``第十三届TeV物理工作组学术研讨会", August 2018, Nankai Univ., Tianjin

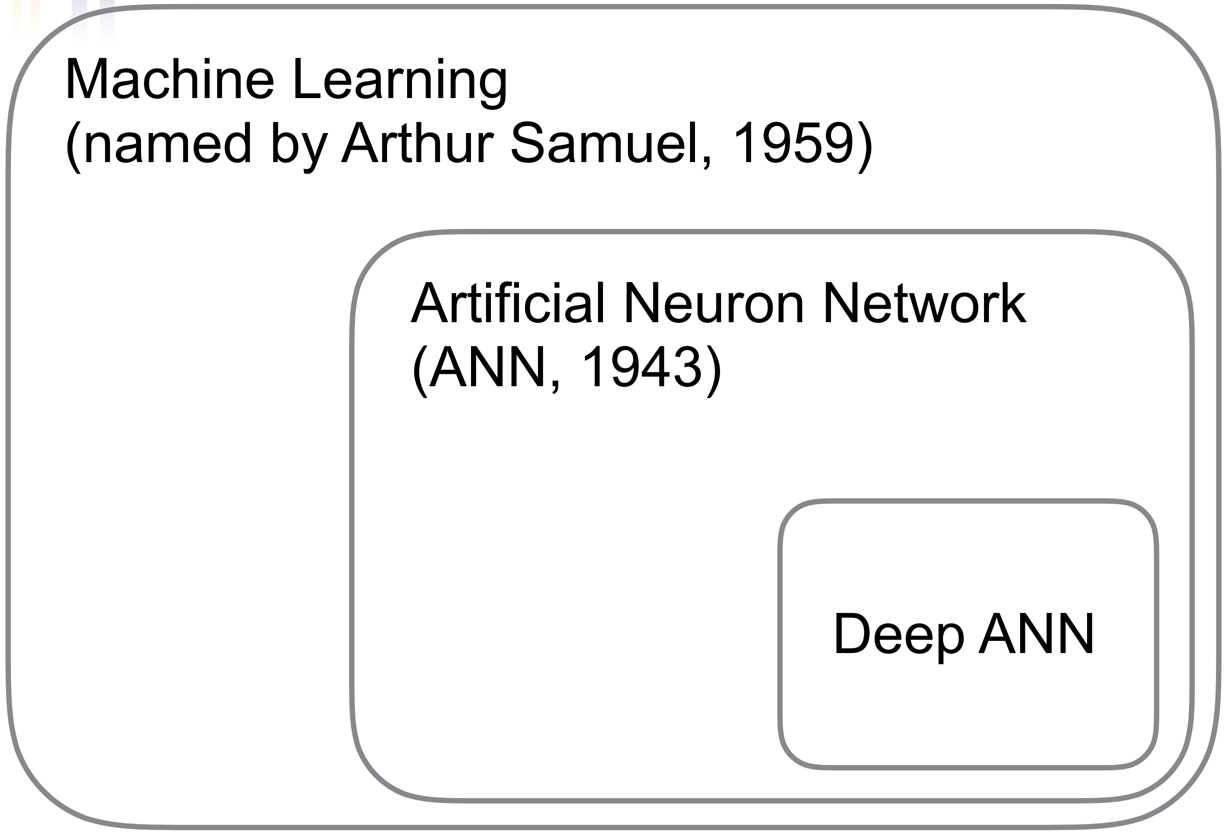
NOVELTY DETECTION MEETS COLLIDER PHYSICS

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

Deep ANN and Collider Physics



Modern deep ANN learning systems developed in last decade is extending its great success in image and speech recognition, self-driving car, etc, to scientific research. This may bring far-reaching influence for collider physics.

- Unlike cut-based method and traditional ML techniques (e.g. BDT) which rely heavily on expert-designed observables to reduce problem dimensionality, deep ANN automatically extracts pertinent features as neurons from data.
- Collider physics in near future: kinematic observable design => algorithm design => high-efficient data mining

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.

Signal Processing 99 (2014) 215–249

Contents lists available at ScienceDirect

SIGNAL PROCESSING

- Detection target is unknown
- Detection target is not necessary to be unique
- No a priori knowledge about targets is needed during the training phase.
- => The detection is ``target'' or ``model''-independent
- If treating the BSM physics as ``novel'' event, we can search for BSM physics model-independently, using the novelty detection techniques

ARTICLE INFO

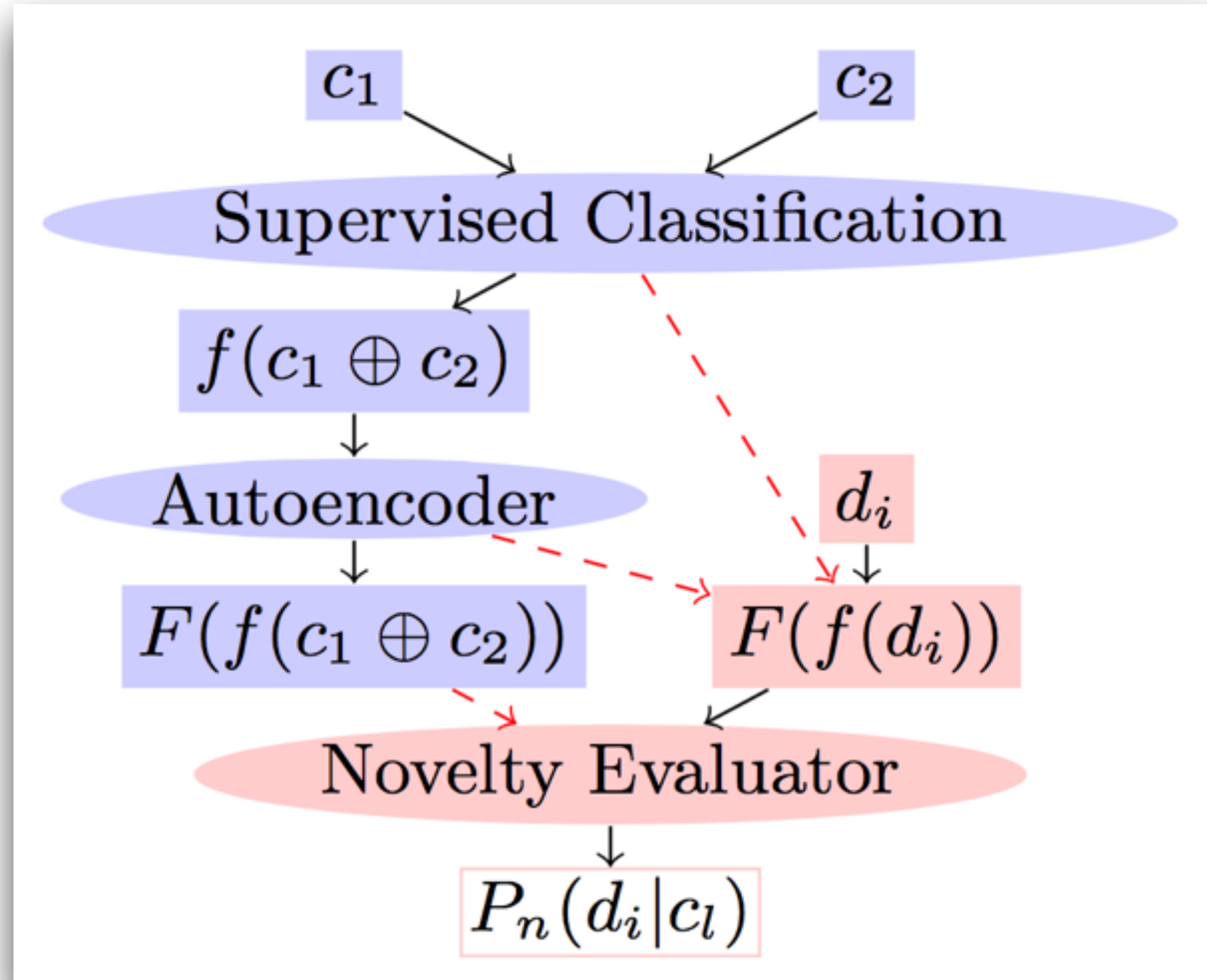
Article history:
Received 17 October 2012
Received in revised form
16 December 2013
Accepted 23 December 2013
Available online 2 January 2014

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.

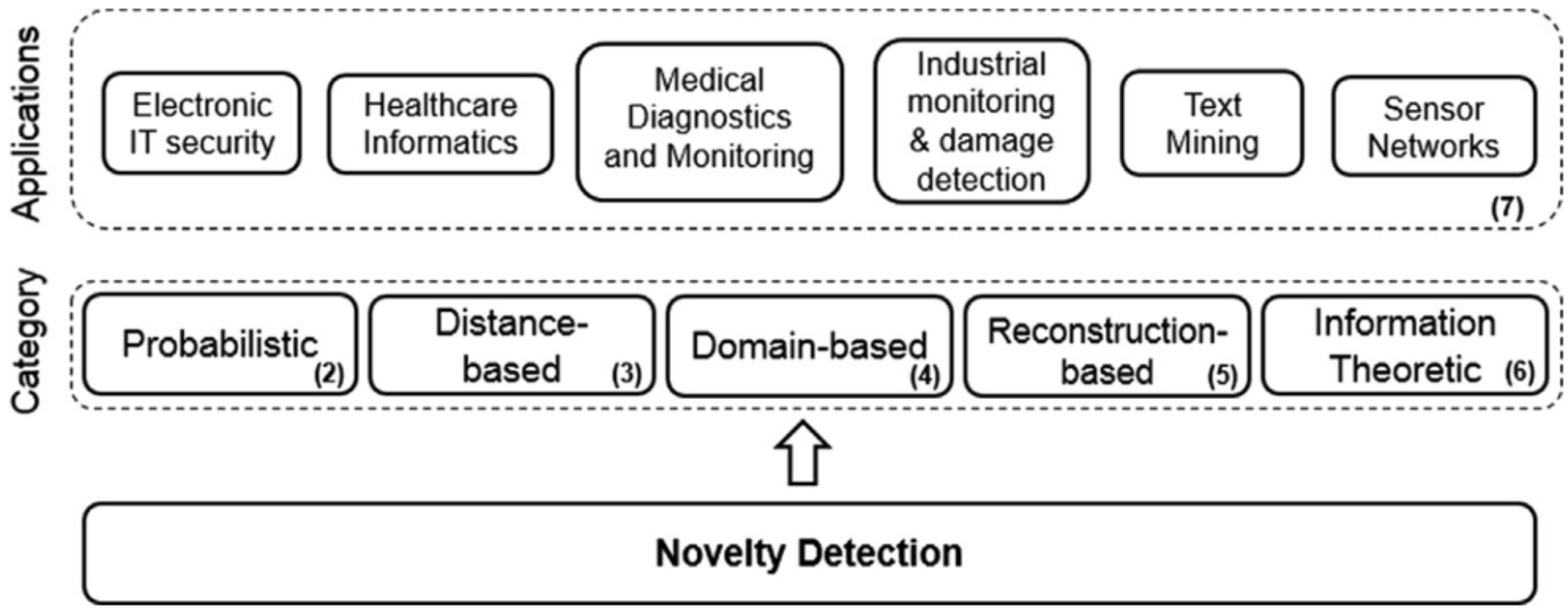
Workflow



- Step 1: (SM) feature learning
- Step 2: dimension reducing of feature space (auto-encoder)
- Step 3: novelty evaluating of testing data
- => Detection sensitivity based on novelty response

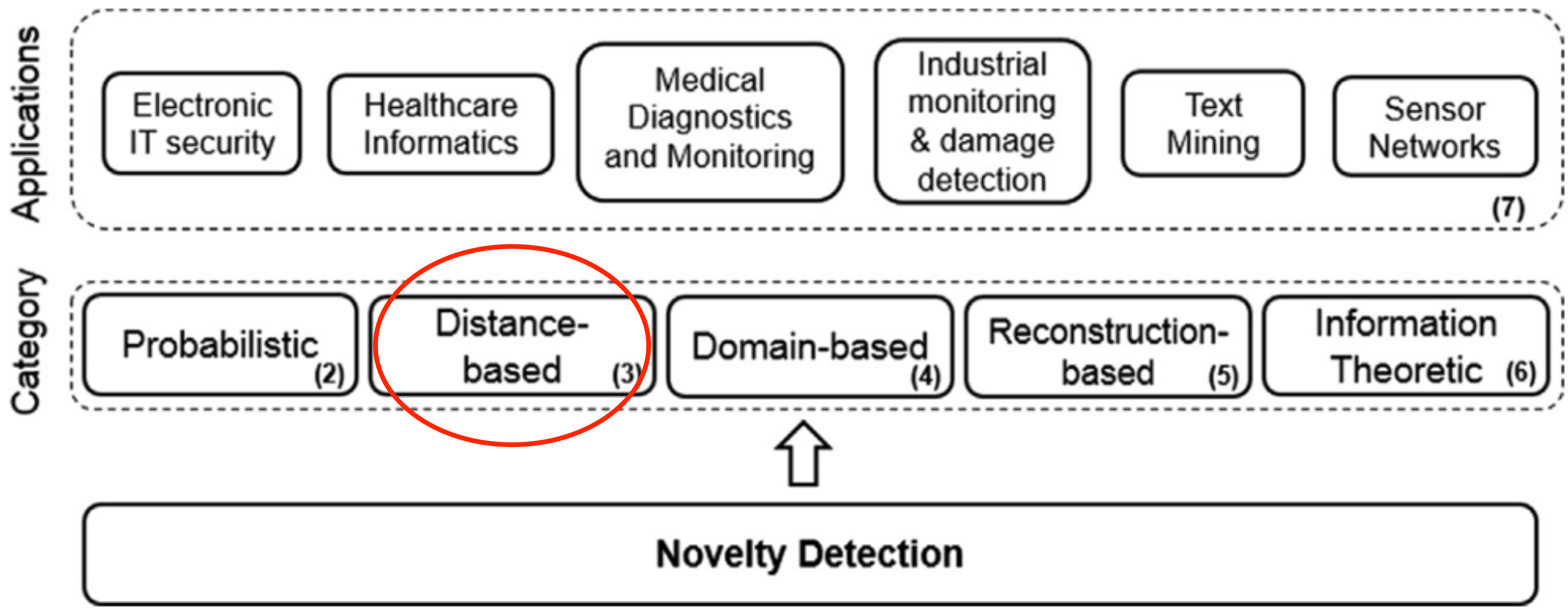
Novelty Evaluators - Application Driven

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



Novelty Evaluators - Application Driven

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



Novelty Evaluators: Traditional Wisdom

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

Novelty measure: range unnormalized

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

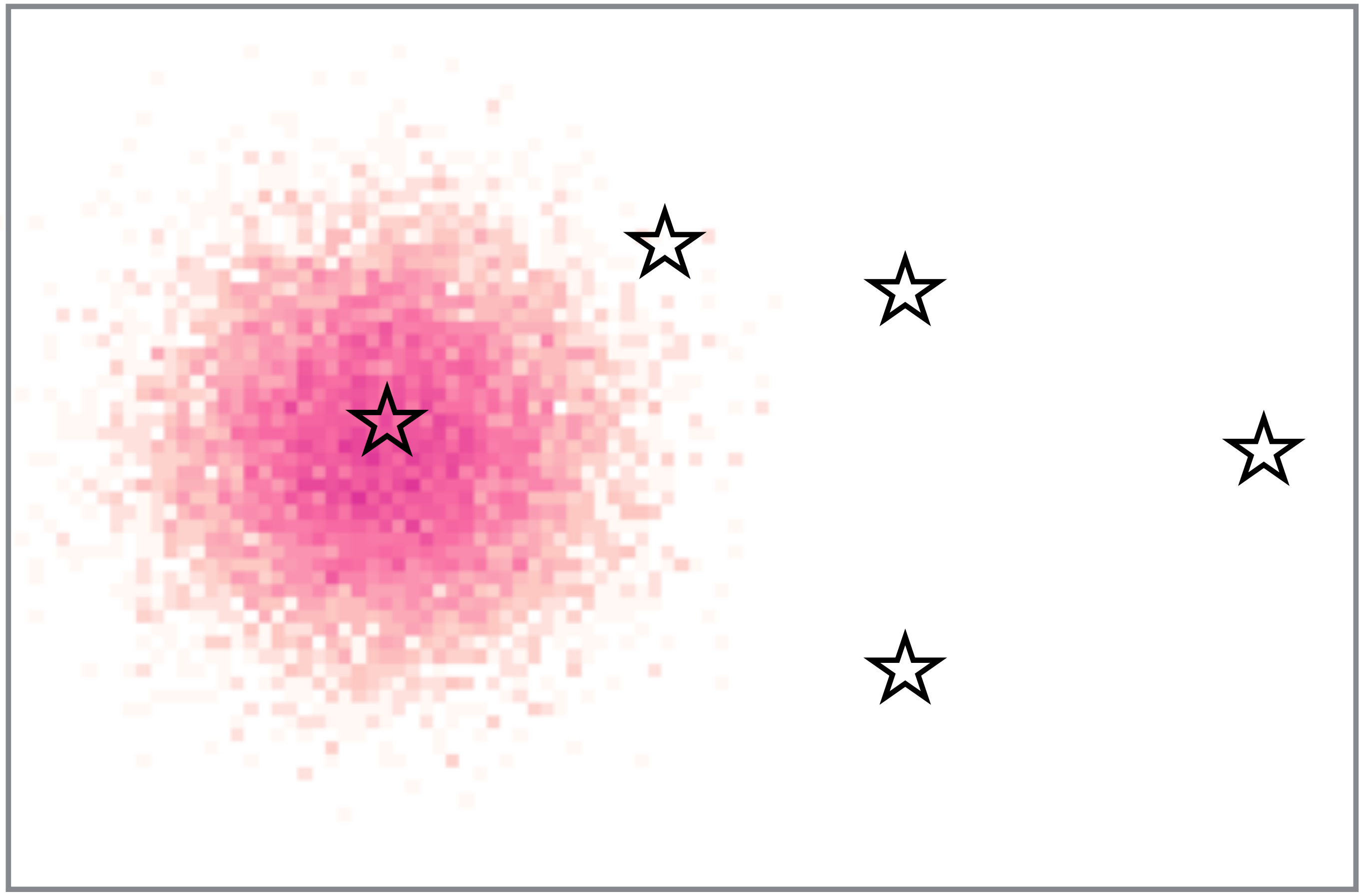
- d_{train} : mean distance of a testing data point to its k nearest neighbors
- $\langle d'_{\mathrm{train}} \rangle$: average of the mean distances defined for its k nearest neighbors
- $\langle d'^2_{\mathrm{train}} \rangle^{1/2}$: standard deviation of the latter
- All quantities are defined wrt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

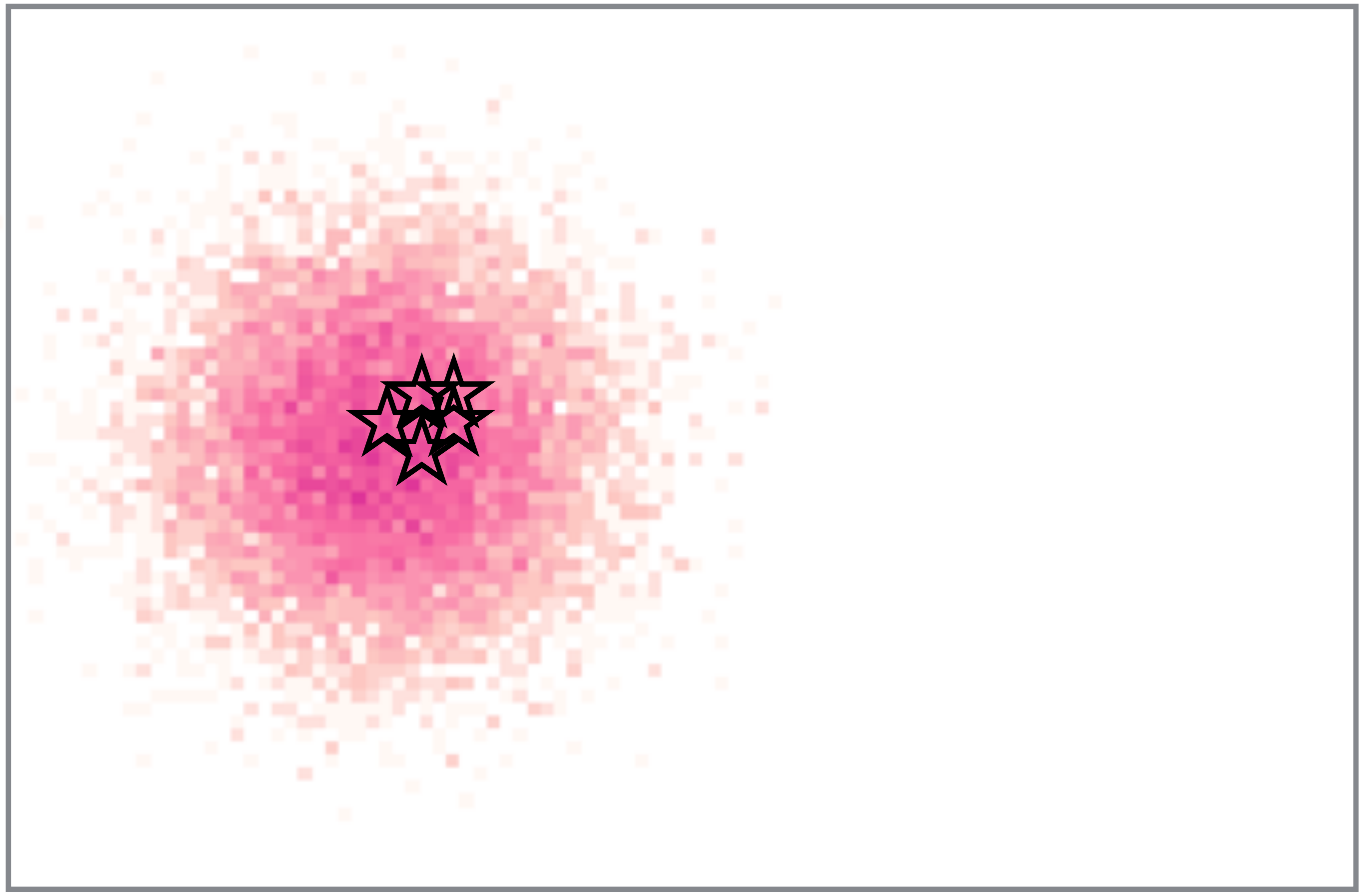
$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}} \qquad \mathcal{O} = \frac{1}{2}\left(1 + \text{erf}\left(\frac{c\Delta}{\sqrt{2}}\right)\right)$$



- Large distance => high score
- Short distance => low score
- => a measure of isolation
- This design ignores the correlation among the testing data with unknown pattern, and may not work well for data analysis in particle physics

Novelty Evaluators: Traditional Wisdom

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



- Resonance, shape, … could be important clustering features for BSM physics detection
- The testing data of unknown pattern with such features are scored low, unless they are away from the training data!
- Why such a design? Application driven, e.g., finger print recognition

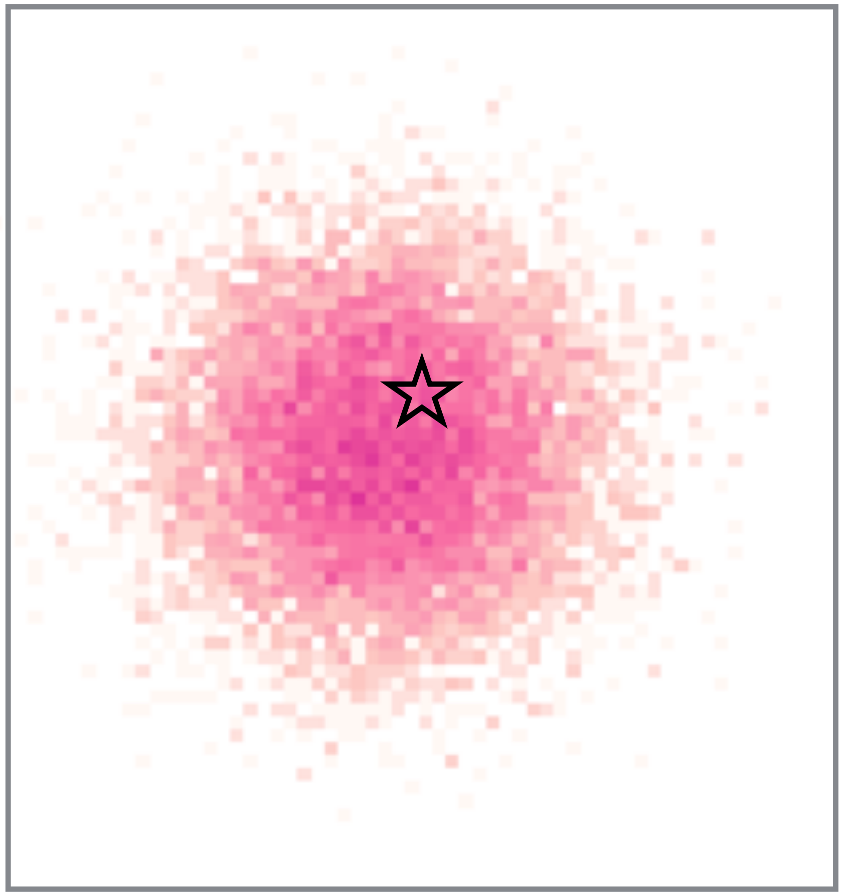
Novelty Evaluators: New Input

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

- 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 is evaluated by comparing local densities of the testing point in the training and testing datasets
- Approximately statistical interpretation : $\Delta_{\text{new}} \propto \left.\frac{S}{\sqrt{B}}\right|_{\text{local bin}}$

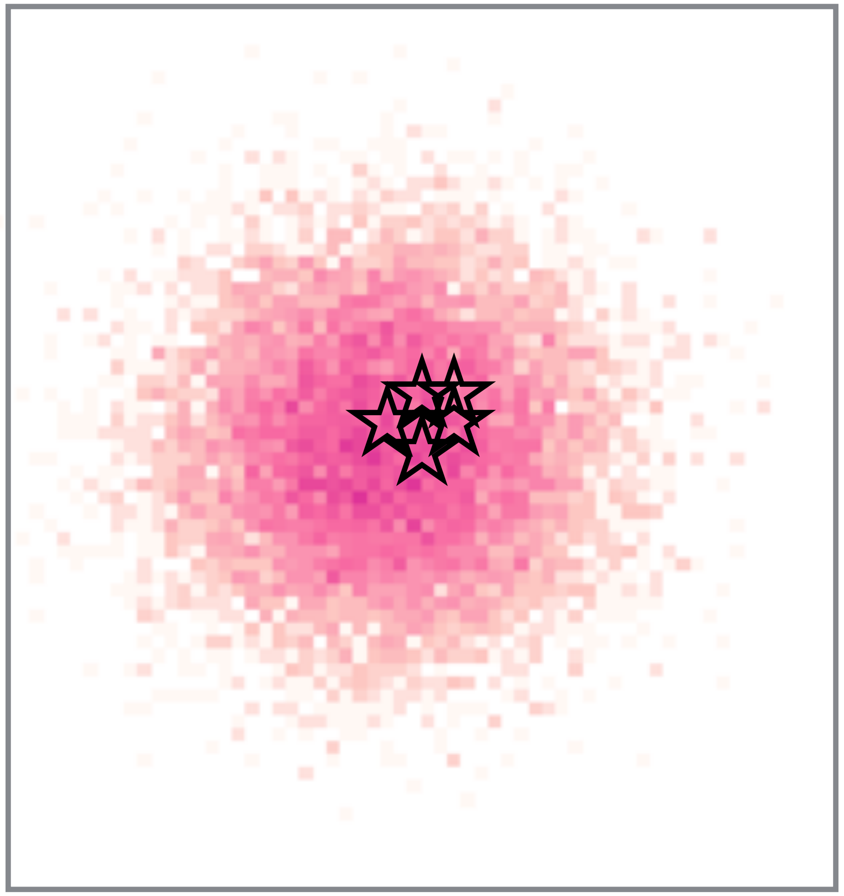
Novelty Evaluators: New Input

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



Training dataset

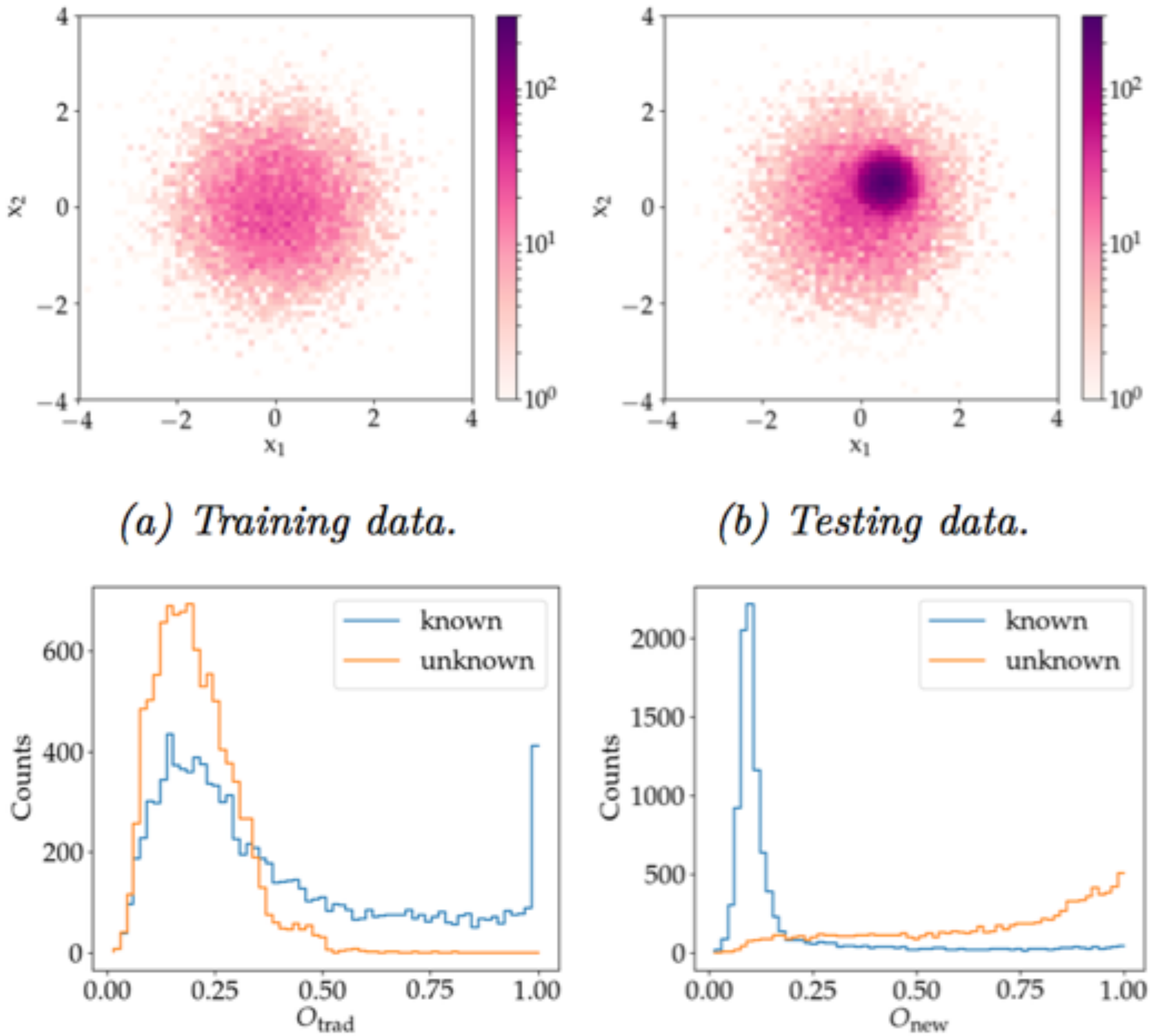
VS



Testing dataset

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

Novelty Evaluators: Performance Comparison



(a) Training data. (b) Testing data.

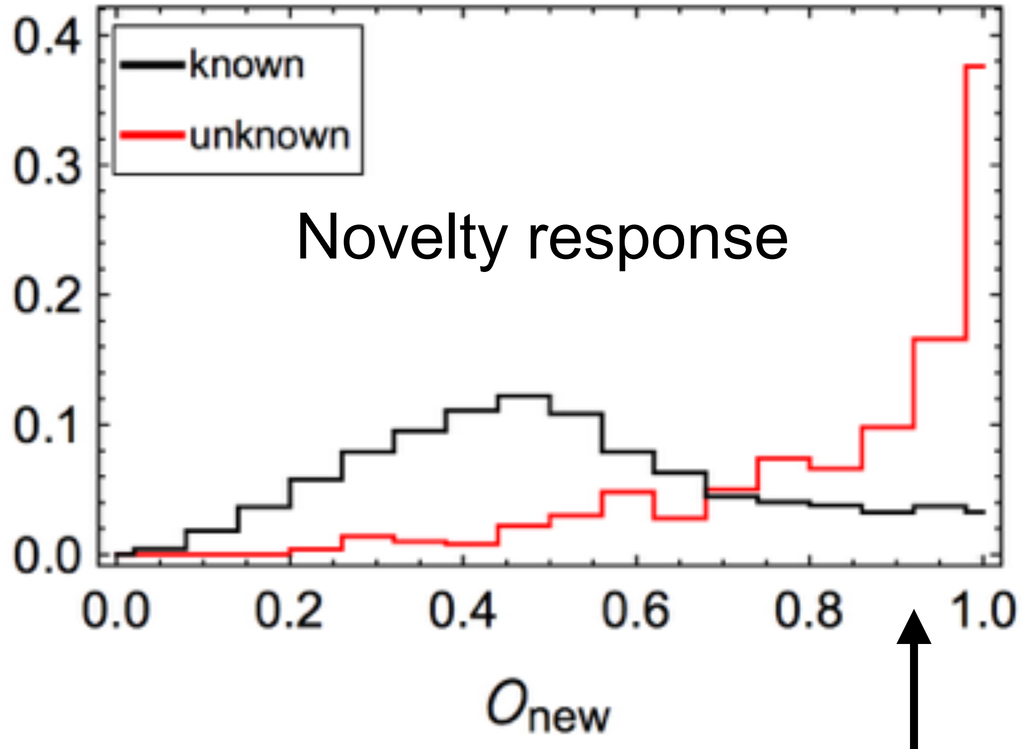
(c) $\mathcal{O}_{trad}$ performance. (d) $\mathcal{O}_{new}$ performance.

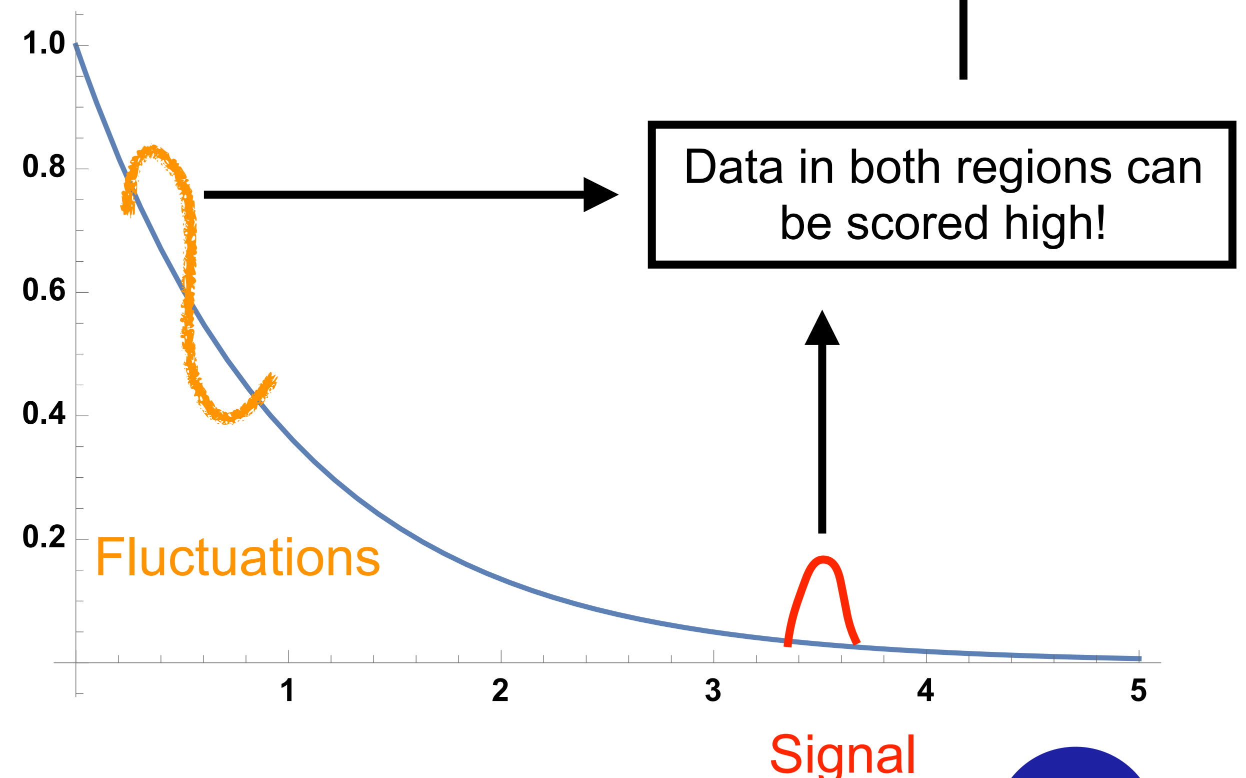
- Consider 2D Gaussian samples
- Training dataset: known pattern only
- Testing dataset: known + unknown patterns
- Compared to O_trad, the novelty response of unknown-pattern data is much stronger for O_new
- => A well-separation between the known- and unknown-pattern data distributions

``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 generically suffers from ``Look Elsewhere Effect (LEE)'', given the size of the parameter space to be searched.

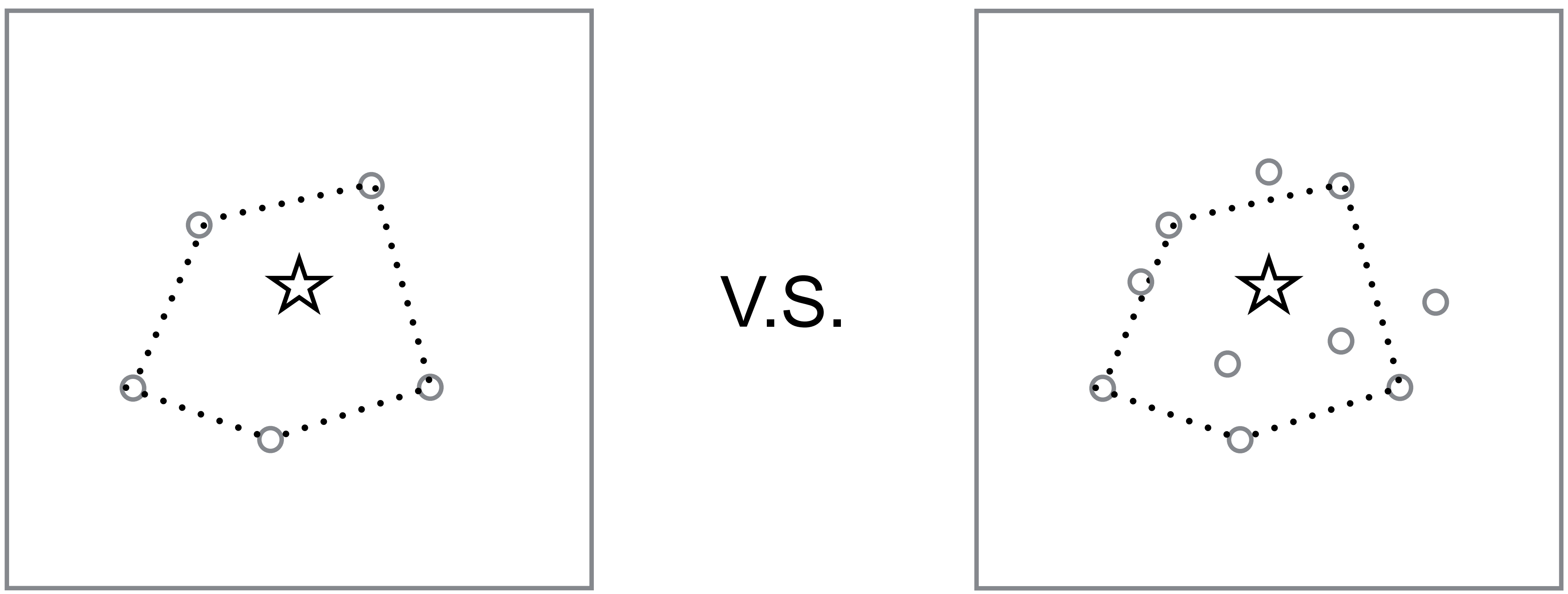




``Look Elsewhere Effect'' - Central Limit Theorem

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.

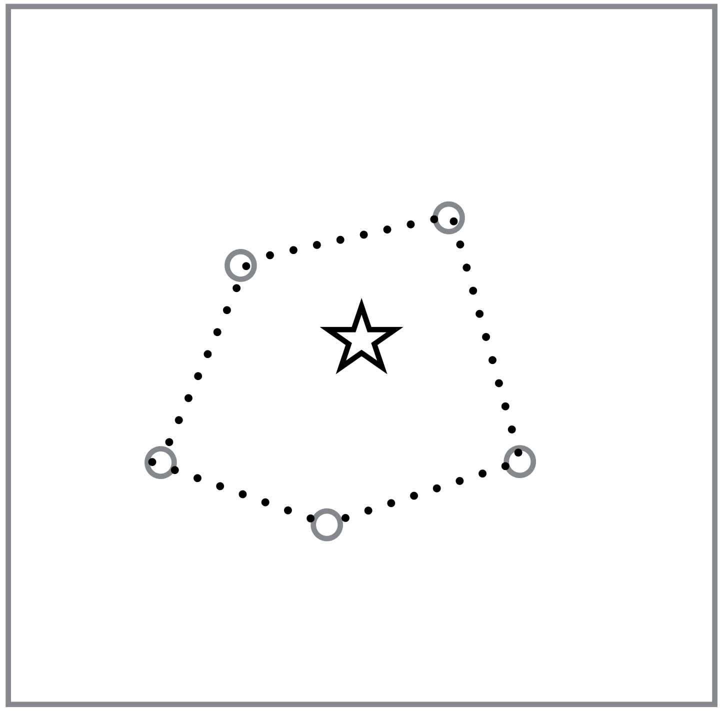


L   2 * L

``Look Elsewhere Effect'' - Central Limit Theorem

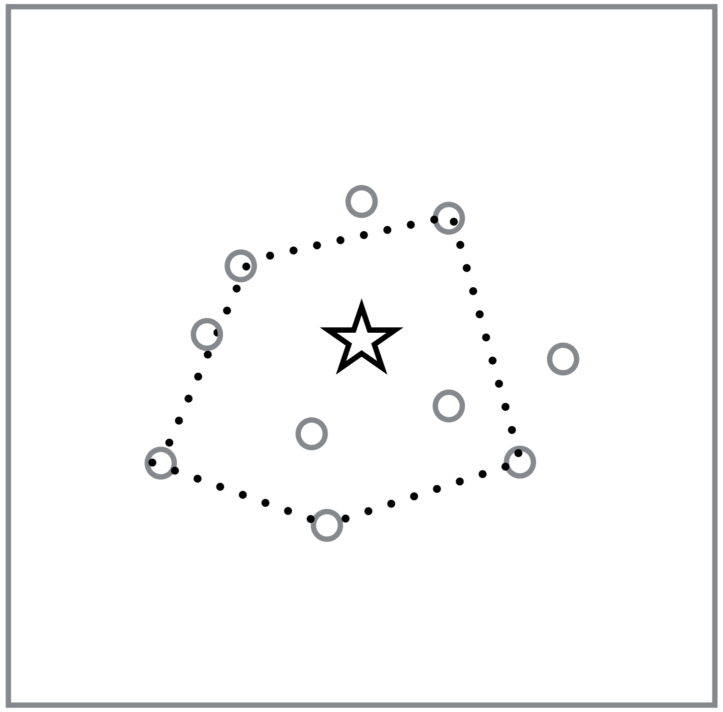
Central Limit Theorem

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



L

V.S.

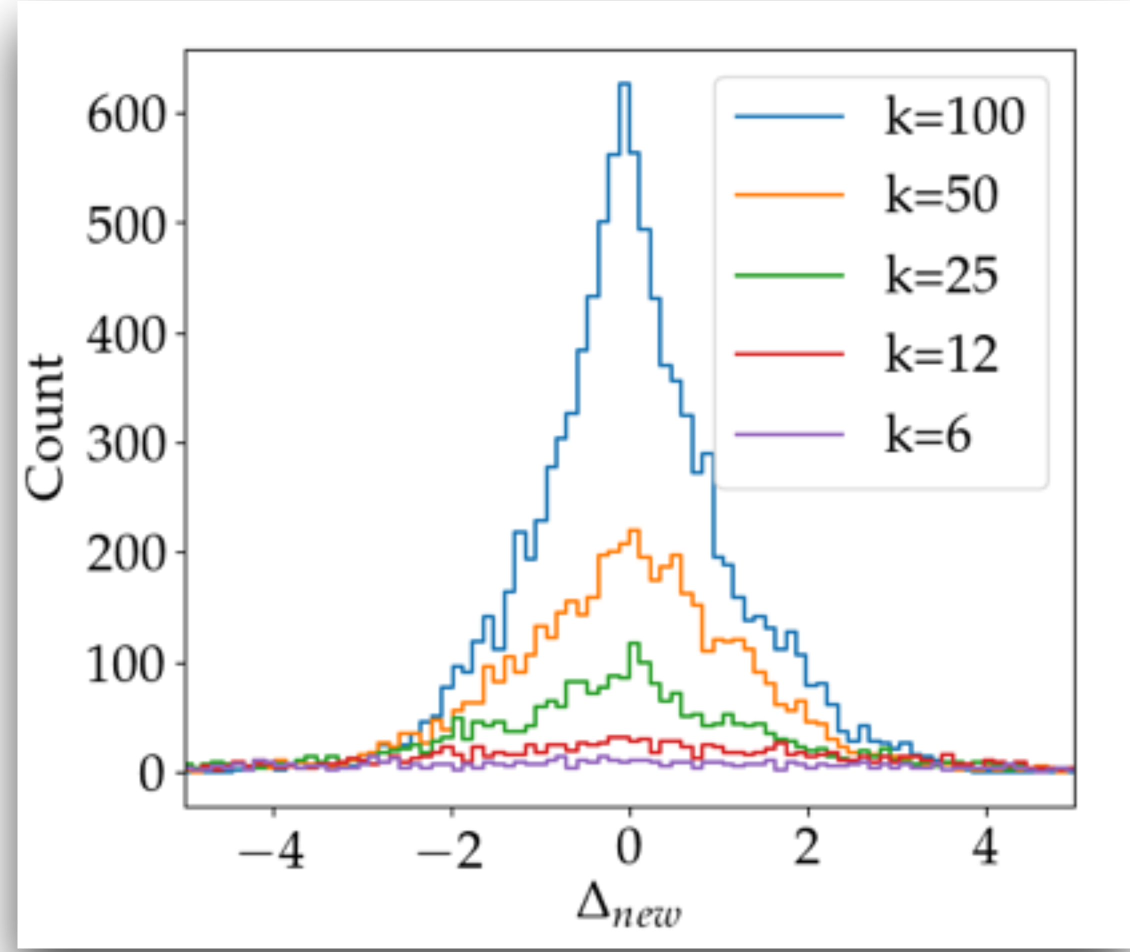


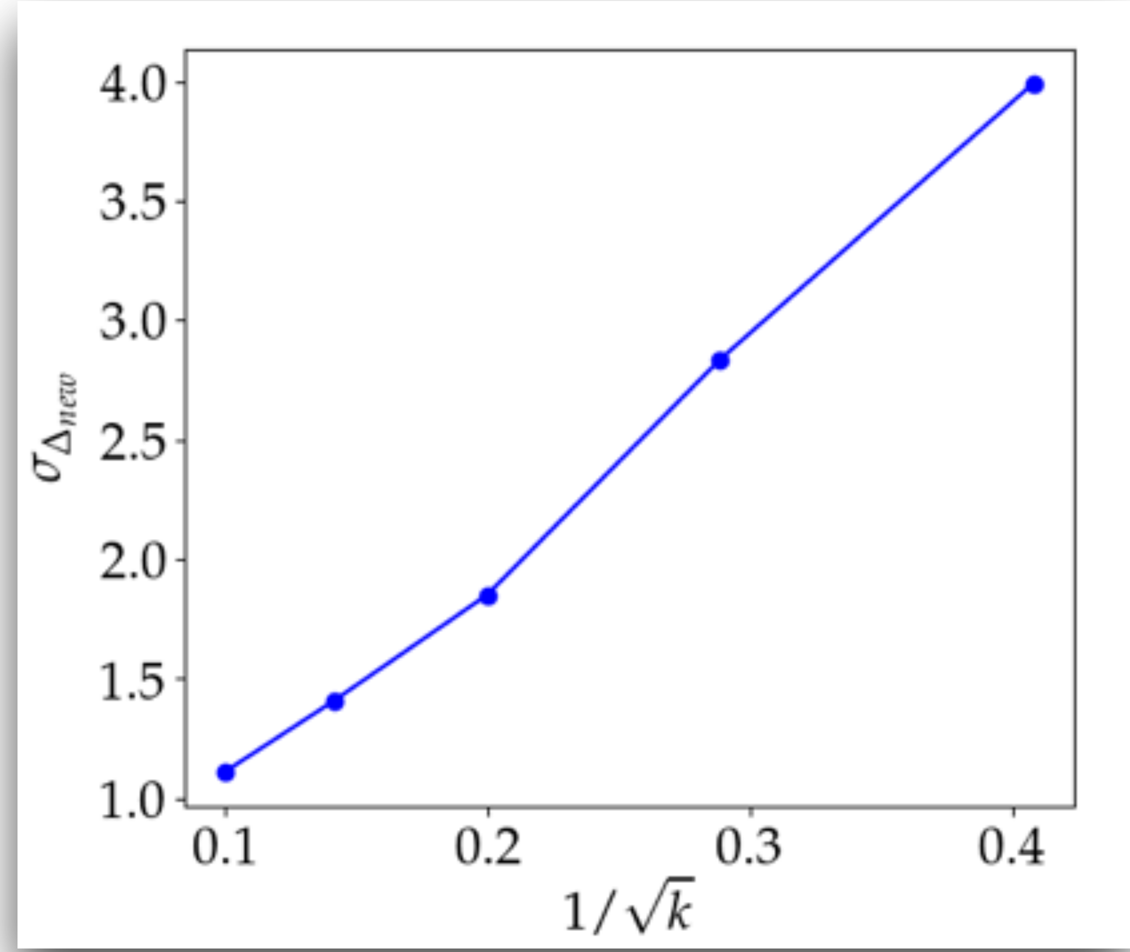
2 * L

``Look Elsewhere Effect'' - Central Limit Theorem

Central Limit Theorem

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

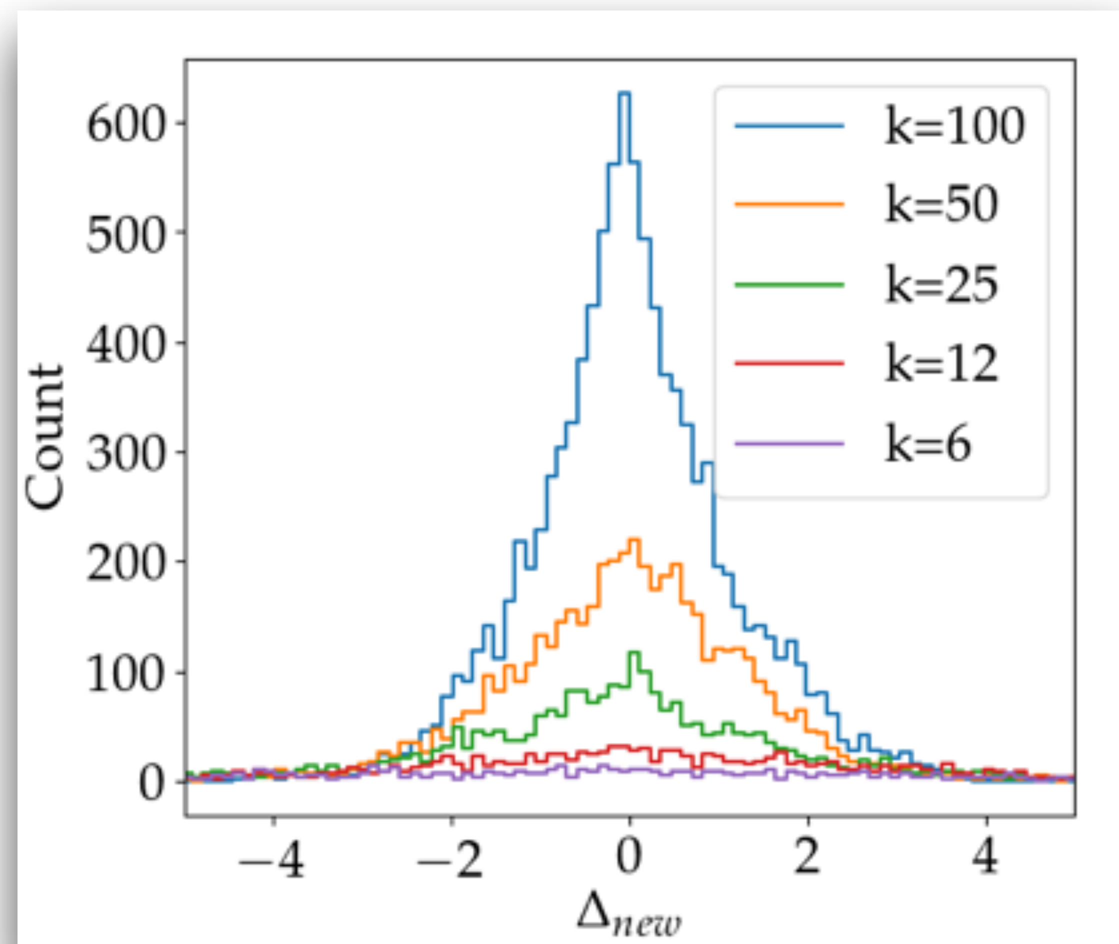


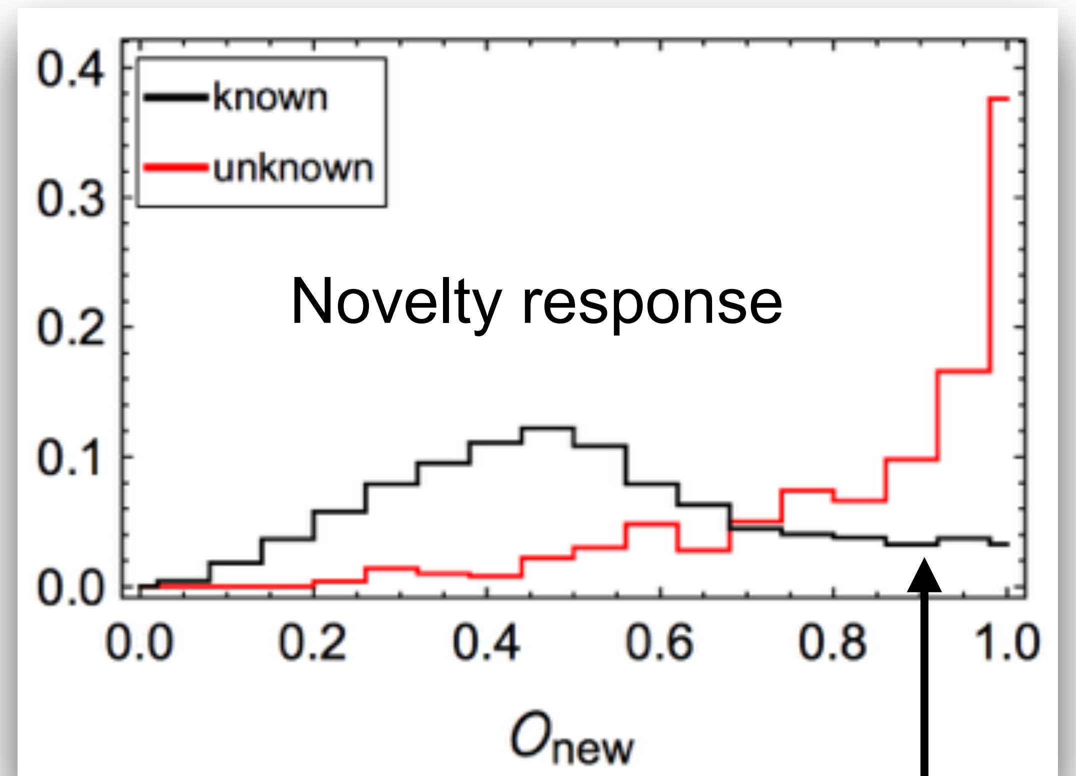


``Look Elsewhere Effect'' - Central Limit Theorem

Central Limit Theorem

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

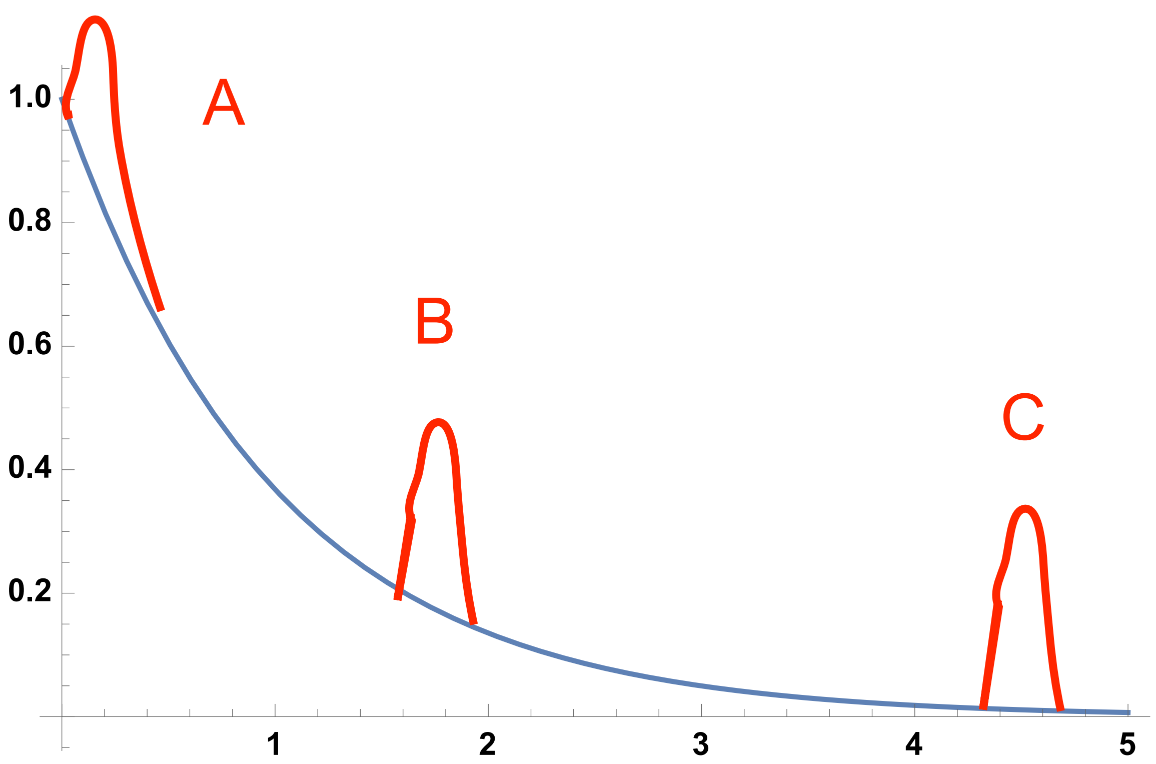




=> The distribution of the data with known patterns will get narrowed, as L increases!

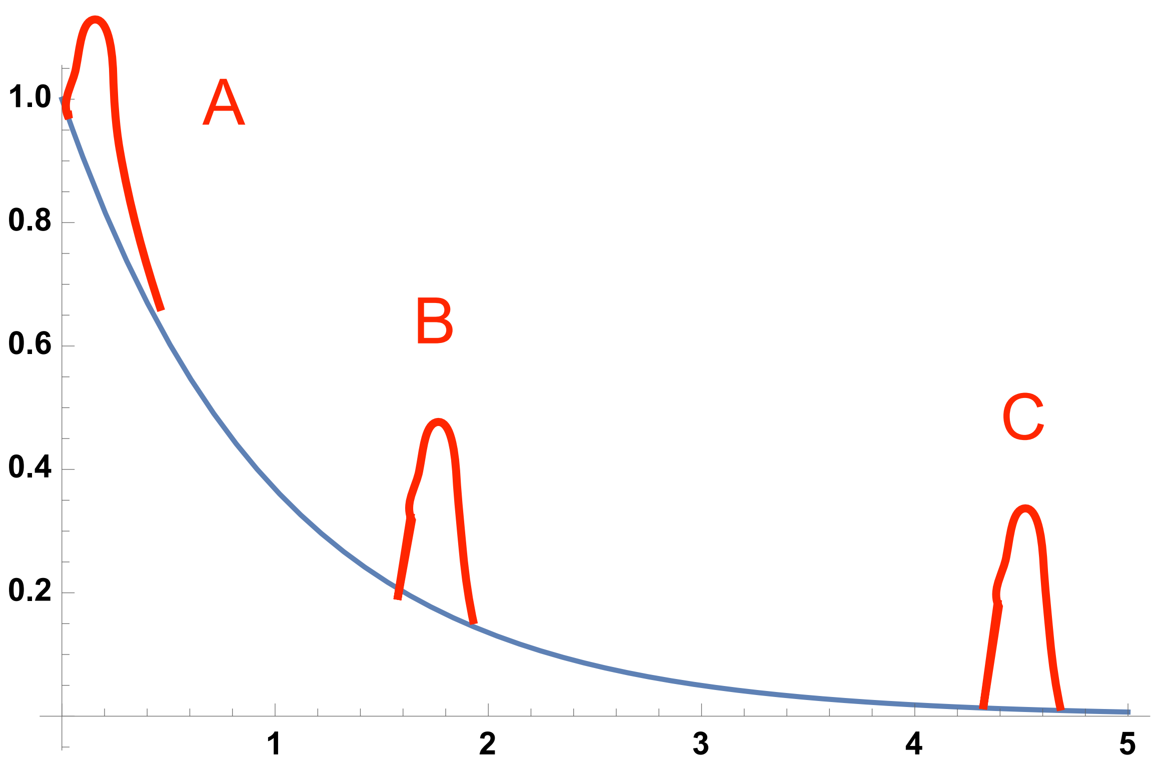
Strategy to Address Large LEE

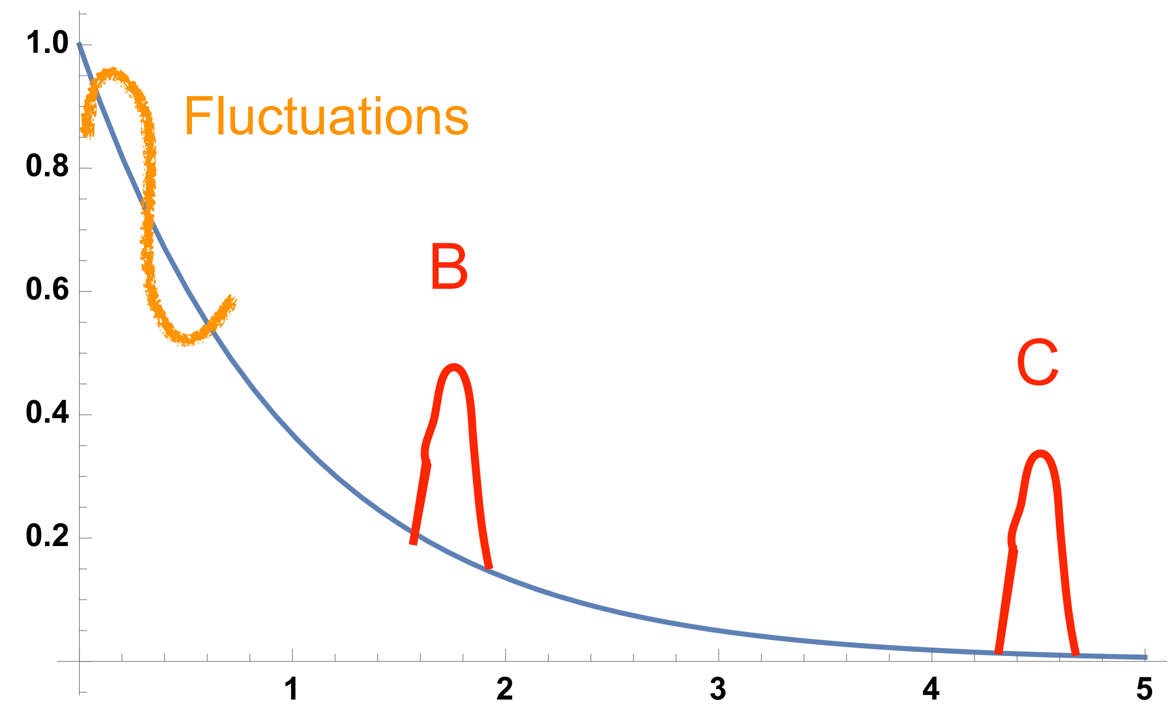
Given the fixed number of background and signal events, which cases have a worse LEE among A, B, C?



Strategy to Address Large LEE

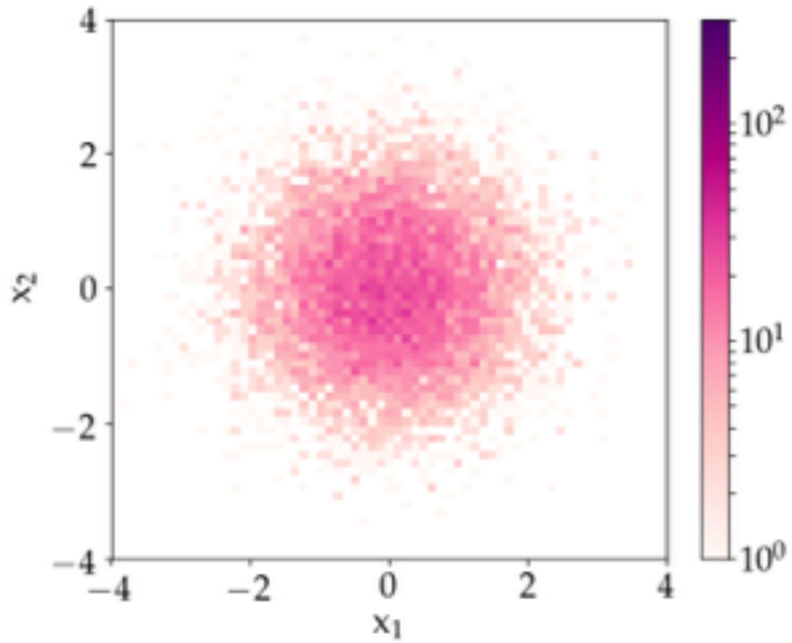
Given the fixed number of background and signal events, which cases have a worse LEE among A, B, C?





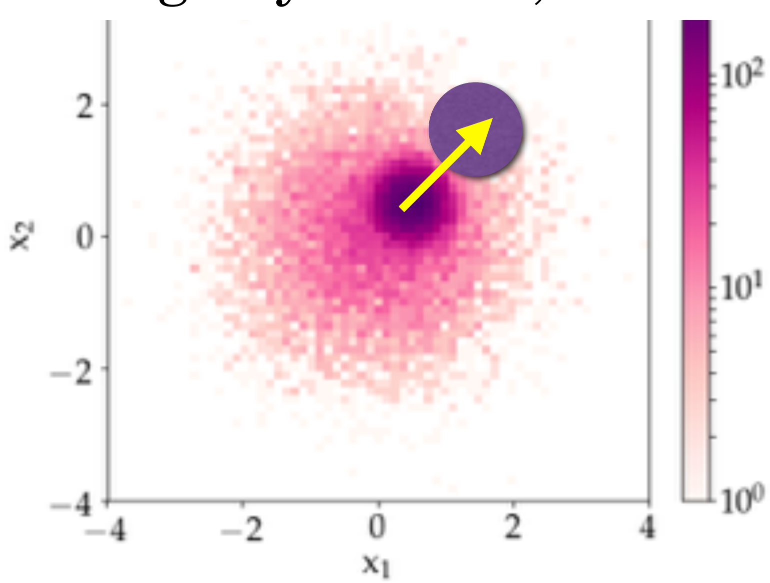
To compensate for high-scoring of known-pattern data from high-density region => $\mathcal{O}_{\mathrm{comb}} = \sqrt{\mathcal{O}_{\mathrm{trad}}\mathcal{O}_{\mathrm{new}}}$

Strategy to Address Large LEE

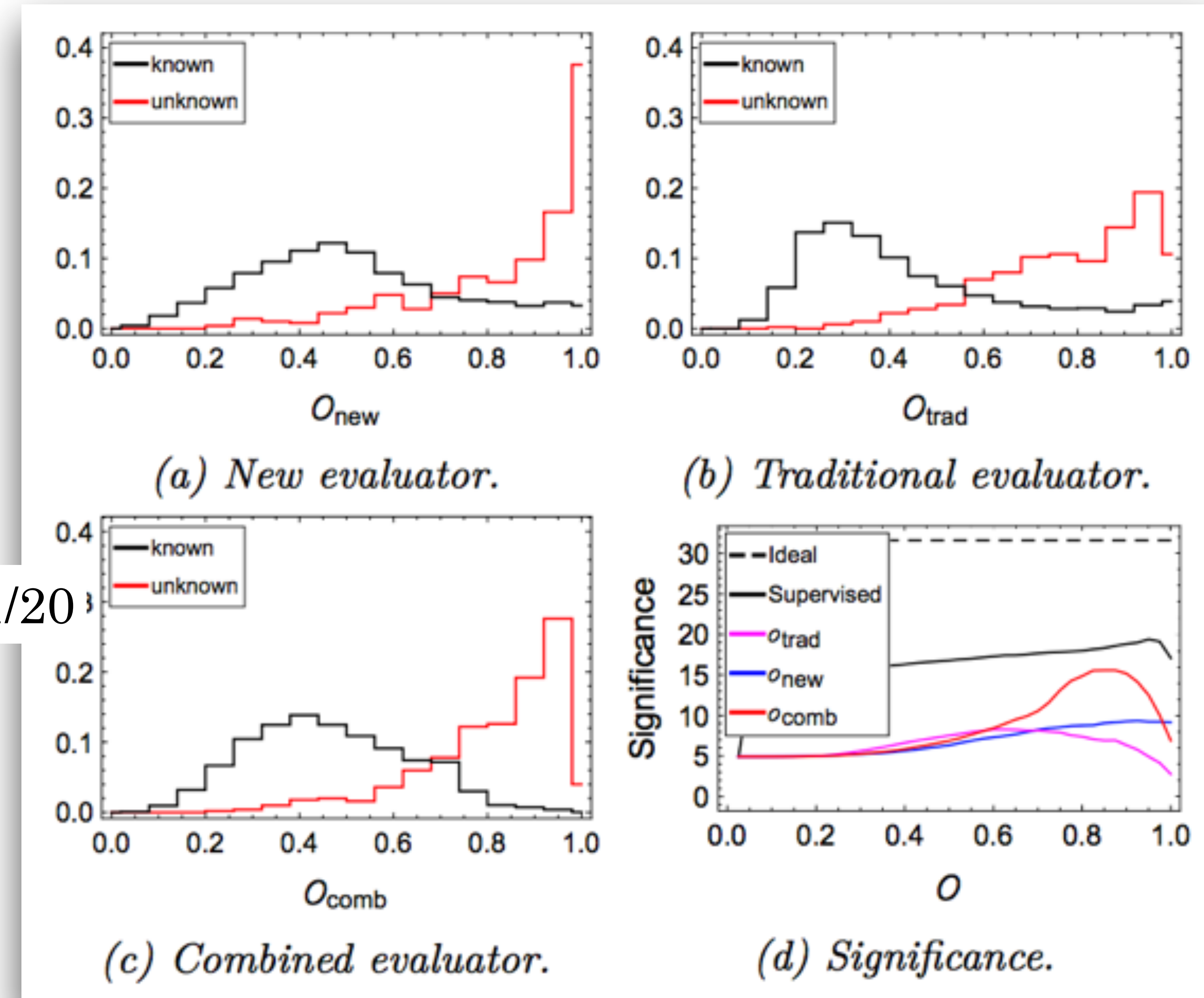


(a) Training data.

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



(b) Testing data.

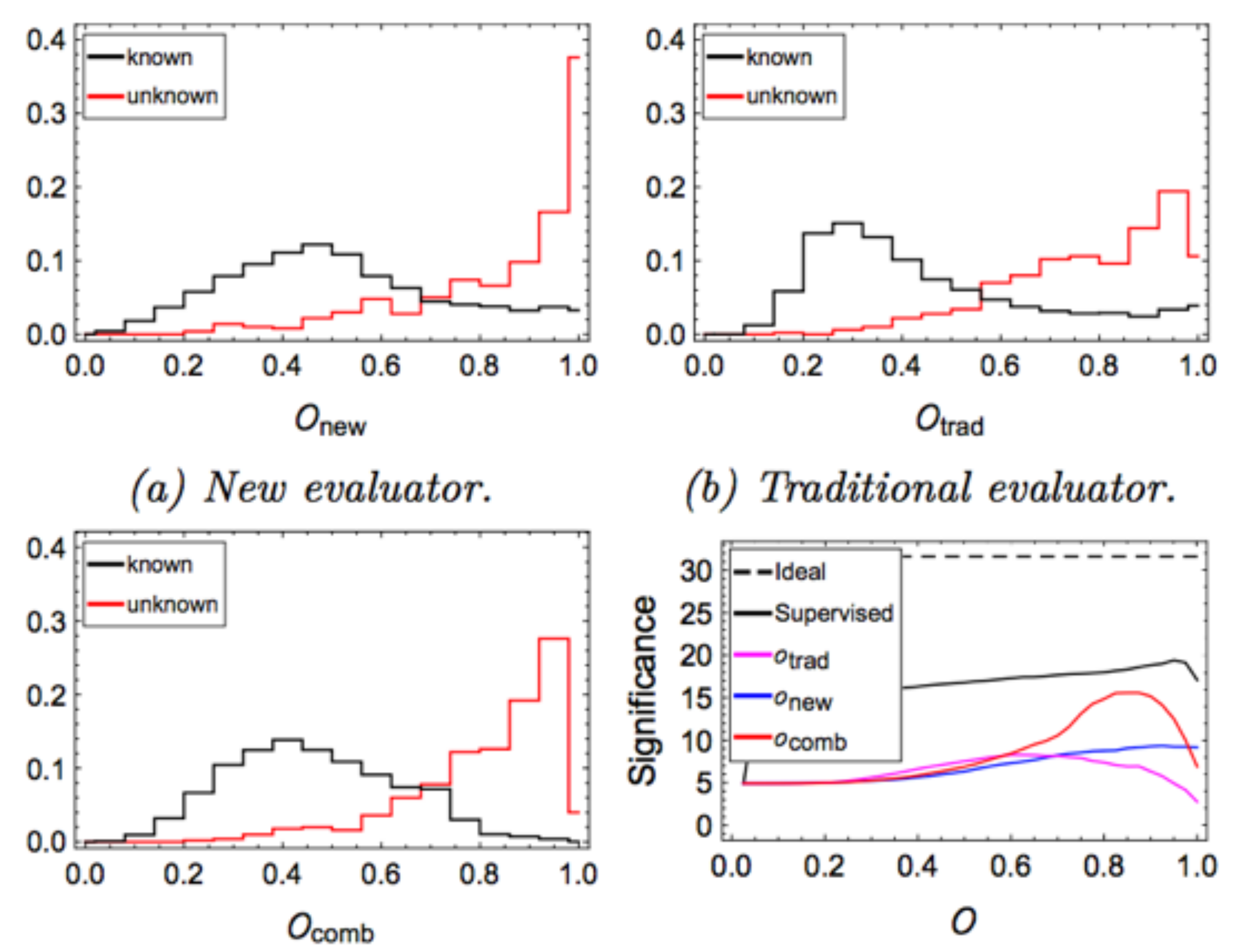


(a) New evaluator. *(b) Traditional evaluator.*

(c) Combined evaluator. *(d) Significance.*

Strategy to Address Large LEE

- Many high-scoring data of known pattern in Fig. a are pushed to the low-scoring end in Fig. c, due to the compensation of O_trad as indicated in Fig. b.
- => ~ 50% improvement in detection sensitivity!



(a) New evaluator. (b) Traditional evaluator.
(c) Combined evaluator. (d) Significance.

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)

- $pp \to \bar{t}_l t_l$, $\sigma = 11.5\,\text{fb}$,
- $pp \to t_l \bar{b} W_l^\pm$, $\sigma = 0.365\,\text{fb}$,
- $pp \to Z_b Z_l$, $\sigma = 0.0765\,\text{fb}$.

$\mathbf{X_1}$: $pp \to \bar{T}T \to W_l^+ W_l^- \bar{b}b$

$\mathbf{X_2}$: $pp \to Z' \to \bar{t}t$

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

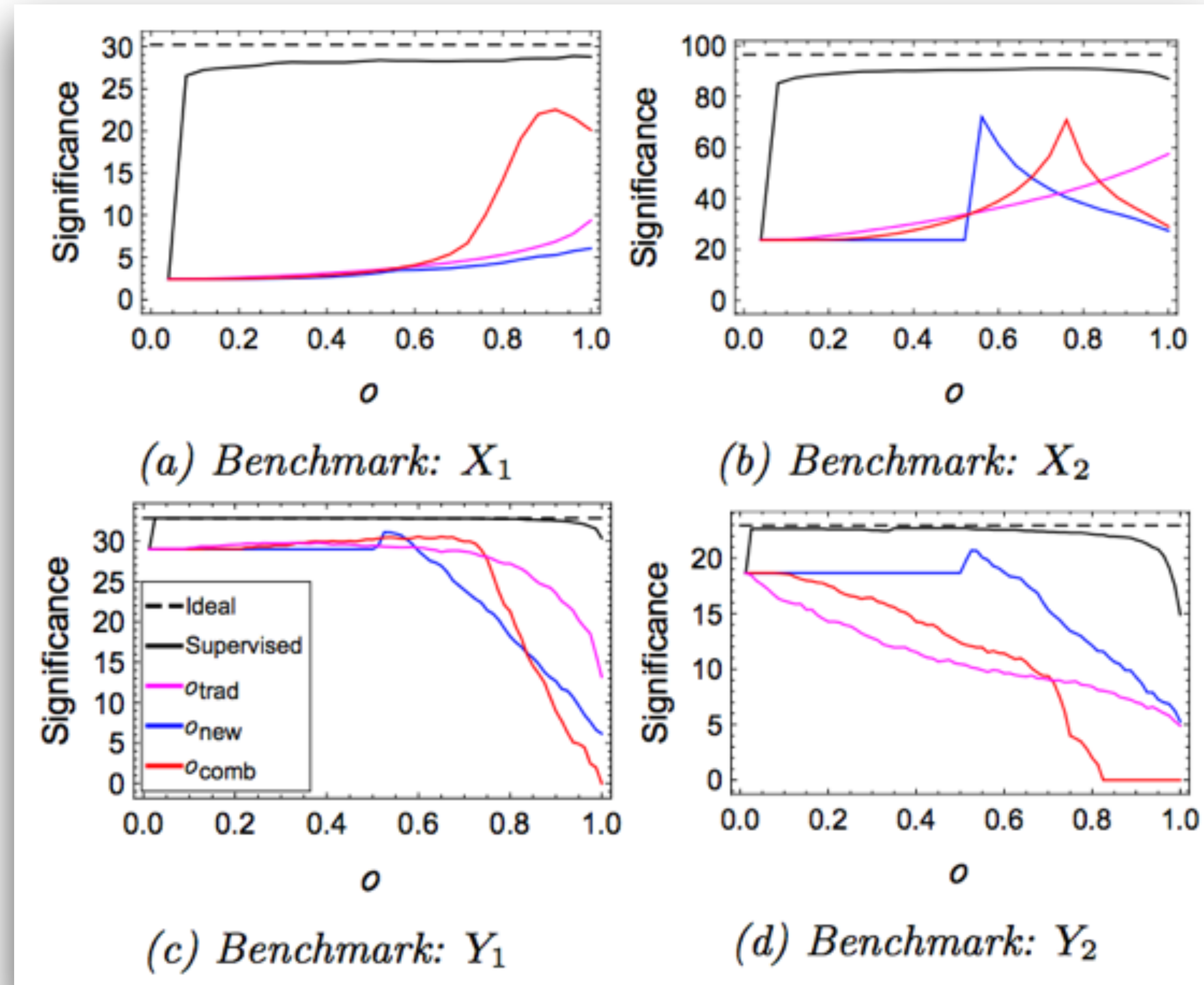
- $e^+e^- \to hZ \to Z^*_{\text{inv}} Z_{\bar{b}b} l^+ l^-$ $\sigma = 0.00686\,\text{fb}$,
- $e^+e^- \to hZ \to Z^*_{\bar{b}b} Z_{\text{inv}} l^+ l^-$ $\sigma = 0.00259\,\text{fb}$.

$\mathbf{Y_1}$: $h \to \tilde{\chi}_1 \tilde{\chi}_2 \to \tilde{\chi}_1 \tilde{\chi}_1 a$.

$\mathbf{Y_2}$: $h \to Za$

	Parameter values	$\sigma(fb)$
$X1$	$m_T = m_{\overline{T}}$ 1.2 TeV, BR$(T \to W_l^+ b) = 50\,\%$	0.152
$X2$	$m_{Z'} = 3\,\text{TeV}$, $g_{Z'} = g_Z$, 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 based on 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 utilize some 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 ``BSM'' scenarios? (Question raised by Junjie Zhu)
- Is it possible to invent a novelty evaluator to exploit multiple measures at once? (Question raised by Aurelio Juste)
- … … … …

Comparison with Recent Efforts

- 1806.02350 (D'Agnolo and Wulzer) and 1807.06038 (Simone and Jacques)
- Similar motivations are shared by all of the three efforts.
- Algorithms: ``differential'' approach (1807.10261) vs. ``integral'' method (1806.02350, 1807.06038)

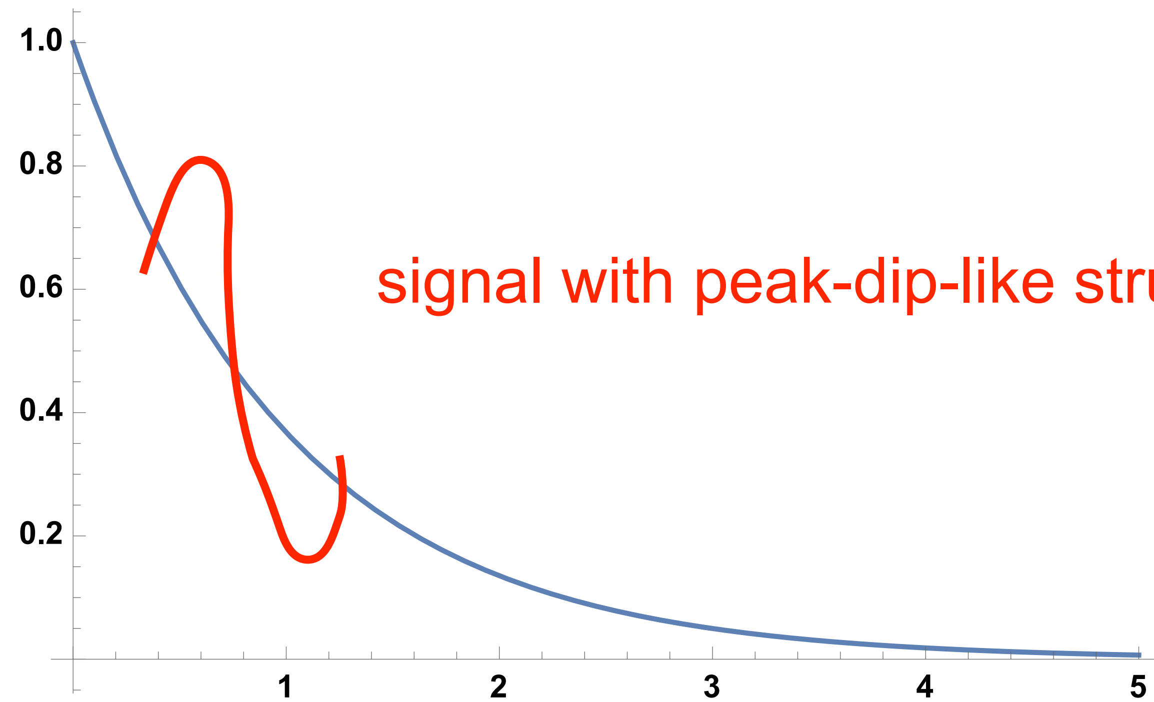
$$t(\mathcal{D}) = 2\,\log\left[\frac{e^{-N(\widehat{\mathbf{w}})}}{e^{-N(\mathrm{R})}}\prod_{x\in\mathcal{D}}\frac{n(x|\widehat{\mathbf{w}})}{n(x|\mathrm{R})}\right]$$

[1806.02350]

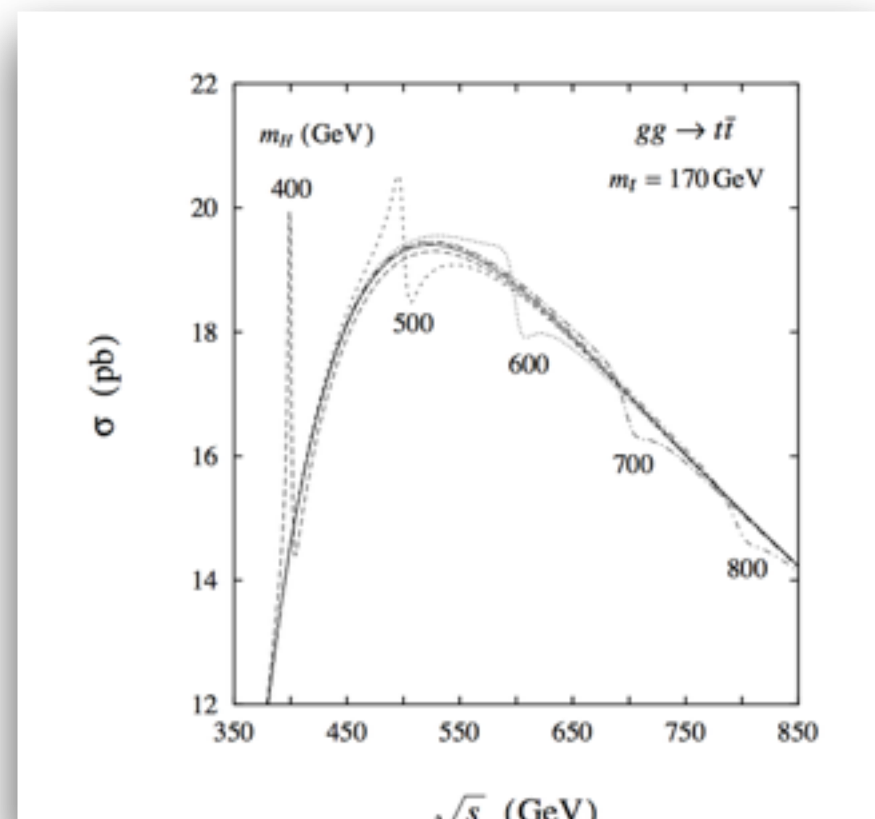
$$\mathrm{TS}(\mathcal{T}) \equiv \log\hat{\lambda}^{1/|\mathcal{T}|} = \frac{1}{N_T}\sum_{j=1}^{N_T}\log\frac{\hat{p}_T(\boldsymbol{x}_j)}{\hat{p}_B(\boldsymbol{x}_j)}$$

[1807.06038]

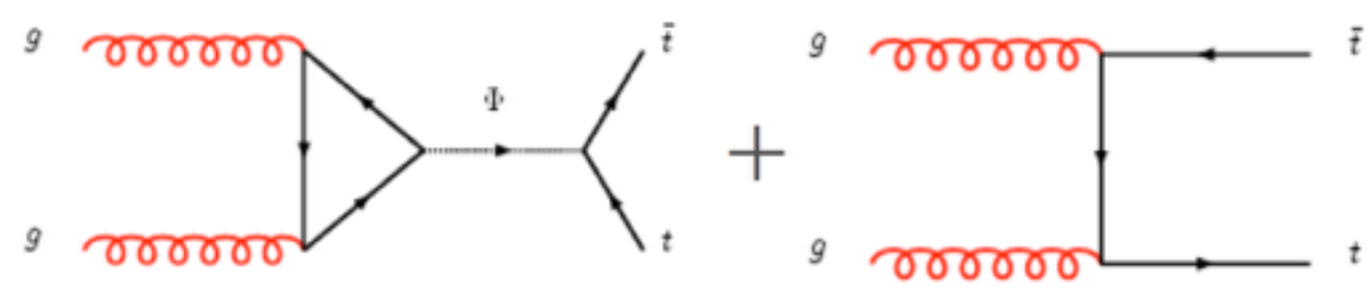
Comparison with Recent Efforts



signal with peak-dip-like structure



A famous example: (pseudo-)scalar di-top resonance



[Dicus, Stange & Willenbrock 1994]

Comparison with Recent Efforts

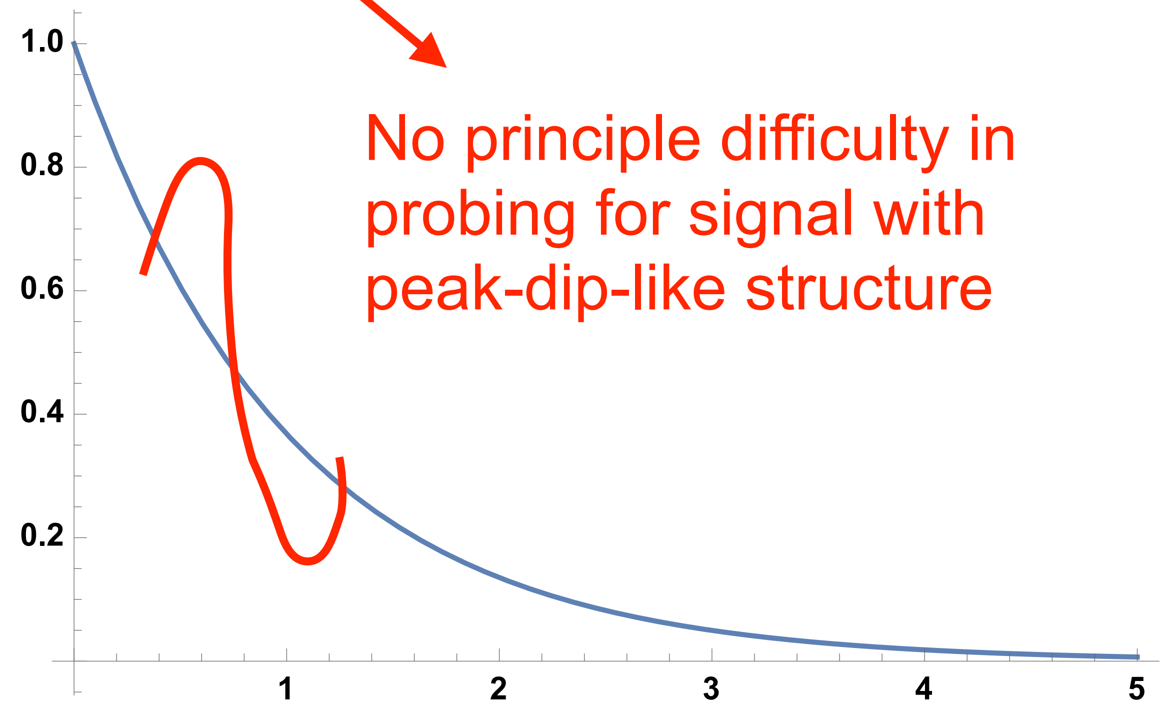
- 1806.02350 (D'Agnolo and Wulzer) and 1807.06038 (Simone and Jacques)
- Similar motivations are shared by all of the three efforts.
- Algorithms: ``differential'' approach (1807.10261) vs. ``integral'' method (1806.02350, 1807.06038)

$$t(\mathcal{D}) = 2\,\log\left[\frac{e^{-N(\widehat{\mathbf{w}})}}{e^{-N(\mathrm{R})}}\prod_{x\in\mathcal{D}}\frac{n(x|\widehat{\mathbf{w}})}{n(x|\mathrm{R})}\right]$$

[1806.02350]

$$\mathrm{TS}(\mathcal{T}) \equiv \log\hat{\lambda}^{1/|\mathcal{T}|} = \frac{1}{N_T}\sum_{j=1}^{N_T}\log\frac{\hat{p}_T(\boldsymbol{x}_j)}{\hat{p}_B(\boldsymbol{x}_j)}$$

[1807.06038]



No principle difficulty in probing for signal with peak-dip-like structure

Comparison with Recent Efforts

- 1806.02350 (D'Agnolo and Wulzer) and 1807.06038 (Simone and Jacques)
- Similar motivations are shared by all of the three efforts.
- Algorithms: ``differential'' approach (1807.10261) vs. ``integral'' method (1806.02350, 1807.06038)

A performance comparison among different algorithms is informative and necessary, and might be pursued in the near future

Summary

- The rapid development of the DNN techniques may bring far-reaching influence for collider physics / particle physics
- We explore the potential role of novelty detection in particle physics
- Complementary to supervised learning, novelty detection allows data to be analyzed model-independently. => A combination of them may lay out a framework for the future data analysis in particle physics
- By properly designing novelty evaluators, encouragingly high sensitivity can be achieved for detecting the BSM physics (at least for benchmarks considered here)
- Following-up project is on-going, in collaboration with experimental colleagues


Thank you!