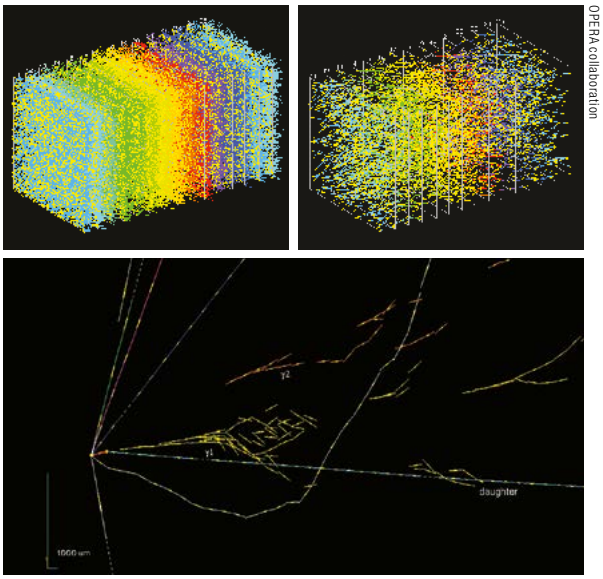
Learning machine learning

The Yandex machine-learning school for high-energy physics is teaming up with experiments at CERN and beyond to train young researchers in the arts of deep learning.



OPERA collaboration

Electromagnetic shower identification for the OPERA experiment, showing a detector element filled with background tracks (top left), the same volume after pre-filtering by OPERA's tracking algorithm (top right), and finally the shower revealed after even more thorough filtering (bottom).

Machine learning, whereby the ability of a computer to perform an intelligent task progressively improves, has penetrated many scientific domains. It allows researchers to tackle problems from a completely new perspective, enabling improvements to things previously thought solved for good. The downside of machine learning is that the field itself is developing so quickly, with new techniques popping up at an incredible rate, that it is hard to keep up. What is needed is some sort of high-level trigger to discriminate between good and bad, and to guide a growing community of users in a systematic way.

Machine-learning techniques are already in wide use in particle physics, and they will only become more prevalent during the coming years of the high-luminosity LHC and future colliders. Online data processing, offline data analysis, fast Monte Carlo generation techniques and detector-upgrade optimisation are just a few examples of the areas that could profit significantly from smarter algorithms (see p31).

The most remarkable growth trend in machine learning today, and one that has also been heavily hyped, concerns so-called deep learning. Although there is no strict boundary, a neural network with less than four layers is considered "shallow", while one with more than 10 layers and many thousands of connections is considered "deep". Using deep-learning algorithms, plus performative computing resources and extremely large datasets, researchers have managed to break important barriers for such tasks as text translation, voice recognition, image segmentation and even to master the game Go. Many of the educational materials one can find on the Internet are thus focused around typical tasks such as image recognition, annotation, segmentation, text processing and pattern generation.

Since most of these are conveyed in computer-science language, there is an obvious language barrier for domain-specific scientists, such as particle physicists, who have to learn a new technique and apply it to their own research. Another complication is that there are a variety of machine-learning methods capable of solving particular problems and a plenitude of tools (i.e. different languages, packages and platforms) out there – almost all of which are online – with which to implement those methods.

Targeting particle physics

As machine learning spreads into new domains such as astrophysics or biology, schools that focus on problems in specific areas are becoming more popular. Historically there are several summer schools for particle physicists focused around data analysis, computing and statistical learning – in particular the CERN School for computing, INFN School of Statistics and the CMS Data Analysis School. But, until 2014, none focused specifically on machine learning. In that year, a series with the straightforward title Machine-Learning school for High-Energy Physics (MLHEP) was launched.

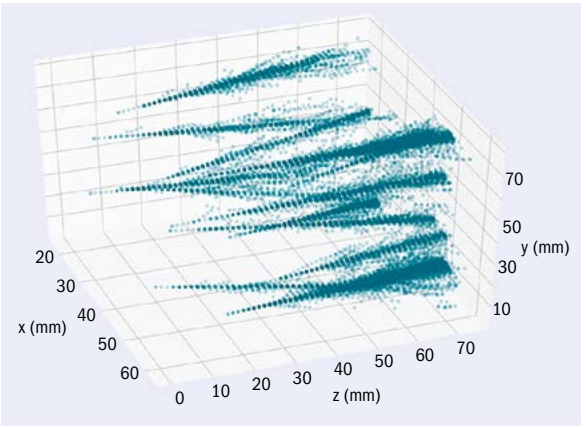
MLHEP grew out of the well-established Yandex School of Data Analysis (YSDA), a non-commercial educational organisation funded by the Russia-based internet firm Yandex. Over the past decade, YSDA has grown to receive several thousand applications per year, out of which around 200 people pass the entrance exams and around 50 graduate in conjunction with leading Russian universities – almost all of them finding data-science positions in the private sector.

In 2015, YSDA joined the LHCb collaboration. The goal was to help optimise LHCb's high-level trigger system to improve its efficiency for selecting B-decay events, and the result of the LHCb-YSDA collaboration was an efficiency gain of up to 60% compared to that obtained during LHC Run I. Another early joint effort between YSDA, CERN and MIT within LHCb was the design of decision-tree algorithms capable of decorrelating their output from

a given variable, such as invariant mass.

The first MLHEP schools in 2015 and 2016 were satellite events at the Large Hadron Collider Physics (LHCP) conference held in St. Petersburg and Lund, respectively. Another key contributor to the school was the faculty of computer science at Russia's Higher School of Economics (HSE), which was founded in 2014 by Yandex. MLHEP 2017 was organised by Imperial College London in the UK, and the 2018 school takes place in Oxford at the beginning of August.

The topics covered during the schools usually start from the basic aspects of machine learning, such as loss functions, optimisation methods, predictive-model quality validation, and stretch towards advanced techniques like generative adversarial networks and Bayesian optimisation. The curriculum is not static, and each year the focus changes to address the most interesting and promising trends in deep learning while providing an overview of various techniques available on the market. At the 2018 school, speakers were invited from both academia and from companies, including Oracle, Nvidia, Yandex and DeepMind.



Simulated overlapping electromagnetic showers in the proposed SHiP detector, which participants of MLHEP-2017 had to recognise against dense background settings using machine-learning methods.

Breaking the language barrier

Some people compare deep learning not with a tool or platform, but with a language that allows a researcher to express computational "sentences" addressing a particular problem.

To reinforce the language analogy, recall that there is no solid theory of deep learning yet; in a sense it is just a bunch of best-practices and approaches that has proven to work in several important cases. A lot of the time during MLHEP classes is therefore devoted to practical exercises. School students are also encouraged to enter a data-science competition that is related to particle physics – e.g. tracking for the Coherent Muon to Electron Transition (COMET) experiment and event selection for the Higgs-boson discovery by the ATLAS and CMS experiments. The competition is published on the machine-learning competition platform kaggle.com at the start of the school, and is open for anyone who wants to get more machine-learning practice.

For summer 2017, the competition was organised together with the OPERA and SHiP collaborations. The goal was to analyse volumes of nuclear emulsions collected by OPERA that contain lots of cosmic-background tracks as well as tracks from electromagnetic showers. These shower-like structures are of interest for OPERA for the analysis of tau-neutrino interactions, so special algorithms have to be developed. Such algorithms are also very relevant to the SHiP experiment, which aims to use emulsion-based detectors for finding hidden-sector particles at CERN. According to some theoretical models, such showers might be closely related to hidden-sector particle interaction with regular matter (e.g. elastic scattering of very weakly-interacting particles off electrons or nuclei), so a separate task would be to discriminate these showers from neutrino interactions. The performance of the algorithms designed by participants was amazing. The winner of the challenge presented his solution at the SHiP collaboration meeting in November 2017 and was invited by OPERA to continue the collaboration.

A major part of the MLHEP curriculum is given by YSDA/HSE lecturers, and guest speakers help to broaden the view on the machine-learning challenges and methods. The school is non-commercial, and its success depends on external contributions from the HSE, YSDA, local organisers and commercial sponsors. For the past two years we have been supported by the Marie Skłodowska-Curie training network AMVA4NewPhysics, which has also sent several PhD students to the school.

The format of the summer school is very productive, allowing students to dive into the topics without distraction. The school materials also remain available at GitHub, allowing students to access them whenever they want. As time goes by, basic machine-learning courses are becoming more readily available online, giving us a chance to introduce more advanced topics every year and to keep up with the rapid developments in this field.

Further reading

bit.ly/mlhep2018.

A Rogozhnikov *et al.* 2014 arXiv:1410.4140.

Résumé

Se former sur l'apprentissage automatique

L'apprentissage automatique permet aux scientifiques de s'attaquer à des problèmes depuis une perspective entièrement nouvelle, et d'approfondir des questions que l'on croyait jusqu'ici résolues pour de bon. Le revers de la médaille est que le domaine lui-même se développe tellement rapidement, avec des nouvelles techniques qui apparaissent à un rythme effréné, qu'il est difficile de suivre son évolution. Au sein de l'apprentissage automatique, le sous-domaine présentant aujourd'hui la croissance la plus remarquable est l'apprentissage approfondi, qui a par ailleurs été très médiatisé. Afin de former des jeunes chercheurs à l'art de l'apprentissage approfondi, l'école d'apprentissage automatique pour la physique des hautes énergies de Yandex s'associe à des expériences au CERN et dans d'autres laboratoires.

Andrey Ustyuzhanin, YSDA, HSE, ICL.