Anomaly detection in 5 mins or less

Snowmass 2022 CompF3 ML Session

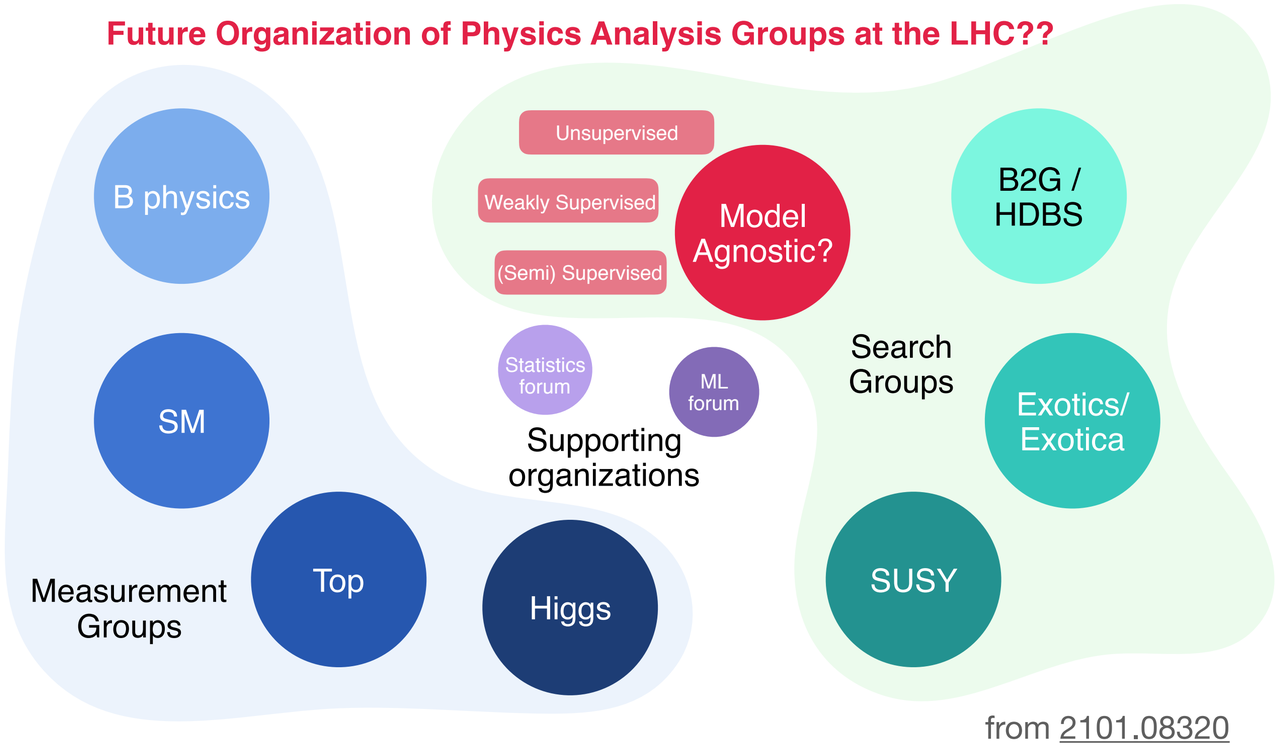
David Shih
July 19, 2022

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

ML for New Physics Searches

The vast majority of LHC searches for new physics are very model specific

ML for New Physics Searches



from 2101.08320

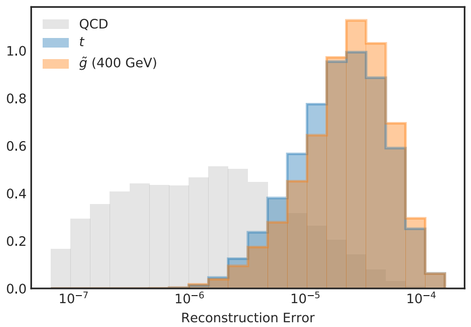
The vast majority of LHC searches for new physics are very model specific

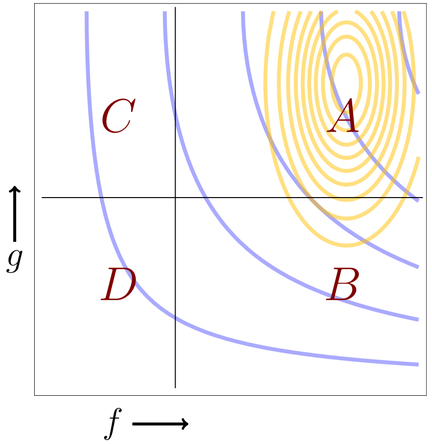
Why aren't there more model-agnostic new physics searches?

ML for Anomaly Detection

- How do we search for new physics in a model-agnostic way?
- Need two ingredients:
  1. Signal sensitivity (anomaly score)
  2. Background estimation

from 1808.08992





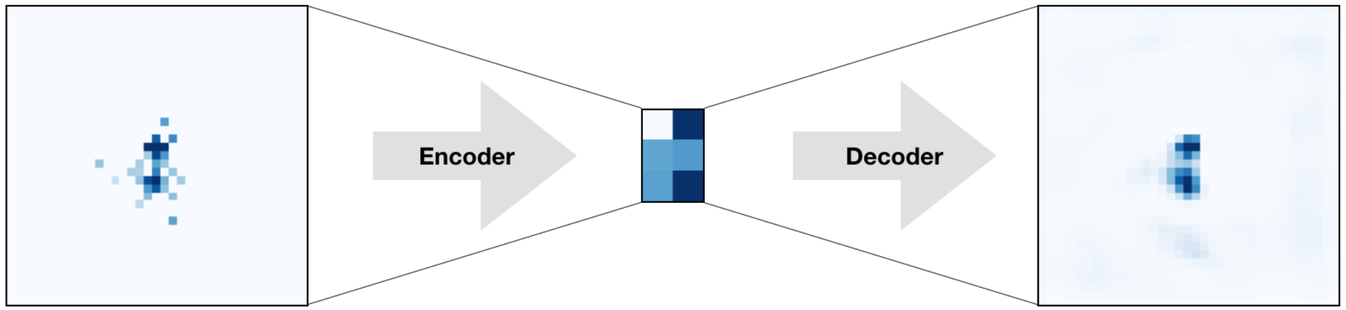
from 2007.14400

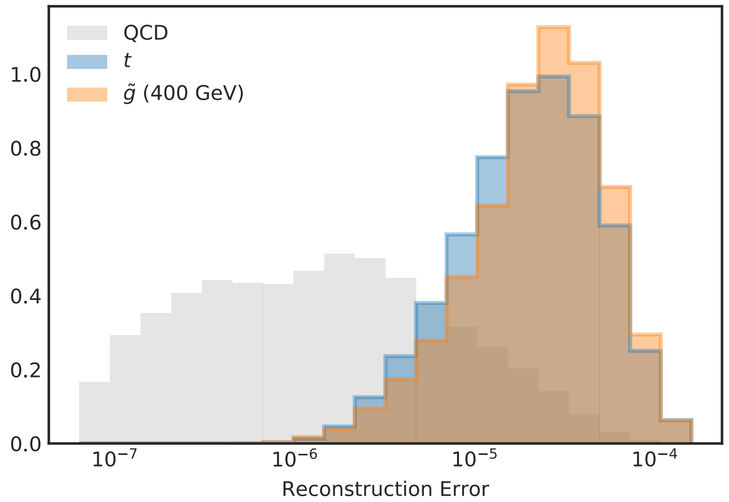
Both should be data-driven for a truly model-agnostic search

ML for Anomaly Detection

Types of Anomaly Scores

- Low p(x) — outliers





Autoencoders

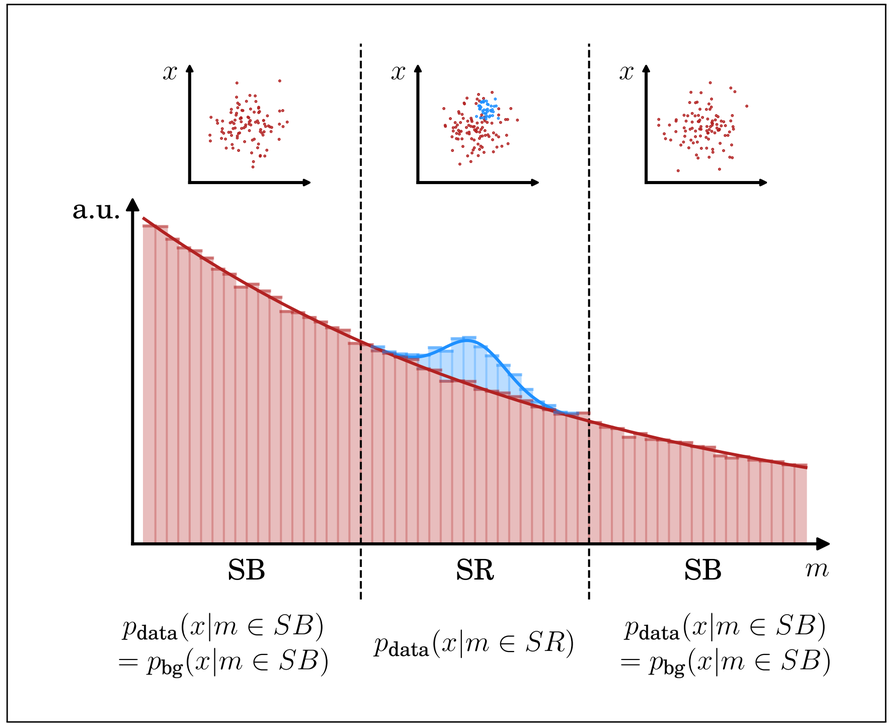
Fully unsupervised

Farina, Nakai & **DS** 1808.08992
Heimel et al 1808.08979
and many more!!

ML for Anomaly Detection

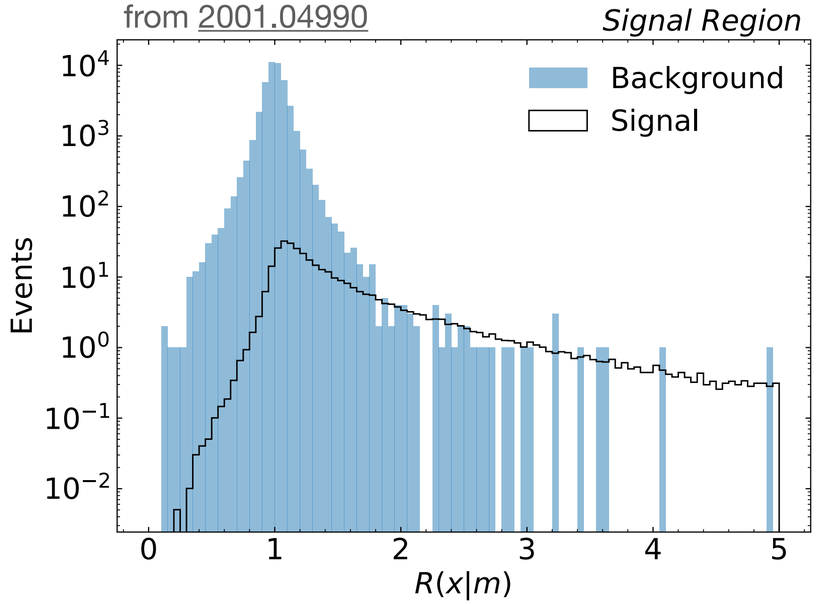
Types of Anomaly Scores

- High $p_{data}(x)/p_{bg}(x)$ — overdensities



from 2109.00546

from 2001.04990



Enhanced bump hunts

Weakly supervised

CWoLa Hunting [Collins, Howe & Nachman 1805.02664, 1902.02634]
ANODE [Nachman & **DS** 2001.04990]
CATHODE [Hallin et al 2109.00546]
CURTAINS [Raine et al 2203.09470]
and more…

LHC Olympics 2020

The LHC Olympics 2020

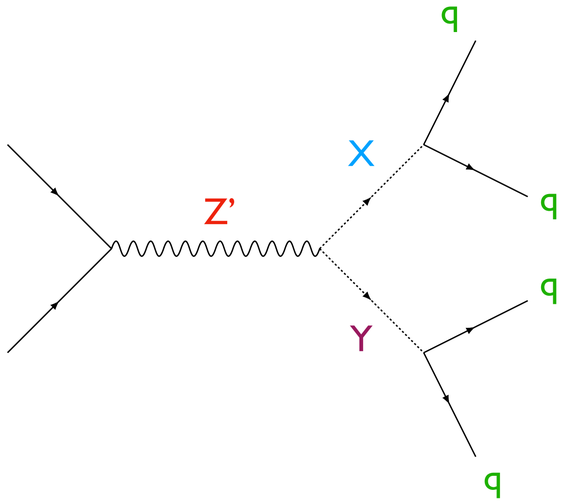
A Community Challenge for Anomaly Detection in High Energy Physics

Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

https://arxiv.org/abs/2101.08320

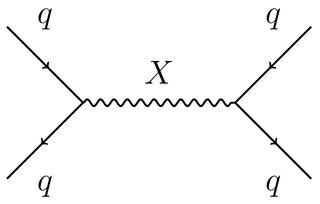
R&D dataset:

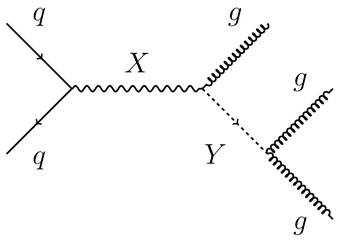
bg: 1M QCD dijet events
signal: up to 100k Z'->XY events
Pythia+Delphes
pT(J1)>1.2 TeV trigger



3 "Black Box" datasets

1. 1M QCD dijets + 834 Z'->XY signal (same topology as R&D, different masses)
2. No signal
3. QCD dijets + 3,000 Z' decaying to dijets or trijets





LHC Olympics 2020

3 “Black Box” datasets

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)
    - **Several successful methods! (based on autoencoders, CWoLa, density estimation…)**

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)
    - **Several successful methods! (based on autoencoders, CWoLa, density estimation…)**
2. No signal

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)
   - **Several successful methods! (based on autoencoders, CWoLa, density estimation…)**
2. No signal
   - **Some approaches found false positives — importance and challenges of background estimation!**

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)
   - **Several successful methods! (based on autoencoders, CWoLa, density estimation…)**
2. No signal
   - **Some approaches found false positives — importance and challenges of background estimation!**
3. QCD dijets + 3,000 Z’ decaying to dijets or trijets

LHC Olympics 2020

3 “Black Box” datasets

1. 1M QCD dijets + 834 Z’->XY signal (same topology as R&D, different masses)
   - **Several successful methods! (based on autoencoders, CWoLa, density estimation…)**
2. No signal
   - **Some approaches found false positives — importance and challenges of background estimation!**
3. QCD dijets + 3,000 Z’ decaying to dijets or trijets
   - **No approaches discovered the signal in BB3**

Outlook

- There is a lot of community interest in anomaly detection and model-agnostic NP searches!
- LHC Olympics 2020 was a very successful challenge, drawing nearly 50 participants from theory, experiment, and beyond (cosmology, computer science)
- Proofs-of-concept are beginning to be ported over to real data

CWoLa Hunting
ATLAS, PRL **125** 131801 (2020)

RNN VAE
ATLAS-CONF-2022-045

- **Many challenges for future R&D, including: feature selection, background estimation, multiple decay modes (BB3), non-resonant signals, …**