Connecting the dots in high-energy physics

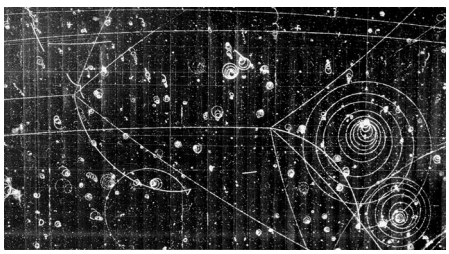
Rebuilding particle trajectories from high-energy proton collisions is an essential step in processing the petabytes of data generated by the Large Hadron Collider at CERN. In search of an order of magnitude speed-up, physicists reached out to the computer science community.

The discovery of the Higgs boson was announced at CERN by the ATLAS and CMS experiments in the summer of 2012. For a decade, I had worked with up to 200 people to develop the software that reduced the petabytes of data from the ATLAS detector into this single bit of information. I was back in my home lab in Orsay when I met a machine learning researcher at the cafeteria. Machine learning was unknown to me at the time, as it was to most particle physicists. By 2014, however, we had organized the HiggsML challenge on Kaggle, with a record-setting 2,000 participants.

While presenting the results of the HiggsML challenge in the aptly named 'Connecting The Dots' workshop at Berkeley, I ended my talk with a few additional slides: what if we were to organize a particle physics tracking challenge? So, what is 'tracking' in particle physics? Well, protons accelerated by the LHC collide to produce a firework of elementary particles. Unlike flares in a real firework, which can be followed by eye, particles are measured by silicon detectors at discrete points. Using these measurements, tracking algorithms need to reconstruct the original trajectories. Current tracking algorithms use a large portion of the computing resources devoted to processing LHC data. As such, this is an important problem, with a richness of possible approaches. Importantly, this problem could also be framed as 'just' a 3D pattern recognition problem.

We started by devising the challenge dataset; we wanted to use simulations in order to have an accurate ground truth. However, existing simulations had many detailed features that obscured the big picture. We decided instead to use ACTS, an open source simulator. This allowed us to make the problem simpler, but not so simple that it became uninteresting.

Then onto evaluation: there are stacks of papers and PhD dissertations on tracking algorithms that are full of plots and tables. But we needed to evaluate a proposed tracking algorithm with just one single number (plus the speed). We finally settled on a very unusual (for us) evaluation score: we match all the proposed tracks to the ground truth ones, spot which points are correctly assigned and which are not, and define the score to be the overall fraction of correctly assigned points. This is reminiscent of the intersection-over-union criterion commonly used for evaluating pattern recognition algorithms.



Bubble chamber images like these could be interpreted by eye to detect particles moving through it, but modern detectors produce increasingly complex images with particle trajectories that are only deciphered by using well-designed algorithms.

By this time, we realized that the challenge, named TrackML, was going to be much more complex than HiggsML, where any off-the-shelf classifier was a reasonable starting point. And not only did we want new algorithms, we wanted them to be fast! To alleviate this difficulty, the challenge was split into two phases: the first 'accuracy' phase would focus on the quality of the algorithm, the second 'throughput' phase would have an additional strong speed incentive.

The accuracy phase ran on Kaggle from May to August 2018[1]. The final leader board shows a peloton of participants preceded by well-detached frontrunners, which, as on a long mountain stage in the Tour de France, indicates that the competition was really difficult. Different algorithms were used, some with clever injections of machine learning, which we acknowledged with an in-kind NVIDIA V100 GPU and invitations to the 2018 Conference on Neural Information Processing Systems or CERN, in addition to monetary prizes for the first three. We studied submissions, producing many plots and tables. We were quite happy to see that algorithms with the best scores were also the ones with the best results according to these plots and tables, meaning that the participants could not 'hack' the evaluation score with algorithms that would turn out to be useless to us.

For the throughput phase, we used the CodaLab platform (managed by Chalearn and University Paris Saclay), which was configured to measure the speed of participants' Python or C++ software on dedicated servers, in addition to the accuracy score as above.

The throughput phase ran from September 2018 to March 2018, with much fewer participants. We suspect that we lost many people when they realized they would have to code in C++ for speed. However, the top three submissions are astonishingly good. Number three, in seven seconds, optimized the winning code of the accuracy phase by injecting new (for particle physicists) machine learning techniques: the key for speed in such a combinatorial problem is to drop branches of the exploration tree as early as possible. The winner and runner-up are in fact established particle physics tracking experts; they had solutions below one second, which is an order of magnitude faster than the current state-of-the-art running on admittedly more complex simulations (needless to say, they didn't have access to any insider information).

So, the story ends with a twist: a scientific community designs a challenge to reach out to computer science, but the competition is won by their own experts. Was it worth it? Absolutely, for three reasons: (i) the diverse machine learning techniques we were exposed to are now on the table, ready to be scrutinized by the community; (ii) by their own admission, the experts enjoyed competing in a lightweight environment with a well-defined single score; and (iii) the TrackML dataset (being released on the CERN Open Data Portal) will serve as a future benchmark. It has in fact already been used to explore the use of quantum computing for tracking. ❐

David Rousseau
Laboratoire de l'Accélérateur Linéaire, Univ. Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France.
e-mail: rousseau@lal.in2p3.fr

Published online: 11 June 2019
https://doi.org/10.1038/s42256-019-0061-0

References

1. Amrouche, S. et al. Preprint at https://arxiv.org/abs/1904.06778 (2019).