Deep Learning for New Particle Searches

Daniel Whiteson, UC Irvine

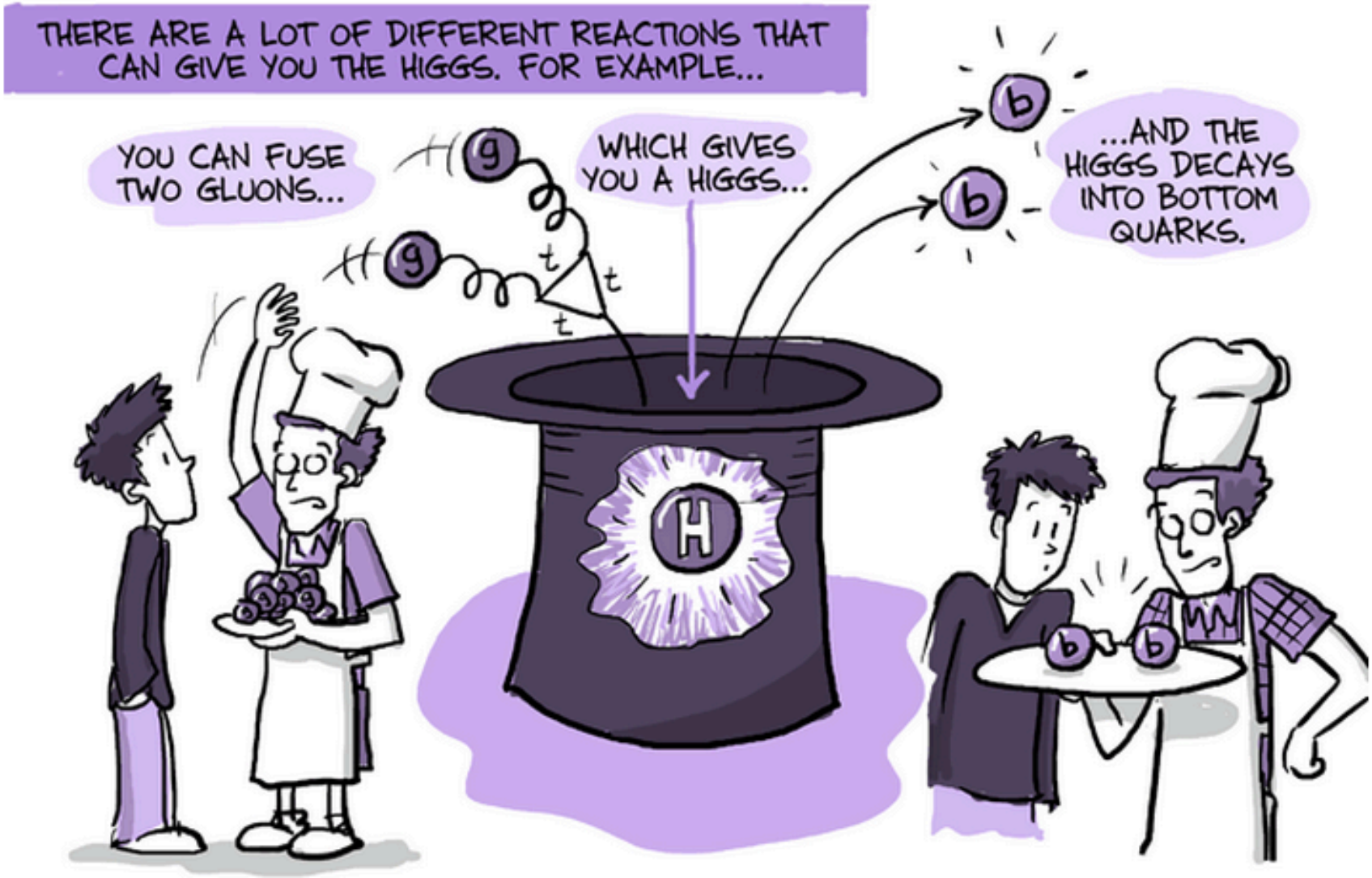
Oct 2017, DNP

What is Deep Learning?

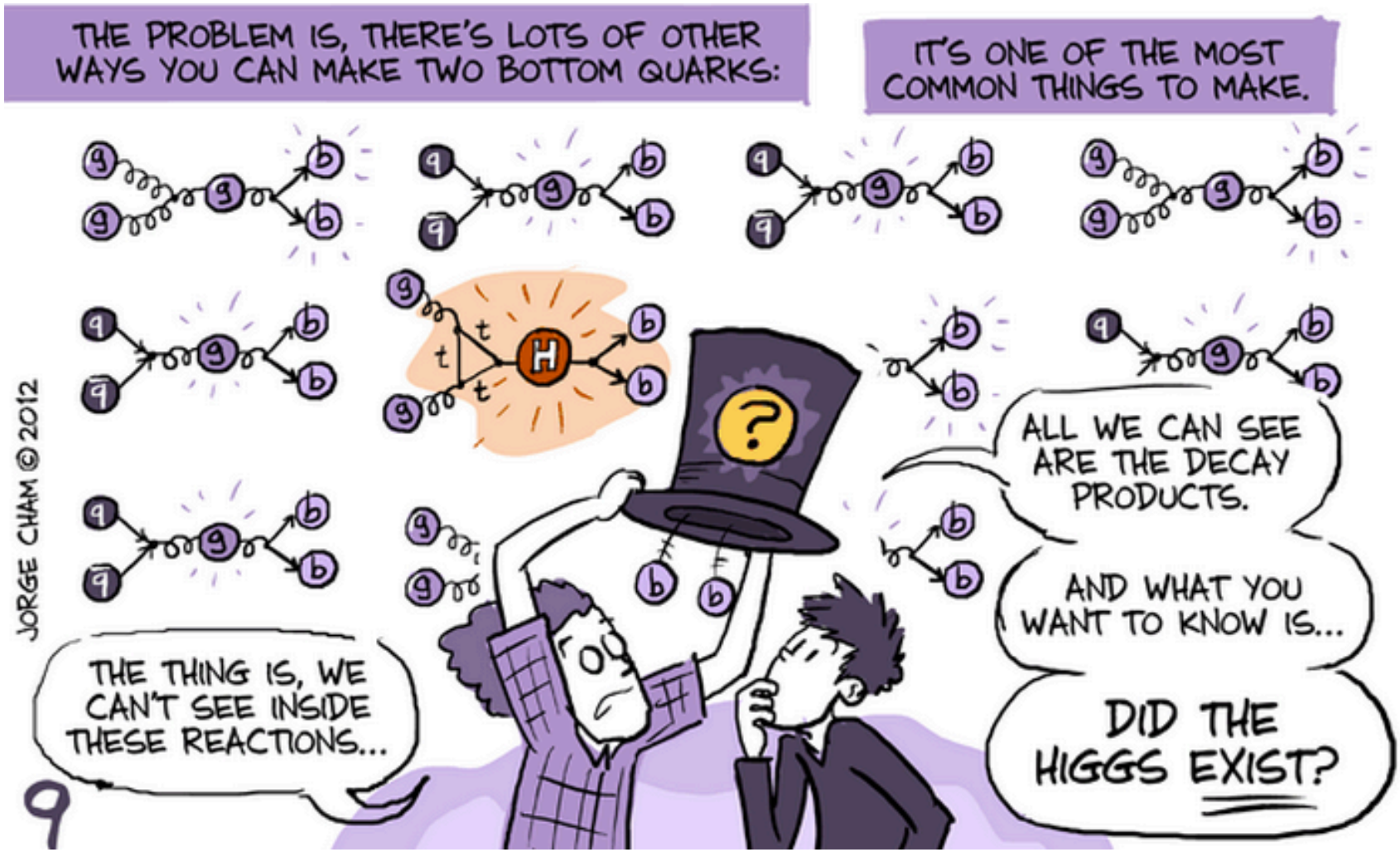
Unambiguous data

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

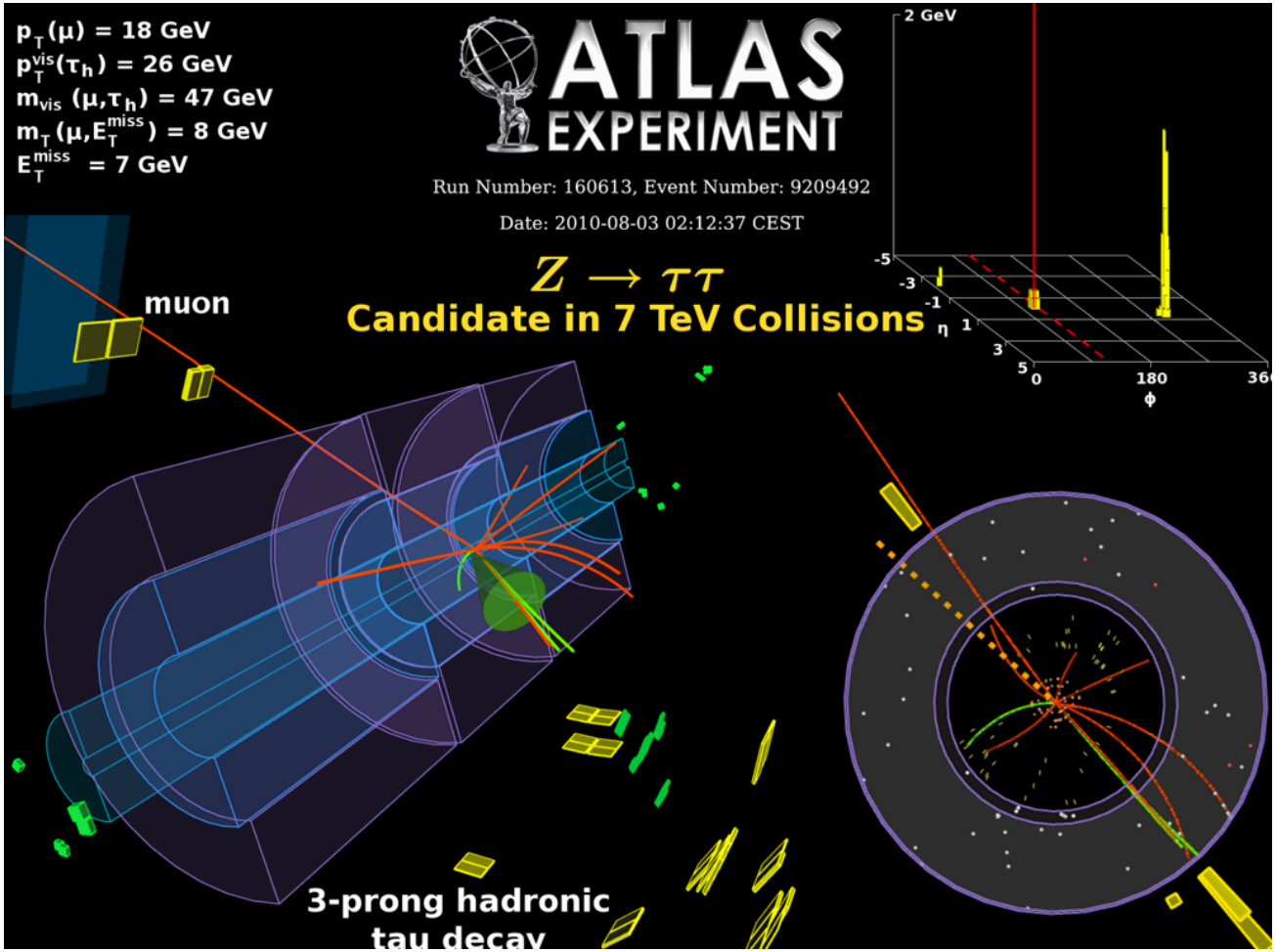
Making a new particle



Backgrounds



Why statistics?

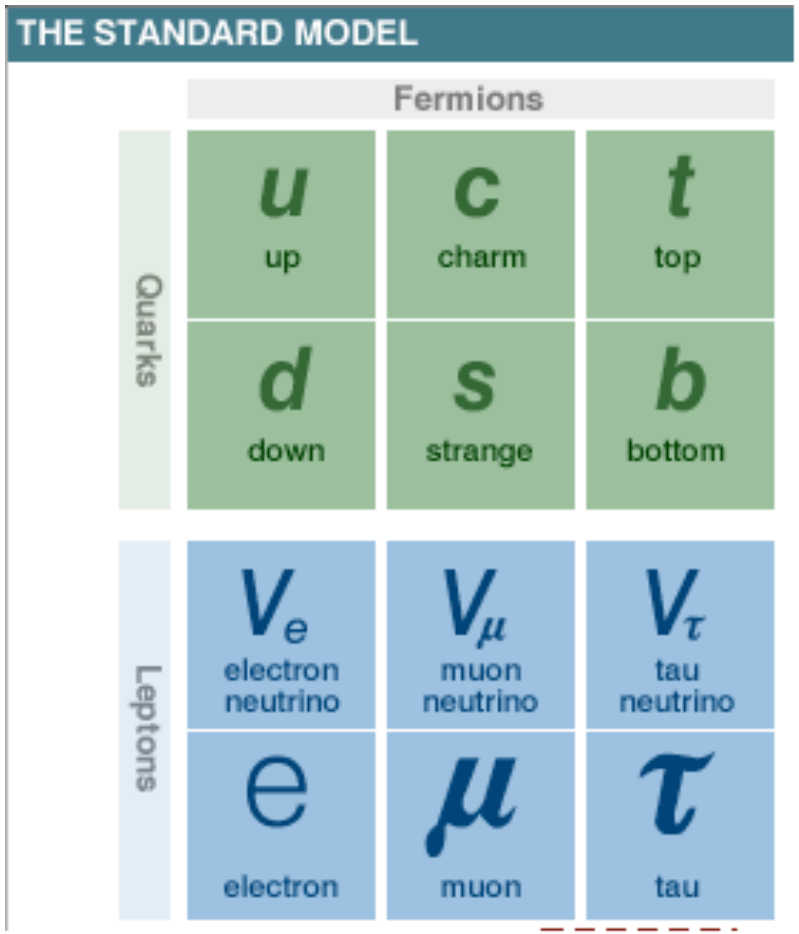


The nature of our data demands it.

Hypothesis testing

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

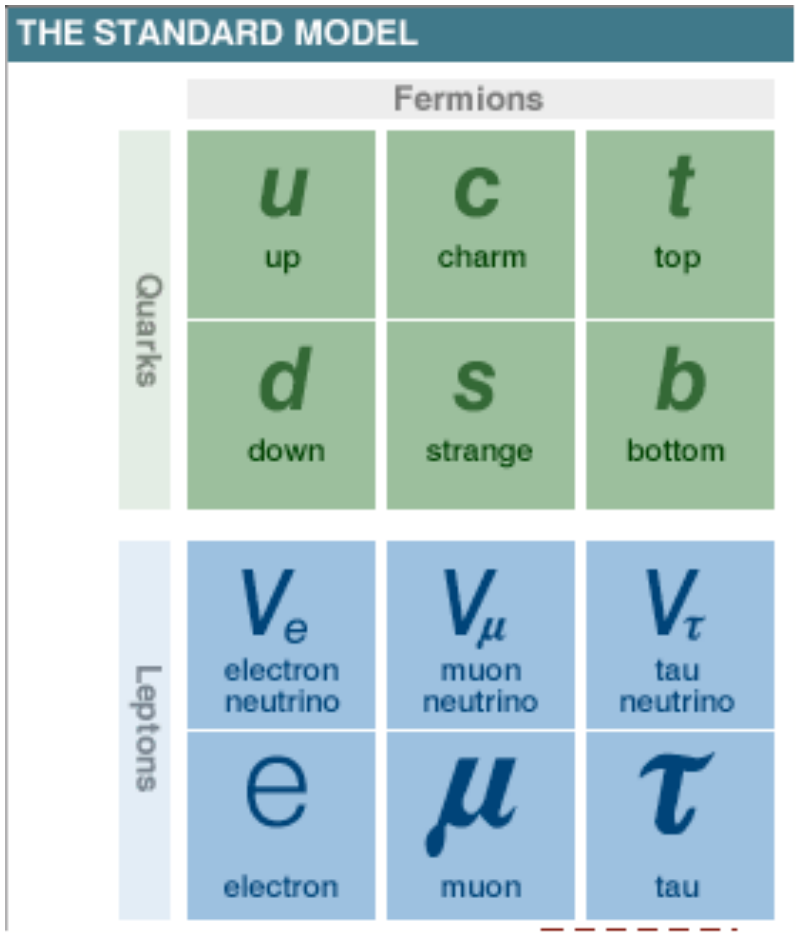
1.



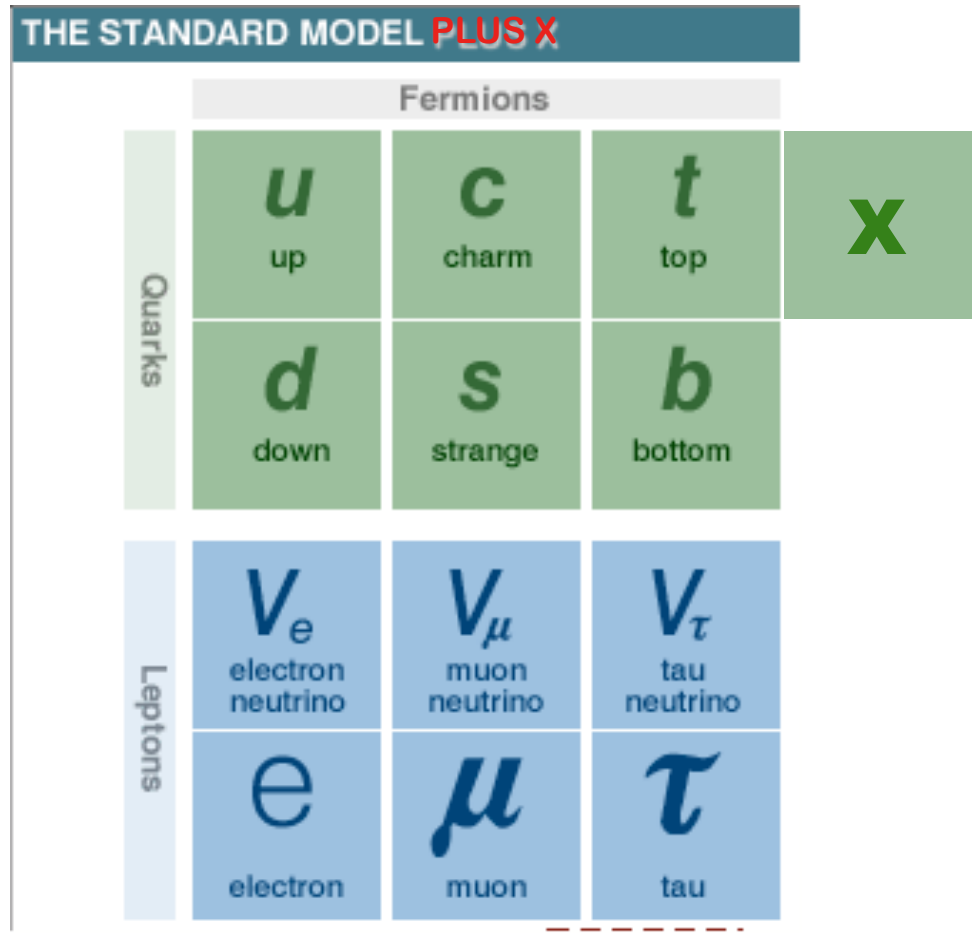
Hypothesis testing

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

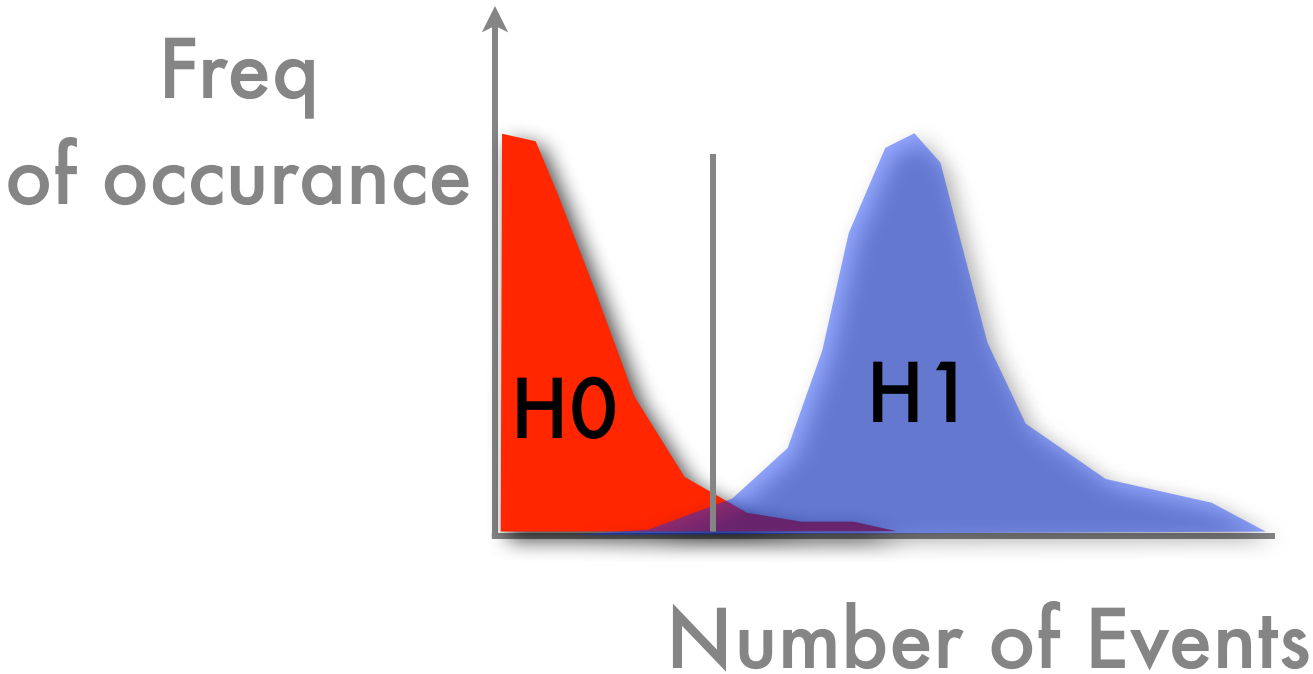
1.



2.

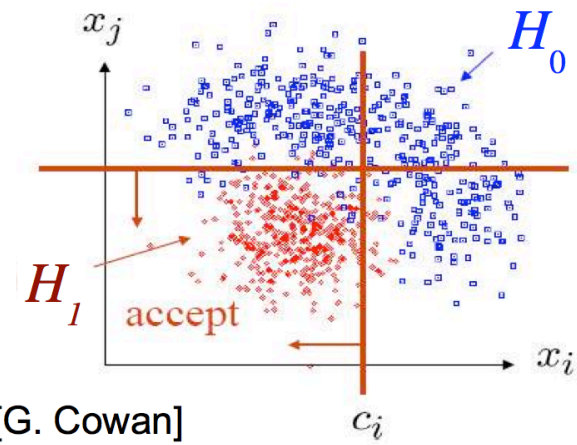


Example

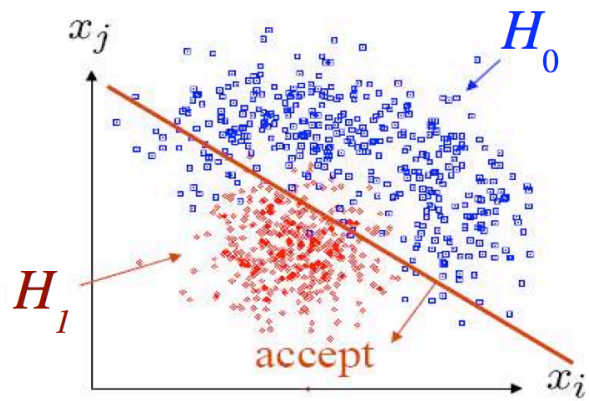


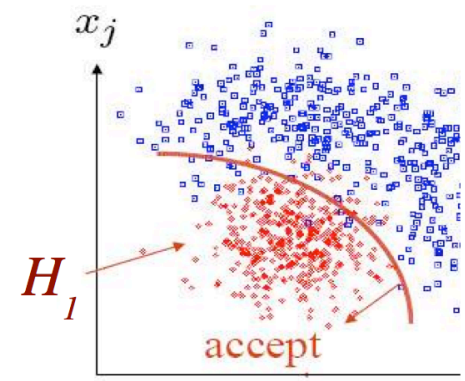
A threshold makes sense.
Choice of position balances
false vs missed discovery

More complicated



[G. Cowan]





Neyman-Pearson

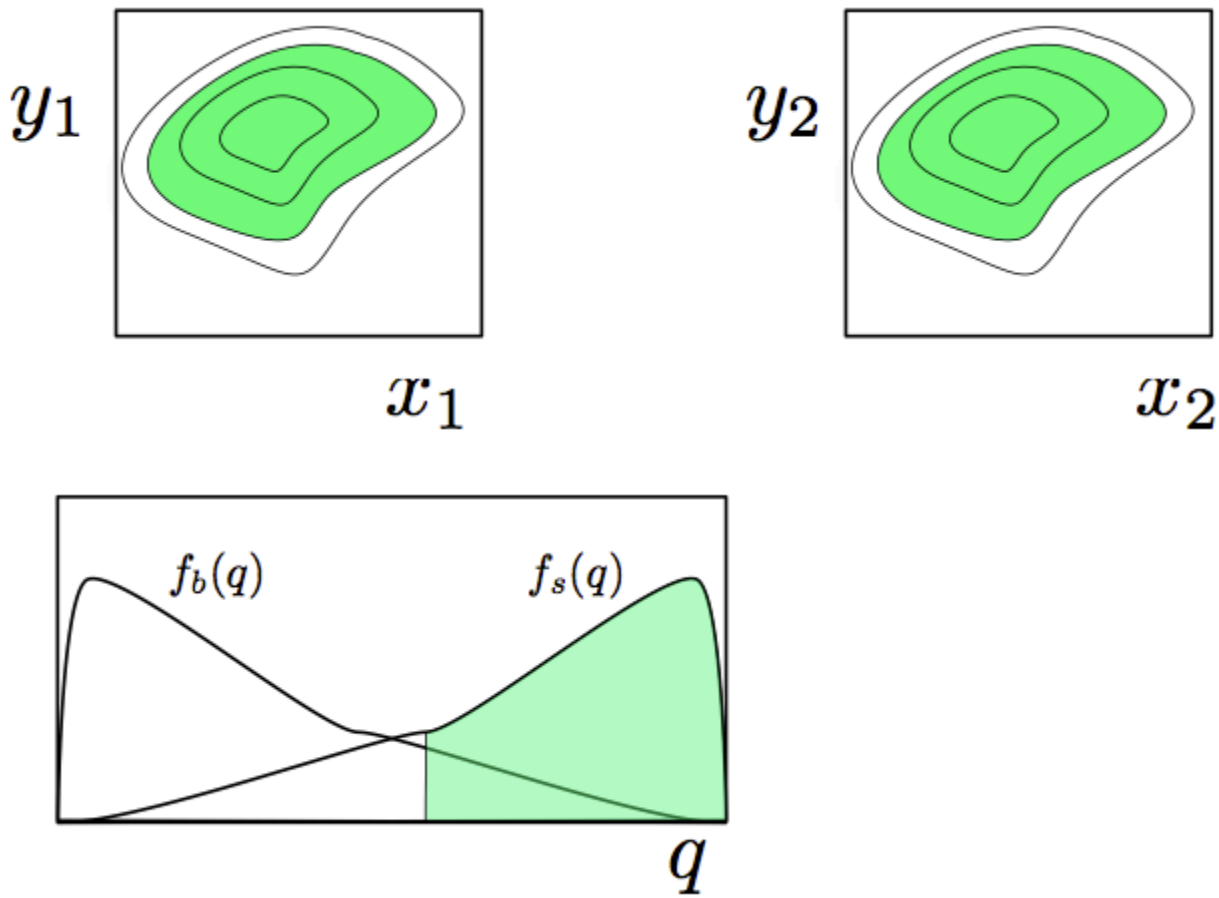
NP lemma says that the best decision boundary is the likelihood ratio:

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

(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

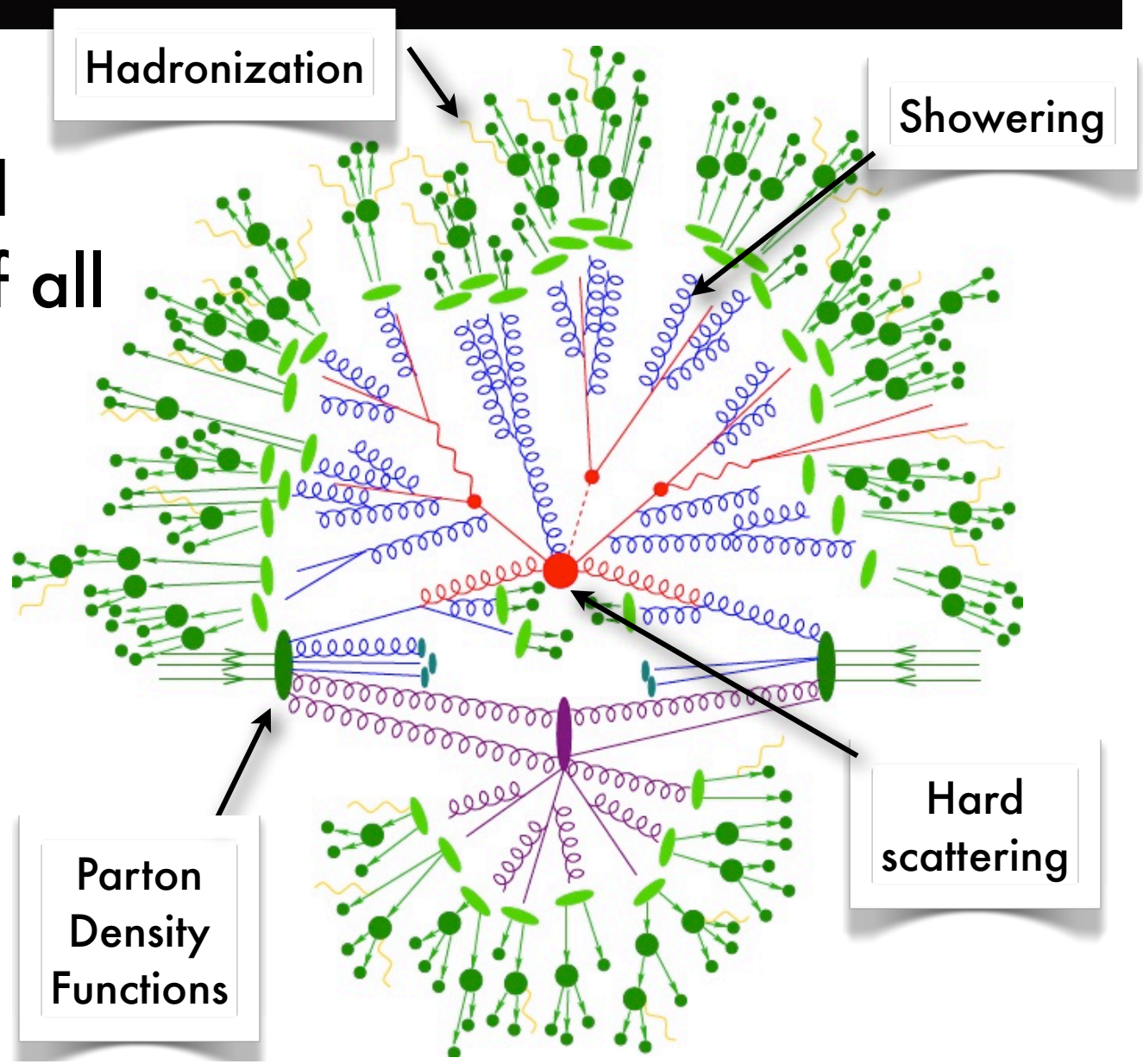
P(data|theory)

In general

We have a good understanding of all of the pieces

Do we have

f(data|theory)?



In general

What would

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

look like?

The dream

f(data | final-state particles P)

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

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

x f(hard scatter products M | theory)

Sum over all possible intermediate P,S,M

The dream

f(hard scatter products M| theory)

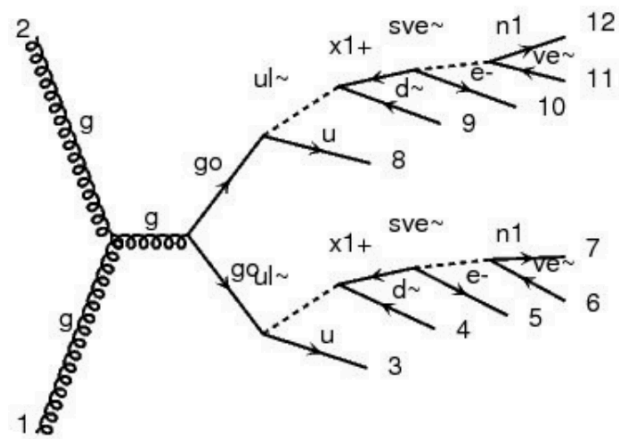
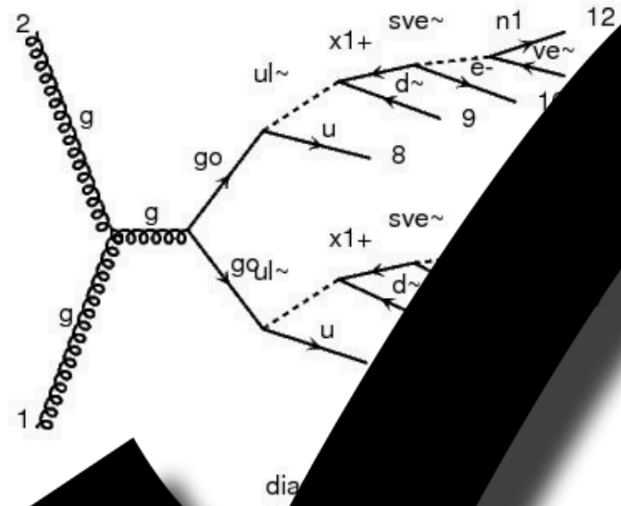


diagram 1

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

The dream

f(hard scatter products M| theory)



The defined
automat ulators exist
for almost any (B)SM theory

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)

<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 f(M|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

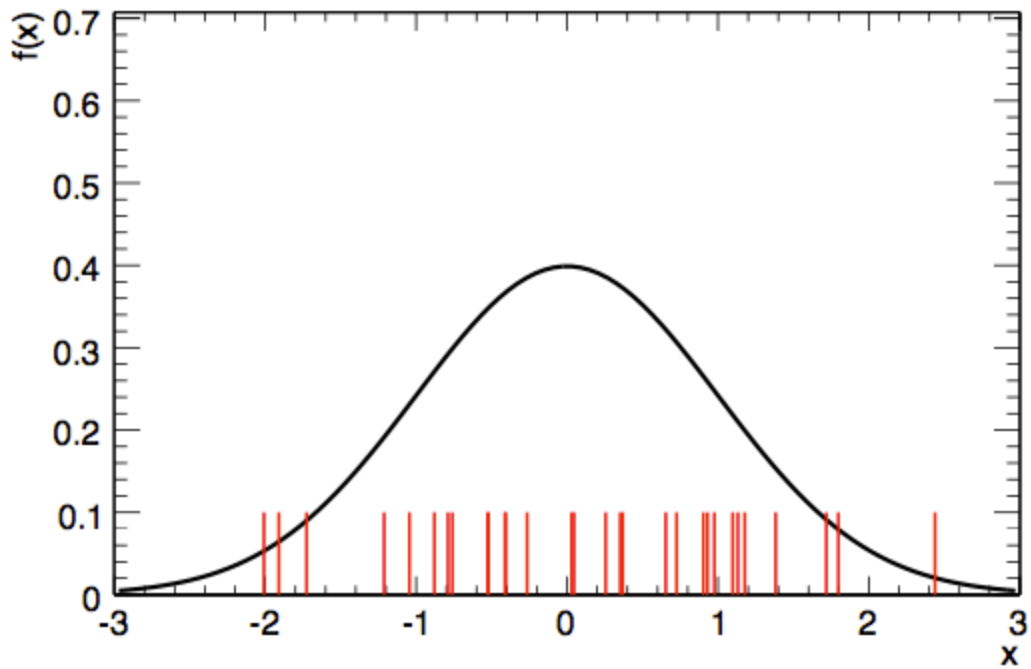
Iterative simulation strategy, no overall PDF

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

The problem

Don't know PDF, have events drawn from PDF

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



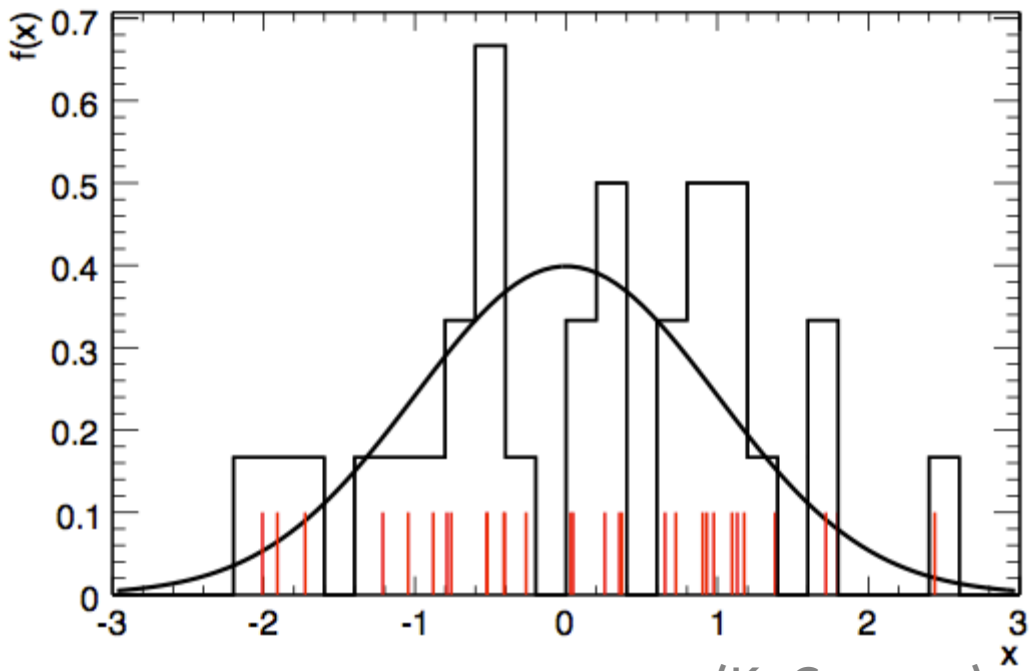
(K. Cranmer)

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

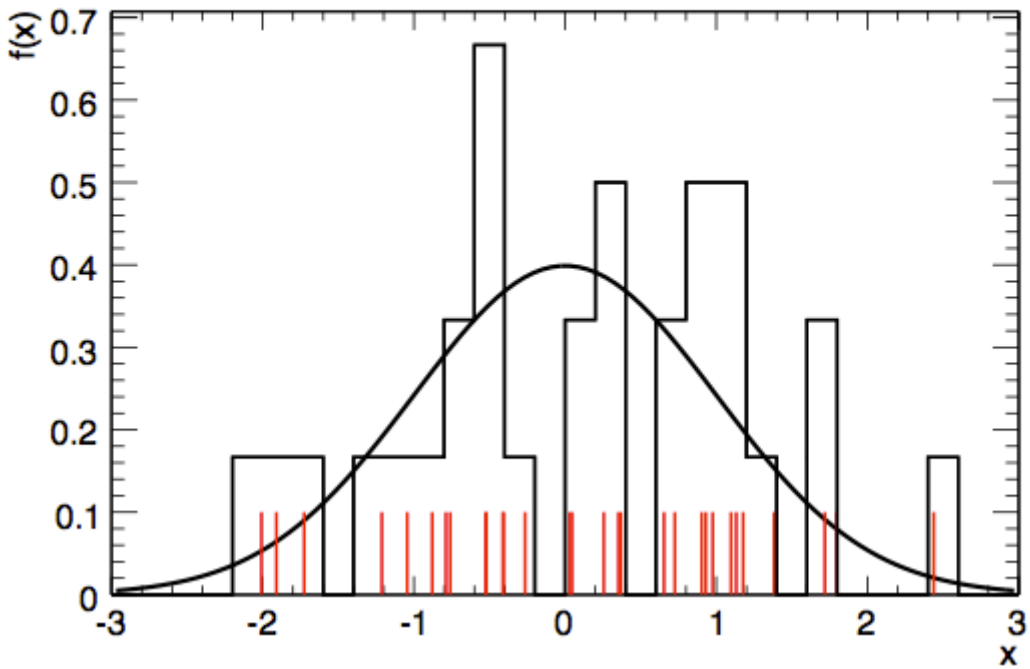
How many events do you need to describe a 1D distribution? O(100)

An n-D distribution?

$O(100^n)$

!!

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



(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 ered 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:

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

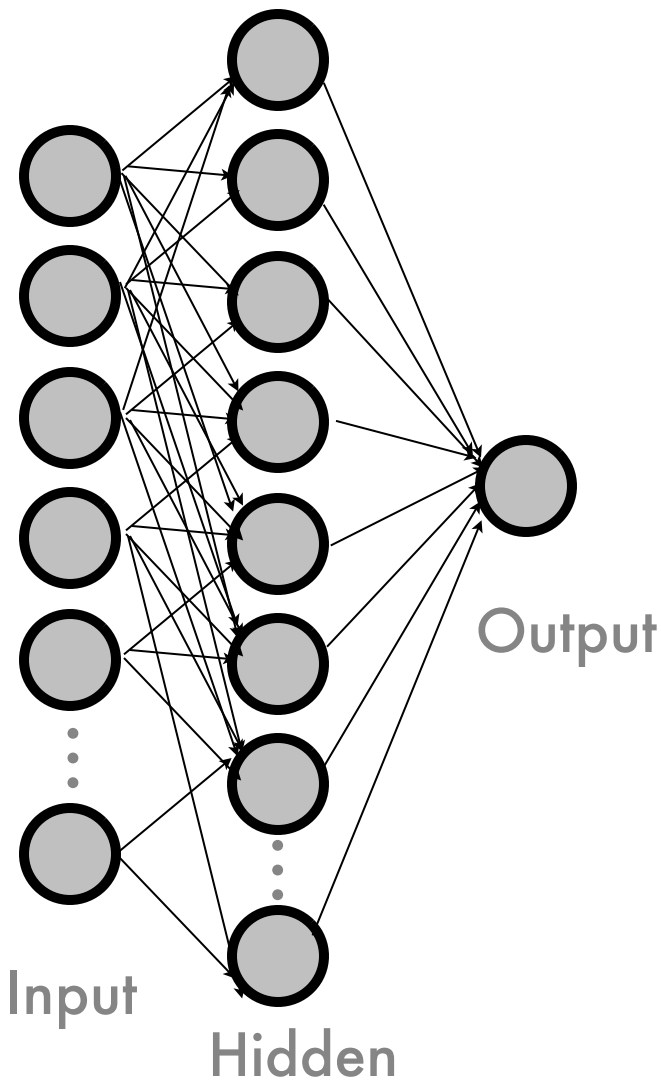
Build f(x)=y(x) out of a pile of convoluted mini-functions

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

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

How complex?

Essentially a functional fit with many parameters



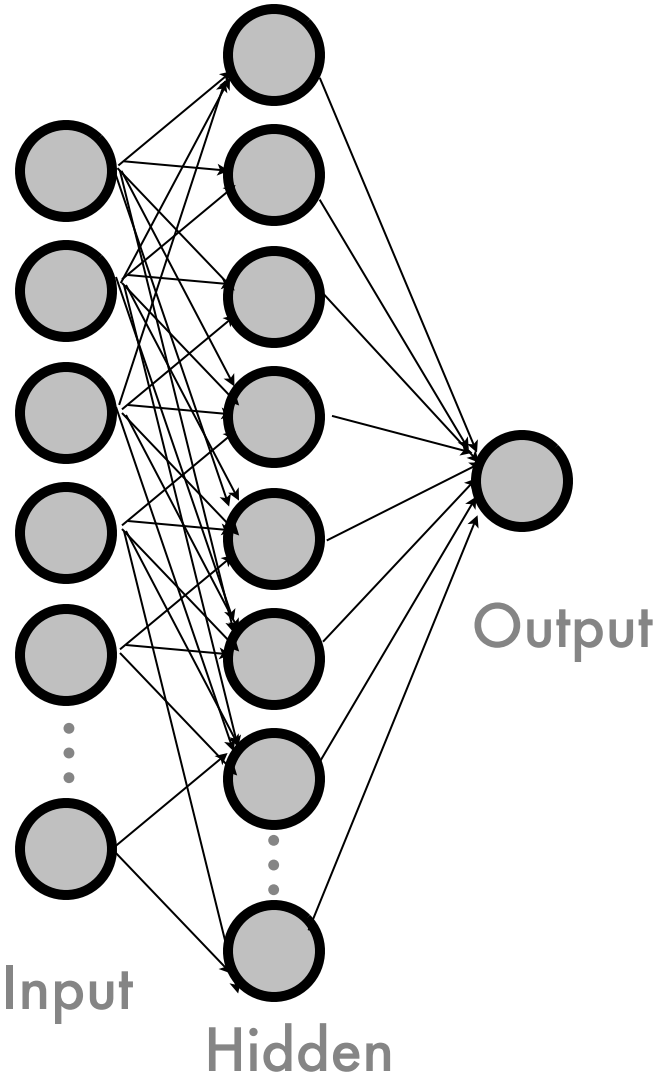
<u>Single hidden layer</u>

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

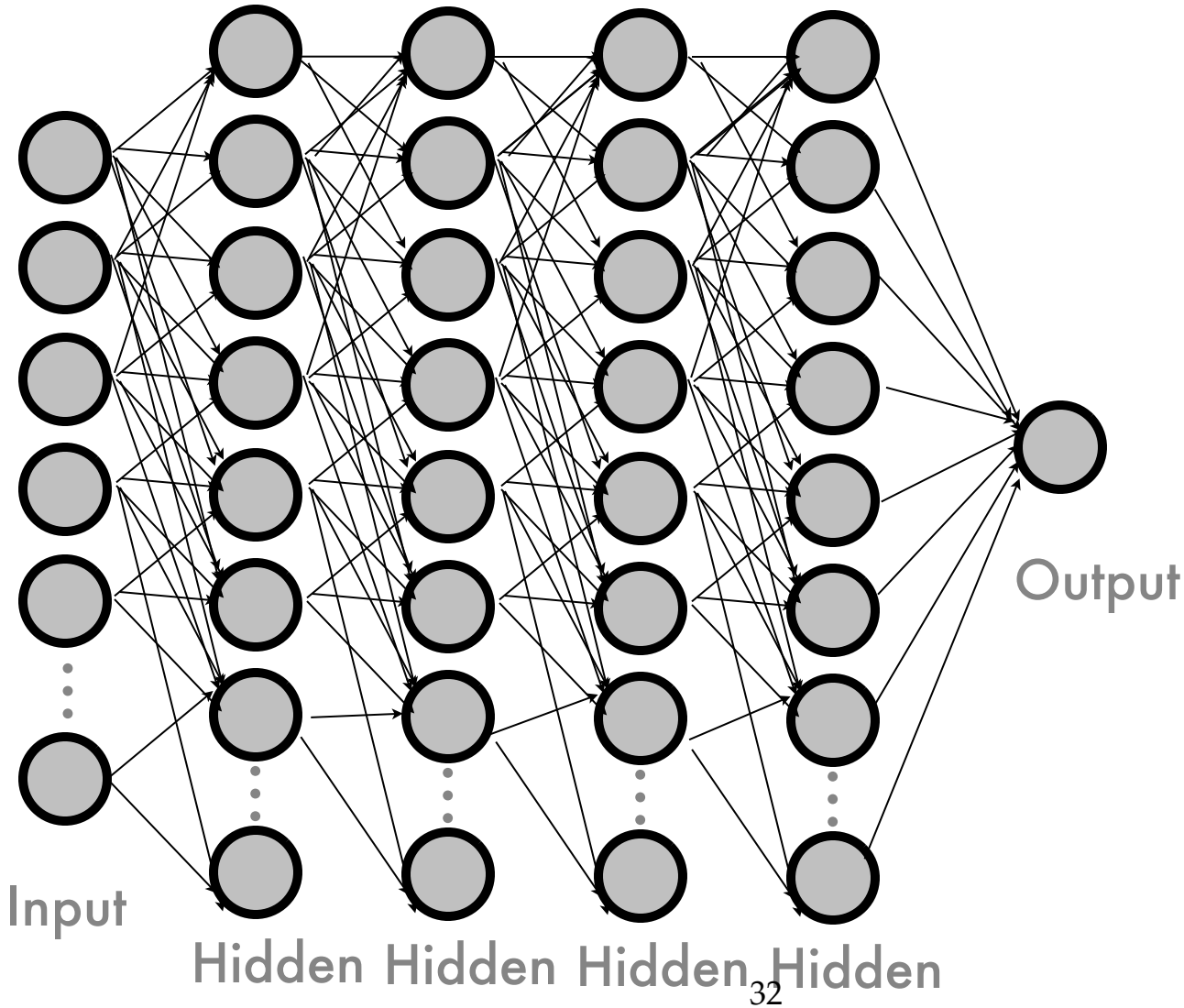
Can't give it too many inputs

Painstaking search through input feature space.

Variable	VBF			Boosted		
	$\tau_{lep}\tau_{lep}$	$\tau_{lep}\tau_{had}$	$\tau_{had}\tau_{had}$	$\tau_{lep}\tau_{lep}$	$\tau_{lep}\tau_{had}$	$\tau_{had}\tau_{had}$
$m_{\tau\tau}^{MMC}$	•	•	•	•	•	•
$\Delta R(\tau,\tau)$	•	•	•		•	•
$\Delta\eta(j_1, j_2)$	•	•	•			
m_{j_1,j_2}	•	•	•			
$\eta_{j_1} \times \eta_{j_2}$		•	•			
p_T^{Total}		•	•			
sum p_T					•	•
$p_T(\tau_1)/p_T(\tau_2)$					•	•
$E_T^{miss}\phi$ centrality		•	•	•	•	•
$x_{\tau 1}$ and $x_{\tau 2}$						•
$m_{\tau\tau,j_1}$				•		
m_{ℓ_1,ℓ_2}				•		
$\Delta\phi_{\ell_1,\ell_2}$				•		
sphericity				•		
$p_T^{\ell_1}$				•		
$p_T^{j_1}$				•		
$E_T^{miss}/p_T^{\ell_2}$				•		
m_T		•			•	
$\min(\Delta\eta_{\ell_1\ell_2,jets})$	•					
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.

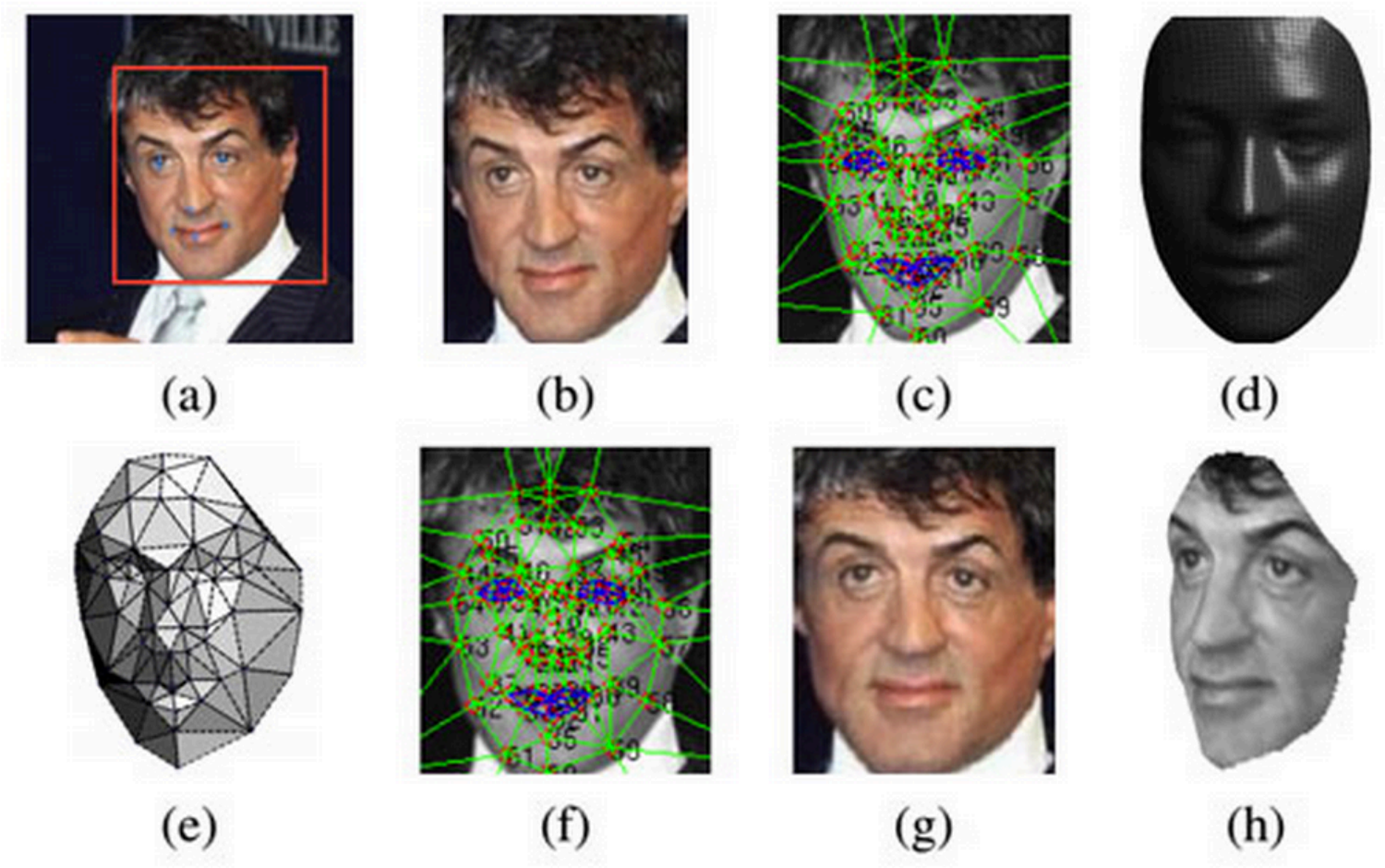
Deep networks



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

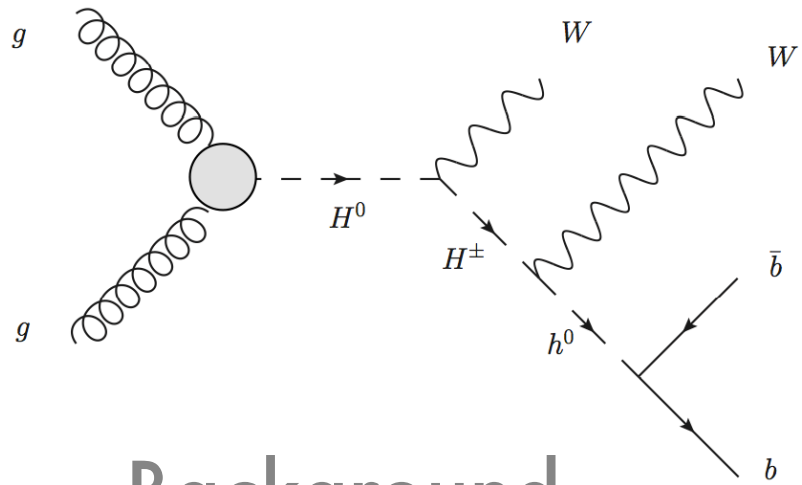
Searching for exotic particles in high-energy physics with deep learning

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

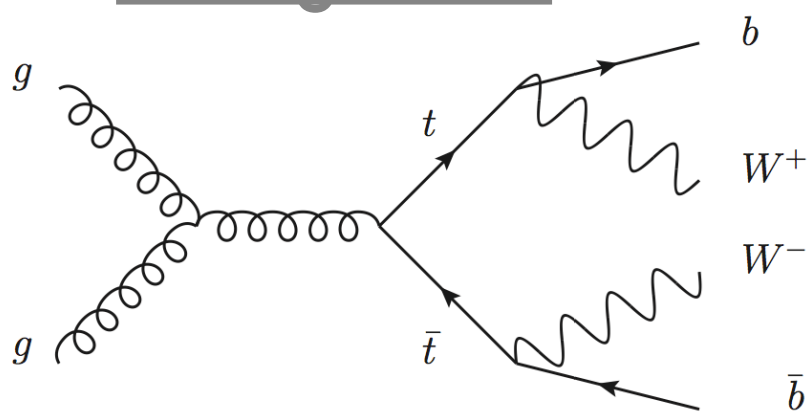
arXiv: 1402.4735

Benchmark problem

Signal



Background



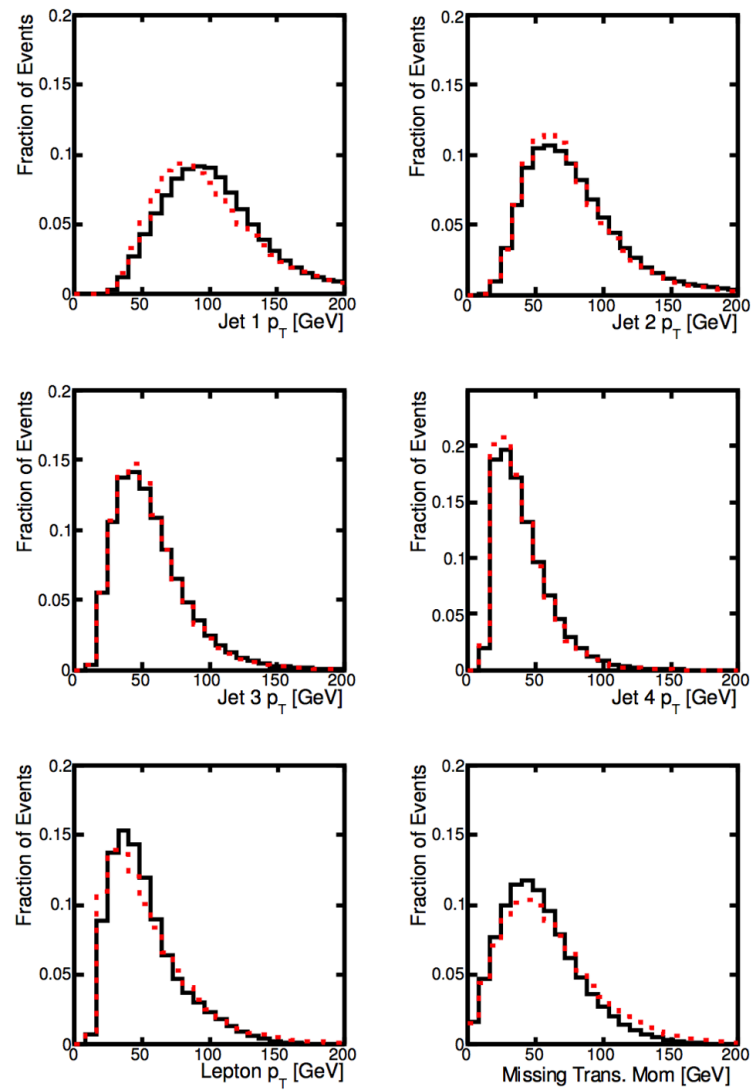
Can deep networks automatically discover useful variables?

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

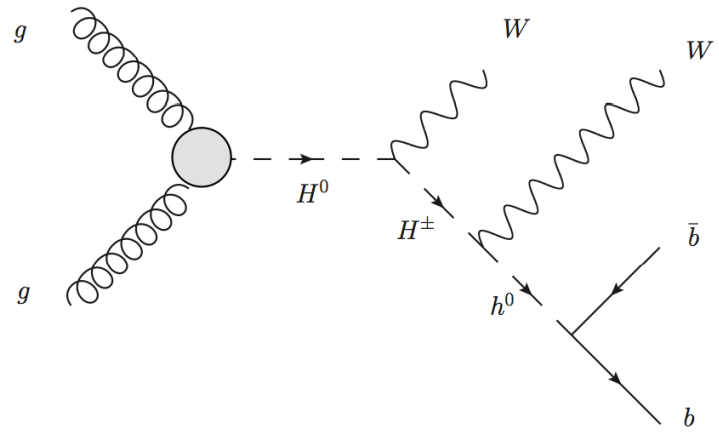


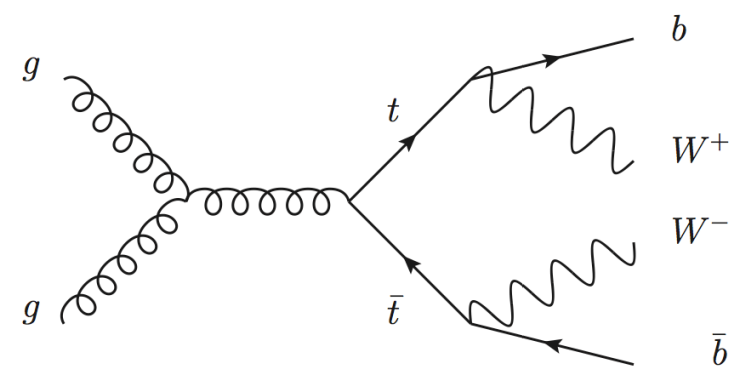
4-vector inputs

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

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

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





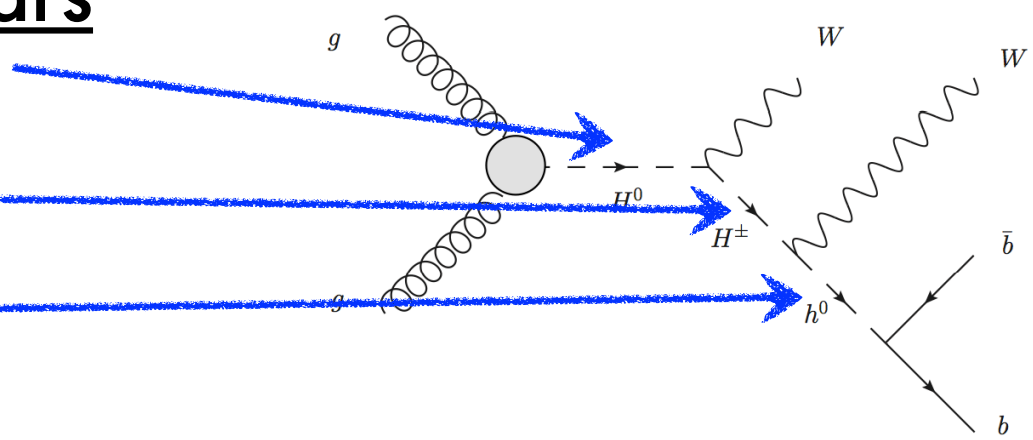
4-vector inputs

7 High-level vars

m(WWbb)

m(Wbb)

m(bb)

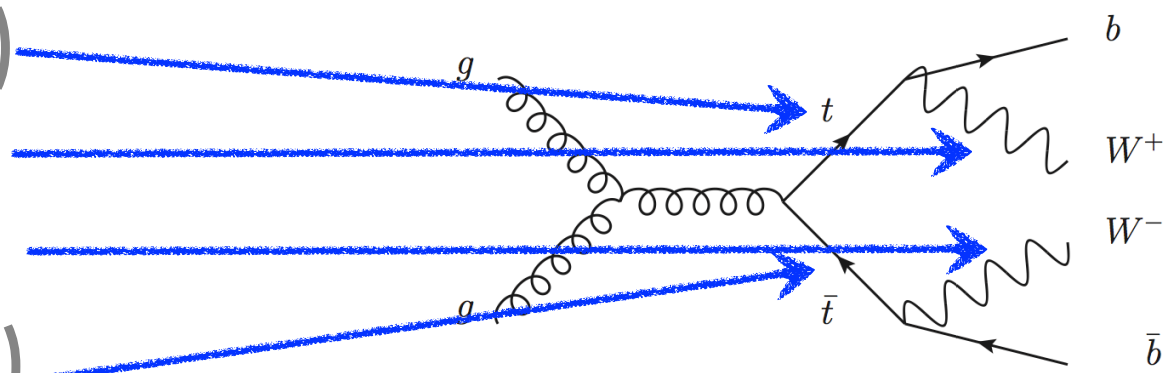


m(bjj)

m(jj)

m(lv)

m(blv)

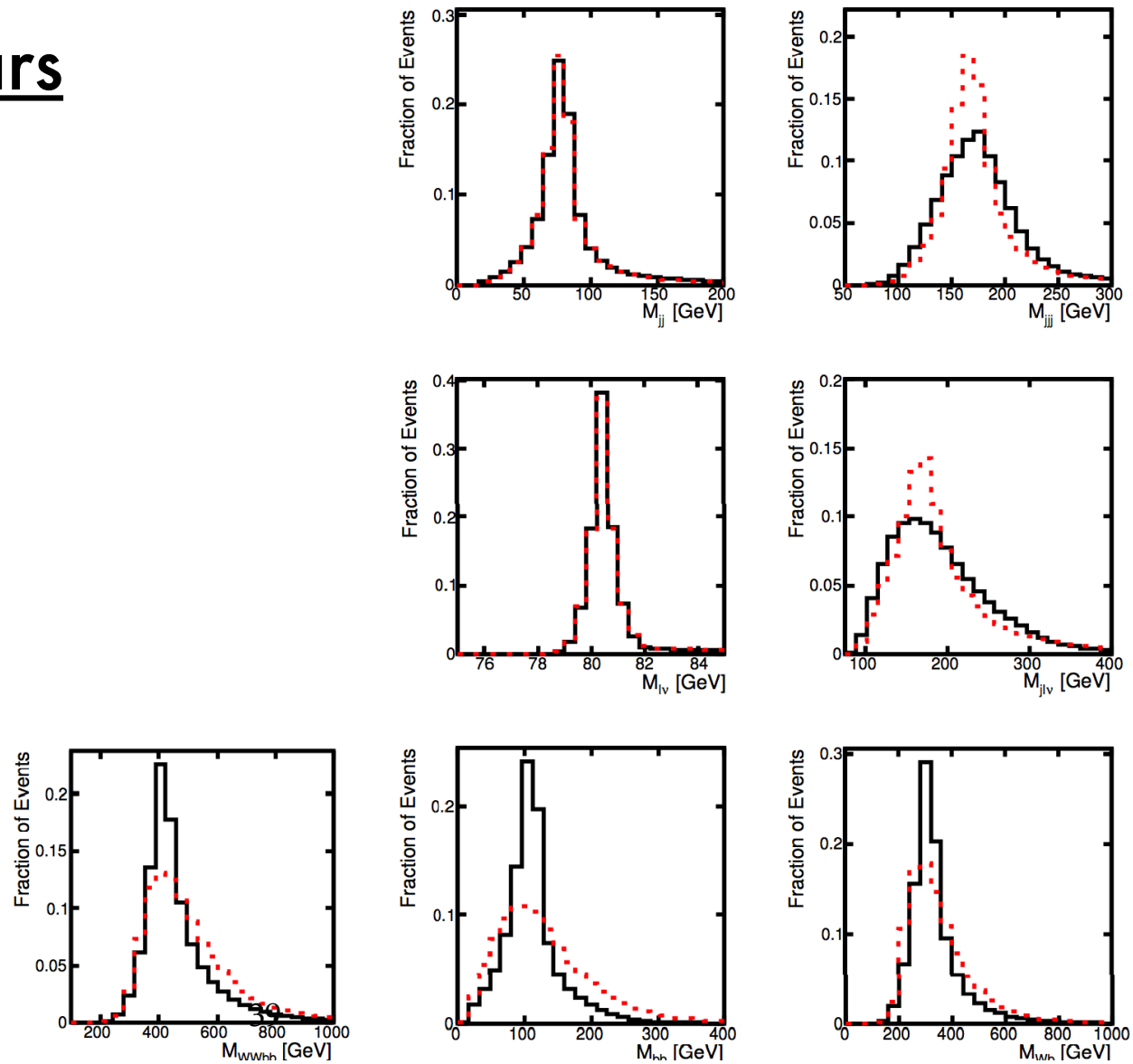


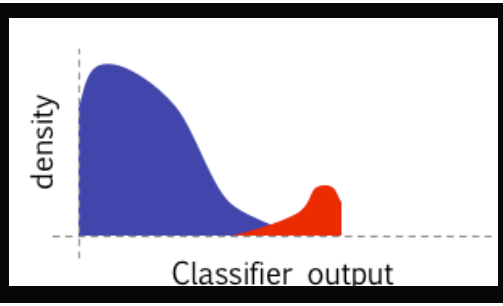
4-vector inputs

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

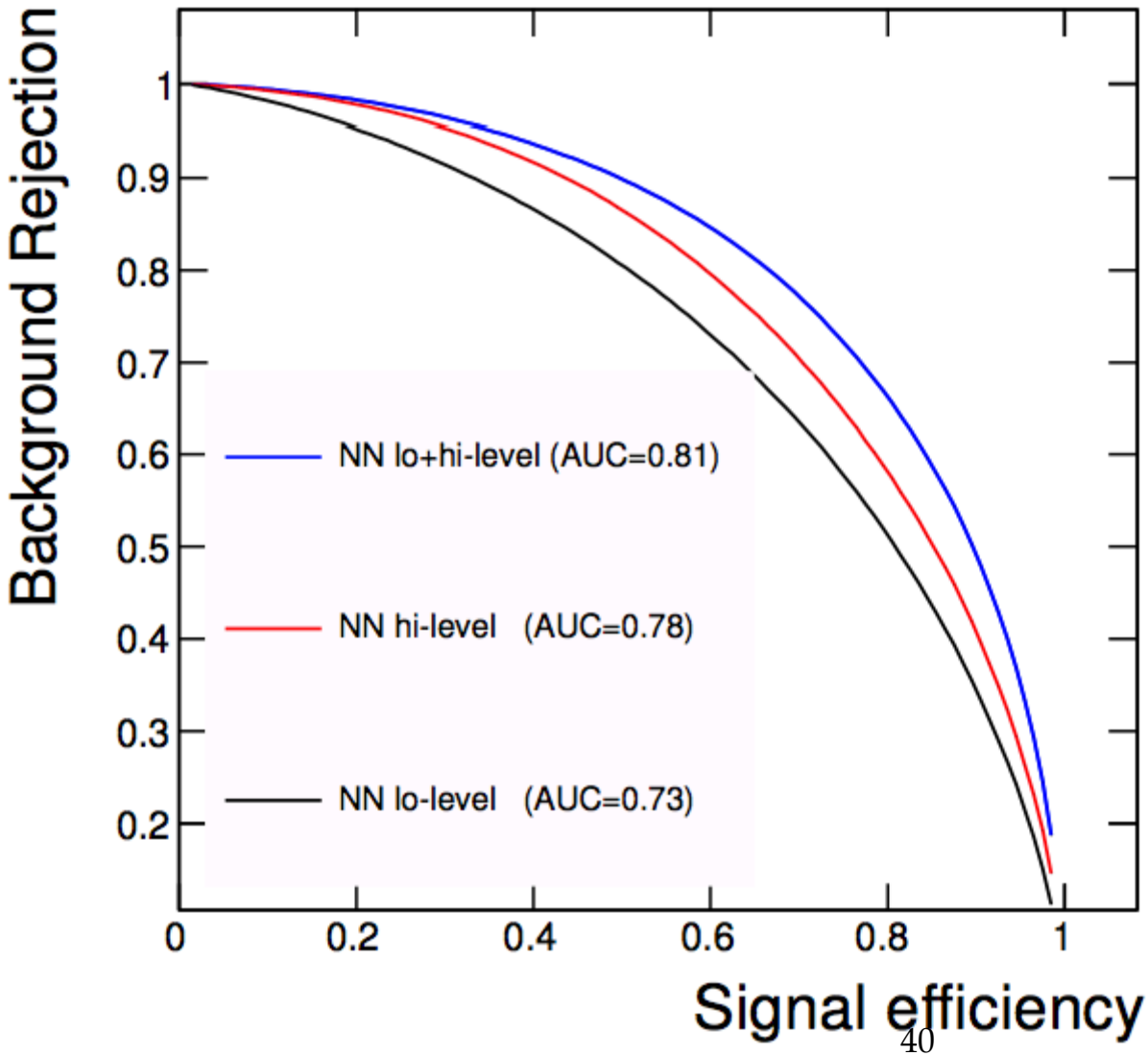
m(WWbb)
m(Wbb)
m(bb)

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





Standard NNs



Results

Adding hi-level boosts performance

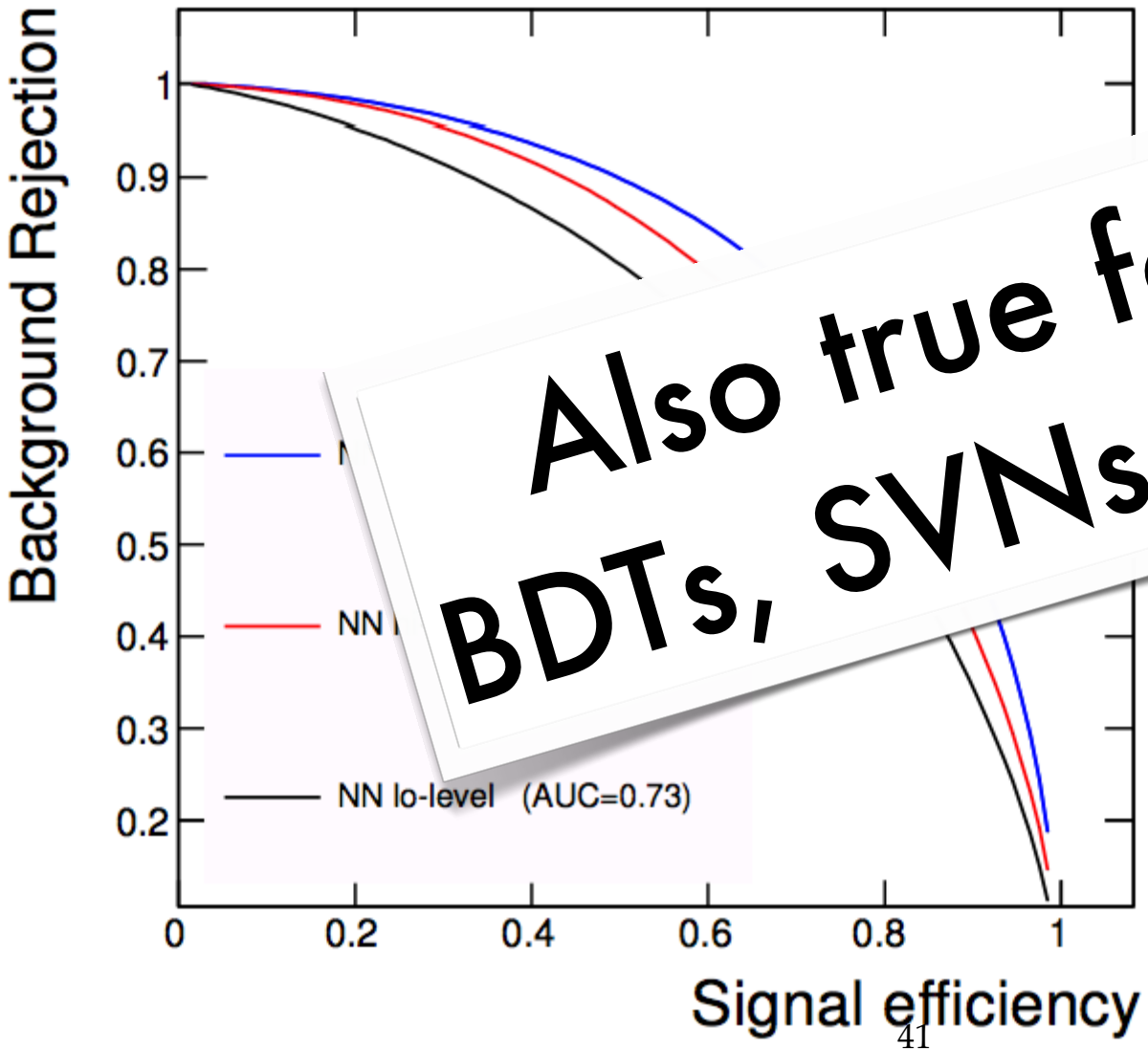
Better: lo+hi-level.

Conclude:

NN can't find hi-level vars.

Hi-level vars do not have all info

Standard NNs



Also true for BDTs, SVNs, etc

Results

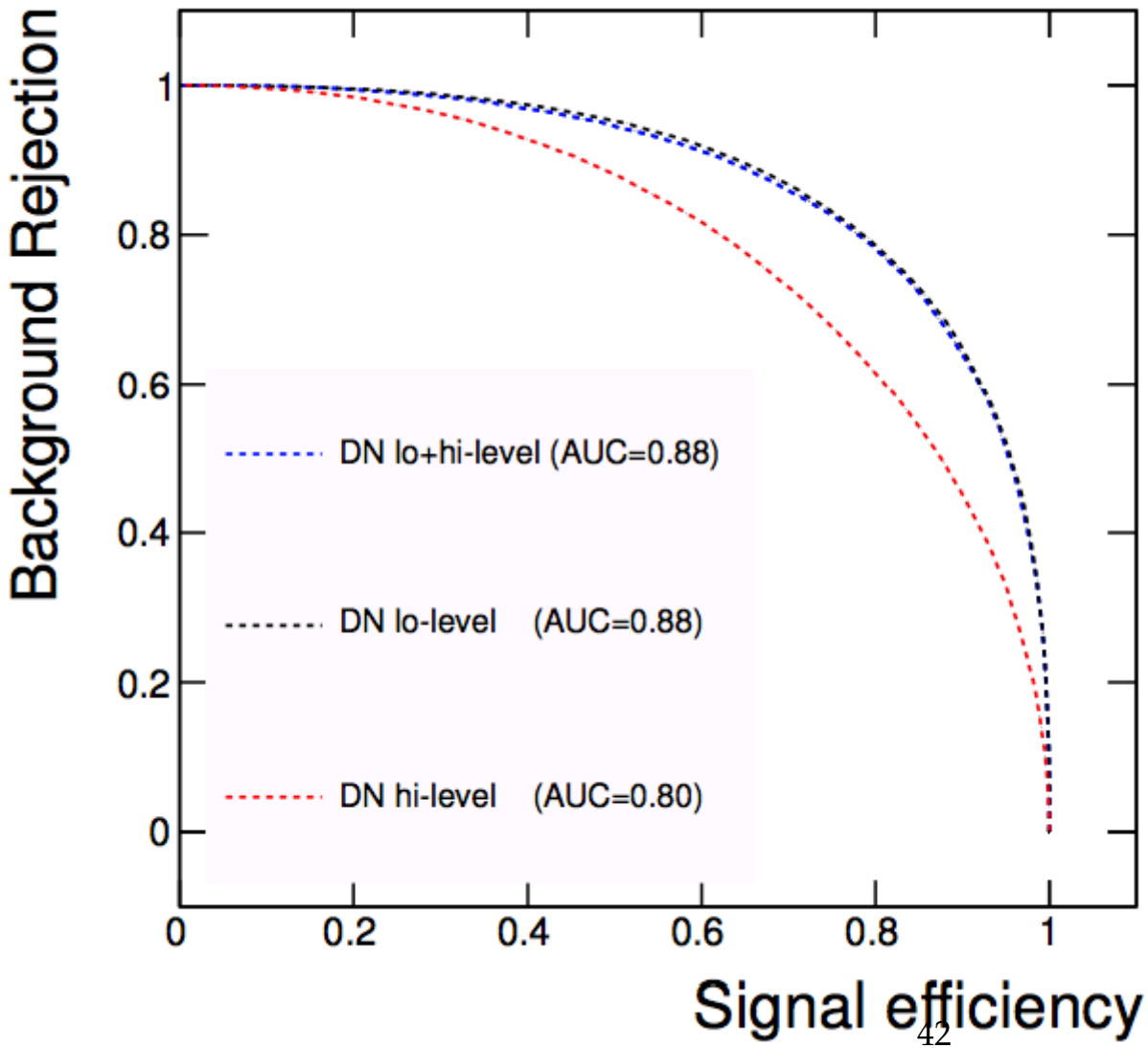
g hi-level
s performance
lo+hi-level.

nclude:

NN can't find
hi-level vars.

Hi-level vars
do not have all info

Deep Networks



Results

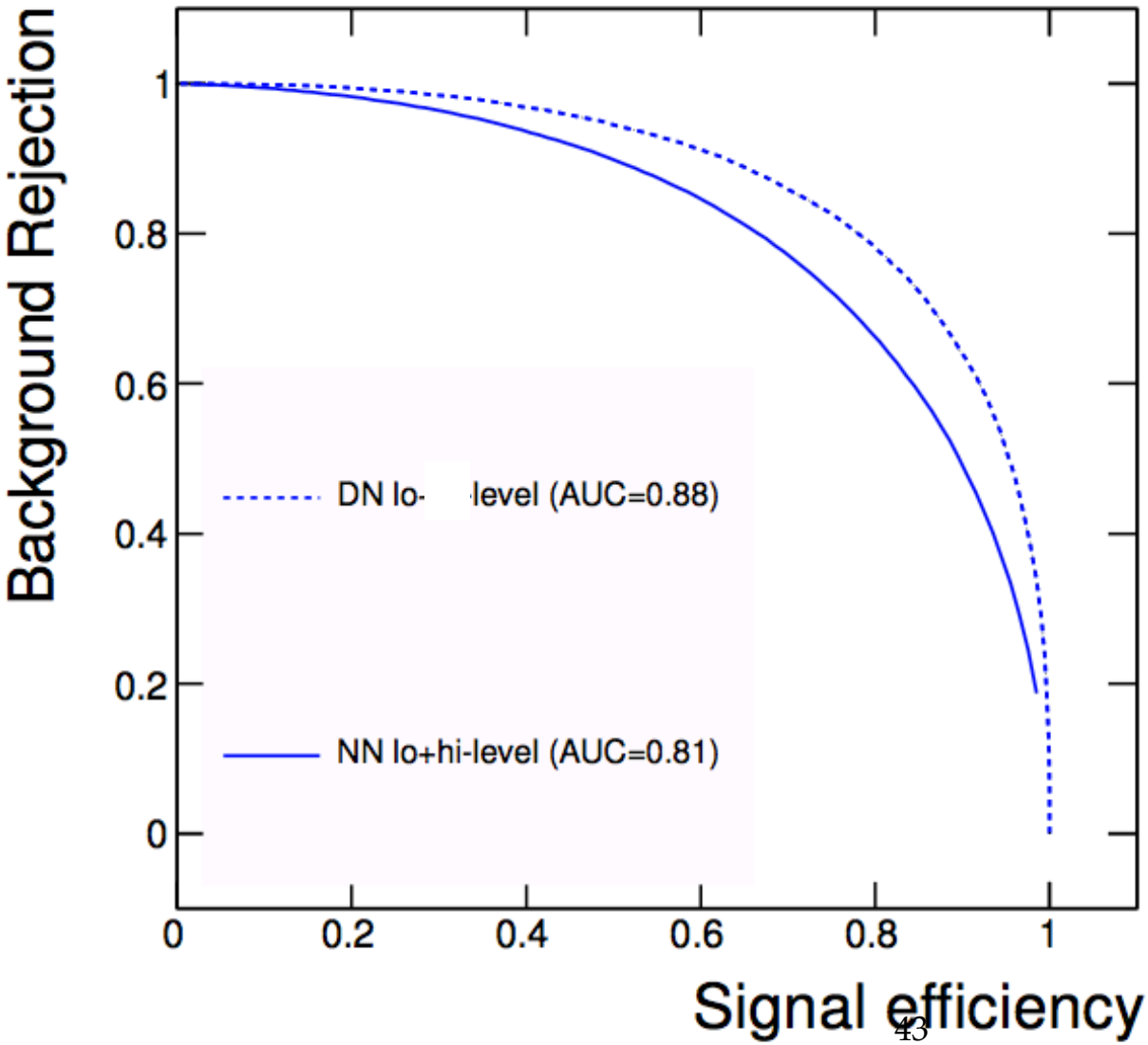
Lo+hi = lo.

Conclude:

DN can find
hi-level vars.

Hi-level vars
do not have all info
are unnecessary

Deep Networks



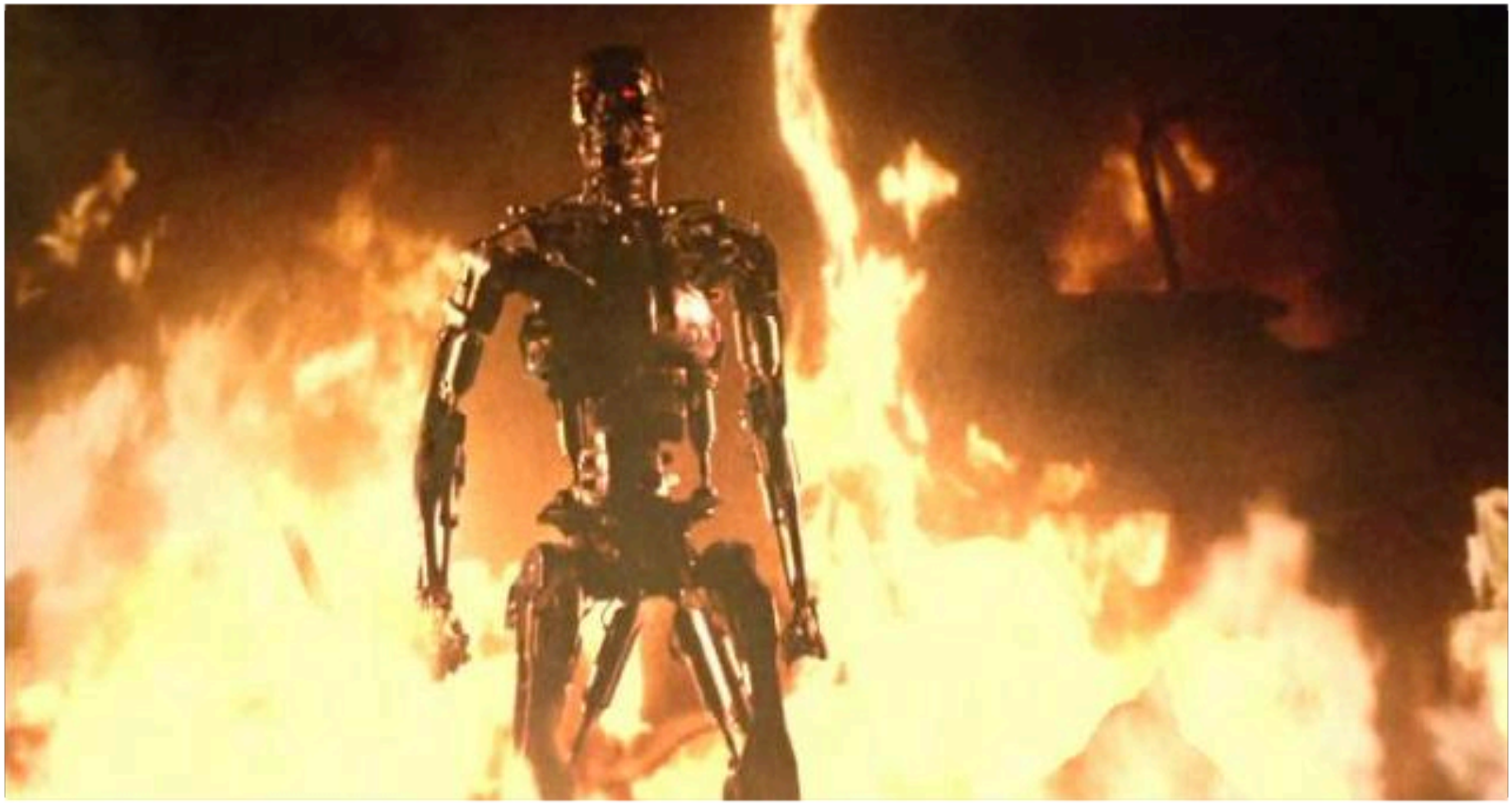
Results

DN > NN

Conclude:

DN does better than human assisted NN

The AIs win



Results

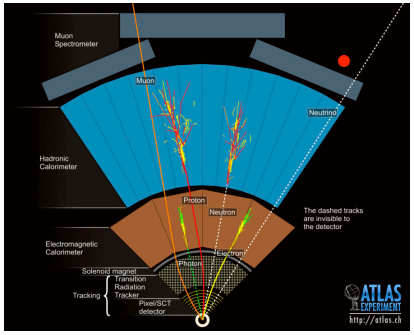
Identified example benchmark where traditional NNs fail to discover all discrimination power.

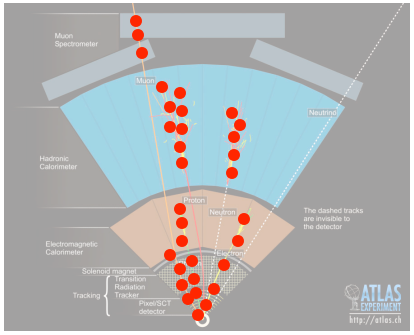
Adding human insight helps traditional NNs.

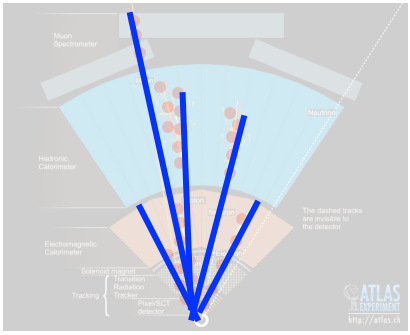
Deep networks succeed without human insight. Outperform human-boosted traditional NNs.

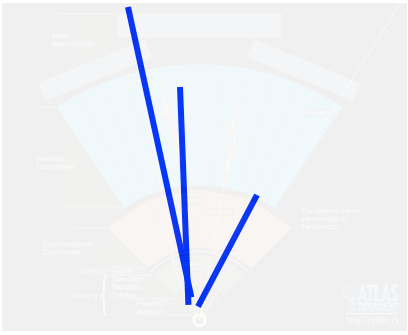
What is possible?

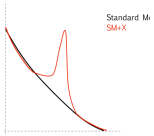
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1





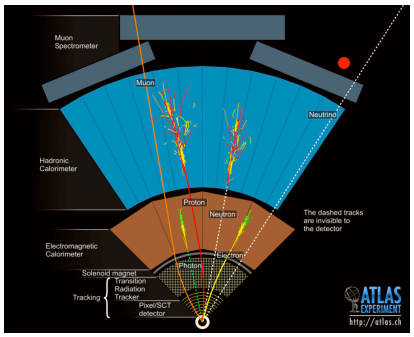


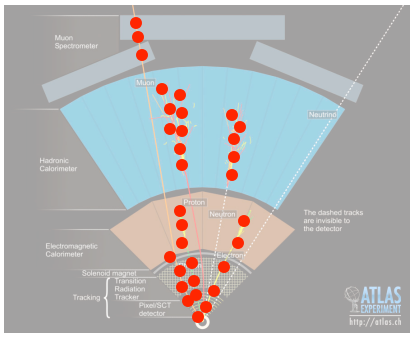


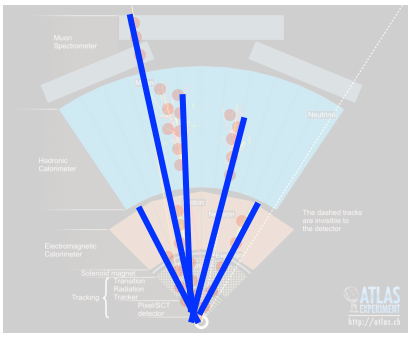


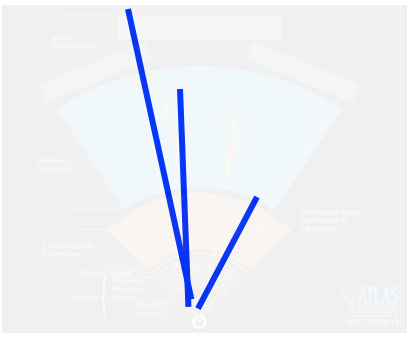
What is possible?

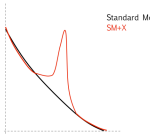
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1







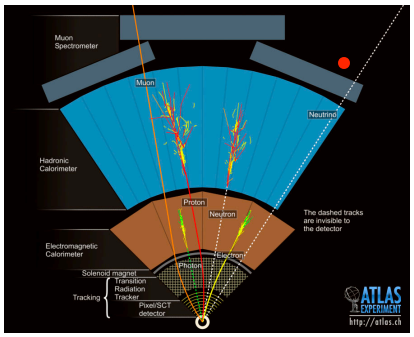


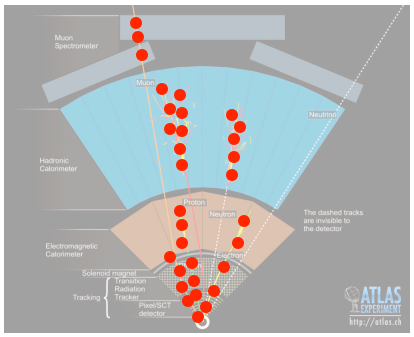


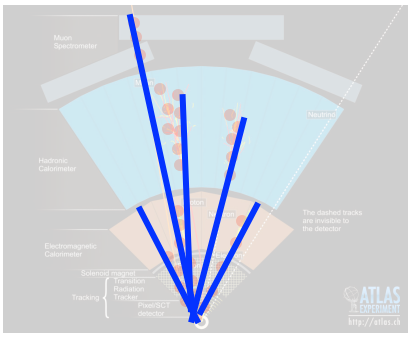
Skip more steps with ML?

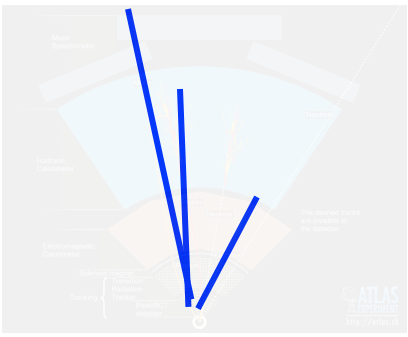
Or this?

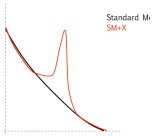
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1





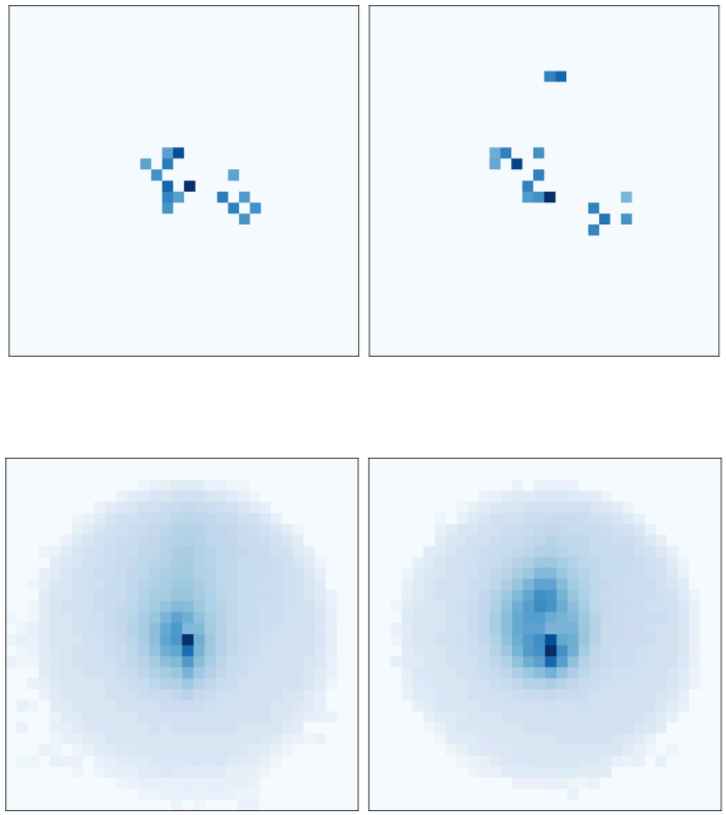


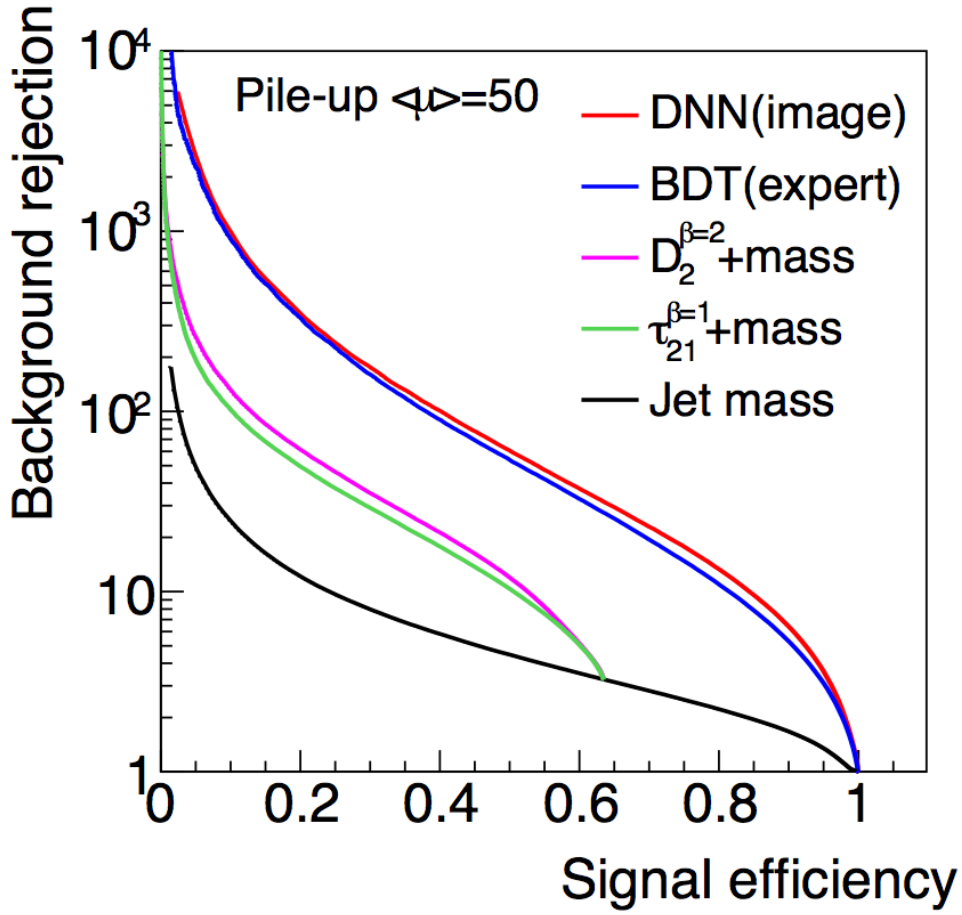




Improve each step with ML?

Jet tagging




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

Conclusions

<u>Deep Learning is a powerful new tool</u>

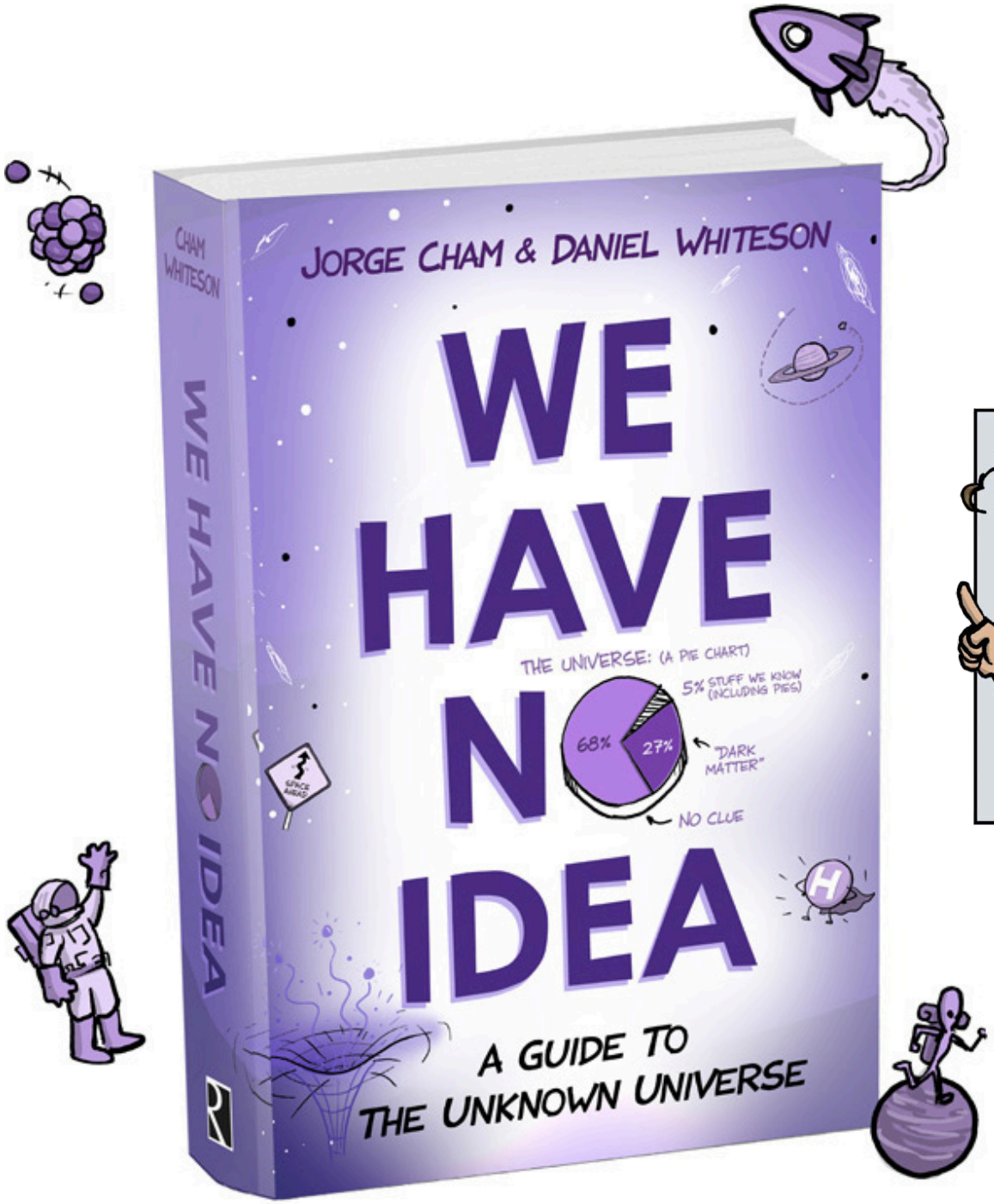
offers faster learning of nonlinear functions

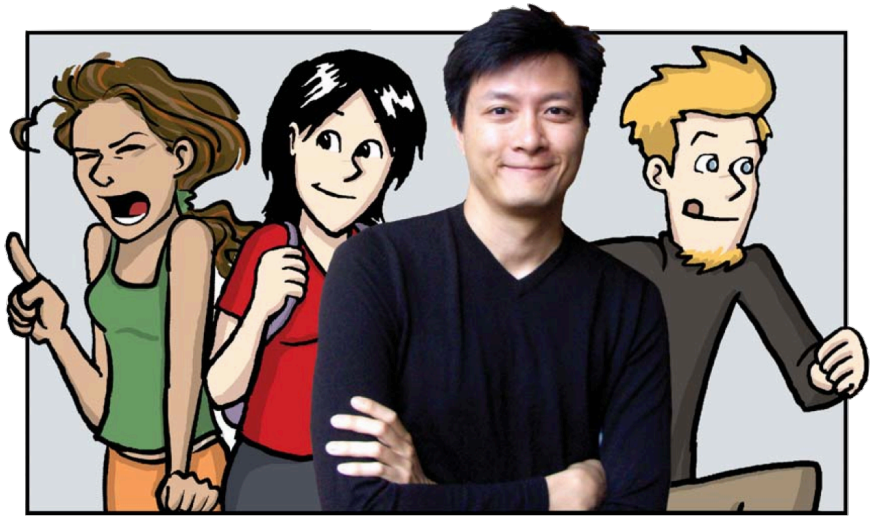
<u>We have many appropriate tasks in HEP</u>

traditional heuristics should be re-examined

<u>No replacement for human intelligence</u>

garbage in will still give garbage out


JORGE CHAM & DANIEL WHITESON
WE HAVE NO IDEA
THE UNIVERSE: (A PIE CHART)
5% STUFF WE KNOW (INCLUDING PIES)
68%
27%
"DARK MATTER"
NO CLUE
A GUIDE TO THE UNKNOWN UNIVERSE
CHAM WHITESON
WE HAVE NO IDEA



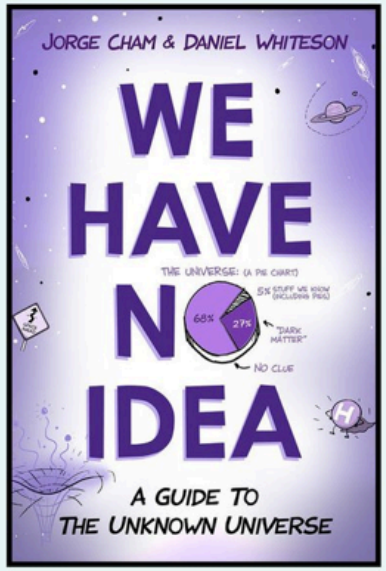
PhD COMICS -COMES TO PITT-

WE HAVE NO IDEA

THURSDAY, OCTOBER 26, 2017
UNIVERSITY CLUB, BALLROOM B
-8 P.M.-

ENJOY AN ENTERTAINING PRESENTATION THAT COMBINES SCIENCE, HUMOR, AND LIVE DRAWING INSPIRED BY THE NEW BOOK **WE HAVE NO IDEA** (JOHN MURRAY PUBLISHERS, 2017). PHD COMICS CREATOR **JORGE CHAM** AND PARTICLE PHYSICIST **DANIEL WHITESON** TEAM UP TO EXPLAIN EVERYTHING WE DON'T KNOW ABOUT THE UNIVERSE, FROM COSMIC RAYS AND DARK MATTER TO TIME TRAVEL AND THE BIG BANG THEORY.

- A BOOK SIGNING WILL FOLLOW THE LECTURE -
- LIGHT REFRESHMENTS WILL BE SERVED -



THIS LECTURE IS FREE AND
- OPEN TO THE PUBLIC -

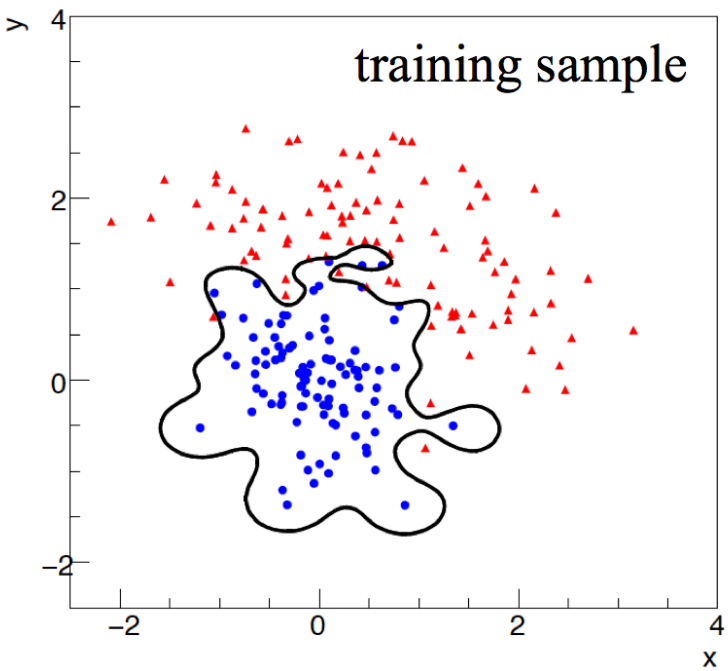
Sponsored by:
Department of Physics and Astronomy • The Humanities Center
The Pittsburgh Particle Physics, Astrophysics, and Cosmology Center

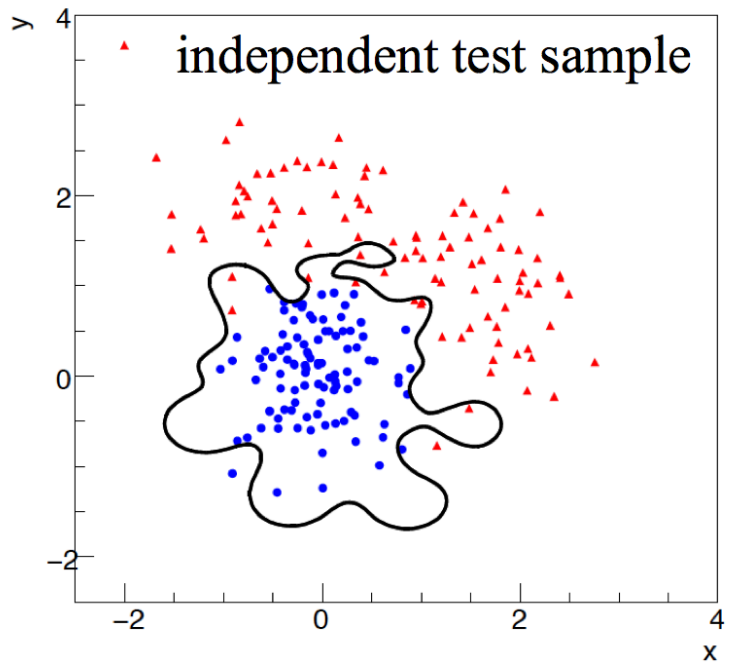
UNIVERSITY OF PITTSBURGH
The DIETRICH School of Arts & Sciences

[redacted]

How much to train?

A complex network, heavily trained will learn the statistical fluctuations of the training examples.





Avoiding overtraining