Deep Learning for High Energy Physics

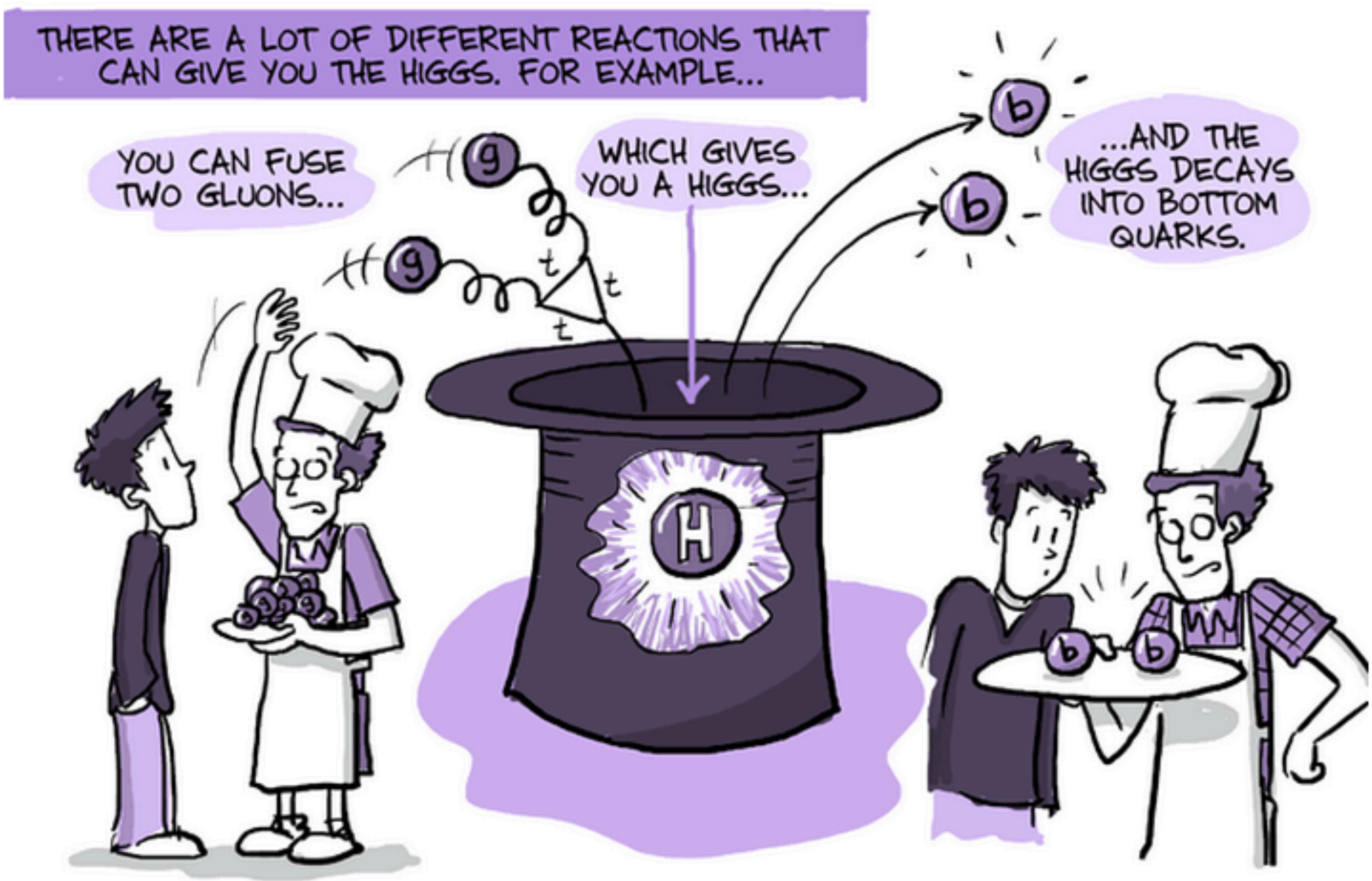
Daniel Whiteson, UC Irvine
Oct 2019, LBL

What is Deep Learning?

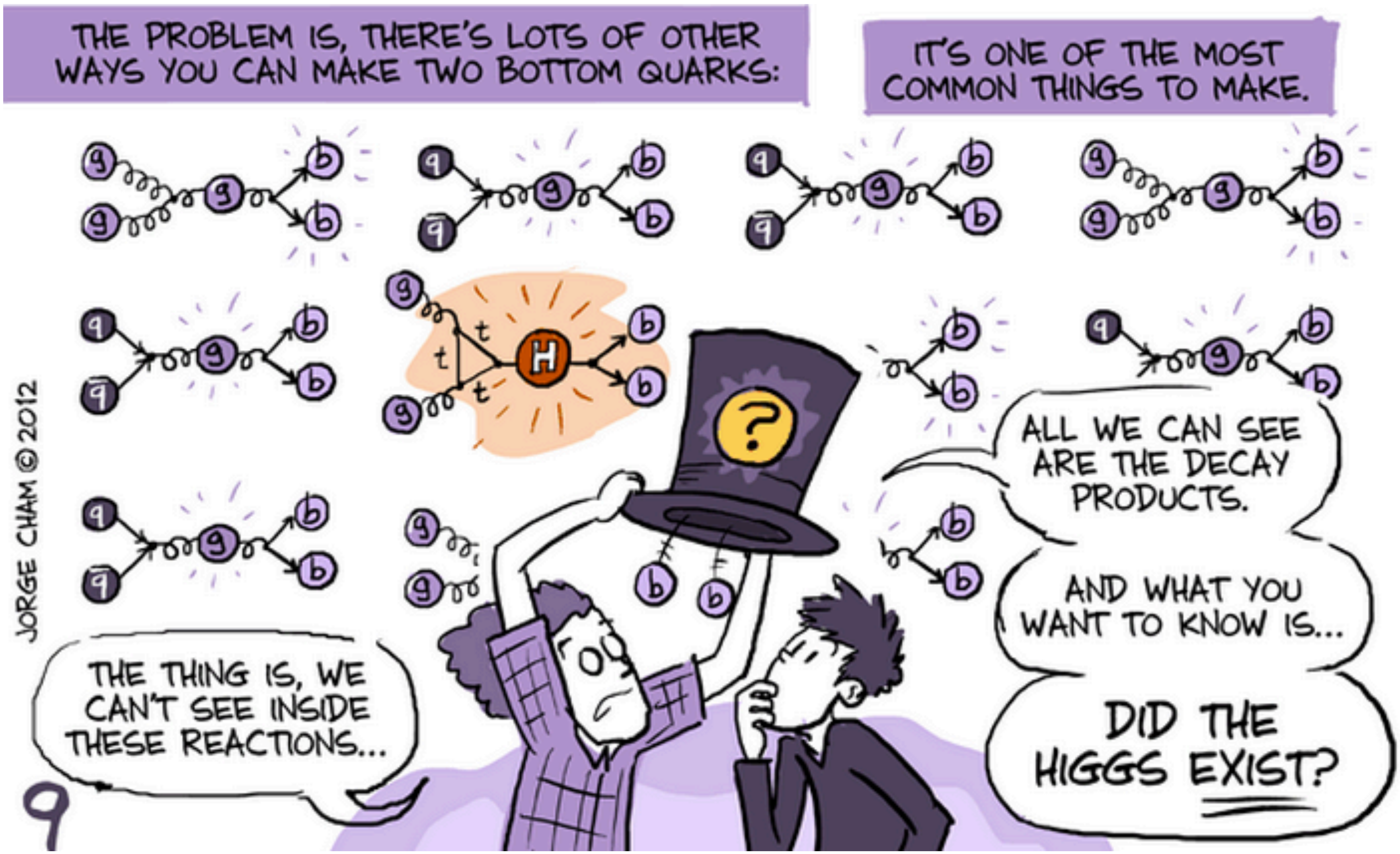
Unambiguous data

Ok, but see:
http://cerncourier.com/cws/article/cern/54388

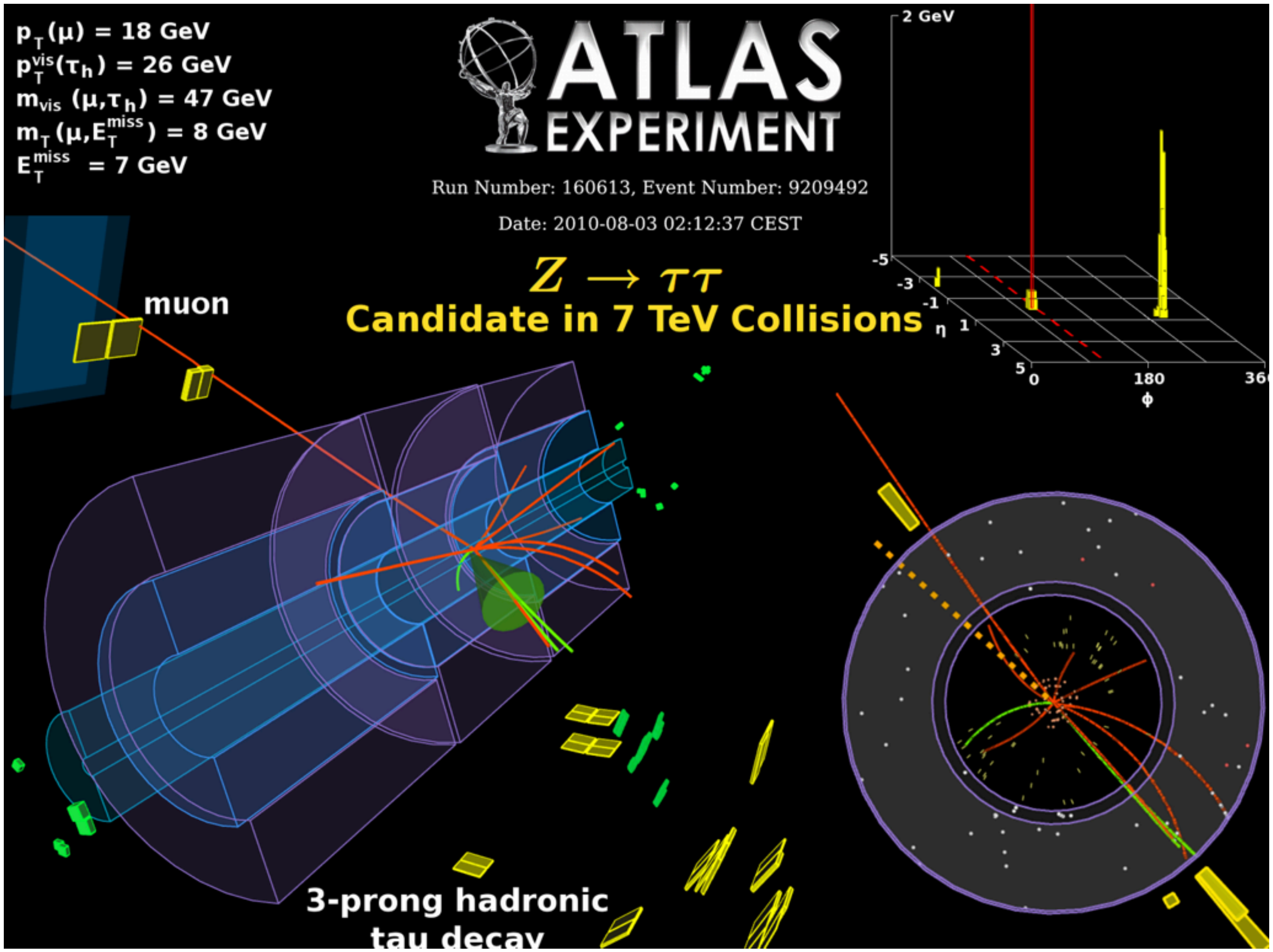
Making a new particle



Backgrounds



Why statistics?

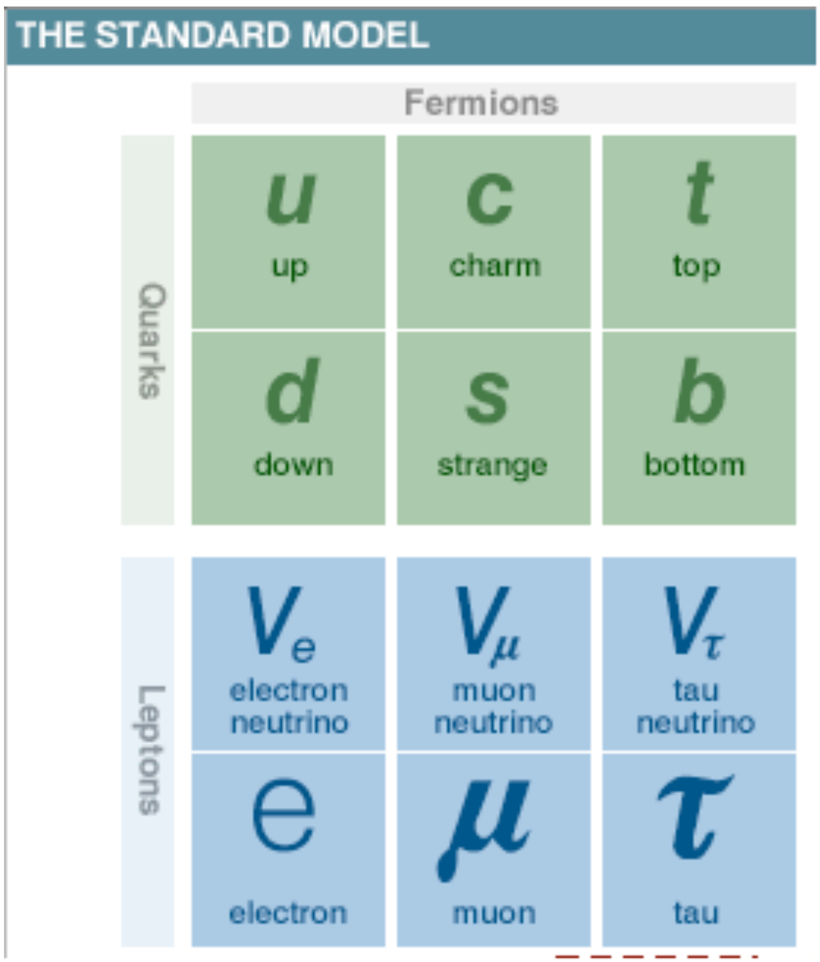


The nature of our data demands it.

Hypothesis testing

To search for a new particle, we compare the predictions of two hypotheses:

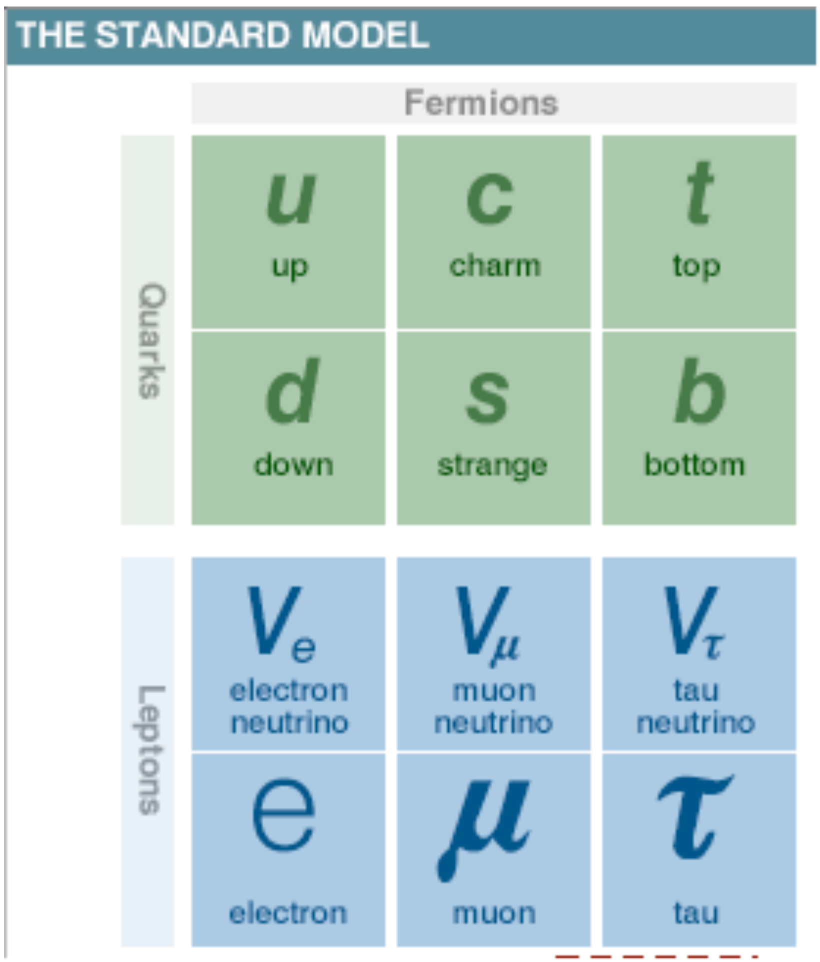
1.



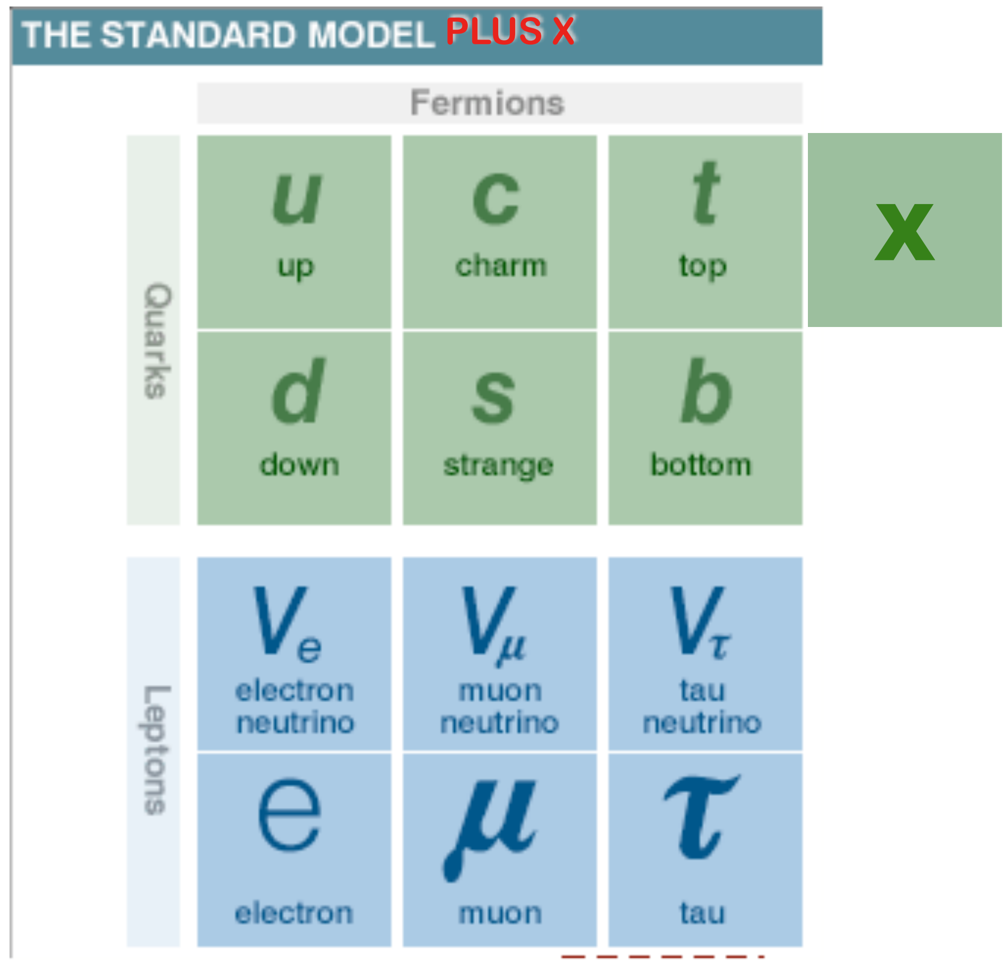
Hypothesis testing

To search for a new particle, we compare the predictions of two hypotheses:

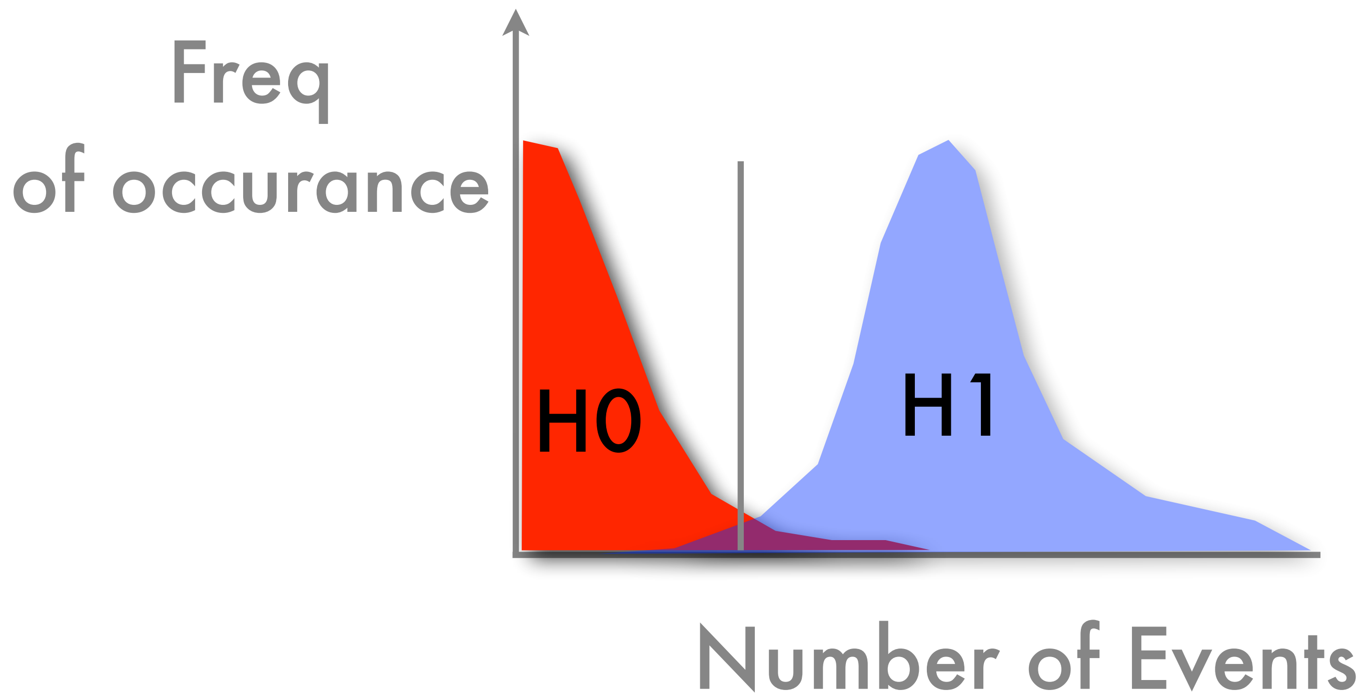
1.



2.

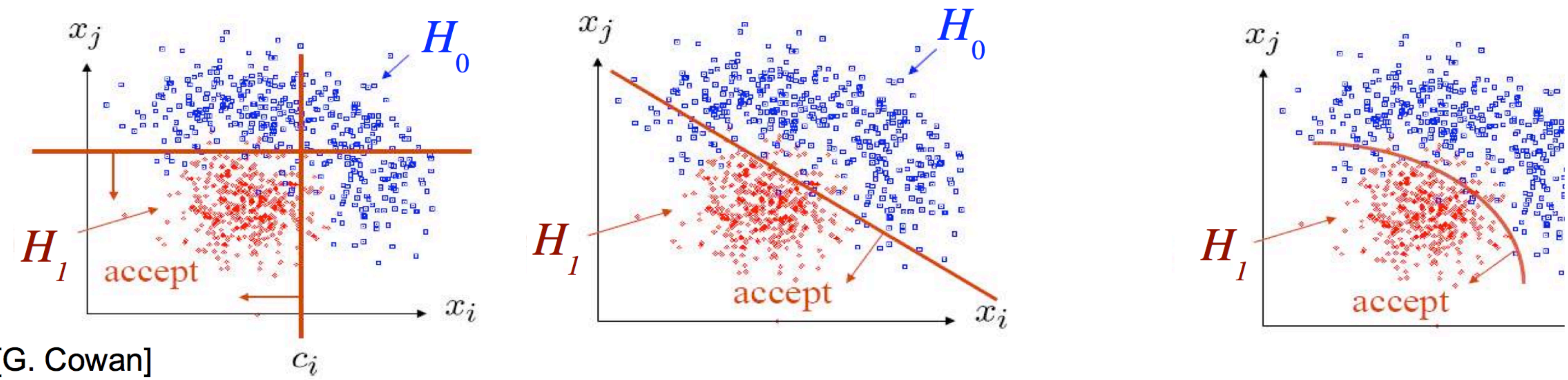


Example



A threshold makes sense.
Choice of position balances
false vs missed discovery

More complicated



[G. Cowan]

Neyman-Pearson

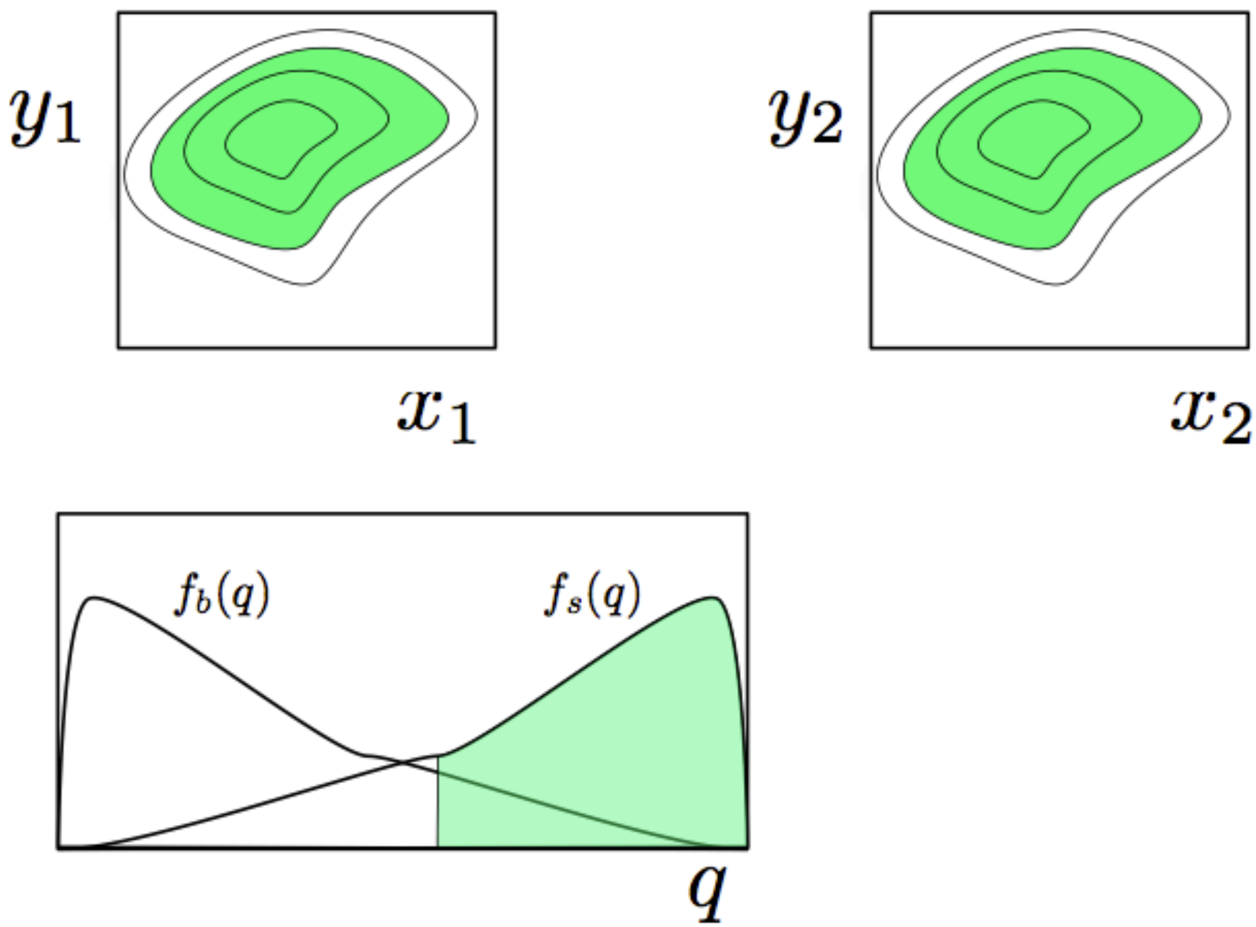
NP lemma says that the best decision boundary is the likelihood ratio:

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

(Gives smallest missed discovery rate for fixed false discovery rate)

What does this do?

Finds a region in variable space



(K. Cranmer)

No problem

Fairly straightforward

if you can calculate

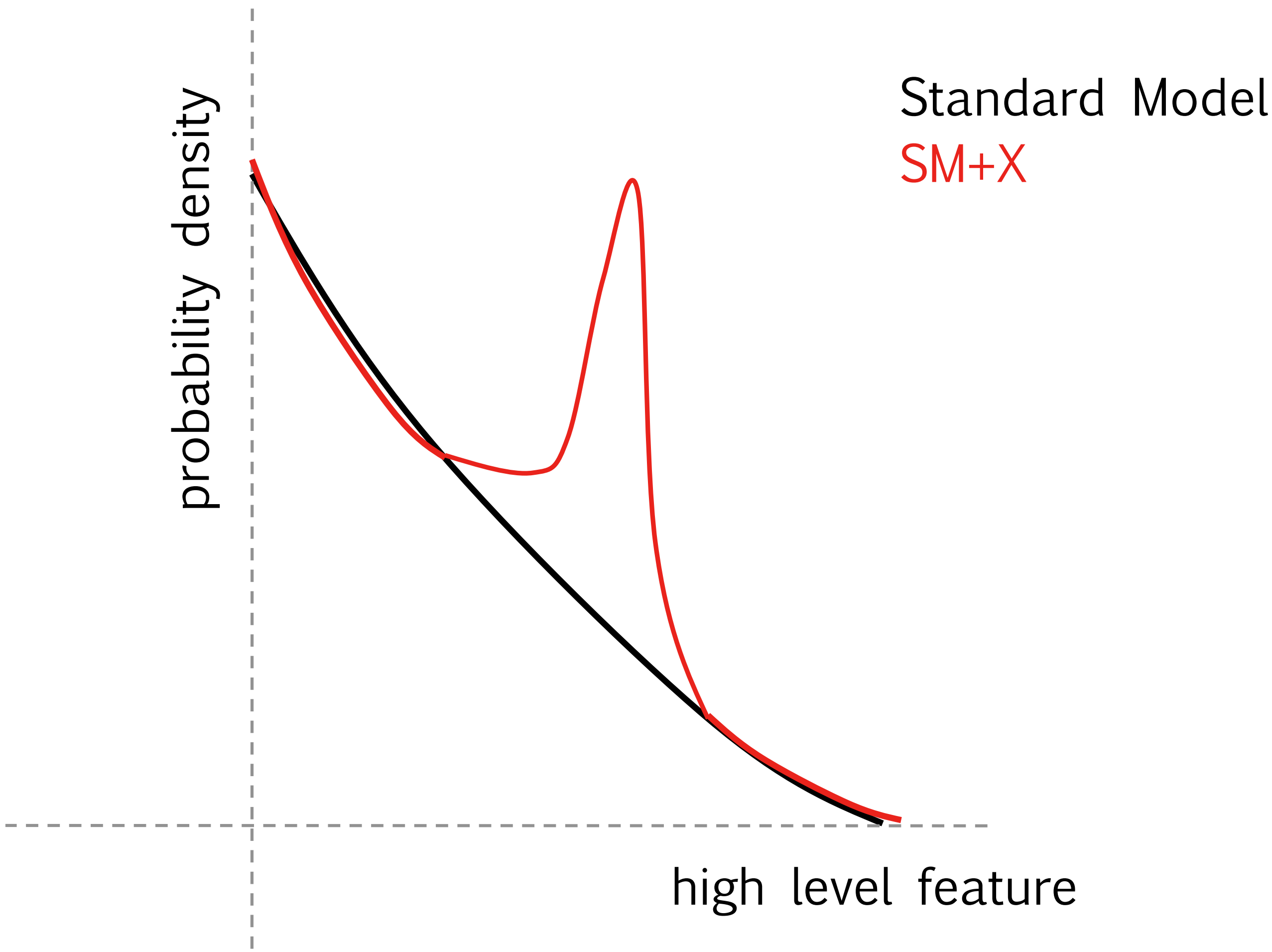
$$\frac{P(x|H_1)}{P(x|H_0)}$$

or generally

P(data|theory)

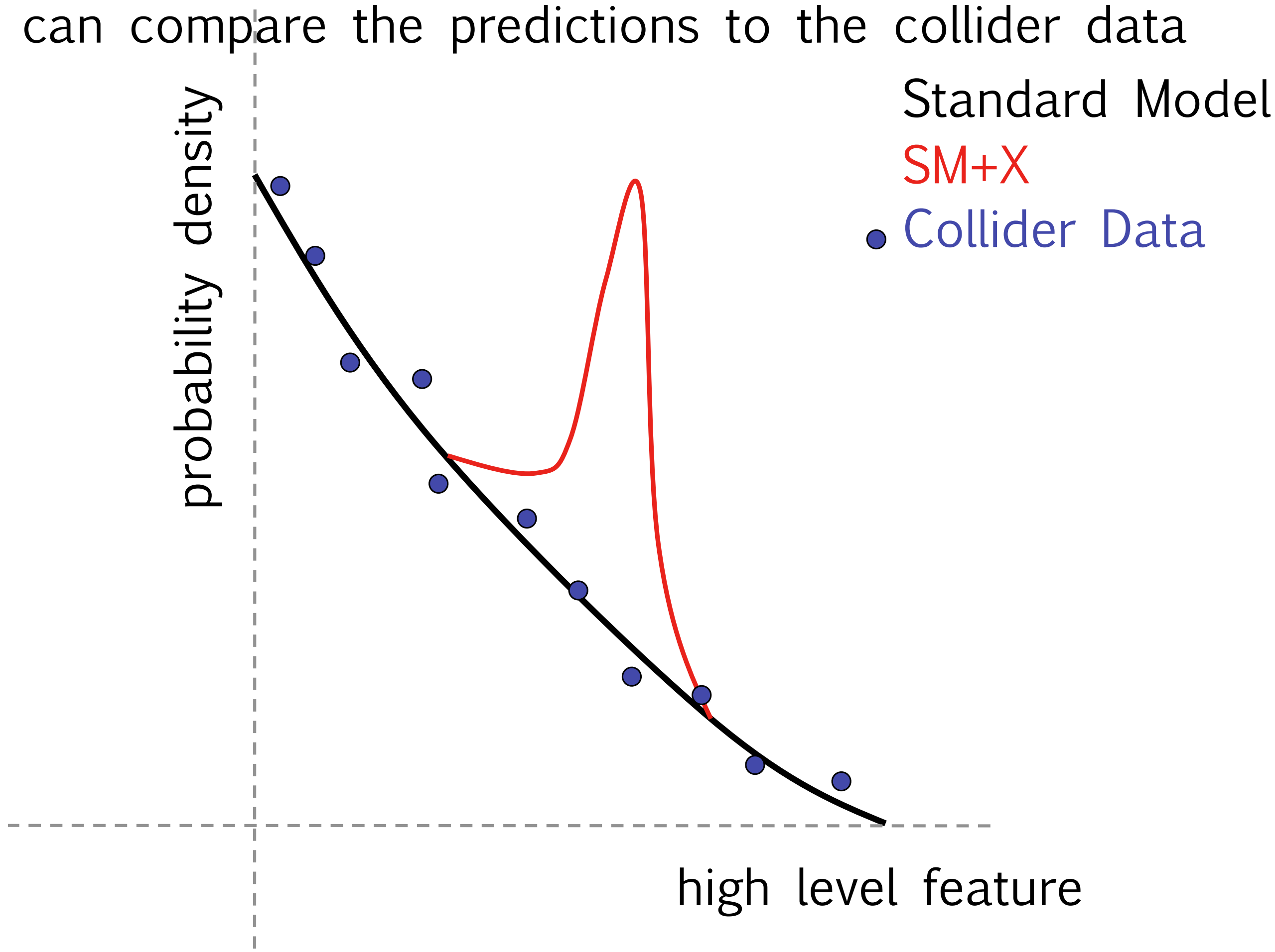
Hypothesis Testing

Sometimes this is easy



Hypothesis Testing

We can compare the predictions to the collider data



Which can tell us which hypothesis is preferred via a likelihood ratio:

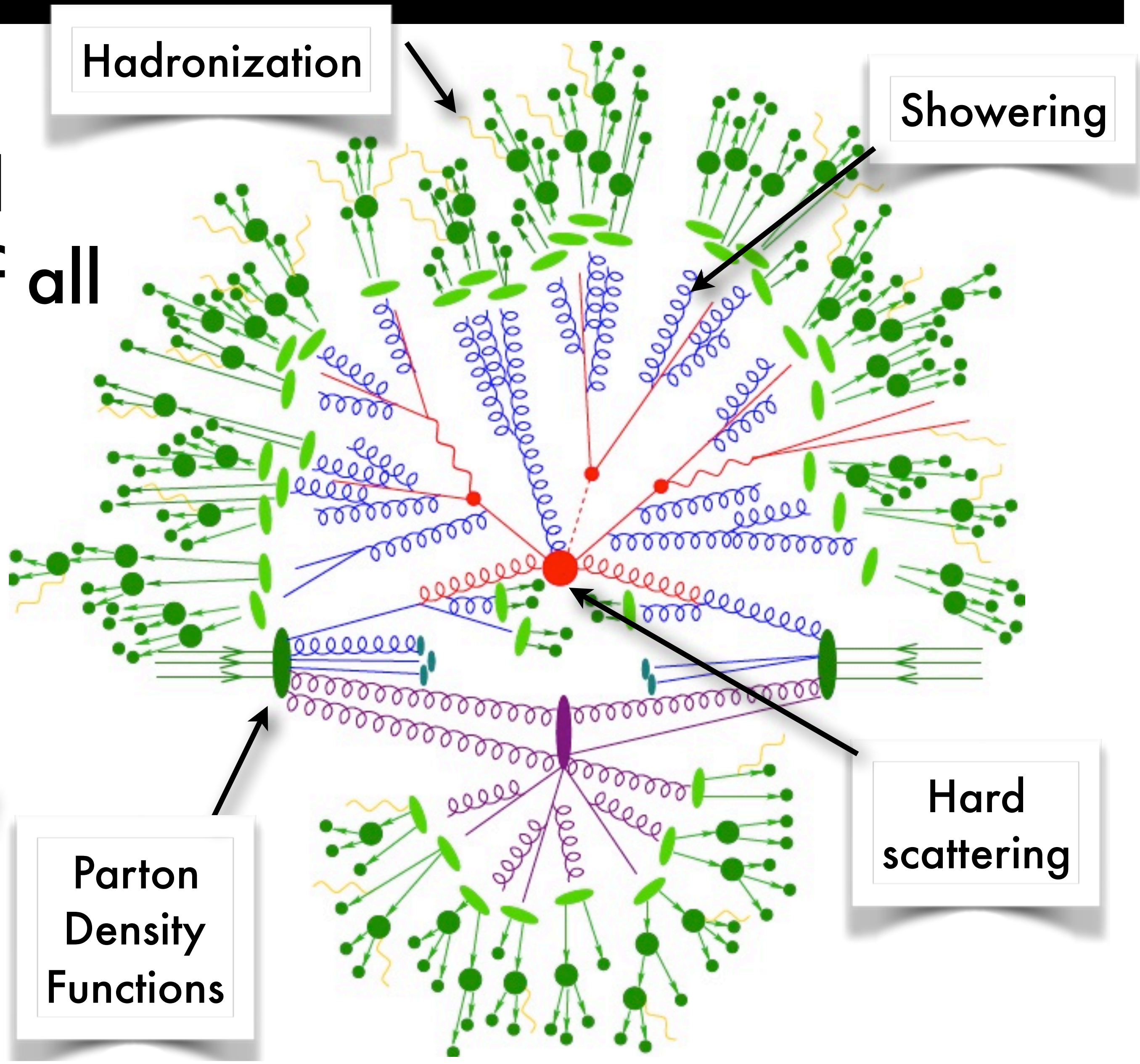
$$\frac{L_{SM+X}}{L_{SM}} = \frac{P(\text{data} \mid SM+X)}{P(\text{data} \mid SM)}$$

In general

We have a good understanding of all of the pieces

Do we have

P(data|theory)?



In general

What would

$P(\text{data}|\text{theory})$

look like?

The dream

Detector Response

p(data | final-state particles P)

Hadronization

x p(final state particles P | showered particles S)

Showering

x p(showered particles S | hard scatter products M)

x p(hard scatter products M | theory)

Hard scattering

Parton Density Functions

Sum over all … e intermediate P,S,M

The dream

p(hard scatter products M | theory)

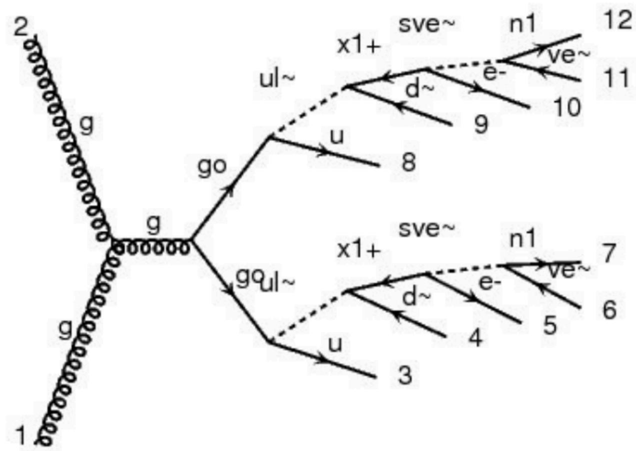
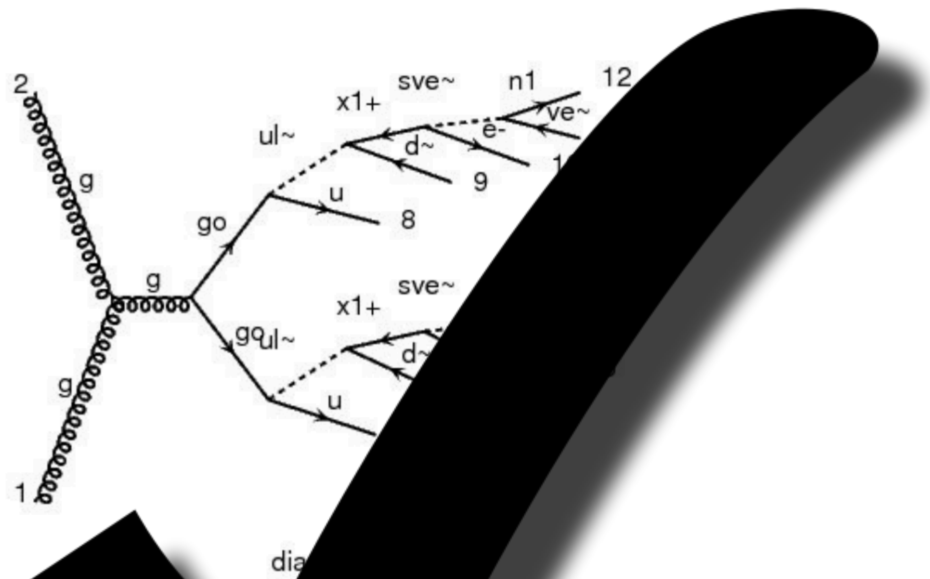


diagram 1

Theory well defined
automatic calculators exist
for almost any (B)SM theory

The dream

p(hard scatter products M | theory)



The … defined

automat… …ulators exist

for almost any (B)SM theory

The nightmare

p(data | final-state particles P)

x p(final state particles P | showered particles S)

x p(showered particles S | hard scatter products M)

<u>We have</u>: solid understanding of microphysics
<u>We need</u>: analytic description of high-level physics

The solution

<u>We have</u>: solid understanding of microphysics
<u>We need</u>: analytic description of high-level physics
<u>But</u>: only heuristic lower-level approaches exist

Iterative simulation strategy, no overall PDF

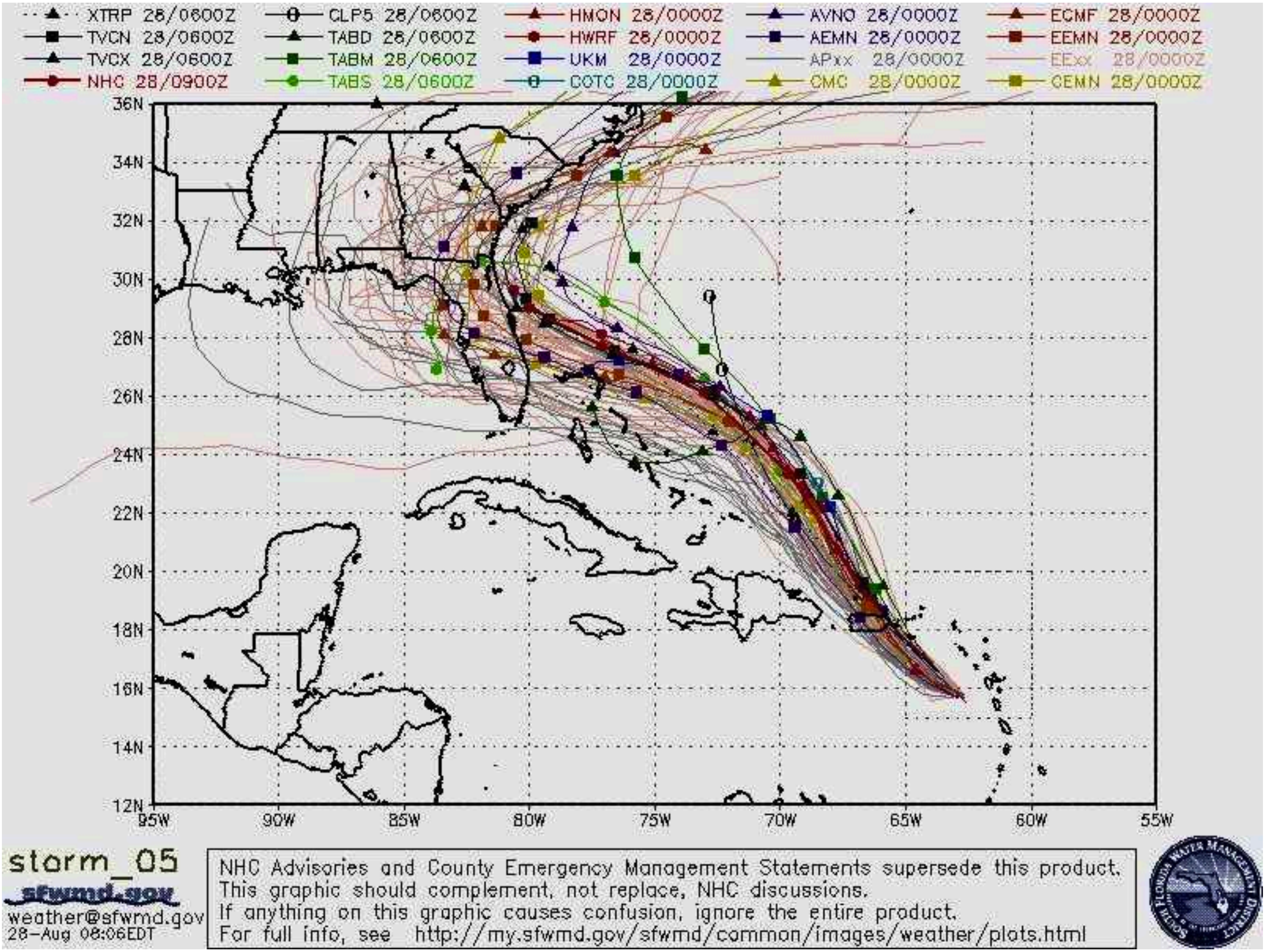
<u>Iterative approach</u>
(1) Draw events from $p(M|\text{theory})$
(2) add random showers
(3) do hadronization
(4) simulate detector

The solution

We have: solid understanding of microphysics
We need: analytic description of high-level physics
But: only heuristic lower-level approaches exist

Iterative simulation strategy, no overall PDF

What do we get
Arbitrarily large samples of events drawn from p(data|theory), but not the PDF itself

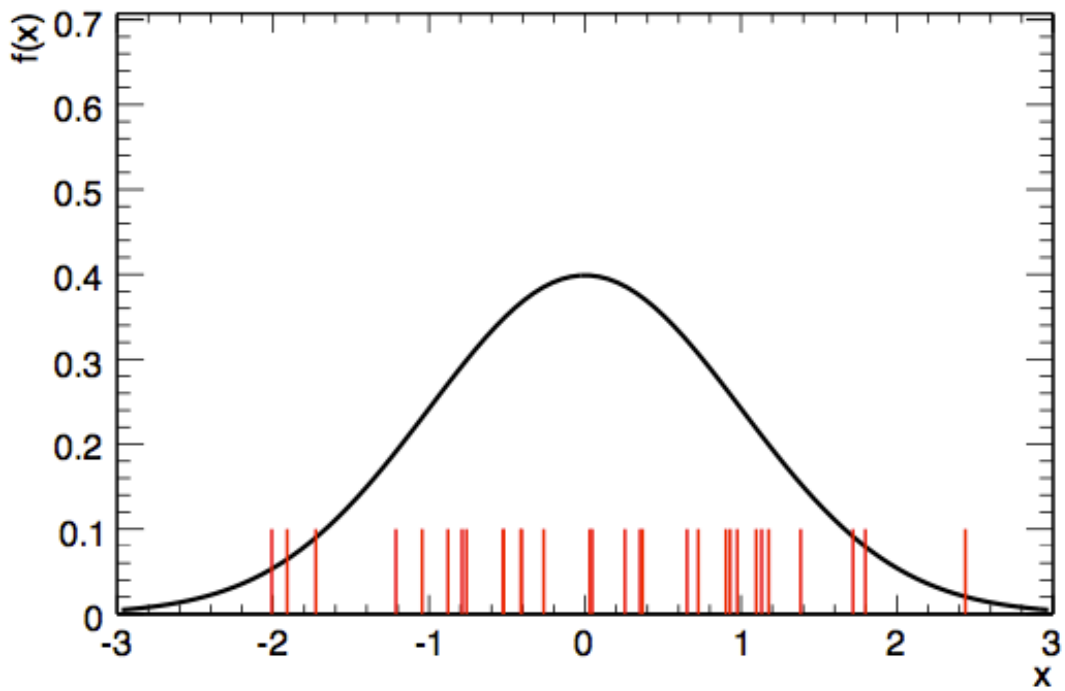


XTRP 28/0600Z
CLP5 28/0600Z
HMON 28/0000Z
AVNO 28/0000Z
ECMF 28/0000Z
TVCN 28/0600Z
TABD 28/0600Z
HWRF 28/0000Z
AEMN 28/0000Z
EEMN 28/0000Z
TVCX 28/0600Z
TABM 28/0600Z
UKM 28/0000Z
APxx 28/0000Z
EExx 28/0000Z
NHC 28/0900Z
TABS 28/0600Z
COTC 28/0000Z
CMC 28/0000Z
CEMN 28/0000Z
36N
34N
32N
30N
28N
26N
24N
22N
20N
18N
16N
14N
12N
95W
90W
85W
80W
75W
70W
65W
60W
55W
storm_05
sfwmd.gov
weather@sfwmd.gov
28-Aug 08:06EDT
NHC Advisories and County Emergency Management Statements supersede this product.
This graphic should complement, not replace, NHC discussions.
If anything on this graphic causes confusion, ignore the entire product.
For full info, see http://my.sfwmd.gov/sfwmd/common/images/weather/plots.html

The problem

Don't know PDF, have events drawn from PDF

$$f_{emp} = \frac{1}{N}\sum_{i}^{N}\delta(x - x_i)$$



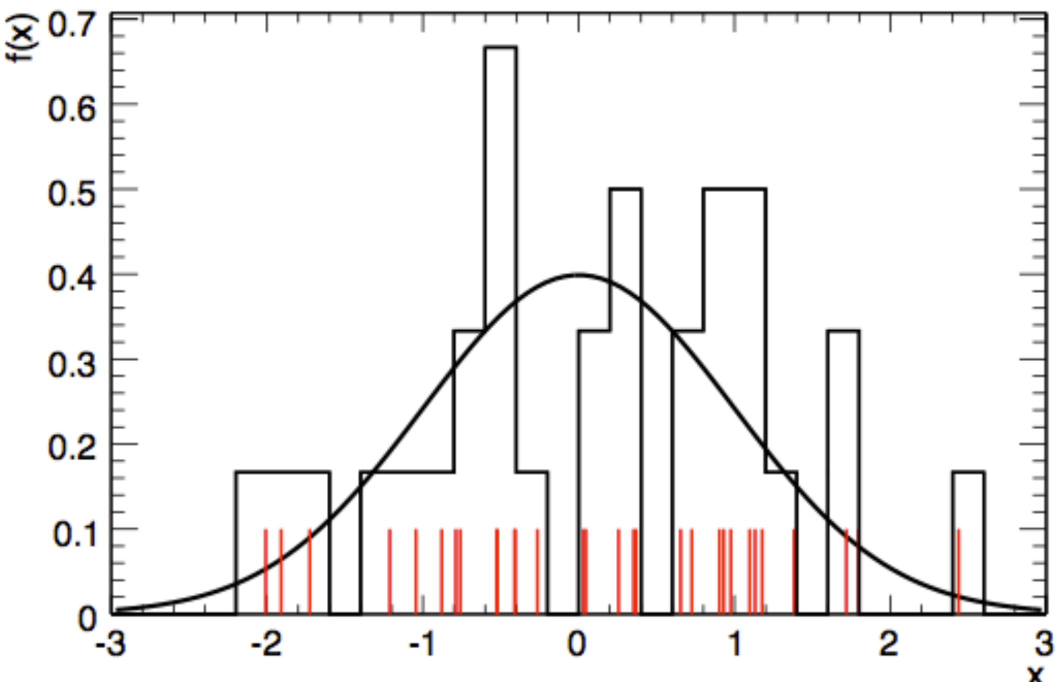
(K. Cranmer)

Need to recreate PDF

MC events to PDF

Simple approach : histogram

$$f_{hist}^{w,s}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$



(K. Cranmer)

Curse of Dimensionality

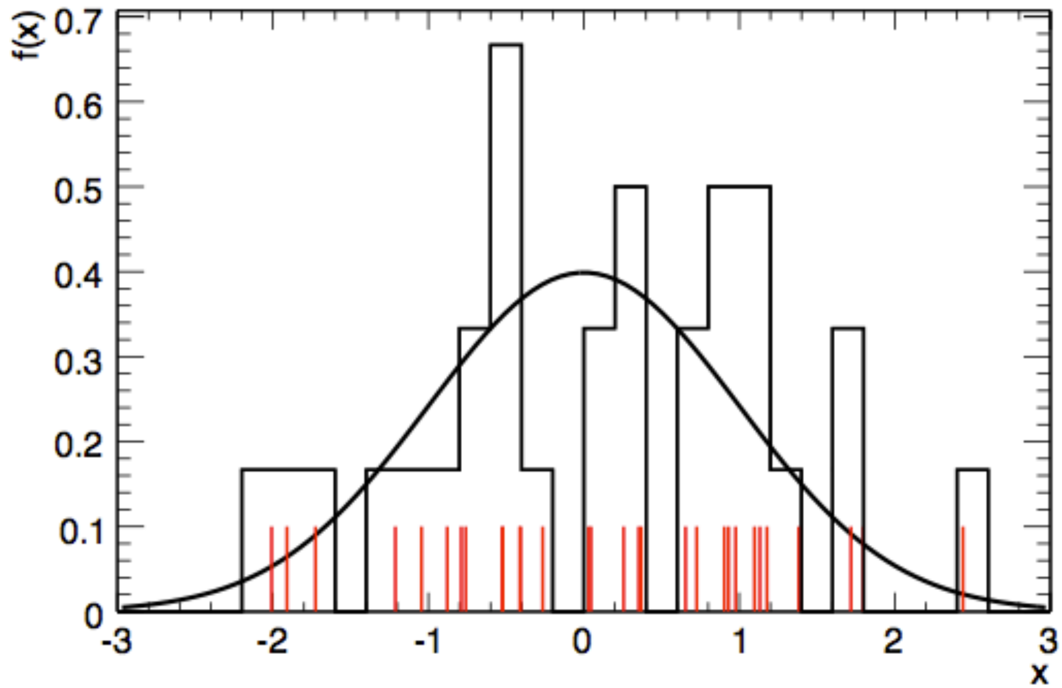
How many events
do you need
to describe a 1D
distribution? O(100)

An n-D distribution?

O(100^n)

!!

$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h^{w,s}_i$$



(K. Cranmer)

The nightmare

f(data|final-state particles P)

x f(final state particles P|showered particles S)

x f(showered particles S|hard scatter products M)

"data" is a 100M-d vector!

The nightmare

f(data | final-state particles P)

x f(final stat... ...red particles S)

x f(showered ... ...atter products M)

" ... ...vector!

Task for ML

Find a function:

$$f(\bar{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^1$$

which contains the same hypothesis testing power as

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

Neural networks

Strategy:

$$f(\bar{x}) : \mathbb{R}^N \rightarrow \mathbb{R}^1$$

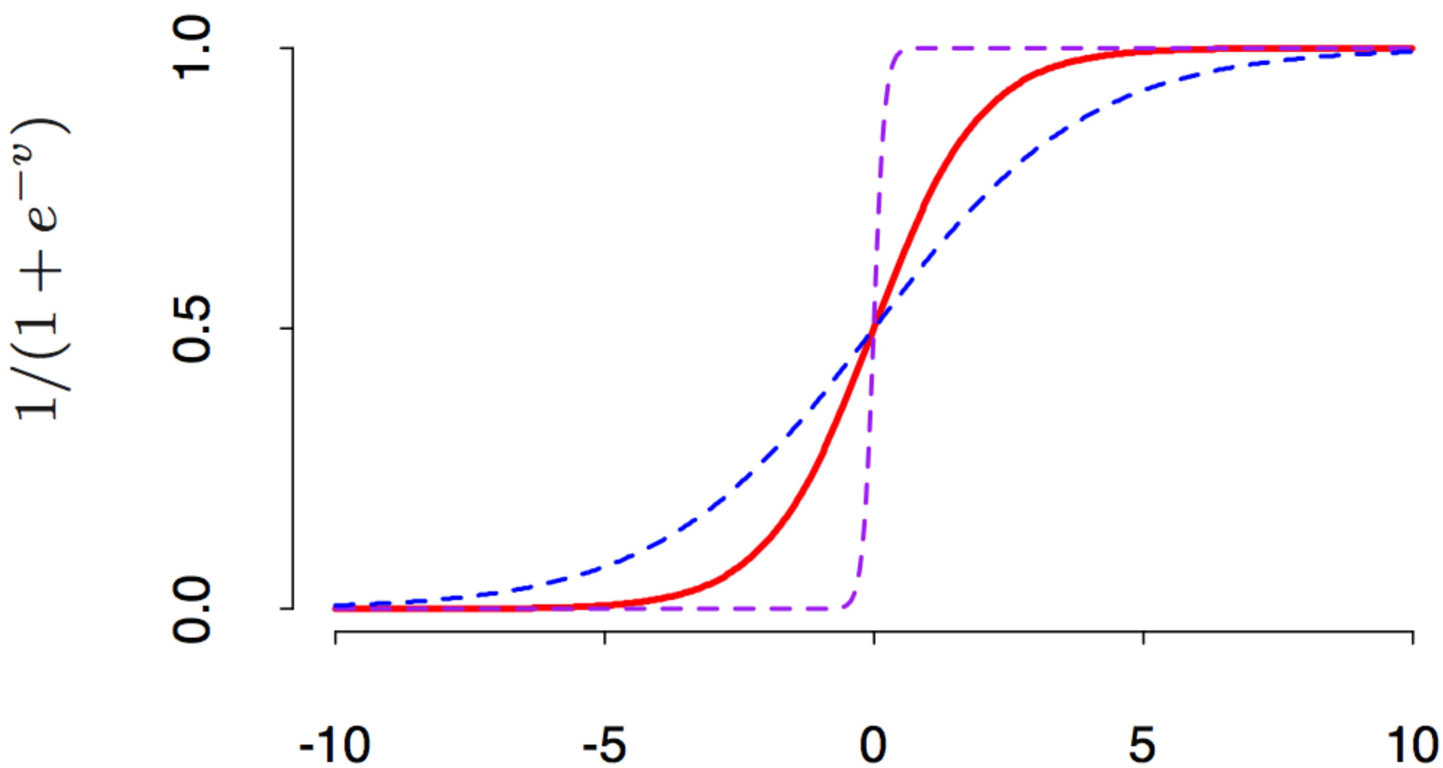
Build f(x)=y(x) out of a pile of convoluted mini-functions

$$y(\vec{x}) = h\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

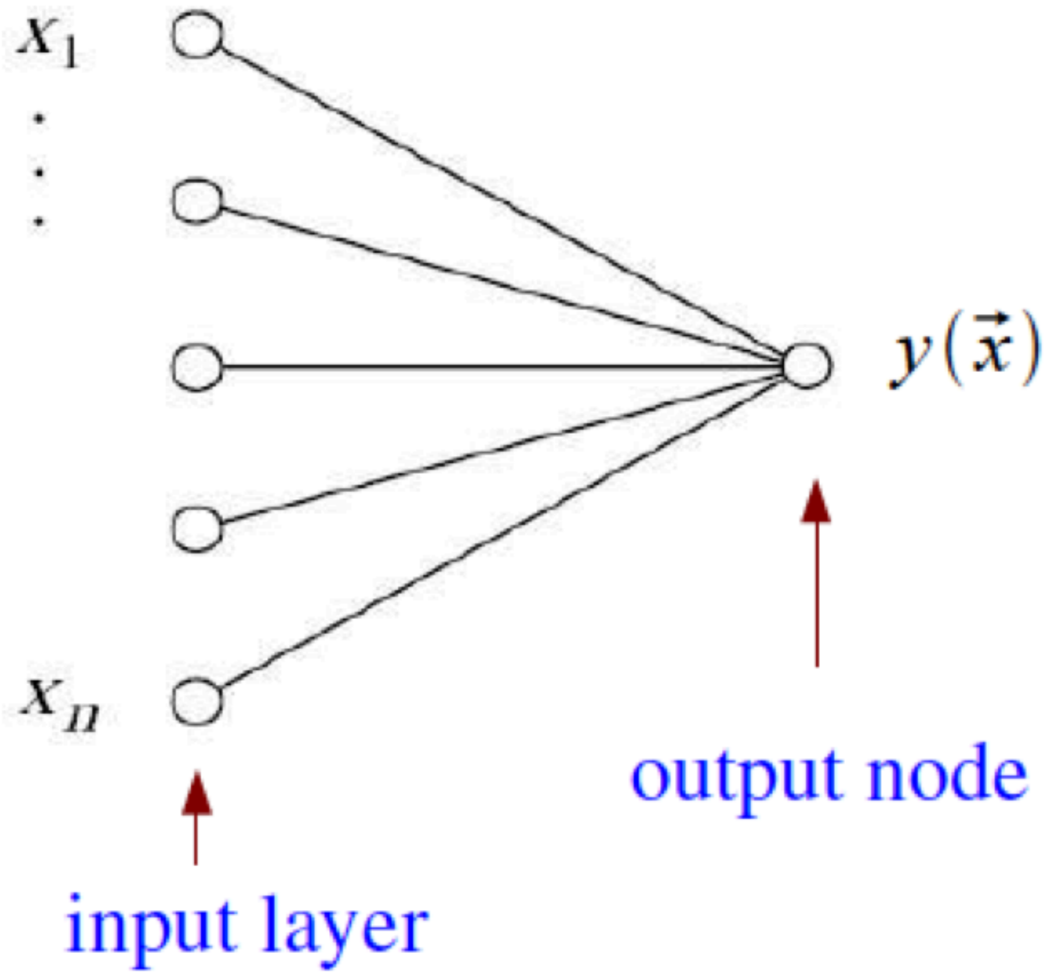
here h() is a non-linear *activation* function and the w factors are unknown parameters

Neuron

Example activation function



Simple visualization



Finding good weights

We have

a weight space $y(\vec{x}) = h\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$

a quality metric $E(\boldsymbol{w})$

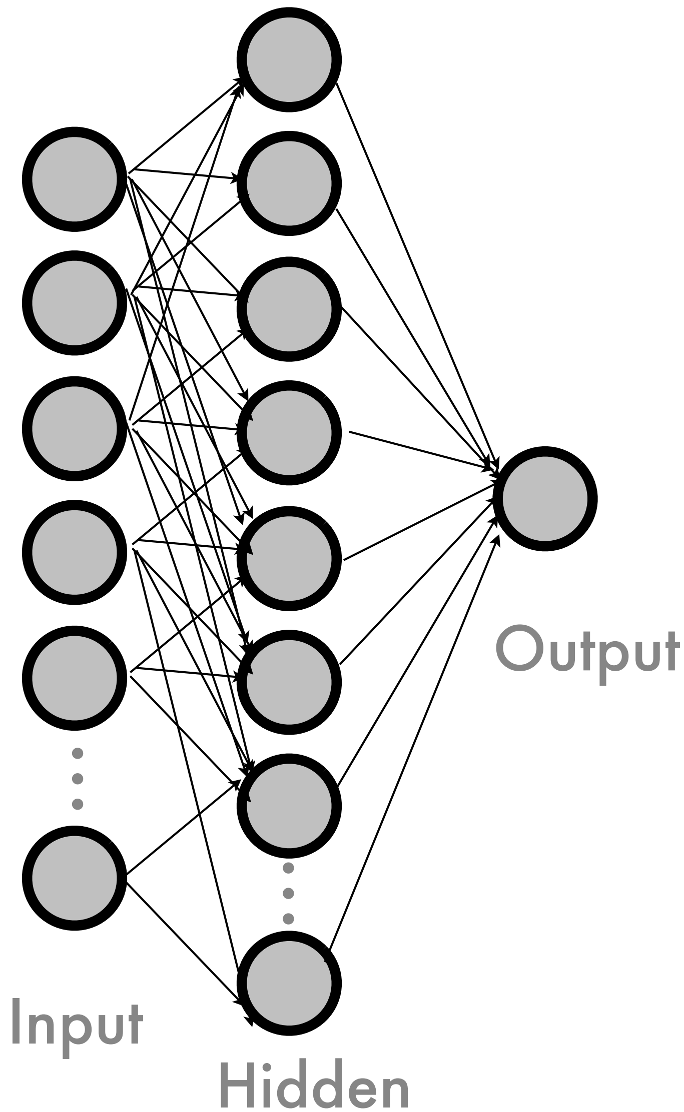
We need

to find the max quality (or min error)

Search the space!

How complex?

Essentially a functional fit with many parameters



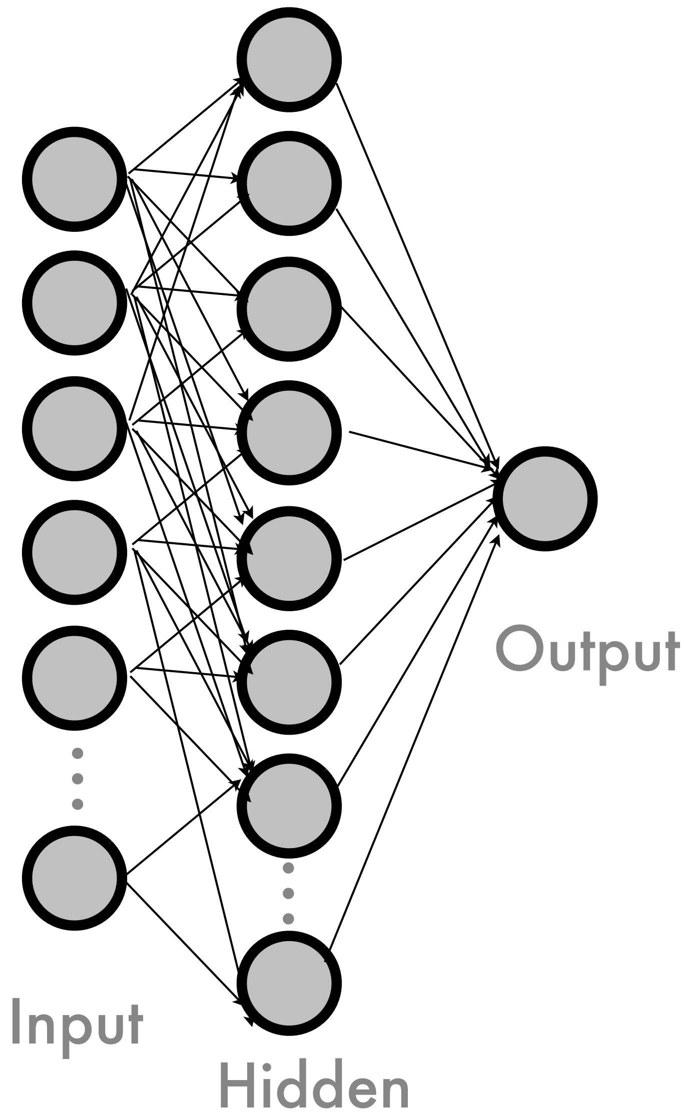
Single hidden layer

In theory any function can be learned with a single hidden layer.

But might require very large hidden layer

Neural Networks

Essentially a functional fit with many parameters



Problem:
Networks with > 1 layer are very difficult to train.

Consequence:
Networks are not good at learning non-linear functions.
(like invariant masses!)

In short:
Can't just throw 4-vectors at NN.

Search for Input

ATLAS-CONF-2013-108

Can’t just use 4v

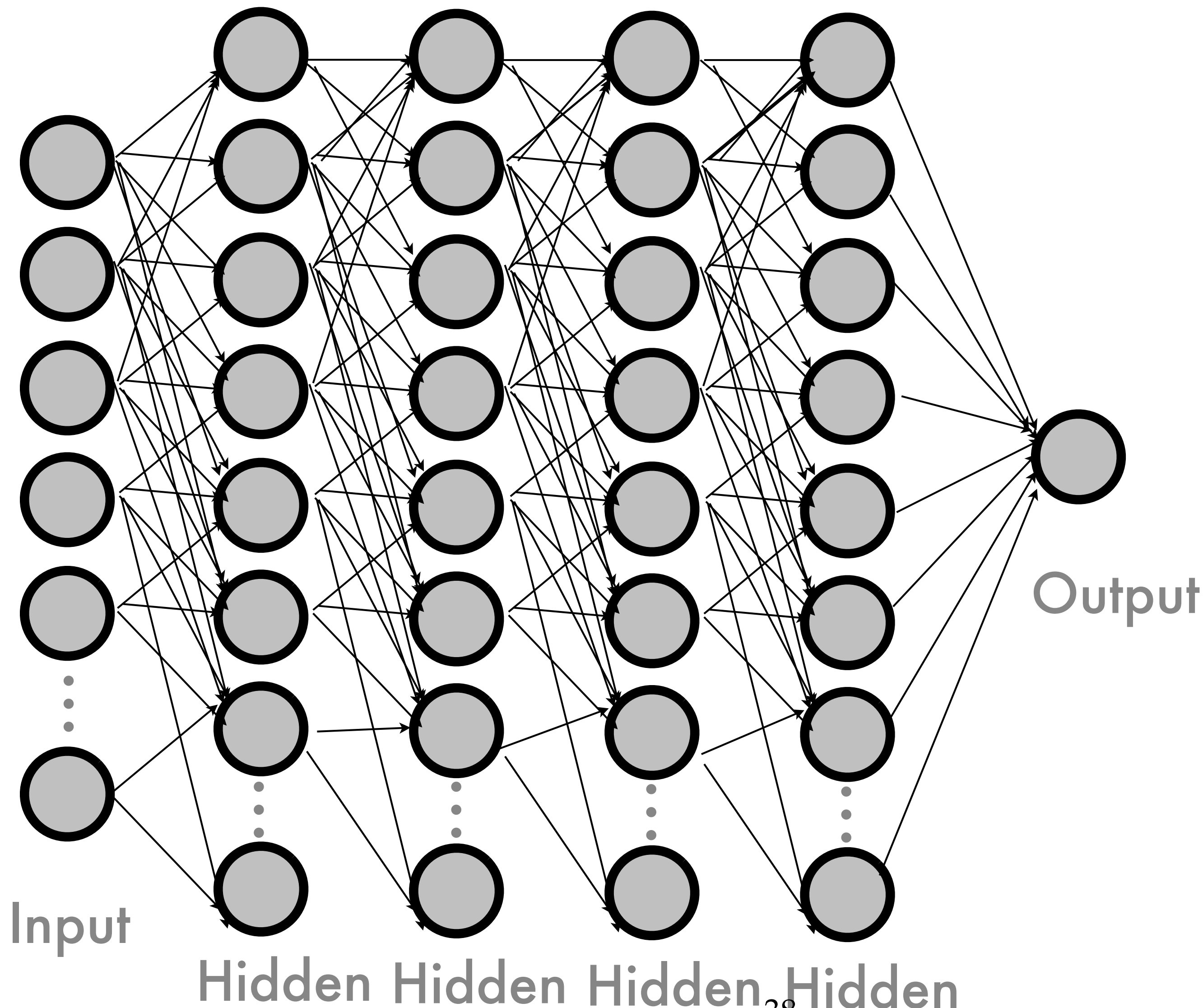
Can’t give it too many inputs

Painstaking search through input feature space.

Variable	VBF $\tau_{lep}\tau_{lep}$	VBF $\tau_{lep}\tau_{had}$	VBF $\tau_{had}\tau_{had}$	Boosted $\tau_{lep}\tau_{lep}$	Boosted $\tau_{lep}\tau_{had}$	Boosted $\tau_{had}\tau_{had}$
$m_{\tau\tau}^{\mathrm{MMC}}$	•	•	•	•	•	•
$\Delta R(\tau,\tau)$	•	•	•		•	•
$\Delta\eta(j_1, j_2)$	•	•	•			
m_{j_1,j_2}	•	•	•			
$\eta_{j_1} \times \eta_{j_2}$		•	•			
$p_{\mathrm{T}}^{\mathrm{Total}}$		•	•			
sum p_{T}					•	•
$p_{\mathrm{T}}(\tau_1)/p_{\mathrm{T}}(\tau_2)$					•	•
$E_{\mathrm{T}}^{\mathrm{miss}}\phi$ centrality		•	•	•	•	•
$x_{\tau1}$ and $x_{\tau2}$						•
$m_{\tau\tau,j_1}$				•		
m_{ℓ_1,ℓ_2}				•		
$\Delta\phi_{\ell_1,\ell_2}$				•		
sphericity				•		
$p_{\mathrm{T}}^{\ell_1}$				•		
$p_{\mathrm{T}}^{j_1}$				•		
$E_{\mathrm{T}}^{\mathrm{miss}}/p_{\mathrm{T}}^{\ell_2}$				•		
m_{T}		•			•	
$\min(\Delta\eta_{\ell_1\ell_2,\mathrm{jets}})$	•					
j_3 η centrality	•					
$\ell_1 \times \ell_2$ η centrality	•					
ℓ η centrality		•				
$\tau_{1,2}$ η centrality			•			

Table 3: Discriminating variables used for each channel and category. The filled circles identify which variables are used in each decay mode. Note that variables such as $\Delta R(\tau,\tau)$ are defined either between the two leptons, between the lepton and τ_{had}, or between the two τ_{had} candidates, depending on the decay mode.

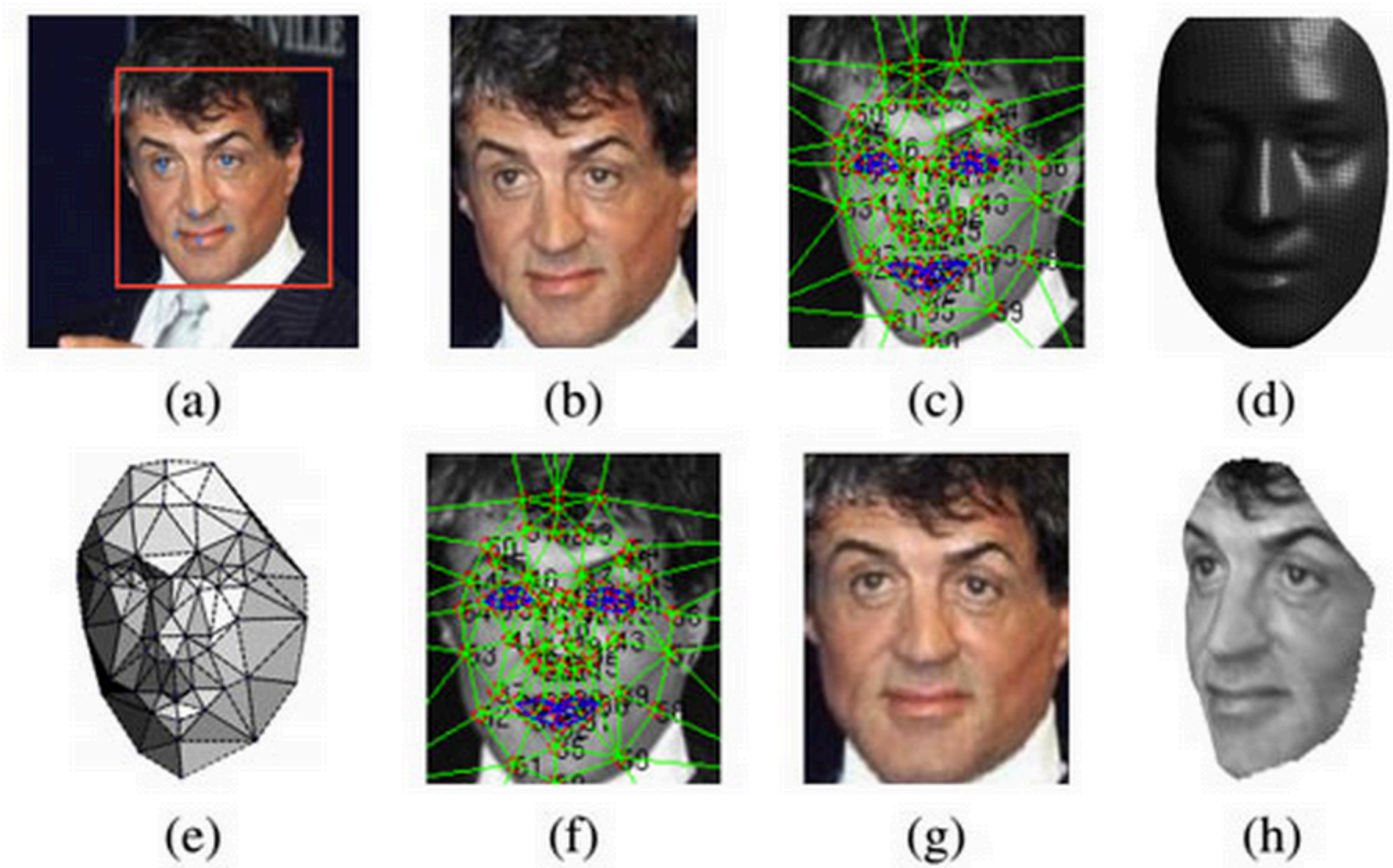
Deep networks



New tools let us train deep networks.

How well do they work?

Real world applications



Head turn: DeepFace uses a 3-D model to rotate faces, virtually, so that they face the camera. Image (a) shows the original image, and (g) shows the final, corrected version.

Paper

nature COMMUNICATIONS

ARTICLE

Received 19 Feb 2014 | Accepted 4 Jun 2014 | Published 2 Jul 2014

DOI: 10.1038/ncomms5308

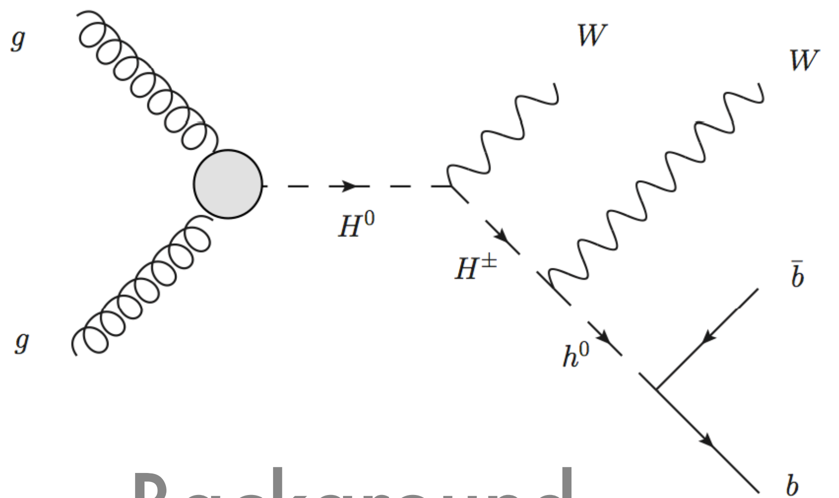
Searching for exotic particles in high-energy physics with deep learning

P. Baldi[1], P. Sadowski[1] & D. Whiteson[2]

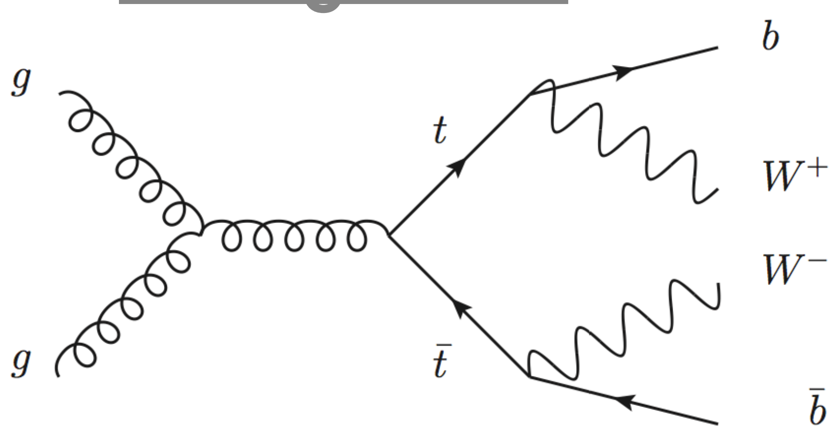
arXiv: 1402.4735

Benchmark problem

Signal



Background



Can deep networks automatically discover useful variables?

4-vector inputs

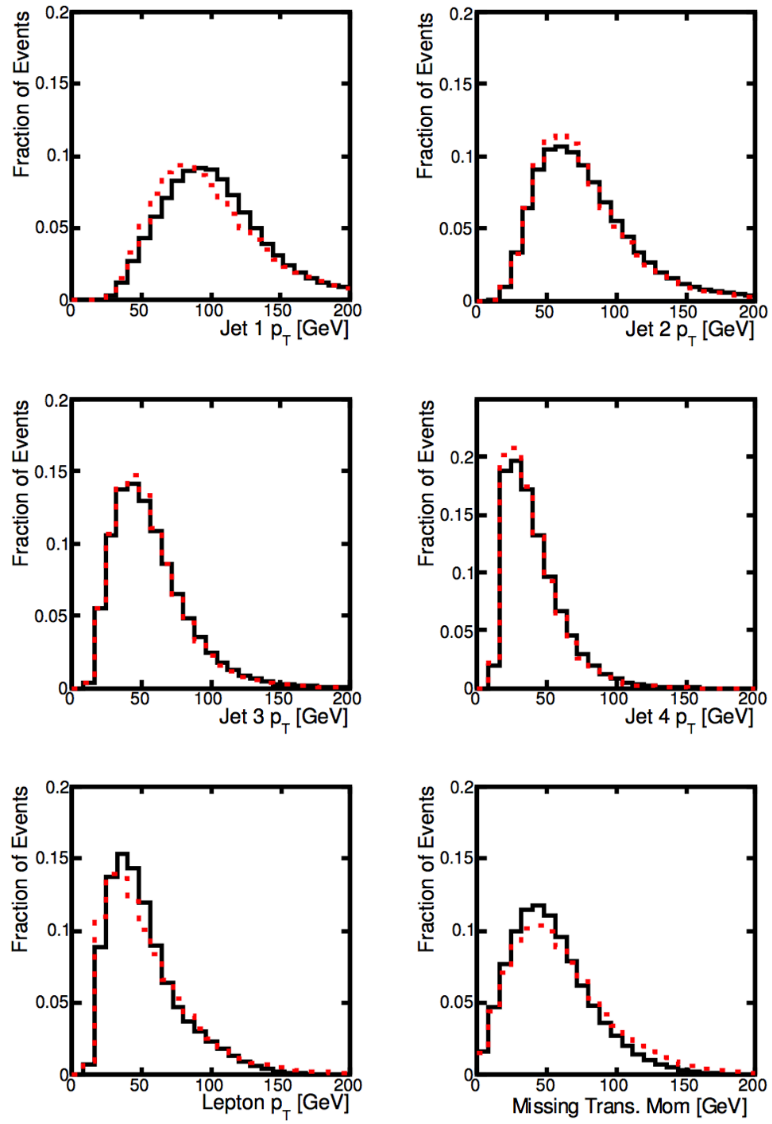
<u>21 Low-level vars</u>

jet+lepton mom. (3x5)

missing ET (2)

jet btags (4)

Not much separation visible in 1D projections



4-vector inputs

<u>7 High-level vars</u>

m(WWbb)

m(Wbb)

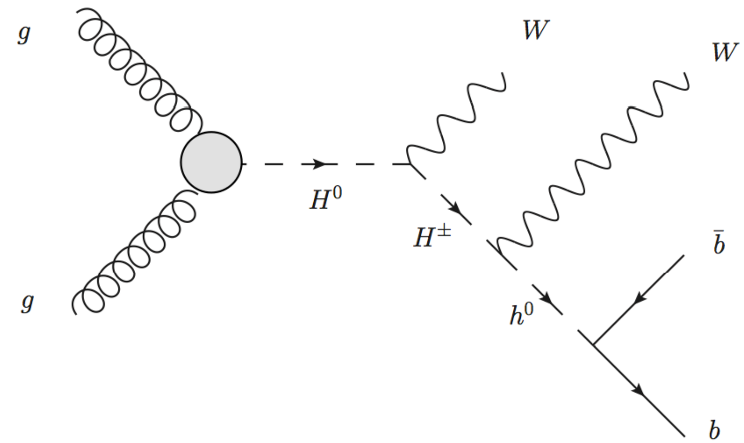
m(bb)

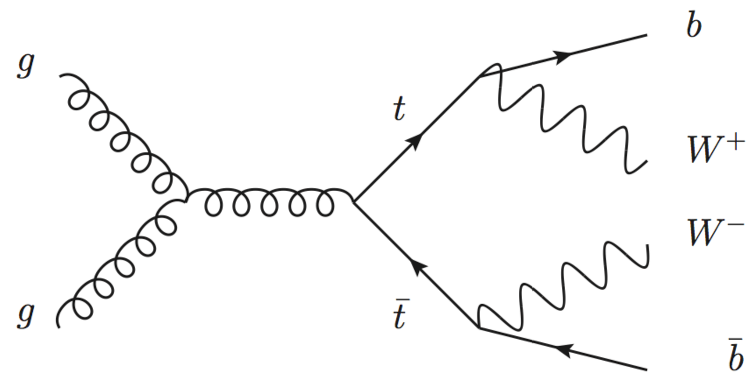
m(bjj)

m(jj)

m(lv)

m(blv)





4-vector inputs

7 High-level vars

m(WWbb)

m(Wbb)

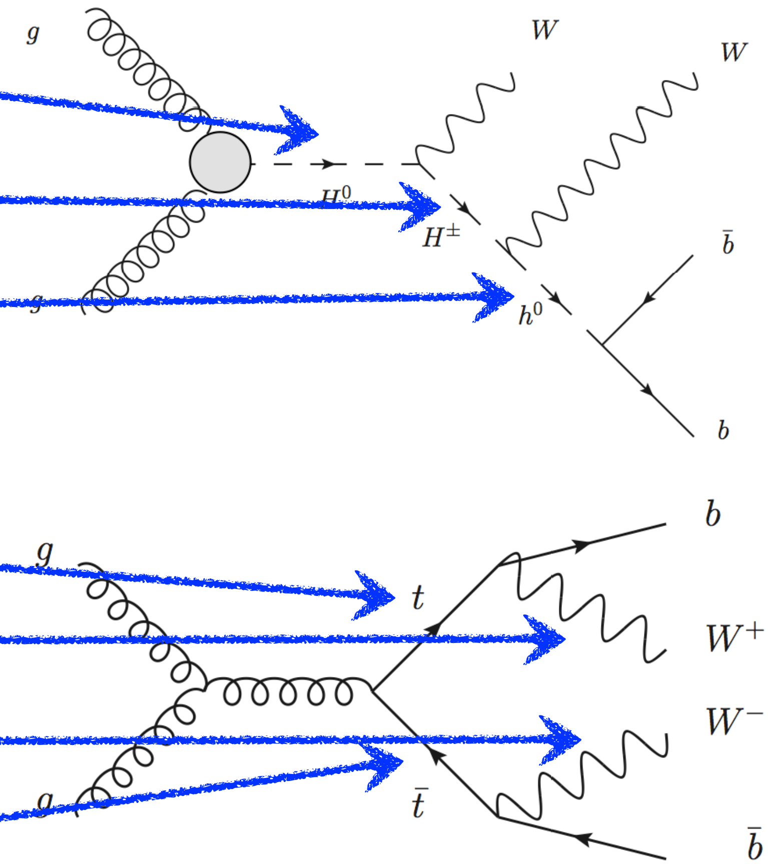
m(bb)

m(bjj)

m(jj)

m(lv)

m(blv)

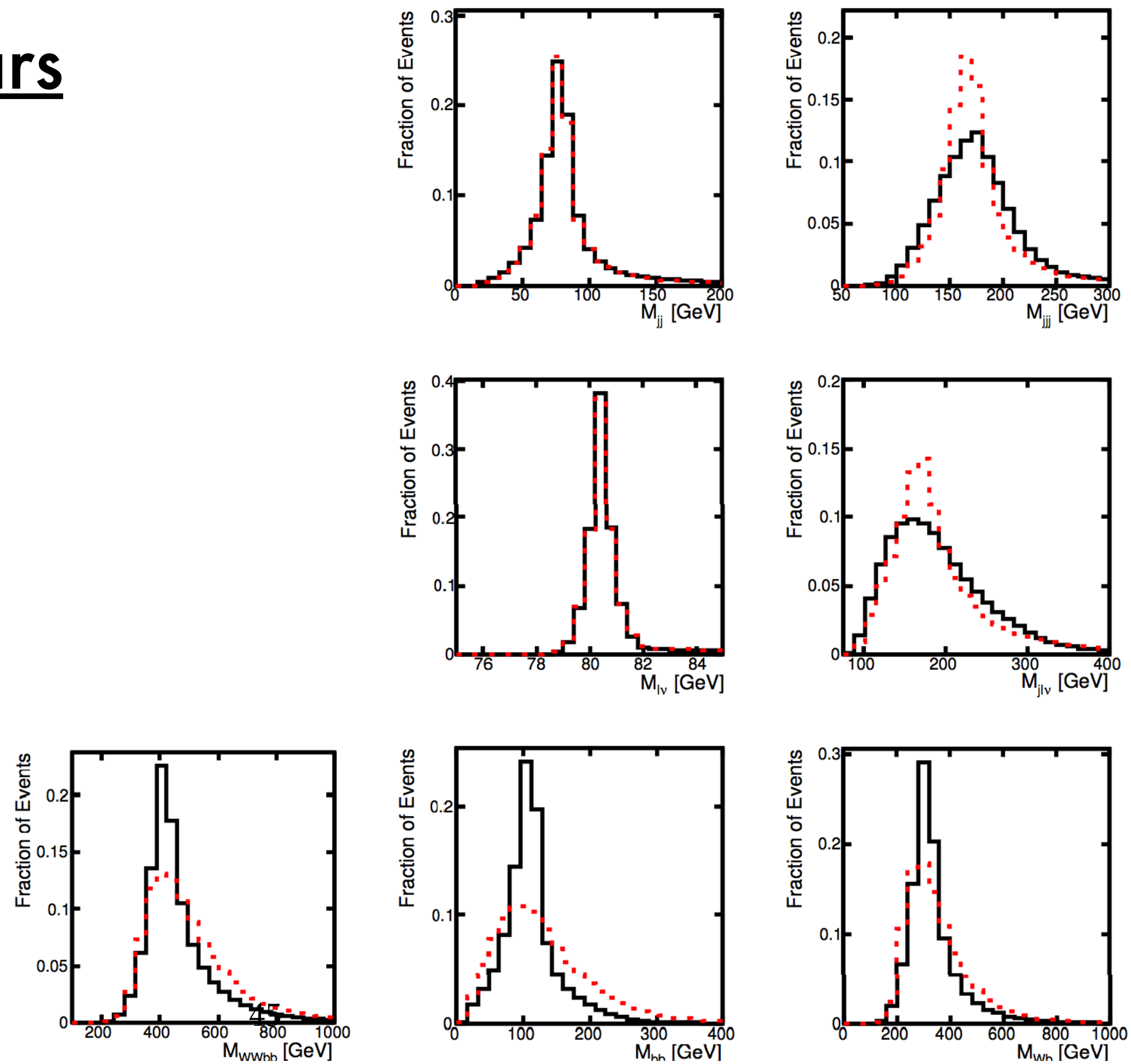


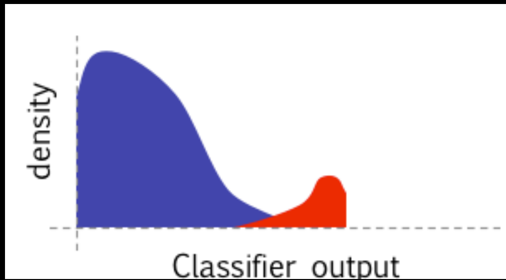
4-vector inputs

7 High-level vars

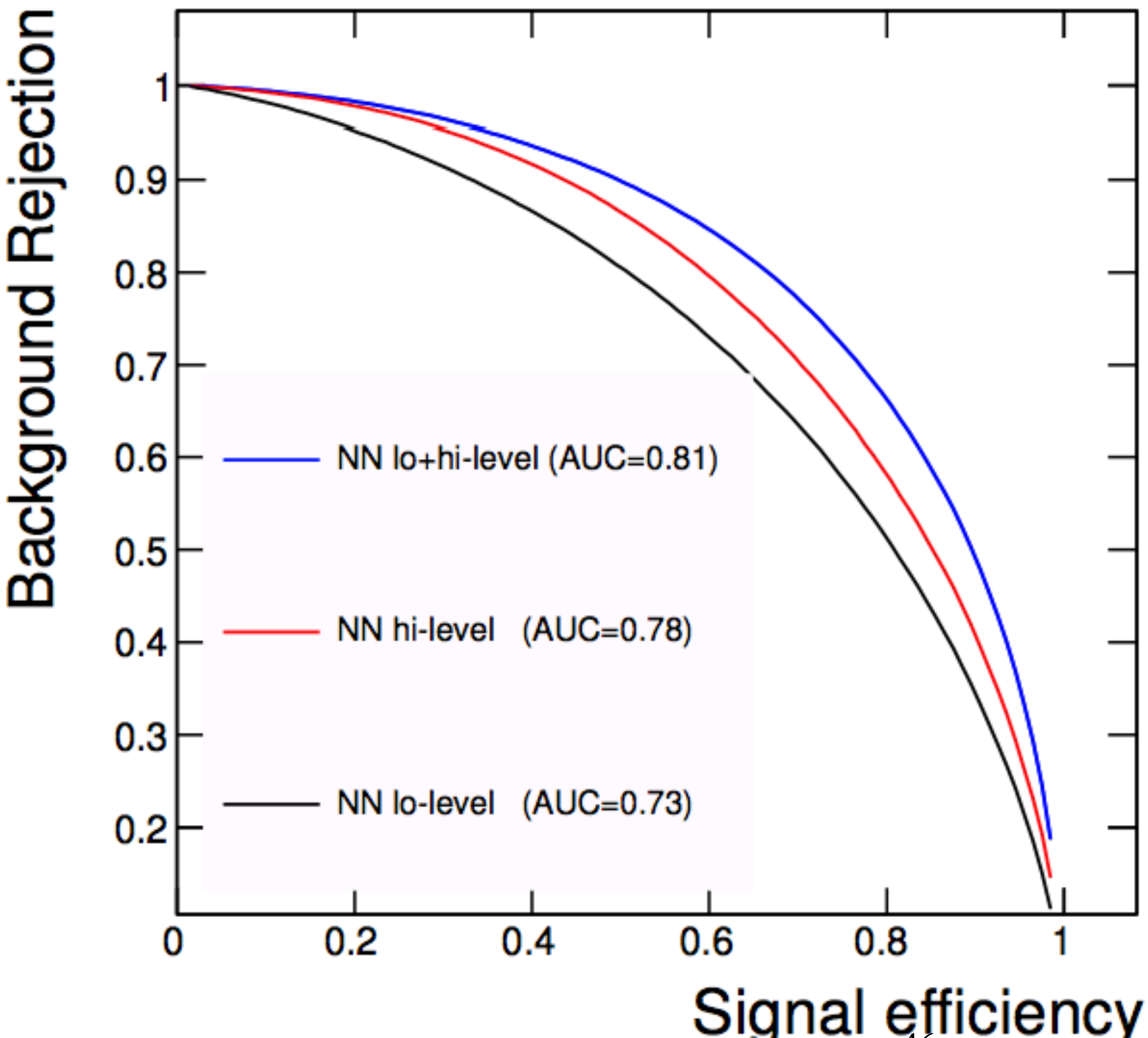
m(WWbb)
m(Wbb)
m(bb)

m(bjj)
m(jj)
m(lv)
m(blv)





Standard NNs



__Results__

Adding hi-level boosts performance

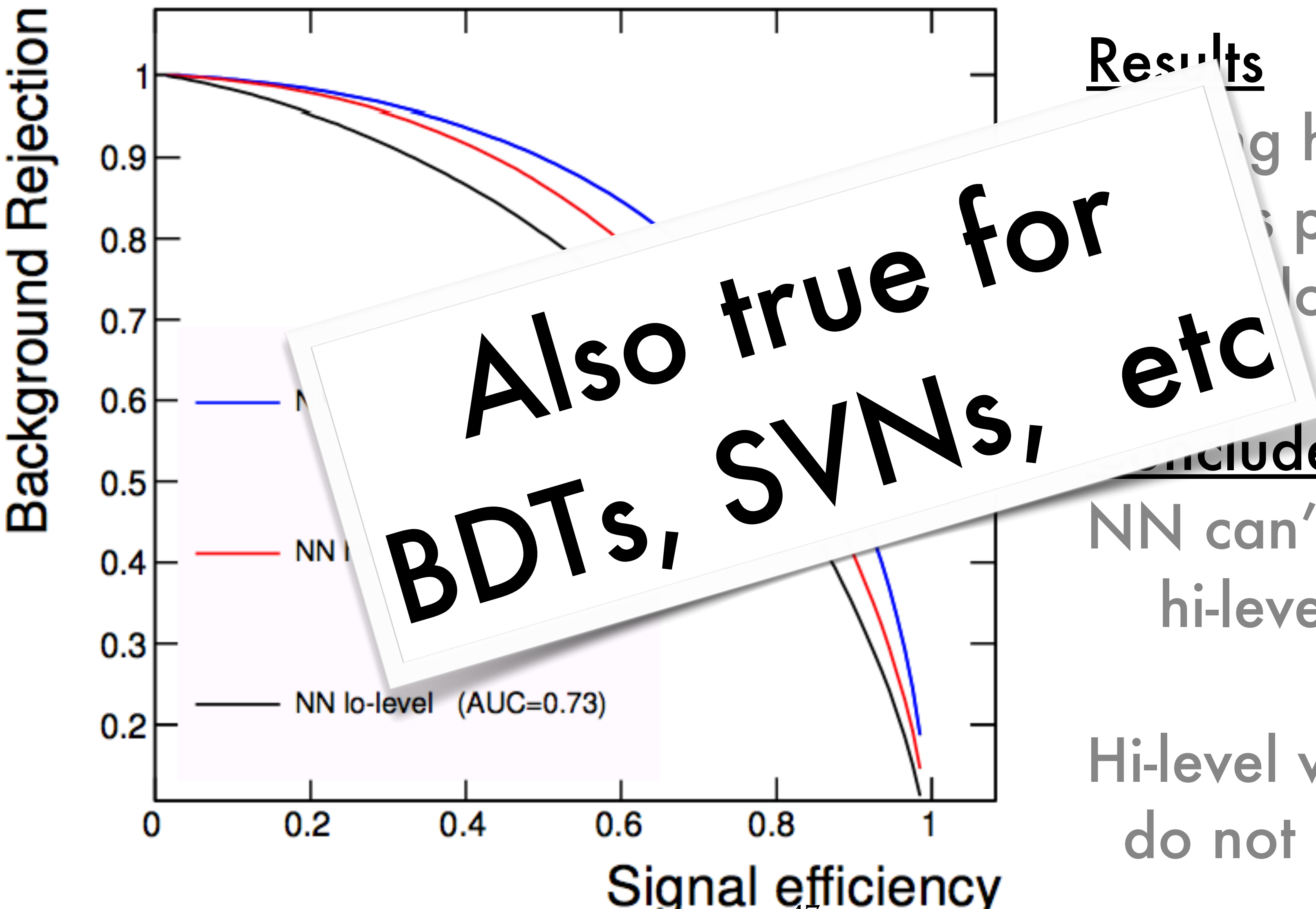
Better: lo+hi-level.

__Conclude:__

NN can't find hi-level vars.

Hi-level vars do not have all info

Standard NNs



Results

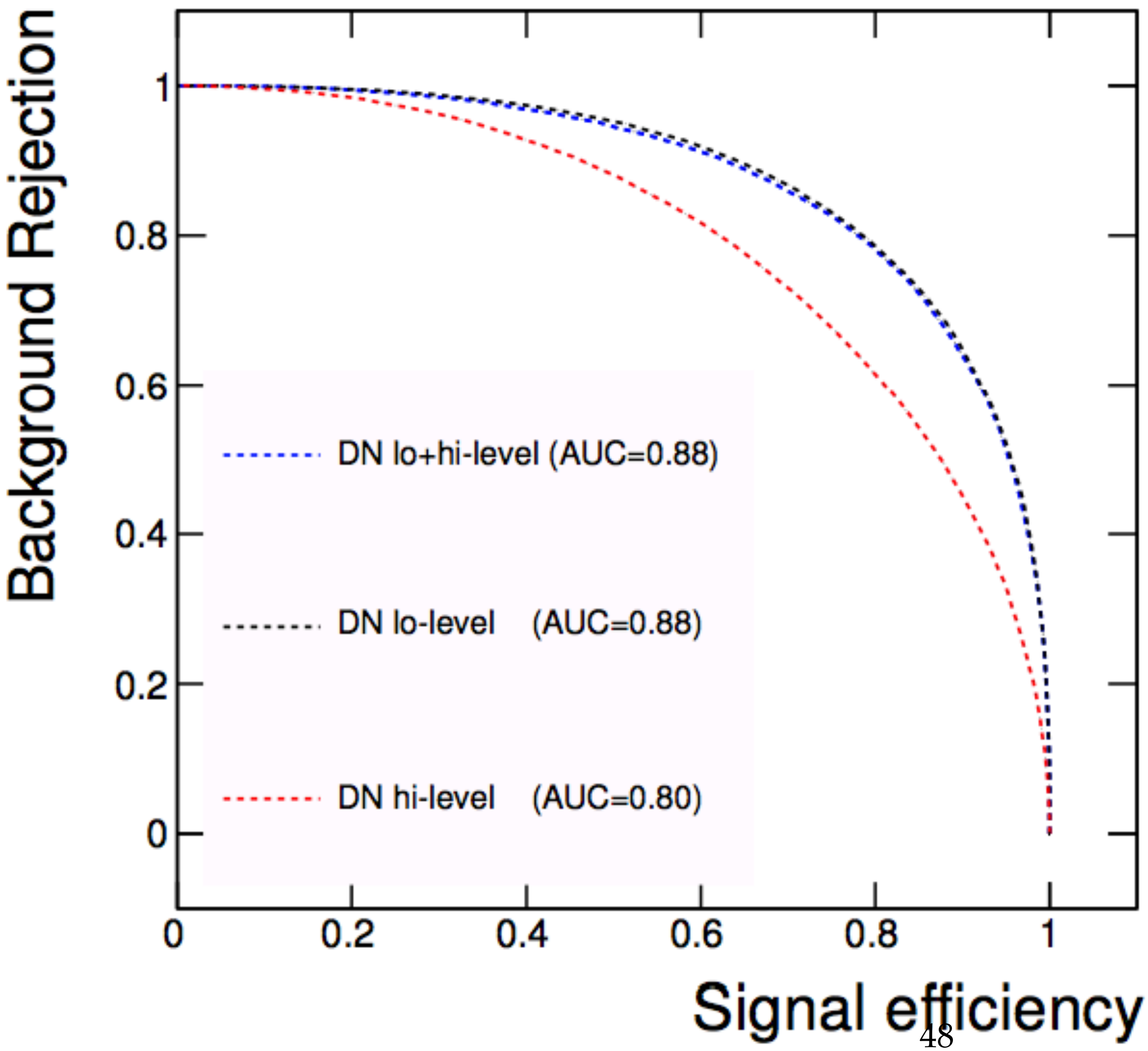
g hi-level
s performance
lo+hi-level.

nclude:

NN can't find
hi-level vars.

Hi-level vars
do not have all info

Deep Networks



Results

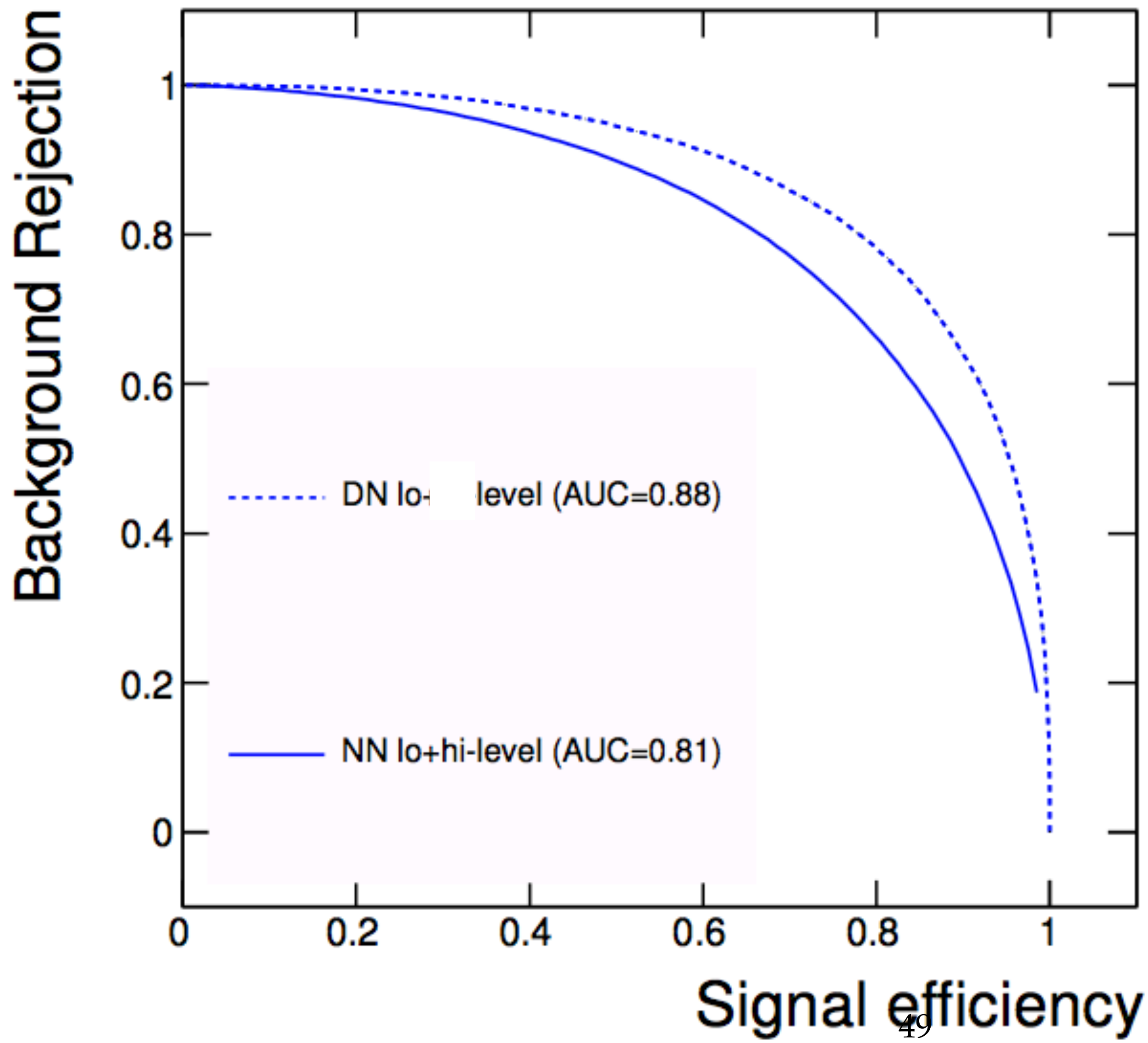
Lo+hi = lo.

Conclude:

DN can find hi-level vars.

Hi-level vars do not have all info are unnecessary

Deep Networks



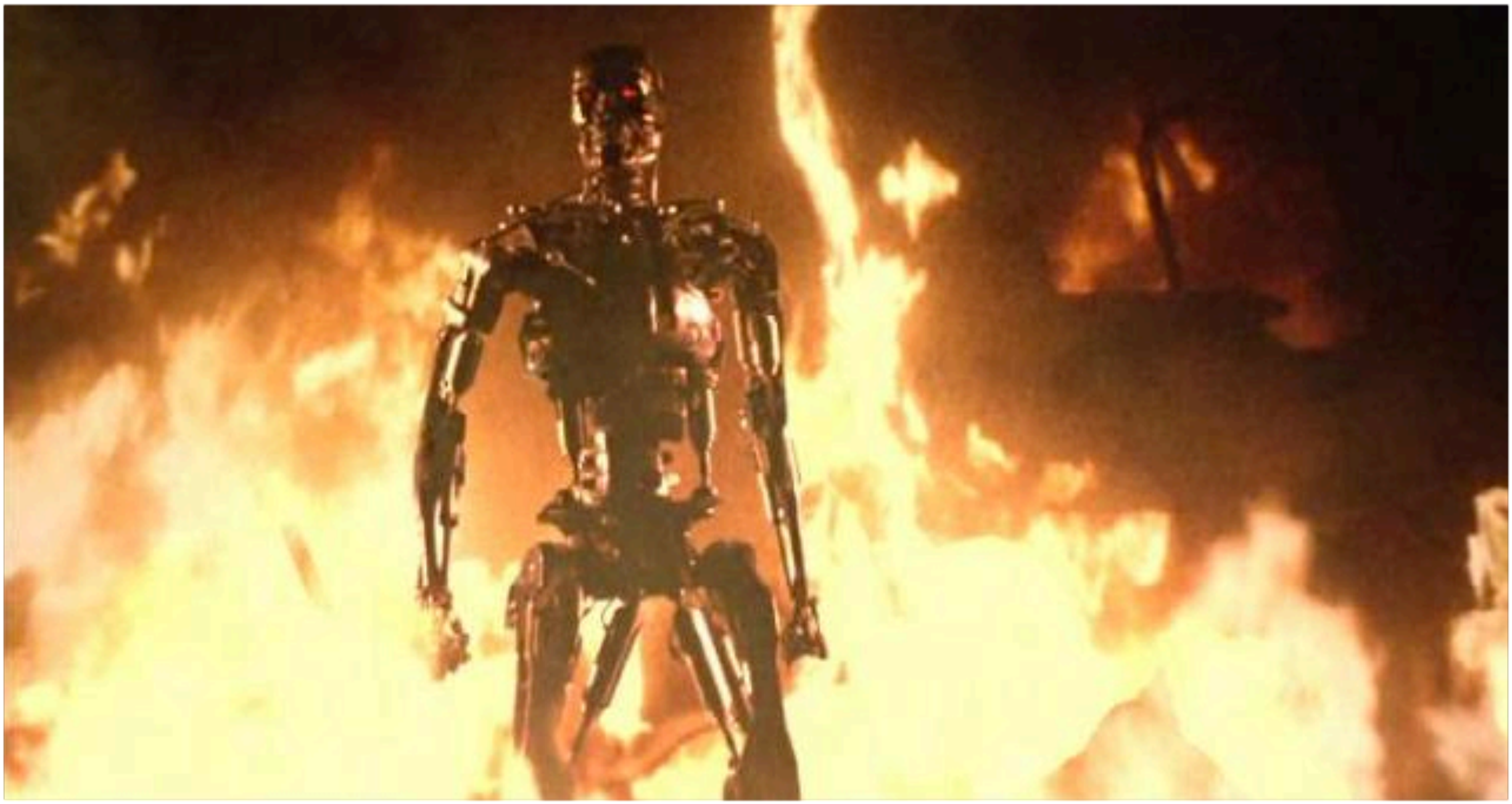
Results

DN > NN

Conclude:

DN does better than human assisted NN

The AIs win



Results

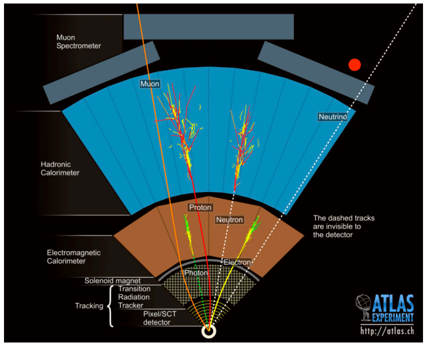
Identified example benchmark where traditional NNs fail to discover all discrimination power.

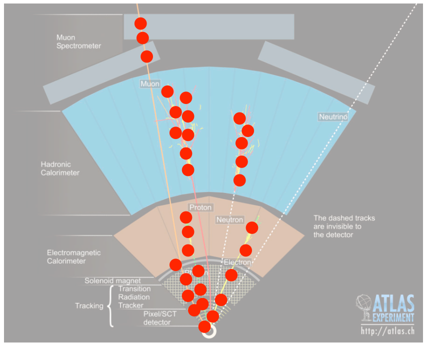
Adding human insight helps traditional NNs.

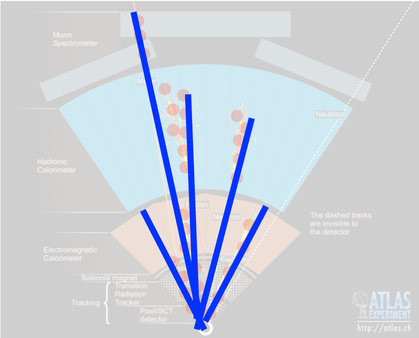
Deep networks succeed without human insight. Outperform human-boosted traditional NNs.

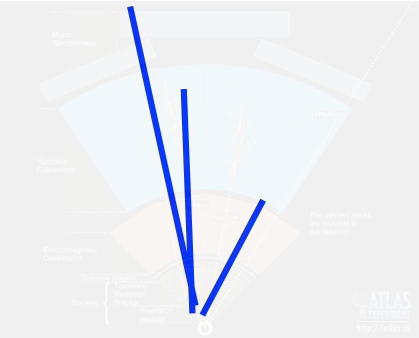
What is possible?

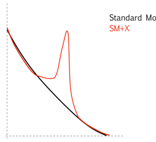
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1





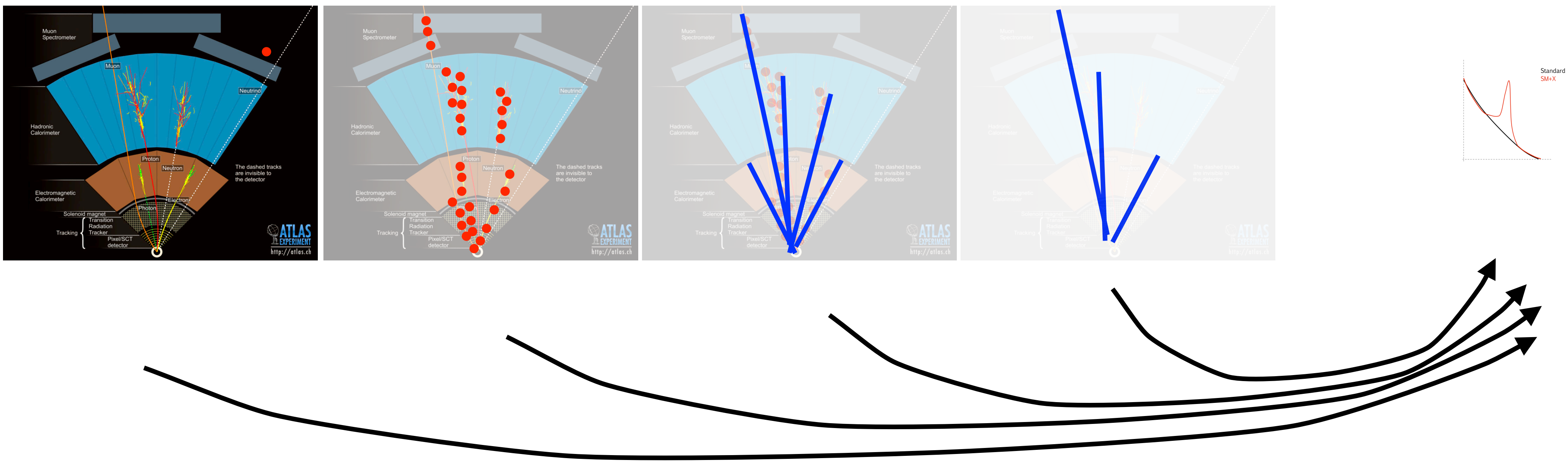






What is possible?

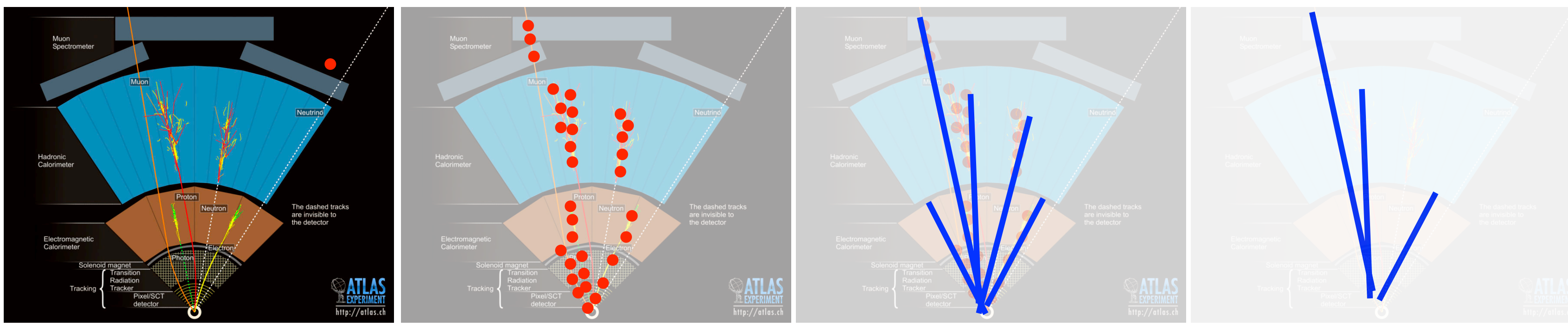
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1

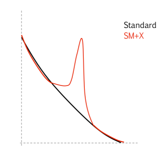


Skip more steps with ML?

Or this?

Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1





Improve each step with ML?

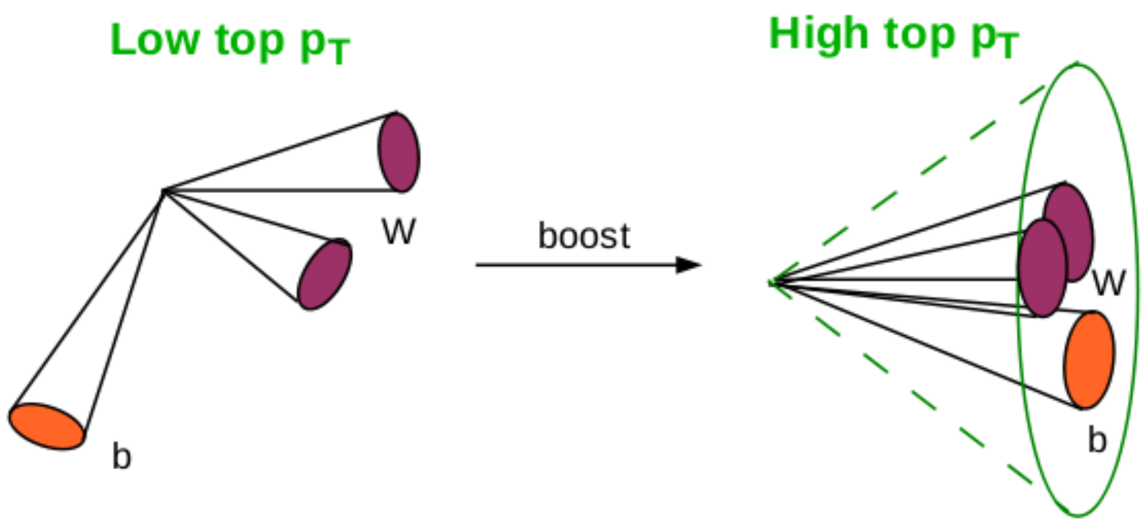
Jets

Jet Substructure Classification in High-Energy Physics with Deep Neural Networks

Pierre Baldi,[1] Kevin Bauer,[2] Clara Eng,[3] Peter Sadowski,[1] and Daniel Whiteson[2]

[1] *Department of Computer Science, University of California, Irvine, CA 92697*
[2] *Department of Physics and Astronomy, University of California, Irvine, CA 92697*
[3] *Department of Chemical and Biomolecular Engineering, University of California, Berkeley CA 94270*

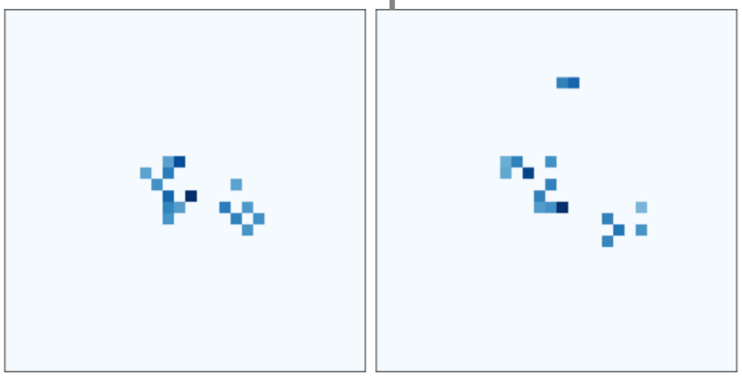
(Dated: April 12, 2016)



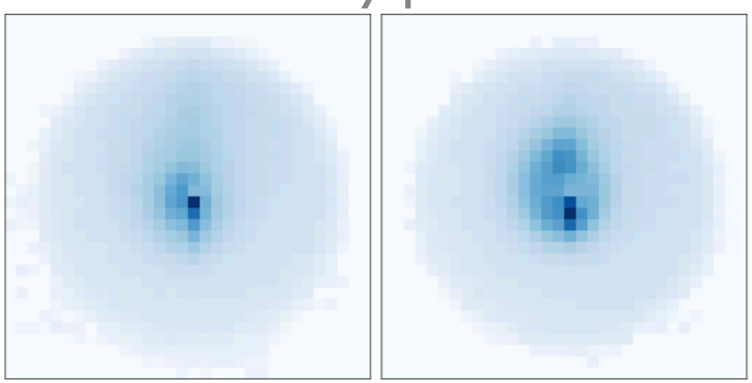
Jet substructure

LL variables

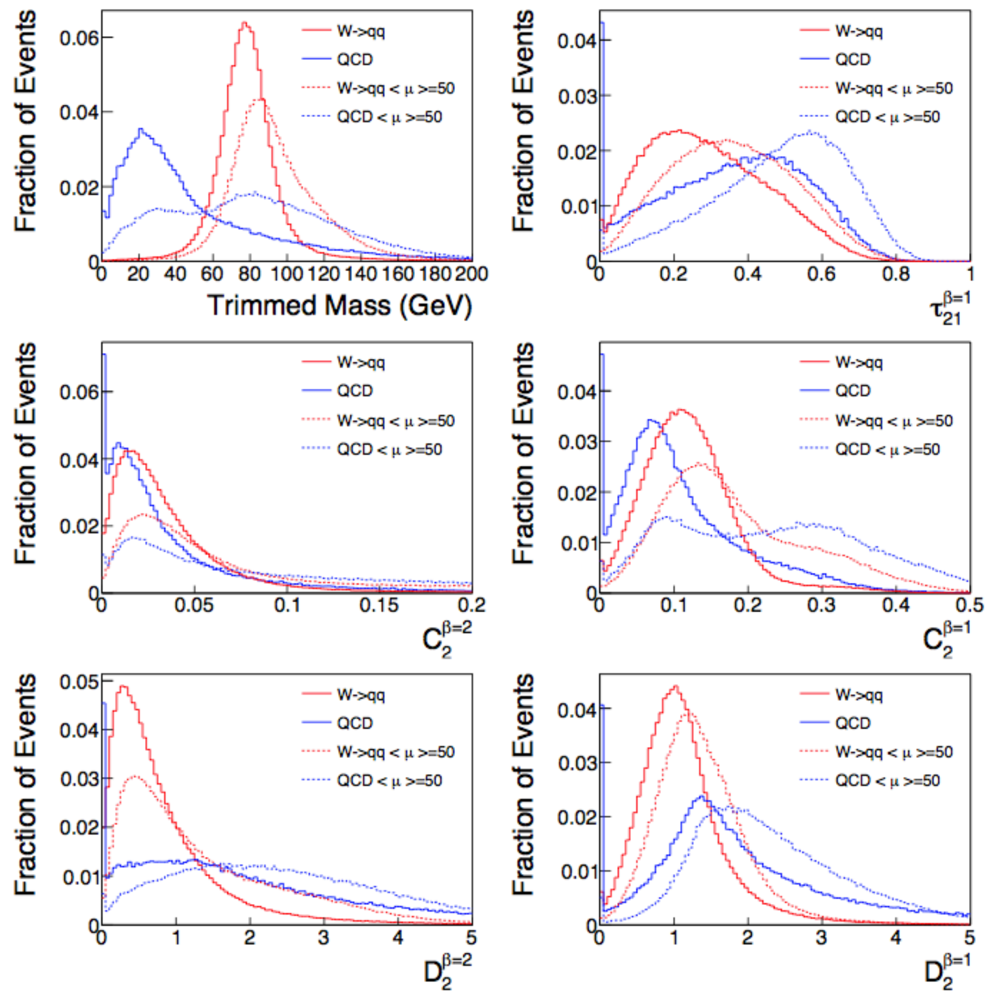
one jet



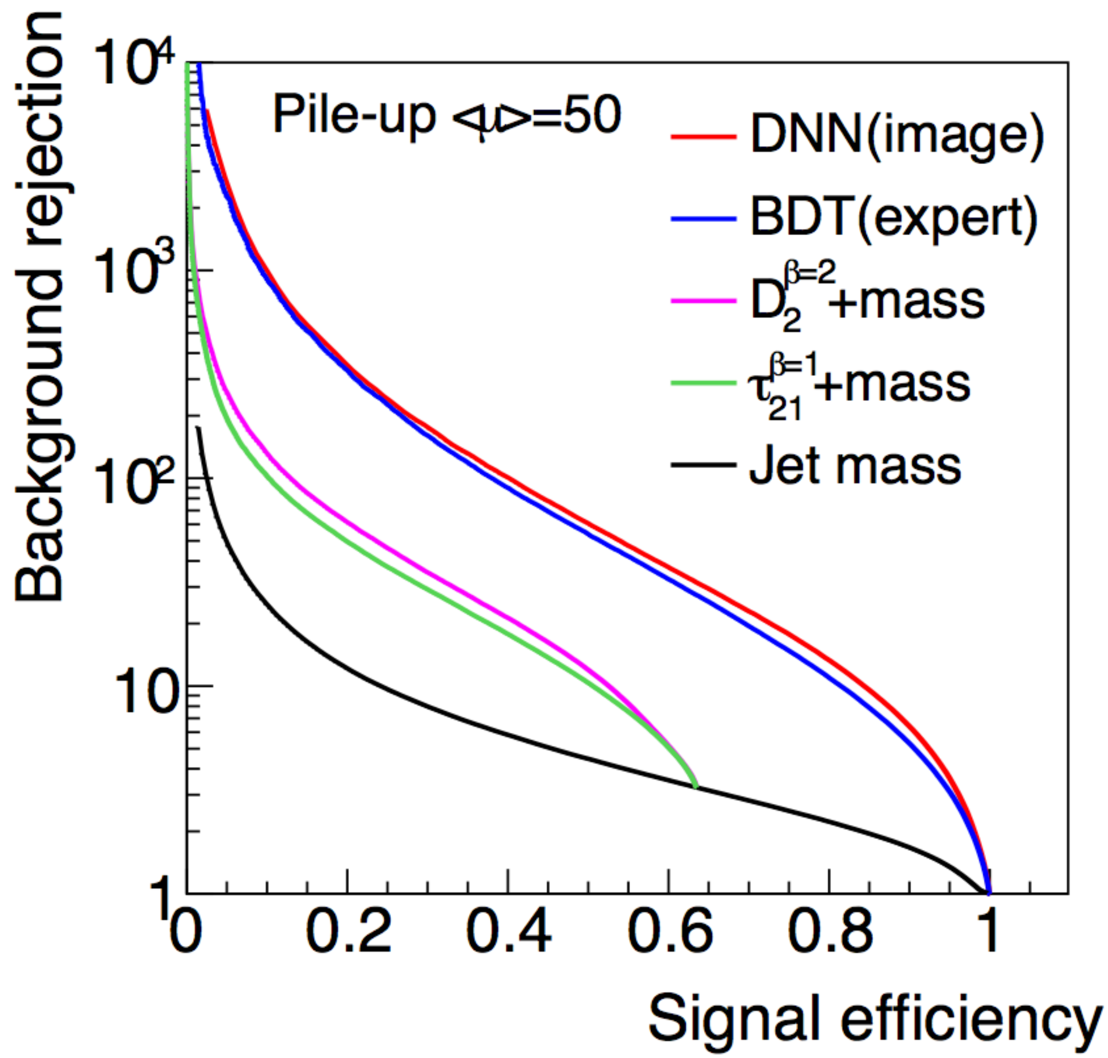
many jets



HL variables



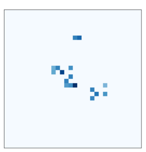
Jet tagging


Background rejection
Signal efficiency
Pile-up <μ>=50
DNN(image)
BDT(expert)
$D_2^{\beta=2}$+mass
$\tau_{21}^{\beta=1}$+mass
Jet mass
10^4
10^3
10^2
10
1
0
0.2
0.4
0.6
0.8
1

How?

What is it doing?

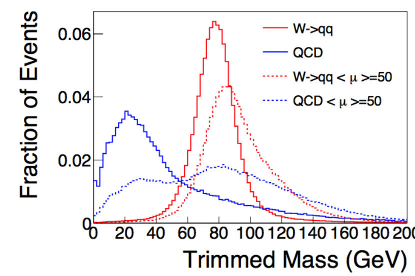
Our low-level (LL) data are often high-dim



LL

expert physics knowl.

HL



We can calculate likelihood ratios in the low-dim HL space often using MC techniques

But HL doesn't always capture the information

Yet we prefer HL

<u>If HL data includes all necessary information...</u>

- It is easier to understand
- Its modeling can be verified
- Uncertainties can be sensibly defined
- It is more compact and efficient

- LL -> HL is physics, so we like it.

Our question

How has the DNN found its solution?

What can we learn from it?

<u>**Residual knowledge:**</u>

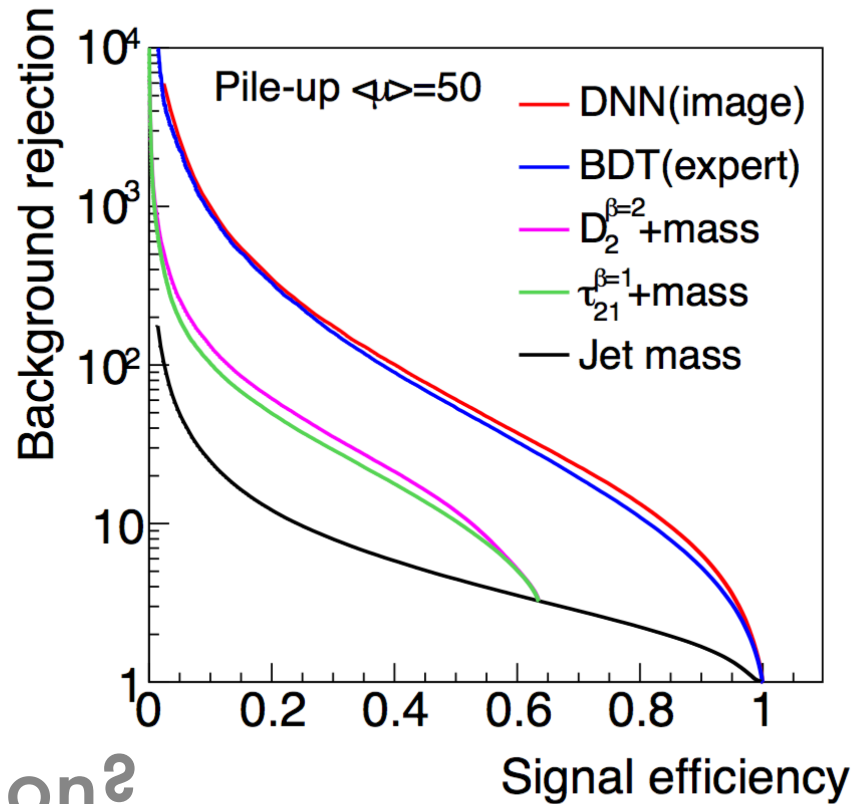
Is there a **new** HL variable?

Can it reveal physics?

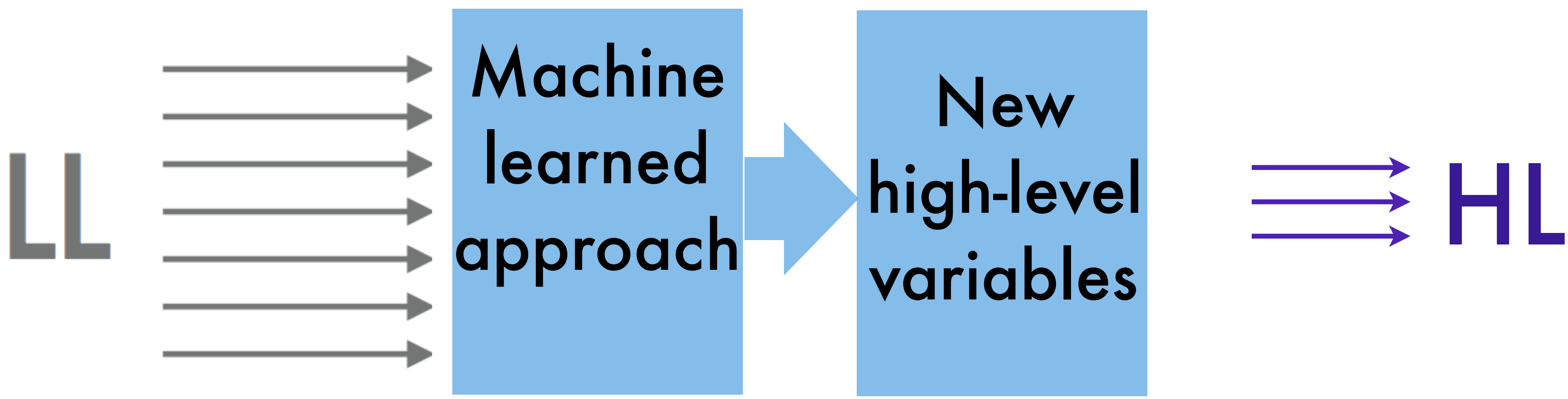
<u>**Translating complete solutions:**</u>

What is the **structure** of its solution?

Has it just rediscovered and optimized the existing HL vars?



Learning from ML



Use LL analysis as a probe, not a final product.

How?

I. Define space of possible human solutions

- provides context for NN solution
- defines problem
- does NN live in this space?
- Can it be compactly represented?
- Yes or No are both interesting!

$$= \sum_a \sum_b \sum_c z_a z_b z_c \theta_{ab} \theta_{ac} \theta_{bc}^2$$

$$z_i = \frac{p_{T_i}}{\sum_i p_{T_j}}$$

How?

I. Define space of possible human solutions

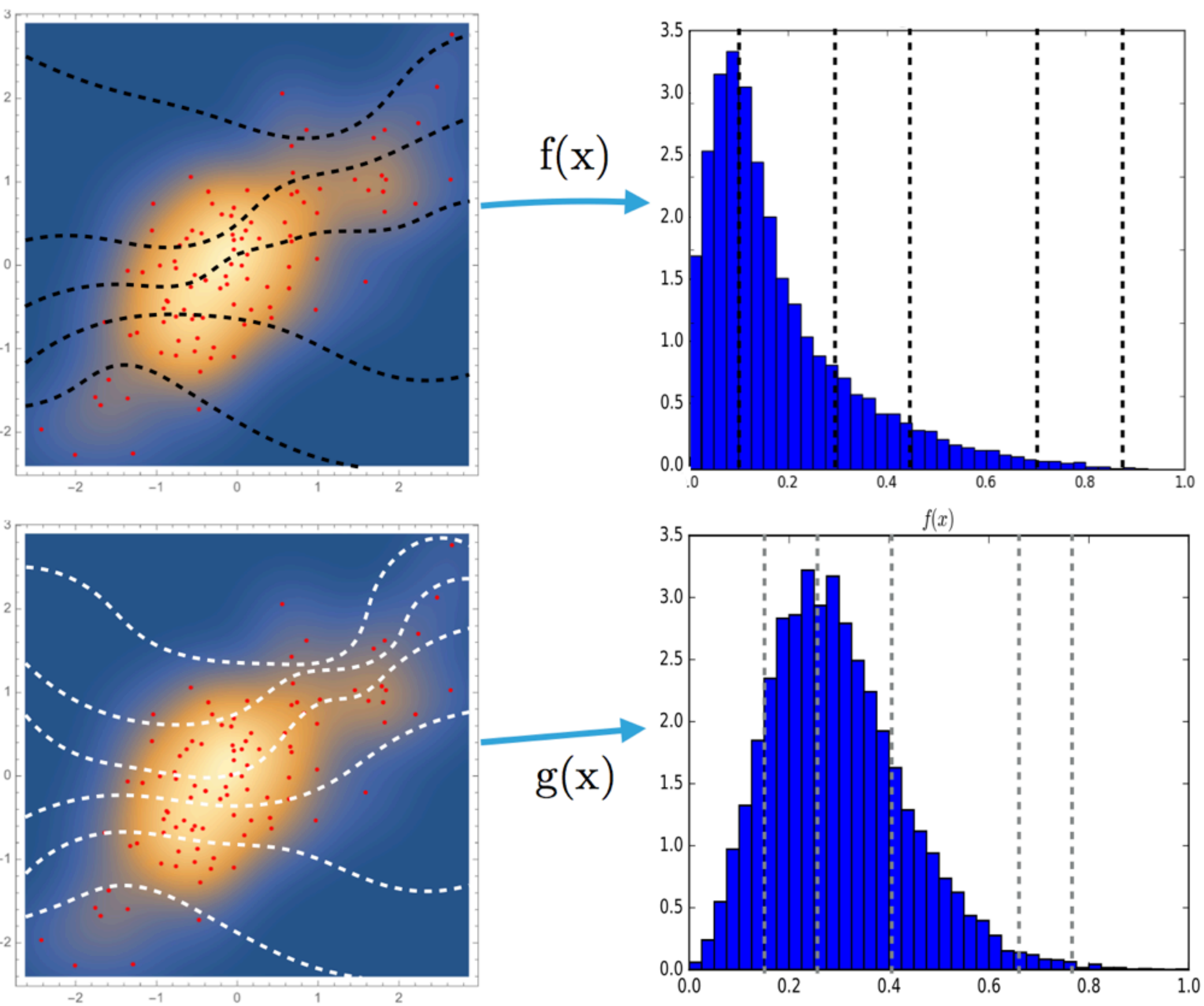
- provides context for NN solution
- defines problem
- does NN live in this space?
- Can it be compactly represented?
- Yes or No are both interesting!

II. Define mapping metric

- how do you compare two solutions?
- can’t use functional identity or linear correlation

Discriminant Similarity

Input space **Output**



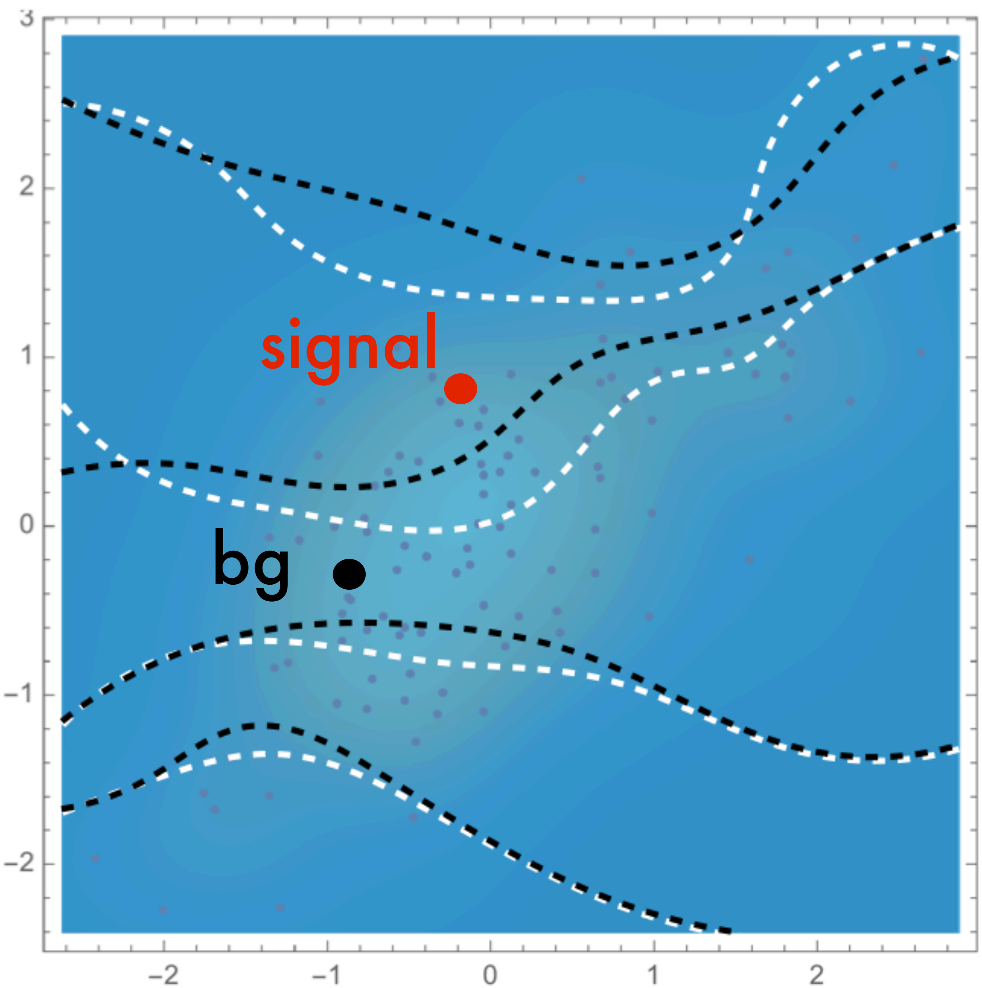
Function sameness

Complete equivalence
not the idea

Any 1:1 transformation
of function has no impact
in our context

Only care about the
ordering of points
not the actual function
values

Discriminant ordering

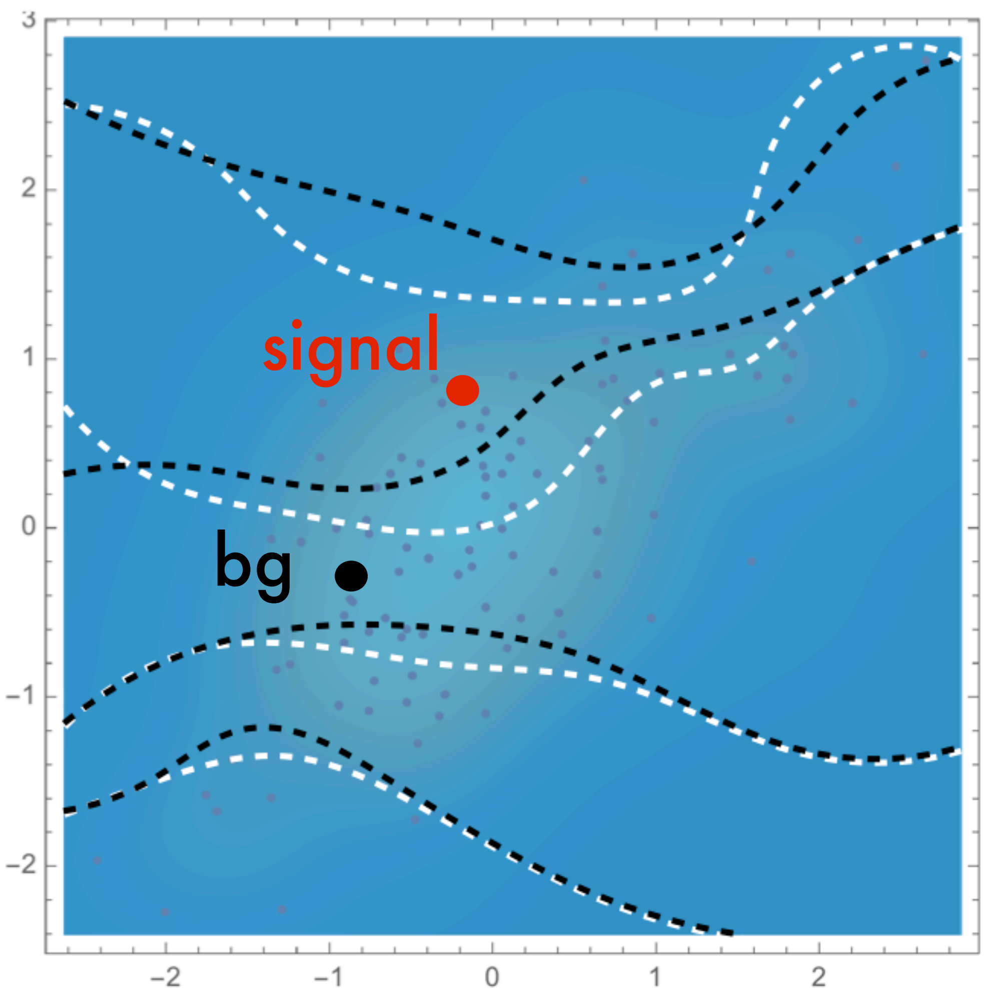


Consider how two functions treat a pair of points

$$f(x_{sig}) - f(x_{bg})$$
$$g(x_{sig}) - g(x_{bg})$$

Do these have the same sign?

Discriminant ordering



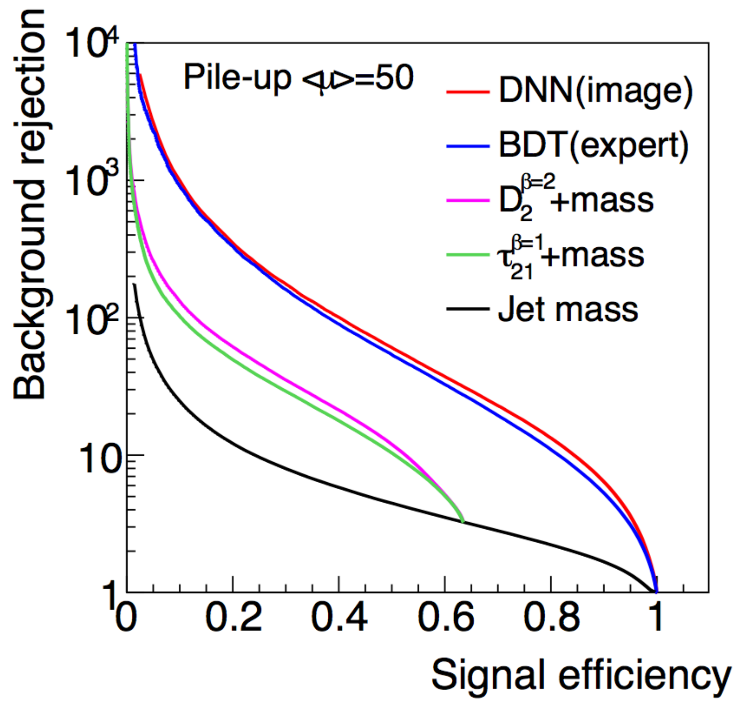
Evaluate how often they give a bg-sig pair the same ordering.

$$\mathrm{DO}(x, x') = \Theta\Big(\big(f(x) - f(x')\big)\big(g(x) - g(x')\big)\Big)$$

Sample the space.

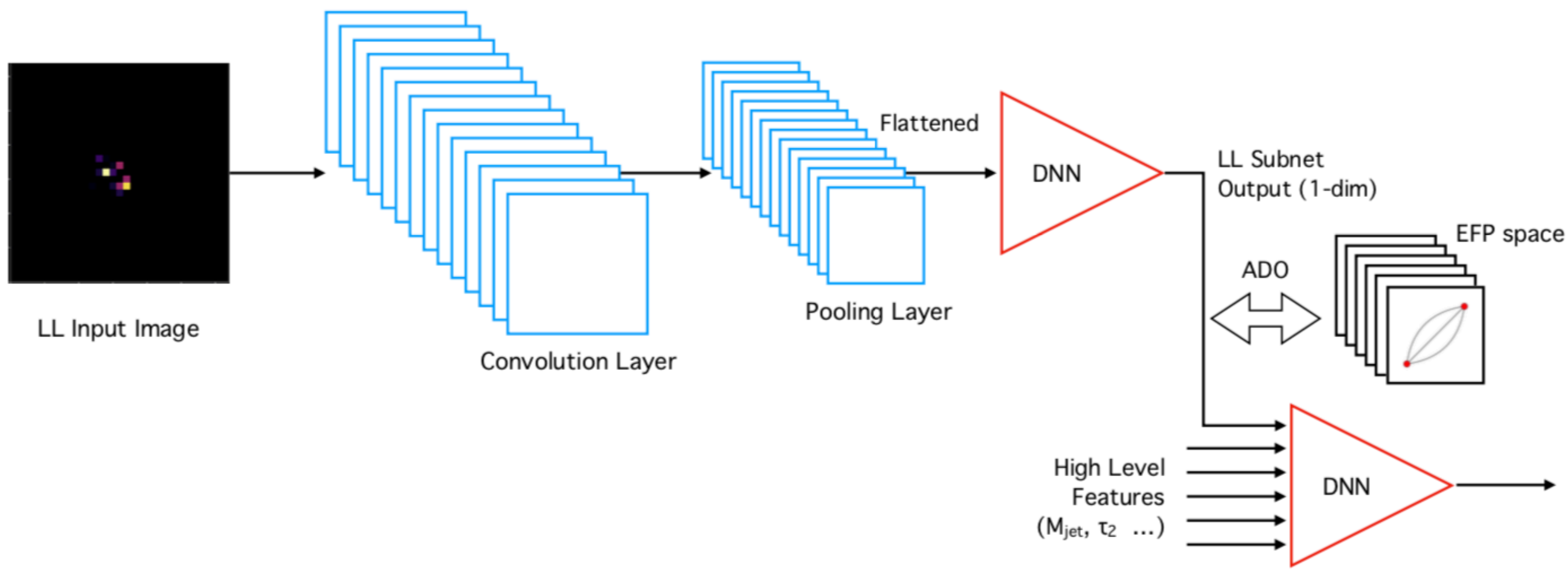
$$\mathrm{ADO} = \int \mathrm{d}x\, \mathrm{d}x'\, p_{\mathrm{sig}}(x)\, p_{\mathrm{bkg}}(x')\, \mathrm{DO}(x, x').$$

The problem



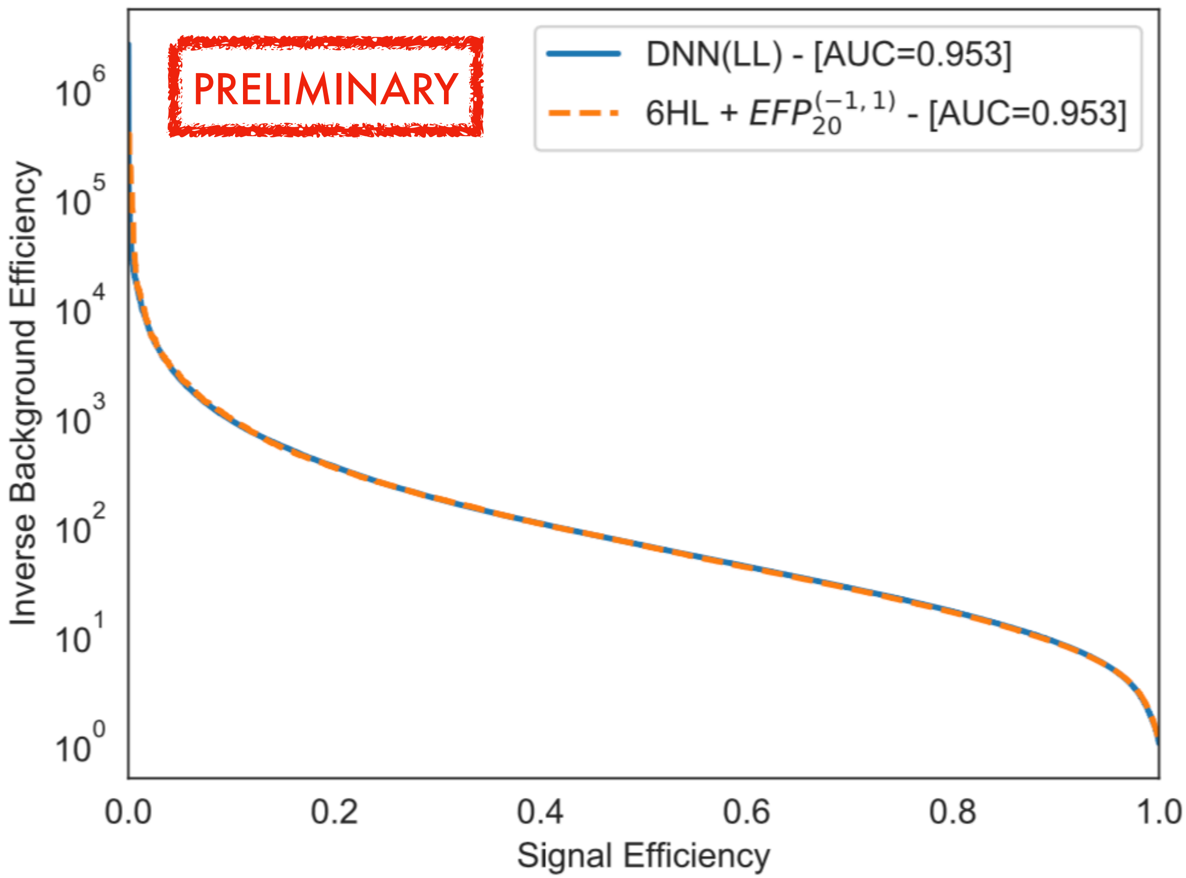
Two approaches:
(1) find the gap
(2) build from scratch

Find the gap

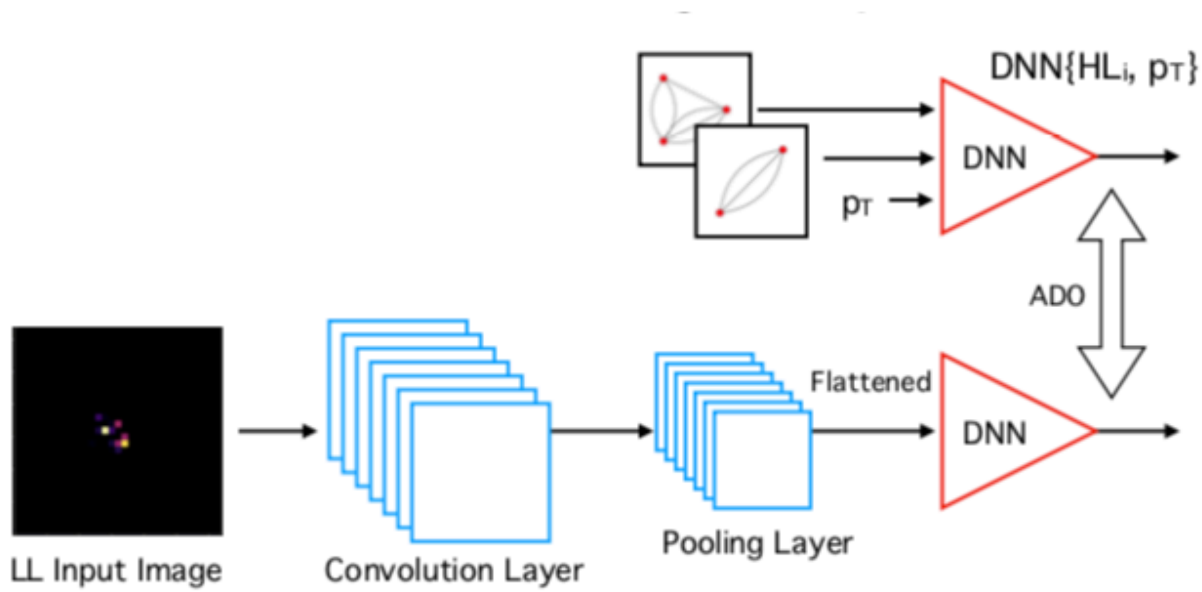


LL Input Image
Convolution Layer
Pooling Layer
Flattened
DNN
LL Subnet
Output (1-dim)
EFP space
ADO
High Level
Features
(M$_{jet}$, τ$_2$ ...)
DNN

It works!


PRELIMINARY
DNN(LL) - [AUC=0.953]
6HL + $EFP_{20}^{(-1,1)}$ - [AUC=0.953]
Inverse Background Efficiency
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0.0
0.2
0.4
0.6
0.8
1.0
Signal Efficiency

Build from scratch

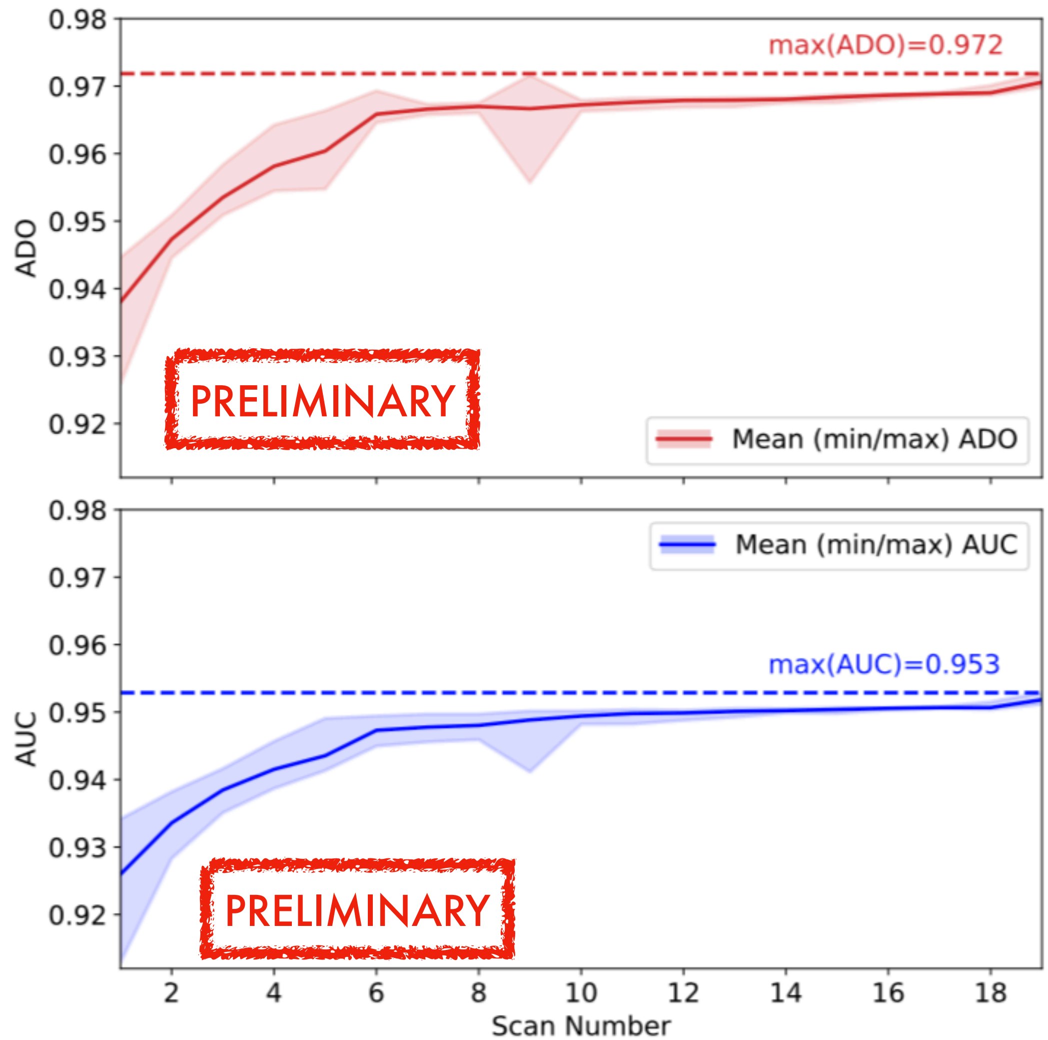


Preliminary

A single point
<u>in this space:</u>

Is very similar
to NN(LL) sol

Captures most of
performance of HL sol

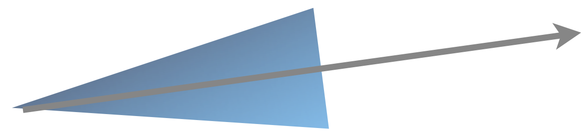


Adding more points
approaches the fulls solution.

Muon isolation

Isolated muons

Muons from jets



<u>Problem</u>

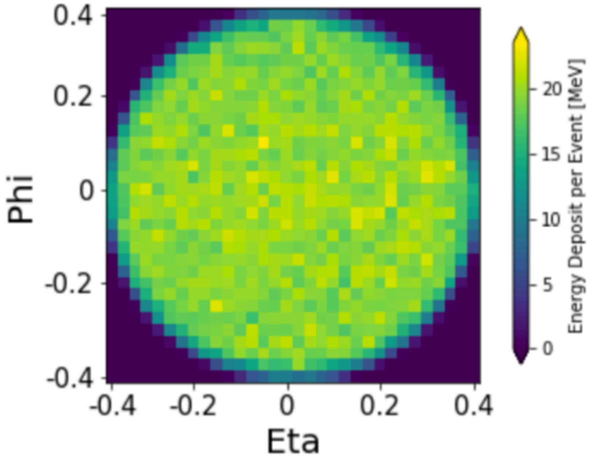
Jet can be soft, not reconstructed

Jets are strongly produced, large background

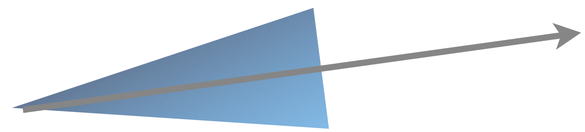
Muon isolation

Isolated muons

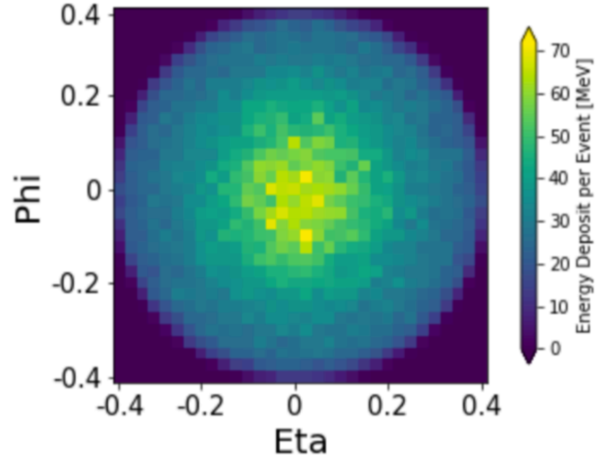
Little cal deposition



Muons from jets



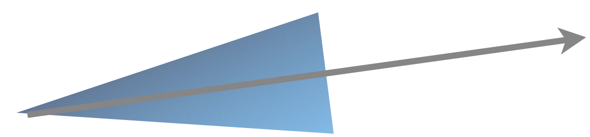
Large cal deposition



Muon isolation

Isolated muons

Muons from jets

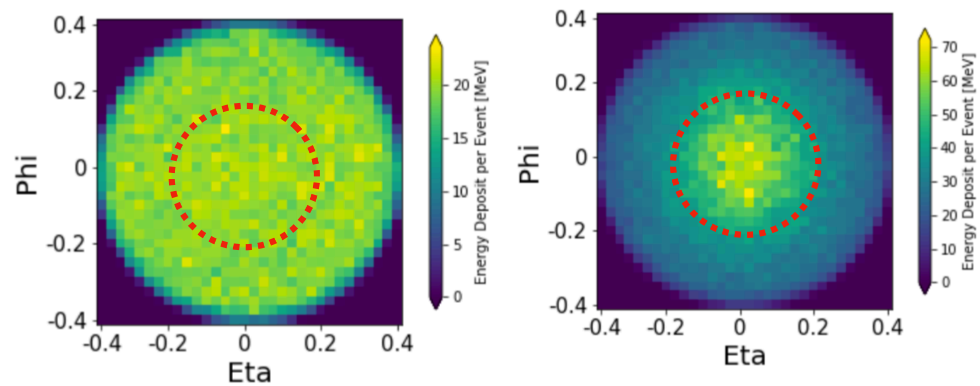


$$I_\mu(R_0) = \sum_{R<R_0} \frac{p_{\mathrm{T}}^{\mathrm{calo}}}{p_{\mathrm{T}}^{\mathrm{muon}}}$$

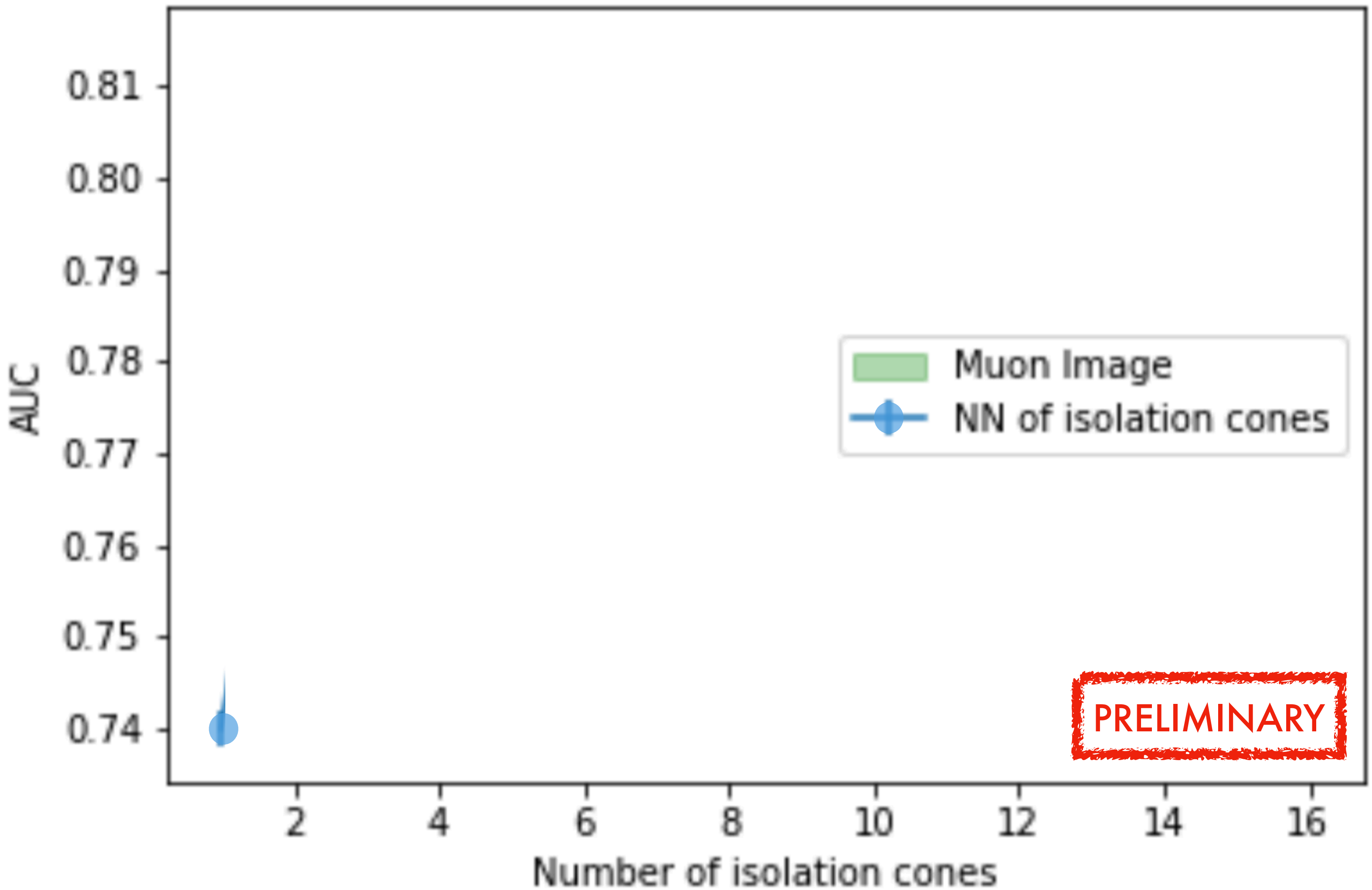
<u>Standard Approach</u>

Calculate "isolation"

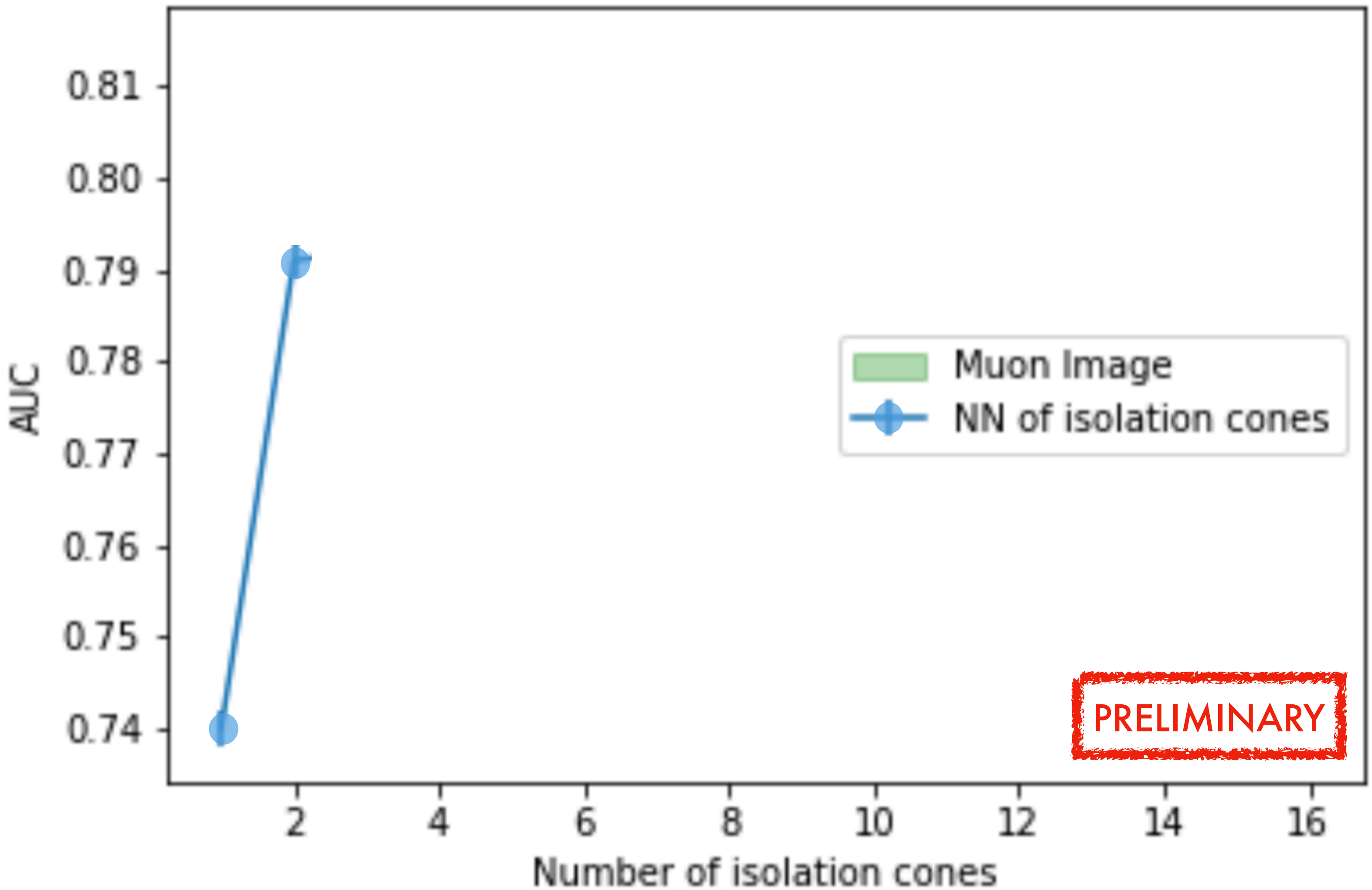
Energy in a cone around muon.



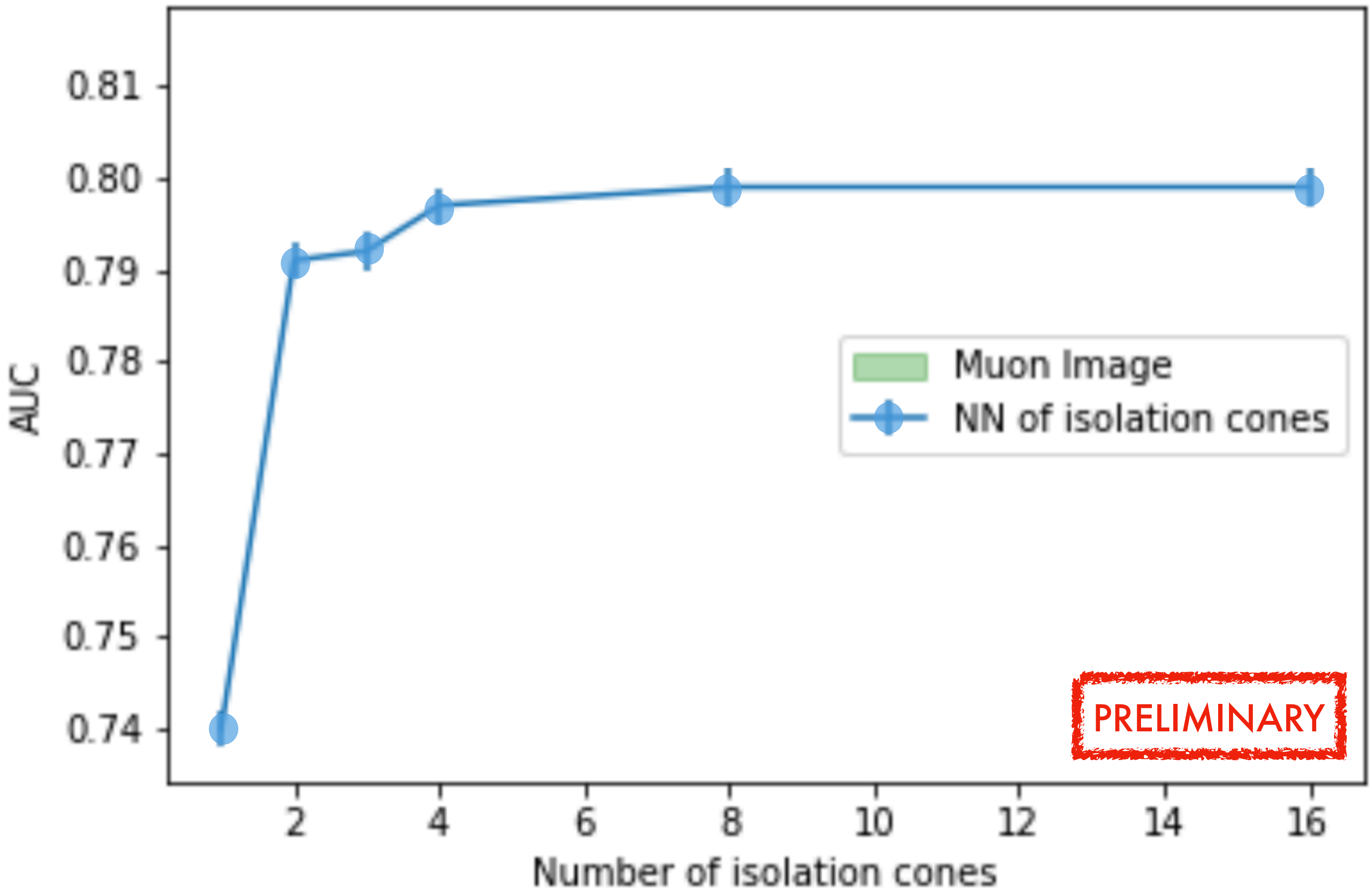
Isolation


0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
AUC
Muon Image
NN of isolation cones
PRELIMINARY
2
4
6
8
10
12
14
16
Number of isolation cones

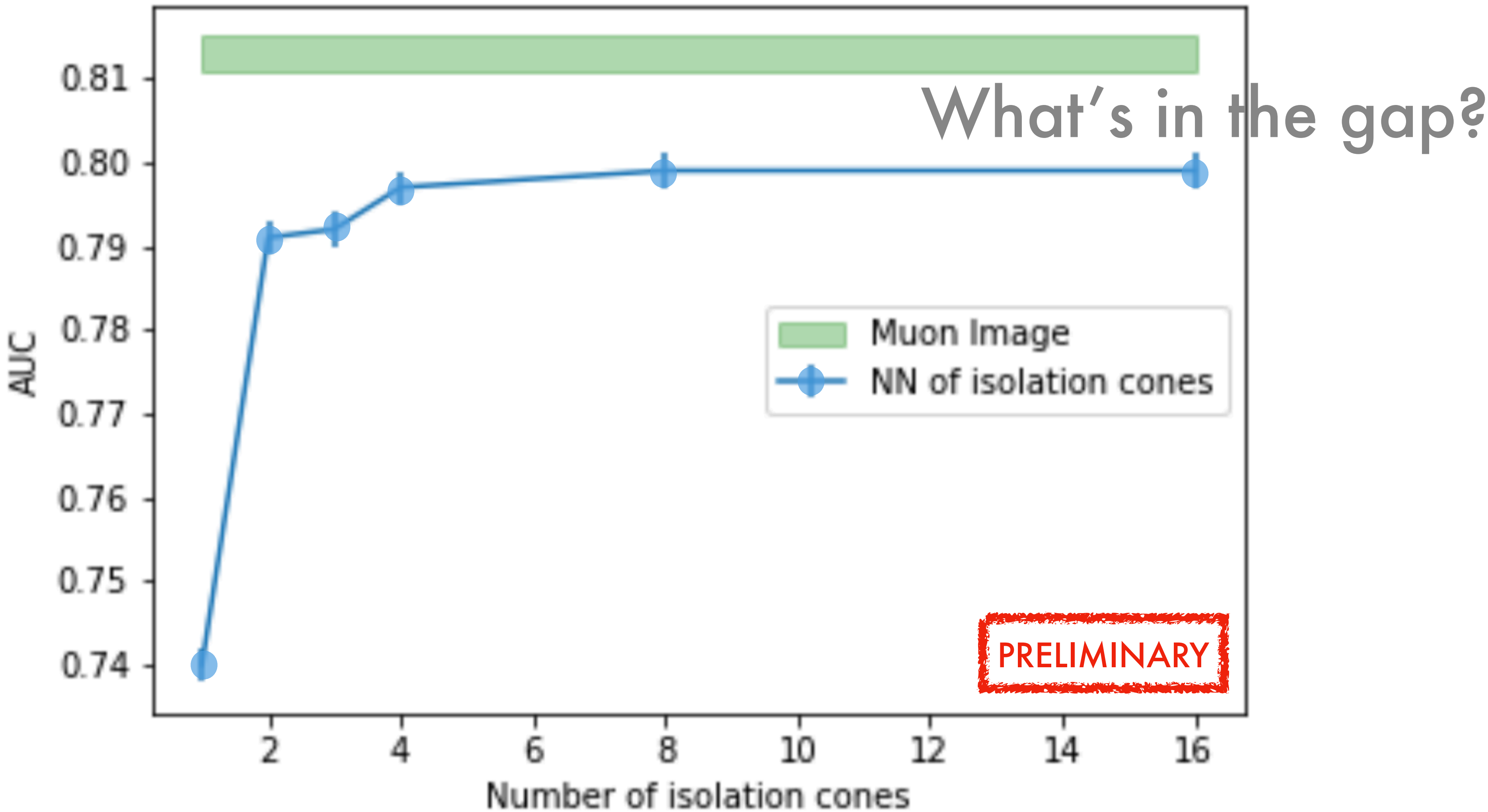
More isolation



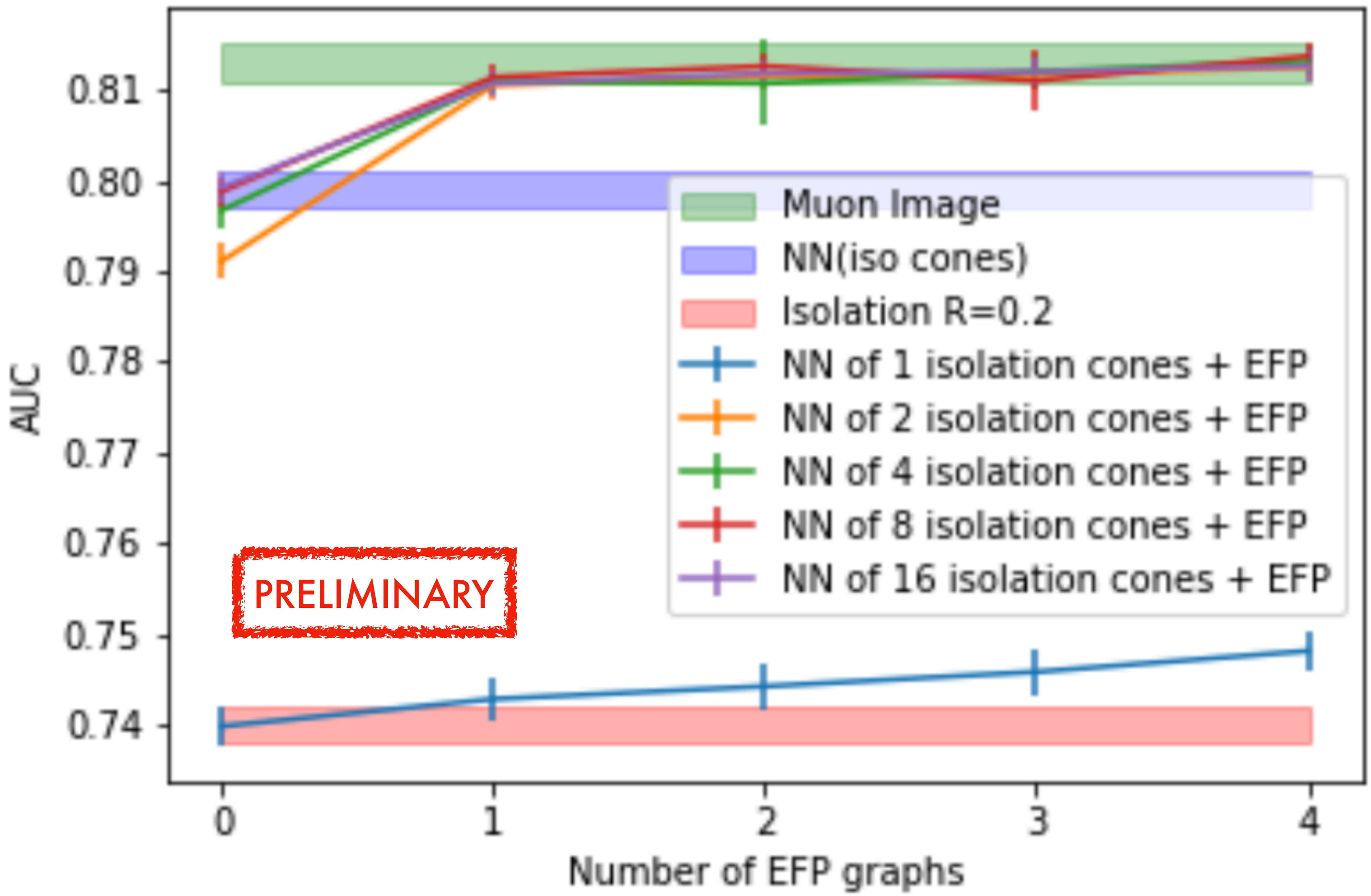
Most isolation



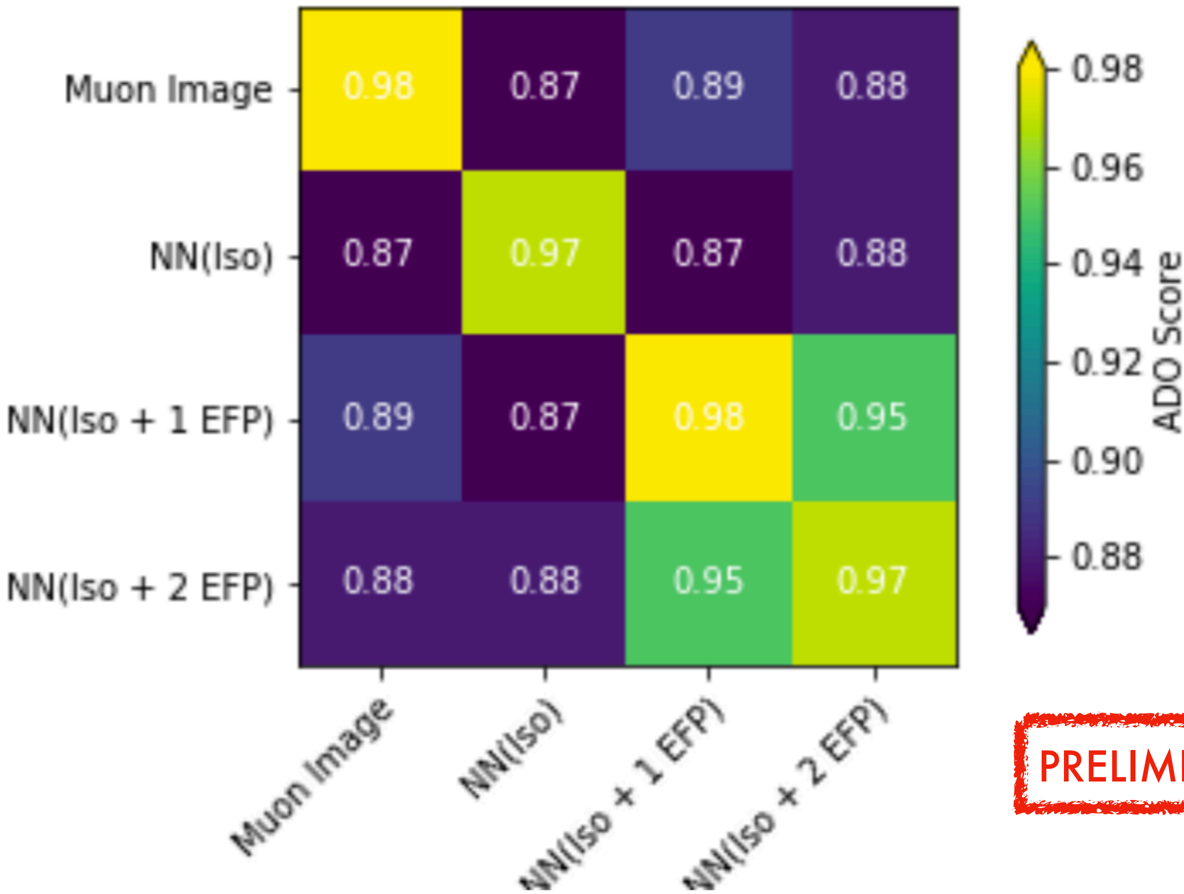
What can ML do?



Close the gap?



Information



PRELIMINARY

Conclusions

Deep Learning is a powerful new tool

offers faster learning of nonlinear functions

We have many appropriate tasks in HEP

traditional heuristics should be re-examined

No replacement for human intelligence

garbage in will still give garbage out