LHC Olympics 2020: Winter Games

Gregor Kasieczka
Hamburg

Benjamin Nachman
LBNL

David Shih
Rutgers / Berkeley / LBNL
Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature.

Why is the Higgs boson so light?

Hierarchy problem

Why do neutrons have no dipole moment?

Strong CP

See also: quantum gravity

Reality

>99% of pictures on the internet

image source: symmetry magazine
Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature.

What is the extra gravitational matter?

Dark Matter

Why do neutrinos have a mass?

Flavor puzzles

See also: dark energy

See also: Where did all the anti-particles go? (Baryogengesis)
Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature.

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model.

Three possibilities:

- Dark matter
- Hierarchy problem
- Baryogenesis
- Strong CP
- Dark energy
- Flavor puzzles
Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature.

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model.

Three possibilities

1. There is nothing new at LHC energies
Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at LHC energies
(2) Patience! (new physics is rare)
Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration of the fundamental structure of nature

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at LHC energies
(2) Patience! (new physics is rare)
(3) We are not looking in the right place
We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

This is the motivation for this challenge!

(3) We are not looking in the right place
Many of the deep questions in fundamental physics can be probed at the LHC.
Many of the deep questions in fundamental physics can be probed at the LHC.

Key challenges (and opportunity!)

Typical collision events at the LHC produce $O(1000+)$ particles

We detect these particles with $O(100\ M)$ readout channels
Searches at the LHC

Run: 302347
Event: 753275626
2016-06-18 18:41:48 CEST
The most powerful* discriminant between two simple hypothesis $H_1$ and $H_2$ is given by the likelihood ratio:

$$R(x) = \frac{\mathcal{L}(x|H_1)}{\mathcal{L}(x|H_2)}$$

**Examples**

- $H_1 =$ specific signal + SM background
  $H_2 =$ SM background
  *conventional, model-specific search strategy*

- $H_1 =$ “data”
  $H_2 =$ SM background
  *model-independent search strategy
  “anomaly detection”*

*This means that for a fixed probability of rejecting $H_1$ when it is true, this has the highest probability of rejecting $H_1$ when $H_2$ is true. Also note that this is in the absence of profiling.*
If one can design an optimal binary classifier (e.g. using a deep neural network) to distinguish between $H_1$ and $H_2$, then the output of the classifier will be the likelihood ratio:

$$L = \sum_{x \in H_1} \log y(x) + \sum_{x \in H_2} \log(1 - y(x))$$

$$\Rightarrow \quad y(x) = \frac{R(x)}{1 + R(x)}$$

$$R(x) = \frac{L(x|H_1)}{L(x|H_2)}$$

“likelihood ratio trick”

Underlies many major recent advances in deep learning, including generative models.
Current Search Paradigm

SUSY = Supersymmetry

- Simple combinations of 4-vectors, e.g., invariant masses
- Quark-proxy 4-vector ("jet")

(well-motivated) theory-biased & low-dimensional observables
Current paradigm for searches

Can we relax model assumptions and explore high-dimensional feature spaces?

(well-motivated) theory-biased & low-dimensional observables

(Clearly, we should still do model-dependent searches as well!)
Suppose you want to search for a new signal process
Model dependence

Signal sensitivity

Standard Model

B. Nachman, D. Shih, 2001.04990
Model dependence

Most searches (train with simulations)

\[ R_S(x) = \frac{\mathcal{L}(x|S_{sim})}{\mathcal{L}(x|B_{sim})} \]

signal and background model dependent

- S: a specific signal model, e.g. supersymmetry
- x: some set of relevant features characterizing each event (e.g. MET, HT, …)
- Rely on simulations of background and signal to construct likelihood ratio.

> 99% of searches at the LHC are of this type
Model dependence

- **Signal model independence**
  - Most searches (train with simulations)
  - Some searches (train signal versus data)

- **Background model independence**

**Standard Model**

**Signal sensitivity**

- e.g. Signal simulation versus calibration data

**Background specificity**
Model dependence

Signal sensitivity

\[ R(x) = \frac{\mathcal{L}(x|\text{data})}{\mathcal{L}(x|B_{\text{sim}})} \]

signal model independent
background model dependent

Idea: compare data vs *simulated* SM background in 1D histograms.

Most searches (train with simulations)
Data versus simulation

Some searches (train signal versus data)

B. Knuteson et al., D0, H1, CDF, CMS (“MUSiC”), ATLAS (“General Search”)
A. De Simone, T. Jacques, 1807.06038, A. Casa, Giovanna, 1809.02977, and others
Model dependence

Idea: compare data vs simulated SM background in 1D histograms.

\[ R(x) = \frac{\mathcal{L}(x|\text{data})}{\mathcal{L}(x|B_{\text{sim}})} \]

signal model independent
background model dependent

Warning: still background model dependent and may focus on the approximations of the simulation

B. Knuteson et al., D0, H1, CDF, CMS (“MUSiC”), ATLAS (“General Search”)
A. De Simone, T. Jacques, 1807.06038, A. Casa, Giovanna, 1809.02977, and others
Signal model independence

Some searches (train signal versus data)

Most searches (train with simulations)

Data versus simulation

Signal sensitivity

Background model independence

Classic: “bump hunt”

$$R(m) = \frac{\mathcal{L}(m|data)}{\mathcal{L}(m|B_{data})}$$

$m$: a single feature

Partially signal and background model independent

Idea: assume signal is localized in $m$ while background is smooth.

Use sidebands $m \notin (m_0-\delta m, m_0+\delta m)$ to interpolate background into signal region $m \in (m_0-\delta m, m_0+\delta m)$. 
Model dependence

Classic method: many discoveries & searches

Classic: “bump hunt”

\[ R(m) = \frac{\mathcal{L}(m | data)}{\mathcal{L}(m | B_{data})} \]

\( m \): a single feature

partially signal and background model independent

Idea: assume signal is localized in \( m \) while background is smooth.

Use sidebands \( m \notin (m_0 - \delta m, m_0 + \delta m) \) to interpolate background into signal region \( m \in (m_0 - \delta m, m_0 + \delta m) \).
Model dependence

Can we develop new methods that also assume as little as possible about the signal and learn from data (no simulation)?

Signal sensitivity

- Some searches (train signal versus data):
  - autoencoders
  - LDA
  - TNT
  - ANODE
  - CWoLa
  - SALAD

- Most searches (train with simulations):
  - Data versus simulation

- Background model independence

- Signal model independence

References:

- M. Farina, Y. Nakai, D. Shih, 1808.08992
- T. Heimel et al. SciPost Phys. 6 (2019) 030, and others
- B. Dillon et al., PRD 100 (2019) 056002
- B. Nachman, D. Shih, 2001.04990
- J. Collins, K. Howe, B. Nachman, PRL 121 (2018) 241803
- A. Andreassen, B. Nachman, D. Shih, 2001.05001
New ideas: brief illustration

Enhancing the bump hunt

Jet 1

Jet 2

\( m = \text{mass of two-jet system} \)

Collisions in/out of page

\( y = \text{many features of the two jets} \)
New ideas: brief illustration

Enhancing the bump hunt

![Graph showing data and simulation comparison]
New ideas: brief illustration

Enhancing the bump hunt

Event/jet Features
New ideas: brief illustration

Enhancing the bump hunt

Event/jet Features
New ideas: brief illustration

Enhancing the bump hunt

Learn $p(\text{features}|\text{ROI})/p(\text{features}|\text{side})$

CWoLa++: 1805.02664, 1902.02634, 2002.12376

Train outside ROI (or not) & pick “weird” events

Autoencoders: 1808.08992, 1808.08979 et al.
New ideas: brief illustration

Enhancing the bump hunt

Event/jet Features

Mask

![Graphs showing data and simulations]

A. Andreassen, BPN, D. Shih, 2001.05001

"Data" (Pythia)
Sim. (Herwig)
Signal

Pythia/Herwig

$m_{jj}$ [GeV]

Normalized to unity

N-subjettiness Ratio $\tau_{21}$ (lead $m_j$)

Normalized to unity

N-subjettiness Ratio $\tau_{21}$ (sublead $m_j$)
New ideas: brief illustration

Enhancing the bump hunt

**Event/jet Features**

- Learn $p(\text{features}|m)$ in sideband (SB) & interpolate
- Learn $p(\text{features}|m)$ in the masked region
- $2001.04990$ (conditional density estimation / ANODE)
- Learn $p_{\text{data}}(\text{features}|m)/p_{\text{MC}}(\text{features}|m)$ in SB
- $2001.05001$ (likelihood-free / SALAD)
New ideas: brief illustration

+More ideas were developed for the LHC Olympics!

It is likely that no one method will cover everything - need many ideas for broad sensitivity. See also P. Martin’s talk for a comparison of autoencoders and CWoLa.
Phase I: R&D  https://doi.org/10.5281/zenodo.2629072

Released spring 2019

Phase II: Black Box I  https://doi.org/10.5281/zenodo.3547721

Winter Olympics @ ML4Jets, NYU, Jan 2020

Phase III: Black Box II-IV  (boxed 2-3 are already on zenodo)

Summer Olympics @ Hamburg, July 18, 2020
We have prepared three black boxes of simulated data:

- 1 million events each
- 4-vectors of every reconstructed particle (hadron) in the event
- Particle ID, charge, etc. not included
- Single R=1 jet trigger $p_T > 1.2$ TeV
- Black boxes are meant to be representative of actual data, meaning they are mostly background and may contain signals of new physics

In addition, a sample of 1M QCD dijet events (produced with Pythia8 and Delphes3.4.1) was provided as a background sample.
A p-value associated with the dataset having no new particles (null hypothesis).

Short answer text

As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).

Short answer text

How many signal events (+uncertainty) are in the dataset (before any selection criteria).

Short answer text

Please consider submitting plots or a Jupyter notebook! (these will be private and used only for the presentation / documentation at the end)

Add file
• 10 groups submitted results on box 1
• 4 of these groups also submitted results on boxes 2 & 3
• A number of additional groups could not finish the challenge in time but got results on the R&D dataset
• 7 of these groups gave talks about their methods and results at the ML4Jets2020 conference
• See the [indico page](#) for details. Note that we did not reveal the answer until the end of the session!
LHC Olympics 2020: Teams (alphabetical order)

- Oz Amram and Cristina Mantilla Suarez (Johns Hopkins)
- Barry Dillon, D. Faroughy, J. Kamenik, M. Swezcz (Institute Jožef Stefan)
- Cosmos Dong (Reed)
- Julien Donini, Ioan-Mihail Dinu, Louis Vaslin (LPClermont, IFIN-HH)
- Felipe F. De Freitas (Beijing), Charanjit K. Khosa, Veronica Sanz (Sussex)
- Gustaaf Brooijmans, Julia Gonski, Alan Kahn, Inês Ochoa, Daniel Williams (Columbia)
- Patrick Komiske, Eric Metodiev, Nilai Sarda, Jesse Thaler (MIT)
- Christopher W. Murphy (Data Science)
- Soroosh Shalileh (HSE, Russia)
- George Stein, Uros Seljak, Biwei Dai (Cosmo, Berkeley)

Congratulations to all teams for braving the challenge!
People tried both supervised and unsupervised methods.

Methods used included:

- Autoencoders
- CWoLa hunting
- PCA outlier detection
- LSTM
- CNN+BDT
- variational RNNs for anti-QCD tagging
- density estimation
- biological neural network
- …
LHC Olympics 2020: Box 1

\[ m_{Z'} = 3.8 \text{ TeV} \]

\[ m_X = 732 \text{ GeV} \]
834 events. Same topology as R&D dataset (not known to participants)
**Results**

(only is arbitrary)

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ResNet + BDT</strong></td>
<td>[slides]</td>
</tr>
<tr>
<td><strong>PCA</strong></td>
<td>Principal component analysis was used as an outlier detector for resonant new physics. Given the popularity of deep learning today, it could be useful to have a shallow baseline to compare against. Features were selected by requiring that they, individually, produce a ROC AUC greater than one-half (a trivial condition to satisfy in supervised learning) on the practice dataset.</td>
</tr>
<tr>
<td><strong>LSTM</strong></td>
<td>I cluster the hadrons into anti-kT R = 0.7 jets, and run the sequence of jets into a simple RNN (supervised learning). The sequence is ordered by the jets' pr from high to low, zero-padding up to 10...I made two LSTM layers, with some dense layers after and a single output identifying signal or not.</td>
</tr>
<tr>
<td><strong>High-level features AE</strong></td>
<td><strong>slides</strong></td>
</tr>
<tr>
<td><strong>Tag N Train</strong></td>
<td><strong>slides</strong></td>
</tr>
<tr>
<td><strong>Density Estimation</strong></td>
<td><strong>slides</strong></td>
</tr>
<tr>
<td><strong>VRNN</strong></td>
<td><strong>slides</strong></td>
</tr>
<tr>
<td><strong>Latent Dirchlet Allocation</strong></td>
<td><strong>slides</strong></td>
</tr>
<tr>
<td><strong>Human NN</strong></td>
<td><em>Look at many histograms</em></td>
</tr>
</tbody>
</table>

**Topic modeling** [slides] *(also preliminary black box I results)*

**VAE** [slides]

**R&D and non-LHCO results**
Results - resonance mass

(order is arbitrary)

Correct answer

<table>
<thead>
<tr>
<th>Method</th>
<th>Resonance Mass [GeV]</th>
<th>Pull</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet + BDT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-level features AE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tag N Train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRNN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent Dirichlet Allocation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human NN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N.B. not everyone reported an uncertainty
Results - number of events

(order is arbitrary)

Correct answer

ResNet + BDT
PCA
LSTM
High-level features AE
Tag N Train
Density Estimation
VRNN
Latent Dirchlet Allocation
Human NN

N.B. not everyone reported an uncertainty

(\text{answer} - \text{true})/\text{uncert}
Results - daughter masses

(order is arbitrary)

N.B. Not everyone reported the daughter masses.
...and the winners are...

We refrained from ranking the results … all methods are an important contribution to this growing research area.

However, two submissions clearly stood out:

**Conditional density estimation for anomaly detection**
George Stein, Uros Seljak, Biwei Dai, He Jia

**Tag N’ Train (CWoLa + Autoencoder)**
Oz Amram and Cristina Mantilla Suarez
Stay tuned for more on the LHCO 2020…

We will be organizing a 1-day mini-workshop on anomaly detection in Hamburg the Saturday before BOOST (July 18).

The answers for Boxes 2 and 3 will be revealed.

We will also discuss plans for a community paper on new methods for anomaly detection and the LHCO2020.

Please come and join us!

http://indico.desy.de/indico/e/anomaly2020
Conclusions and outlook

These are exciting times for anomaly detection in HEP.

Many new approaches making use of unsupervised ML are being developed by theorists and experimentalists.

Model independent searches have a bright future at the LHC. Maybe this is how we will finally discover the new physics!

These methods also have potential applications beyond HEP. There have already been connections with other fields in the winter olympics!