Learning through weak supervision

Bryan Ostdiek
Machine learning for phenomenology workshop. IPPP, Durham
April 5, 2018

Labeled data

Unlabeled data

Supervised Learning

- Classification
- Numerical Predictions
- etc

Unsupervised Learning

- Clustering
- Anomaly Detection
- GAN
- etc

Hybrid?

- Learning from label proportions
- Classification without labels

Weak supervision

References

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

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

"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]

References

"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] CWoLa

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

Outline

1. Introduction / Toy model

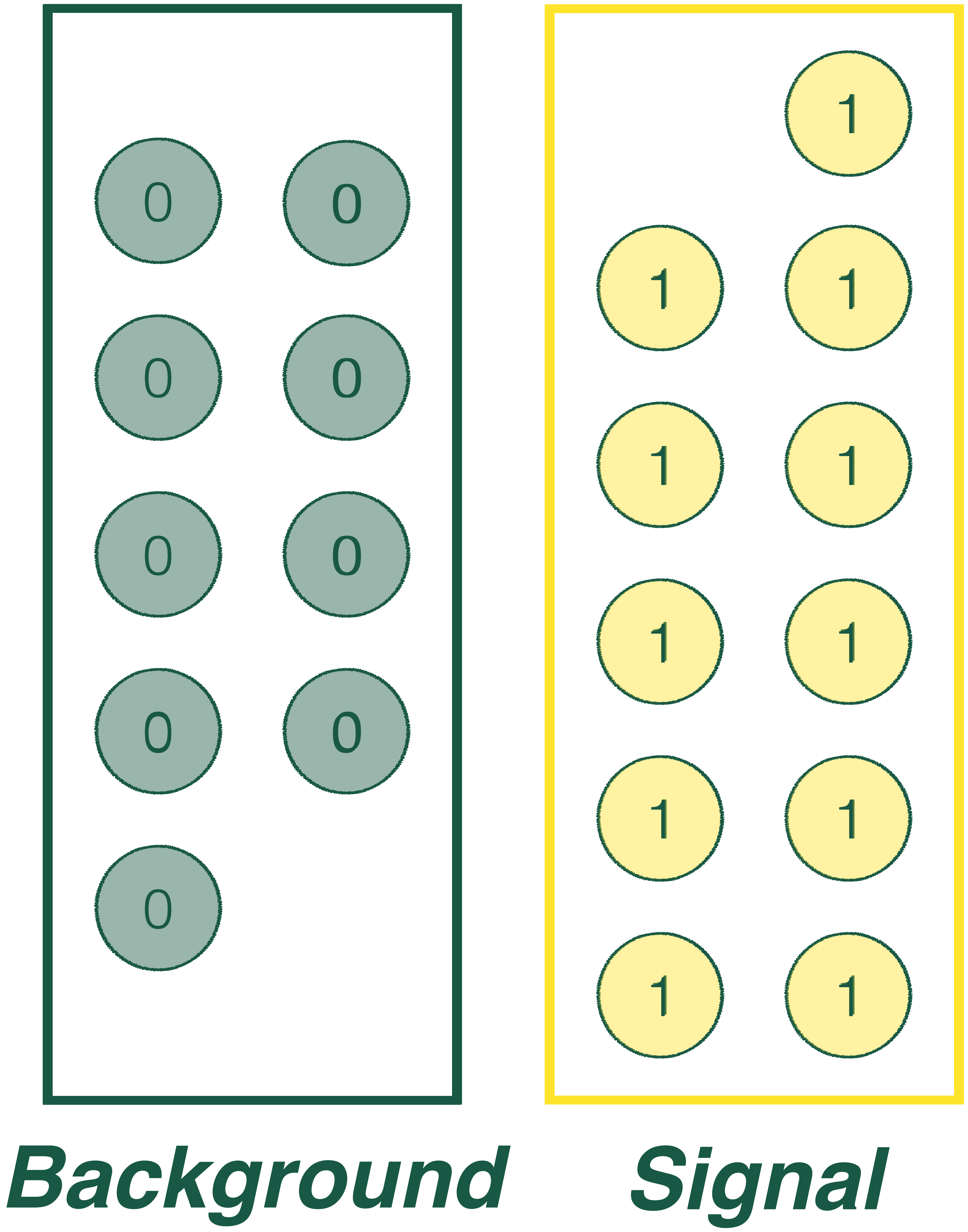
- What is weak supervision?
- How can it work?
- Is it robust?

2. LHC Scenario

- Higher input dimension
- Application to unseen data
- Affects of mis-modeling
- Combination of Full and Weak

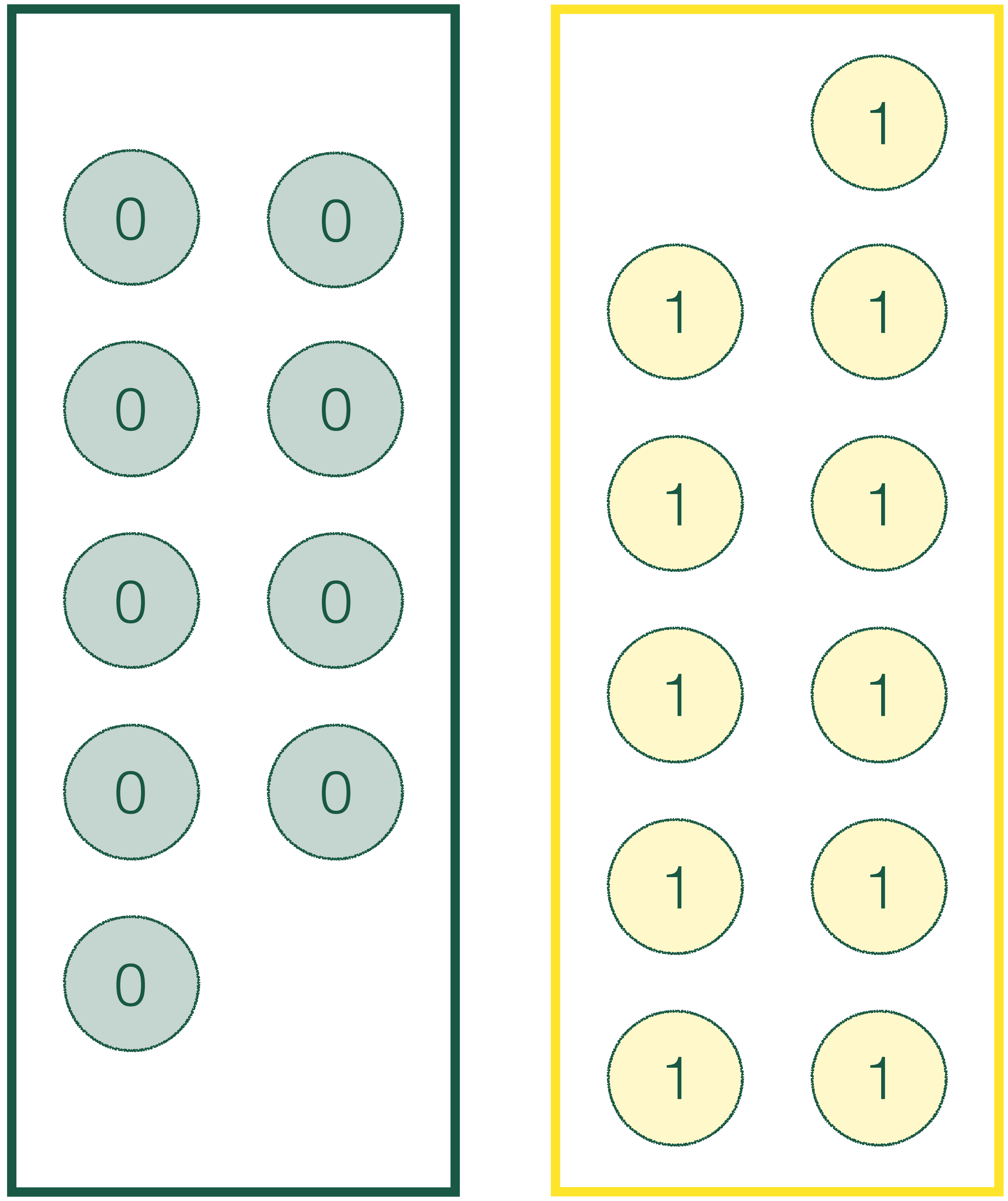
Problem: (How) can we make a classifier without event-by-event truth-level labels

Fully supervised



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

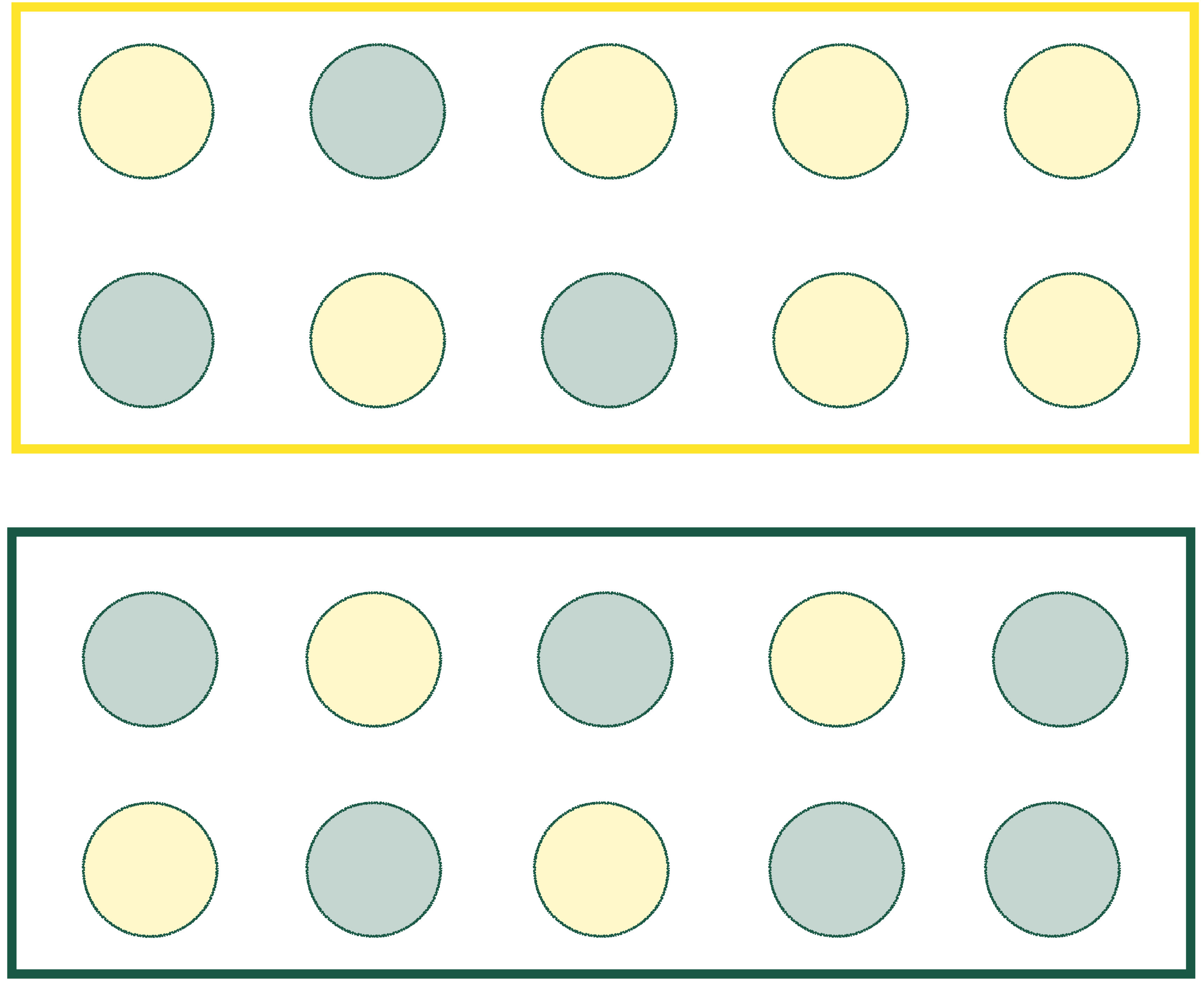
Fully supervised



Background ***Signal***

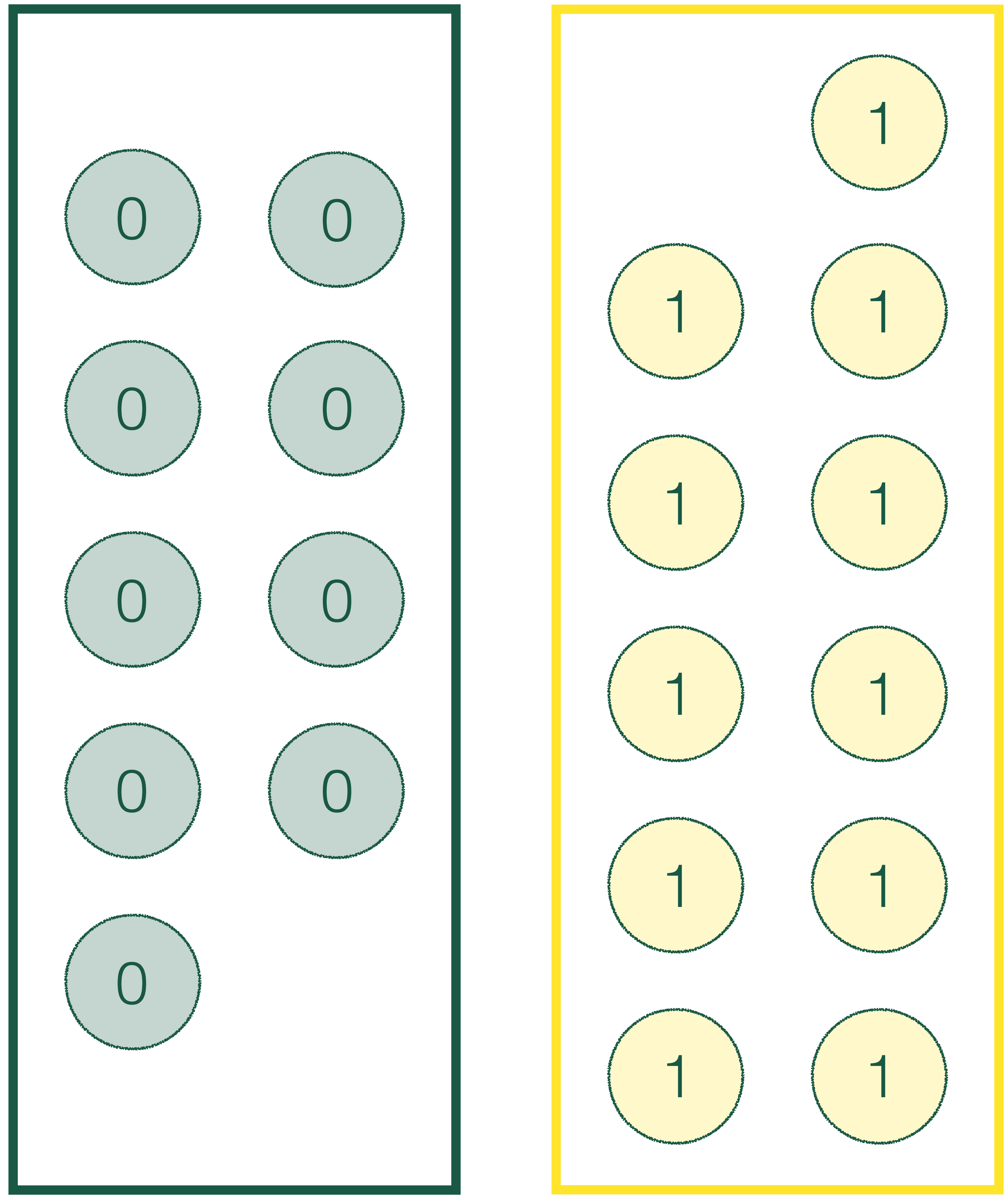
Group A

Group B



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

Fully supervised



Background ***Signal***

Label the portions

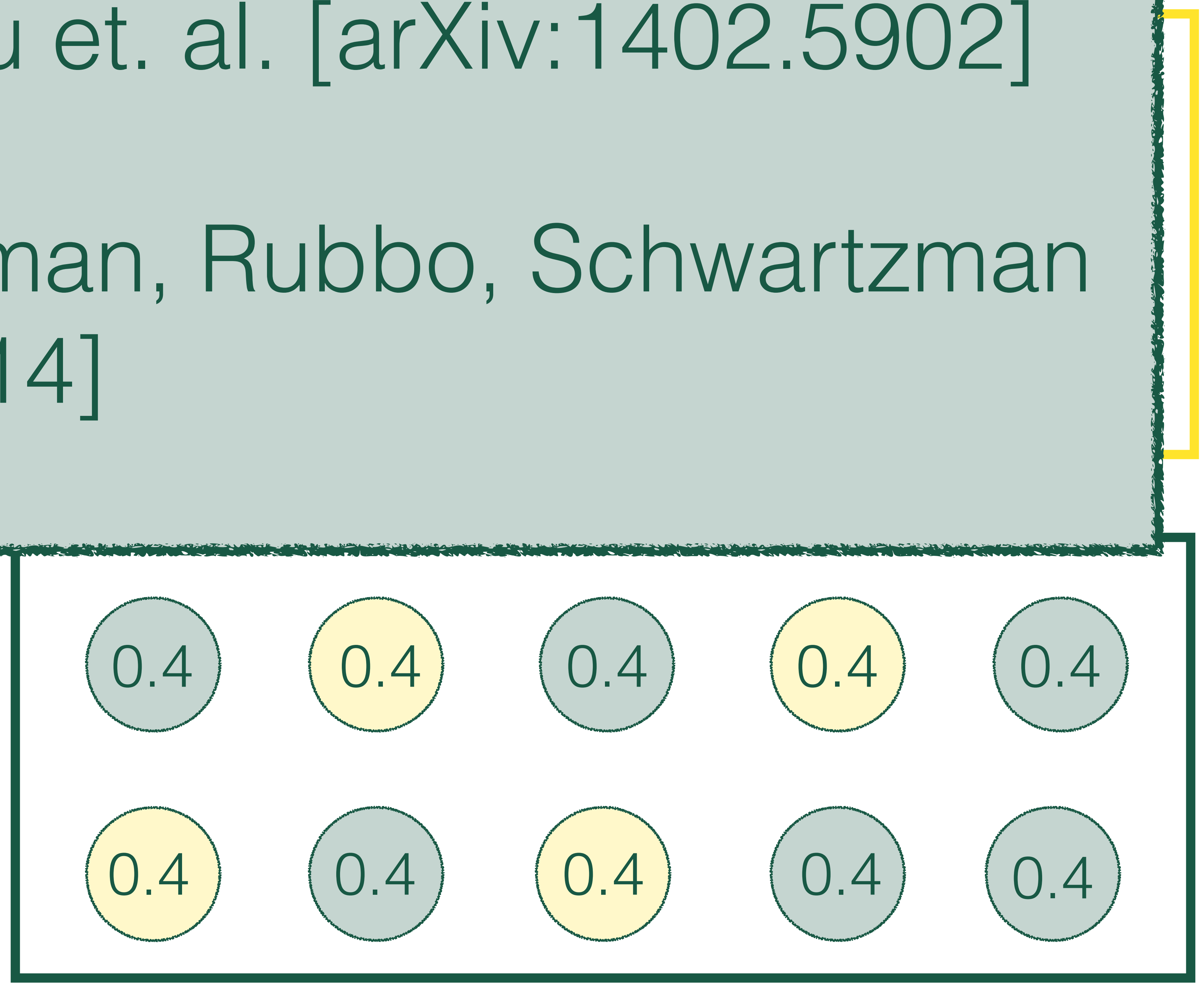
Problem: (How) can we make a classifier without event-by-event truth-level labels

In CS lit, mixed training sets are called “bags” and method is Learning from Label Proportions (LLP), see e.g. Dietterich, Lathrop, Lozano-Prez [1997]; Amores [2013]; Yu et. al. [arXiv:1402.5902]

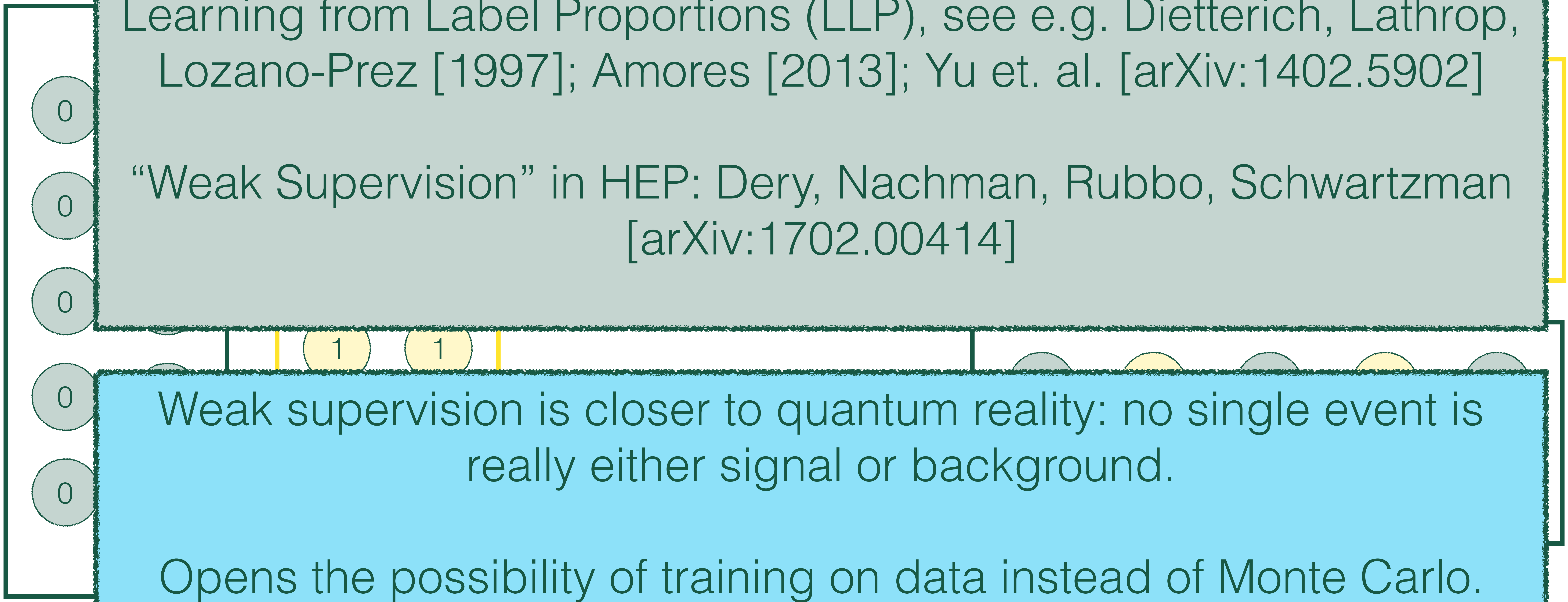
“Weak Supervision” in HEP: Dery, Nachman, Rubbo, Schwartzman [arXiv:1702.00414]

Background ***Signal***

Group B



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



In CS lit, mixed training sets are called “bags” and method is Learning from Label Proportions (LLP), see e.g. Dietterich, Lathrop, Lozano-Prez [1997]; Amores [2013]; Yu et. al. [arXiv:1402.5902]

“Weak Supervision” in HEP: Dery, Nachman, Rubbo, Schwartzman [arXiv:1702.00414]

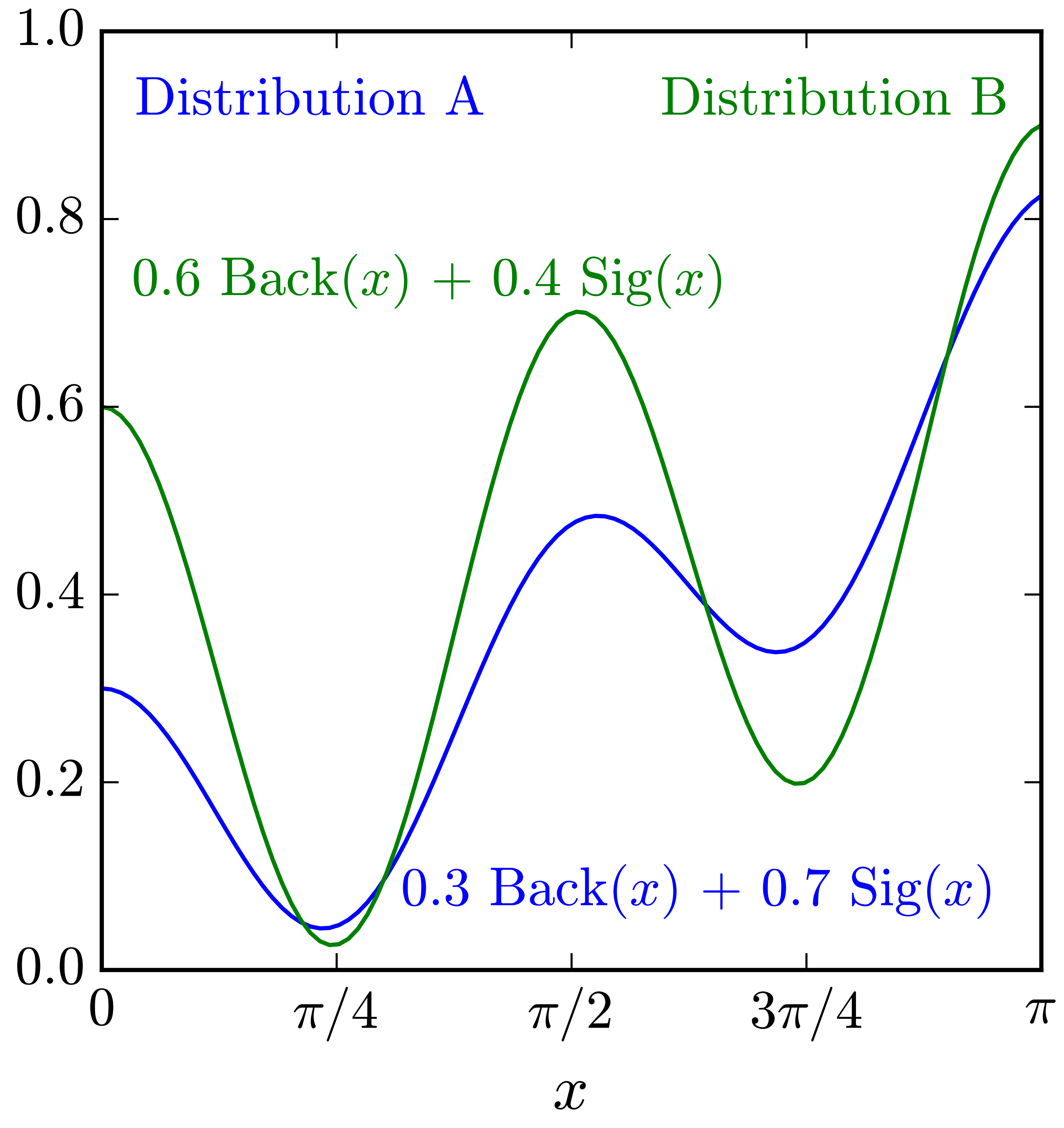
Weak supervision is closer to quantum reality: no single event is really either signal or background.

Opens the possibility of training on data instead of Monte Carlo.

Background ***Signal***

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

How can this work?

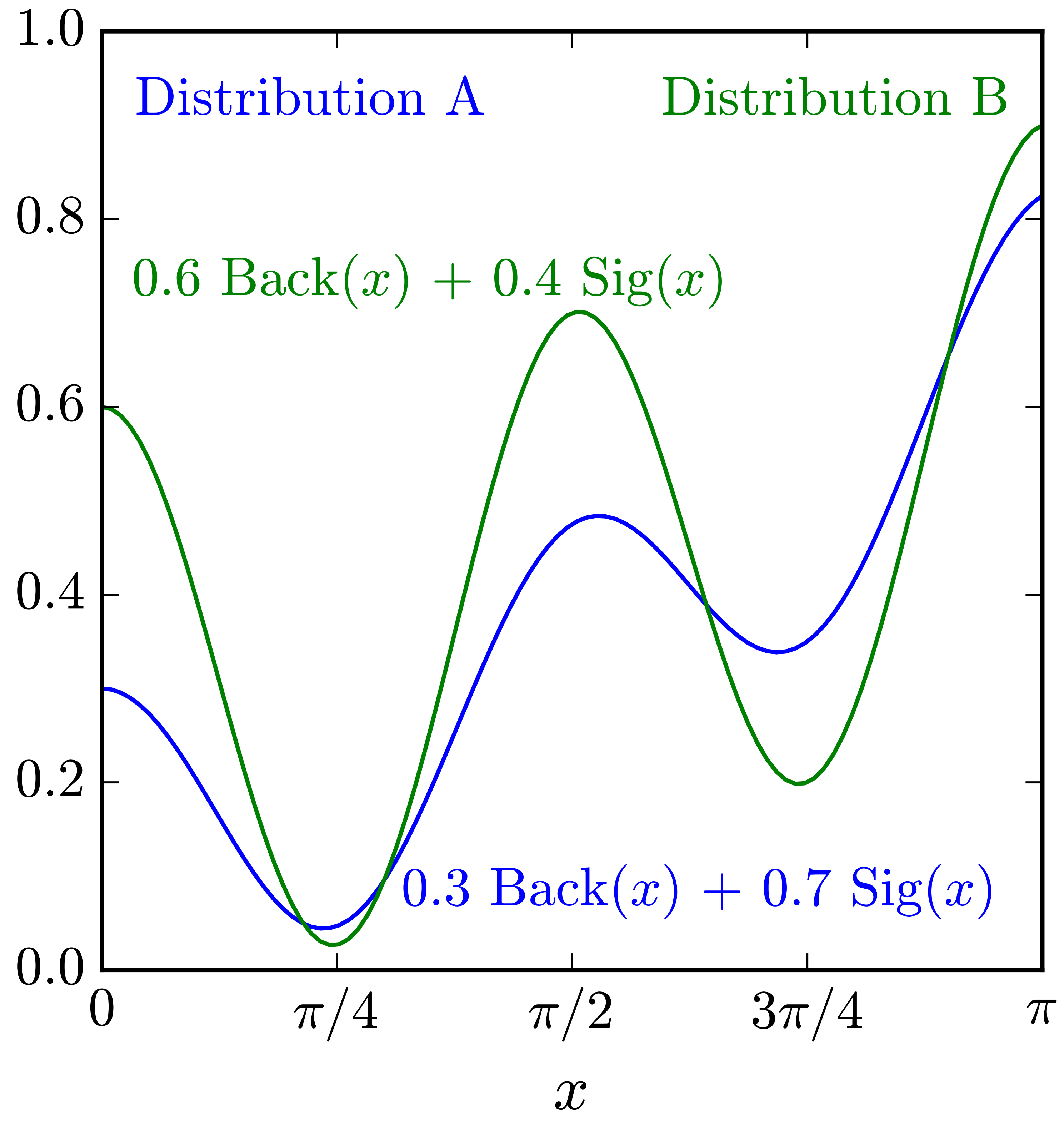


Problem: (How) can we make a classifier without event-by-event truth-level labels

How can this work?

Group A

Group B



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

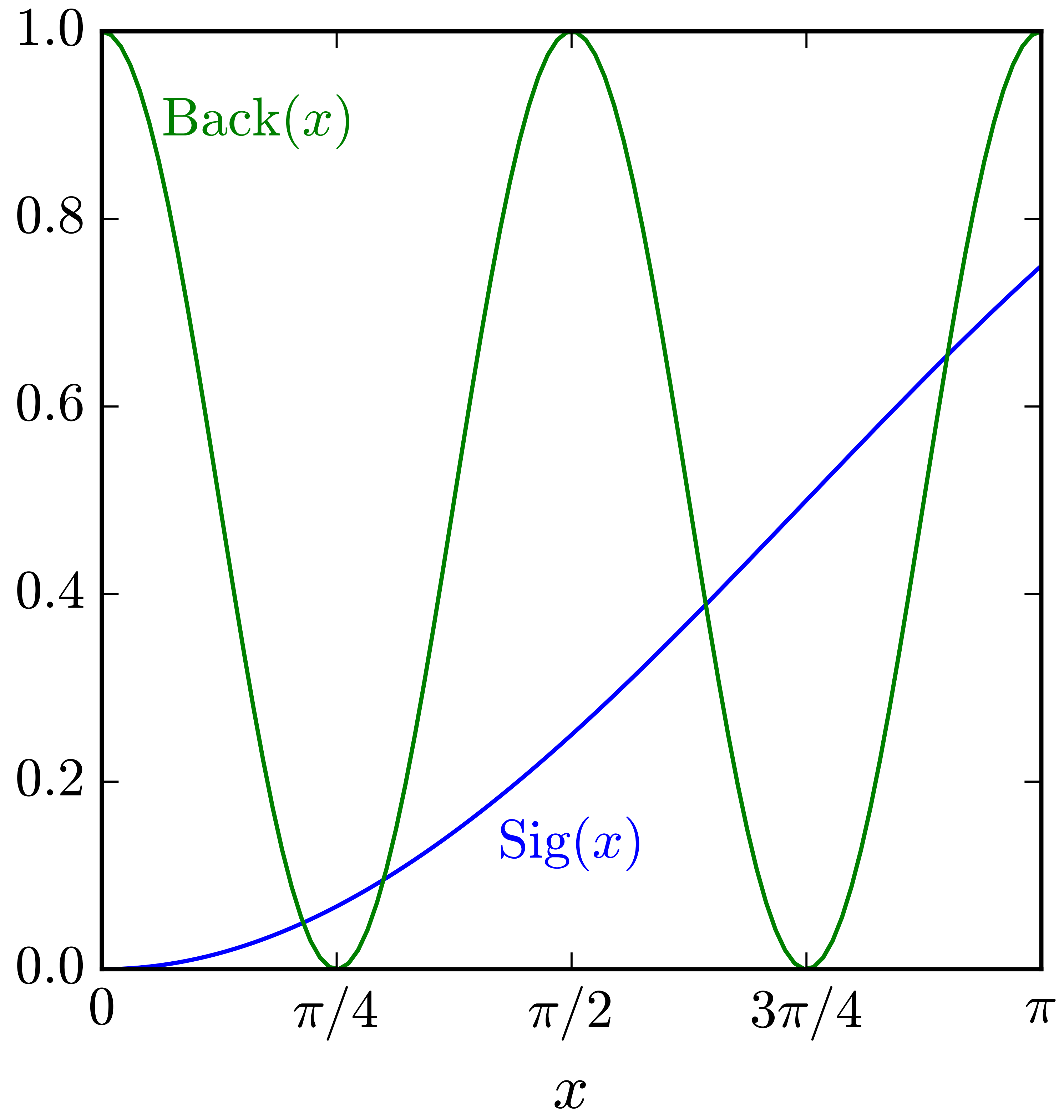
$$\mathrm{Back}(x) = \frac{1}{3}\left(-4A(x) - 7B(x)\right)$$

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

How can this work?

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

How can this work?

Make a histogram of the multi-dimensional data

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

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

Group A

Group B

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

How can this work?

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

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

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

How can this work?

Group A

Group B

Make a histogram of the multi-dimensional data

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

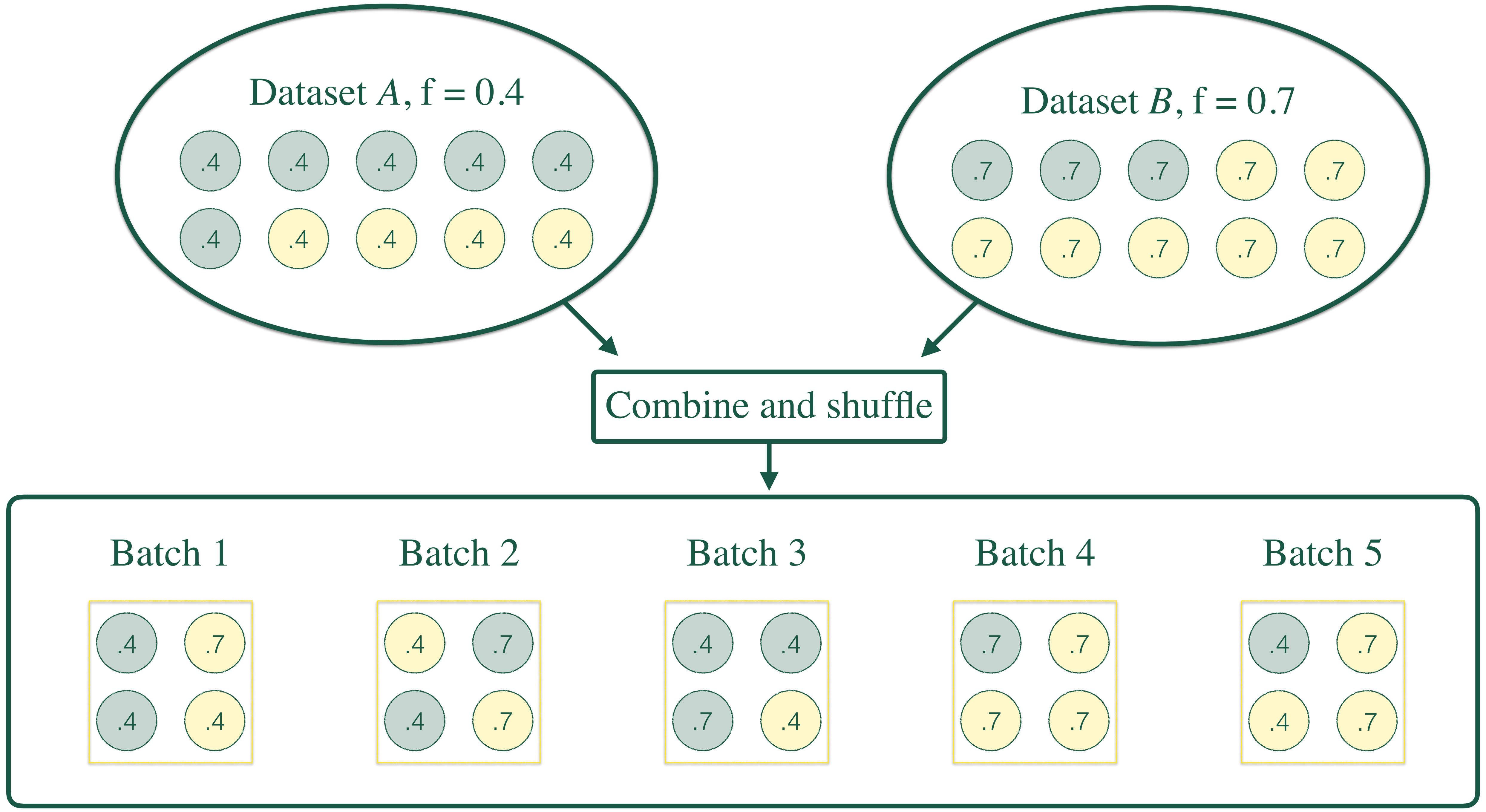
Invert

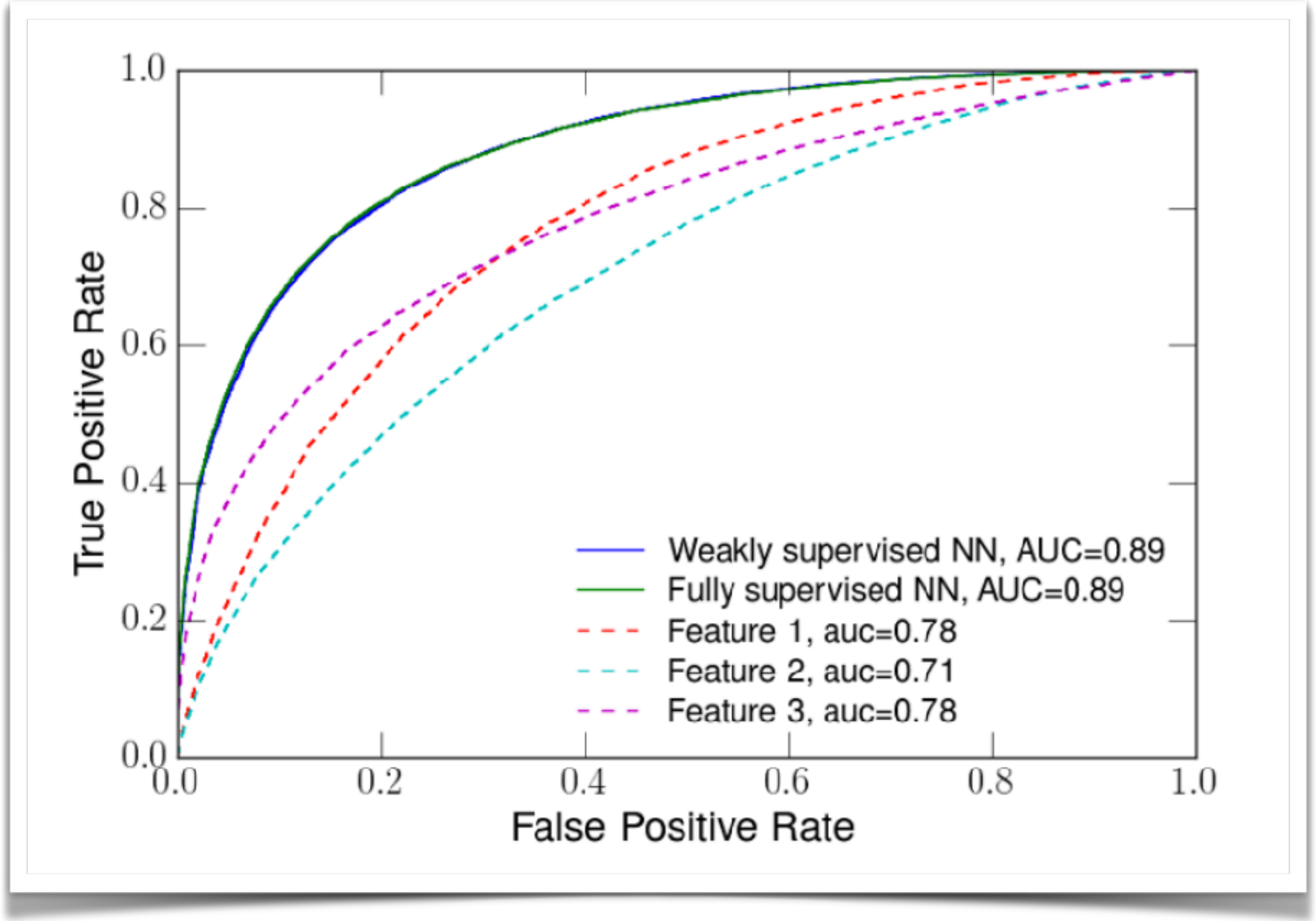
$$h_{0,i} = \frac{y_A \; h_{B,i} - y_B \; h_{A,i}}{y_A - y_B}$$
$$h_{1,i} = \frac{(1-y_B) h_{A,i} - (1-y_A) h_{B,i}}{y_A - y_B}$$

Machine learning helps with:

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

Weak supervision - DNRS

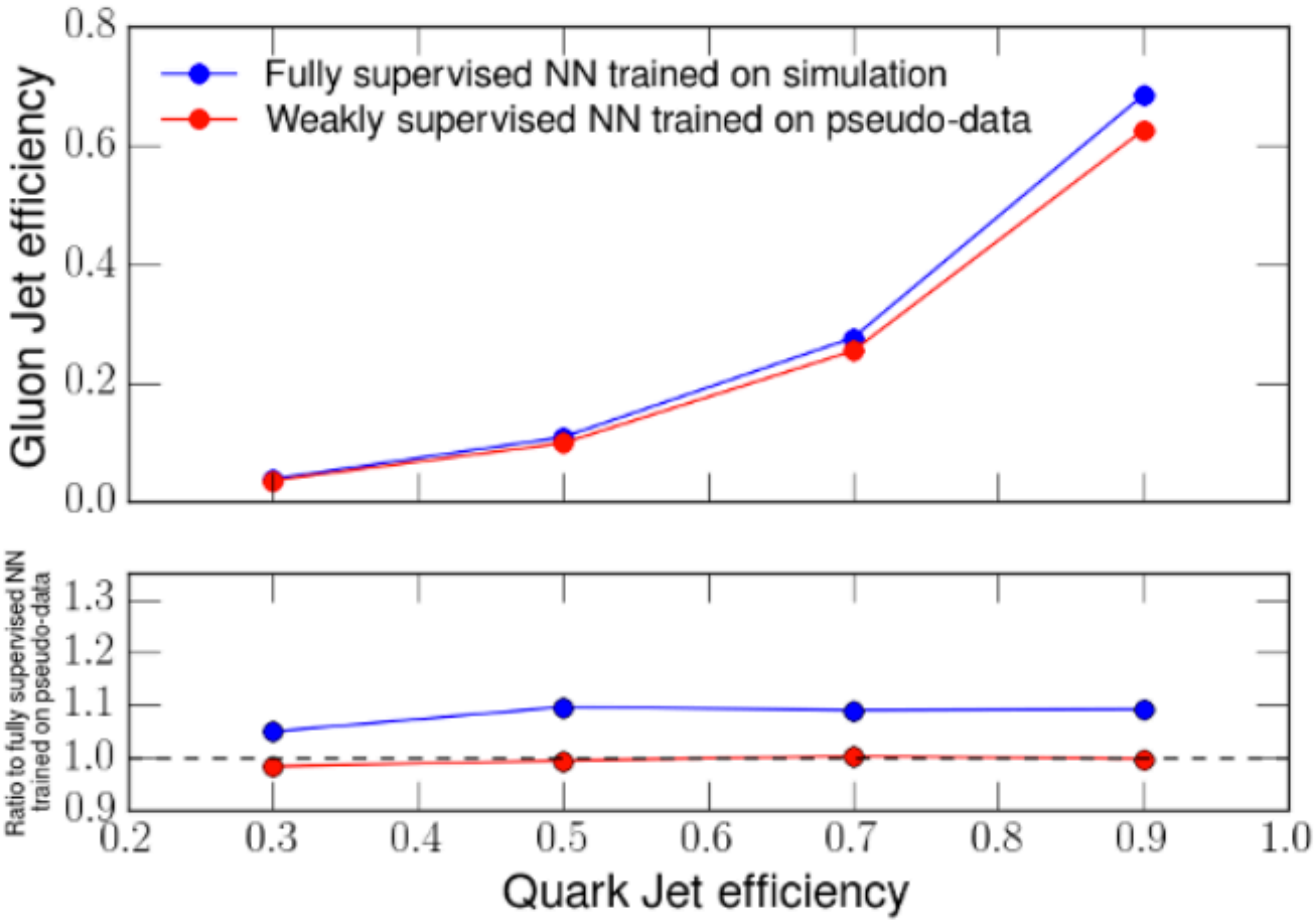




$$\ell_{DNRS} = \sum_{batches} |\langle f_{t,i} \rangle - \langle y_{p,i} \rangle|$$

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 - DNRS



- Training directly on 'data' mitigates effects of mismodeling.
- Only depends on ratios, not modeled distribution.

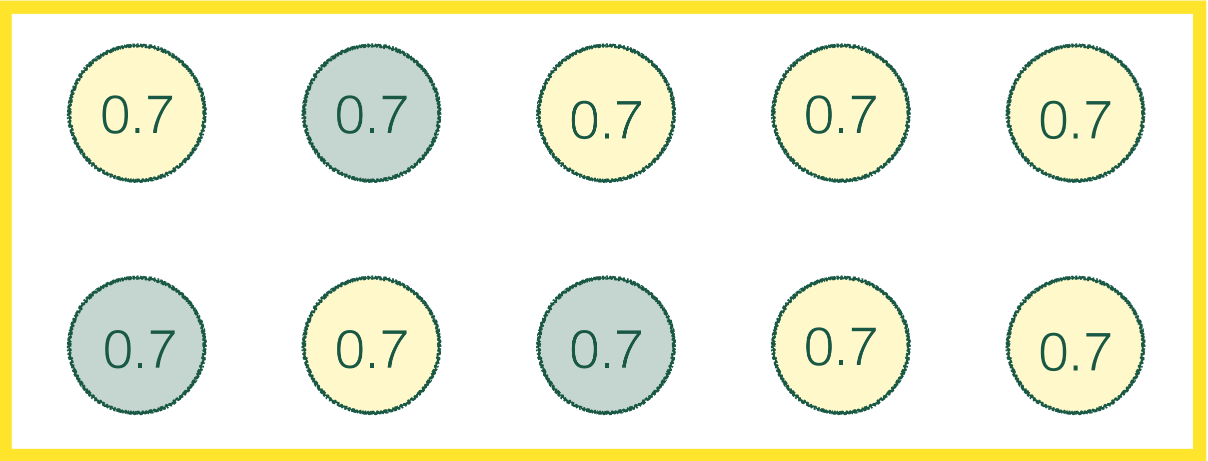
L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, JHEP **1705**, 145 (2017) doi:10.1007/JHEP05(2017)145 [arXiv:1702.00414 [hep-ph]]

Weak supervision - DNRS

Only depends on ratios, not modeled distribution.

Weak supervision - DNRS

Only depends on ratios, not modeled distribution.



Group A

Group B

What if there are uncertainties on the ratios?

Weak supervision - DNRS

Only depends on ratios, not modeled distribution.

Group A

Group B

What if there are uncertainties on the ratios?

Cohen, Freytsis, and BO [arXiv:1706.09451]

Condition for when label errors do not affect classifier

Metodiev, Nachman, and Thaler [arXiv:1708.02949]

Possible to do classification with arbitrary labels

CWoLa

What if there are uncertainties on the ratios?

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]

What if there are uncertainties on the ratios?

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:

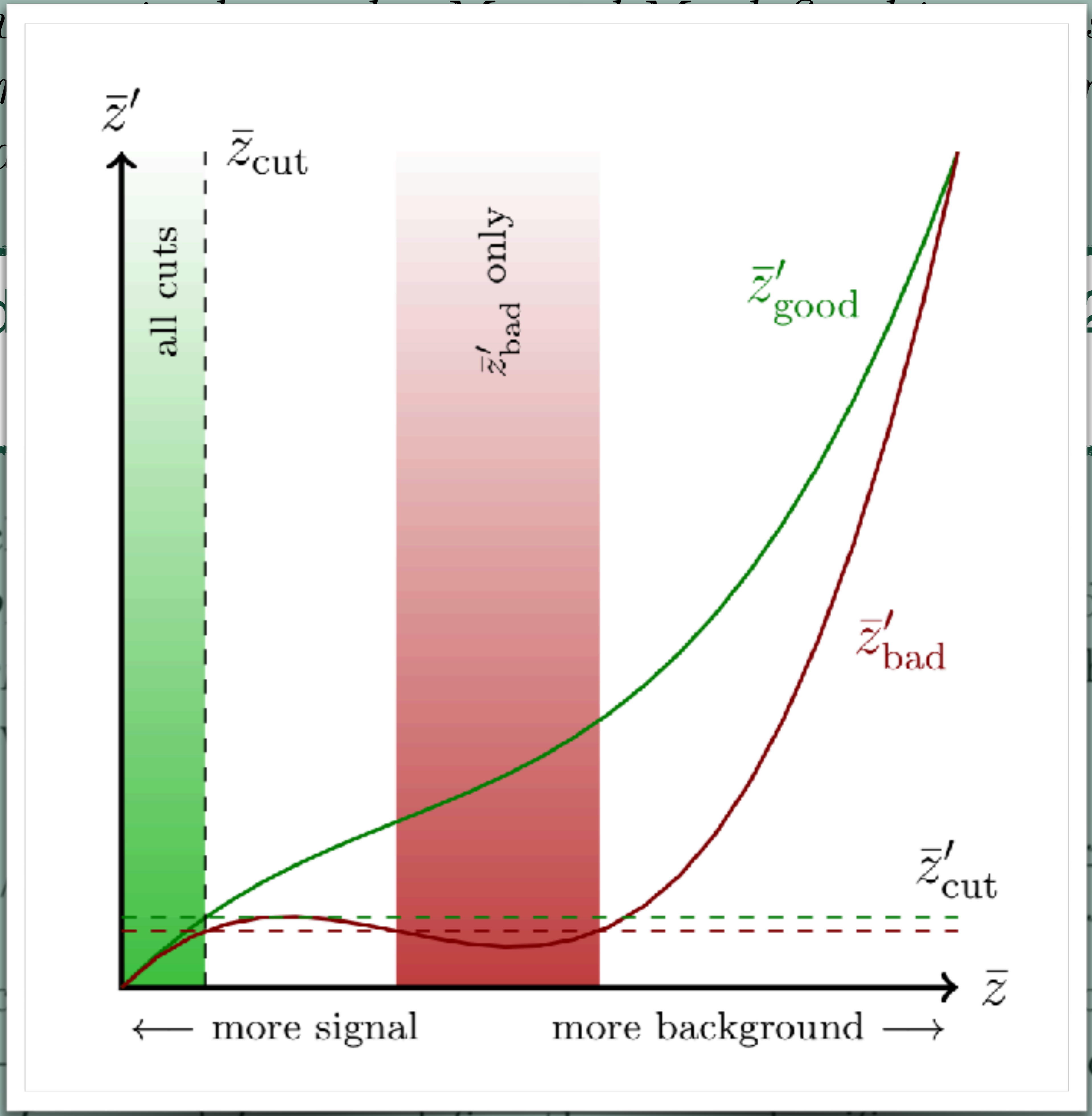
$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1 - f_1)\, p_B}{f_2\, p_S + (1 - f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1 - f_1)}{f_2\, L_{S/B} + (1 - f_2)},$$

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. $\square$

What if there are uncertainties on the ratios?

Theorem 1 *Giv… s of pure samples*
S and B with sig… ned to distinguish
M_1 from M_2 is …

Metod… 2949]



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

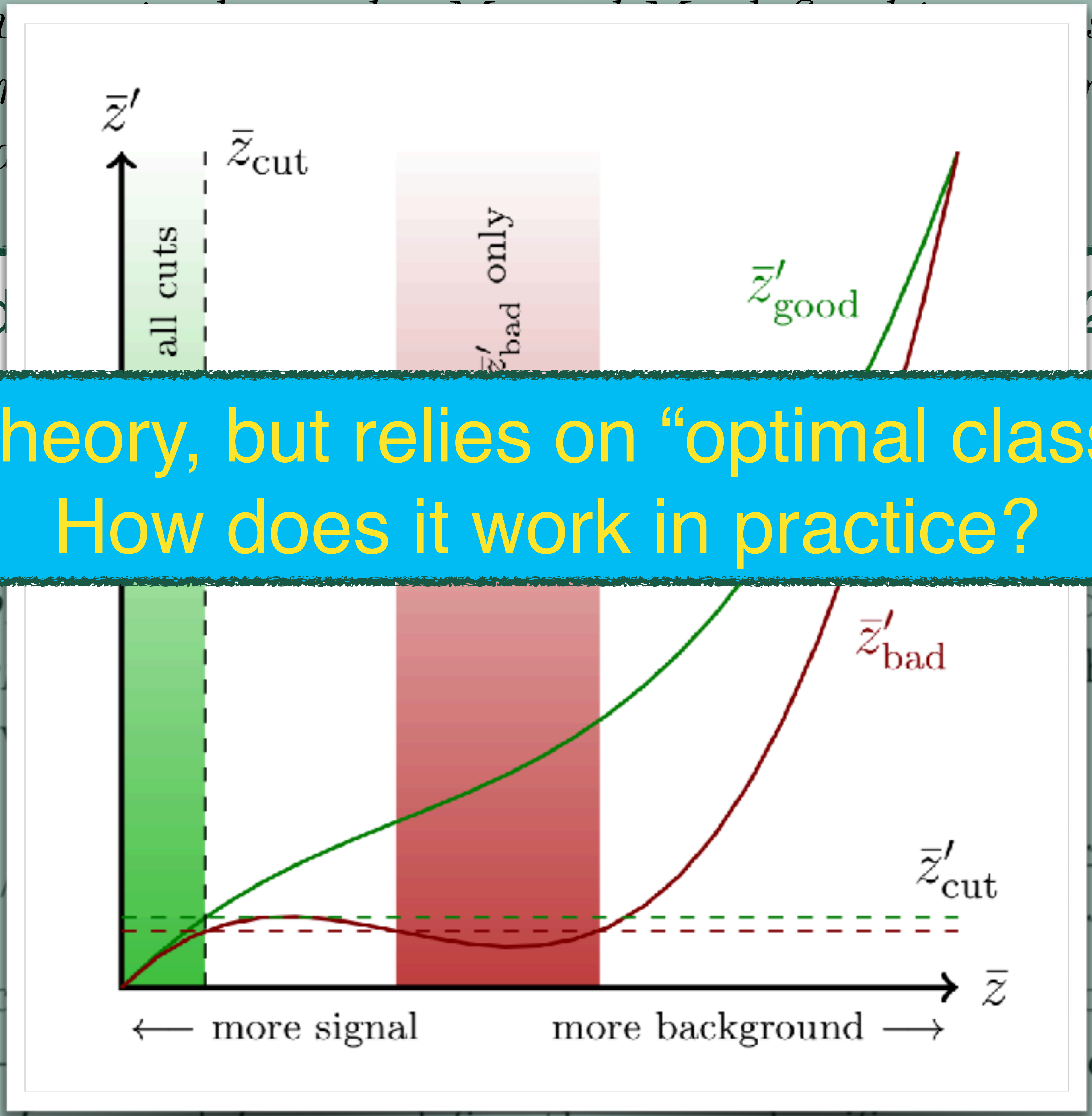
$$L_{M_1}\ldots \ldots f_1) \ldots f_2),$$

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

What if there are uncertainties on the ratios?

Theorem 1 *Gi… s of pure samples S and B with sig… ned to distinguish M_1 from M_2 is …*

Metod… 2949]



Nice theory, but relies on "optimal classifier."
How does it work in practice?

Proof. Th… ikelihood ratio $L_{M_1/M_2}(x) = p$… distinguish examples drawn from p_S and p… here p_B has support, we can relate these t…

$$L_{M_1/…\ f_1) … f_2)},$$

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

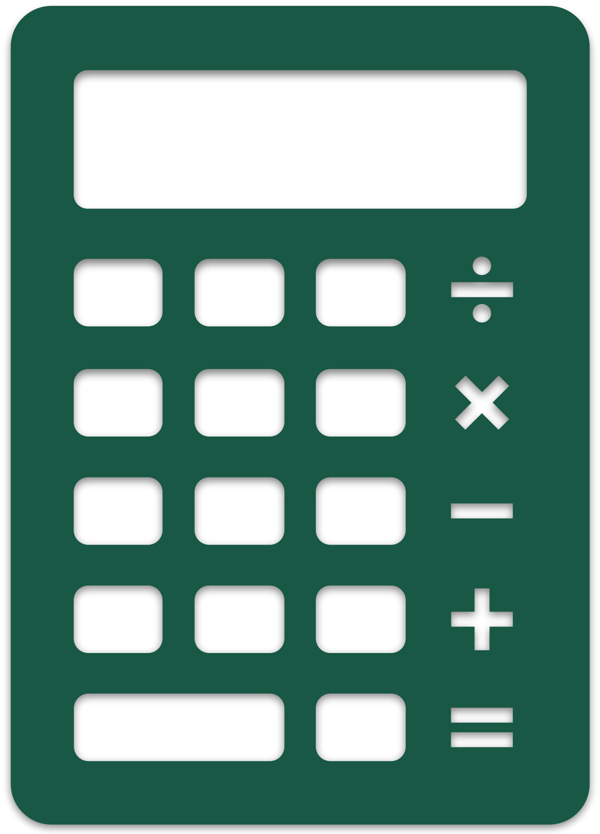
Technical Aspects

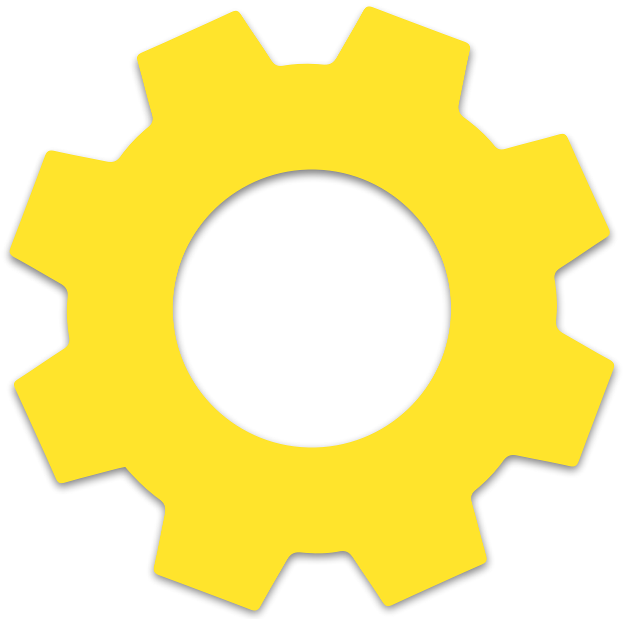
Learning implemented with Keras.
TensorFlow backend
scikit-learn used to compute metrics.
Particle physics events generated with MadGraph + pythia + Delphes.

Choice	Toy Models	BSM Scenario
Loss function	BCE	BCE
n_{input}	3	11
Hidden Nodes	30	30
Activation	Sigmoid	Sigmoid
Initialization	Normal	Normal
Learning algorithm	ADAM	SGD
Learning rate	0.0015	0.01
Batch size	32	64
Epochs	100	20

Technical Aspects

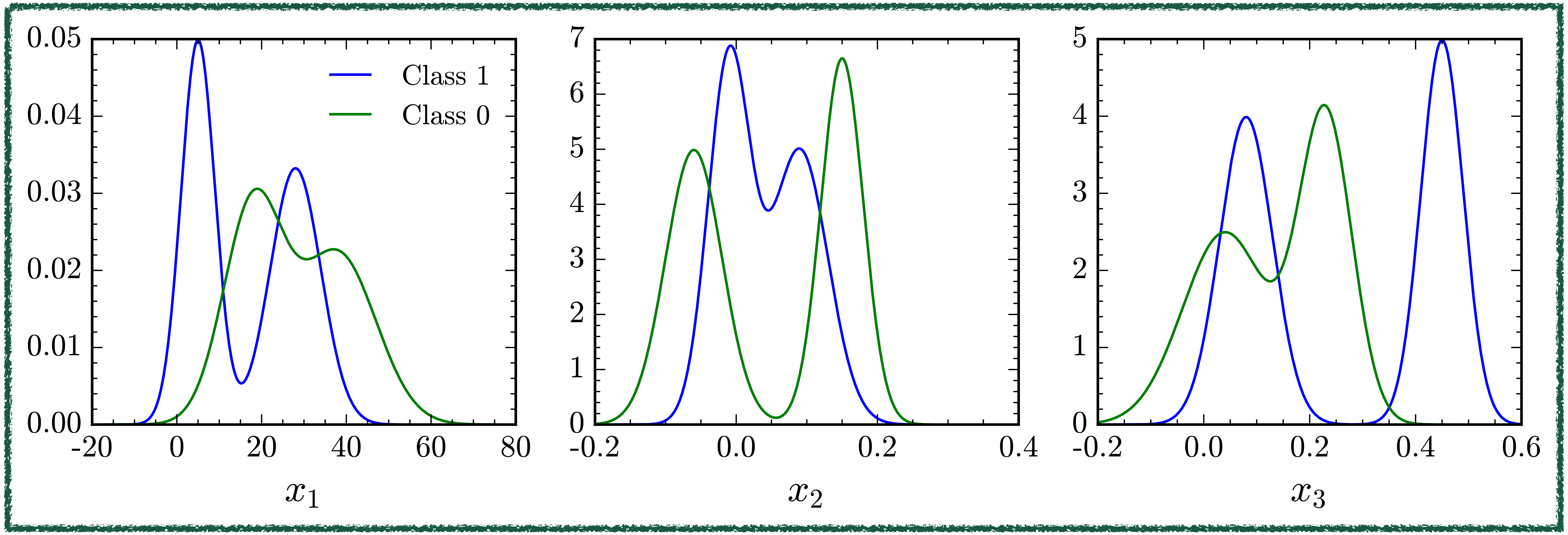
Learning implemented with Keras.
TensorFlow backend
scikit-learn used to compute metrics.
Particle physics events generated with MadGraph + pythia + Delphes.





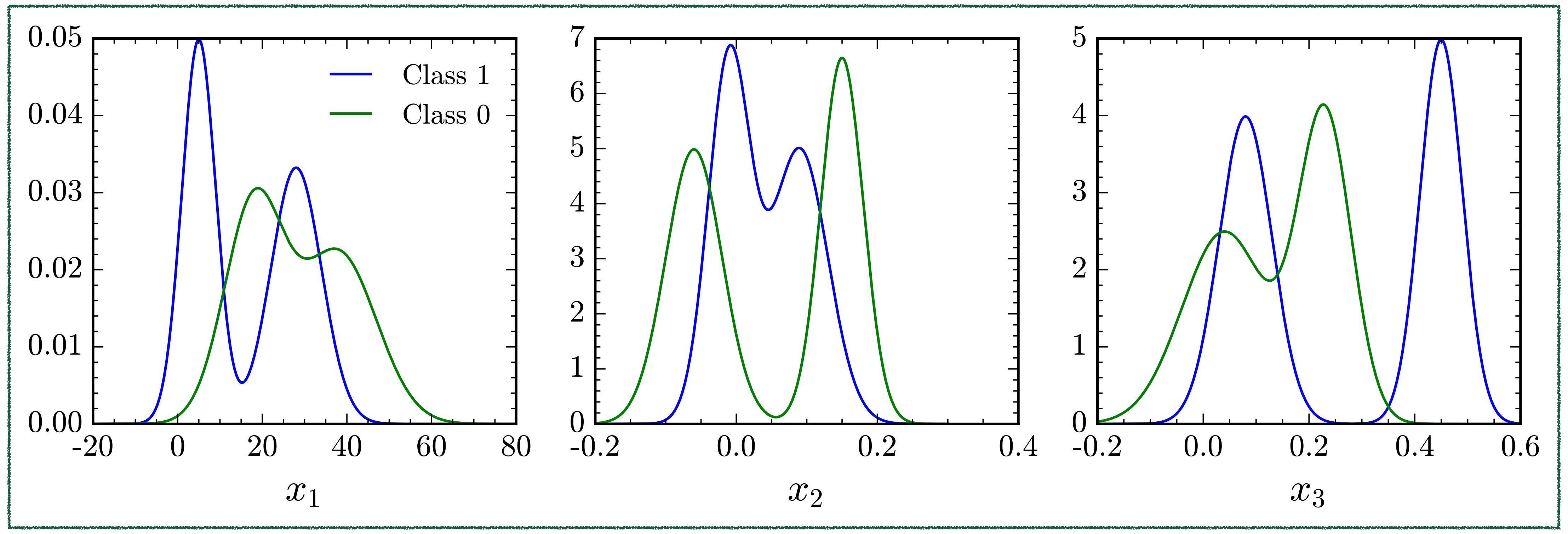
Choice	Toy Models	BSM Scenario
Loss function	BCE	BCE
n_{input}	3	11
Hidden Nodes	30	30
Activation	Sigmoid	Sigmoid
Initialization	Normal	Normal
Learning algorithm	ADAM	SGD
Learning rate	0.0015	0.01
Batch size	32	64
Epochs	100	20

Toy model with 3 inputs

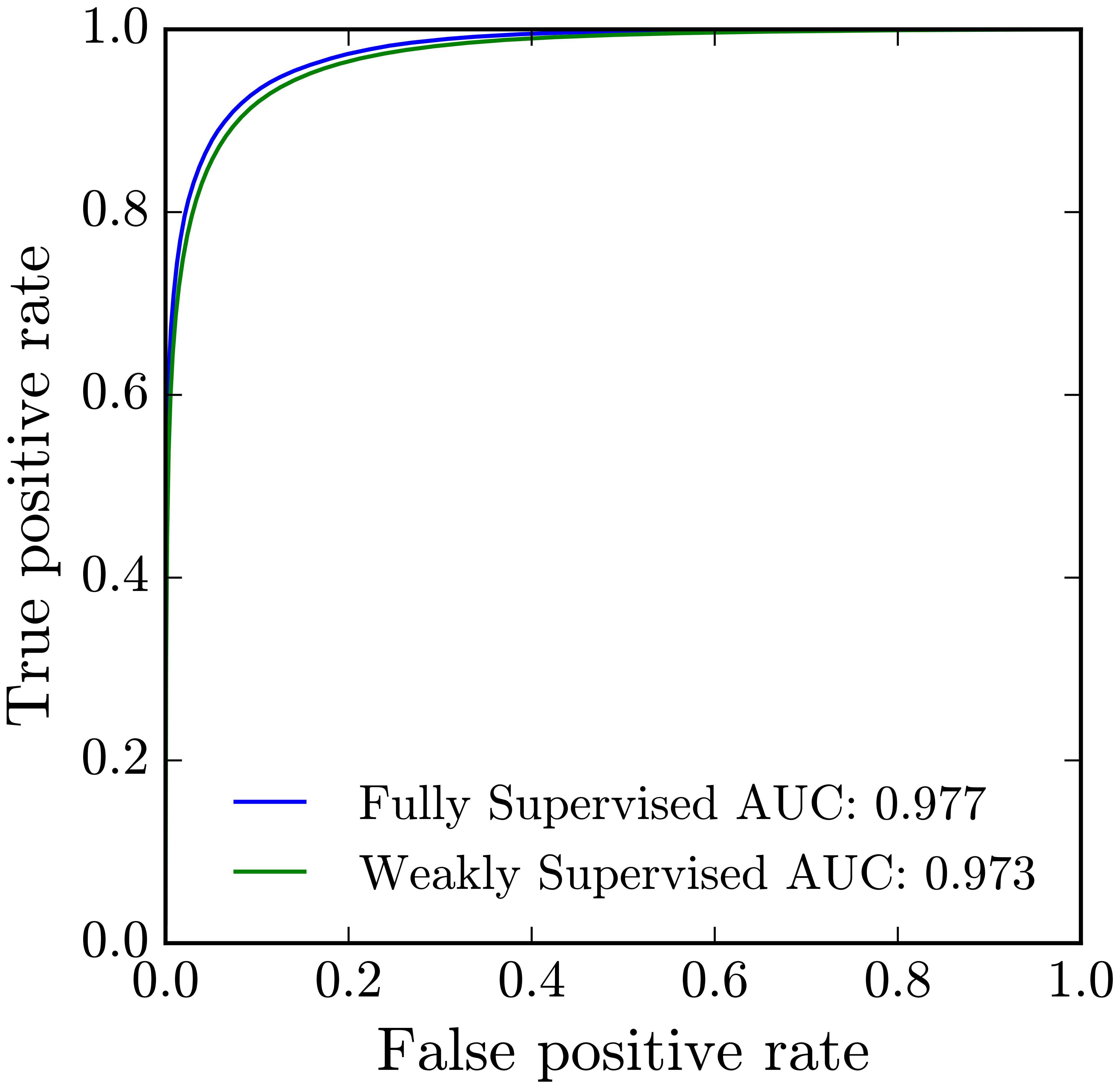


200,000 samples with 70% signal
200,000 samples with 40% signal
Test on 200,000 samples with 55% signal

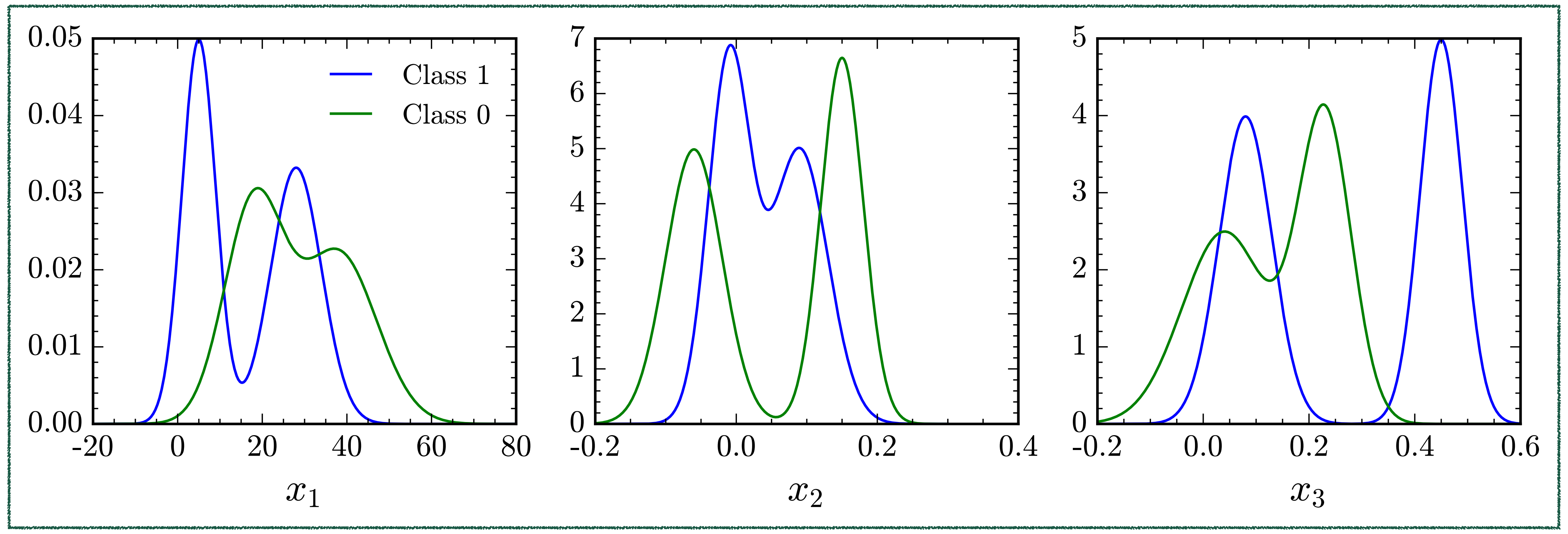
Toy model with 3 inputs



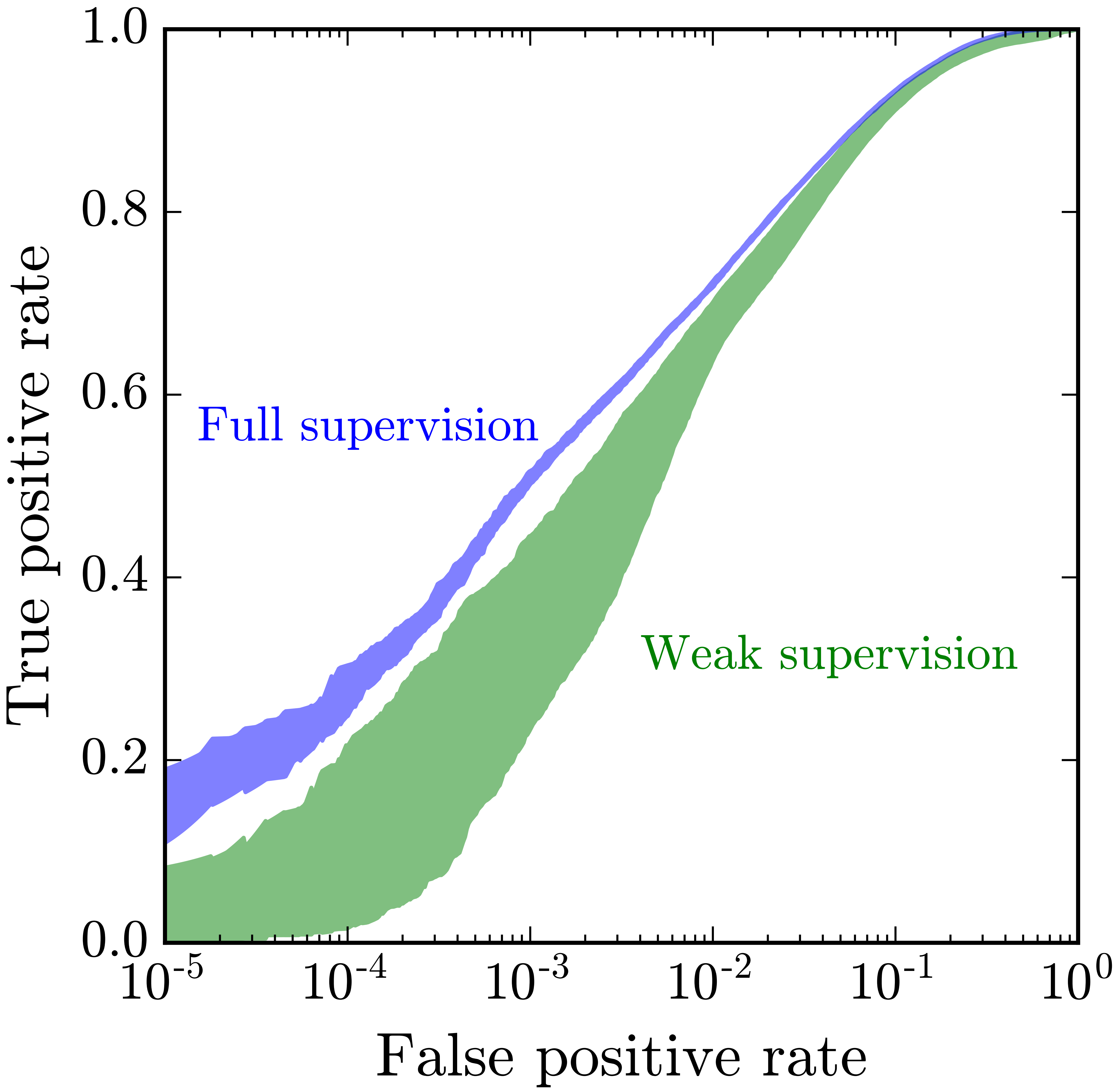
- Each event labeled with 0.4 or 0.7
- Using a loss function other than DNRS (LLP) works
- Weak and full supervision yield similar results



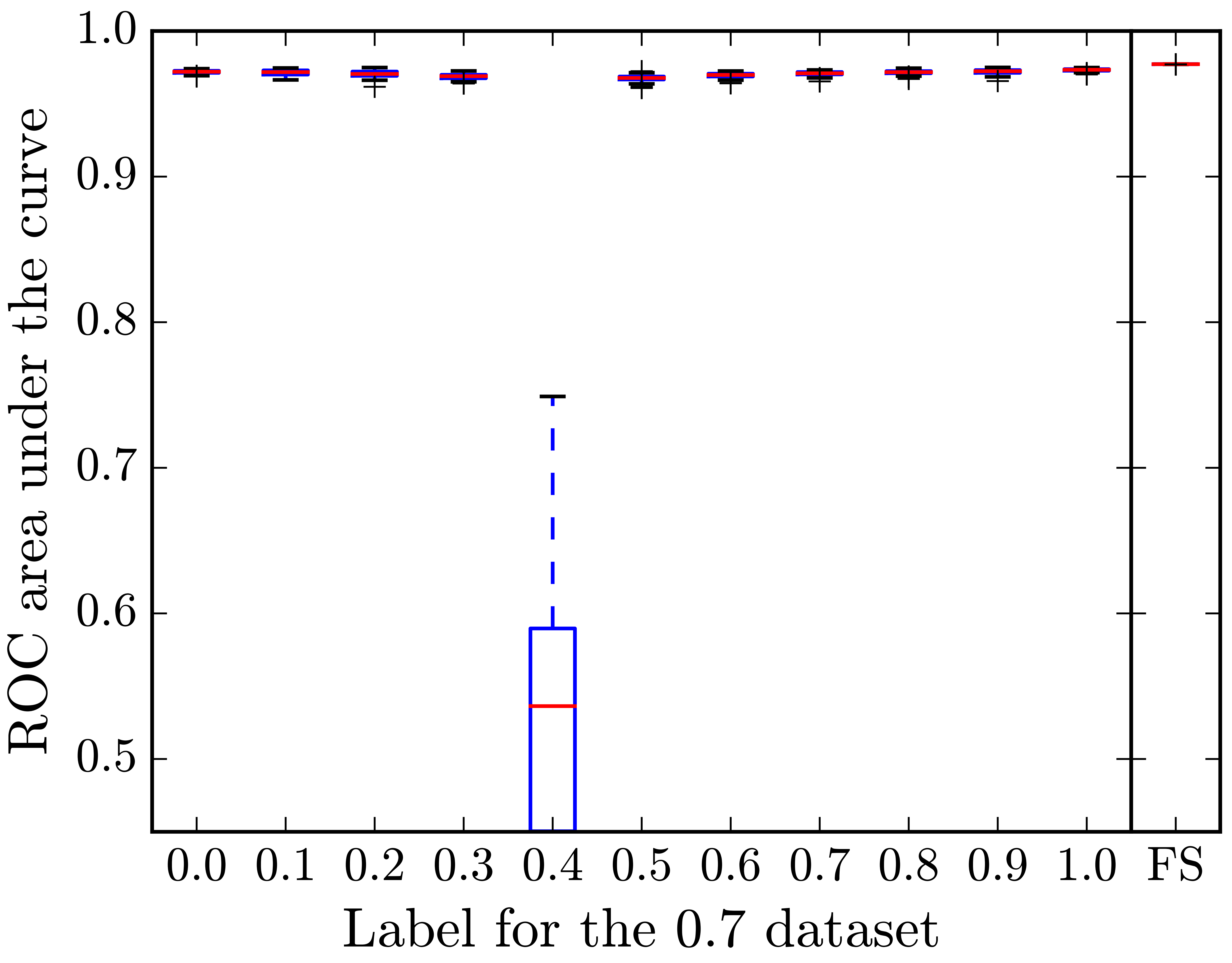
Toy model with 3 inputs



- Each event labeled with 0.4 or 0.7
- Using a loss function other than DNRS (LLP) works
- Weak and full supervision yield similar results



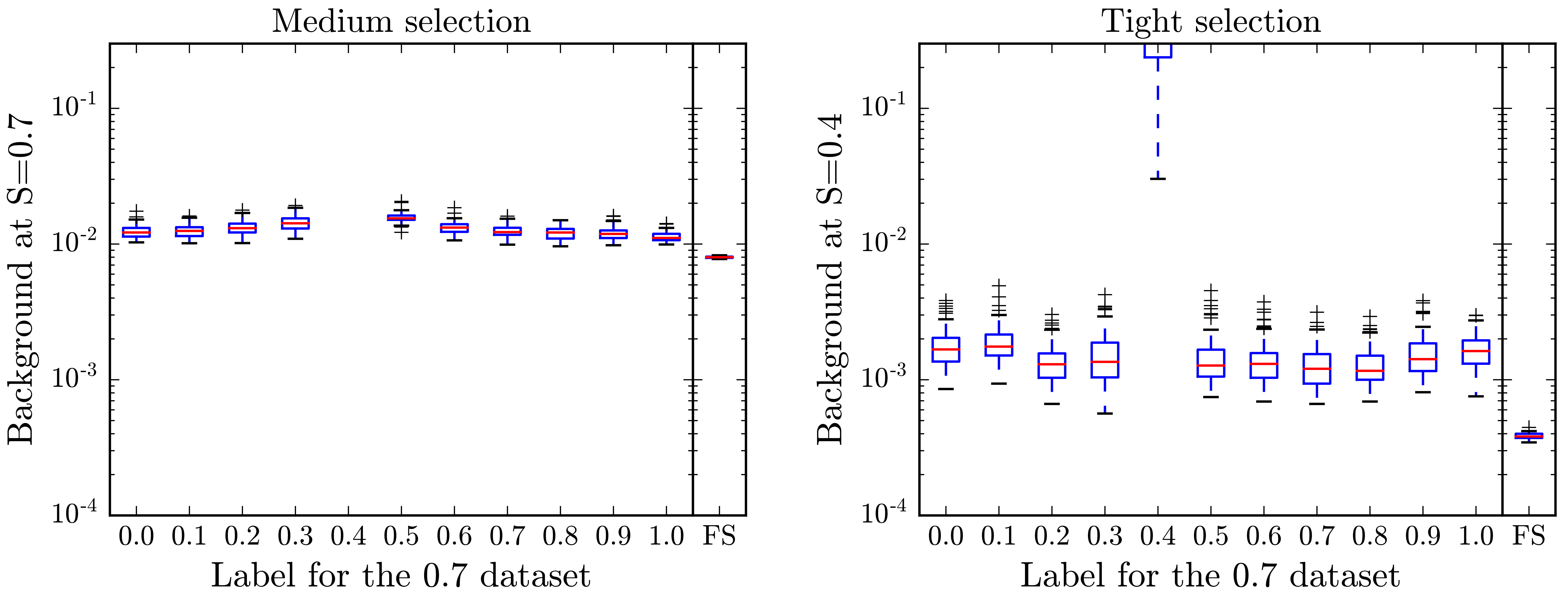
Toy model with 3 inputs



- Keep ratios fixed at 0.4 and 0.7
- Label the 0.7 set with a different number
- Test on original test set

Almost no dependence on the label!

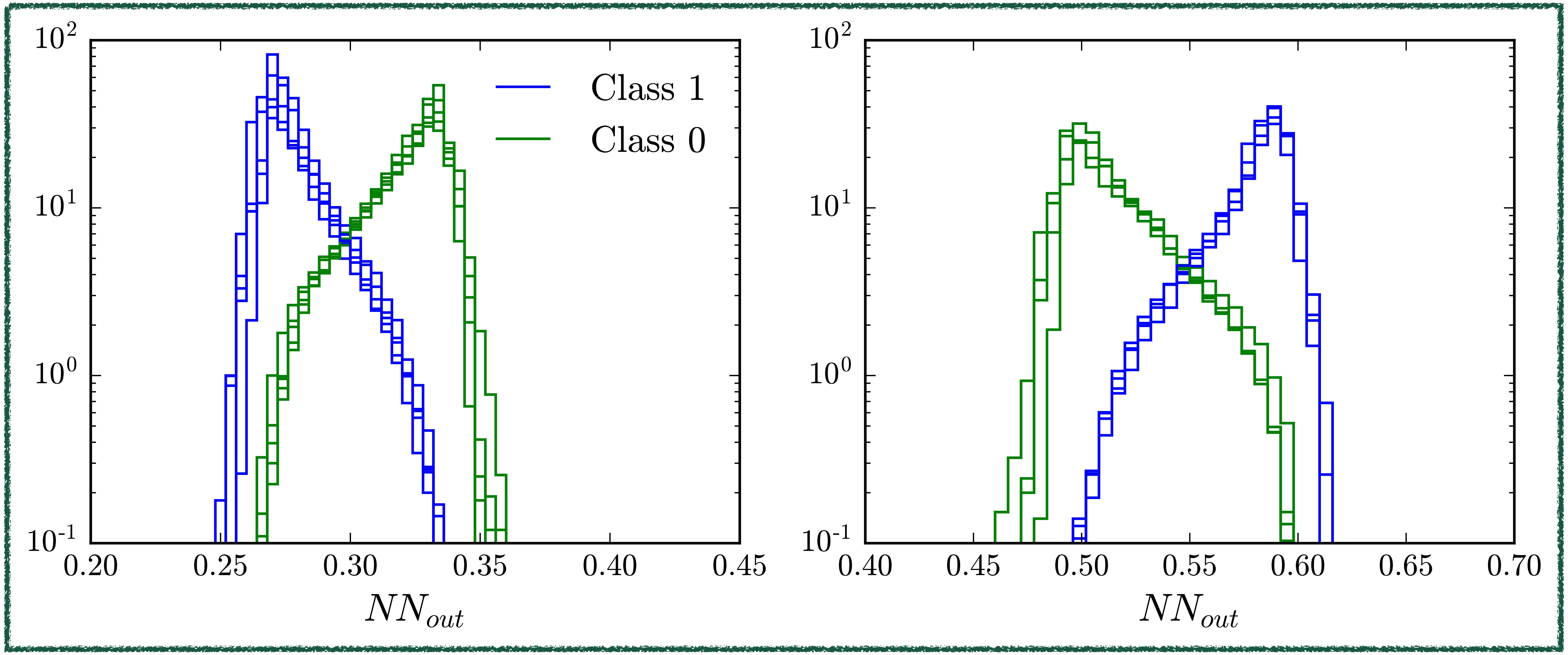
Toy model with 3 inputs



Toy model with 3 inputs

How can the weak networks achieve similar results with the wrong information?

Examine the output of the networks



Outline

1. Introduction / Toy model

- What is weak supervision?
- How can it work?
- Is it robust?

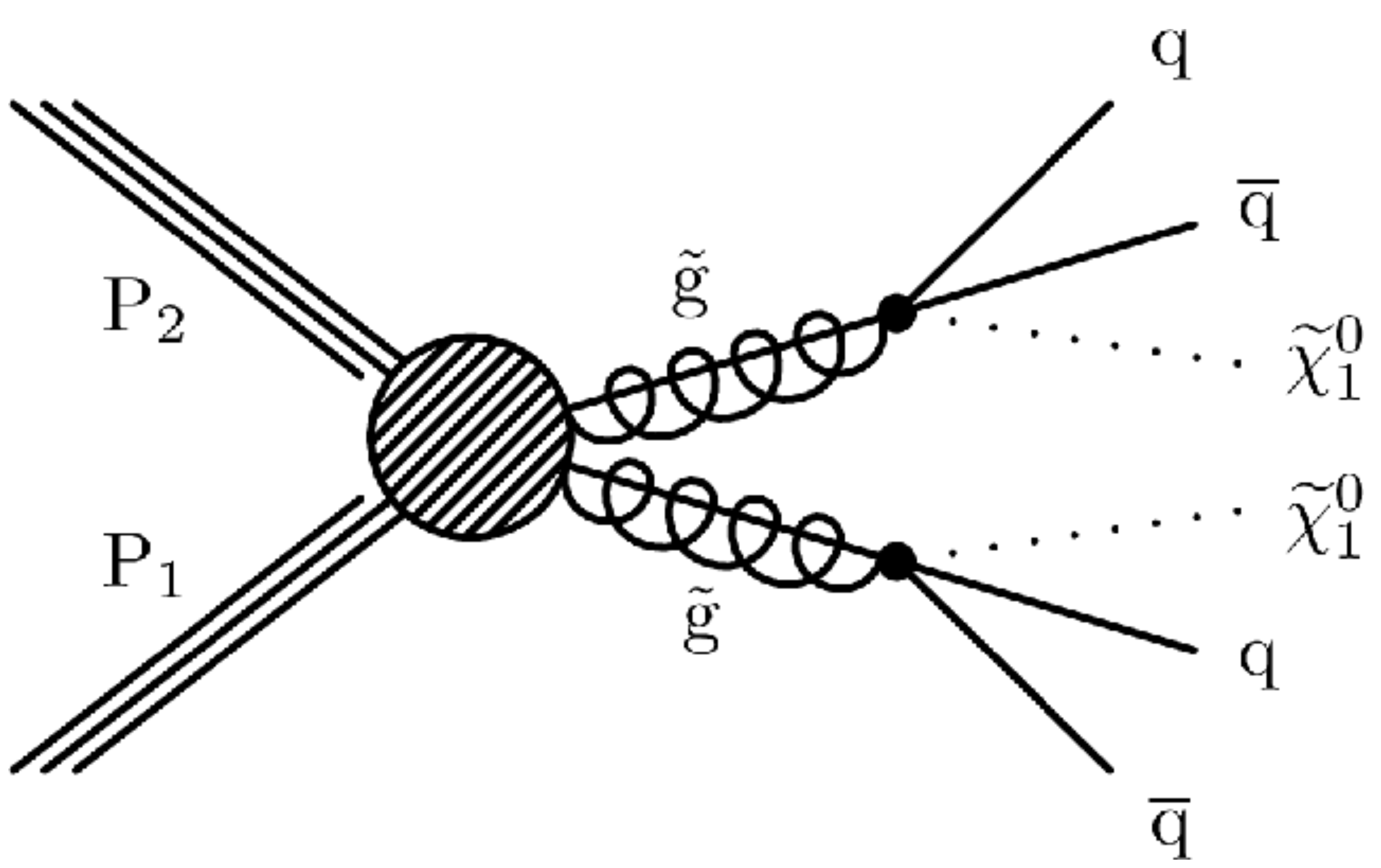
Outline

1. Introduction / Toy model
    - What is weak supervision?
    - How can it work?
    - Is it robust?
2. LHC Scenario
    - Higher input dimension
    - Application to unseen data
    - Affects of mis-modeling
    - Combination of Full and Weak

LHC Scenario

JETS + MET

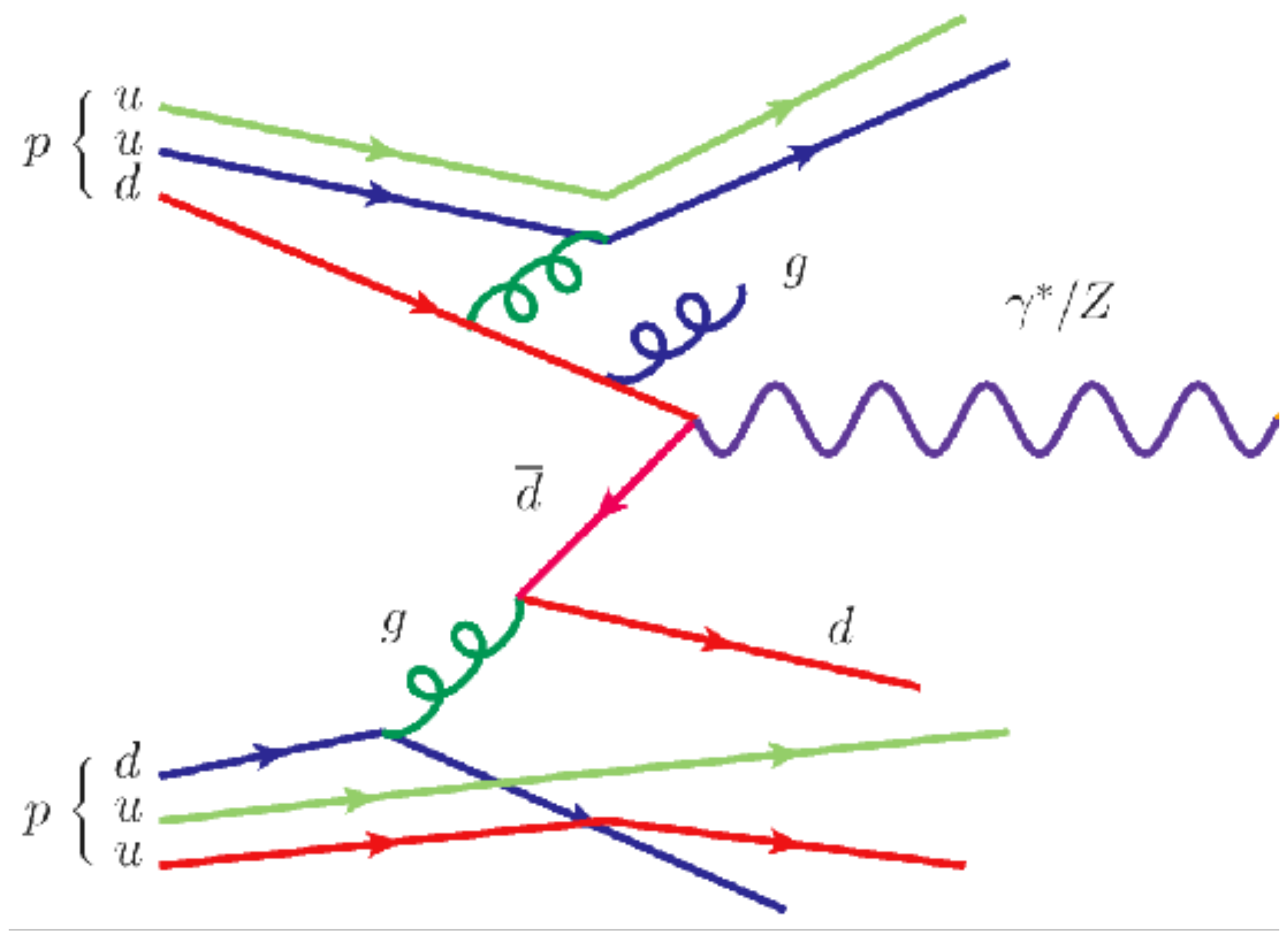
gluino



$$\sigma = 0.3 \text{ fb}$$

vs

Z + jets



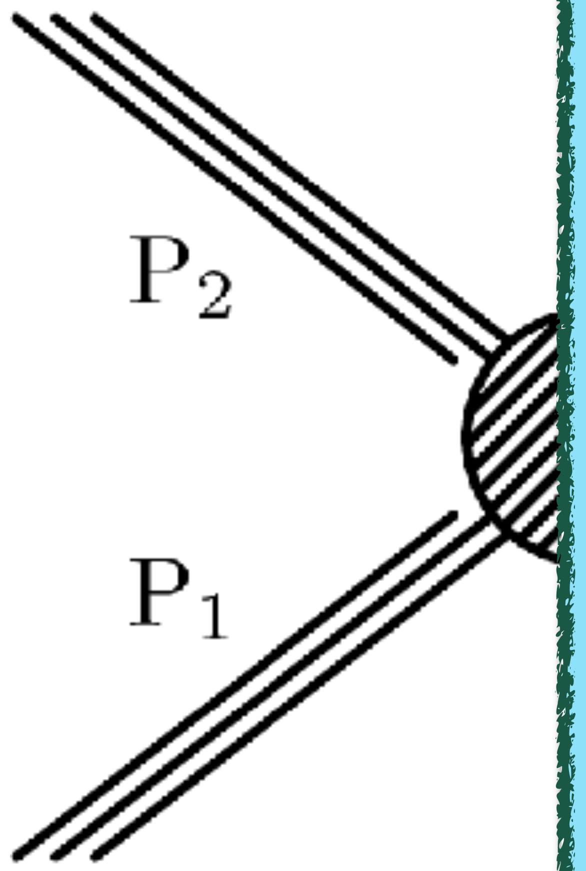
$$\sigma = 28800 \text{ fb}$$

LHC Scenario

Jets + MET

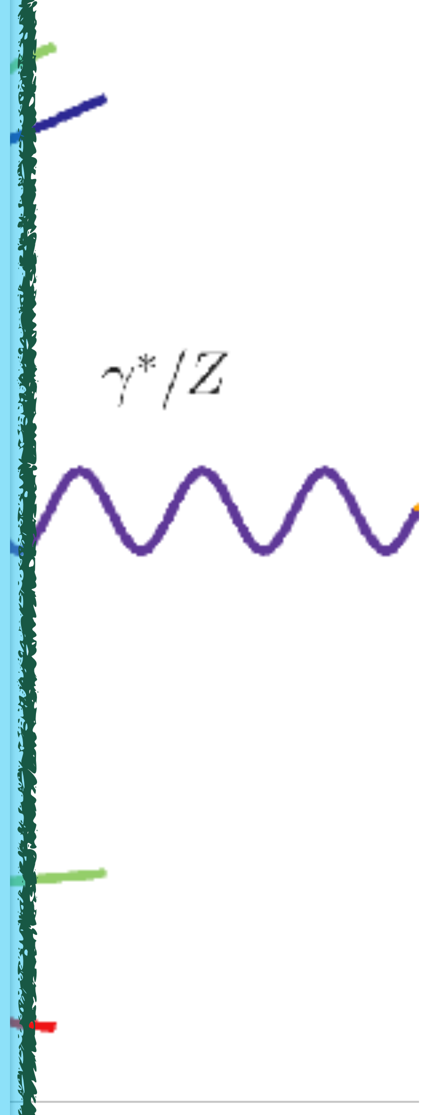
gluino

Z + jets



- Not trying to get the best, realistic search possible
  - Characterizing weak supervision
- Generate 1 million background, 500k signal
- After basic cuts, 264K and 473K events left

$$p_T(j_1) > 200 \text{ GeV} \qquad E_T^{miss} > 200 \text{ GeV}$$

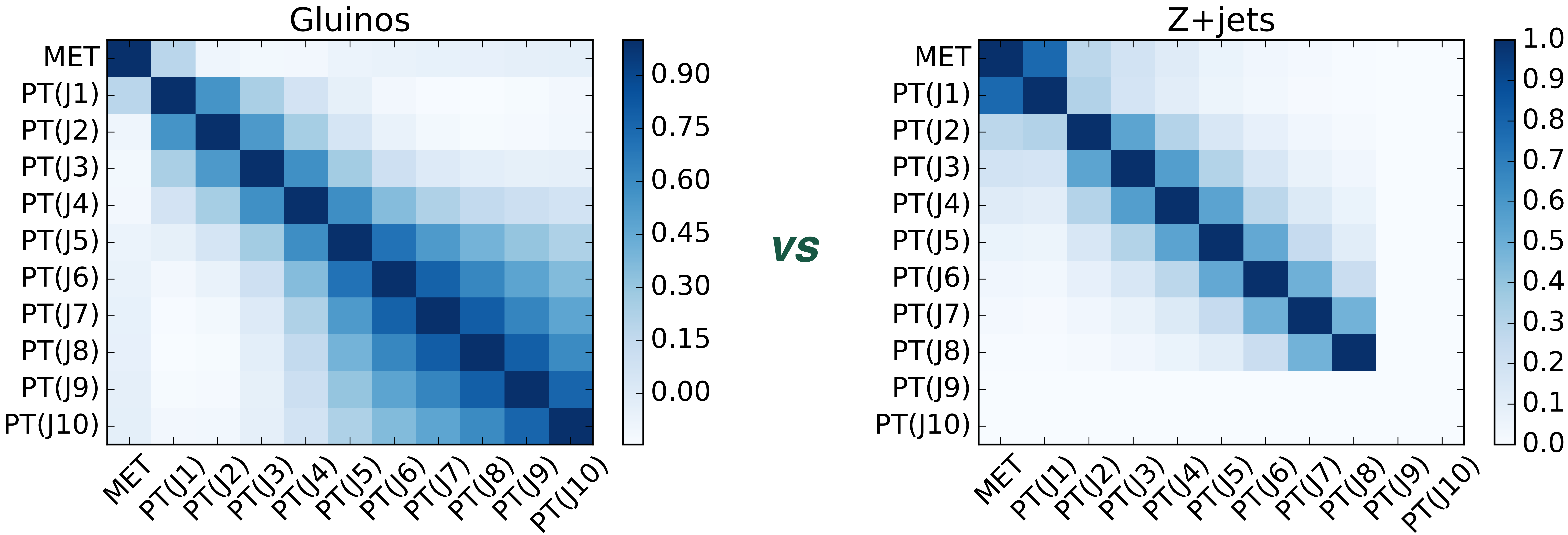


$\sigma = 0.3$ fb

$\sigma = 28800$ fb

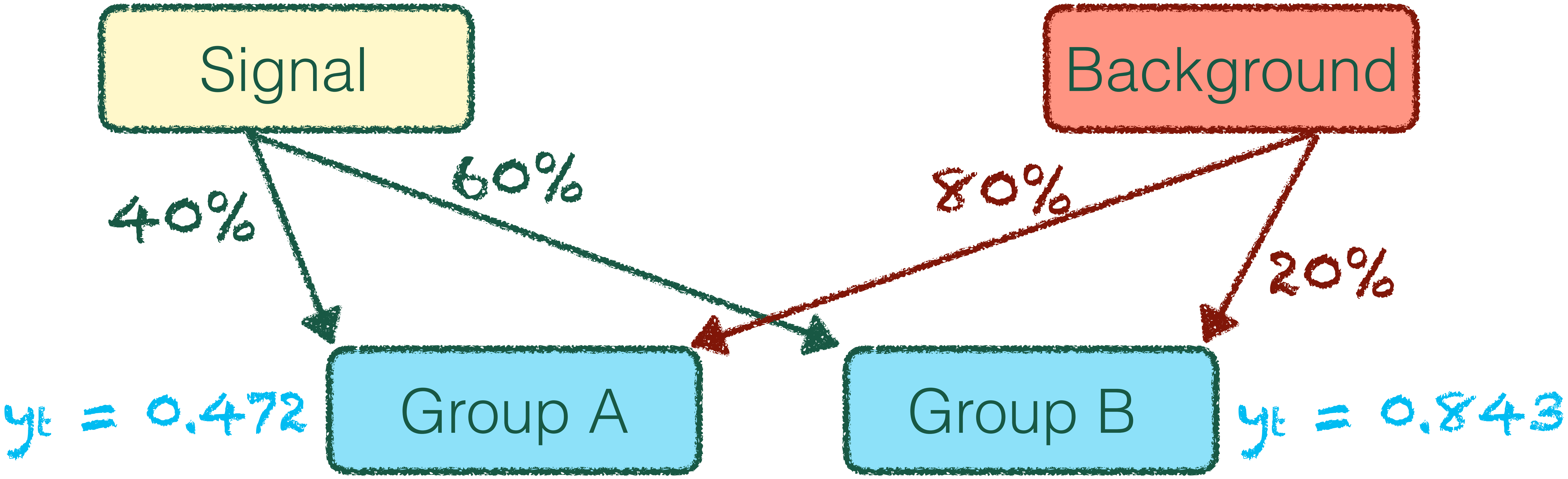
LHC Scenario

Jets + MET



Correlations between input parameters

LHC Scenario



Network	AUC	Signal efficiency
Full	0.99992393(31)	0.999373(17)
Weak	0.9998978(35)	0.999286(30)

Metrics for training networks to distinguish gluino pair production with decays to 1st generation quarks from the dominant $Z+$jet background. The signal efficiency is given for a background acceptance of 0.01.

LHC Scenario

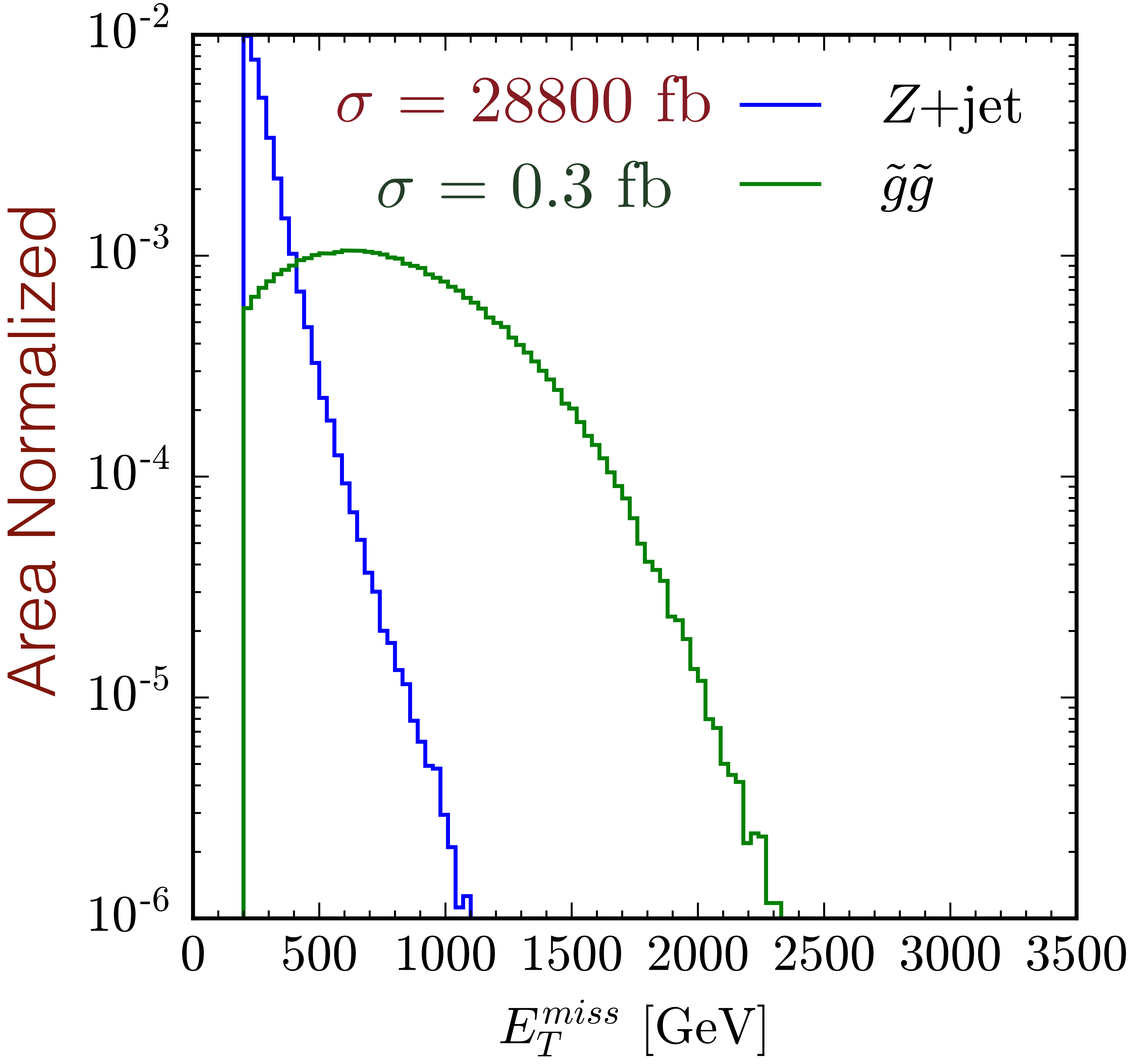
Network	AUC	Signal efficiency
Full	0.99992393(31)	0.999373(17)
Weak	0.9998978(35)	0.999286(30)

Metrics for training networks to distinguish gluino pair production with decays to 1st generation quarks from the dominant $Z+$jet background. The signal efficiency is given for a background acceptance of 0.01.

Can we trust this performance?

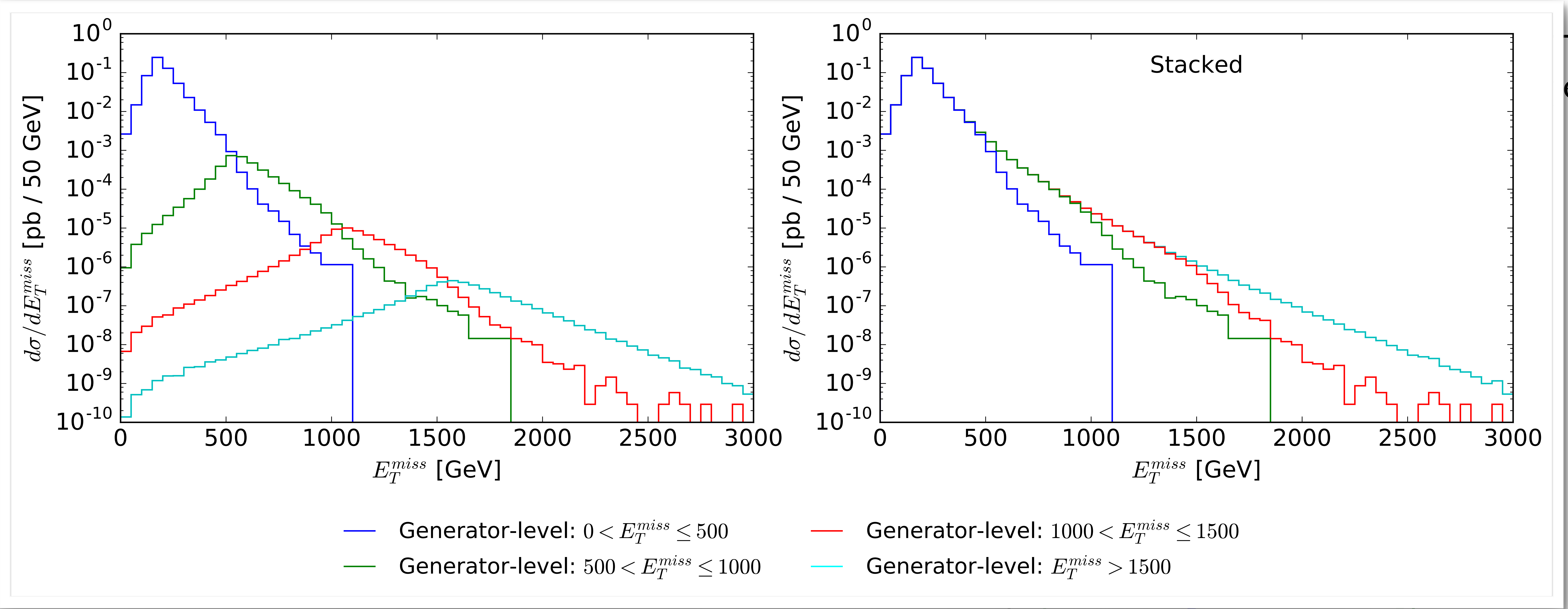
Not enough backgrounds (effective luminosity for backgrounds much smaller)

Need much more pure samples than can be obtained with the given number of background events



LHC Scenario

Me
t

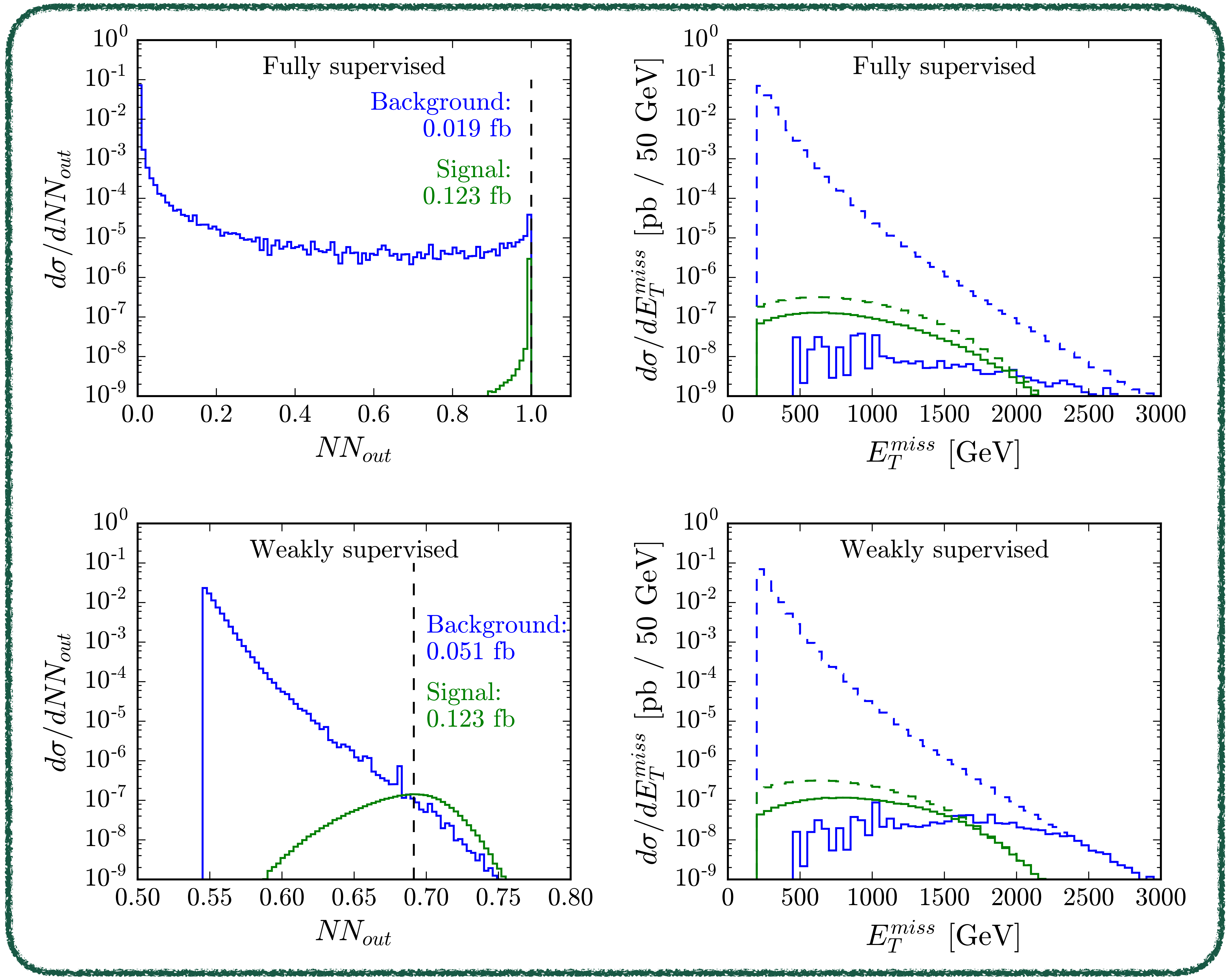


Not
lum

Need much more
be obtained with t
background event

Evaluate networks trained on “bad” background using new background samples

LHC Scenario



Cuts on network to give same signal count for both full and weak supervision.

Dashed = before cut
Solid = after cut

LHC Scenario

How to mimic the effects of mis-modeling?

LHC Scenario

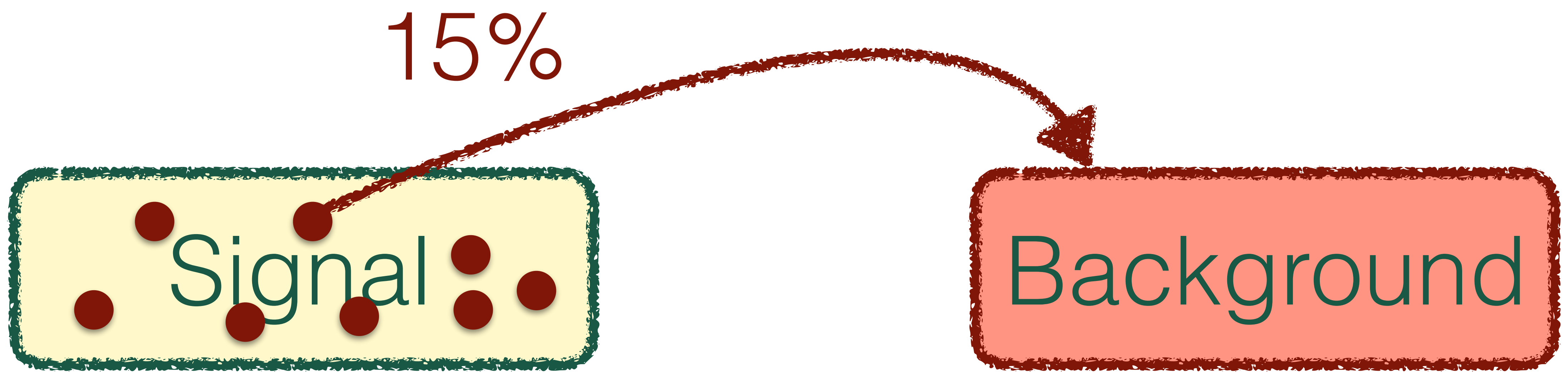
How to mimic the effects of mis-modeling?

Signal

Background

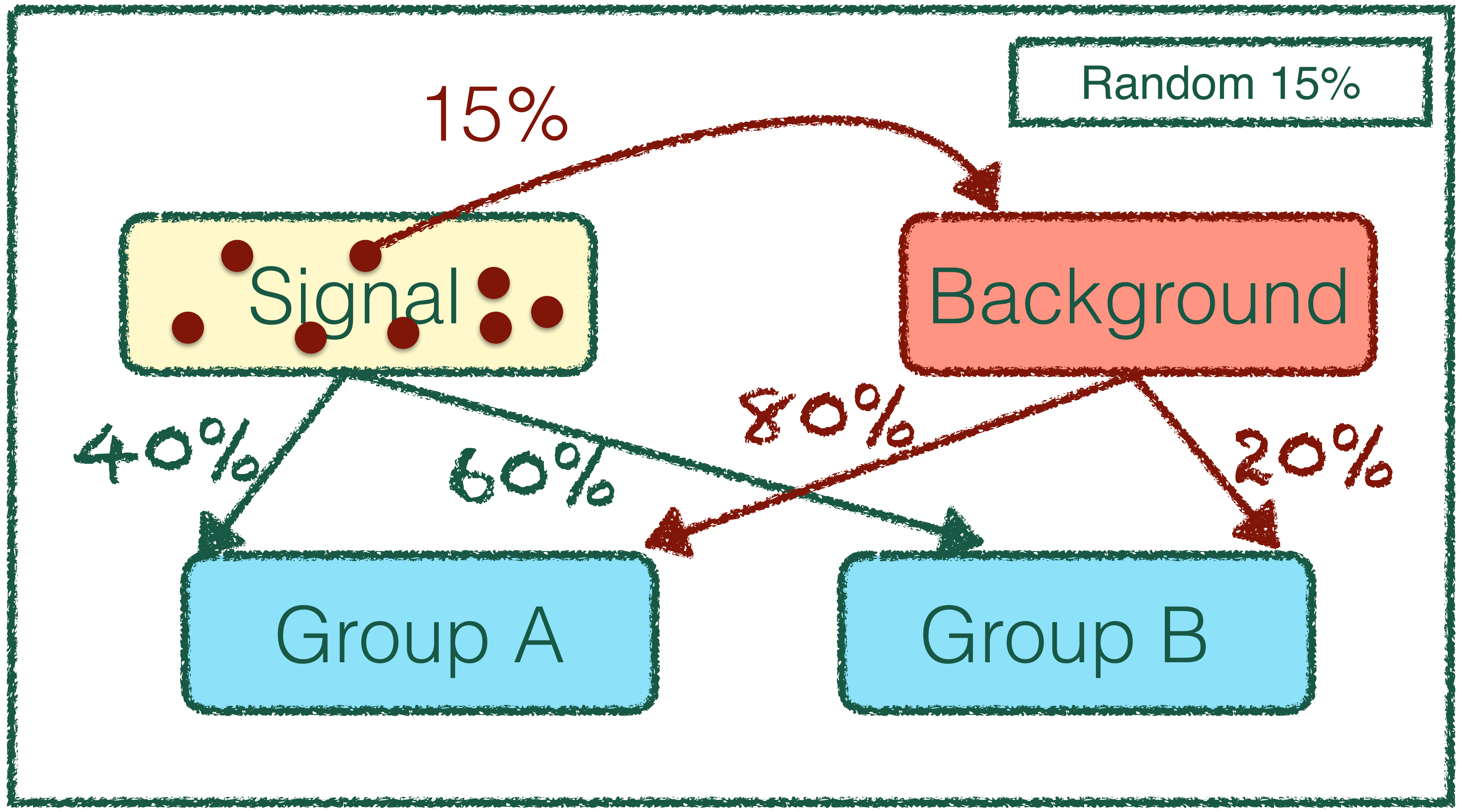
LHC Scenario

How to mimic the effects of mis-modeling?



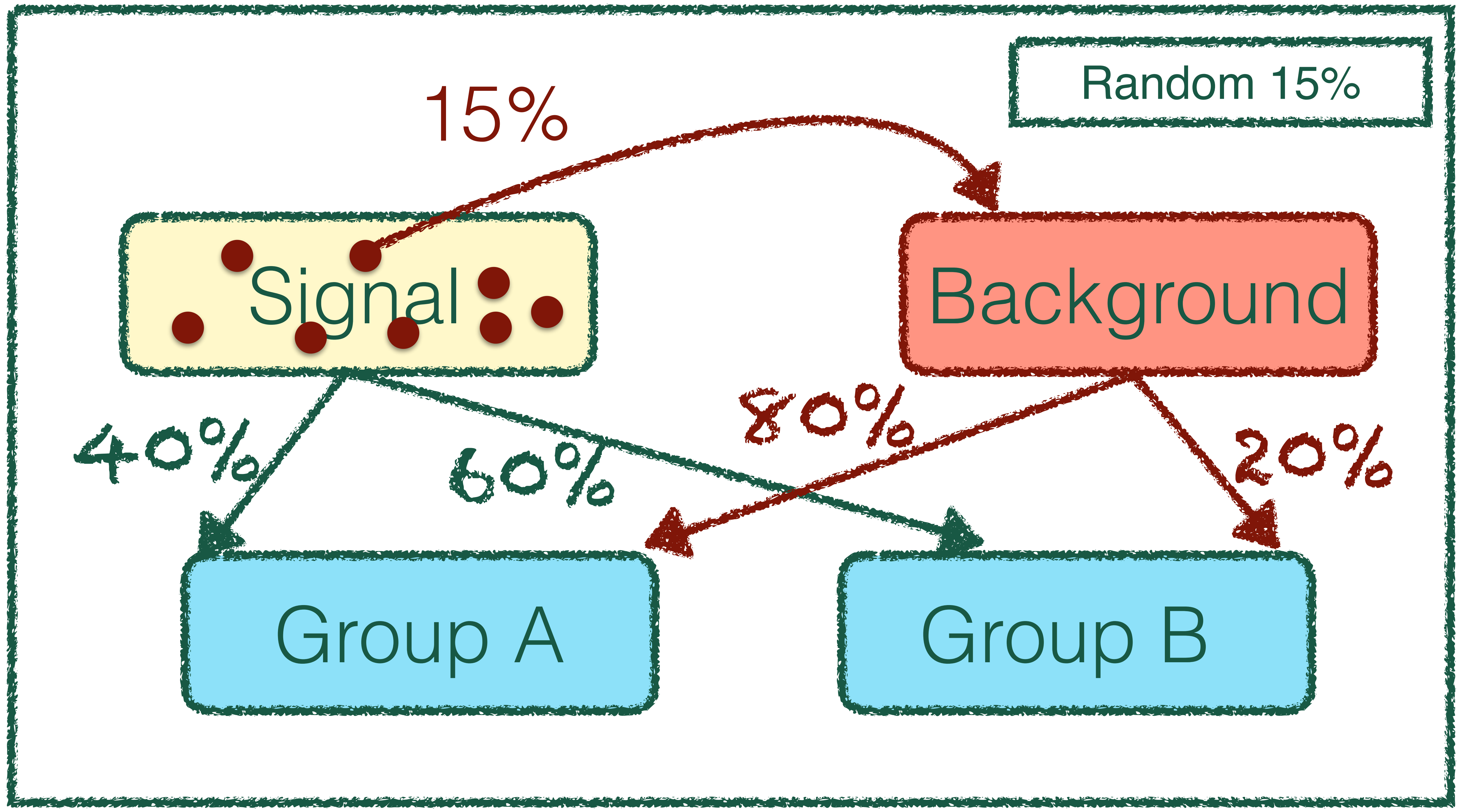
LHC Scenario

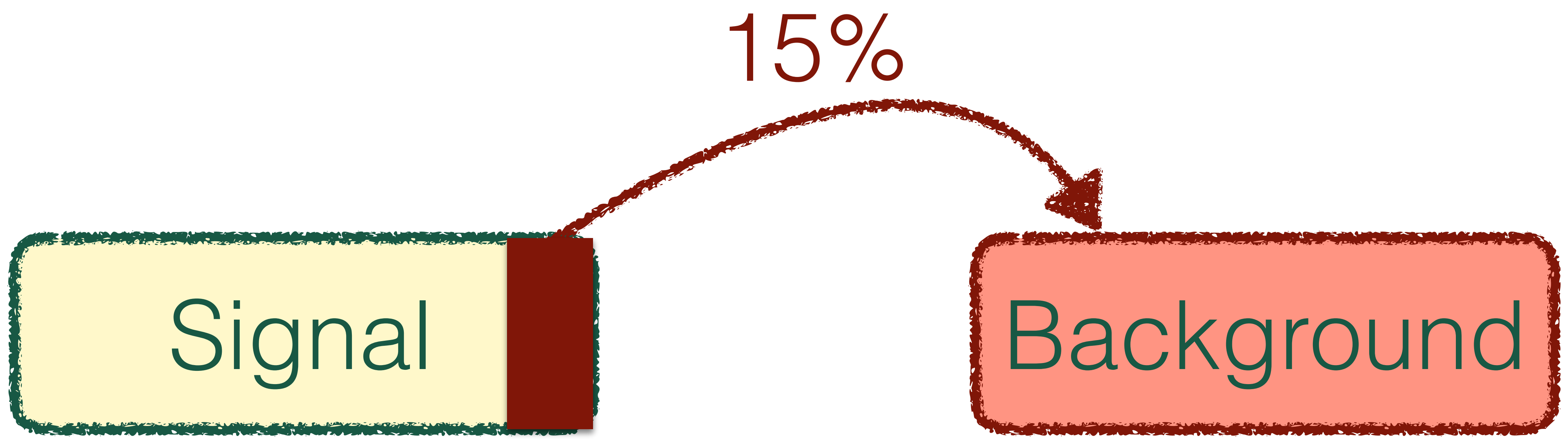
How to mimic the effects of mis-modeling?



LHC Scenario

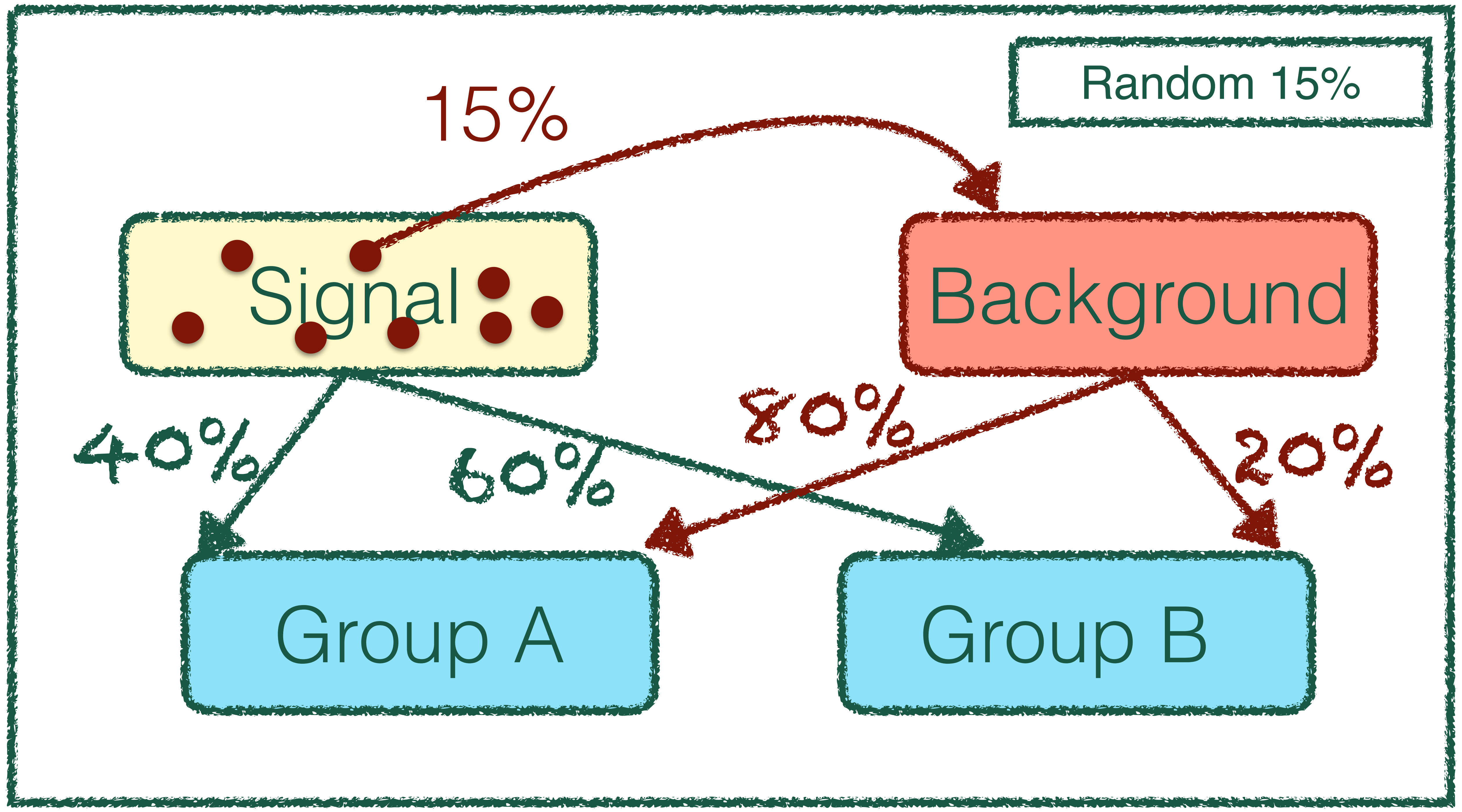
How to mimic the effects of mis-modeling?

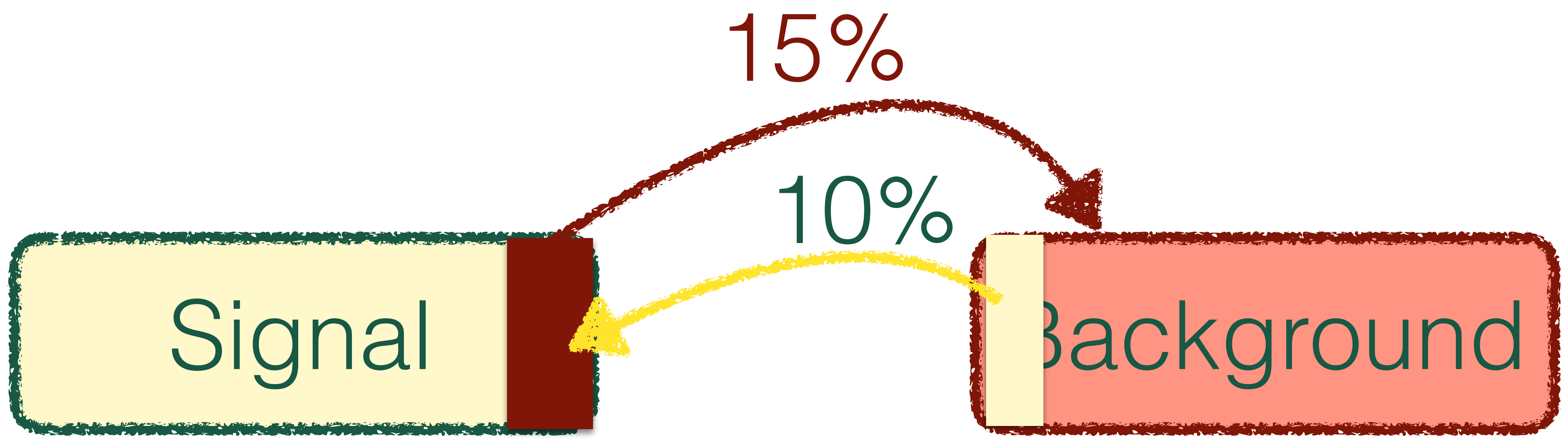




LHC Scenario

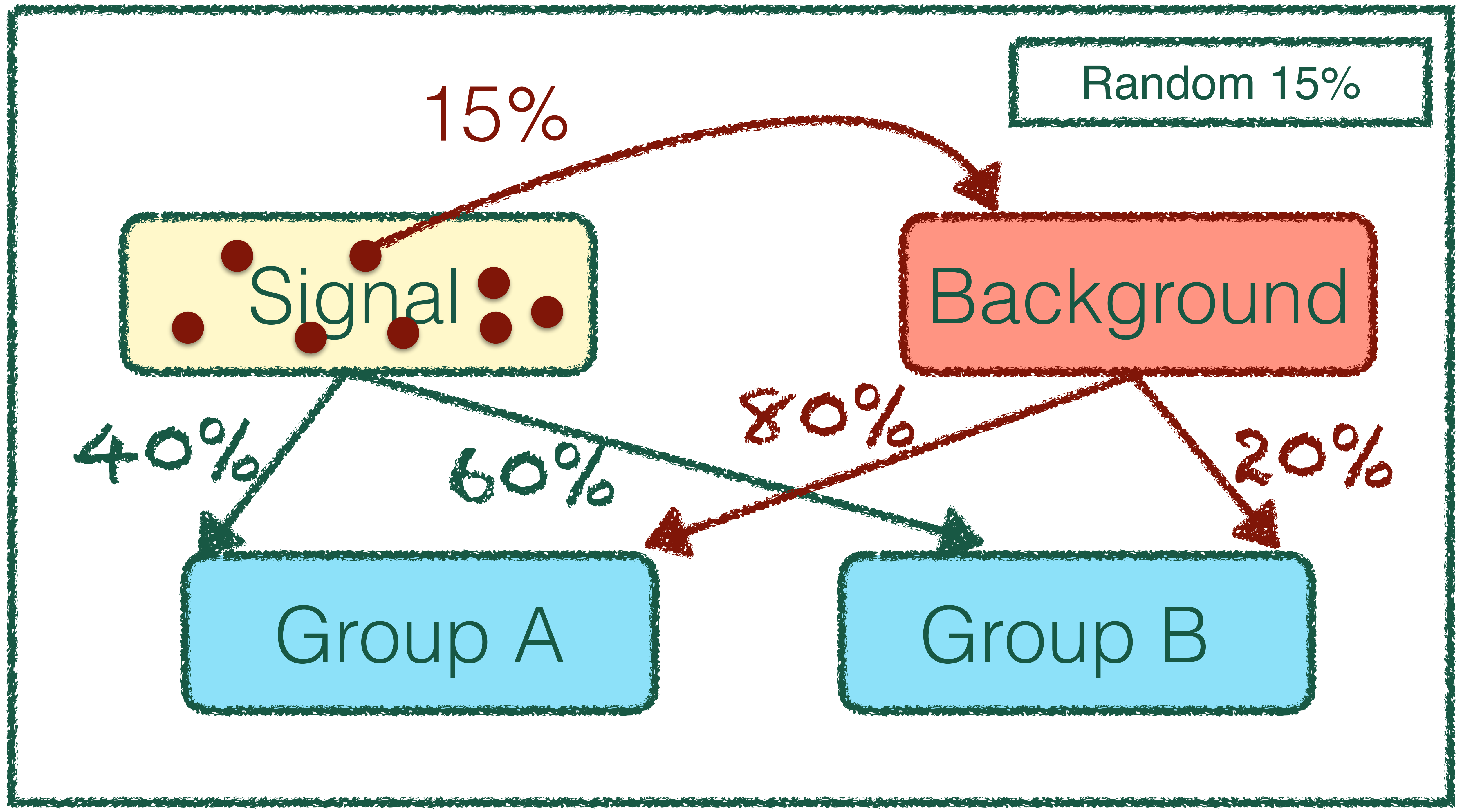
How to mimic the effects of mis-modeling?

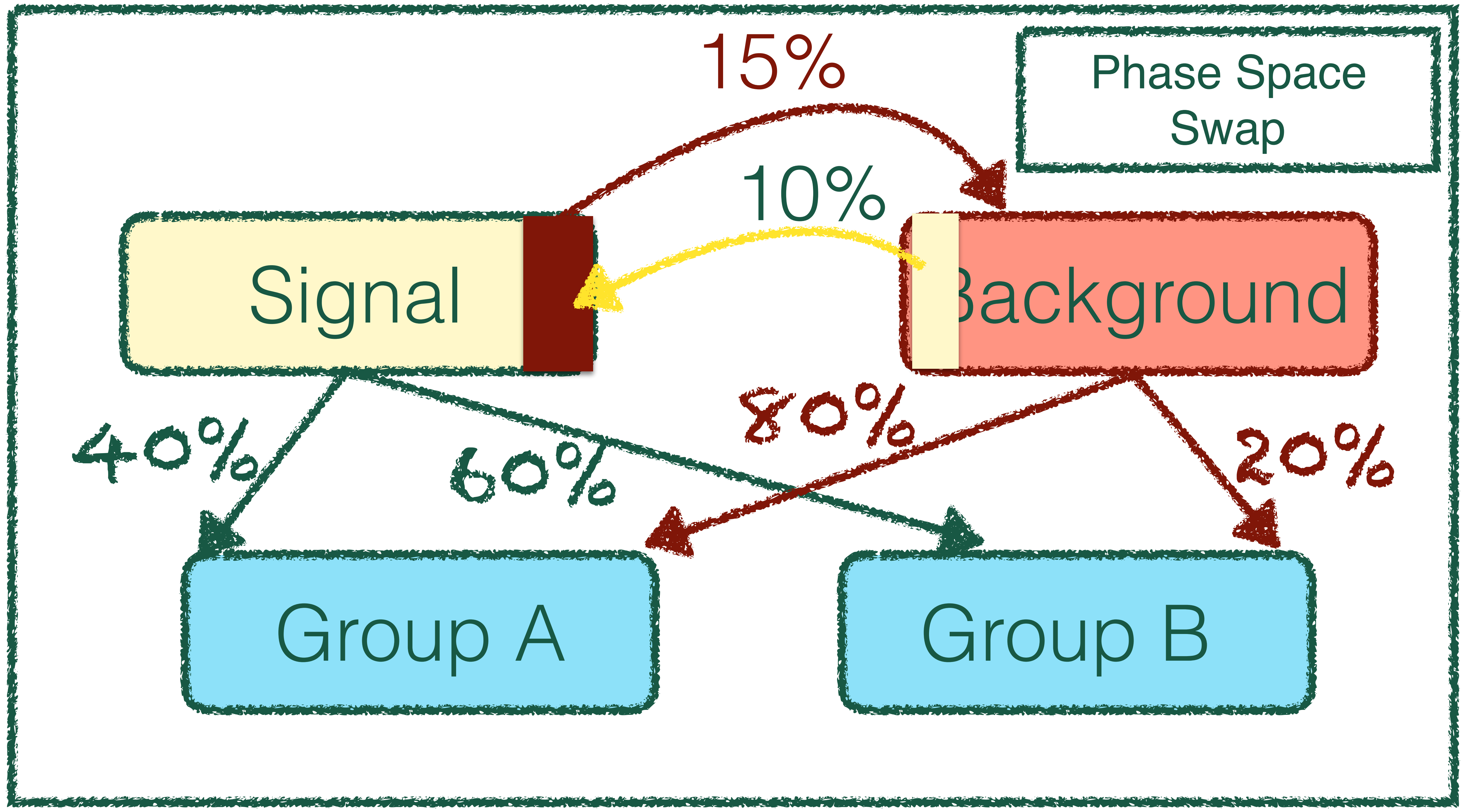




LHC Scenario

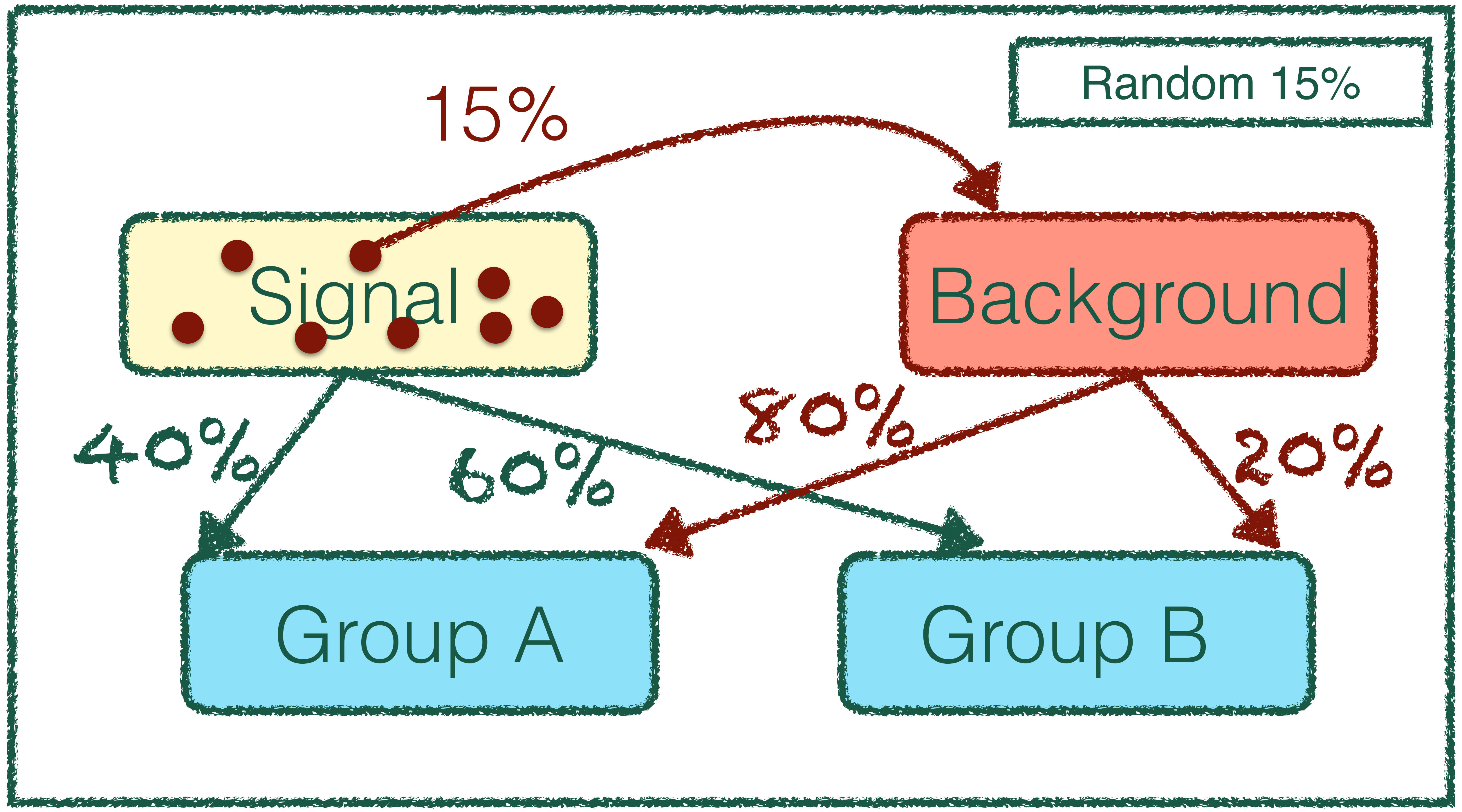
How to mimic the effects of mis-modeling?

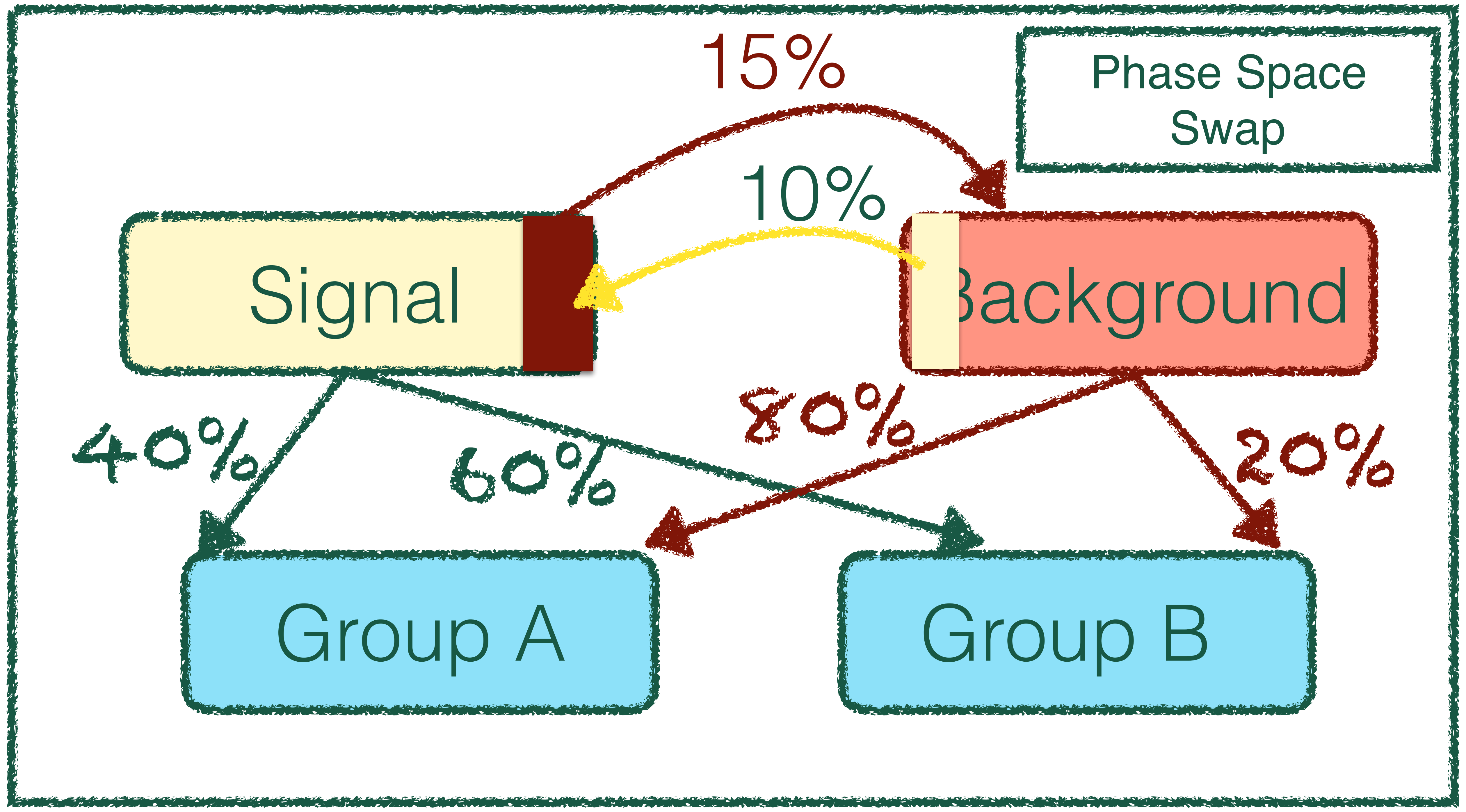




LHC Scenario

How to mimic the effects of mis-modeling?



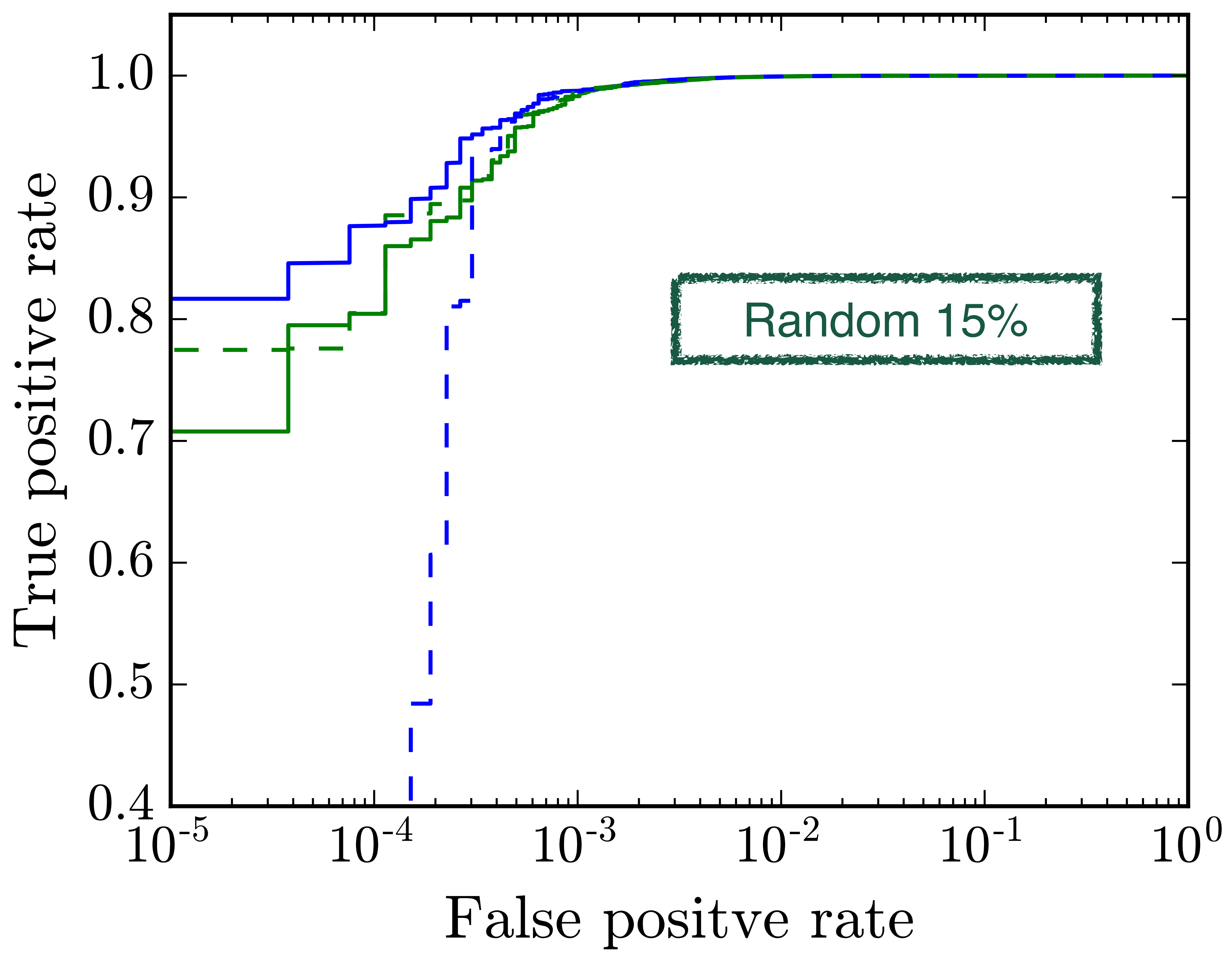


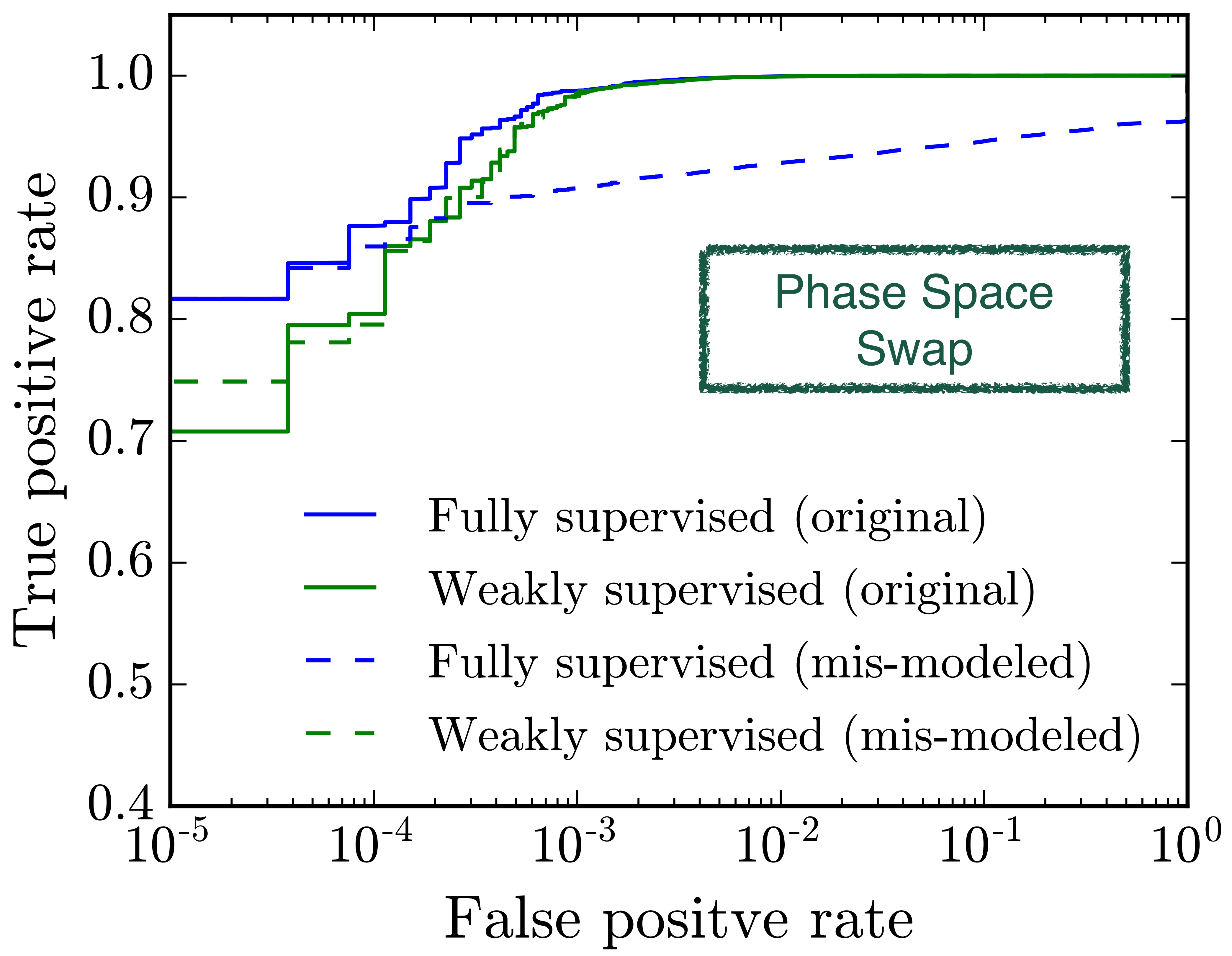
Dataset	f_t label		
	Original	Random 15%	Phase space swap
A	0.472	0.374 (0.585)	0.416 (0.593)
B	0.843	0.782 (0.769)	0.810 (0.747)

LHC Scenario

How to mimic the effects of mis-modeling?

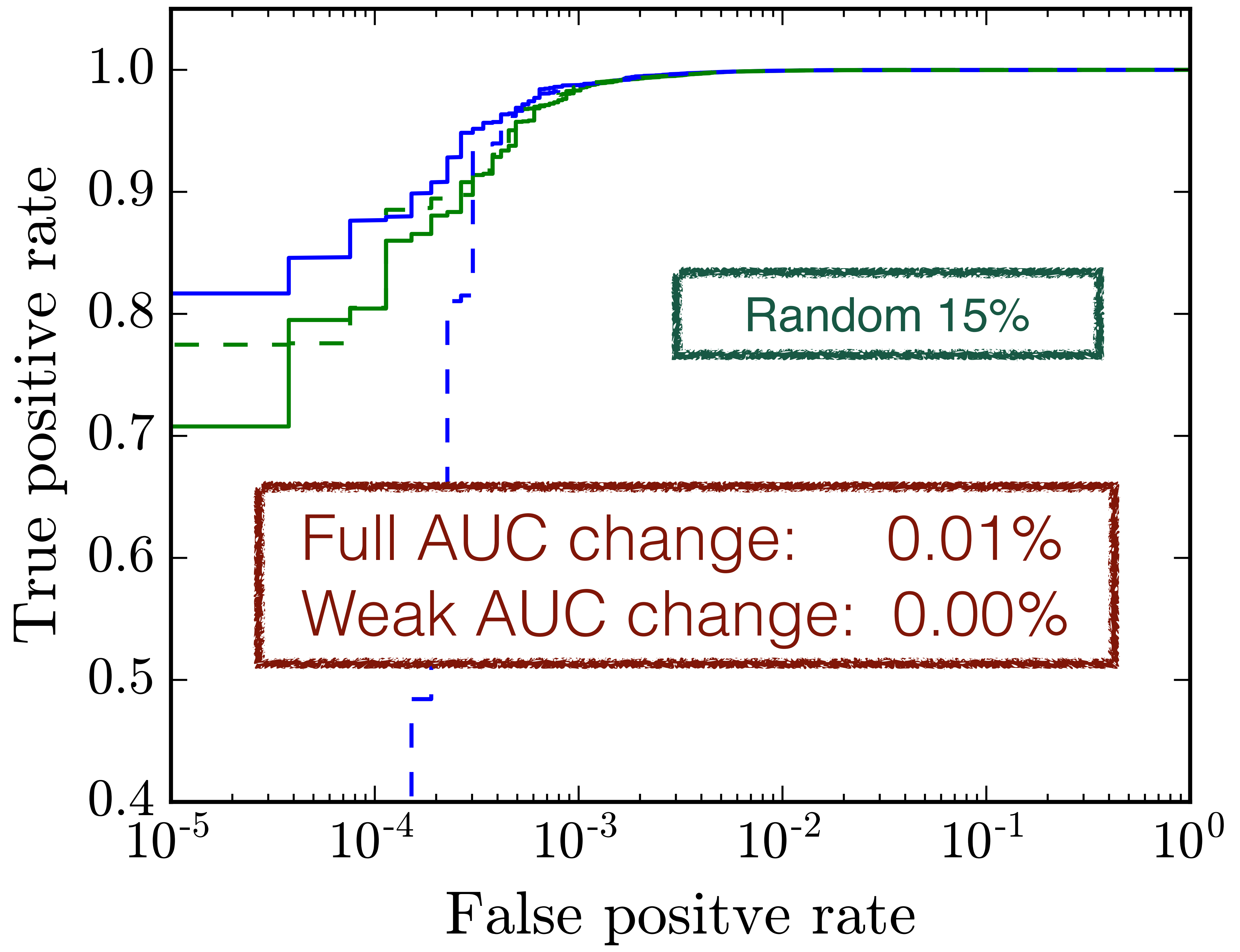
LHC Scenario

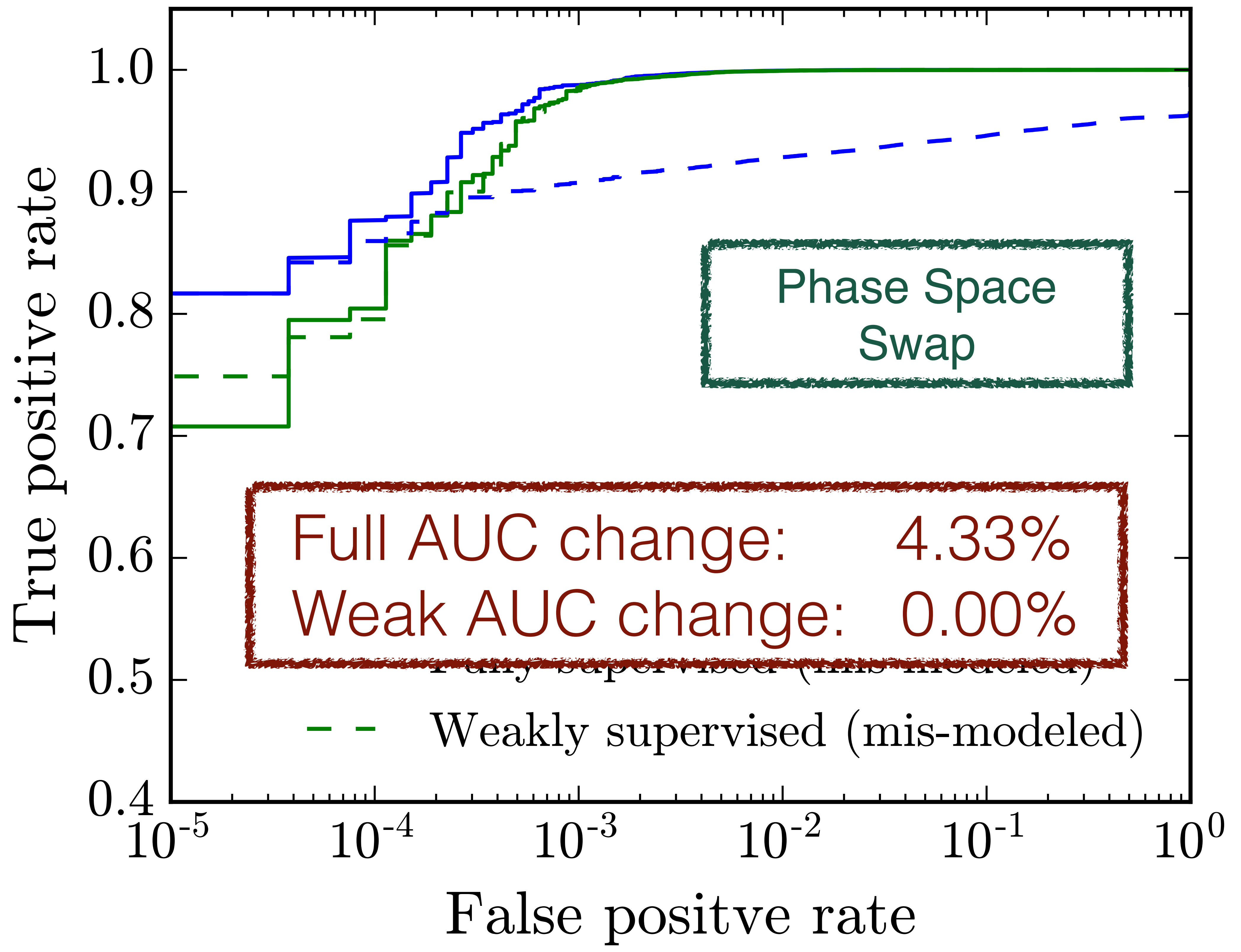




	f_t label		
Dataset	Original	Random 15%	Phase space swap
A	0.472	0.374 (0.585)	0.416 (0.593)
B	0.843	0.782 (0.769)	0.810 (0.747)

LHC Scenario





Dataset	f_t label		
	Original	Random 15%	Phase space swap
A	0.472	0.374 (0.585)	0.416 (0.593)
B	0.843	0.782 (0.769)	0.810 (0.747)

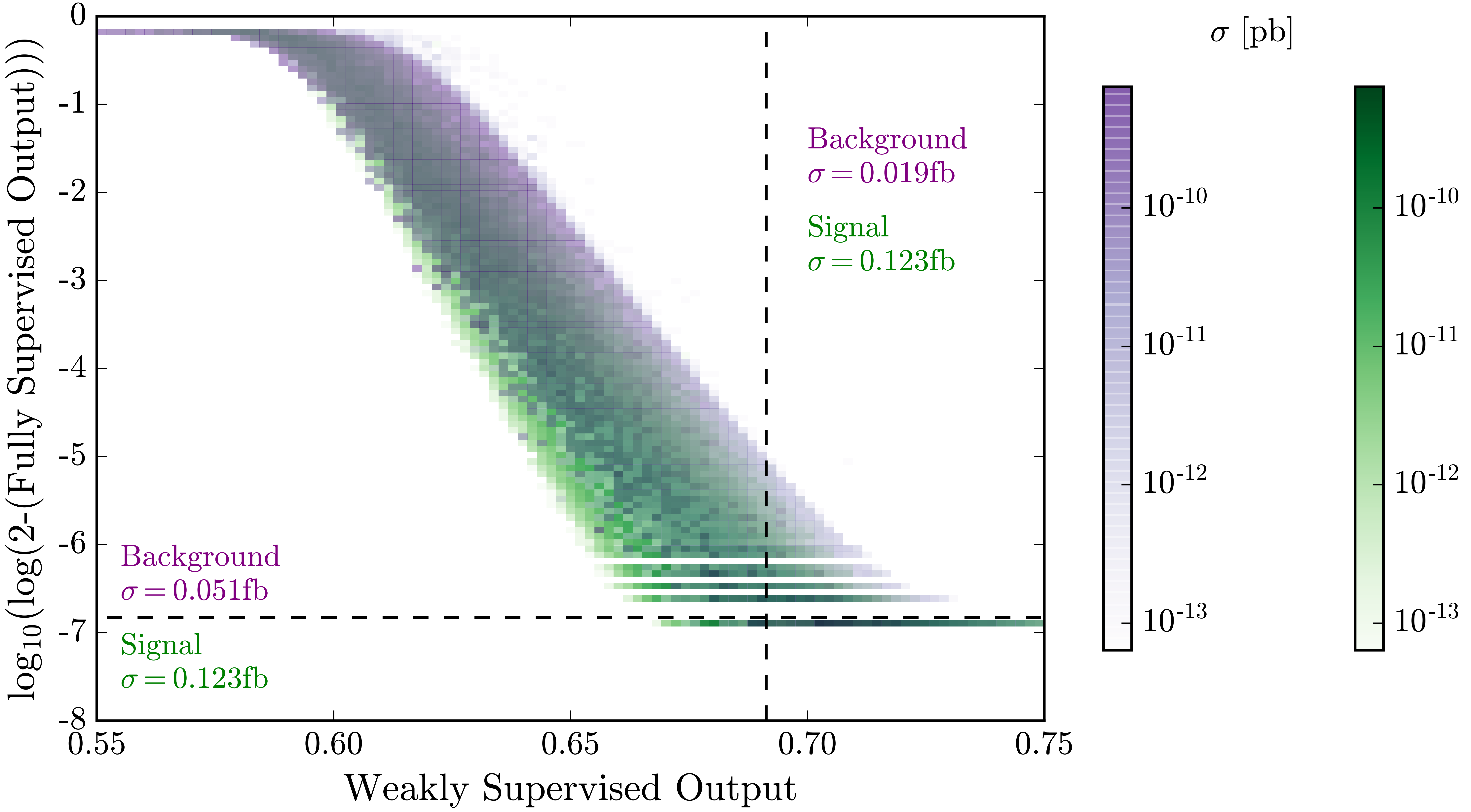
LHC Scenario

Both full and weak supervision capable of great classification

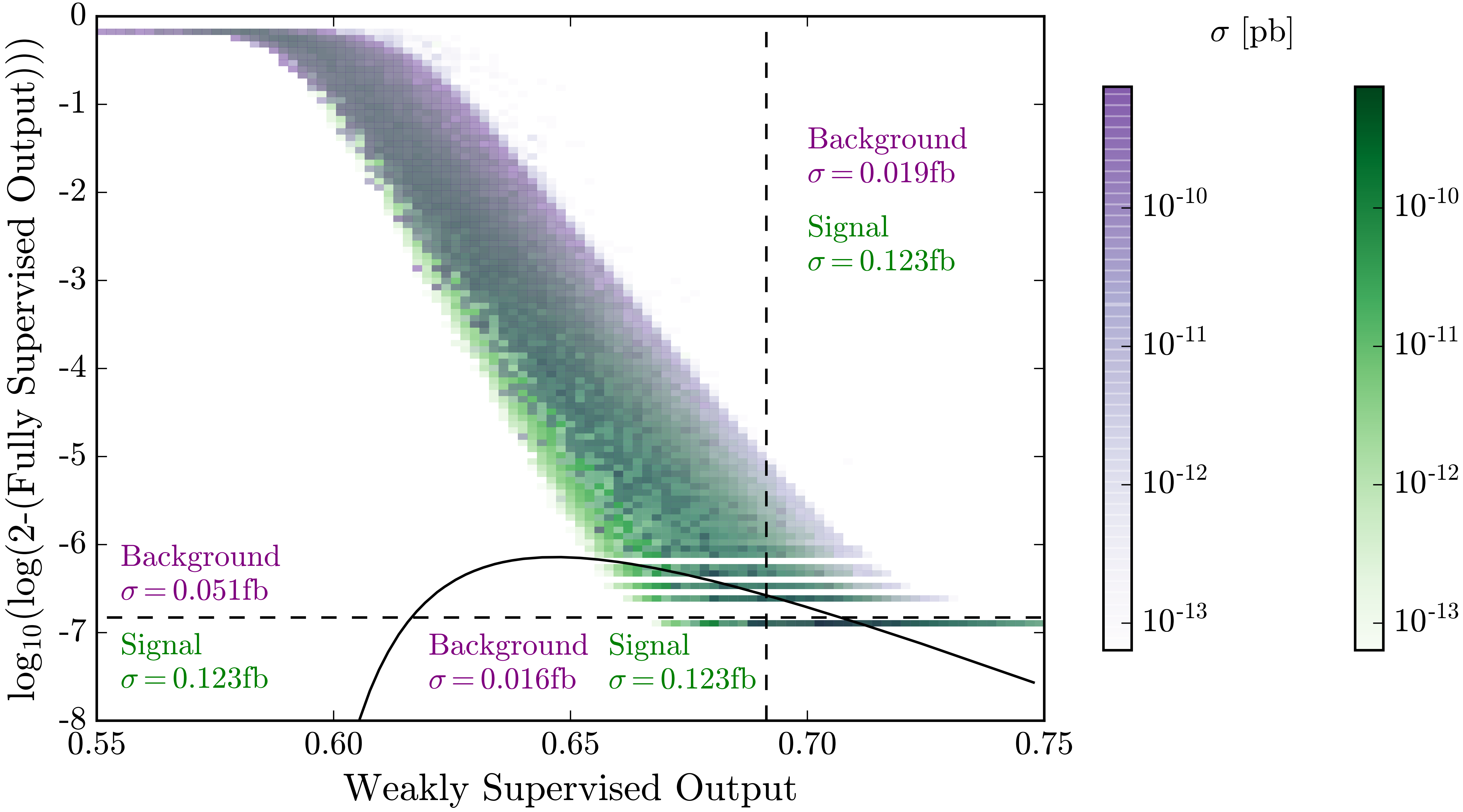
Same network architecture, training method, and loss function are used

Event-by-event do both networks yield same classification?

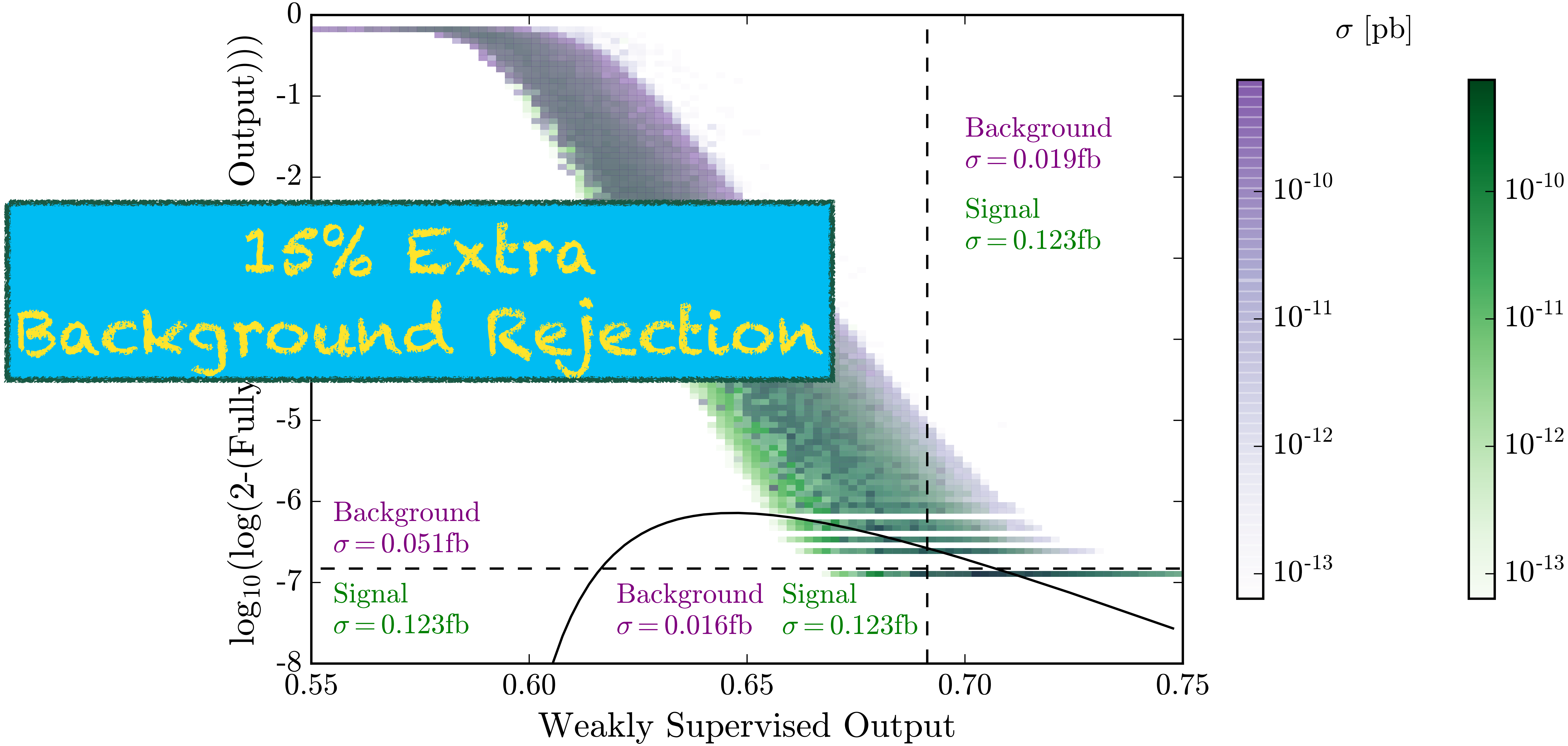
LHC Scenario



LHC Scenario

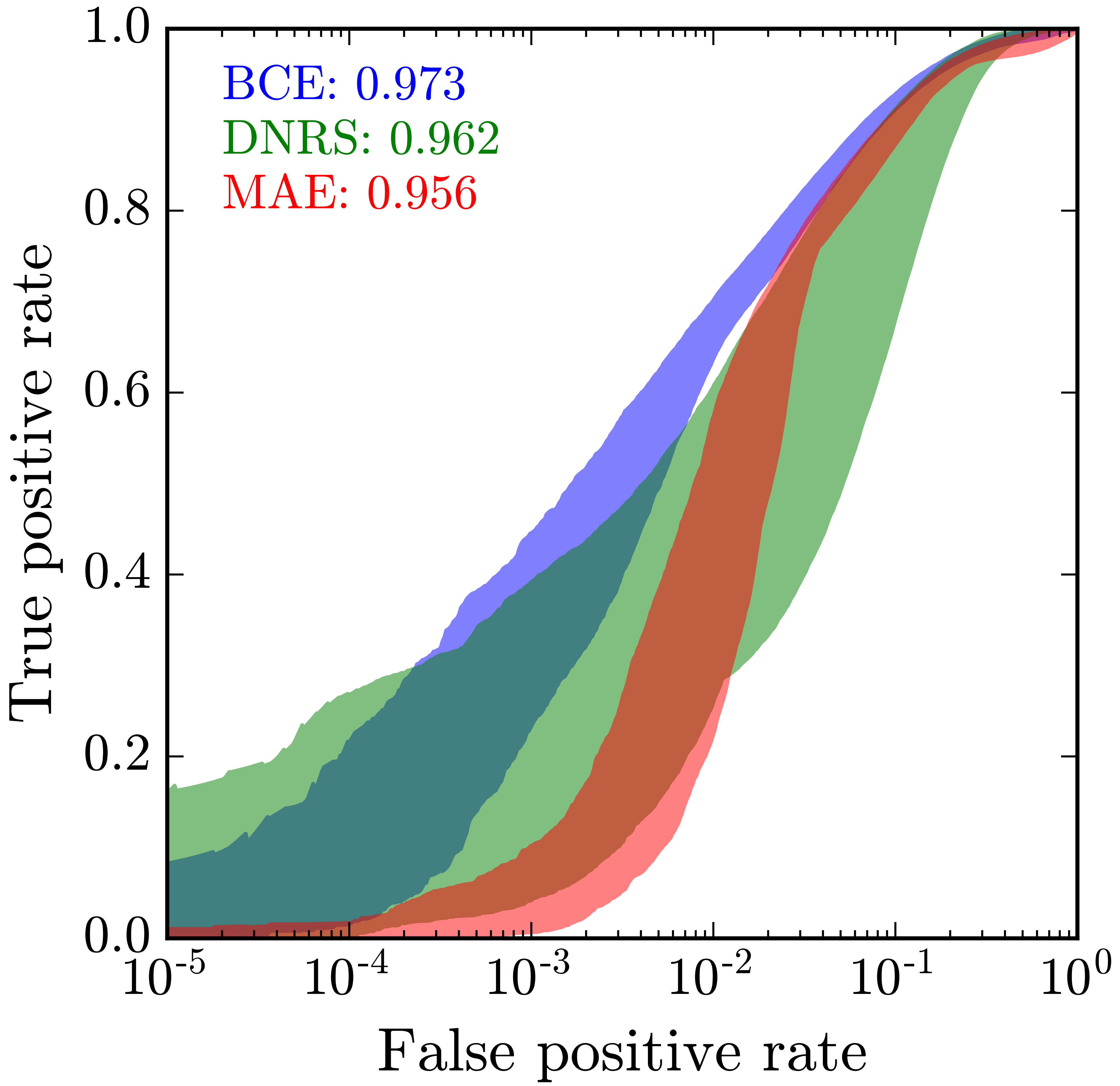


LHC Scenario



LHC Scenario

Choice of loss function matters



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

TABLE I. The essential pros (✓), cons (✗), and open questions (?) of the CWoLa and LLP weak supervision paradigms.

Komiske, Metodiev, Nachman, and Schwartz [1801.10158]

Conclusion

Highlights

- Weak supervision: training on mixed data sets.
- Closer model of quantum reality.
- Robust to mismodeling.
- Analytic arguments.

Open Questions

- Why does weak supervision not match full supervision performance?
- Can weak supervision be done on multi-class problems?
- Particular LHC scenarios which would work?
- How to validate?
- Can this work with small amounts of data?