



Novelty Detection in HEP Data Analysis

Based on arXiv: 1807.10261

in collaboration with Jan Hajer, Tao Liu and He Wang

Ying-Ying Li
HKUST



Supervised Learning in HEP Data Analysis

- High Energy Physics (HEP) is a big data science and has a long history of using supervised ML for data analysis: BDT, neural network, etc.
 - a. neural network for top search @D0 (1990);
 - b. MiniBooNe first used BDT and compared it with neural network (2004);
 - c. BDT becomes more popular in HEP data analysis, e.g. in TOP2018, more than 50% of the results presented are based on BDT analysis.
- These supervised ML methods are highly efficient in analyzing signal events with complex topologies.

Un-supervise the Learning? —Novelty Detection

- Given new physics scenarios may share similar final states, we hope that they can be searched for simultaneously and more efficiently
 - Case I: di-top partner production vs Z' production (decay to top pair).
 - Case II: Exotic Higgs decays (rich decay modes)
- Also, given the null results at LHC, new physics could be very unexpected.
- Supervised learning is model dependent, incapable for these tasks.

=> Novelty detection is ``model''-independent, complementary to supervised learning,
Allows us to detect new physics without a priori knowledge about it.



ELSEVIER

Contents lists available at ScienceDirect

Signal Processing

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



ARTICLE INFO

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

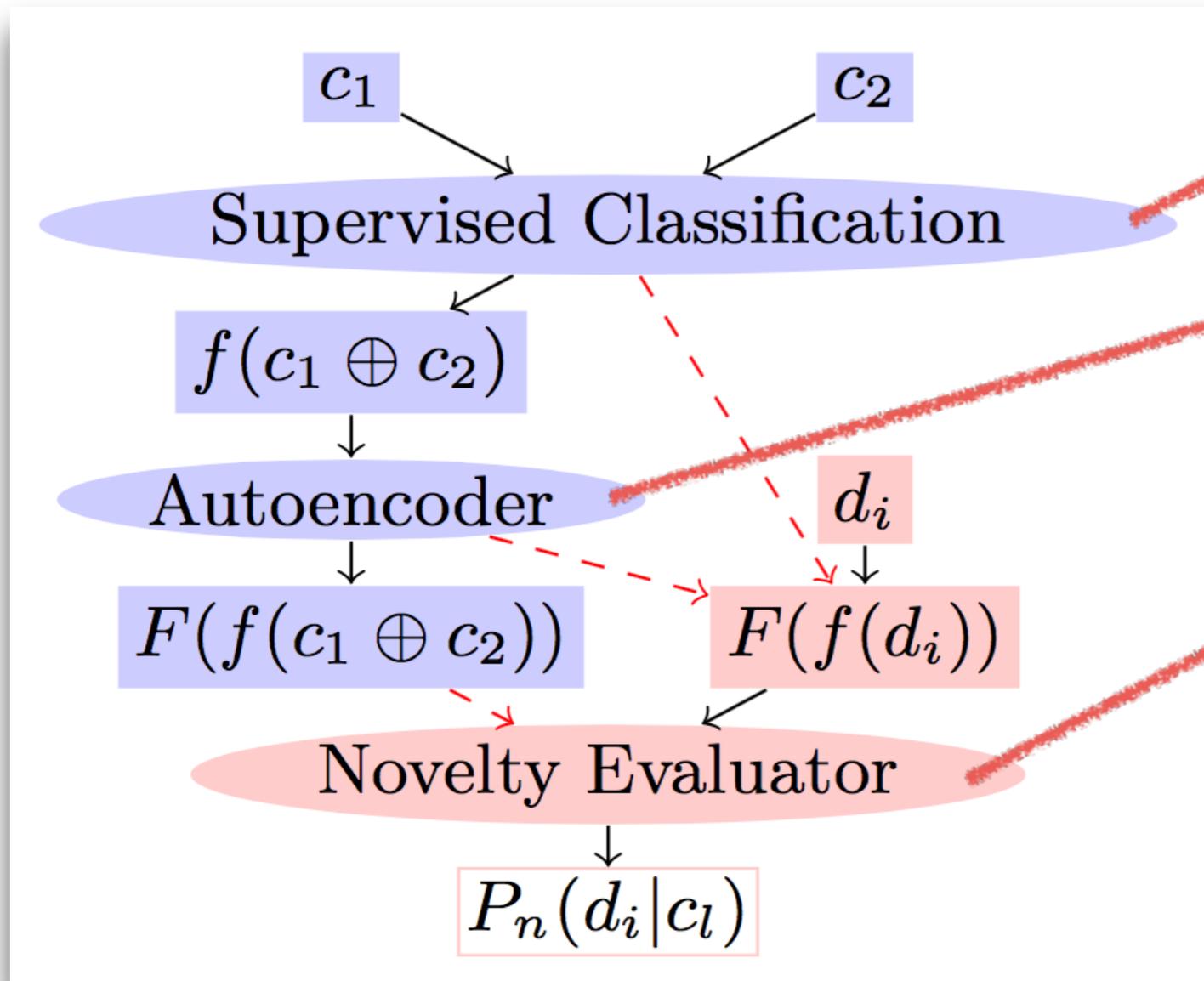
Machine learning

ABSTRACT

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.

© 2014 Published by Elsevier B.V.

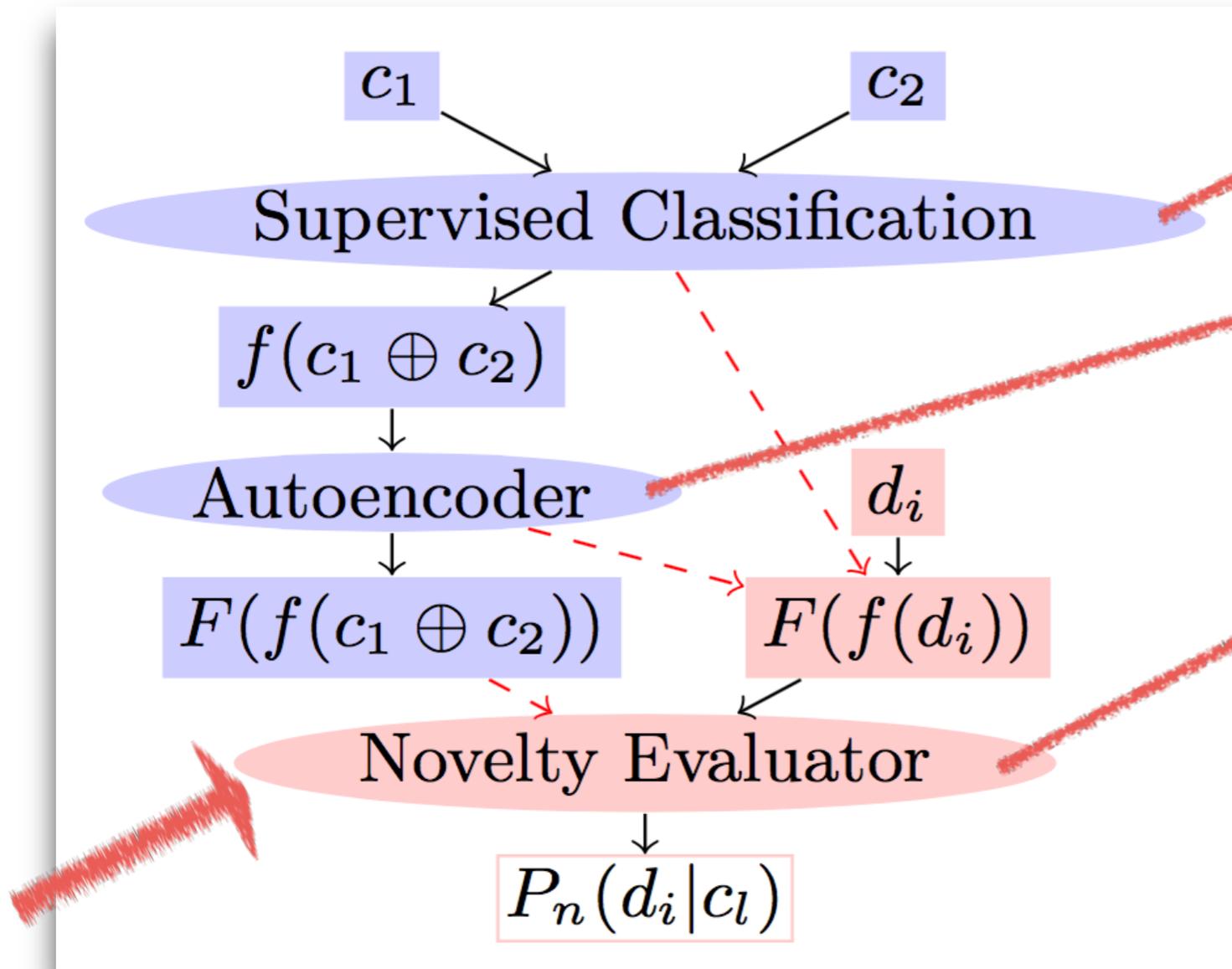
Workflow for Novelty Detection



- Step 1: (SM/background) feature learning
- Step 2: dimension reducing of feature space (**auto-encoder**)
- Step 3: novelty evaluating of testing data
- Analyze detection sensitivity based on novelty response of testing data

Allow us to detect new physics model-independently!

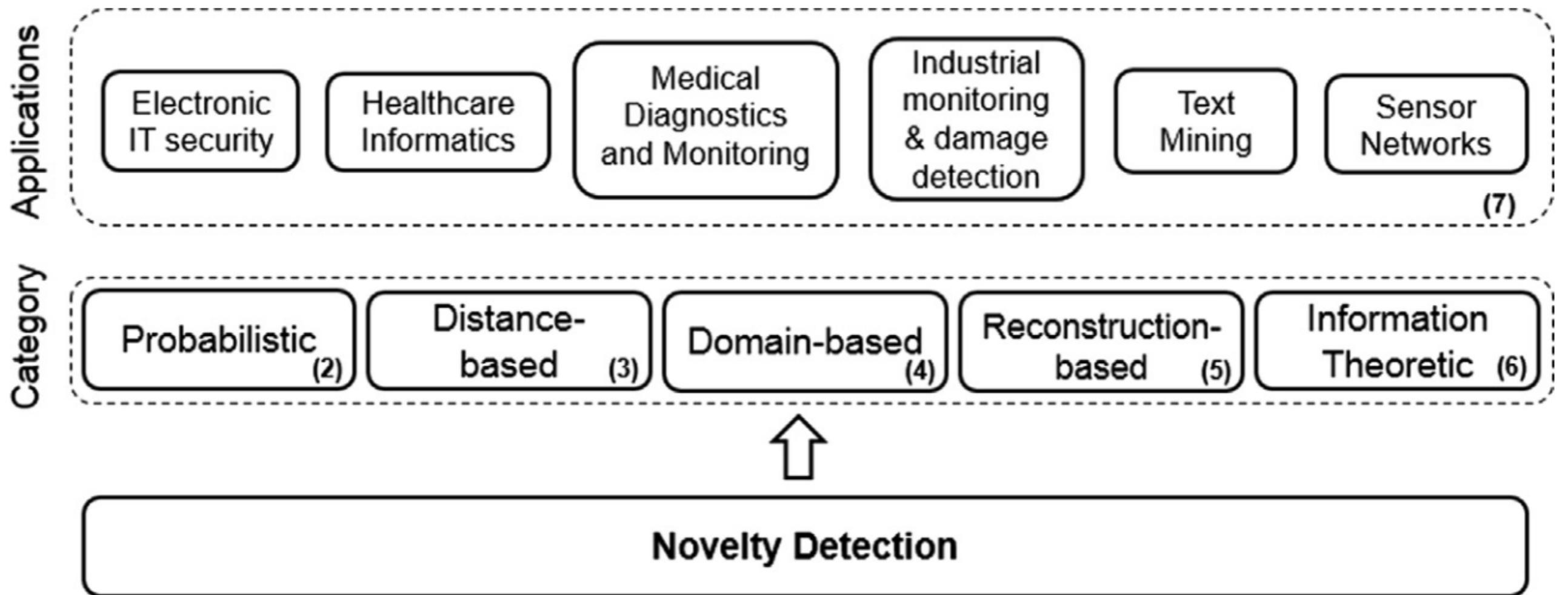
Workflow for Novelty Detection



- Step 1: (SM/background) feature learning
- Step 2: dimension reducing of feature space (**auto-encoder**)
- Step 3: novelty evaluating of testing data
- Analyze detection sensitivity based on novelty response of testing data

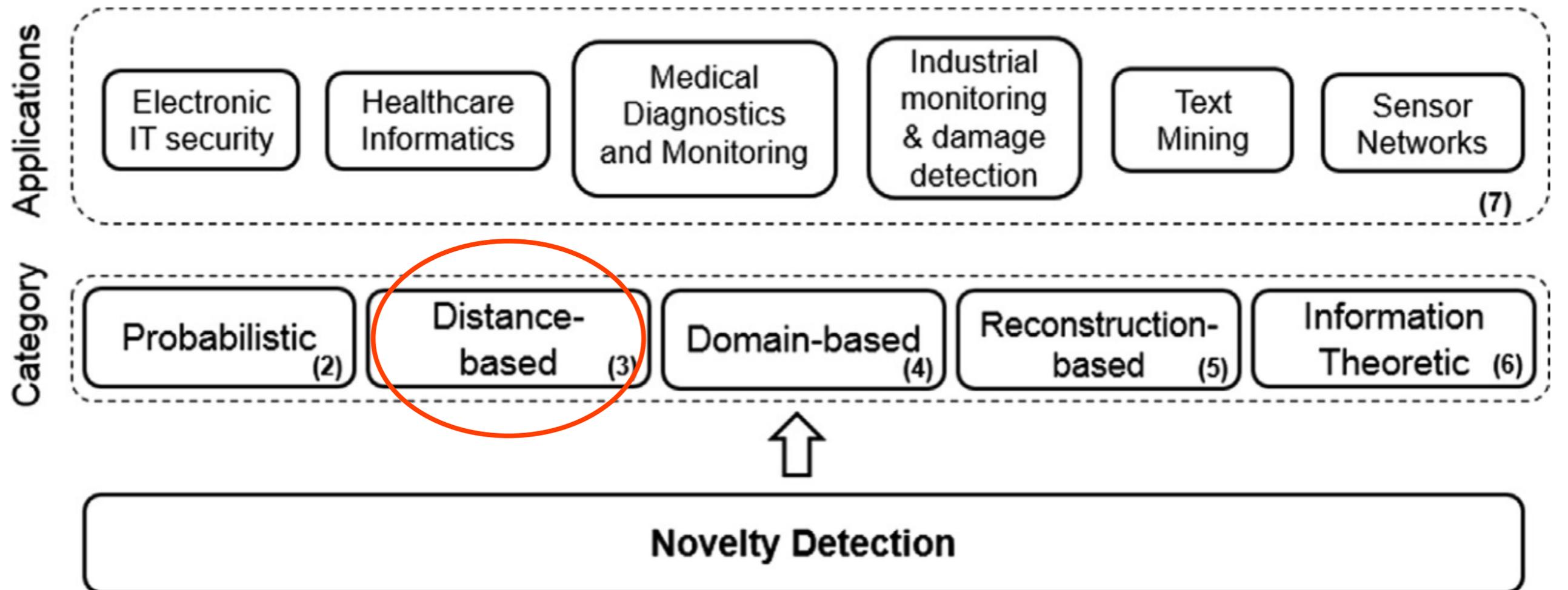
Allow us to detect new physics model-independently!

The history of novelty detection is basically a history of developing novelty evaluators or evaluation approaches



[Pimentel, Clifton, Tarassenko, 2014]

The history of novelty detection is basically a history of developing novelty evaluators or evaluation approaches



[Pimentel, Clifton, Tarassenko, 2014]

Traditional Novelty Measure

[H. Kriegel, P. Kroger, E. Schubert, and A. Zimek, 2009]

[R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, 2013]

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

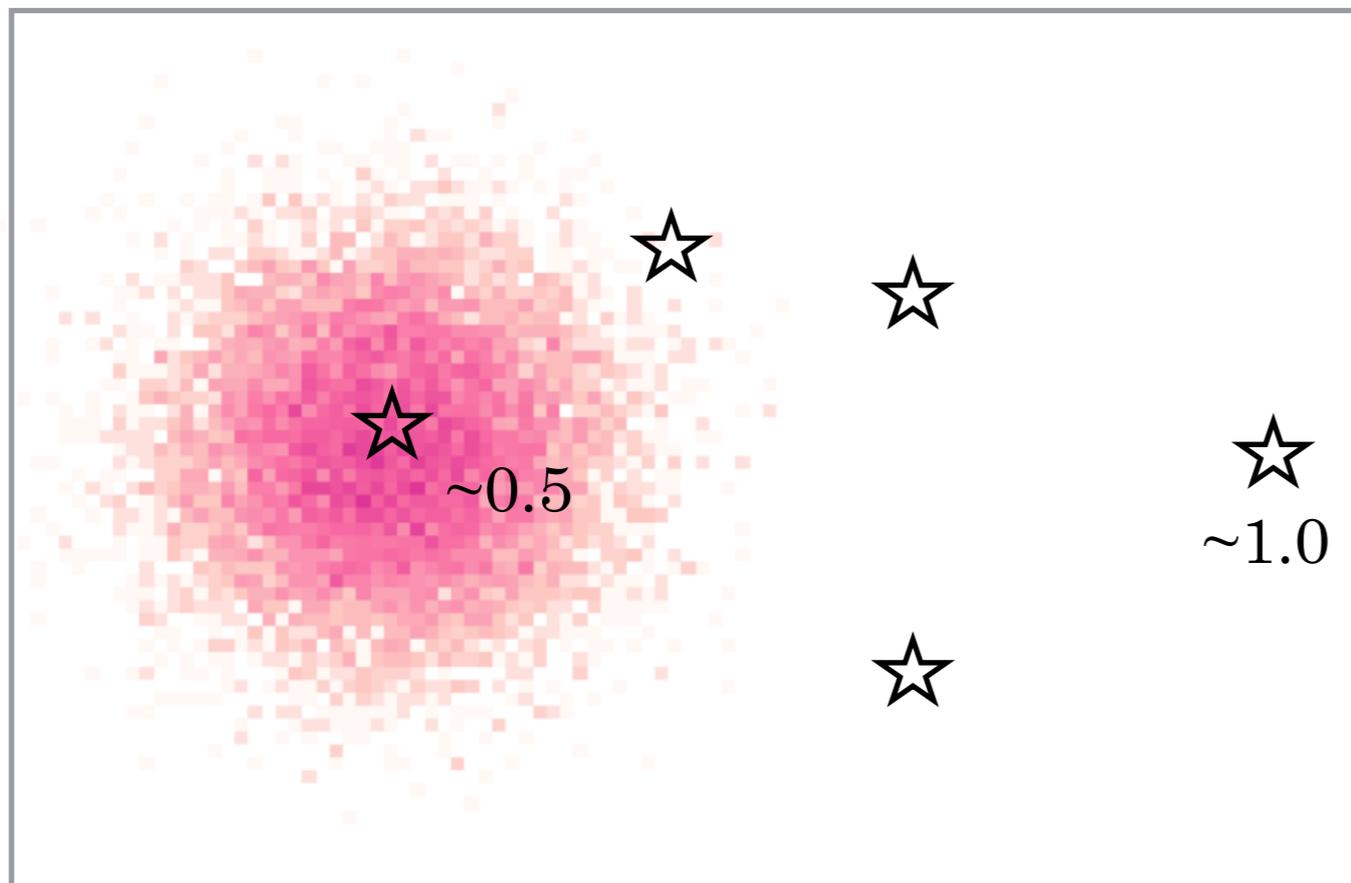
- d_{train} : mean distance of a testing data point to its k nearest neighbors
- $\langle d'_{\text{train}} \rangle$: average of the mean distances defined for its k nearest neighbors
- $\langle d'^2_{\text{train}} \rangle^{1/2}$: standard deviation of the latter
- All quantities are defined wrt the training dataset

Traditional Novelty Measure

[H. Kriegel, P. Kroger, E. Schubert, and A. Zimek, 2009]

[R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, 2013]

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



a measure of isolation from the training sample

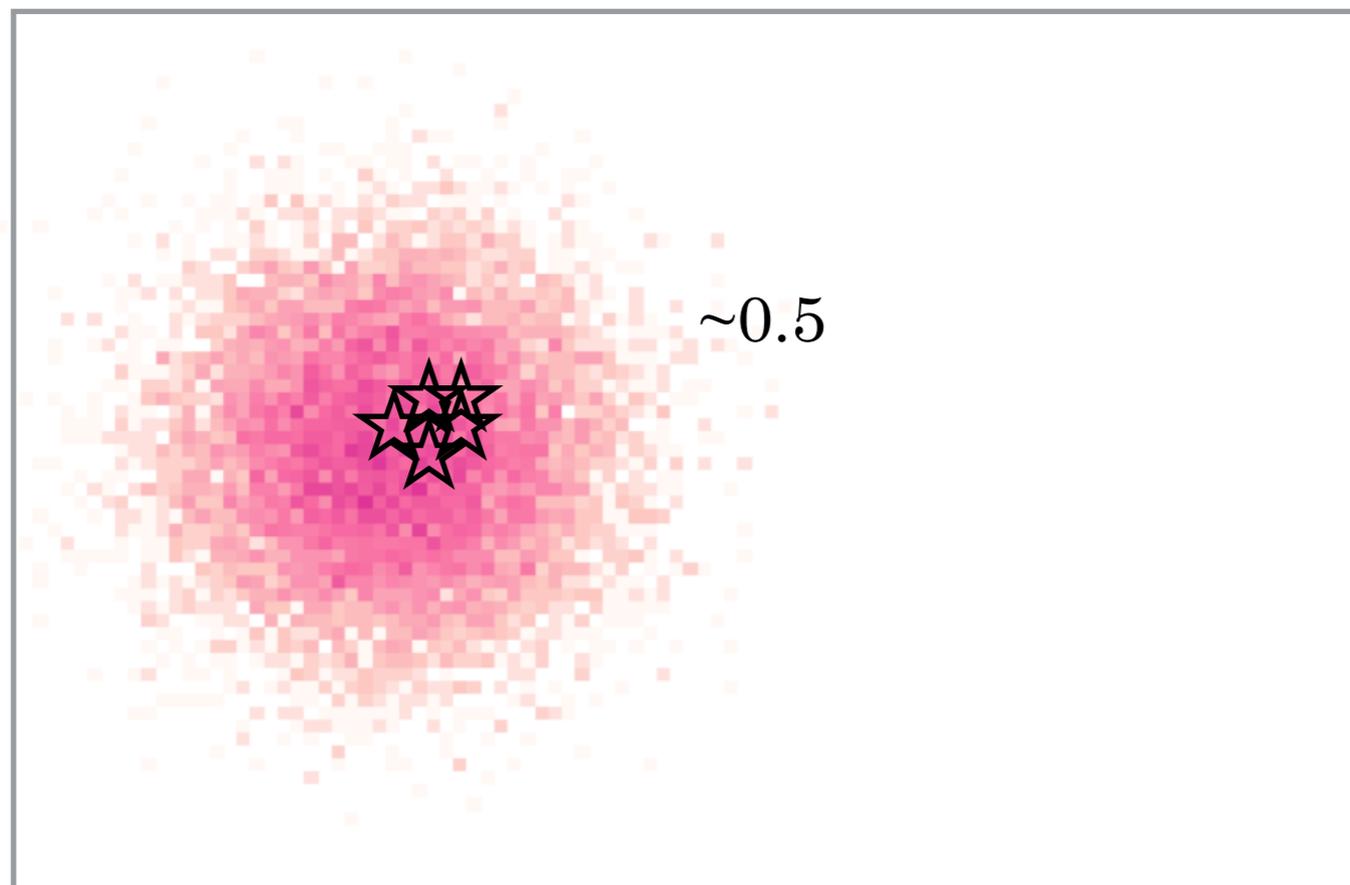
Traditional Novelty Measure

[H. Kriegel, P. Kroger, E. Schubert, and A. Zimek, 2009]

[R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, 2013]

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

$$\mathcal{O} = \frac{1}{2} \left(1 + \text{erf} \left(\frac{c\Delta}{\sqrt{2}} \right) \right)$$



- However, this design is insensitive to the clustering of the testing data with unknown pattern
- Recall: the clustering features such as resonance, shape, etc., could be important for BSM physics detection
- The testing data of unknown pattern with such features are scored low, unless they are away from the training data distribution!

New Novelty Measure

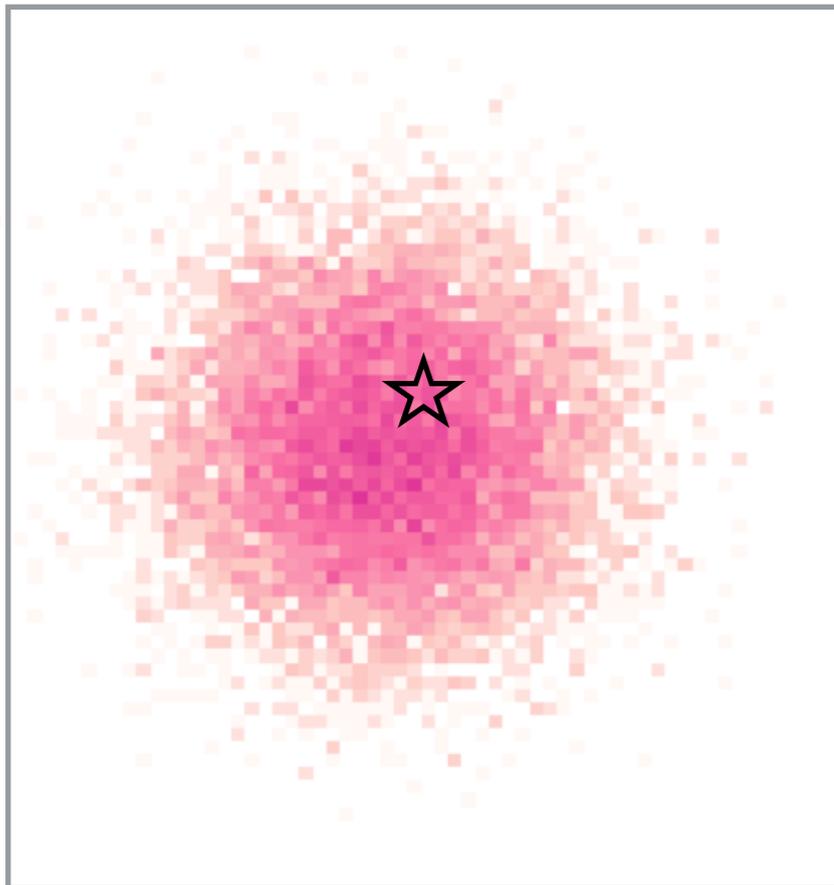
$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}} \quad \Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$

- d_{train} : mean distance of a testing data point to its k nearest neighbors in the training dataset
- d_{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 datasets
- Approximately statistical interpretation : $\Delta_{\text{new}} \propto \frac{S}{\sqrt{B}} \Big|_{\text{local bin}}$

New Novelty Measure

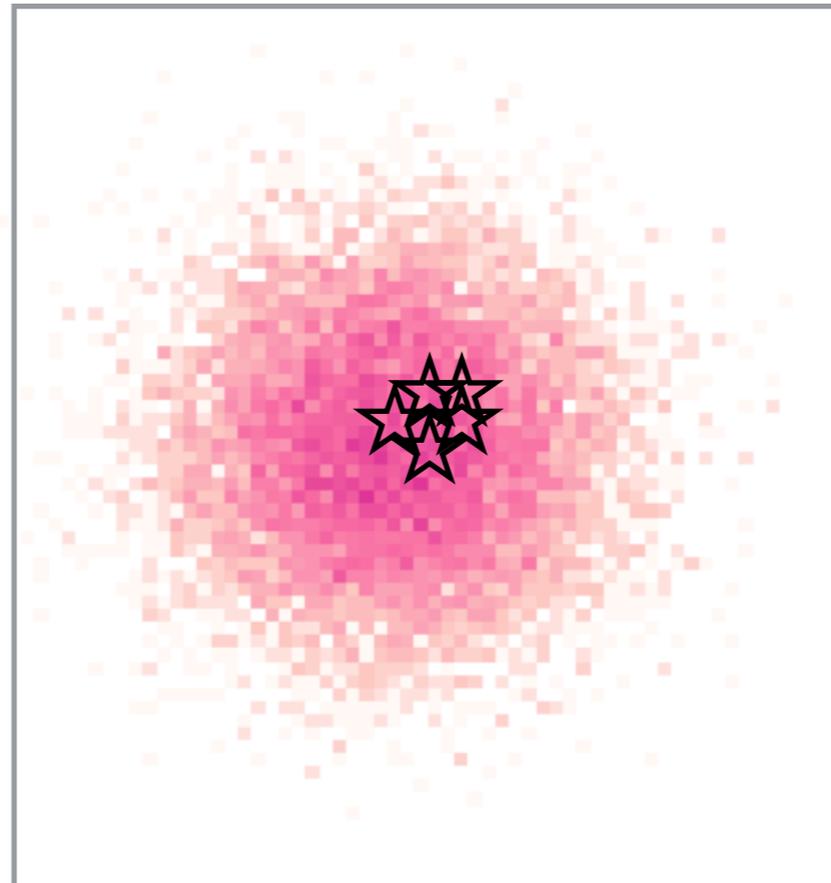
$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}}$$

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$



Training dataset

VS



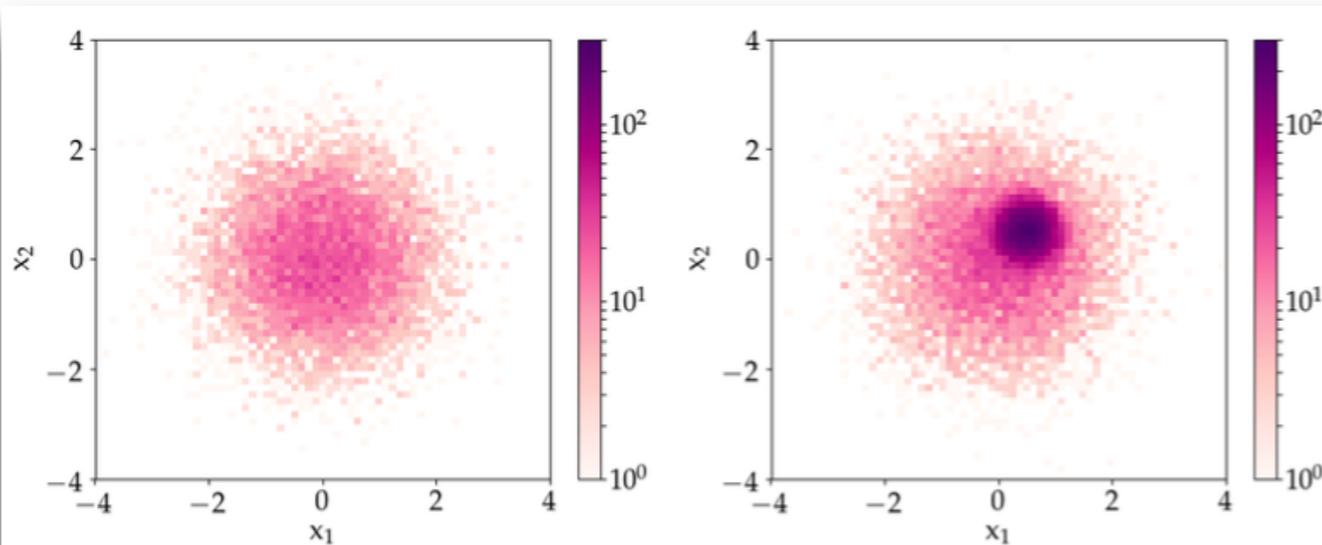
Testing dataset

- Big density difference => high score
- Small density difference => low score
- => **a measure of clustering**

New Novelty Measure

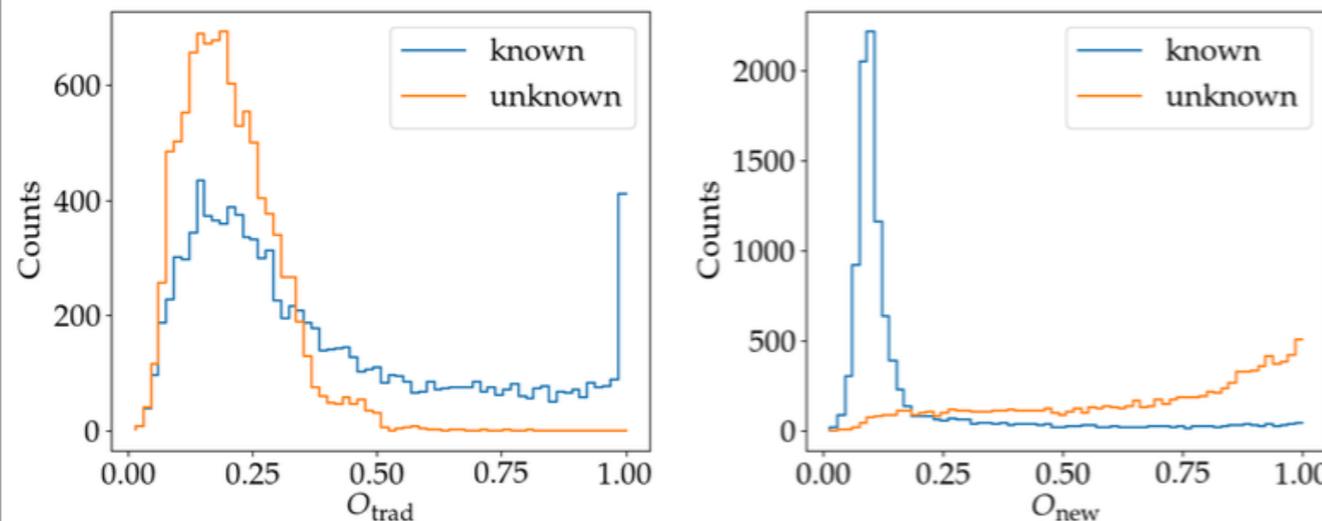
$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}}$$

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$



(a) Training data.

(b) Testing data.



(c) O_{trad} performance.

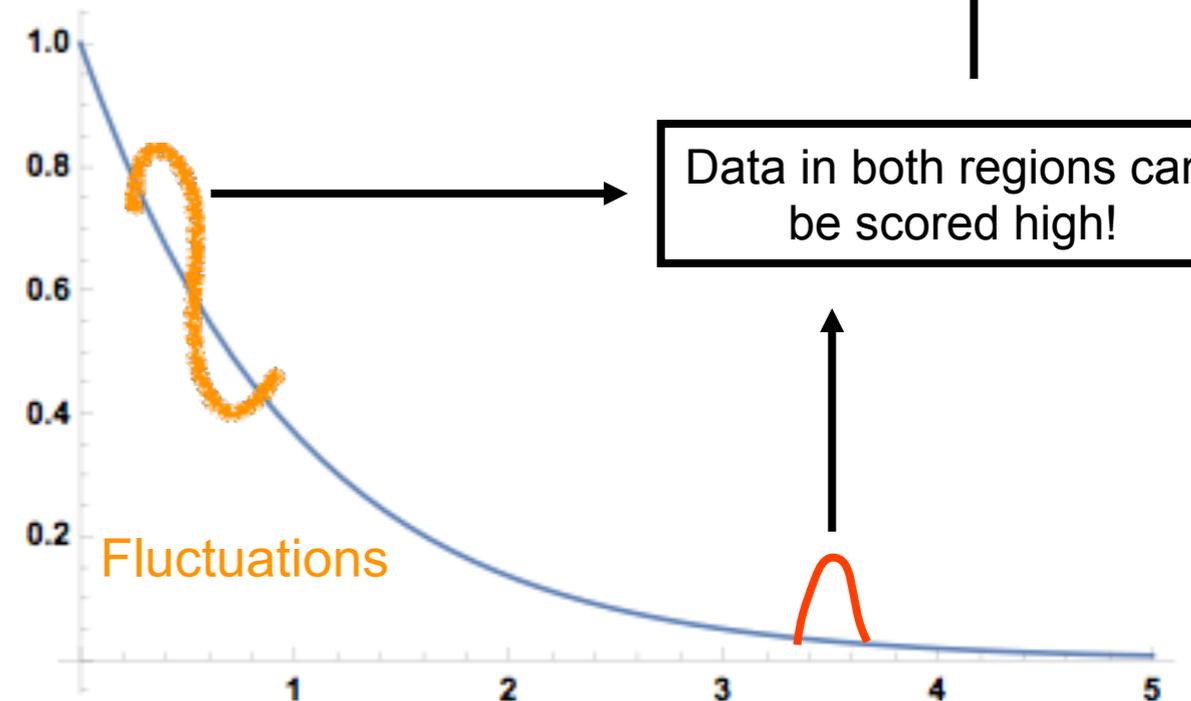
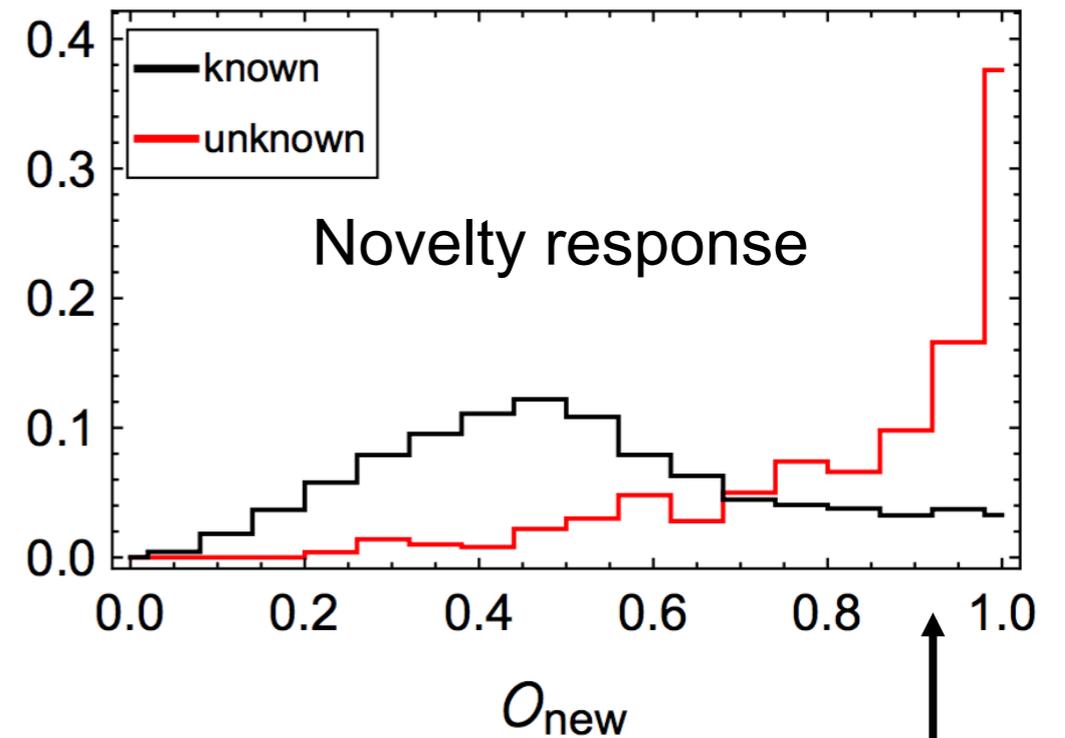
(d) O_{new} performance.

- Consider 2D Gaussian samples
- Training dataset: known pattern only
- Testing dataset: known + unknown patterns
- Compared to O_{trad} , the novelty response of unknown-pattern data is much stronger for O_{new}
- => A well-separation between the known- and unknown-pattern data distributions

Look Elsewhere Effect

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$

Without a priori knowledge on the BSM physics, novelty detection might suffer from a large “Look Elsewhere Effect (LEE)”, given the feature space to probe!



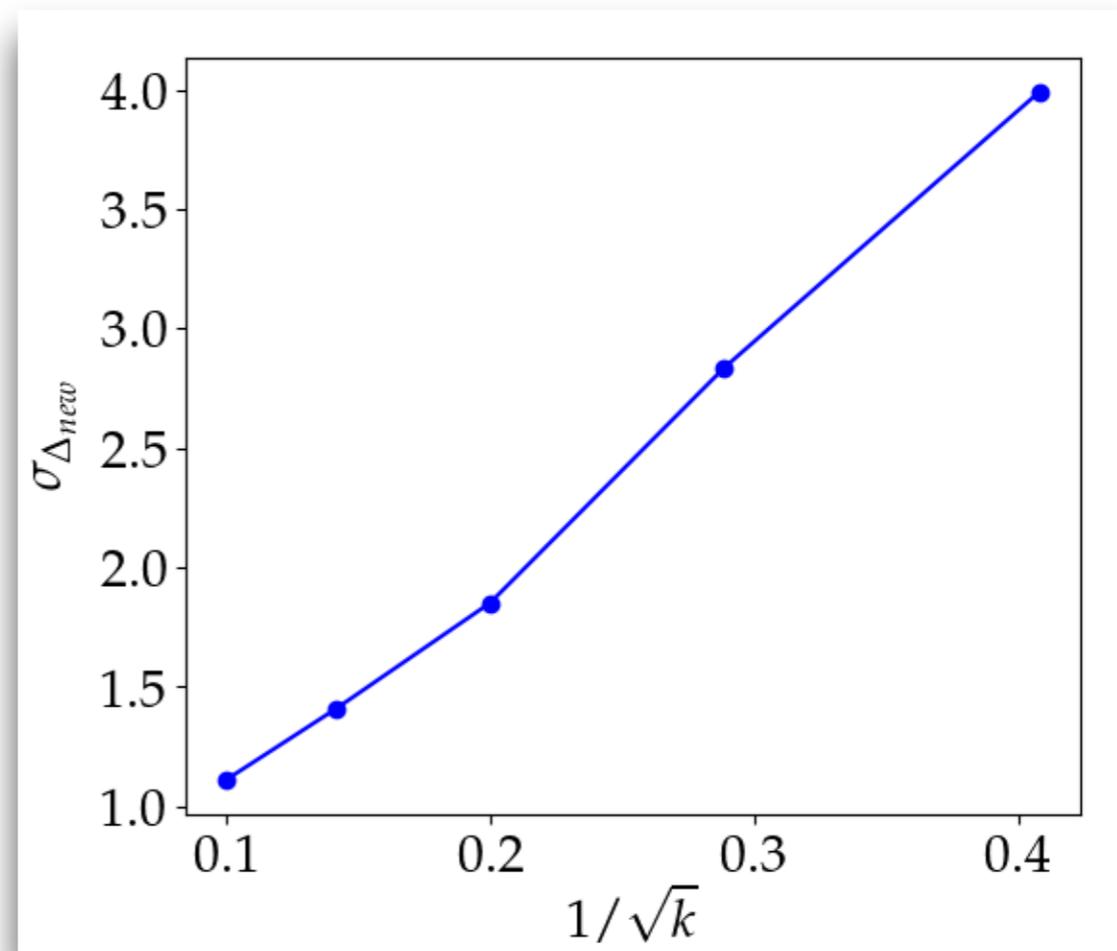
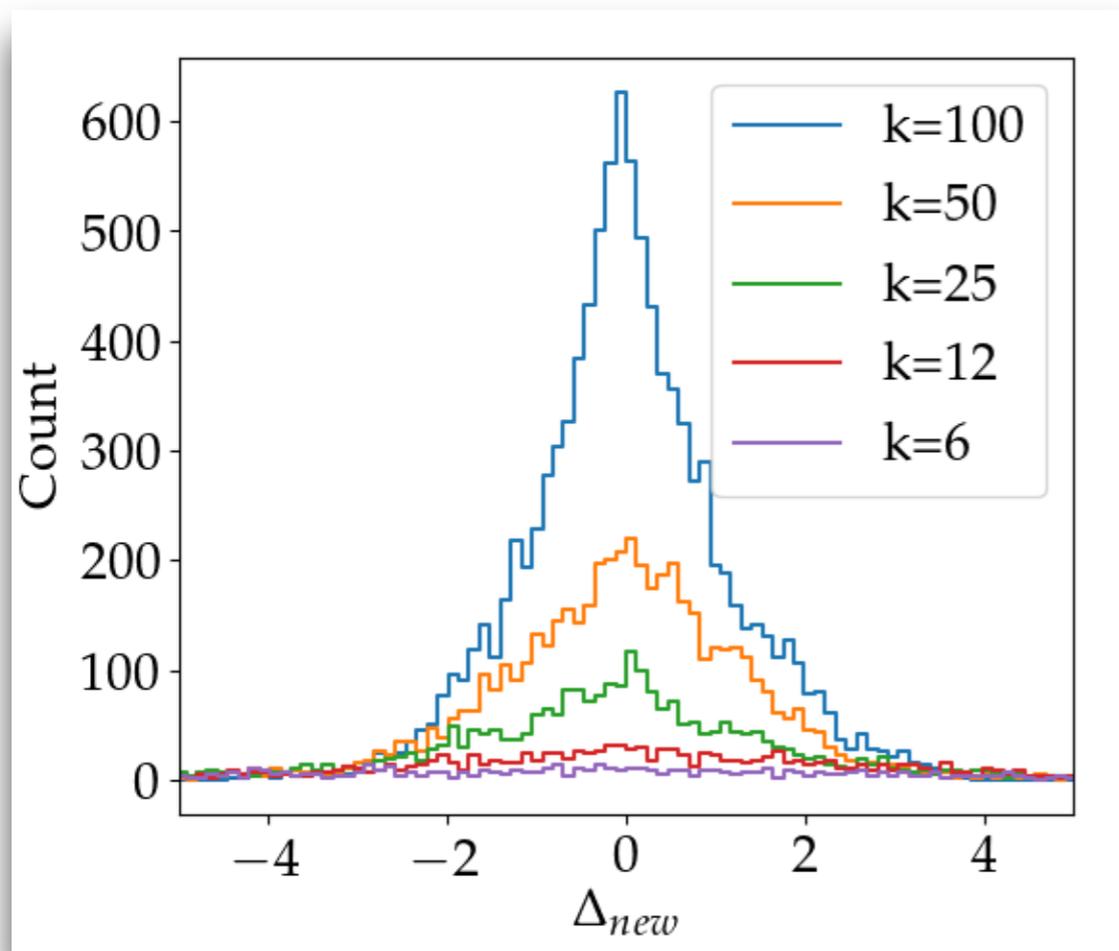
Strategies to Address LEE

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$

The influence of fluctuations for detection sensitivity can be compensated for as the luminosity L increases, if k scales with L .

Central Limit Theorem

The standard deviation of Δ_{new} scales with $1/\sqrt{k}$ or $1/\sqrt{L}$, for the testing data with known patterns only.



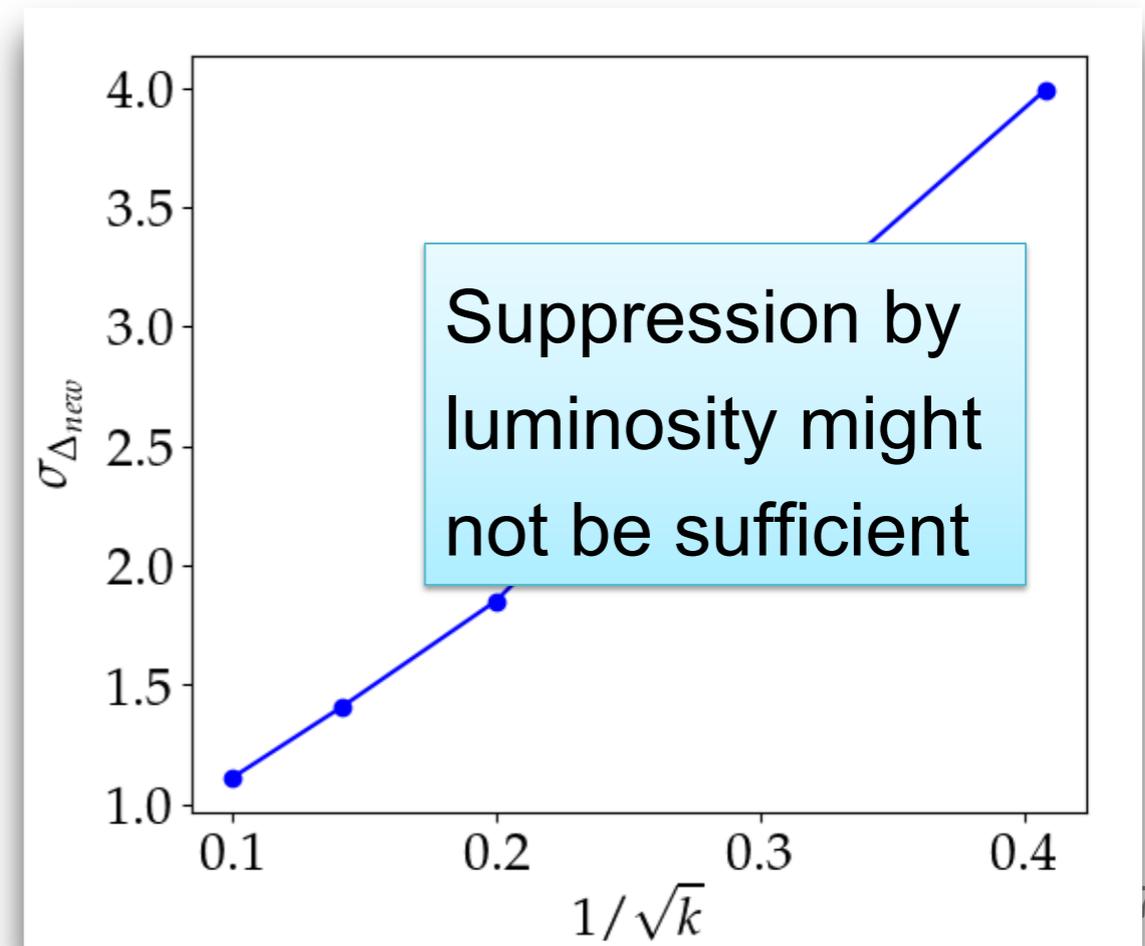
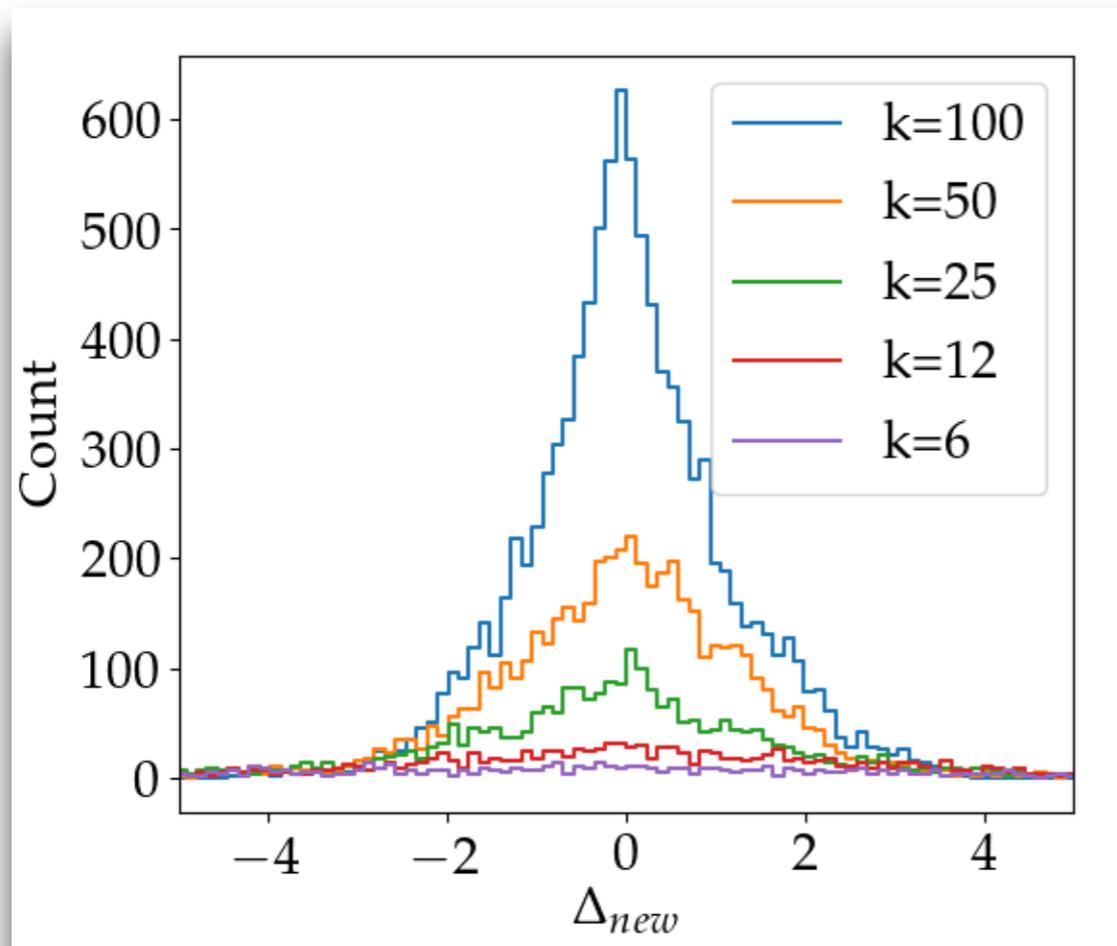
Strategies to Address LEE

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}$$

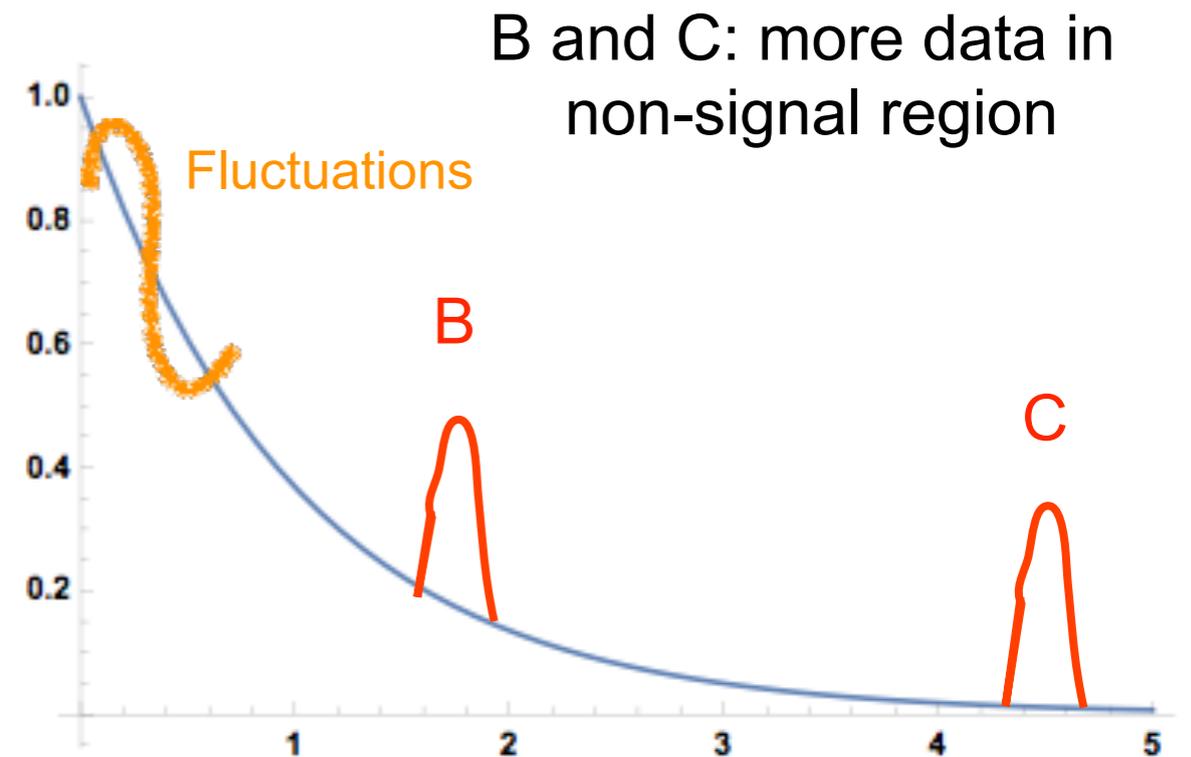
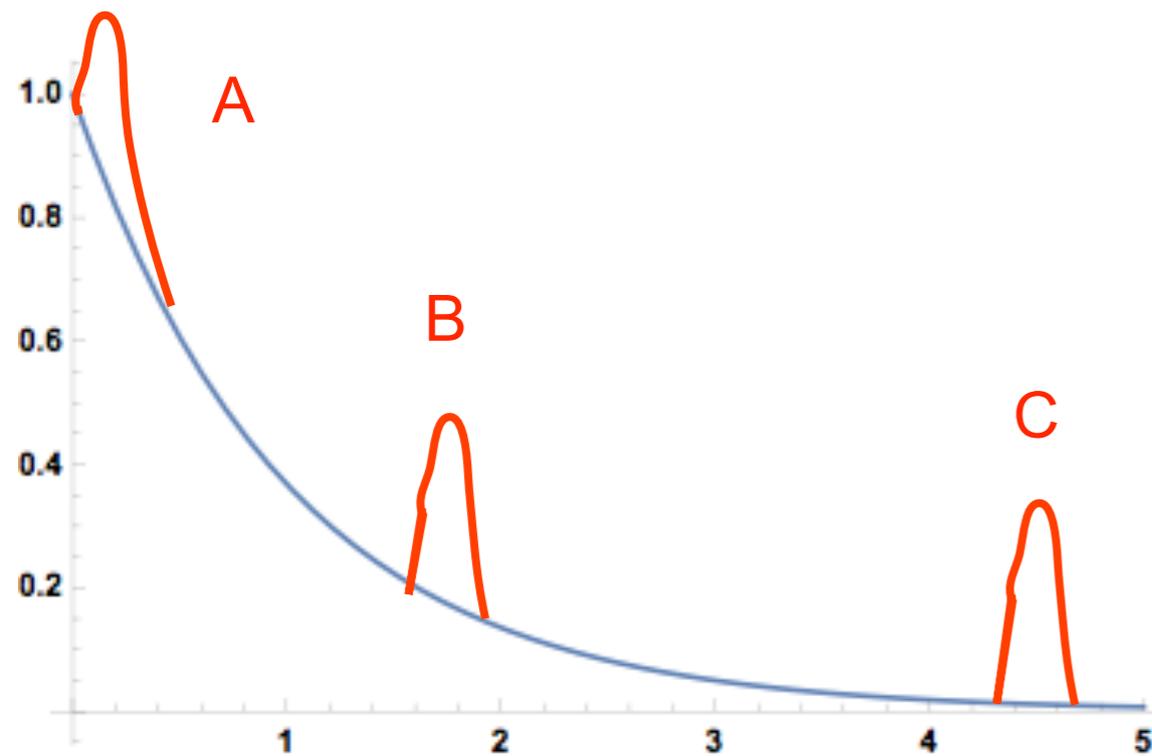
The influence of fluctuations for detection sensitivity can be compensated for as the luminosity L increases, if k scales with L .

Central Limit Theorem

The standard deviation of Δ_{new} scales with $1/\sqrt{k}$ or $1/\sqrt{L}$, for the testing data with known patterns only.



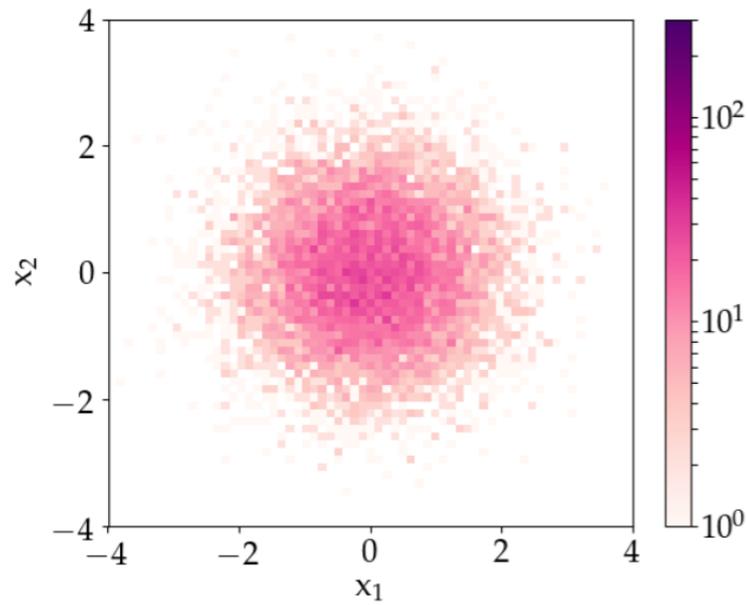
Strategies to Address LEE



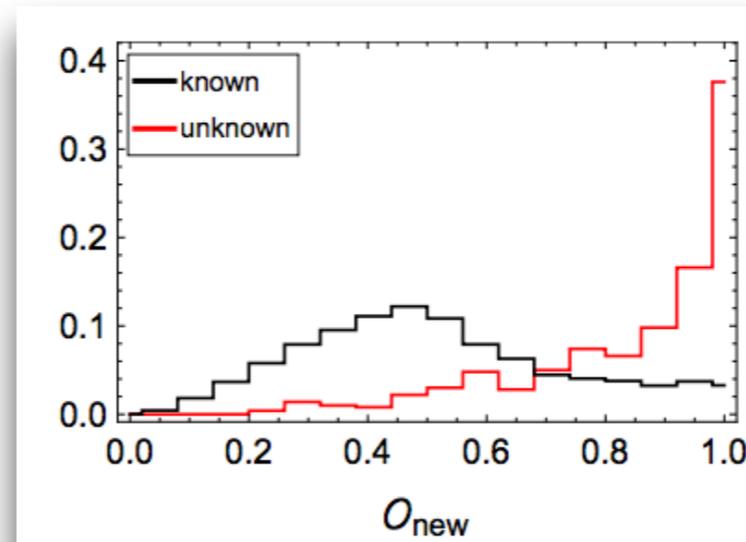
To compensate for high-scoring (by \mathcal{O}_{new}) of known-pattern data from high-density region

$$\Rightarrow \mathcal{O}_{\text{comb}} = \sqrt{\mathcal{O}_{\text{trad}} \mathcal{O}_{\text{new}}}$$

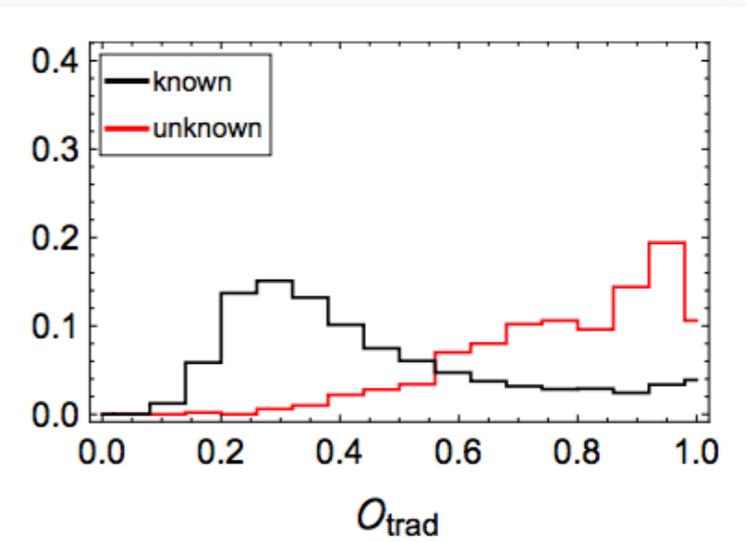
Strategies to Address LEE



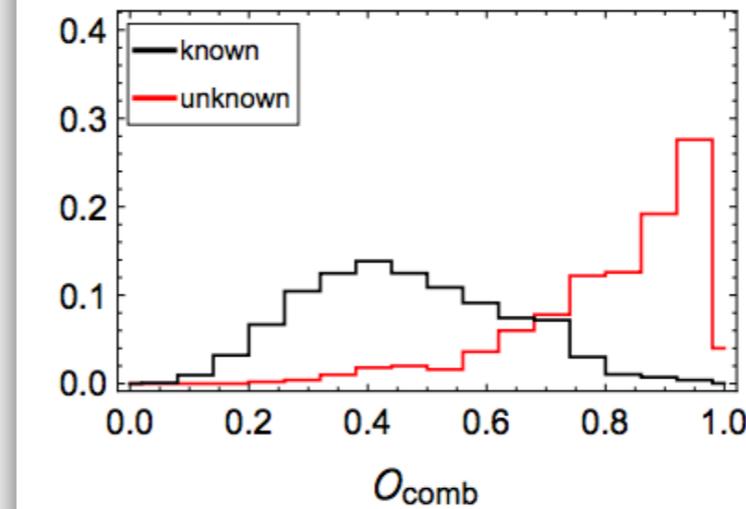
(a) Training data.



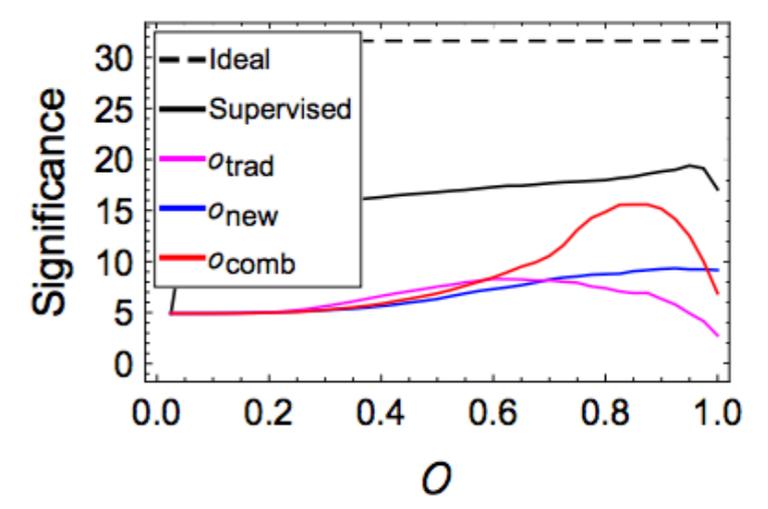
(a) New evaluator.



(b) Traditional evaluator.

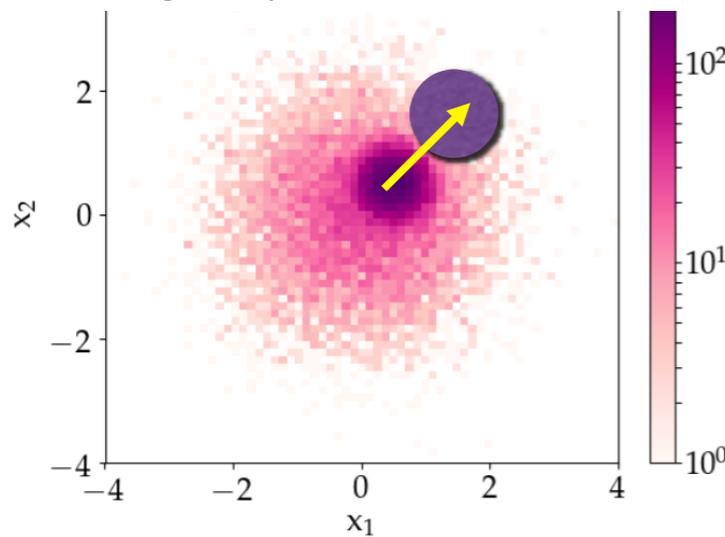


(c) Combined evaluator.



(d) Significance.

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



(b) Testing data.

O_{comb} based analysis yields more than 50% improvement in detection sensitivity!

Benchmark Analysis

Analysis one: di-top (leptonic) production at LHC (the SM cross sections have been scaled by a factor 1/2000, for simplification)

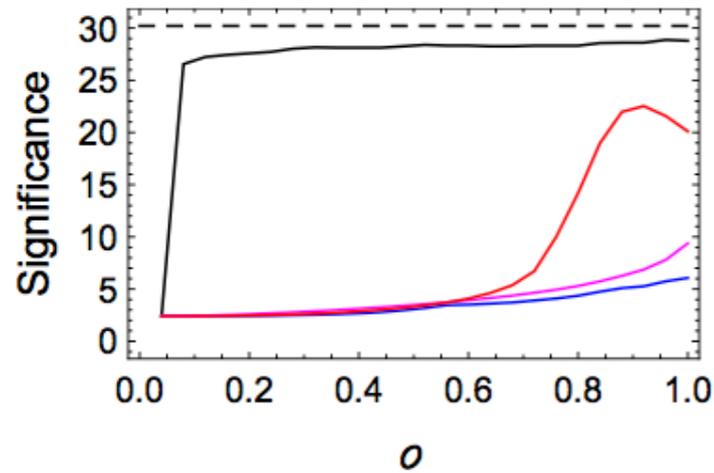
- $pp \rightarrow \bar{t}_l t_l$, $\sigma = 11.5 \text{ fb}$, $\mathbf{X}_1: pp \rightarrow \bar{T}T \rightarrow W_l^+ W_l^- \bar{b}b$
- $pp \rightarrow t_l \bar{b} W_l^\pm$, $\sigma = 0.365 \text{ fb}$,
- $pp \rightarrow Z_b Z_l$, $\sigma = 0.0765 \text{ fb}$. $\mathbf{X}_2: pp \rightarrow Z' \rightarrow \bar{t}t$

Analysis two: exotic Higgs decays at e+e- collider

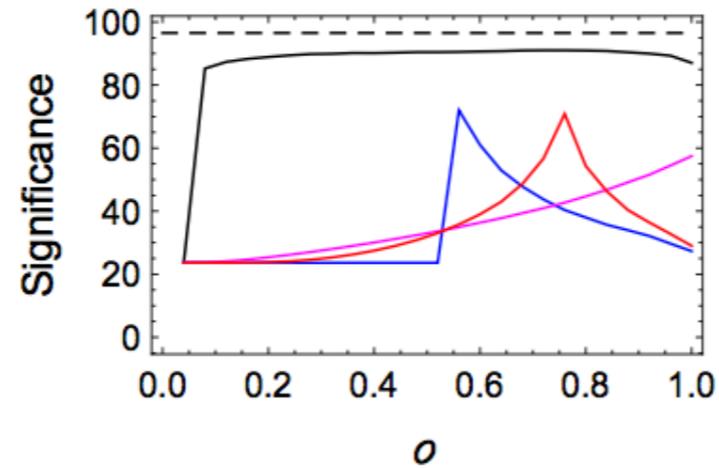
- $e^+e^- \rightarrow hZ \rightarrow Z_{\text{inv}}^* Z_{\bar{b}b} l^+ l^-$ $\sigma = 0.00686 \text{ fb}$, $\mathbf{Y}_1: h \rightarrow \tilde{\chi}_1 \tilde{\chi}_2 \rightarrow \tilde{\chi}_1 \tilde{\chi}_1 a$.
- $e^+e^- \rightarrow hZ \rightarrow Z_{\bar{b}b}^* Z_{\text{inv}} l^+ l^-$ $\sigma = 0.00259 \text{ fb}$. $\mathbf{Y}_2: h \rightarrow Za$

	Parameter values	$\sigma(\text{fb})$
X1	$m_T = m_{\bar{T}} = 1.2 \text{ TeV}$, $\text{BR}(T \rightarrow W_l^+ b) = 50\%$	0.152
X2	$m_{Z'} = 3 \text{ TeV}$, $g_{Z'} = g_Z$, $\text{BR}(Z' \rightarrow \bar{t}t) = 16.7\%$	1.55
Y1	$m_{N_1} = \frac{m_{N_2}}{9} = \frac{m_a}{4} = 10 \text{ GeV}$, $\text{BR}(h \rightarrow \bar{b}b E_T^{\text{miss}}) = 1\%$	0.108
Y2	$m_a = 25 \text{ GeV}$, $\text{BR}(h \rightarrow \bar{b}b E_T^{\text{miss}}) = 1\%$	0.053

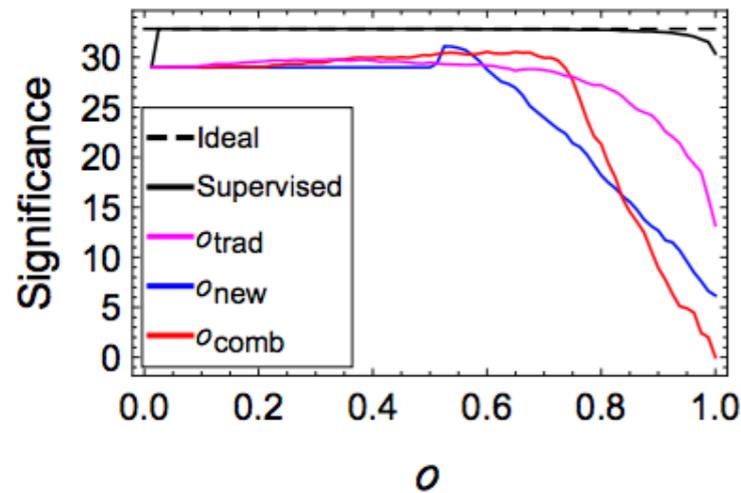
Benchmark Analysis



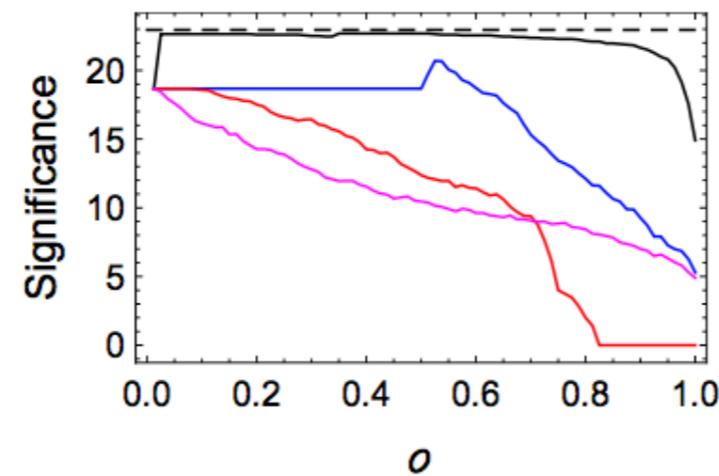
(a) Benchmark: X_1



(b) Benchmark: X_2



(c) Benchmark: Y_1



(d) Benchmark: Y_2

- X_1 : well-modeled by the Gaussian sample!
- X_2 : O_{comb} less efficient due to one-order larger S/B
- X_3 and X_4 : O_{new} performs universally better than the others, due to large S/B
- The sensitivities based on the algorithm designed are not far below the ones set by supervised learning

Summary and Outlook

- Proposed workflow for novelty detection
- New novelty evaluators are proposed to address the clustering and LLE effect;

Follow-up study to parton showers and full detector simulations.

Interesting questions to study:

- Understand which features are captured when auto-encoder doing dimension reduction;
- Invent a novelty evaluator to exploit multiple measures at once
- ...

in collaboration with Jiang, Juste Rozas, Liu and Tu;

Thank you!

