UNVEILING BSM HIDDEN PHYSICS WITH MACHINE LEARNING

Tao Liu

The Hong Kong University of Science and Technology

Following the discovery of an SM-like Higgs boson at the Large Hadron Collider the chief focus is on the observation of new phenomena beyond the Standard Model.

Such observations would provide clear guiding principles for the future of the entire field that are, given the present discussions on future colliders all over the world, more crucial than ever before.

So far, inclusive and model dependent searches have not provided evidence of new resonances, indicating that these could be driven by more subtle topologies, hidden by large backgrounds. Phenomenologists have found many classes of New Physics that are difficult to test with current LHC analyses.

In this light, it is important to keep investigating what theories could be further explored. In addition, we need to elaborate on methodologies that display less model dependencies. The use of Machine Learning may play a critical role here.

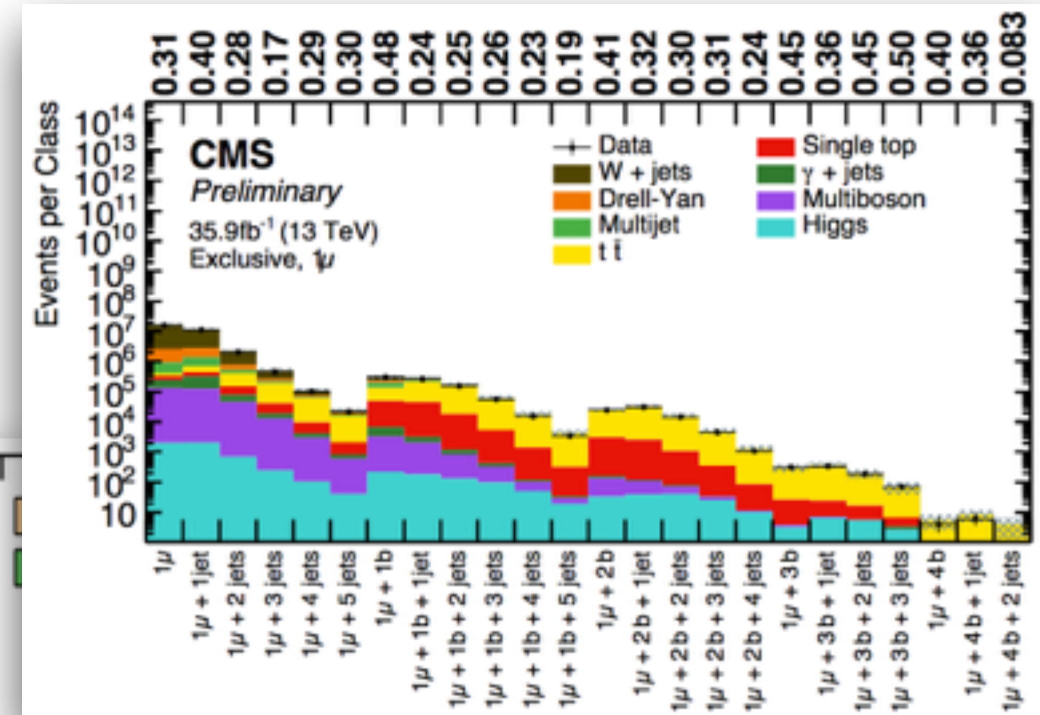
The opportunity to test a wide range of New Physics opened up very recently: CERN announced on the 11th of December 2020 a new open data policy, which "will make scientific research more accessible to the community". This opens up the testing ground for new search strategies.

The aim of this workshop is to bring together theory and experiment to identify novel signatures connected to such 'hidden' New Physics, to devise new methodologies, and to establish new search strategies.

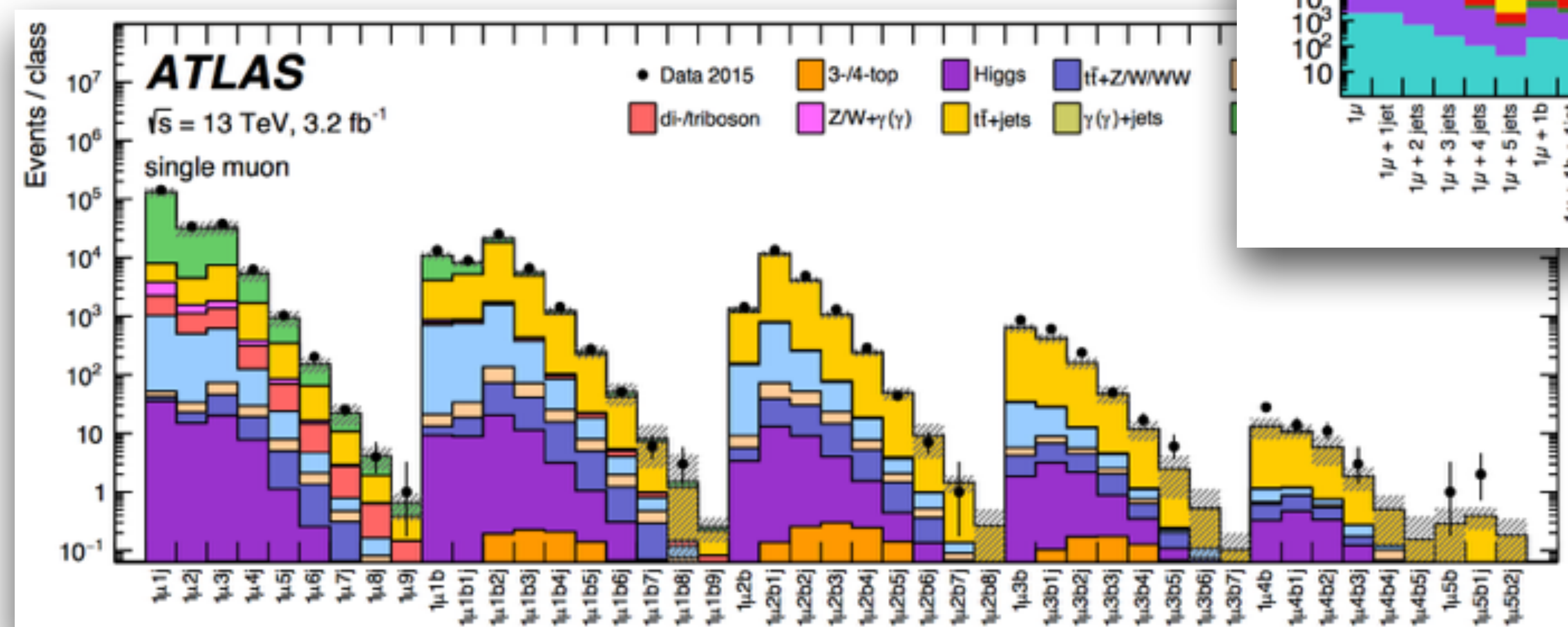
LEP - Tevatron - LHC Epochs

Compare data with the simulated background for each channel over its full/sub phase space

Recently correlated to DNN by [D'Agnolo et. al., 2018,Simone et. al., 2018]



[CMS-PAS-EXO-19-008]



[ATLAS,1807.07447]

But, wanna coverage the analysis to the signal-rich region efficiently

=> Benefit sensitivity study (especially data-driven) and the subsequent NP identification

Novelty (Anomaly) Detection

ELSEVIER

journal homepage: www.elsevier.com/locate/sigpro

Review

A review of novelty detection

Marco A.F. Pimentel*, David A. Clifton, Lei Clifton, Lionel Tarassenko

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK

A R T I C L E I N F O

Article history:
Received 17 October 2012
Received in revised form
16 December 2013
Accepted 23 December 2013
Available online 2 January 2014

Keywords:
Novelty detection
One-class classification
Machine learning

A B S T R A C T

Novelty detection is the task of classifying test data that differ in some respect from the data that are available during training. This may be seen as "one-class classification", in which a model is constructed to describe "normal" training data. The novelty detection approach is typically used when the quantity of available "abnormal" data is insufficient to construct explicit models for non-normal classes. Application includes inference in datasets from critical systems, where the quantity of available normal data is very large, such that "normality" may be accurately modelled. In this review we aim to provide an updated and structured investigation of novelty detection research papers that have appeared in the machine learning literature during the last decade.

Novelty detection - a task in ML closely related to this goal

Novelty (Anomaly) Evaluation

ELSEVIER journal homepage: www.elsevier.com/locate/sigpro

Review

A review of novelty detection

CrossMark

Marco A.F. Pimentel *, David A. Clifton, Lei Clifton, Lionel Tarassenko

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK

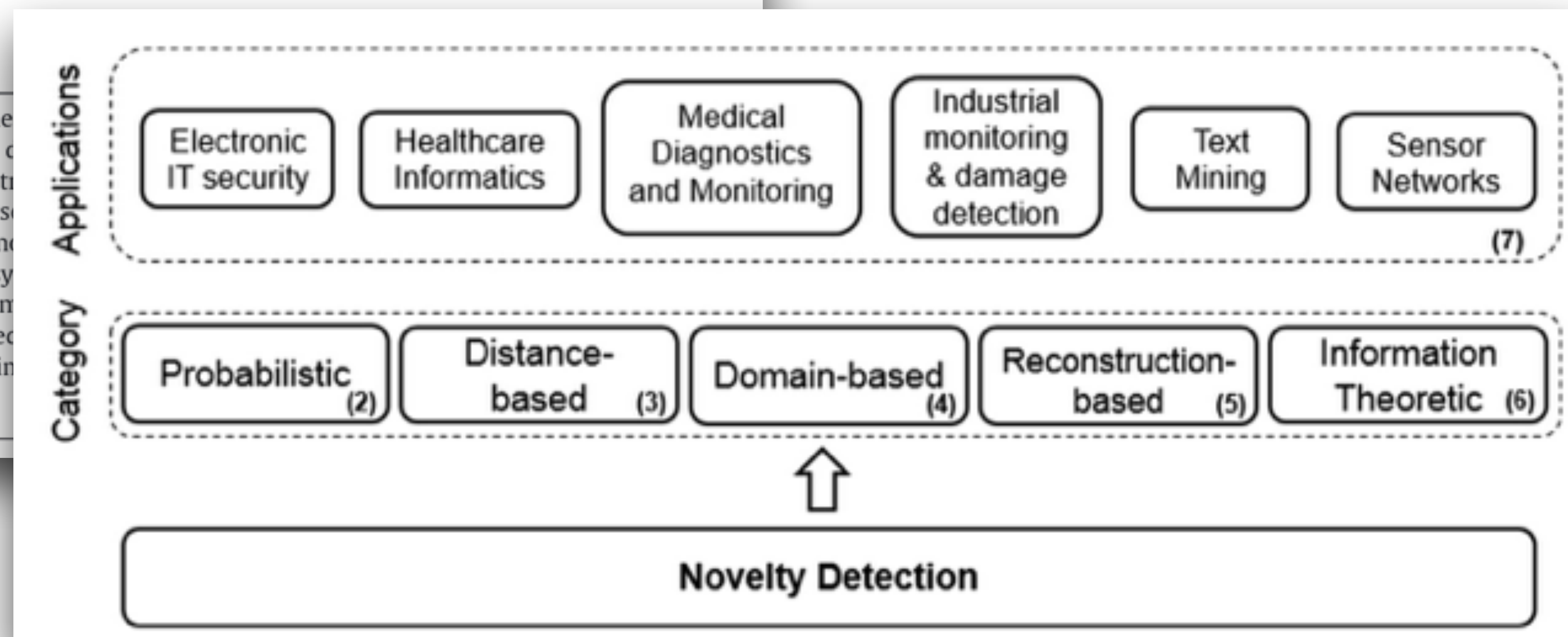
A R T I C L E I N F O

Article history:
Received 17 October 2012
Received in revised form
16 December 2013
Accepted 23 December 2013
Available online 2 January 2014

Keywords:
Novelty detection
One-class classification
Machine learning

A B S T R A C T

Novelty detection is the
data that are available
which a model is constr
approach is typically us
to construct explicit m
datasets from critical sy
such that "normality" m
updated and structure
appeared in the machin

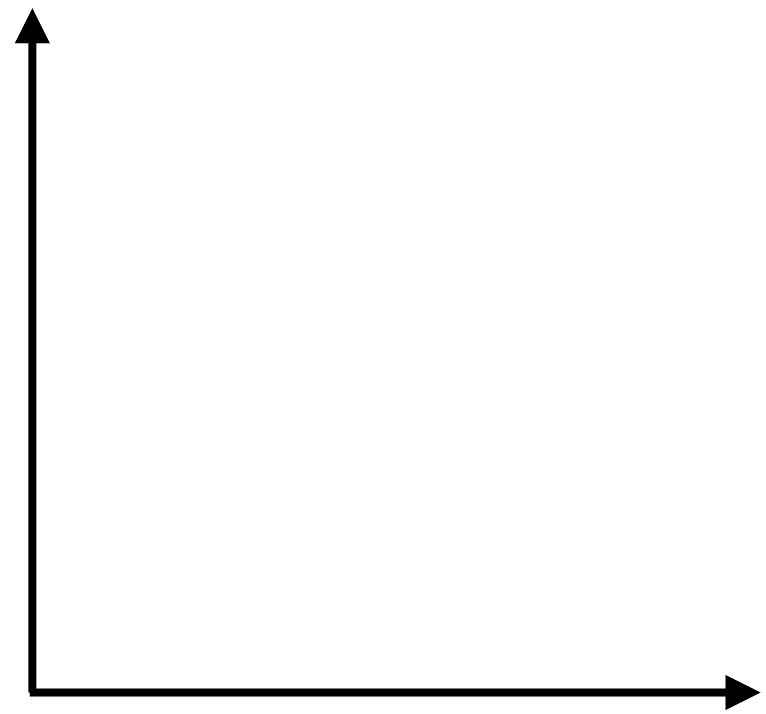


The history of addressing ``novelty detection'' is essentially one of developing novelty evaluators/evaluating methods for the testing sample

These evaluators/evaluating methods in principle can be applied to both semi-supervised and fully unsupervised learning which are distinguished by how to model the backgrounds (simulation-based or data-driven)

Broad Classification of Novelty Evaluators

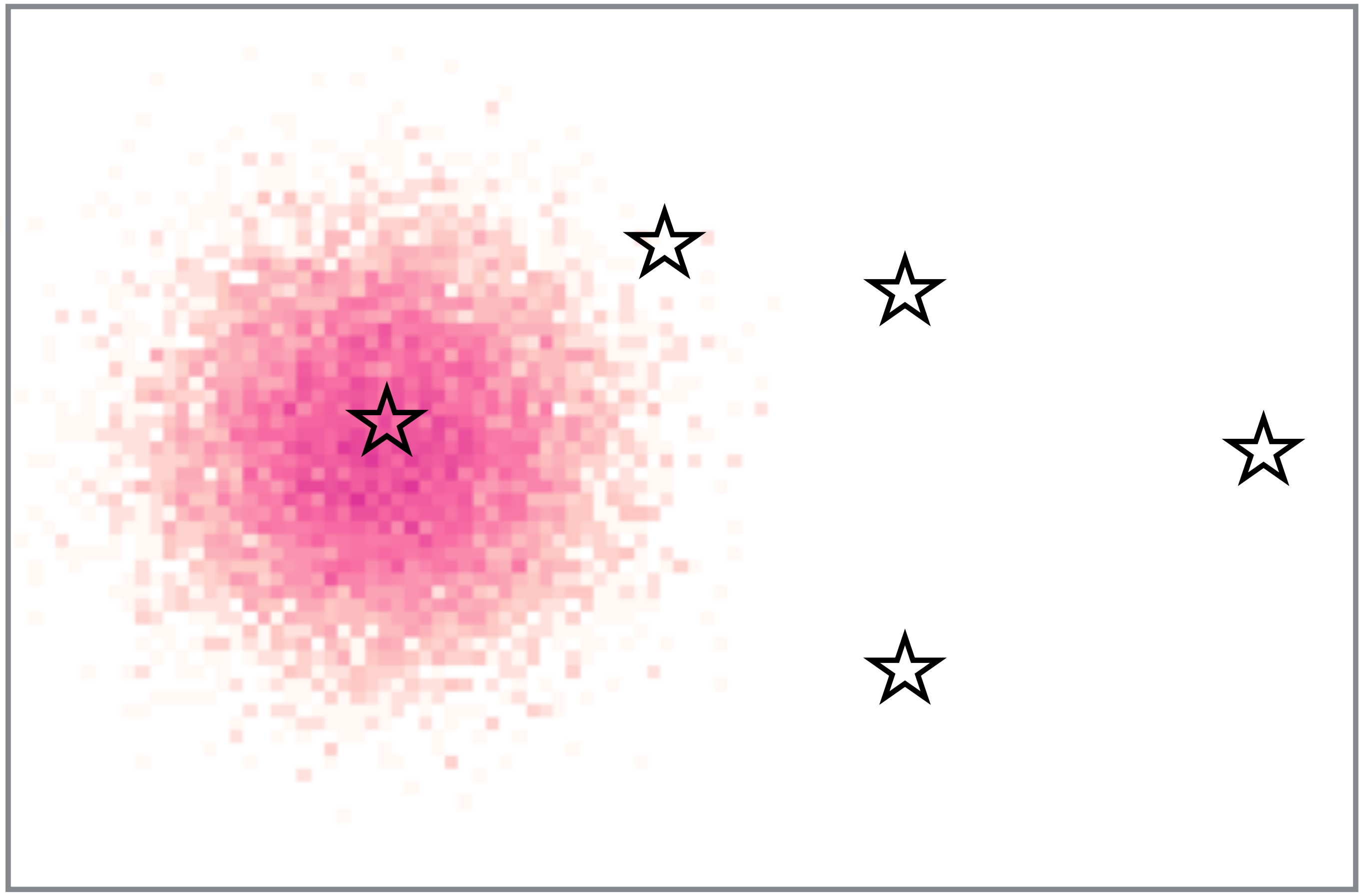
Clustering (density)-based



Isolation-based

[J. Hajer, Y. Li, TL, H. Wang; arXiv:1807.10261]

Isolation-based



The novelty of a given testing point (☆) is evaluated according to its distance to or isolation from the distribution of the ``known-pattern'' data in the feature space (the ``known-pattern'' data distribution could be either from simulation or from the control-region data).

The other testing points in the sample are irrelevant.

Isolation-based I: k-Nearest Neighbors

$$\Delta_{\text{iso}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}} \qquad \mathcal{O} = \frac{1}{2}\left(1 + \text{erf}\left(\frac{c\Delta}{\sqrt{2}}\right)\right)$$

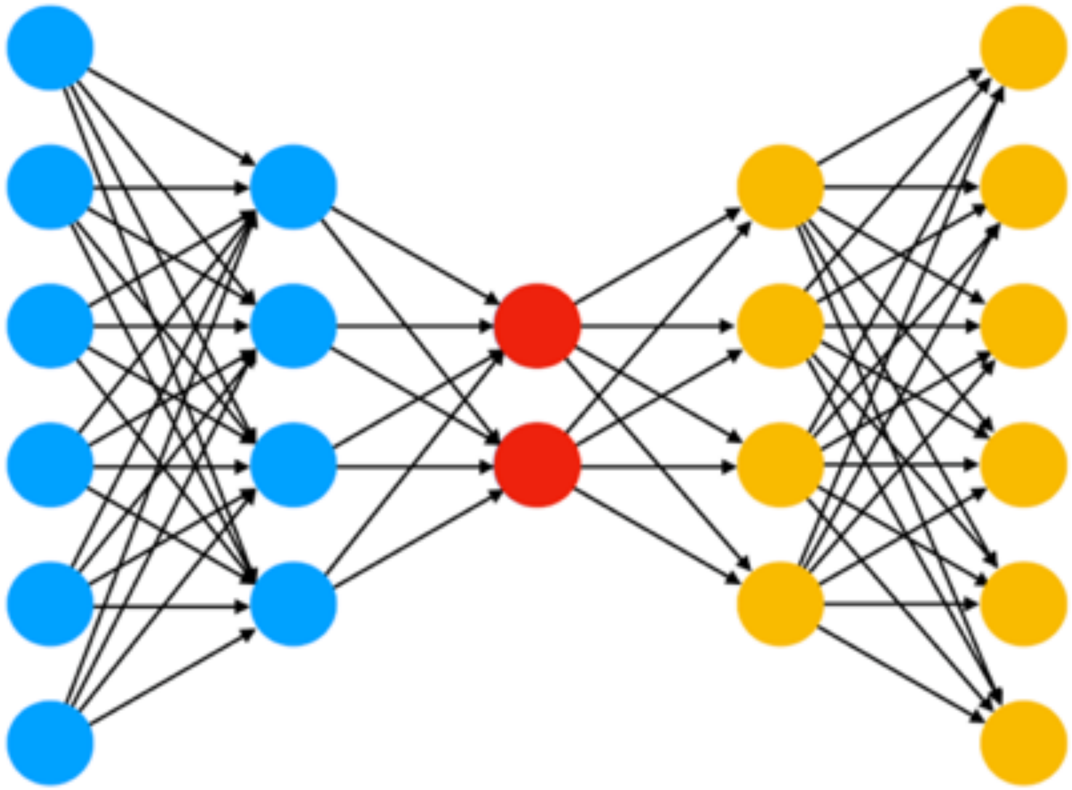
Novelty measure: range unnormalized

Novelty evaluator: $0 \leq \mathcal{O} \leq 1$

- d_{train} : mean distance of a testing data point to its k nearest neighbors
- $\langle d'_{\text{train}} \rangle$: average of the mean distances defined for its k nearest neighbors
- $\langle d'^2_{\text{train}} \rangle^{1/2}$: standard deviation of the latter
- All quantities are defined w.r.t. the training sample

[J. Hajer, Y. Li, TL, H. Wang; arXiv:1807.10261]

Isolation-based II: Reconstruction Error

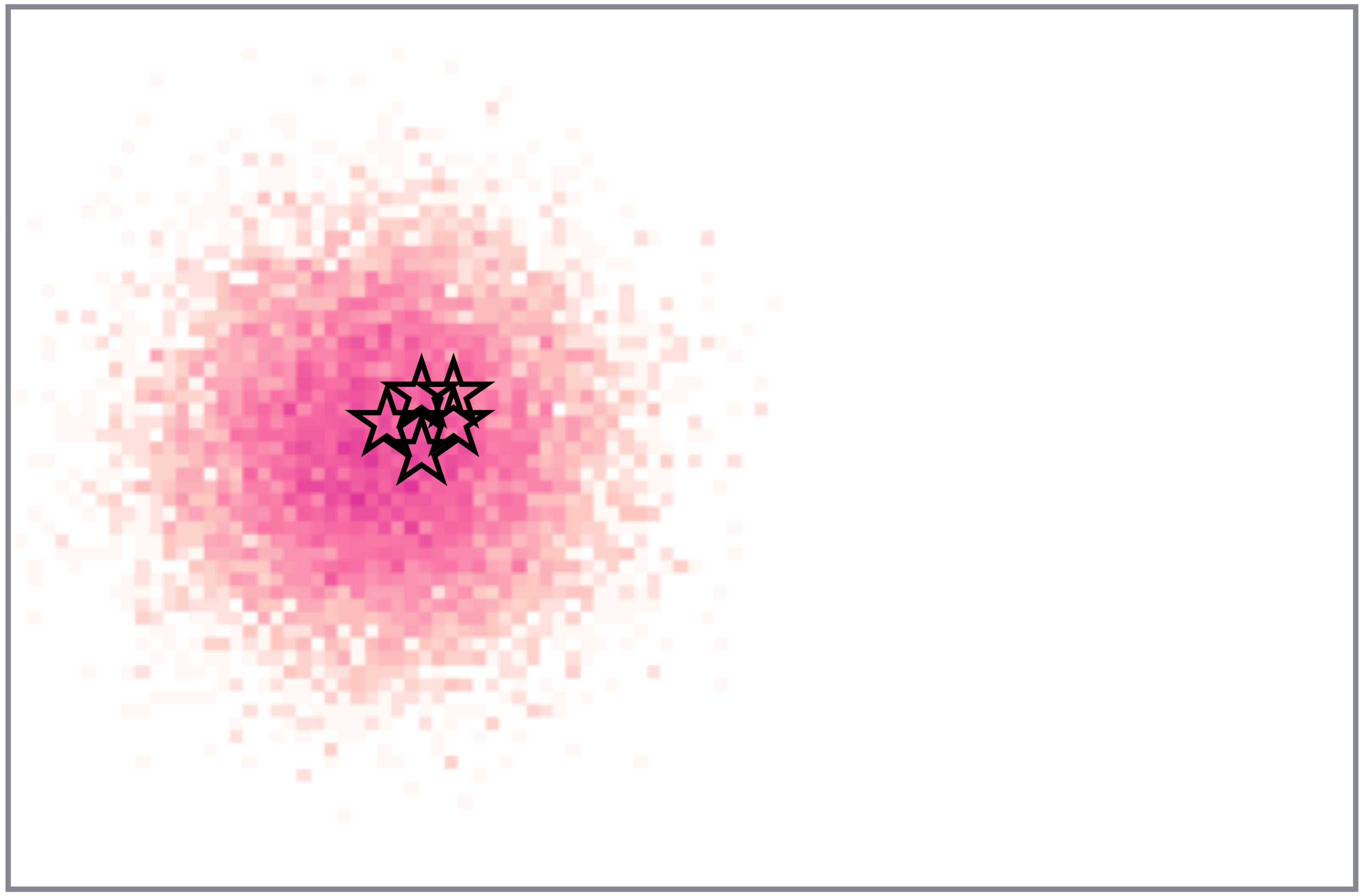


$$L = |x - x'|^2$$

[T. Heimel et. al.; M. Farina et. al; 2018]

- Reconstruction error: a measure of the distance of the testing point to the distribution of the known-pattern data
- In essence, isolation-based

Clustering (density)-based



The novelty of a given testing point is evaluated according to the data clustering around this point on top of the known-pattern data distribution in the feature space.

The testing data around this given testing point are relevant.

Clustering (Density)-based I: k-Nearest Neighbors

[J. Hajer, Y. Li, TL, H. Wang; arXiv:1807.10261]

$$\Delta_{\rm iso} = \frac{d_{\rm train} - \langle d'_{\rm train}\rangle}{\langle d'^2_{\rm train}\rangle^{1/2}}$$

$$\boxed{\Delta_{\rm clu} = \frac{d_{\rm test}^{-m} - d_{\rm train}^{-m}}{d_{\rm train}^{-m/2}}}$$

- $d_{\rm train}$: mean distance of a testing data point to its k nearest neighbors in the training dataset
- $d_{\rm test}$: mean distance of a testing data point to its k nearest neighbors in the testing dataset
- m: dimension of the feature space
- Novelty response is evaluated by comparing local densities of the testing point in the training and testing samples

Clustering (Density)-based II: ANODE

arXiv.org > hep-ph > arXiv:2001.04990

High Energy Physics – Phenomenology

Anomaly Detection with Density Estimation

Benjamin Nachman, David Shih

(Submitted on 14 Jan 2020)

We leverage recent breakthroughs in neural density estimation to propose a n

More specifically, ANODE attempts to learn two densities: $p_{\text{data}}(x|m)$ and $p_{\text{background}}(x|m)$ for $m \in \text{SR}$. Then, classification is performed with the likelihood ratio

$$\boxed{R(x|m) = \frac{p_{\text{data}}(x|m)}{p_{\text{background}}(x|m)}.} \tag{3.1}$$

In the ideal case that $p_{\text{data}}(x|m) = \alpha\, p_{\text{background}}(x|m) + (1-\alpha)\, p_{\text{signal}}(x|m)$ for $0 \leq \alpha \leq 1$ and $m \in \text{SR}$, Eq. 3.1 is the optimal test statistic for identifying the presence of signal. In the absence of signal, $R(x|m) = 1$, so as long as $p_{\text{signal}}(x|m) \neq p_{\text{background}}(x|m)$, this leads to a zero-background search.

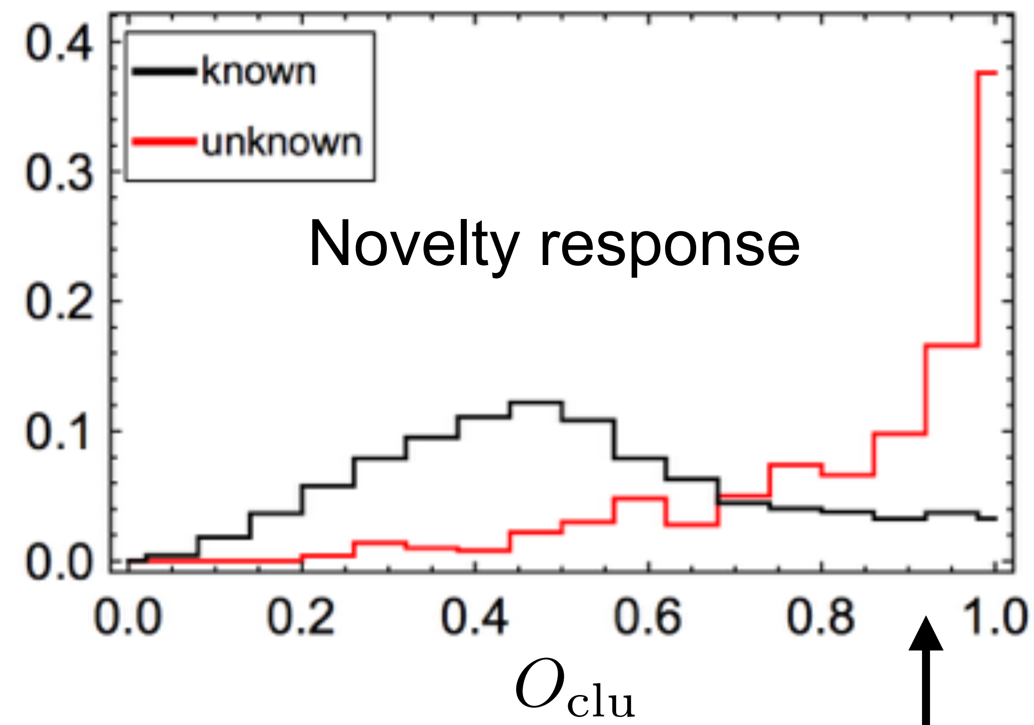
Novelty response is evaluated by comparing probability densities of the testing point in the ``background'' and ``data'' samples

Synergy of Both

[J. Hajer, Y. Li, TL, H. Wang; arXiv:1807.10261]

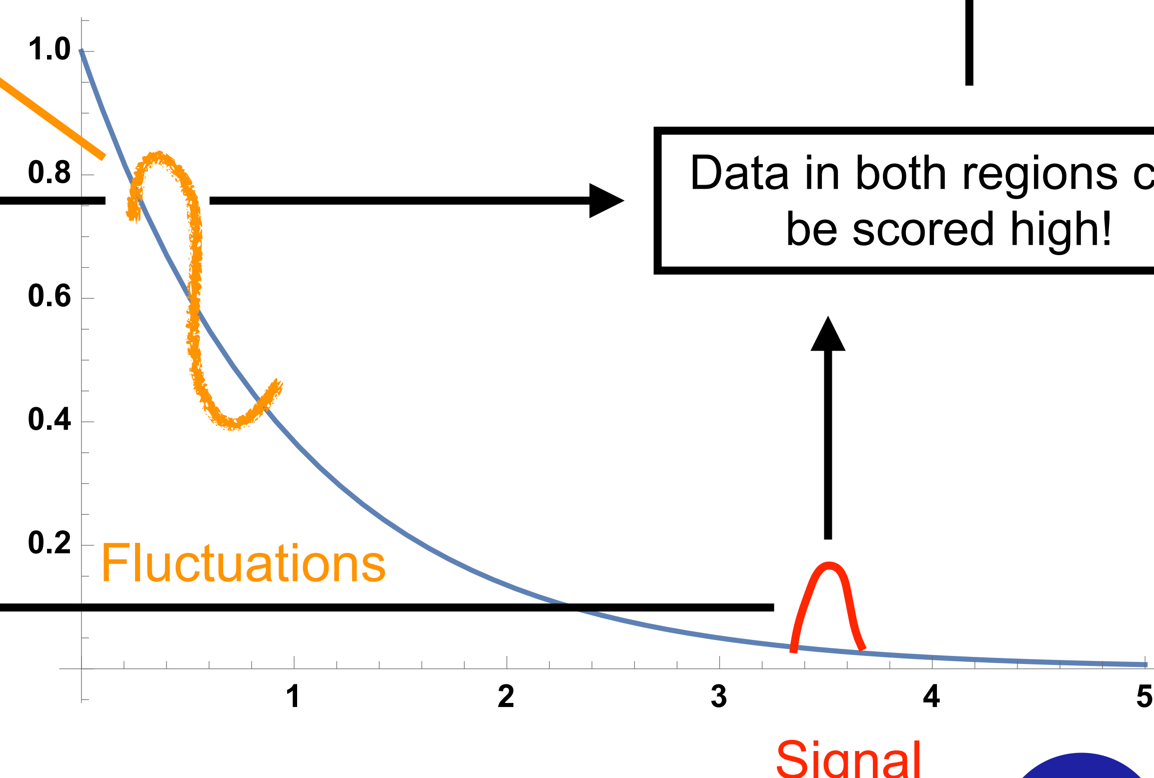
$$\mathcal{O}_{\rm syn} = \sqrt{\mathcal{O}_{\rm iso}\mathcal{O}_{\rm clu}}$$

Expect the O_clu high scoring for the known-pattern data from non-signal bulk to be compensated for by their low-scoring from O_iso



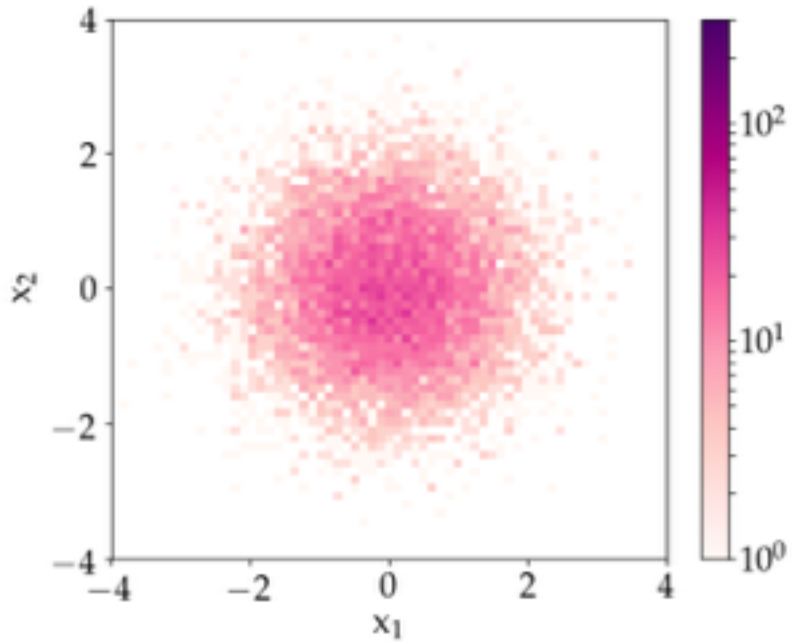
Scored low by O_iso

Scored high by O_iso



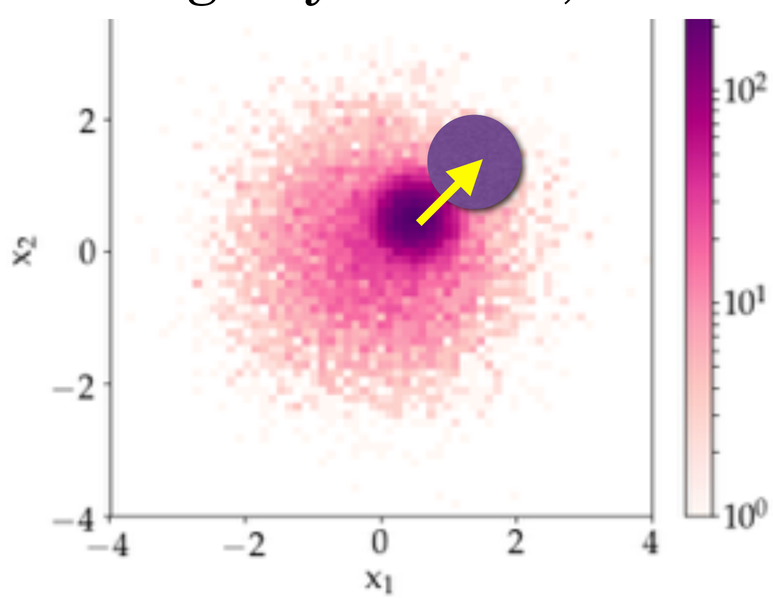
Data in both regions can be scored high!

Synergy of Both

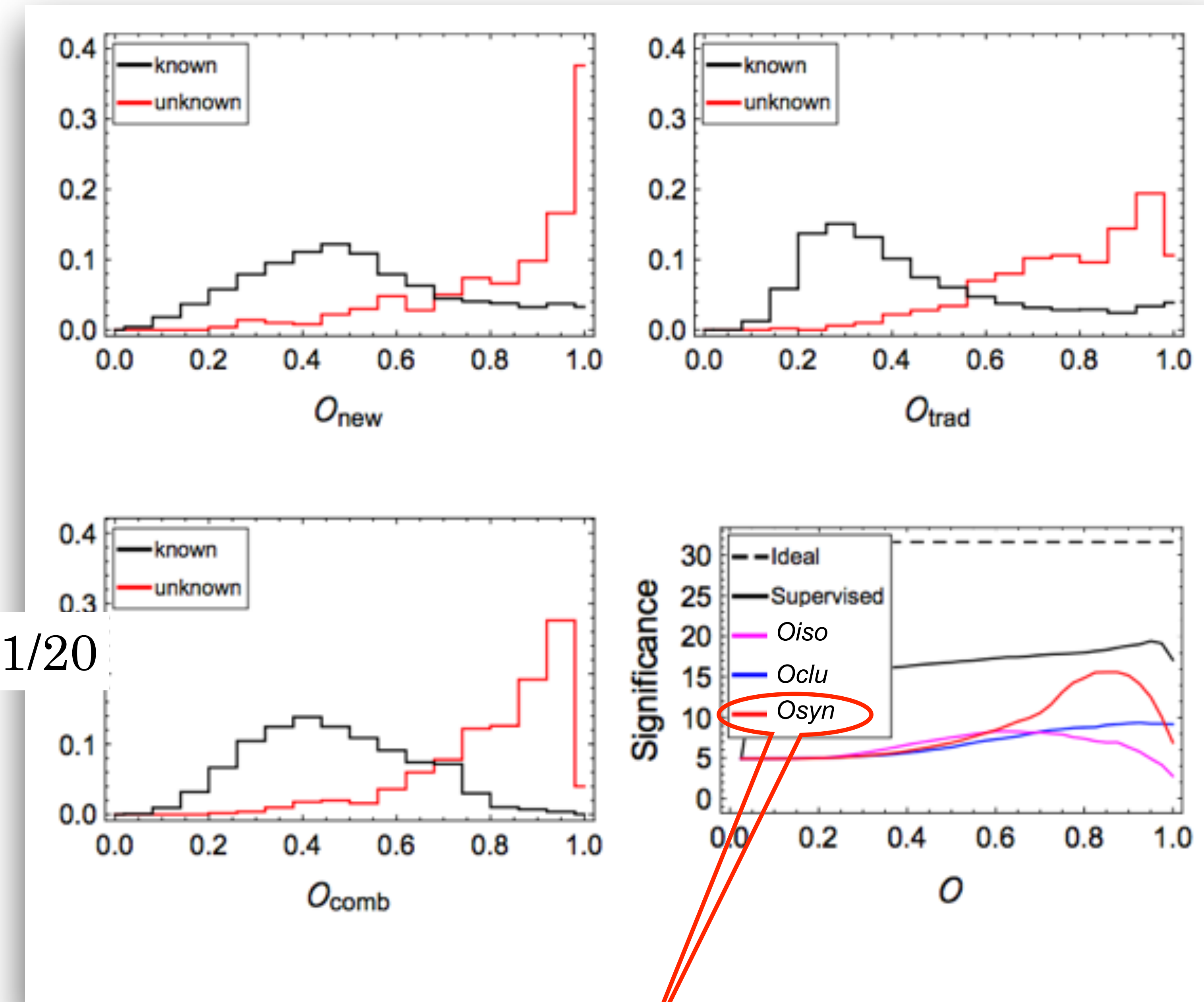


(a) Training data.

Center slightly shifted, with S/B = 1/20

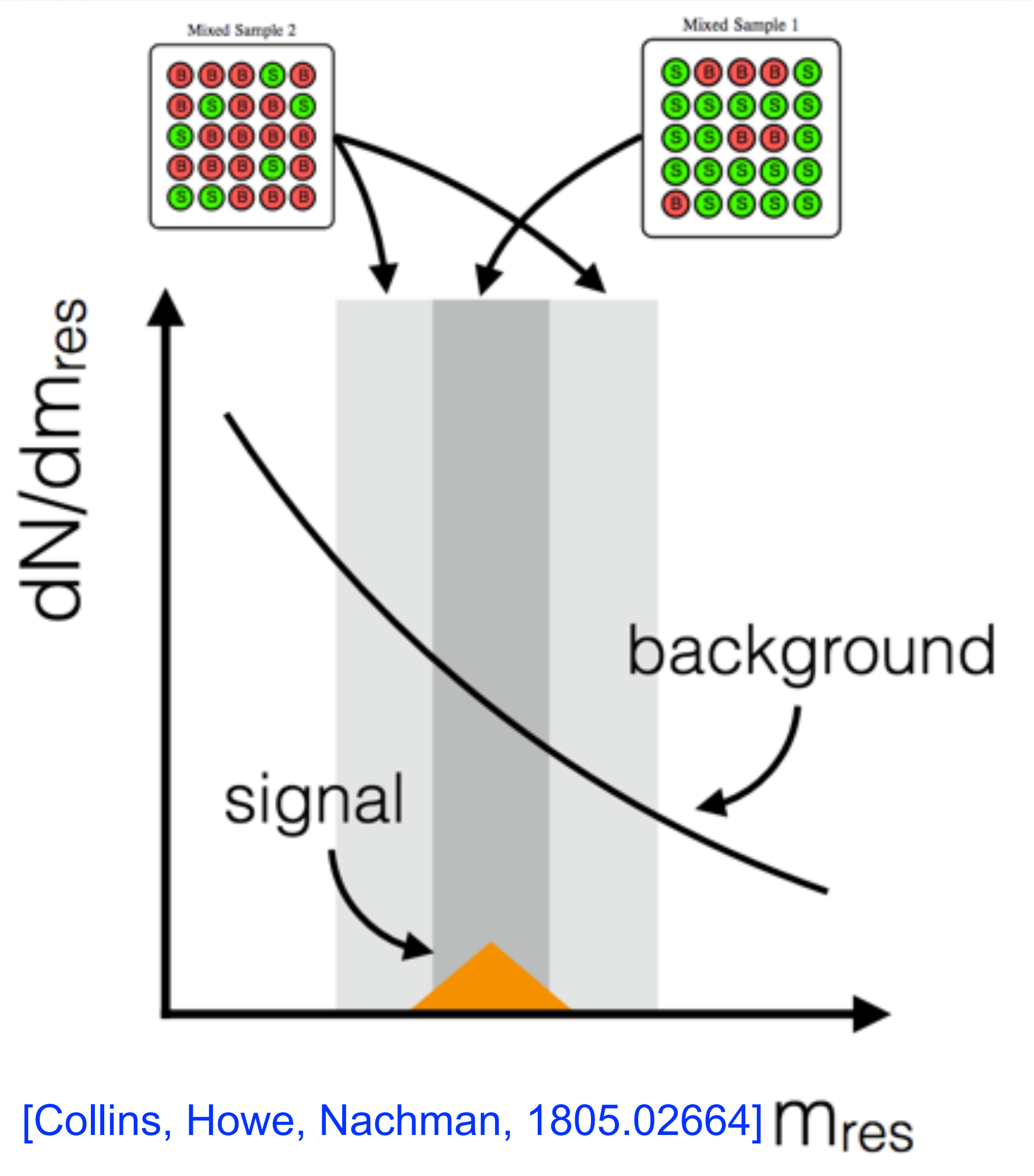


(b) Testing data.



Significance curve for the synergy-based evaluator

Weakly-supervised Learning/CWoLa



[Collins, Howe, Nachman, 1805.02664]

- Assign labels to each event in the mixed samples (0 => M0; 1 => M1) and then performs supervised learning
- S0/B0 =/ S1/B1: optimize S0/B0 vs S1/B1 <=> optimize S vs B
- If S0/B0 => 0, S1/B1 => 1, then reduced to fully supervising learning

Novelty evaluation (=> mixed samples in favor of S [high-score bin] and B [low-score bin], respectively) can be combined with the WSL algorithm, to further improve the sensitivity performance of the classifier for novelty detection

First application to Real Data

arXiv.org > hep-ex > arXiv:2005.02983

Search...

Help | Advance

High Energy Physics – Experiment

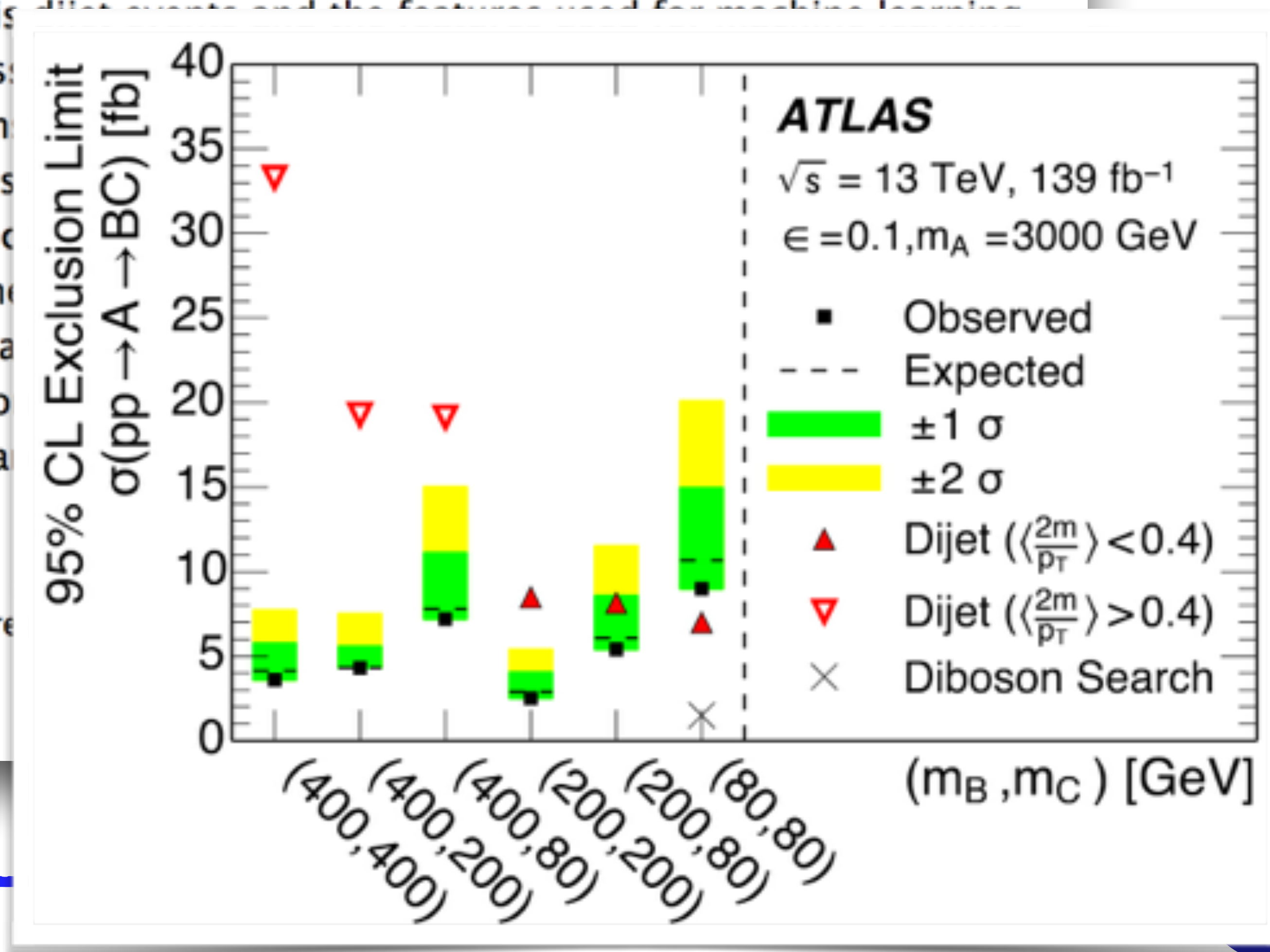
[Submitted on 6 May 2020]

Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector

ATLAS Collaboration

This Letter describes a search for resonant new physics using a machine-learning anomaly detection procedure that does not rely on a signal model hypothesis. Weakly supervised learning is used to train classifiers directly on data to enhance potential signals. The targeted topology is
are the masses of the two jets. The resulting analysis is es
$m_A \sim \mathcal{O}(\text{TeV})$, $m_B, m_C \sim \mathcal{O}(100\ \text{GeV})$ and B, C are recon
associated with a large trials factor in the scan of the mass
collision data set of 139 fb^{-1} recorded by the ATLAS detec
There is no significant evidence of a localized excess in th
TeV. Cross-section limits for narrow-width A, B, and C pa
$m_A = 3$ TeV and $m_B \gtrsim 200$ GeV, a production cross sectio
level, depending on m_C. For certain masses, these limits a
the inclusive dijet search.

Comments: 32 pages in total, author list starting page 16, 3 figure
Subjects: **High Energy Physics – Experiment (hep-ex)**
Report number: CERN-EP-2020-062



New Idea Testing

Despite this progress, most of these new ideas are proof-of-concept =>
Needs to test their performance in a more realistic environment

Many new approaches inspired by (or at least demonstrated on) the **LHC Olympics 2020 Data Challenge**

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]

[Shih, HKIAS 2021 HEP program]

Opportunity from Open Data!

Despite this progress, most of these new ideas are proof-of-concept =>
Needs to test their performance in a more realistic environment

CERN Accelerating science

ABOUT NEWS SCIENCE

News › Press release › Topic: Knowledge sharing

Voir en français

CERN announces new open data policy in support of open science

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

11 DECEMBER, 2020

of an SM-like Higgs boson at the Large Hadron Collider the chief focus is on
ıenomena beyond the Standard Model.

provide clear guiding principles for the future of the entire field that are,
ions on future colliders all over the world, more crucial than ever before.

el dependent searches have not provided evidence of new resonances,
d be driven by more subtle topologies, hidden by large backgrounds.
ound many classes of New Physics that are difficult to test with current LHC

nt to keep investigating what theories could be further explored. In addition,
methodologies that display less model dependencies. The use of Machine
Learning may play a critical role here.

The opportunity to test a wide range of New Physics opened up very recently: CERN announced on the 11th of December 2020 a new open data policy, which "will make scientific research more accessible to the community". This opens up the testing ground for new search strategies.

The aim of this workshop is to bring together theory and experiment to identify novel signatures connected to such 'hidden' New Physics, to devise new methodologies, and to establish new search strategies.

Opportunity from Open Data!

arXiv.org > hep-ex > arXiv:2005.01598

Search...

Help | Advanced S

High Energy Physics – Experiment

[Submitted on 4 May 2020]

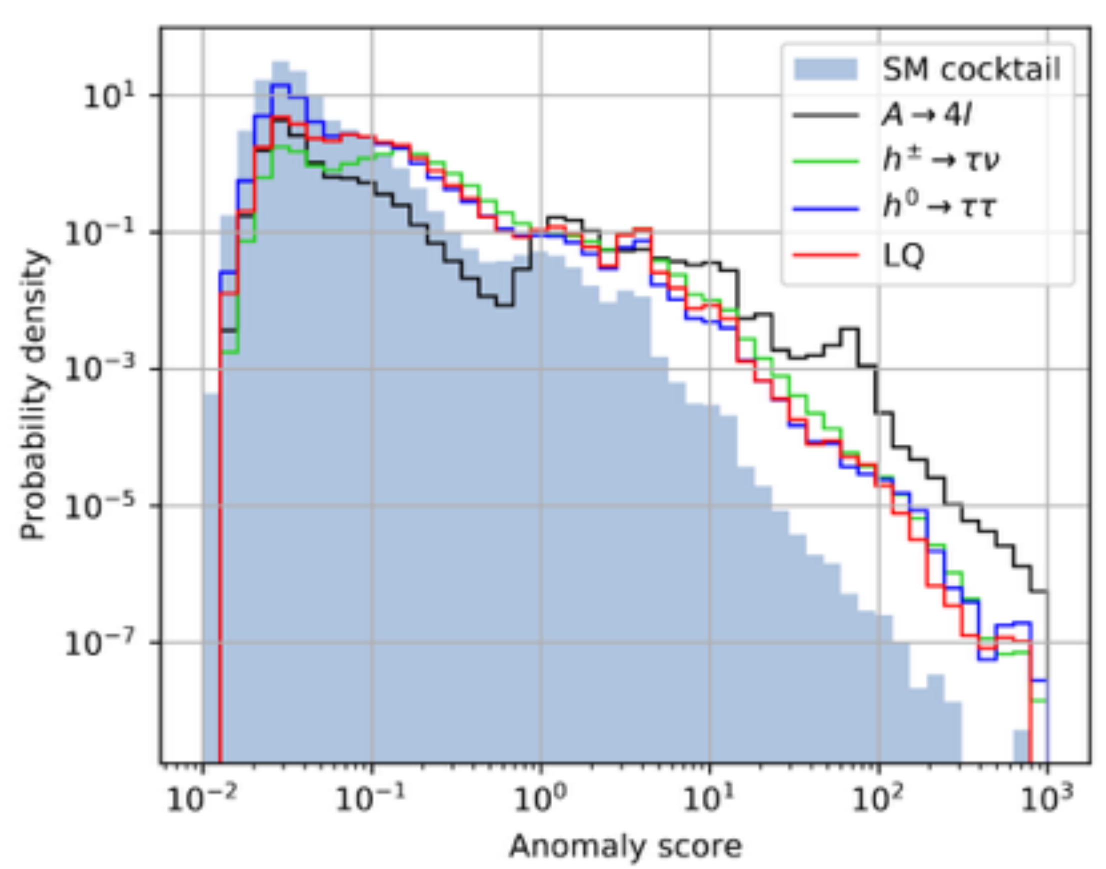
Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark

Oliver Knapp, Guenther Dissertori, Olmo Cerri, Thong Q. Nguyen, Jean-Roch Vlimant, Maurizio Pierini

We apply an Adversarially Learned Anomaly Detection (ALAD) algorithm to the problem of detecting new physics processes in proton-proton collisions at the Large Hadron Collider. Anomaly detection based on ALAD matches performances reached by Variational Autoencoders with a substantial improvement in some cases. Training the ALAD
CMS Open Data, we show how a data-driven anomaly detection a
in real life, re-discovering the top quark by identifying the main
experimental signature at the LHC.

Comments: 16 pages, 9 figures

Subjects: **High Energy Physics – Experiment (hep-ex)**; Machine Learning (
Phenomenology (hep-ph)



Summary

- The null results for the NP searches at LHC strongly call for strategies of searching for highly unexpected signals
- The ML techniques developed for novelty/anomaly detection in last decades come in handy
- Open data, committed for open science, provides a great opportunity to fill up the gap between proof of concept of these ideas and real data analysis … …

Thank you!

大學教育資助委員會
University Grants Committee

GRF under grant No. 16312716, 16302117 and 16304315