Weakly supervised classifiers

learning from data and proportions

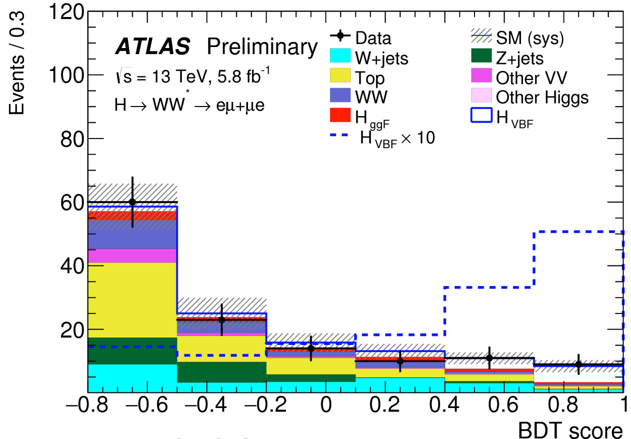
L. Dery (Stanford), B. Nachman (LBNL), F. Rubbo (SLAC), A. Schwartzman (SLAC)

U.S. DEPARTMENT OF ENERGY
Office of Science

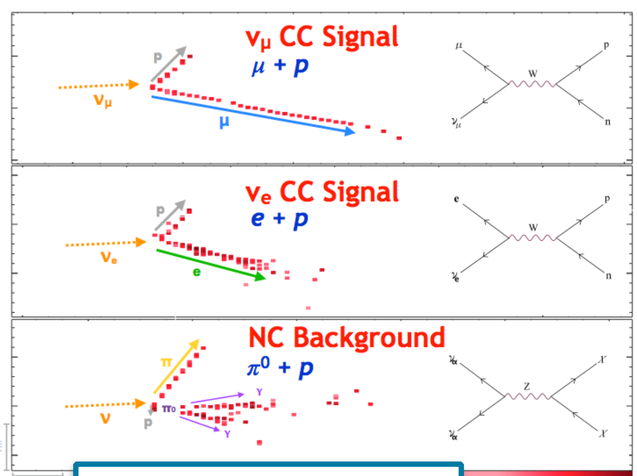
SLAC NATIONAL ACCELERATOR LABORATORY

Classification in HEP

Discriminating signal events from backgrounds

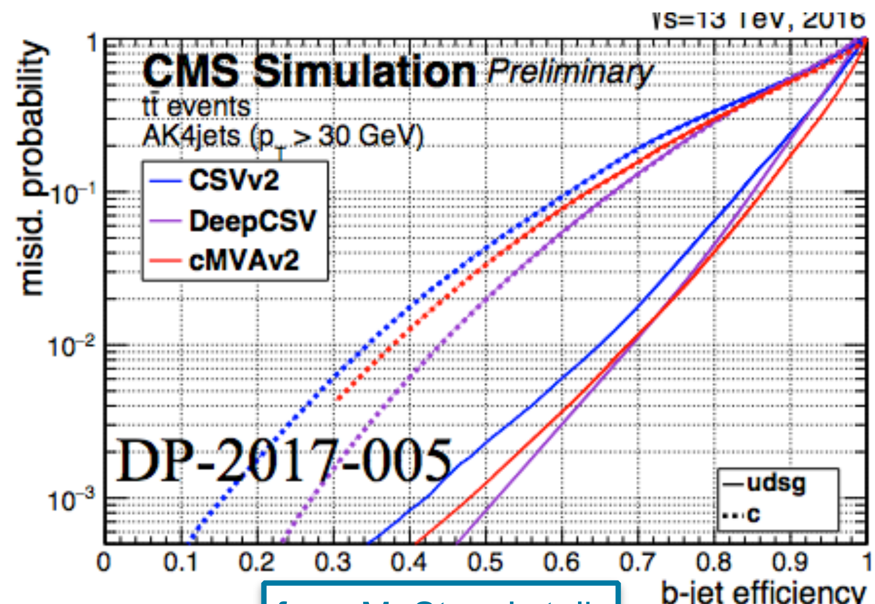


ATLAS-CONF-2016-112



from A. Radovic's talk

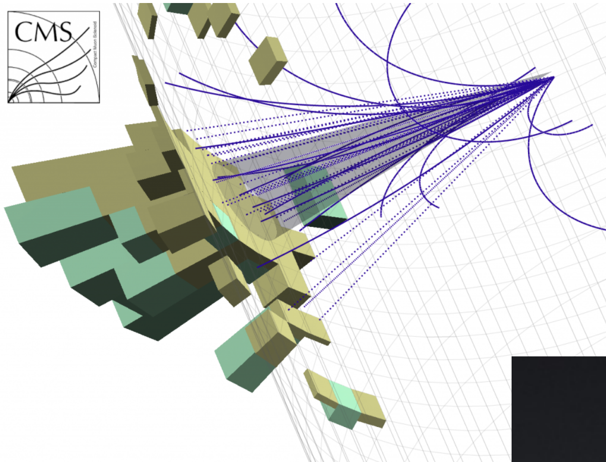
Classifying reconstructed objects

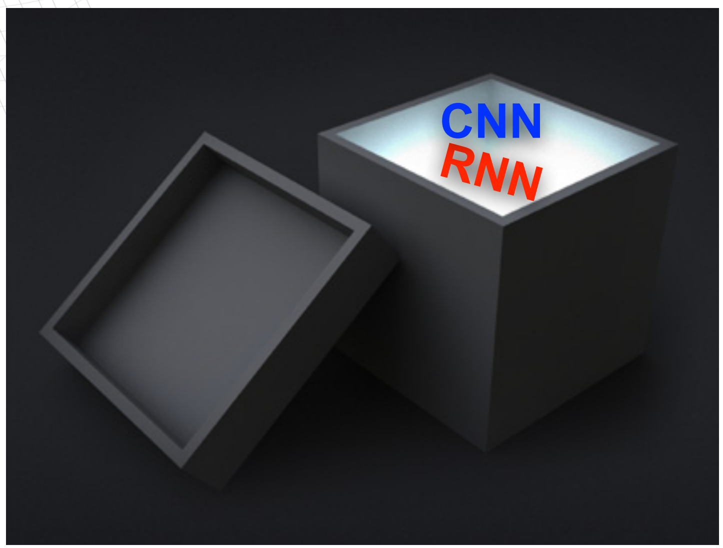


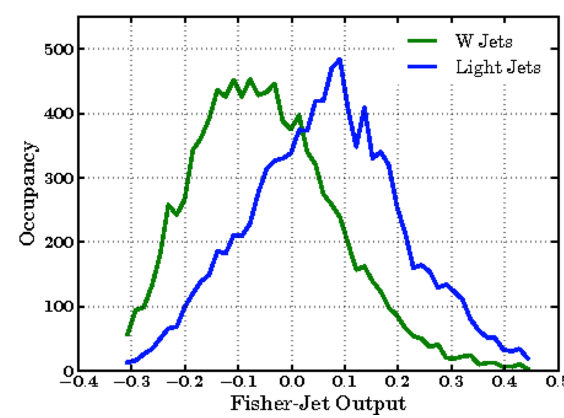
from M. Stoye's talk

& more…

Jet classification example

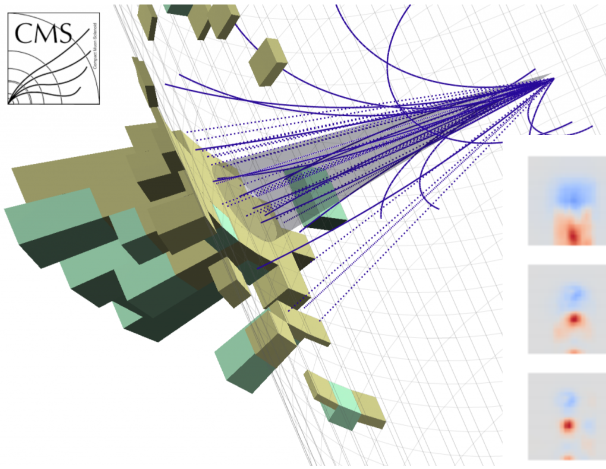


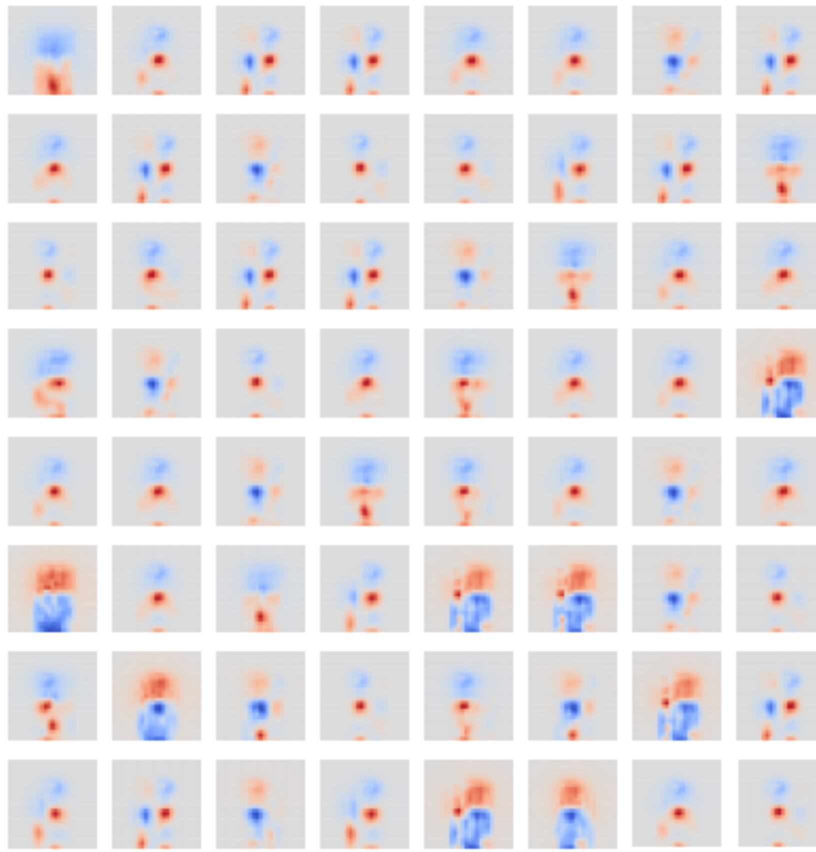


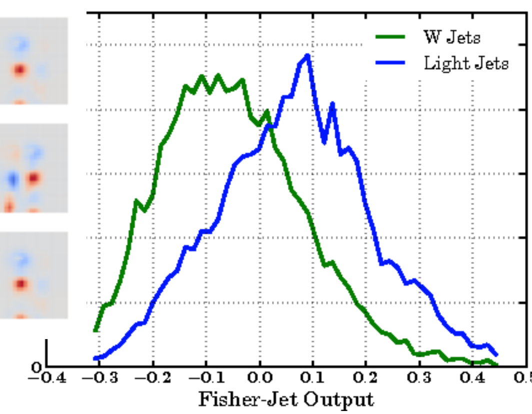


Jet classification example

1511.05190

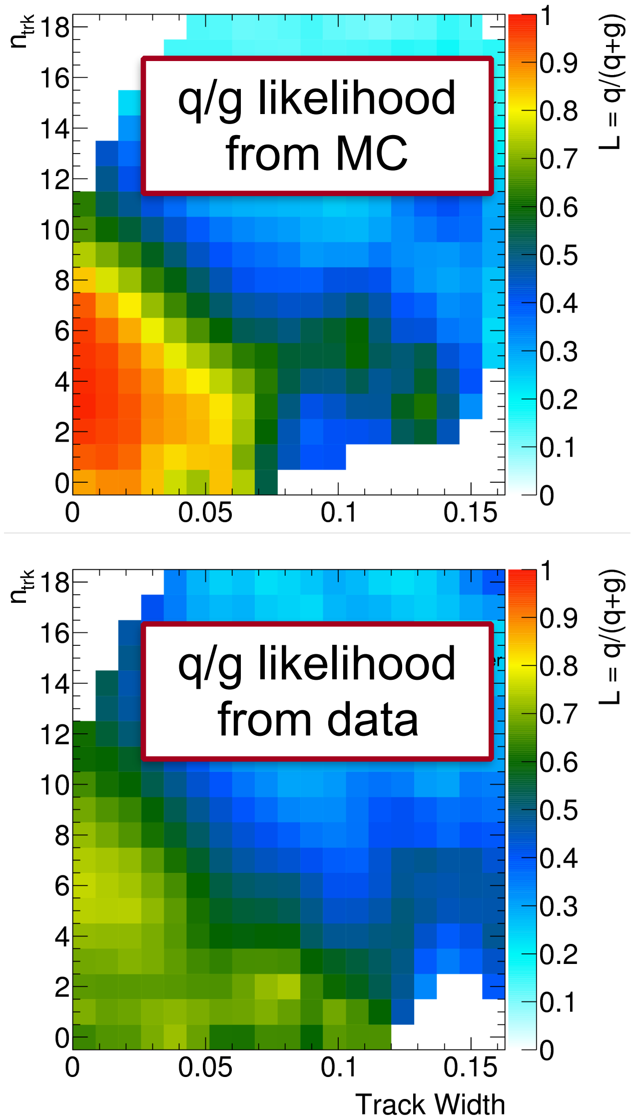




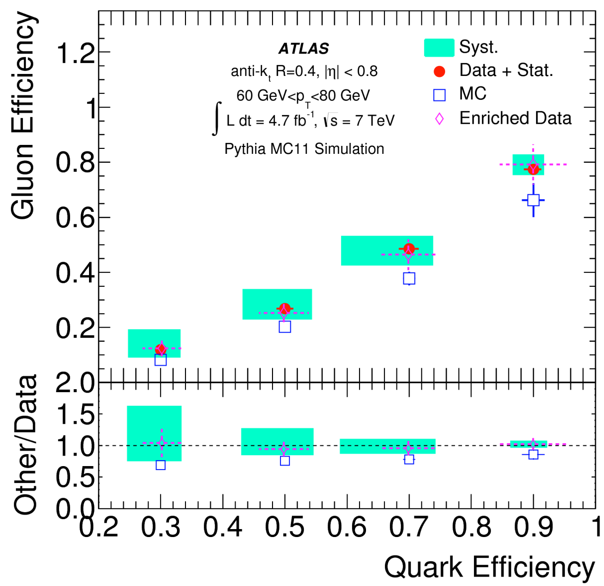


*or other high-dimensional representations (embedding,M-body, etc..)

Learning from simulation vs learning from data



- Modeling of multi-dimensional soft QCD features (e.g. n_{track},w_{track}) is challenging for MC.



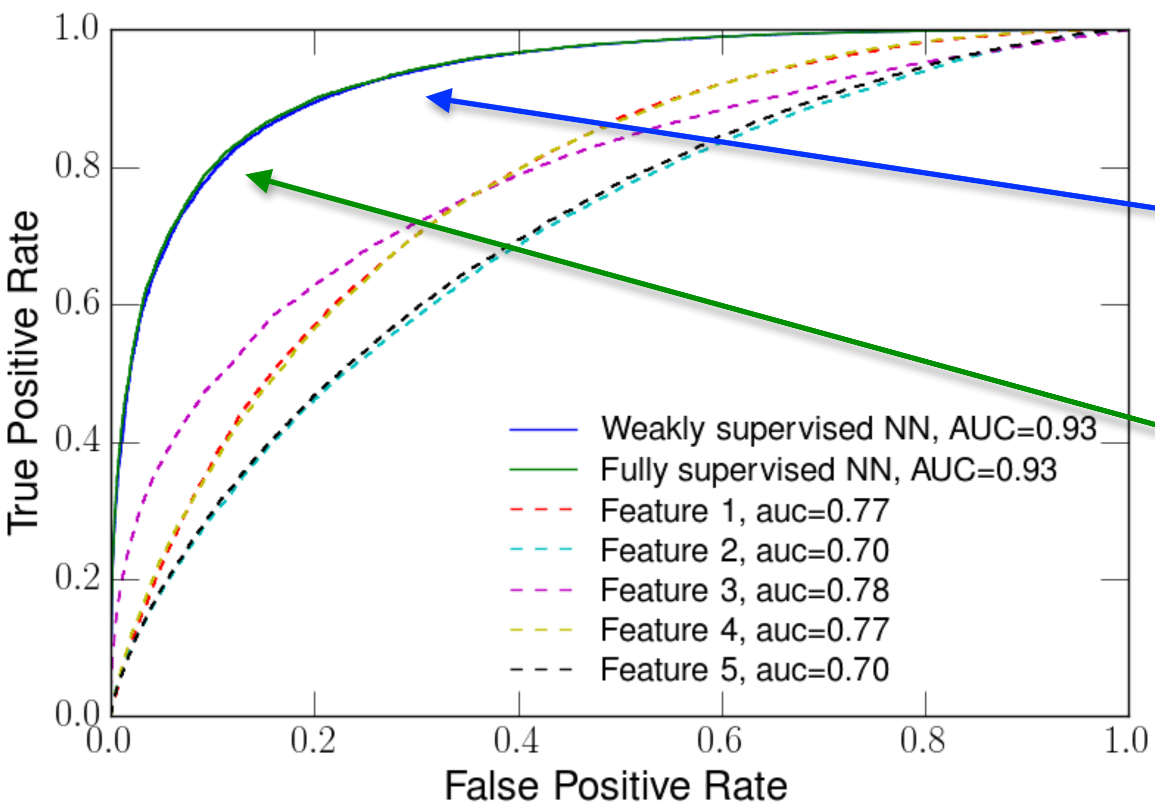
1405.6583

- Expect further strain at higher dimensionality (e.g. images with thousands of pixels!)

Training on data

- Classifier is always suboptimal if distribution of training and test samples are different.
- Data is the perfect event “simulation”: exactly the same distribution as in the test sample.
- N.B.: doesn’t impact uncertainties, only the “central value” of the performance (i.e. how optimal is the discrimination in data)!
- N.B.2: for many applications simulation is very good and its distribution is close to data.

Learn directly from unlabeled data!



Weakly supervised classifier trained without using labels

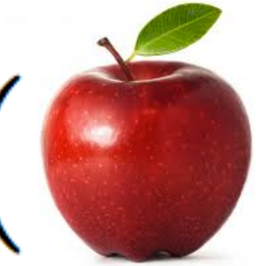
Traditional fully supervised classifier

Traditional full supervision

Labeled training set ("simulation")

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

instance label:
0:pear 1:apple

Classification $f_{\text{full}}\Bigg($  $\Bigg)$ = 0.97

Weak supervision

unlabeled training data

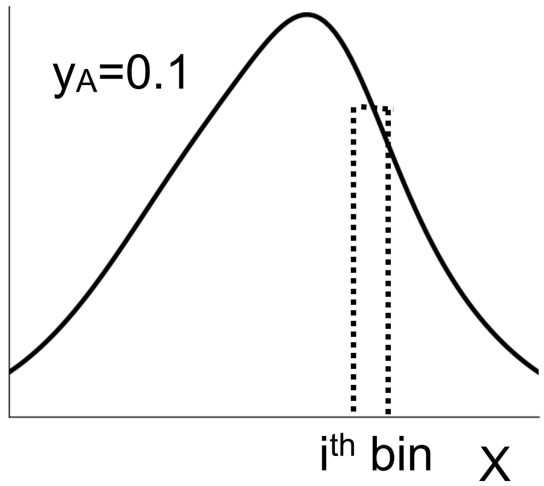
$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell \left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y \right)$$

average composition for each barrel

Classification $f_{\text{weak}}\Bigg($  $\Bigg)$ = 0.97

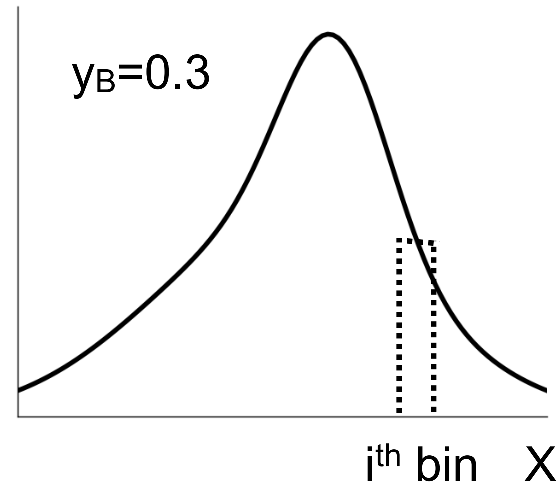
Weak supervision - analytically

unlabeled data sample A



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

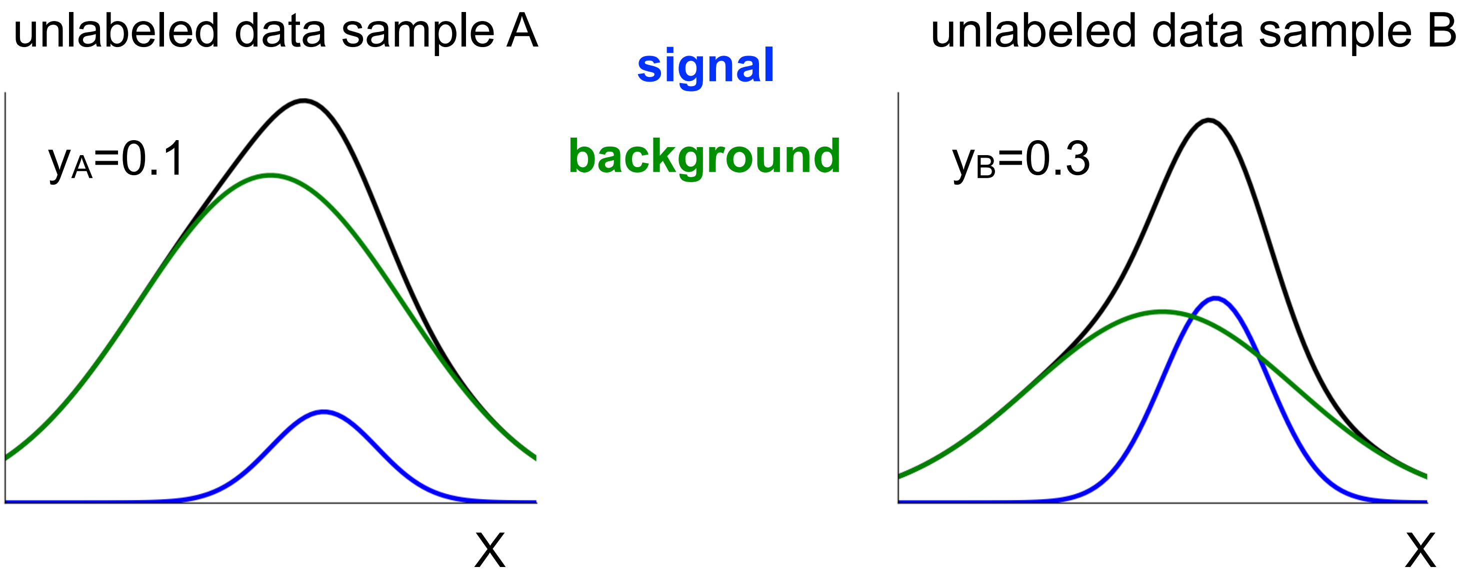
unlabeled data sample B



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

- Given two independent unlabeled data samples, and the corresponding proportion of signal, we can extract the signal and background distributions.

Weak supervision - analytically



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

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

- Given two independent unlabeled data samples, and the corresponding proportion of signal, we can extract the signal and background distributions.

—> build Likelihood Ratio discriminant.

Weak supervision

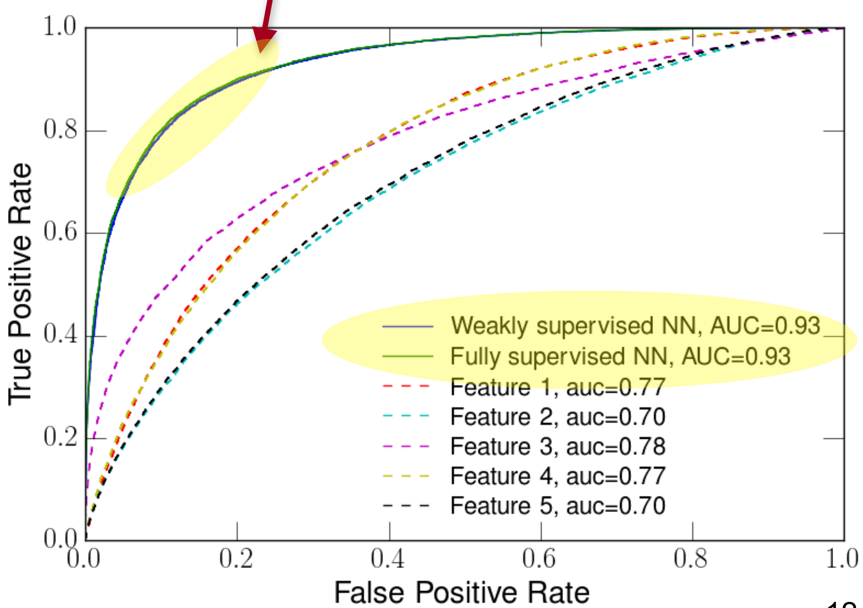
- The analytic approach requires binning and becomes quickly unmanageable as the feature space grows.

- ML approach directly looks for discriminant, without extracting explicitly n-dimensional feature distributions for S and B.

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

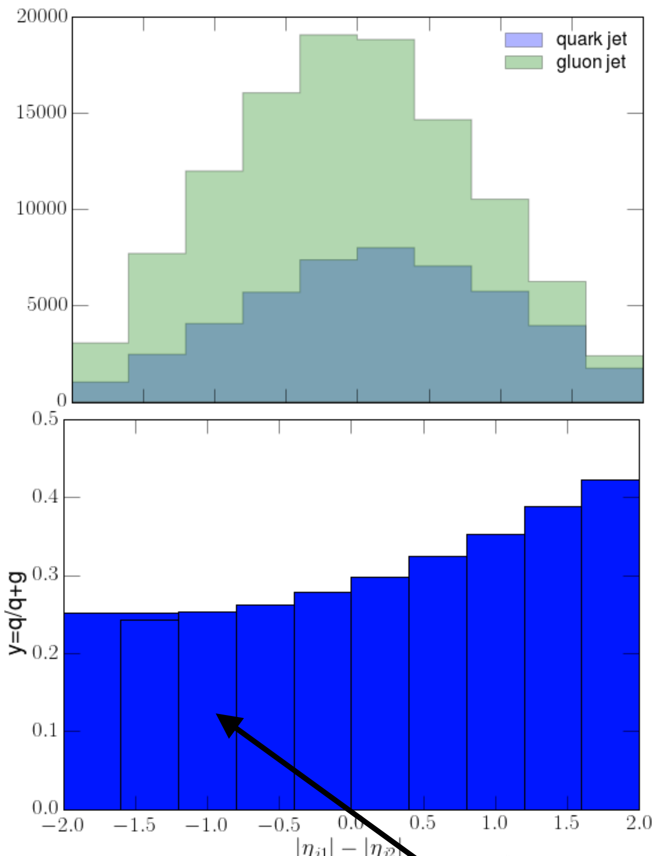
$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

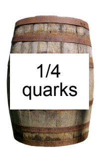
Same discrimination as fully supervised



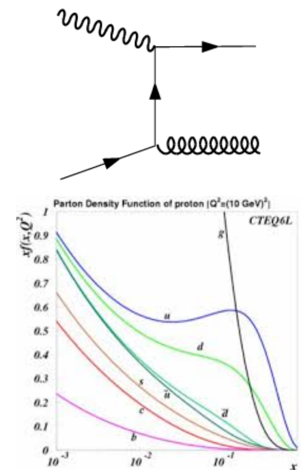
Weak supervision - q/g tagging

|η[j1]| - |η[j2]| in dijet events

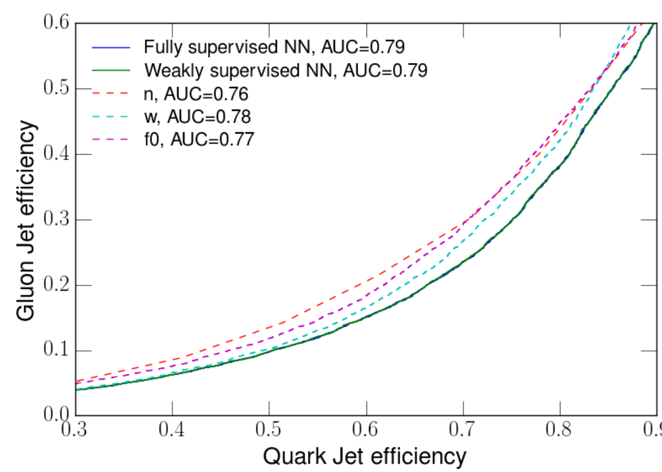




Each bin is a “barrel” of jets with known proportion

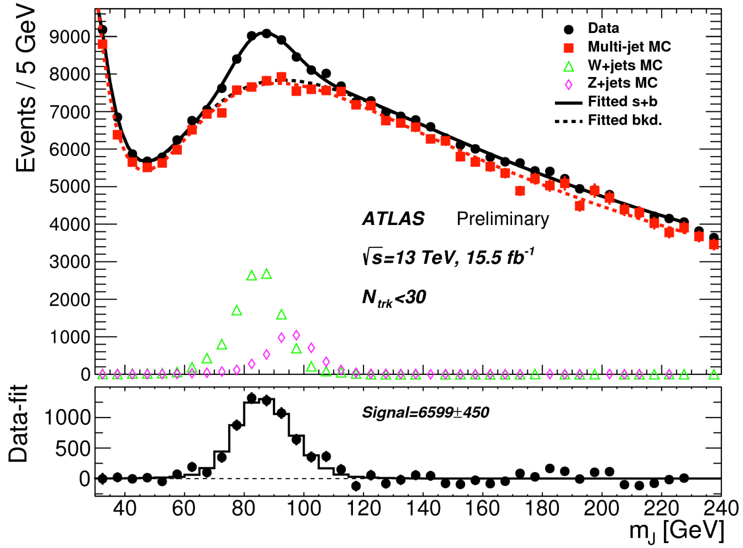


Leverage precise description of ME and PDF (MC/theory) to extract discrimination from soft QCD features (from data!)

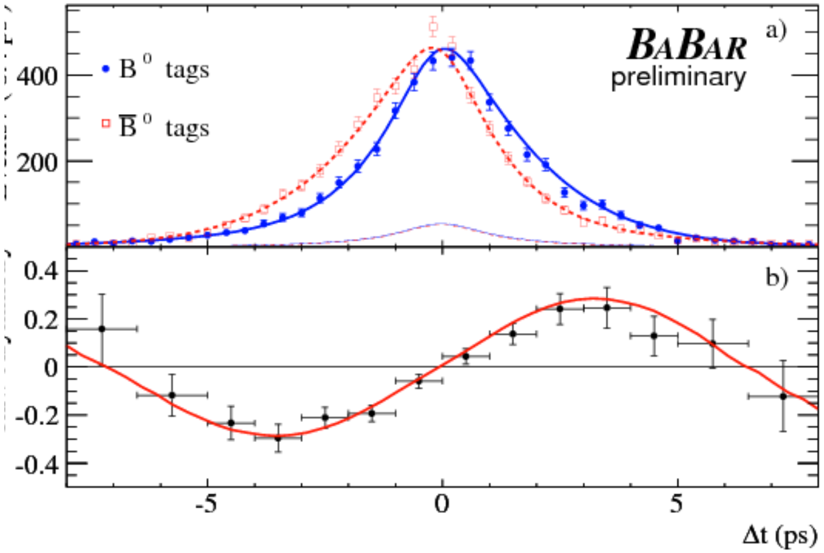


Summary

- **Weak supervision** is a new paradigm leveraging the **class proportions** in high-level observables in order to use **unlabeled data** to extract **discriminating information** from poorly modeled or unknown **low-level observables**.
- Multiple potential applications in HEP



ATLAS-CONF-2016-055/

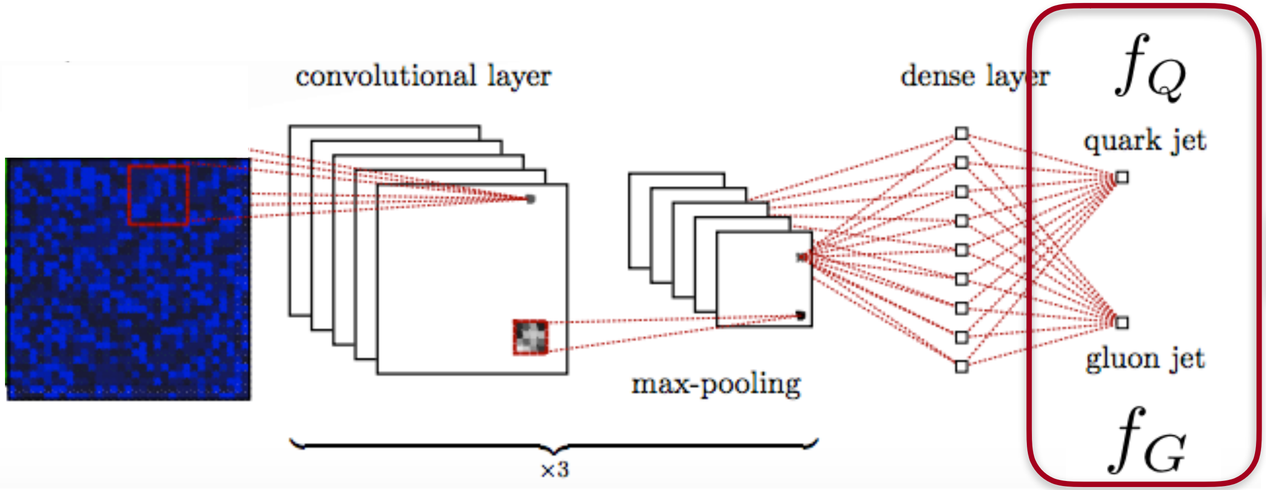


SLAC-PUB-13402

Next step: scaling to higher dimensionality

Quark/gluon jet tagging with jet images (grayscale) and CNN

Fully supervised network:



Komiske *et al.* 1612.01551

softmax activation

$$f_Q = 1 - f_G$$

$$Q : t = 1 \quad G : t = 0$$

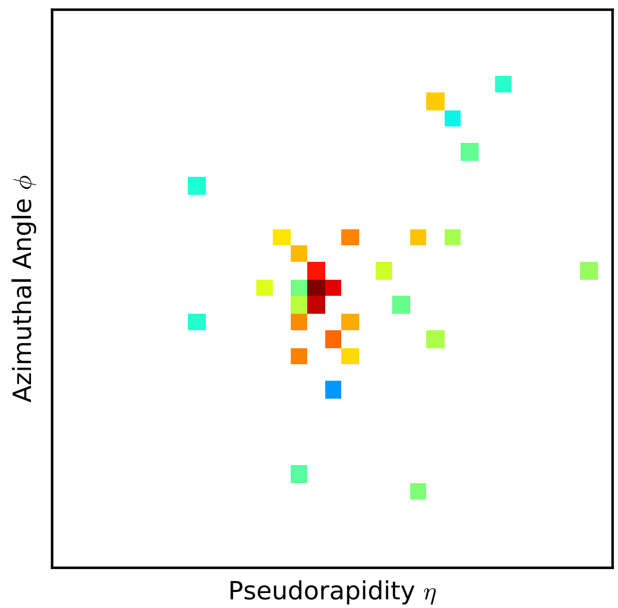
$$\boxed{-t \cdot \log f_Q - (1-t) \cdot \log f_G}$$

cross-entropy loss

First look at weak supervision on same architecture in “ideal” conditions:
50 samples with proportions in **[0,1]** (regularly spaced)

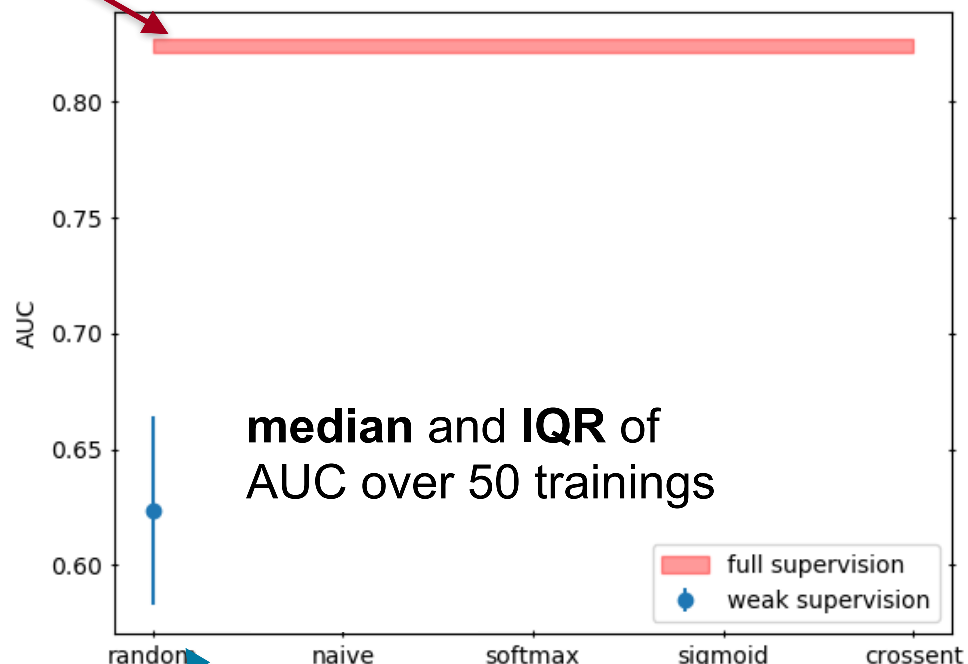
[Work in progress with Dery, Komiske, Metodiev, Nachman, Schwartz]

Jet image + weak supervision



33x33=1089 input features

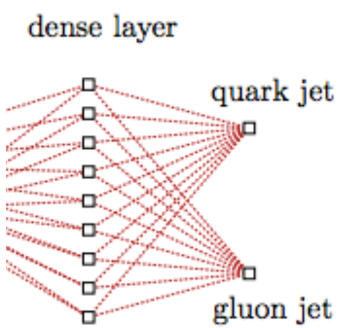
fully supervised CNN



randomly initialized CNN
(untrained)

[Work in progress with Dery, Komiske, Metodiev, Nachman, Schwartz]

Jet image + weak supervision

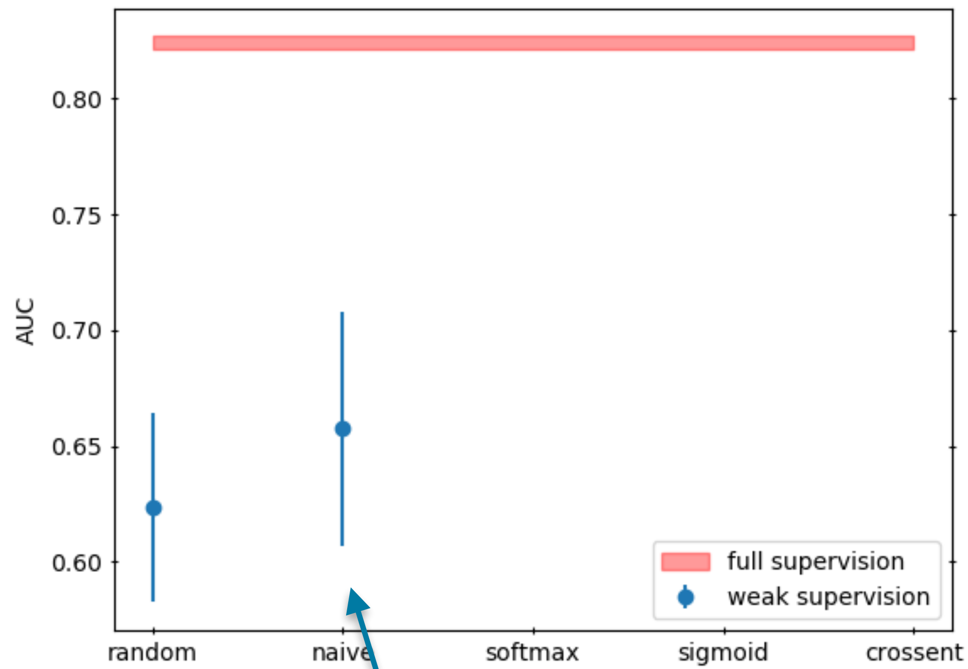


$$f_Q, y_Q = \frac{N_Q}{N_Q + N_G}$$

$$f_G, y_G = 1 - y_Q$$

$$L = \left(\sum \frac{f_Q}{N} - y_Q\right)^2$$

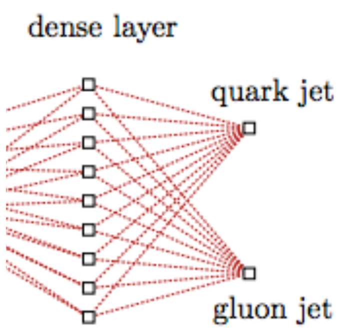
less constraint for “gluon” weights (asymmetric gradient)



naive squared loss

[Work in progress with Dery, Komiske, Metodiev, Nachman, Schwartz]

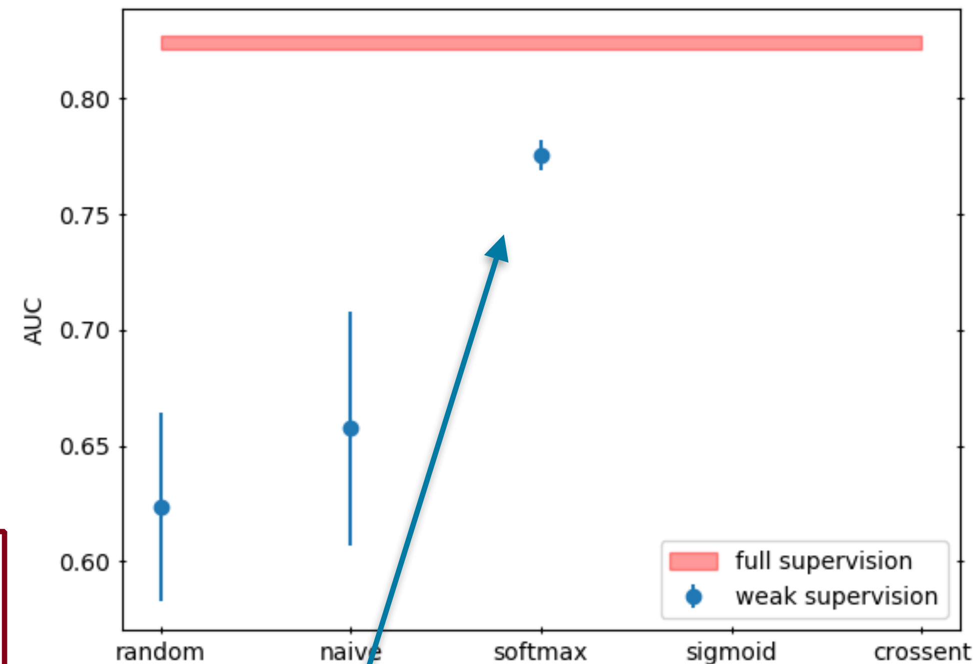
Jet image + weak supervision



$$f_Q, y_Q = \frac{N_Q}{N_Q + N_G}$$

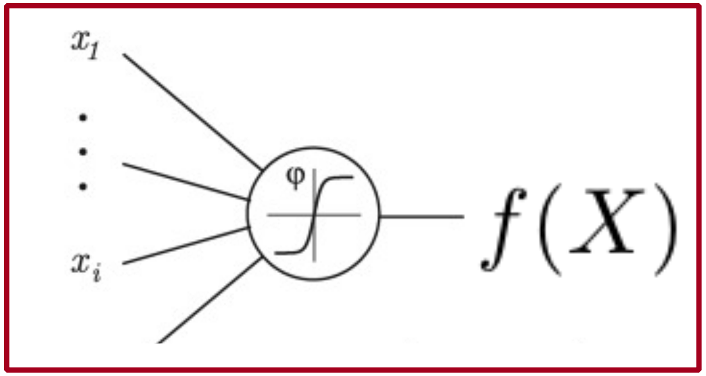
$$f_G, y_G = 1 - y_Q$$

$$L = \left(\sum \frac{f_Q}{N} - y_Q\right)^2 \boxed{+ \left(\sum \frac{f_G}{N} - y_G\right)^2}$$

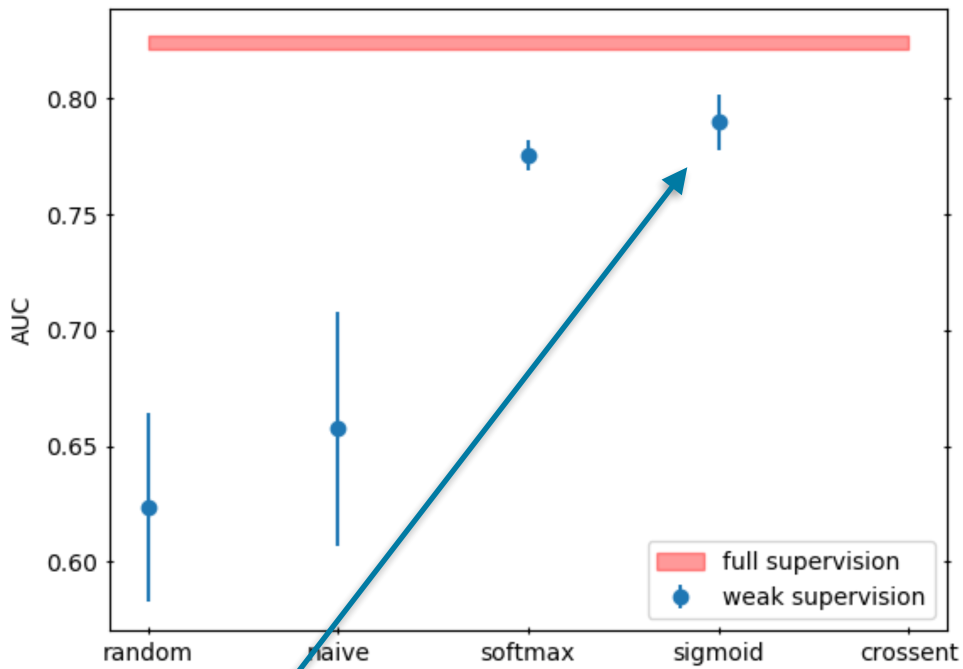


symmetric squared loss with softmax activation

Jet image + weak supervision

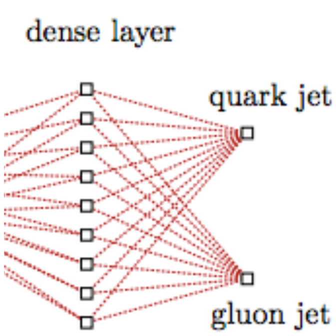


$$L = \left(\sum \frac{f}{N} - y_Q \right)^2$$



squared loss with sigmoid activation

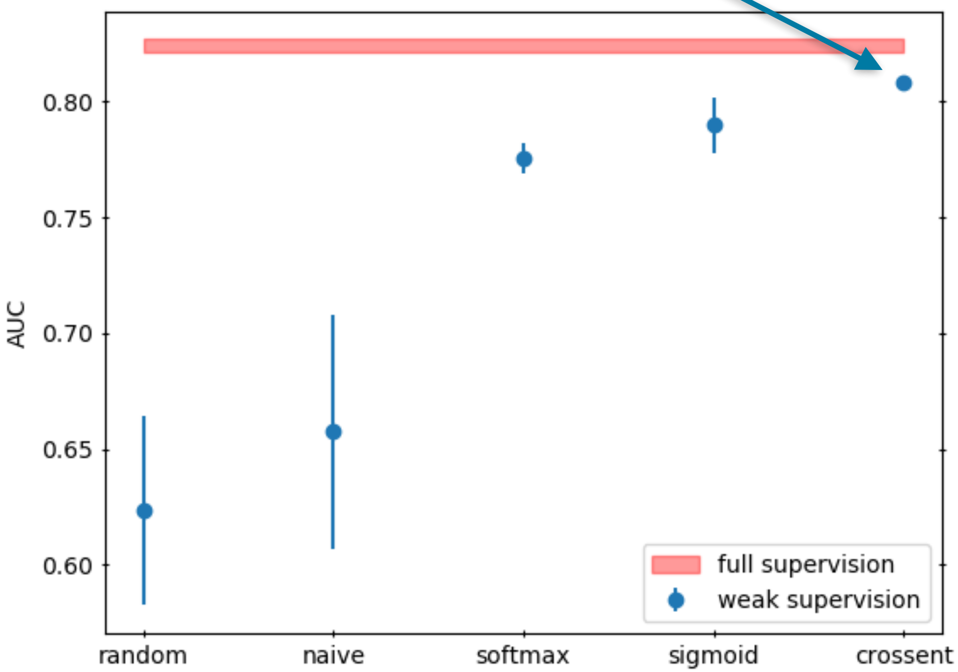
Jet image + weak supervision



$$f_Q, y_Q = \frac{N_Q}{N_Q + N_G}$$

$$f_G, y_G = 1 - y_Q$$

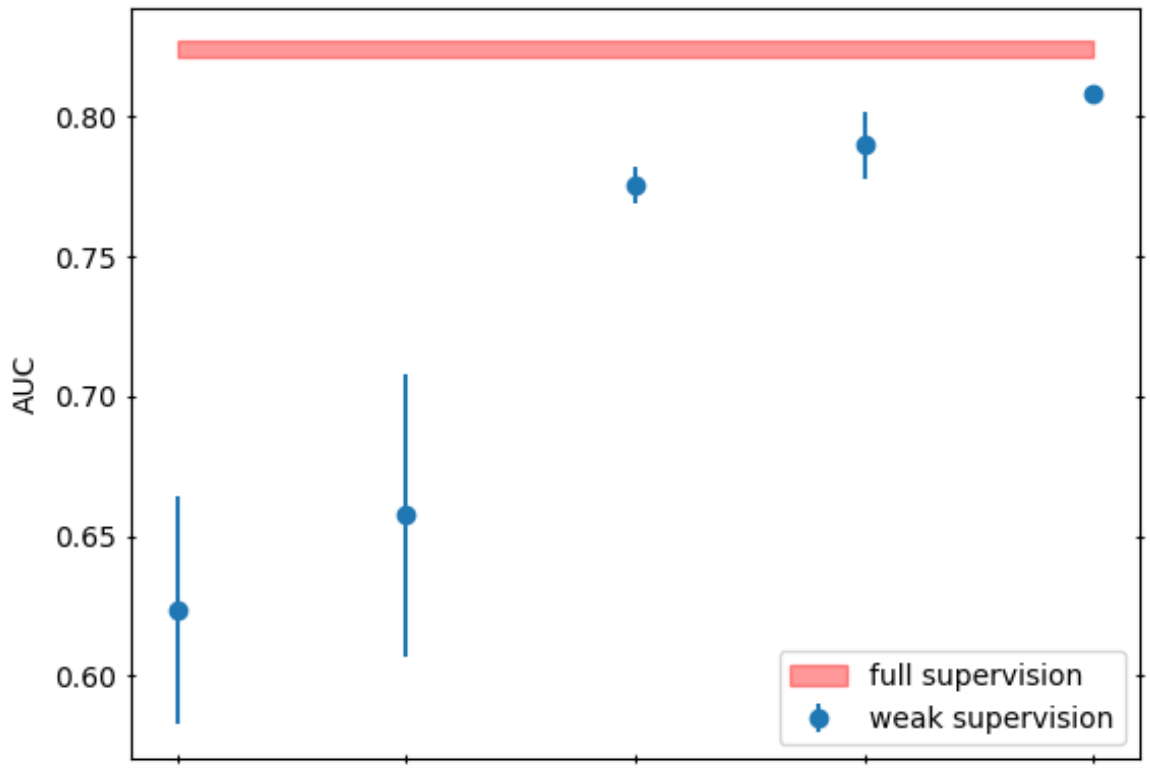
"weak cross-entropy" with softmax activation



$$L = -y_Q \log\left(\sum \frac{f_Q}{N} / y_Q\right) - y_G \log\left(\sum \frac{f_G}{N} / y_G\right)$$

[Work in progress with Dery, Komiske, Metodiev, Nachman, Schwartz]

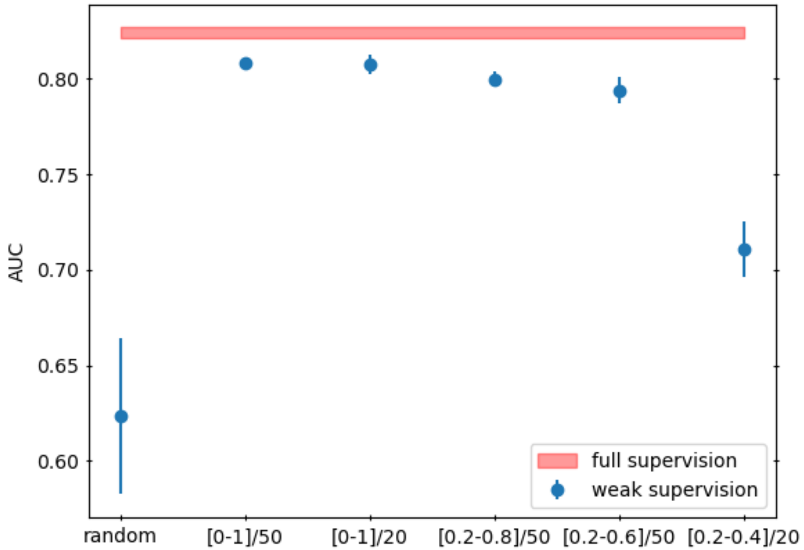
Jet image + weak supervision



Loss	N/A	square	square symmetric	square symmetric	cross-entropy
Activation	N/A	softmax	softmax	sigmoid	softmax

Conclusions and next steps

- First implementation of **weak supervision+CNN** shows promising results for jet image classification with **unlabeled training data**.
- Careful choice for activation and loss function provide important handles to close gap wrt full supervision performance.
- Plan to investigate impact of size and structure of training data
- Architecture choices possibly play a role (e.g. “wider” networks)



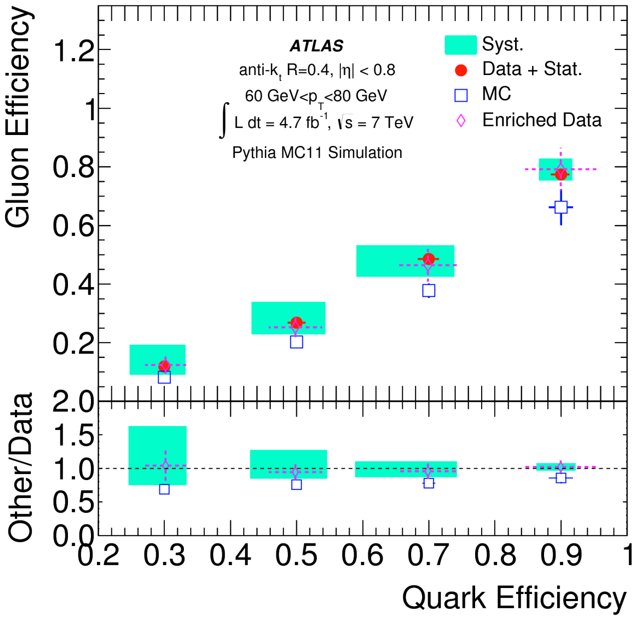
References

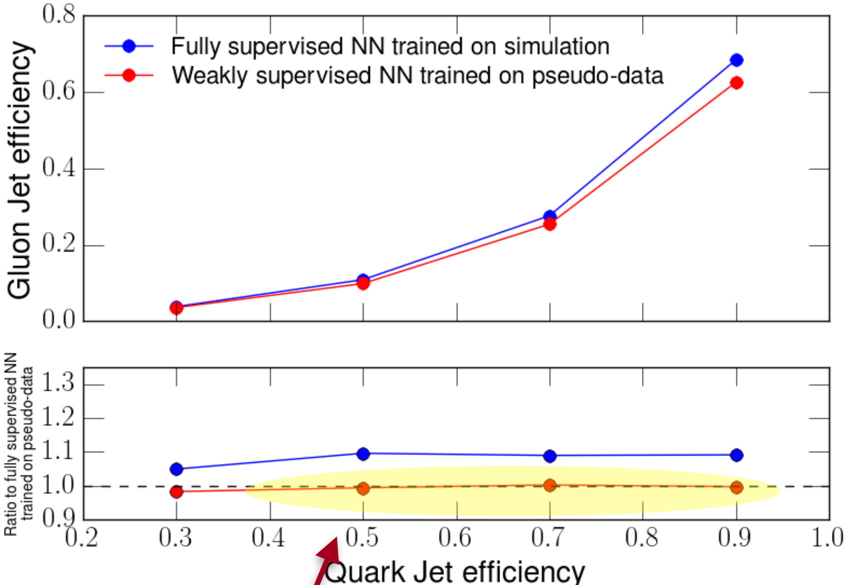
- Jet-Images: Computer Vision Inspired Techniques for Jet Tagging - https://arxiv.org/abs/1407.5675
- Jet-Images — Deep Learning Edition - https://arxiv.org/abs/1511.05190
- Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s}=7$ TeV with the ATLAS detector - https://arxiv.org/abs/1405.6583
- Weakly Supervised Classification in High Energy Physics - https://arxiv.org/abs/1702.00414

Backup

Weak supervision

- Weak supervision allows training directly on data
- Learns only <u>real</u> features, from being exposed to discriminant features in data.

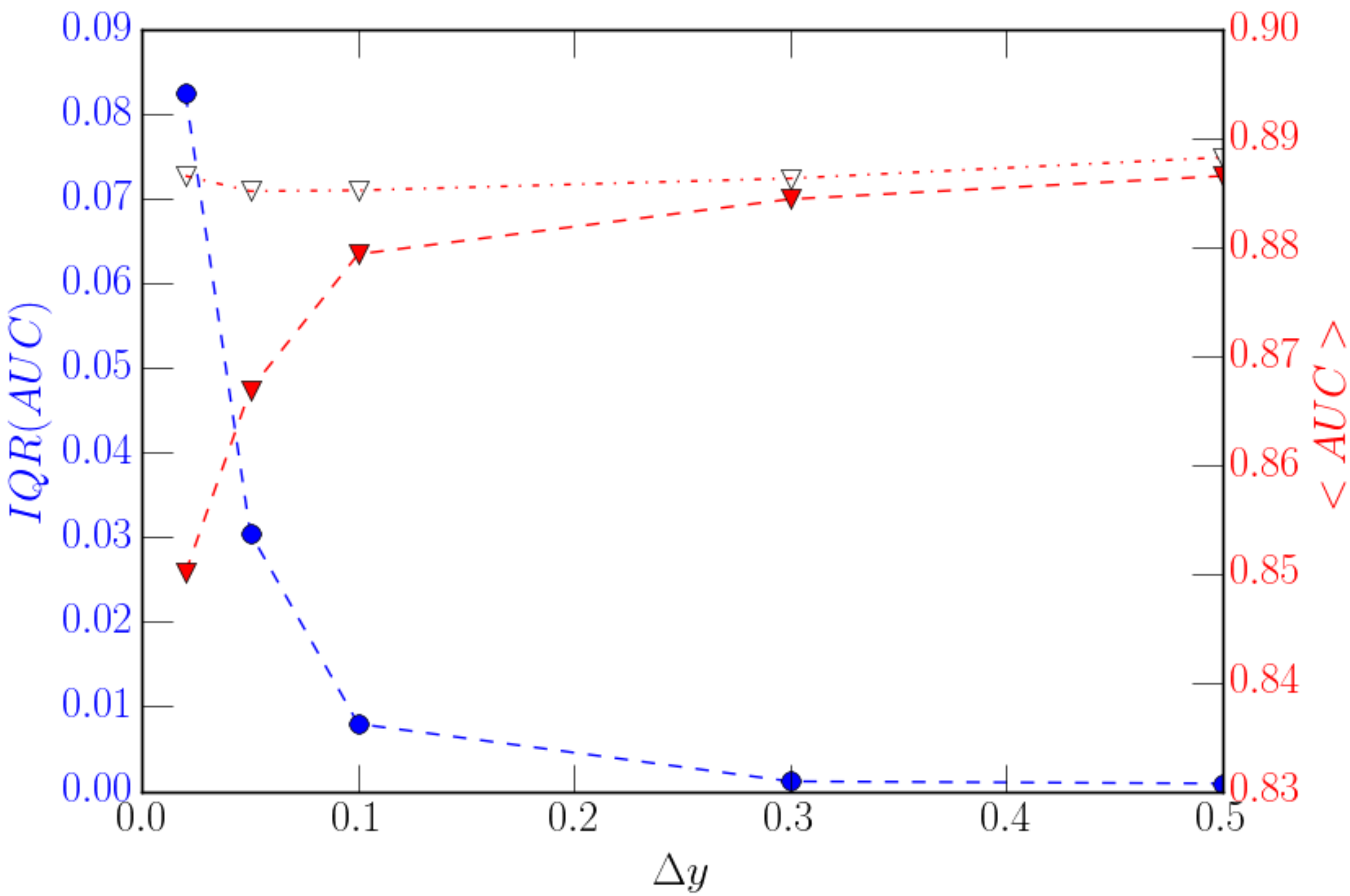




1405.6583

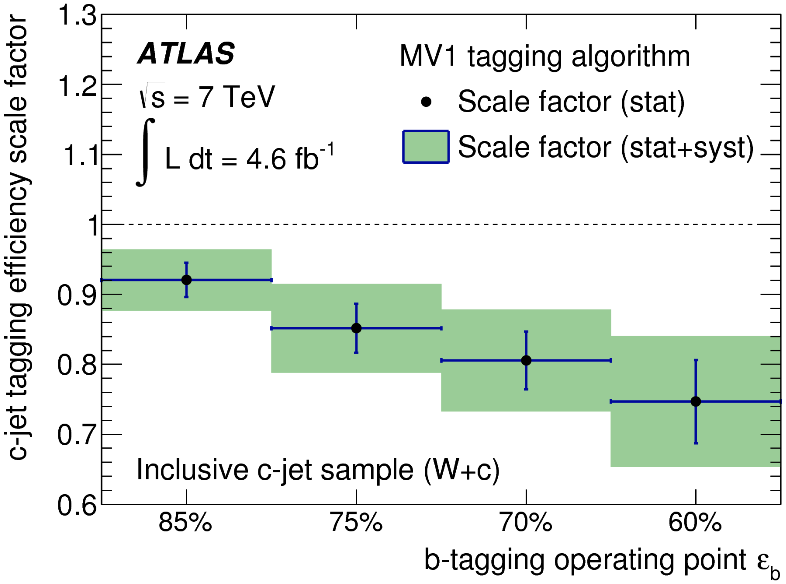
Same performance as ideal classifier, trained on labeled data

Stability



0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
$IQR(AUC)$
0.90
0.89
0.88
0.87
0.86
0.85
0.84
0.83
$< AUC >$
0.0
0.1
0.2
0.3
0.4
0.5
Δy

data-MC SFs



2016 JINST 11 P04008

Cumulative data-simulation scale factor - CMS Tagger, CMS Combined Tagger

$\lvert\eta\rvert < 1.0$			
Selection	**MadGraph**	**POWHEG**	**MC@NLO**
CMS Tagger WP0	0.985 ± 0.073	1.173 ± 0.092	1.033 ± 0.081
CMS Combined Tagger WP3	0.891 ± 0.118	1.063 ± 0.146	0.933 ± 0.129

$1.0 < \lvert\eta\rvert < 2.4$			
Selection	**MadGraph**	**POWHEG**	**MC@NLO**
CMS Tagger WP0	0.644 ± 0.100	0.704 ± 0.110	0.768 ± 0.118
CMS Combined Tagger WP3	0.685 ± 0.199	0.906 ± 0.277	0.802 ± 0.230

CMS-PAS-JME-13-007

[Work in progress with Dery, Komiske, Metodiev, Nachman, Schwartz]

Jet image + weak supervision