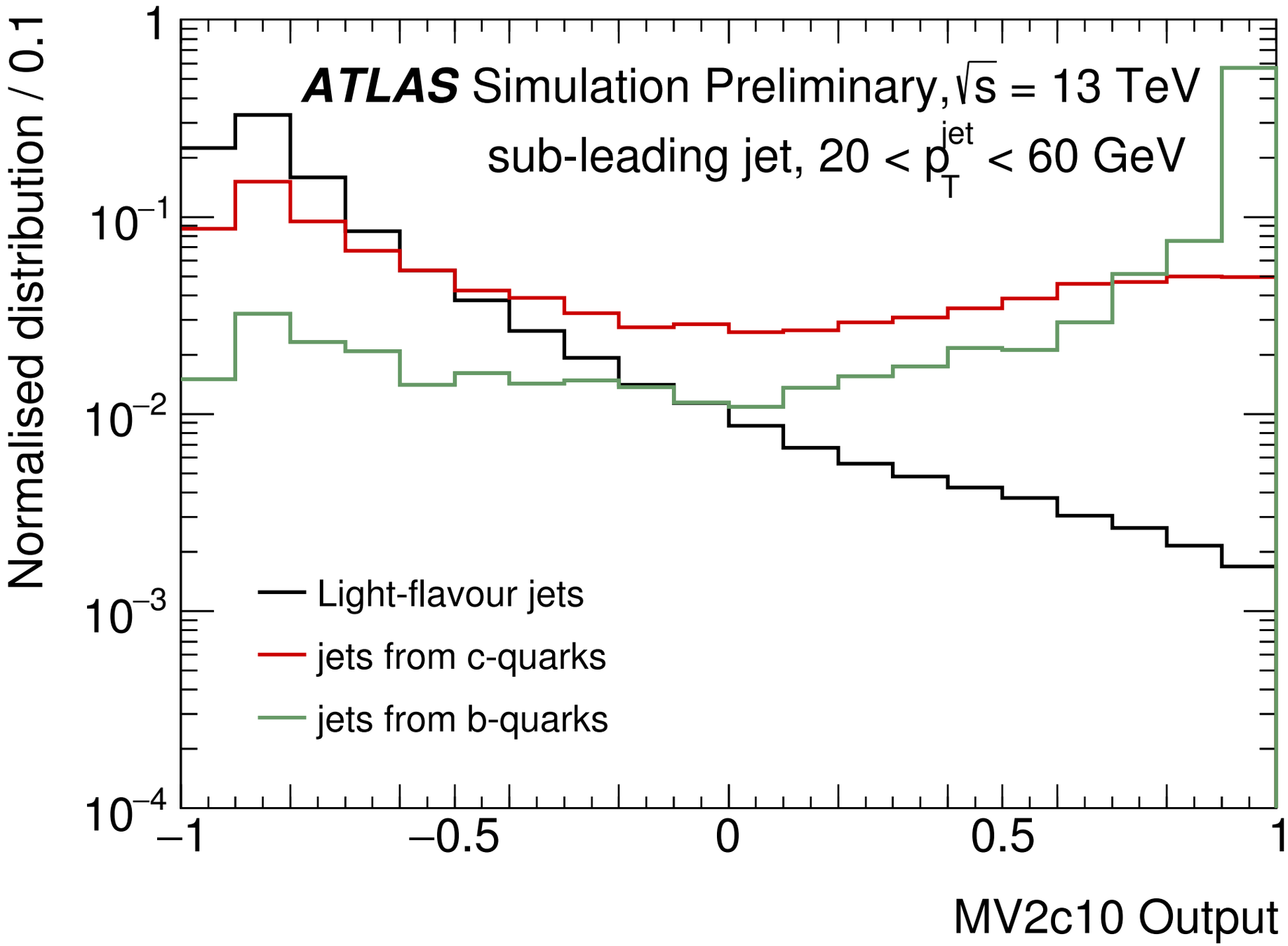
Learning when you know (basically) nothing

Benjamin Nachman

Lawrence Berkeley National Laboratory

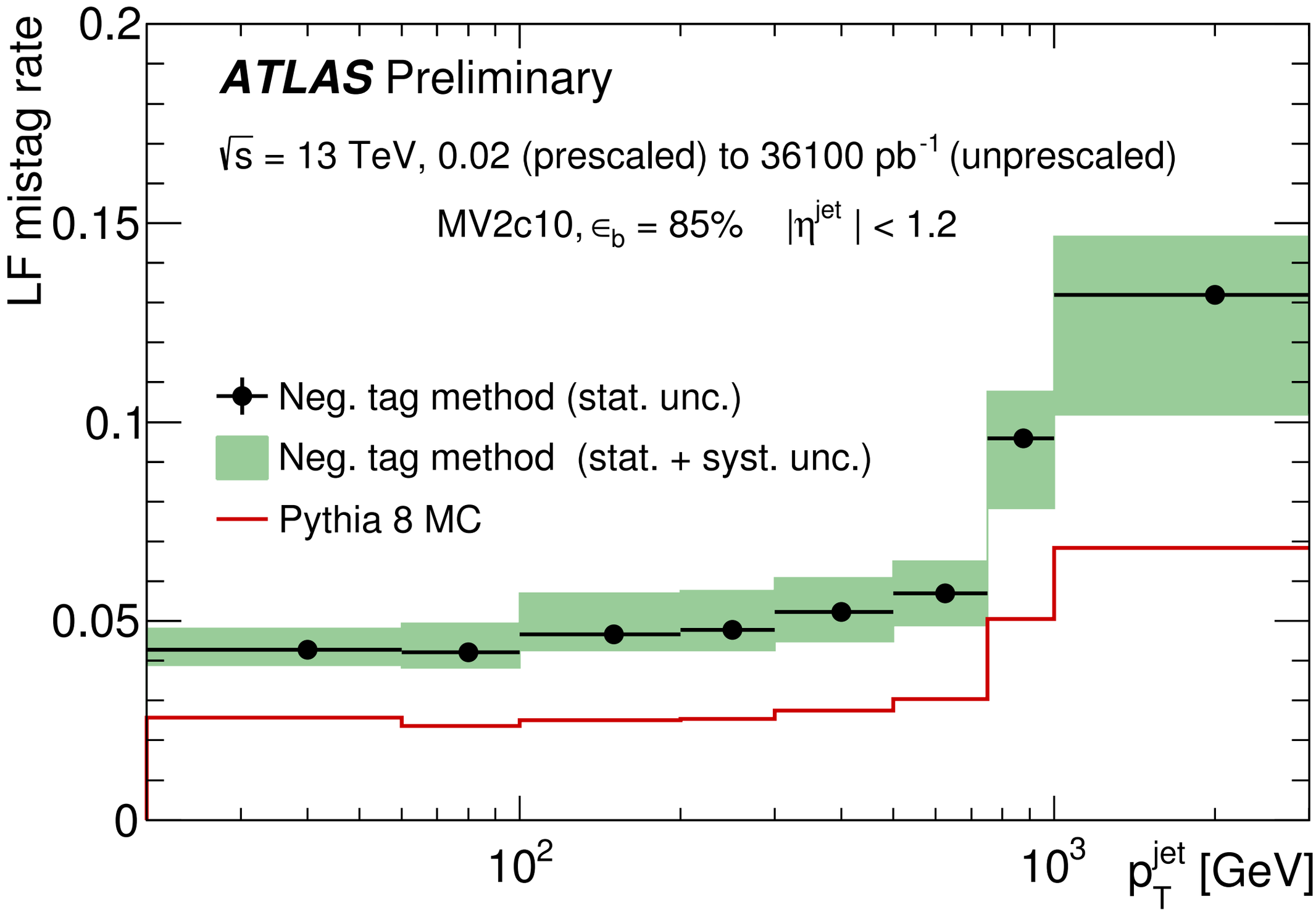
Outline: Background ⇢ LLP
⇢ CWoLA ⇢ The future

Motivation



Current paradigm: Train on simulation -> Test on data

Motivation



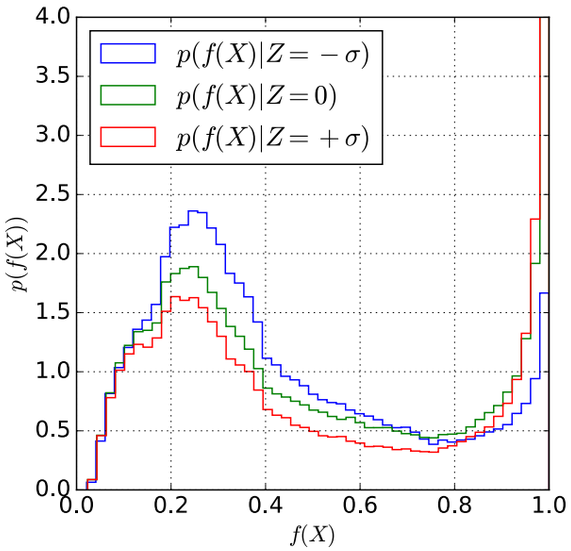
Current paradigm: Train on simulation **-> Test on data**

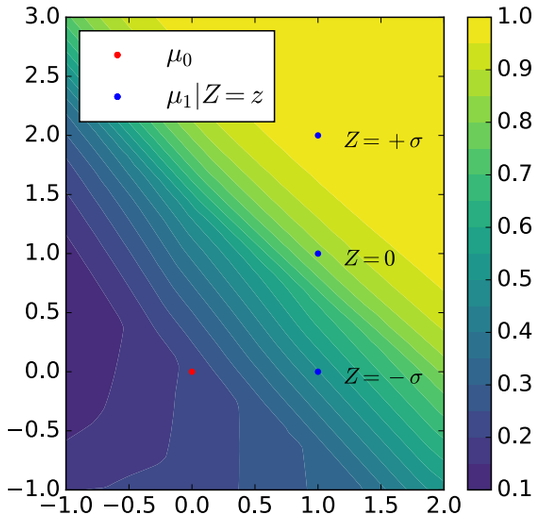
Previous proposals

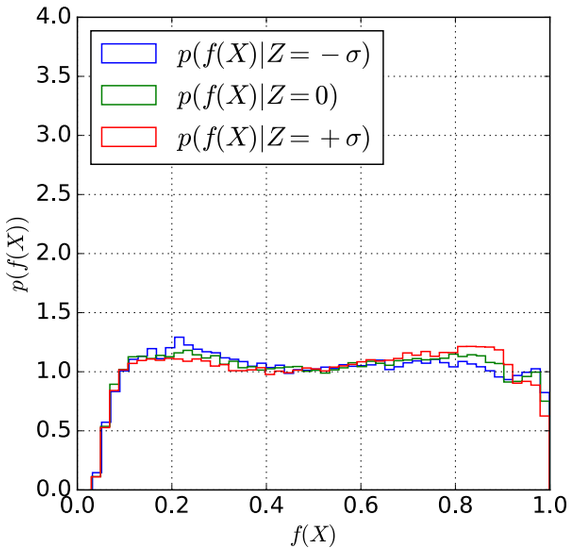
Use adversaries to penalize learning data versus simulation

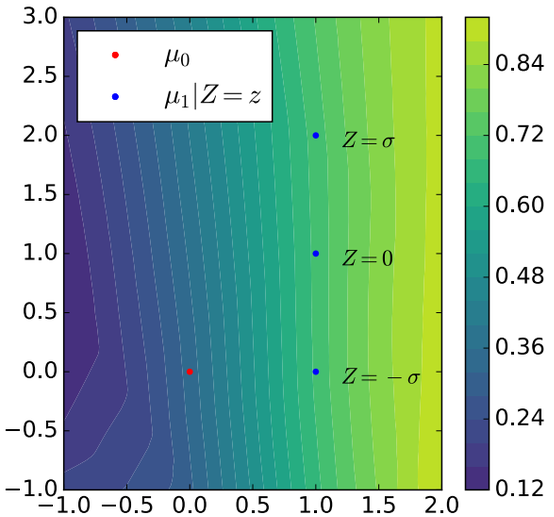
Louppe et al.

1611.01046







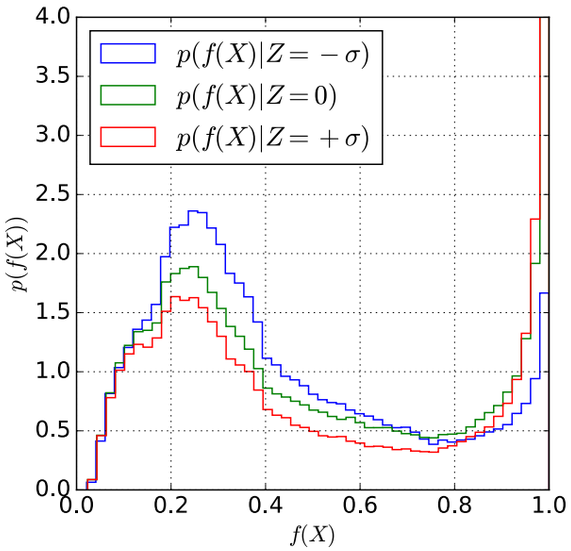


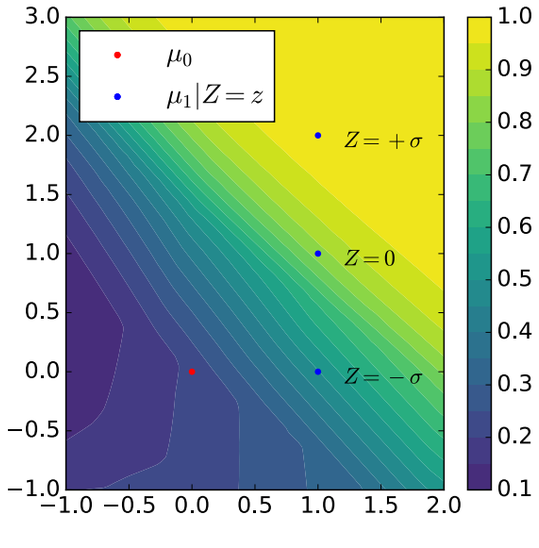
Previous proposals

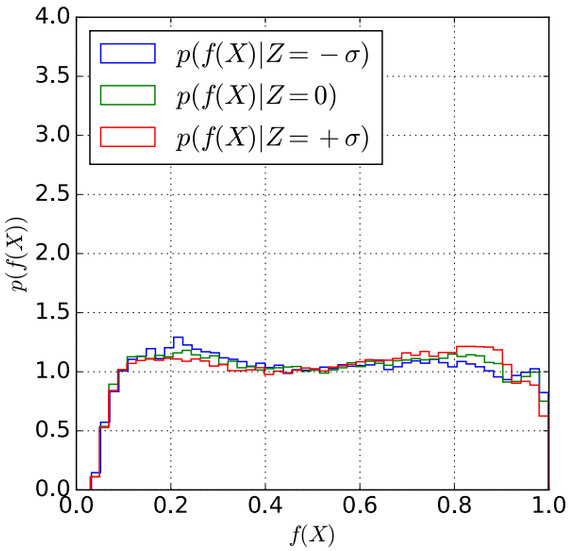
Use adversaries to penalize learning data versus simulation

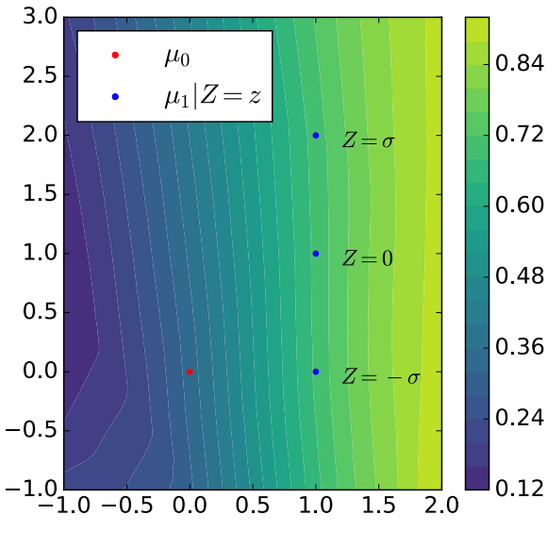
Louppe et al.

1611.01046









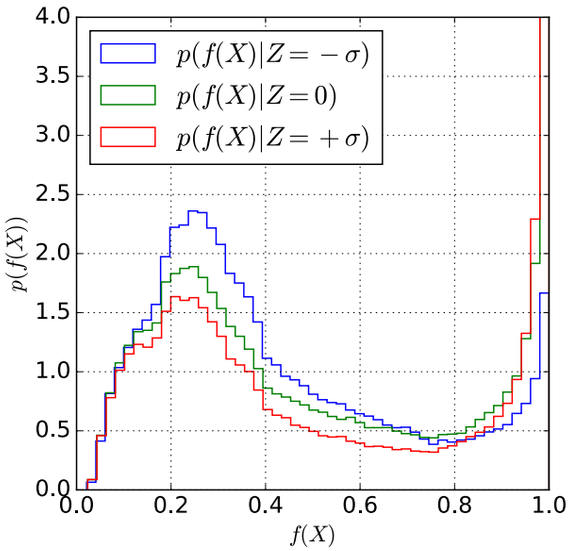
What if there is discriminating information in the data that is not in the MC? The resulting classifier will be sub-optimal.

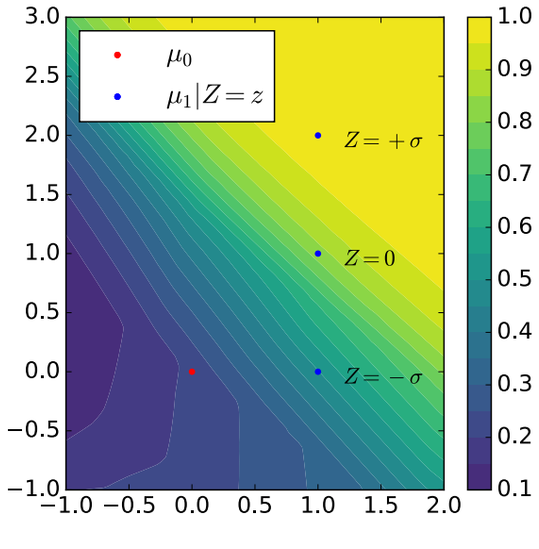
Previous proposals

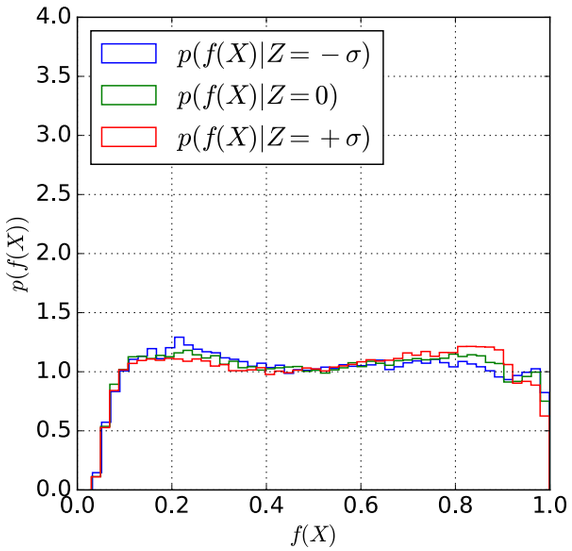
Use adversaries to penalize learning data versus simulation

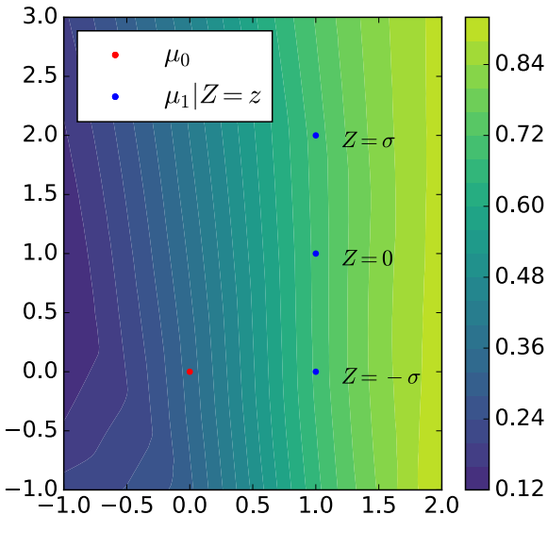
Louppe et al.

1611.01046









What if there is discriminating information in the data that is not in the MC? The resulting classifier will be sub-optimal.

Solution: train directly on (unlabeled) data!

New Proposal #1: Weak supervision

Weakly Supervised Classification in High Energy Physics

Lucio Mwinmaarong Dery, Benjamin Nachman, Francesco Rubbo, Ariel Schwartzman

(Submitted on 1 Feb 2017 (v1), last revised 3 Jul 2017 (this version, v4))

As machine learning algorithms become increasingly sophisticated to exploit subtle features of the data, they often become mo dependent on simulations. This paper presents a new approach called weakly supervised classification in which class proportio the only input into the machine learning algorithm. Using one of the most challenging binary classification tasks in high energy physics - quark versus gluon tagging - we show that weakly supervised classification can match the performance of fully super algorithms. Furthermore, by design, the new algorithm is insensitive to any mis-modeling of discriminating features in the data simulation. Weakly supervised classification is a general procedure that can be applied to a wide variety of learning problems to performance and robustness when detailed simulations are not reliable or not available.

Comments: 8 pages, 4 figures
Subjects: **High Energy Physics - Phenomenology (hep-ph)**; Data Analysis, Statistics and Probability (physics.data-an); Machine L (stat.ML)
Journal reference: JHEP 05 (2017) 145
DOI: 10.1007/JHEP05(2017)145
Cite as: **arXiv:1702.00414 [hep-ph]**
(or **arXiv:1702.00414v4 [hep-ph]** for this version)

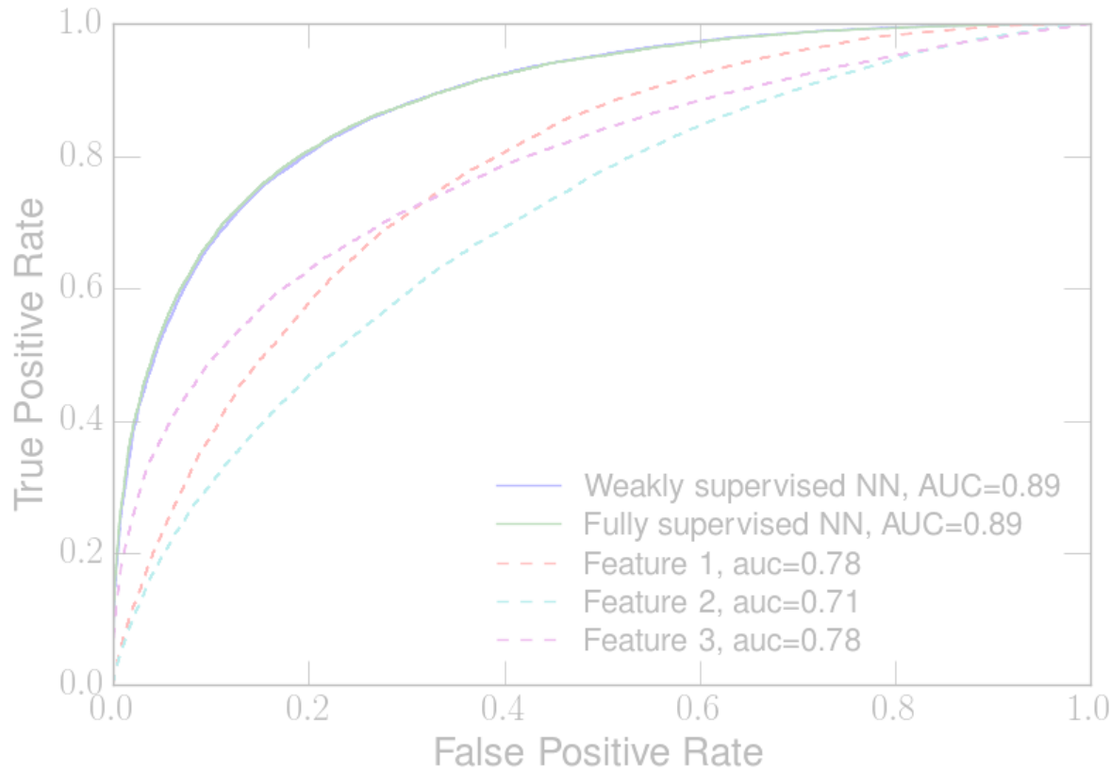
New Proposal #1: Weak supervision

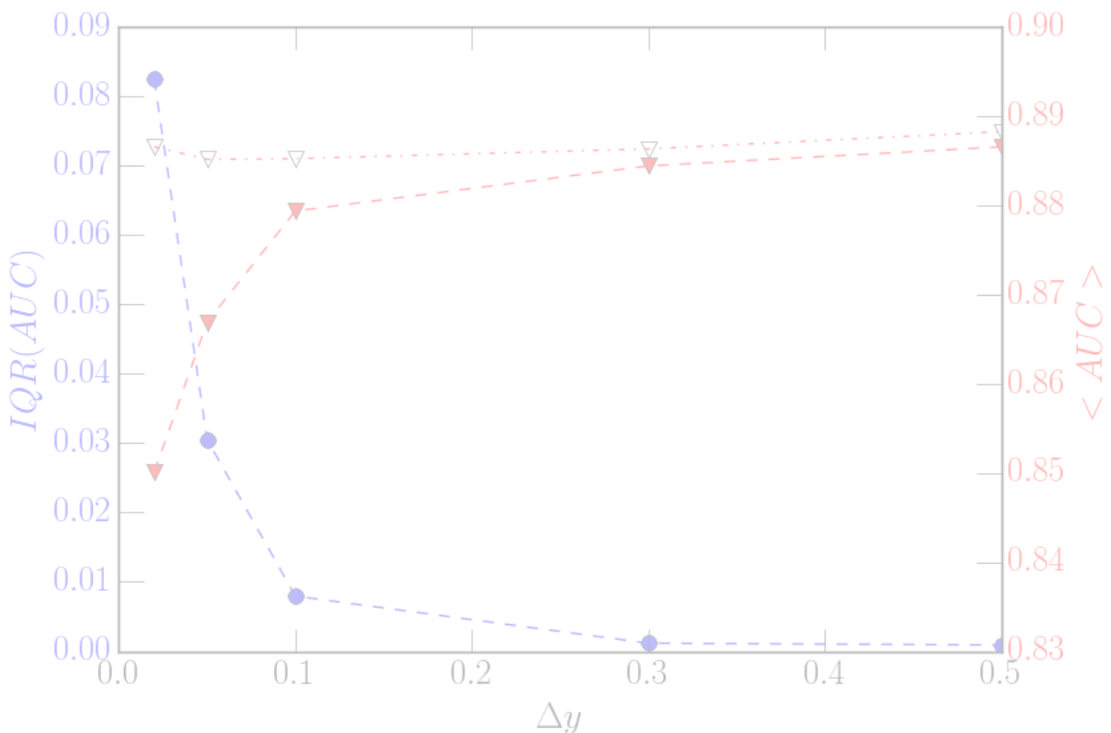
$$f_{\text{full}} = \text{argmin}_{f':\mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell\left(f'(x_i) - t_i\right)$$

$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

intuition:

$$h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i}$$
$$h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i},$$





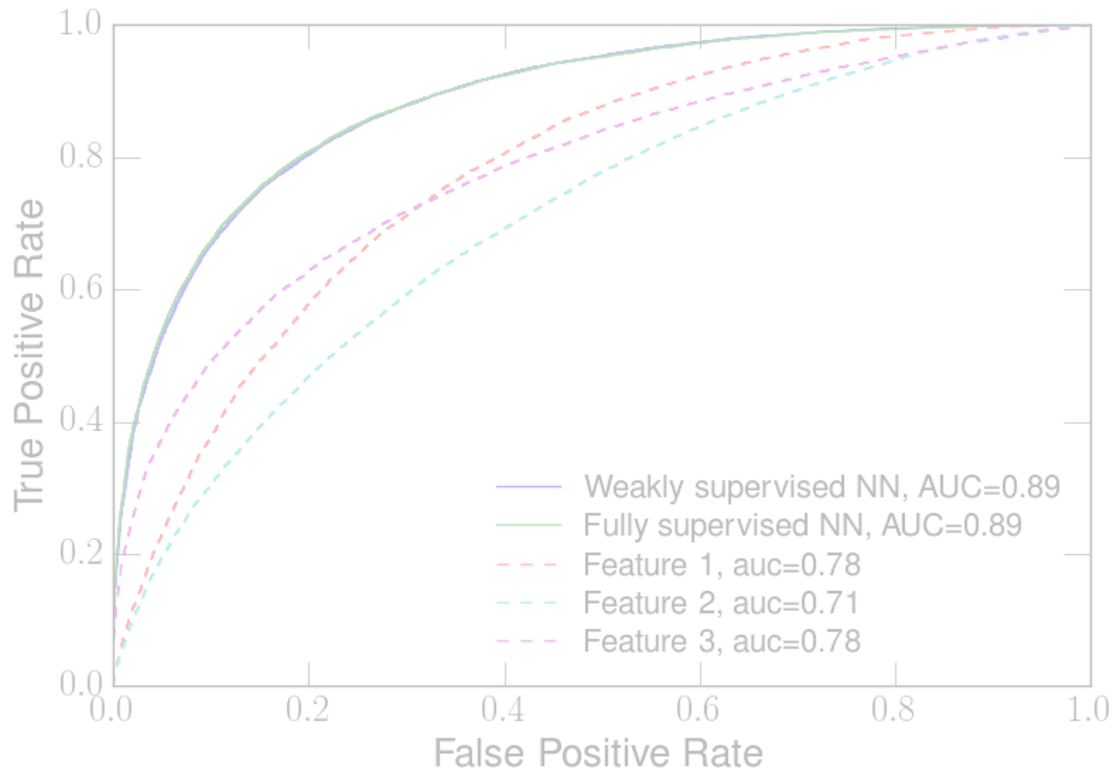
New Proposal #1: Weak supervision

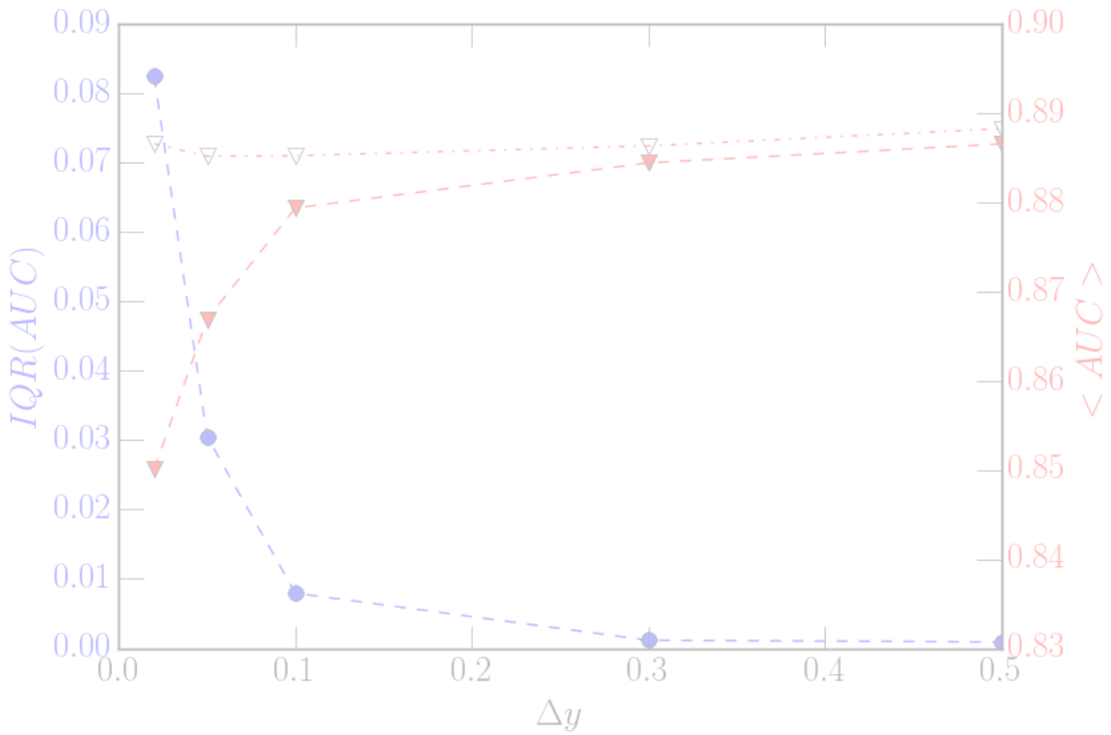
$$f_{\text{full}} = \text{argmin}_{f':\mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell\left(f'(x_i) - t_i\right)$$

$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

intuition:

$$h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i}$$
$$h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i},$$





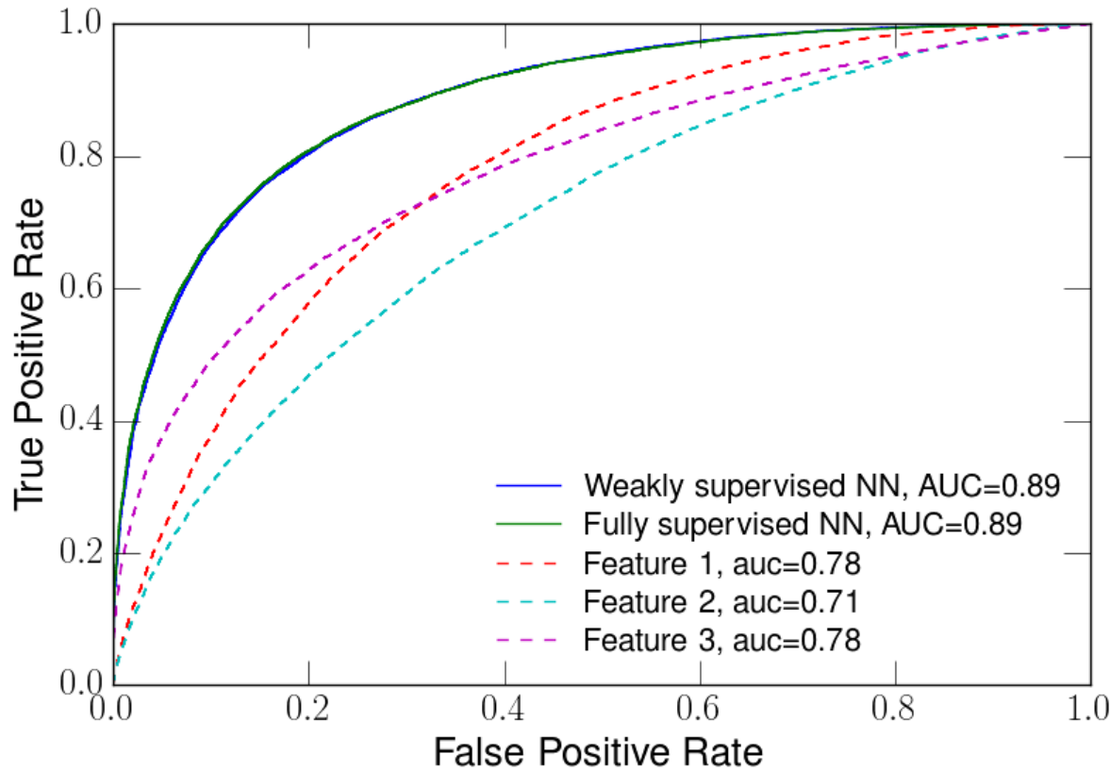
New Proposal #1: Weak supervision

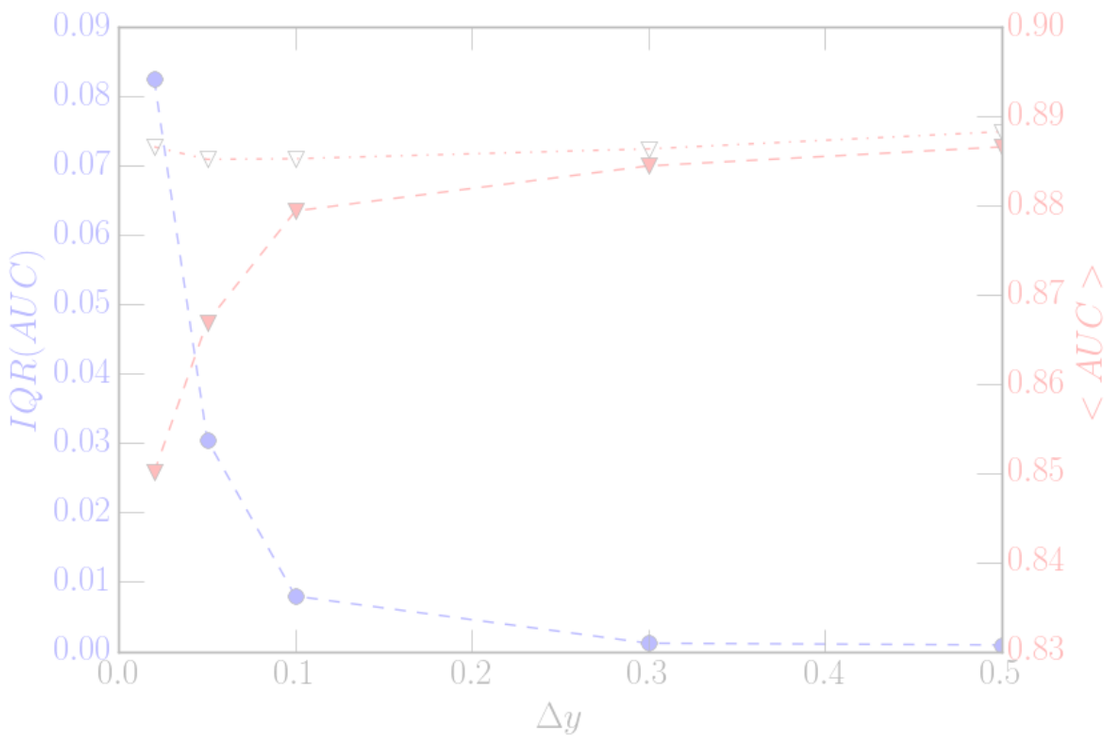
$$f_{\text{full}} = \text{argmin}_{f':\mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell\left(f'(x_i) - t_i\right)$$

$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

intuition:

$$h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i}$$
$$h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i},$$





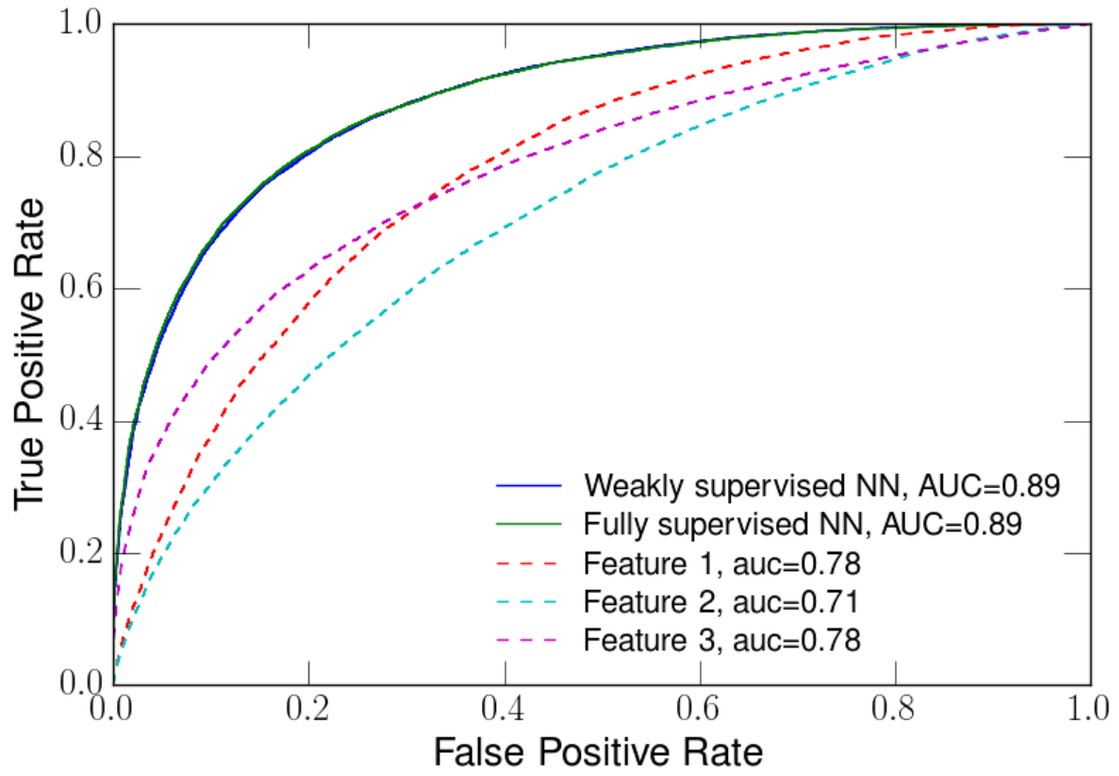
New Proposal #1: Weak supervision

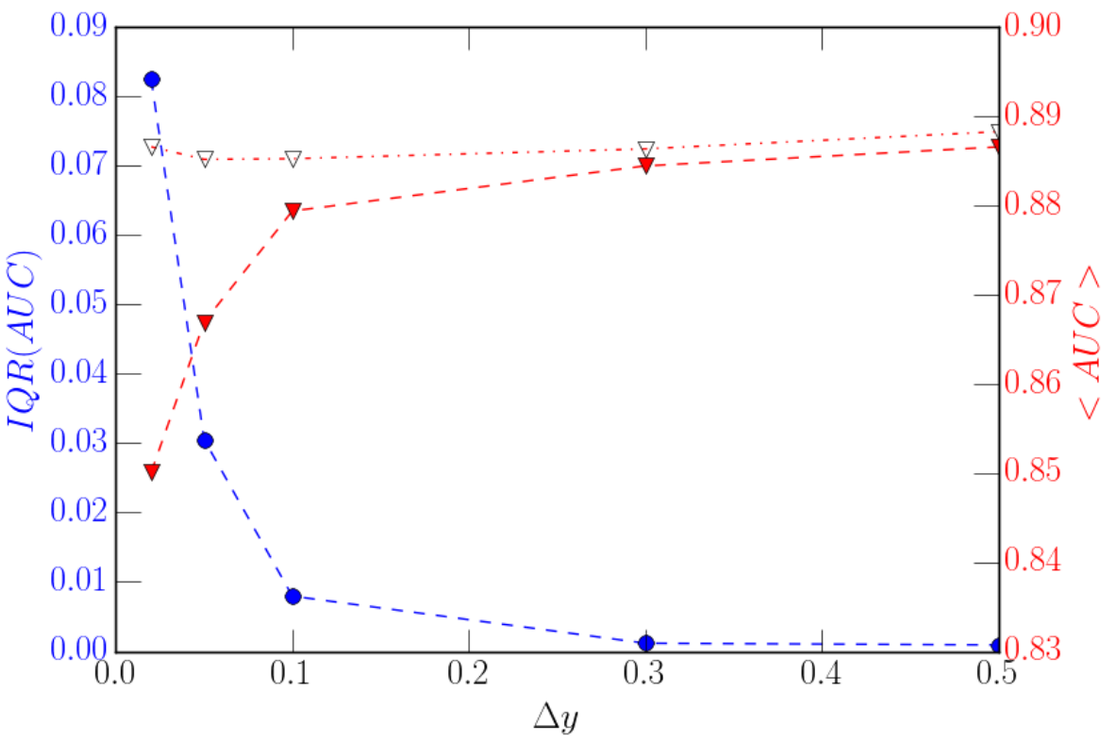
$$f_{\text{full}} = \text{argmin}_{f':\mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell\left(f'(x_i) - t_i\right)$$

$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

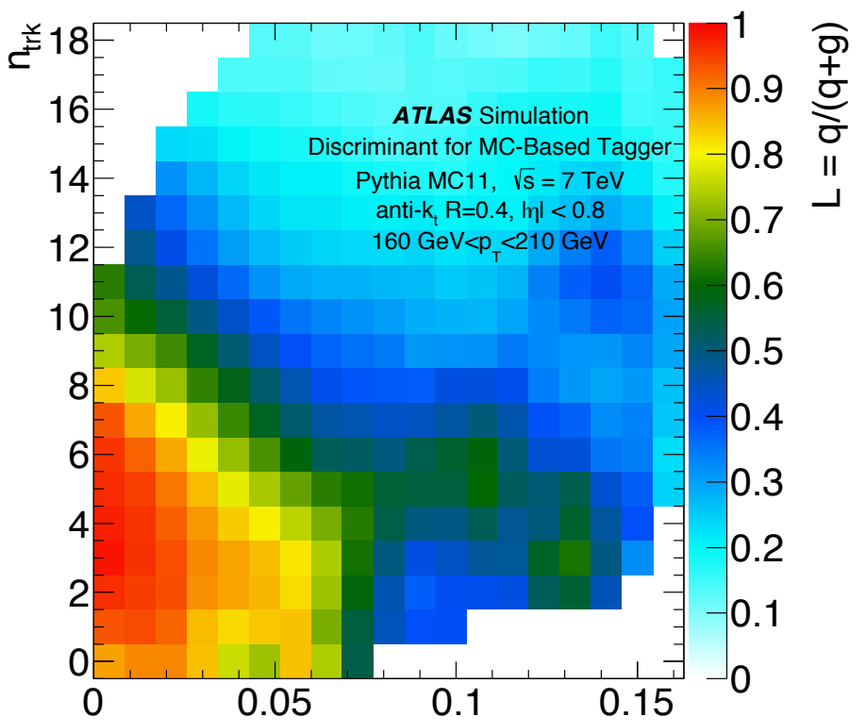
intuition:

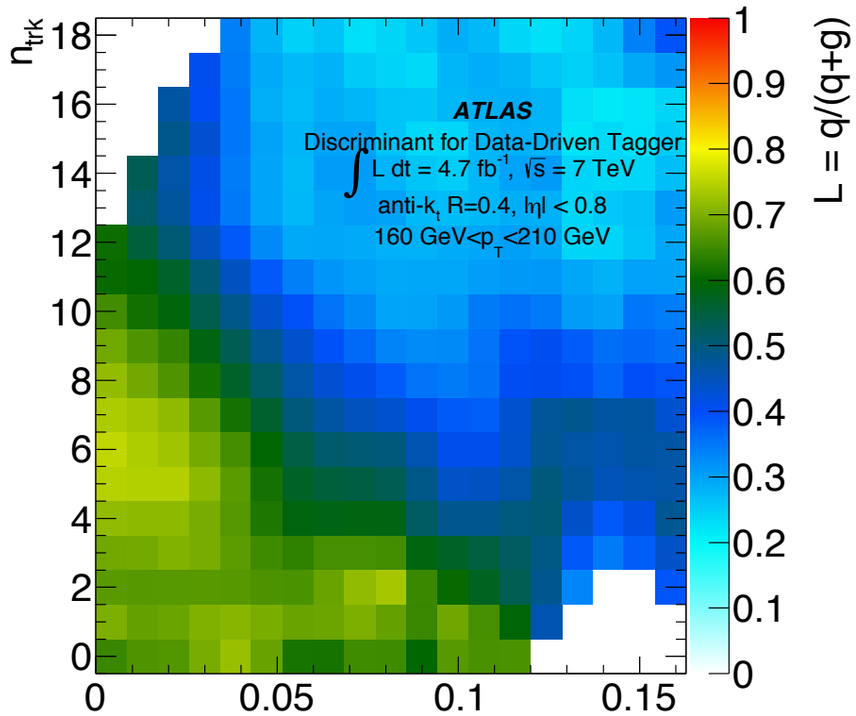
$$h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i}$$
$$h_{B,i} = y_B h_{1,i} + (1 - y_B) h_{0,i},$$

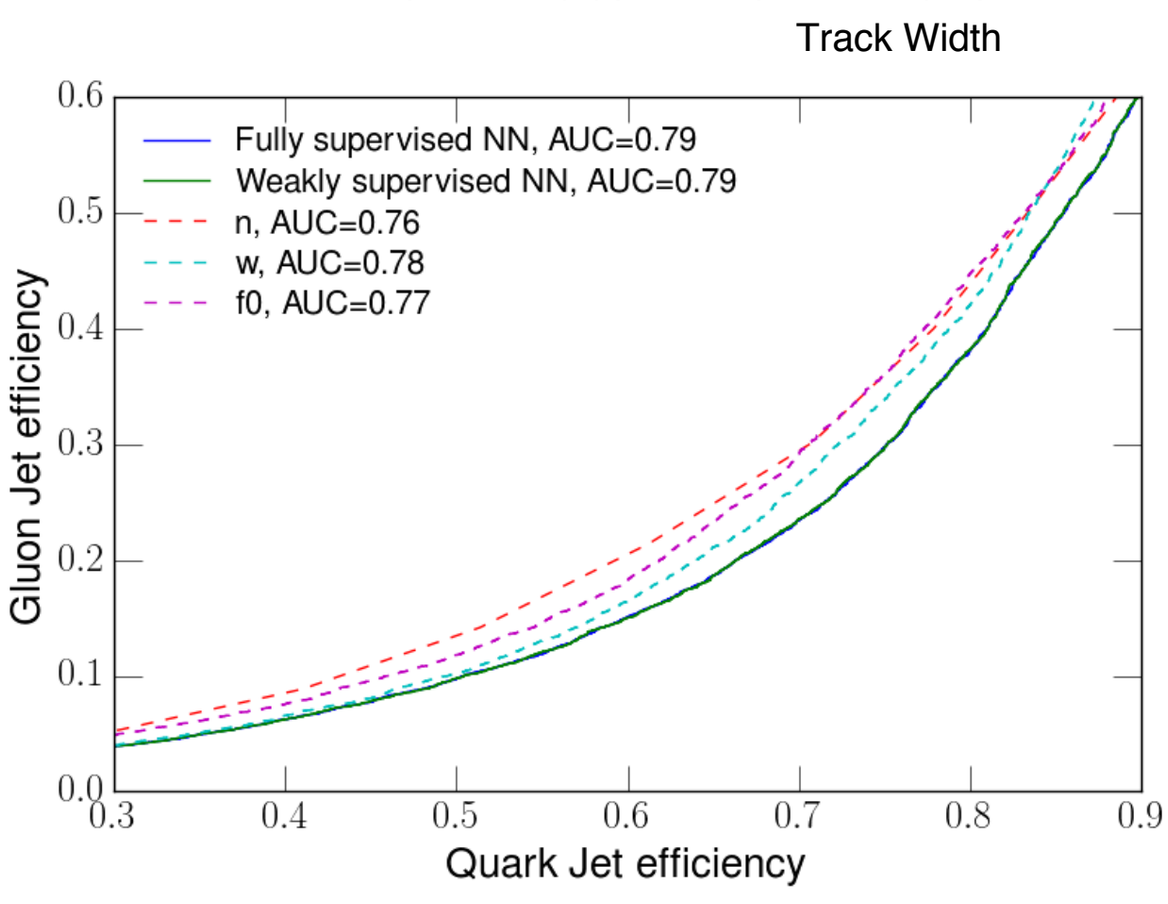


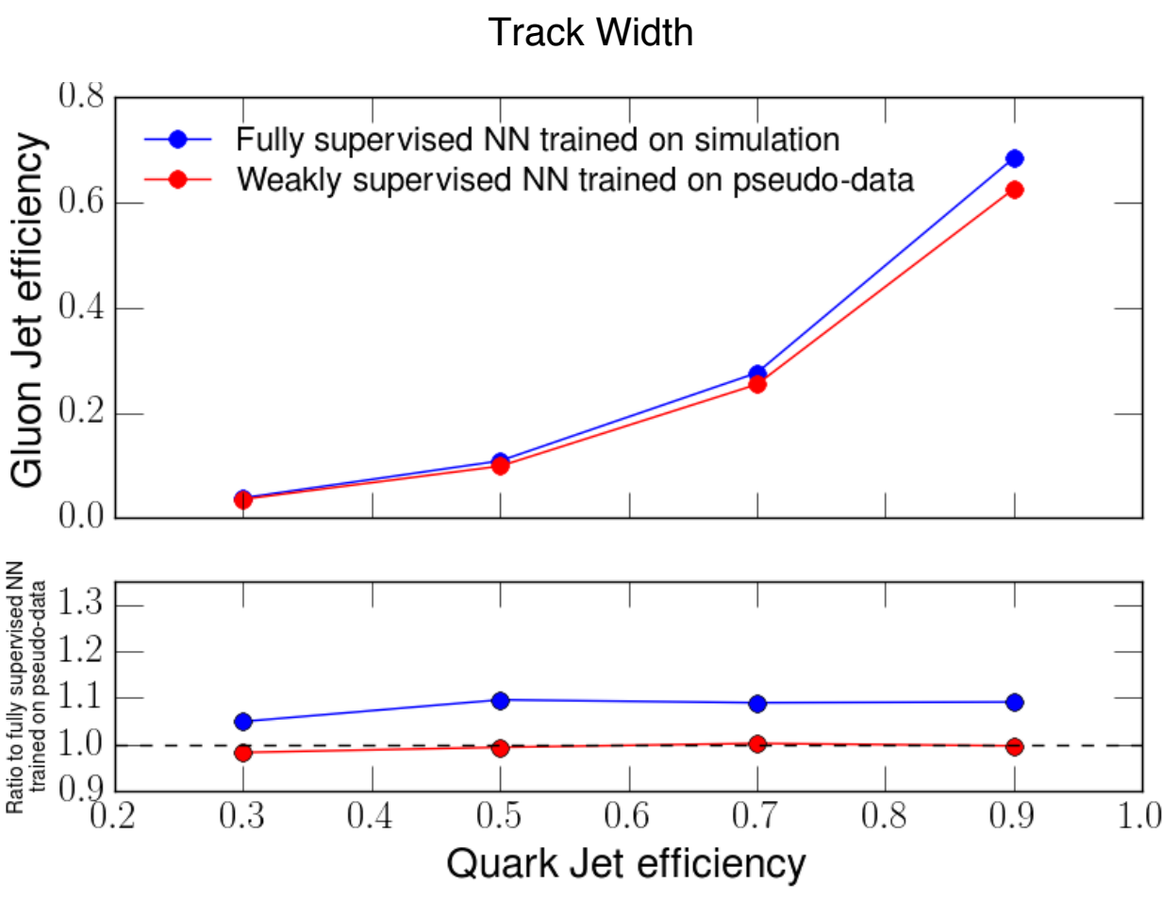


New Proposal #1: Weak supervision









New Proposal #2: Class. w/o labels (CWoLa)

Classification without labels: Learning from mixed samples in high energy physics

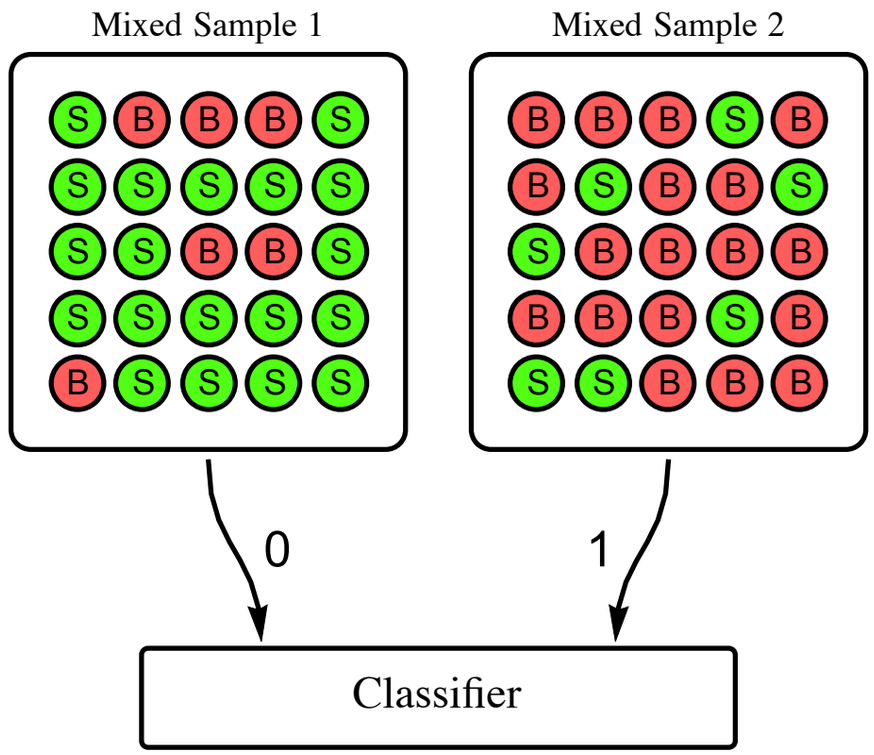
Eric M. Metodiev, Benjamin Nachman, Jesse Thaler

(Submitted on 9 Aug 2017 (v1), last revised 8 Sep 2017 (this version, v2))

Modern machine learning techniques can be used to construct powerful models for difficult collider physics problems. In many applications, however, these models are trained on imperfect simulations due to a lack of truth-level information in the data, which risks the model learning artifacts of the simulation. In this paper, we introduce the paradigm of classification without labels (CWoLa) in which a classifier is trained to distinguish statistical mixtures of classes, which are common in collider physics. Crucially, neither individual labels nor class proportions are required, yet we prove that the optimal classifier in the CWoLa paradigm is also the optimal classifier in the traditional fully-supervised case where all label information is available. After demonstrating the power of this method in an analytical toy example, we consider a realistic benchmark for collider physics: distinguishing quark- versus gluon-initiated jets using mixed quark/gluon training samples. More generally, CWoLa can be applied to any classification problem where labels or class proportions are unknown or simulations are unreliable, but statistical mixtures of the classes are available.

Comments: 16 pages, 5 figures; v2: intro extended and references added
Subjects: **High Energy Physics - Phenomenology (hep-ph)**; High Energy Physics - Experiment (hep-ex); Machine Learning (stat.ML)
Report number: MIT--CTP 4922
Cite as: arXiv:1708.02949 **[hep-ph]**
(or arXiv:1708.02949v2 **[hep-ph]** for this version)

New Proposal #2: Class. w/o labels (CWoLa)



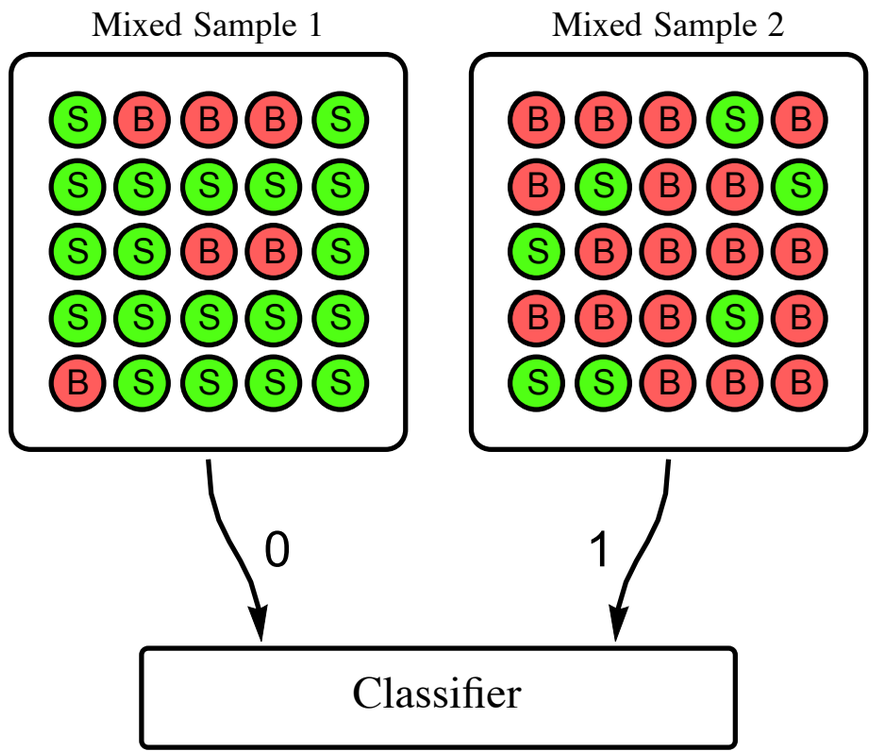
Idea: classify mixed sample 1 from mixed same 2 and then apply it to distinguish S from B.

Theorem 1. *Given mixed samples M_1 and M_2 defined in terms of pure samples S and B using Eqs. (2.3) and (2.4) with signal fractions $f_1 > f_2$, an optimal classifier trained to distinguish M_1 from M_2 is also optimal for distinguishing S from B.*

Proof. The optimal classifier to distinguish examples drawn from p_{M_1} and p_{M_2} is the likelihood ratio $L_{M_1/M_2}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$. Similarly, the optimal classifier to distinguish examples drawn from p_S and p_B is the likelihood ratio $L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. Where p_B has support, we can relate these two likelihood ratios algebraically:

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}, \tag{2.6}$$

New Proposal #2: Class. w/o labels (CWoLa)



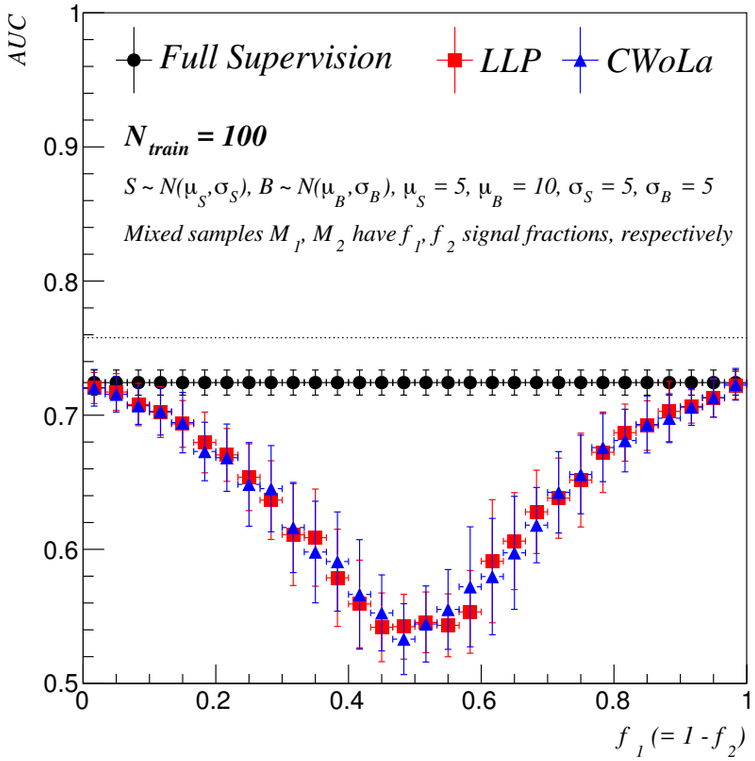
Idea: classify mixed sample 1 from mixed same 2 and then apply it to distinguish S from B.

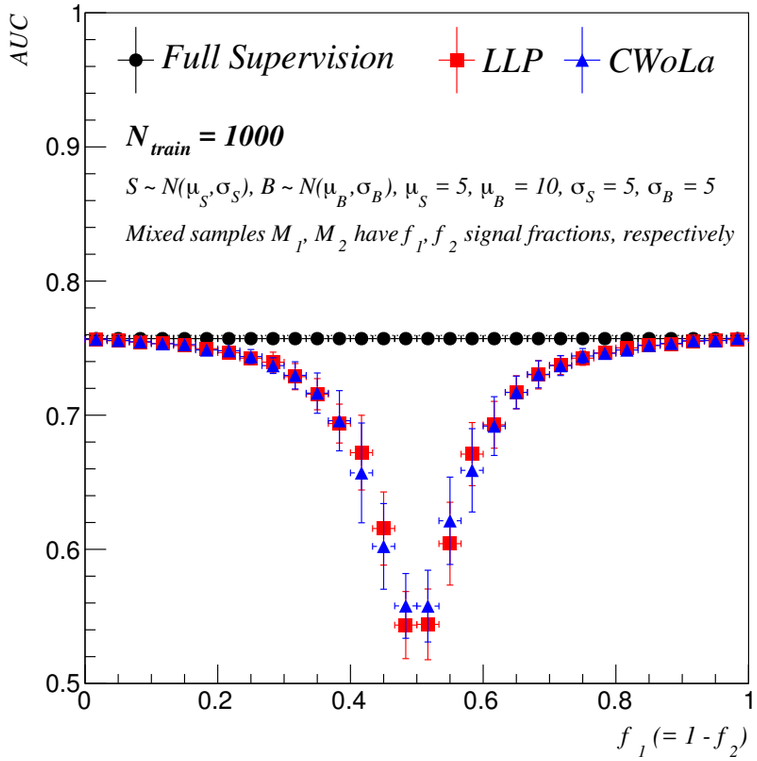
Theorem 1. *Given mixed samples M_1 and M_2 defined in terms of pure samples S and B using Eqs. (2.3) and (2.4) with signal fractions $f_1 > f_2$, an optimal classifier trained to distinguish M_1 from M_2 is also optimal for distinguishing S from B.*

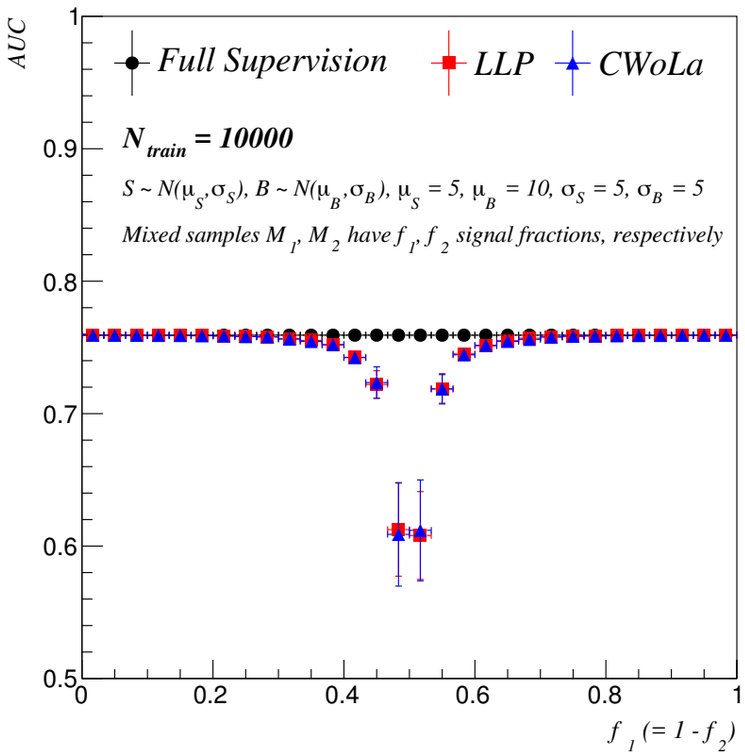
Proof. The optimal classifier to distinguish examples drawn from p_{M_1} and p_{M_2} is the likelihood ratio $L_{M_1/M_2}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$. Similarly, the optimal classifier to distinguish examples drawn from p_S and p_B is the likelihood ratio $L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. Where p_B has support, we can relate these two likelihood ratios algebraically:

$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}, \tag{2.6}$$

New Proposal #2: Class. w/o labels (CWoLa)



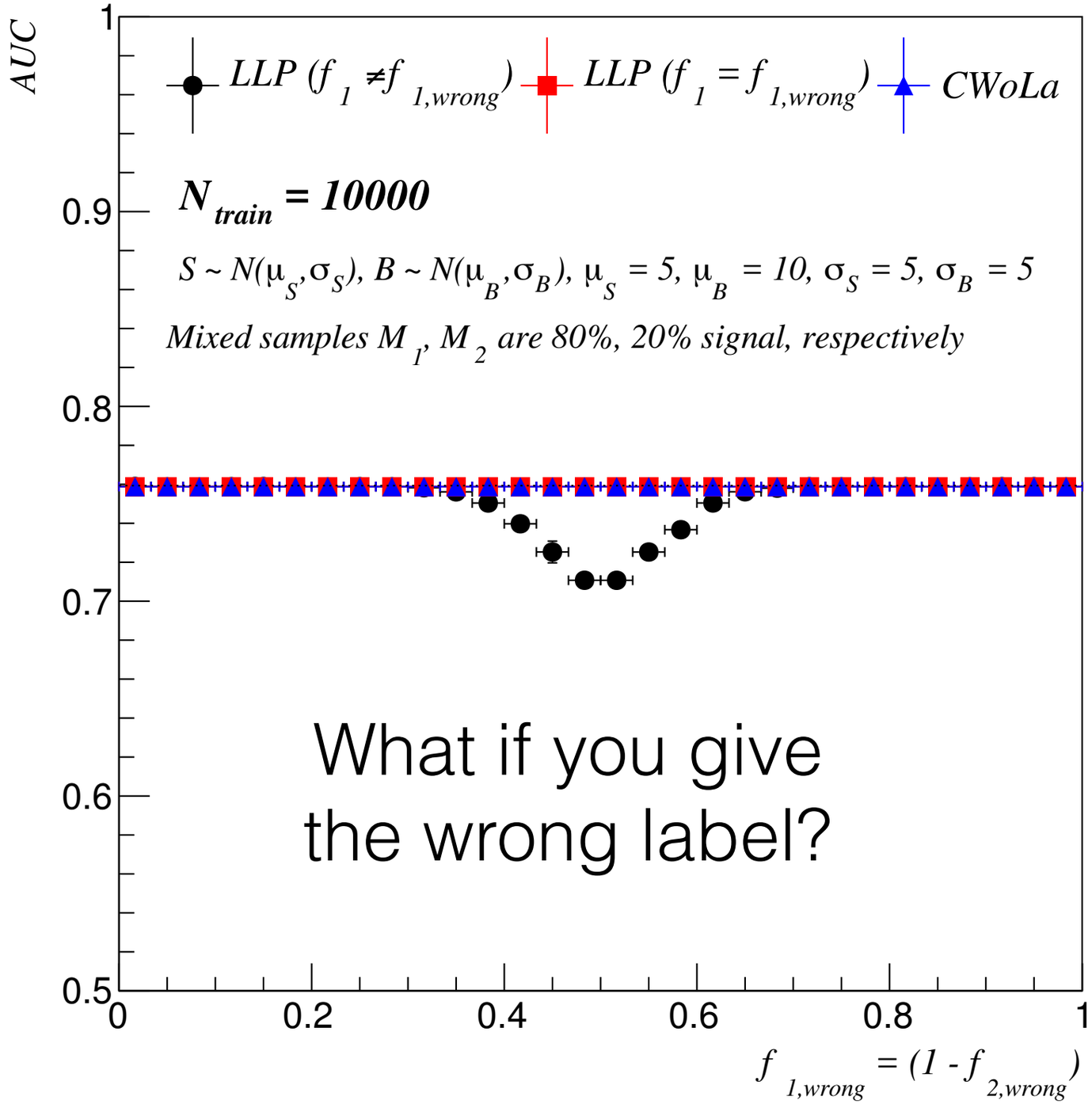




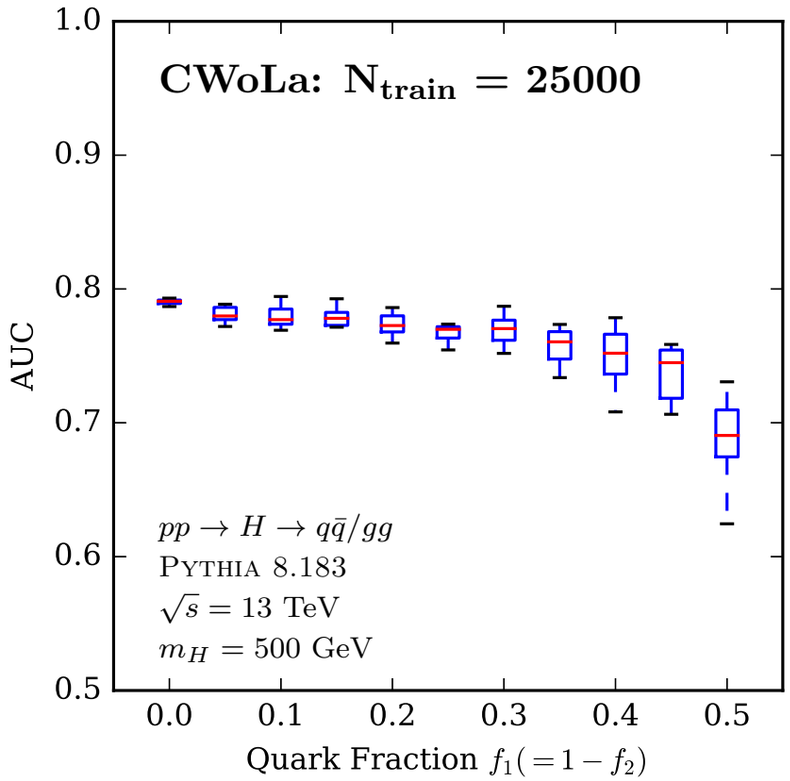
Increase the number of training events ->

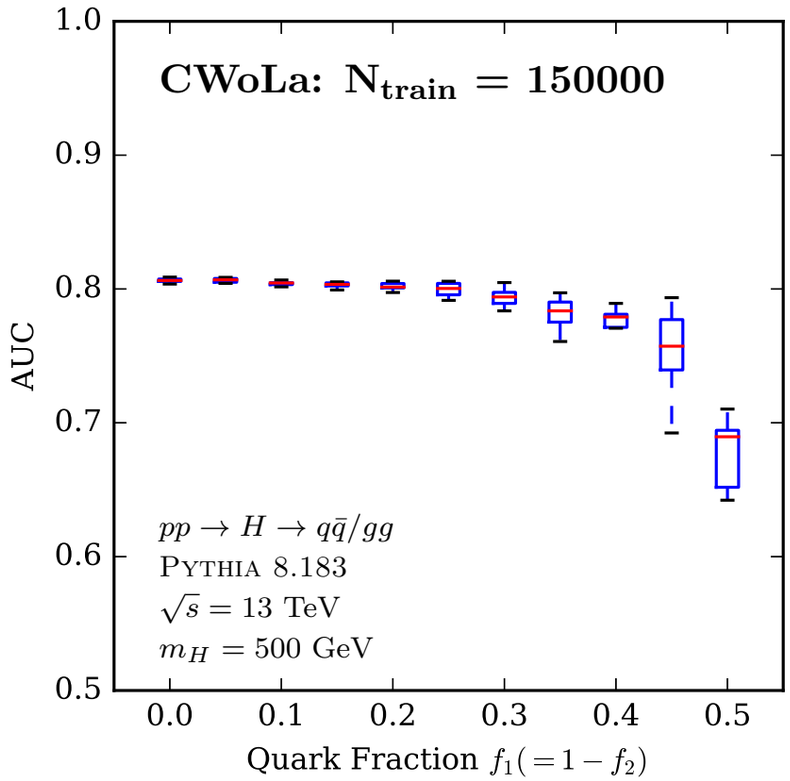
(LLP = weak supervision)

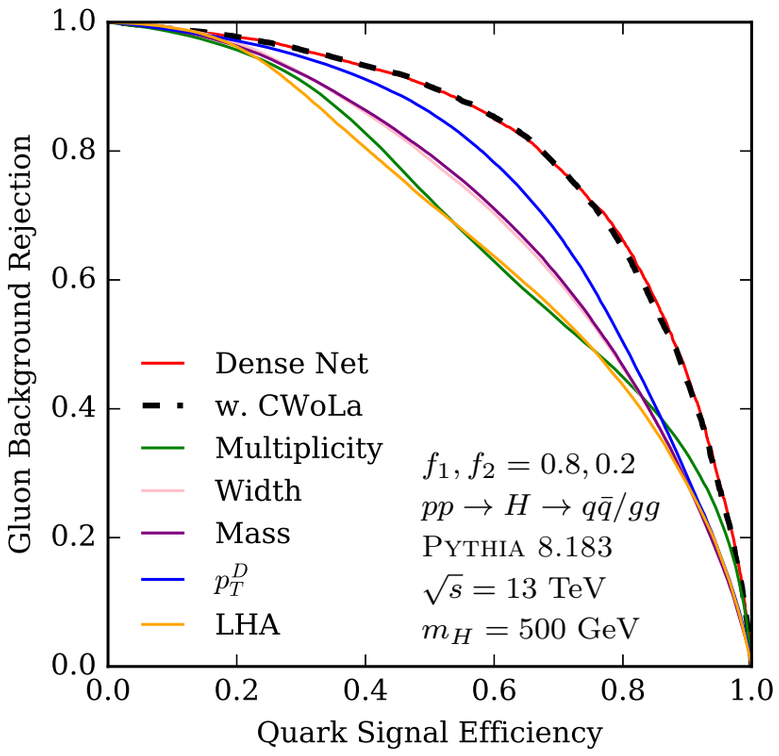
New Proposal #2: Class. w/o labels (CWoLa)

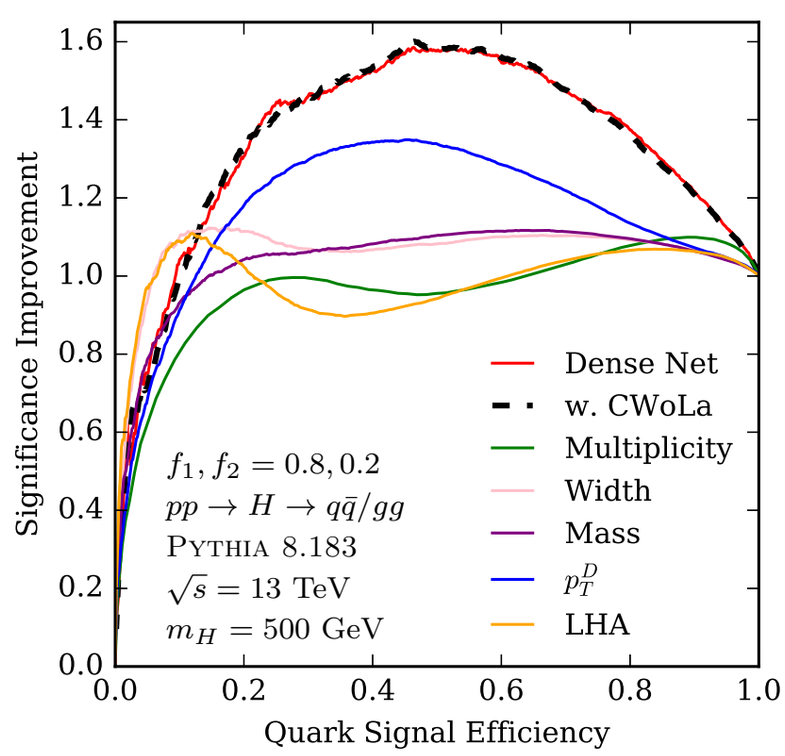


New Proposal #2: Class. w/o labels (CWoLa)









Future outlook

We have developed two methods for training classifiers directly on (unlabeled) data.

LLP: non-trivial to train, applies to any number of samples, small dependence on knowledge of fractions.

CWoLa: easy to train, only works for 2 samples, no dependence on fractions, can’t directly compute ROC

The performance has been demonstrated in ‘simple’ cases - we are now working on applying these techniques in more complex scenarios.

Future outlook

L. Dery, P. Komiske, E. Metodiev, **BPN**, F. Rubbo, M. Schwartz