Can we trust scientific discoveries made using machine learning?

February 15 2019

Rice University statistician Genevera Allen will discuss research to improve the accuracy and reproducibility of scientific discoveries made by machine learning in both a press briefing and general session at the 2019 AAAS Annual Meeting. Credit:Tommy LaVergne/Rice University

Rice University statistician Genevera Allen says scientists must keep

questioning the accuracy and reproducibility of scientific discoveries made by machine-learning techniques until researchers develop new computational systems that can critique themselves.

Allen, associate professor of statistics, computer science and electrical and computer engineering at Rice and of pediatrics-neurology at Baylor College of Medicine, will address the topic in both a press briefing and a general session today at the 2019 Annual Meeting of the American Association for the Advancement of Science (AAAS).

"The question is, 'Can we really trust the discoveries that are currently being made using machine-learning techniques applied to large data sets?'" Allen said. "The answer in many situations is probably, 'Not without checking,' but work is underway on next-generation machine-learning systems that will assess the uncertainty and reproducibility of their predictions."

Machine learning (ML) is a branch of statistics and computer science concerned with building computational systems that learn from data rather than following explicit instructions. Allen said much attention in the ML field has focused on developing predictive models that allow ML to make predictions about future data based on its understanding of data it has studied.

"A lot of these techniques are designed to always make a prediction," she said. "They never come back with 'I don't know,' or 'I didn't discover anything,' because they aren't made to."

She said uncorroborated data-driven discoveries from recently published ML studies of cancer data are a good example.

"In precision medicine, it's important to find groups of patients that have genomically similar profiles so you can develop drug therapies that are

targeted to the specific genome for their disease," Allen said. "People have applied machine learning to genomic data from clinical cohorts to find groups, or clusters, of patients with similar genomic profiles.

"But there are cases where discoveries aren't reproducible; the clusters discovered in one study are completely different than the clusters found in another," she said. "Why? Because most machine-learning techniques today always say, 'I found a group.' Sometimes, it would be far more useful if they said, 'I think some of these are really grouped together, but I'm uncertain about these others.'"

Allen will discuss uncertainty and reproducibility of ML techniques for data-driven discoveries at a 10 a.m. press briefing today, and she will discuss case studies and research aimed at addressing uncertainty and reproducibility in the 3:30 p.m. general session, "Machine Learning and Statistics: Applications in Genomics and Computer Vision." Both sessions are at the Marriott Wardman Park Hotel.

Provided by Rice University