UNIVERSITY OF OREGON

# Simulation Independent Methods: Overview


bostdiek@uoregon.edu

# Talks from last year

| | |
|---|---|
| **Introduction and Overview (15'+5')** | *Marat Freytsis* |
| *2-100, Lawrence Berkeley National Laboratory* | 09:00 - 09:20 |
| **"Planing" to expose what the machine is learning (15'+5')** | *Bryan Ostdiek* |
| *2-100, Lawrence Berkeley National Laboratory* | 09:20 - 09:40 |
| **Weak Supervision in High Dimensions (15'+5')** | *Eric Metodiev* |
| *2-100, Lawrence Berkeley National Laboratory* | 09:40 - 10:00 |
| **Building an anti-QCD tagger (15'+5')** | *Jack Collins* |
| *2-100, Lawrence Berkeley National Laboratory* | 10:00 - 10:20 |
| **Adversarial Approaches (15'+5')** | *Kyle Cranmer* |
| *2-100, Lawrence Berkeley National Laboratory* | 10:20 - 10:40 |

Other systematics mitigation

Review weak supervision

Alternative approach to decorrelation

# Learning overview

Labeled data

Unlabeled data

Supervised Learning

- Classification
- Regression
- etc

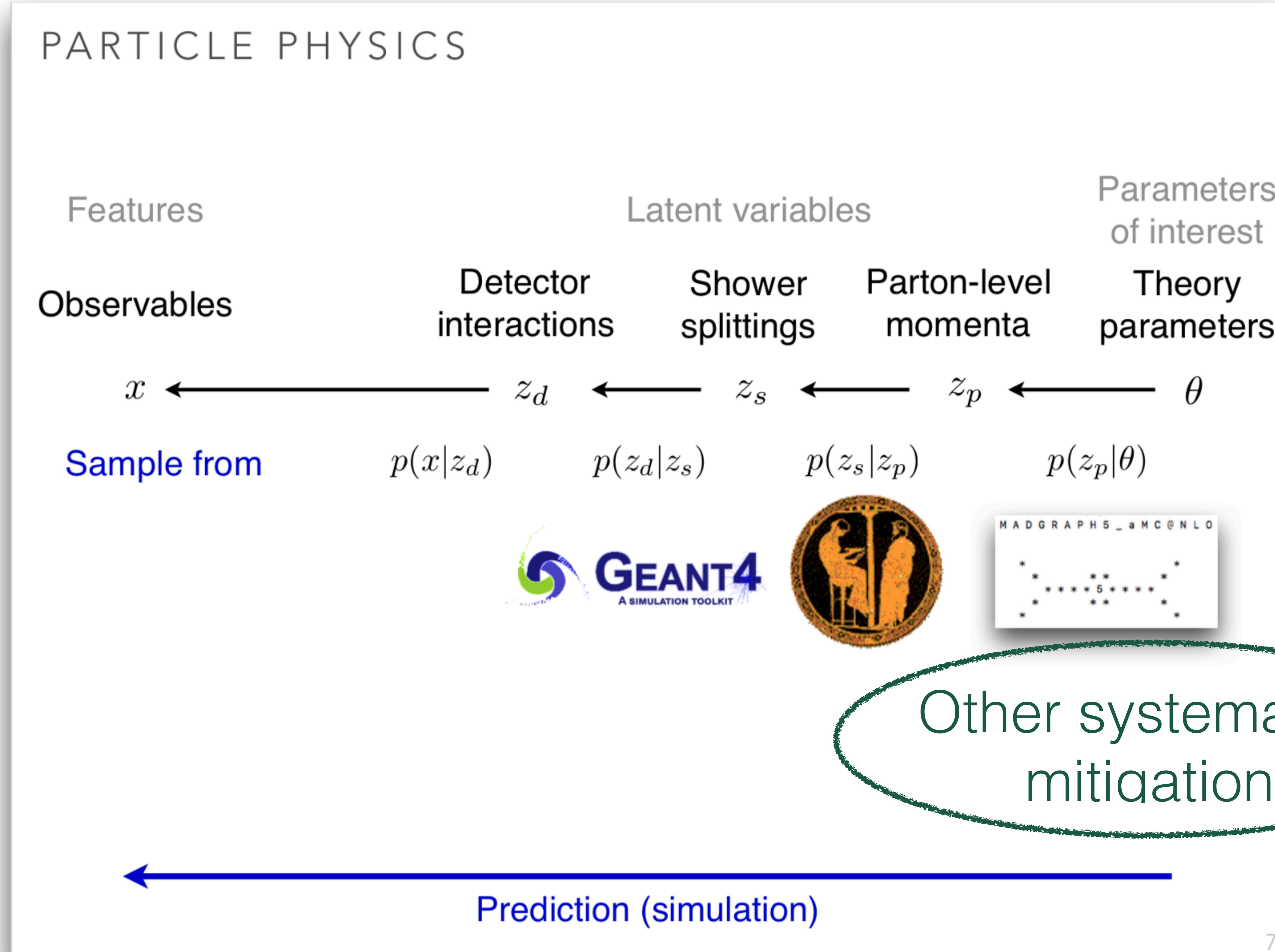
Unsupervised Learning

- Clustering
- Anomaly detection
- GAN
- etc

Hybrid?

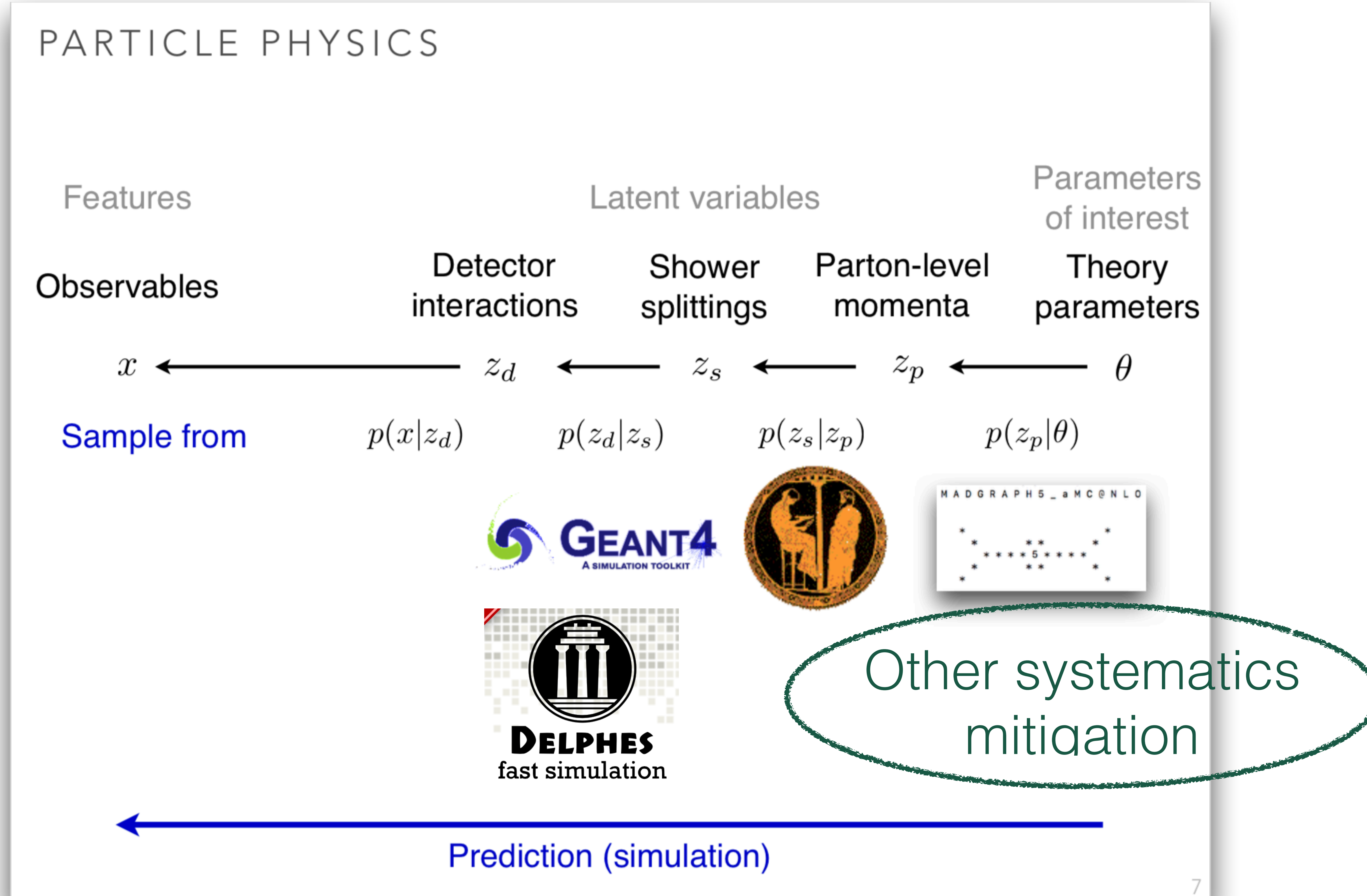
- Learning from label proportions
- Classification without labels

Weak supervision




Other systematics mitigation

Stolen from Kyle Cranmer




Stolen from Kyle Cranmer

# Information geometry

[1506.02169]
[1612.05261]
[1712.02350]

| Observables | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters |
| --- | --- | --- | --- | --- |

$$x \longleftarrow z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$$

Sample from $p(x|z_d)$ $p(z_d|z_s)$ $p(z_s|z_p)$ $p(z_p|\theta)$

GEANT4
A SIMULATION TOOLKIT

MADGRAPH5_aMC@NLO

DELPHES
fast simulation

Prediction (simulation)

7

Stolen from Kyle Cranmer

Information geometry

[1506.02169]
[1612.05261]
[1712.02350]

Observables | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters

Likelihood-free inference

[1805.00013]
[1805.00020]
[1805.12244]
[1808.00973]

DELPHES
fast simulation

Prediction (simulation)

7

Stolen from Kyle Cranmer

Information geometry

[1506.02169]
[1612.05261]
[1712.02350]

Observables | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters

Likelihood-free inference

[1805.00013]
[1805.00020]
[1805.12244]
[1808.00973]

Including uncertainties/ systematics in loss function

[1806.04743]
[1802.03537]

Stolen from Kyle Cranmer

# Weak supervision

# Weak supervision

"Weakly Supervised Classification in High Energy Physics," Dery, Nachman, Rubbo, and Schwartzman. [1702.00414 ]

LLP

"(Machine) Learning to Do More with Less," Cohen, Freytsis, and **BO.** [1706.09451]

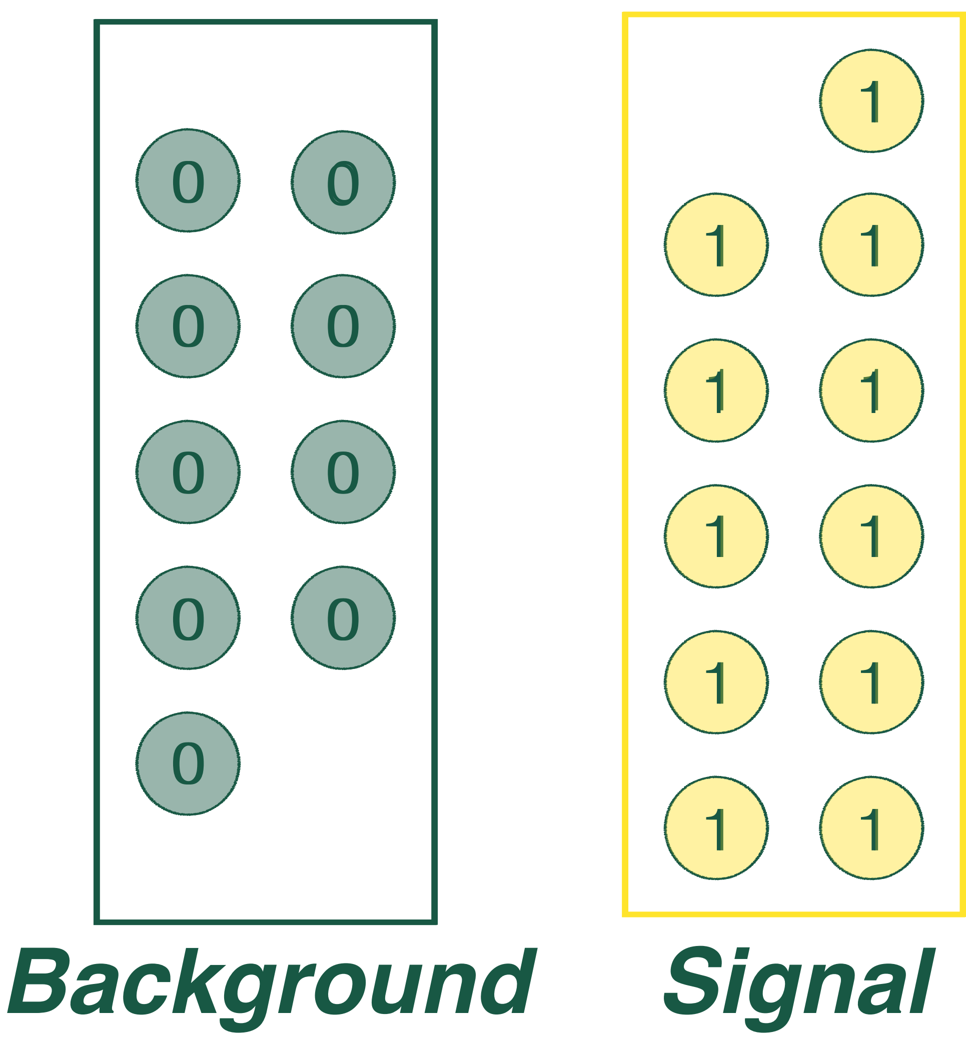
CWoLa

"Classification without labels: Learning from mixed samples in high energy physics," Metodiev, Nachman, and Thaler. [1708.02949]

"Learning to Classify from Impure Samples," Komiske, Metodiev, Nachman, and Schwartz. [1801.10158]

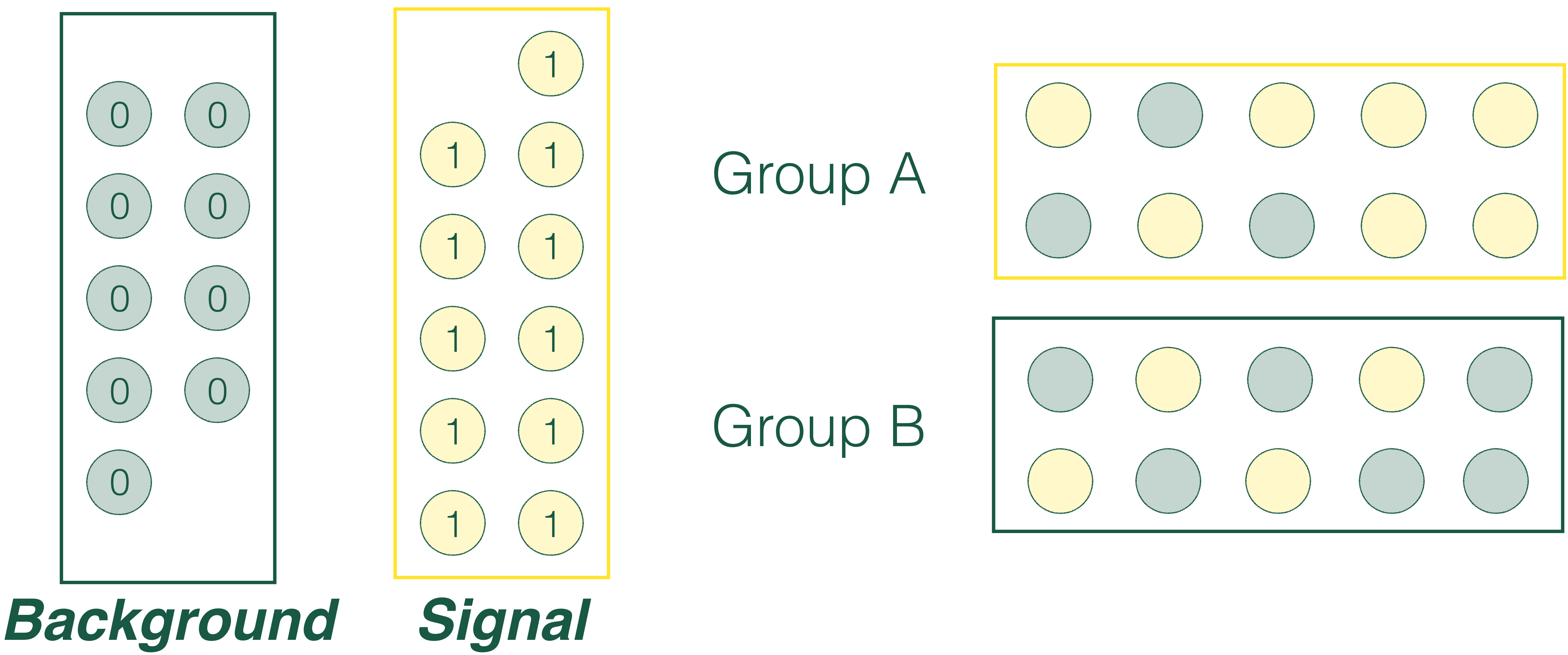
# Problem: (How) can we make a classifier without event-by-event truth-level labels

Fully supervised

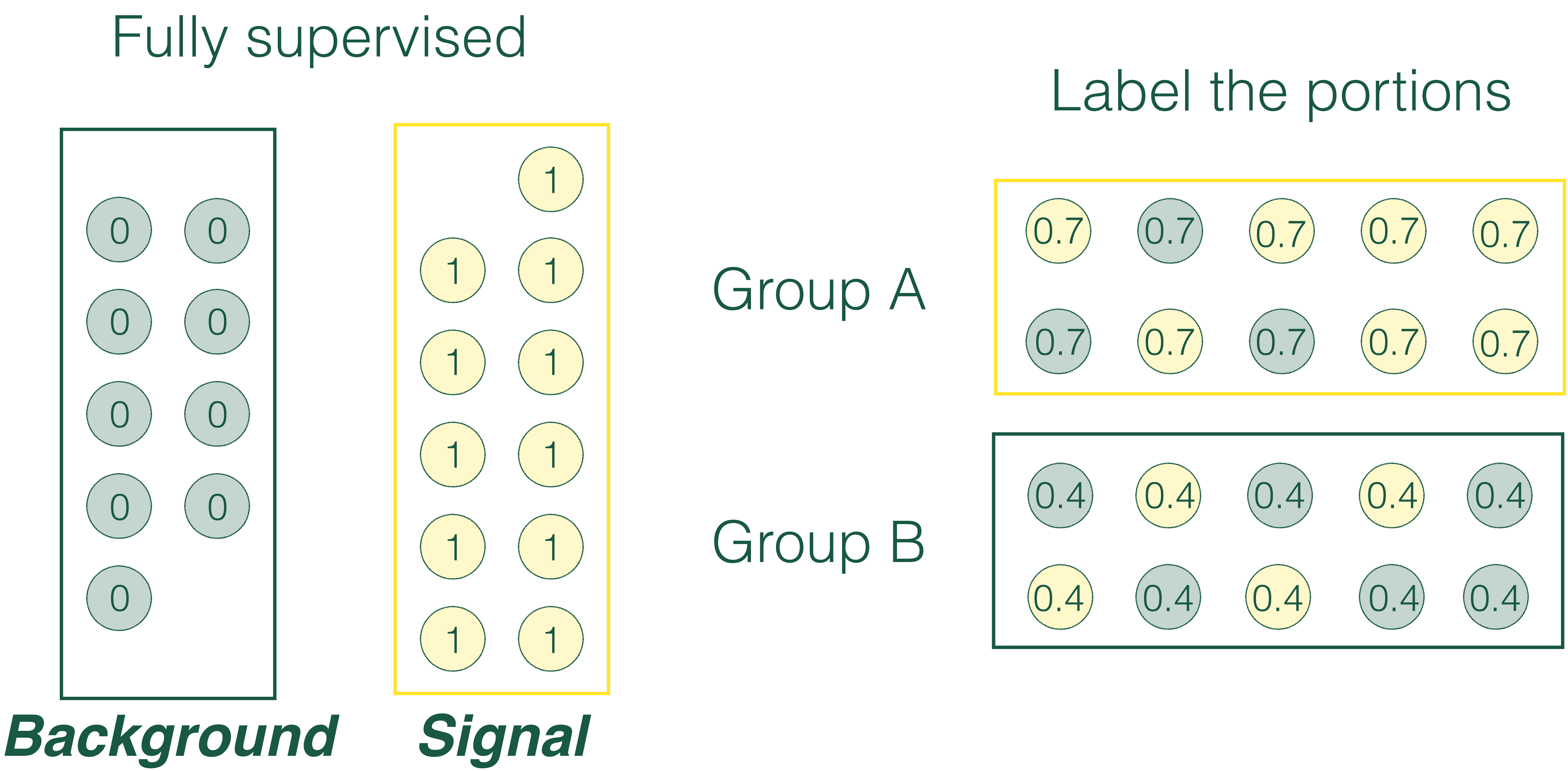


# Problem: (How) can we make a classifier without event-by-event truth-level labels

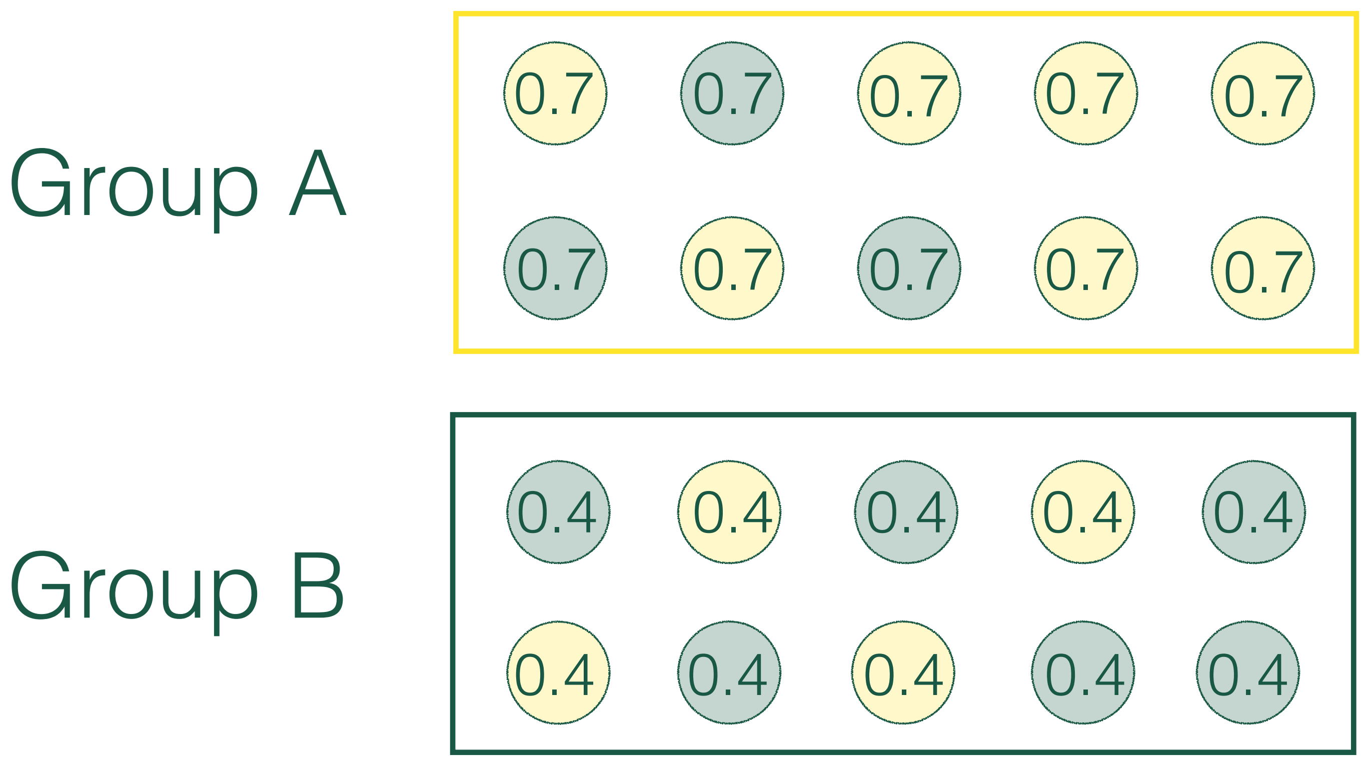
Fully supervised

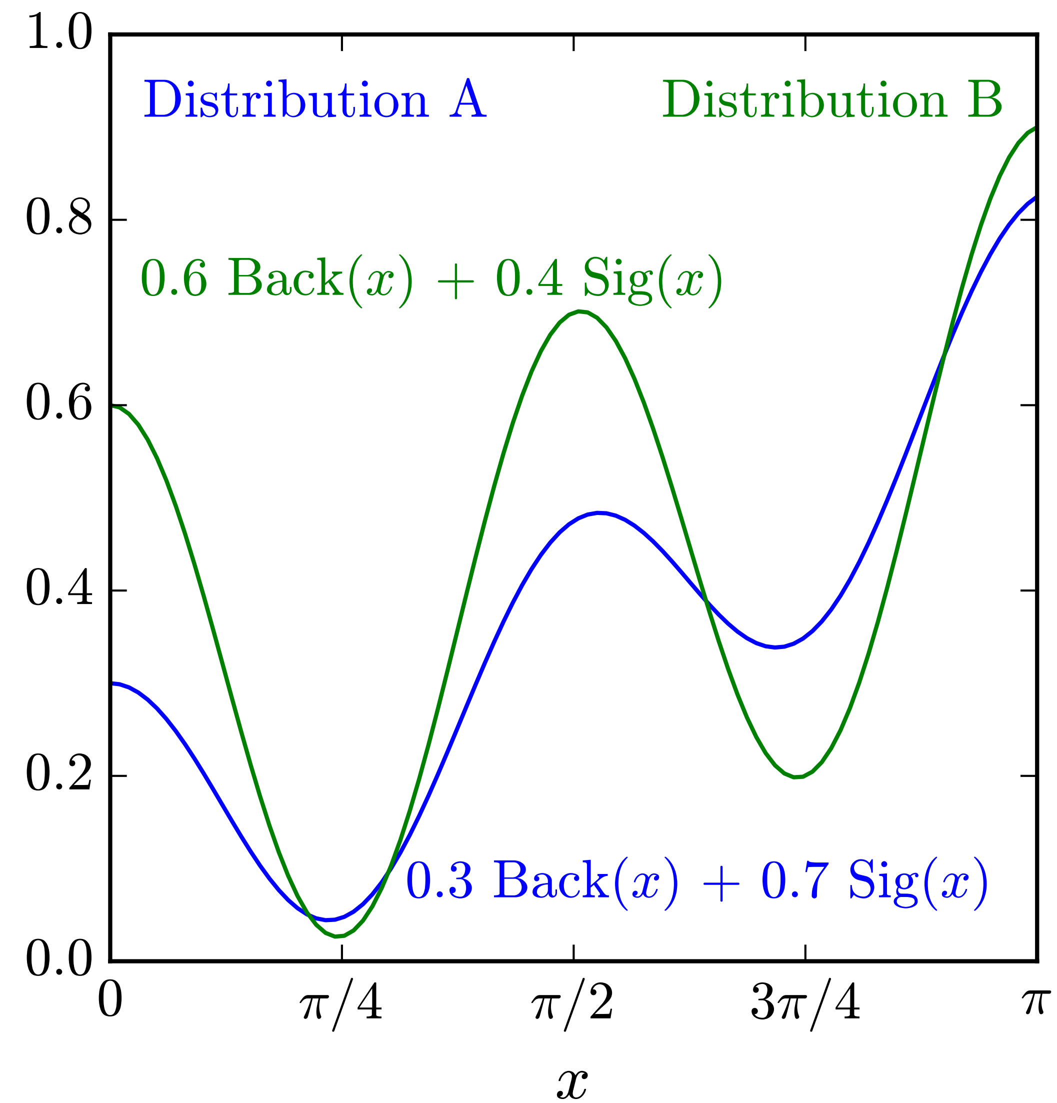


# Problem: (How) can we make a classifier without event-by-event truth-level labels



# Problem: (How) can we make a classifier without event-by-event truth-level labels

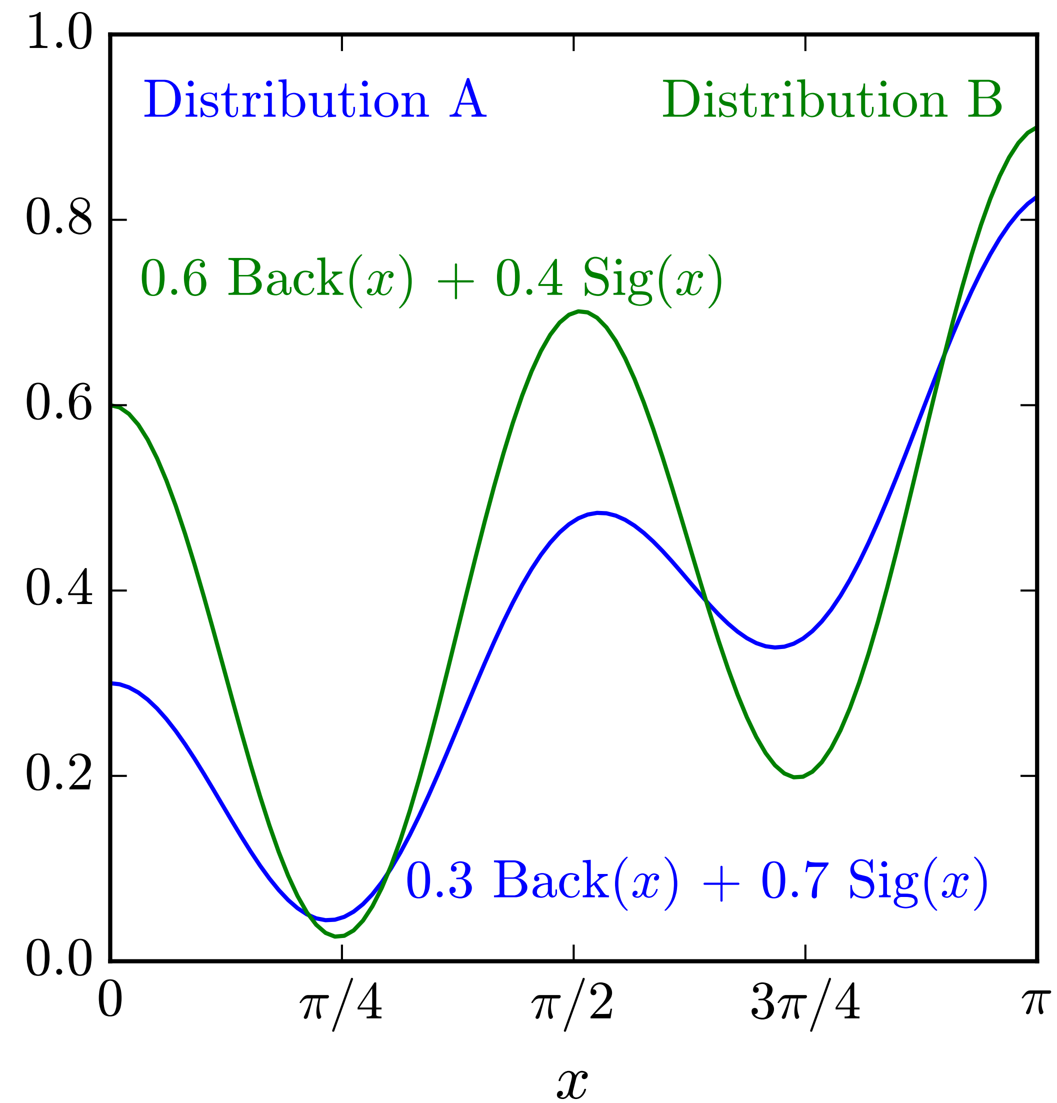


# Problem: (How) can we make a classifier without event-by-event truth-level labels

Group A

Group B

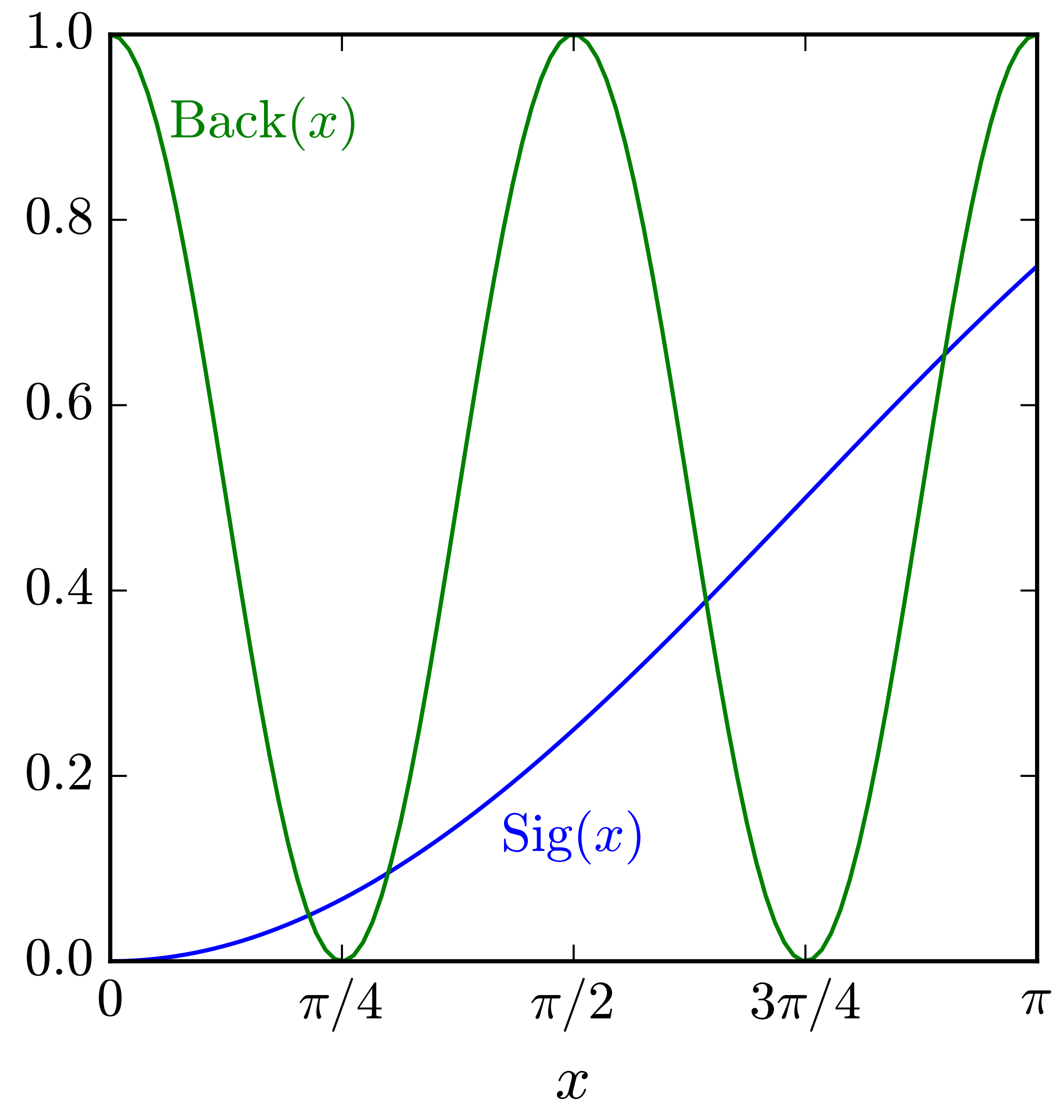

$$\mathrm{Sig}(x) = 2A(x) - B(x)$$

$$\mathrm{Back}(x) = \frac{1}{3}\left(-4A(x) - 7B(x)\right)$$

# Problem: (How) can we make a classifier without event-by-event truth-level labels

Group A


Group B


$$\mathrm{Sig}(x) = 2A(x) - B(x)$$

$$\mathrm{Back}(x) = \frac{1}{3}\left(-4A(x) - 7B(x)\right)$$

# Problem: (How) can we make a classifier without event-by-event truth-level labels

Group A

Group B


Make a histogram of the multi-dimensional data

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

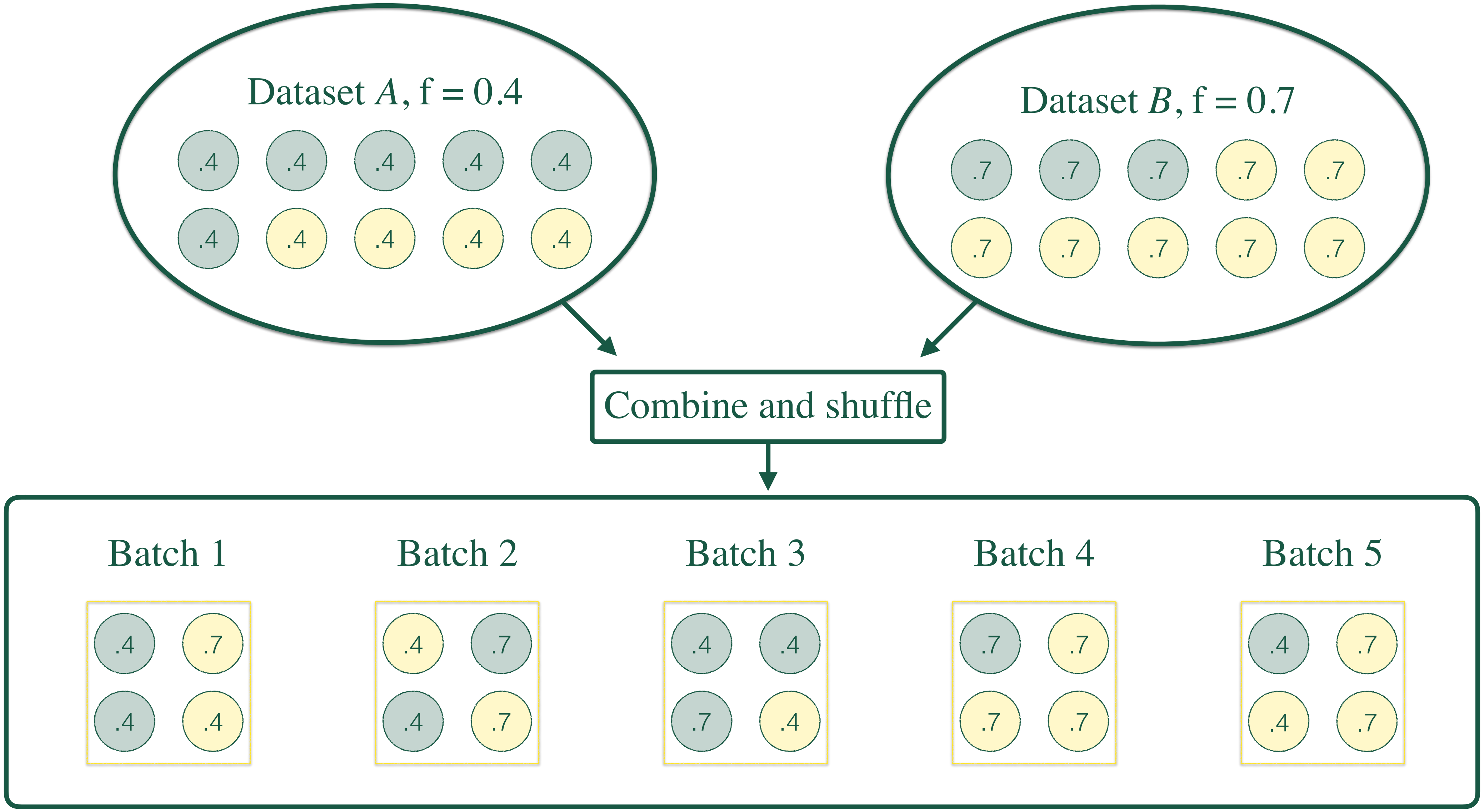
Invert

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

Machine learning helps with:

- Large dimensionality
- Over-constrained (more groups)
- Finite statistics

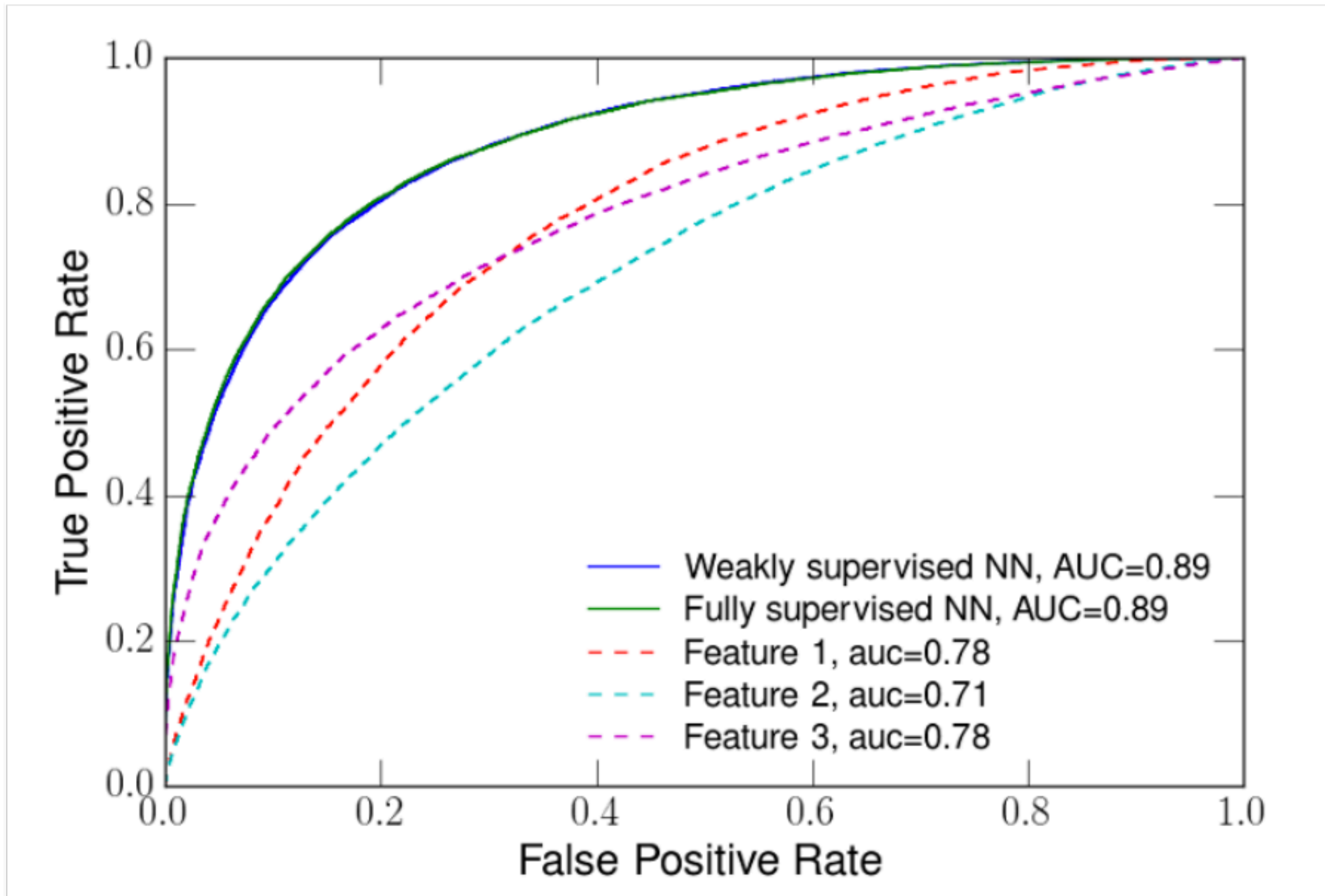
# Weak supervision - LLP




$$\ell_{LLP} = \sum_{batches} |\langle f_{t,i}\rangle - \langle y_{p,i}\rangle|$$

L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, JHEP **1705**, 145 (2017) doi:10.1007/JHEP05(2017)145 [arXiv:1702.00414 [hep-ph]]

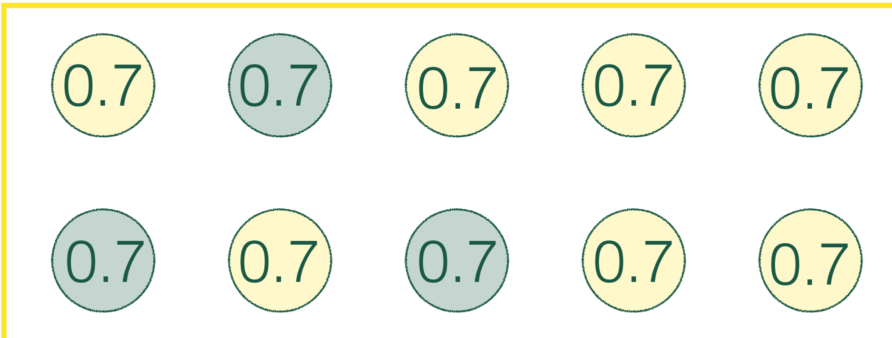
# Weak supervision - LLP




$$\ell_{LLP} = \sum_{batches} |\langle f_{t,i} \rangle - \langle y_{p,i} \rangle|$$

L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, JHEP **1705**, 145 (2017) doi:10.1007/JHEP05(2017)145 [arXiv:1702.00414 [hep-ph]]

# Weak supervision





**What if there are uncertainties on the ratio?**


Cohen, Freytsis, BO [1706.09451]

Label errors don't affect classifier

Metodiev, Nachman, Thaler [1708.02949]

Possible to do classification with arbitrary labels

# Weak supervision

**Theorem 1** *Given mixed samples $M_1$ and $M_2$ defined in terms of pure samples $S$ and $B$ 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 $B$.*

Metodiev, Nachman, and Thaler [arXiv:1708.02949]

# Weak supervision

**Theorem 1** *Given mixed samples $M_1$ and $M_2$ defined in terms of pure samples $S$ and $B$ 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 $B$.*

Metodiev, Nachman, and Thaler [arXiv:1708.02949]

*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:

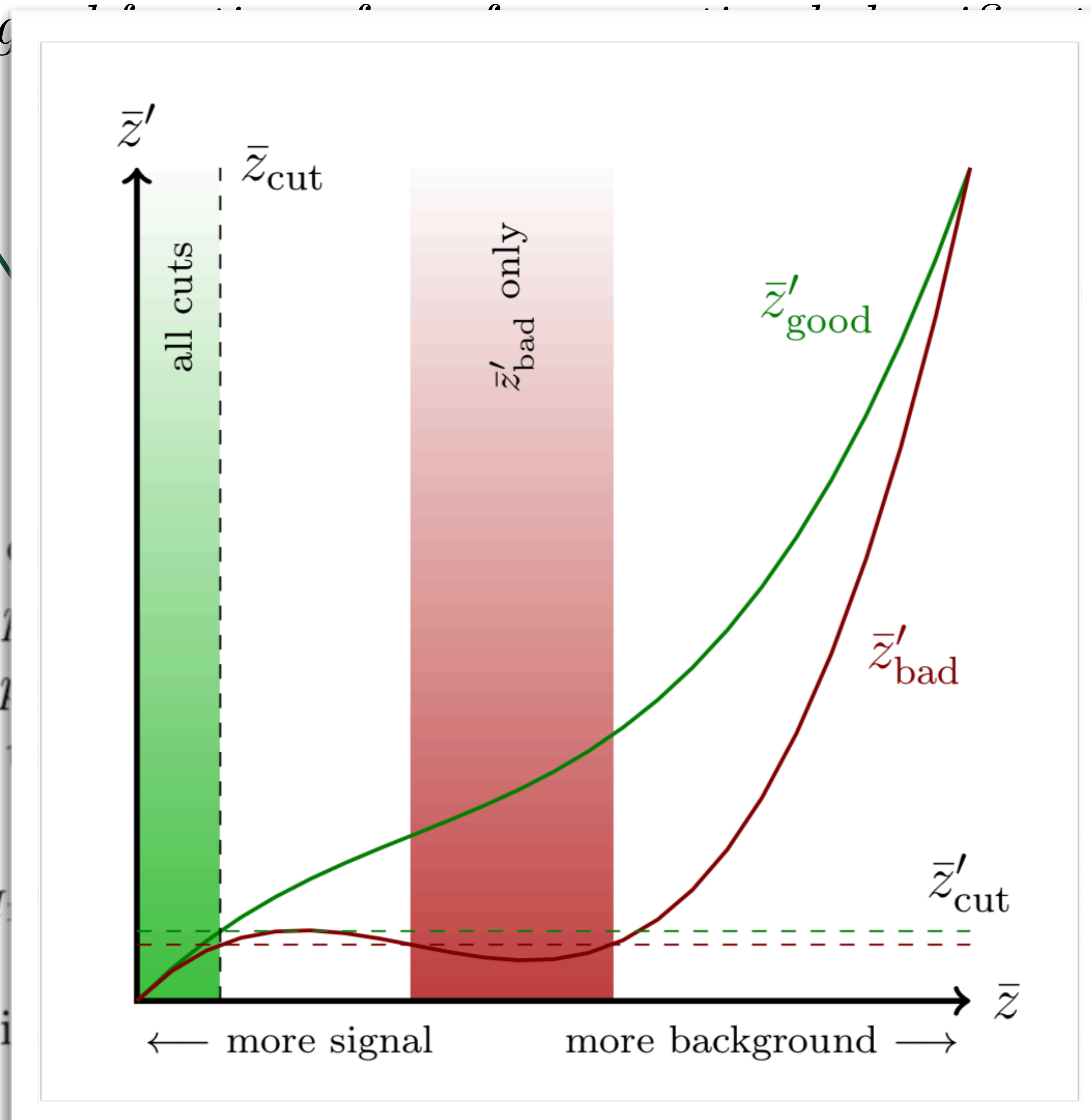
$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)},$$

which is a monotonically increasing rescaling of the likelihood $L_{S/B}$ as long as $f_1 > f_2$, since $\partial_{L_{S/B}} L_{M_1/M_2} = (f_1 - f_2)/(f_2 L_{S/B} - f_2 + 1)^2 > 0$. If $f_1 < f_2$, then one obtains the reversed classifier. Therefore, $L_{S/B}$ and $L_{M_1/M_2}$ define the same classifier. $\square$

# Weak supervision

**Theorem 1** *Given mixed samples $M_1$ and $M_2$ defined in terms of pure samples $S$ and $B$ with sig… …ained to distinguish $M_1$ from $M_2$ is*

## Metodiev, N… …:1708.02949]

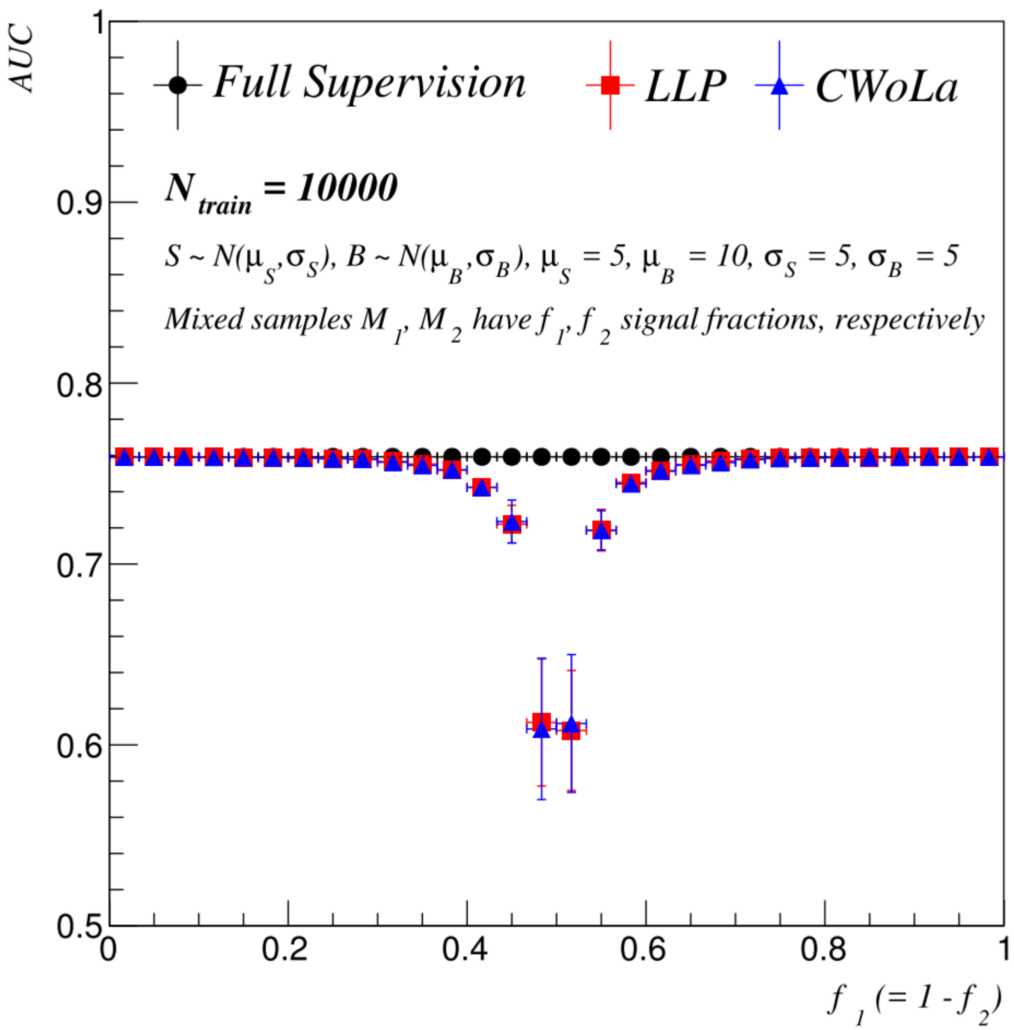



*Proof.* The optimal … …nd $p_{M_2}$ is the likelihood ratio $L_{M_1/M_2}(\vec{x}) =$ … …o distinguish examples drawn from $p_S$ and … …Where $p_B$ has support, we can relate these …

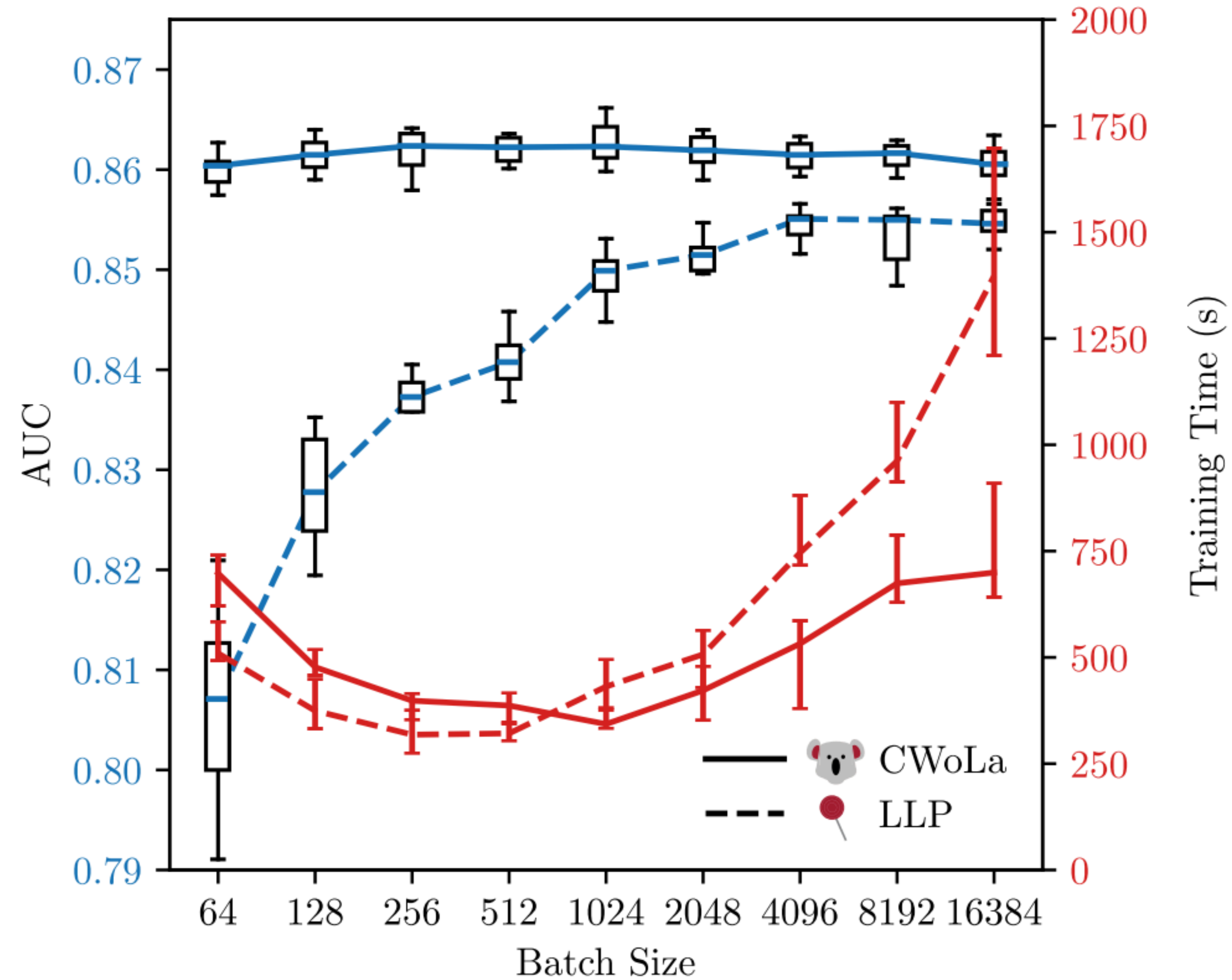
$$L_{M} \ldots \frac{\ldots - f_1)}{\ldots - f_2)},$$

which is a monotoni… …s long as $f_1 > f_2$, since $\partial_{L_{S/B}} L_{M_1/M_2} = (f_1$ … …e obtains the reversed classifier. Therefore, $L_{S/B}$ and $L_{M_1/M_2}$ define the same classifier. □

# Weak supervision



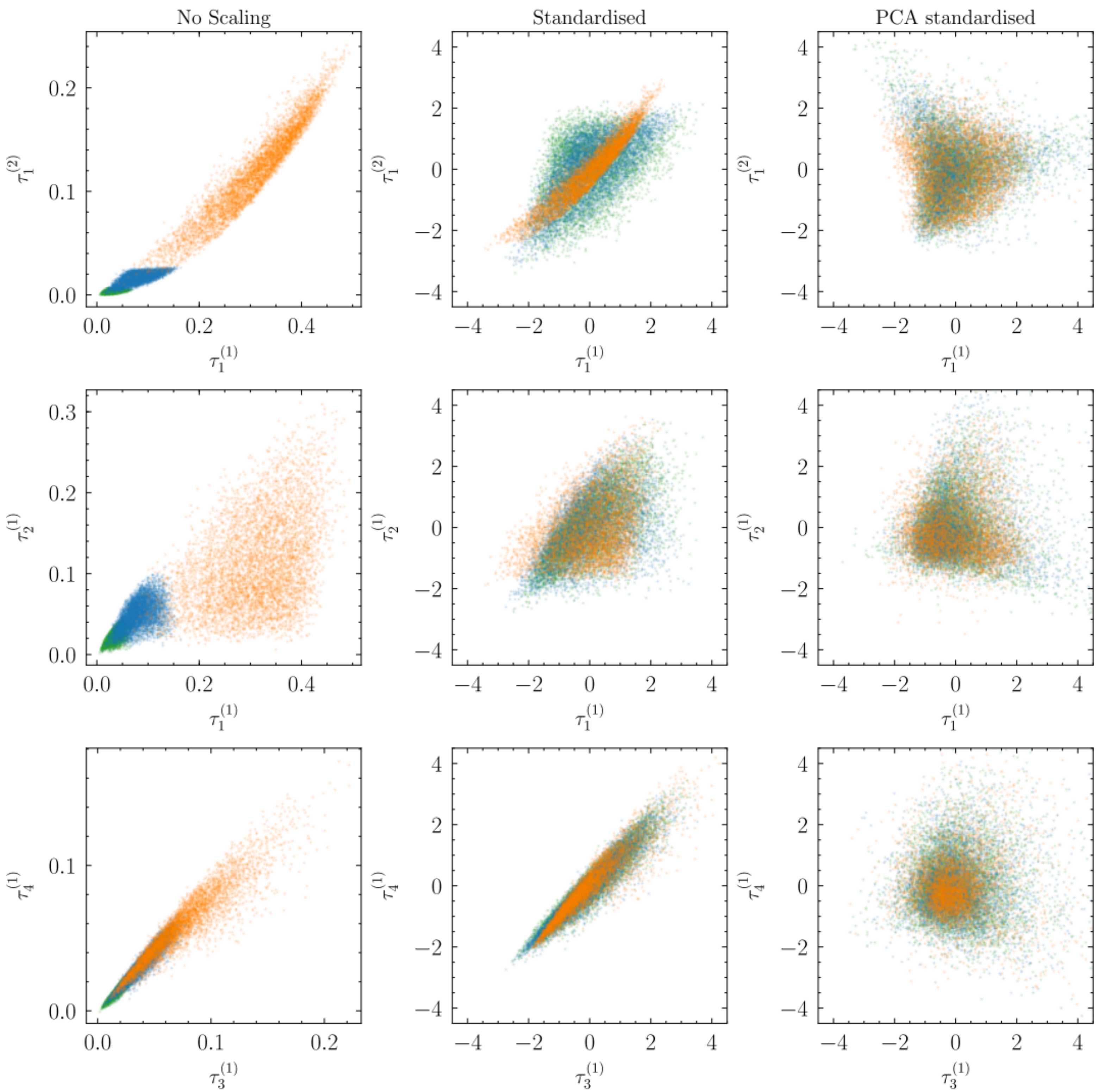

Metodiev, Nachman, and Thaler. [1708.02949]




Komiske, Metodiev, Nachman, and Schwartz. [1801.10158]

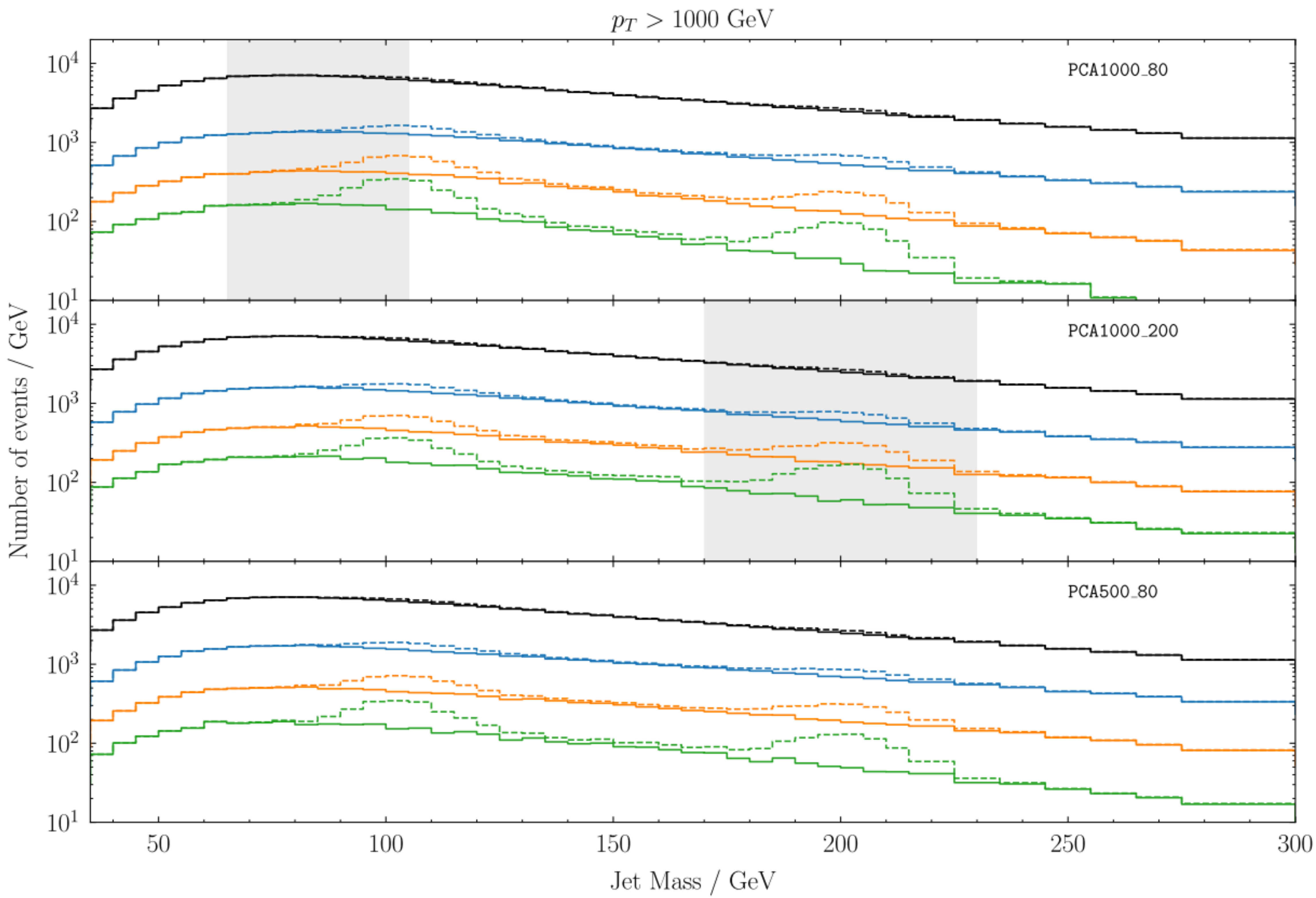
# Alternative approach to decorrelation

# Alternative approach to decorrelation




Aguilar-Saavedra, Collins, and Mishra. [1709.01087]

# Alternative approach to decorrelation




Aguilar-Saavedra, Collins, and Mishra. [1709.01087]

| | | |
|---|---|---|
| 10:00 | **Introduction and overview (20'+10')** *One West (WH1W), Fermilab* | *Bryan Ostdiek* 10:00 - 10:30 |
| | **Disentangling Jet Categories at Colliders (20'+5')** *One West (WH1W), Fermilab* | *Eric Metodiev* 10:35 - 11:00 |
| 11:00 | **JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics (20'+5')** *One West (WH1W), Fermilab* | *Anders Andreassen* 11:05 - 11:30 |
| 12:00 | **Lunch** *One West (WH1W), Fermilab* | 11:30 - 13:00 |
| 13:00 | **CWoLa Hunting: Enhancing the Bump Hunt with Machine Learning (20'+5')** *One West (WH1W), Fermilab* | *Jack Collins* 13:00 - 13:25 |
| | **QCD or What: Deep autoencoder based searches for new physics (20'+5')** *One West (WH1W), Fermilab* | *Gregor Kasieczka* 13:30 - 13:55 |
| 14:00 | **Searching for New Physics with Autoencoders (20'+5')** *One West (WH1W), Fermilab* | *Marco Farina* 14:00 - 14:25 |
| | **Novelty Detection Meets Collider Physics (20'+5')** *One West (WH1W), Fermilab* | *Tao Liu* 14:30 - 14:55 |

https://indico.cern.ch/event/745718/timetable/#20181116

## Open questions, concrete & speculative

### Marat's overview last year

- performance for multi-component classification?
  - does CWoLa even have a multi-component generalization?
- how do the optimality arguments change at finite statistics?
- can we propagate uncertainties on inputs through the network?
  - would this be useful?
- can we invert any of these result to see what our models get wrong
- can we go even weaker?
  - *e.g.*, Hopfield networks and generalizations
  - can solve certain classification tasks unsupervised
  - some use in astrophysics, nearly no collider proposals to date
- ...

- What can theorists do to help weak supervision get implemented in the experiments?
- Are there easier ways to de-correlate than adversarial training?
- If taggers get nearly identical ROC curves, but some de-correlate well, and others do not, can we learn physics from that?