``Machine Learning for Jet Physics'', Nov 2018, FNAL

Novelty Detection Meets Collider Physics

Tao Liu

The Hong Kong University of Science and Technology

Based on arXiv: 1807.10261

in collaboration with Jan Hajer, Ying-Ying Li and He Wang

Signal Processing 99 (2014) 215–249

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.

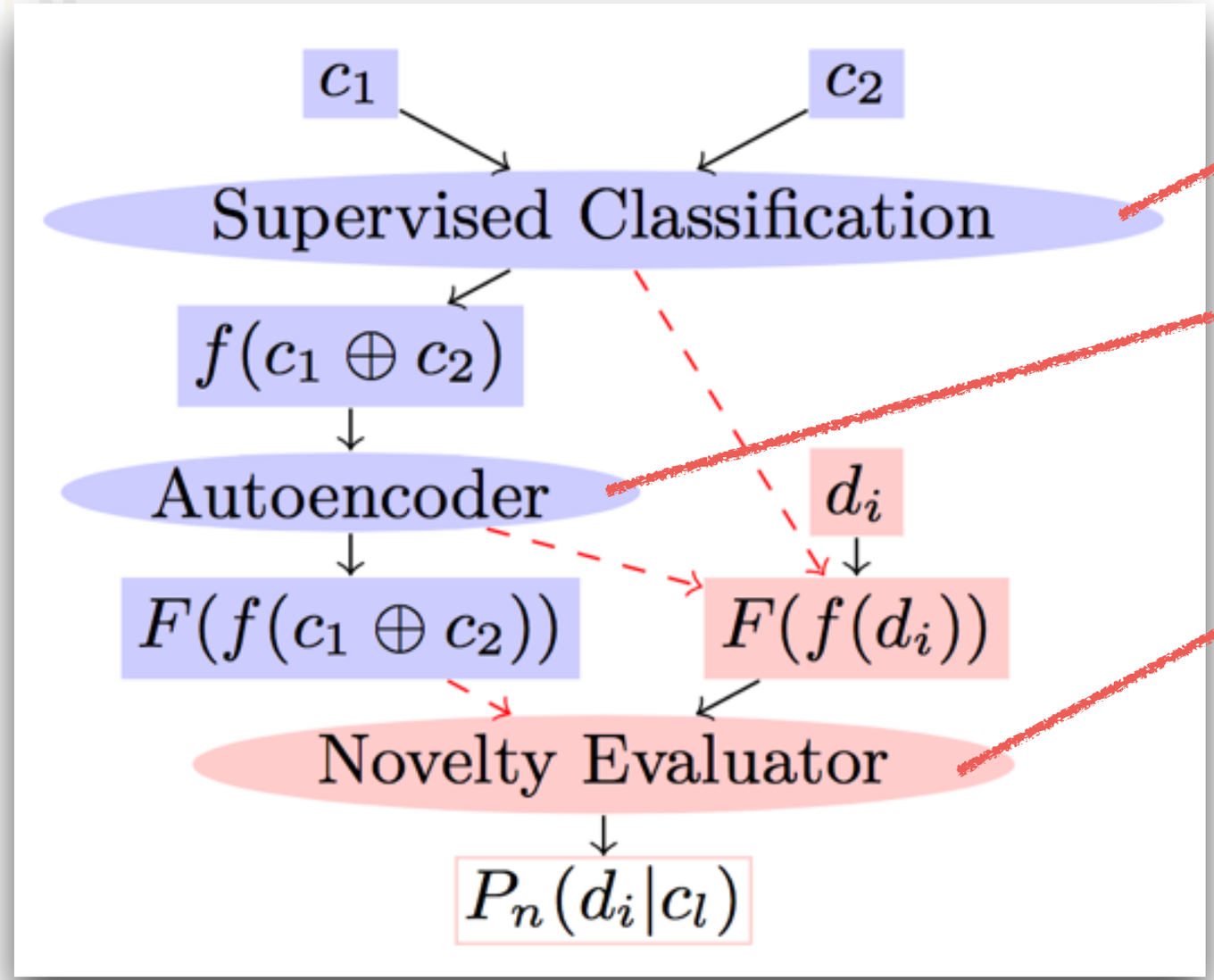
Novelty Detection

- The detection targets/signals are of unknown patterns.
- No a priori knowledge about them is applied to machine training.

=> Novelty detection is ``target'' or ``model''-independent, different from supervised learning!

- If defining, e.g., BSM physics, as the target, we can search for BSM physics model-independently, using the DNN techniques of novelty-detection
- Especially useful if we don't know what we are searching for

Workflow

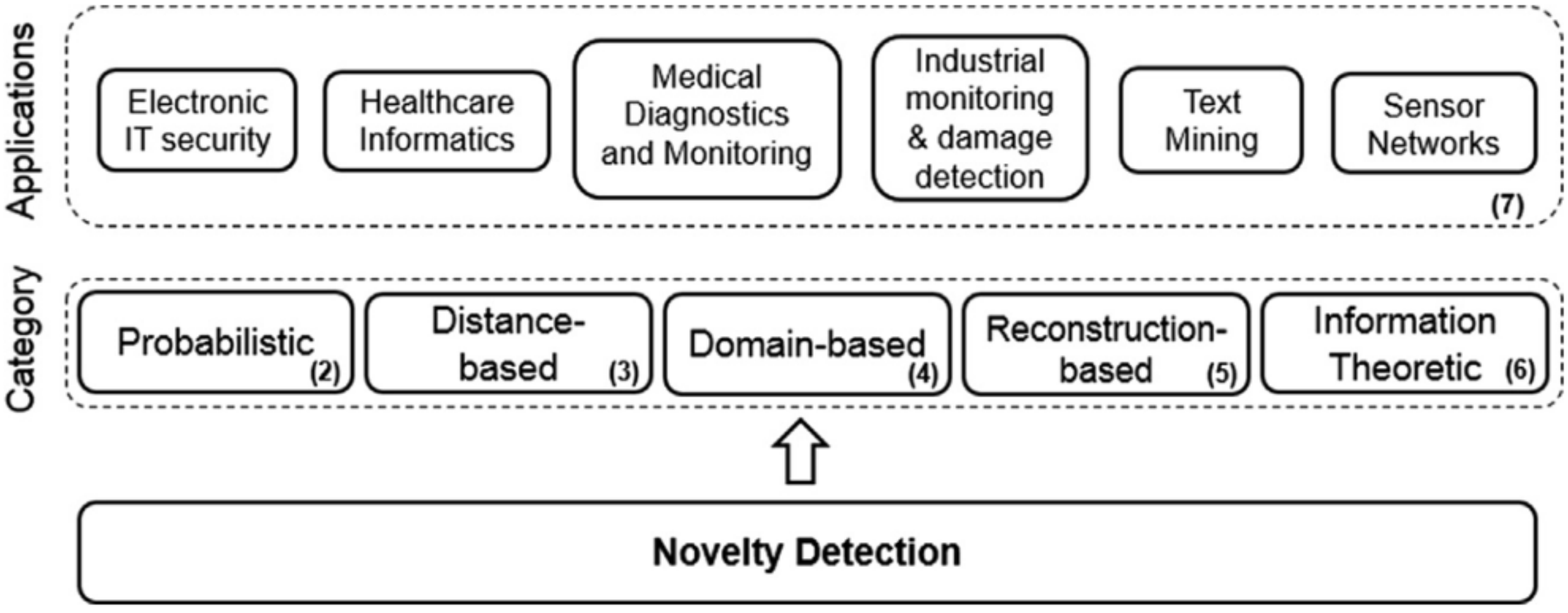


- Step 1: (SM/background) feature learning
- Step 2: dimension reducing of feature space (auto-encoder)
- Step 3: novelty evaluating of testing data
- Analyze detection sensitivity based on novelty response of testing data

With this algorithm, new physics can be searched for without a priori knowledge!

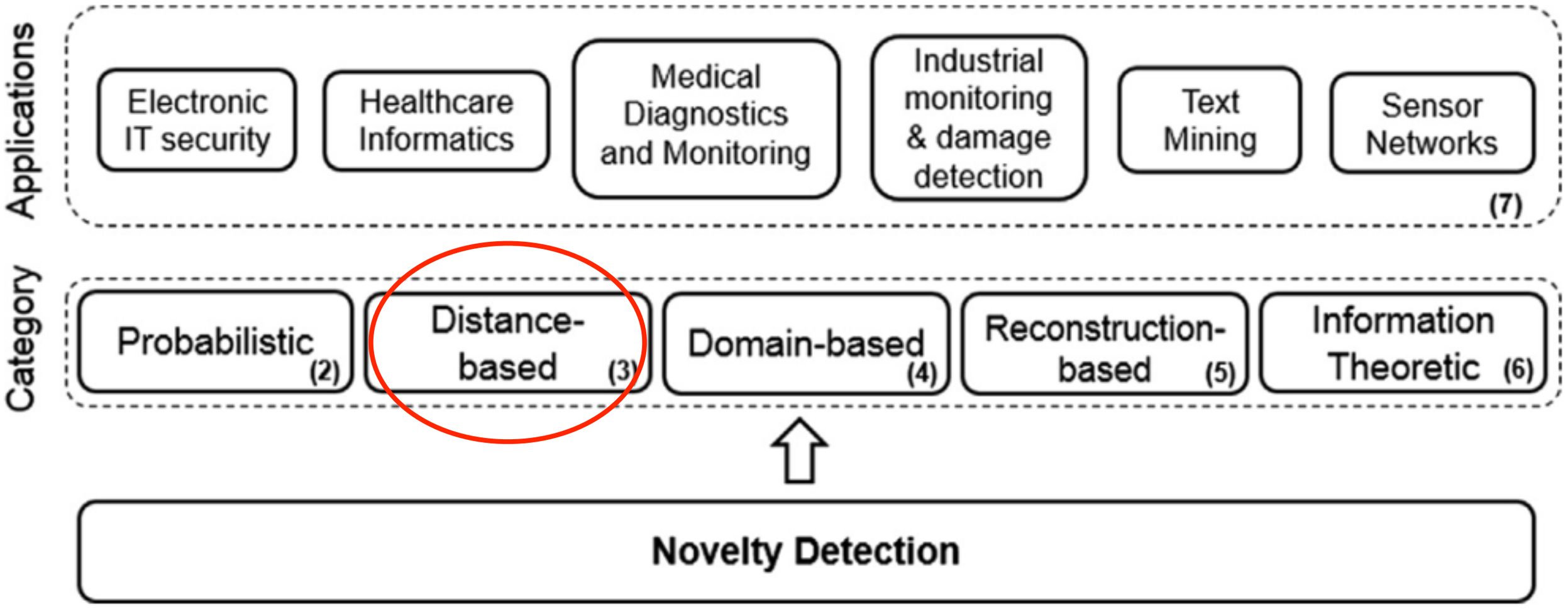
Novelty Evaluators - Application Driven

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



Novelty Evaluators - Application Driven

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



Novelty Evaluators: Traditional Wisdom

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

Novelty measure: range unnormalized

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

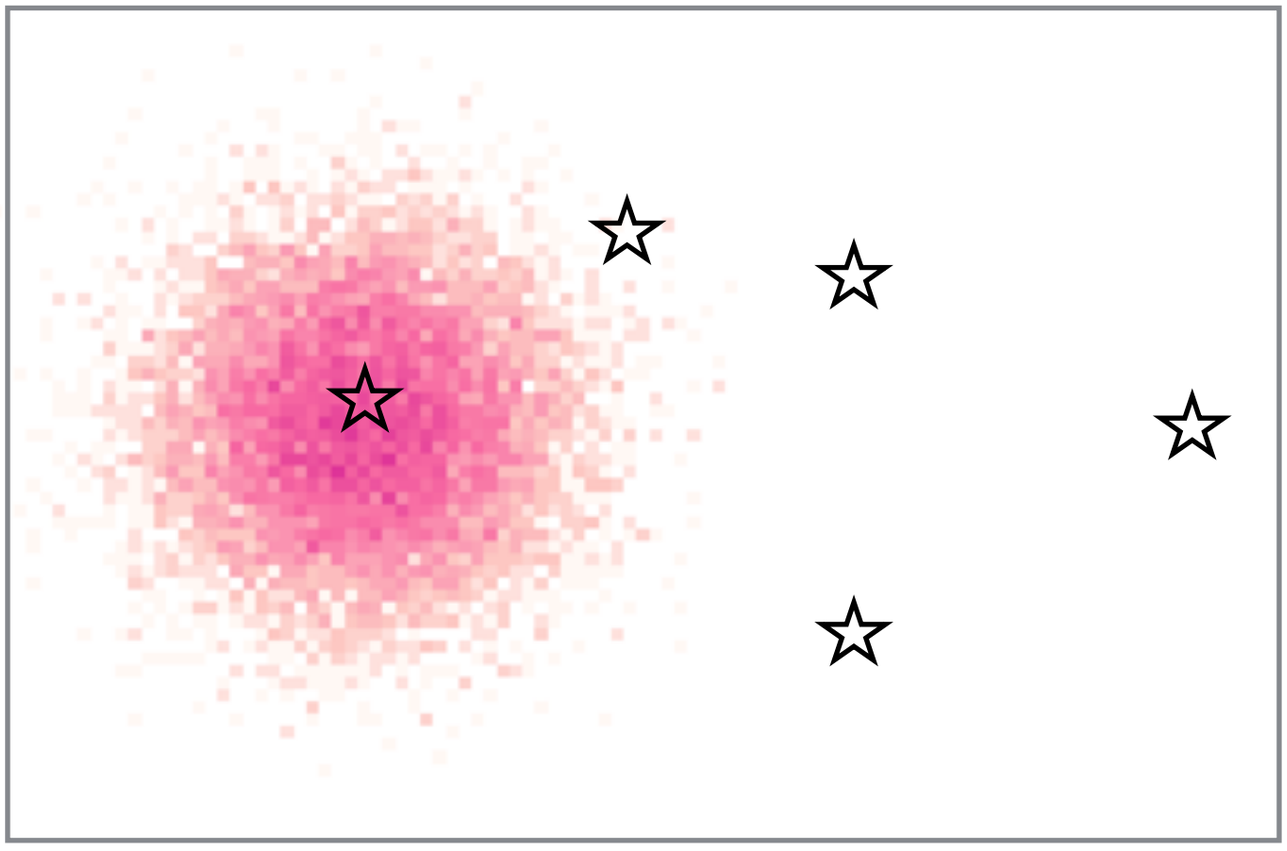
- d_{train} : mean distance of a testing data point to its k nearest neighbors
- $\langle d'_{\text{train}} \rangle$: average of the mean distances defined for its k nearest neighbors
- $\langle d'^2_{\text{train}} \rangle^{1/2}$: standard deviation of the latter
- All quantities are defined wrt the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

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

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

Novelty Evaluators: Traditional Wisdom

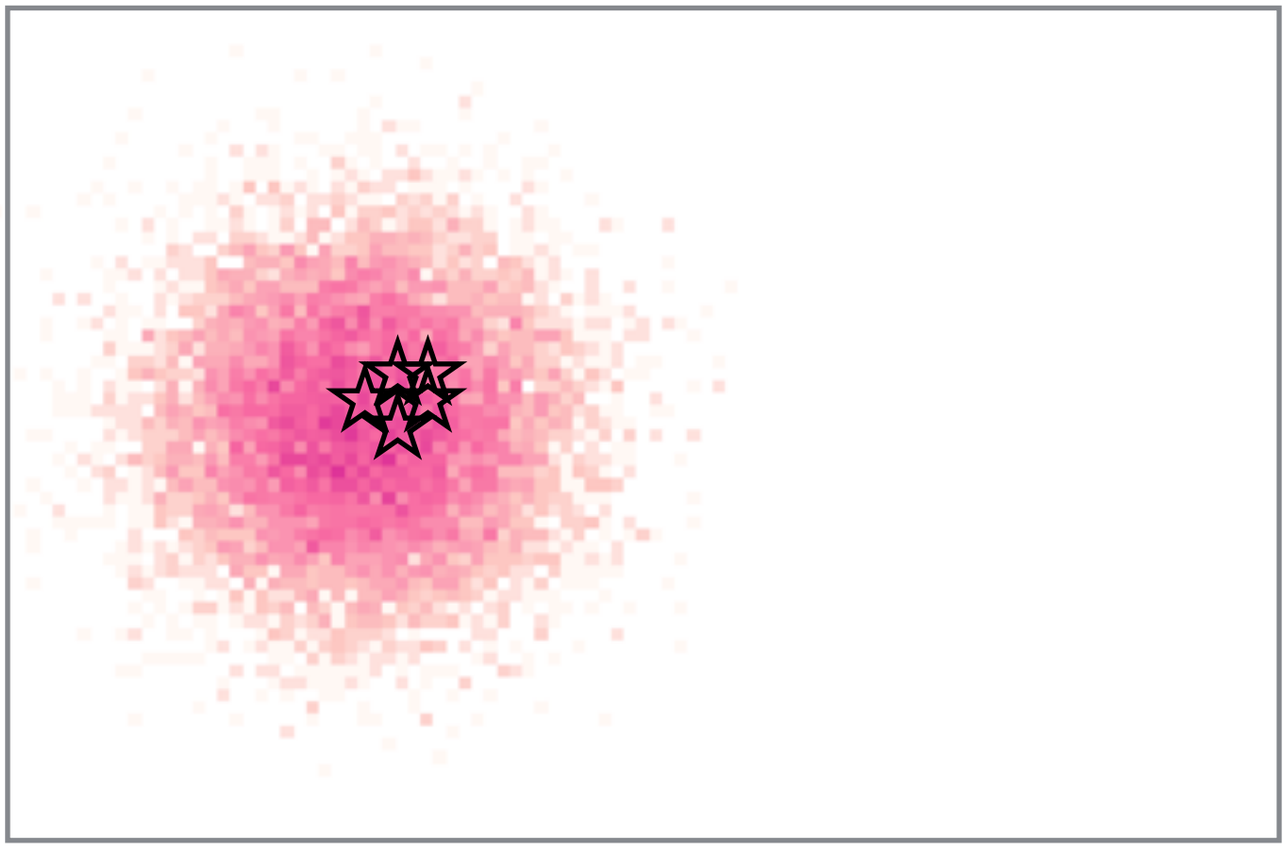
$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}} \qquad \mathcal{O} = \frac{1}{2}\left(1 + \text{erf}\left(\frac{c\Delta}{\sqrt{2}}\right)\right)$$



- Large distance => high score
- Short distance => low score
- => a measure of isolation
- Successful while being applied to, e.g., finger print or face recognition!

Novelty Evaluators: Traditional Wisdom

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



- 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!

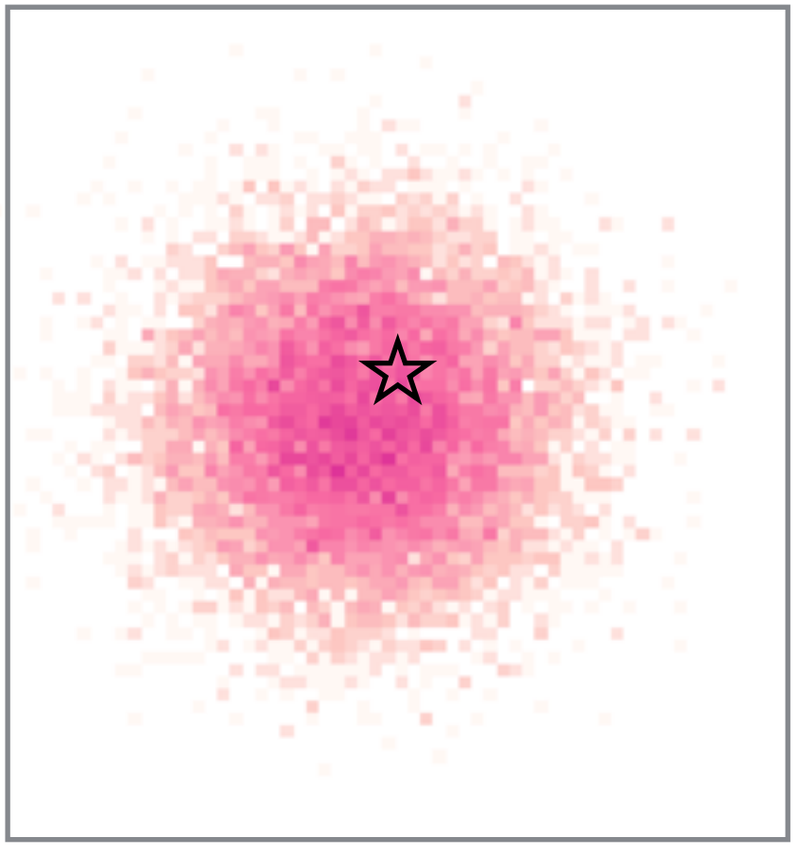
Novelty Evaluators: New Input

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

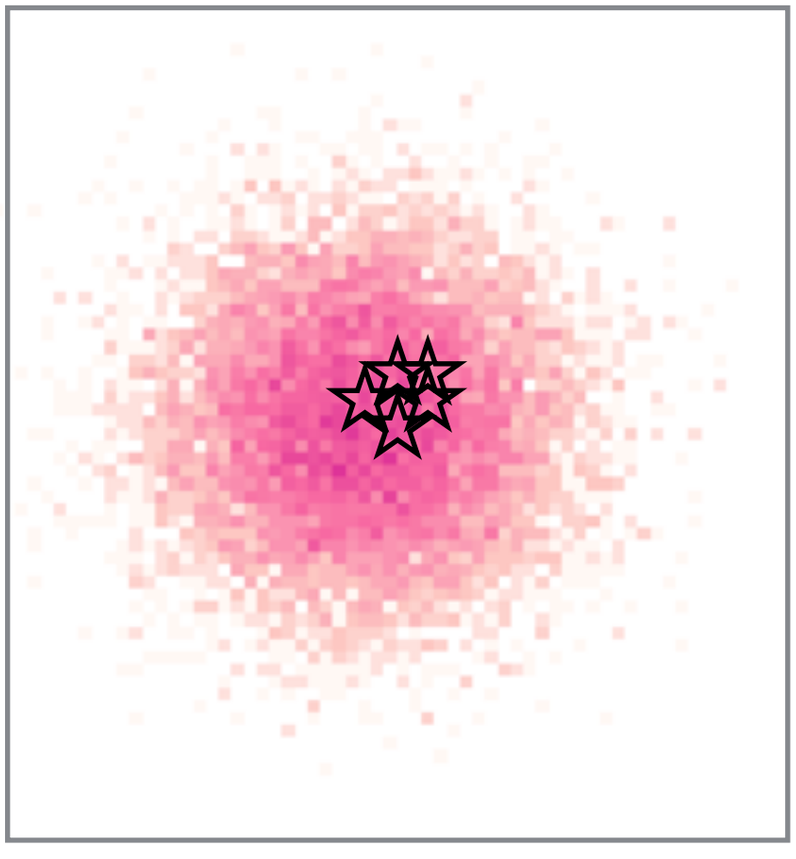
- d_{train}: mean distance of a testing data point to its k nearest neighbors in the training dataset
- d_{test} : mean distance of a testing data point to its k nearest neighbors in the testing dataset
- m: dimension of the feature space
- Novelty response is evaluated by comparing local densities of the testing point in the training and testing datasets
- Approximately statistical interpretation : $\Delta_{\text{new}} \propto \left. \frac{S}{\sqrt{B}} \right|_{\text{local bin}}$

Novelty Evaluators: New Input

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



VS

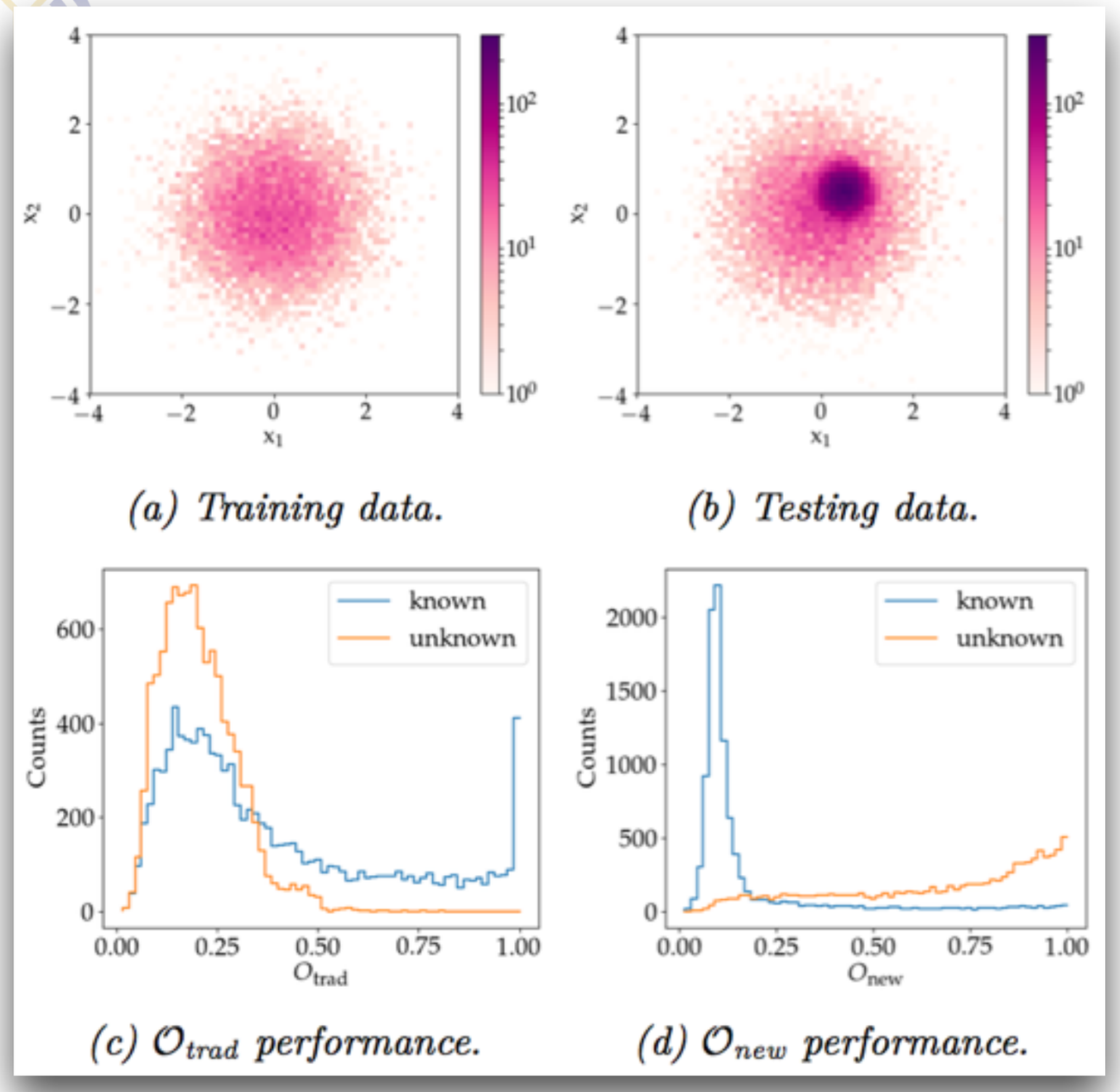


Training dataset

Testing dataset

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

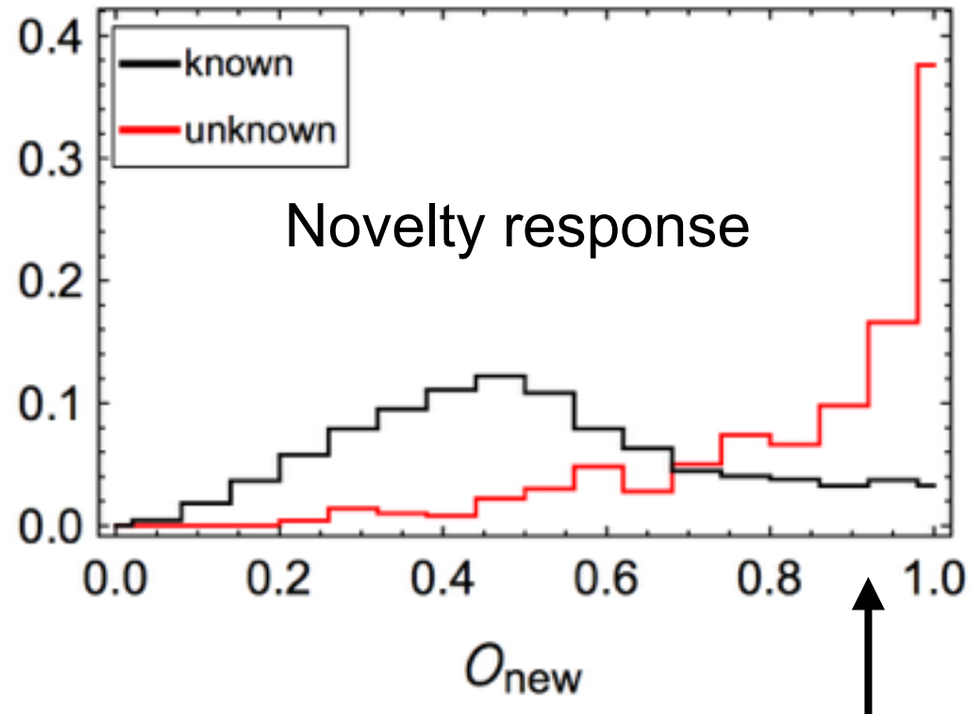
Novelty Evaluators: Performance Comparison

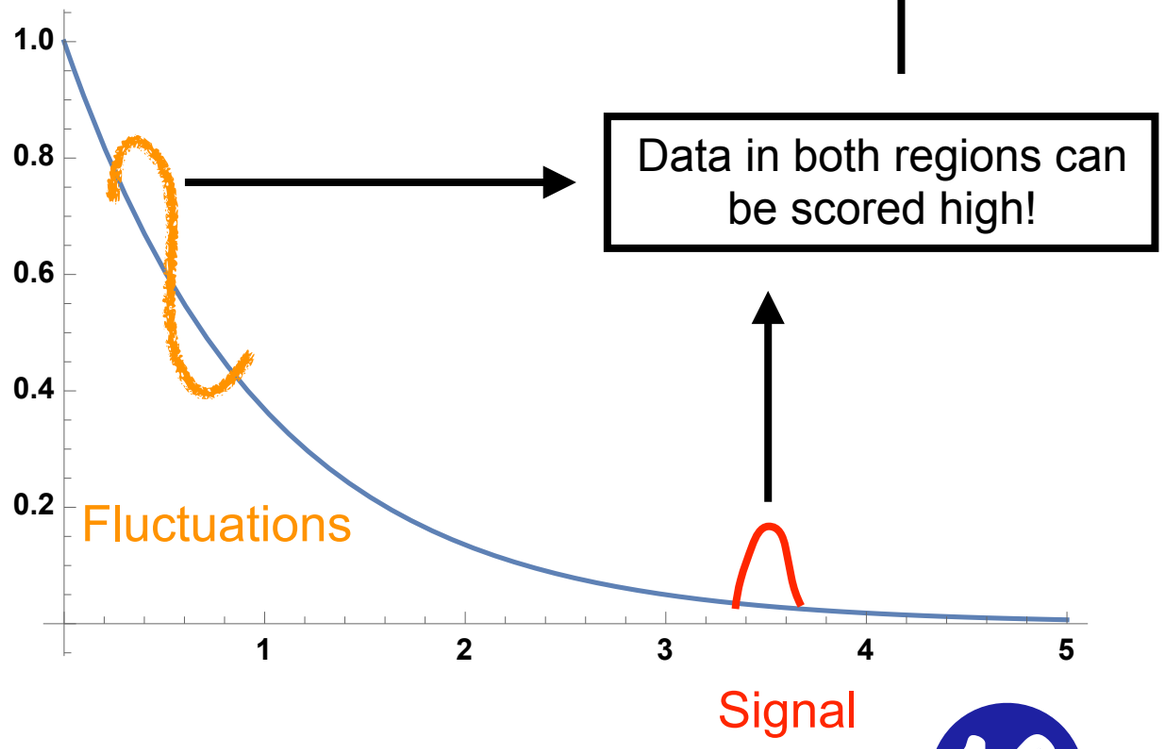


- Consider 2D Gaussian samples
- Training dataset: known pattern only
- Testing dataset: known + unknown patterns
- Compared to O_trad, the novelty response of unknown-pattern data is much stronger for O_new
- => A well-separation between the known- and unknown-pattern data distributions

``Look Elsewhere Effect''

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



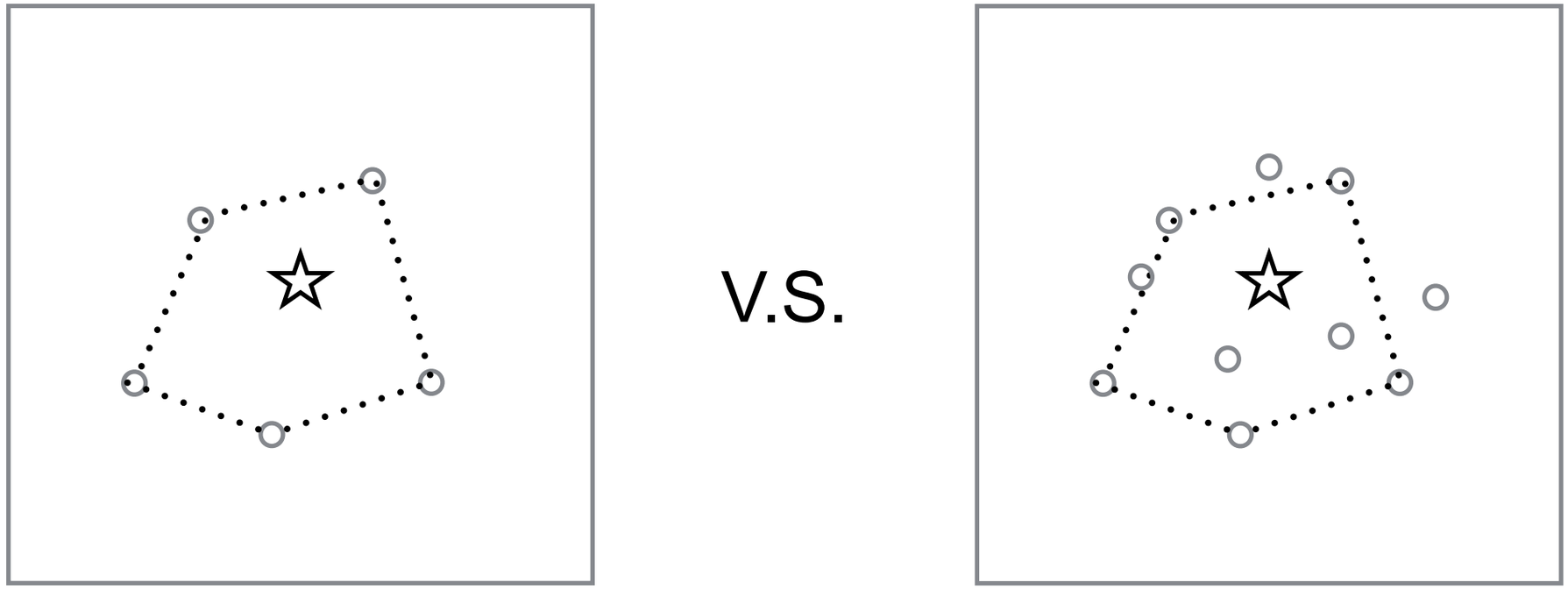


Without a priori knowledge on the BSM physics, novelty detection might suffer from a large ``Look Elsewhere Effect (LEE)'', given the feature space to probe!

``Look Elsewhere Effect'' - Central Limit Theorem

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

This can be understood since more and more data are used to calculate d_{test} in the local bin which is barely changed.



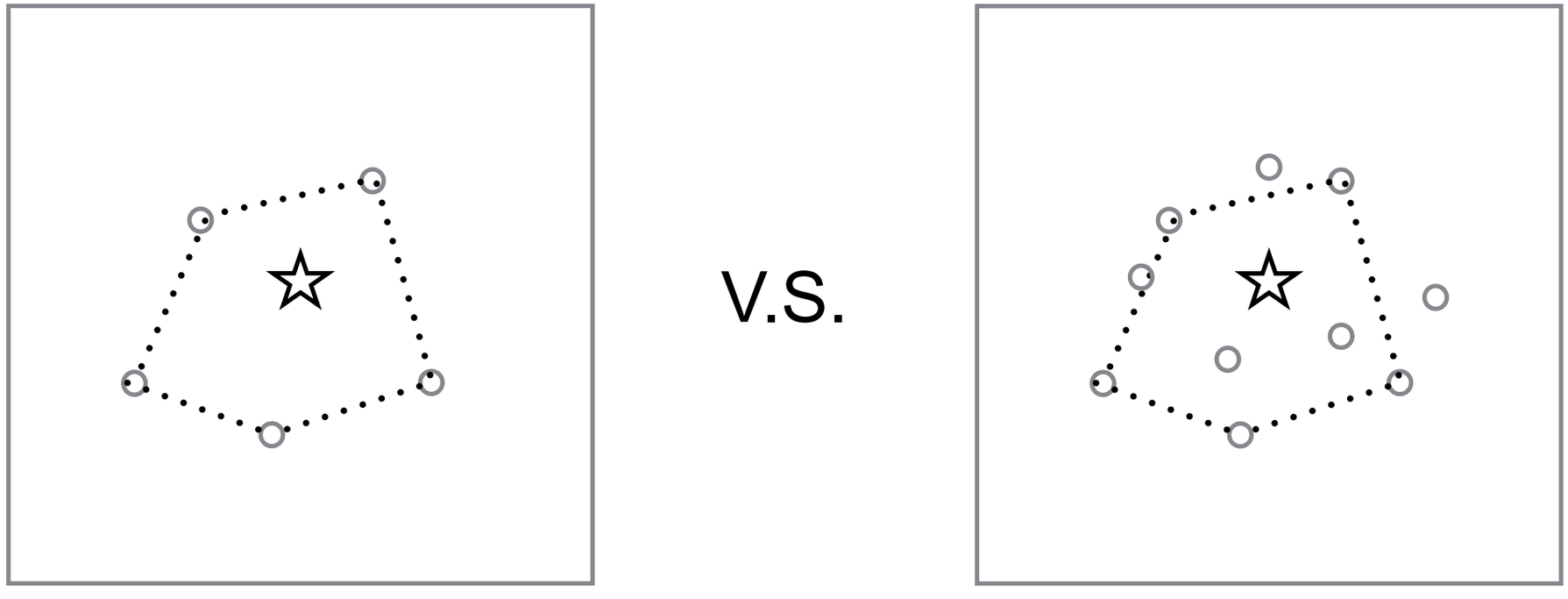
L

2 * L

``Look Elsewhere Effect'' - Central Limit Theorem

Central Limit Theorem

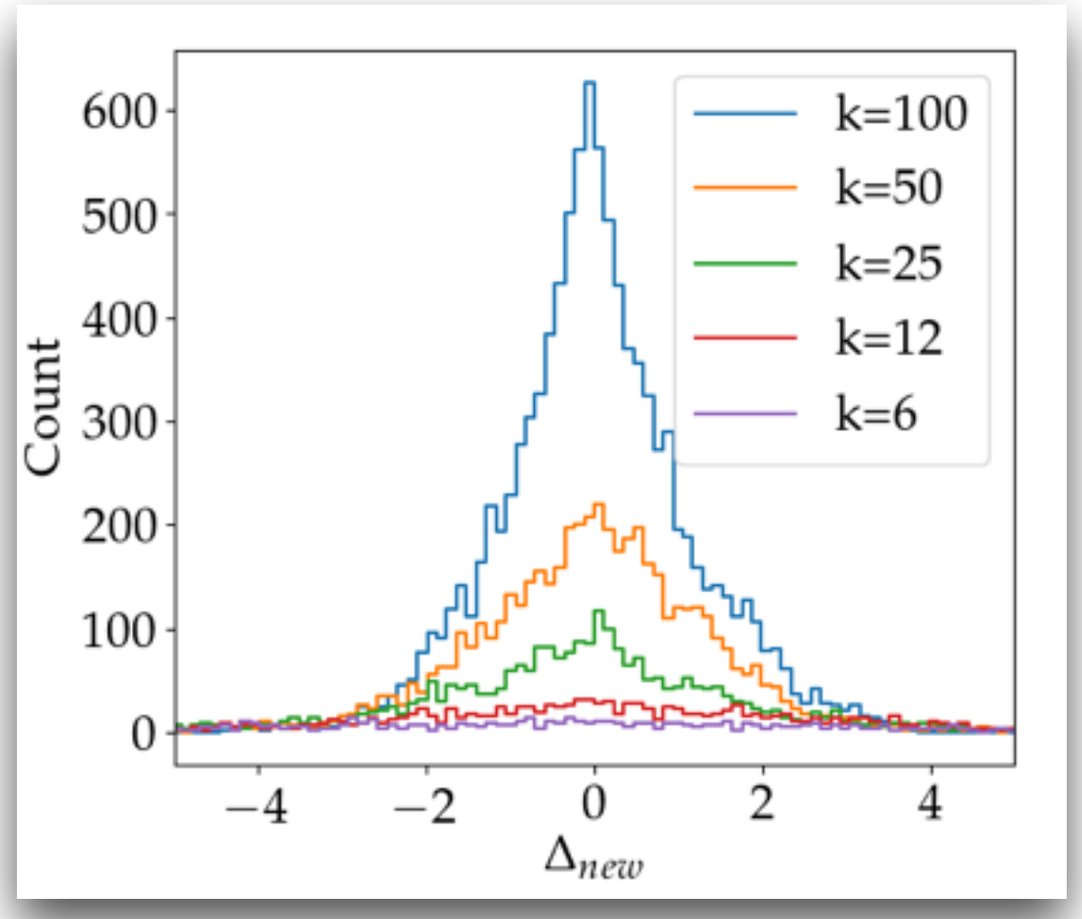
The standard deviation of the novelty response based on Delta_new scales with 1/sqrt{k} or 1/sqrt{L}, for the testing data with known patterns only.

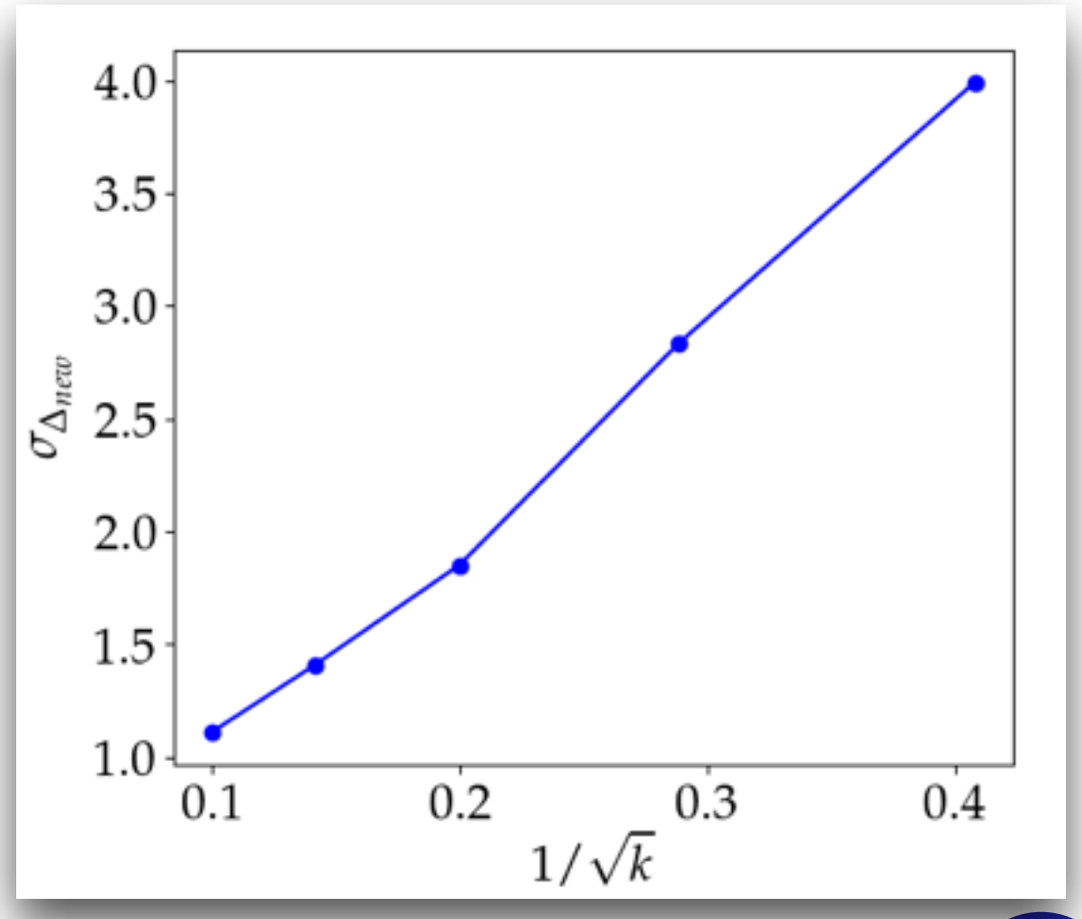


``Look Elsewhere Effect'' - Central Limit Theorem

Central Limit Theorem

The standard deviation of the novelty response based on Delta_new scales with 1/sqrt{k} or 1/sqrt{L}, for the testing data with known patterns only.

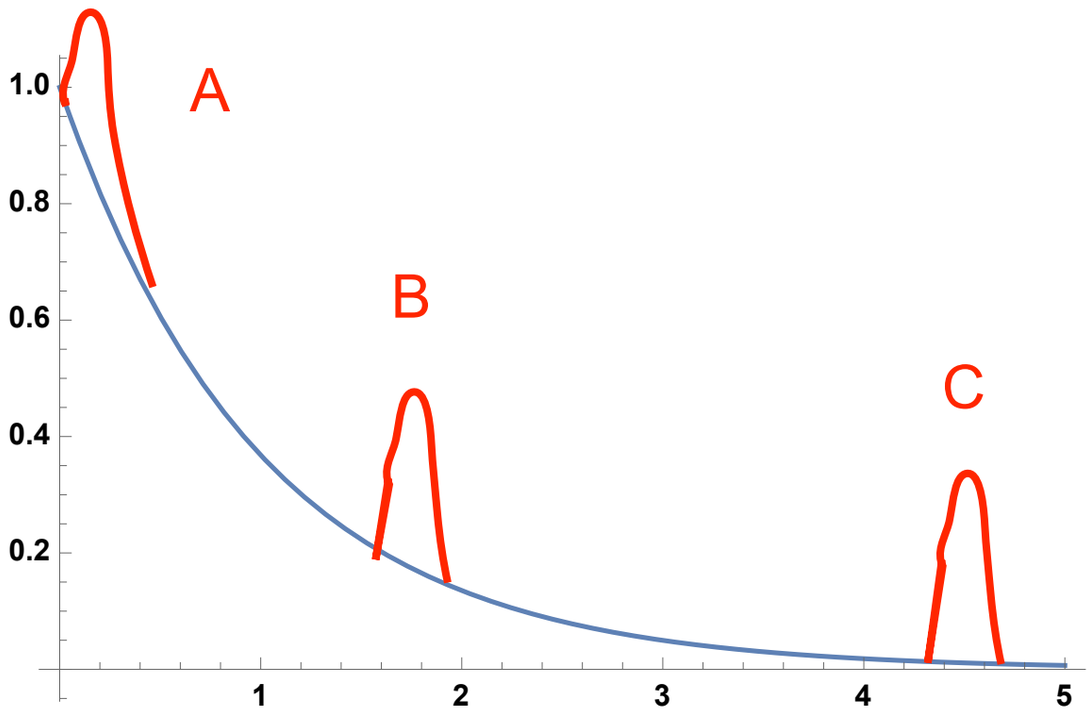




Strategy to Address Large LEE

The suppression by luminosity might not be sufficient if S/B is small.

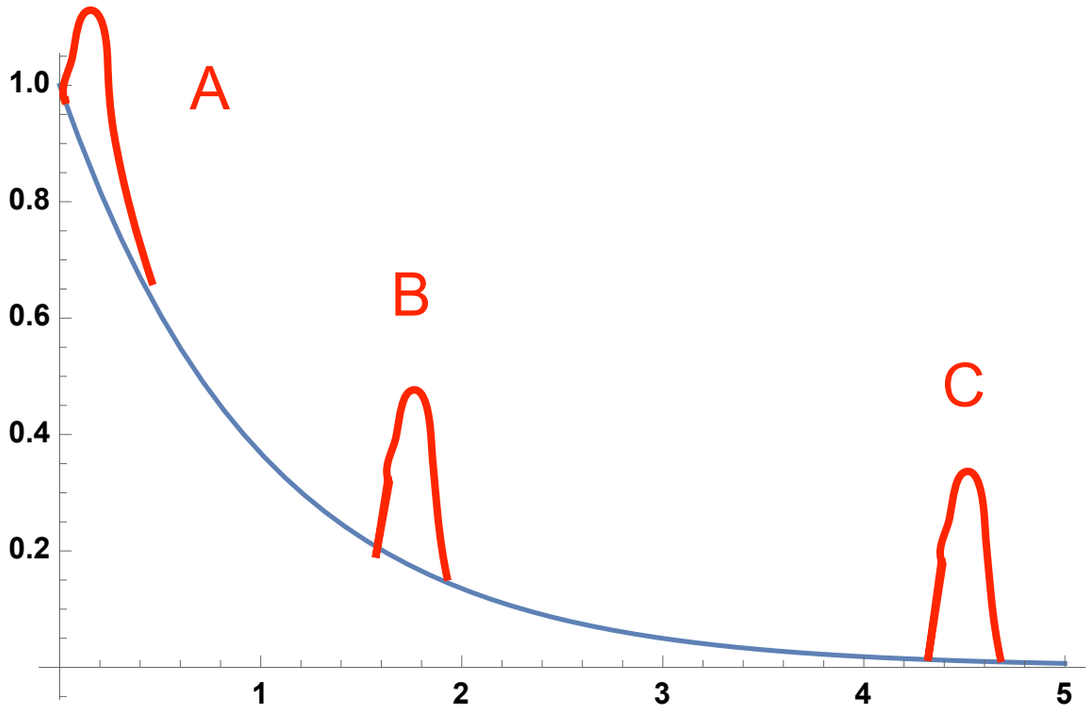
To find a way to address this problem, consider three cases A, B and C (given the fixed number of background and signal events): which ones suffer more from LEE?

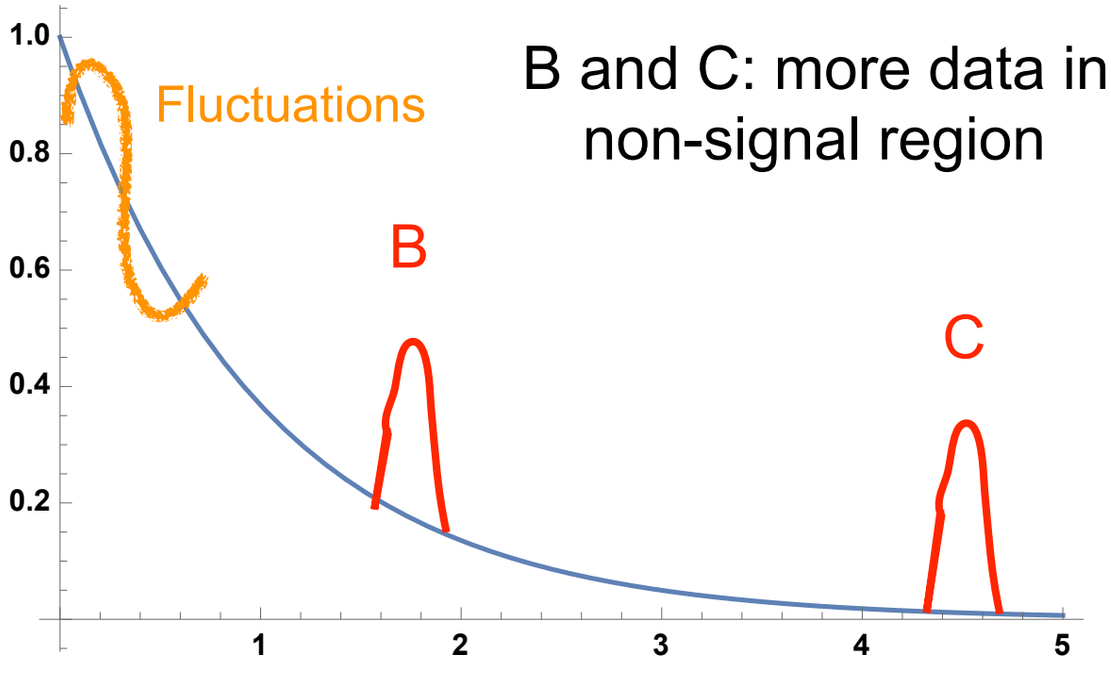


Strategy to Address Large LEE

The suppression by luminosity might not be sufficient if S/B is small.

To find a way to address this problem, consider three cases A, B and C (given the fixed number of background and signal events): which ones suffer more from LEE?



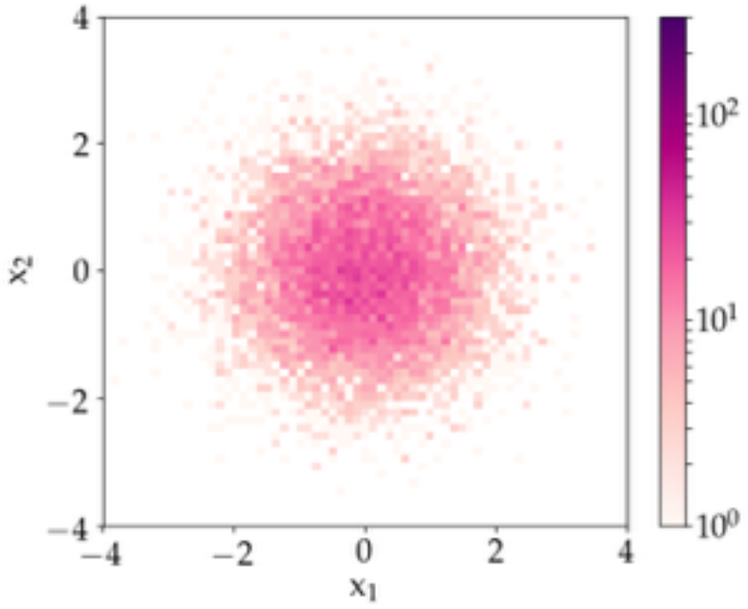


B and C: more data in non-signal region

To compensate for high-scoring (by O_new) of known-pattern data from high-density region

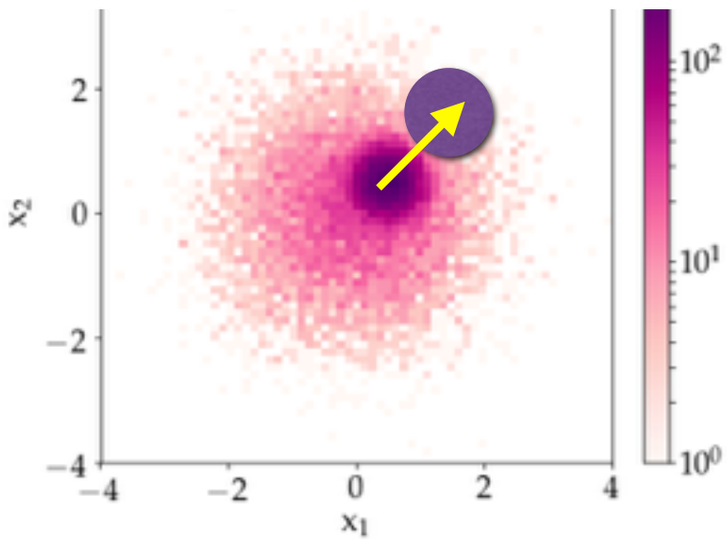
=> $\mathcal{O}_{\rm comb} = \sqrt{\mathcal{O}_{\rm trad}\mathcal{O}_{\rm new}}$

Strategy to Address Large LEE

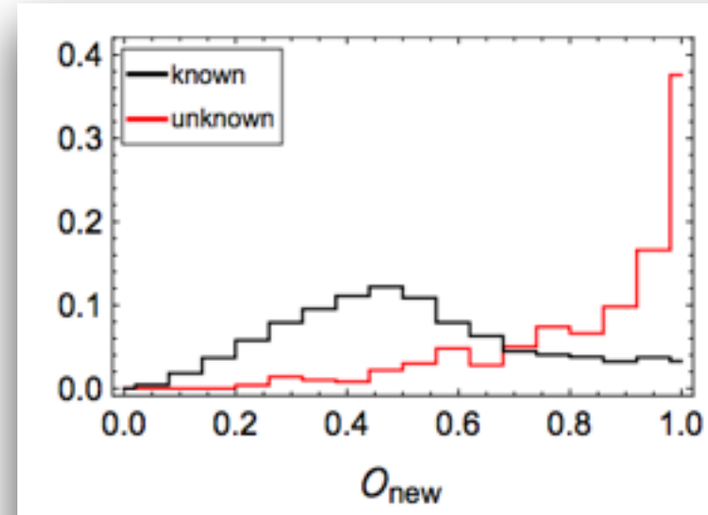


(a) Training data.

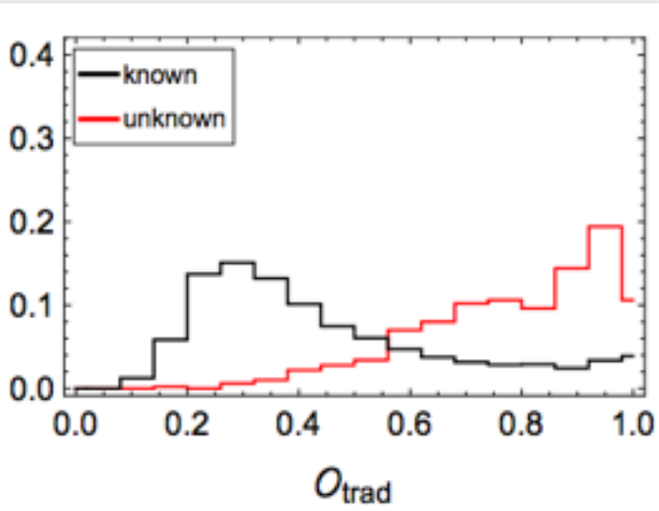
Center slightly shifted, with S/B = 1/20



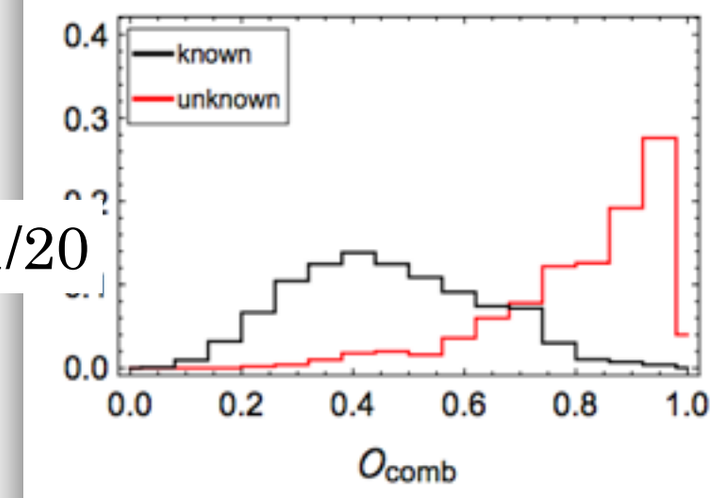
(b) Testing data.



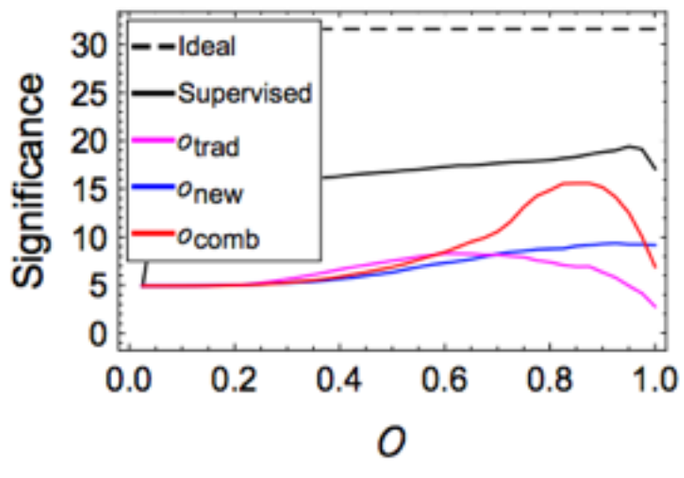
(a) New evaluator.



(b) Traditional evaluator.



(c) Combined evaluator.



(d) Significance.

O_comb based analysis yields more than 50% improvement in detection sensitivity!

Parton-level Benchmark Study

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

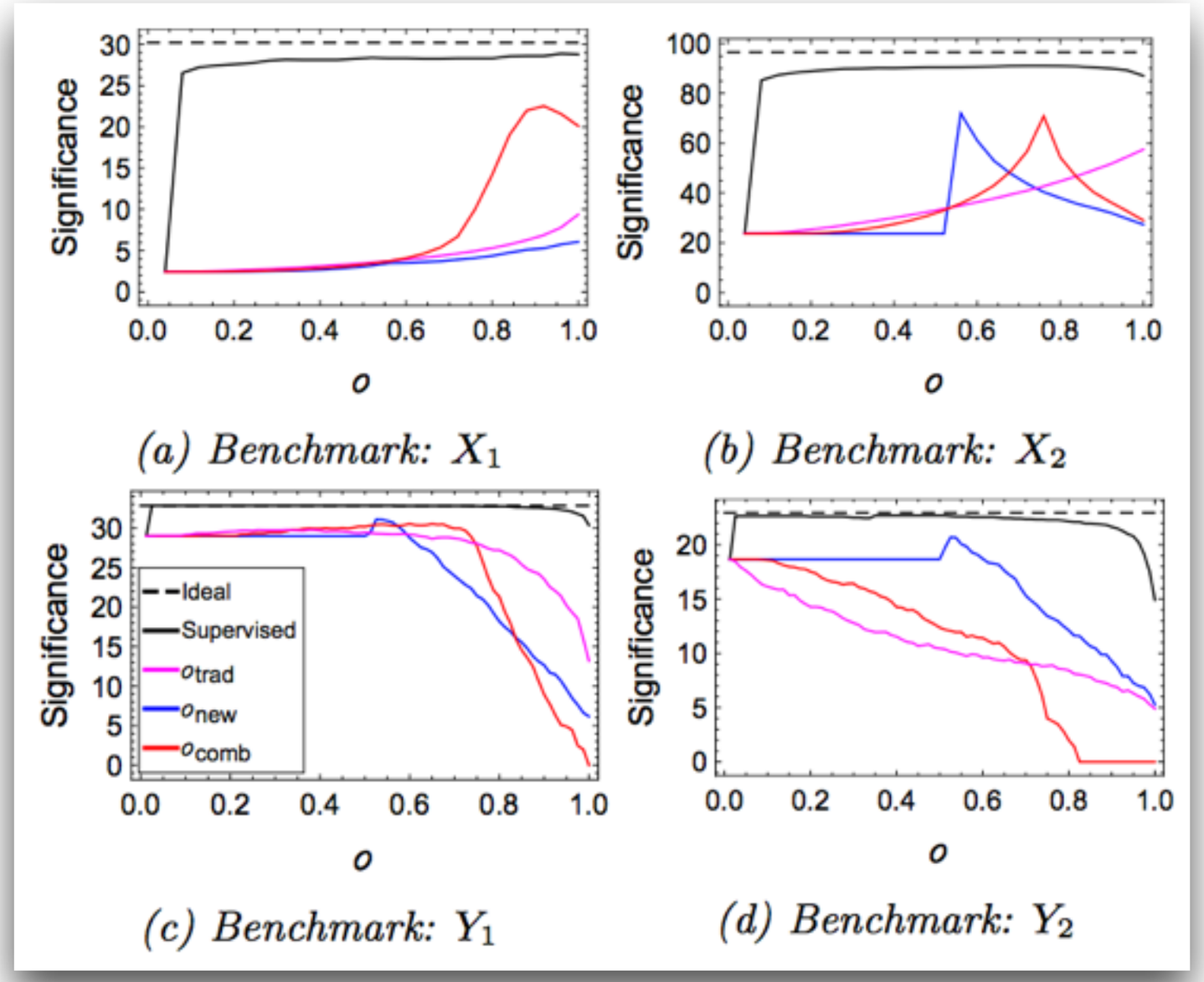
- $pp \to \bar{t}_l t_l$, $\sigma = 11.5\,\text{fb}$, $\mathbf{X_1}$: $pp \to \overline{T}T \to W_l^+ W_l^- \bar{b}b$
- $pp \to t_l \bar{b} W_l^\pm$, $\sigma = 0.365\,\text{fb}$,
- $pp \to Z_b Z_l$, $\sigma = 0.0765\,\text{fb}$. $\mathbf{X_2}$: $pp \to Z' \to \bar{t}t$

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

- $e^+e^- \to hZ \to Z^*_{\text{inv}} Z_{\bar{b}b} l^+ l^-$ $\sigma = 0.00686\,\text{fb}$, $\mathbf{Y_1}$: $h \to \tilde{\chi}_1 \tilde{\chi}_2 \to \tilde{\chi}_1 \tilde{\chi}_1 a$.
- $e^+e^- \to hZ \to Z^*_{\bar{b}b} Z_{\text{inv}} l^+ l^-$ $\sigma = 0.00259\,\text{fb}$. $\mathbf{Y_2}$: $h \to Za$

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

Parton-level Benchmark Study



- X1: well-modeled by the Gaussian sample!
- X2: O_comb less efficient due to one-order larger S/B
- X3 and X4: 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

Wishlist of Questions to Address

- Optimize the algorithm (e.g., if it is possible to reduce sensitivity discrepancy between novelty detection and supervised learning by utilizing dynamical learning mechanism)

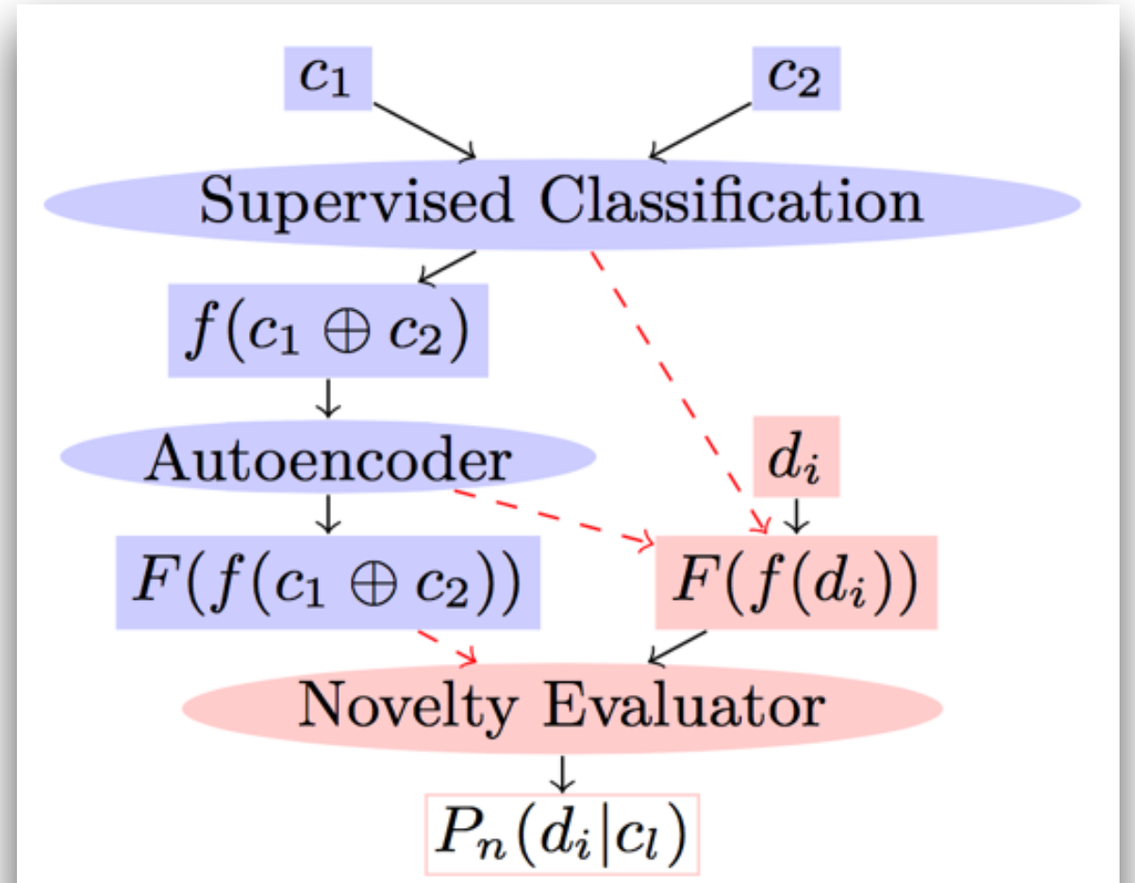
- Test the algorithm at more realistic level (hadron level)

- What is its sensitivity performance if we treat some SM processes to measure as the targets of novelty detection? (Question raised by Junjie Zhu)

- Is it possible to invent a novelty evaluator to exploit multiple measures at once? (Question raised by Aurelio Juste)

- … … … …

Summary



- Rapid development of the DNN techniques is bringing far-reaching influence to particle physics
- A combination of supervised learning and novelty detection may lay out the framework for future data analysis
- By properly designing novelty evaluators (clustering sensitive, LEE suppressed, etc.), encouragingly high sensitivity could be achieved in detecting the BSM physics
- Follow-up project is ongoing, in collaboration with ATLAS experimentalists, dedicated to filling the gap between proof of concept and real data analysis


Thank you!