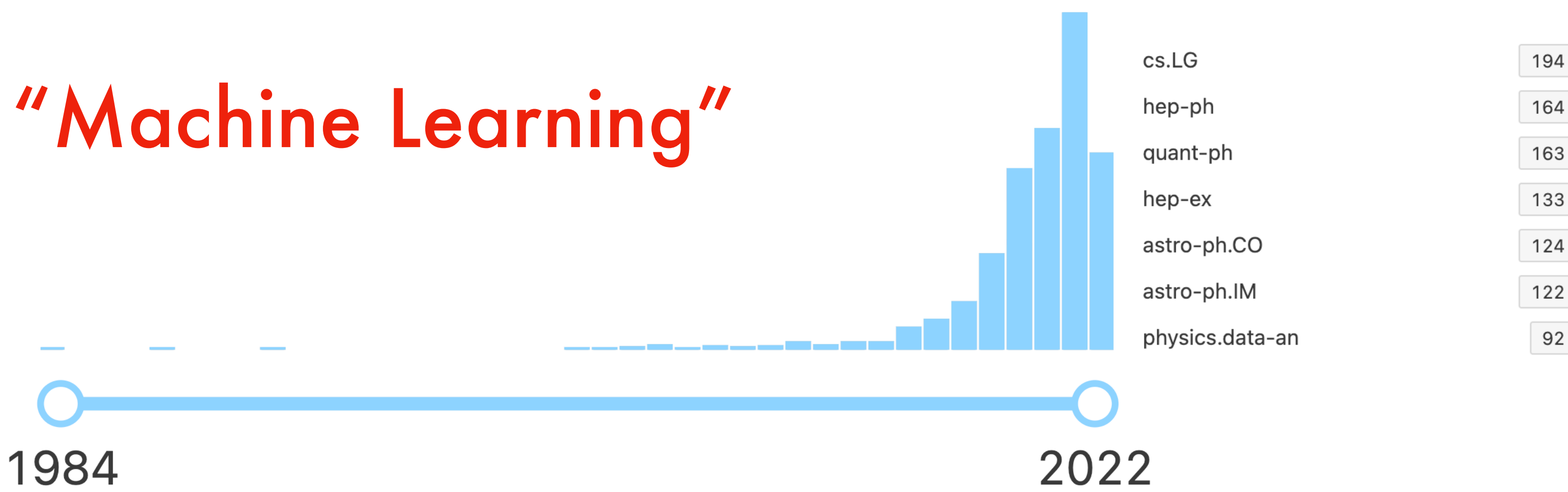
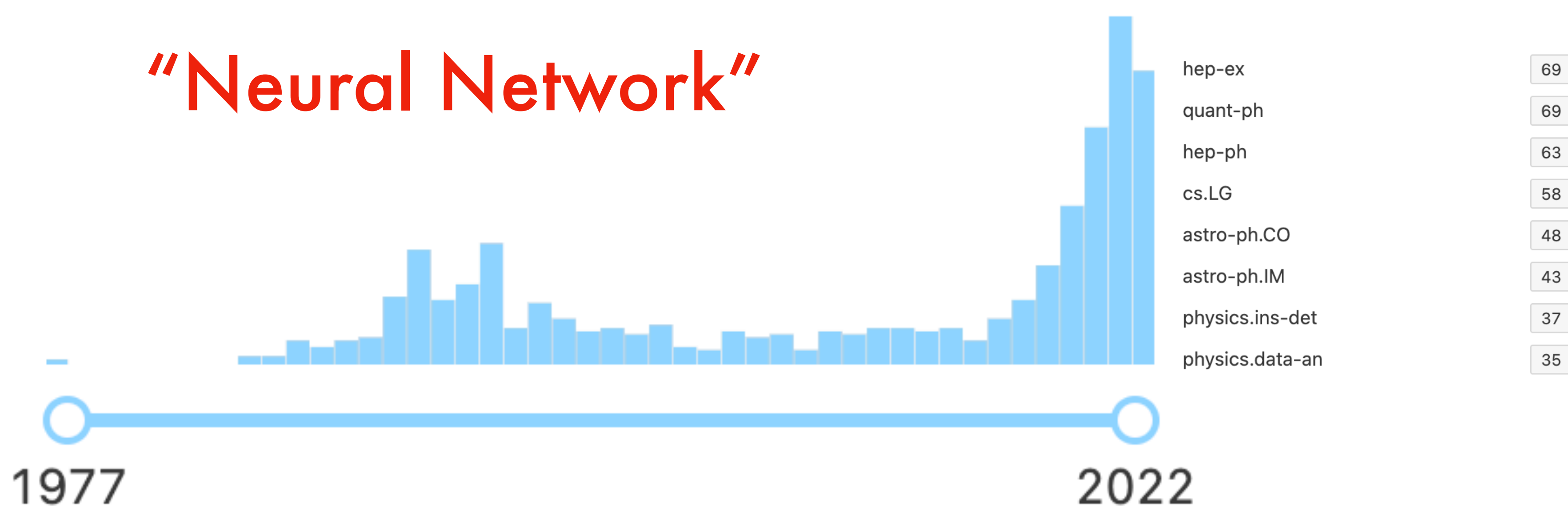
What's new about Machine Learning?

Daniel Whiteson, UC Irvine
Jul 2022 / Snowmass in Seattle

It's everywhere!





What’s new about ML?

Graphic from D. Guest

Early days of HEP

EARLY PHYSICISTS

ML in HEP is not new

1997:

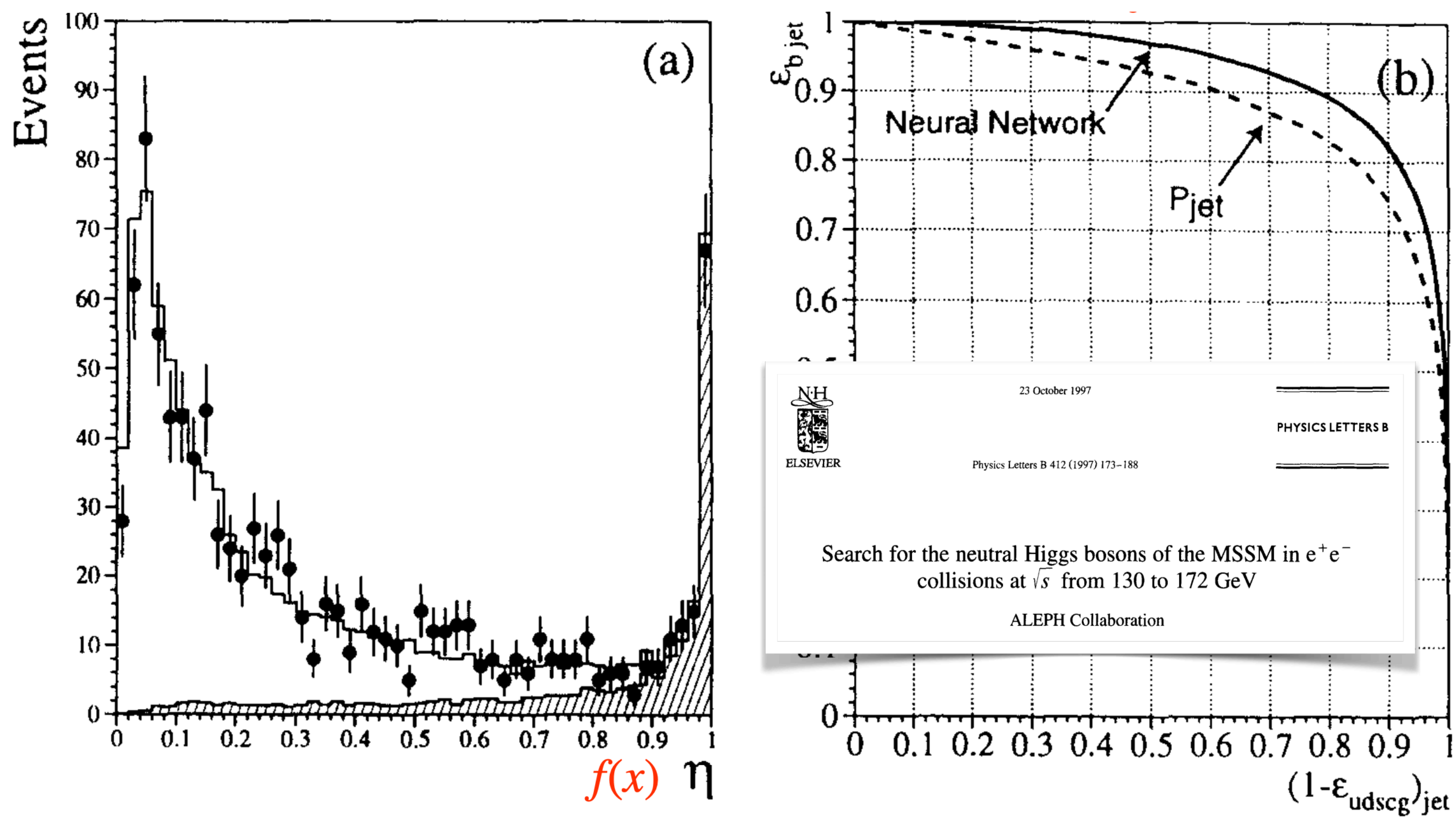


Fig. 2. (a) The output η of the neural network b tag for radiative returns to the Z for 161 GeV $q\bar{q}$ Monte Carlo (histogram) compared to the data at 161 GeV (points). The shaded region shows the contribution from generated b-jets. (b) The performance of the neural network b tag (solid line) for Monte Carlo events, presented in terms of the efficiency for identifying b-jets versus the efficiency for rejecting light quark jets. The performance of the single most powerful b tagging input variable to the neural network is shown for comparison (dashed curve).

Example from K. Cranmer

Is it something new?

Is modern ML something new,
or just more of the same?

Is it something new?

Is modern ML something new, or just more of the same?

Daniel Whiteson
@DanielWhiteson

Is recent (> ~2013 deep learning moment) ML in particle physics "more of the same" or "qualitatively something new".

Is it something new?

Is modern ML something new, or just more of the same?

Daniel Whiteson
@DanielWhiteson

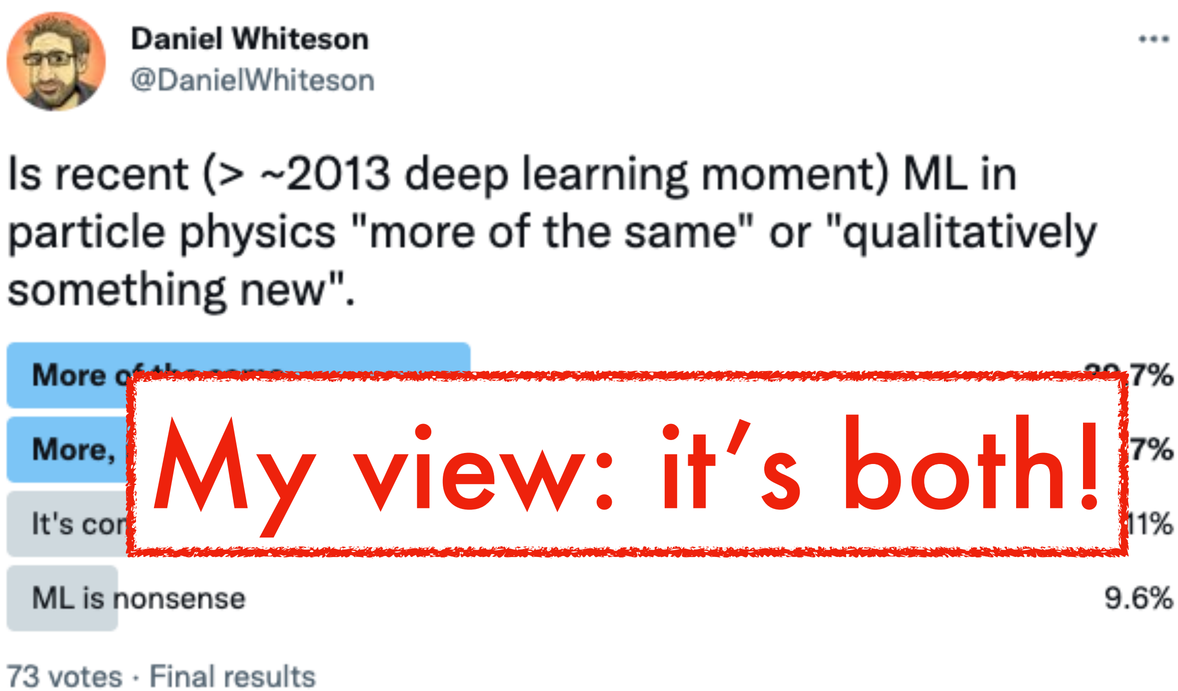
Is recent (> ~2013 deep learning moment) ML in particle physics "more of the same" or "qualitatively something new".

More of the same	**39.7%**
More, not the same	**39.7%**
It's complicated(comment)	11%
ML is nonsense	9.6%

73 votes · Final results

Is it something new?

Is modern ML something new,
or just more of the same?



My view: it's both!

Outline

1. Much much more of the same
2. Something qualitatively new

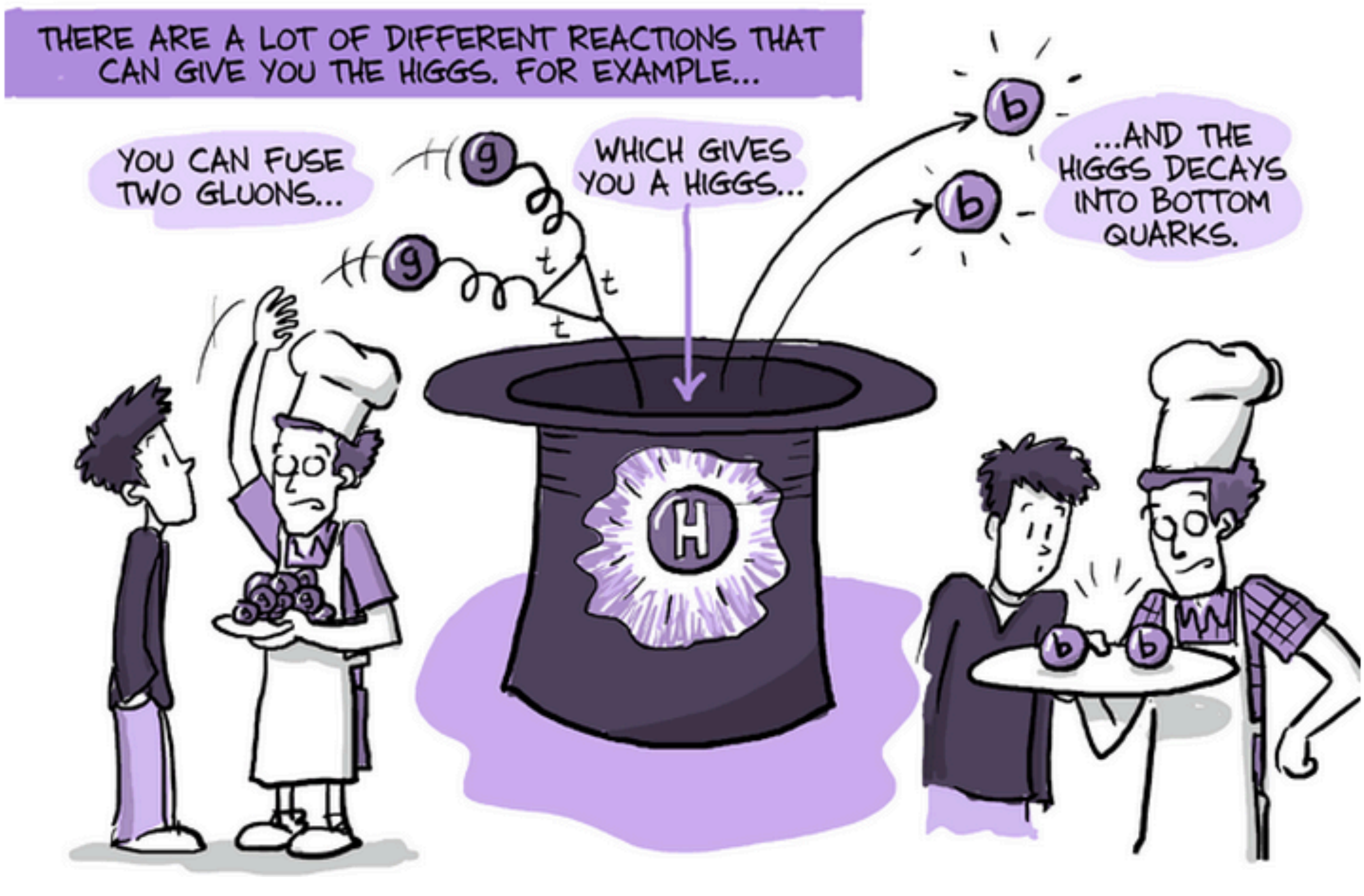
Traditional role of ML

Why do we need machine learning?

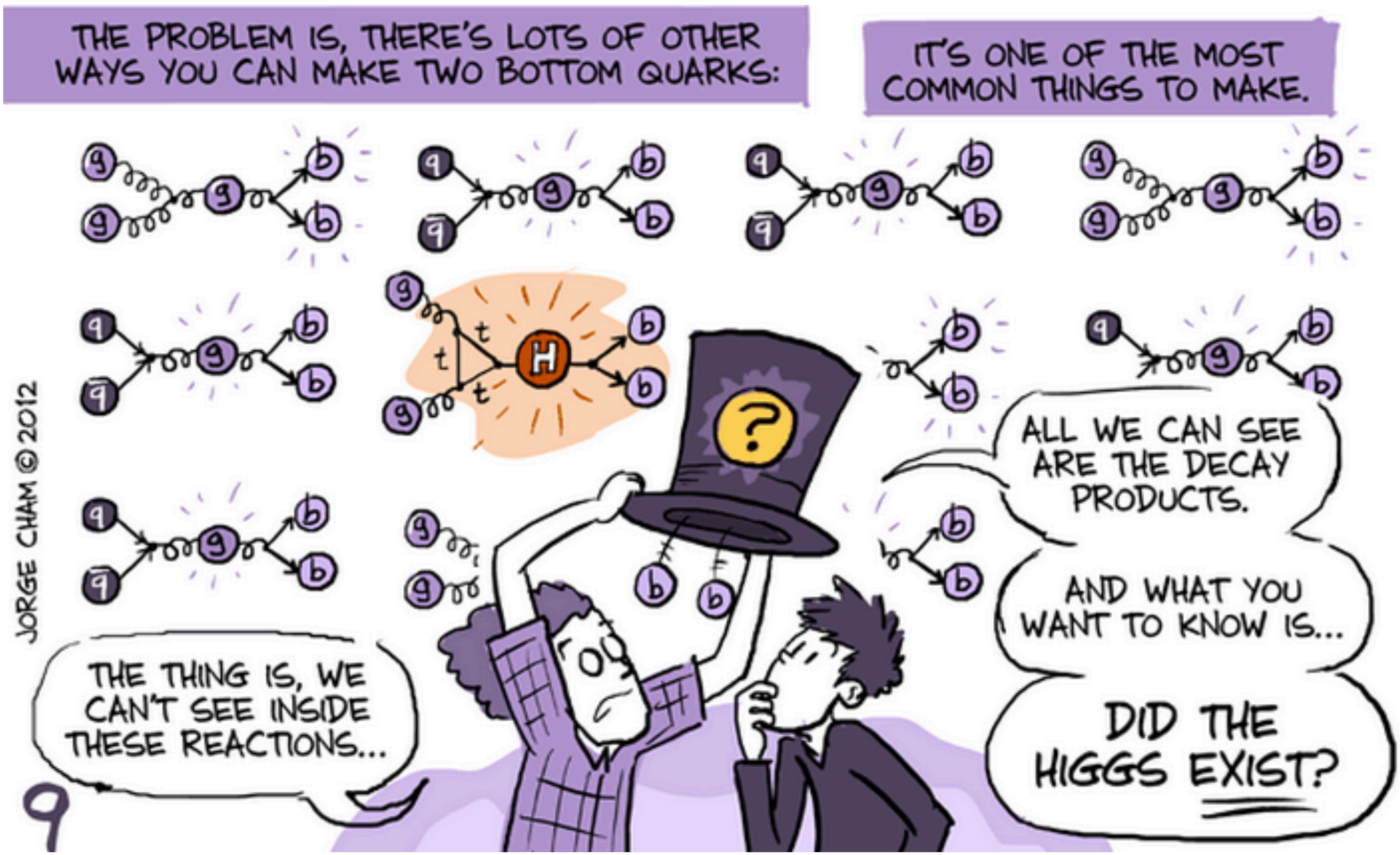
Traditional role of ML

Why do we need machine learning?

Making a new particle



Backgrounds



Neyman-Pearson

NP lemma says that the best statistic is the likelihood ratio:

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

data

theory

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

Functional space

No problem

If you can calculate:

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

For which you need:

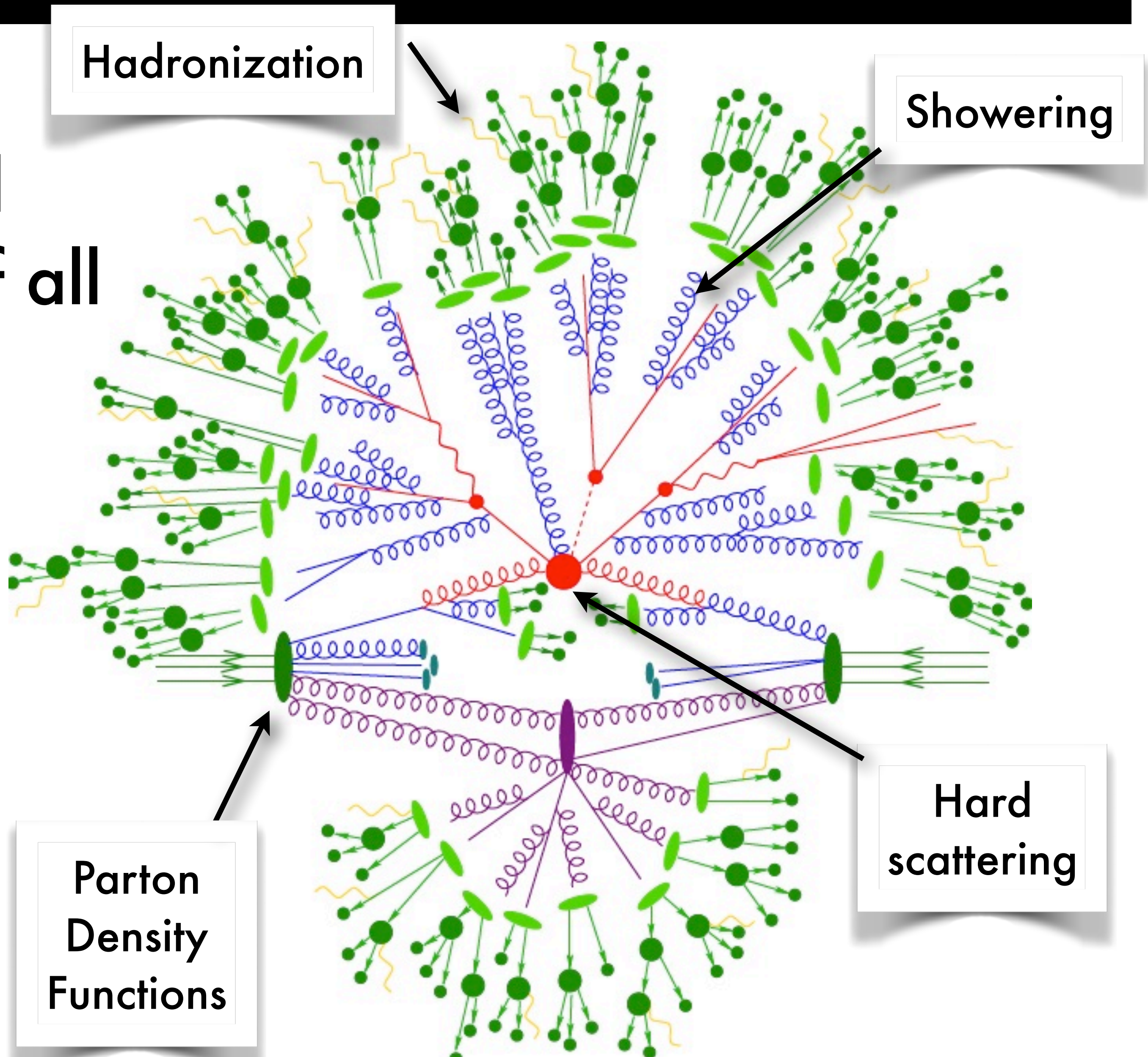
$P(data|theory)$

In general

We have a good understanding of all of the pieces

Do we have

P(*data*|*theory*)?



In general

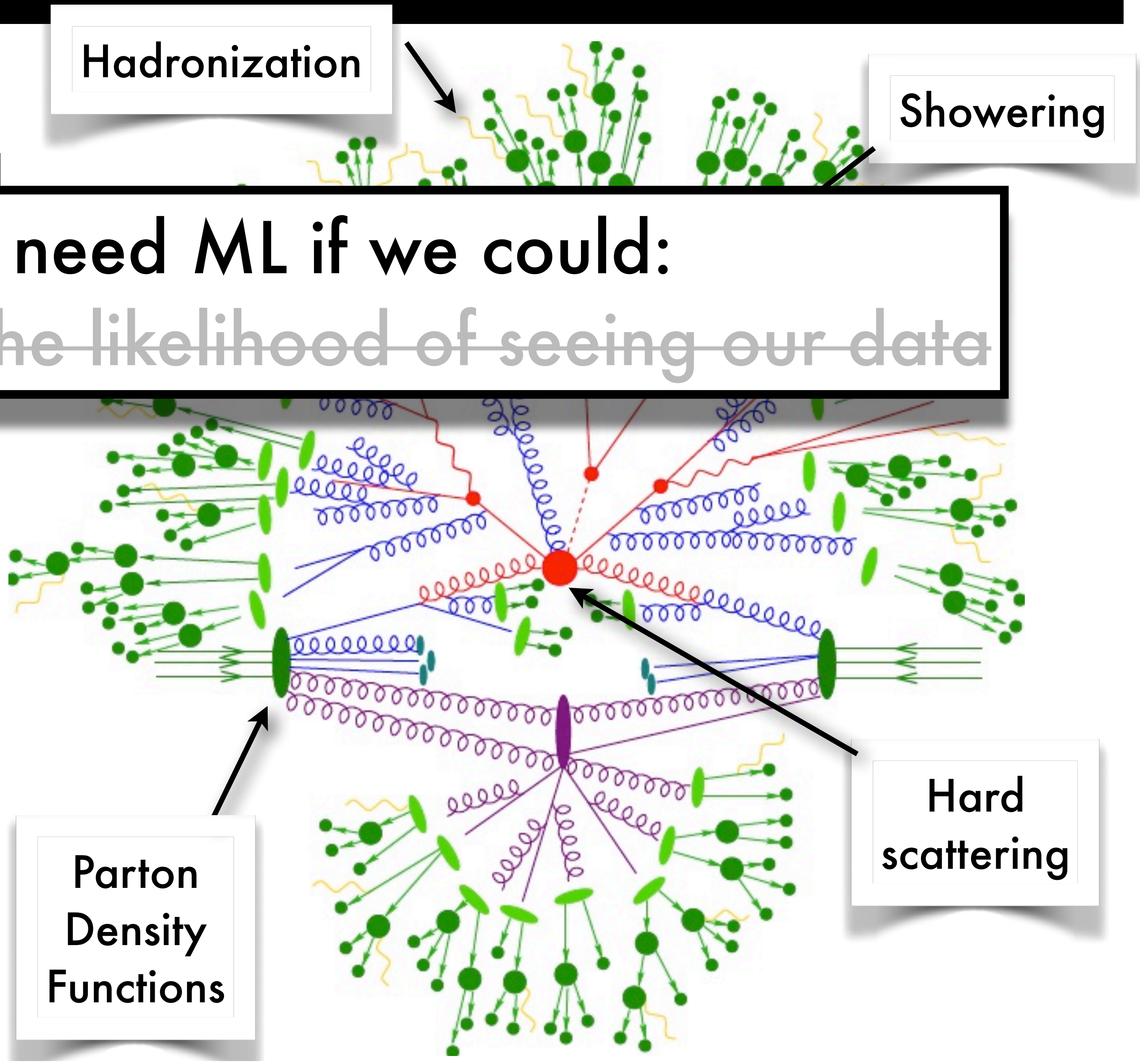
We have a good
und
of t

We wouldn't need ML if we could:

- ~~Express the likelihood of seeing our data~~

Do we have

P(*data* | *theory*)?

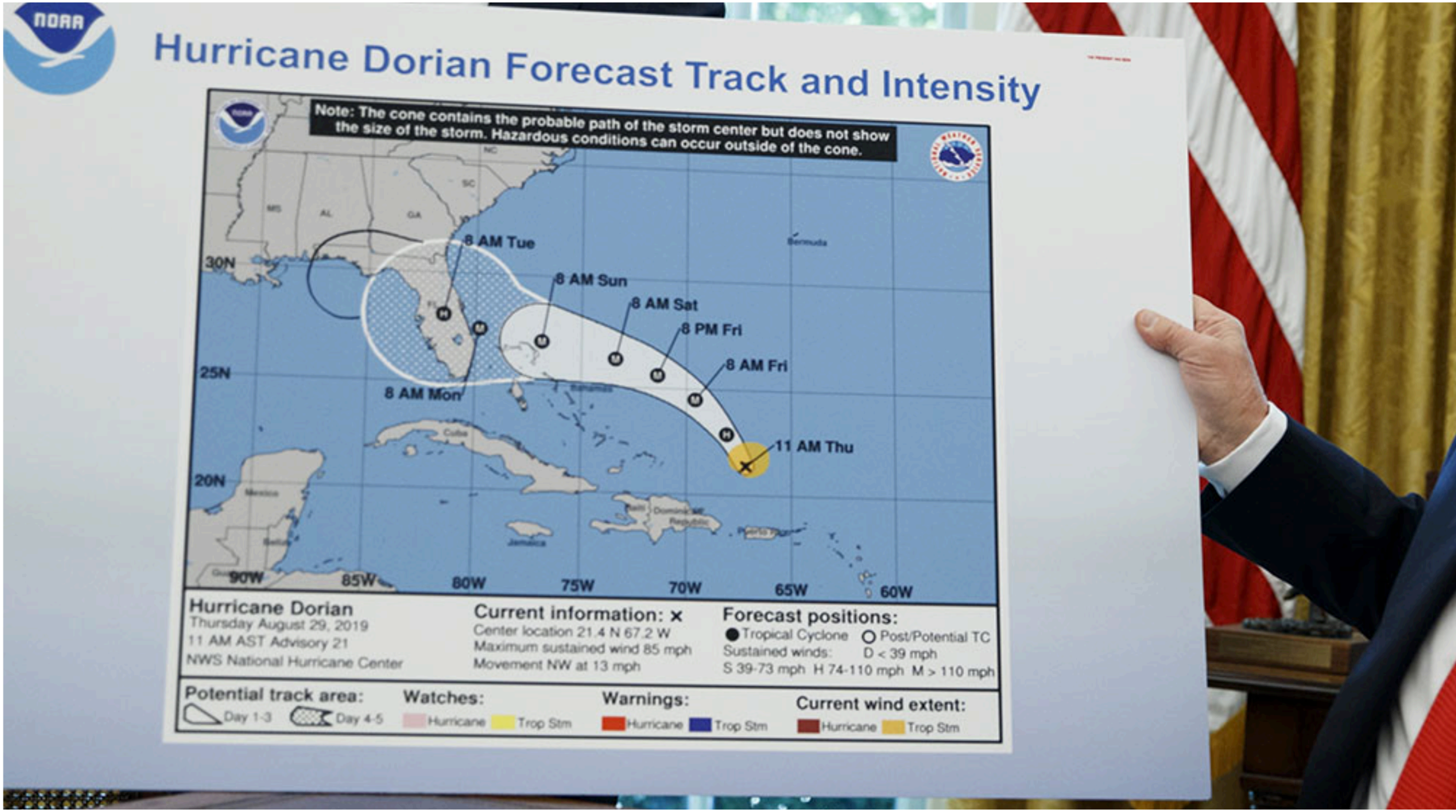


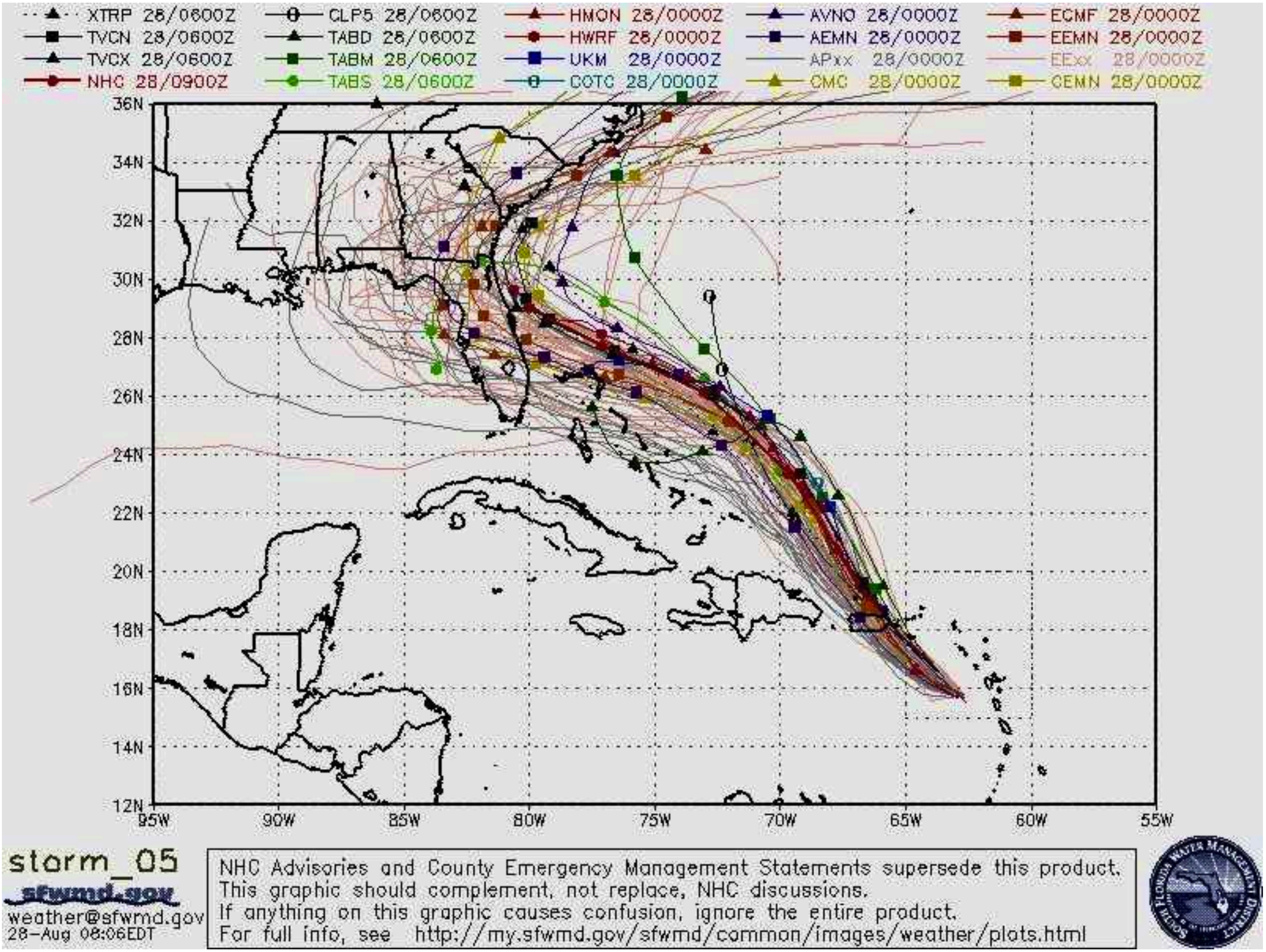
Darn

We can’t calculate

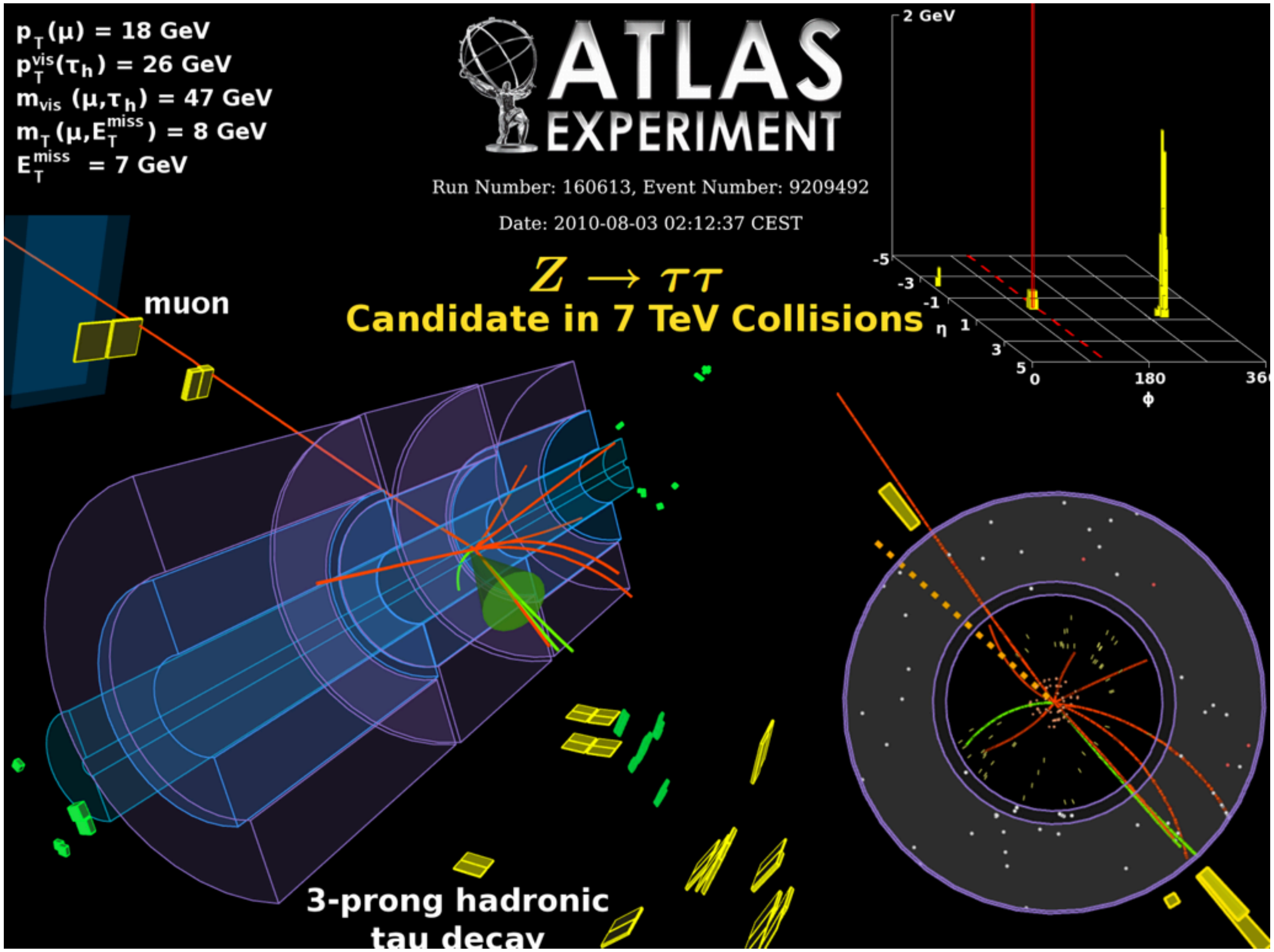
P(data|theory)

.... but we can simulate it!


NOAA
Hurricane Dorian Forecast Track and Intensity
Note: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.
8 AM Tue
8 AM Sun
8 AM Sat
8 PM Fri
8 AM Fri
8 AM Mon
11 AM Thu
30N
25N
20N
90W
85W
80W
75W
70W
65W
60W
Hurricane Dorian
Thursday August 29, 2019
11 AM AST Advisory 21
NWS National Hurricane Center
Current information: x
Center location 21.4 N 67.2 W
Maximum sustained wind 85 mph
Movement NW at 13 mph
Forecast positions:
Tropical Cyclone
Post/Potential TC
Sustained winds:
D < 39 mph
S 39-73 mph
H 74-110 mph
M > 110 mph
Potential track area:
Day 1-3
Day 4-5
Watches:
Hurricane
Trop Stm
Warnings:
Hurricane
Trop Stm
Current wind extent:
Hurricane
Trop Stm

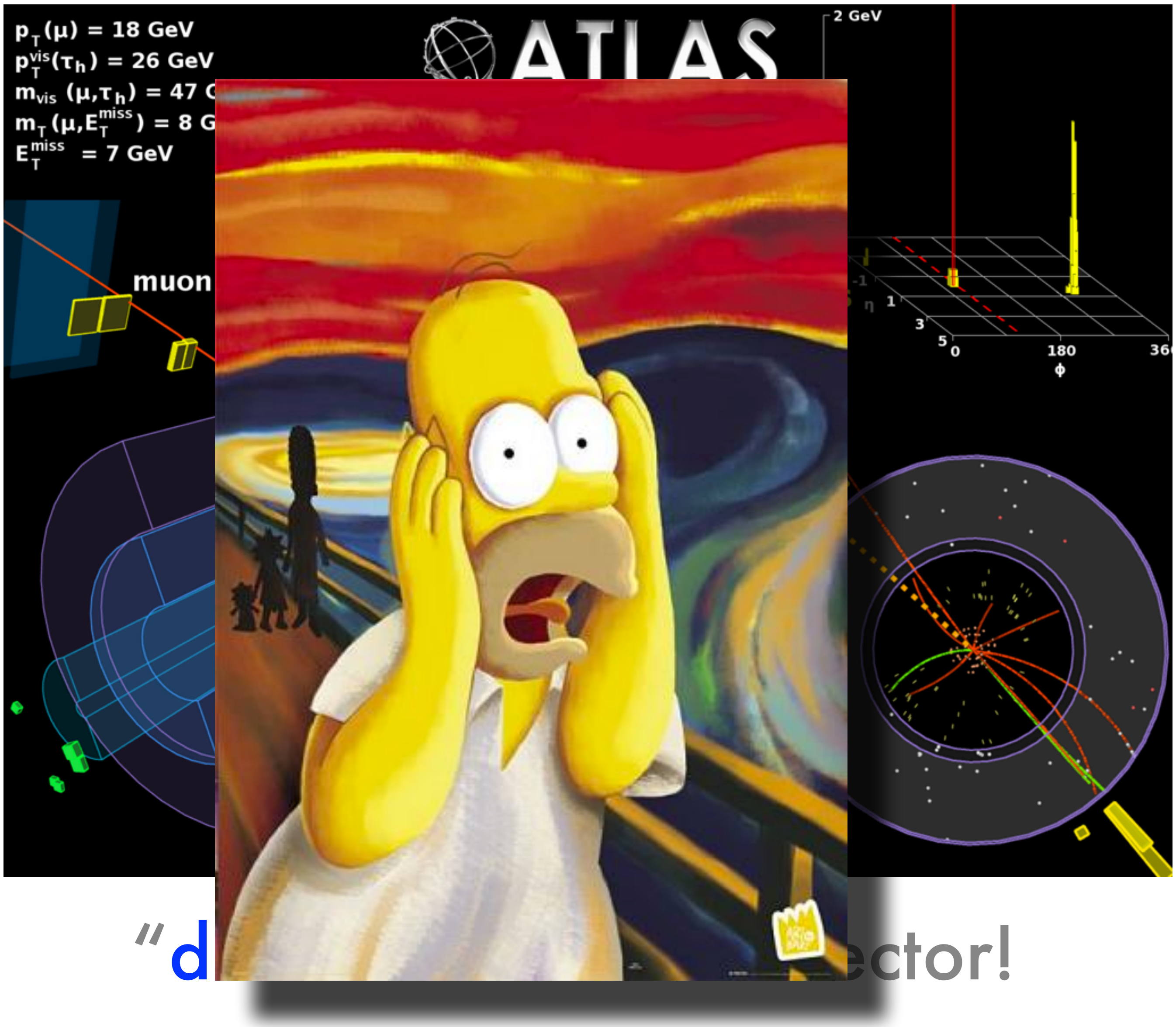

XTRP 28/0600Z
TVCN 28/0600Z
TVCX 28/0600Z
NHC 28/0900Z
CLP5 28/0600Z
TABD 28/0600Z
TABM 28/0600Z
TABS 28/0600Z
HMON 28/0000Z
HWRF 28/0000Z
UKM 28/0000Z
COTC 28/0000Z
AVNO 28/0000Z
AEMN 28/0000Z
APxx 28/0000Z
CMC 28/0000Z
ECMF 28/0000Z
EEMN 28/0000Z
EExx 28/0000Z
CEMN 28/0000Z
36N
34N
32N
30N
28N
26N
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 nightmare



"data" is a 100M-d vector!

The nightmare



"d ... ector!

The nightmare

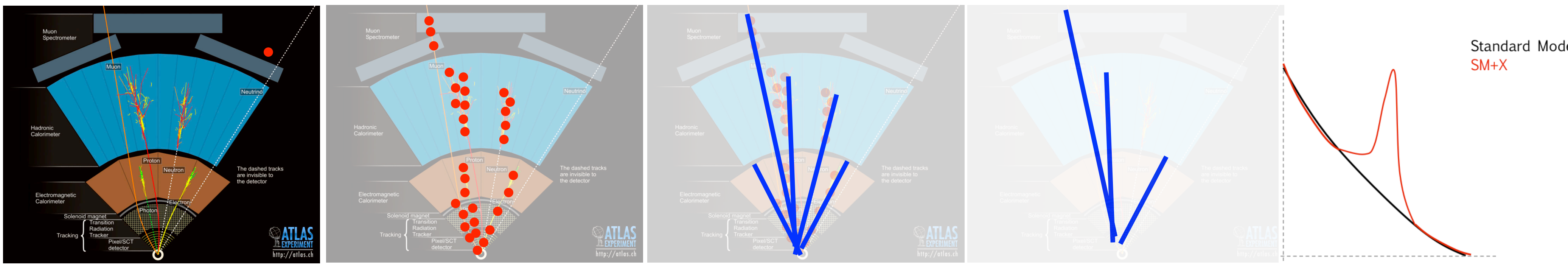
We wouldn't need ML if we could:

- ~~Express the likelihood of seeing our data~~
- ~~Access infinite computing resources~~
- ~~Develop infinitely fast simulation~~

"d ... ector!

Summary statistics

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

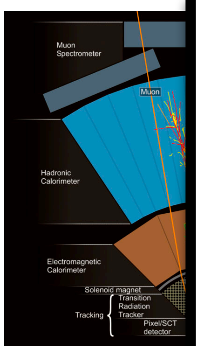


We don't need to analyze the raw data

...If we could summarize it perfectly

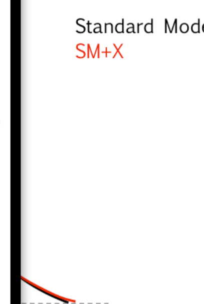
Summary statistics

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



We wouldn't need ML if we could:

- ~~Express the likelihood of seeing our data~~
- ~~Access infinite computing resources~~
- ~~Develop infinitely-fast simulation~~
- ~~Derive perfect summary statistics~~



...If we could summarize it perfectly

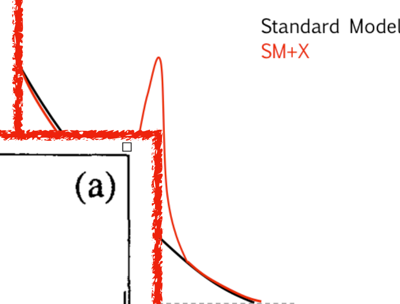
Summary statistics

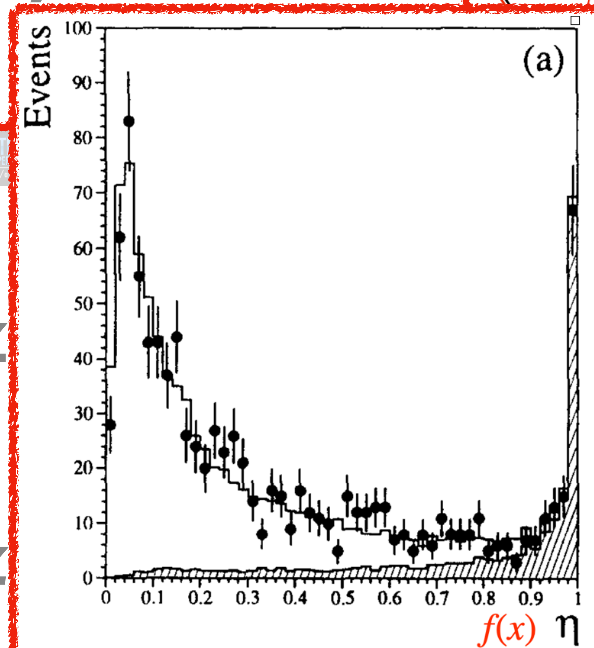
Raw	Sparsified	Reco	Select	Ana
1e7	1e3	100	50	1

Traditional ML (< 2012)

Mostly allowed for >1 analysis feature

Combine a few features





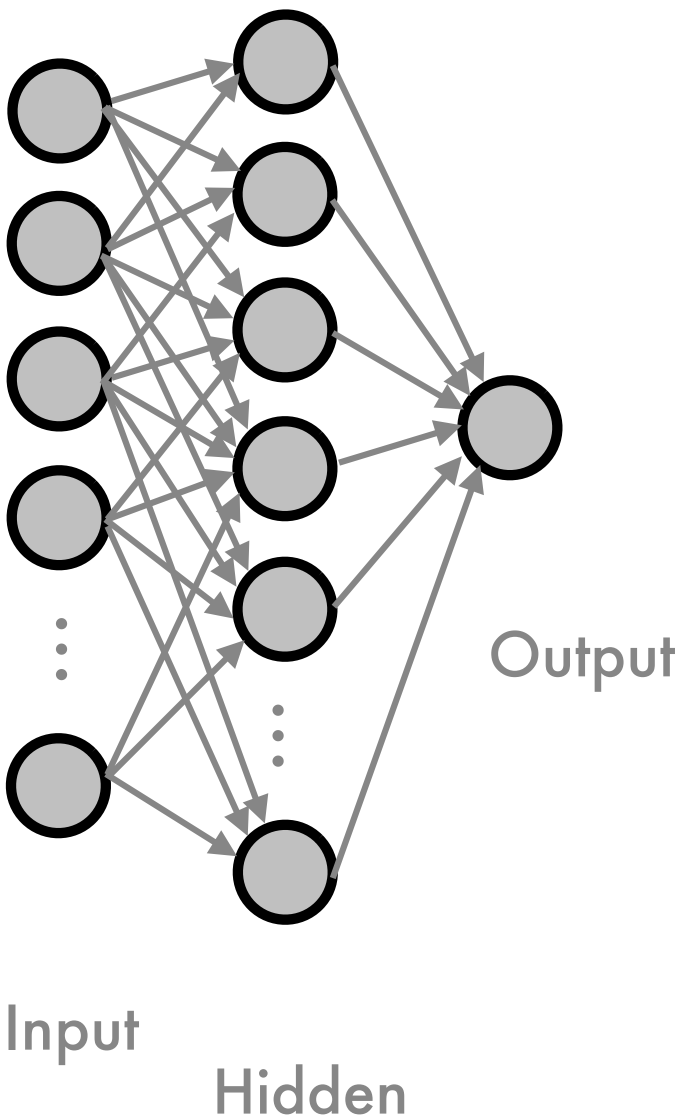
We don't need to analyz

...If we could summariz

Functional space

How complex?

Essentially a functional fit with many parameters

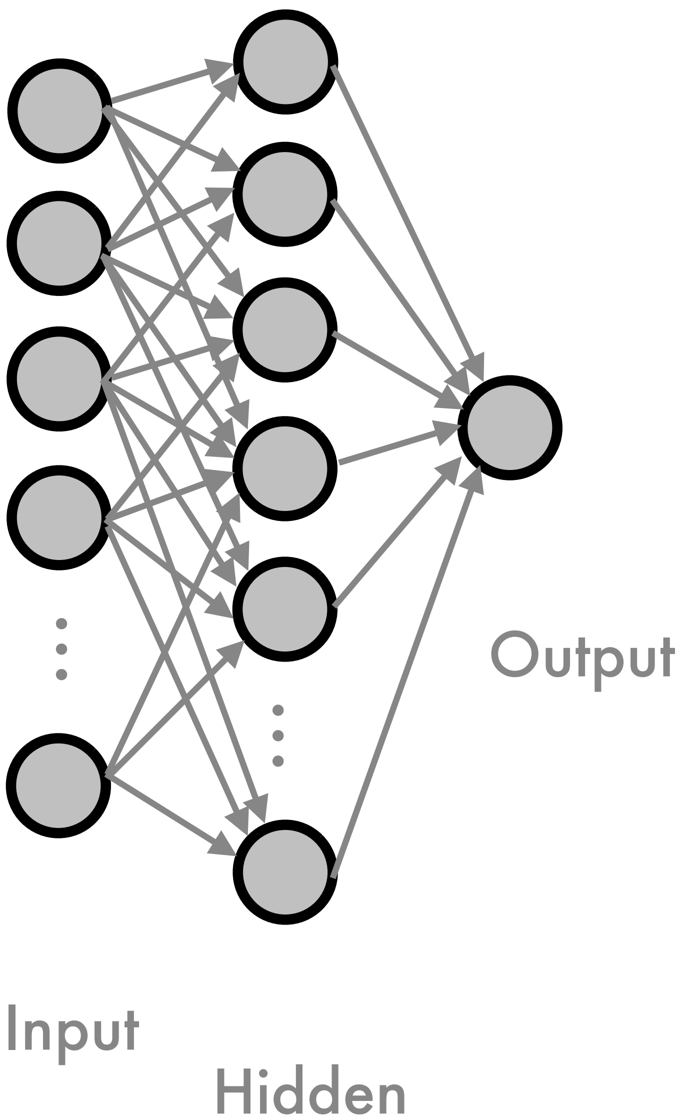


<u>Single hidden layer</u>

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

How complex?

Essentially a functional fit with many parameters

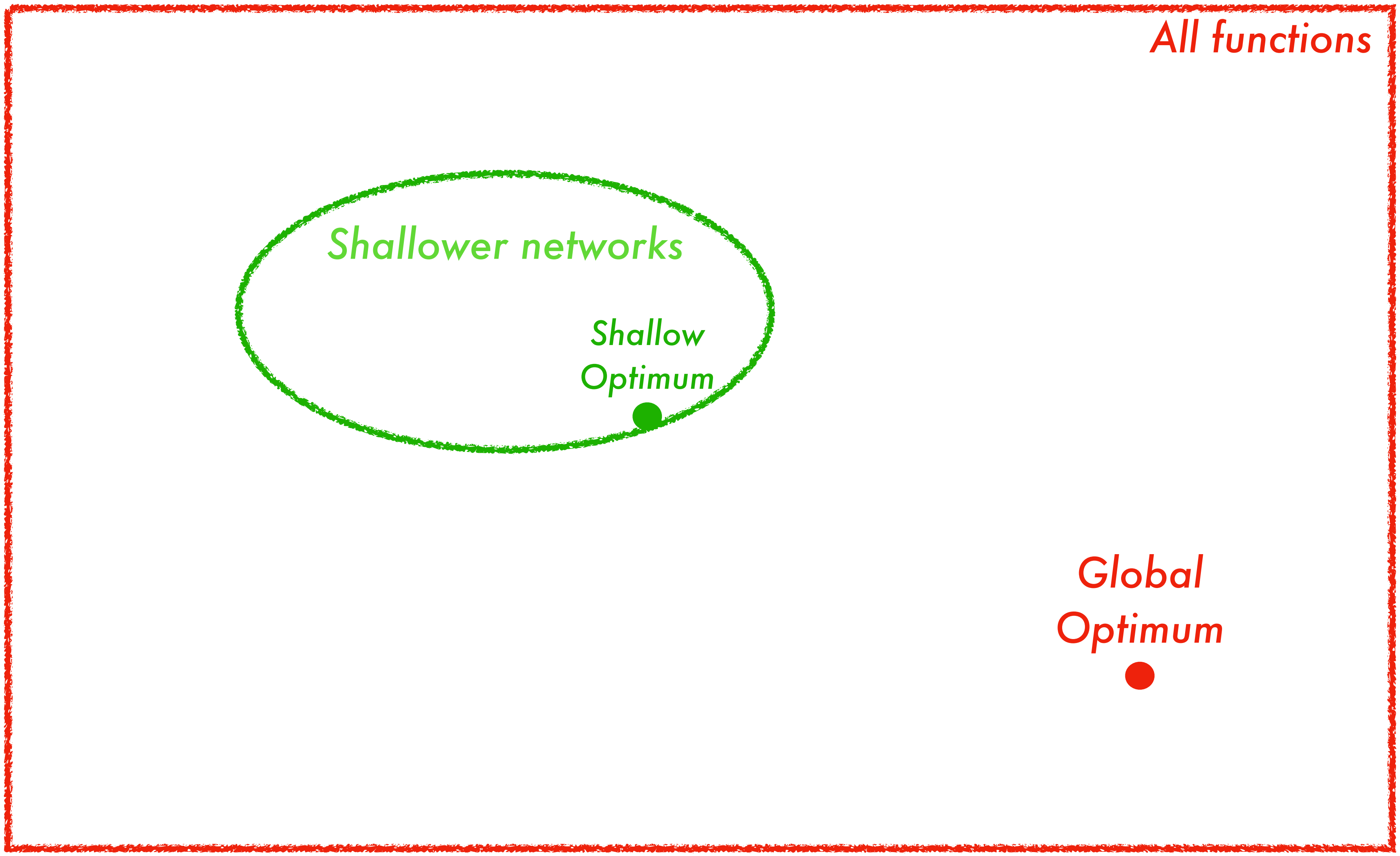


Single hidden layer

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

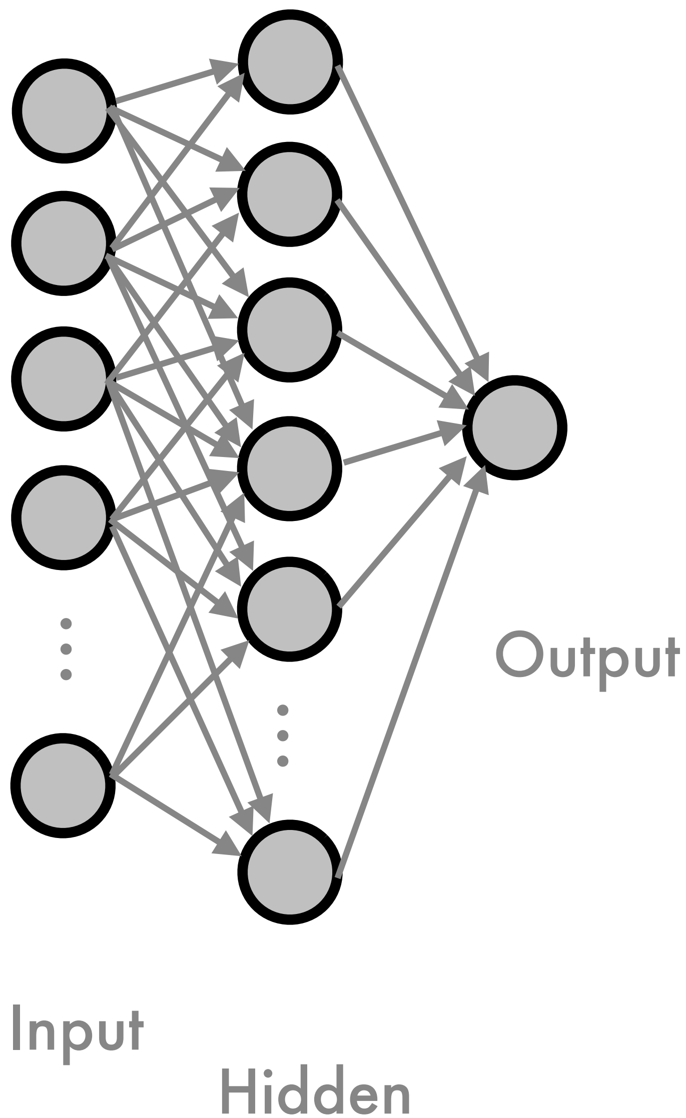
But might require very large hidden layer

Shallow space



Neural Networks

Essentially a functional fit with many parameters



Consequence:

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

In short:

Couldn't just throw data at NN.

Search for Input

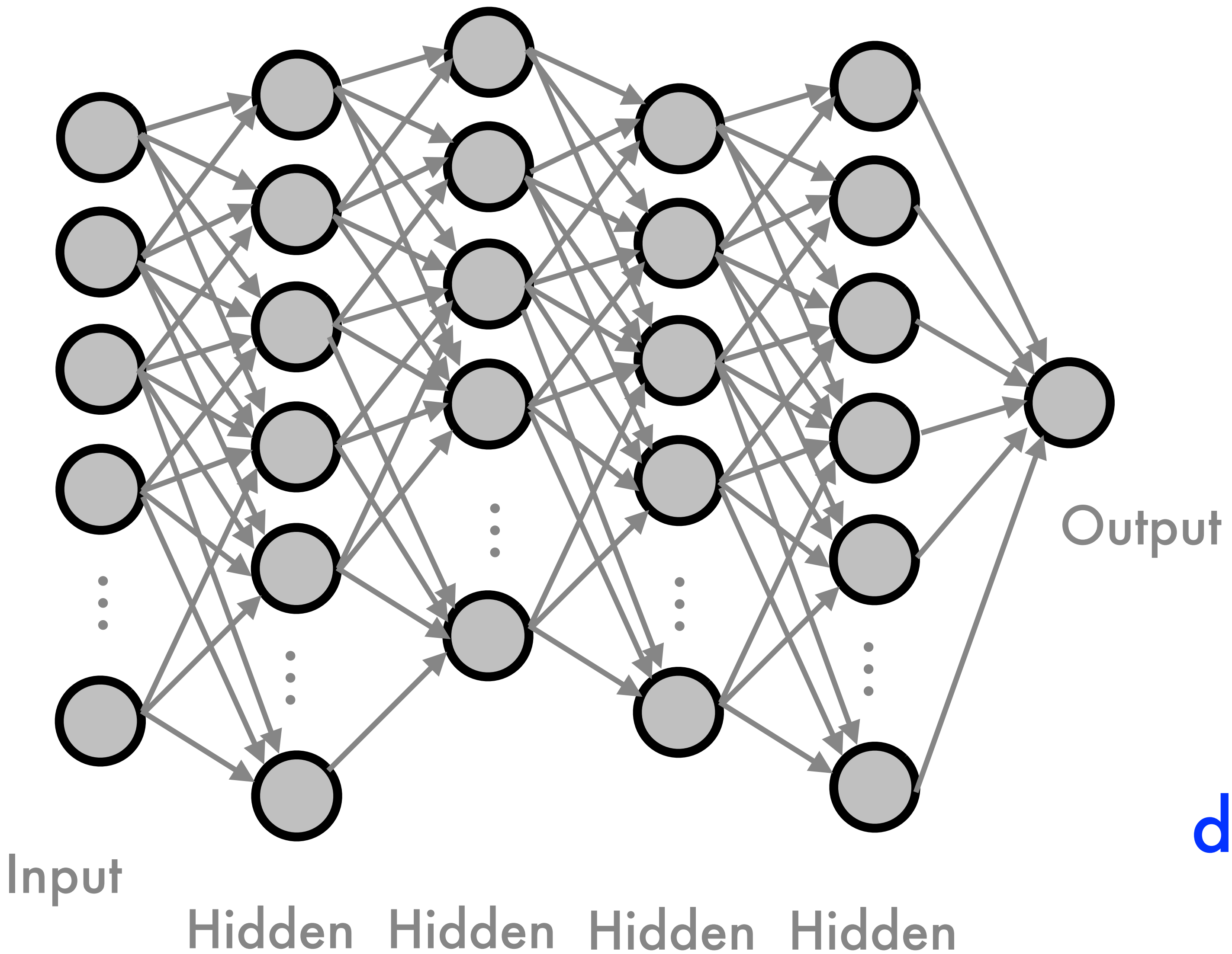
No low-level inputs

Limited input size

Painstaking search through input space.

Variable	VBF			Boosted		
	$\tau_{\text{lep}}\tau_{\text{lep}}$	$\tau_{\text{lep}}\tau_{\text{had}}$	$\tau_{\text{had}}\tau_{\text{had}}$	$\tau_{\text{lep}}\tau_{\text{lep}}$	$\tau_{\text{lep}}\tau_{\text{had}}$	$\tau_{\text{had}}\tau_{\text{had}}$
$m_{\tau\tau}^{\text{MMC}}$	•	•	•	•	•	•
$\Delta R(\tau,\tau)$	•	•	•		•	•
$\Delta\eta(j_1,j_2)$	•	•	•			
m_{j_1,j_2}	•	•	•			
$\eta_{j_1}\times\eta_{j_2}$		•	•			
$p_{\text{T}}^{\text{Total}}$		•	•			
sum p_{T}					•	•
$p_{\text{T}}(\tau_1)/p_{\text{T}}(\tau_2)$					•	•
$E_{\text{T}}^{\text{miss}}\,\phi$ 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				•		
$p_{\text{T}}^{\ell_1}$				•		
$p_{\text{T}}^{j_1}$				•		
$E_{\text{T}}^{\text{miss}}/p_{\text{T}}^{\ell_2}$				•		
m_{T}		•			•	
$\min(\Delta\eta_{\ell_1\ell_2,\text{jets}})$	•					
$j_3\ \eta$ centrality	•					
$\ell_1\times\ell_2\ \eta$ centrality	•					
$\ell\ \eta$ centrality		•				
$\tau_{1,2}\ \eta$ centrality			•			

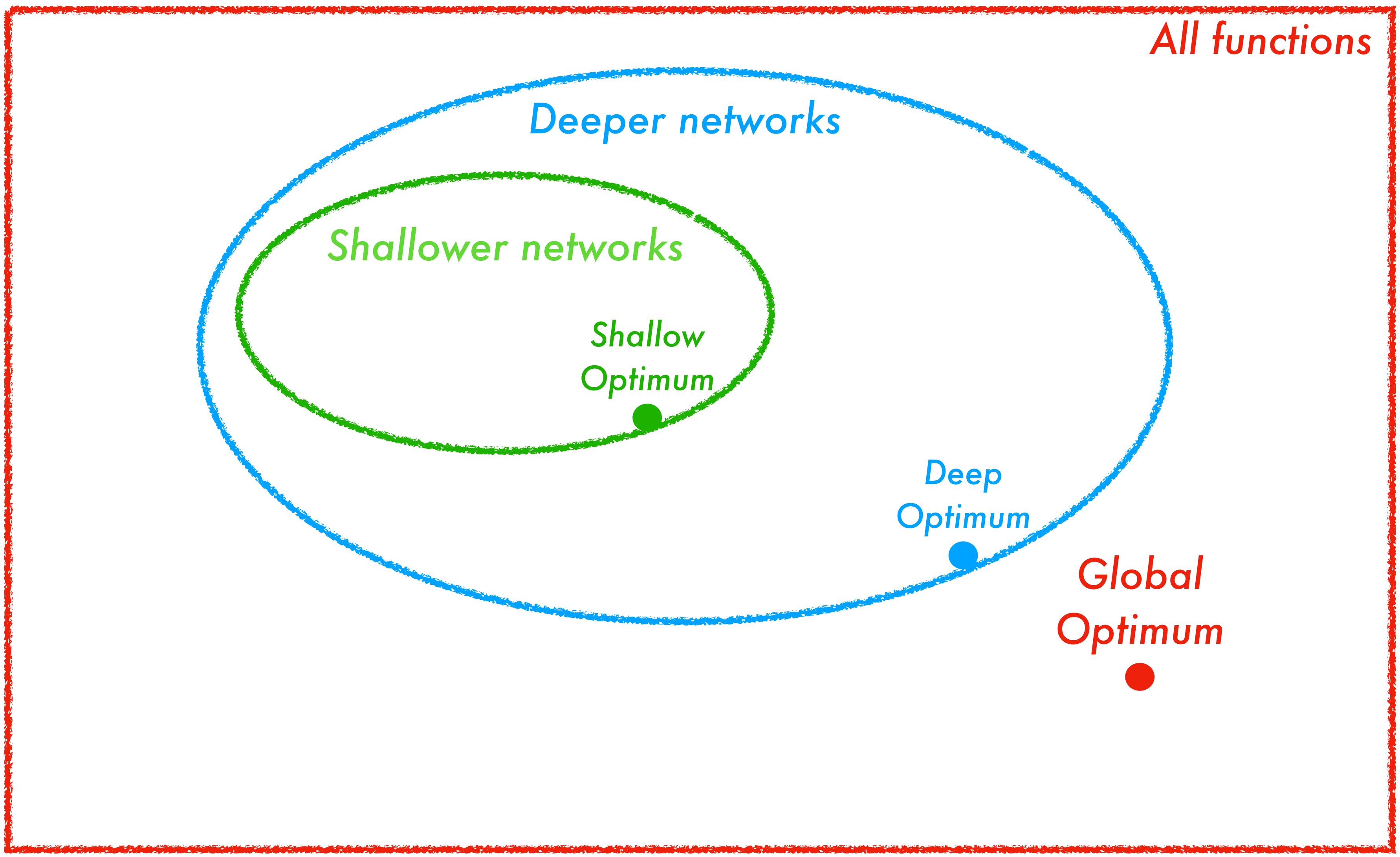
Deep networks



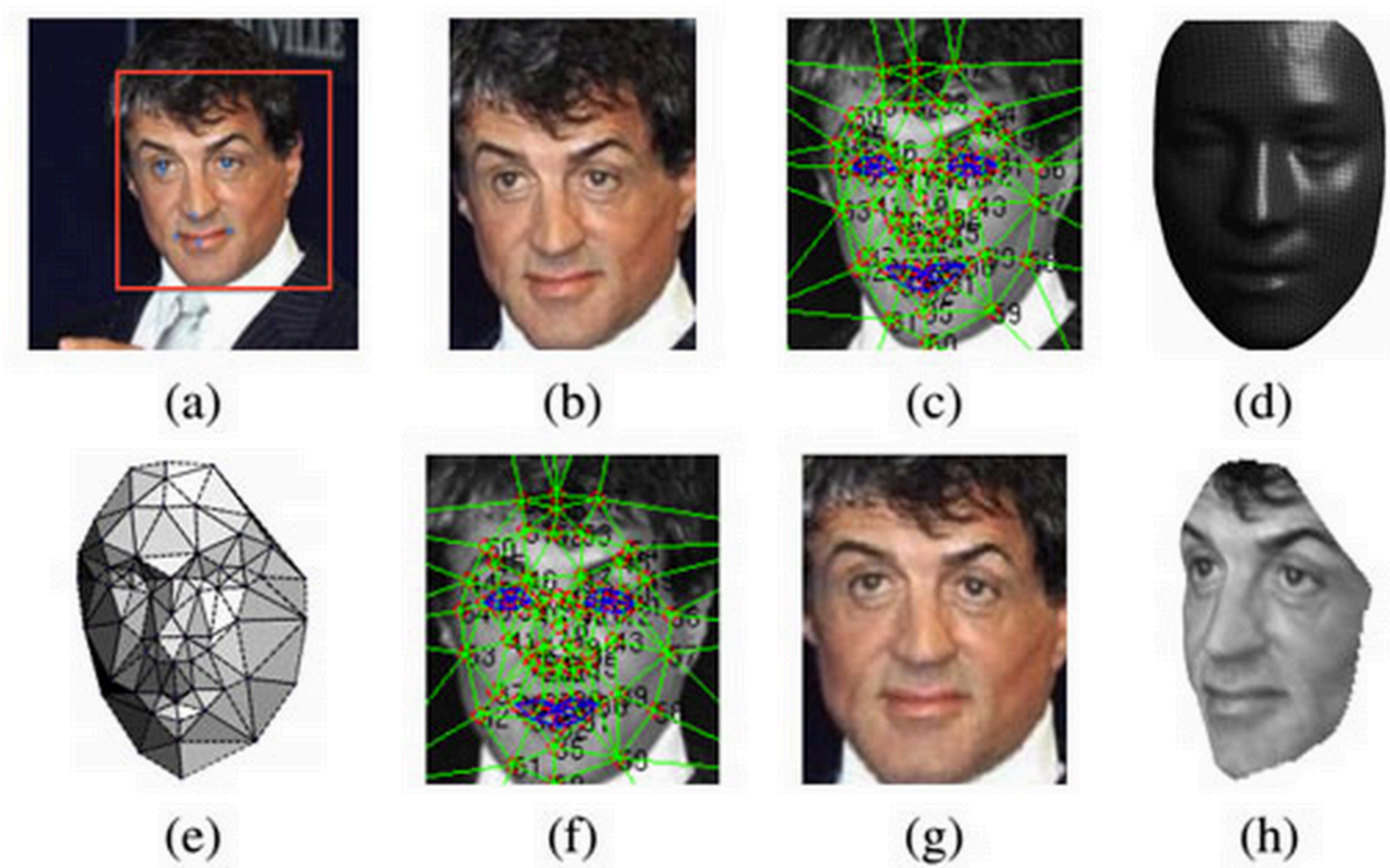
New tools let us train deep networks.

How well do they work?

Expanding space



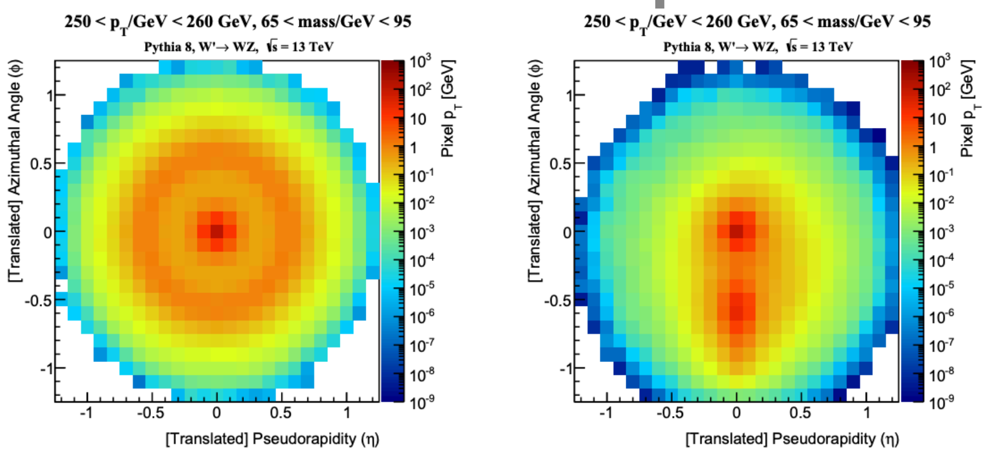
Real world applications



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.

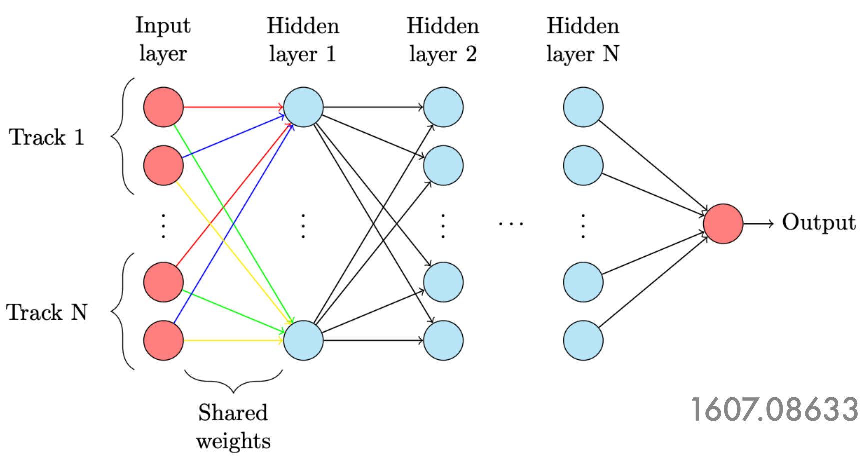
Low level data

Calorimeter pixels



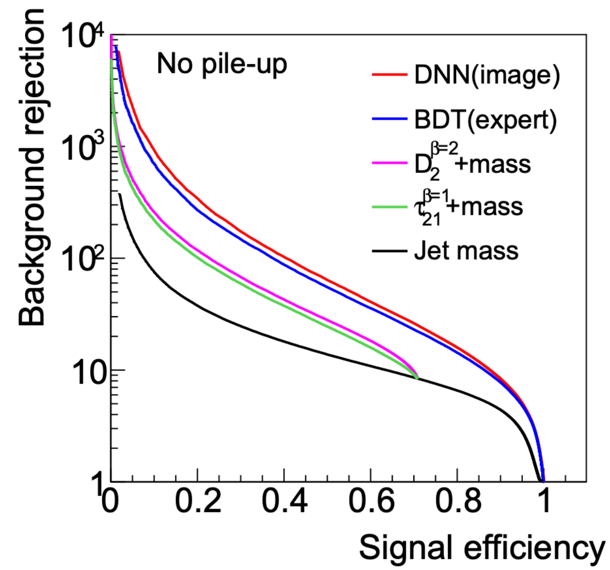
1511.05190

Lists of tracks



1607.08633

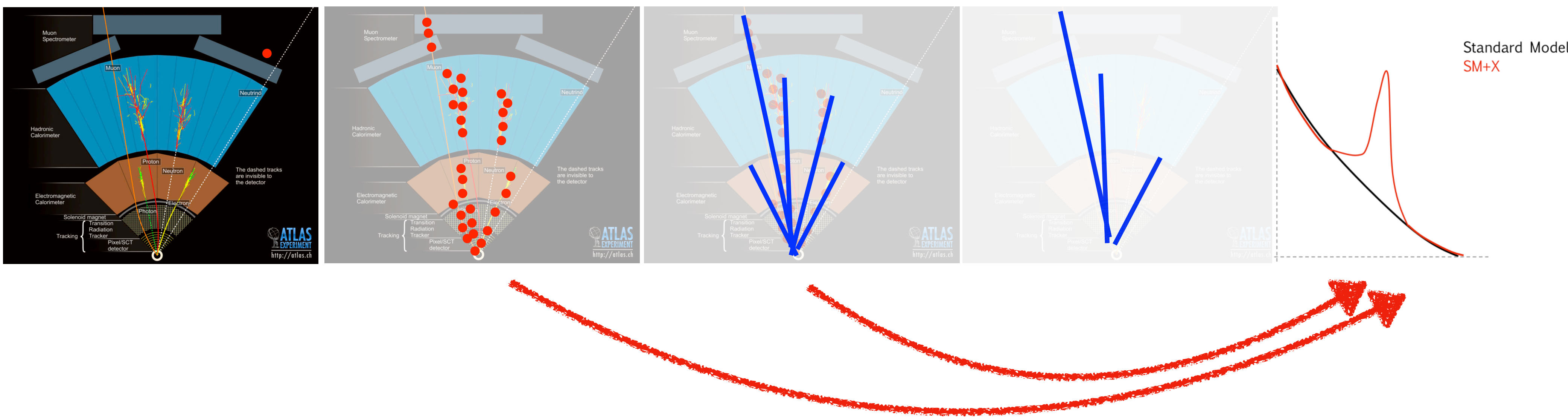
Networks beat experts



1603.09349

Summary statistics

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



Networks can handle higher dimensionality
And lower-level data

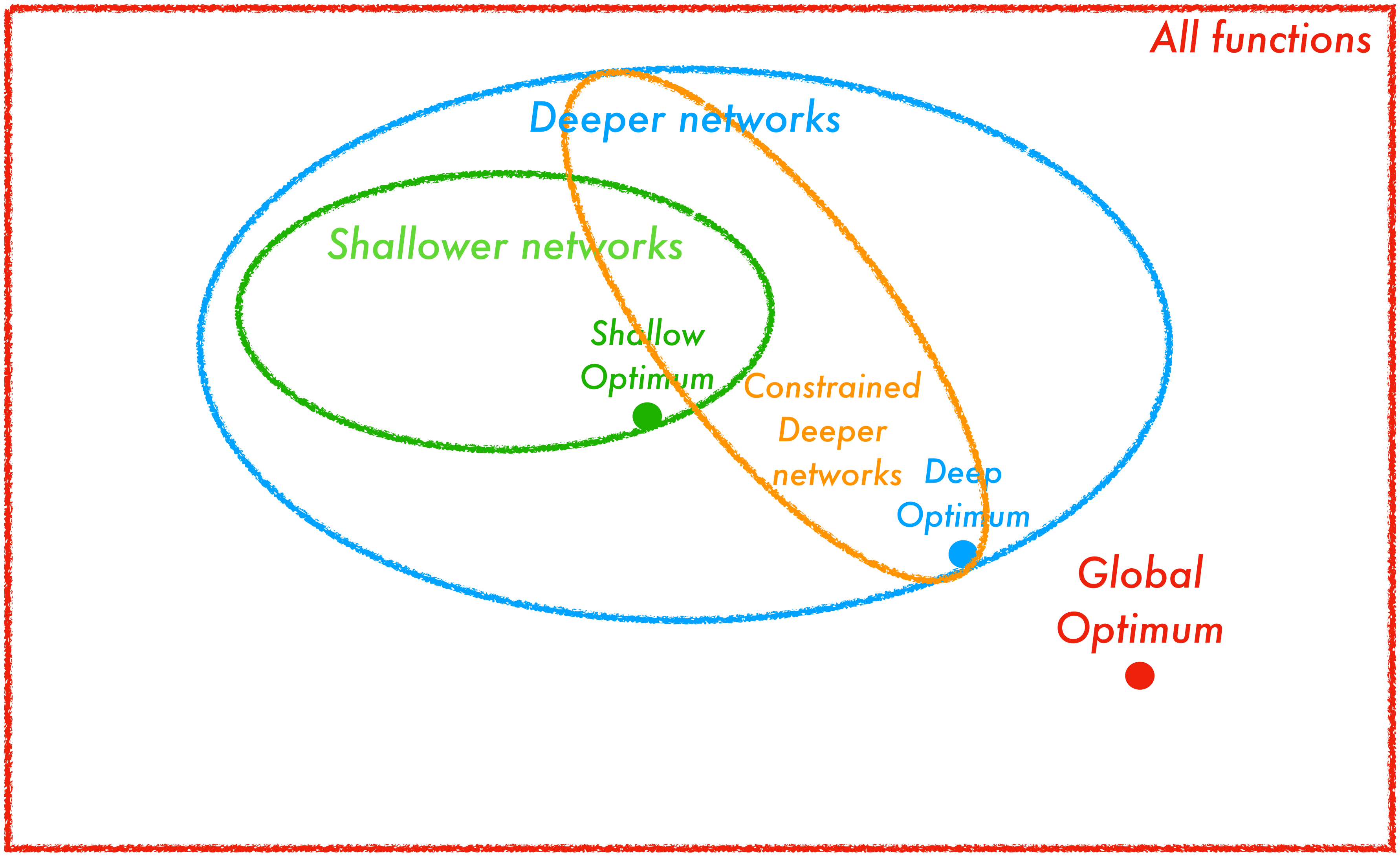
The new frontier

Expertise is not obsolete!

If you know something about the problem,
don't use a completely general solution.
Engineer your network structure!

e.g, network structures which respect symmetries

Constraining space

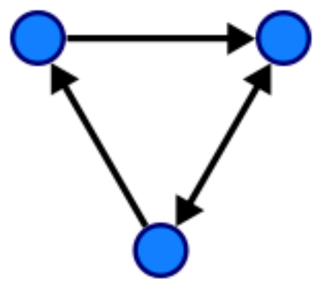


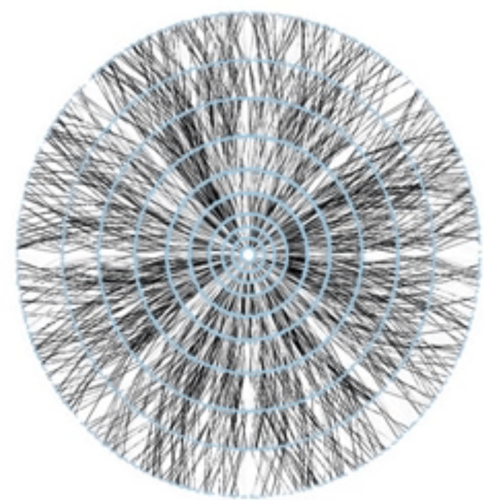
Outline

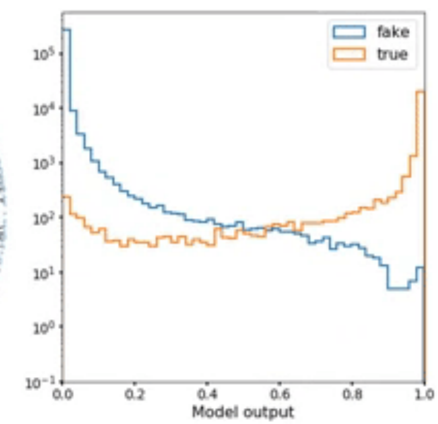
1. Much much more of the same
2. Something qualitatively new

Graph networks

Represent structured data



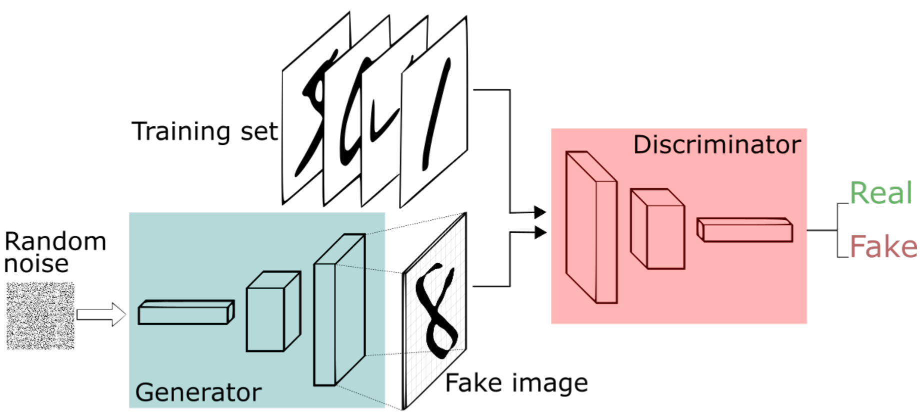


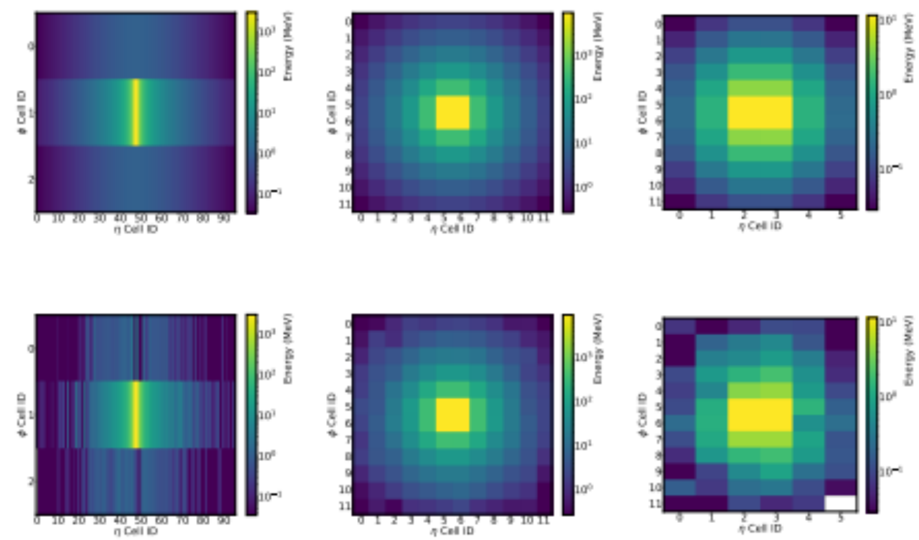


Generative models

Do more than classify

Generate data from noise



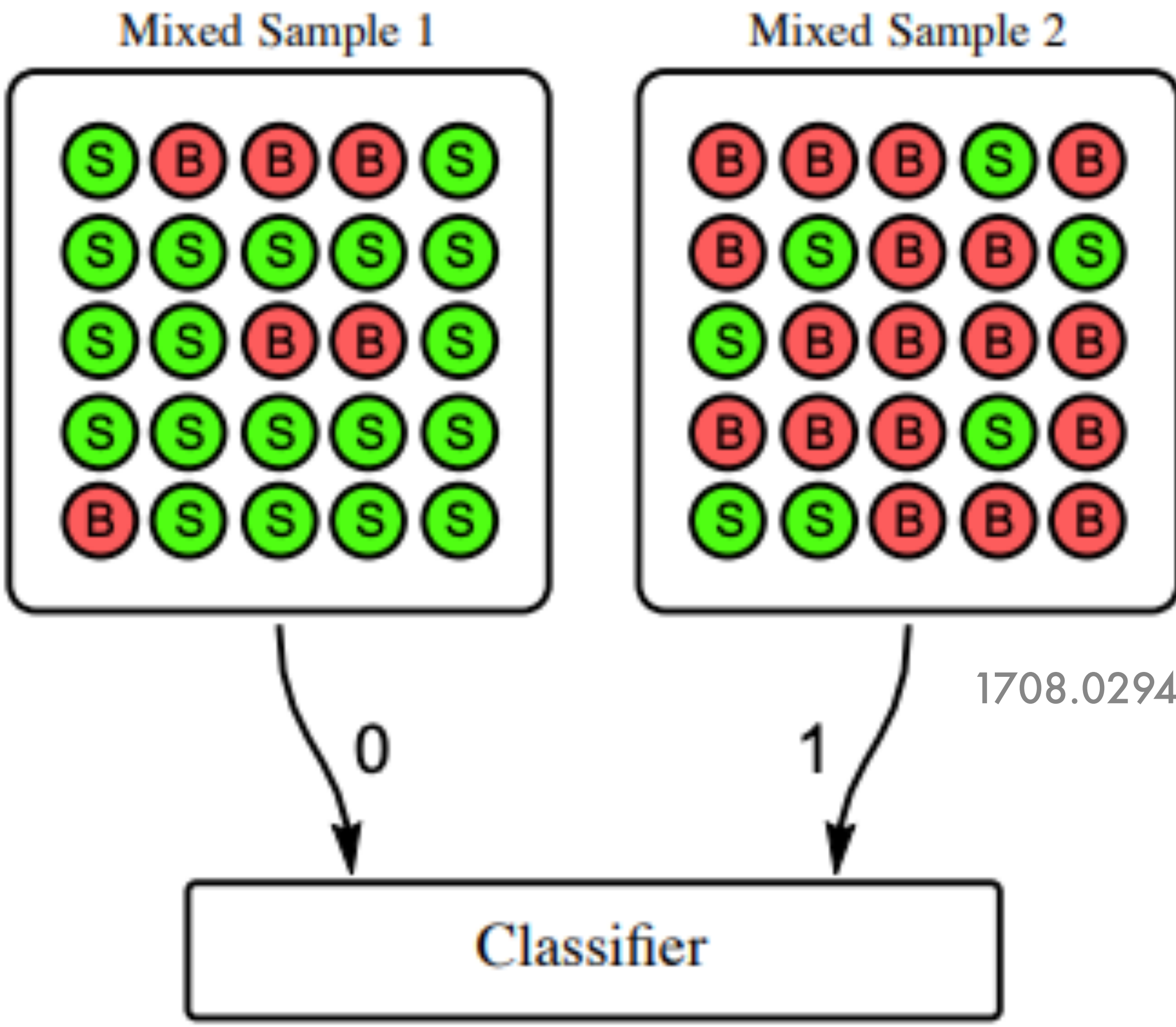


1712.10321

Optimal transport: new ways to compare distributions

2101.08944: Learn the detector from data!

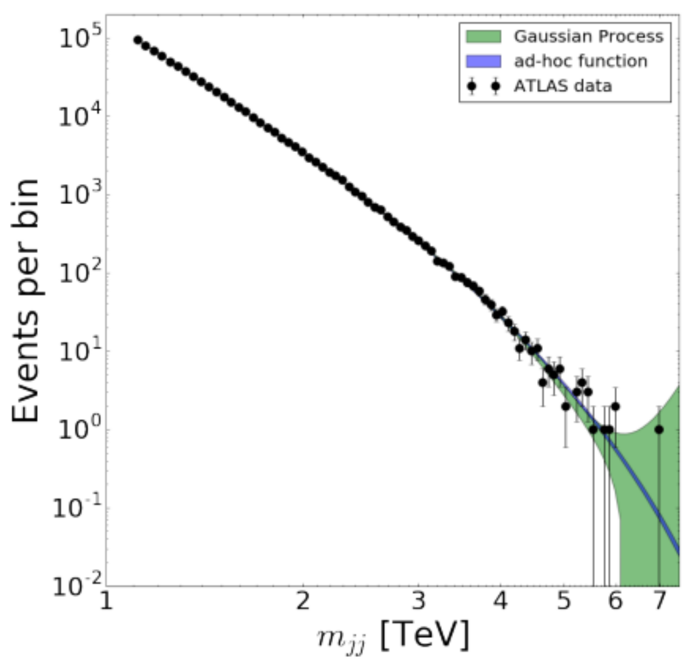
Away from supervision



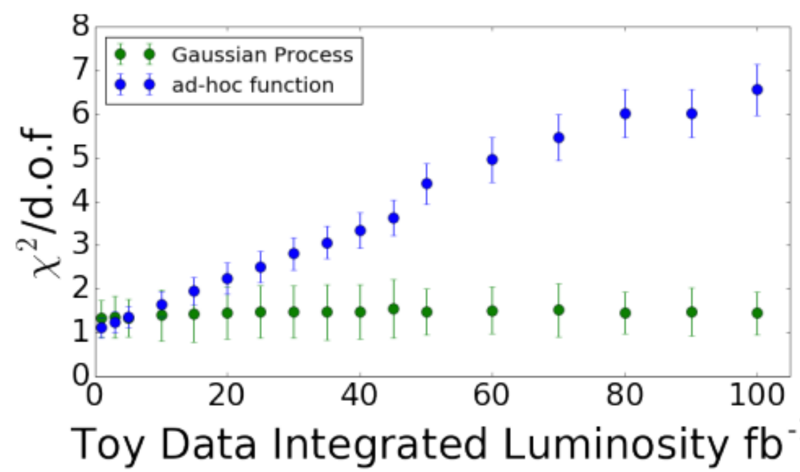
1708.02949

Background fitting

Away from ad-hoc background shapes:



1709.05681



ML for design

Optimize everything

Automatic Differentiation

Numerical gradients $\Delta L / \Delta\phi$ hopeless in trillion-D, need exact gradients $\partial L / \partial\phi$

Automatic Differentiation: careful application of *chain rule to computer programs*

```
import jax
import jax.numpy as jnp

def func(x):
  y = x
  for i in range(4):
    y += x[0]**2 + jnp.sin(x[1]) + jnp.exp(-x[2])
  y = y.sum()
  return y
```

exact gradients!

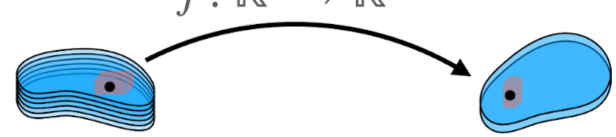
```
gfunc = jax.value_and_grad(func)
gfunc(jnp.array([2.,3.,-2]))

(DeviceArray(141.36212, dtype=float32),
 DeviceArray([ 49.        , -10.8799095, -87.66867  ], dtype=float32))
```

TensorFlow JAX PYTORCH

... but also C++, Fortran, ...

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$$



$$y = f(x) \qquad dy = J_f \, dx$$

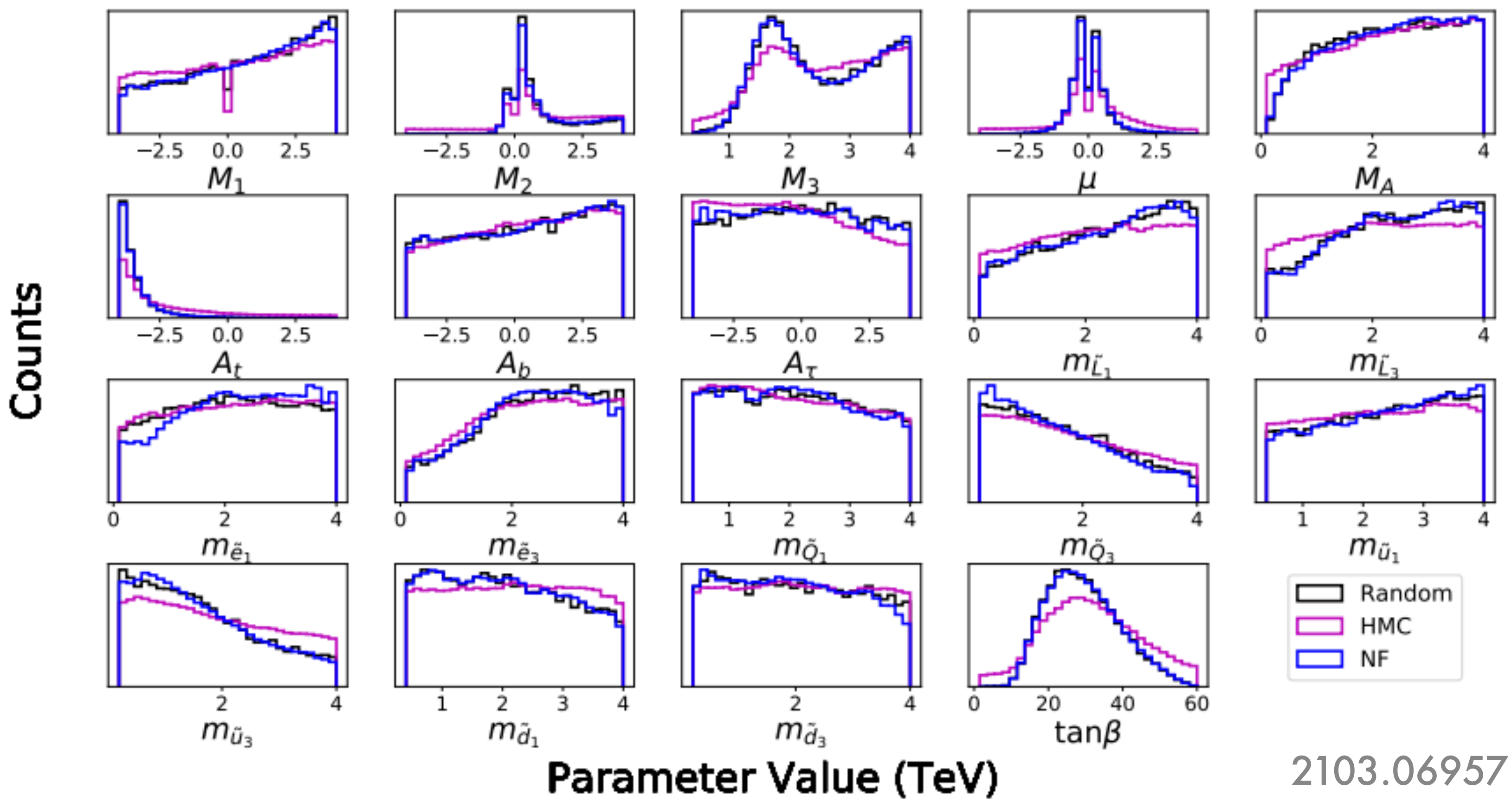
$$J_f = \frac{\partial(y_1, \ldots, y_m)}{\partial(x_1, \ldots, x_n)}$$

L. Heinrich

See also: 1806.04743

ML for Theory!

How do we search large spaces?



2103.06957

String theory applications: 1707.00655,1903.11616

Summary

<u>Modern ML</u>

Much more flexible and capable

Tackling previously intractable problems

<u>Many creative new ideas</u>

Widening in scope

Attacking new problems