Weak Supervision in High Dimensions

Machine Learning for Jet Physics Workshop, 2017

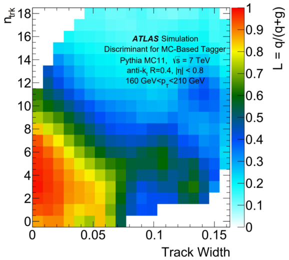
Eric M. Metodiev

Center for Theoretical Physics

Massachusetts Institute of Technology

Work with Patrick T. Komiske, Francesco Rubbo, Benjamin Nachman, and Matthew D. Schwartz

December 13, 2017



Why learn from data?

Weak Supervision in HEP

Lessons from High Dimensions

Simulation vs. Data

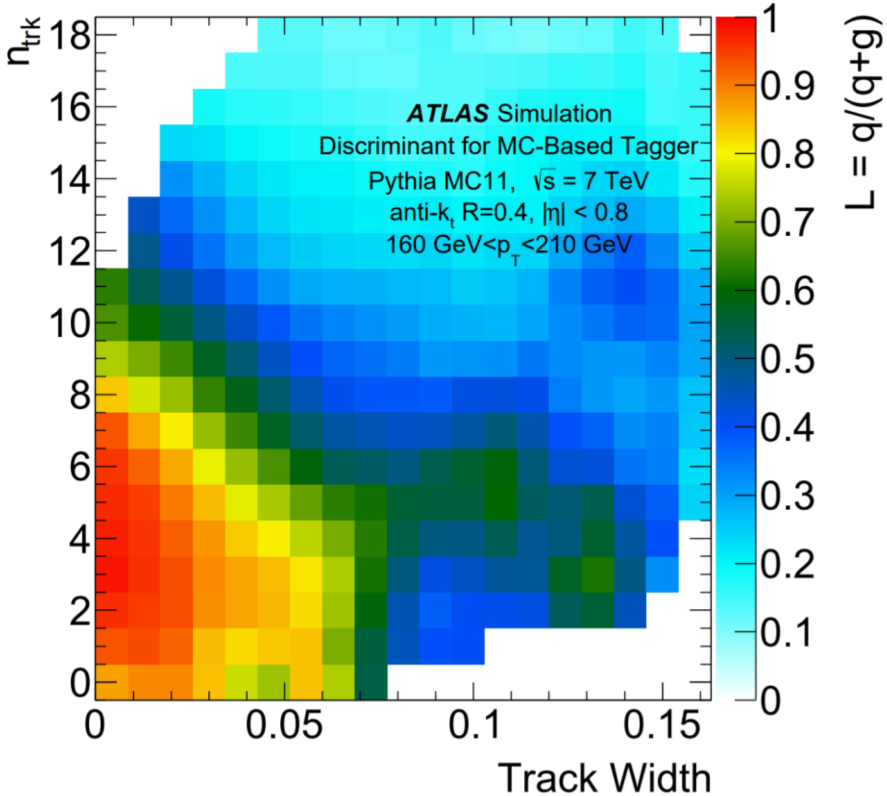
Quark/Gluon Discrimination

Using two features: width and ntrk.

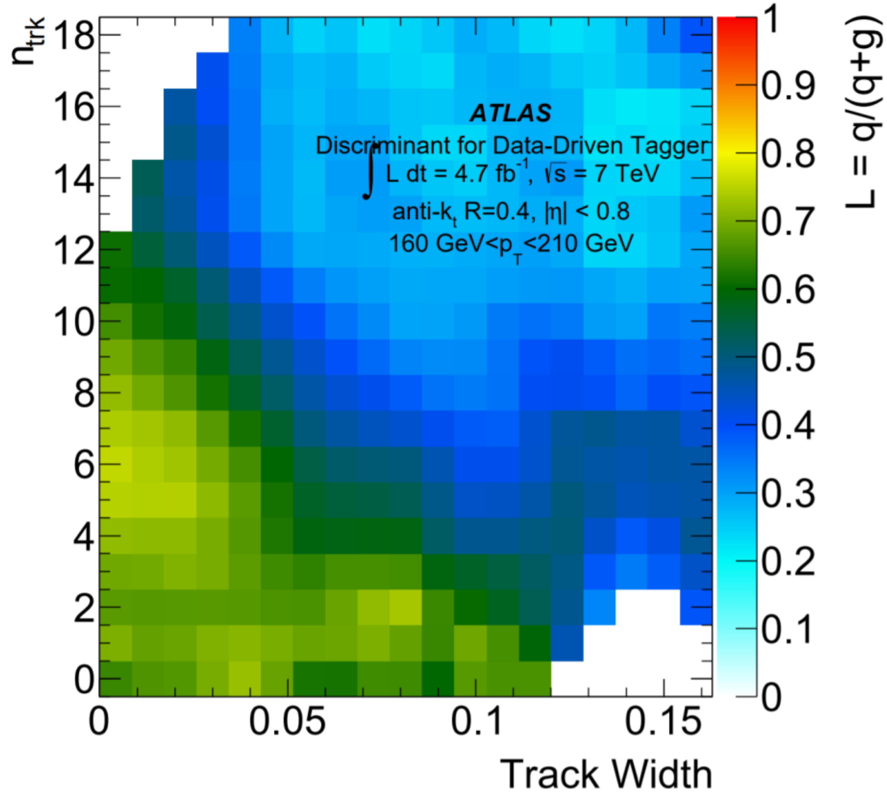
Signal (Q) vs. Background (G) likelihood ratio

[ATLAS Collaboration, arXiv: 1405.6583]

Simulation



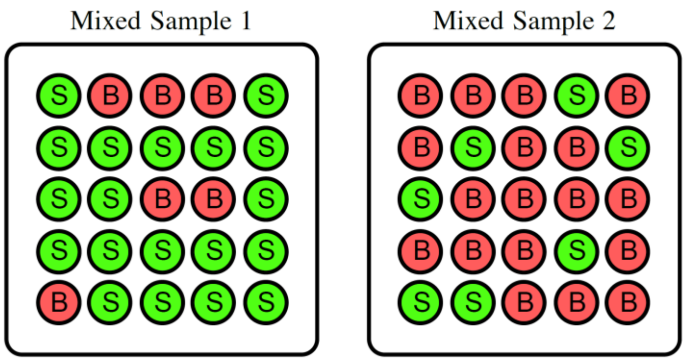
Data



Mixed Samples

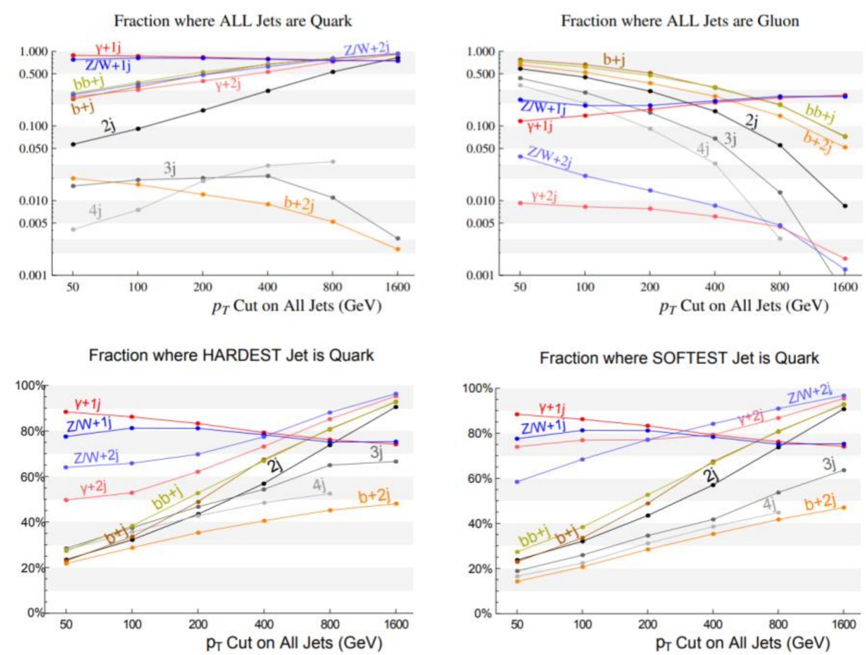
Data does not have pure labels, but does have mixed samples!

Some caveats apply. See e.g. [P. Gras, *et al.*, arXiv: 1704.03878]



$$p_{M_a}(x) = f_a\, p_S(x) + (1 - f_a)\, p_B(x)$$

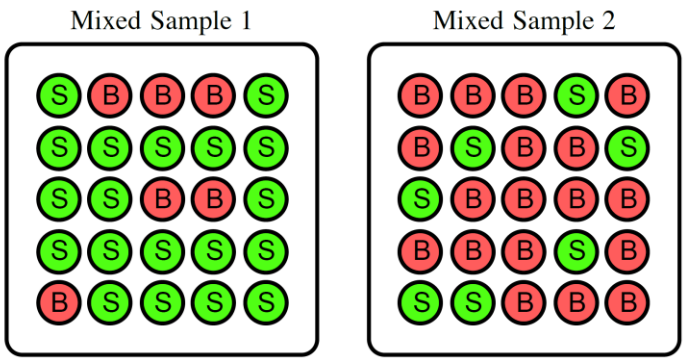
Fractions of quark and gluon jets studied in detail in:
[J. Gallicchio and M.D. Schwartz, arXiv: 1104.1175]



Mixed Samples

Data does not have pure labels, but does have mixed samples!

Some caveats apply. See e.g. [P. Gras, *et al.*, arXiv: 1704.03878]



$$p_{M_a}(x) = f_a\, p_S(x) + (1 - f_a)\, p_B(x)$$

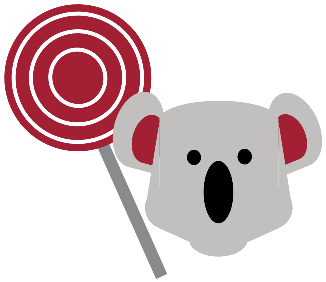
Criteria to use Weak Supervision:

Sample Independence: The *same* signal and background in all the mixtures.

Different Purities: $f_a \neq f_b$ for some a and b.

(Known fractions): The fractions f_a are known.

Why learn from data?

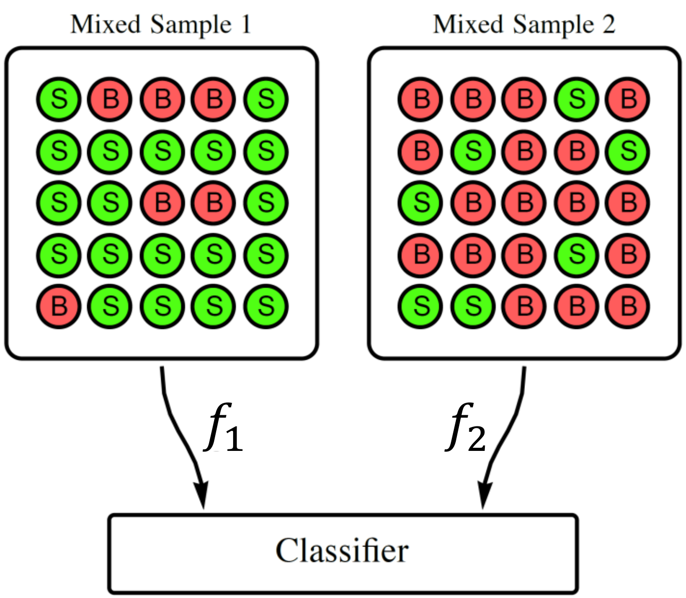


Weak Supervision in HEP

Lessons from High Dimensions

Learning from Label Proportions (LLP) (LoLiProp?)

[L. Dery, *et al.*, arXiv: 1702.00414]

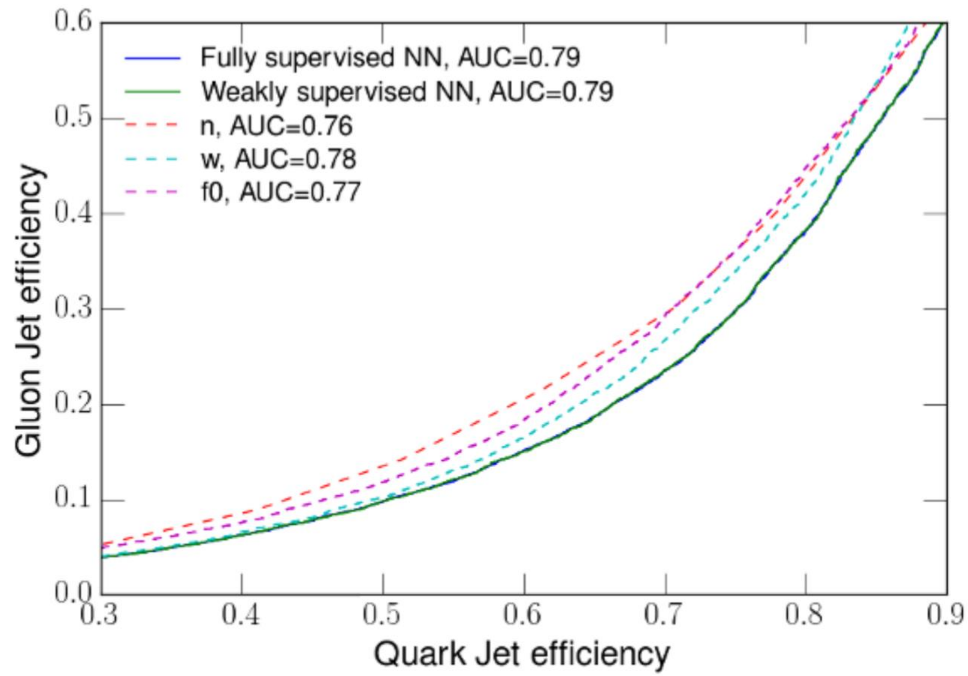


$$\ell_{\text{LLP}} = \sum_a \ell\left(f_a, \frac{1}{N_a}\sum_{i=1}^{N_a} h(x)\right)$$

$\ell_{MSW}, \ell_{CE}, \ldots$

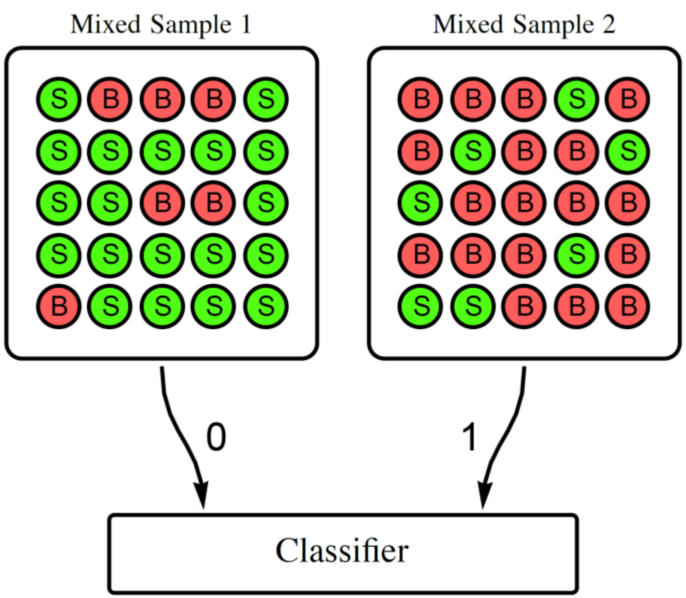
Q/G WS with **3** inputs works

[L. Dery, *et al.*, arXiv: 1702.00414]



Classification Without Labels (CWoLa, “koala”)

[EMM, B. Nachman, and J. Thaler, arXiv: 1708.02949]
[T. Cohen, M. Freytsis, and B. Ostdiek, arXiv: 1706.09451]
See also: [G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, arXiv:1303.1208]



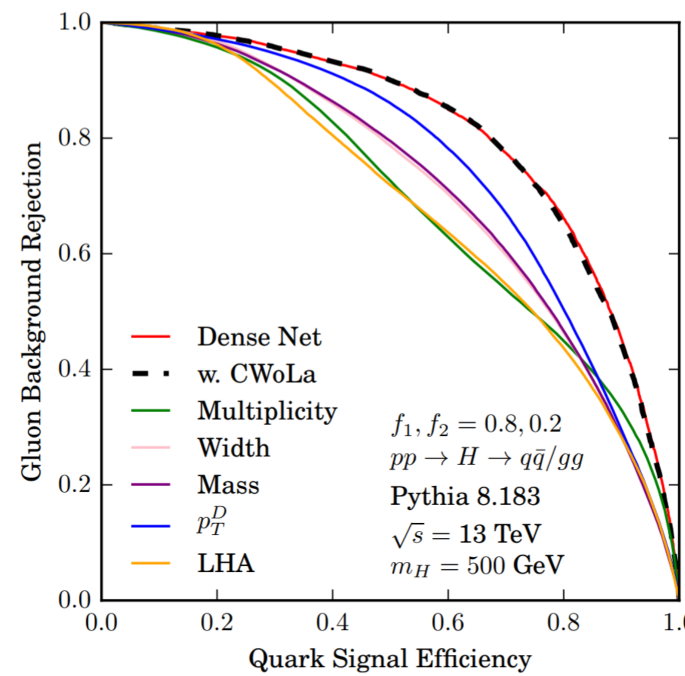
No label proportions needed during training!

Smoothly connected to the fully supervised case as $f_1, f_2 \to 0{,}1$

Note: Need small test sets with known signal fractions to determine the ROC.

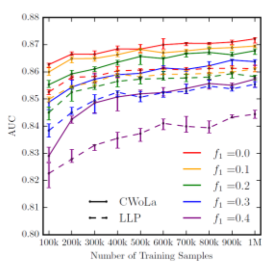
Q/G WS with **5** inputs works

[EMM, B. Nachman, and J. Thaler, arXiv: 1708.02949]



Why learn from data?

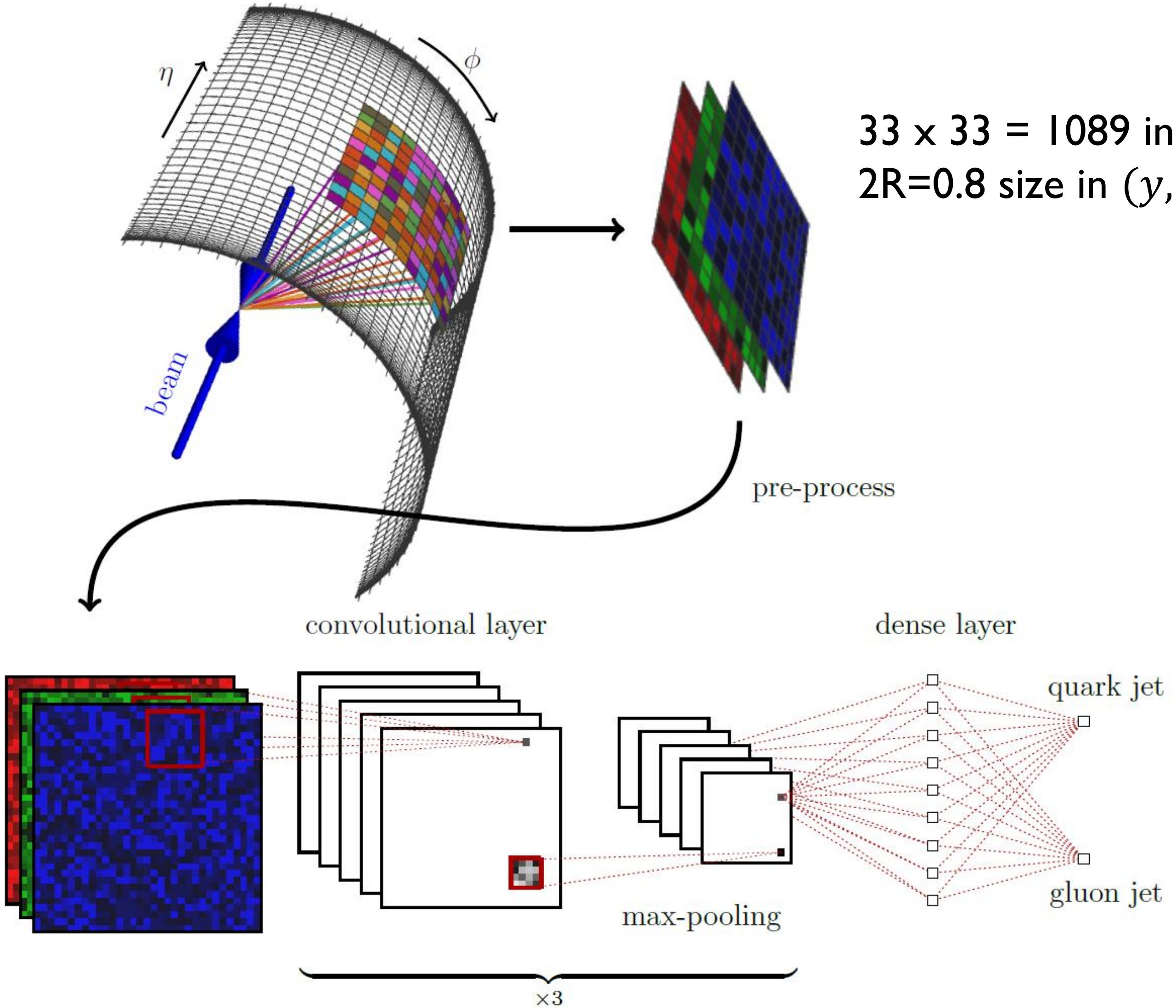
Weak Supervision in HEP



Lessons from High Dimensions

Convolutional Net for QG

CNN as in:
[P. Komiske, E. Metodiev, M.D. Schwartz, arXiv:1612.01551]

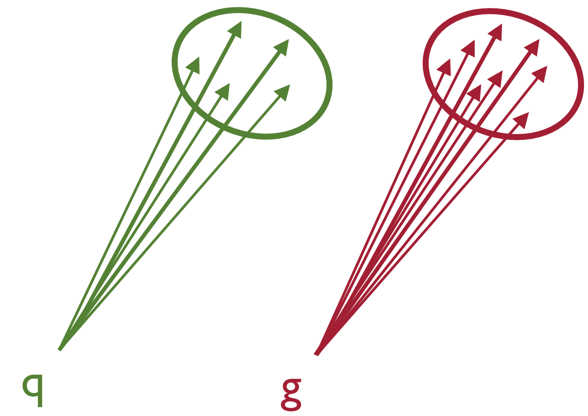


33 x 33 = 1089 inputs,
2R=0.8 size in (y, ϕ)

Only used pT-channel images

Defaults

Jet Generation



Z + q/g
Pythia 8.226, $\sqrt{s}$ =13 TeV
R=0.4 anti-kT central jets
pT in [250 GeV, 275 GeV]
Artifical q/g mixtures

CNN Training

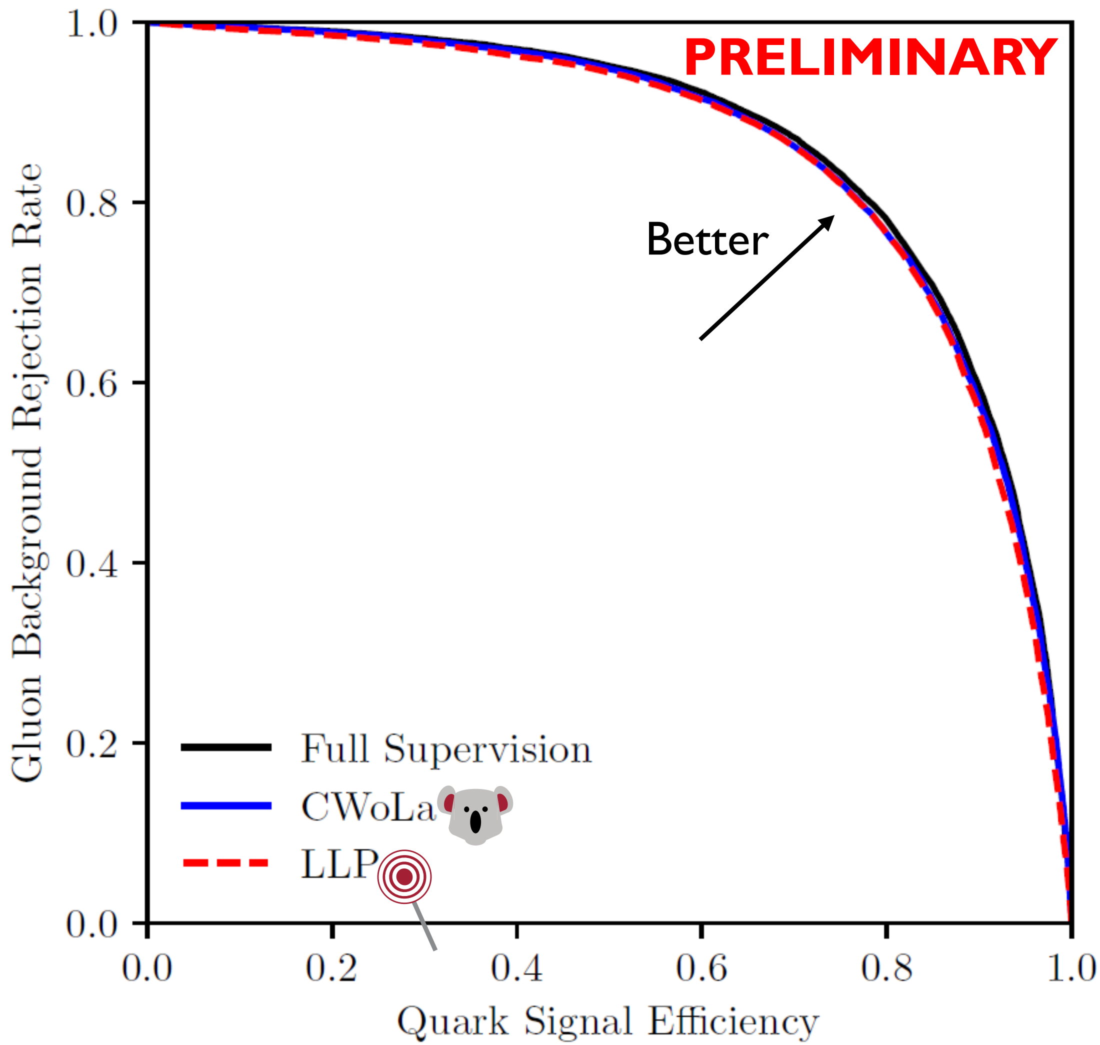
Keras and TensorFlow
300k/50k/50k train/test/val data
Mixed sample fractions $f_1 = 0.2$ and $f_2 = 0.8$
Batch size 400 for CWoLa and 4k for LLP
ELU activation and cross-entropy loss functions

Training until validation accuracy failed to improve for 10 epochs
Repeat each training 10x for statistics

Training on mixed samples

Q/G weak supervision with jet images works!

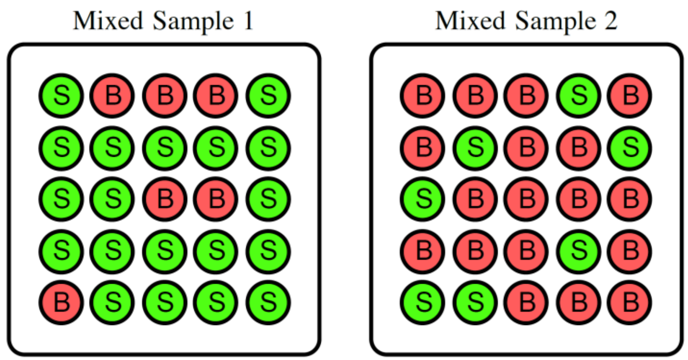
Lesson should be true for complex models more generally.



What about naturally mixed samples?

Z + jet:
$f_q = 0.88$

dijets:
$f_q = 0.37$



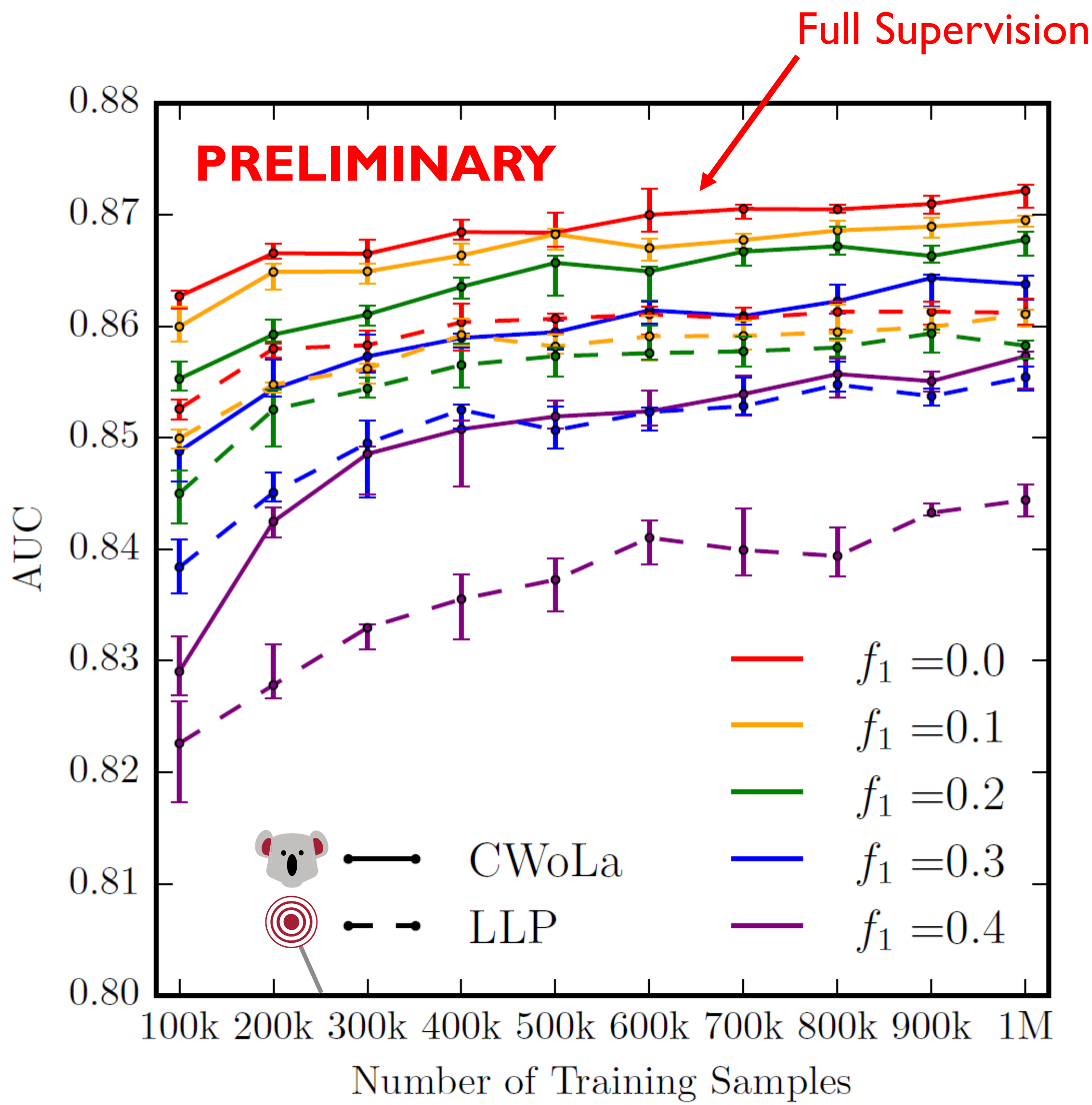
Learning	Sample	AUC
CWoLa	Z+jet vs. dijets	0.8619 ± 0.0014
CWoLa	Artificial $Z + q/g$	0.8621 ± 0.0019
LLP	Z+jet vs. dijets	0.8535 ± 0.0022
LLP	Artificial $Z + q/g$	0.8509 ± 0.0028

Restrict to artificially mixed samples to have fine control of the fractions.

Purity and Number of Data

Two mixed samples: $f_1, 1 - f_1$

Purity/Data plot can characterize tradeoffs in a weak learning method



Batch Size and Training Time

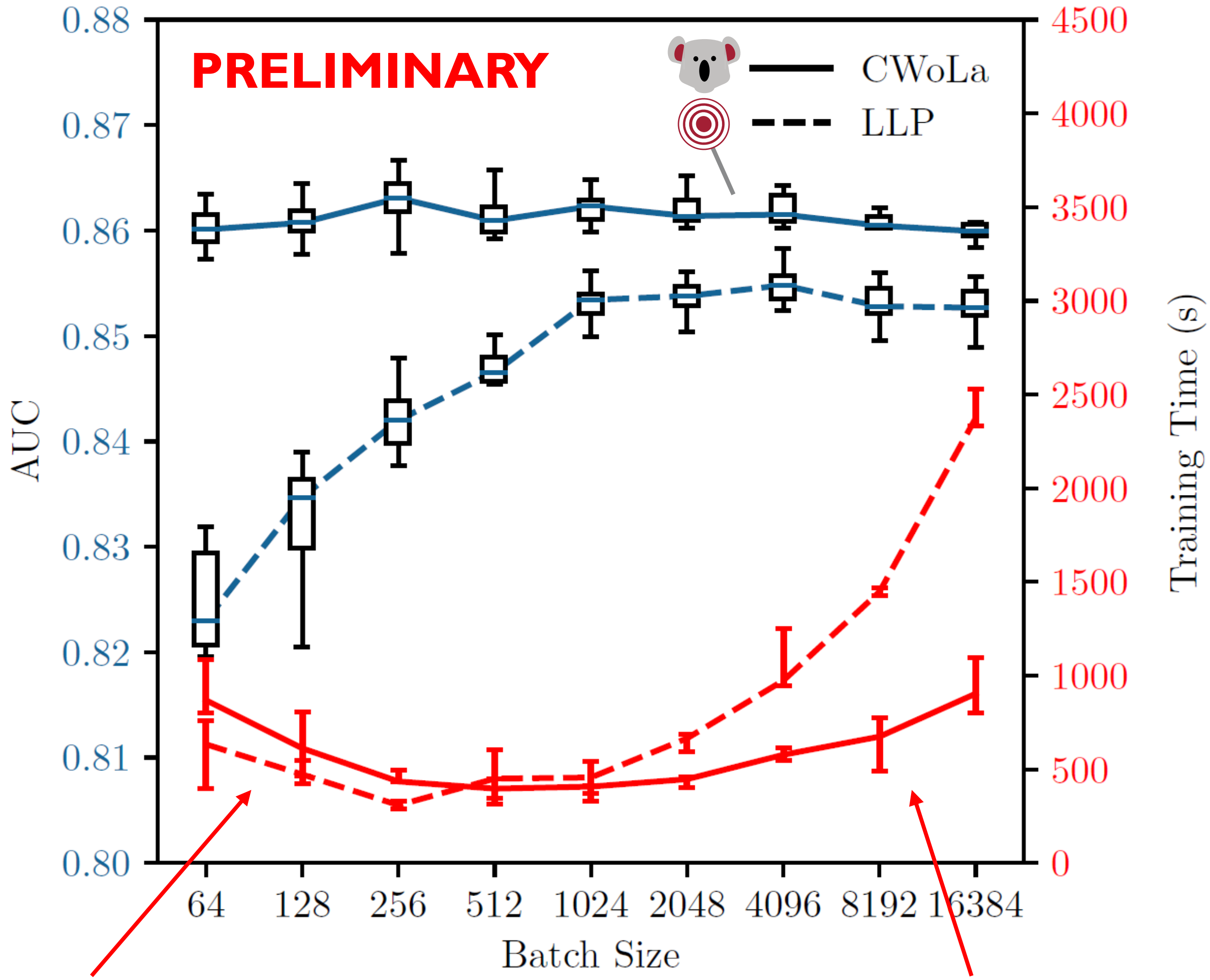
Batch size

Usual parameter for CWoLa

Need large batch size for LLP

Batch Size > 1000

$$\ell_{\mathrm{LLP}} = \sum_a \ell\left(f_a, \frac{1}{N_a}\sum_{i=1}^{N_a} h(x)\right)$$



time/epoch increases

# of epochs increases

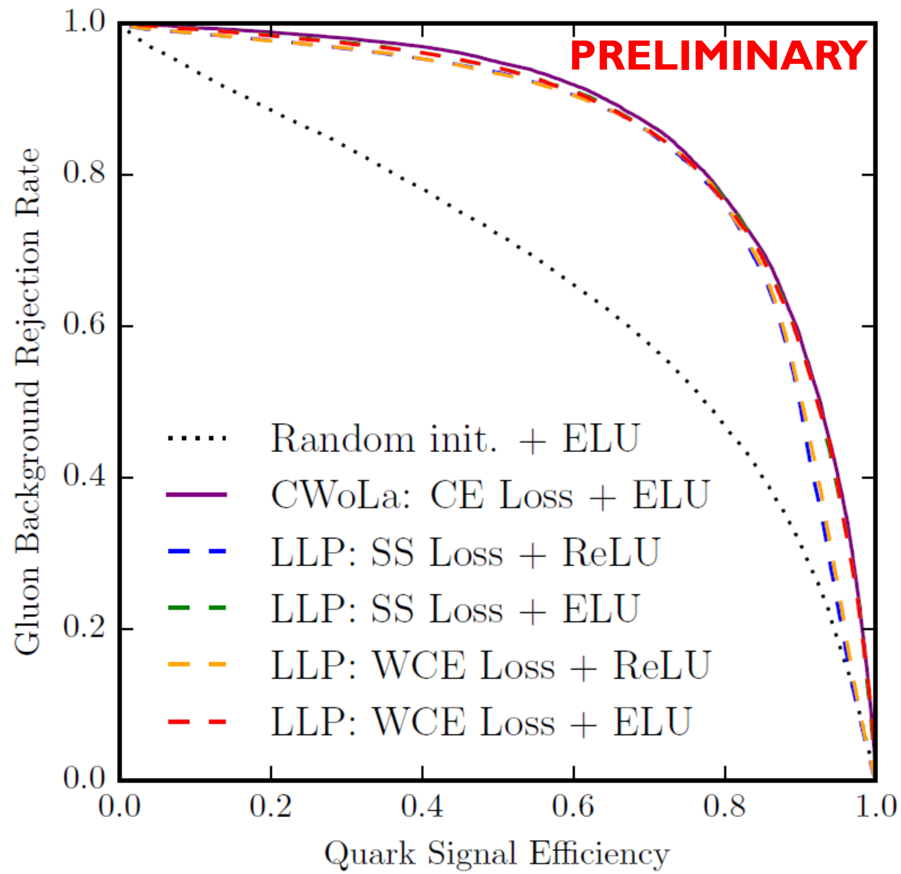
Loss and Activation Functions

- LLP:

  ELU activations help significantly over ReLU activations.

  Weak crossentropy loss helps over weak MSE loss.

  Include the softmax in the loss (not model) to avoid underflow.



Conclusions

Weak supervision methods work for training complex classifiers.

Have several different methods that utilize different information.

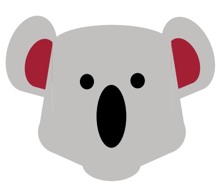
Which to use depends on the specific application.

LLP:

Requires specialized loss functions and care

Utilizes fraction information

Can make use of multiple fractions



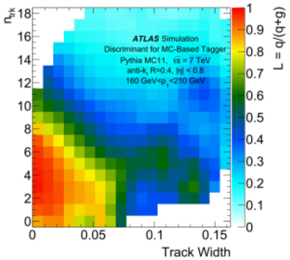
CWoLa:

Can use with any fully supervised technique

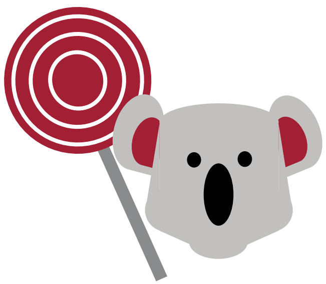
Does not require fraction information

Only works with two mixed samples

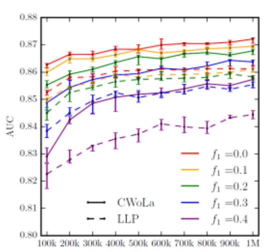
The End



Why learn from data?



Weak Supervision in HEP



Lessons from High Dimensions

Multiple Mixture Fractions