Collider Searches & Unsupervised Learning

Andrea De Simone
andrea.desimone@sissa.it

Amir Farbin
amir.farbin@cern.ch

Erzsébet Merényi
erzsebet@rice.edu

IDEA:

Employ/develop ML techniques to find
New Physics in LHC data (without specifying it)

GOAL:

find presence of New Physics in subset of LHC data, by detecting samples not falling in any “known” class (“never-seen-before” processes)

> Dataset

- Dataset is dominated by known processes (Standard Model '**background**').
  Interested in few **signal** events (1 in 10^6-10^{11})

- Features:
  - low-level: 'raw' 4-vectors for each particle in event
  - high-level: mass, transverse momentum, missing energy…

- Publicly available samples: Snowmass 2013, HepSim (bkg)

- Need to generate (MC) data for signal processes

- Data pre-processing:
  - compute features
  - convert to ML format (e.g. HDF5)

> 3 Complementary Approaches

- **Feature learning** (A. Farbin)

  Learn optimal high-level features (w/ autoencoders)

  raw ⟶ learned features ⟶ clustering/anomaly det./stat. tests

- **Statistical Tests of Distributions** (A. De Simone)

  Check compatibility of high-dim data vs simulated background

  (w/ Nearest Neighbors)

  high-level features ⟶ two-sample test of data vs bkg

  high-level features ⟶ characterize discrepant regions

- **Structures in Data** (E. Merényi)

  Detect structures in data with SOM-based clustering

  raw ⟶ Self-Organizing Maps (SOM)

  raw ⟶ learned features ⟶ SOM

  compare two-sample tests ⟷ SOM

> Implementation

- No challenge, we prefer a **project**

- Coordinators set up the guidelines and organize the participants.

- Split participants into 3 working groups.
  Each group works on the same full dataset to find anomalous data.

- Share results on google drive/slack.

- **Synergy** among groups to use each other’s results.