

# Recent progress in ML applications for LHC phenomenology

Benjamin Nachman

*Lawrence Berkeley National Laboratory*

[bpnachman.com](http://bpnachman.com)

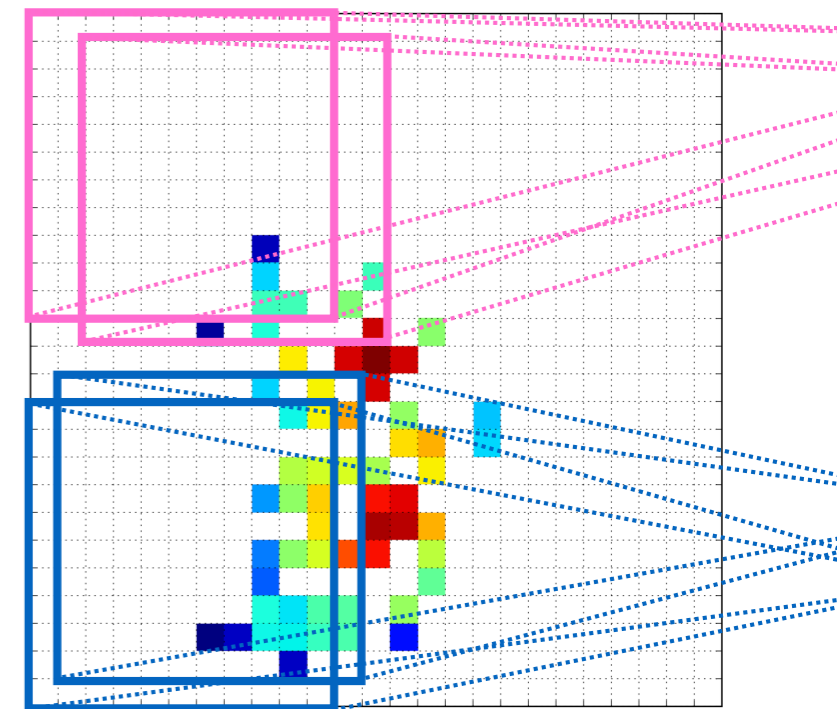
[bpnachman@lbl.gov](mailto:bpnachman@lbl.gov)



@bpnachman



bnachman

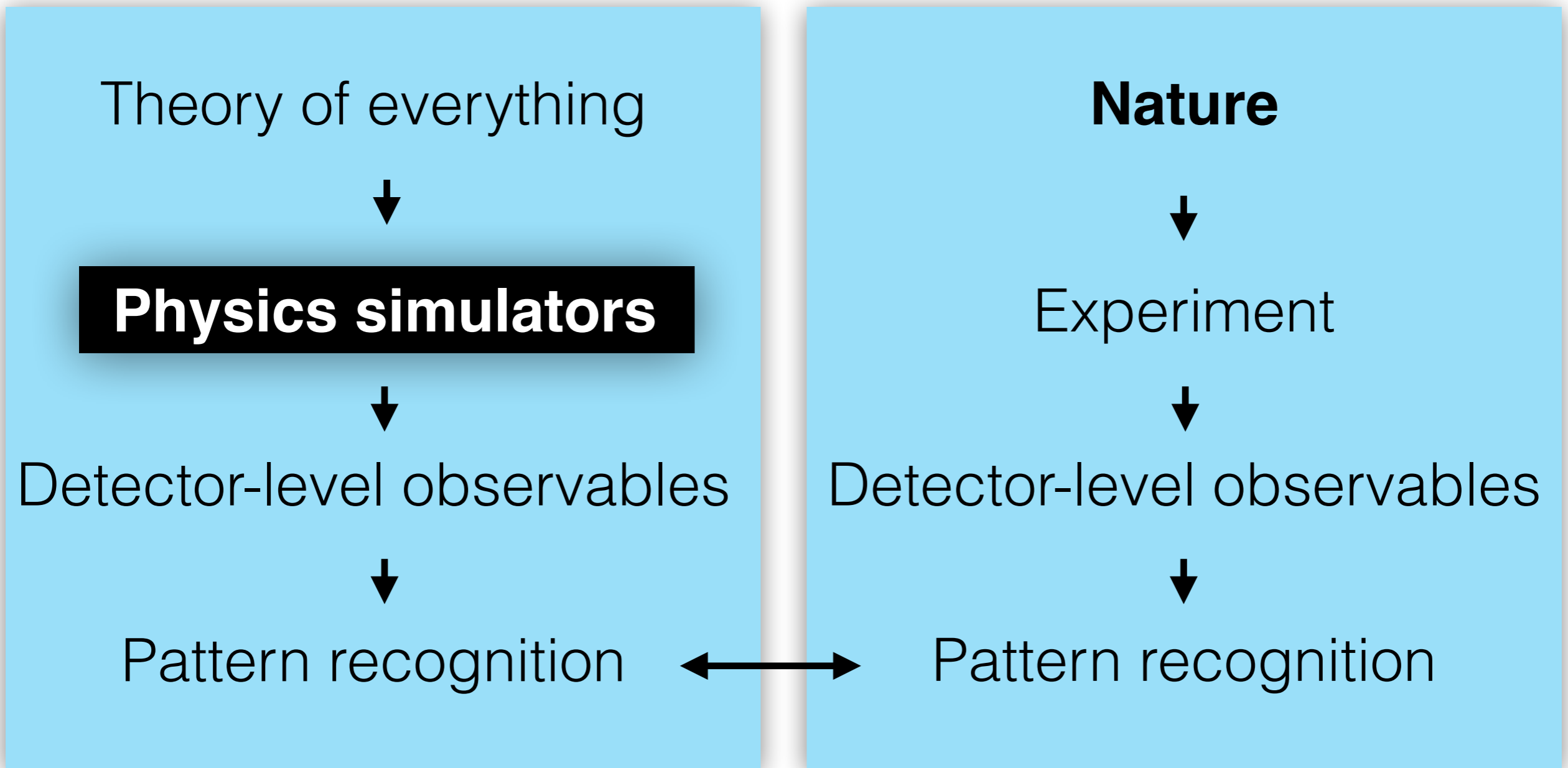


Virtual Houches

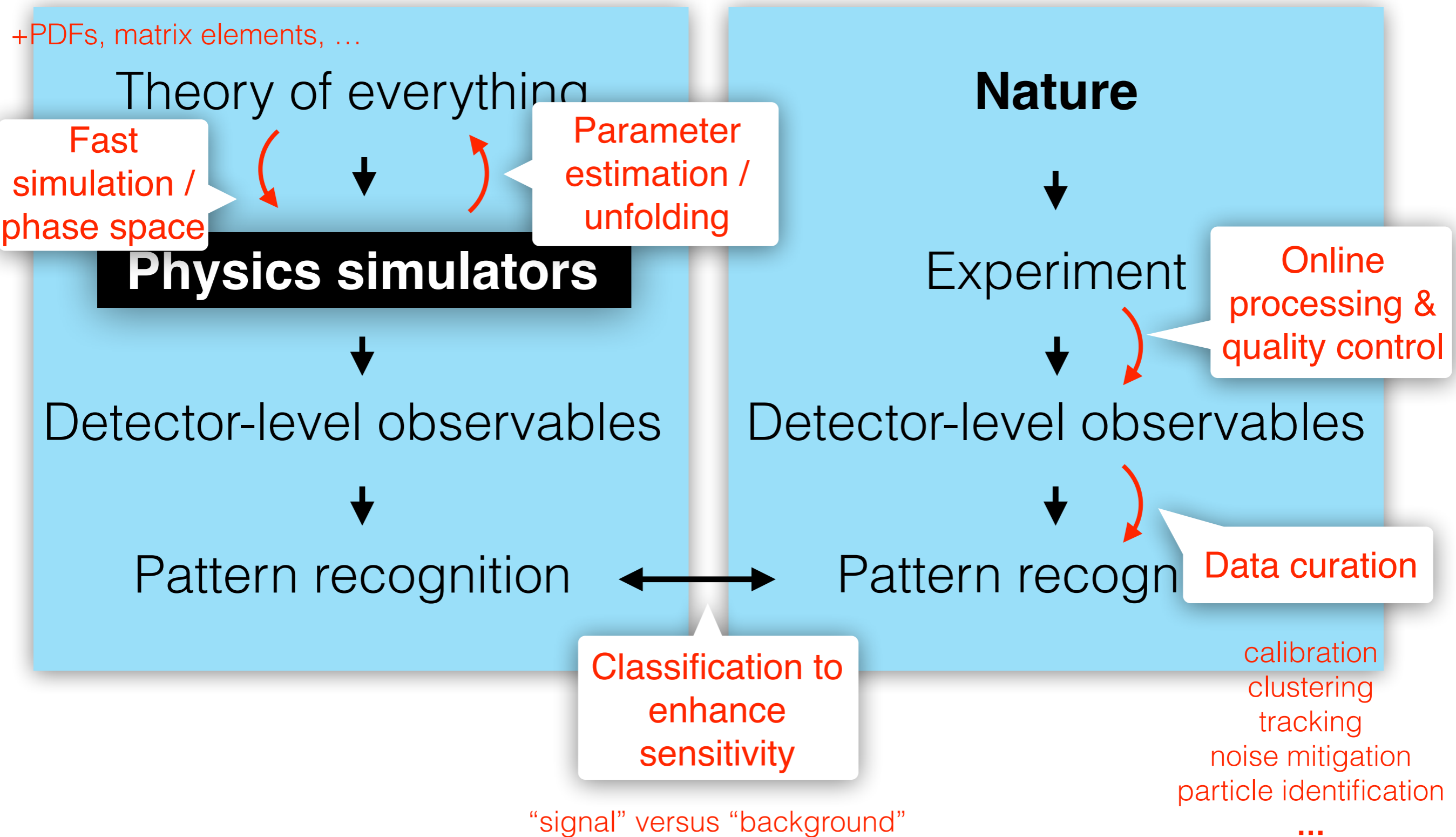
June 17, 2021

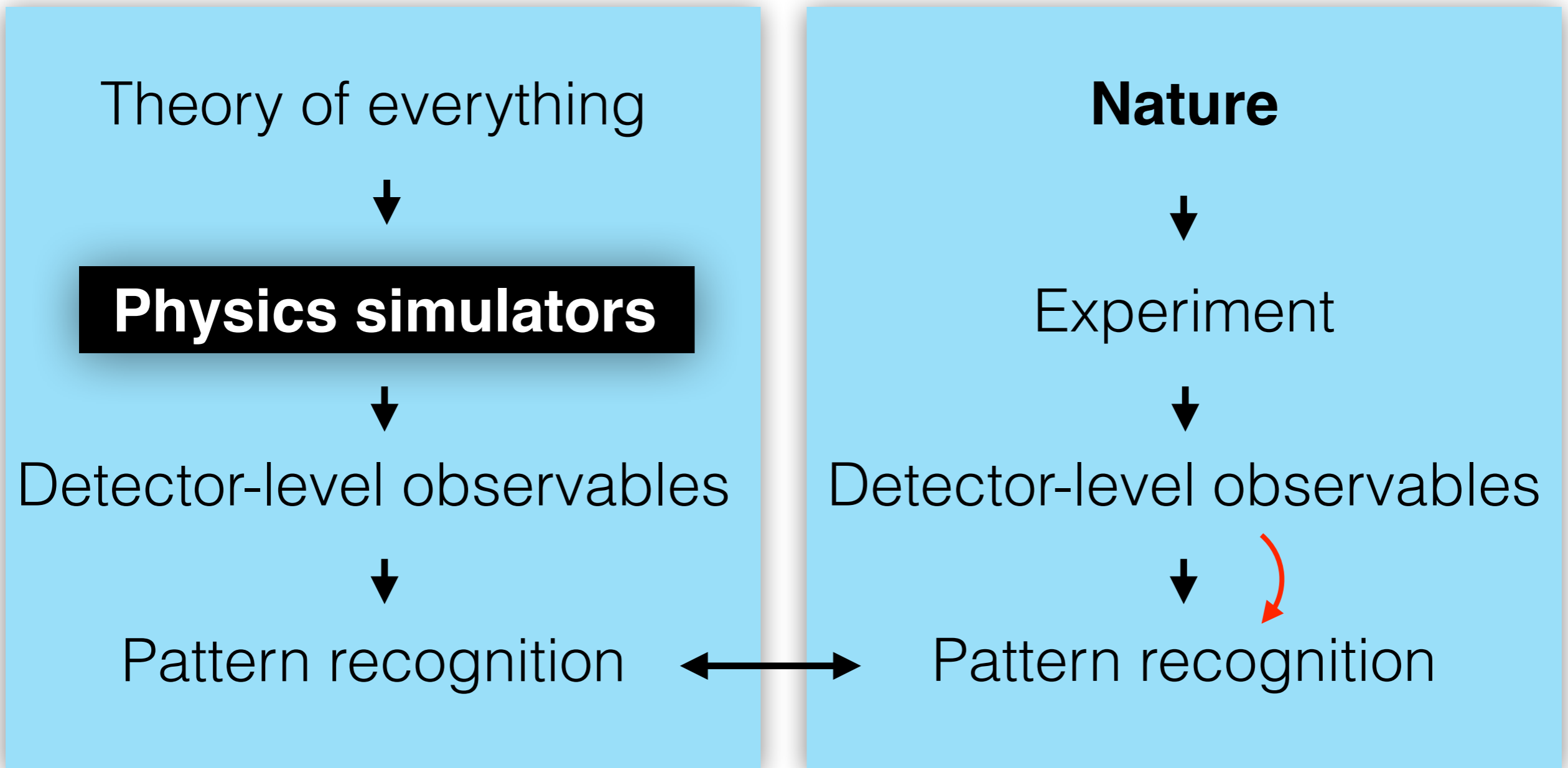
# Data analysis in particle physics

2



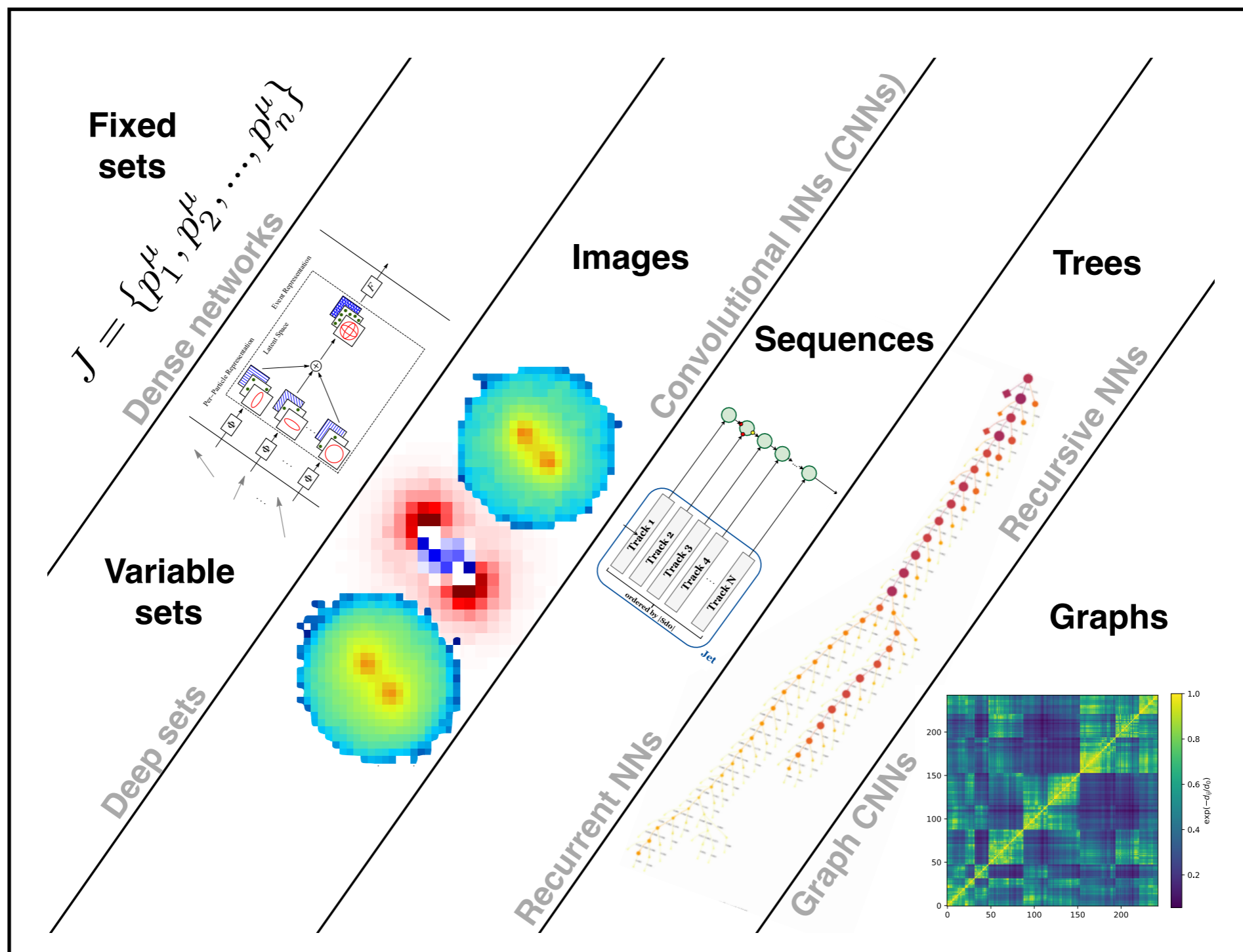
# Data analysis in particle physics + ML





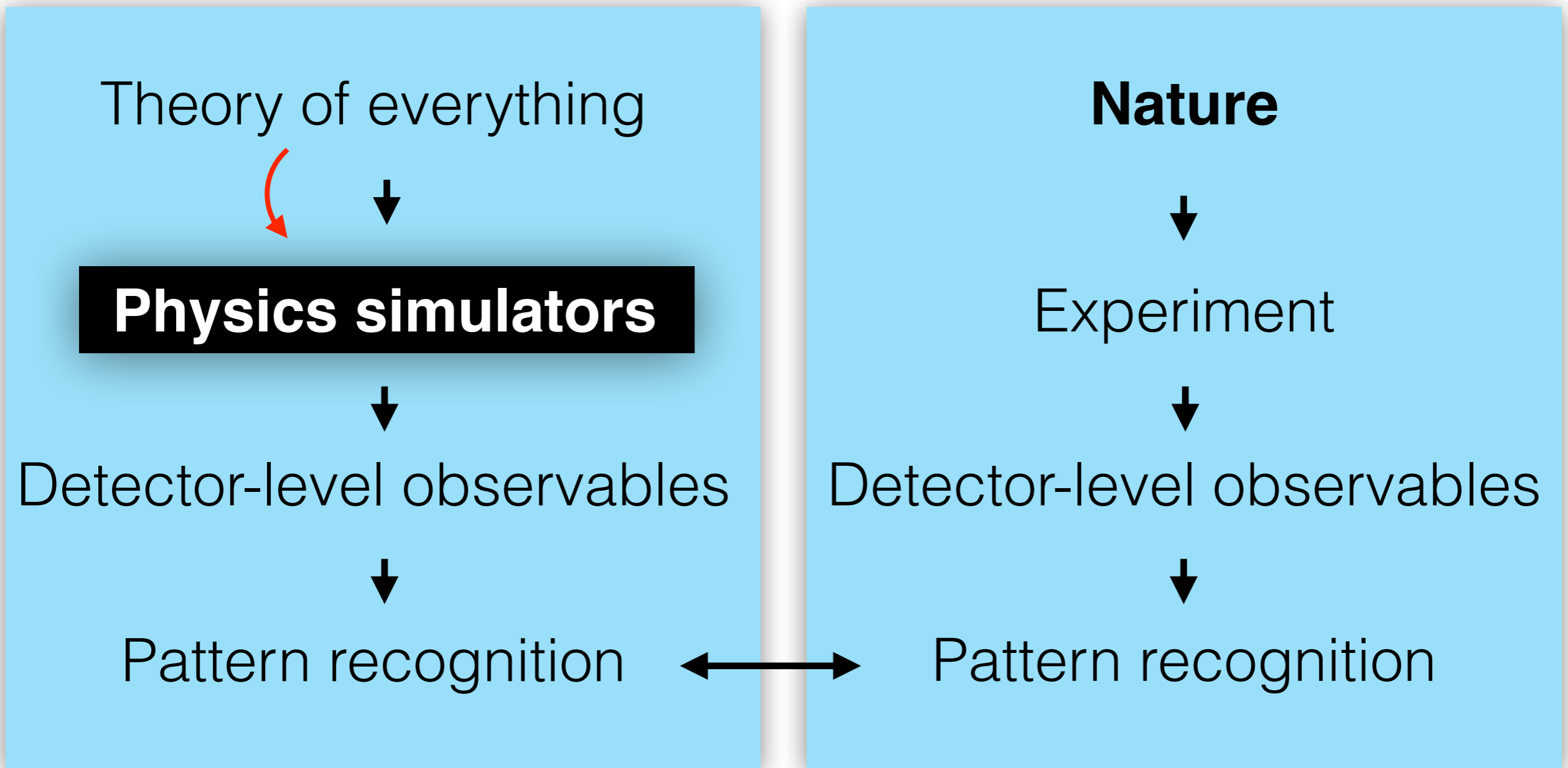
This is where most machine learning is being applied.

# Representing our data



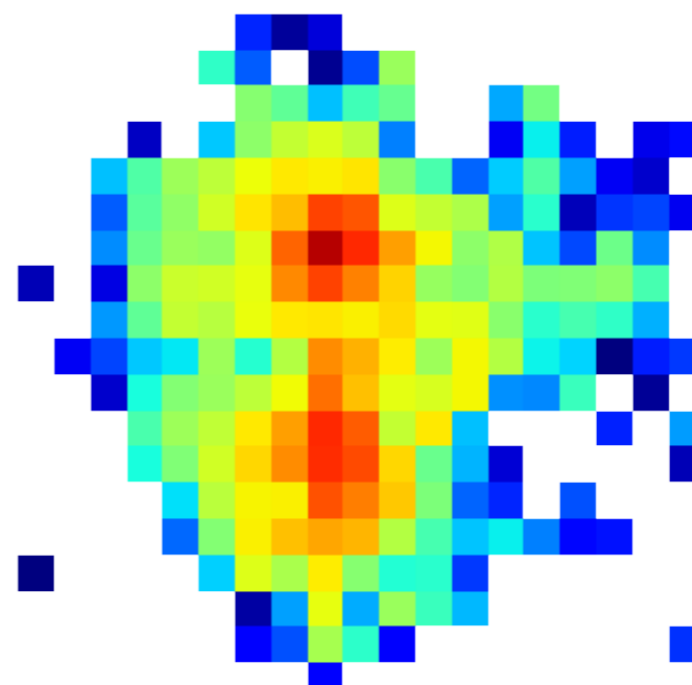
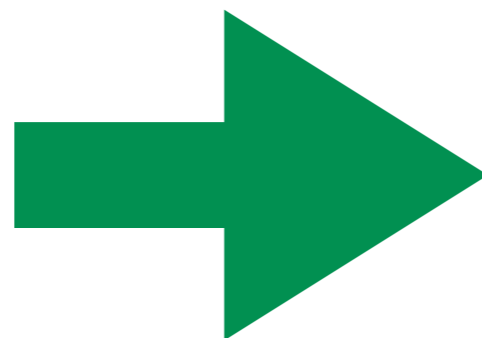
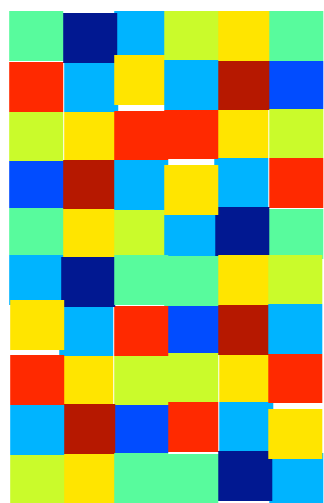
# Data analysis in particle physics

6



The growing toolkit of generative models are being developed to accelerate or augment simulations.

A **generator** is nothing other than a function that maps random numbers to structure.



# Deep Generative Models in HEP



Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches

*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*



# Deep Generative Models in HEP



Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches

*N.B. being comprehensive  
with citations would fill up  
the slide - please see my link  
to the Living Review at the  
end for a comprehensive list*

*N.B. being comprehensive with citations would  
fill up the slide - please see my link to the Living  
Review at the end for a comprehensive list*

# Deep Generative Models in HEP

10

Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches

*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*

# Deep Generative Models in HEP

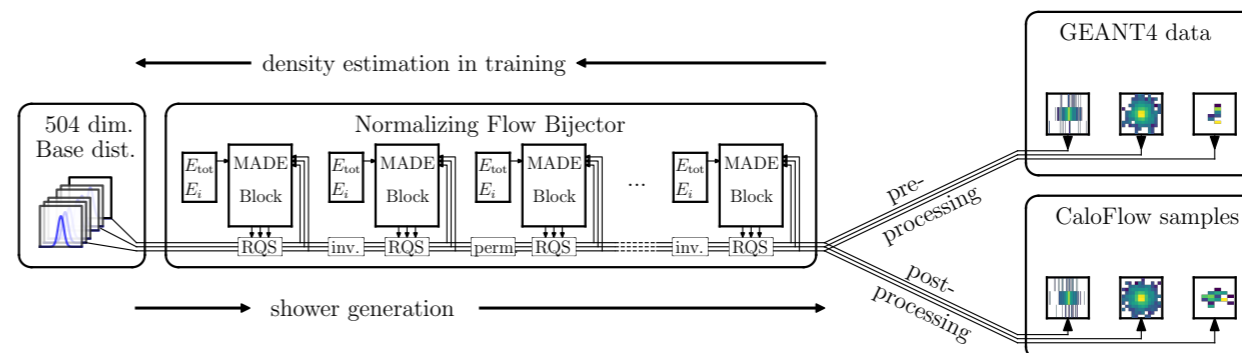
Speeding up slow simulation

Generating Phase space

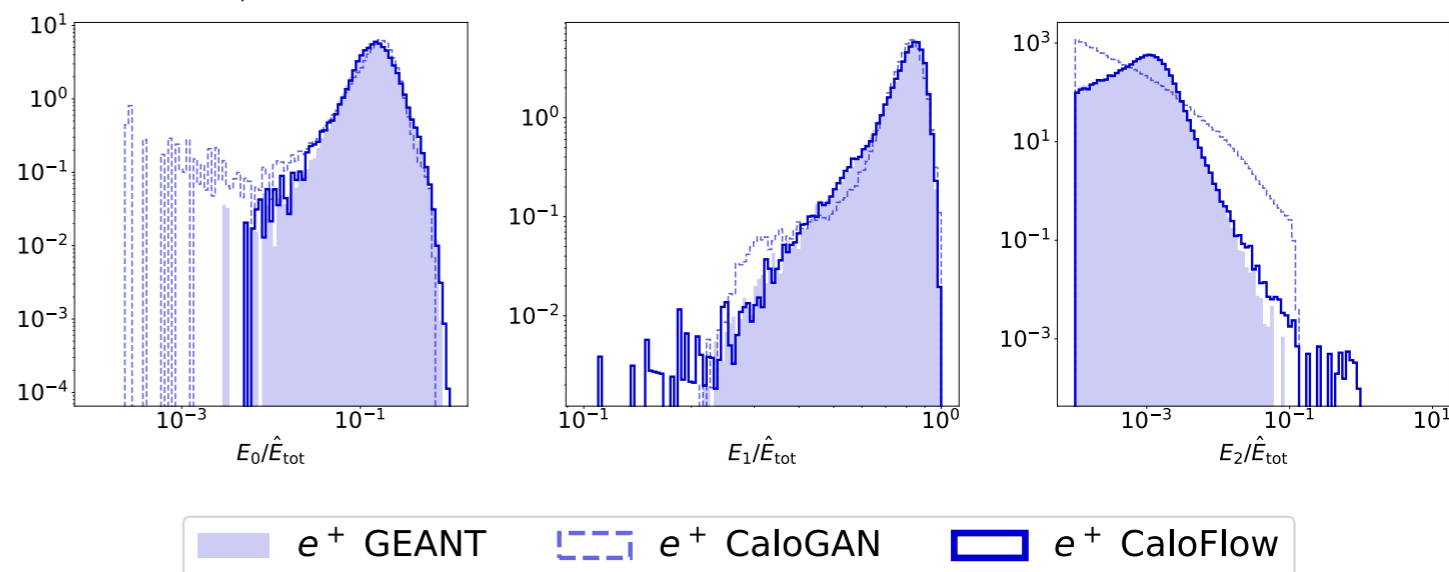
Estimating SM backgrounds

Measurements and Inference

BSM searches



GAN, Flow are NNs



CaloFlow: Krause and Shih, 2106.05285

CaloGAN: Paganini, Oliveira, Nachman, 1705.02355

not quite a fair comparison, but the state-of-the-art accuracy is highly non-trivial and very impressive!

*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*

# Deep Generative Models in HEP

12

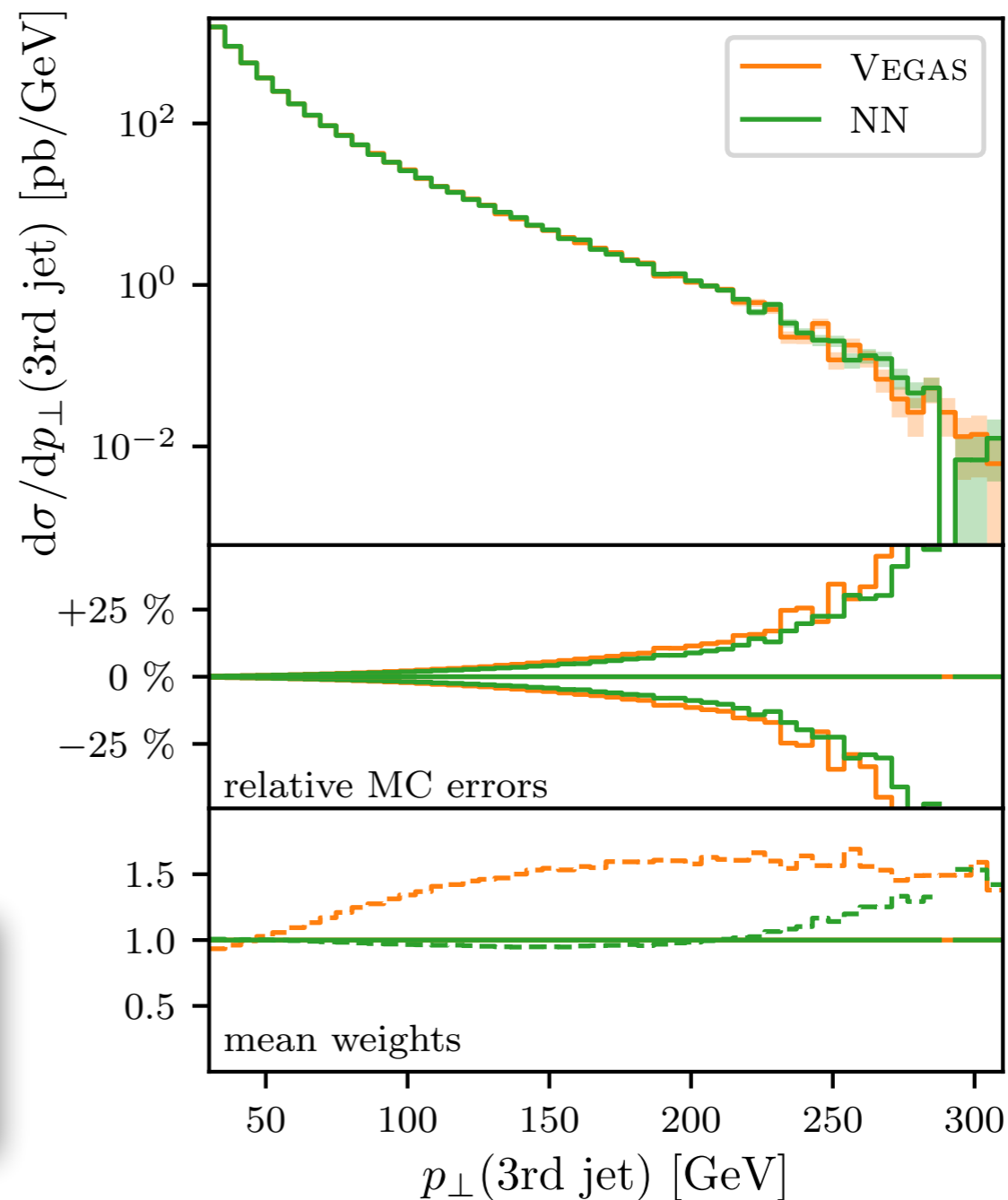
Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches



Bothmann et al., 2001.05478

*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*

# Deep Generative Models in HEP

13

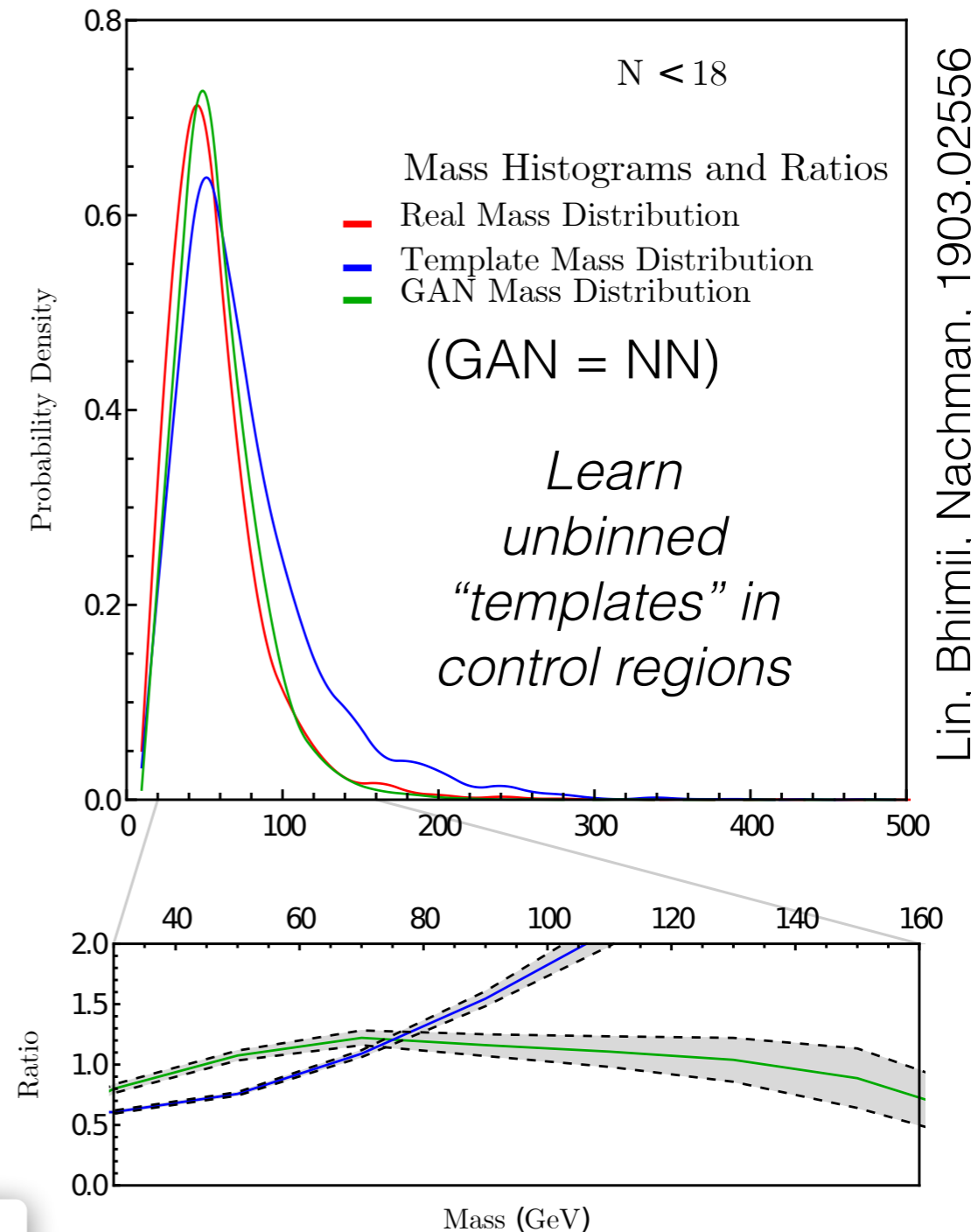
Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches



*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*

# Deep Generative Models in HEP

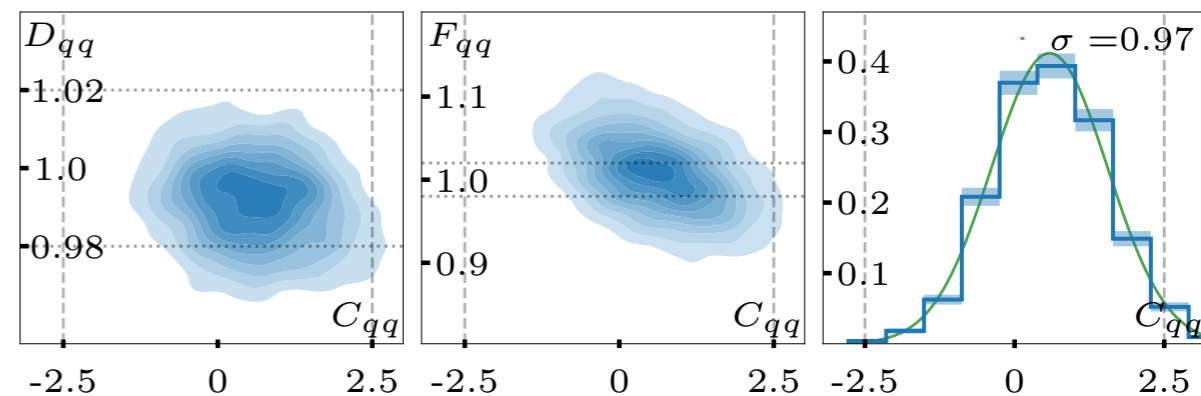
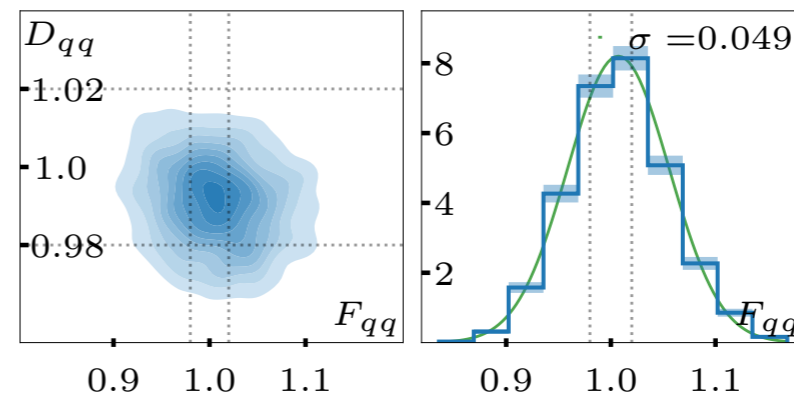
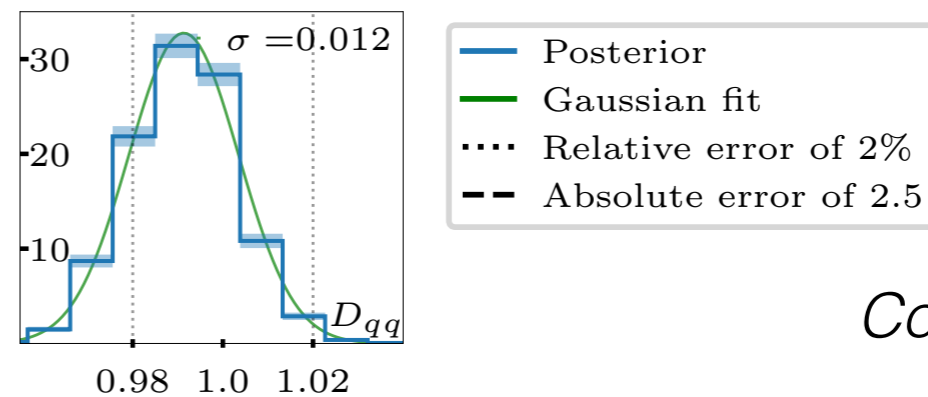
Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

Measurements  
and Inference

BSM searches



*Coefficients of  
splitting  
functions with  
invertible NNs*

Bieringer et al., 2012.09873

*N.B. being comprehensive with citations would fill up the slide - please see my link to the Living Review at the end for a comprehensive list*

# Deep Generative Models in HEP

15

Speeding up  
slow simulation

Generating  
Phase space

Estimating SM  
backgrounds

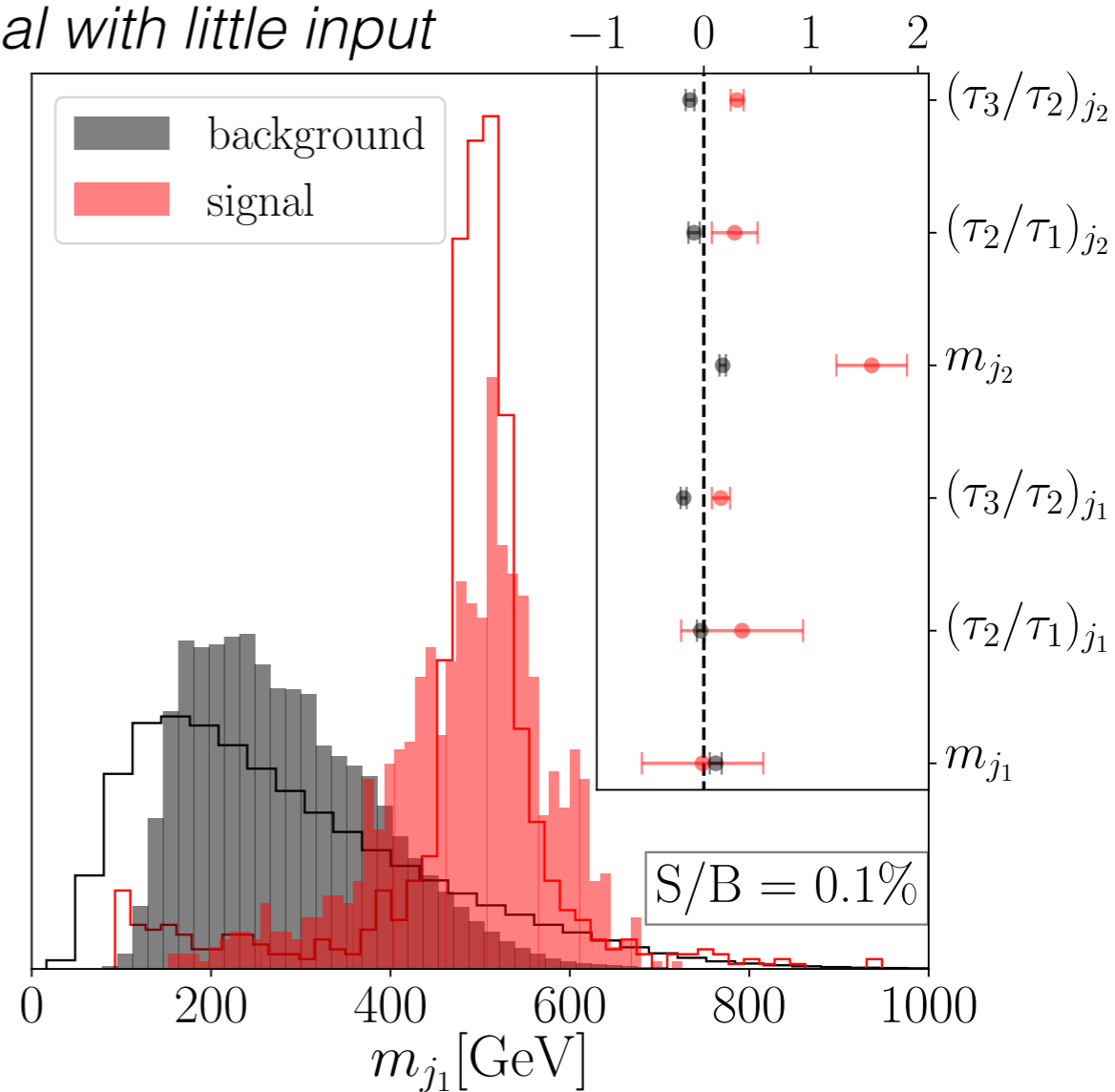
Measurements  
and Inference



See also the LHC  
Olympics 2020

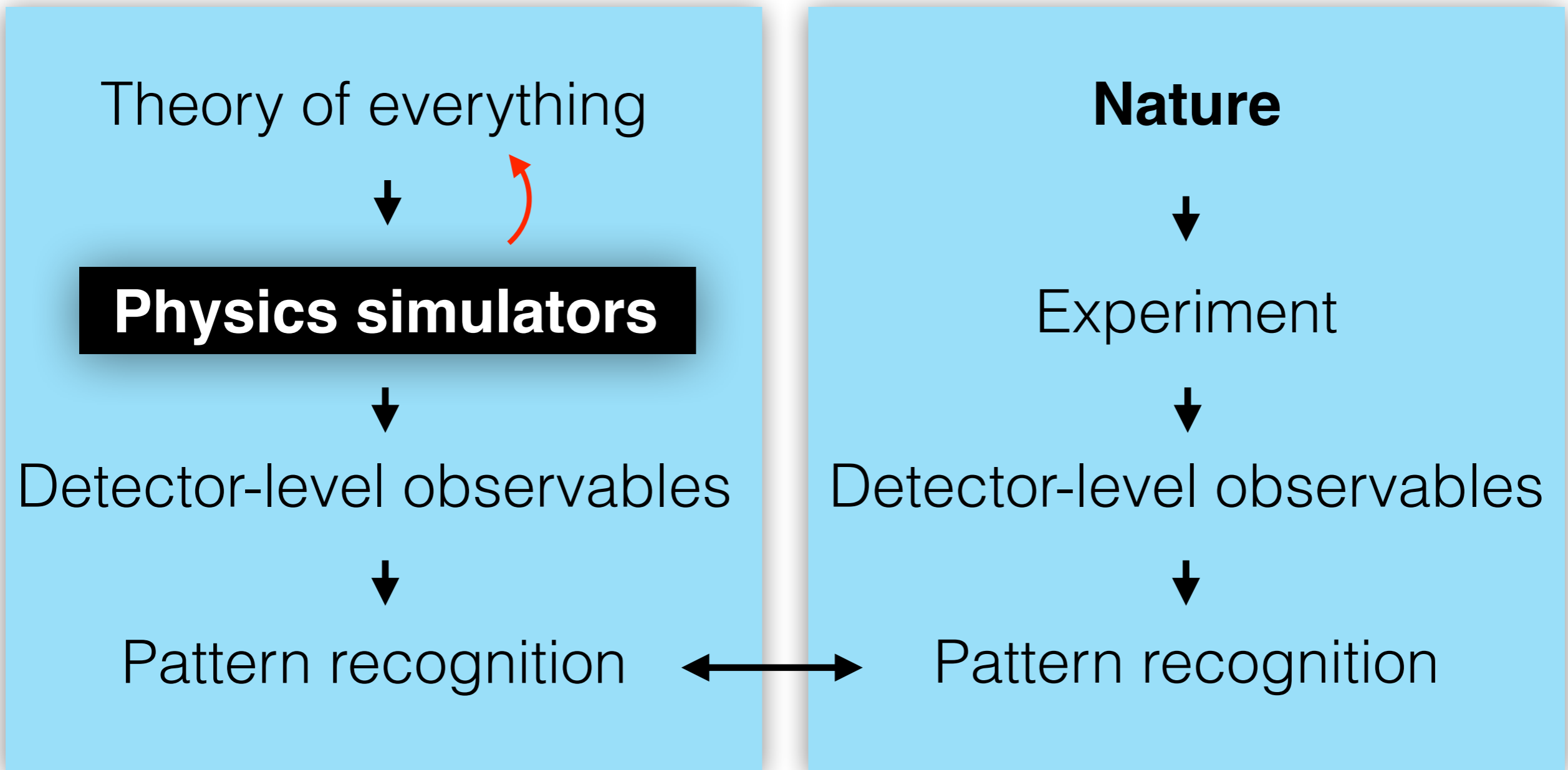
BSM searches

*Extract features of  
signal with little input*



Bortolato et al., 2103.06595

*N.B. being comprehensive with citations would  
fill up the slide - please see my link to the Living  
Review at the end for a comprehensive list*



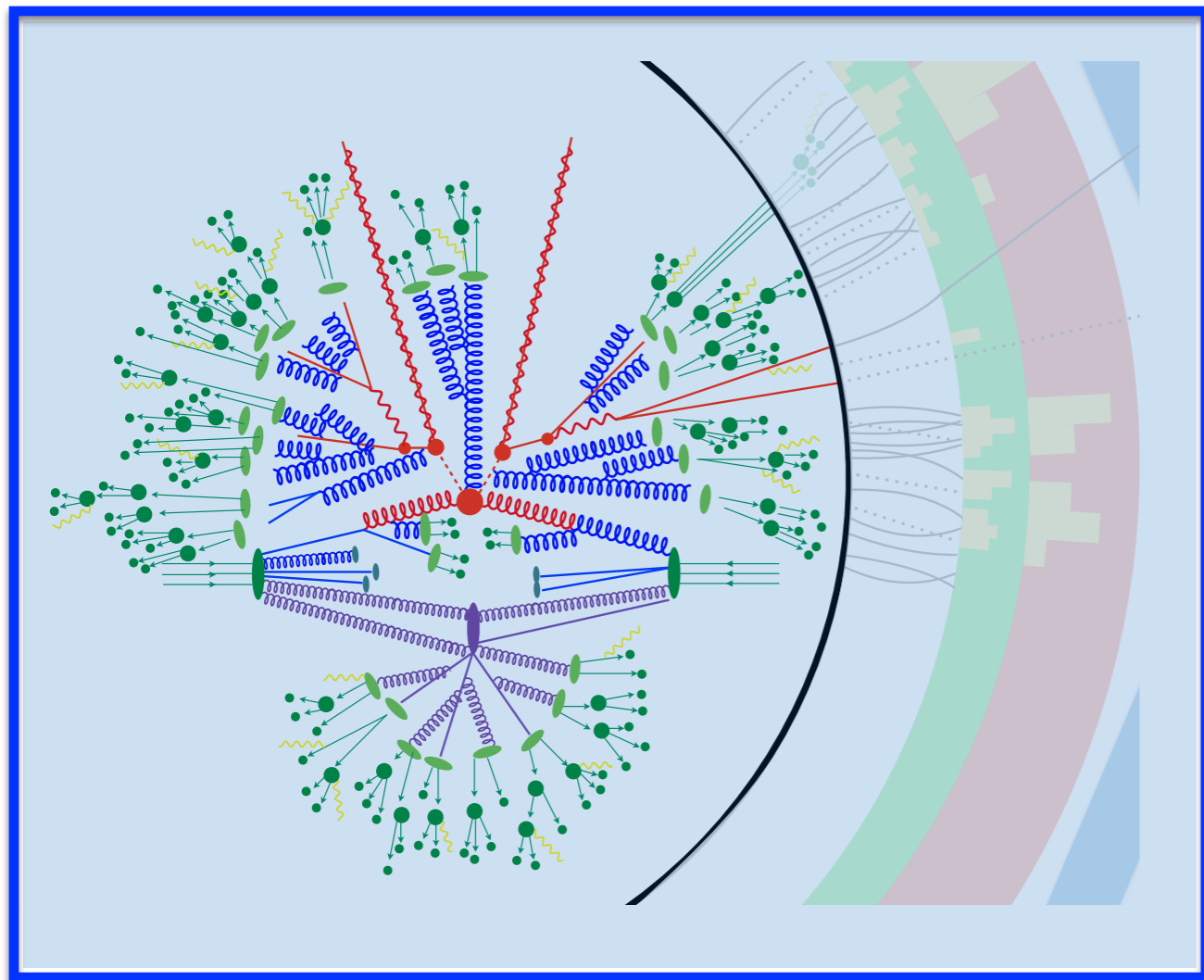
Simulators are a unique and powerful aspect of particle physics, but, they do not allow us to go “backwards” !!



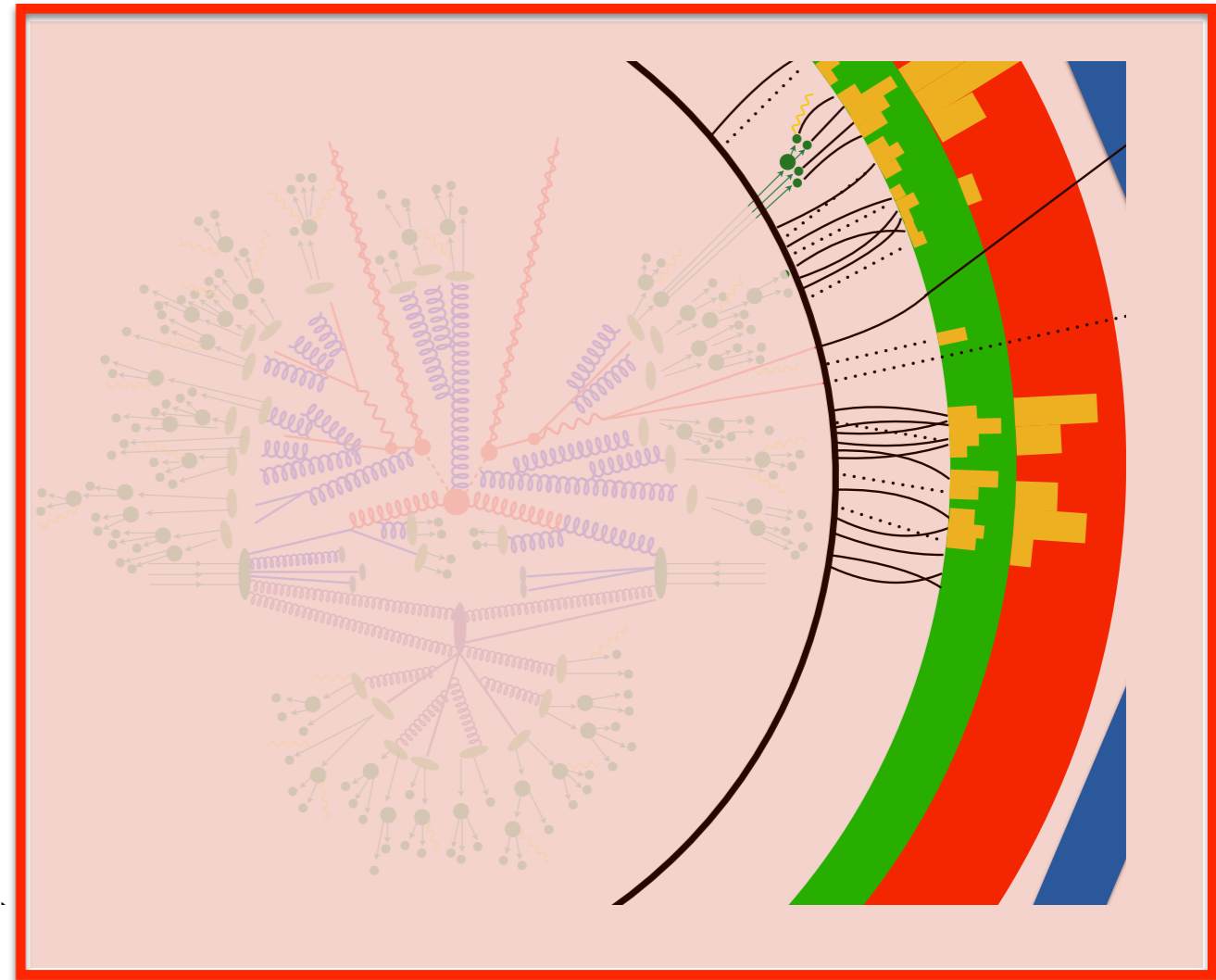
# The Inference Challenge

17

**Want this**



**Measure this**



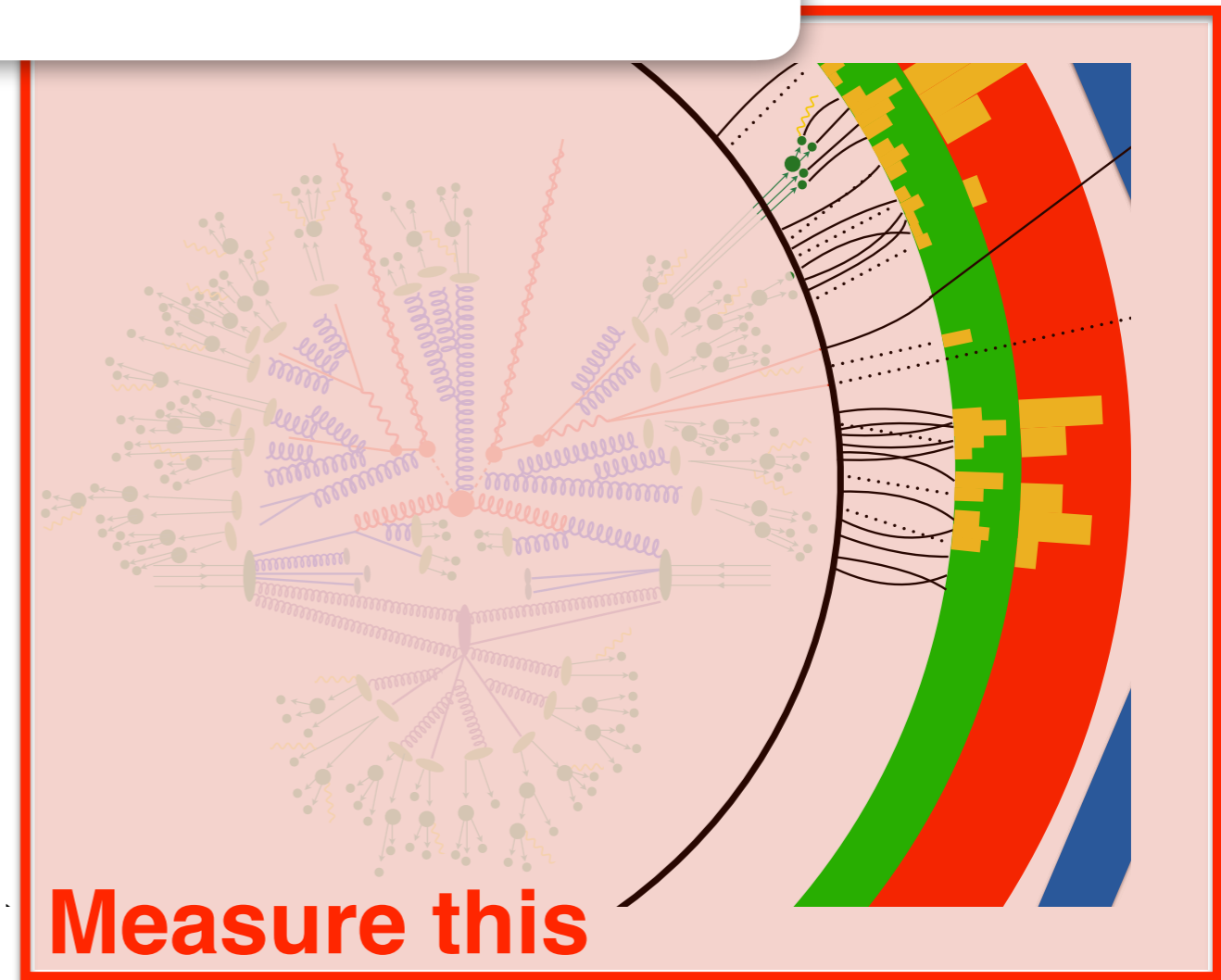
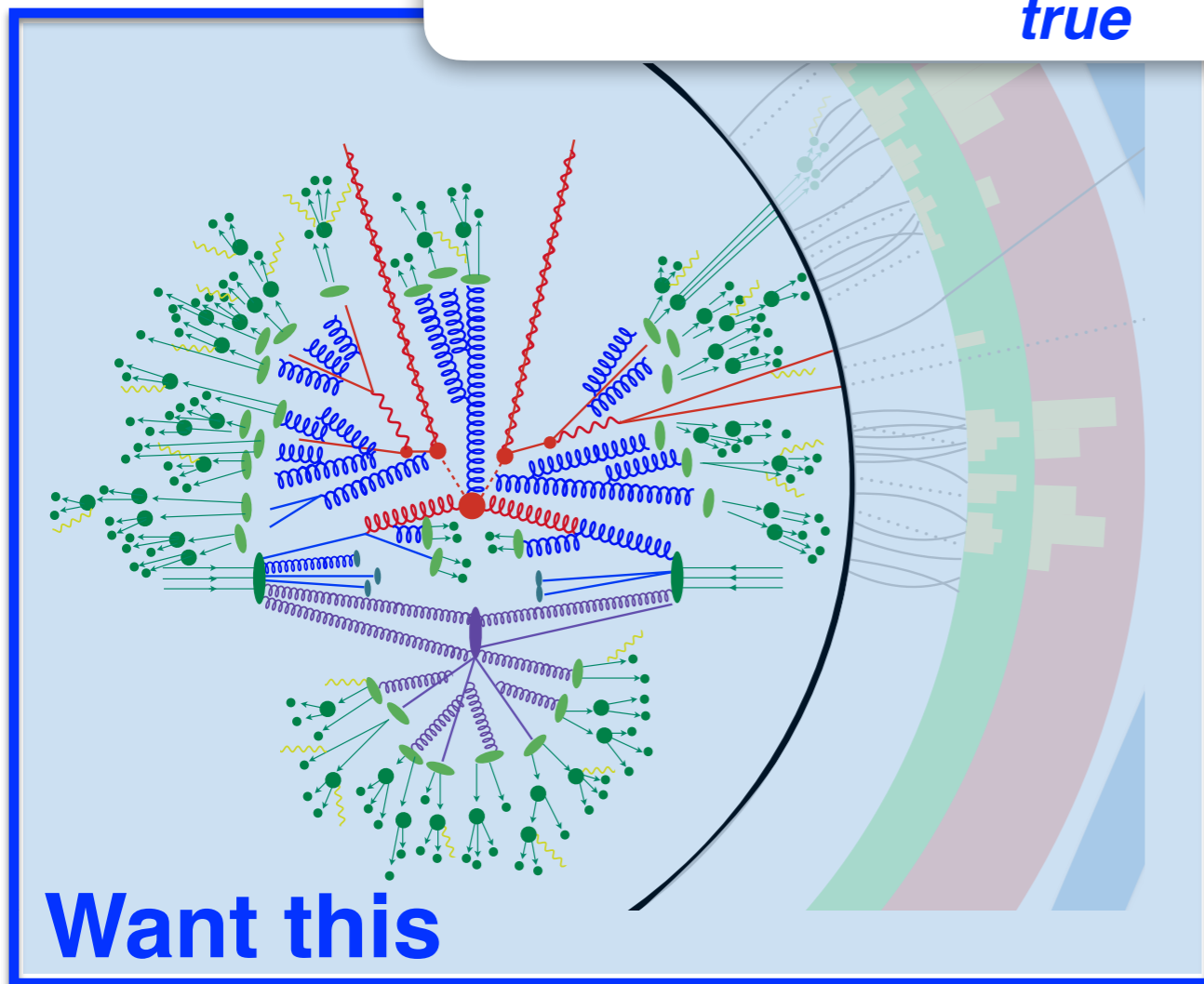
(or sometimes the parameters of the model that generate this)

# The Inference Challenge

18

If you know  $p(\textit{meas.} / \textit{true})$ , could do maximum likelihood, i.e.

$$\textit{unfolded} = \underset{\textit{true}}{\operatorname{argmax}} p(\textit{measured} / \textit{true})$$



(or sometimes the parameters of the model that generate this)

# The Inference Challenge

19

If you know  $p(\textit{meas.} \mid \textit{true})$ , could do maximum likelihood, i.e.

$$\textit{unfolded} = \underset{\textit{true}}{\operatorname{argmax}} p(\textit{measured} \mid \textit{true})$$



Challenge: **measured** is hyperspectral and **true** is hypervariate ...  $p(\textit{meas.} \mid \textit{true})$  is **intractable** !

# The Inference Challenge

20

If you know  $p(\textit{meas.} \mid \textit{true})$ , could do maximum likelihood, i.e.

$$\textit{unfolded} = \underset{\textit{true}}{\operatorname{argmax}} p(\textit{measured} \mid \textit{true})$$



Challenge: **measured** is hyperspectral and **true** is hypervariate ...  $p(\textit{meas.} \mid \textit{true})$  is **intractable** !

However: we have **simulators** that we can use to sample from  $p(\textit{meas.} \mid \textit{true})$

→ **Simulation-based (likelihood-free) inference** !

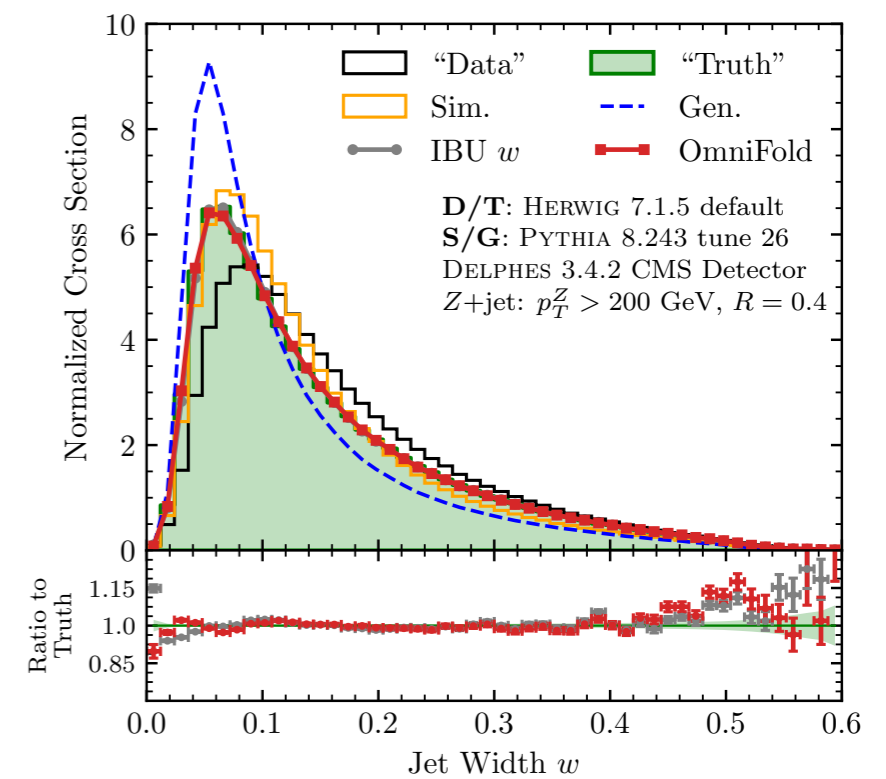
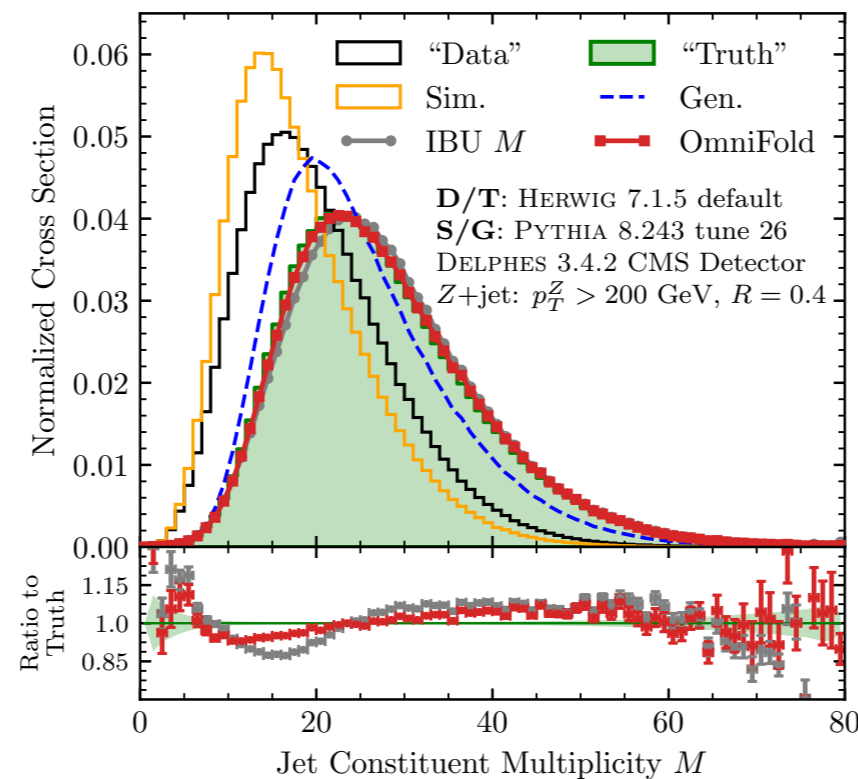
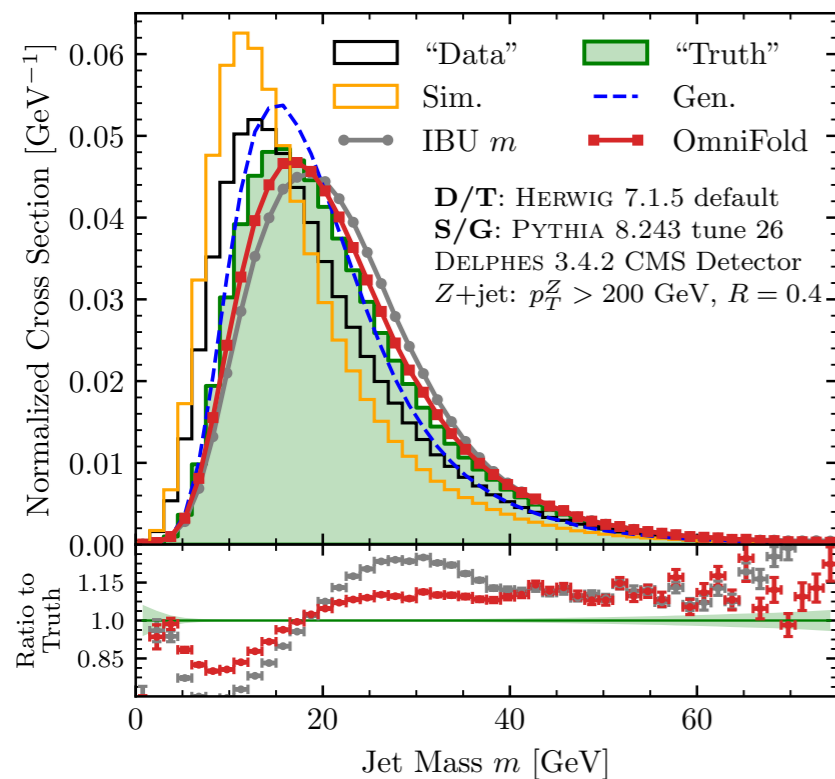
...an area of machine learning where particle physics is making key contributions!

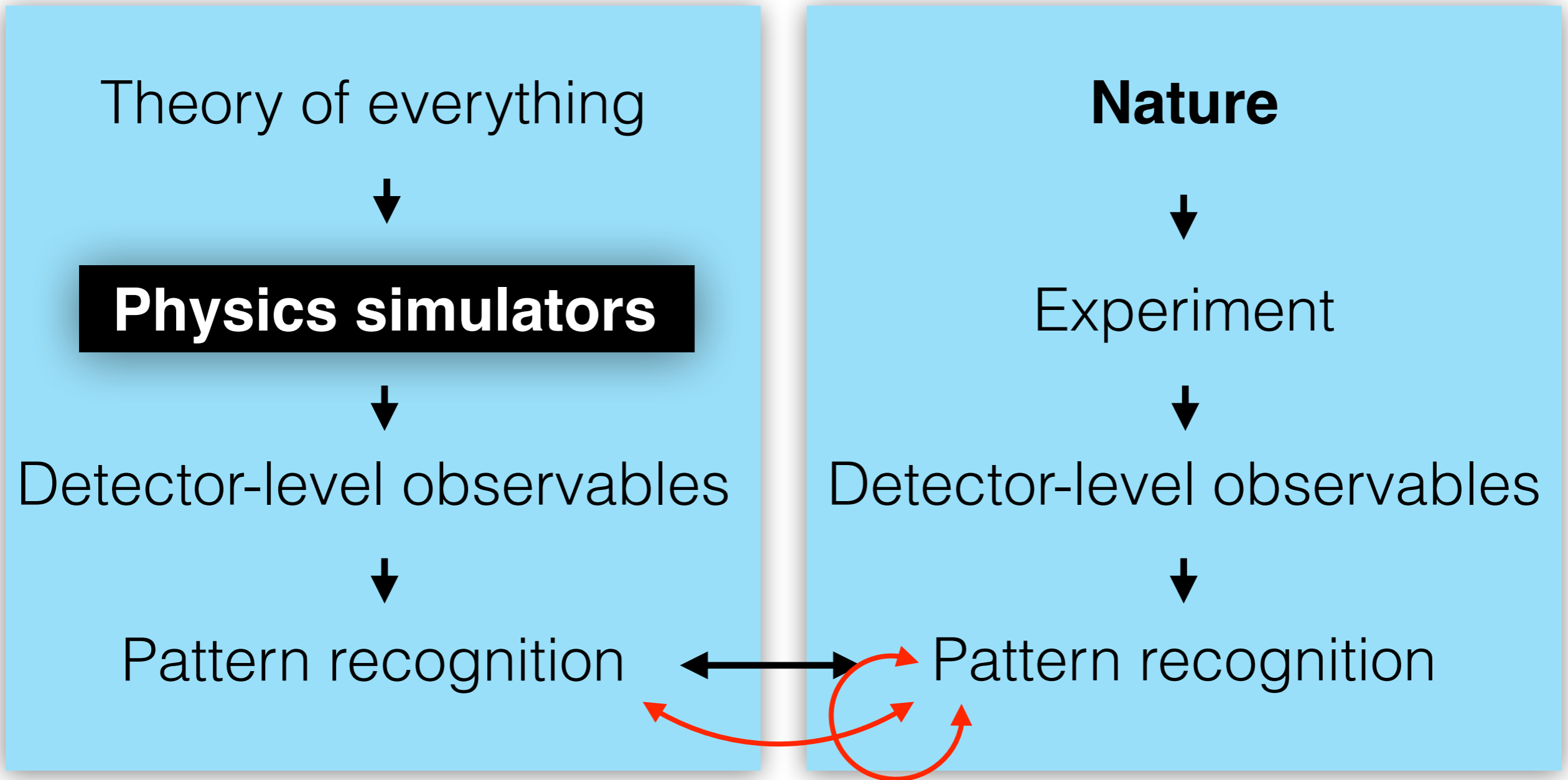
# Example: Unfolding

21

*What if we could unfold all particles simultaneously?  
We could then compute observables (and their bins)  
AFTER doing the measurement (!)*

*...stick around for the second part of this  
session for more discussions on this point*





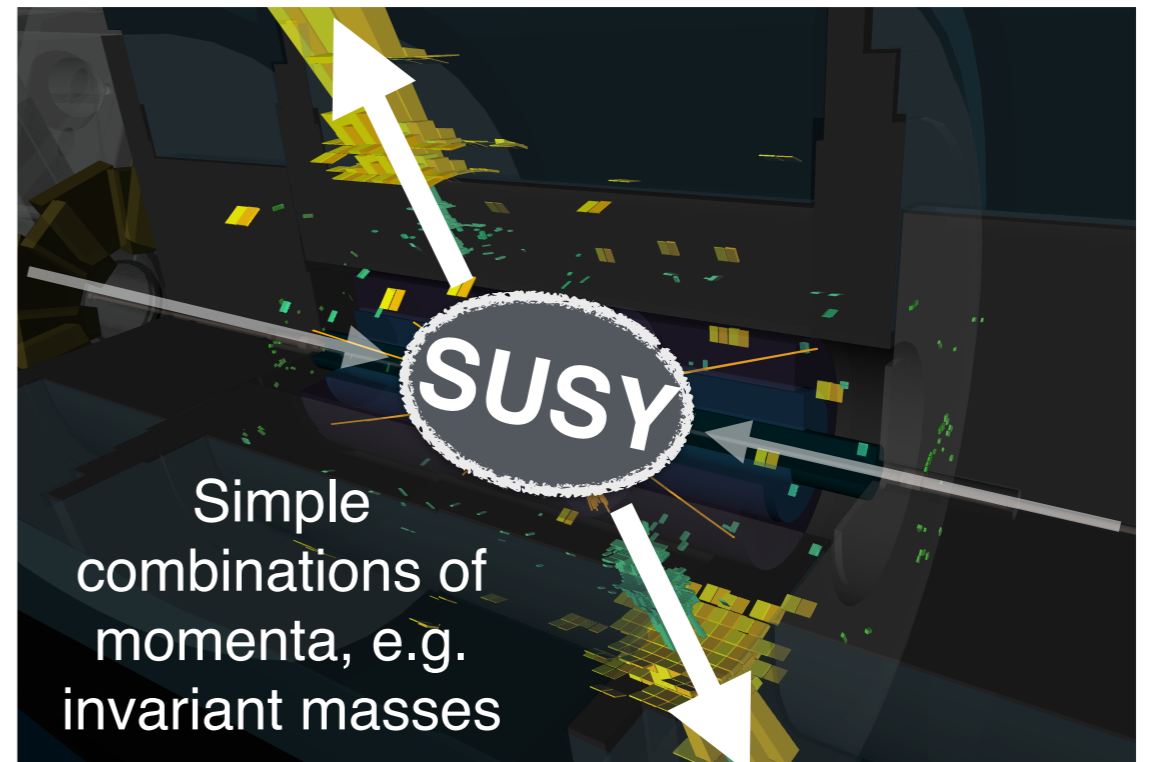
Anomaly detection

# Current Search Paradigm

23



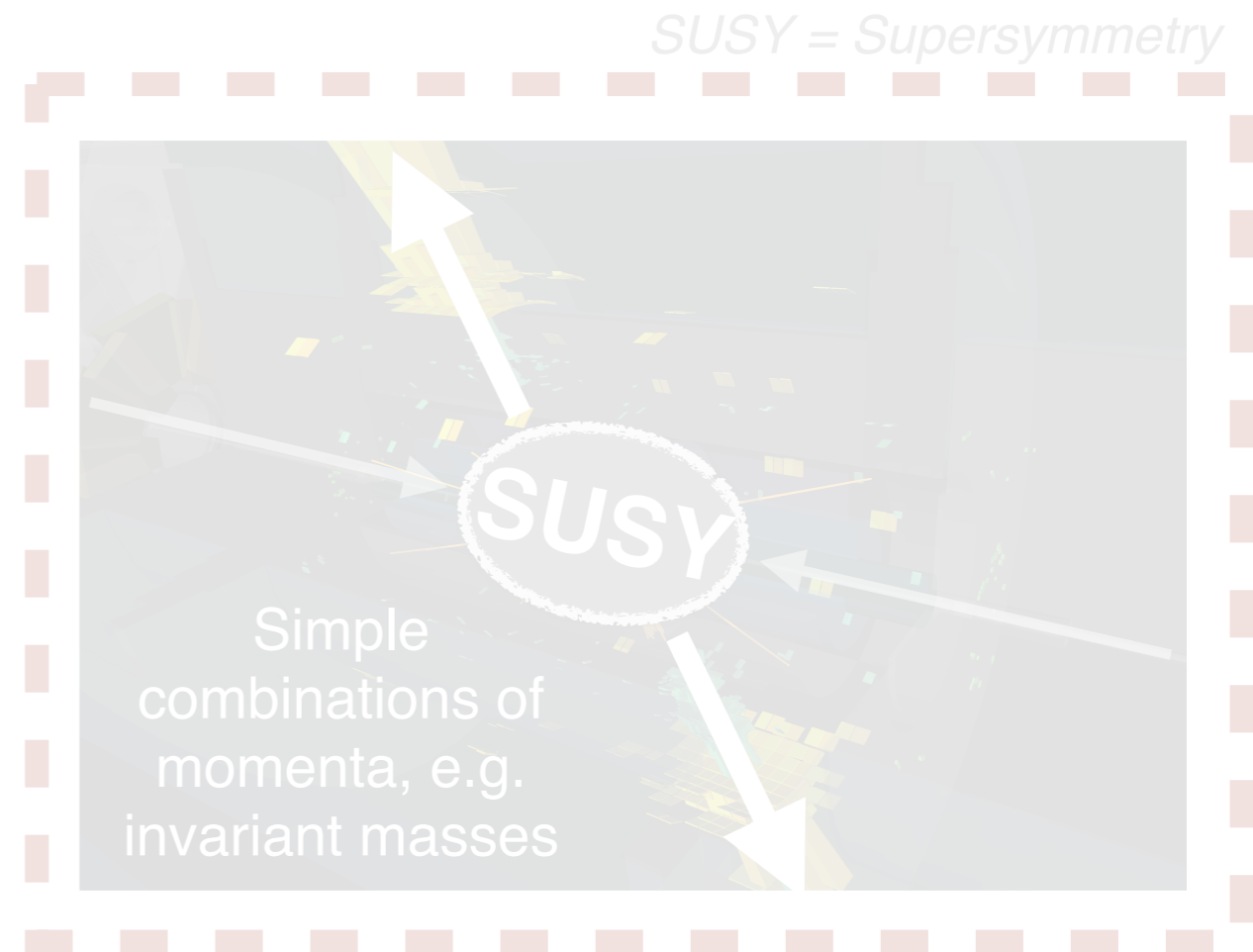
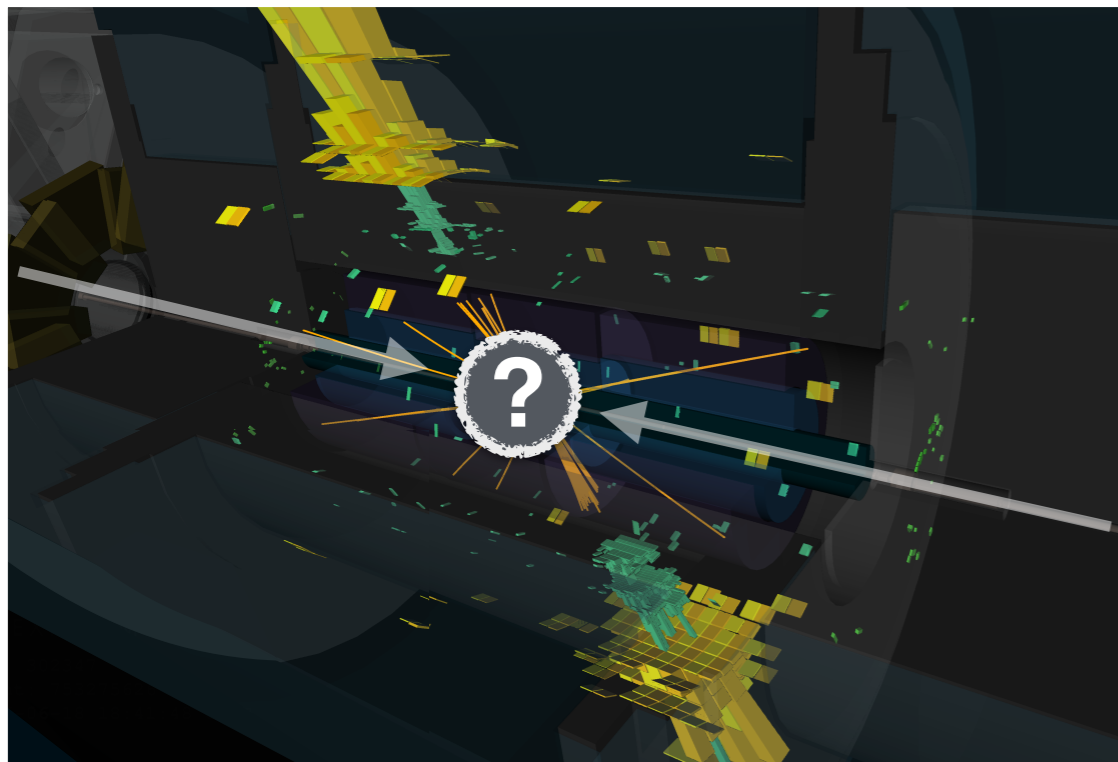
*SUSY = Supersymmetry*



(well-motivated) theory-biased  
& low-dimensional observables

# Current Search Paradigm

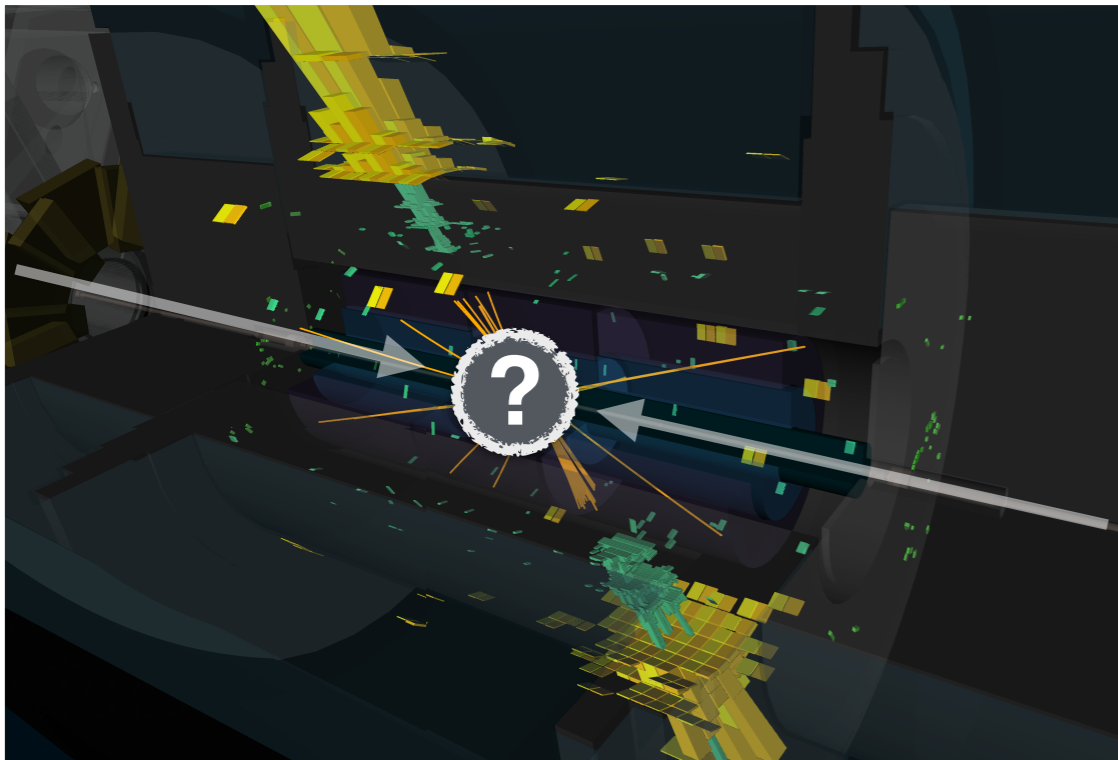
24



Can we relax model assumptions and explore high-dimensional feature spaces?

(well-motivated) theory-biased & low-dimensional observables





**What if we are not looking in the right place for the new phenomena?!**

Can we relax model assumptions and explore high-dimensional feature spaces?

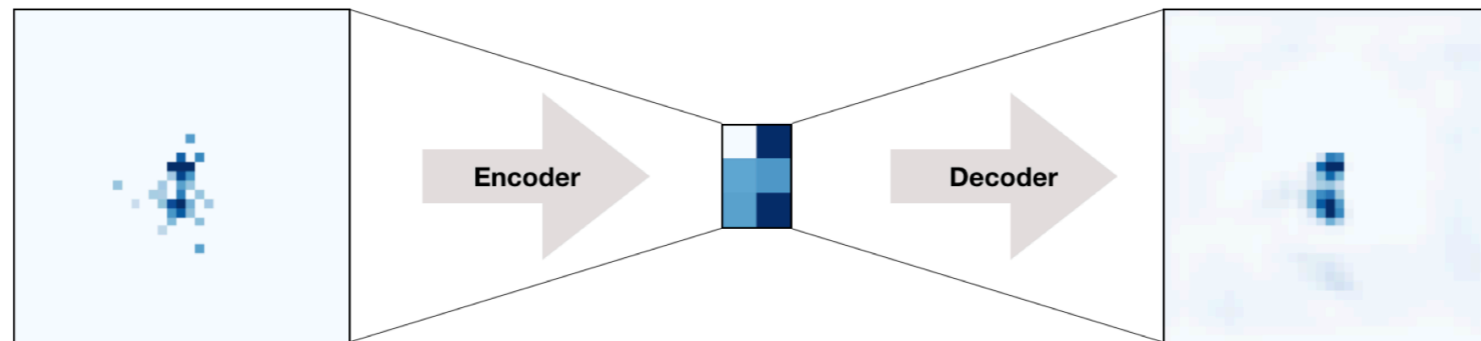
Supervision refers to the type of label information provided to the ML during training.

**Unsupervised** = no labels  
**Weakly-supervised** = noisy labels  
**Semi-supervised** = partial labels  
**Supervised** = full label information

These categories are not exact  
and the boundaries are not rigid!

**Unsupervised** = no labels

Typically, the goal of these methods is to look for events with low  $p(\text{background})$



One strategy (autoencoders) is to try to compress events and then uncompress them. When  $x = \text{uncompress}(\text{compress}(x))$ , then  $x$  probably has low  $p(x)$ .

**Weakly-supervised** = noisy labels

Typically, the goal of these methods is to look for events with high  $p(\textit{possibly signal-enriched})/p(\textit{possibly signal-depleted})$

e.g. Classification Without Labels (CWoLa), events in a signal region are labeled “signal” and events in a sideband are labeled “background”. These labels are “noisy” but a classifier trained with them can detect the presence of a signal.

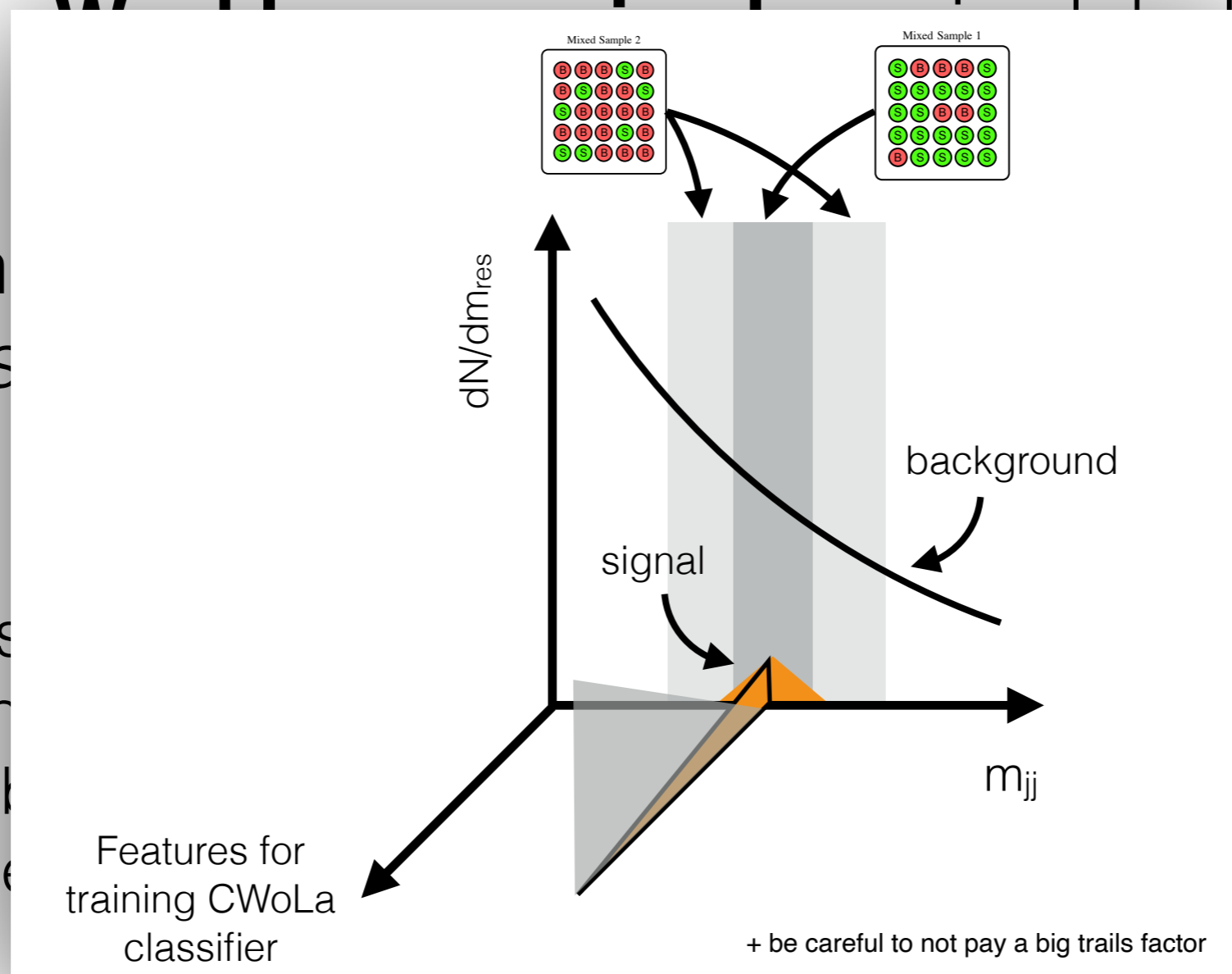
# Solutions: Weakly-supervised

Typically, the  
high  $p(\text{poss})$

or events with  
*signal-depleted*)

e.g. Clas  
region  
labeled “  
traine

in a signal  
band are  
a classifier  
signal.

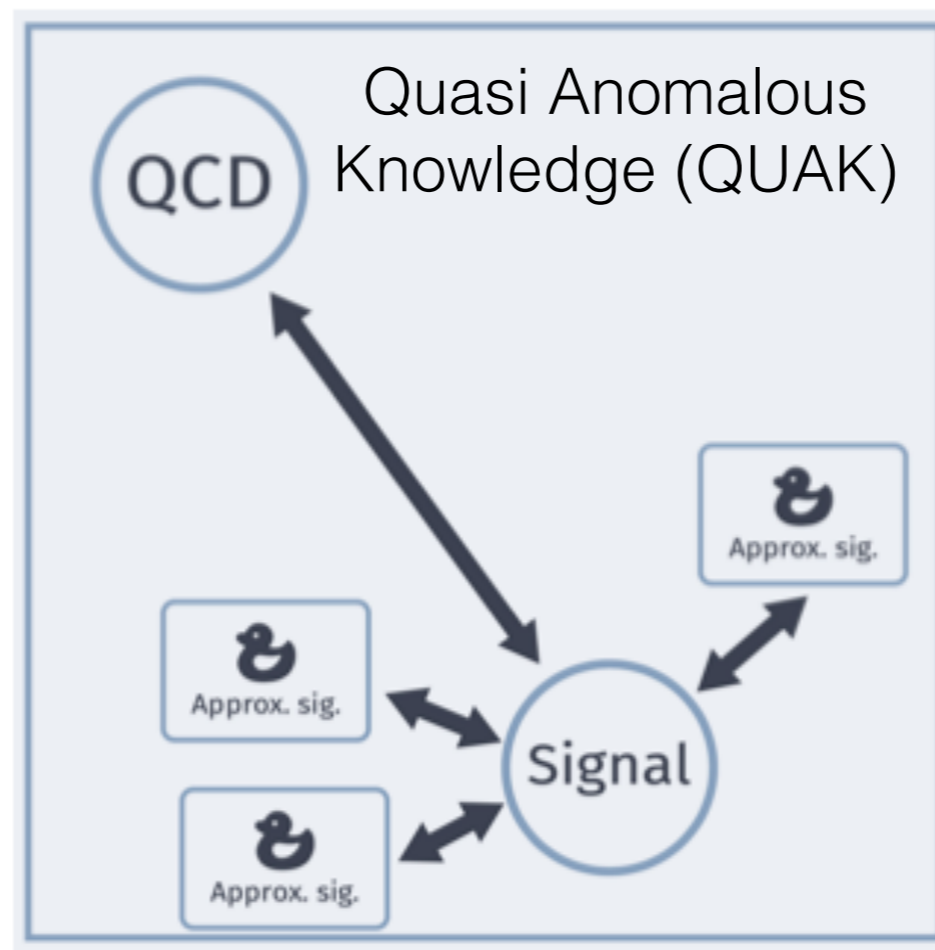


# Solutions: Semi-supervised

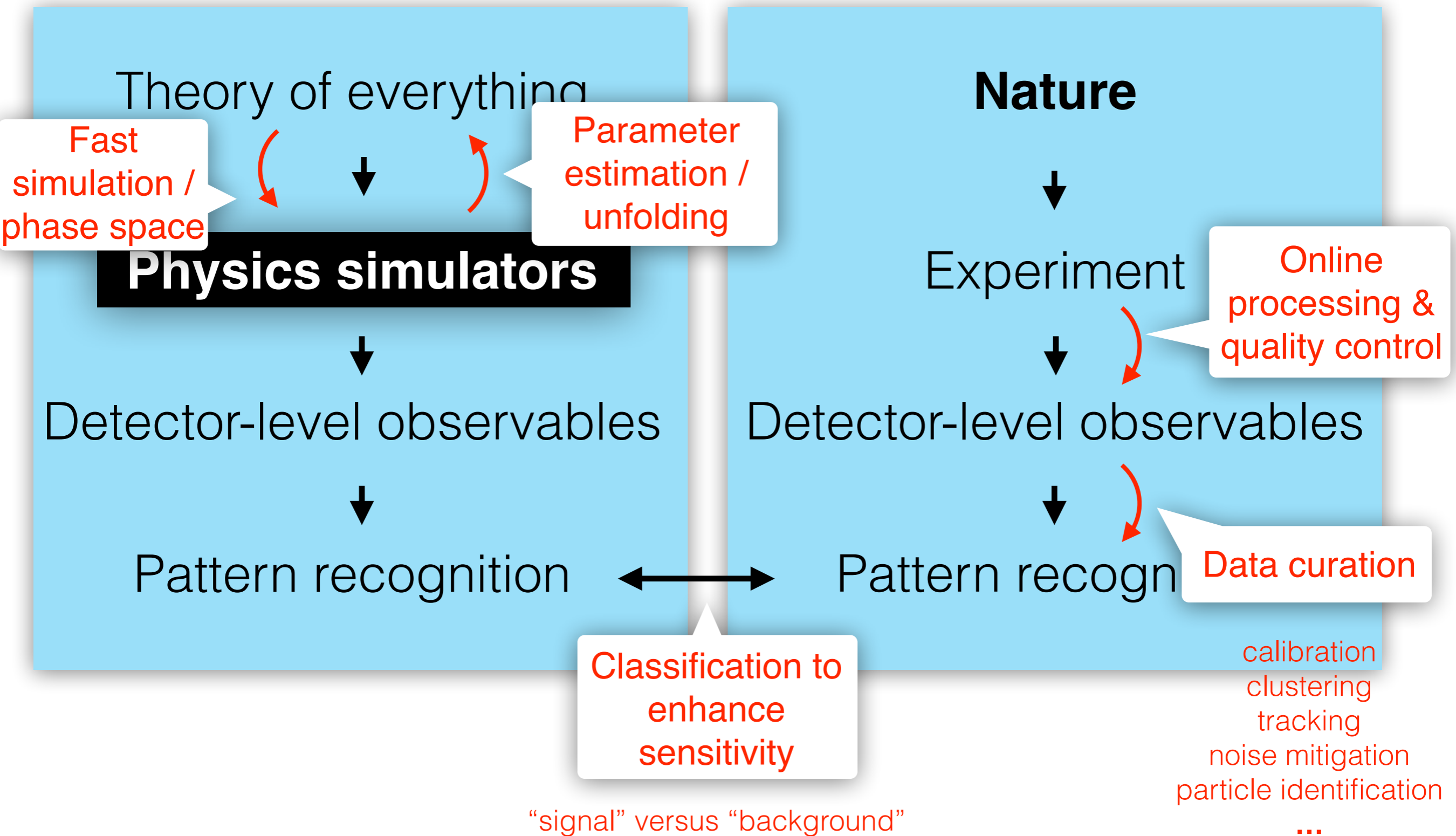
30

**Semi-supervised** = partial labels

Typically, these methods use some signal simulations to build signal sensitivity

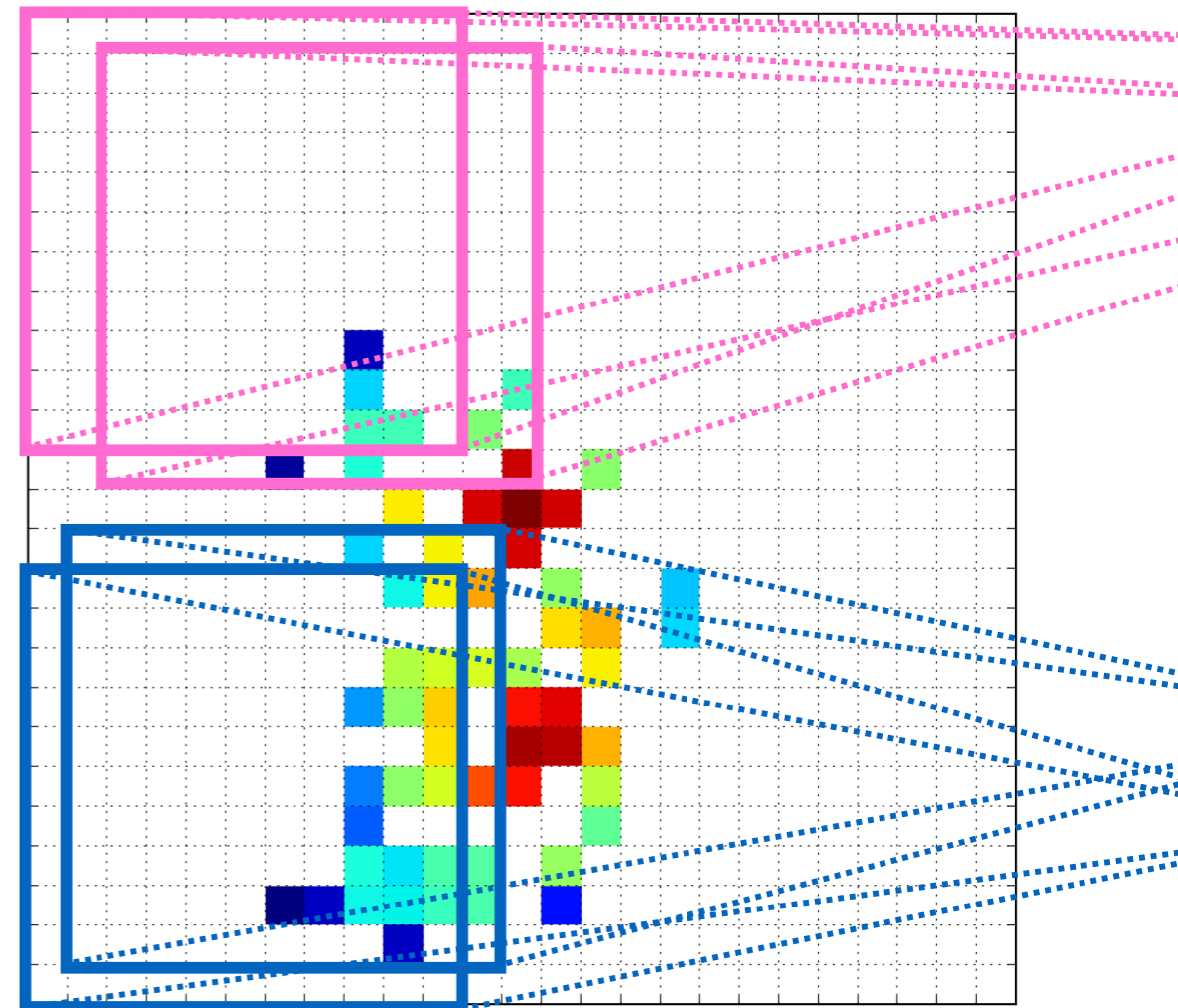


# Overview: Particle physics and ML



Deep learning has a great potential to **enhance**, **accelerate**, and **empower** discoveries in particle physics.

We have some unique challenges that require dedicated solutions.



With these new tools, we will be able to fully exploit the data in their natural high dimensionality enhancing the potential for **discovery!**



I only covered a small fraction of the exciting work from the hep-ph and hep-ex (friends in stat.ml++) communities!

*My apologies to everyone who's awesome plot(s) I could not show...*

enhancing the potential for **discovery!**

# Upcoming ML Workshops



**ML4Jets hybrid**  
July 6-8 2021

INSTITUTE FOR THEORETICAL PHYSICS



UNIVERSITÄT HEIDELBERG  
ZUKUNFT SEIT 1386

Local Organizers  
Anja Butter  
Barry Dillon  
Ullrich Köthe  
Tilman Plehn  
Hans-Christian Schultz-Coulon

International Organization Committee  
Kyle Cranmer (NYU)  
Ben Nachman (LBNL)  
Maurizio Pierini (CERN)  
Tilman Plehn (Heidelberg)  
Jesse Thaler (MIT)

<https://indico.cern.ch/event/980214>

Photo: Eyetonic / Fotolia / Adobe Stock; Composition Anke Heinzlmann



**MITP VIRTUAL WORKSHOP**

**Machine Learning for Particle Physics**  
21 June – 2 July 2021

 <https://indico.mitp.uni-mainz.de/event/199>

  
Mainz Institute for Theoretical Physics

# Where can I learn more?

35

## [HEPML-LivingReview](#)

### A Living Review of Machine Learning for Particle Physics

*Modern machine learning techniques, including deep learning, is rapidly being applied, adapted, and developed for high energy physics. The goal of this document is to provide a nearly comprehensive list of citations for those developing and applying these approaches to experimental, phenomenological, or theoretical analyses. As a living document, it will be updated as often as possible to incorporate the latest developments. A list of proper (unchanging) reviews can be found within. Papers are grouped into a small set of topics to be as useful as possible. Suggestions are most welcome.*

[download](#) [review](#)

The purpose of this note is to collect references for modern machine learning as applied to particle physics. A minimal number of categories is chosen in order to be as useful as possible. Note that papers may be referenced in more than one category. The fact that a paper is listed in this document does not endorse or validate its content - that is for the community (and for peer-review) to decide. Furthermore, the classification here is a best attempt and may have flaws - please let us know if (a) we have missed a paper you think should be included, (b) a paper has been misclassified, or (c) a citation for a paper is not correct or if the journal information is now available. In order to be as useful as possible, this document will continue to evolve so please check back before you write your next paper. If you find this review helpful, please consider citing it using `\cite{hepmlivingreview}` in HEPML.bib.

<https://iml-wg.github.io/HEPML-LivingReview/>

<https://github.com/iml-wg/HEPML-LivingReview>

<https://arxiv.org/abs/2102.02770>

Questions?



# Unbinned differential cross section measurements: *towards a common format*

Benjamin Nachman

*Lawrence Berkeley National Laboratory*

[bpnachman.com](http://bpnachman.com)

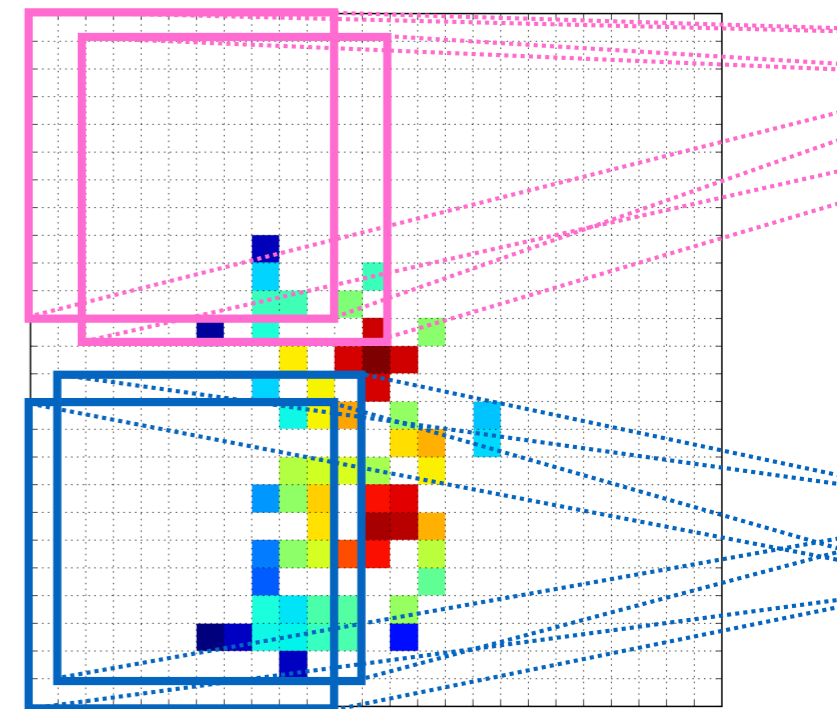
[bpnachman@lbl.gov](mailto:bpnachman@lbl.gov)



@bpnachman



bnachman

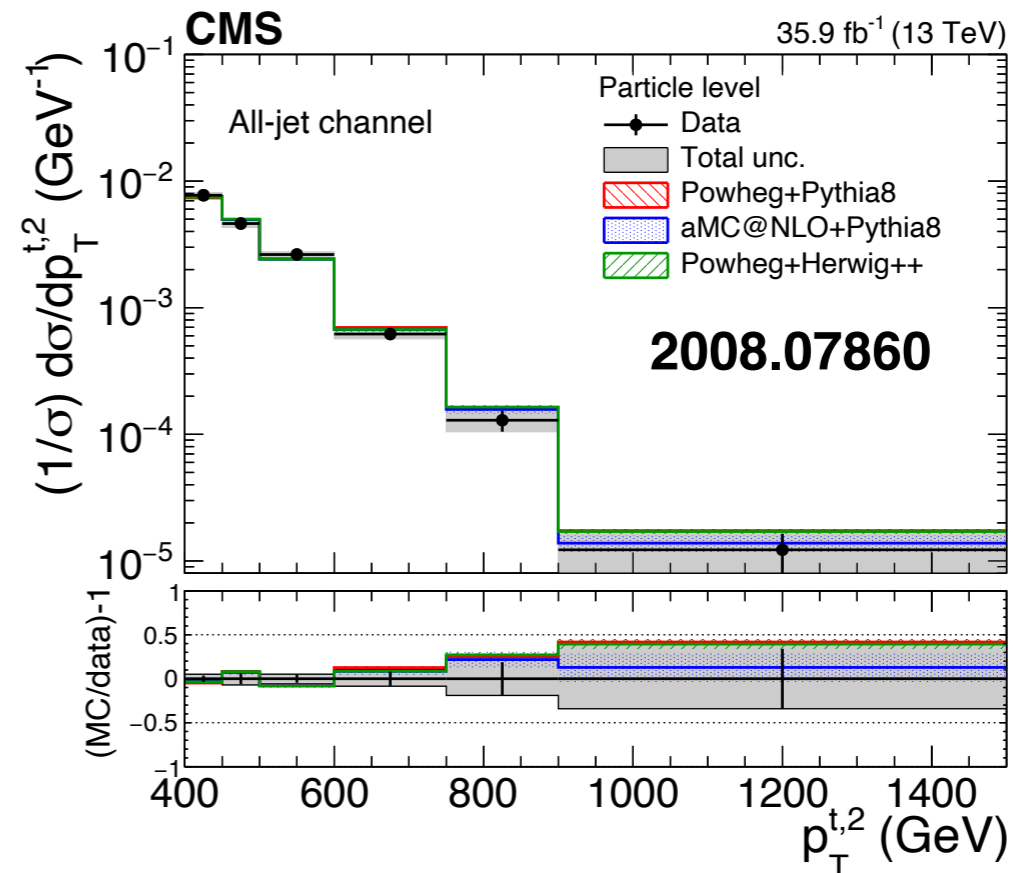
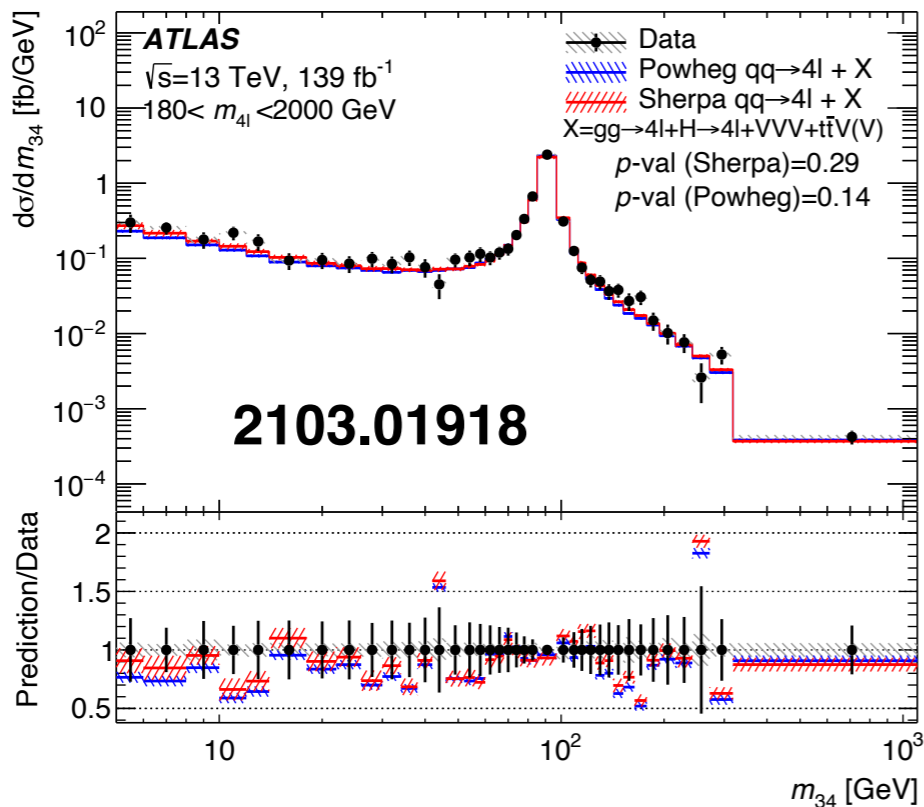
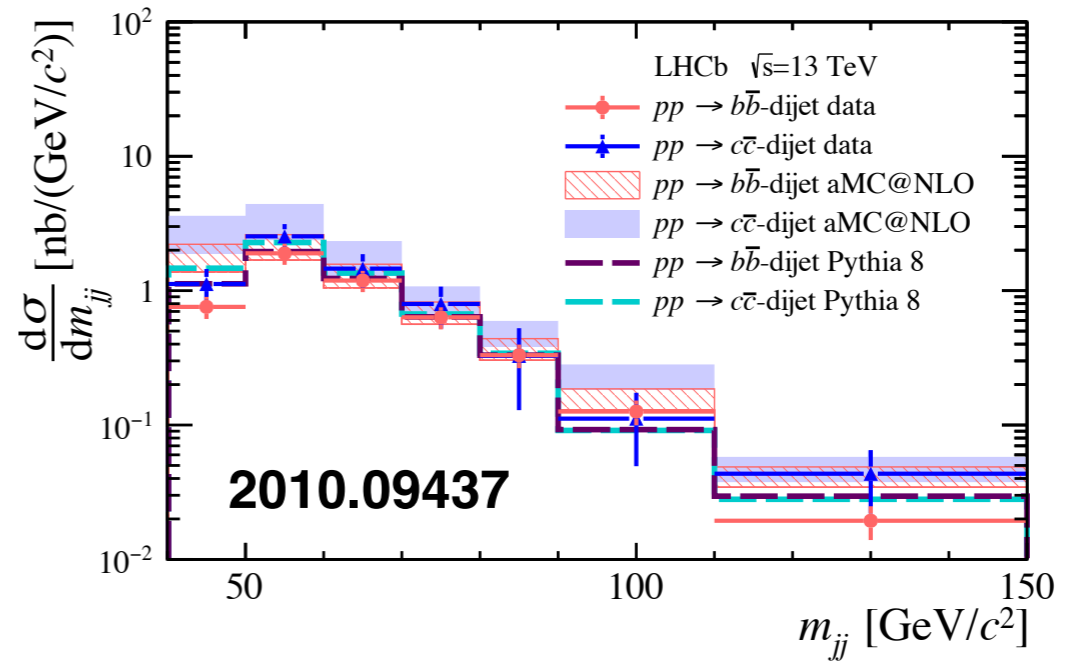
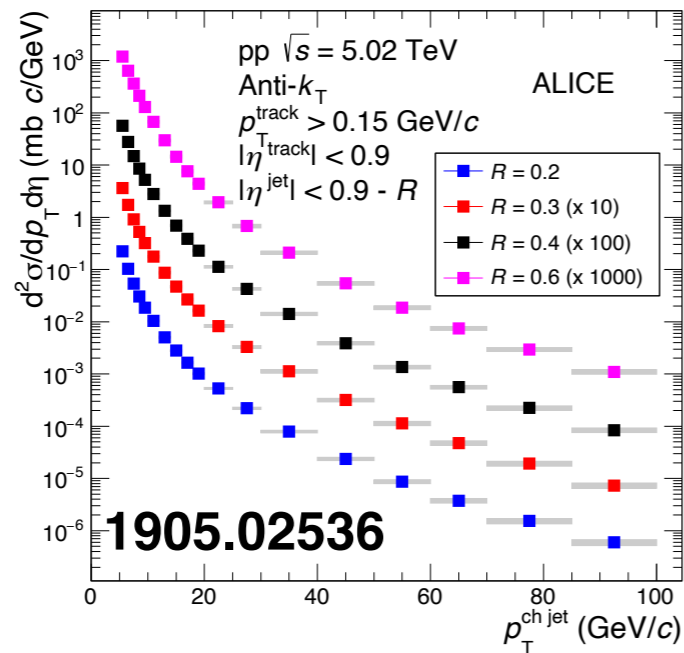


Virtual Houches

June 17, 2021

# Differential Cross Section Measurements

38



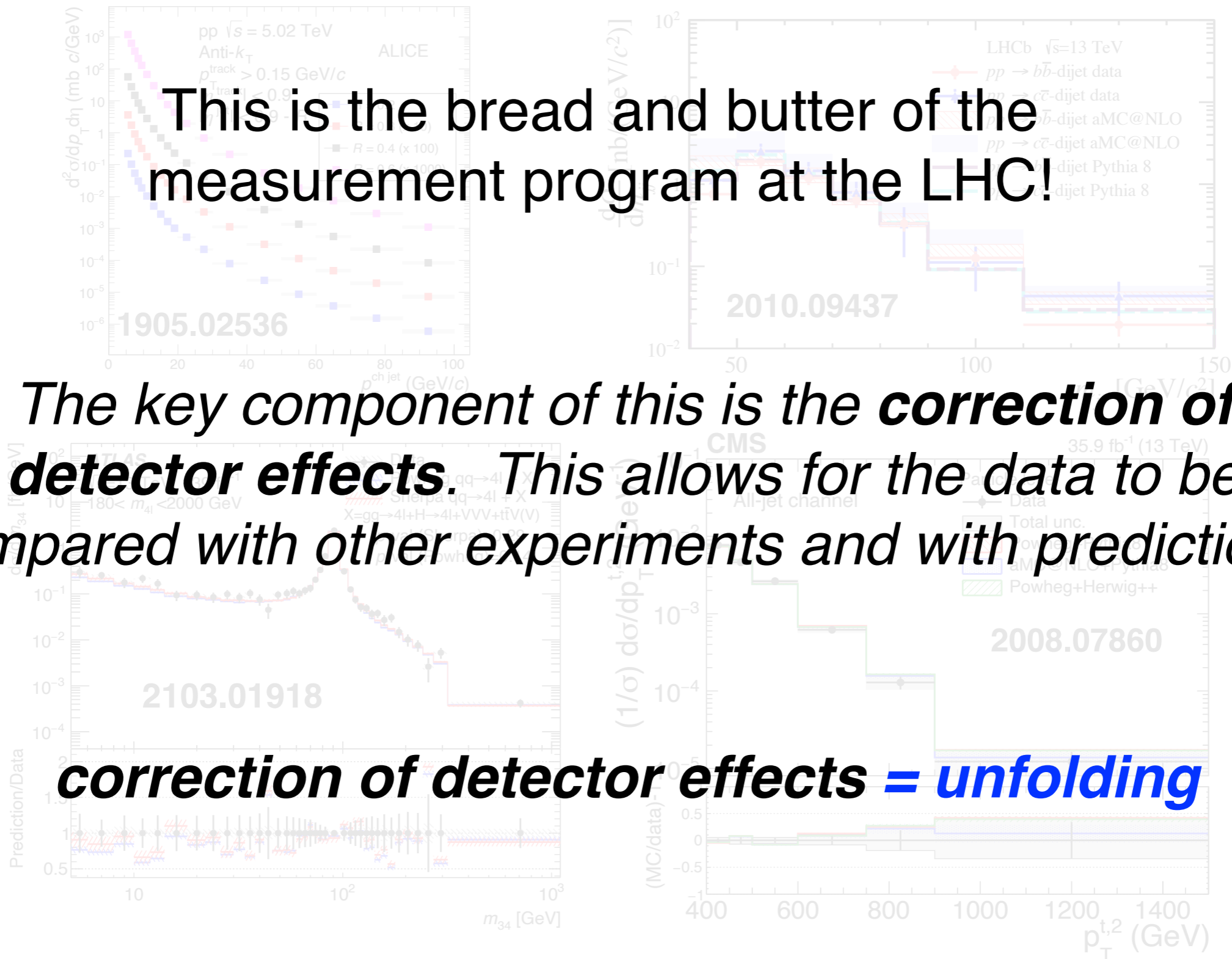
# Differential Cross Section Measurements

39

This is the bread and butter of the measurement program at the LHC!

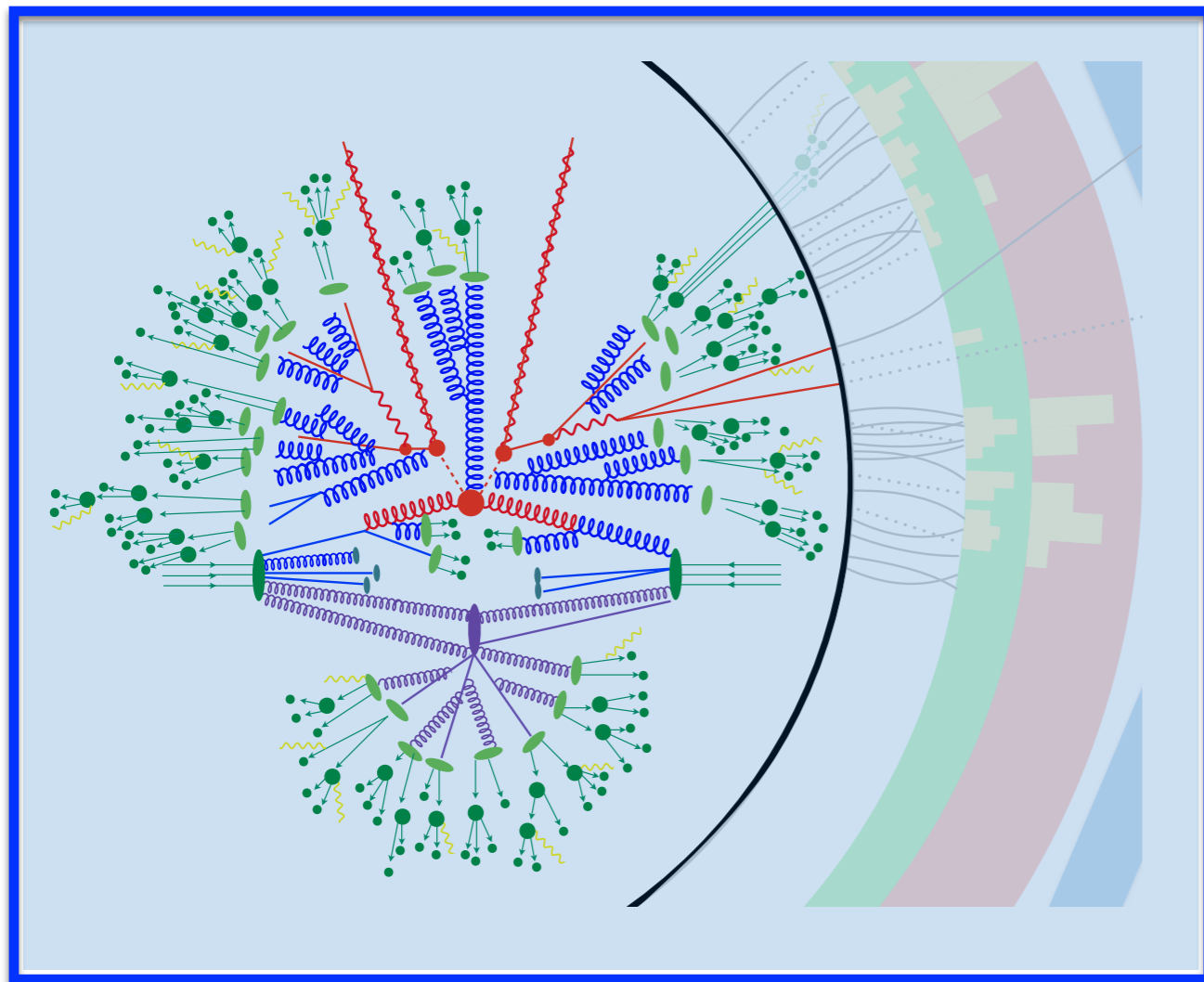
The key component of this is the **correction of detector effects**. This allows for the data to be compared with other experiments and with predictions.

**correction of detector effects = unfolding**

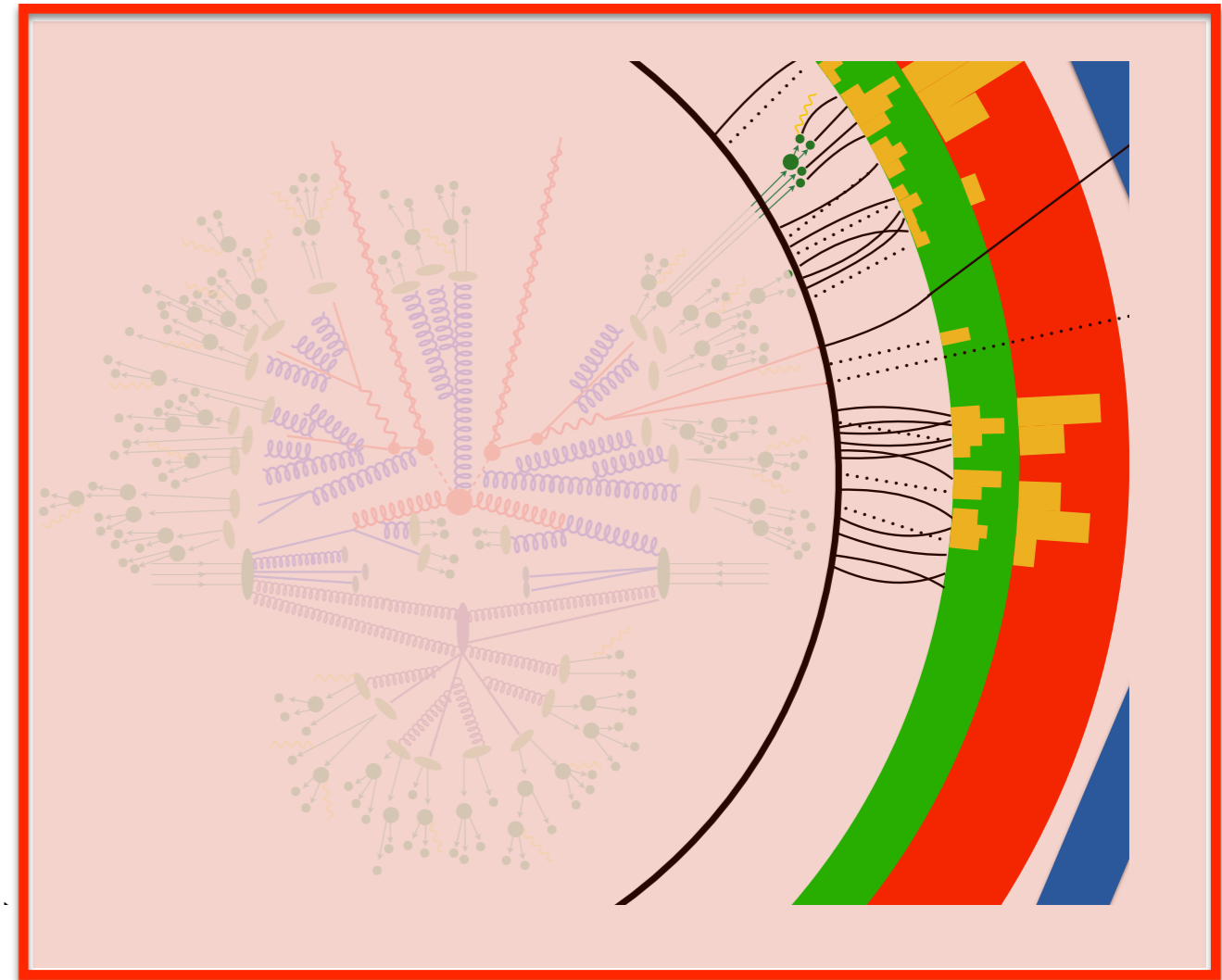


# The Unfolding Challenge

Want this



Measure this





# The Unfolding Challenge

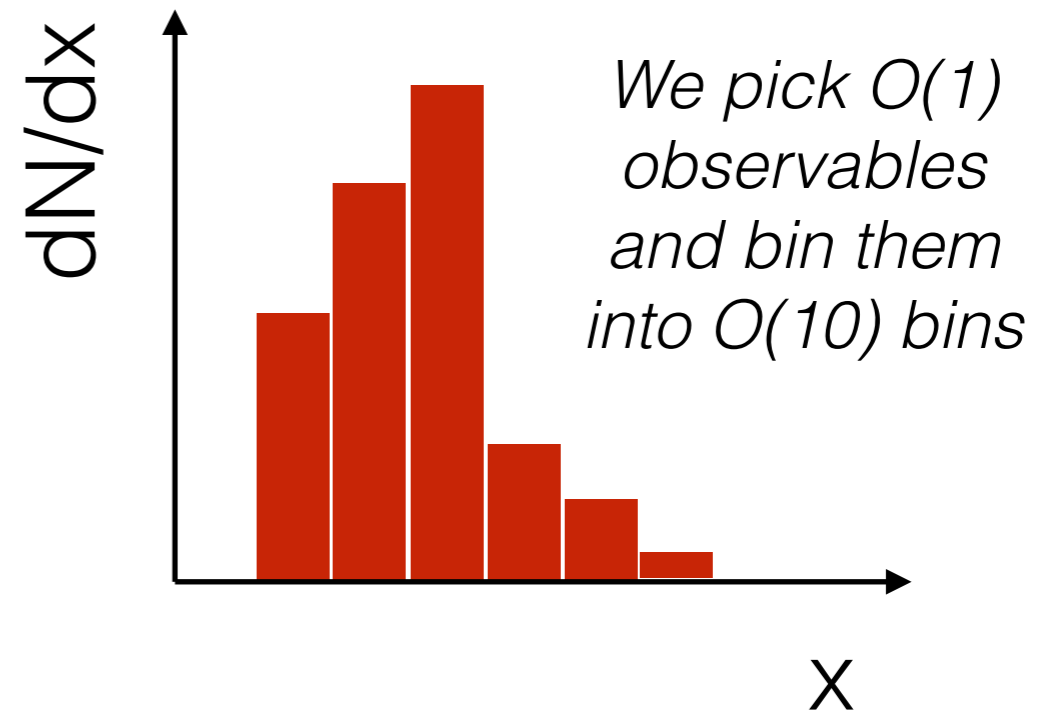
41

Want this



Measure this

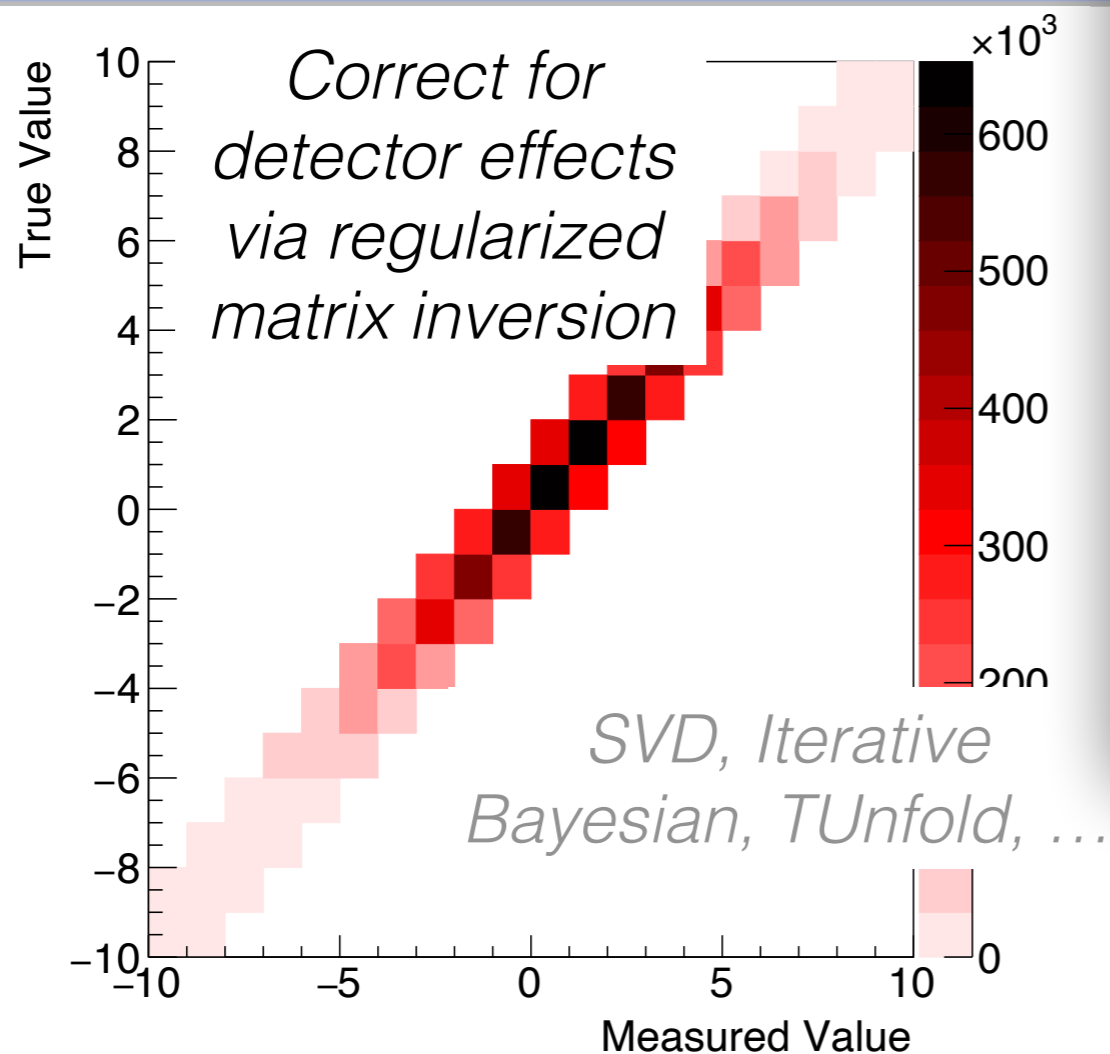
Usual solution:



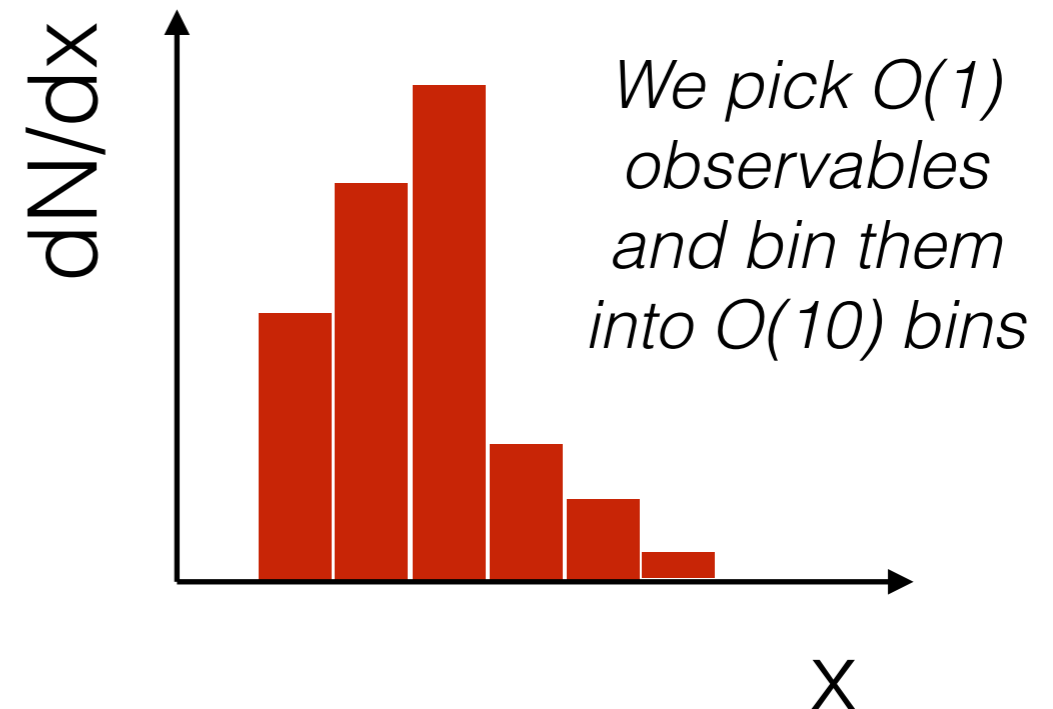
# The Unfolding Challenge

42

Want this



Usual solution:



# The Unfolding Challenge

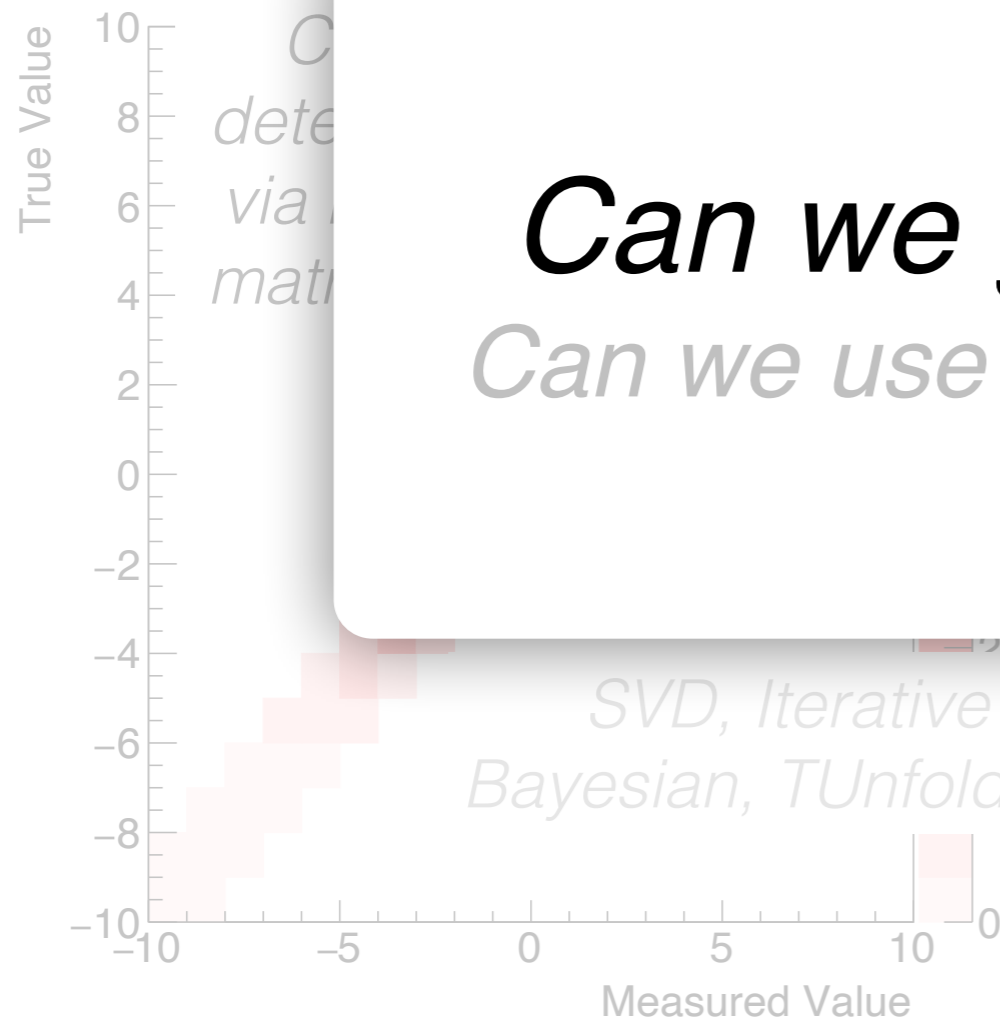
43

Want this



Usual solution:

***Can we go unbinned?***  
*Can we use many dimensions?*



*SVD, Iterative  
Bayesian, TUnfold, ...*

Measure this

*pick  $O(1)$   
observables  
bin them  
 $O(10)$  bins*

X



There were some early proposals for unbinned unfolding\* but as far as I am aware, they were not used for any measurements.

However, recent innovations in machine learning and resulted in new methods for unbinned unfolding, which are being used for data analysis+ (!)

**The goal of this discussion is to propose a common way for publishing unbinned results to maximize their science potential**

We need input from both experimentalist and theorists (!)

\*see L. Lindemann and G. Zech, NIM A 354 (1995) 516 & related

+see <https://www-h1.desy.de/h1/www/publications/htmlsplit/H1prelim-21-031.long.html>

# Publishing Binned Results

45

How do we publish binned results?

The screenshot displays the HEPData website interface. At the top, the HEPData logo is centered, with the tagline "Repository for publication-related High-Energy Physics data" below it. A search bar is present, with the text "Search on 9427 publications and 96116 data tables." above it. The search bar contains the placeholder text "Search for a paper, author, experiment, reaction" and a "Search" button. To the right of the search bar is an "Advanced" link. Below the search bar, an example search query is provided: "e.g. reaction  $P P \rightarrow L Q L Q X$ , title has 'photon collisions', collaboration is LHCf or D0." Below the search bar, the section "Data from the LHC" is displayed, featuring four icons representing different experiments: ATLAS, ALICE, CMS, and LHCb. Each icon is accompanied by a "View Data" button.

HEPData  
Repository for publication-related High-Energy Physics data

Search on 9427 publications and 96116 data tables.

Search for a paper, author, experiment, reaction Search Advanced

e.g. reaction  $P P \rightarrow L Q L Q X$ , title has "photon collisions", collaboration is LHCf or D0.

Data from the LHC

ATLAS View Data

ALICE View Data

CMS View Data

LHCb View Data

# Publishing Binned Results

46

◀ Hide Publication Information

Properties of  $g \rightarrow b\bar{b}$  at small opening angles in  $pp$  collisions with the ATLAS detector at  $\sqrt{s} = 13$  TeV

The ATLAS collaboration

Aaboud, Morad , Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abidinov, Ovsat , Abeloos, Baptiste , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola

Phys.Rev.D 99 (2019) 052004, 2019.

<https://doi.org/10.17182/hepdata.85697>

Journal INSPIRE Resources

Rivet Analysis

Abstract (data abstract)  
CERN-LHC.

The fragmentation of high energy gluons at small opening angles is largely unconstrained by present measurements. Gluon splitting to b-quark pairs is a unique probe into the properties of gluon fragmentation because identified b-tagged jets provide a proxy for the quark daughters of the initial gluon. In this study, key differential distributions related to the  $g \rightarrow b\bar{b}$  process are measured using 33/fb of  $\sqrt{s}=13$  TeV  $pp$  collision data recorded by the ATLAS experiment at the LHC in 2016. Jets constructed from charged-particle tracks, clustered with the jet anti-kt algorithm with radius parameter  $R = 0.2$ , are used to probe angular scales below the  $R = 0.4$  jet radius. The observables are unfolded to particle level in order to facilitate direct

Download All

View Analyses

Filter 4 data tables

## Table 1

Data from Figure 6A  
10.17182/hepdata.85697.v1/t1  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$ , as a function of  $\Delta R(b, b)$  - the angle in  $\eta$  and  $\phi$  between...

## Table 2

Data from Figure 6B  
10.17182/hepdata.85697.v1/t2  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta\theta_{gpp,gb}/\pi$ , the angle between production (gpp) and decay (gbb) planes ( $\Delta\theta_{gpp,gb}$ ).

## Table 3

Data from Figure 6C  
10.17182/hepdata.85697.v1/t3  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/dz(p_T)$ , as a function of  $z(p_T) = p_{T,2}/(p_{T,1} + p_{T,2})$ .

## Table 4

Data from Figure 6D  
10.17182/hepdata.85697.v1/t4  
Normalized differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\log(m_{bb}/p_T)$ , as a function of  $\log(m_{bb}/p_T)$  for  $m_{bb}$  the invariant mass of the two b-jets.

## Table 1

Data from Figure 6A

Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$ , as a function of  $\Delta R(b, b)$  - the angle in  $\eta$  and  $\phi$  between the two b-tagged jets.

cmenergies

13000.0

observables

DSIG/DDR

<https://www.hepdata.net/rec>

JSON

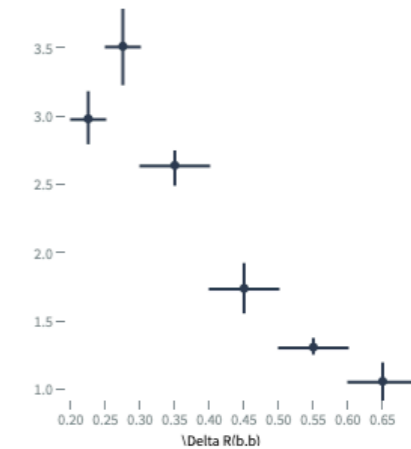
reactions

PP -> g + X

g -> b + b

RE	PP -> g < b b > X
SQRT(S)	13000.0 GEV
$\Delta R(b, b)$	$(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$
0.2 - 0.25	2.98 $\pm 0.0016$ stat $^{+0.076}_{-0.019}$ sys,Calorimeterjetenergy $^{+0.041}_{-0.042}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.25 - 0.3	3.51 $\pm 0.0019$ stat $^{+0.062}_{-0.1}$ sys,Calorimeterjetenergy $^{+0.034}_{-0.034}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.3 - 0.4	2.64 $\pm 0.0017$ stat $^{+0.02}_{-0.089}$ sys,Calorimeterjetenergy $^{+0.021}_{-0.018}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.4 - 0.5	1.74 $\pm 0.0018$ stat $^{+0.039}_{-0.024}$ sys,Calorimeterjetenergy $^{+0.0092}_{-0.0094}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.5 - 0.6	1.31 $\pm 0.0013$ stat $^{+0.036}_{-0.067}$ sys,Calorimeterjetenergy $^{+0.013}_{-0.013}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.6 - 0.7	1.06 $\pm 0.0013$ stat $^{+0.018}_{-0.026}$ sys,Calorimeterjetenergy $^{+0.014}_{-0.017}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>

Visualize



Sum errors  Log Scale (X)  Log Scale (Y)

# Publishing Binned Results

47

[Hide Publication Information](#)  
**Properties of  $g \rightarrow b\bar{b}$  at small opening angles in  $pp$  collisions with the ATLAS detector at  $\sqrt{s} = 13$  TeV**  
 The ATLAS collaboration  
 Aaboud, Morad , Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abidinov, Ovsat , Abeloos, Baptiste , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola  
**Phys.Rev.D 99 (2019) 052004, 2019.**  
<https://doi.org/10.17182/hepdata.85697>

[Journal](#) [INSPIRE](#) [Resources](#)  
[Rivet Analysis](#)

**Abstract (data abstract)**  
 CERN-LHC.  
 The fragmentation of high energy gluons at small opening angles is largely unconstrained by present measurements. Gluon splitting to b-quark pairs is a unique probe into the properties of gluon fragmentation because identified b-tagged jets provide a proxy for the quark daughters of the initial gluon. In this study, key differential distributions related to the  $g \rightarrow b\bar{b}$  process are measured using 33/fb of  $\sqrt{s}=13$  TeV pp collision data recorded by the ATLAS experiment at the LHC in 2016. Jets constructed from charged-particle tracks, clustered with the jet anti-k<sub>t</sub> algorithm with radius parameter R = 0.2, are used to probe angular scales below the R = 0.4 jet radius. The observables are unfolded to particle level in order to facilitate direct

[Download All](#)  
[View Analyses](#)  
 Filter 4 data tables

**Table 1** [10.17182/hepdata.85697.v1/t1](#)  
 Data from Figure 6A  
 Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$ , as a function of  $\Delta R(b, b)$  - the angle in  $\eta$  and  $\phi$  between...

**Table 2**  
 Data from Figure 6B  
 Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta\theta_{gpp,gb}/\pi$ , the angle between production (gpp) and decay (gb) planes ( $\Delta\theta_{gpp,gb}$ ).

**Table 3**  
 Data from Figure 6C  
 Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/dz(p_T)$ , as a function of  $z(p_T) = p_{T,2}/(p_{T,1} + p_{T,2})$ .

**Table 4**  
 Data from Figure 6D  
 Normalized differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d \log(m_{bb}/p_T)$ , as a function of  $\log(m_{bb}/p_T)$  for  $m_{bb}$  the invariant mass of the two b-jets.

**Table 1** [10.17182/hepdata.85697.v1/t1](#)  
 Data from Figure 6A  
 Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$ , as a function of  $\Delta R(b, b)$  - the angle in  $\eta$  and  $\phi$  between the two b-tagged jets.

**cmenergies** 13000.0  
**observables** DSIG/DDR  
**reactions** P P -> g + X, g -> b + b

RE	P P -> g < b b > X
<b>SQRT(S)</b>	13000.0 GEV
<b><math>\Delta R(b, b)</math></b>	$(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$
0.2 - 0.25	2.98 $\pm$ 0.0016 stat $^{+0.076}_{-0.019}$ sys,Calorimeterjetenergy $^{+0.041}_{-0.042}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.25 - 0.3	3.51 $\pm$ 0.0019 stat $^{+0.062}_{-0.1}$ sys,Calorimeterjetenergy $^{+0.034}_{-0.034}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.3 - 0.4	2.64 $\pm$ 0.0017 stat $^{+0.02}_{-0.089}$ sys,Calorimeterjetenergy $^{+0.021}_{-0.018}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.4 - 0.5	1.74 $\pm$ 0.0018 stat $^{+0.039}_{-0.024}$ sys,Calorimeterjetenergy $^{+0.0092}_{-0.0094}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.5 - 0.6	1.31 $\pm$ 0.0013 stat $^{+0.036}_{-0.067}$ sys,Calorimeterjetenergy $^{+0.013}_{-0.013}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>
0.6 - 0.7	1.06 $\pm$ 0.0013 stat $^{+0.018}_{-0.026}$ sys,Calorimeterjetenergy $^{+0.014}_{-0.017}$ sys,Flavor tagging + 4 more errors <a href="#">Show all</a>

**Visualize**  
  
 Sum errors  Log Scale (X)  Log Scale (Y)

YAML with resource files

YAML

YODA

ROOT

CSV

# Publishing Binned Results

Properties of  $g \rightarrow b\bar{b}$  at small opening angles in  $pp$  collisions with the ATLAS detector at  $\sqrt{s} = 13$  TeV

The ATLAS collaboration

Aaboud, Morad , Aad, Georges , Abbott, Brad , Abbott, Dale Charles , Abidinov, Ovsat , Abeloos, Baptiste , Abhayasinghe, Deshan Kavishka , Abidi, Syed Haider , Abouzeid, Ossama , Abraham, Nicola

Phys.Rev.D 99 (2019) 052004, 2019.

https://doi.org/10.17182/hepdata.85697

Journal INSPIRE Resources

Rivet Analysis

Abstract (data abstract)  
CERN-LHC.

The fragmentation of high energy gluons at small opening angles is largely unconstrained by present measurements. Gluon splitting to b-quark pairs is a unique probe into the properties of gluon fragmentation because identified b-tagged jets provide a proxy for the quark daughters of the initial gluon. In this study, key differential distributions related to the  $g \rightarrow b\bar{b}$  process are measured using 33/fb of  $\sqrt{s}=13$  TeV  $pp$  collision data recorded by the ATLAS experiment at the LHC in 2016. Jets constructed from charged-particle tracks, clustered with the jet anti-kt algorithm with radius parameter  $R = 0.2$ , are used to probe angular scales below the  $R = 0.4$  jet radius. The observables are unfolded to particle level in order to facilitate direct

Download All

View Analyses

Filter 4 data tables

Table 1  
Data from Figure 6A  
10.17182/hepdata.85697.v1/t1  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta R(b, b)$ , as a function of  $\Delta R(b, b)$  - the angle in  $\eta$  and  $\phi$  between...

Table 2  
Data from Figure 6B  
10.17182/hepdata.85697.v1/t2  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\Delta\theta_{gpp,gb}/\pi$ , the angle between production (gpp) and decay (gb) planes ( $\Delta\theta_{gpp,gb}$ ).

Table 3  
Data from Figure 6C  
10.17182/hepdata.85697.v1/t3  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/dz(p_T)$ , as a function of  $z(p_T) = p_{T,2}/(p_{T,1} + p_{T,2})$ .

Table 4  
Data from Figure 6D  
10.17182/hepdata.85697.v1/t4  
Normalised differential cross section,  $(1/\sigma_{fid})d\sigma_{fid}/d\log(m_{bb}/p_T)$ , as a function of  $\log(m_{bb}/p_T)$  for  $m_{bb}$  the invariant mass of the two b-jets.

```
dependent_variables:
- header: {name: '$(1/\sigma_{\text{fid}})d\sigma_{\text{fid}}/d\Delta R(b,b)$'}
  qualifiers:
  - {name: RE, value: P P --> g < b b > X}
  - {name: Sqrt(S), units: GEV, value: '13000.0'}
  values:
  - errors:
    - {label: stat, symerror: '0.0016'}
    - asymerror: {minus: '-0.019', plus: '0.076'}
      label: sys,Calorimeterjetenergy
    - asymerror: {minus: '-0.042', plus: '0.041'}
      label: sys,Flavortagging
    - asymerror: {minus: '-0.0069', plus: '0.03'}
      label: sys,Tracking
    - asymerror: {minus: '-0.1', plus: '0.1'}
      label: sys,Backgroundfit
    - asymerror: {minus: '-0.00099', plus: '0.00099'}
      label: sys,UnfoldingMethod
    - asymerror: {minus: '-0.15', plus: '0.15'}
      label: sys,Theoreticalmodeling
    value: '2.98'
  - errors:
    - {label: stat, symerror: '0.0019'}
    - asymerror: {minus: '-0.1', plus: '0.062'}
      label: sys,Calorimeterjetenergy
    - asymerror: {minus: '-0.034', plus: '0.034'}
      label: sys,Flavortagging
    - asymerror: {minus: '-0.0092', plus: '0.022'}
      label: sys,Tracking
    - asymerror: {minus: '-0.13', plus: '0.13'}
      label: sys,Backgroundfit
    - asymerror: {minus: '-0.0015', plus: '0.0015'}
      label: sys,UnfoldingMethod
    - asymerror: {minus: '-0.23', plus: '0.23'}
      label: sys,Theoreticalmodeling
    value: '3.51'
```

- YAML with resource files
- YAML
- YODA
- ROOT
- CSV

```
independent_variables:
- header: {name: '$\Delta R(b,b)$'}
  values:
  - {high: 0.25, low: 0.2}
  - {high: 0.3, low: 0.25}
  - {high: 0.4, low: 0.3}
  - {high: 0.5, low: 0.4}
  - {high: 0.6, low: 0.5}
  - {high: 0.7, low: 0.6}
```

YAML files with metadata, bin contents, and uncertainties



# How to represent unbinned data?

49

If the data can be fit with a function, you could publish the function (e.g. if it is a NN, you could publish the architecture and weights).

Another natural representation that doesn't require a function fit is to publish data sampled from the unfolded result.



*My proposal is based on this idea.*

As in HEPData, I propose there is a “submission” YAML file with the same measurement metadata.

Each submeasurement\* also has some metadata & points to a data file. In HEPData, the data file is itself a YAML file.

The files will have data with the “shape”  $[(M+1) \times N(k+1)]$

...where  $N$  is the number of sampled events  
and  $M$  is the number of systematic uncertainties  
and  $k$  is the number of dimensions per event

\*this could be a single observable, or many observables

The files will have data with the “shape”  $[(M+1) \times N(k+1)]$

Each event has  $k$  floats\* and 1 event weight

There are  $N$  events

This is repeated for each of the  $M$  systematic uncertainties

*For representations that don't have weights, the weights will be set to 1. For representations that only use weights, there will be  $M$  copies of the original array.*

I have not thought deeply about file formats (numpy, root, hdf5) and would be happy to hear opinions.

\*For variable-length measurements, perhaps should use variable-length arrays like [awkward](#) for storage

The submission YAML should give metadata about which uncertainties are included.

For statistical uncertainties, there should be  $Q$  replicas and the uncertainty in a given bin is computed by taking the standard deviation over replicas.

For systematic uncertainties, the difference between the nominal and varied bin content is the uncertainty.

There should be warnings in metadata and/or inflated uncertainties in regions of phase space that should not be studied with the data.

# Proposal - where to store?

53

Zenodo is a very natural location. Maybe the submission YAML can also be hosted on HEPData and linked to Zenodo for each searching?

# Proposal - example

54

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
In [2]: x = np.random.normal(0,1,10000)
```

```
In [3]: w = np.abs(x)**0.2
w_syst_up = np.abs(x)**0.3
w_syst_dn = np.abs(x)**0.1
```

```
In [4]: plt.hist(x,bins=np.linspace(-3,3,10),alpha=0.5)
plt.hist(x,bins=np.linspace(-3,3,10),weights=w,histtype="step",color="black")
n_syst_up,b=np.histogram(x,bins=np.linspace(-3,3,10),weights=w_syst_up)
n_syst_dn,_=np.histogram(x,bins=np.linspace(-3,3,10),weights=w_syst_dn)
for i in range(len(b)-1):
    plt.fill_between([b[i],b[i+1]],n_syst_dn[i],n_syst_up[i],color="black",alpha=0.3)
plt.xlabel("X")
```

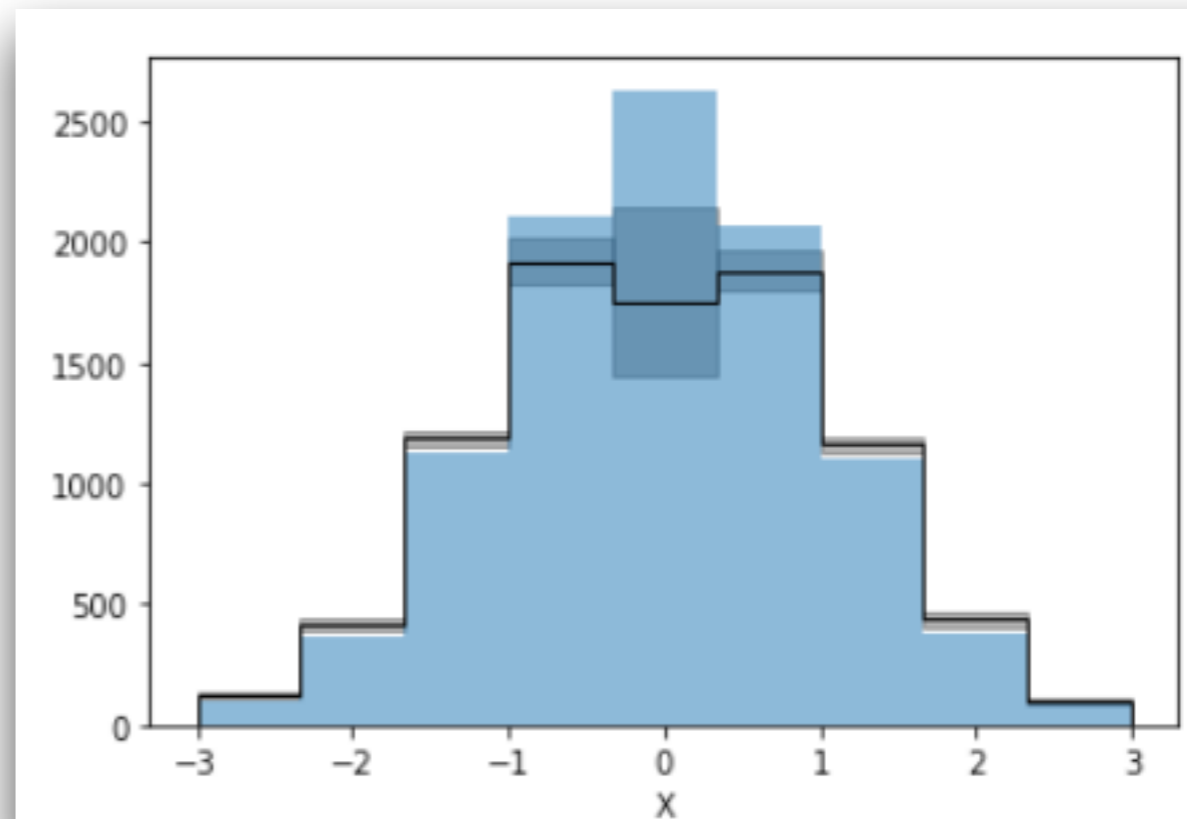
```
In [5]: d = {"nominal":x,
            "nominalw":w,
            "syst_up":x,
            "syst_upw":w_syst_up,
            "syst_dn":x,
            "syst_dnw":w_syst_dn}
```

```
In [6]: df = pd.DataFrame(data=d)
```

```
In [7]: df
```

```
Out[7]:
```

	nominal	nominalw	syst_up	syst_upw	syst_dn	syst_dnw
0	0.731914	0.939490	0.731914	0.910622	0.731914	0.969273
1	0.146232	0.680783	0.146232	0.561711	0.146232	0.825096
2	-0.629654	0.911634	-0.629654	0.870423	-0.629654	0.954795
3	0.581001	0.897089	0.581001	0.849675	0.581001	0.947148
4	-0.321038	0.796730	-0.321038	0.711159	-0.321038	0.892597
...	...	...	...	...	...	...



Discussion!

