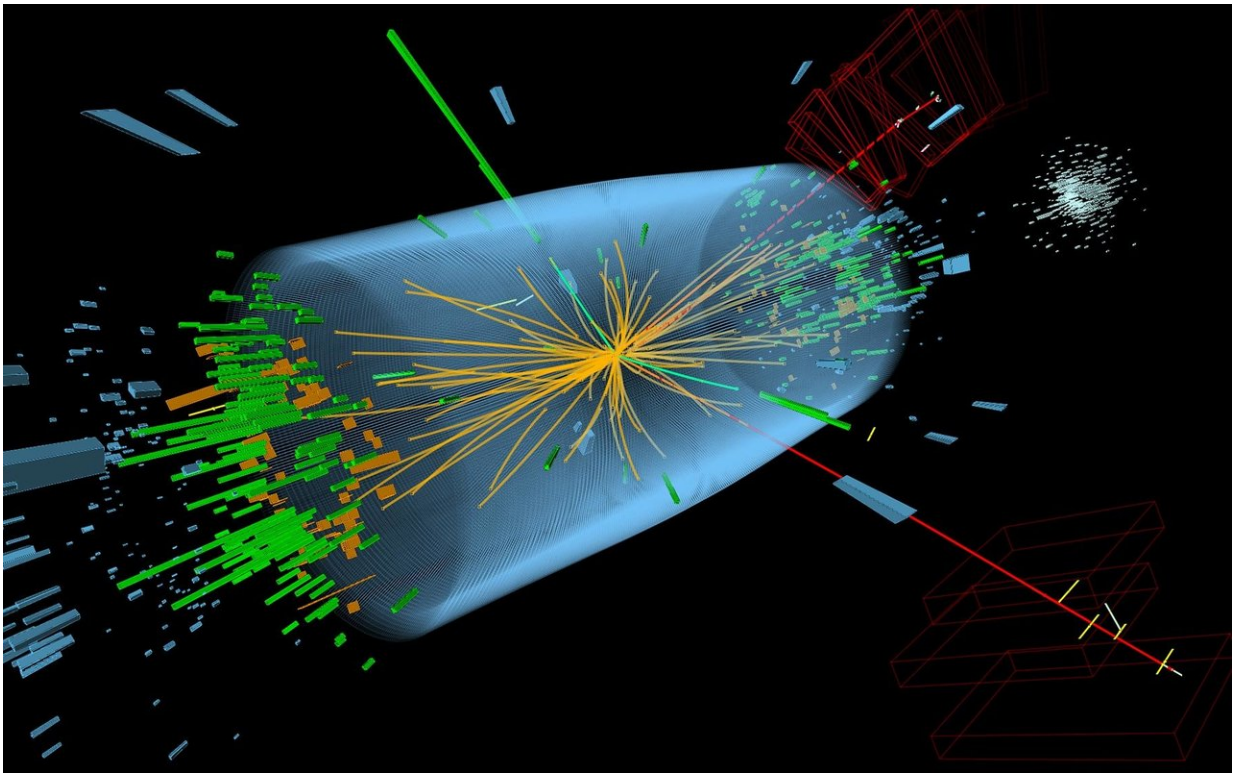
The big problem of small data: A new approach

October 18 2018



To demonstrate that DEFT can be applied to a variety of small datasets, CSHL scientists used it to analyze data from the CMS Higgs Boson detector. Of 60 particle impressions, DEFT estimated that up to six were from real events. (Pictured: A 3D perspective of a Higgs Boson event recorded in 2012. Impressions are characterized by green towers and red lines.) Credit: McCauley, T; Taylor, L; CERN

Big Data is all the rage today, but Small Data matters too! Drawing reliable conclusions from small datasets, like those from clinical trials for rare diseases or in studies of endangered species, remains one of the trickiest obstacles in statistics. Now, Cold Spring Harbor Laboratory (CSHL) researchers have developed a new way to analyze small data, one inspired by advanced methods in theoretical physics, but available as easy-to-use software.

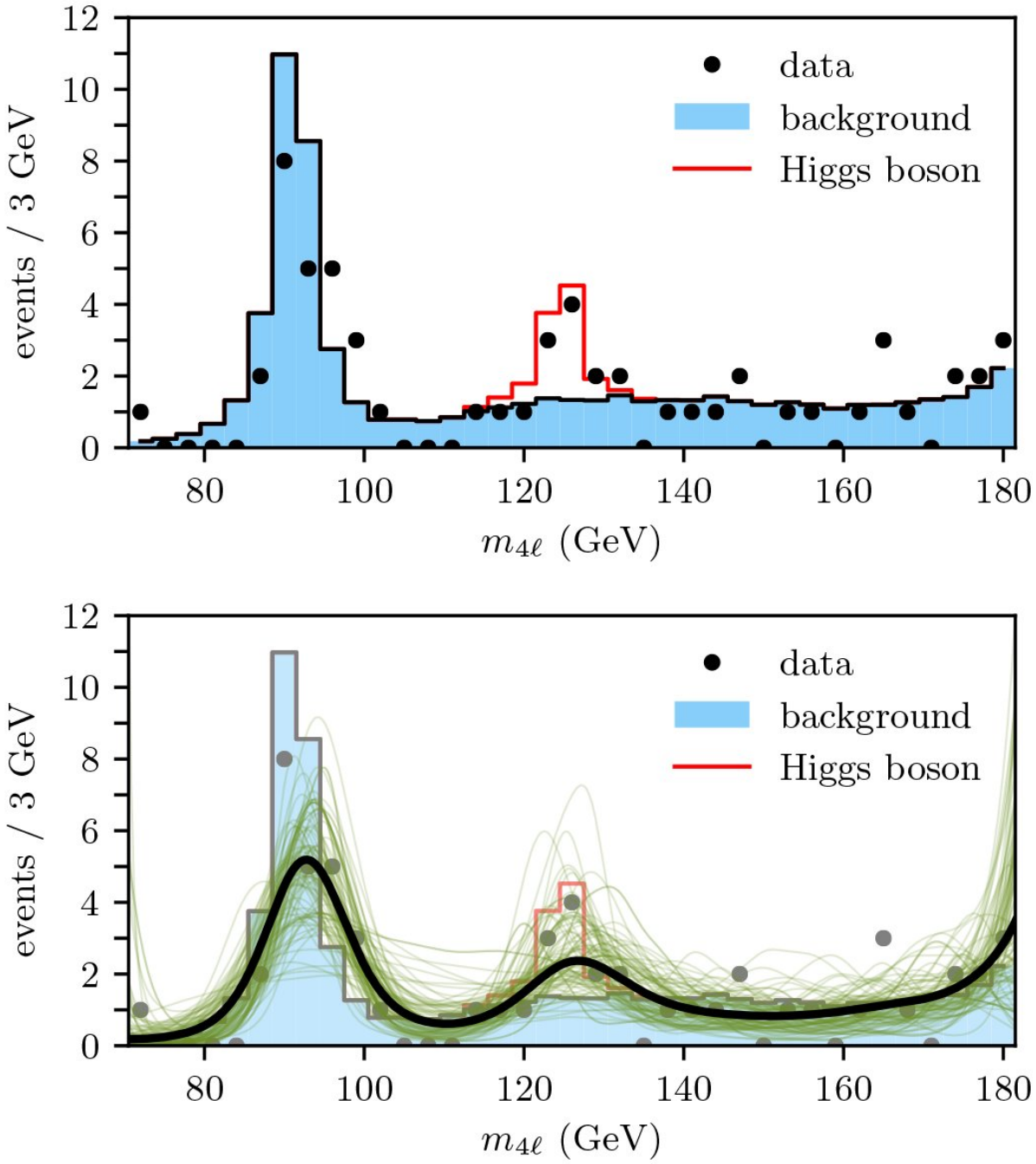
"Dealing with small datasets is a fundamental part of doing science," CSHL Assistant Professor Justin Kinney explained. The challenge is that, with very little data, it's not only hard to come to a conclusion; it's also hard to determine how certain your conclusions are.

"It's important to not only produce the best guess for what's going on, but also to say, 'This guess is probably correct,'" said Kinney.

A good example is clinical drug trials.

"When each data point is a patient, you will always be dealing with small datasets, and for very good reasons," he said. "You don't want to test a treatment on more people than you have to before determining if the drug is safe and effective. It's really important to be able to make these decisions with as little data as possible."

Quantifying that certainty has been difficult because of the assumptions that common statistical methods make. These assumptions were necessary back when standard methods were developed, before the computer age. But these approximations, Kinney notes, "can be catastrophic" on small datasets.



Top: Number of Higgs Boson particle events expected based on Standard Model simulations.

Bottom: DEFT was used to smoothly predict (black) how many 4-lepton decay events were indicators of a true Higgs Boson event within a margin of uncertainty (green). Credit: Kinney Lab/CSHL

Now, Kinney's lab has crafted a modern computational approach called Density Estimation using Field Theory, or DEFT, that fixes these shortcomings. DEFT is freely available via an open source package called SUFTware.

In their recent paper, published in *Physical Review Letters*, Kinney's lab demonstrates DEFT on two datasets: national health statistics compiled by the World Health Organization, and traces of subatomic particles used by physicists at the Large Hadron Collider to reveal the existence of the Higgs boson particle.

Kinney says that being able to apply DEFT to such drastically diverse "real-world" situations —despite its computations being inspired by theoretical physics—is what makes the new approach so powerful.

"Flexibility is a really good thing... We're now adapting DEFT to problems in survival analysis, the type of statistics used in clinical trials," Kinney said. "Those new capabilities are going to be added to SUFTware as we continue developing this new approach to statistics."

More information: SUFTware

Provided by Cold Spring Harbor Laboratory