LHC Olympics 2020: Winter Games

Gregor Kasieczka
Hamburg

Benjamin Nachman
LBNL

David Shih
Rutgers / Berkeley / LBNL

BERKELEY LAB

BERKELEY EXPERIMENTAL PARTICLE PHYSICS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Emmy Noether-Programm
DFG Deutsche Forschungsgemeinschaft

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Questions in fundamental physics

Theoretical and experimental questions motivate a deep exploration **of the fundamental structure of nature**

Why is the Higgs boson so light?

Hierarchy problem

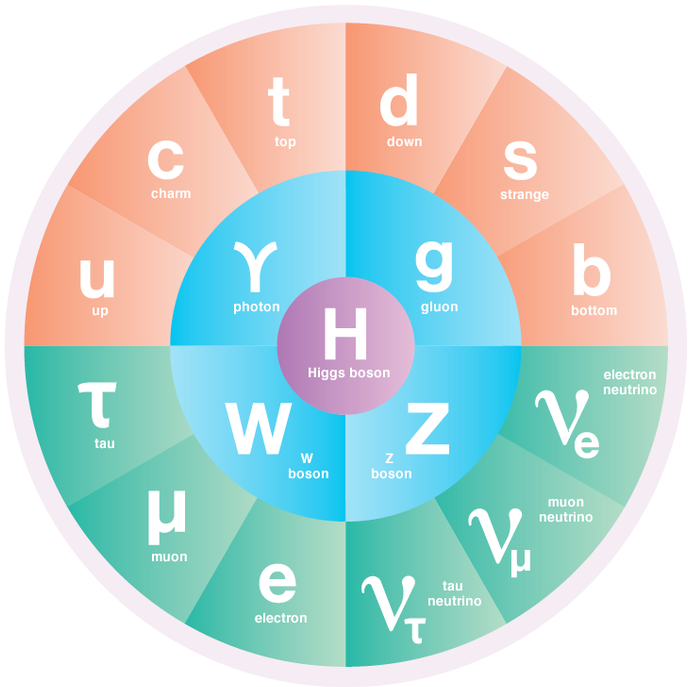
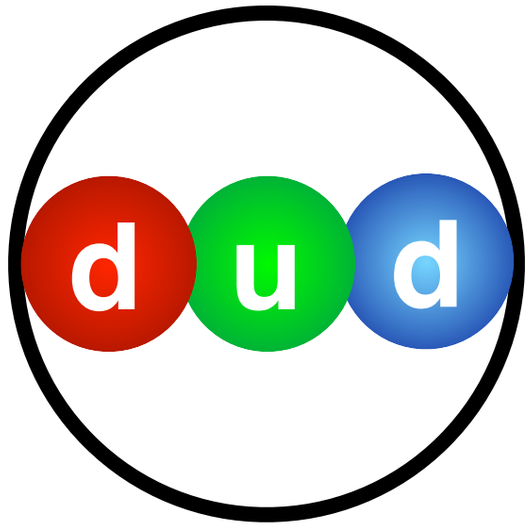


image source: symmetry magazine

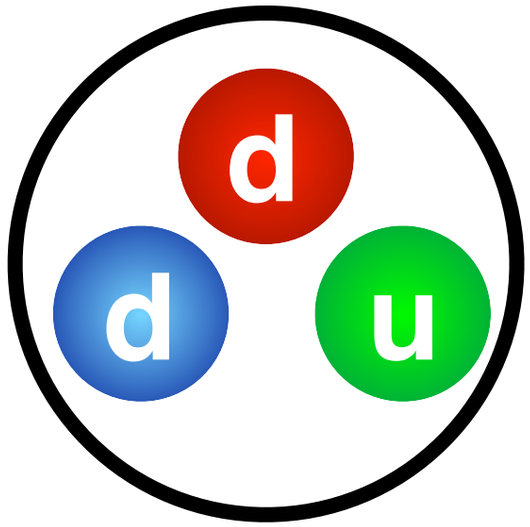
See also: quantum gravity

Why do neutrons have no dipole moment?

Strong CP



Reality



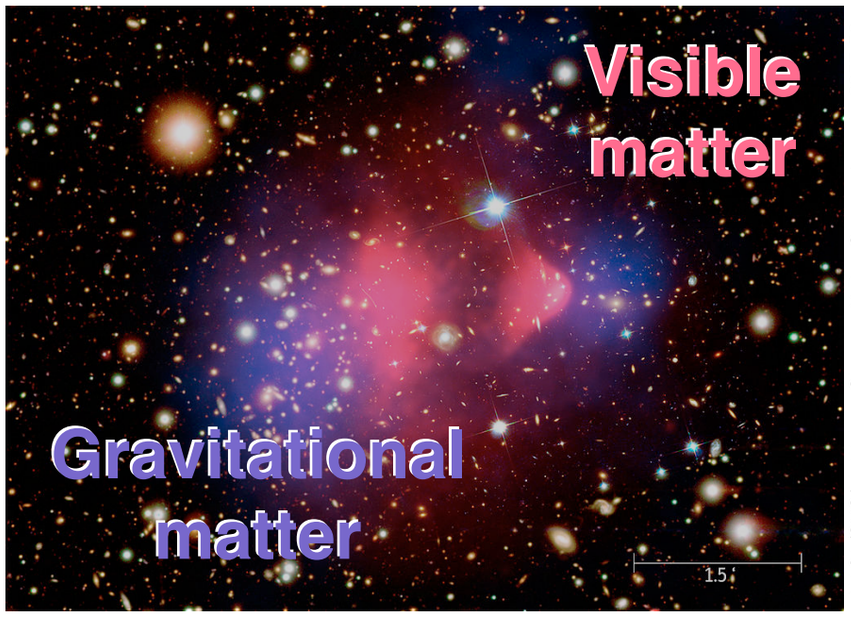
>99% of pictures on the internet

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

What is the extra gravitational matter?

Dark Matter



NASA/CXC/M. Weiss - Chandra X-Ray Observatory: 1E 0657-56

See also: dark energy

Why do neutrinos have a mass?

Flavor puzzles

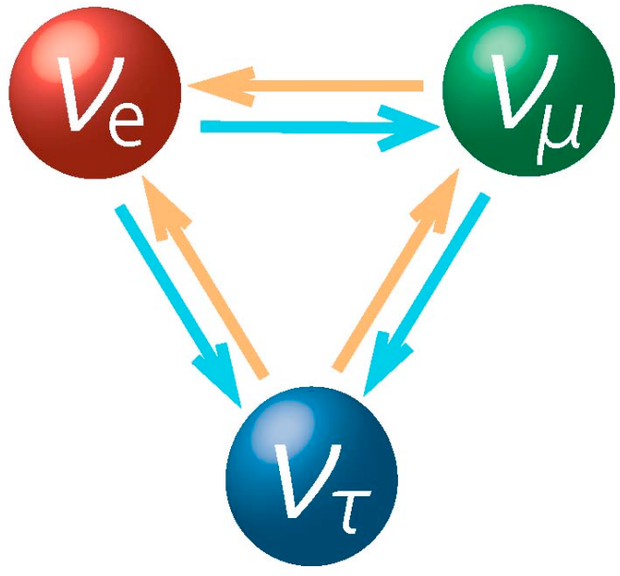


image source: T2K collaboration

See also: Where did all the anti-particles go? (Baryogengesis)

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Dark energy

Flavor puzzles

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Dark energy

Flavor puzzles

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at LHC energies

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Dark energy

Flavor puzzles

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at LHC energies

(2) Patience! (new physics is rare)

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Dark energy

Flavor puzzles

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

(1) There is nothing new at LHC energies

(2) Patience! (new physics is rare)

(3) We are not looking in the right place

Questions in fundamental physics

Theoretical and **experimental** questions motivate a deep exploration **of the fundamental structure of nature**

Dark matter

Hierarchy problem

Strong CP

Dark energy

Flavor puzzles

Baryogengesis

We have performed thousands of hypothesis tests & have no significant evidence for physics beyond the Standard Model

Three possibilities

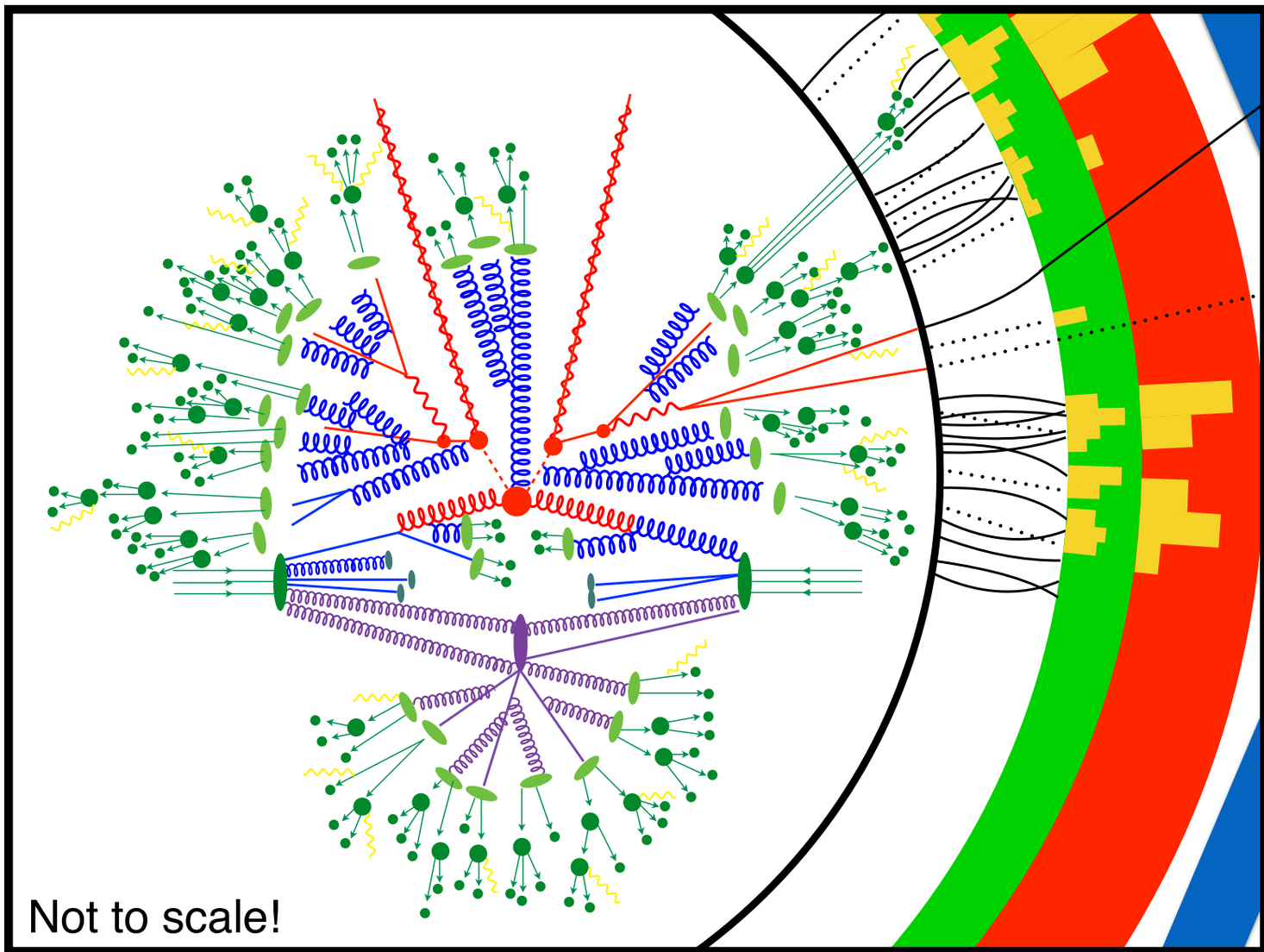
This is the motivation for this challenge!

(3) We are not looking in the right place

Large Hadron Collider

Many of the deep questions in fundamental physics can be probed at the LHC.

Image inspired by JHEP 02 (2009) 007



Large Hadron Collider

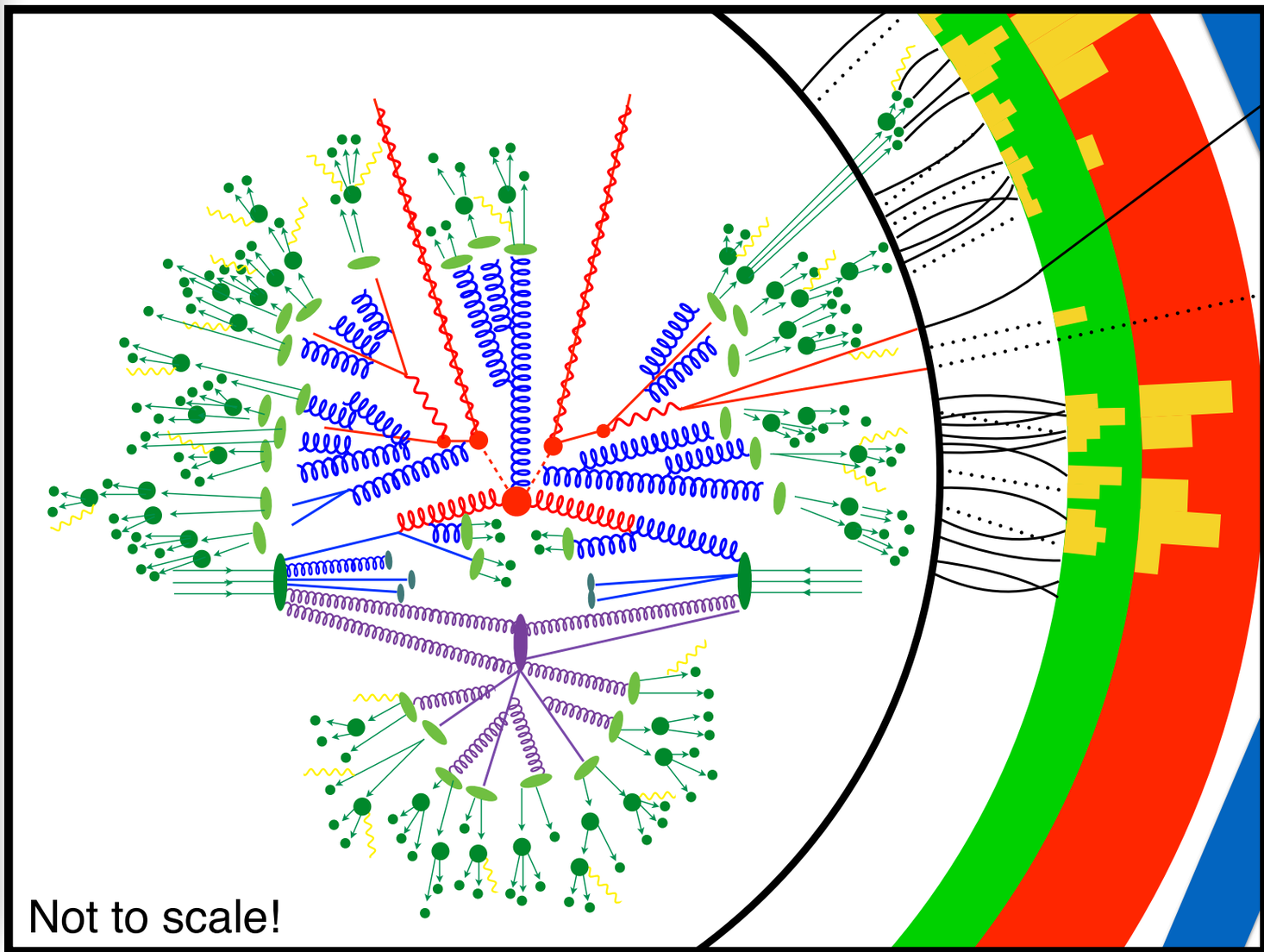
Many of the deep questions in fundamental physics can be probed at the LHC.

Key challenges (and opportunity!)

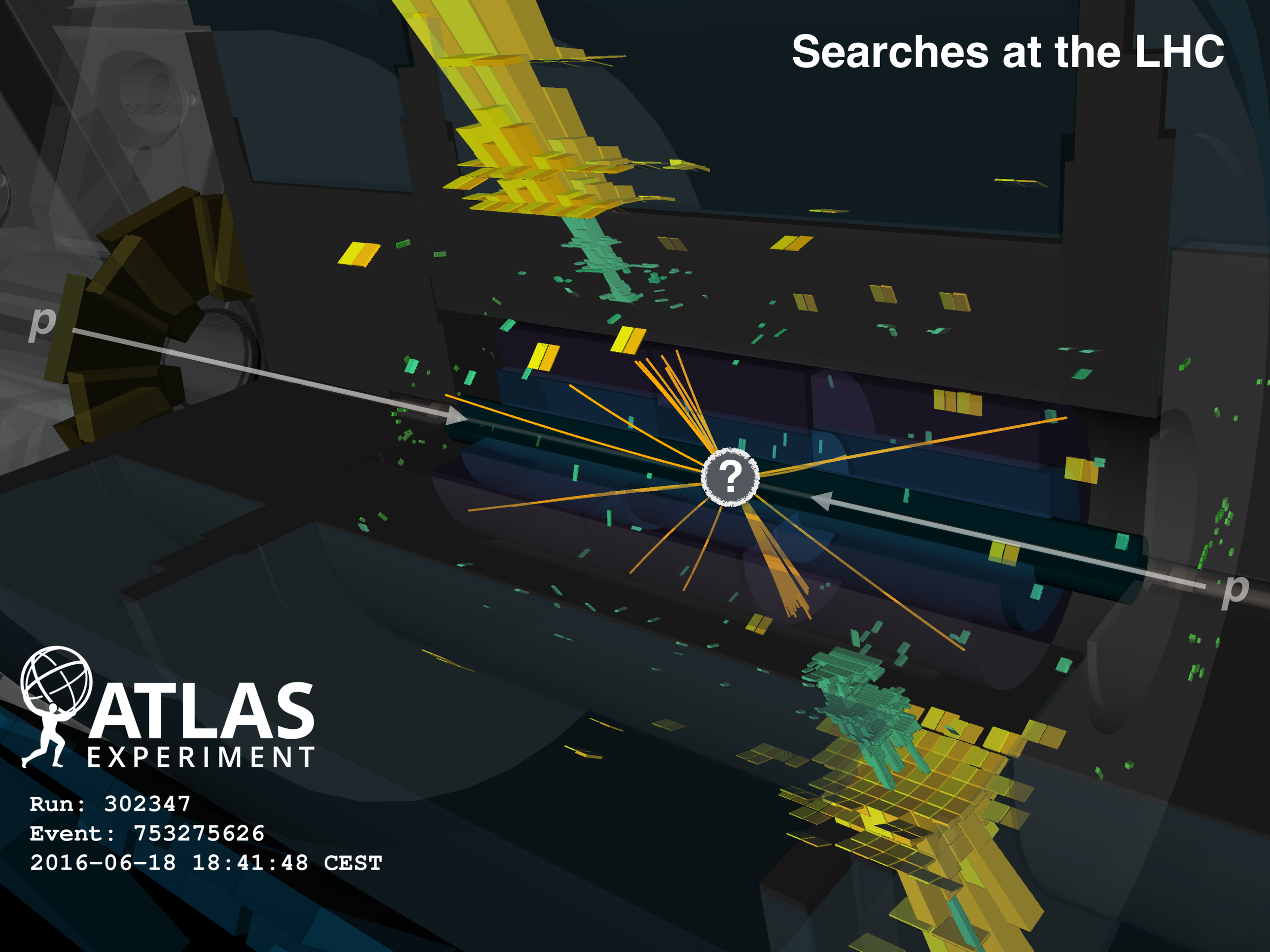
Typical collision events at the LHC produce **O(1000+)** particles

We detect these particles with **O(100 M)** readout channels

Image inspired by JHEP 02 (2009) 007



Searches at the LHC



Optimality and Neyman-Pearson

The most powerful* discriminant between two simple hypothesis H_1 and H_2 is given by the likelihood ratio:

$$R(x) = \frac{\mathcal{L}(x|H_1)}{\mathcal{L}(x|H_2)}$$

Examples

- H_1 = specific signal + SM background
  H_2 = SM background

  conventional, model-specific search strategy

- H_1 = "data"
  H_2 = SM background

  model-independent search strategy "anomaly detection"

*This means that for a fixed probability of rejecting H_1 when it is true, this has the highest probability of rejecting H_1 when H_2 is true. Also note that this is in the absence of profiling.

Connection with classification

If one can design an optimal binary classifier (e.g. using a deep neural network) to distinguish between H_1 and H_2, then the output of the classifier will be the likelihood ratio:

$$L = \sum_{x \in H_1} \log y(x) + \sum_{x \in H_2} \log(1 - y(x))$$

$$\Rightarrow \quad y(x) = \frac{R(x)}{1 + R(x)}$$

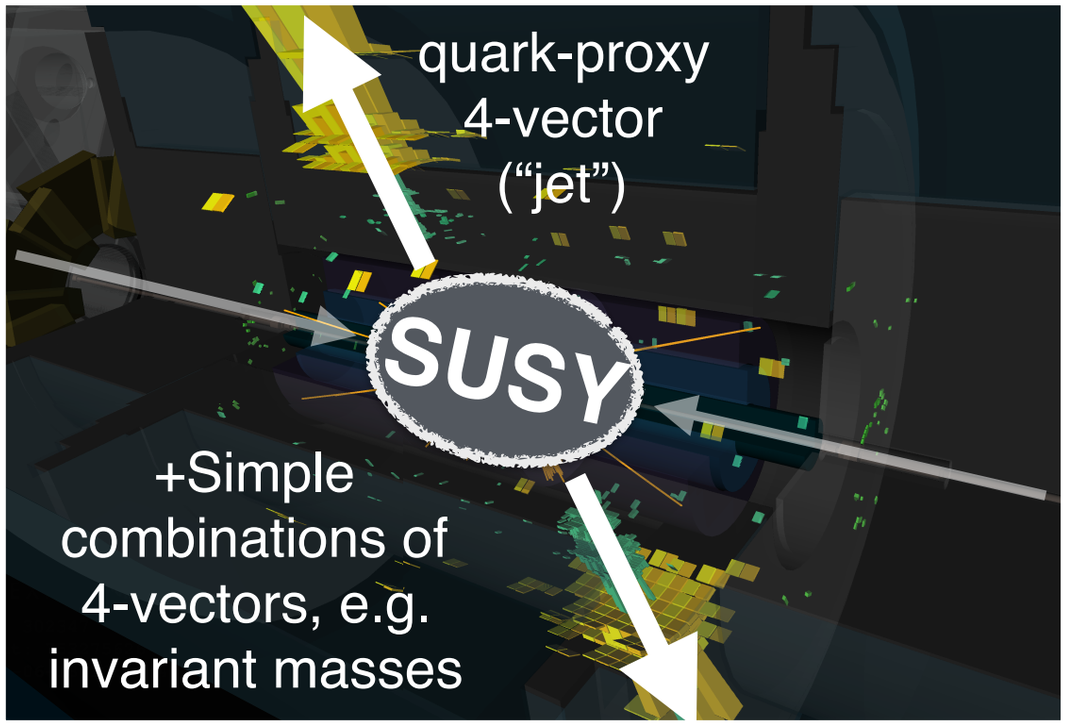
$$R(x) = \frac{\mathcal{L}(x|H_1)}{\mathcal{L}(x|H_2)}$$

"likelihood ratio trick"

Underlies many major recent advances in deep learning, including generative models.

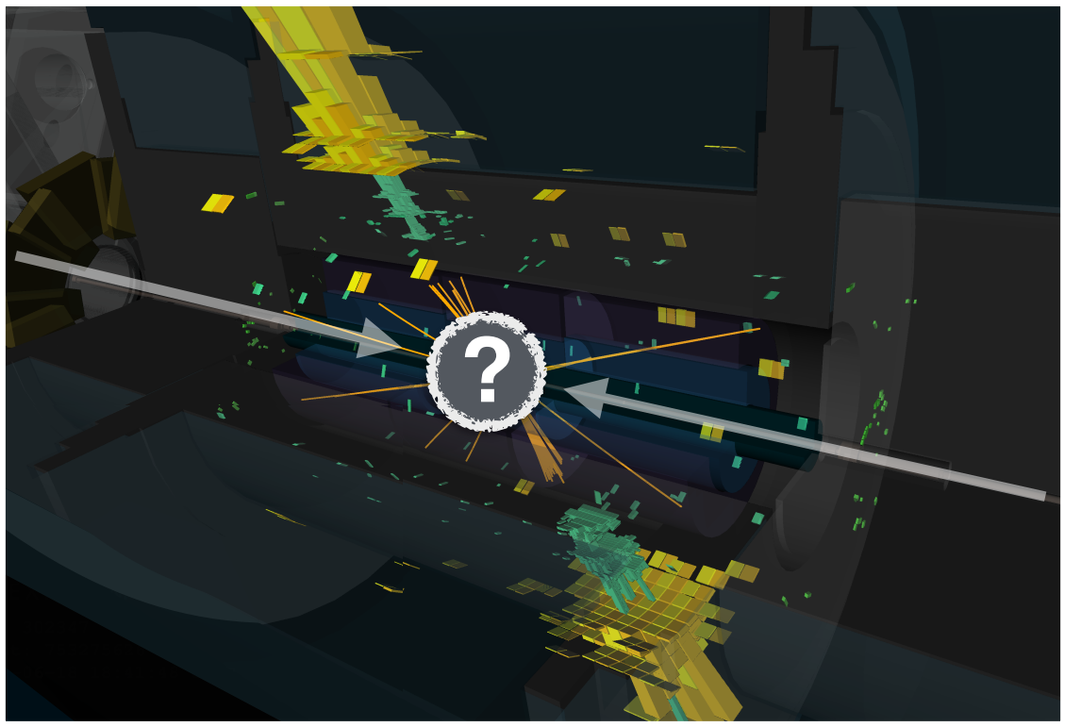
Current Search Paradigm

SUSY = Supersymmetry



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

Current paradigm for searches



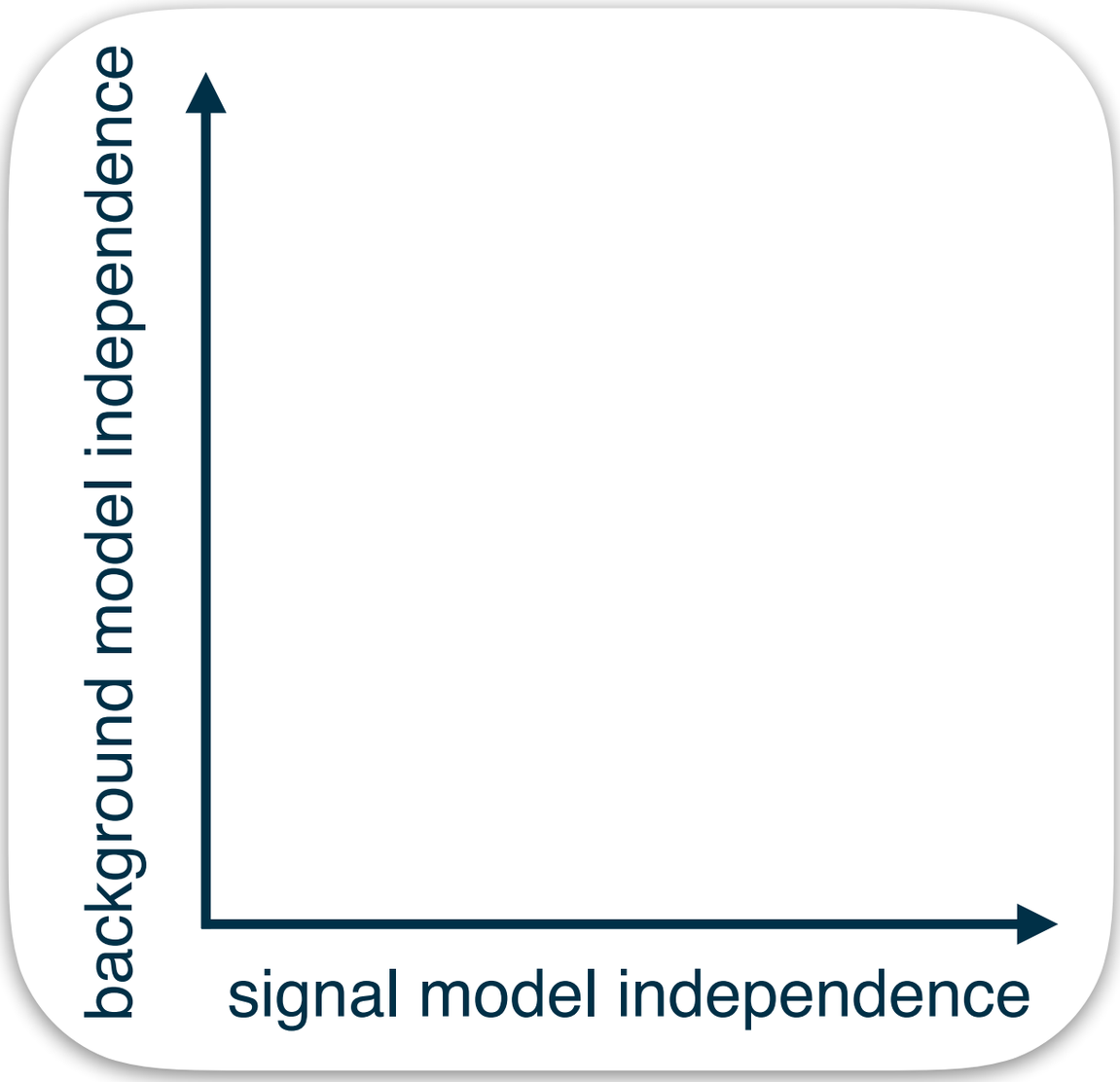
SUSY = Supersymmetry

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

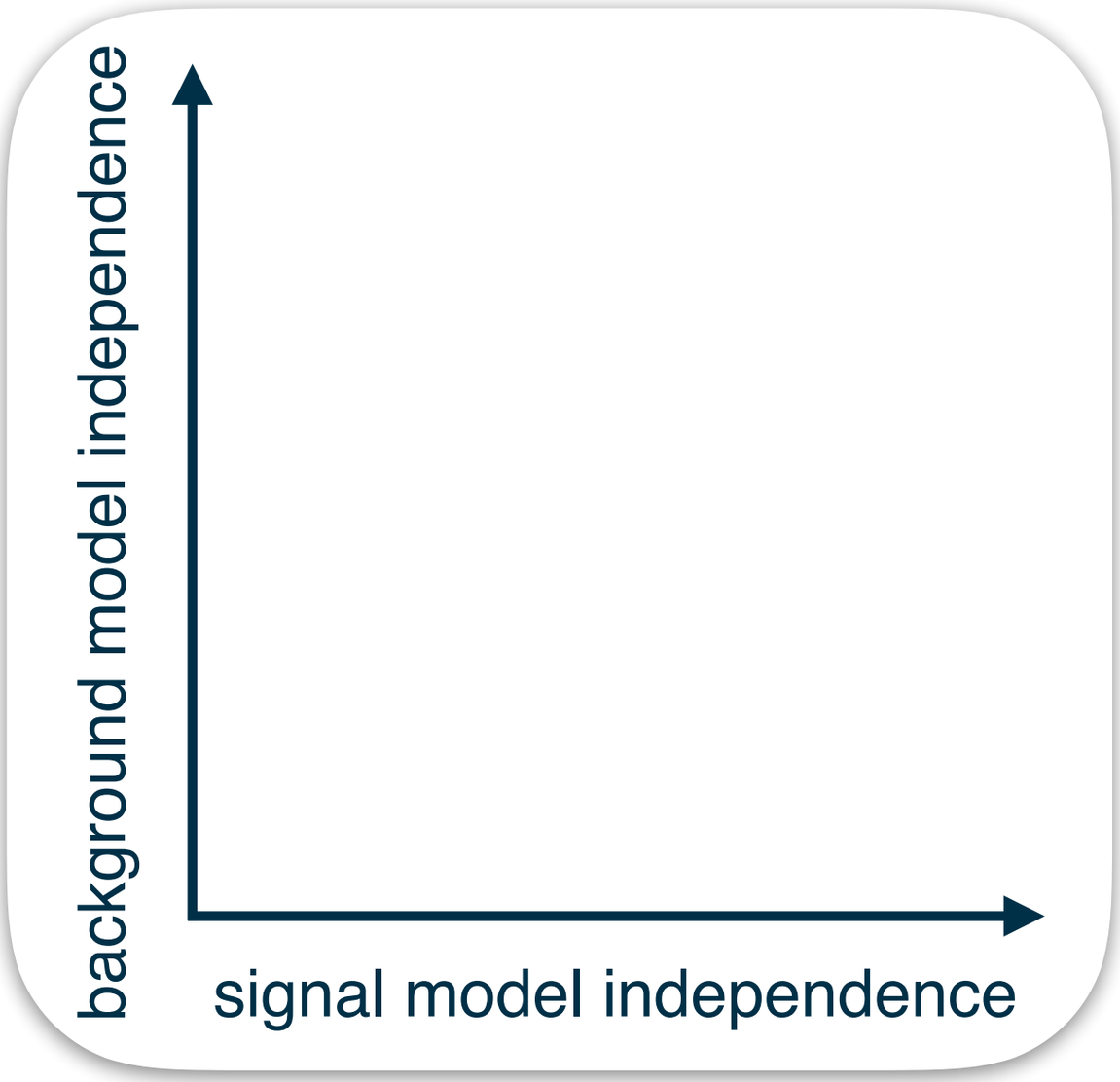
(clearly, we should still do model-dependent searches as well!)

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

Model dependence



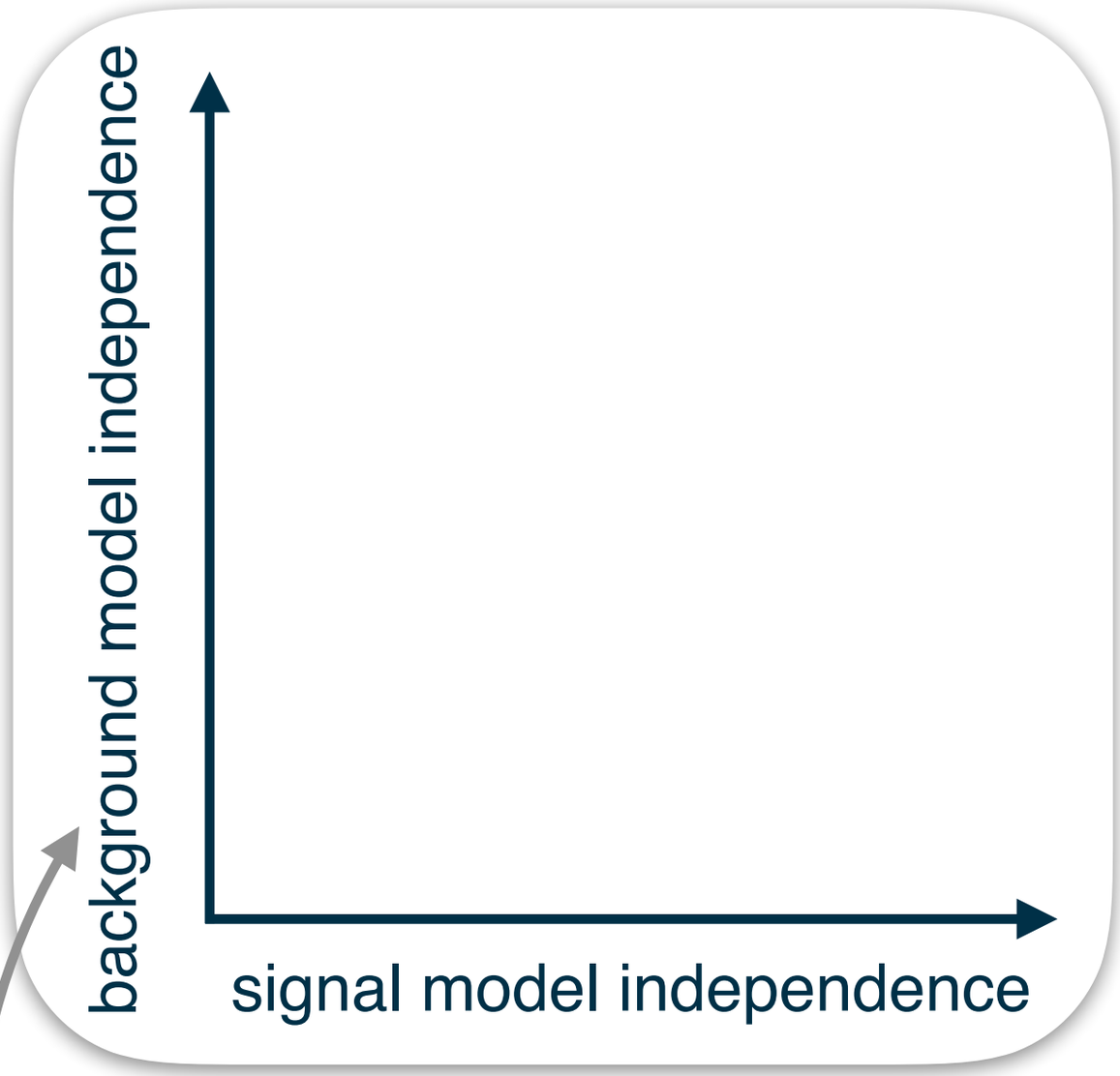
Signal sensitivity



Background specificity

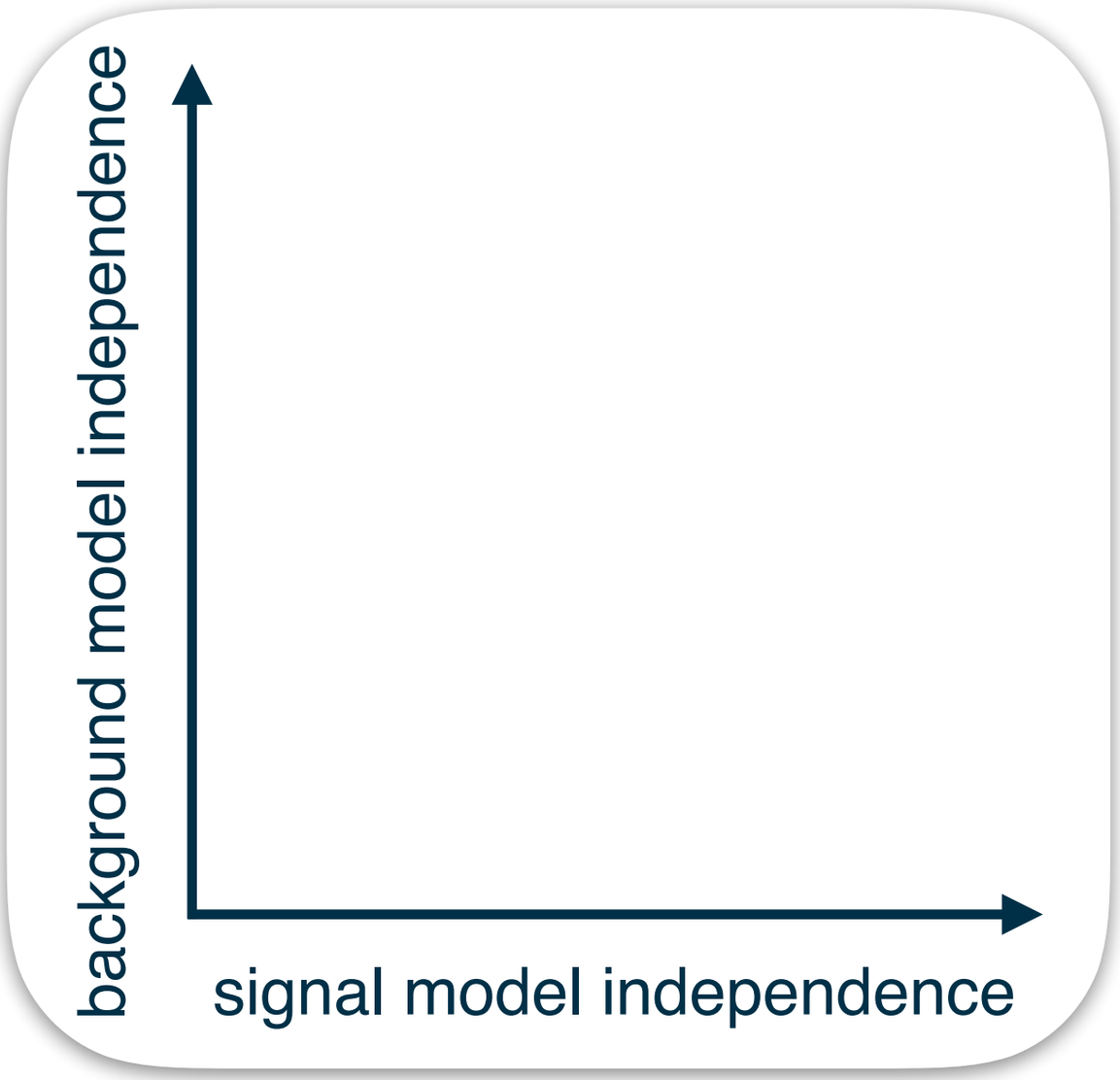
Suppose you want to search for a new signal process

Model dependence



Signal sensitivity

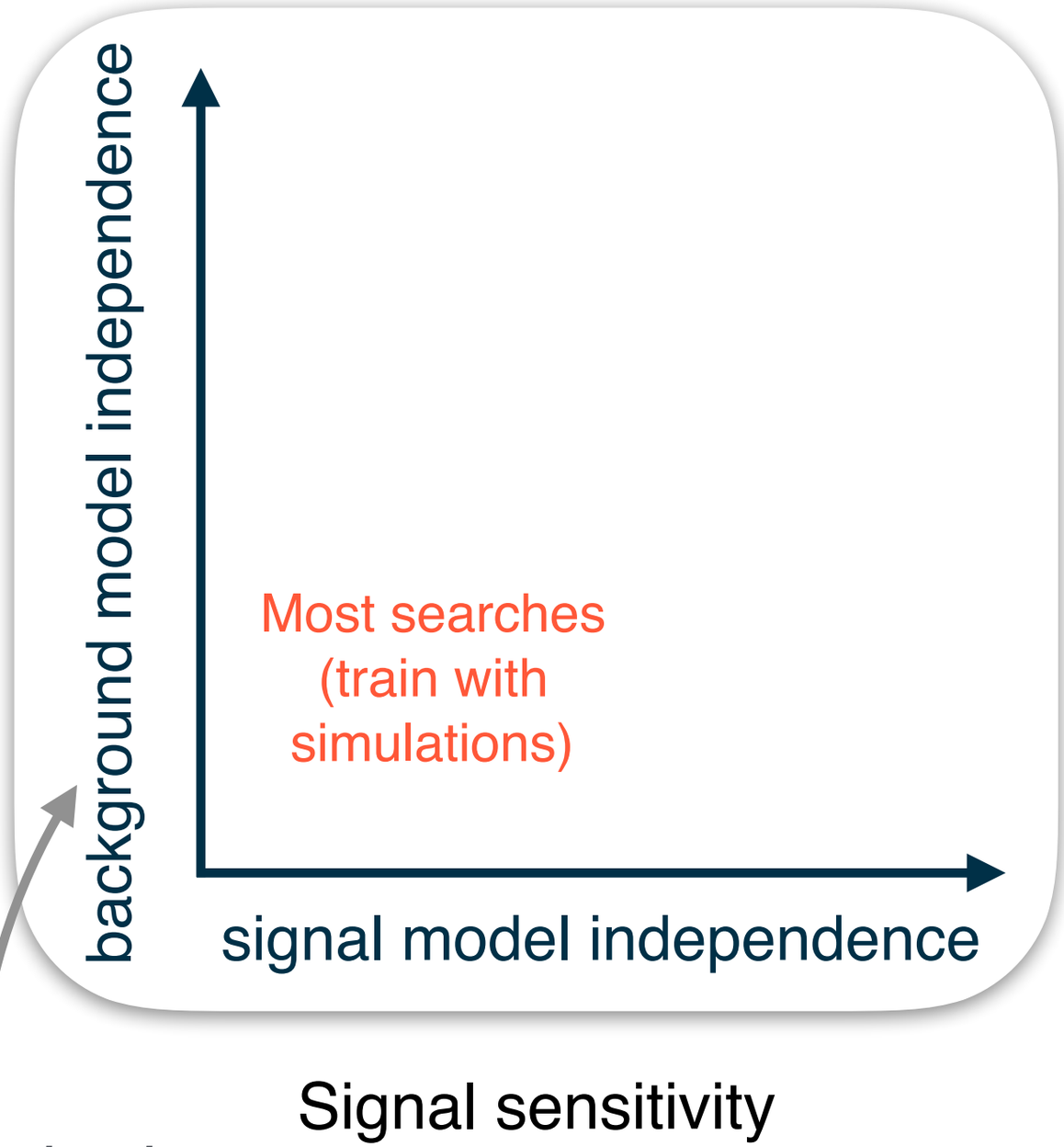
Standard Model



Background specificity

B. Nachman, D. Shih, 2001.04990

Model dependence



$$R_S(x) = \frac{\mathcal{L}(x|S_{sim})}{\mathcal{L}(x|B_{sim})}$$

signal and background
model *dependent*

- S: a *specific* signal model, e.g. supersymmetry
- x: some set of relevant features characterizing each event (e.g. MET, HT,…)
- Rely on simulations of background and signal to construct likelihood ratio.

> 99% of searches at the LHC are of this type

Model dependence

background model independence

Some searches (train signal versus data)

Most searches (train with simulations)

signal model independence

Signal sensitivity

Standard Model

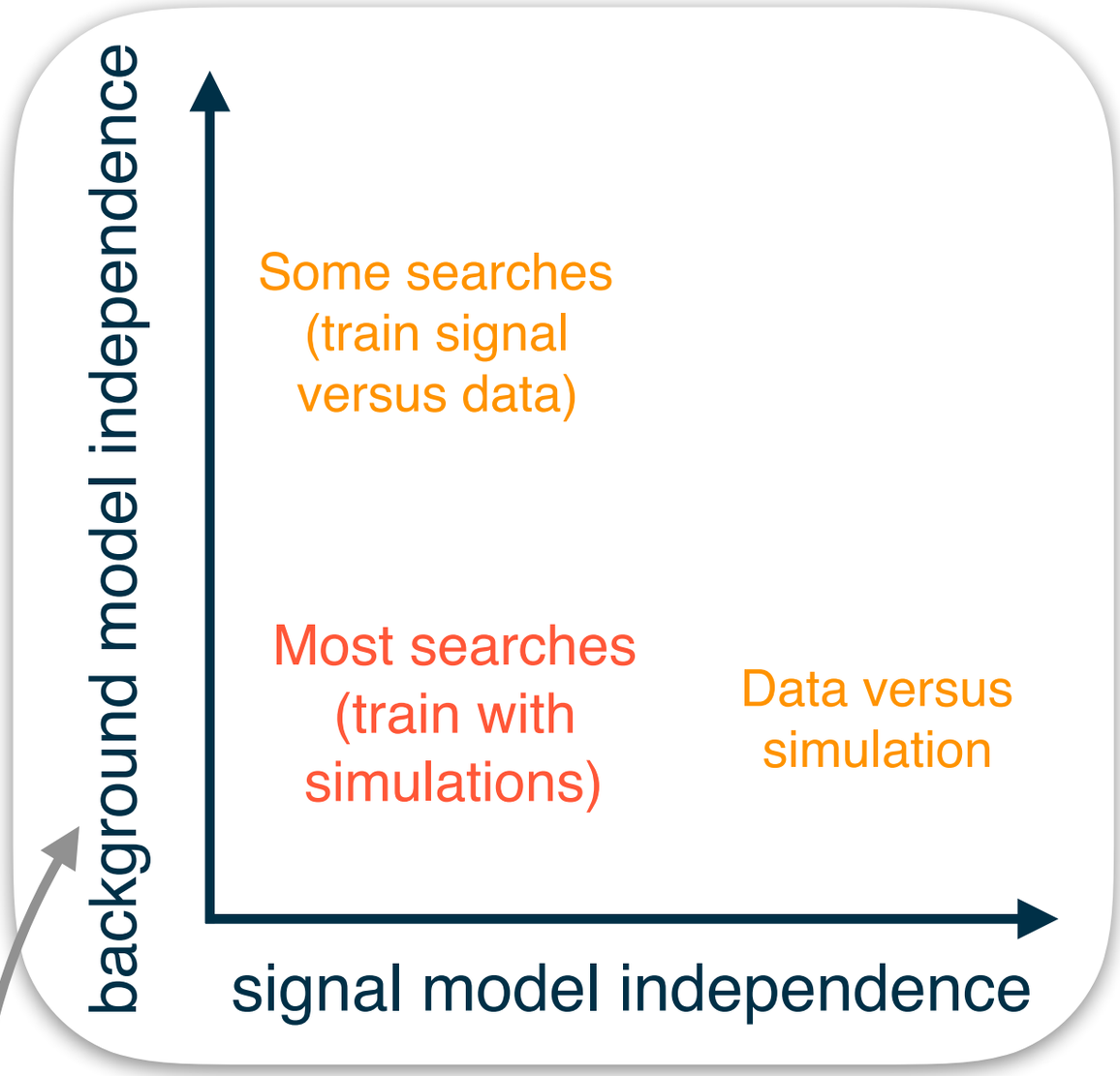
background model independence

e.g. Signal simulation versus calibration data

signal model independence

Background specificity

Model dependence



Signal sensitivity

Standard Model

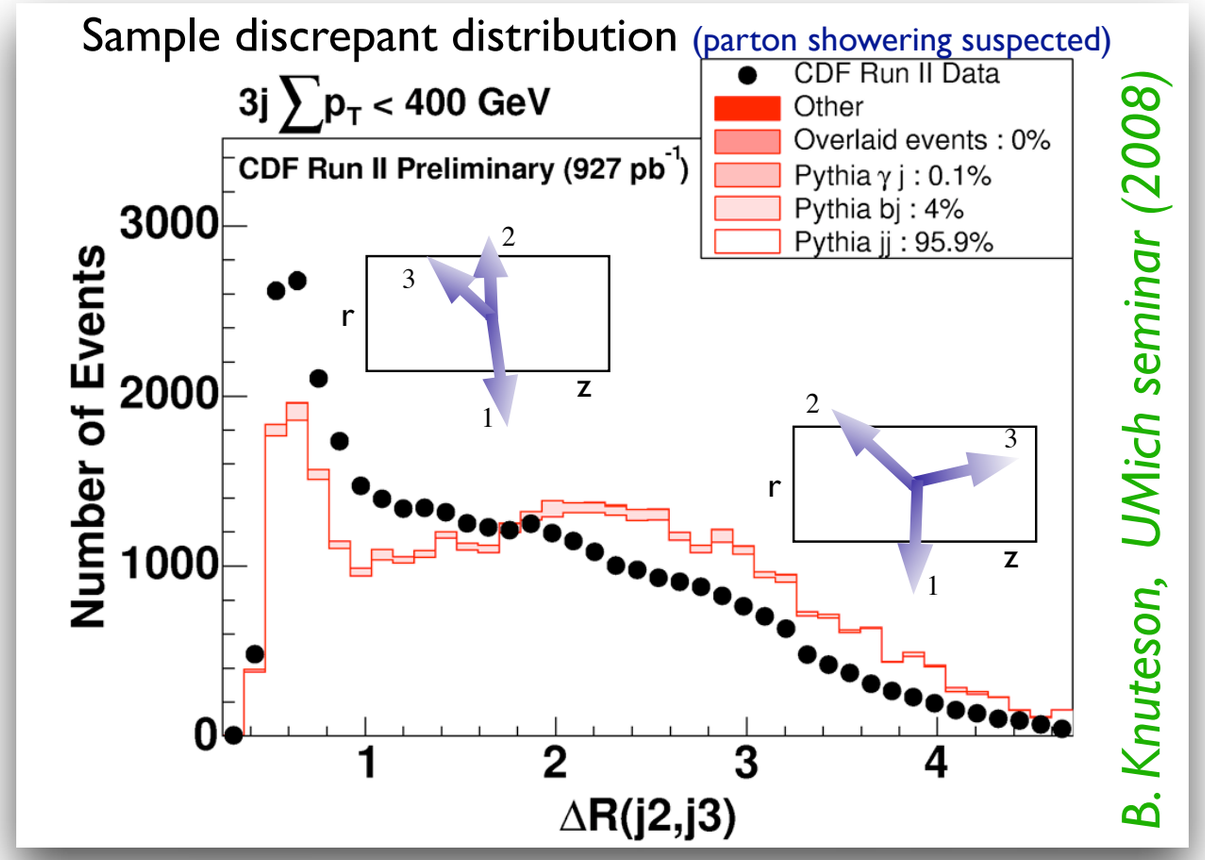
$$R(x) = \frac{\mathcal{L}(x|data)}{\mathcal{L}(x|B_{sim})}$$

signal model *independent*
background model *dependent*

Idea: compare data vs ***simulated*** SM background in 1D histograms.

B. Knuteson et al., D0, H1, CDF, CMS (“MUSiC”), ATLAS (“General Search”)
A. De Simone, T. Jacques, 1807.06038, A. Casa, Giovanna, 1809.02977, and others
R. T. D’Agnolo and A. Wulzer, PRD 99 (2019) 015014, R. T. D’Agnolo et al. 1912.12155

Model dependence



Warning: still background model dependent and may focus on the approximations of the simulation

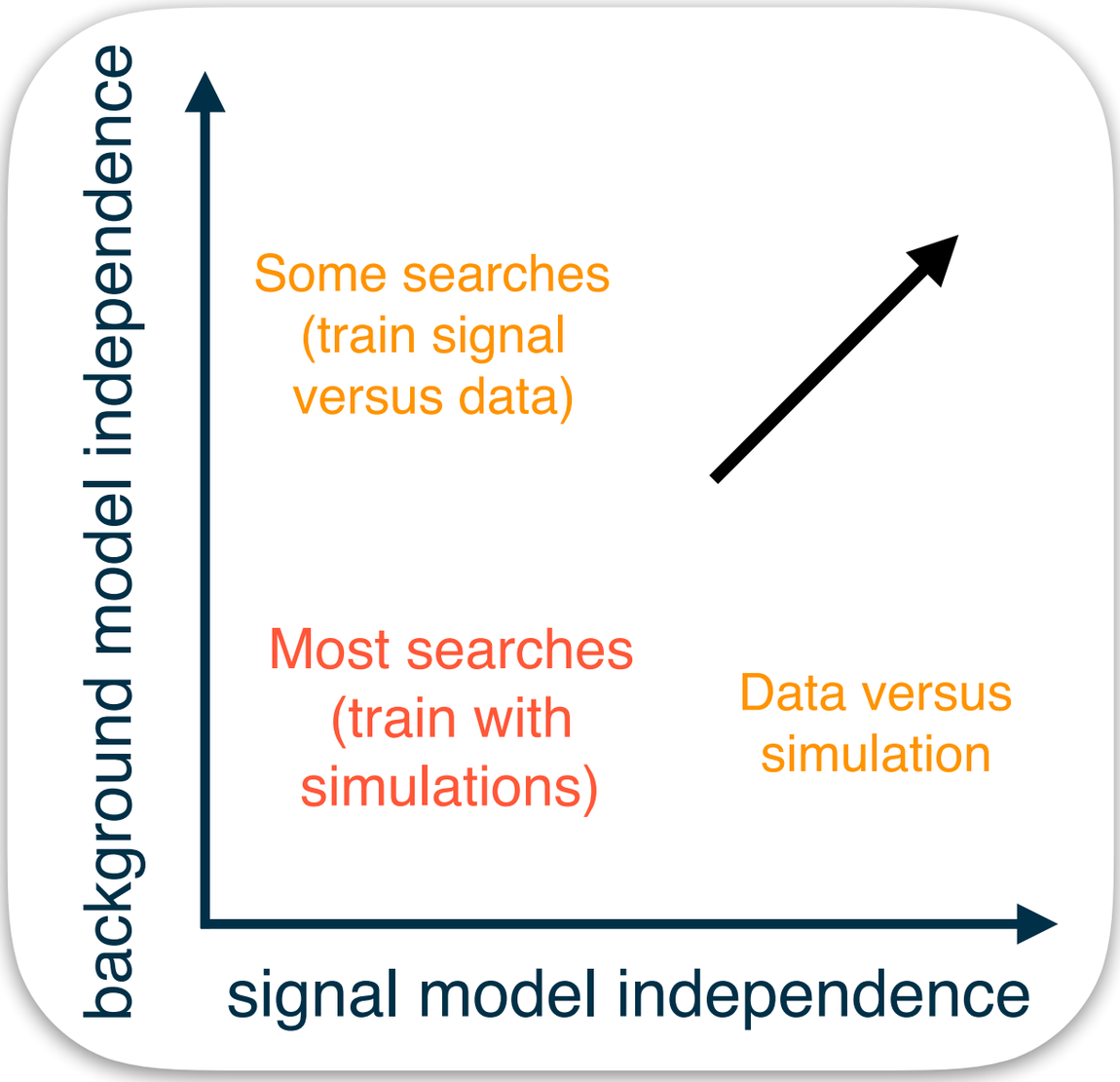
$$R(x) = \frac{\mathcal{L}(x|data)}{\mathcal{L}(x|B_{sim})}$$

signal model *independent*
background model *dependent*

Idea: compare data vs ***simulated*** SM background in 1D histograms.

B. Knuteson et al., D0, H1, CDF, CMS (“MUSiC”), ATLAS (“General Search”)
A. De Simone, T. Jacques, 1807.06038, A. Casa, Giovanna, 1809.02977, and others
R. T. D’Agnolo and A. Wulzer, PRD 99 (2019) 015014, R. T. D’Agnolo et al. 1912.12155

Model dependence



Signal sensitivity

Classic: "bump hunt"

$$R(m) = \frac{\mathcal{L}(m|data)}{\mathcal{L}(m|B_{data})}$$

m: a single feature

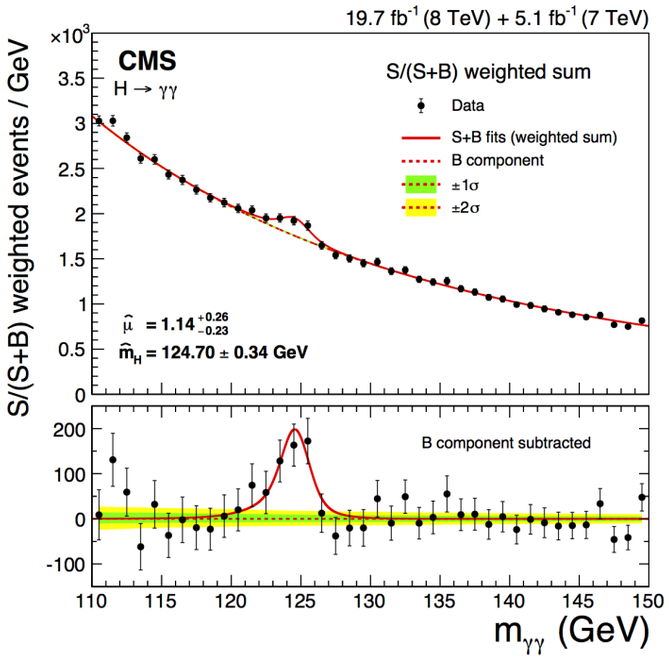
partially signal and background model *independent*

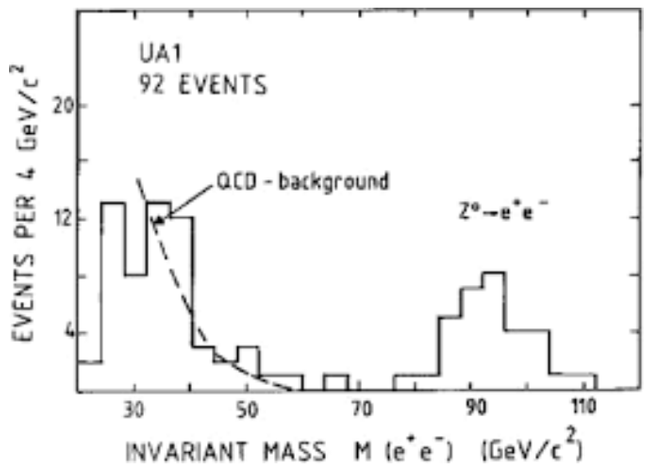
Idea: assume signal is localized in *m* while background is smooth.

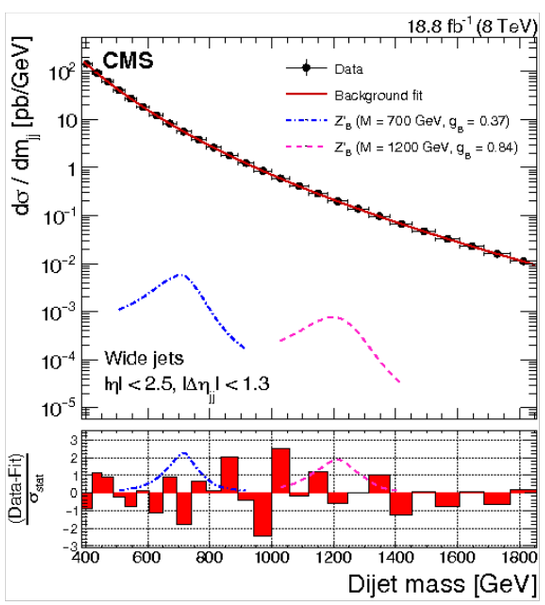
Use ***sidebands*** $m \notin (m_0-\delta m, m_0+\delta m)$ to interpolate background into ***signal region*** $m \in (m_0-\delta m, m_0+\delta m)$.

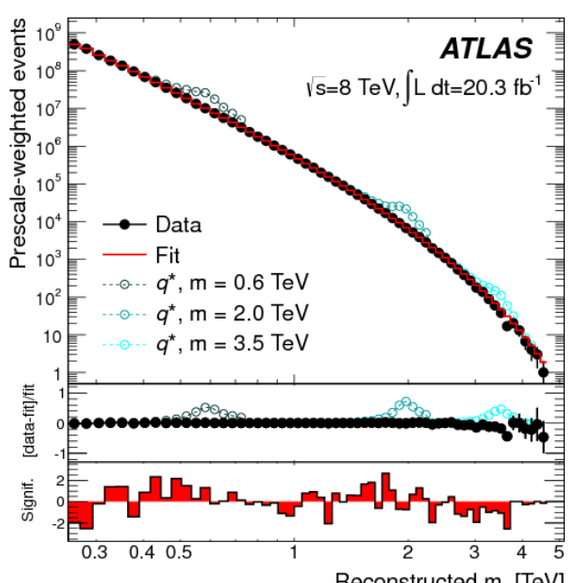
Model dependence

Classic method: many discoveries & searches









Classic: "bump hunt"

$$R(m) = \frac{\mathcal{L}(m|data)}{\mathcal{L}(m|B_{data})}$$

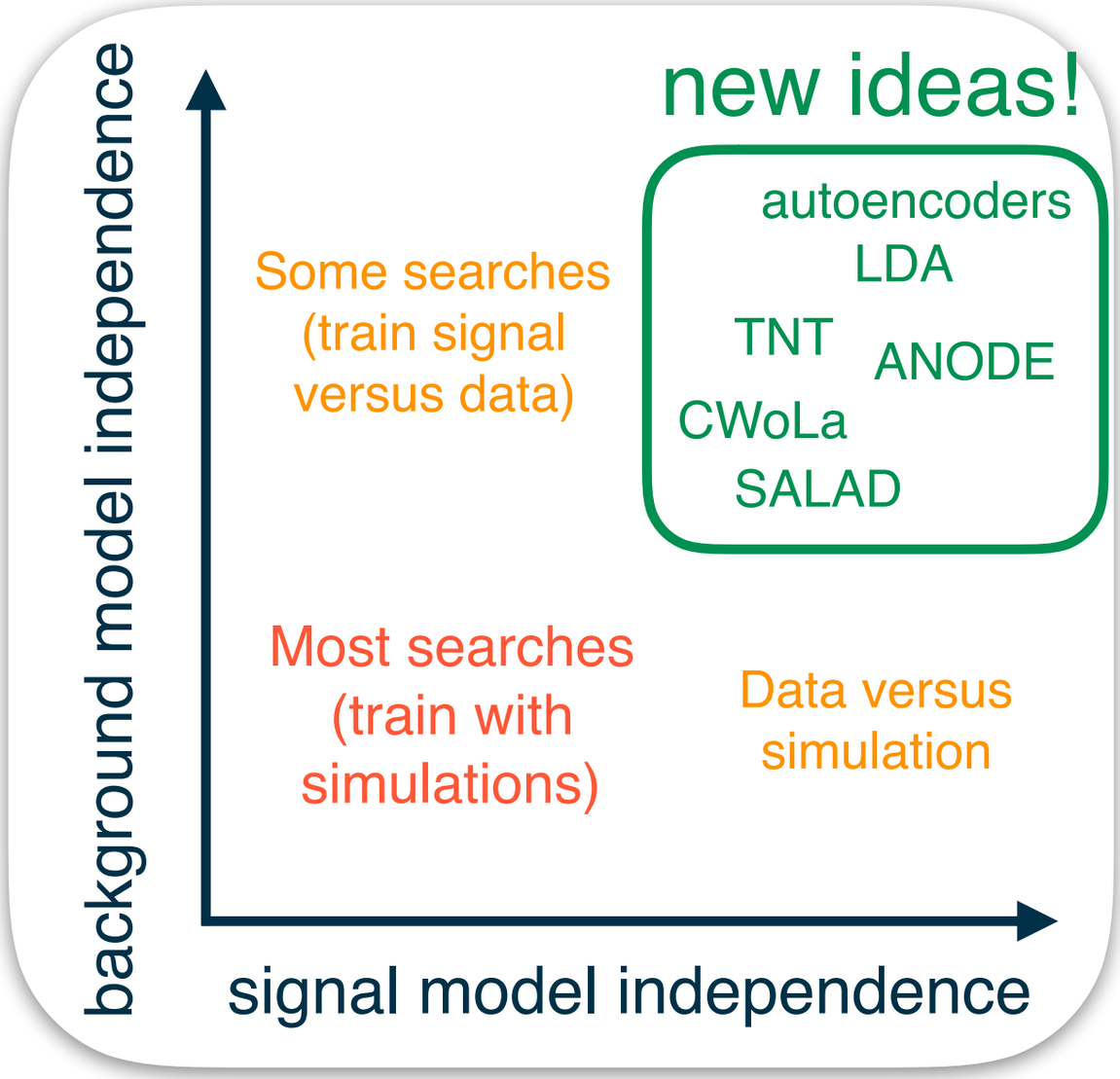
m: a single feature

partially signal and background model *independent*

Idea: assume signal is localized in *m* while background is smooth.

Use ***sidebands*** $m \notin (m_0-\delta m, m_0+\delta m)$ to interpolate background into ***signal region*** $m \in (m_0-\delta m, m_0+\delta m)$.

Model dependence



Signal sensitivity

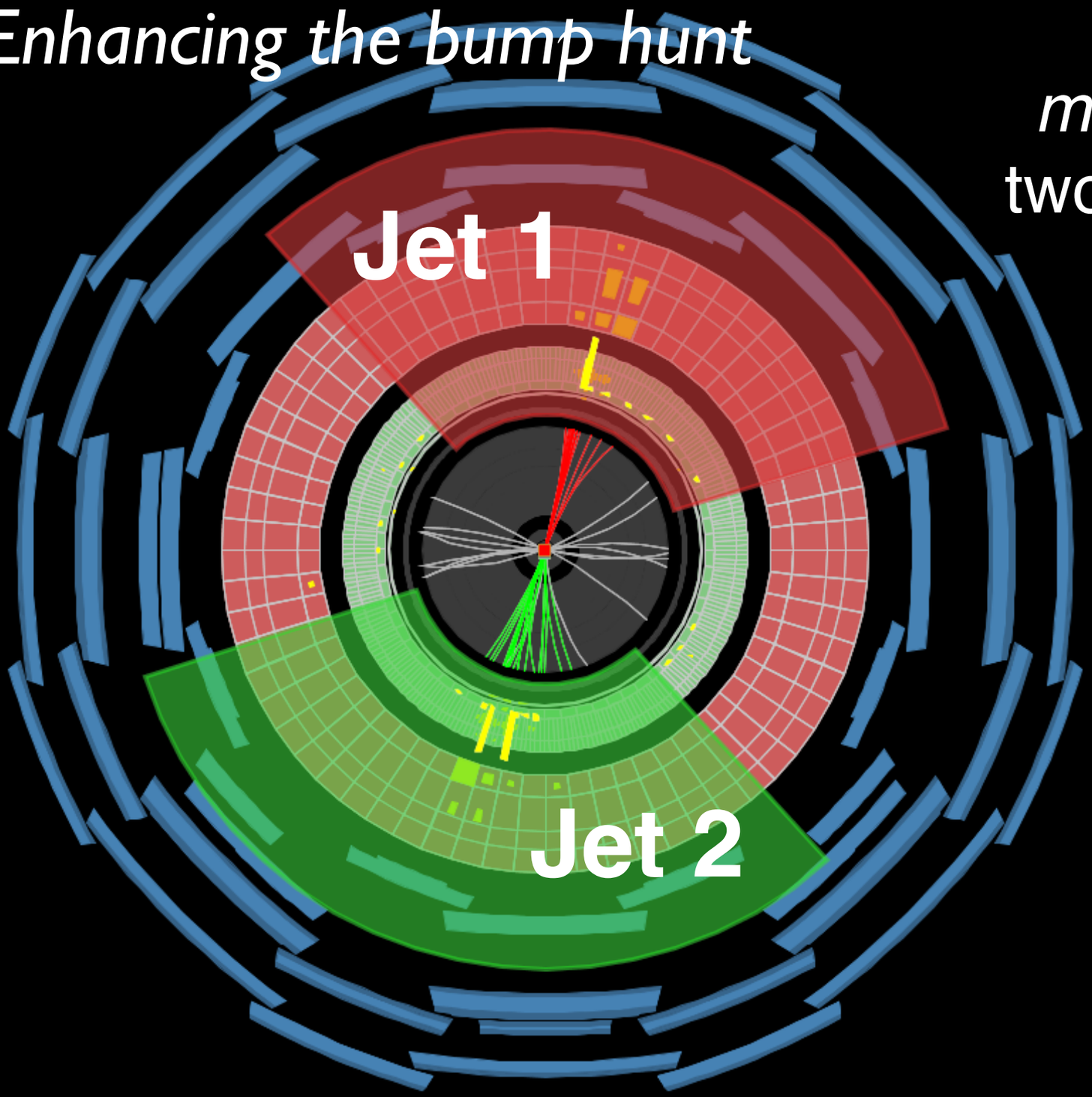
Can we develop new methods that also assume as little as possible about the signal and learn from data (no simulation)?

M. Farina, Y, Nakai, D. Shih, 1808.08992,
T. Heimel et al. SciPost Phys. 6 (2019) 030, and others
B. Dillon et al., PRD 100 (2019) 056002
B. Nachman, D. Shih, 2001.04990

O. Amram, C. Suarez, 2002.12376
J. Collins, K. Howe, B. Nachman, PRL 121 (2018) 241803
A. Andreassen, B. Nachman, D. Shih, 2001.05001

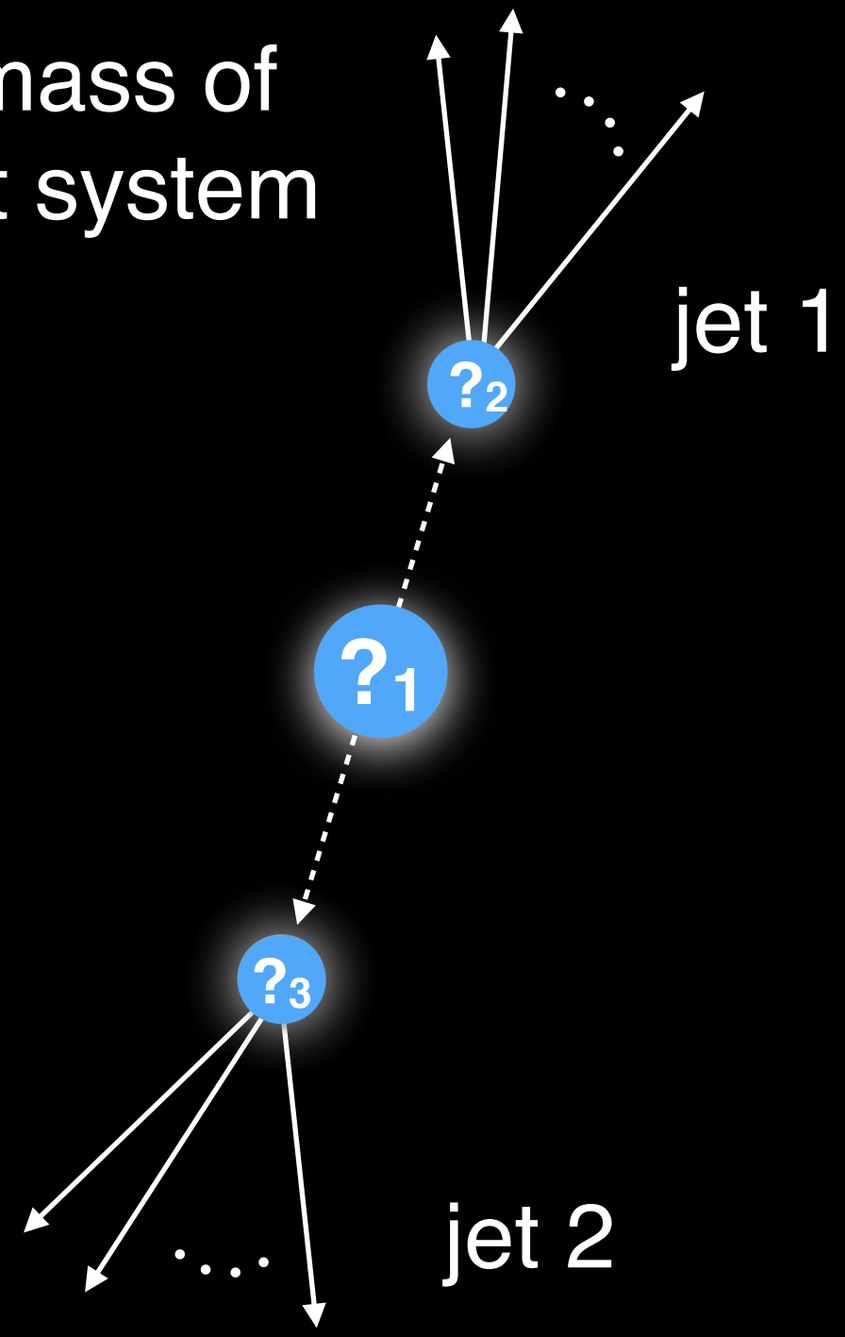
New ideas: brief illustration

Enhancing the bump hunt



collisions in/out of page

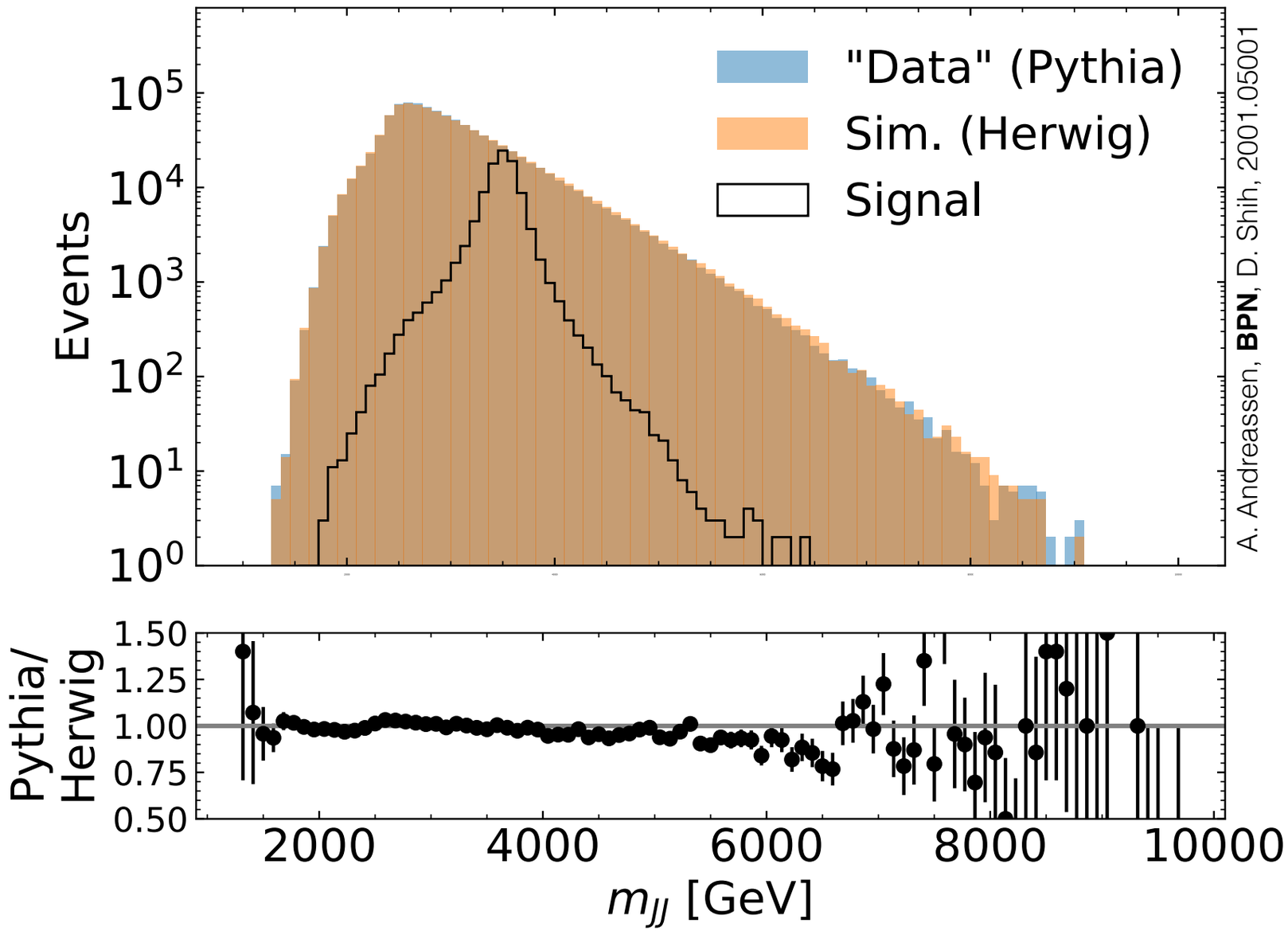
m = mass of two-jet system

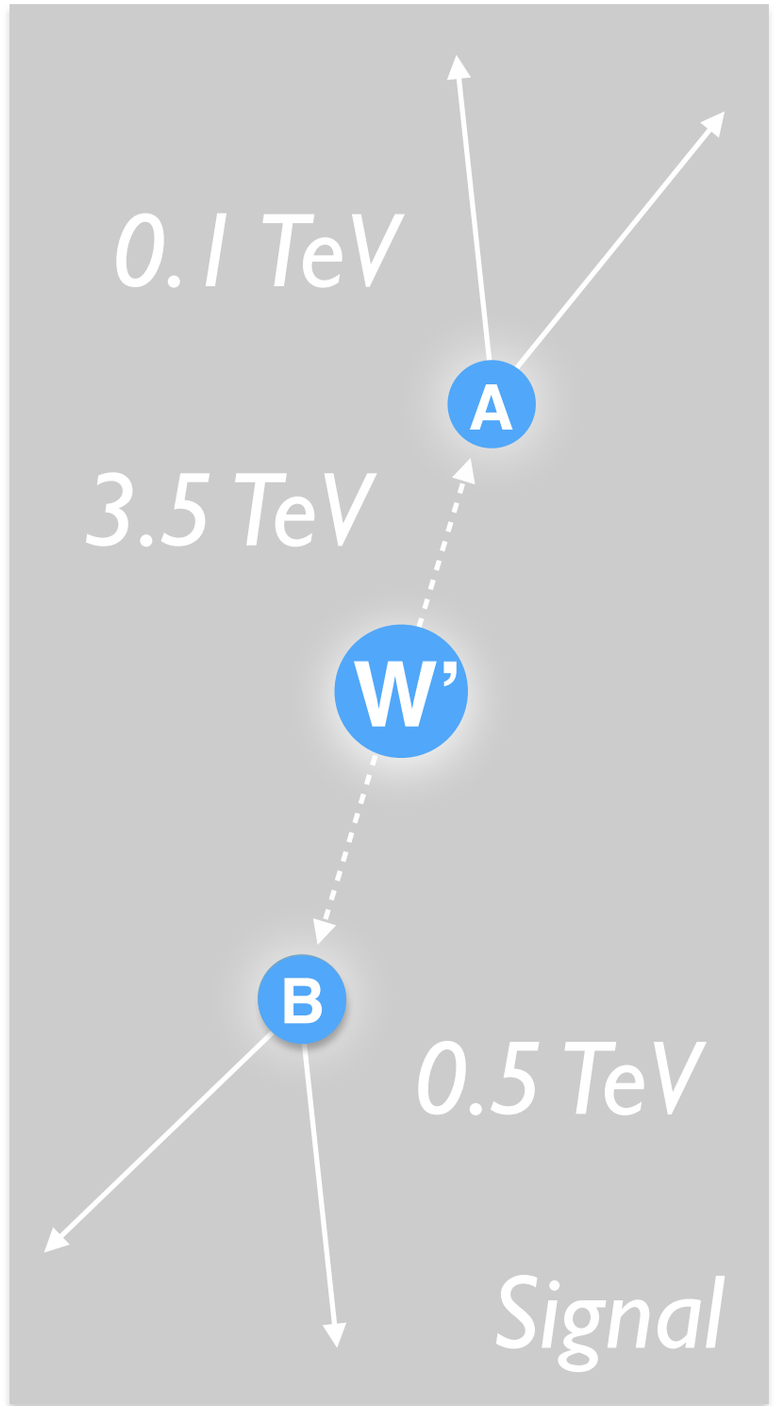


y = many features of the two jets

New ideas: brief illustration

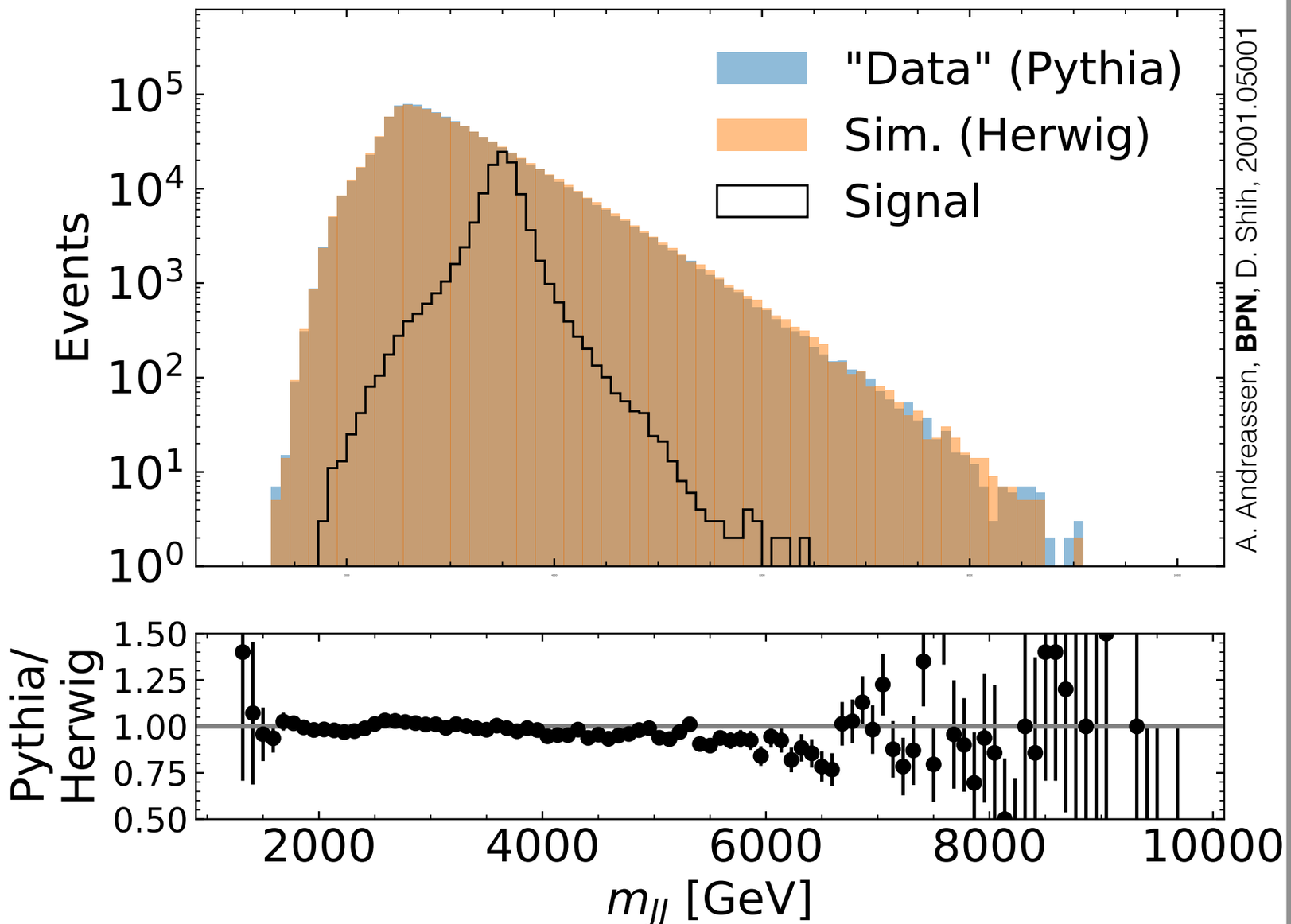
Enhancing the bump hunt



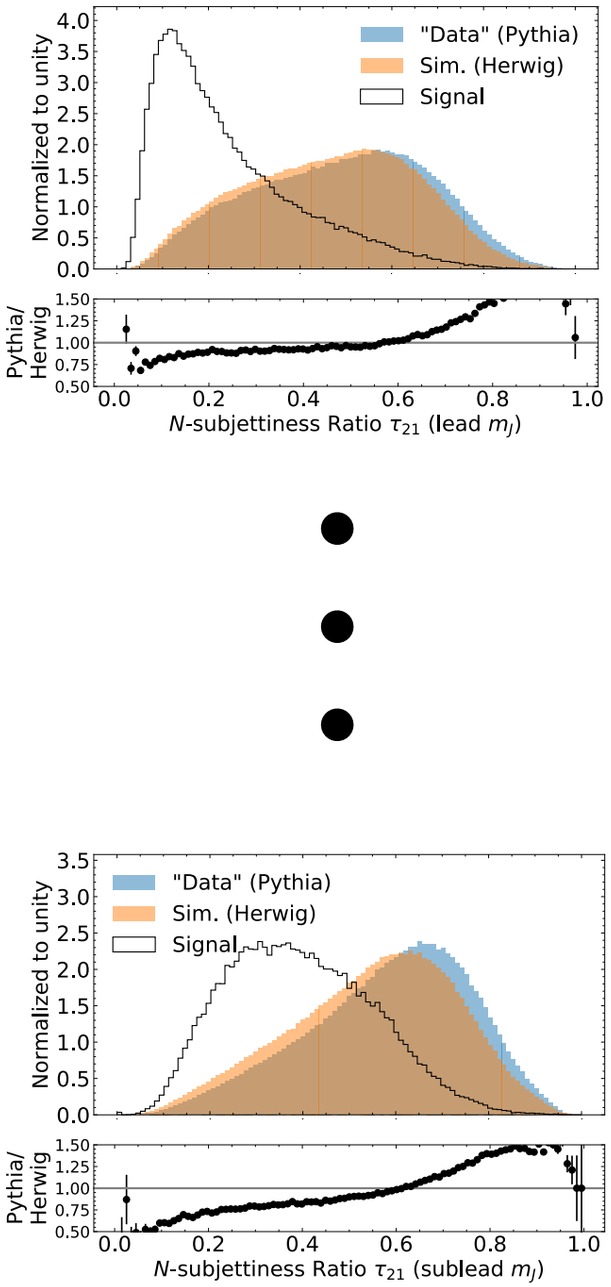


New ideas: brief illustration

Enhancing the bump hunt

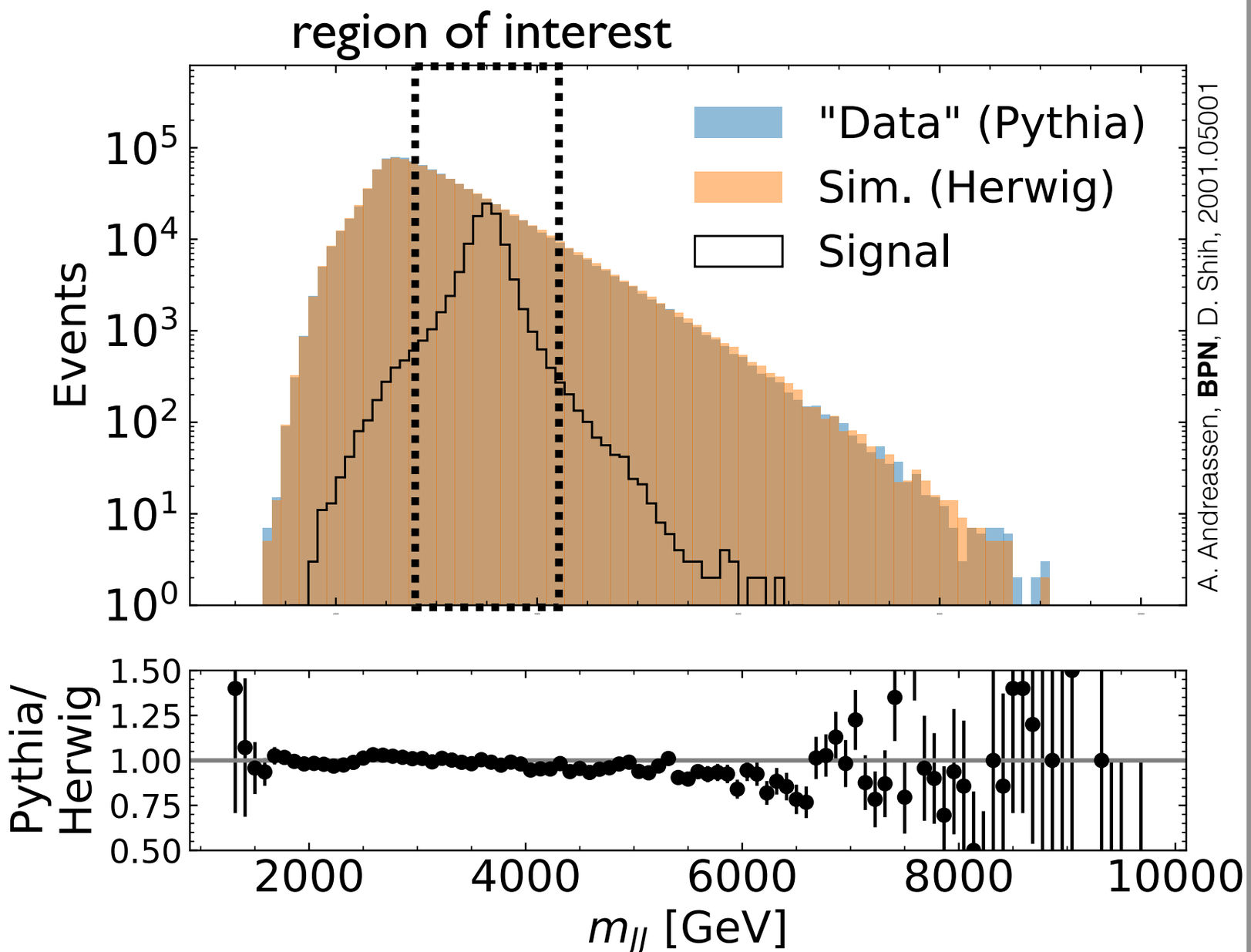


Event/jet Features

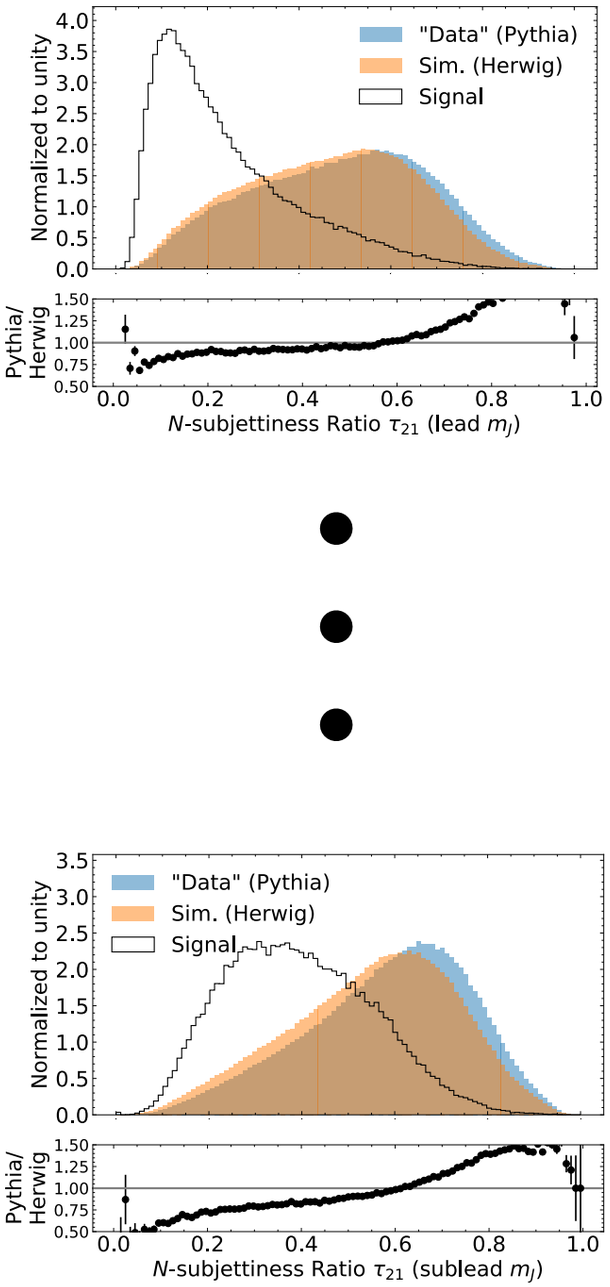


New ideas: brief illustration

Enhancing the bump hunt

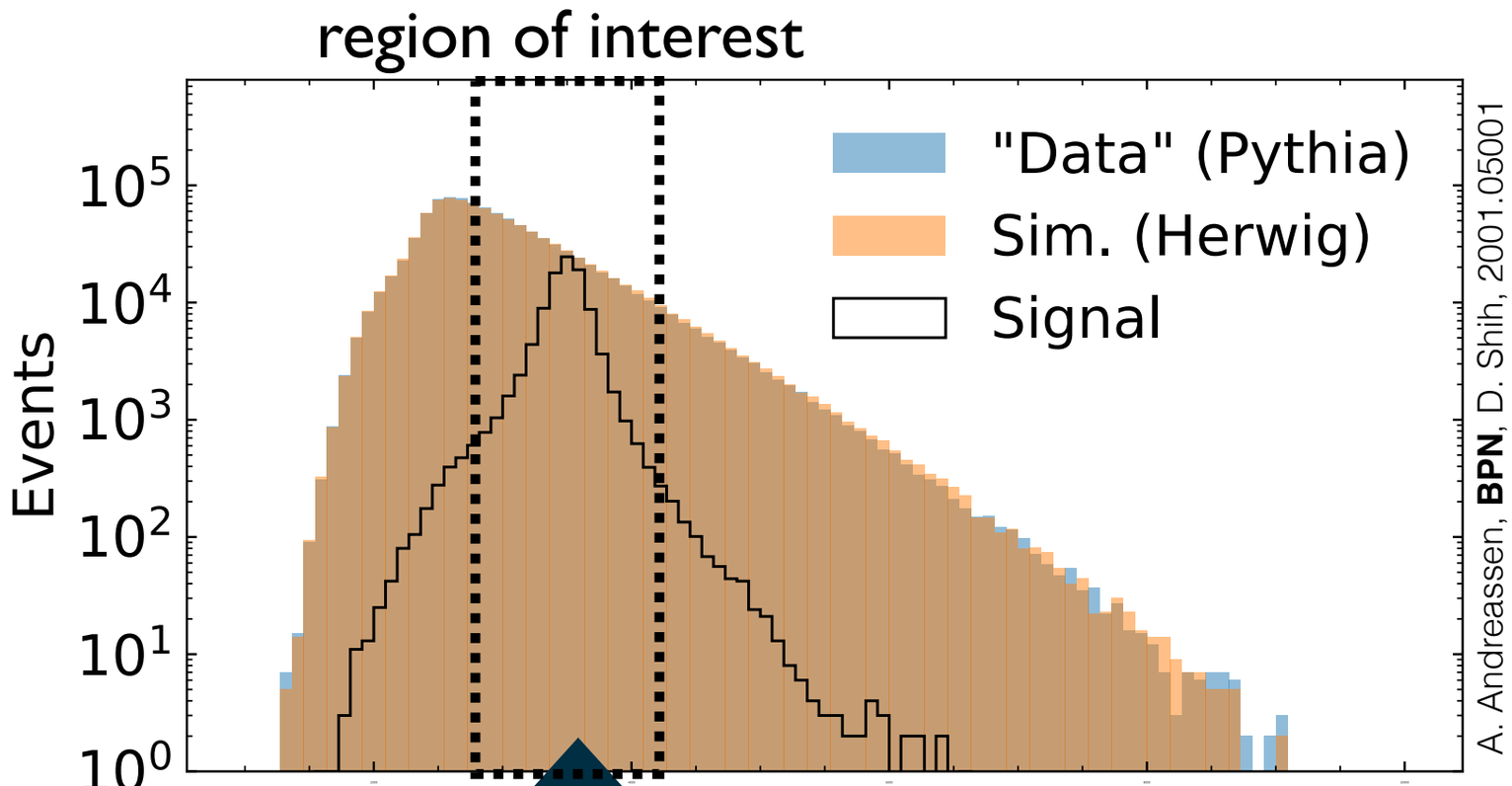


Event/jet Features



New ideas: brief illustration

Enhancing the bump hunt



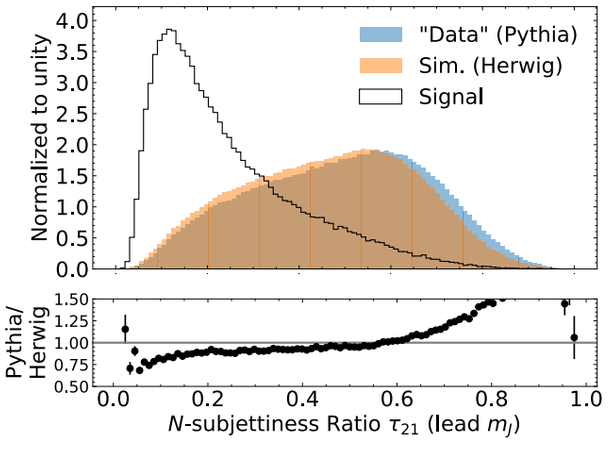
Learn p(features|ROI)/p(features|side)

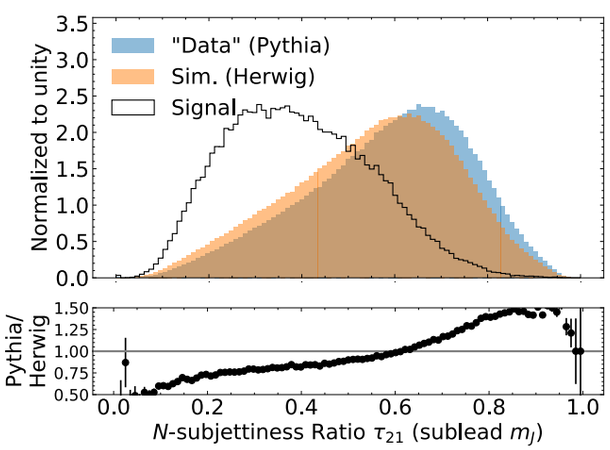
CWoLa++: 1805.02664, 1902.02634, 2002.12376

Train outside ROI (or not) & pick "weird" events

Autoencoders: 1808.08992, 1808.08979 et al.

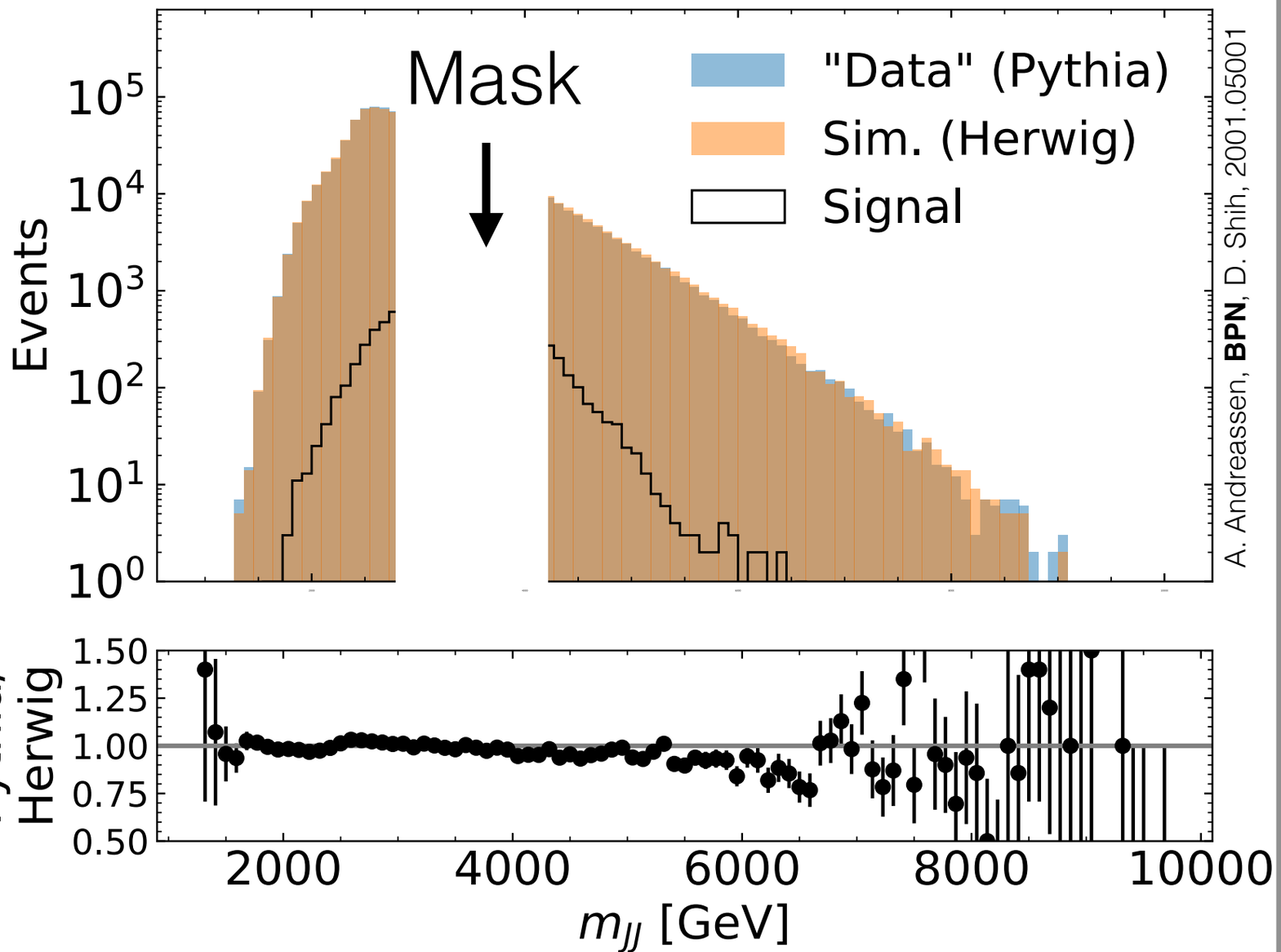
Event/jet Features



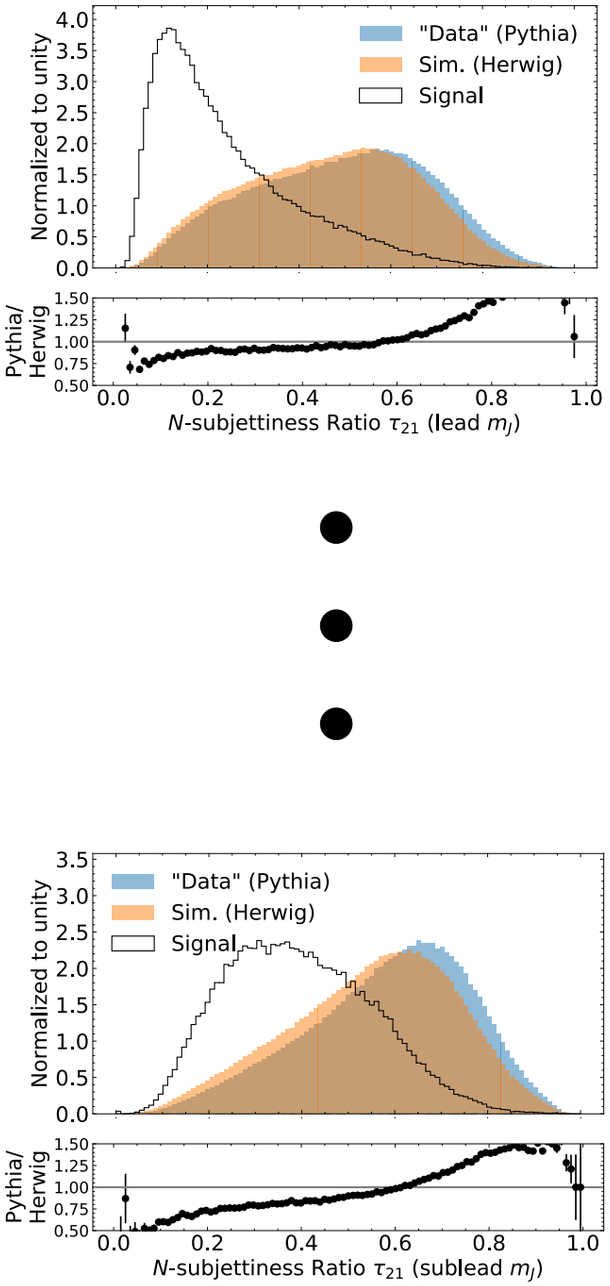


New ideas: brief illustration

Enhancing the bump hunt

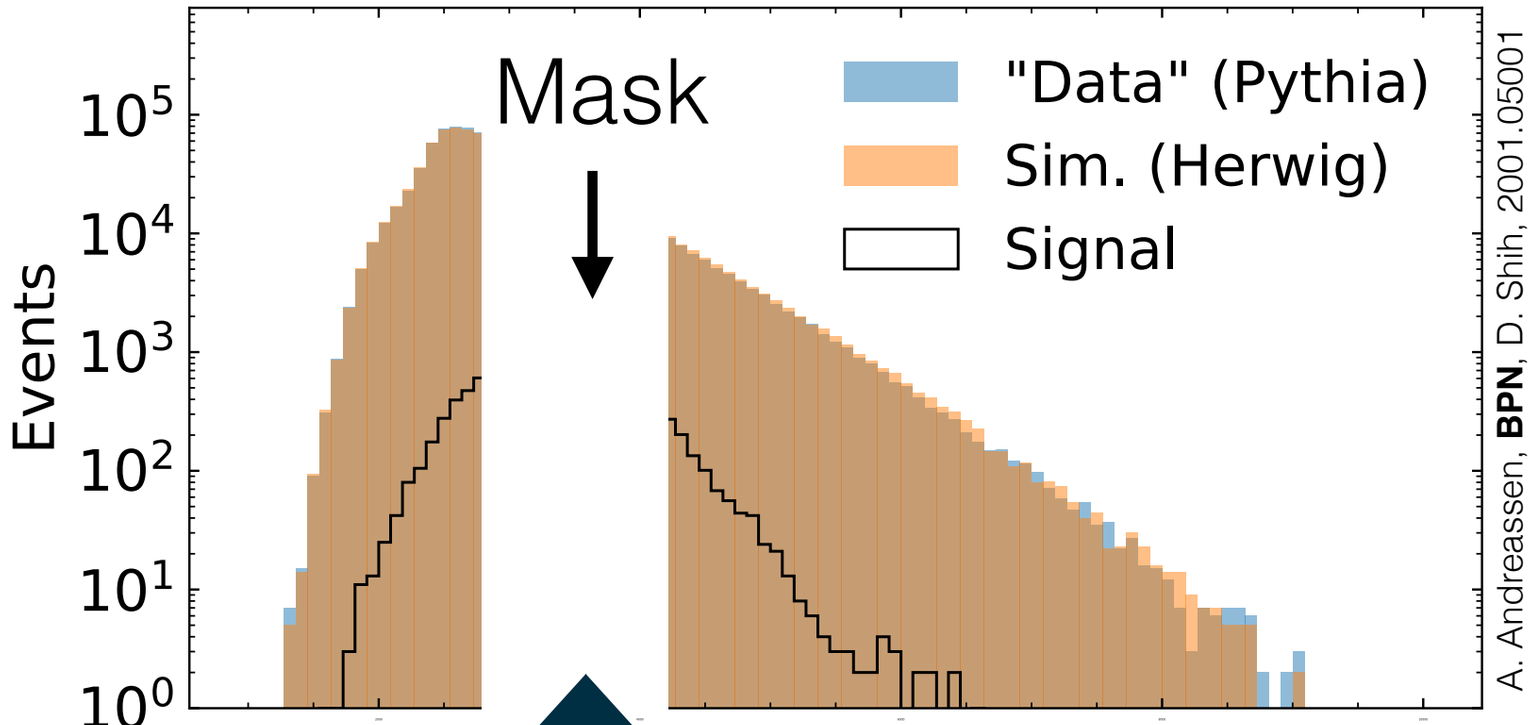


Event/jet Features



New ideas: brief illustration

Enhancing the bump hunt



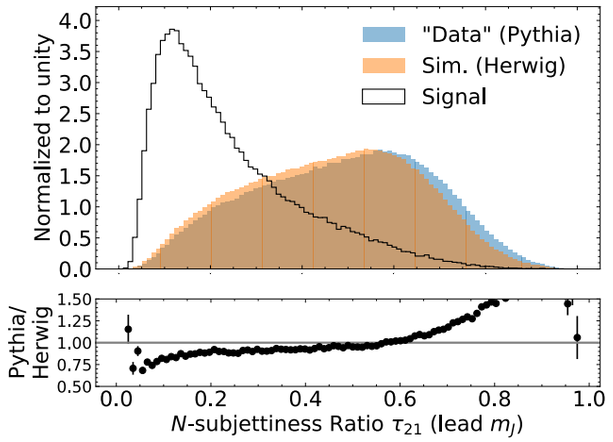
Learn p(features|m) in sideband (SB) & interpolate
+ learn p(features|m) in the masked region

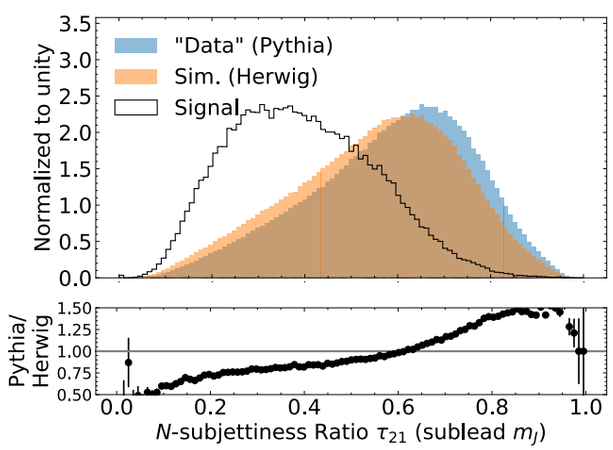
2001.04990 (conditional density estimation / ANODE)

Learn p_{data}(features|m)/p_{MC}(features|m) in SB

2001.05001 (likelihood-free / SALAD)

Event/jet Features

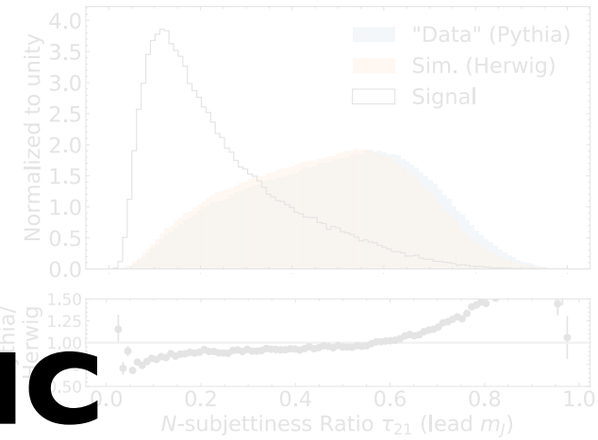




New ideas: brief illustration

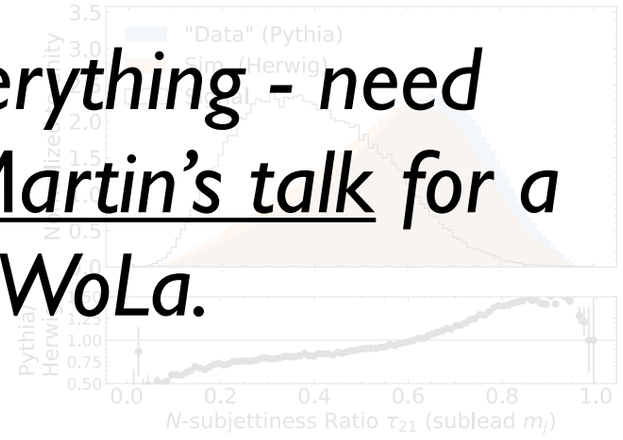
Enhancing the bump hunt

Event/jet Features



+More ideas were developed for the LHC Olympics !

It is likely that no one method will cover everything - need many ideas for broad sensitivity. See also P. Martin's talk for a comparison of autoencoders and CWoLa.



LHC Olympics 2020

Phase I: R&D https://doi.org/10.5281/zenodo.2629072

Released spring 2019

Phase II: Black Box 1 https://doi.org/10.5281/zenodo.3547721

Winter Olympics @ ML4Jets, NYU, Jan 2020

Phase III: Black Box II-IV (boxed 2-3 are already on zenodo)

Summer Olympics @ Hamburg, July 18, 2020

LHC Olympics 2020: Black Boxes

We have prepared three black boxes of simulated data:

- 1 million events each
- 4-vectors of every reconstructed particle (hadron) in the event
- Particle ID, charge, etc. not included
- Single R=1 jet trigger p_T>1.2 TeV
- Black boxes are meant to be representative of actual data, meaning they are mostly background and may contain signals of new physics

In addition, a sample of 1M QCD dijet events (produced with Pythia8 and Delphes3.4.1) was provided as a background sample.

LHC Olympics 2020: Submission

A p-value associated with the dataset having no new particles (null hypothesis).

Short answer text

As complete a description of the new physics as possible. For example: the masses and decay modes of all new particles (and uncertainties on those parameters).

Short answer text

How many signal events (+uncertainty) are in the dataset (before any selection criteria).

Short answer text

Please consider submitting plots or a Jupyter notebook! (these will be private and used only for the presentation / documentation at the end)

Add file

https://docs.google.com/forms/d/e/1FAIpQLScw323fa9qpLbdMvGtr2YeqcGTjE5Zm18-umIDiPldi_cWxVA/viewform

LHC Olympics 2020: Teams

- 10 groups submitted results on box 1
- 4 of these groups also submitted results on boxes 2 & 3
- A number of additional groups could not finish the challenge in time but got results on the R&D dataset
- 7 of these groups gave talks about their methods and results at the ML4Jets2020 conference
- See the indico page for details. Note that we did not reveal the answer until the end of the session !

LHC Olympics 2020: Teams (alphabetical order)

- Oz Amram and Cristina Mantilla Suarez (Johns Hopkins)
- Barry Dillon, D. Faroughy, J. Kamenik, M. Swezc (Institue Jožef Stefan)
- Cosmos Dong (Reed)
- Julien Donini, Ioan-Mihail Dinu, Louis Vaslin (LPClermont, IFIN-HH)
- Felipe F. De Freitas (Beijing), Charanjit K. Khosa, Veronica Sanz (Sussex)
- Gustaaf Brooijmans, Julia Gonski, Alan Kahn, Inês Ochoa, Daniel Williams (Columbia)
- Patrick Komiske, Eric Metodiev, Nilai Sarda, Jesse Thaler (MIT)
- Christopher W. Murphy (Data Science)
- Soroosh Shalileh (HSE, Russia)
- George Stein, Uros Seljak, Biwei Dai (Cosmo, Berkeley)

Congratulations to all teams for braving the challenge !

LHC Olympics 2020: Strategies

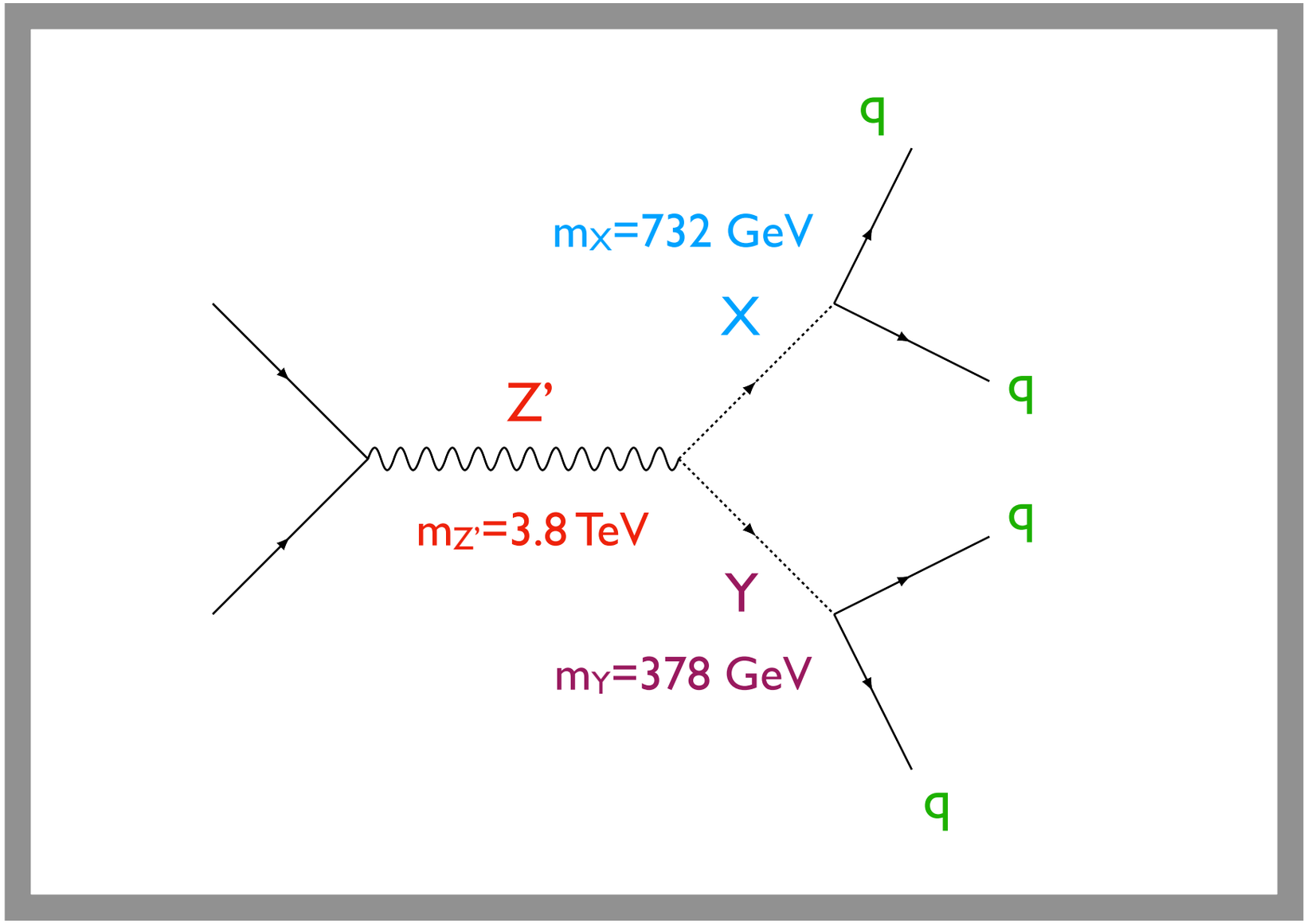
People tried both supervised and unsupervised methods.

Methods used included:

- Autoencoders
- CWoLa hunting
- PCA outlier detection
- LSTM
- CNN+BDT
- variational RNNs for anti-QCD tagging
- density estimation
- biological neural network
- …

LHC Olympics 2020: Box 1

LHC Olympics 2020: Box 1



834 events. Same topology as R&D dataset
(not known to participants)

Results

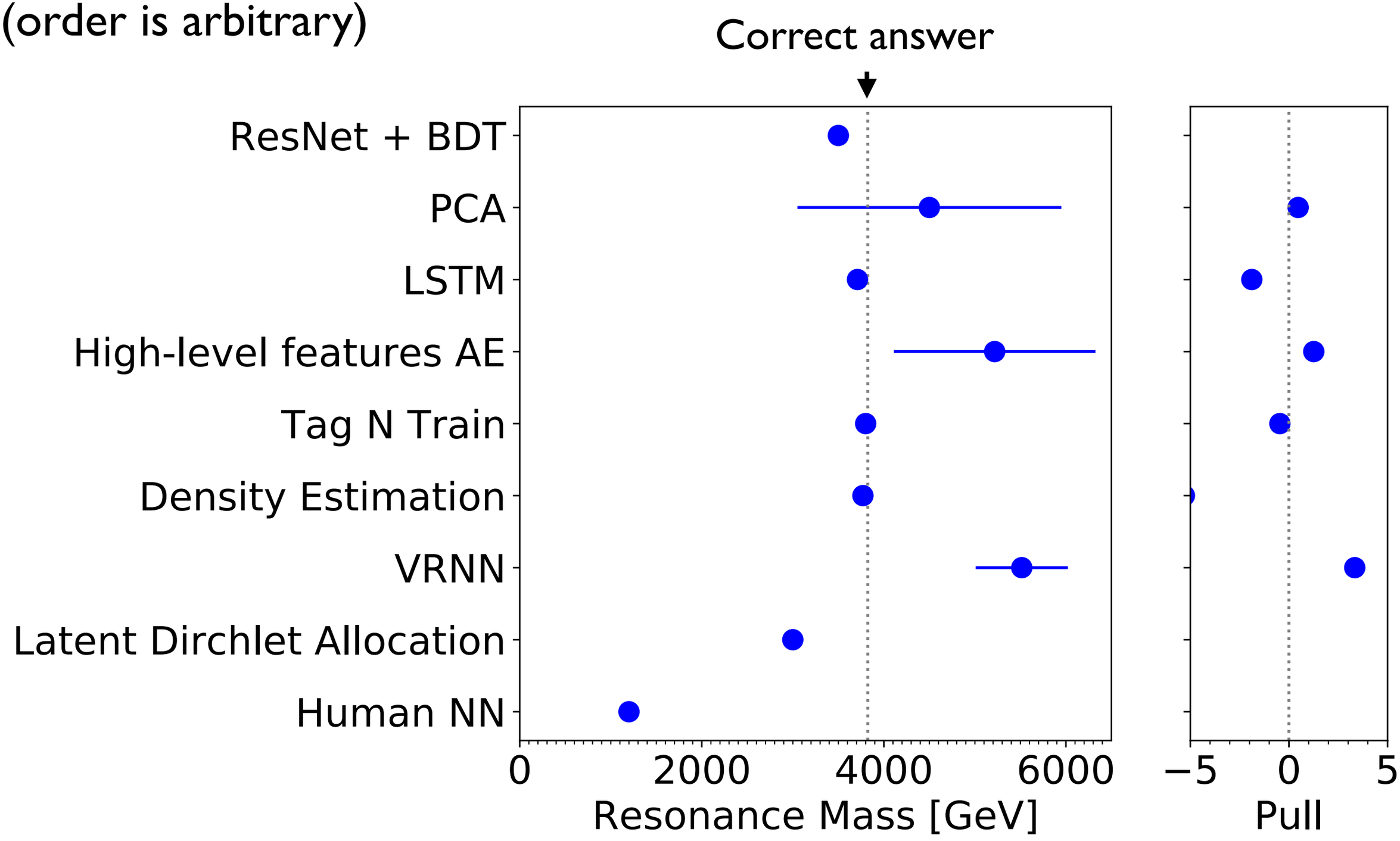
(order is arbitrary)

ResNet + BDT	[slides]
PCA	*Principal component analysis was used as an outlier detector for resonant new physics. Given the popularity of deep learning today, it could be useful to have a shallow baseline to compare against. Features were selected by requiring that they, individually, produce a ROC AUC greater than one-half (a trivial condition to satisfy in supervised learning) on the practice dataset.*
LSTM	*I cluster the hadrons into anti-kT R = 0.7 jets, and run the sequence of jets into a simple RNN (supervised learning). The sequence is ordered by the jets' p_T from high to low, zero-padding up to 10…I made two LSTM layers, with some dense layers after and a single output identifying signal or not.*
High-level features AE	[slides]
Tag N Train	[slides]
Density Estimation	[slides]
VRNN	[slides]
Latent Dirchlet Allocation	[slides]
Human NN	*Look at many histograms*

Topic modeling	[slides] *(also preliminary black box 1 results)*	*R&D and non-LHCO results*
VAE	[slides]	

Results - resonance mass

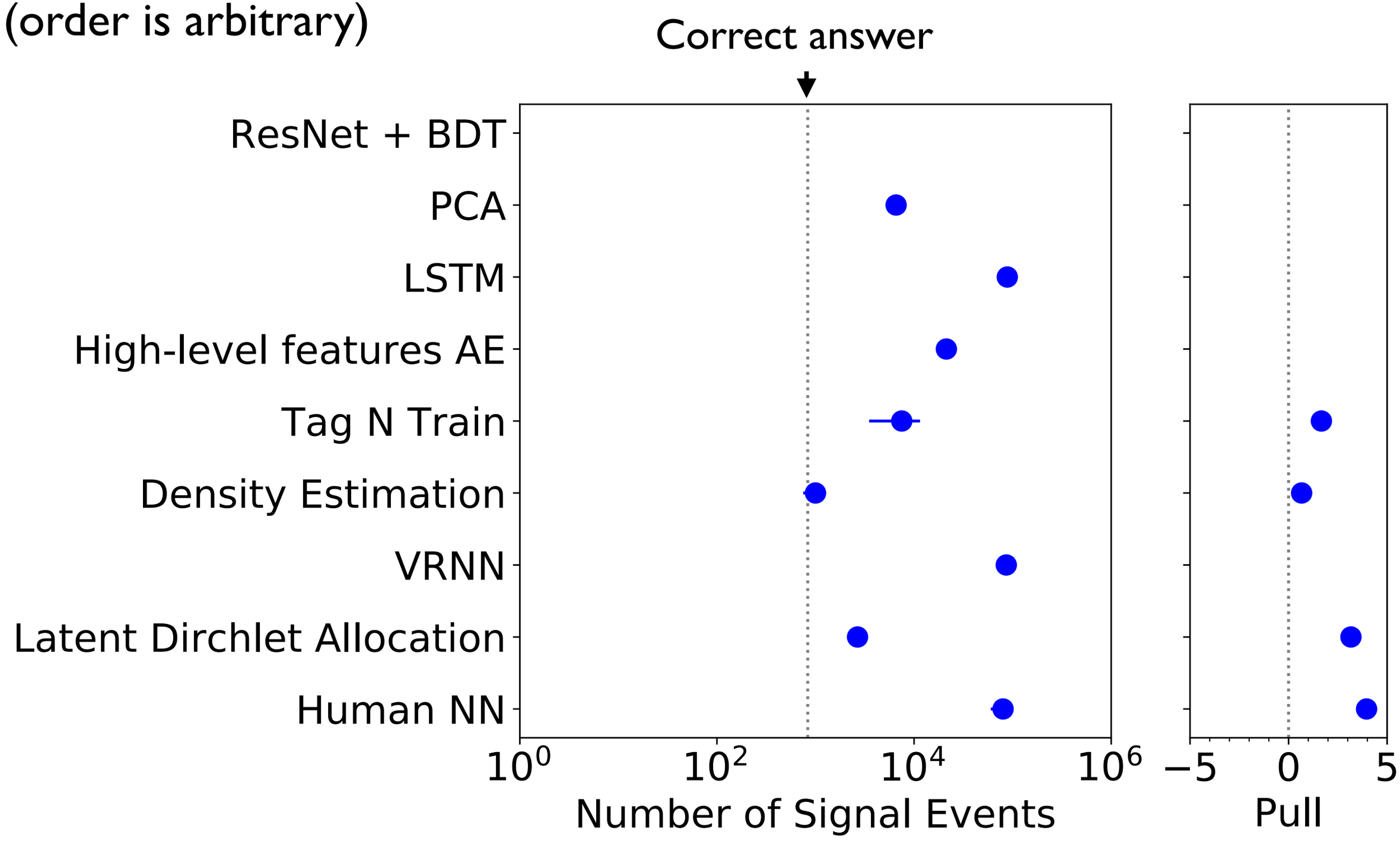
(order is arbitrary)



N.B. not everyone reported an uncertainty

Results - number of events

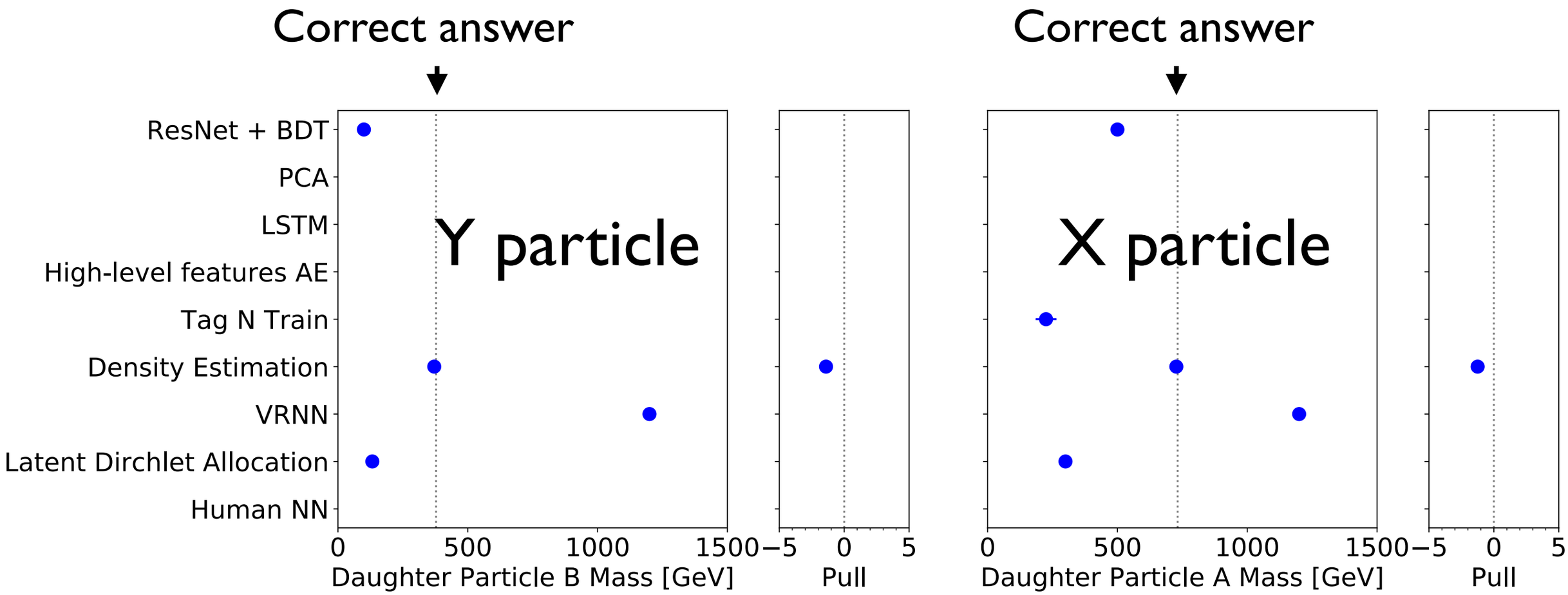
(order is arbitrary)



N.B. not everyone reported an uncertainty

Results - daughter masses

(order is arbitrary)



N.B. Not everyone reported the daughter masses.

…and the winners are…

We refrained from ranking the results … all methods are an important contribution to this growing research area.

However, two submissions clearly stood out:

Conditional density estimation for anomaly detection

George Stein, Uros Seljak, Biwei Dai, He Jia

BERKELEY CENTER *for* COSMOLOGICAL PHYSICS

Tag N' Train (CWoLa + Autoencoder)

Oz Amram and Cristina Mantilla Suarez

JOHNS HOPKINS UNIVERSITY

Summer Olympics

Stay tuned for more on the LHCO 2020…

We will be organizing a 1-day mini-workshop on anomaly detection in Hamburg the Saturday before BOOST (July 18).

The answers for Boxes 2 and 3 will be revealed.

We will also discuss plans for a community paper on new methods for anomaly detection and the LHCO2020.

Please come and join us!

http://indico.desy.de/indico/e/anomaly2020

Conclusions and outlook

These are exciting times for anomaly detection in HEP.

Many new approaches making use of unsupervised ML are being developed by theorists and experimentalists.

Model independent searches have a bright future at the LHC. Maybe this is how we will finally discover the new physics !

These methods also have potential applications beyond HEP. There have already been connections with other fields in the winter olympics !