

Berkeley lab cosmologists are top contenders in machine learning challenge

March 20 2020



2020 LHC Olympics logo. Credit: Gregor Kasieczka, Benjamin Nachman, David Shih

In searching for new particles, physicists can lean on theoretical predictions that suggest some good places to look and some good ways to find them: It's like being handed a rough sketch of a needle hidden in a



haystack.

But blind searches are a lot more complicated, like hunting in a haystack without knowing what you are looking for.

To find what conventional computer algorithms and scientists may overlook in the huge volume of data collected in particle collider experiments, the particle physics community is turning to machine learning, an application of artificial intelligence that can teach itself to improve its searching skills as it sifts through a haystack of data.

In a machine learning challenge dubbed the 2020 Large Hadron Collider (LHC) Olympics, a team of cosmologists from the U.S. Department of Energy's Lawrence Berkeley National Laboratory (Berkeley Lab) developed a code that best identified a mock signal hidden in simulated particle-collision data.

Cosmologists? That's right.

"It was totally unexpected for us to perform so well," said George Stein, a Berkeley Lab and UC Berkeley postdoctoral researcher who participated in the challenge with Uros Seljak, a Berkeley Lab cosmologist, UC Berkeley professor, and co-director of the Berkeley Center for Cosmological Physics, of which Stein is a member.

Ten teams, composed mostly of particle physicists, competed in the competition, which ran from Nov. 19, 2019, to Jan. 12, 2020.

Stein led the adaptation of a code that two other student researchers had developed under Seljak's direction. The competition was launched by the organizers of the Machine Learning for Jets 2020 (ML4Jets2020) conference. Jets are narrow cones of particles produced in particlecollision experiments that particle physicists can trace back to measure



the properties of their particle sources.

The competition results were announced during the conference, which was held at New York University Jan. 15-17.

Ben Nachman, a Berkeley Lab postdoctoral researcher who is part of a group that works on ATLAS—a large detector at CERN's LHC—served as one of the event and contest organizers. David Shih, a physics and astronomy professor at Rutgers University now on a sabbatical at Berkeley Lab, and Gregor Kasieczka, a professor at the University of Hamburg in Germany, were co-organizers.

While some computing competitions allow participants to submit and test their codes multiple times to gauge whether they are getting closer to the correct results, the 2020 LHC Olympics competition gave teams just one shot to submit a solution.

"The cool thing is that we didn't use an off-the-shelf tool," Seljak said. "We used a tool that we had developed for our research."

He noted, "In my group we had been working on unsupervised machine learning. The idea is that you want to describe data where the data have no labels."

The tool that the team used is called sliced iterative optimal transport. "It's a form of deep learning, but a form where we do not optimize everything at once," Seljak said. "Instead, we do it iteratively," in stages.

The code is so efficient that it can run on a simple desktop or laptop computer. It was developed for a statistical approach known as Bayesian evidence.

Seljak said, "Suppose you are looking at anomalies in a planet's transit



time," the time it takes for the planet to pass in front of a larger object from your viewpoint—like watching from Earth as Mercury moves in front of the sun.

"One solution requires that there be an extra planet," he said, "and the other solution requires an extra moon, and they are both a good fit to the data, but they have very different parameters. How do I compare these two solutions?"

The Bayesian approach is to compute the evidence for both solutions and see which solution has a higher probability of being true.

"This kind of example comes up all of the time," Seljak said, and his team's code is designed to speed up the complex calculations required by conventional methods. "We were trying to improve upon something unrelated to particle physics, and we realized this could be used as a general machine learning tool."

He added, "Our solution is particularly useful for so-called anomaly detection: looking for very tiny signals in data that are somehow different than its other data."

In the 2020 LHC Olympics competition, participants first received a sample set of data that called out particle signal data from some background data—both the needle and the haystack—that allowed participants to test their codes.

Then they received the actual "black box" contest data: just the haystack. They were tasked to find a different and entirely unknown kind of particle signal hidden in the background data, and to specifically describe the signal events that their methods turned up.

Competition co-organizers Shih and Nachman noted that they had



personally been working on an anomaly-detection method that uses a very similar approach (called "conditional density estimation") to the technique developed by Seljak and Stein that was entered in the competition.

Seljak and Stein consulted with a number of particle physicists at the lab, including Nachman, Shih, and graduate student Patrick McCormack. They discussed, among other topics, how the high-energy physics community typically analyzes datasets like those used in the competition, but for the actual "black box" challenge Seljak and Stein were on their own.

As the competition was drawing toward a close, Stein said, "We thought we found something about a week before the deadline."

Stein and Seljak submitted their results a few days before the conference, "but as we are not particle physicists, we were not planning to participate at the conference," Seljak said.

Then, Stein received an email from the conference organizers, who asked him to fly out and present a talk on the team's solution later that week. The organizers didn't share the results of the competition until all of the speakers had presented their results.

"My talk was originally first, and then shortly before the start of the session they moved me to last. I didn't know if that was a good thing," Stein said.

The code that the Berkeley Lab team entered picked up about 1,000 events, with an error margin of plus or minus 200, and the correct response was 843 events. Their code was the clear winner in that category.



Several teams were close in estimating the energy level, or "resonance mass," of the signal, and the Berkeley Lab team was closest in its estimate of the resonance mass for a secondary signal stemming from the main signal.

At the conference, Stein noted, "There was a huge interest in the overall approach we took. It made waves."

Oz Amram, another competitor in the contest, quipped in a Twitter post, "The result of the LHC Olympics ... is that cosmologists are better at our job than we are." But contest organizers did not formally announce a winner.

Nachman, one of the event organizers, said, "Even though George and Uros clearly outperformed the other competitors, in the end it is likely that no one algorithm will cover every possibility—so we will need a diverse set of approaches to achieve broad sensitivity."

He added, "Particle physics has entered an interesting time where every prediction for <u>new particles</u> we have tested at the Large Hadron Collider has so far turned out to be not realized in nature—except the Standard Model of particle physics. While it is essential to continue the program of model-driven searches, we also have to develop a parallel program to be model-agnostic. That is the motivation for this challenge."

Seljak said that his team is planning to publish a paper that details its machine learning code.

"We are definitely planning to apply this to many astrophysics problems," he said. "We will look for interesting applications—anything with glitches or transients, anything anomalous. We will work to speed up the code and make it more powerful. These kinds of approaches can really help."



Provided by Lawrence Berkeley National Laboratory

Citation: Berkeley lab cosmologists are top contenders in machine learning challenge (2020, March 20) retrieved 27 April 2024 from <u>https://techxplore.com/news/2020-03-berkeley-lab-cosmologists-contenders-machine.html</u>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.