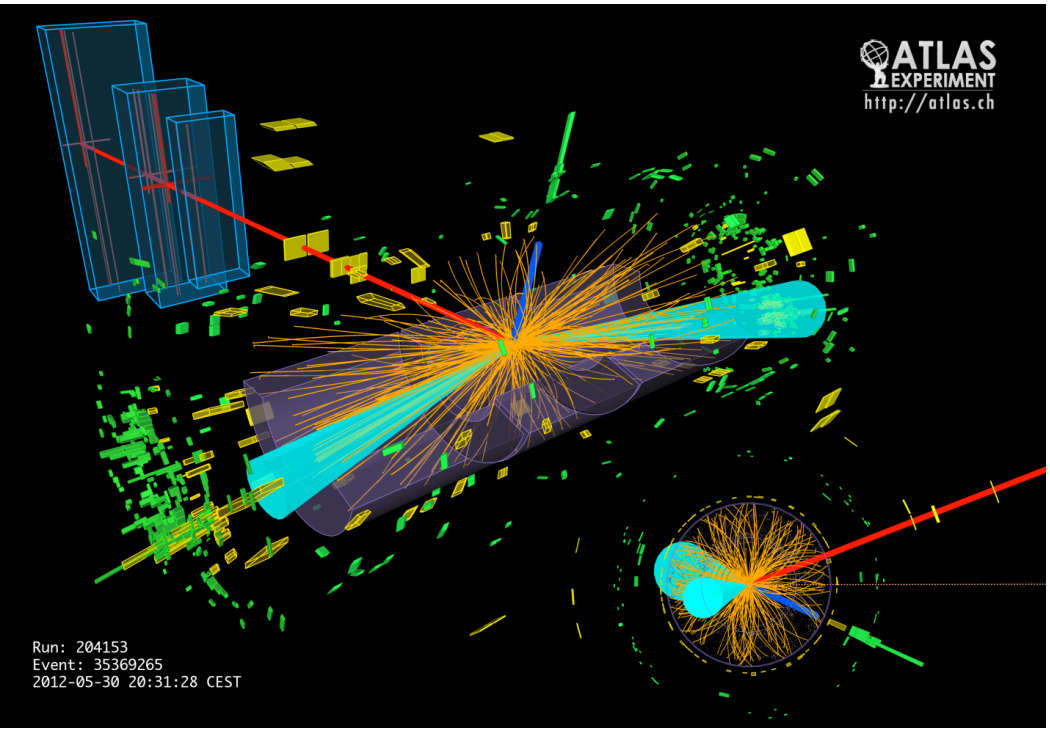
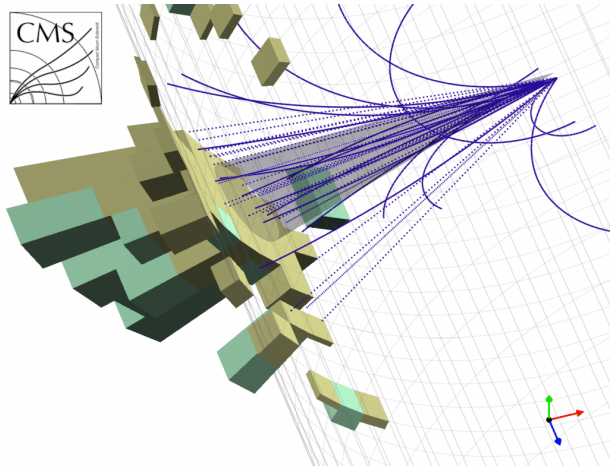
Weakly supervised classifiers
learning from data and proportions

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

LHC detectors as cameras

The LHC experiments are O(100) Megapixel 3D fast cameras
—> High resolution “pictures” of proton-proton collisions.

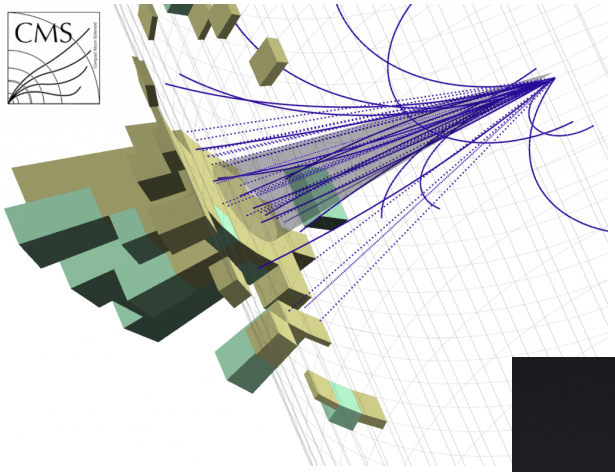


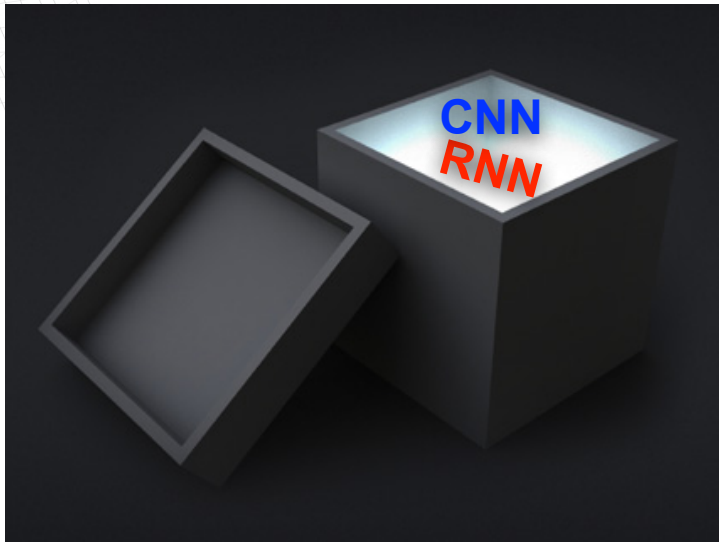


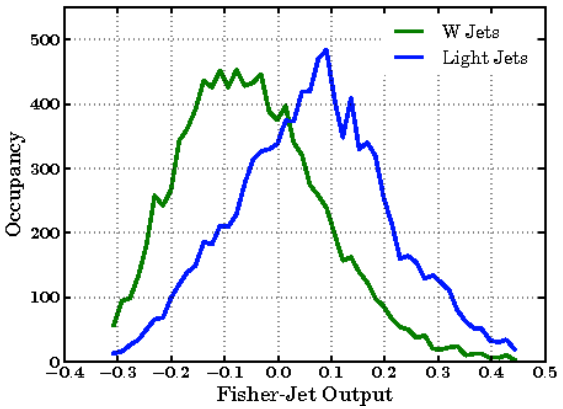
E.g. Jets are recorded as densely packed tracks and calorimeter “images”.

Broad effort aiming at outperforming Physics-motivated feature extraction by using low-level inputs (e.g. calorimeter “pixels”) for ML algorithms.

Jet classification example

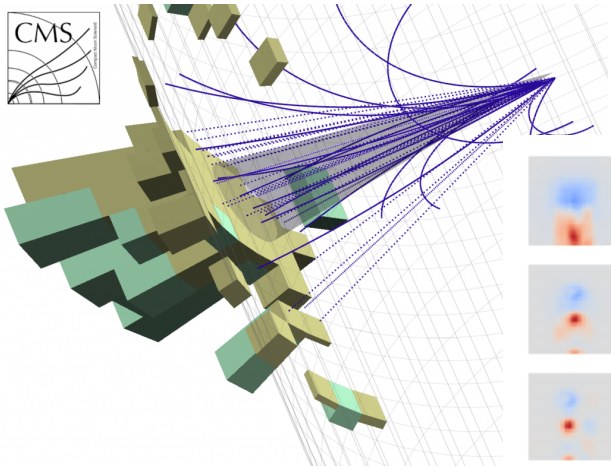




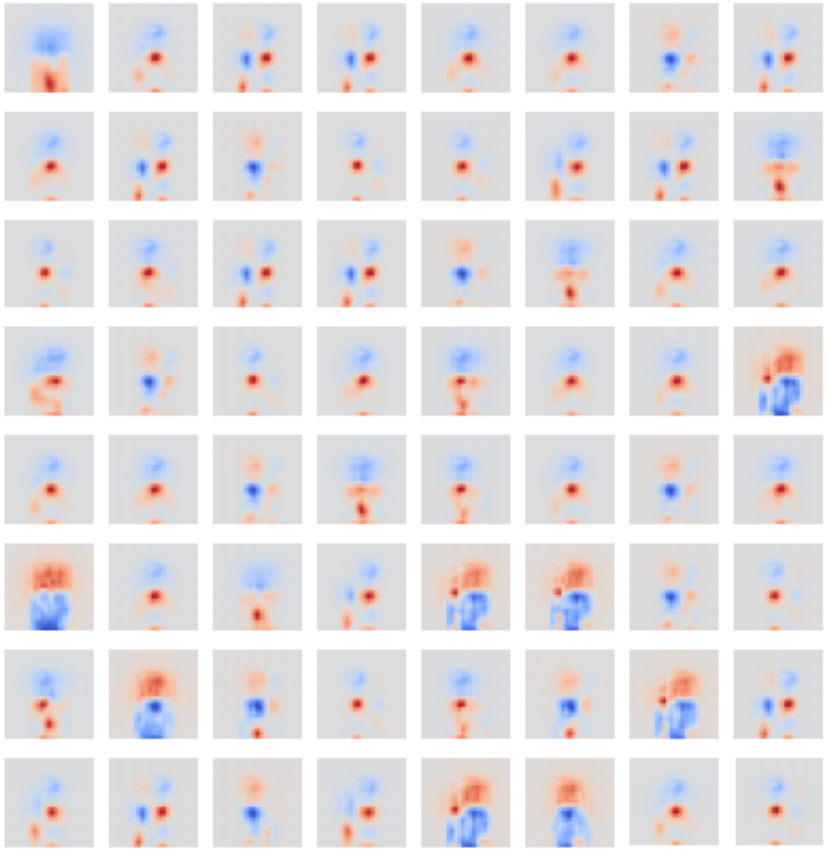


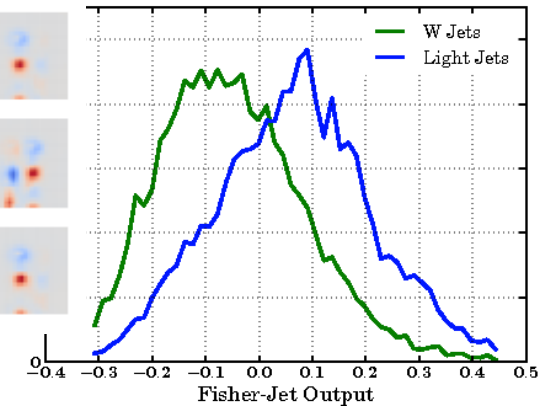
Jet classification example

SLAC

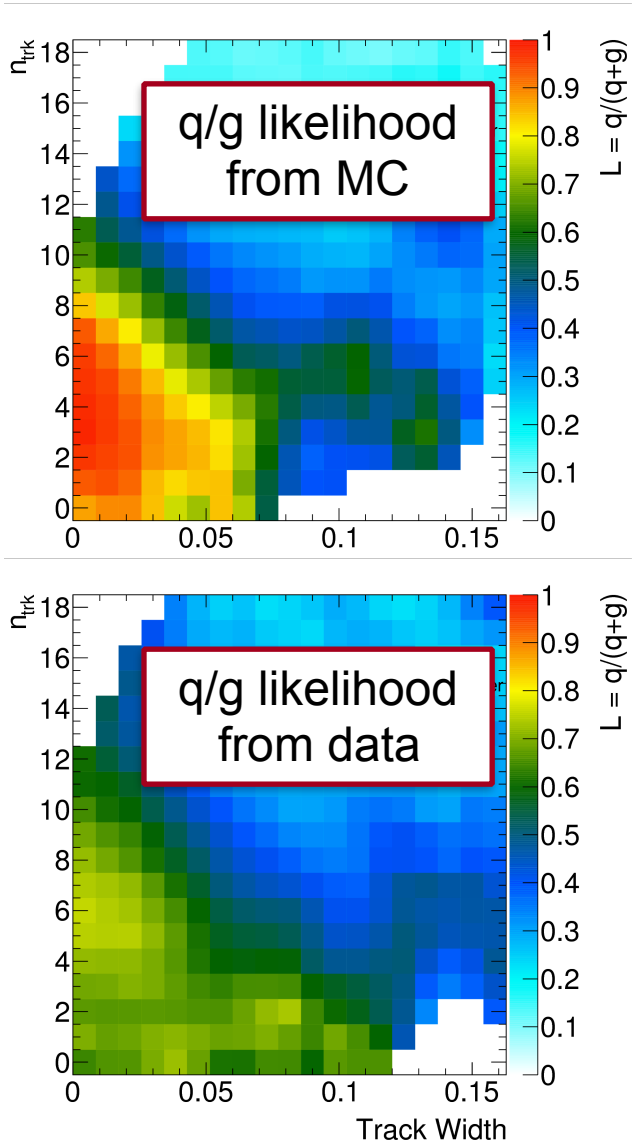


1511.05190

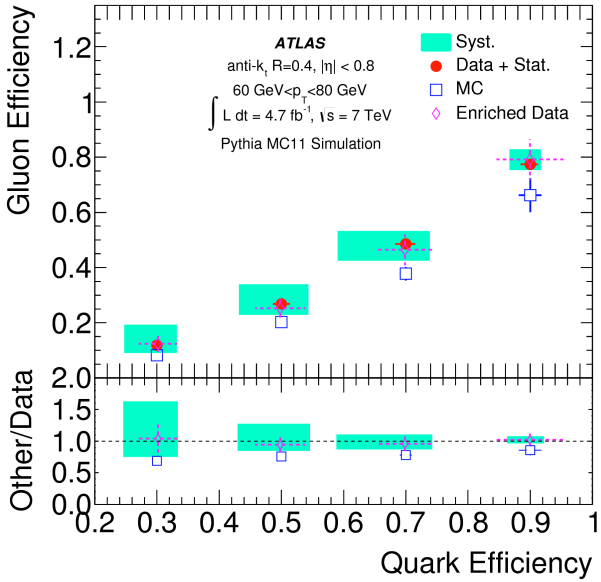




Learning from simulation vs learning from data



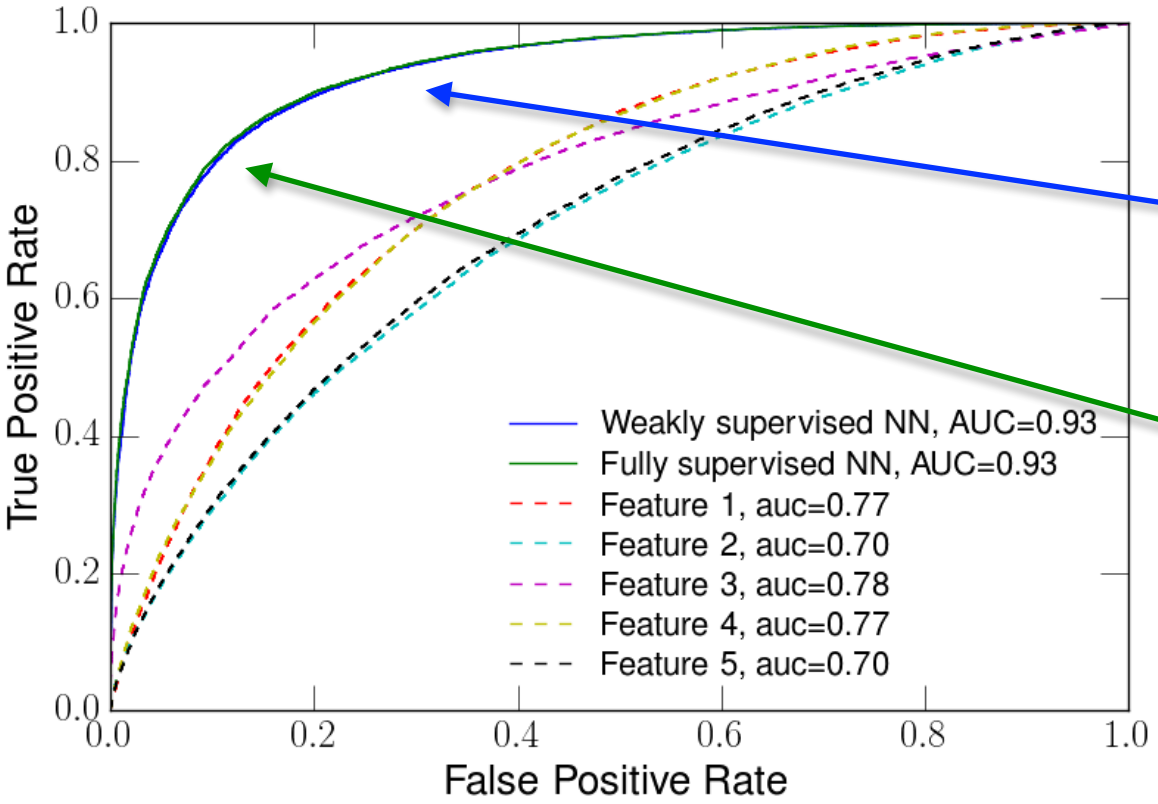
- Modeling of multi-dimensional soft QCD features (e.g. n_{track},w_{track}) is challenging for MC.



1405.6583

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

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_{full} () = 0.97

Weak supervision

unlabeled training data

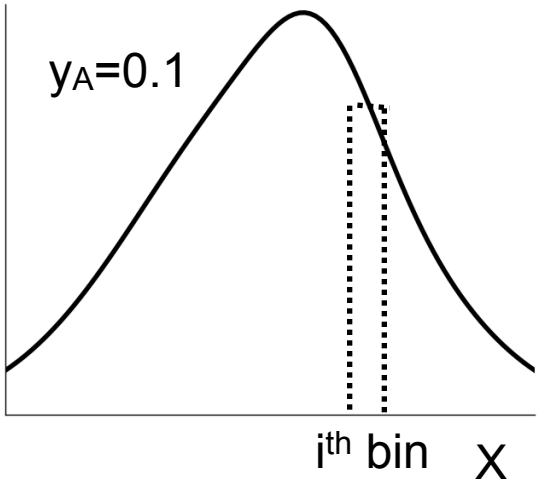
$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell \left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y \right)$$

average composition for each barrel

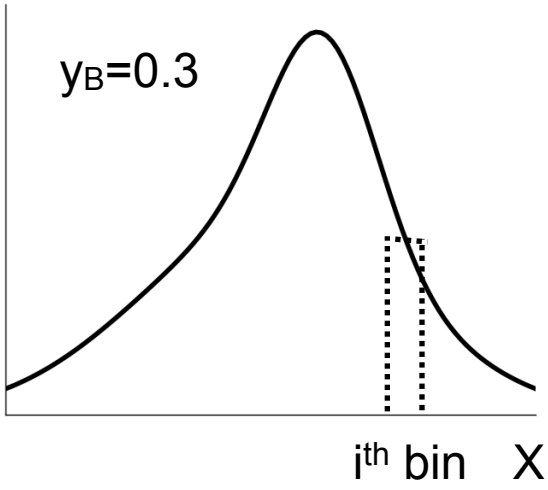
Classification f_{weak}  = 0.97

Weak supervision - analytically

unlabeled data sample A



unlabeled data sample B

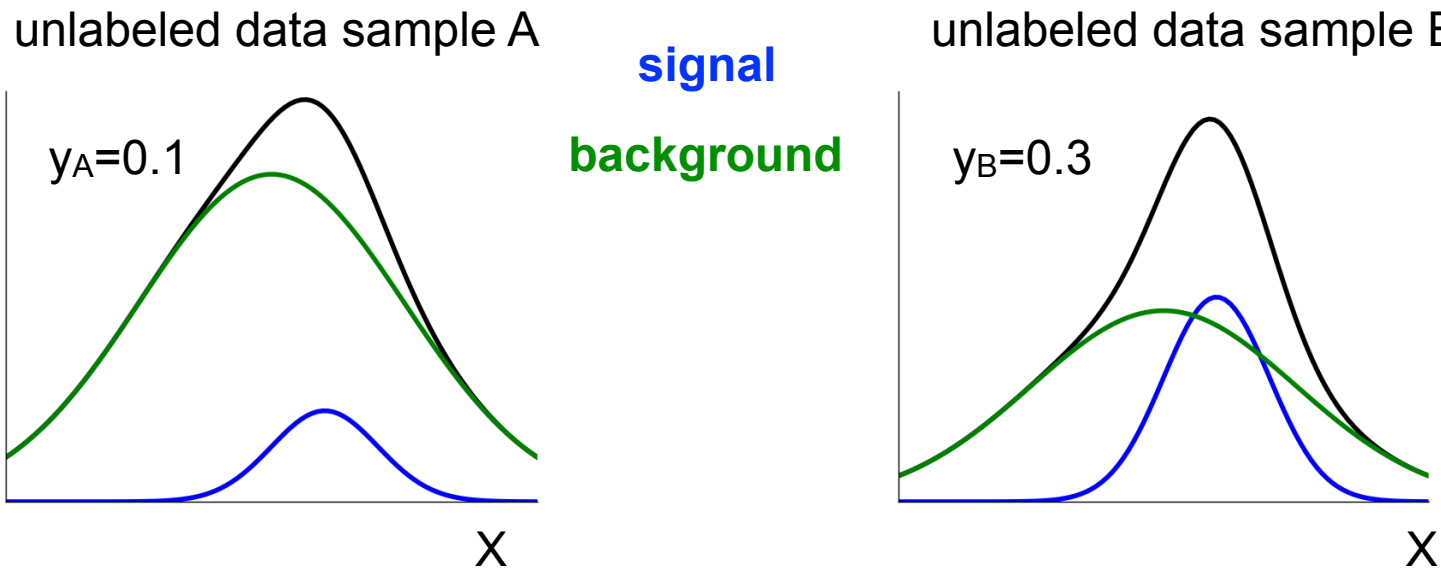


$$h_{A,i} = y_A h_{1,i} + (1 - y_A) h_{0,i}$$

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

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

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

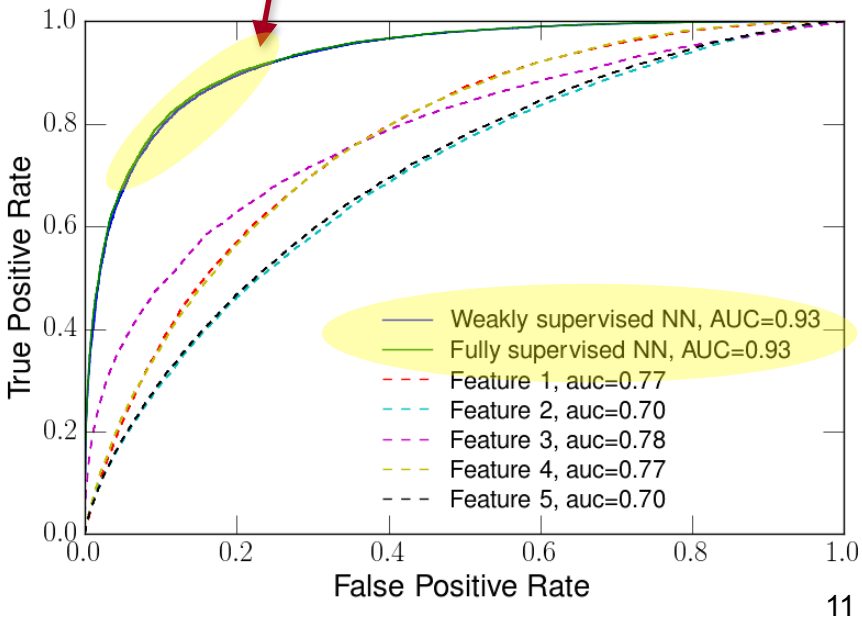
- The analytic approach requires binning and becomes quickly unmanageable as the feature space grows.

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

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

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

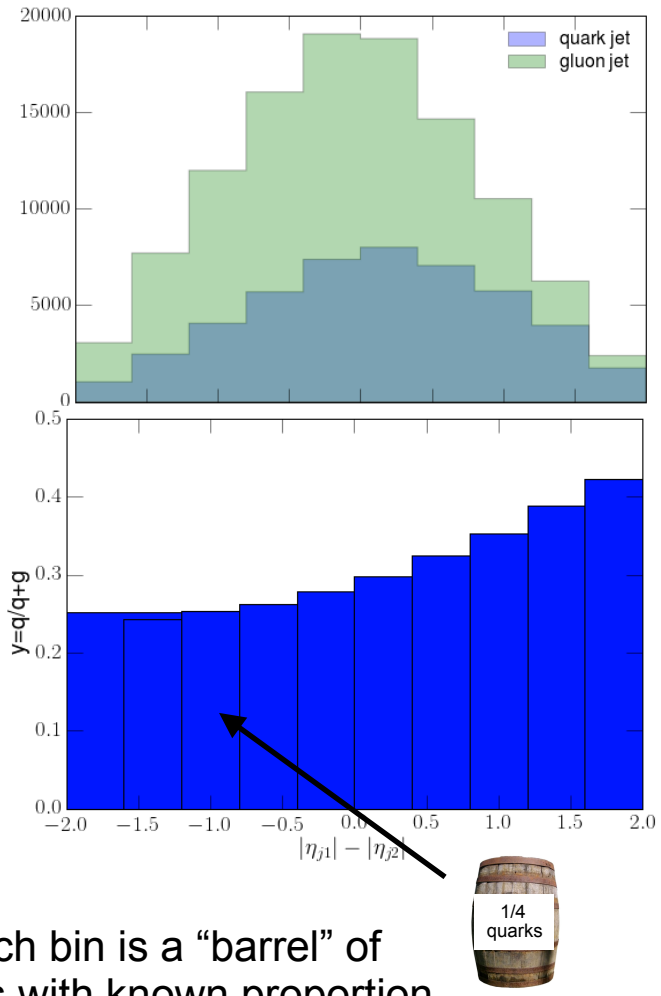
Same discrimination as fully supervised

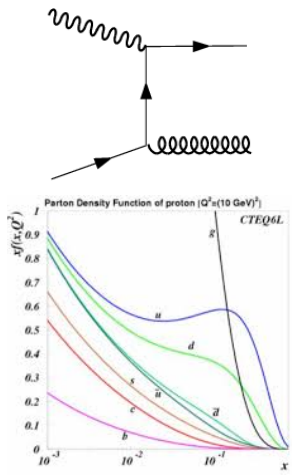


Weak supervision - q/g tagging

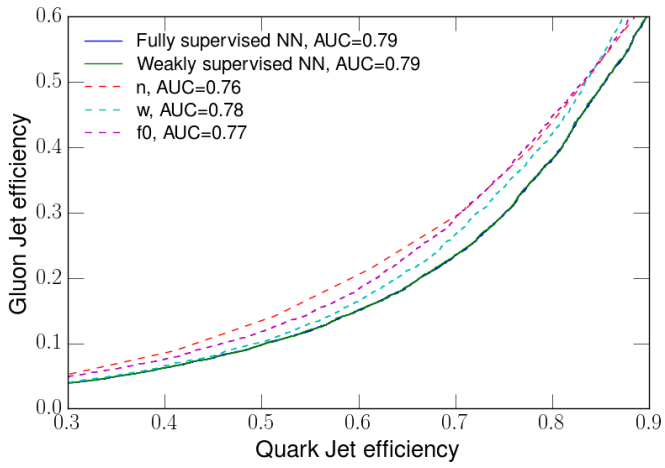
SLAC

$|\eta_{j1}| - |\eta_{j2}|$ in dijet events





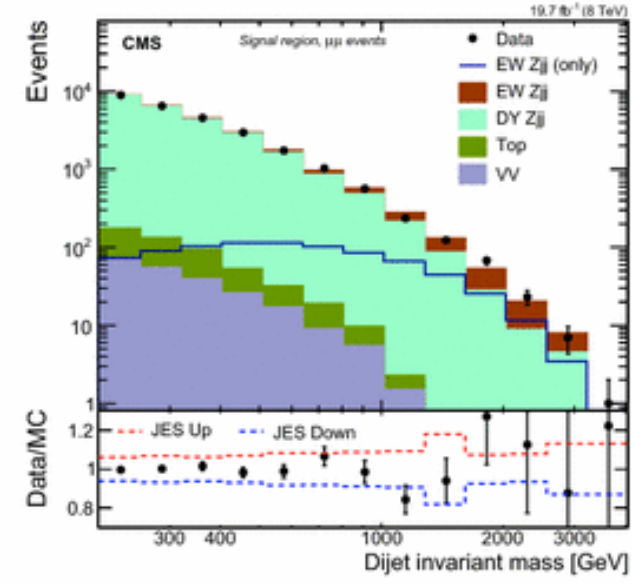
Leverage precise description of ME and PDF (MC/theory) to extract discrimination from soft QCD features (from data!)



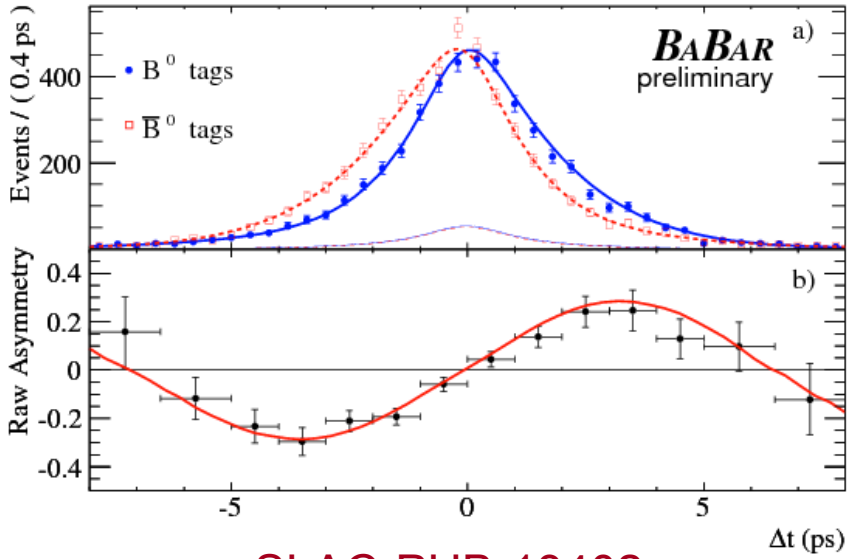
Each bin is a “barrel” of jets with known proportion

Conclusion

- **Weak supervision** is a new paradigm 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**.



Eur. Phys. J. C (2015) 75: 66



SLAC-PUB-13402

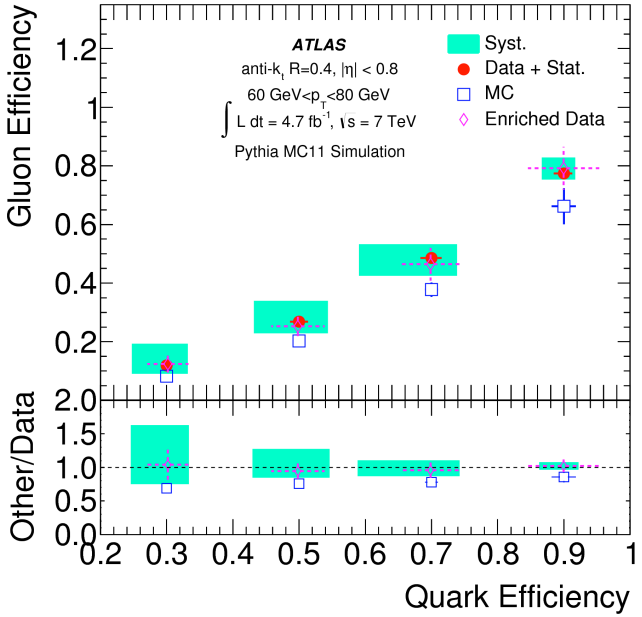
References

- 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 √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

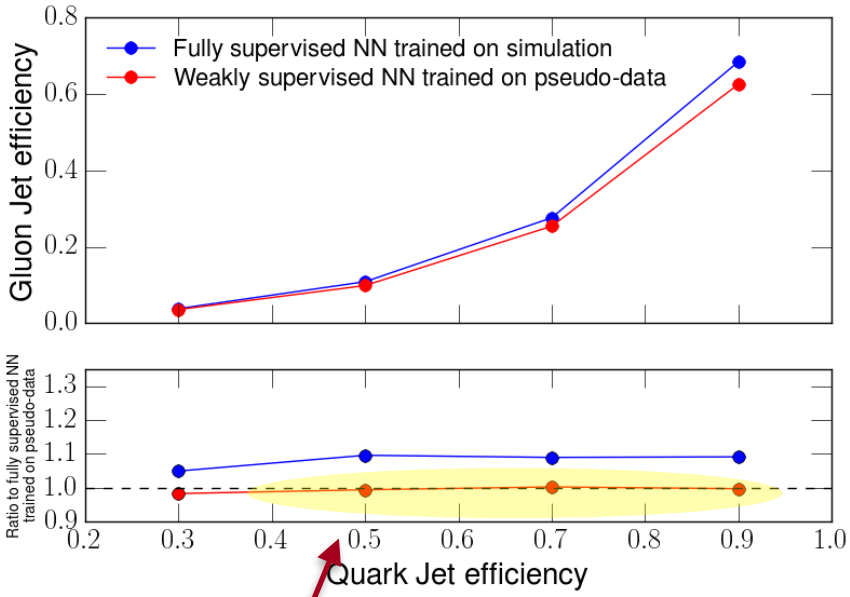
Backup

Weak supervision

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

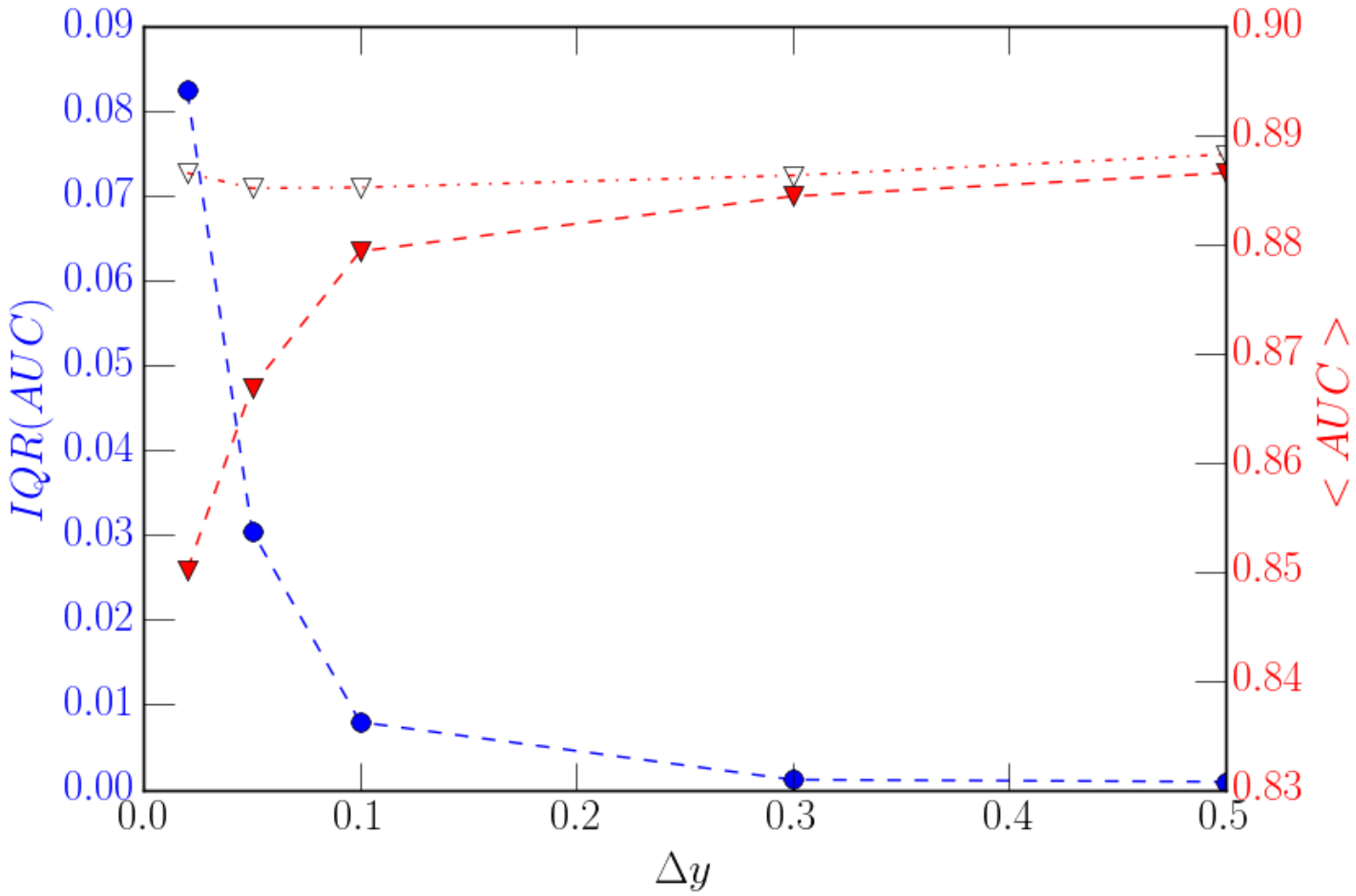


1405.6583



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