly the same origin and therefore there is a correspondence between arc length and time ordering. Given the coordinates of the impact of particles on detectors (3D points), the problem is to "connect the dots" or rather the points, i.e. return all sets of points belonging to alleged particle trajectories. From the machine learning point of view, the problem can be treated as:

*trackml.contact@gmail.com

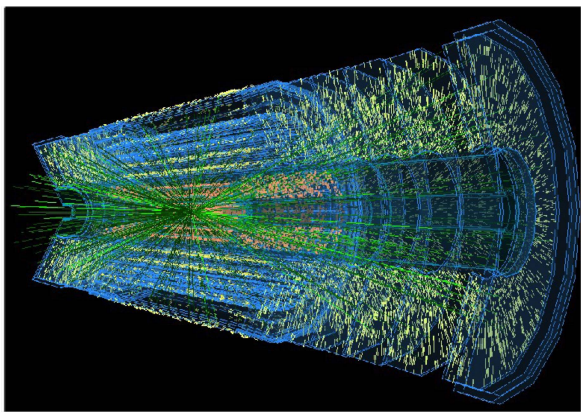

Figure 1: An etched-out high-multiplicity collision image in the future detector, measurements in yellow, trajectories are green.

- **A latent variable problem:** A data generating process first drew at random particles with given characteristics (momentum, charge, mass), then drew points along a trajectory originating near a collision focus (with uncertainties including diffusion/scattering and imperfections of detectors). "Particle memberships" are the latent variables to be inferred. This is similar to a clustering problem.

- **A tracking problem:** Using the correspondence between arc length and time ordering, one can treat the trajectories as time series and use tracking techniques, including Kalman filters.

- **A pattern de-noising problem:** Considering the collision snapshot as a 3D image, through the data acquisition process, the original trajectory lines were degraded into dotted lines with just a dozen points per line (the human eye cannot see the lines); the problem can therefore be thought of as signal enhancement of an "in-painting" problem (filling in missing data).

Therefore this challenge offers an interesting new puzzle to the Computational Intelligence community, while addressing pressing needs of the Physics community.

Indeed, current methods employed for tracking particles in the LHC experiments at CERN will be soon outdated: By 2025, there will be a major upgrade of the LHC to fulfill its rich physics program: understanding the characteristics of the Higgs boson, searching for the elusive dark matter, or elucidating the dominance of matter over anti-matter in the observable Universe. The number of proton collisions will be increased 10-fold progressively until 2025 so that the number of particles per proton bunch collision will also increase from about 1000 to 10,000. In addition, the ATLAS and CMS experiments plan a 10-fold increase of the readout rate. The explosion in combinatorial complexity is mainly due to the increase of the probability of confusion between tracks. It will have to be dealt with with a flat budget at best. The projection of CPU computing power gain with the already highly optimized production software leaves at least a 10-fold gap.

The HEP (High Energy Physics) experiments have embraced Machine Learning, originally for supervised classification as a routine tool in the final analysis stage, and in the past few years for exploring more diverse applications. The preliminary attempts of applying Machine Learning to particle physics pattern recognition-tracking indicate a strong potential [1]. Considering the success of the Higgs Boson ML Challenge [2], the HEP-ML collaboration for this challenge can be expected to produce high impact results. The algorithms exposed during the challenge, if promising, will be reused within the LHC experiments. To facilitate it a second phase of the challenge (not part of this proposal) will be run a few months later, with a similar metrics coupled this time to a strong CPU speed incentive.

1.1.2 Prior work

The field of particle tracking is well developed and there is in particular a yearly conference called "Connecting The Dots" (the 2017 edition [1] at LAL-Orsay was organised by one of us) where experts share new techniques on the problem. While early methods included mathematical transformations such as the Hough transform, the methods offering the best speed/accuracy tradeoff have concentrated on variants of Kalman filters in recent years, combined with various local pattern recognition methods. For an in depth review of the pre-Machine Learning state of the art, see [3].

In a TrackMLRamp hackathon we recently organized [4], proposing a simplified and smaller 2D version of the problem, several promising machine learning and neural network solutions have emerged, including LSTM (Long Short-Term Memory). Optimization methods such as MCTS (Monte Carlo Tree Search) were also successfully used. We see many more opportunities for the computational intelligence community to contribute, as the problem relates to representation learning [5] as in [6], to combinatorial optimization as in [7], neural-network based clustering [8], and even to time series prediction [9](even though the time information is lost, it can safely be assumed that particles were coming from the center of the detector and have successively crossed the nested layers of the detector). A possible approach is to efficiently exploit the a priori knowledge about geometrical constraints [10],

[1] https://ctdwit2017.lal.in2p3.fr

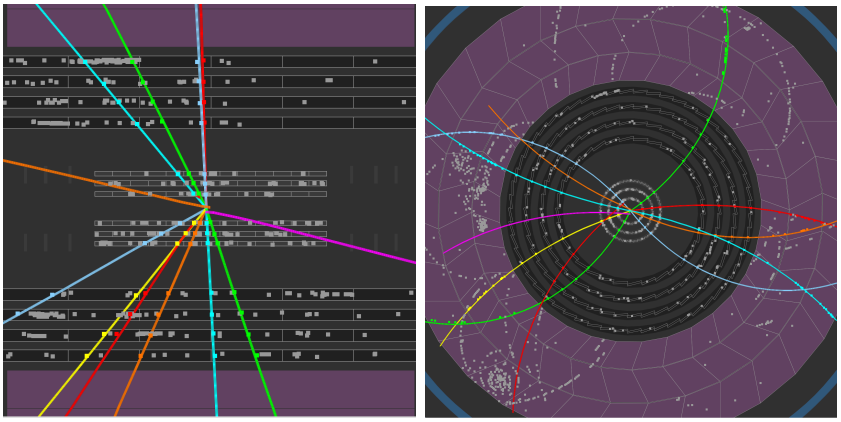

Figure 2: Projection of the tracks in the longitudinal and transverse planes, for low multiplicity events in the current detector

with recent work in the generative, *e.g.* [11] and [12] as well as discriminative approaches *e.g.* [13] for combining structural priors and nonlinear state estimation with neural network.

1.2 Challenge organization

We propose a challenge that aims at exercising the latest research advances from the pattern recognition, and more generally machine learning, community in devising fast and accurate particle tracking algorithms. The methods will be evaluated on a very large dataset simulated with a realistic simulator anticipating the new LHC architecture to be deployed by 2025. Thus the ground truth of particle trajectories will be known. The data created for the challenge will representative of the real HEP experimental scenarios.

1.2.1 The task

The participants should associate 3D points together to form tracks. While the task can be formally stated as a clustering problem, the ratio between the number of clusters ($\sim$ 10K) and their size ($\sim$ 10 points), is highly unusual, and drastically limits the performance of off the shelf clustering algorithms. Typically, at least 90% of the true tracks should be recovered.

The tracks are slightly distorted arc of helices with axes parallel to a fixed direction, and pointing approximately to the interaction center. On figure 2, the arcs appear as lines on the longitudinal projection and circles on the transverse one. Robustness with respect to these distortions and approximate pointing are enforced by the metric and are a de facto requirement. This task correspond to the first step of particle physics tracking, which is to attribute the points to the track they associated to the same true track, bent in a magnetic field. Further steps, like deriving the parameters of the track trajectory given the 3D points, is not part of the competition.

1.2.2 Protocol

The first phase the "Accuracy phase" of the challenge (which is the object of this proposal) will be run as a typical Kaggle challenge where the participants do not upload software but *solution files*. As usual, the dataset is partitioned into training, public test and private test. The challenge platform will be Kaggle (the Kaggle competition site is being set up, but is not yet public).

A solution is a list of associated points, each association being an assumed track, with an arbitrary unique numbering. The participant will develop their code (without any restriction on the language or libraries), and train their models on their own computing resources. They will apply their resulting evaluation algorithm to a test file (with hidden ground truth) by uploading the solution (list of points associated together) to the challenge platform. The challenge platform will compute a score value of the metric for the test sets and display the score on the public test set on a leaderboard. The final

ranking will be based on the private score only in order to avoid overfitting the public test set. It will not be disclosed until the end of the challenge. The precise numbers of leaderboard submissions per day and final submissions will be finalized with the platform provider.

The competition is organized in two tracks. Participants competing for a prize in any of the two tracks will be requested to release their software and to self-assess the CPU usage for both training and testing.

- The *Performance* track will be solely based on the previously described metric in section 1.3.2.
- The *Algorithm* track will be based on a jury (with experts in particle physics tracking algorithms and machine learning) which will select the submission with the most promising balance between the score, speed and originality.

A second phase of the challenge called the "Throughput" phase will be run from mid-June 2017 to October 2017. This second phase will be focused on the evaluation speed of the algorithms exposed during the first phase (the training speed will not be constrained), while maintaining a similar accuracy. The speed will be evaluated by the challenge platform. This second phase is not part of the proposal; however we count on our participation to WCCI to advertise the second phase.

1.2.3 Schedule

The challenge will run from February 1st to May 15th. Fact sheets describing the algorithms to be submitted by May 27th. The results will be announced early June.

The second phase of the challenge (not part of this proposal), focused on speed, will be organised during summer 2018.

A final workshop will be organised at CERN in spring 2019, where winners of both phases of the challenge will be invited.

1.2.4 Web site

A very preliminary website is `https://sites.google.com/site/trackmlparticle/` (the Kaggle site is not public yet.

1.3 Data, evaluation and toolkits

1.3.1 Data

A dataset consisting of a simulation of a typical full Silicon LHC detector will be made available, listing for each event the measured 3D points coordinates, and the list of 3D points associated to each true track (ground truth), en event corresponding to the tracks of one collision. The simulation engine uses the ACTS[2] [14] simulator both fast (1s per event) and accurate. Realistic collisions yielding 10.000 tracks per event have been simulated with a sufficient level of details to make the task almost as difficult as for real events: points are measured with a precision of approximately 50 microns, some tracks are grouped in dense "jets" (increasing the possibility of confusion), multiple scattering distorts the tracks, points are some times missing, some tracks stops early.

The data set is large in order to allow the training of data intensive Machine Learning methods. The orders of magnitude are : 10^5 events, with each 10^4 tracks, for a total dataset size of 100 GBytes. The events are independent. The public and private evaluation datasets need to be much smaller, about 100 events (100 MBytes), but are large enough to evaluate the metrics within a per mille of statistical uncertainty.

The dataset has been generated for the purpose of the challenge, and can be publicly released; all copyrights and privacy of experimental data and software have been respected.

[2] https://gitlab.cern.ch/acts/acts-core

1.3.2 Evaluation

A perfect algorithm will uniquely associate correctly each points to the track it belongs to. An imperfect algorithm will miss some tracks altogether, miss one or more points for an otherwise valid track, associate wrong points to an otherwise valid track, find tracks from random association of points, find multiple variants of the same track.

The metric used for the challenge is a weighted version of the *adjusted Rand index* clustering metric [15, 16]. The overall metric is the average over the events. The same metric was used during the TrackMLRamp hackathon[4] already mentioned and was demonstrated to be both robust and consistent with different metrics used in the field.

1.3.3 Helpers

A set of helpers have been designed in order to facilitate accessing the challenge. They will be made available on the challenge platform from the first day:

Two light events (with just 100 tracks instead of 10.000) will allow first steps and easier development of auxiliary tools.

1.3.4 Starting kit

The starting kit includes iPython notebooks implementing a complete workflow, various python procedures for accessing the data as well as creating and evaluating solutions and the baseline algorithms. A 3D web based visualization tool is under development.

Two baselines will be provided.

- DBScan: this clustering algorithm demonstrates non-trivial performance, although far from the requested ones. The main goal is to provide a simple method to demonstrate the workflow.
- Hough transform; where the 3D hit space is mapped onto a track parameter space, where maxima (corresponding to tracks) are found, and then moved back to the original 3D space to associate the points. This technique has a linear complexity; however it does not allow to reach the maximum efficiency.

1.4 Visibility

1.4.1 Anticipated number of participants

Based on the previous HEP-ML challenges that we have organised, and CERN implication, we anticipate more than 500 of participating teams, between 500 and 1000.

1.4.2 Dissemination

It is foreseen to advertise the competition through the channels associated with the HEP-ML collaboration channels (see next section). The promotion of the challenge will also be relayed through CERN and LHC experiments social media, with hundred of thousands followers.

1.5 Brief bio of proposers

The team organising this competition has experience with the HiggsML challenge in 2014 on Kaggle[3], which, with close to 2000 participants, was the most successful Kaggle competition at the time, as well as with the Flavour of Physics challenge in 2015[4], with close to 800 participants.

Furthermore, many collaboration and workshops between Machine Learning and High Energy Physics have taken place meanwhile, for example DS@LHC at CERN in October 2015[5], Heavy Flavour

[3] `https://www.kaggle.com/c/higgs-boson`

[4] `https://www.kaggle.com/c/flavours-of-physics`

[5] `http://indico.cern.ch/event/395374/`

Data Mining Workshop at Zürich 2016[6], DS@HEP at Fermilab[7], Hammers and Nails 2017[8], as well as Connecting The Dots series[9]. Most of them have been co-organised by members of the team.

The team is mostly composed of particle physics tracking experts working on the ATLAS (SA,PC,SF,TG,HG,MK,DR,AS), CMS (VI,JRV) and LHCb (VG,AU) experiments at CERN, and machine learning specialists (CG,IG,AU,MH).

Due to the large number of co-organizers, we give only three representative bios.

- **Isabelle Guyon** is chaired professor in Data Science at the Université Paris-Saclay, member of the TAU INRIA team, specialized in statistical data analysis, pattern recognition and machine learning. She is one of the co-founders of the ChaLearn Looking at People (LAP) challenge series. She co-organised the HiggsML challenge and she pioneered applications of the Microsoft Kinect to gesture recognition. Her areas of expertise include computer vision and and bioinformatics. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She worked on early applications of Convolutional Neural Networks (CNN) to handwriting recognition in the 1990s. She is also the primary inventor of SVM-RFE, a variable selection technique based on SVM. The SVM-RFE paper has thousands of citations and is often used as a reference method against which new feature selection methods are benchmarked. She also authored a seminal paper on feature selection that received thousands of citations. She organized many challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. Isabelle Guyon holds a Ph.D. degree in Physical Sciences of the University Pierre and Marie Curie, Paris, France. She is president of Chalearn, a non-profit dedicated to organizing challenges, vice-president of the Unipen foundation, adjunct professor at New-York University, action editor of the Journal of Machine Learning Research, editor of the Challenges in Machine Learning book series of Microtome, program chair of the NIPS 2016 conference, and general chair of the NIPS 2017 conference. Isabelle Guyon has contributed several competitions to IJCNN/WCCI, including:
  - WCCI 2006 Performance Prediction challenge: `http://www.modelselect.inf.ethz.ch/`
  - IJCNN 2007 Agnostic Learning vs. Prior Knowledge challenge: `http://www.agnostic.inf.ethz.ch/`
  - WCCI 2008 Causality challenge: `http://www.causality.inf.ethz.ch/challenge.php`
  - WCCI 2010 Active Learning challenge: `http://www.causality.inf.ethz.ch/activelearning.php`
  - IJCNN 2011 Unsupervised and Transfer Learning challenge: `http://www.causality.inf.ethz.ch/unsupervised-learning.php`
  - IJCNN 2013 Cause-effect pair challenge: `http://www.causality.inf.ethz.ch/cause-effect.php`
  - WCCI 2014 Neural Connectomics challenge `http://connectomics.chalearn.org/`
  - IJCNN 2015-16 AutoML challenge: `https://www.codalab.org/AutoML`
  - IJCNN 2017 ML Explainability challenge: `http://chalearnlap.cvc.uab.es/challenge/23/description/`
- **David Rousseau** is a senior physicist at LAL-Orsay, a joint CNRS/IN2P3, Université Paris-Sud/Paris-Saclay laboratory, in the ATLAS LHC collaboration at CERN. He gained is PhD in

[6] `https://indico.cern.ch/event/433556/`

[7] `https://indico.fnal.gov/event/13497/`

[8] `https://www.weizmann.ac.il/conferences/SRitp/Summer2017/`

[9] `https://indico.physics.lbl.gov/indico/event/149/`, `https://indico.hephy.oeaw.ac.at/event/86/`, `https://ctdwit2017.lal.in2p3.fr`

Particle Physics from the Université Aix-Marseille II (France) in 1992; his PhD analysis was among the very firsts making in depth use of the precise tracking made possible by the silicon detectors (a budding technology at the time) on the ALEPH experiment at CERN. He joined the ATLAS experiment in 1998, in the R&D team developing the tracking detector. From 2000 to 2010, he has been responsible for the development of the reconstruction algorithms suite (including tracking), of the experiment, and from 2010 to 2012 of all the offline software. After the Higgs boson discovery in July 2012 by the ATLAS and CMS experiments, he's switched to the promotion of advanced Machine Learning techniques in High Energy Physics. He co-organised the HiggsML challenge in 2014 and a number of HEP and ML workshops as indicated above. He is currently co-coordinator of the ATLAS Machine Learning group.

- **Andreas Salzburger** Andreas Salzburger, senior CERN staff member and member of the ATLAS experiment since 2001. He gained his Particle Physics PhD from university of Innsbruck (Austria) in 2018 with a thesis devoted to track simulation and reconstruction in the ATLAS Experiment. He has been author and architect of core ATLAS track reconstruction and Monte Carlo Simulation software. He researches on software and algorithmic solutions for fast and precise pattern recognition modules, using classical combinatorial and data science inspired approaches; tracking detectors, their design and performance aspects for physics analyses in high energy physics. He is lecturer on particle physics, track reconstruction and simulation techniques at Innsbruck university and international graduate schools. Former tracking performance and event reconstruction group leader at the ATLAS experiment, he is currently ACTS (A common Tracking Software) project leader and ATLAS upgrade software coordinator, at the forefront of the algorithm developments for the future ATLAS upgrades. He has overseen the production of the dataset for the challenge.

2 Competition category

We propose a **Category C** competition.

3 Sponsorship

Academic sponsorships by Université Paris-Saclay and by CERN Openlab have been secured, which should be enough to cover the basic organisation cost. Application for industry sponsorship have been or are being submitted in parallel to Azure and Amazon WS, as well as to partners of CERN Openlab Intel, Nvidia and IBM. This additional sponsorship will allow us to propose significant prize money, and to sponsor training resources for participants.

References

[1] S. Farrell *et al.*, "The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking," *EPJ Web Conf.*, vol. 150, p. 00003, 2017.

[2] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs boson machine learning challenge." in *HEPML@ NIPS*, 2014, pp. 19–55. [Online]. Available: http://www.jmlr.org/proceedings/papers/v42/cowa14.pdf

[3] T. Cornelissen, M. Elsing, S. Fleischmann, W. Liebig, and E. Moyse, "Concepts, Design and Implementation of the ATLAS New Tracking (NEWT), ATL-SOFT-PUB-2007-007," 2007.

[4] S. Amrouche *et al.*, "Track reconstruction at LHC as a collaborative data challenge use case with RAMP," *EPJ Web Conf.*, vol. 150, p. 00015, 2017.

[5] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *26th NIPS*, 2013, pp. 809–817.

[7] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, "Neural combinatorial optimization with reinforcement learning," 2016. [Online]. Available: https://arxiv.org/abs/1611.09940

[8] G. Aad *et al.*, "A neural network clustering algorithm for the ATLAS silicon pixel detector," *Journal of Instrumentation*, vol. 9, no. 09, 2014.

[9] P. Ondruska and I. Posner, "Deep tracking: Seeing beyond seeing using recurrent neural networks," in *13th AAAI Conference on Artificial Intelligence*, 2016, pp. 3361–3368.

[10] R. Urtasun, "Incorporating structure in deep learning," in *4th ICLR*, 2016, keynote. [Online]. Available: http://videolectures.net/iclr2016_urtasun_incoporating_structure/

[11] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep kalman filters," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1511.05121

[12] M. Johnson *et al.*, "Composing graphical models with neural networks for structured representations and fast inference," in *29th NIPS*, 2016.

[13] T. Haarnoja *et al.*, "Backprop kf: Learning discriminative deterministic state estimators," in *29th NIPS*, 2016, pp. 4376–4384.

[14] C. Gumpert, A. Salzburger, M. Kiehn, J. Hrdinka, and N. Calace, "ACTS: from ATLAS software towards a common track reconstruction software," *J. Phys. Conf. Ser.*, vol. 898, no. 4, p. 042011, 2017.

[15] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[16] M. Meilă, "Comparing clusterings—an information based distance," *J. Multivar. Anal.*, vol. 98, no. 5, pp. 873–895, 2007.