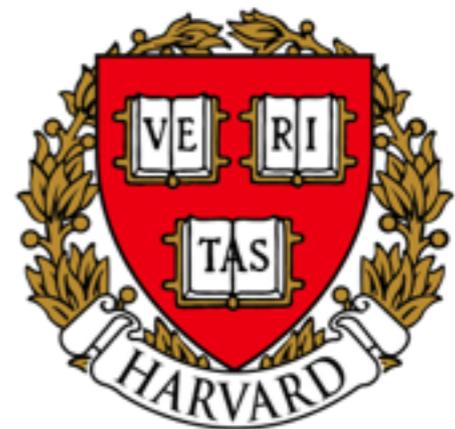


# Learning to Classify from Impure Samples with High-Dimensional Data

based on Phys. Rev. D 98, 011502(R)

Patrick Komiske, Eric Metodiev,  
Benjamin Nachman, Matt Schwartz

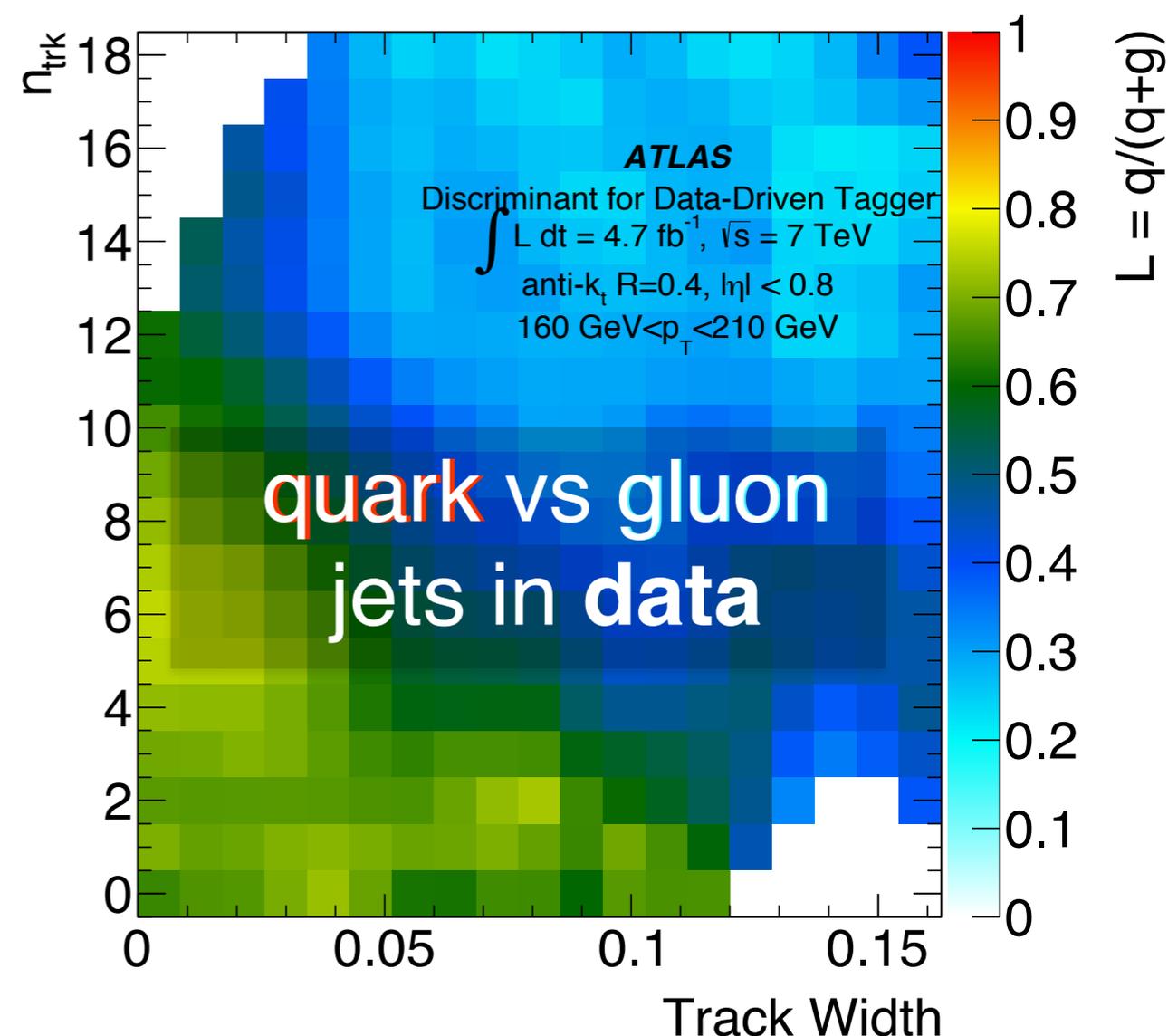
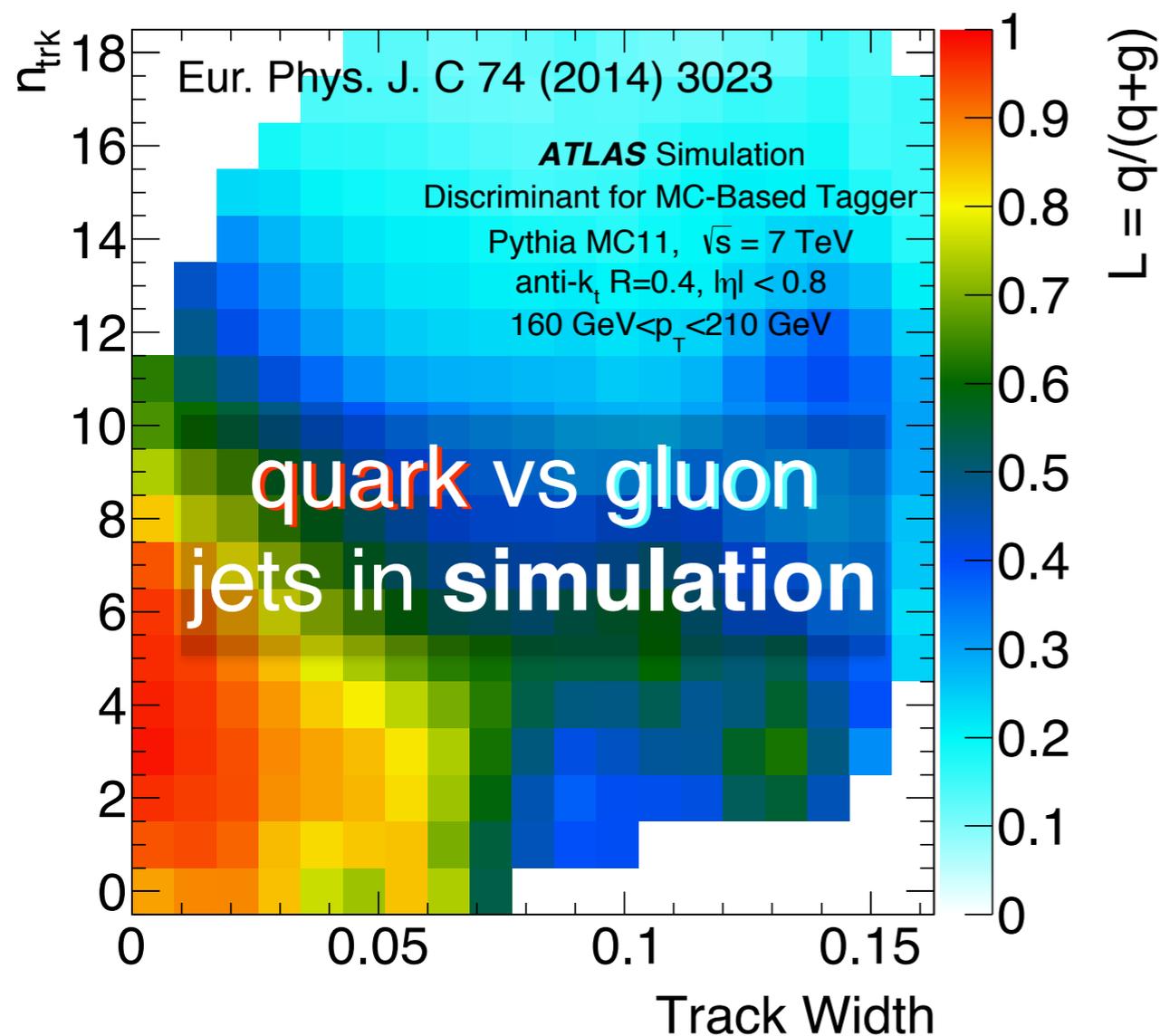
*...building on work also in collaboration with Lucio Dery,  
Francesco Rubbo, Ariel Schwartzman, Jesse Thaler*



# Motivation



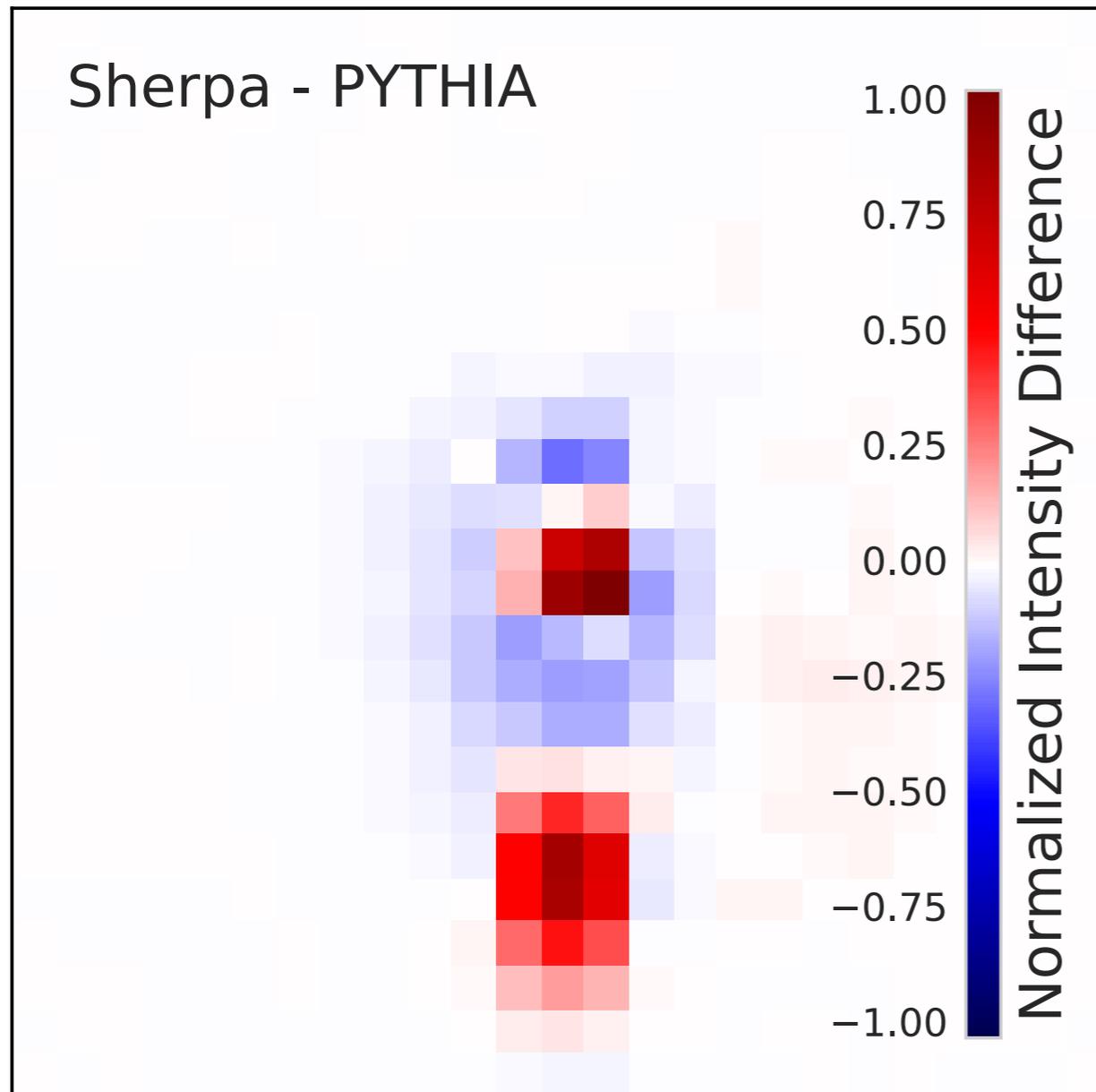
Usual paradigm: train in simulation, test on data.



If data and simulation differ, this is **sub-optimal!**

# Motivation, continued

3



Especially important for **deep learning** using subtle features → hard to model!

*W boson radiation pattern - same physics, different simulators!*

J. Barnard, E. Dawe, M. Dolan, N. Rajcic,  
Phys. Rev. D 95 (2017) 014018

# Solution 1: Use class proportions



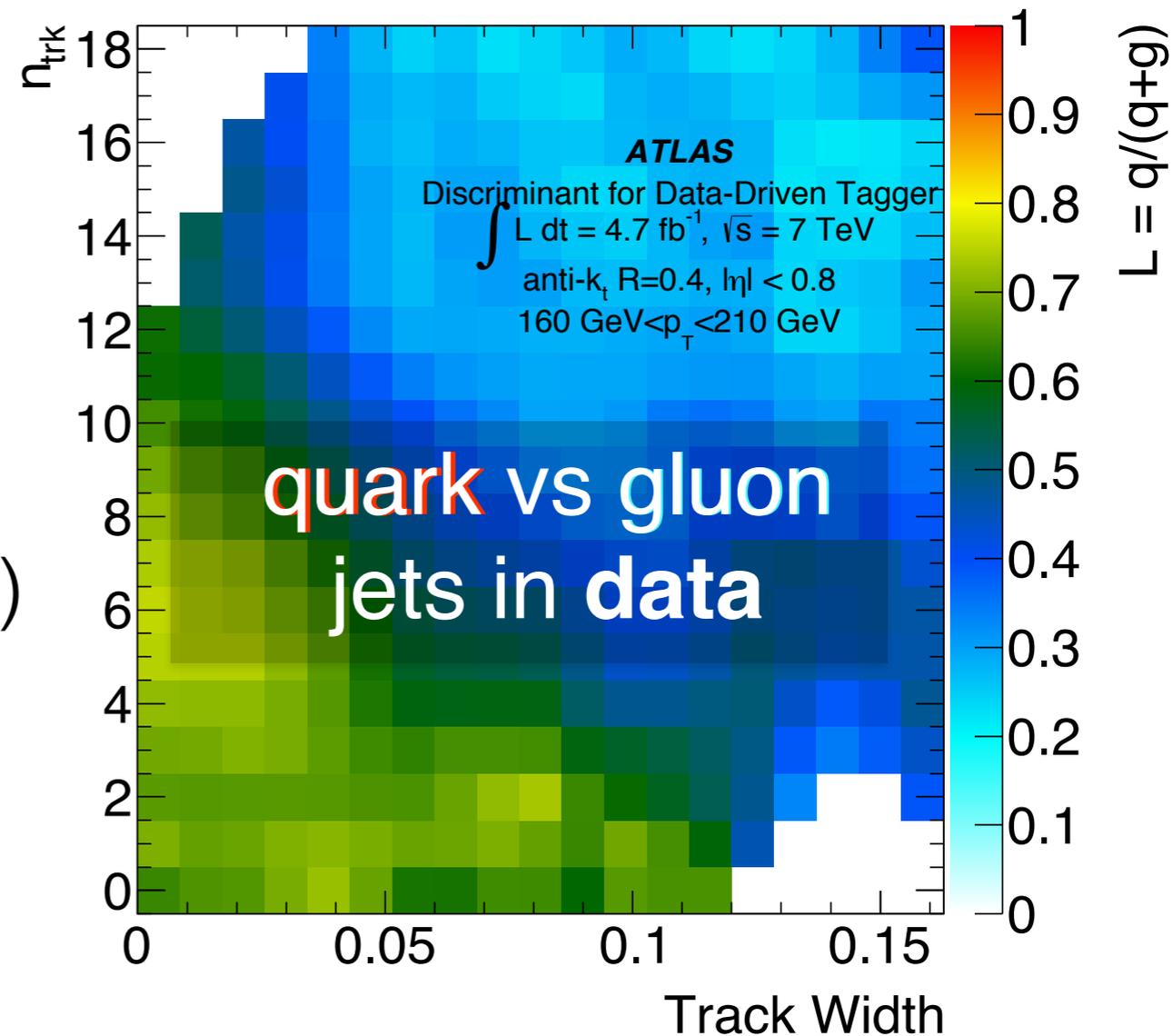
How did we make this plot?

$$\text{dijets} = f_q \times Q + (1-f_q) \times G$$

$$\text{Z+jets} = g_q \times Q + (1-g_q) \times G$$

two equations, two unknowns (Q, G)

We often know  $f$ ,  $g$   
(from ME + PDF) much better than  
full radiation pattern inside jets.



This doesn't work well when you have more than 2  
observables because the templates become sparse.

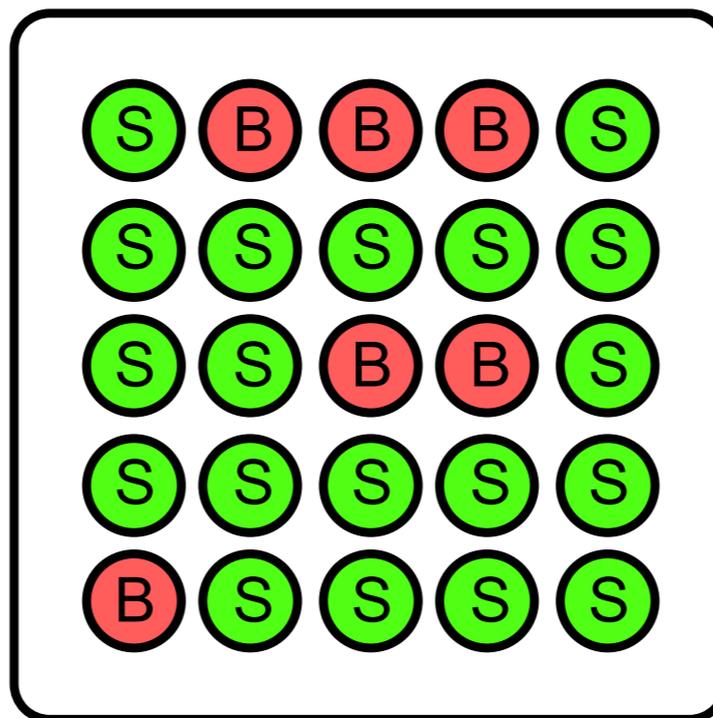
# Method 1: Learn from Proportions



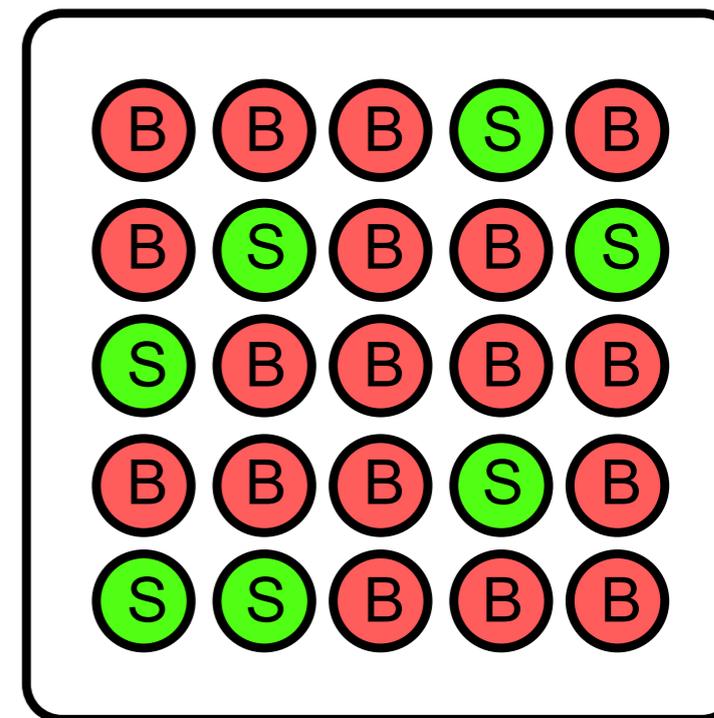
$$f_{\text{full}} = \operatorname{argmin}_{f': \mathbb{R}^n \rightarrow \{0,1\}} \sum_{i=1}^N \ell(f'(x_i) - t_i)$$

$\xrightarrow{\text{loss fcn.}}$        $\xrightarrow{\text{labels}}$

Mixed Sample 1



Mixed Sample 2



## LoLiProp

*Learning from Label Proportions*

Solution: Train using class proportions.  
Work “on average”

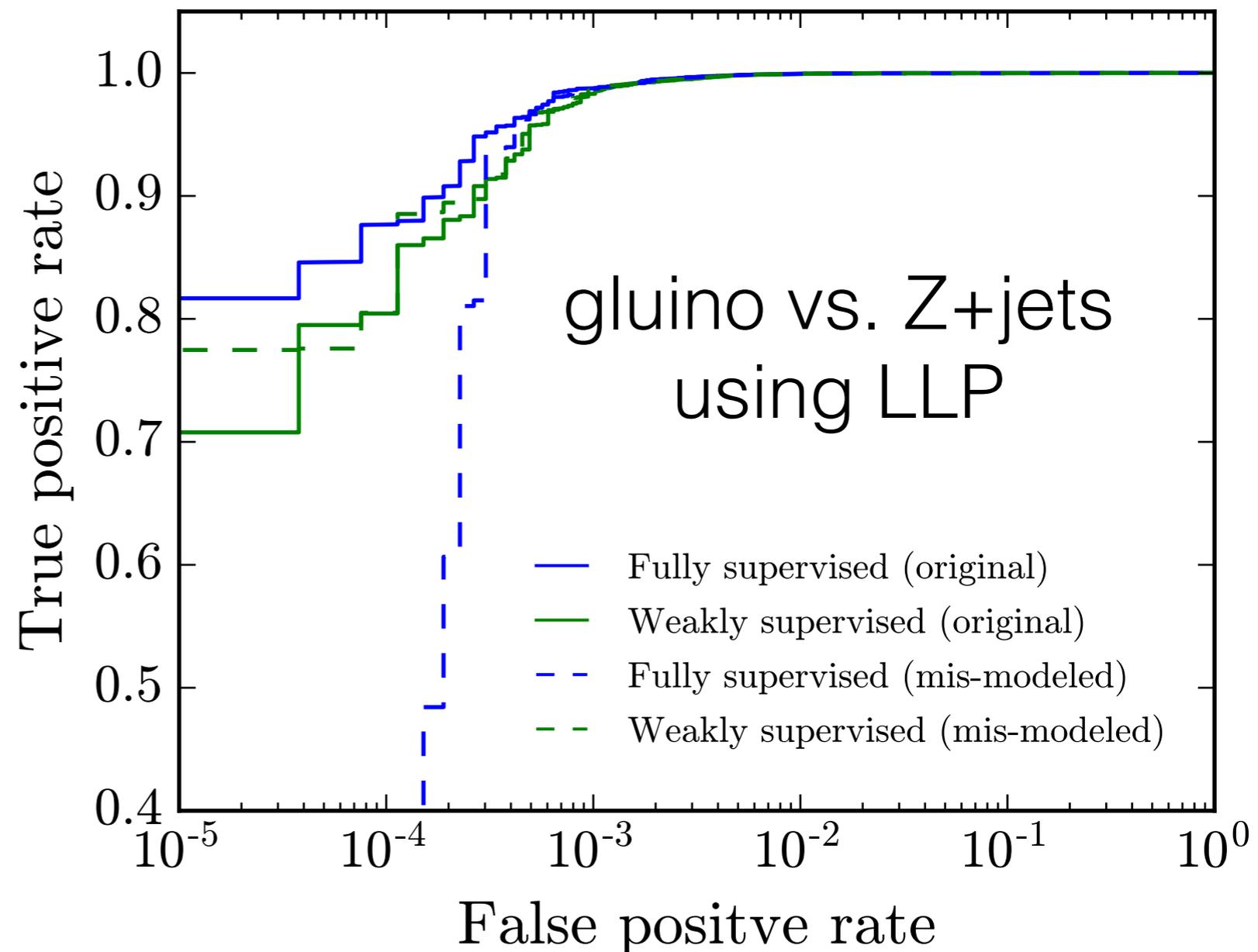
$$f_{\text{weak}} = \operatorname{argmin}_{f': \mathbb{R}^n \rightarrow [0,1]} \ell \left( \sum_{i=1}^N \frac{f'(x_i)}{N} - y \right)$$

$\xrightarrow{\text{proportions}}$

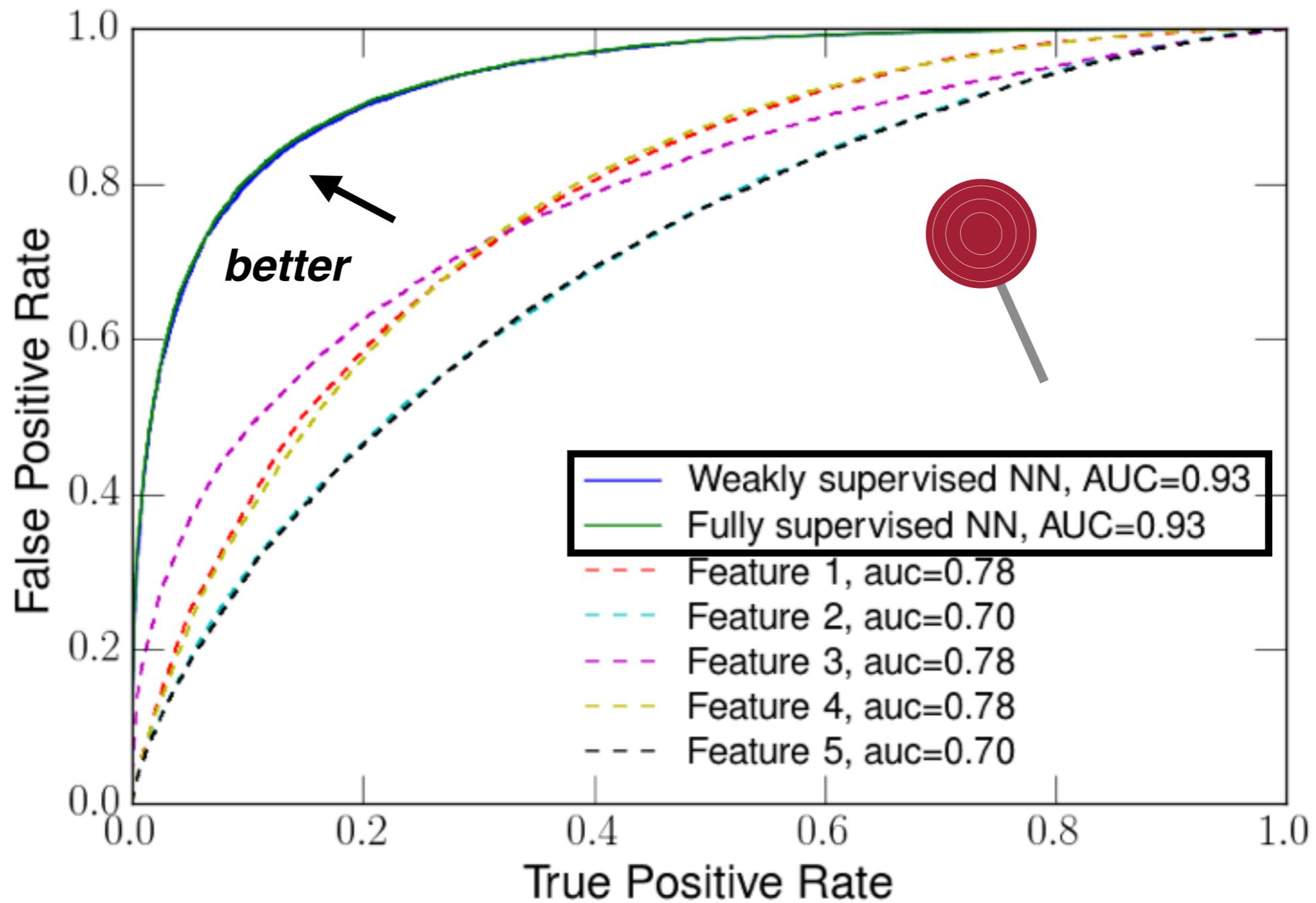
# N.B. Don't need 100% fraction accuracy

6

Even though the proportions are required as input, if they are slightly wrong, you can end up with the correct classifier.



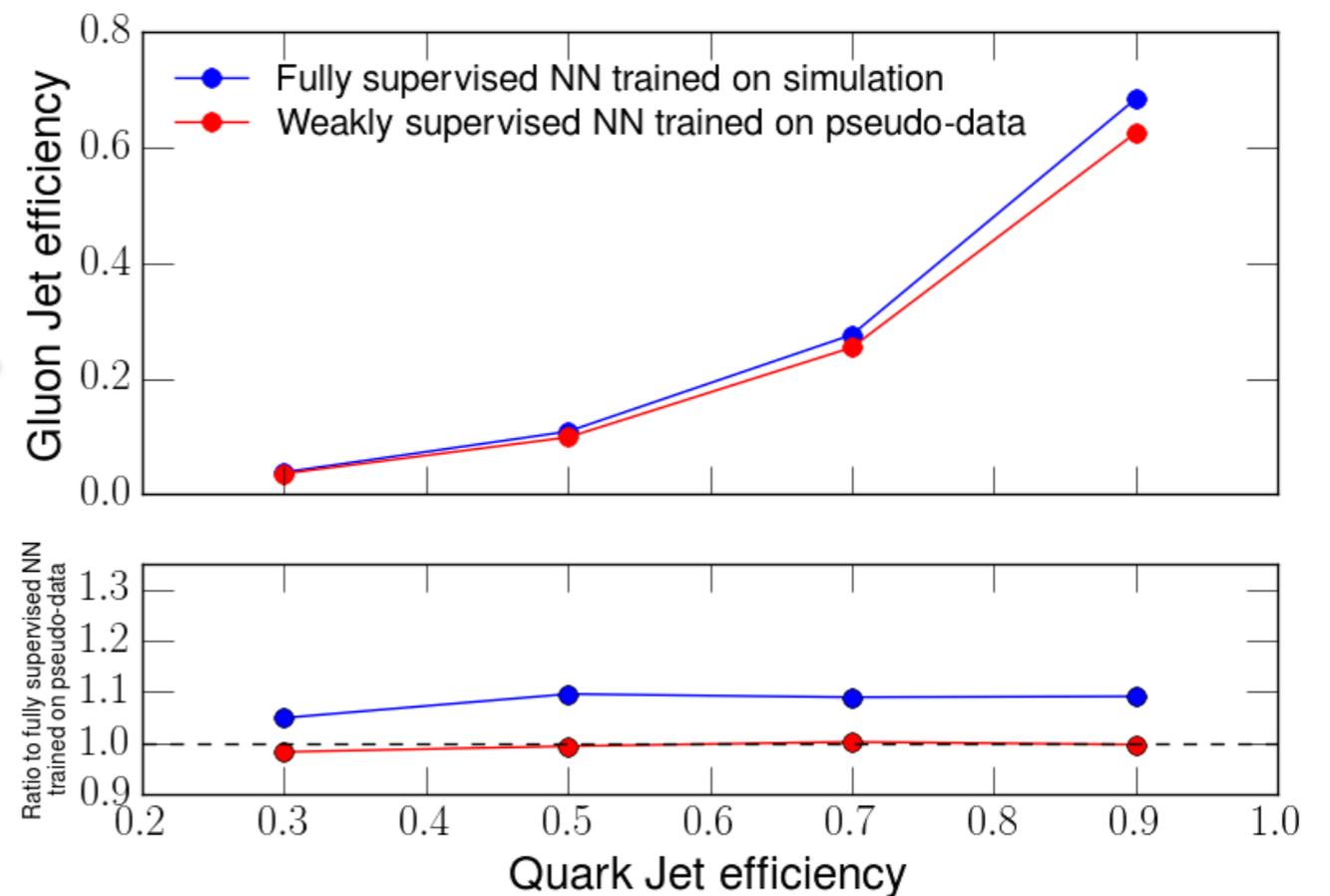
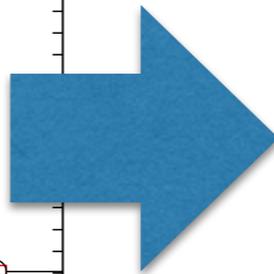
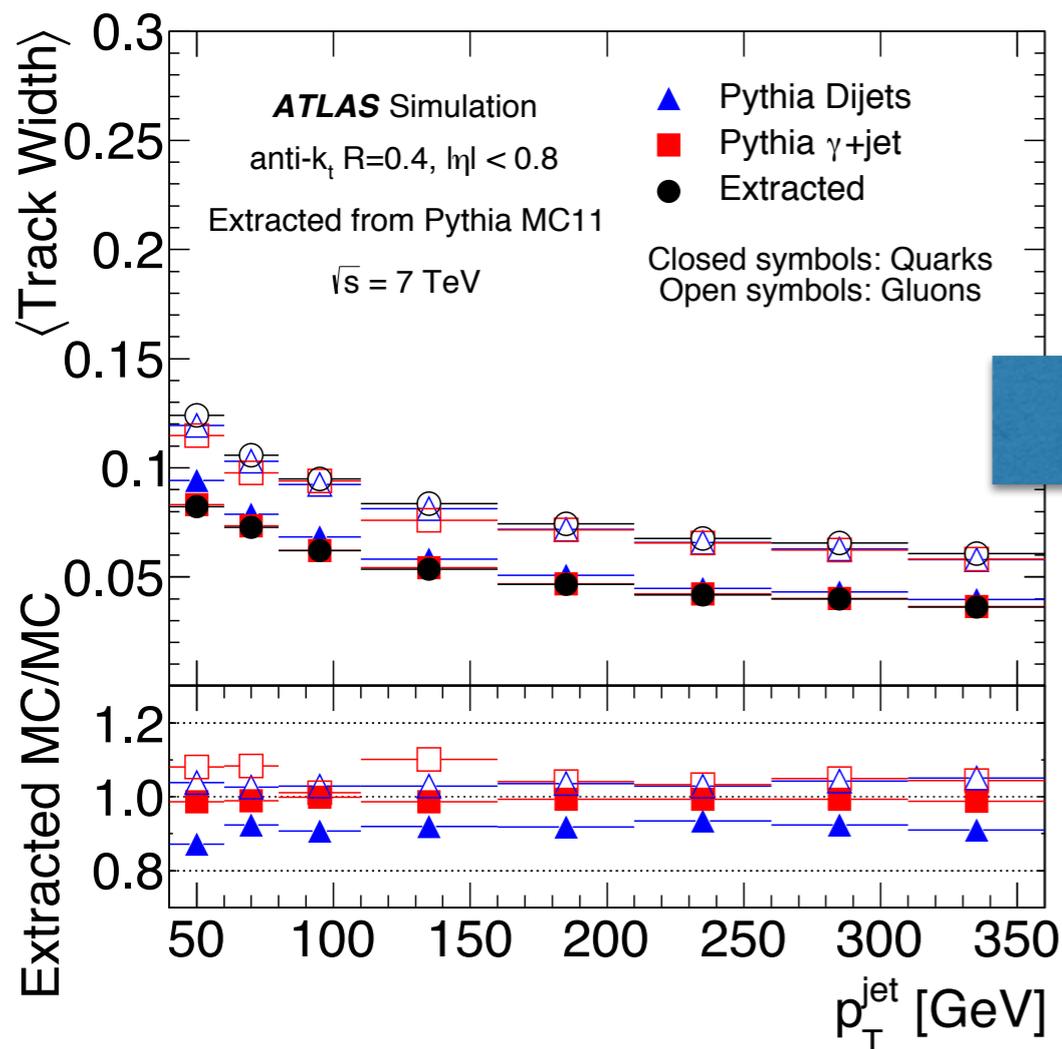
# Works in low-dimensions



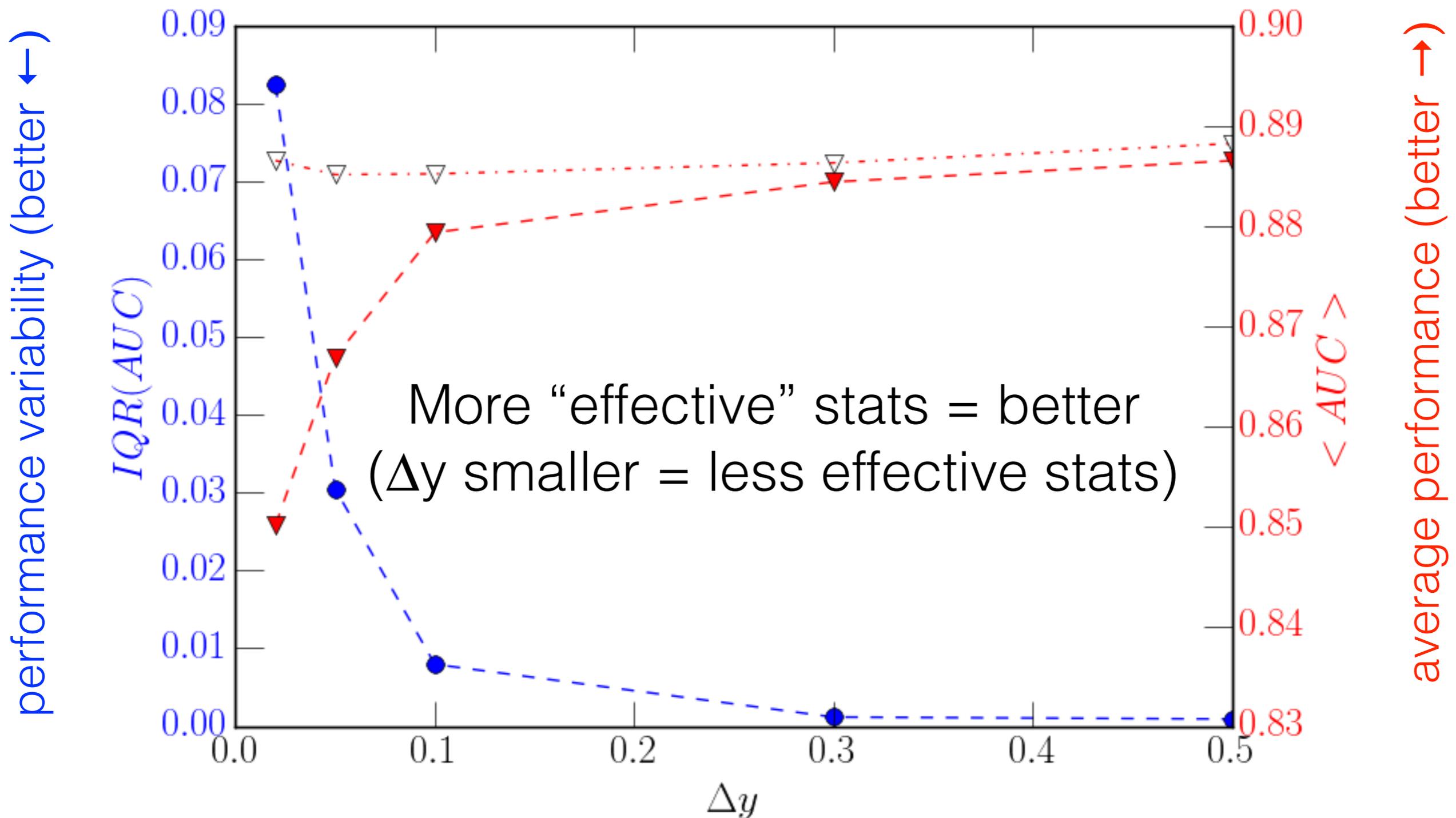
# Works in low-dimensions ... for q/g



Given the data/MC disagreement from the first slide, this is what you might expect in terms of the performance difference.



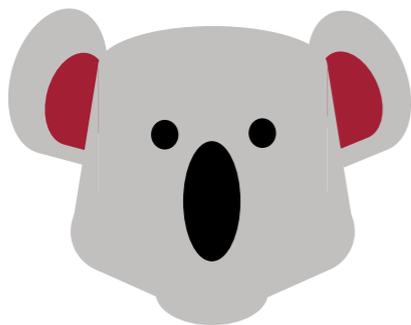
# A note about training statistics



how different are the proportions for the two mixed samples

# Method 2: Learning without Proportions

10

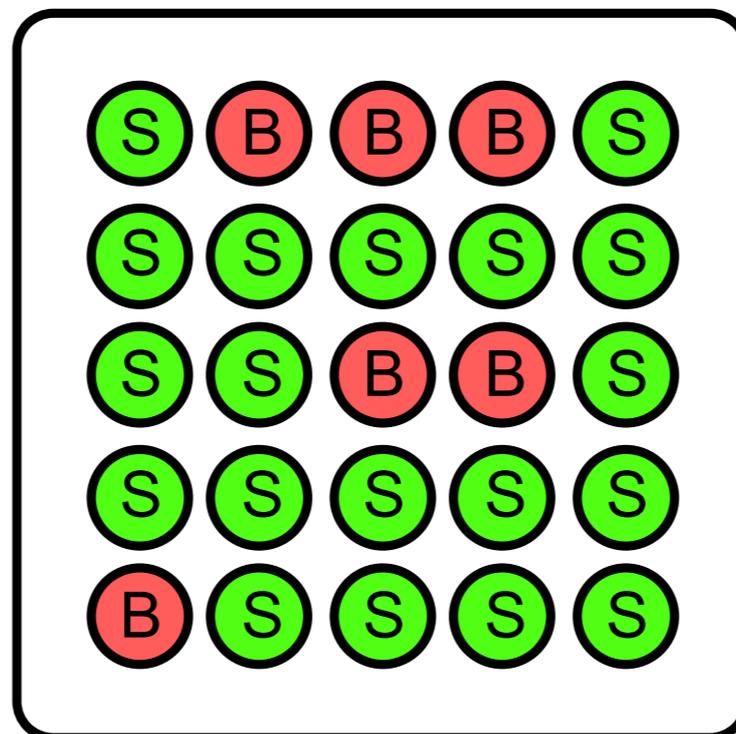


## CWoLa

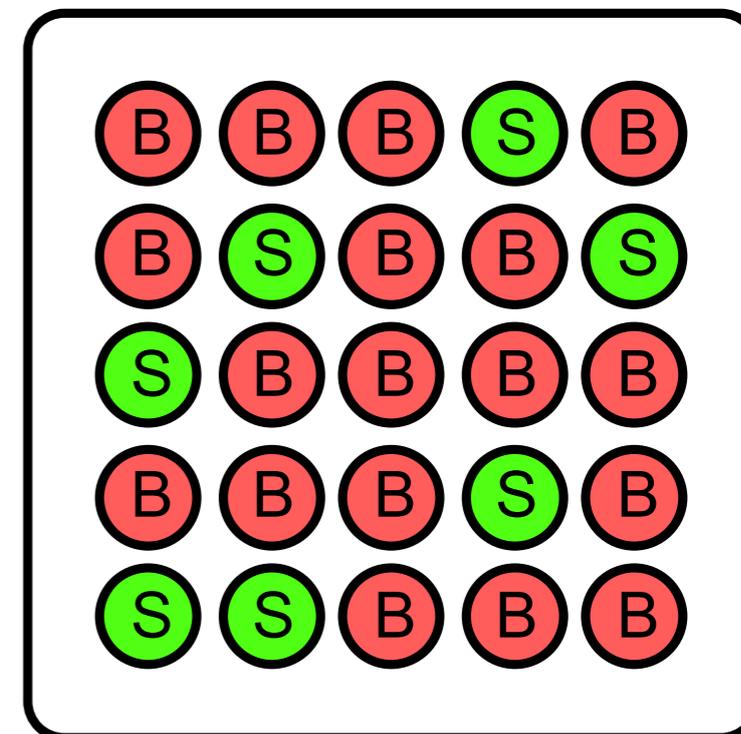
*Classification  
Without Labels*

Solution: Train  
**directly on data** using  
mixed samples

Mixed Sample 1



Mixed Sample 2



0

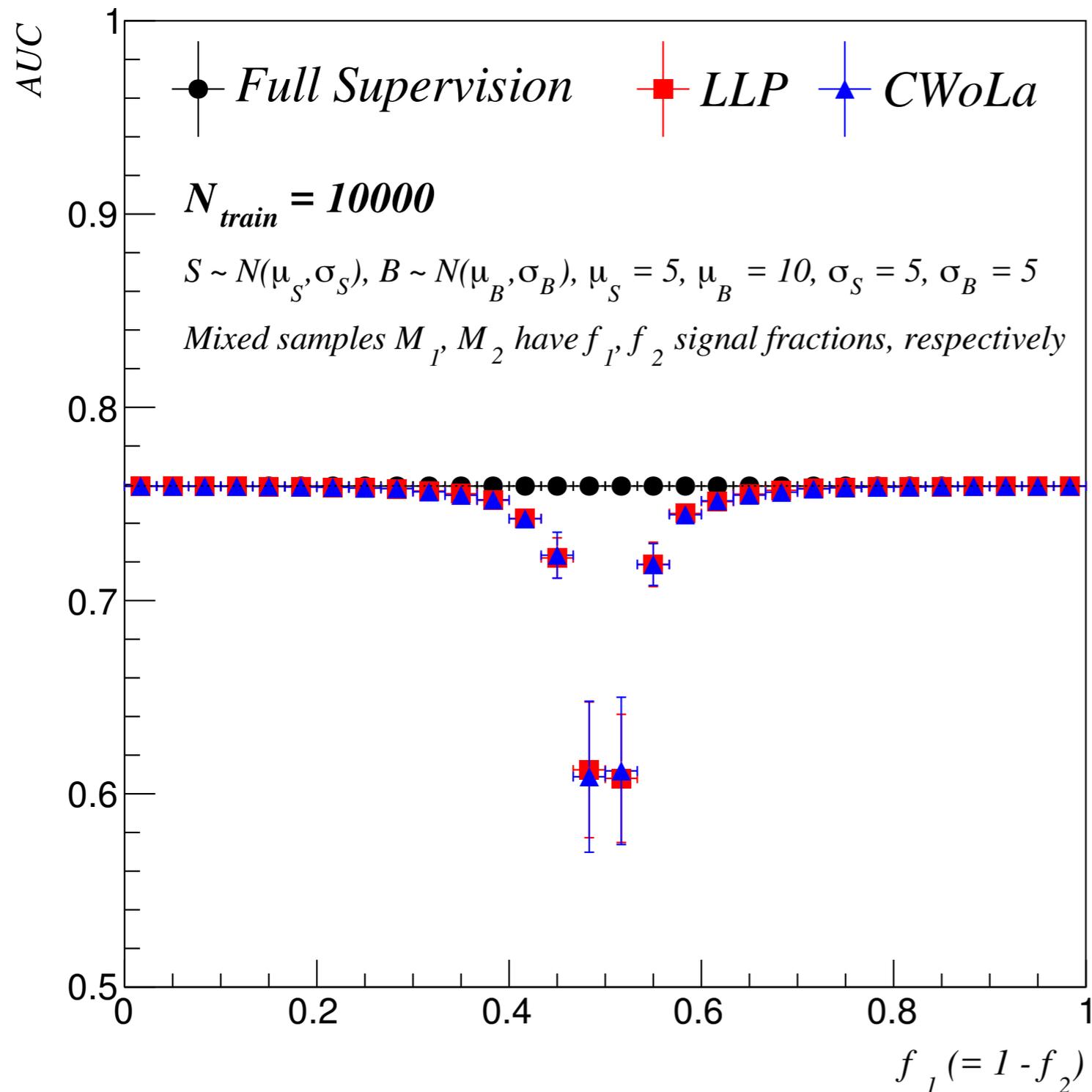
1

Classifier



# A note about training statistics

12



As with LLP, need sufficient effective statistics

Can't learn when the two proportions are the same.

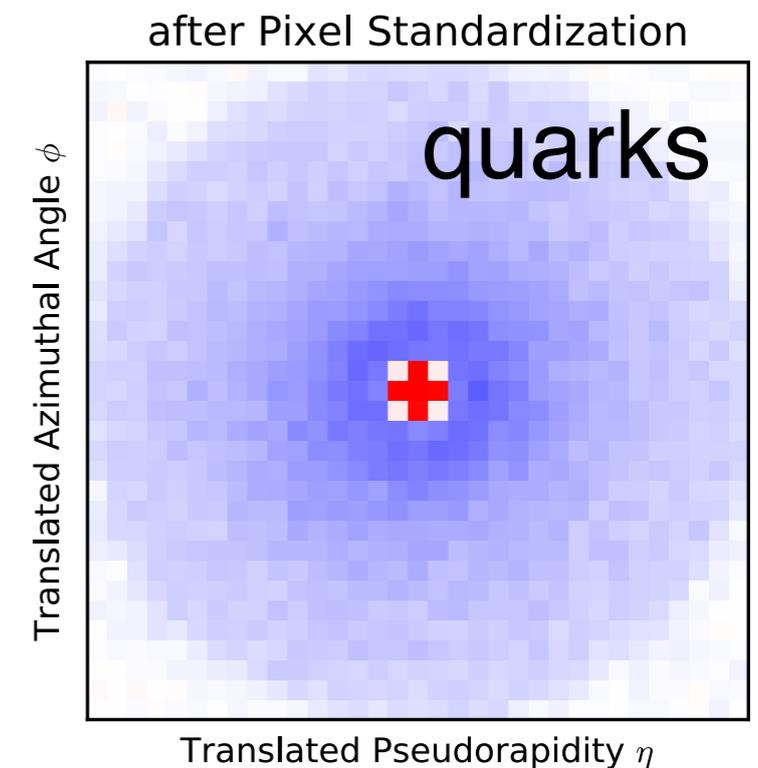
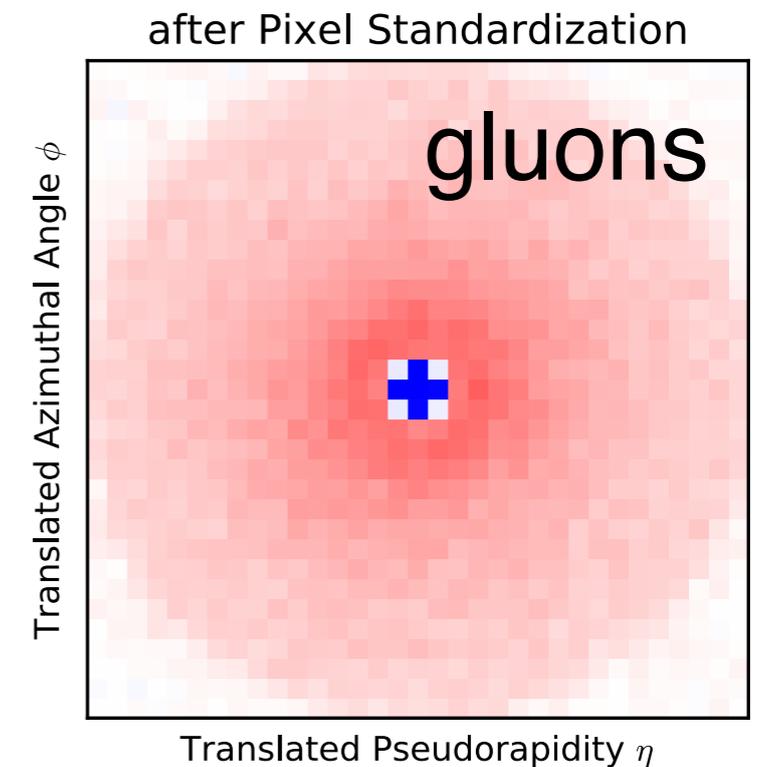
<b>Property</b>	 <b>LLP</b>	 <b>CWoLa</b>
Compatible with any trainable model	✓	✓
No training modifications needed	✗	✓
Training does not need fractions	✗	✓
Smooth limit to full supervision	✗	✓
Works for $> 2$ mixed samples	✓	?

# Next step: what about high dim.?

14

There are many 0(1)-dimensional ML problems for jets, but since the full radiation pattern is higher dimensional, need to go to bigger!

We'll use jet images as a testing ground, still focusing on quarks versus gluons.



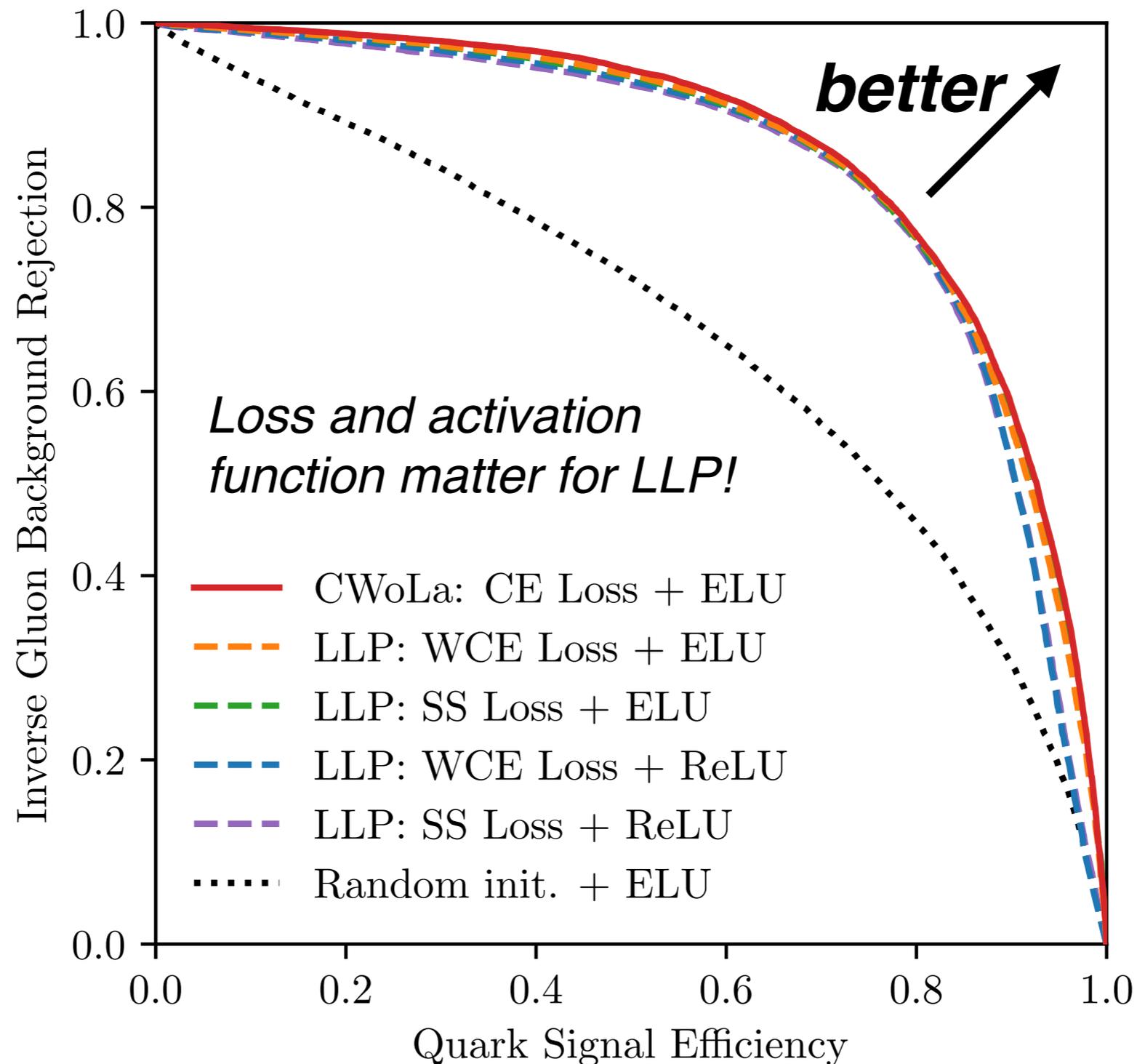
The CWoLa approach works out-of-the box - can use well-tested CNN architecture with usual cross-entropy loss.

On the other hand, LLP requires significant work on the technical implementation / optimization.

$$\ell_{\text{WMSE}} = \sum_a \left( f_a - \frac{1}{N} \sum_{i=1}^N h(\mathbf{x}_i) \right)^2 \quad \ell_{\text{WCE}} = \sum_a \text{CE} \left( f_a, \frac{1}{N} \sum_{i=1}^N h(\mathbf{x}_i) \right)$$

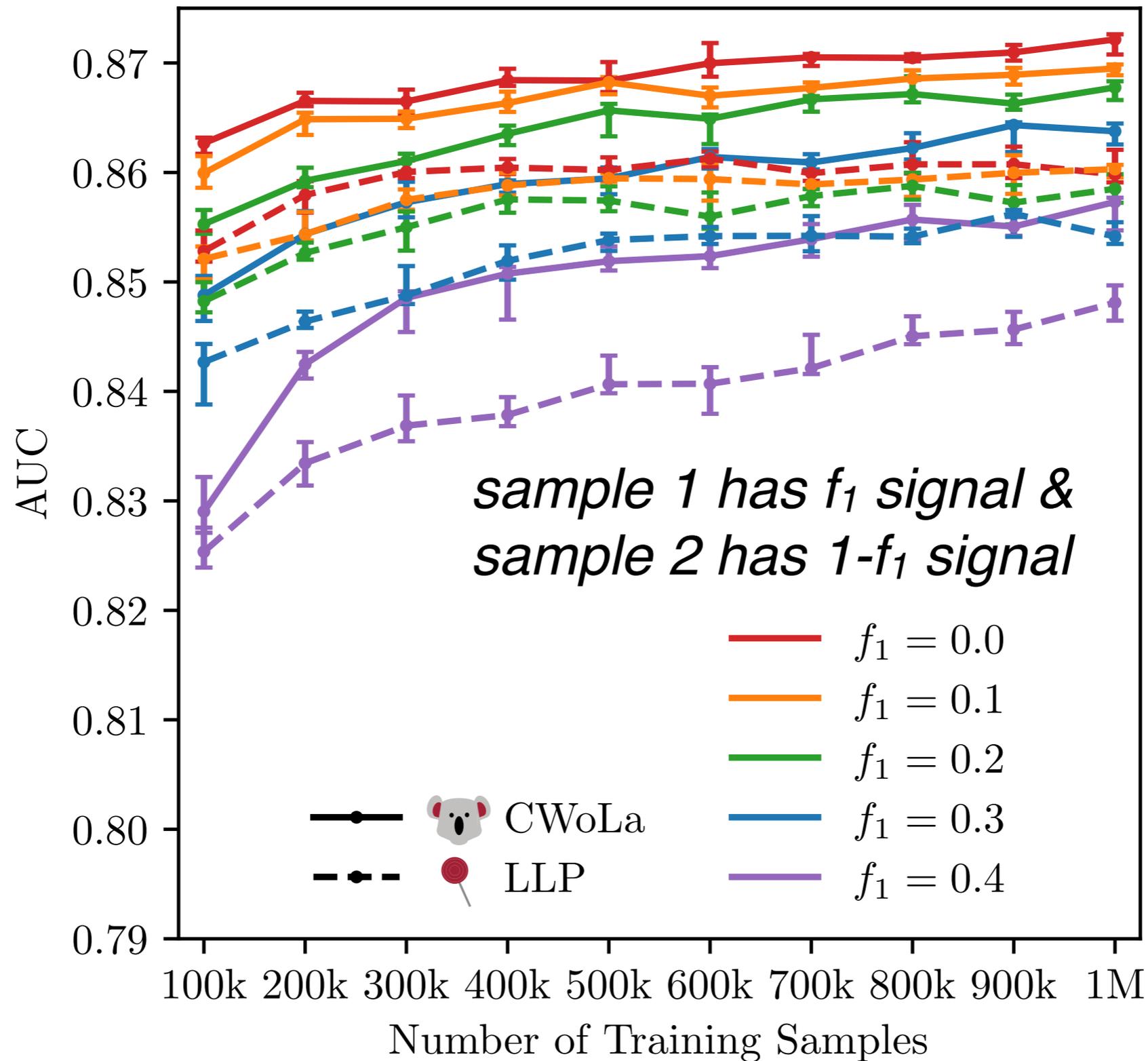
# Works in many-dimensions!

16



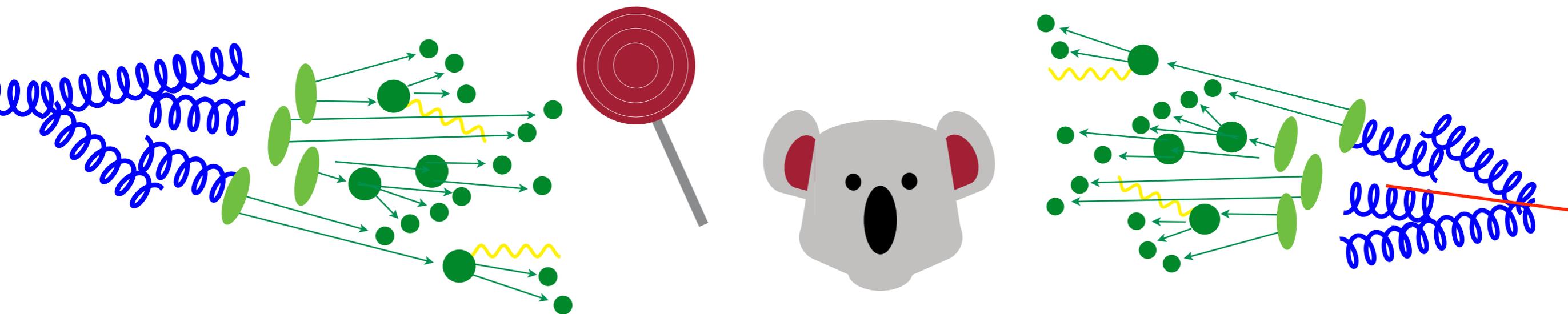
# A note about training statistics

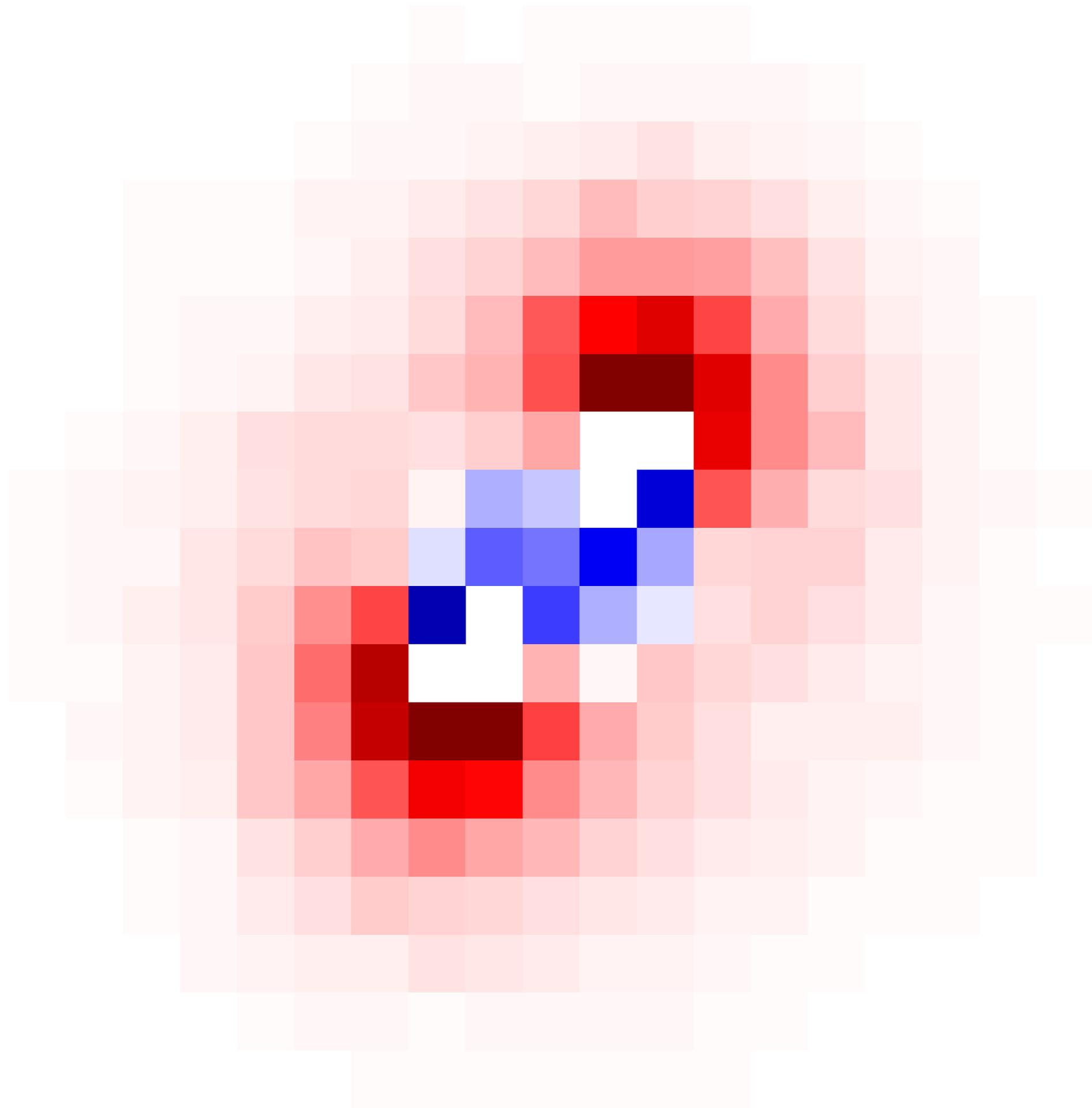
17



Weak supervision is a new & exiting paradigm for training classifiers. We can learn directly from nature!

This is particularly important for (VBF) jet physics, where there are concerns about mis-modeling. We have shown that the methods work even for high-dimensional data.





Fin.

# Backup: Topology Dependence

Learning	Sample	AUC
CWoLa	$Z + \text{jet vs. dijets}$	$0.8626 \pm 0.0020$
	Artificial $Z + q/g$	$0.8621 \pm 0.0019$
LLP	$Z + \text{jet vs. dijets}$	$0.8544 \pm 0.0019$
	Artificial $Z + q/g$	$0.8549 \pm 0.0018$

# Timing

