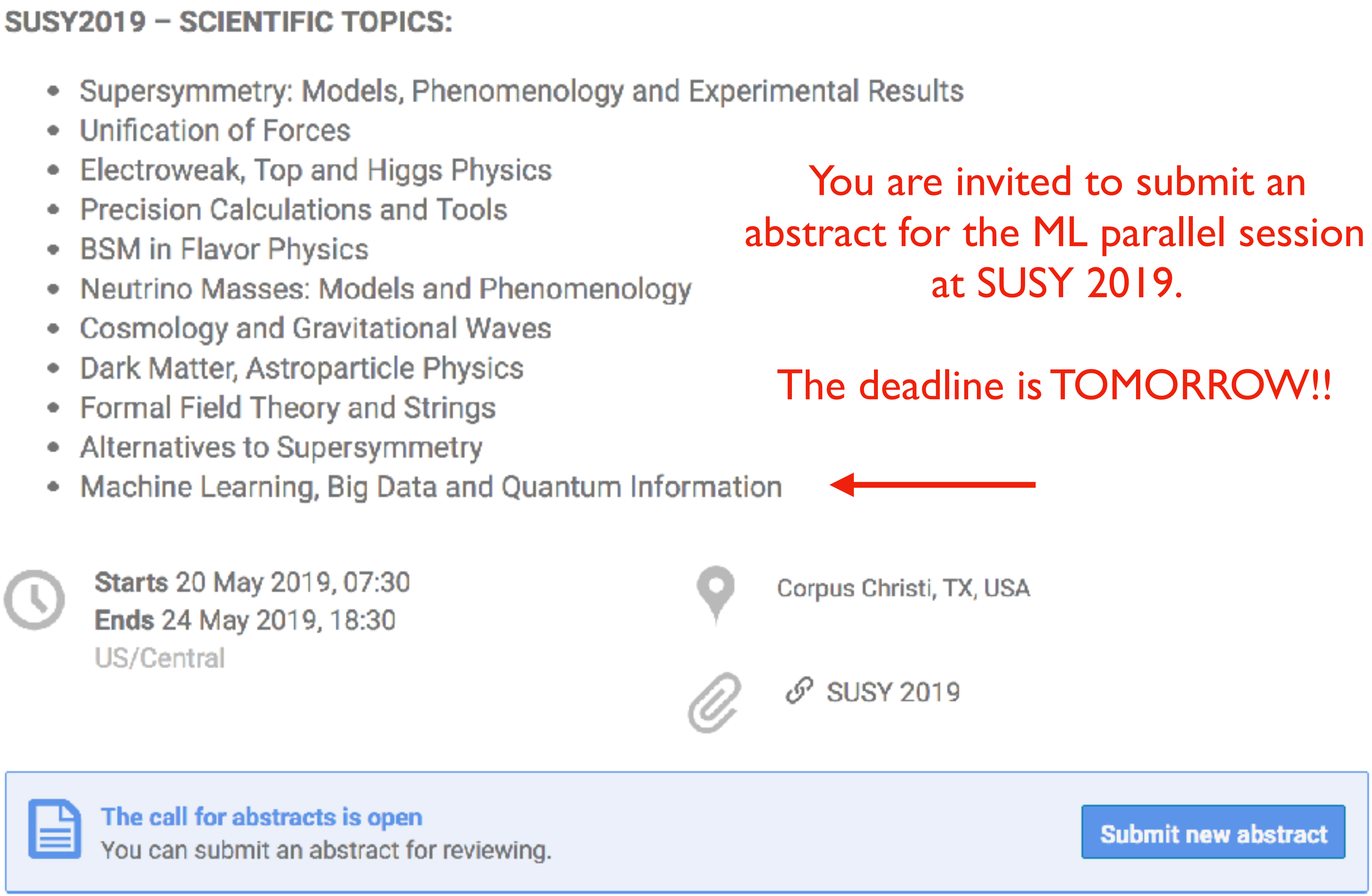
Boosting New Physics Searches with Deep Learning

David Shih
NHETC, Rutgers University

Accelerating the Search for Dark Matter with Machine Learning
ICTP, Trieste

April 9, 2019

Announcement



SUSY2019 – SCIENTIFIC TOPICS:

- Supersymmetry: Models, Phenomenology and Experimental Results
- Unification of Forces
- Electroweak, Top and Higgs Physics
- Precision Calculations and Tools
- BSM in Flavor Physics
- Neutrino Masses: Models and Phenomenology
- Cosmology and Gravitational Waves
- Dark Matter, Astroparticle Physics
- Formal Field Theory and Strings
- Alternatives to Supersymmetry
- Machine Learning, Big Data and Quantum Information

Starts 20 May 2019, 07:30
Ends 24 May 2019, 18:30
US/Central

Corpus Christi, TX, USA

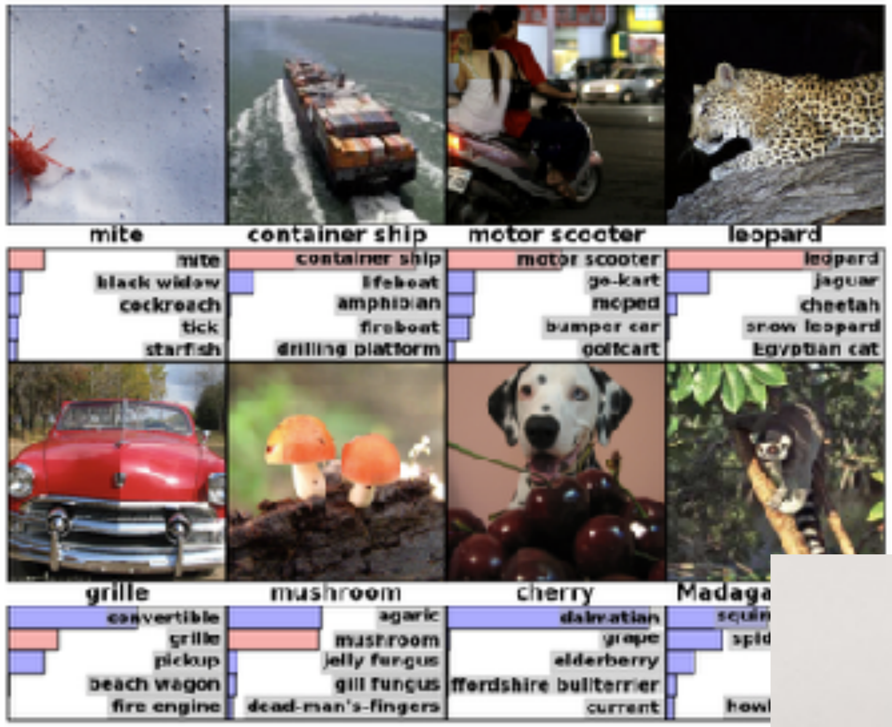
SUSY 2019

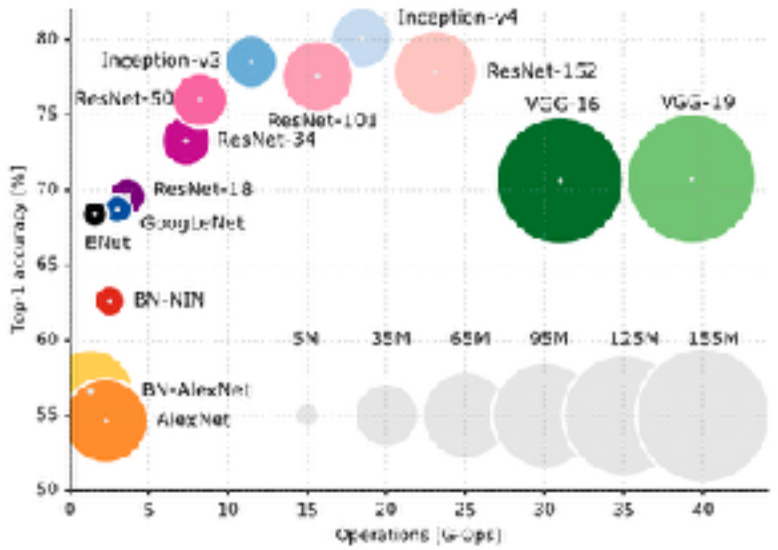
The call for abstracts is open
You can submit an abstract for reviewing.

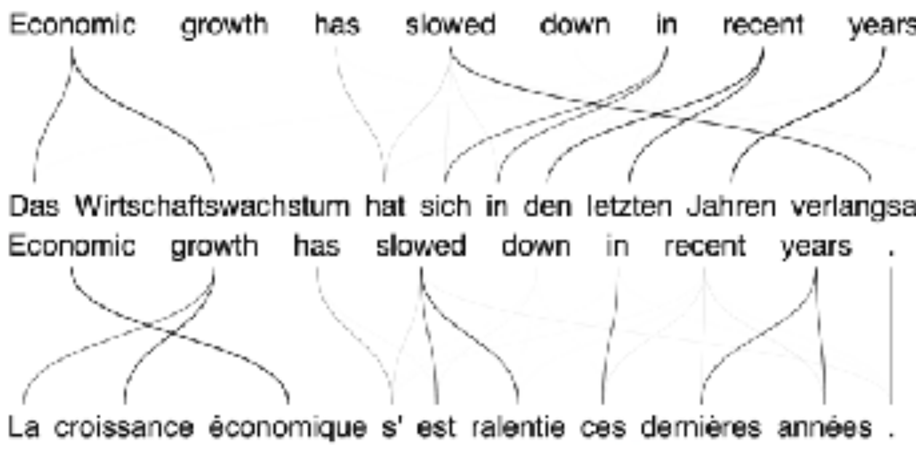
You are invited to submit an abstract for the ML parallel session at SUSY 2019.

The deadline is TOMORROW!!

The AI Revolution is Here





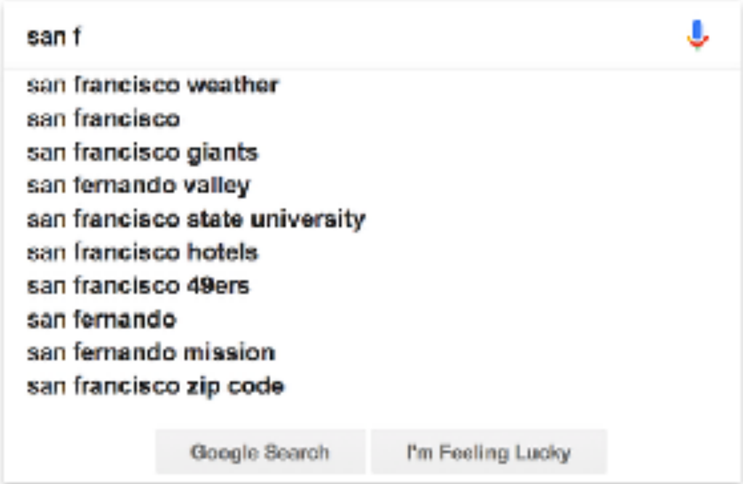


Economic growth has slowed down in recent years

Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsa

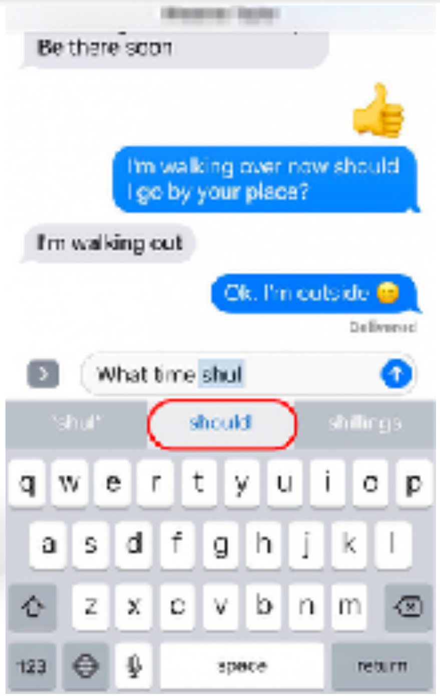
Economic growth has slowed down in recent years .

La croissance économique s' est ralentie ces dernières années .



san f

san francisco weather
san francisco
san francisco giants
san fernando valley
san francisco state university
san francisco hotels
san francisco 49ers
san fernando
san fernando mission
san francisco zip code



Be there soon

I'm walking over now should I go by your place?

I'm walking out

Ok. I'm outside

Delivered

What time shul

The AI Revolution is Here

So many stunning real world successes in recent years.

Driven by:

- Growth in computational power
- Improvements in algorithms
- Increased quantity and quality of data

Prerequisite for deep learning: large, complex, and well-understood datasets.

Many real world applications are limited by the quality and quantity of the data.

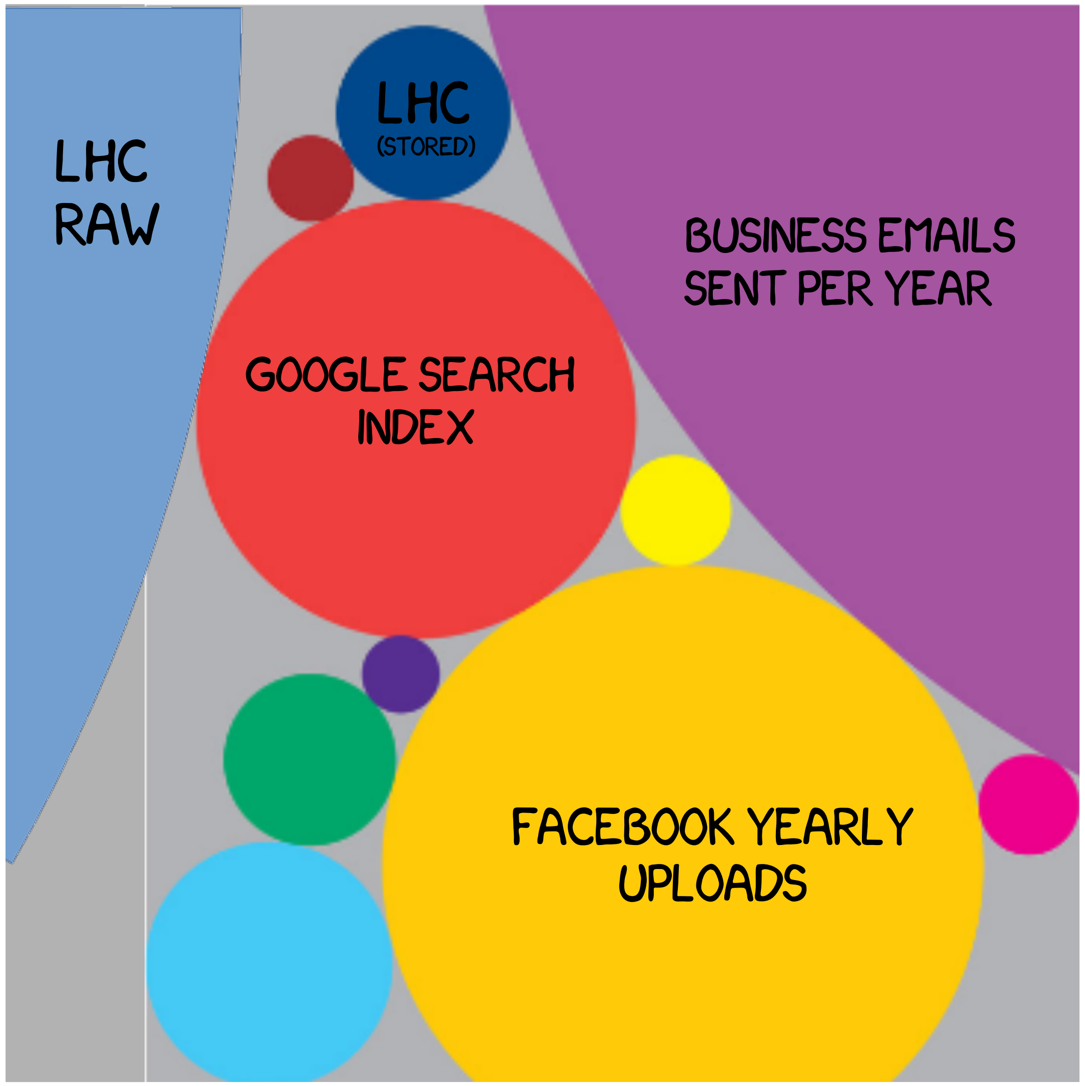
Big Data and Deep Learning

The LHC is the perfect setting for deep learning!

The data is

- large (billions of events on tape)
- complex (hundreds of particles per event)
- well-understood (Standard Model of particle physics).

Also, it is relatively easy to generate realistic simulated data.
(Madgraph, Pythia, Herwig, Delphes, GEANT,…)



https://www.wired.com/2013/04/bigdata/

Pasquale Musella, ETH-Zurich seminar

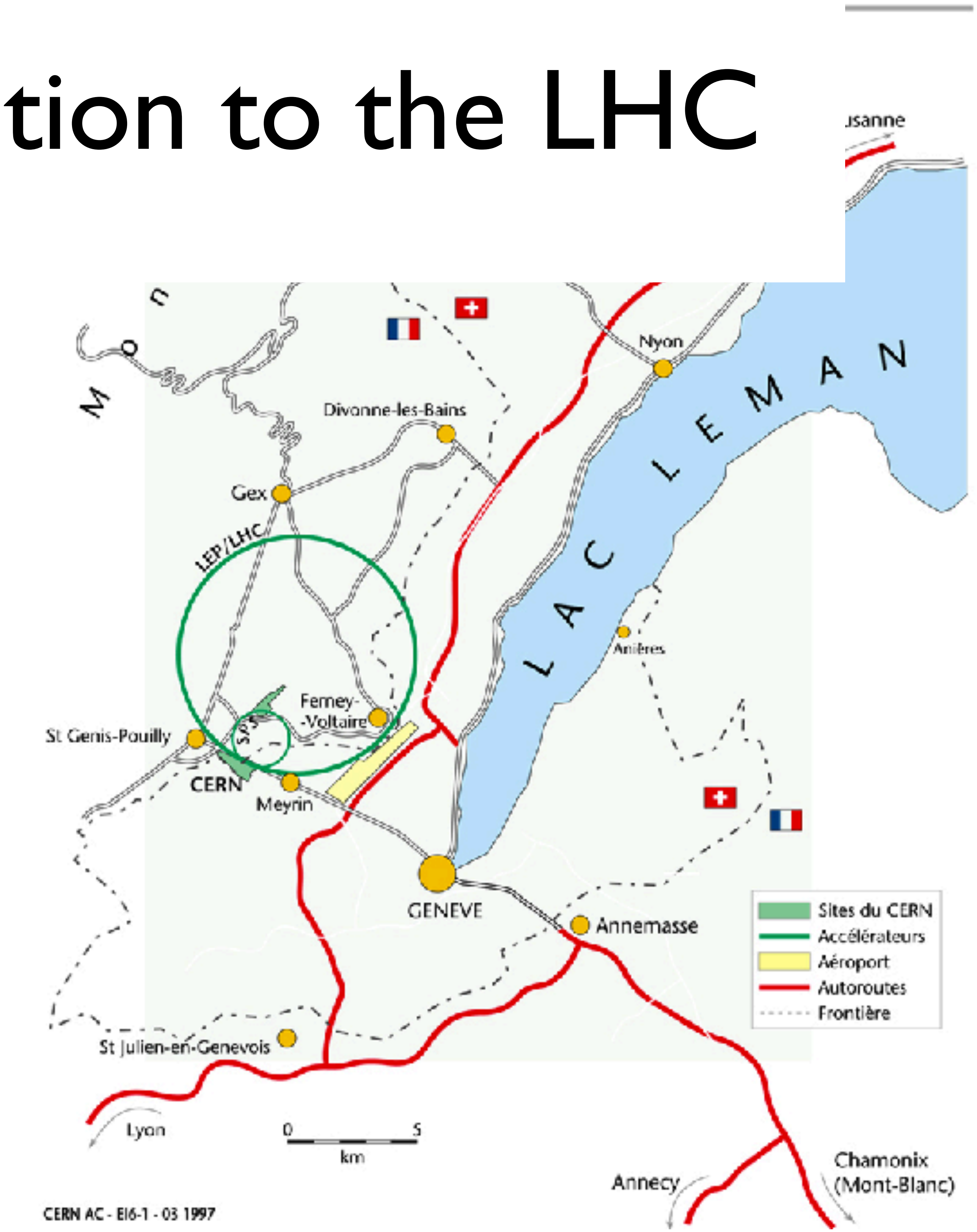
A brief introduction to the LHC

An introduction to the LHC

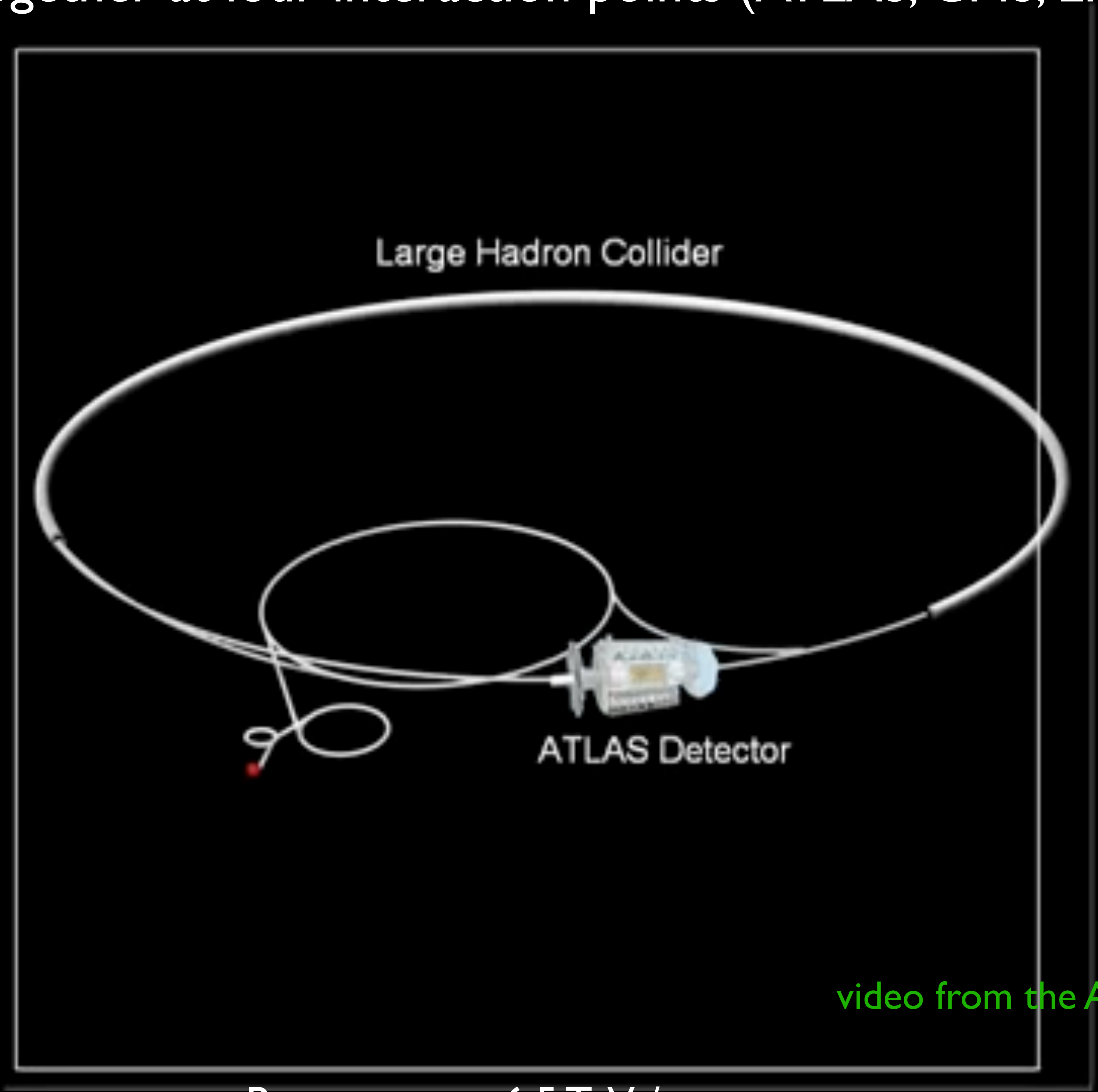
The Large Hadron Collider is the largest and highest-energy particle accelerator in the world.

It is part of CERN, located at the border of France and Switzerland, near the city of Geneva.

- 27 km long tunnel
- 100 m underground
- ~ $10 billion
- ~5,000 scientists from ~200 countries



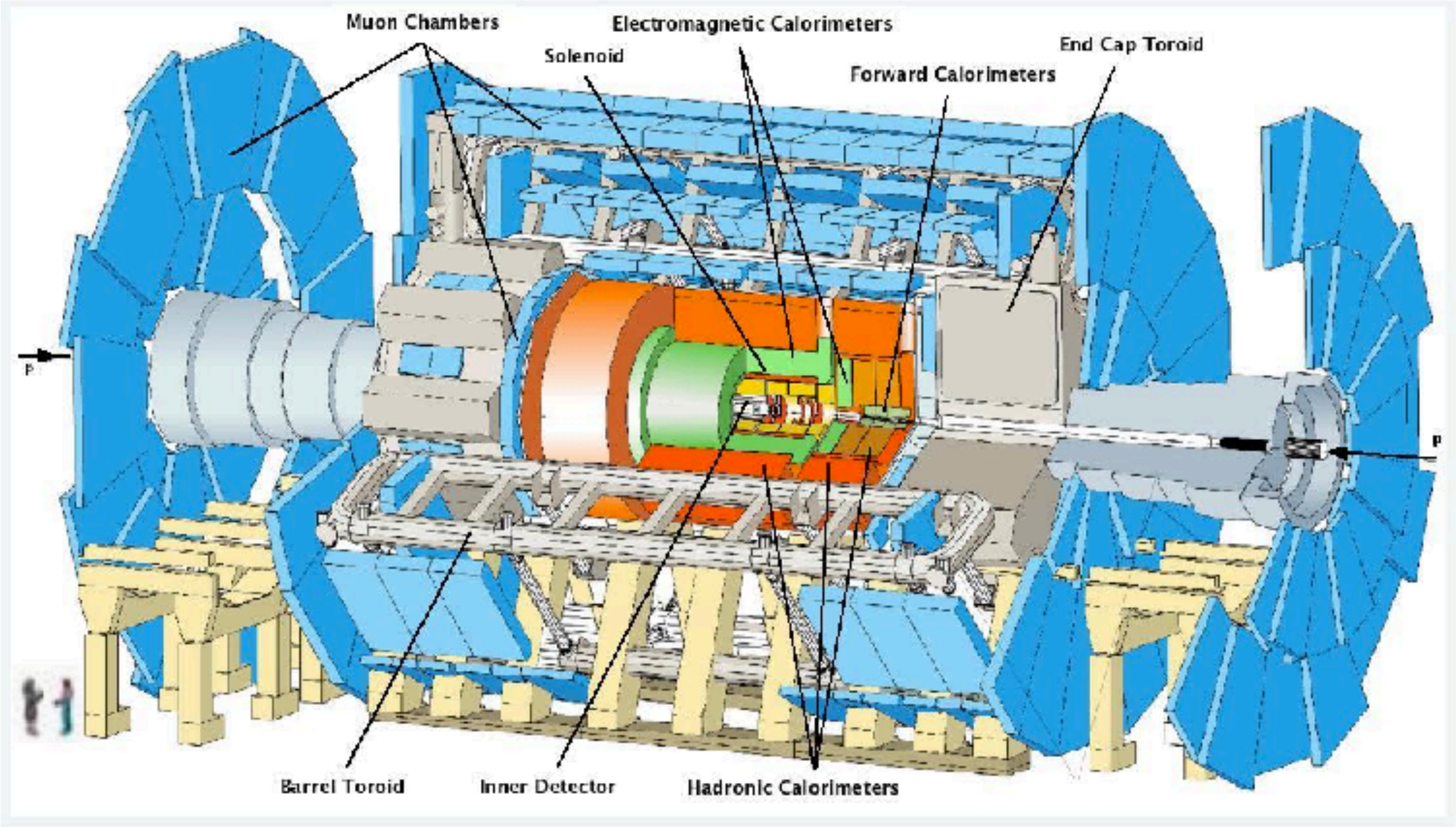
At the LHC, protons are accelerated to 99.9999991% of the speed of light, and collided together at four interaction points (ATLAS, CMS, LHCb, ALICE)



video from the ATLAS experiment

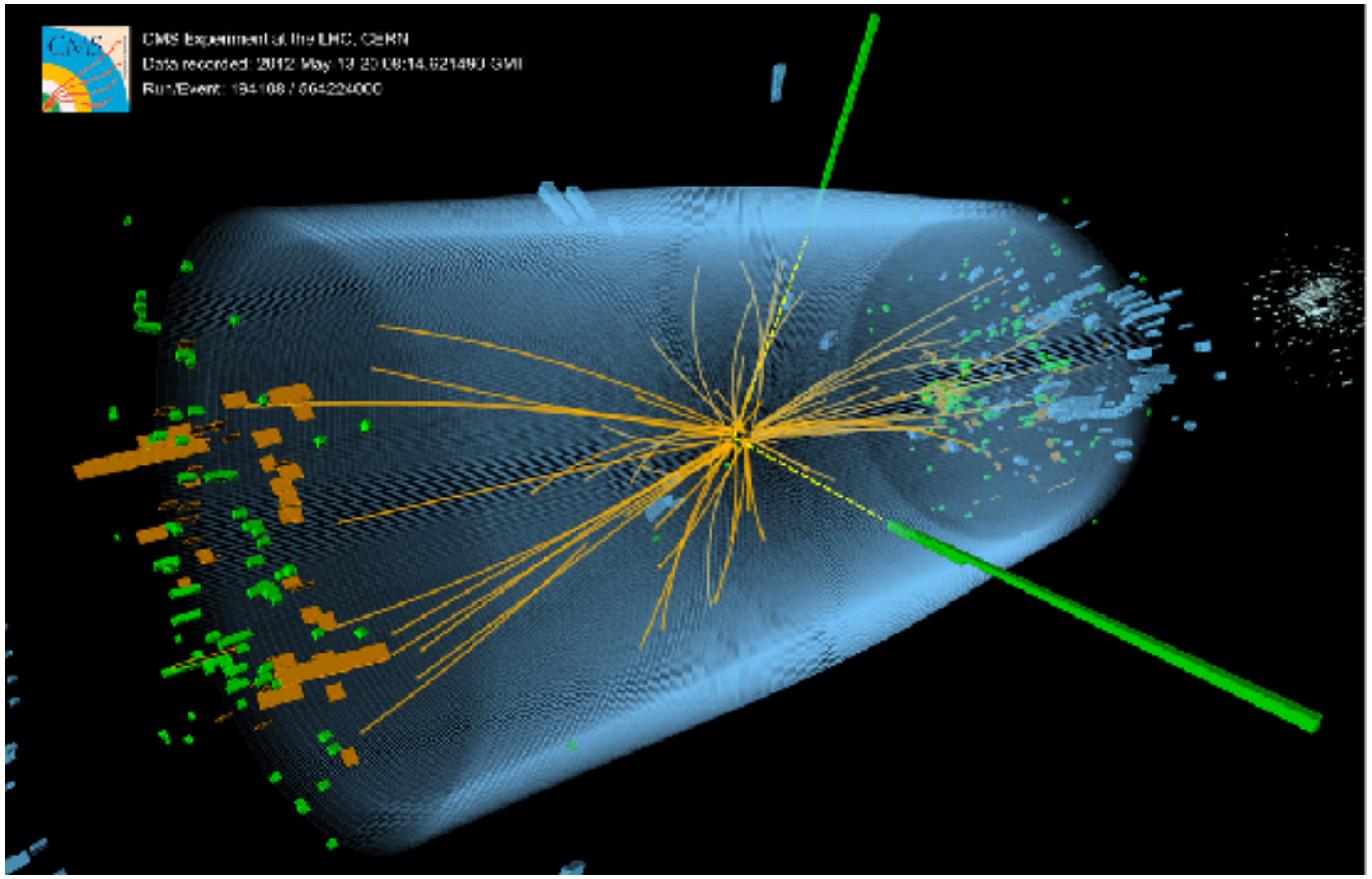
Beam energy: 6.5 TeV / proton
~ 300 trillion protons (in ~3000 bunches) in each beam
25 ns bunch spacing

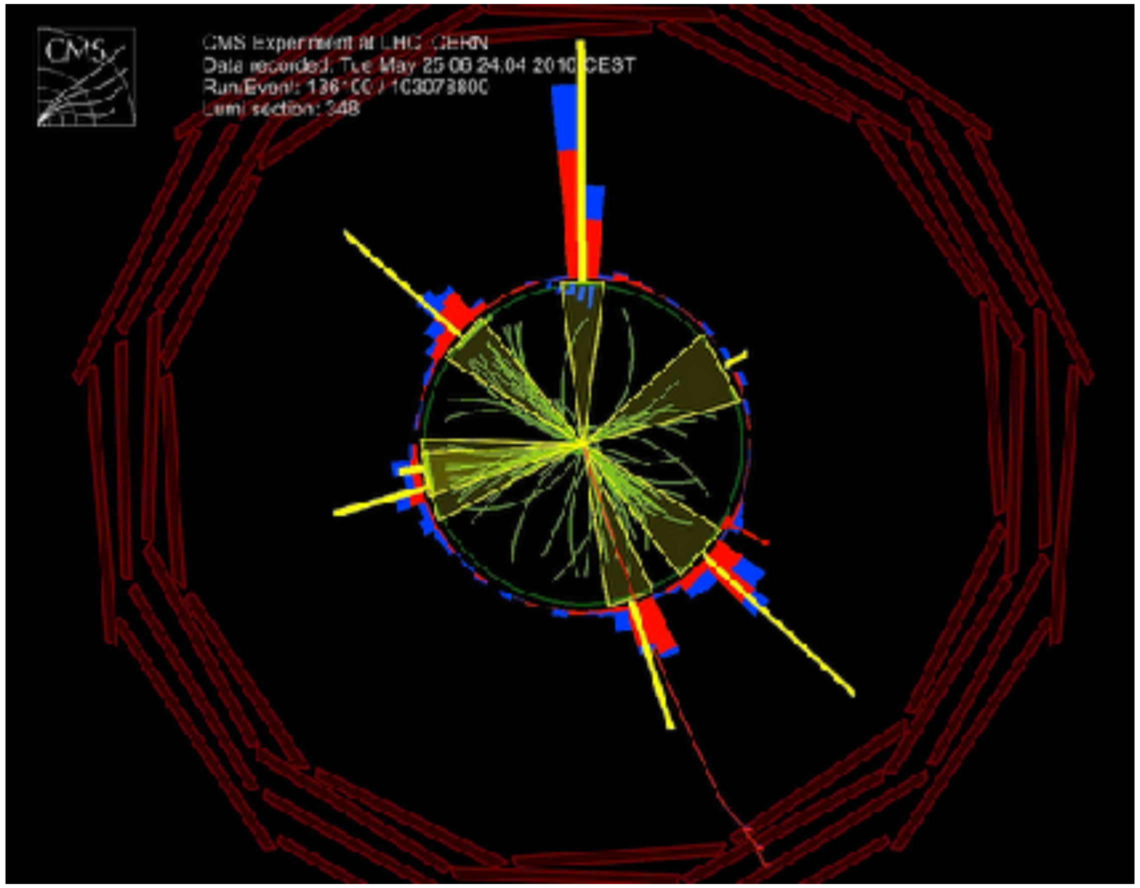
An LHC Detector



Detector is cylindrical (symmetric around beam axis)

Collision events at the LHC


CMS Experiment at the LHC, CERN
Data recorded: 2012 May 13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000

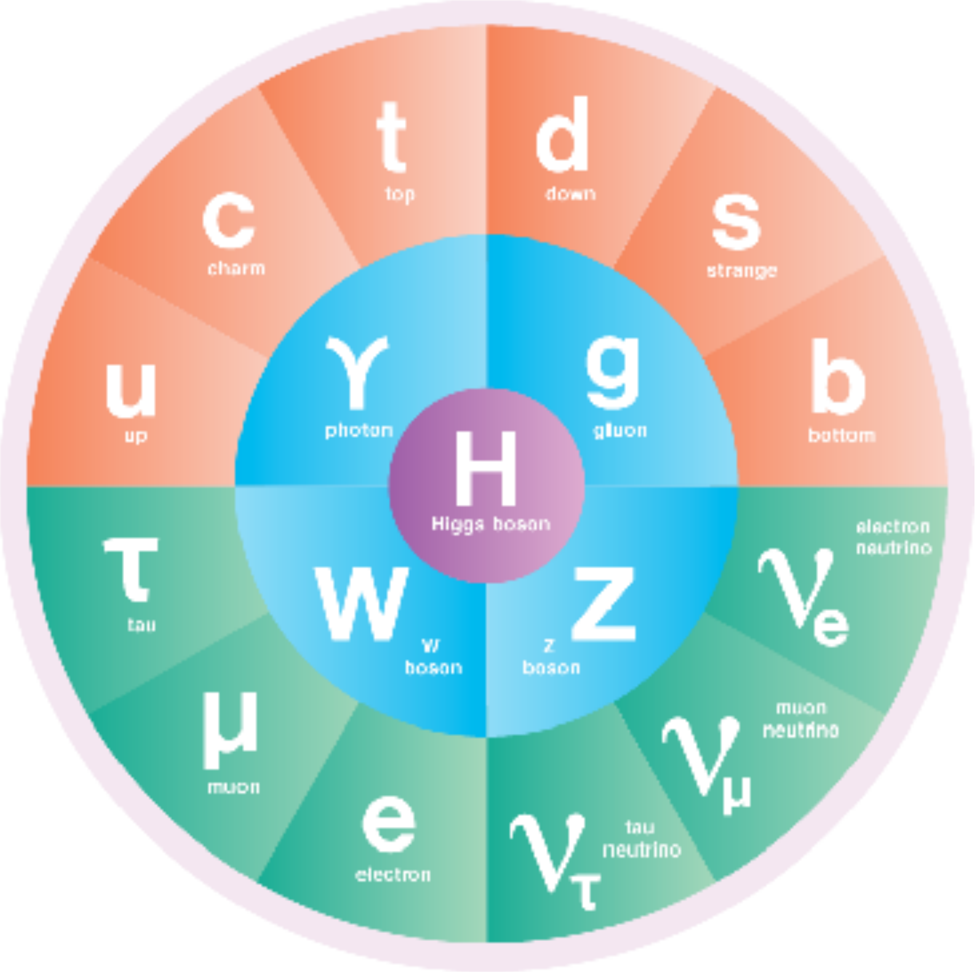

CMS Experiment at LHC, CERN
Data recorded: Tue May 25 06:24:04 2010 CEST
Run/Event: 136100 / 103073800
Lumi section: 348

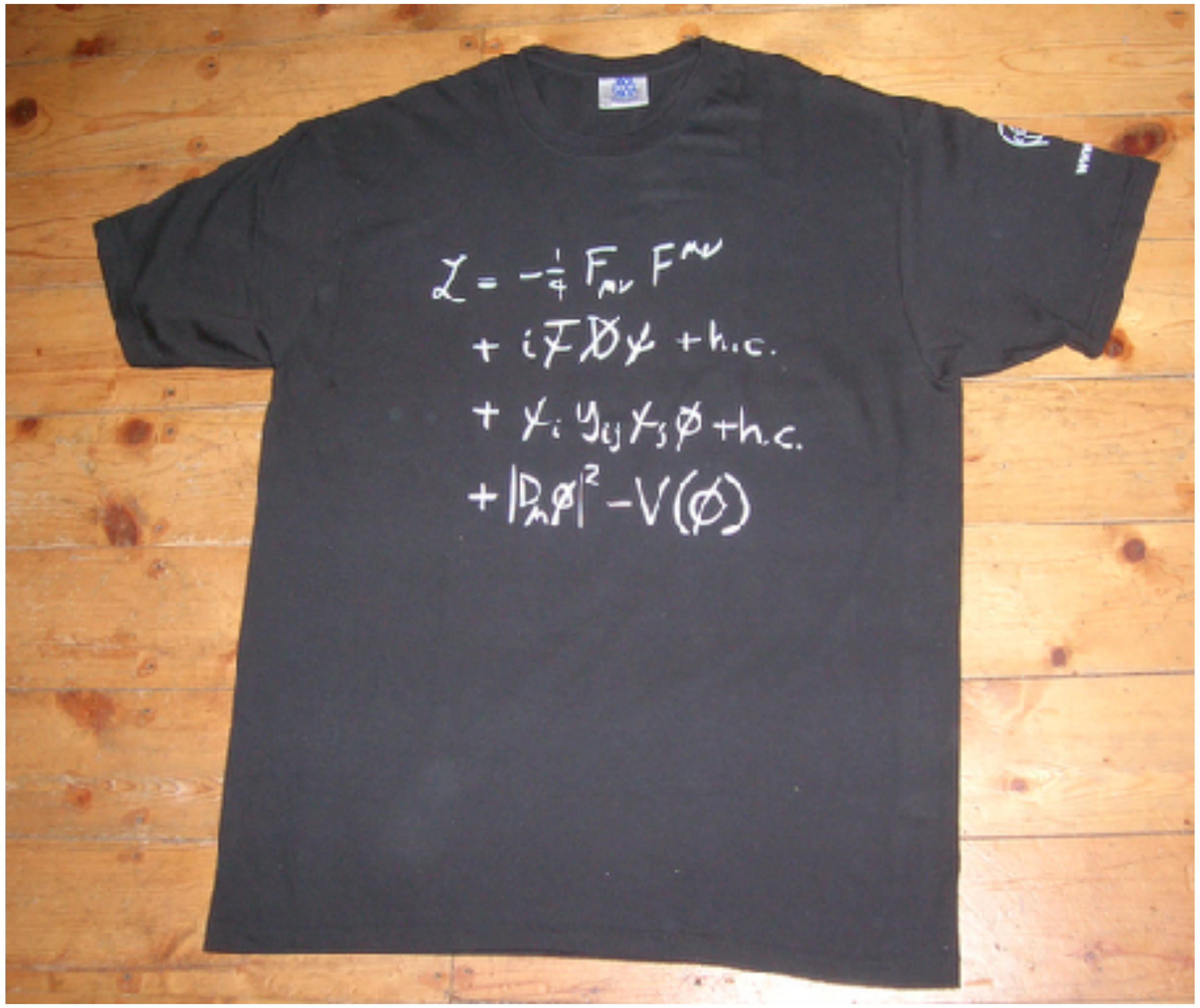
raw event rate ~ GHz => ~ 100 Hz after “triggering”

data rate: ~ 1 GB/s ~ several PB/year

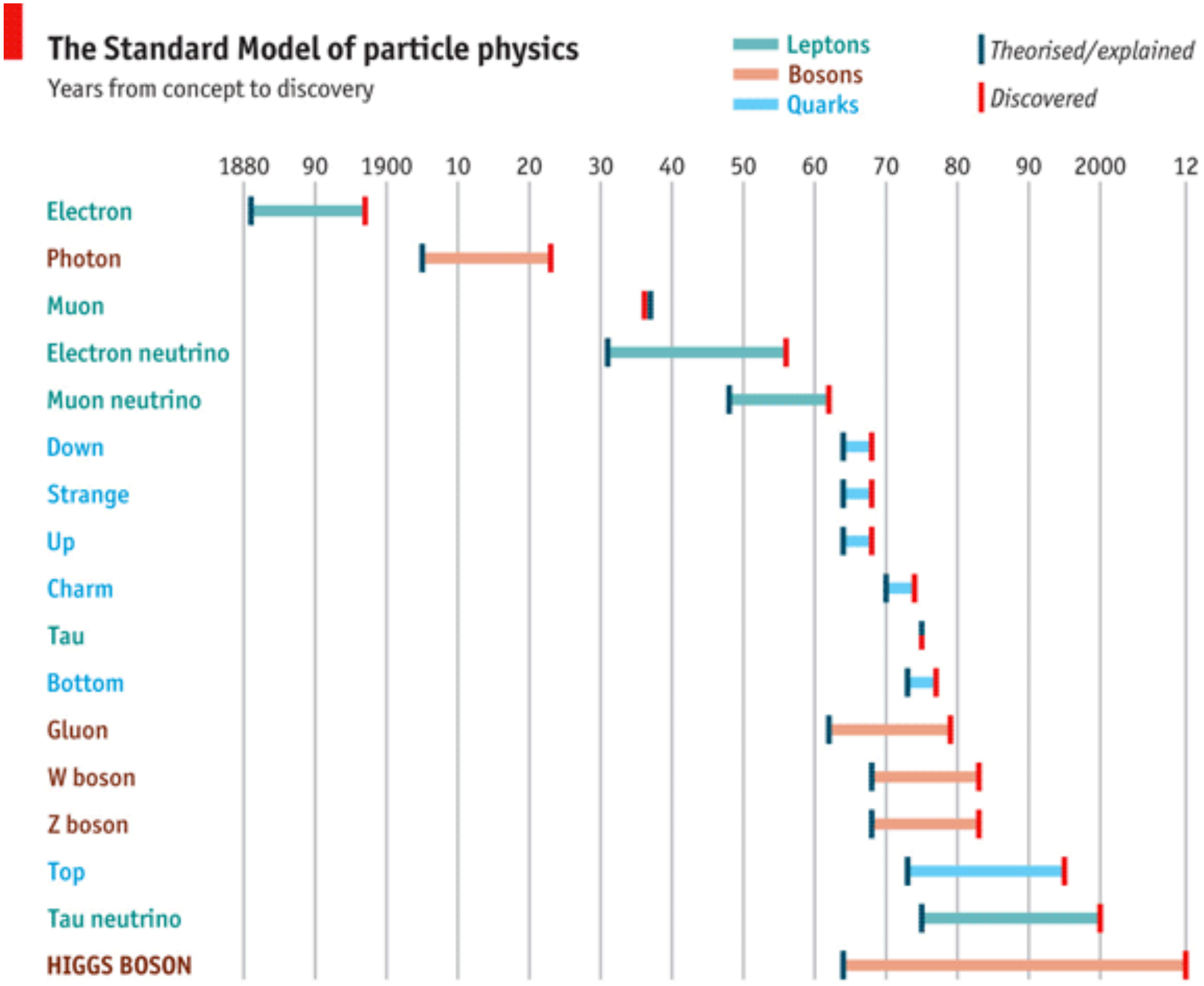
What is all this for?

The Standard Model of Particle Physics





$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$
$+ i \bar{\psi} \not{D} \psi + h.c.$
$+ \psi_i y_{ij} \psi_j \phi + h.c.$
$+ |D_\mu \phi|^2 - V(\phi)$



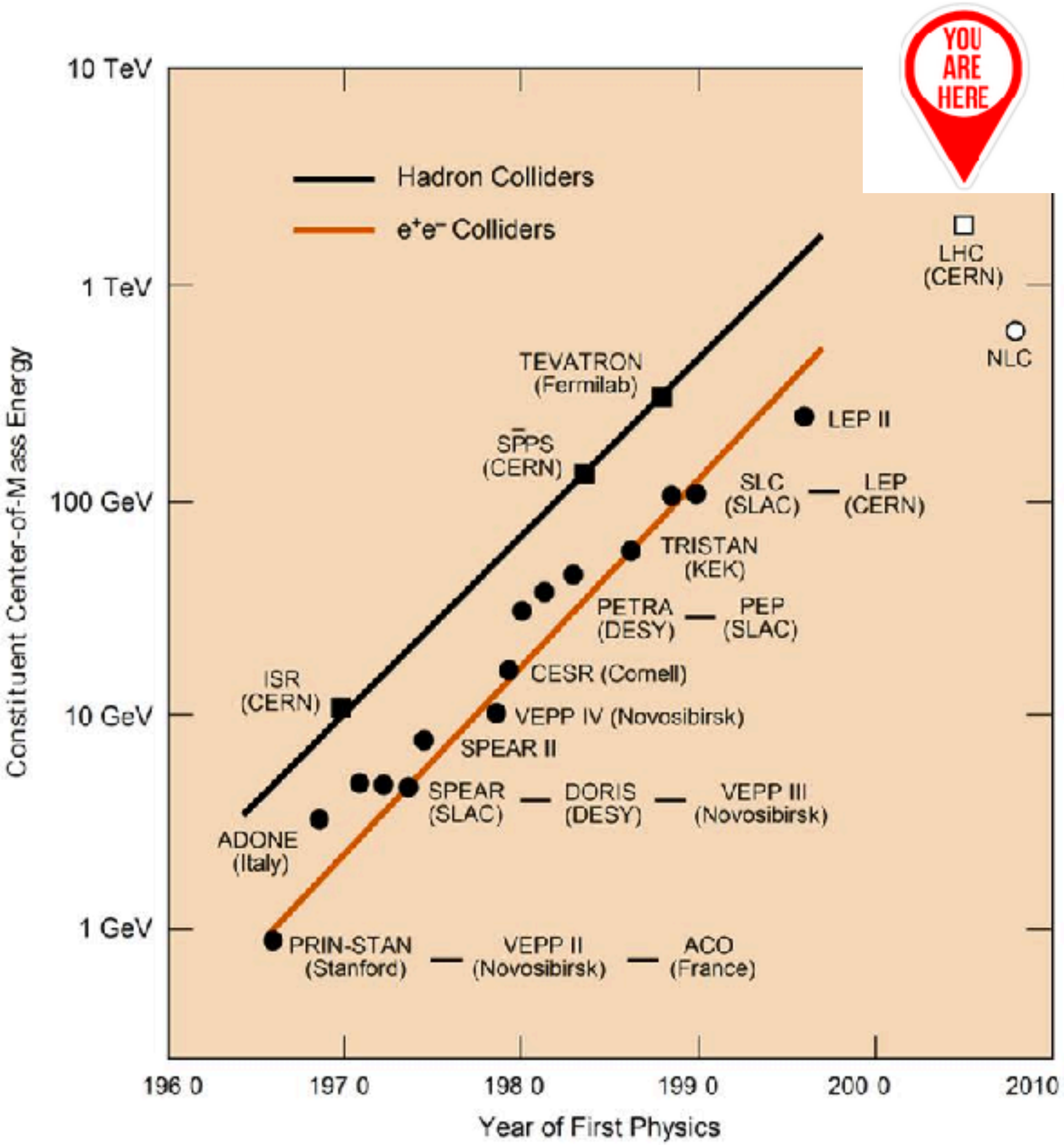
Source: *The Economist*

Was established in the 1970s…

… and people have been trying (and failing) to break it ever since.

What else is there beyond the
Standard Model?

What is the next layer of fundamental
matter and interactions?



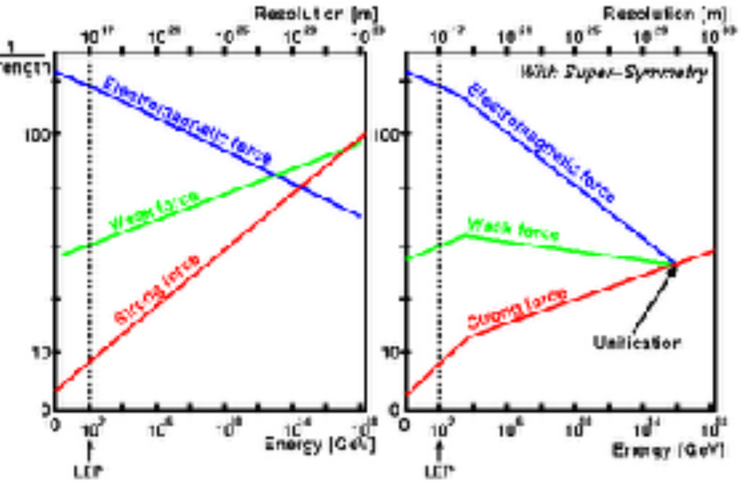
The main tool in the search for new physics beyond the SM is the particle collider.

By smashing together elementary particles at higher and higher energies, we hope to create new particles.

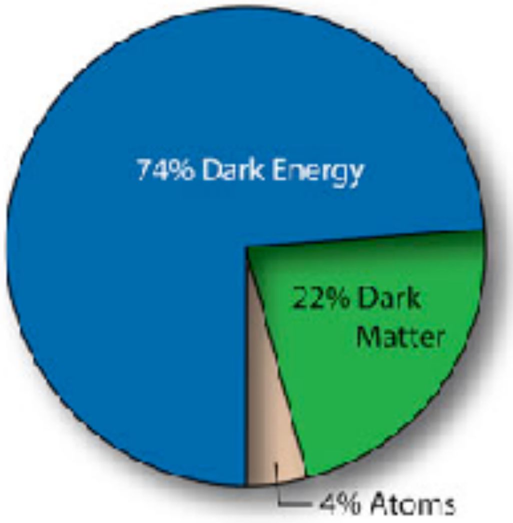
We attempt to “see” these new particles by studying the collision debris with very powerful detectors.

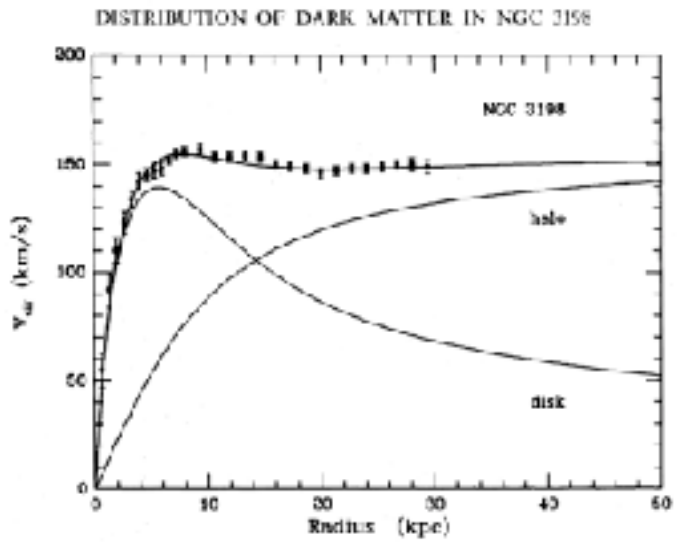
We know there's new physics out there…

grand unification

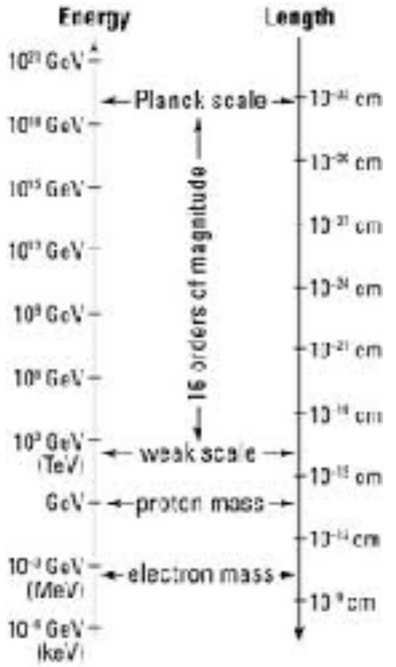


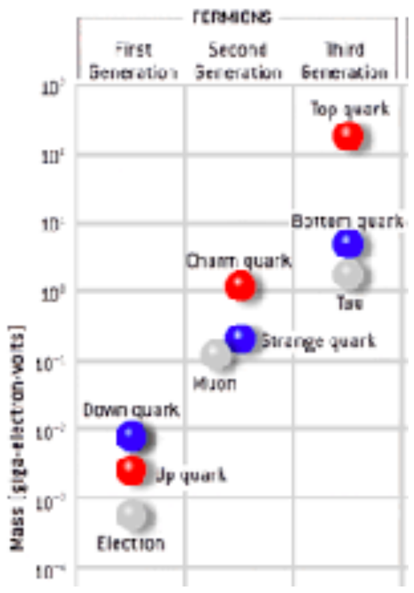
dark matter

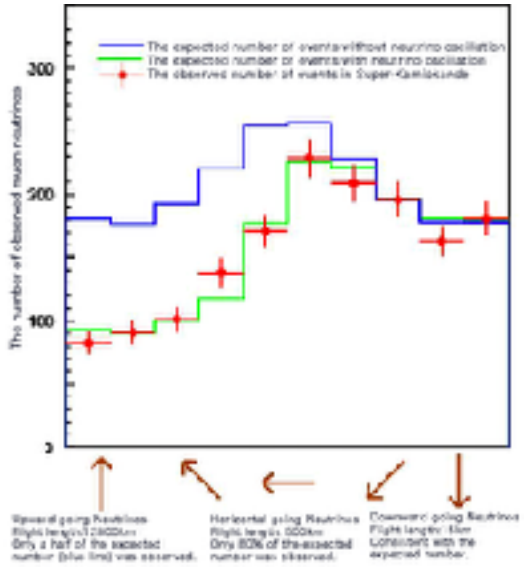


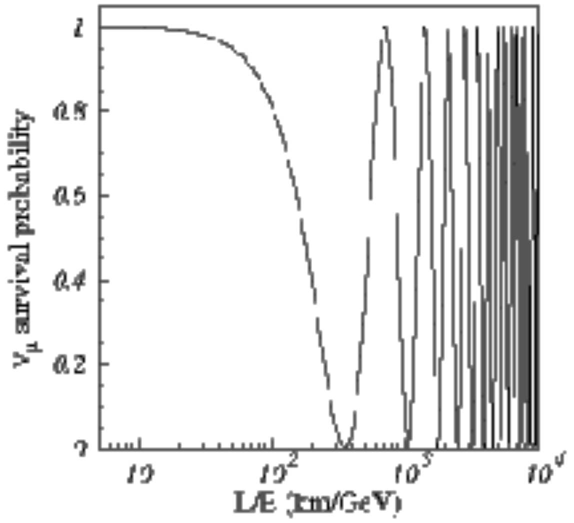


hierarchy problem









$$\mathcal{L} \supset \theta \, \frac{\alpha_s}{8\pi} G_{\mu\nu} \tilde{G}^{\mu\nu}$$

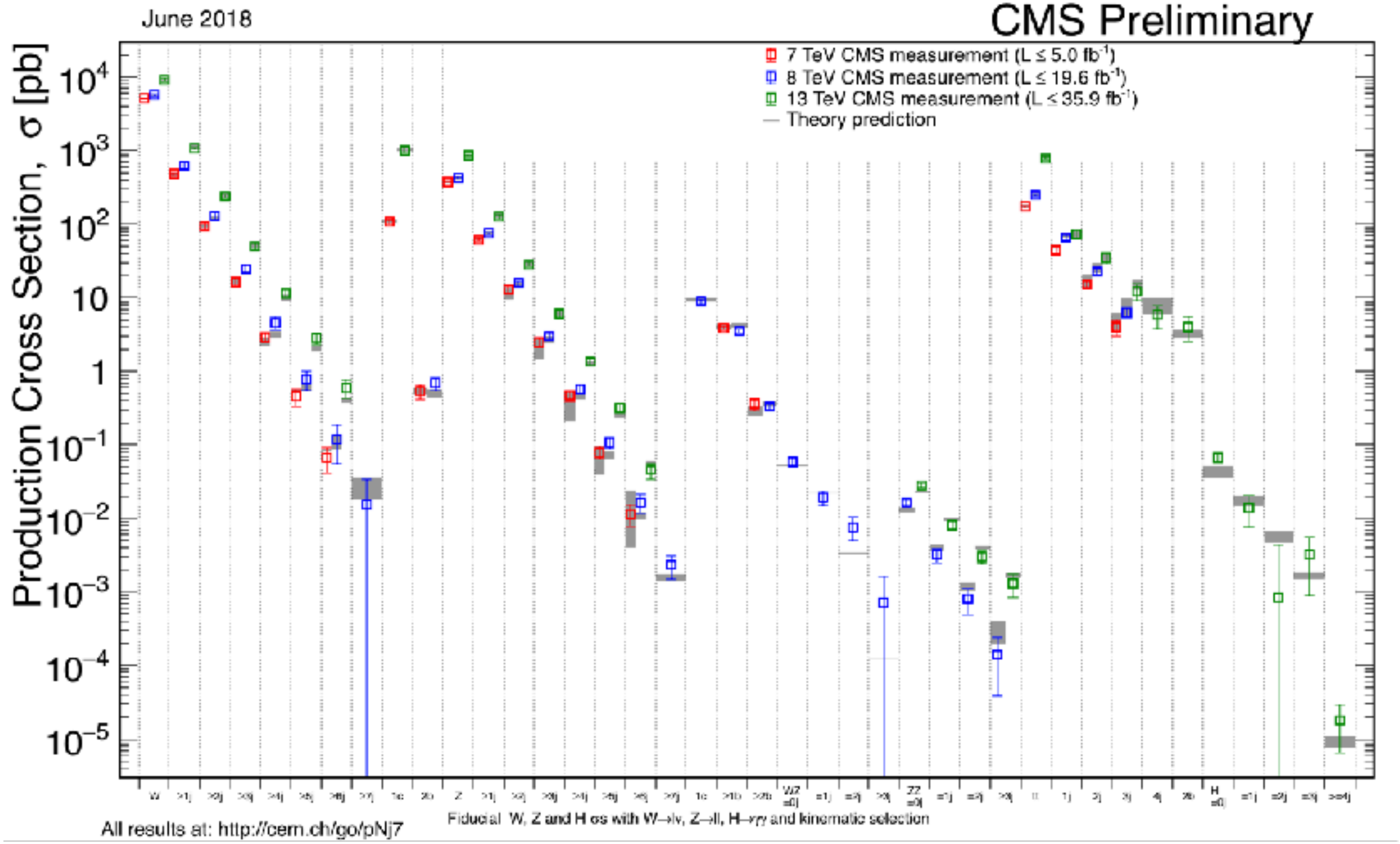
$$\theta \lesssim 10^{-10}$$

flavor puzzle

neutrino masses

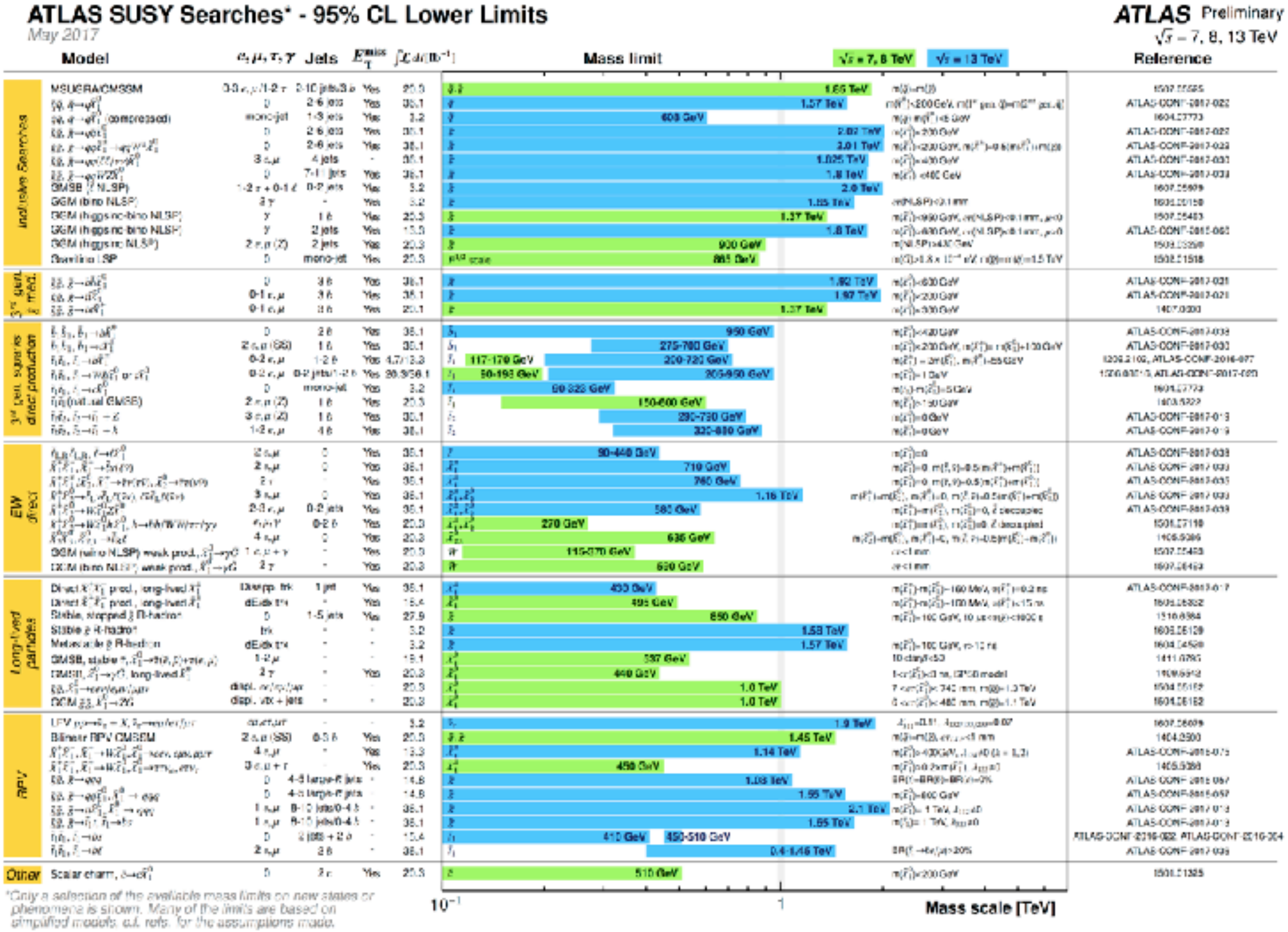
strong CP problem

But no sign of it yet at the LHC…



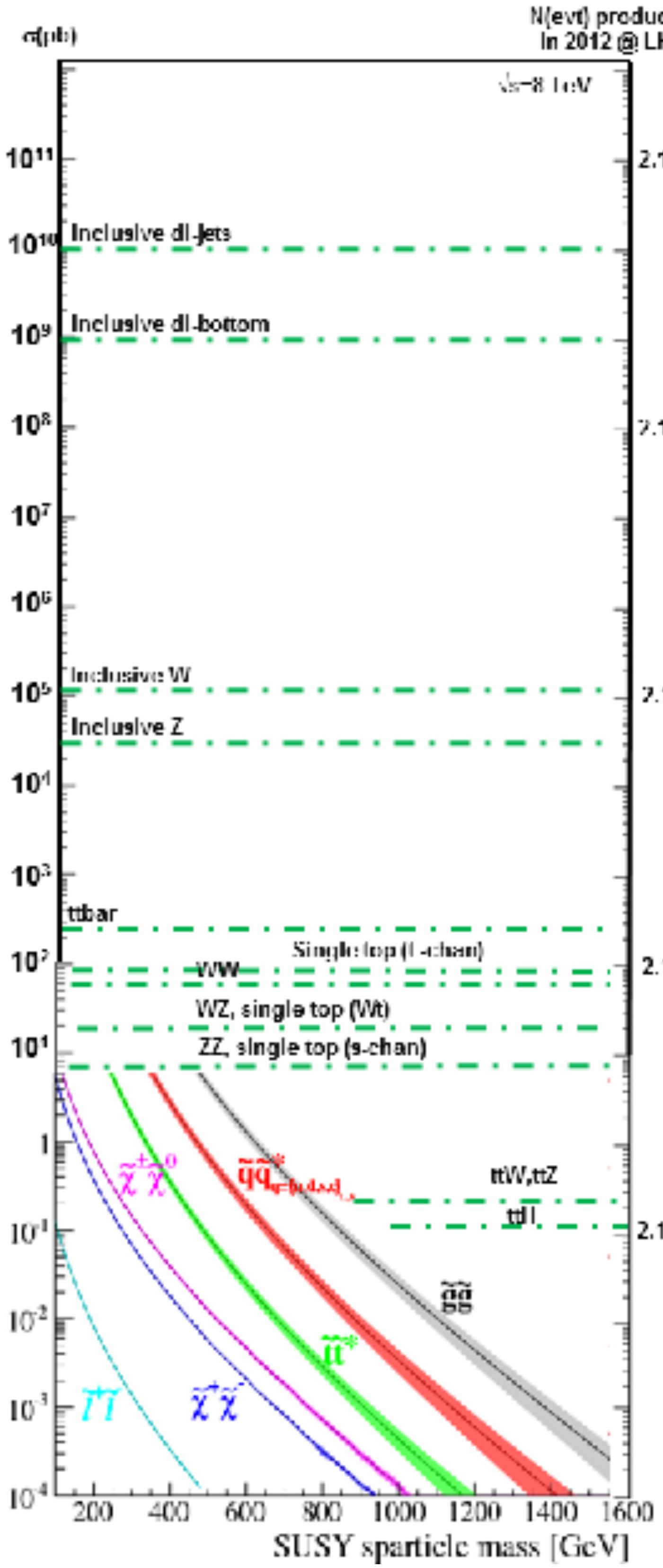
Precision measurements of SM processes.
Agreement between theory and experiment across ~9 orders of magnitude.

But no sign of it yet at the LHC…



Countless searches for new physics beyond the SM.
So far no concrete evidence, only lower limits on the NP scale.

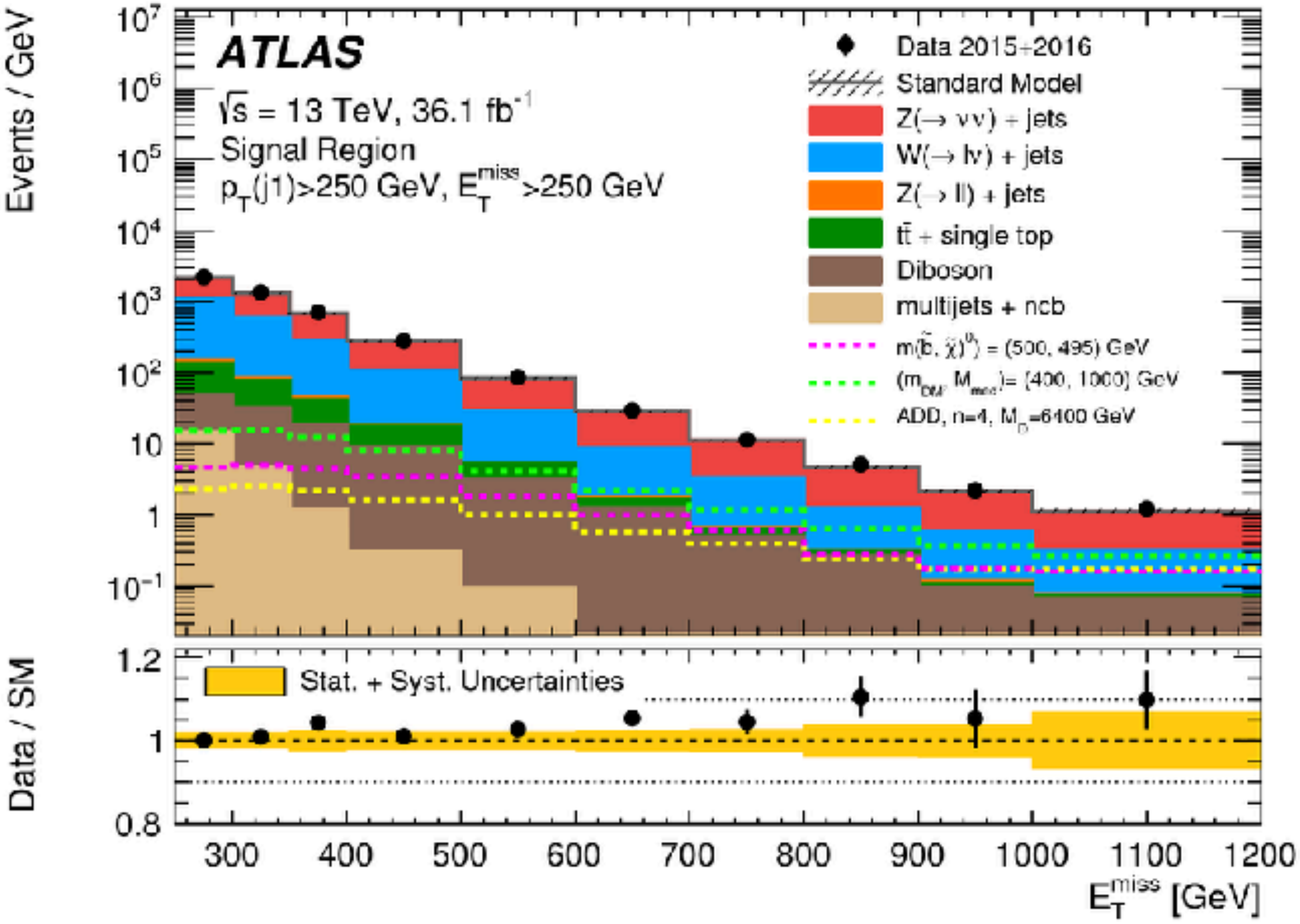
What does a typical search for new physics look like at the LHC?



Typical new physics production rates are many, many orders of magnitude smaller than SM processes.

Need a way to improve signal to noise to have any hope of seeing new physics.

What does a typical search for new physics look like at the LHC?



- Identify a "signal region" in the phase space, motivated by some model, where one expects S/N to be greatly enhanced.
- Estimate SM background using combination of simulations and data-driven methods (control regions)
- Compare data to SM prediction: announce discovery significance or set a limit on the model

This generally assumes we know what we're looking for.

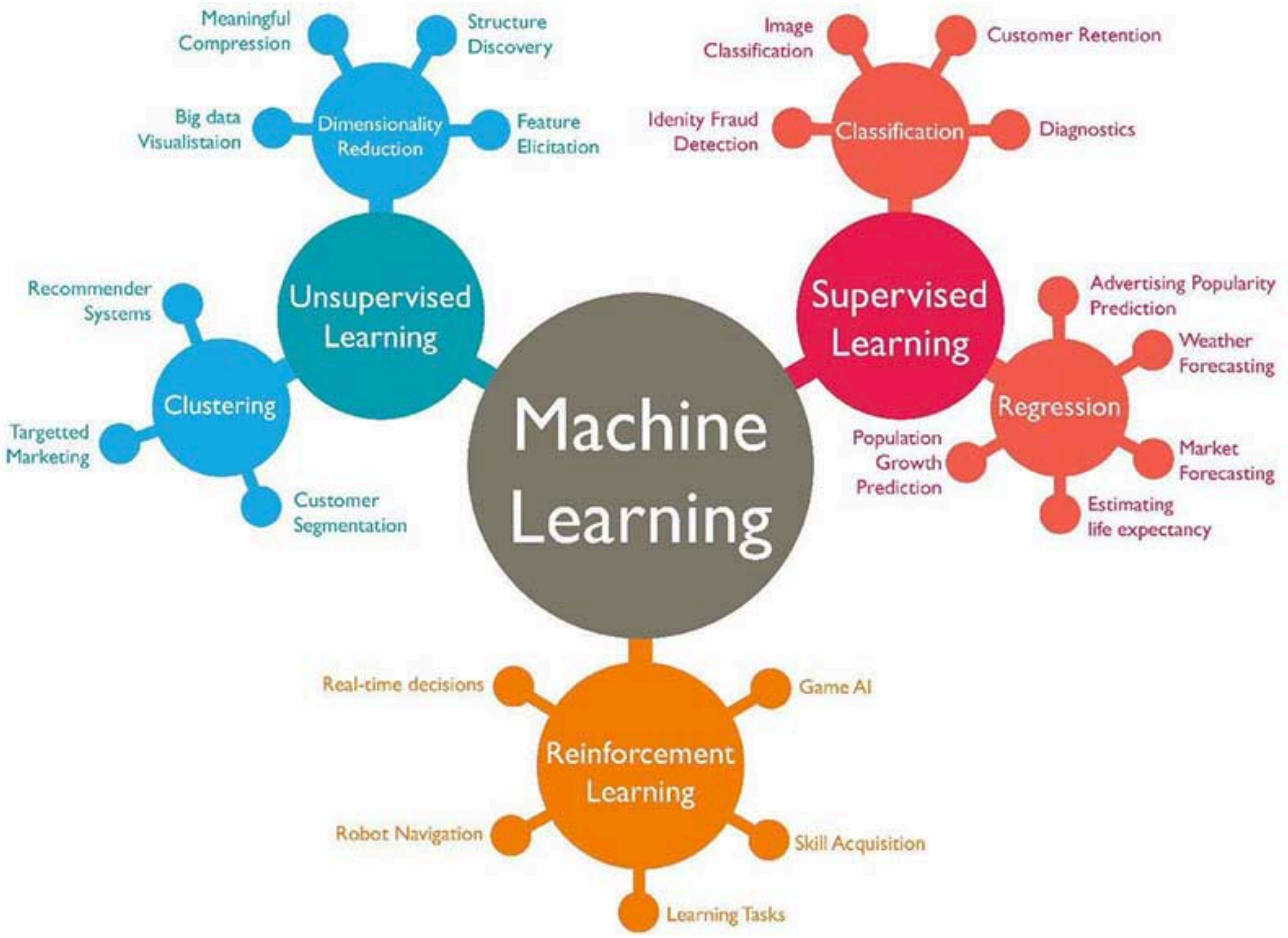
➡ ML can still help in this case, by improving S/N — supervised learning, classification, regression

What if we don't know what we're looking for? Can we find the unexpected signal buried underneath all this raw data?

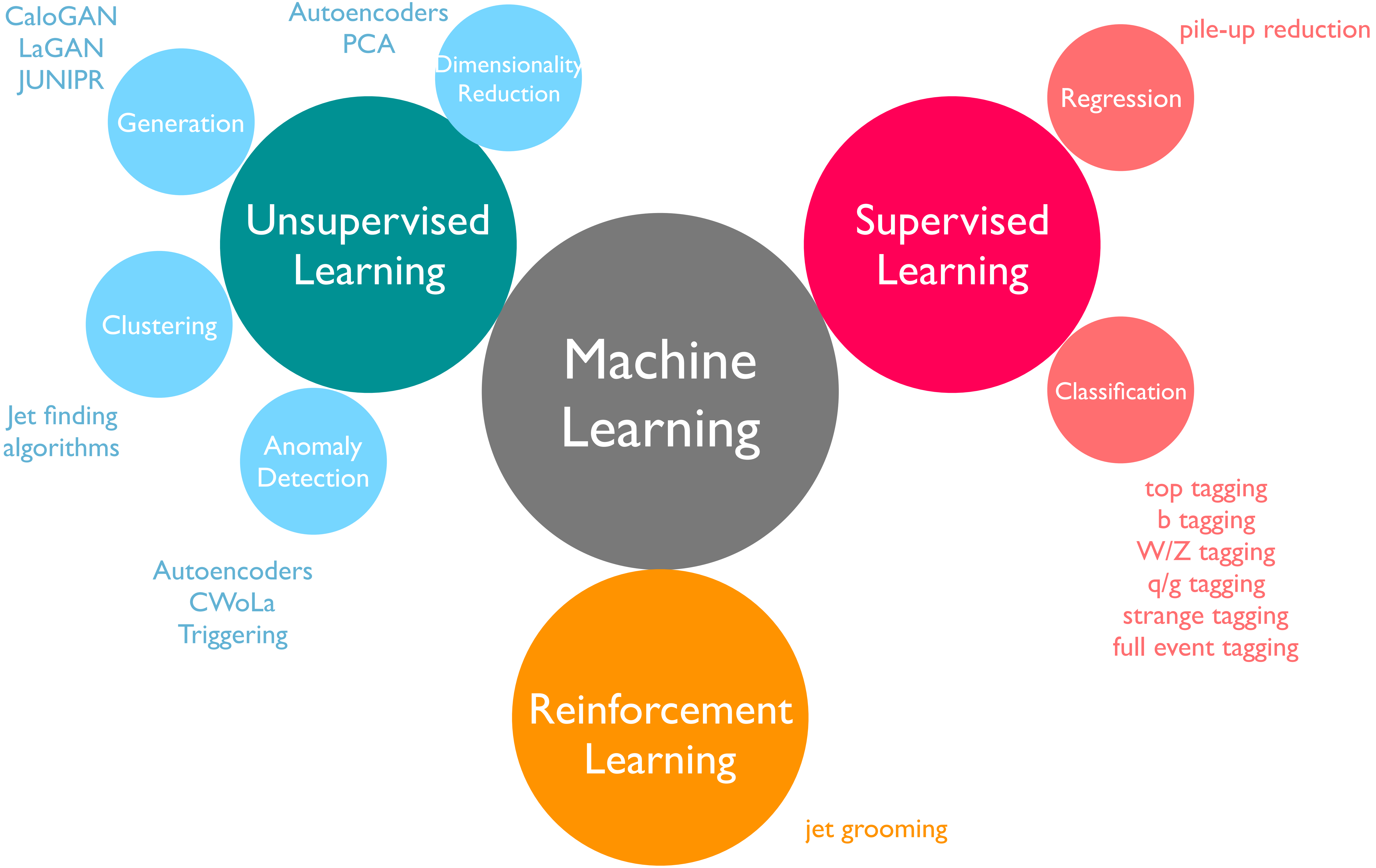
➡ ML can help in this case — unsupervised learning, clustering, anomaly detection

A promising path forward:
Adapt sophisticated ML tools developed for real-world applications in order to improve data analysis at the LHC

The Landscape of ML



The Landscape of ML @ LHC



Machine Learning

Unsupervised Learning

Generation

CaloGAN
LaGAN
JUNIPR

Dimensionality Reduction

Autoencoders
PCA

Clustering

Jet finding algorithms

Anomaly Detection

Autoencoders
CWoLa
Triggering

Supervised Learning

Regression

pile-up reduction

Classification

top tagging
b tagging
W/Z tagging
q/g tagging
strange tagging
full event tagging

Reinforcement Learning

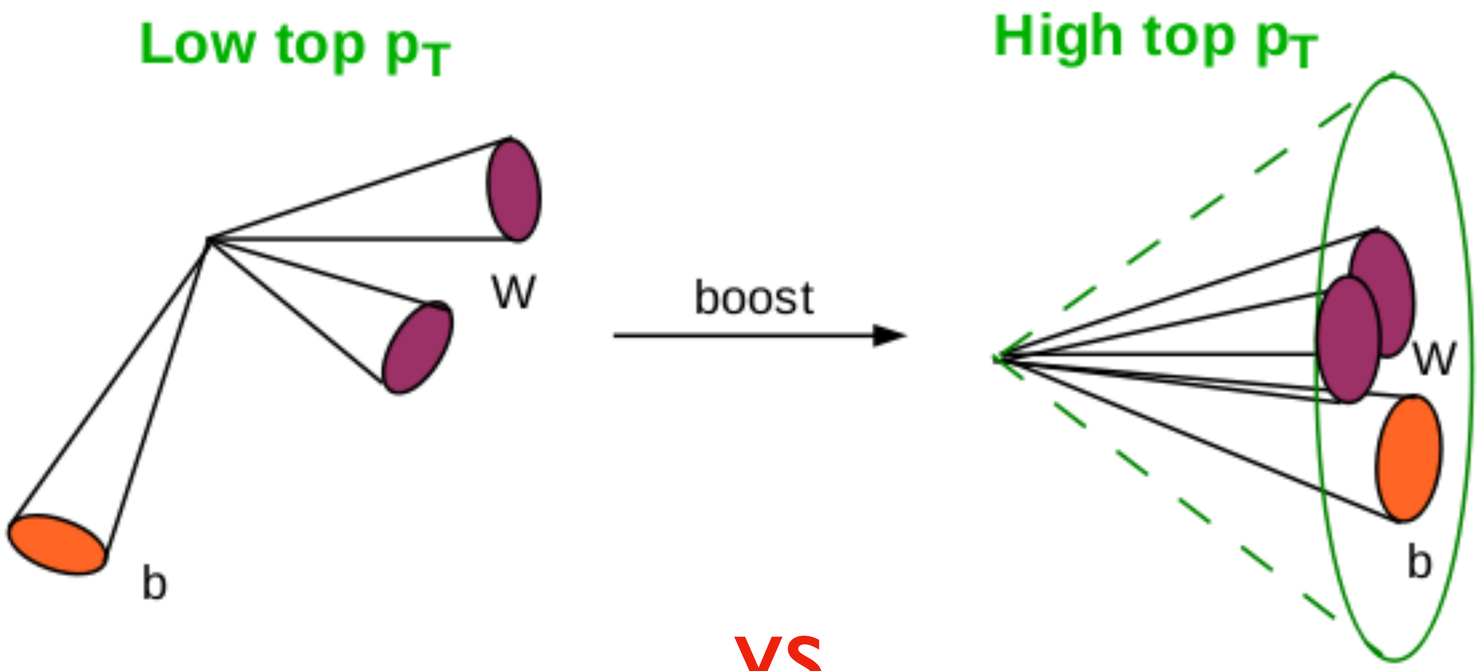
jet grooming

Recent progress in ML @ LHC

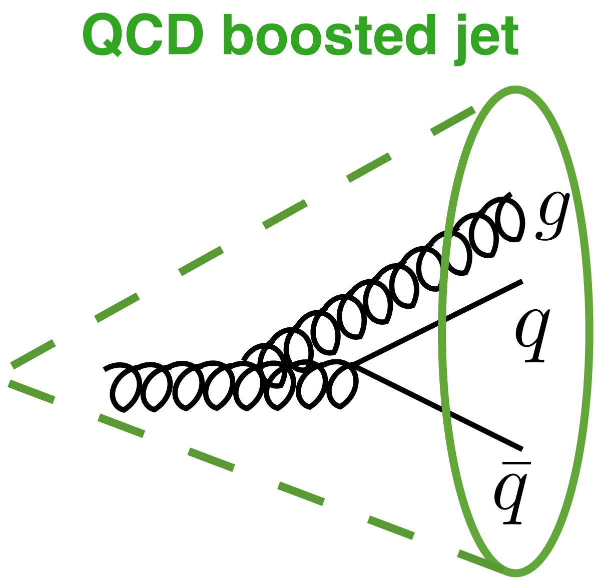
- Huge performance gains, especially for object classification
- Exploring the possibilities of learning physics directly from the data
- Developing new and unconventional ways of searching for new physics

In the rest of this talk, I will focus on some recent works that touch upon these points.

A benchmark problem: boosted top tagging

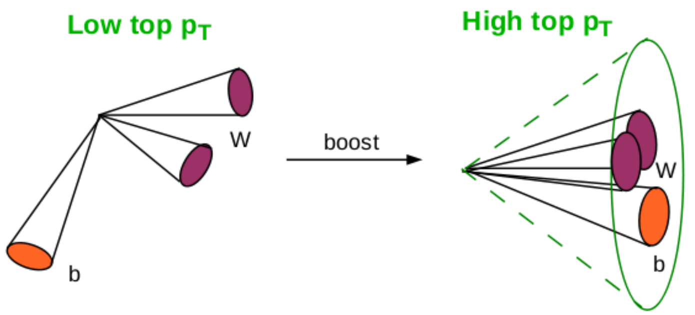


vs.

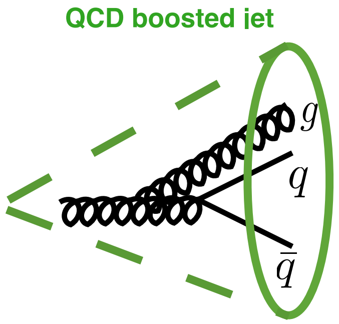


How to differentiate between these two types of jets?

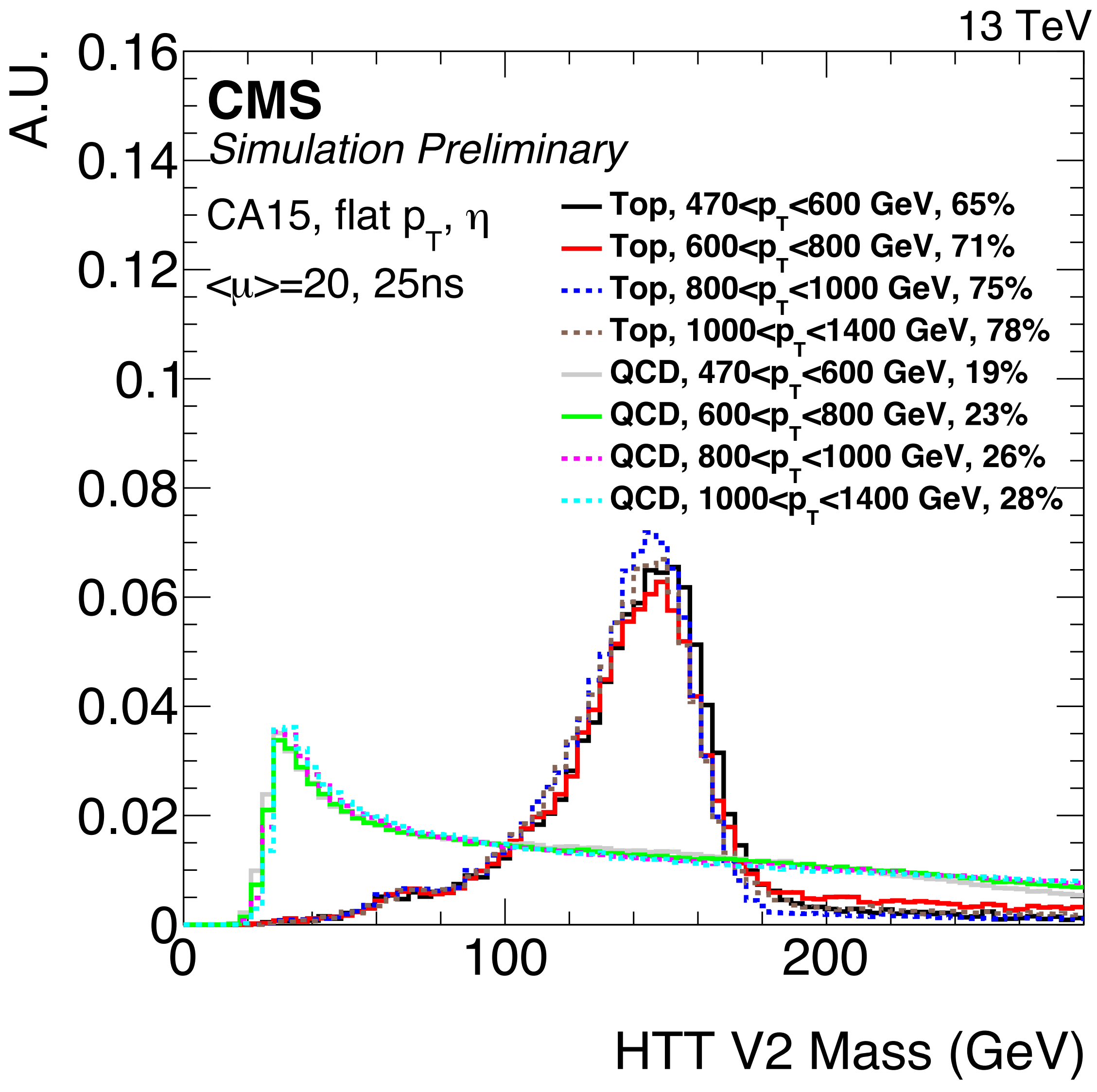
This is a straightforward supervised classification problem in ML.



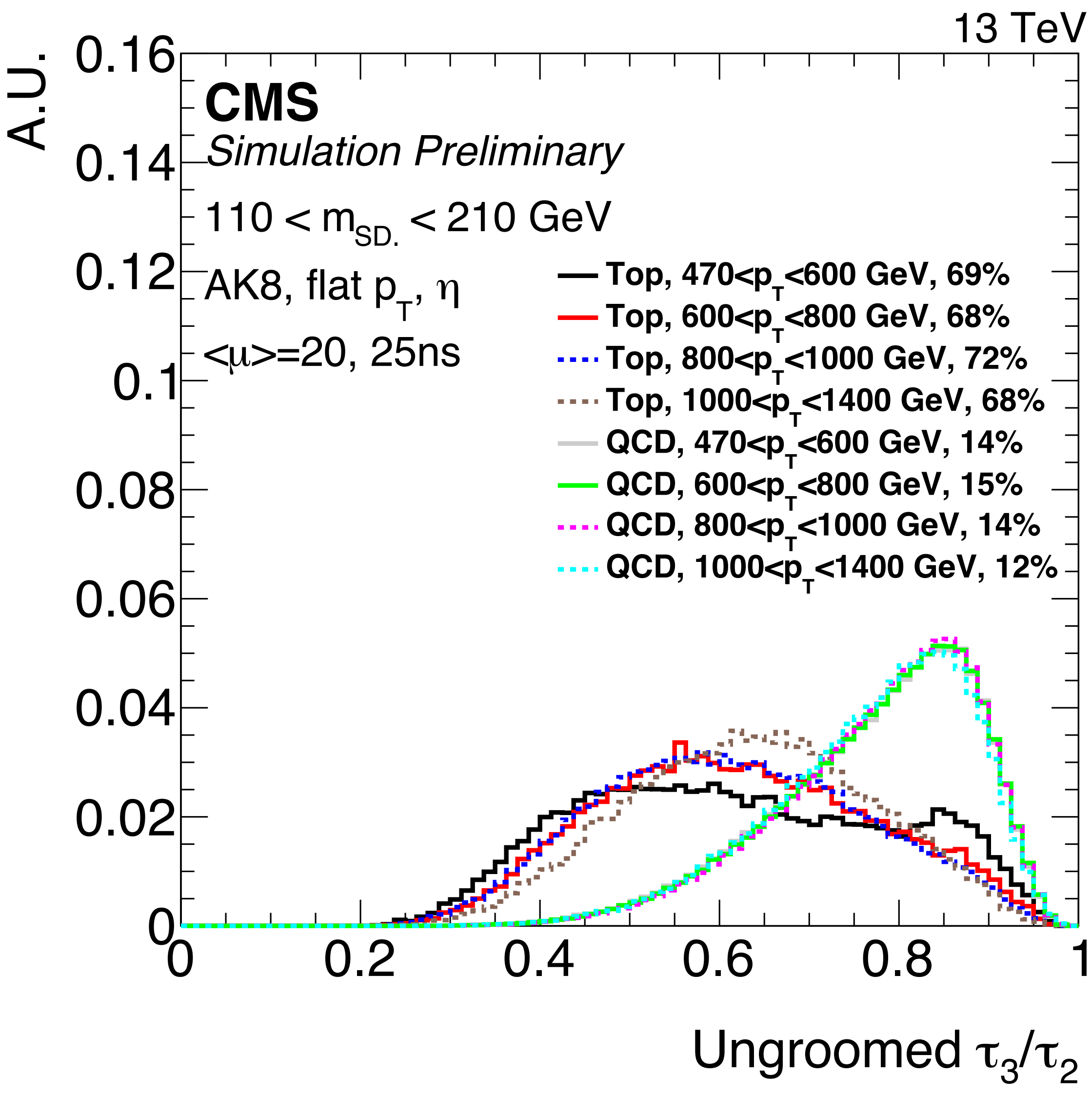
vs



Some obvious ideas:



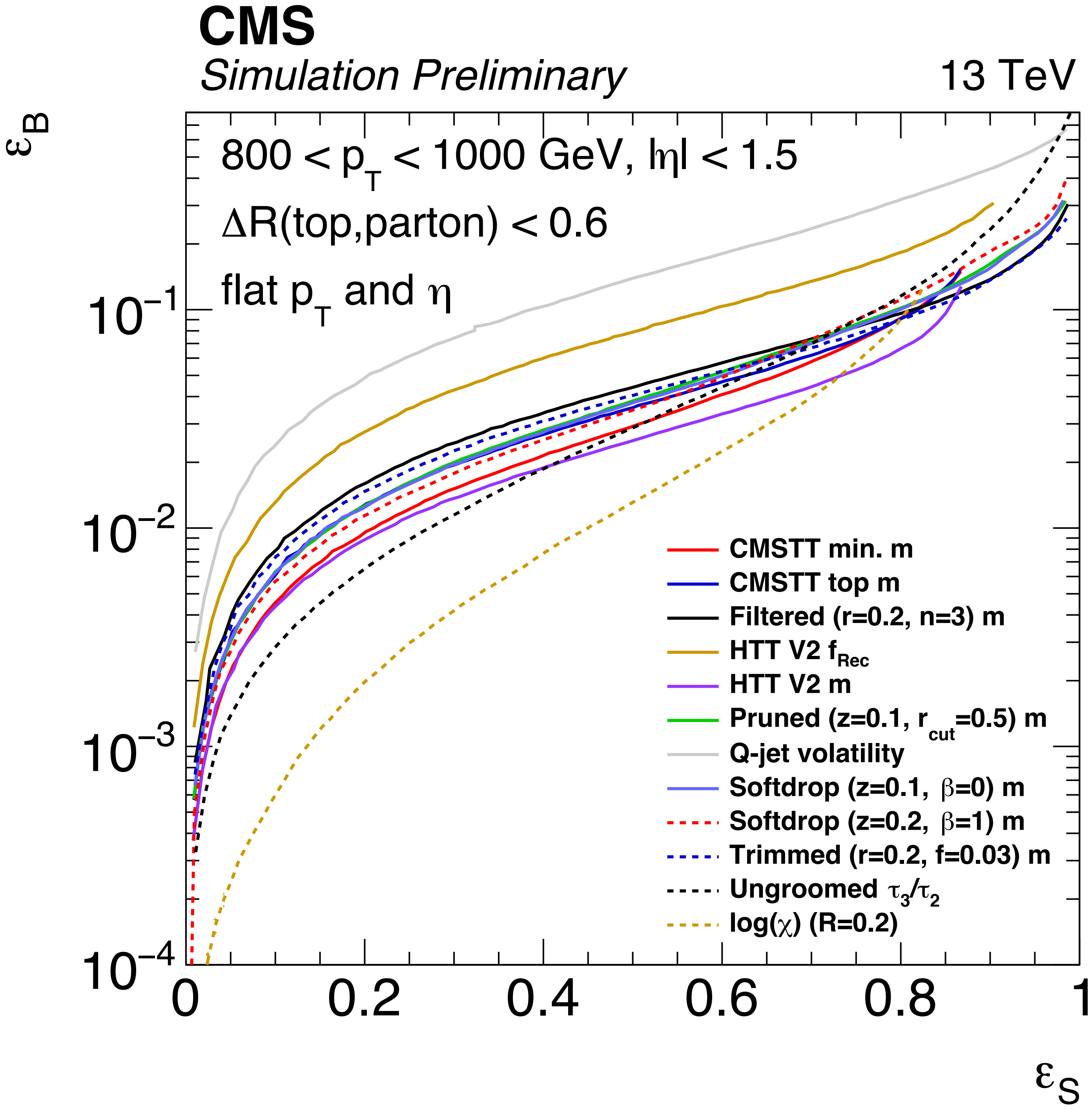
jet mass (m_{top} vs 0)



jet substructure (3 vs 1)

State of the art with cuts on kinematic quantities:

QCD jet mistag rate



“ROC curve”

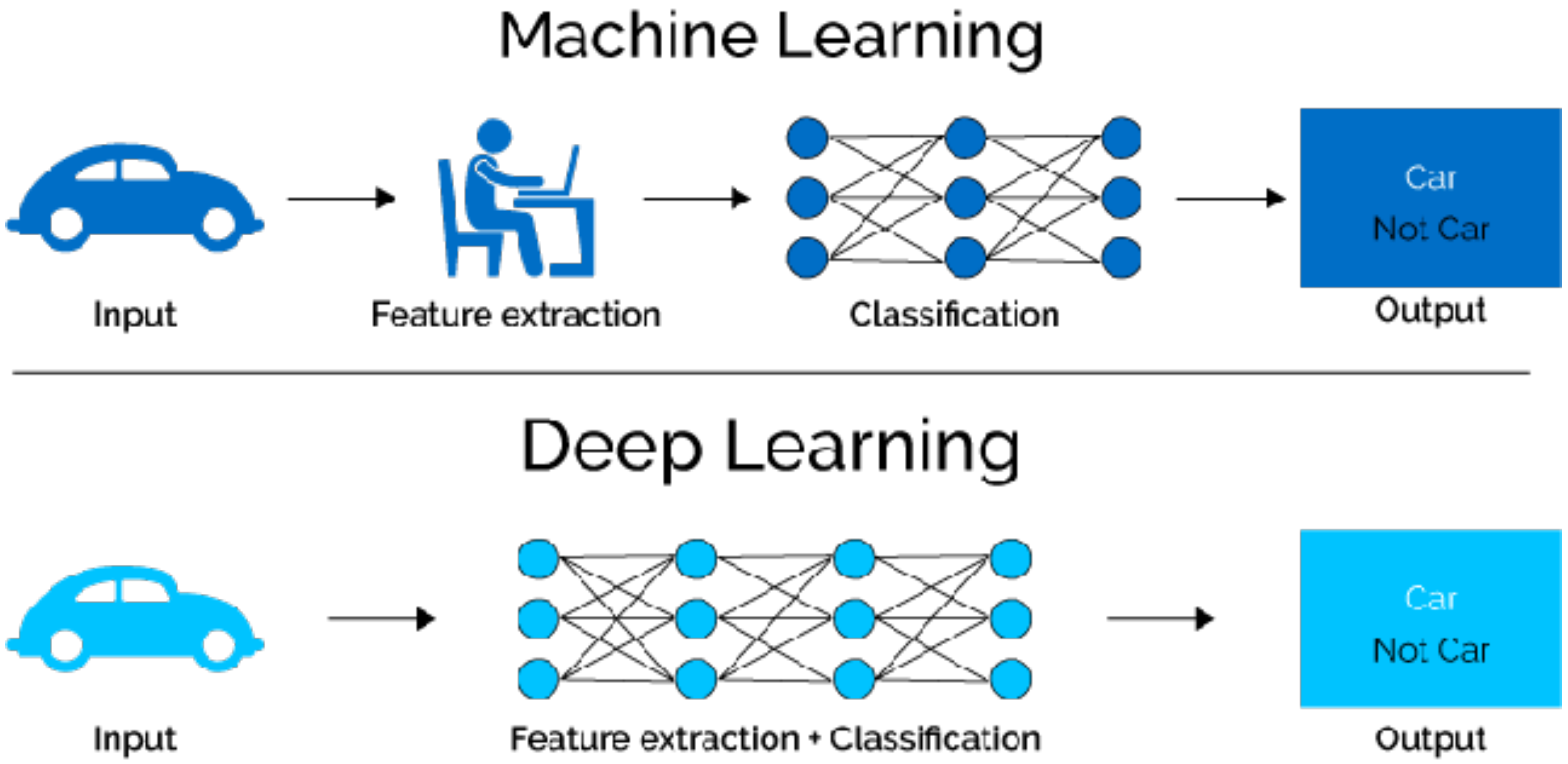
top tagging efficiency

Can deep learning do better??

Automated Feature Engineering

By training on raw, low-level inputs, deep learning can achieve much better performance.

Deep neural networks automate and optimize the process of "feature engineering".



From towardsdatascience.com

Jet constituents → Deep learning algorithm → Top or QCD

Data Representations

Although deep learning capable of building features from raw data, how we represent the data can still matter a lot.

In the case of jets, some popular options are

- Four vectors (DNNs)
- Sequences (RNNs, LSTMs)
- Binary trees (RecNNs)
- Graphs (point clouds)
- Images (CNNs)

Jet Images

Can think of a jet as an image in eta and phi, with

- Pixelation provided by calorimeter towers
- Pixel intensity = pT recorded by each tower

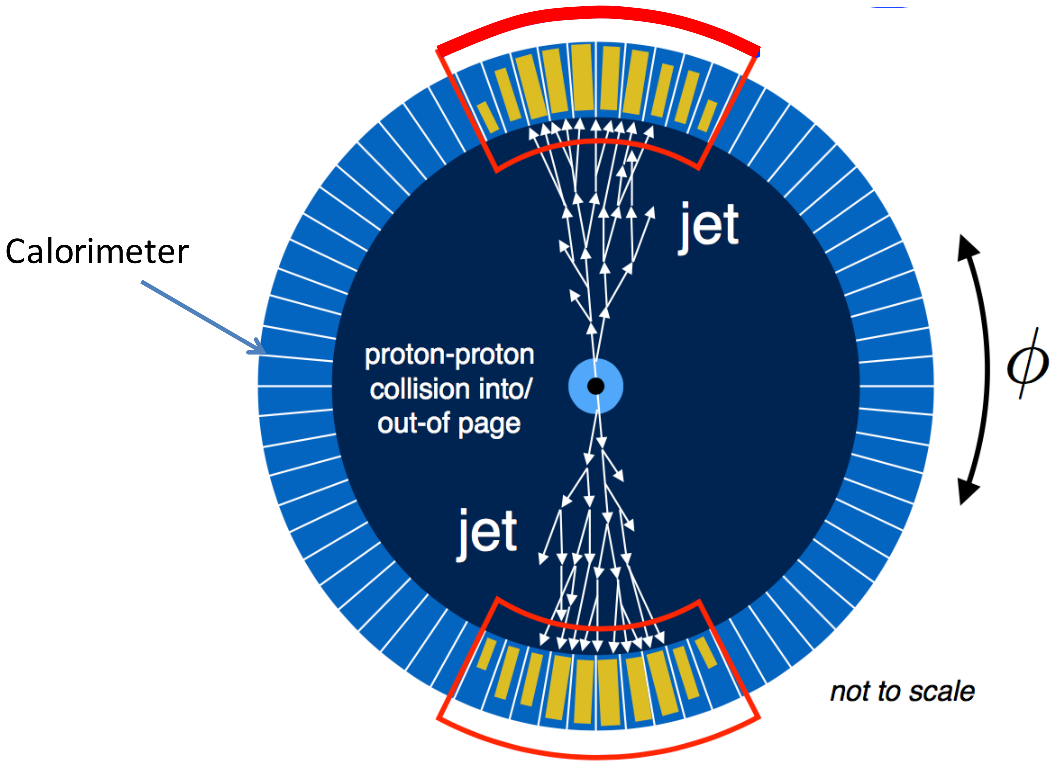
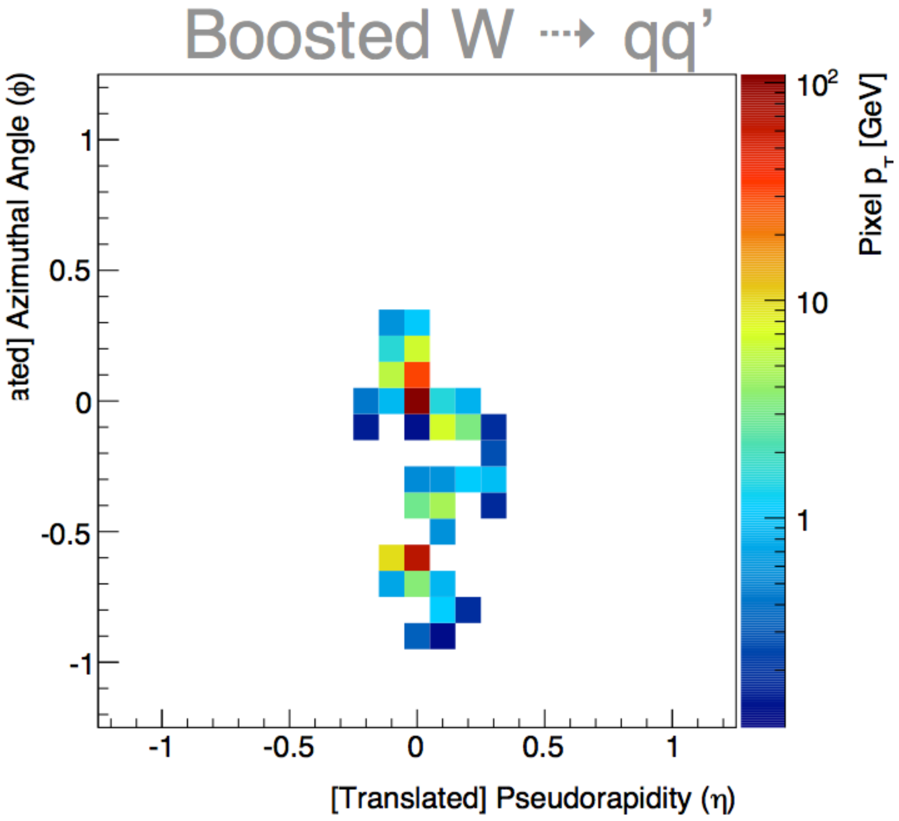


Figure credit: B. Nachman



Should be able to apply "off-the-shelf" NNs developed for image recognition to classify jets at the LHC! de Oliveira et al 1511.05190

Top Tagging with CNNs

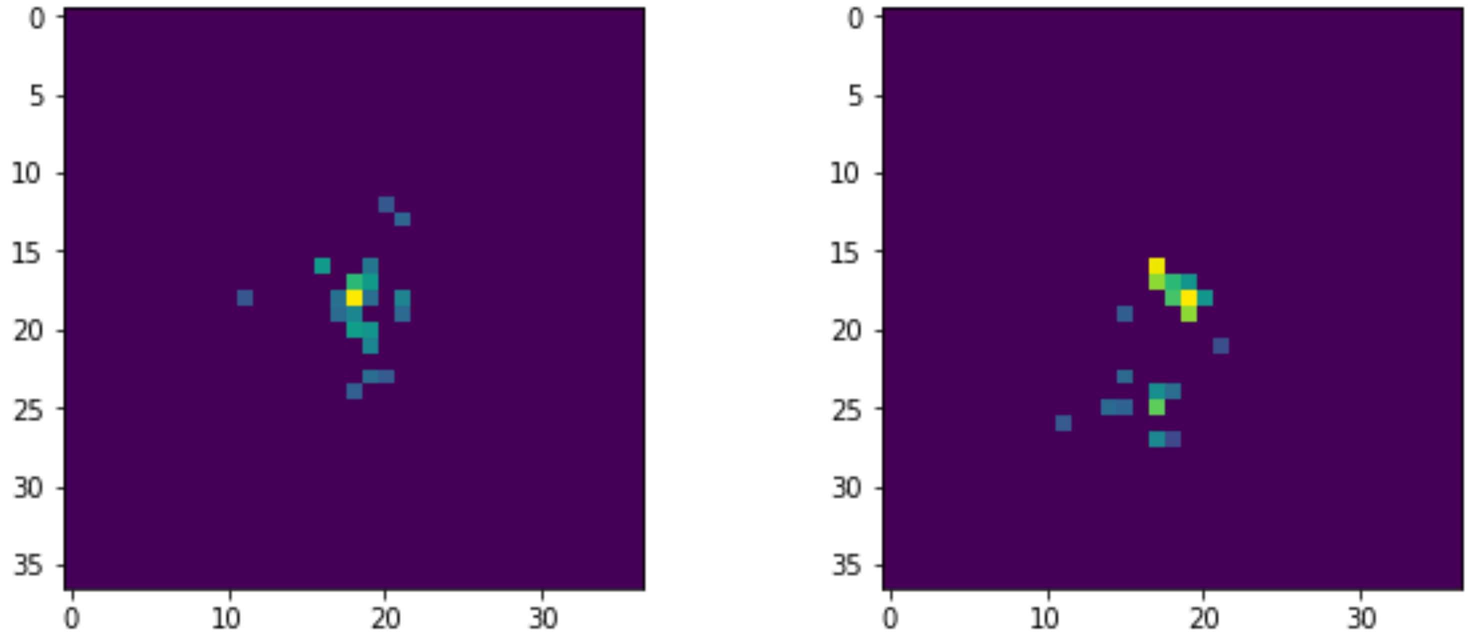
Macaluso & DS 1803.00107

	CMS
Jet sample	13 TeV $p_T \in (800, 900)$ GeV, $\lvert\eta\rvert < 1$ Pythia 8 and Delphes particle-flow match: $\Delta R(t, j) < 0.6$ merge: $\Delta R(t, q) < 0.6$ 1.2M + 1.2M
Image	37×37 $\Delta\eta = \Delta\phi = 3.2$
Colors	$(p_T^{neutral}, \, p_T^{track}, \, N_{track}, \, N_{muon})$

Individual images very sparse

QCD

Tops



Building on previous "DeepTop" tagger of Kasieczka et al 1701.08784

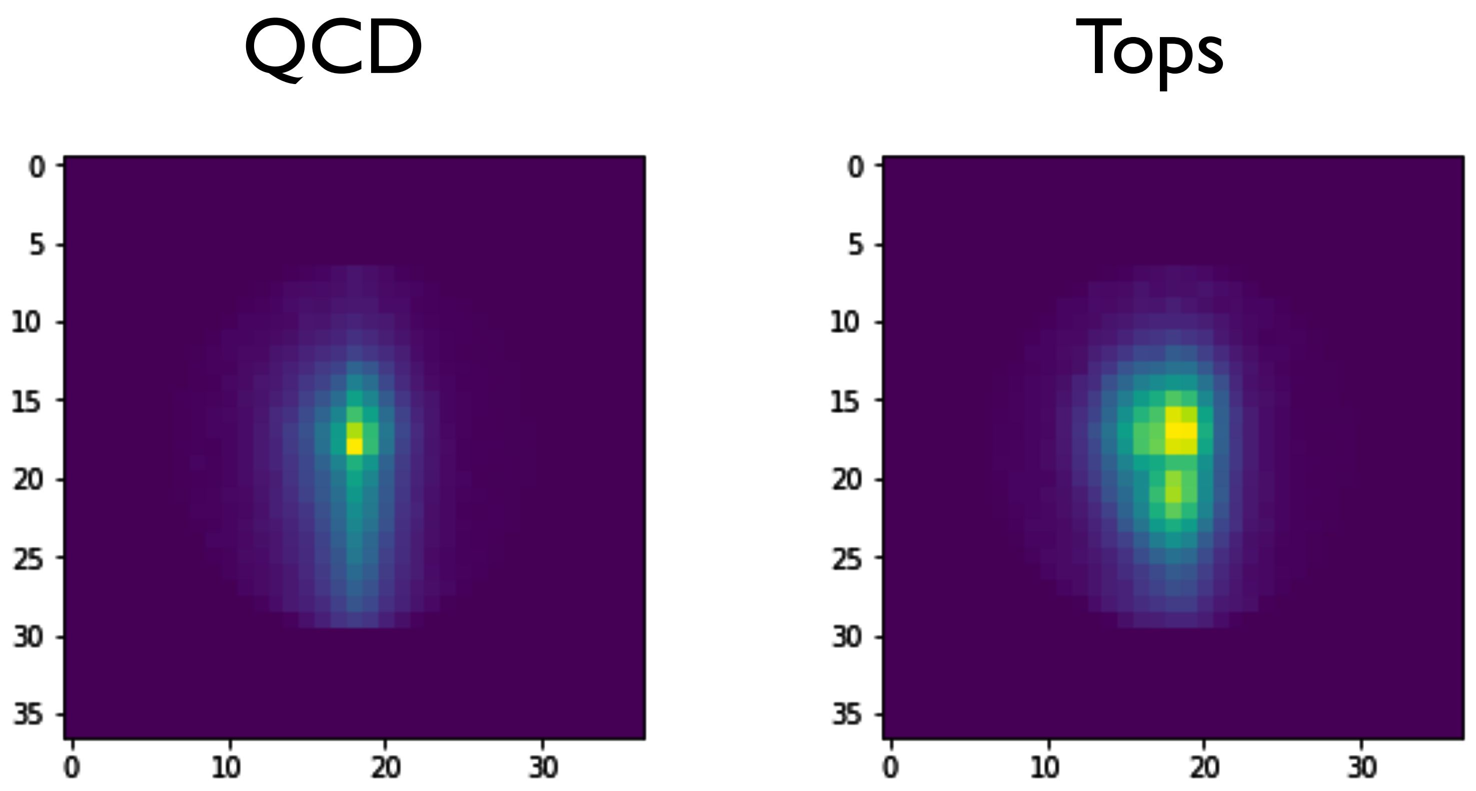
Other approaches also promising (DNNs, RecNNs, RNNs, LSTMs, GNNs, …)

Top Tagging with CNNs

Macaluso & DS 1803.00107

	CMS
Jet sample	13 TeV $p_T \in (800, 900)$ GeV, $\lvert\eta\rvert < 1$ Pythia 8 and Delphes particle-flow match: $\Delta R(t, j) < 0.6$ merge: $\Delta R(t, q) < 0.6$ 1.2M + 1.2M
Image	37×37 $\Delta\eta = \Delta\phi = 3.2$
Colors	$(p_T^{neutral}, p_T^{track}, N_{track}, N_{muon})$

Average images clearly different


QCD

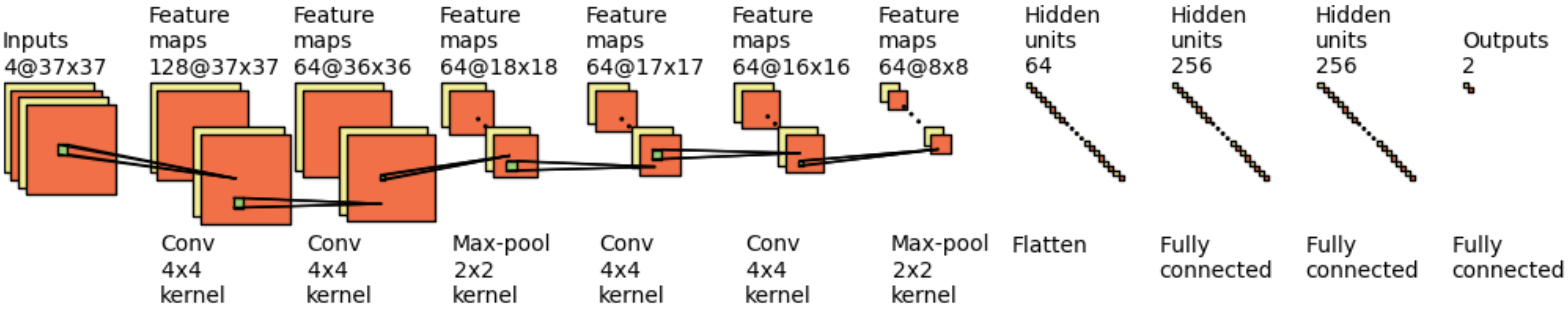
Tops

Building on previous “DeepTop” tagger of Kasieczka et al 1701.08784

Other approaches also promising (DNNs, RecNNs, RNNs, LSTMs, GNNs, …)

Top Tagging with CNNs

Macaluso & DS 1803.00107



AdaDelta

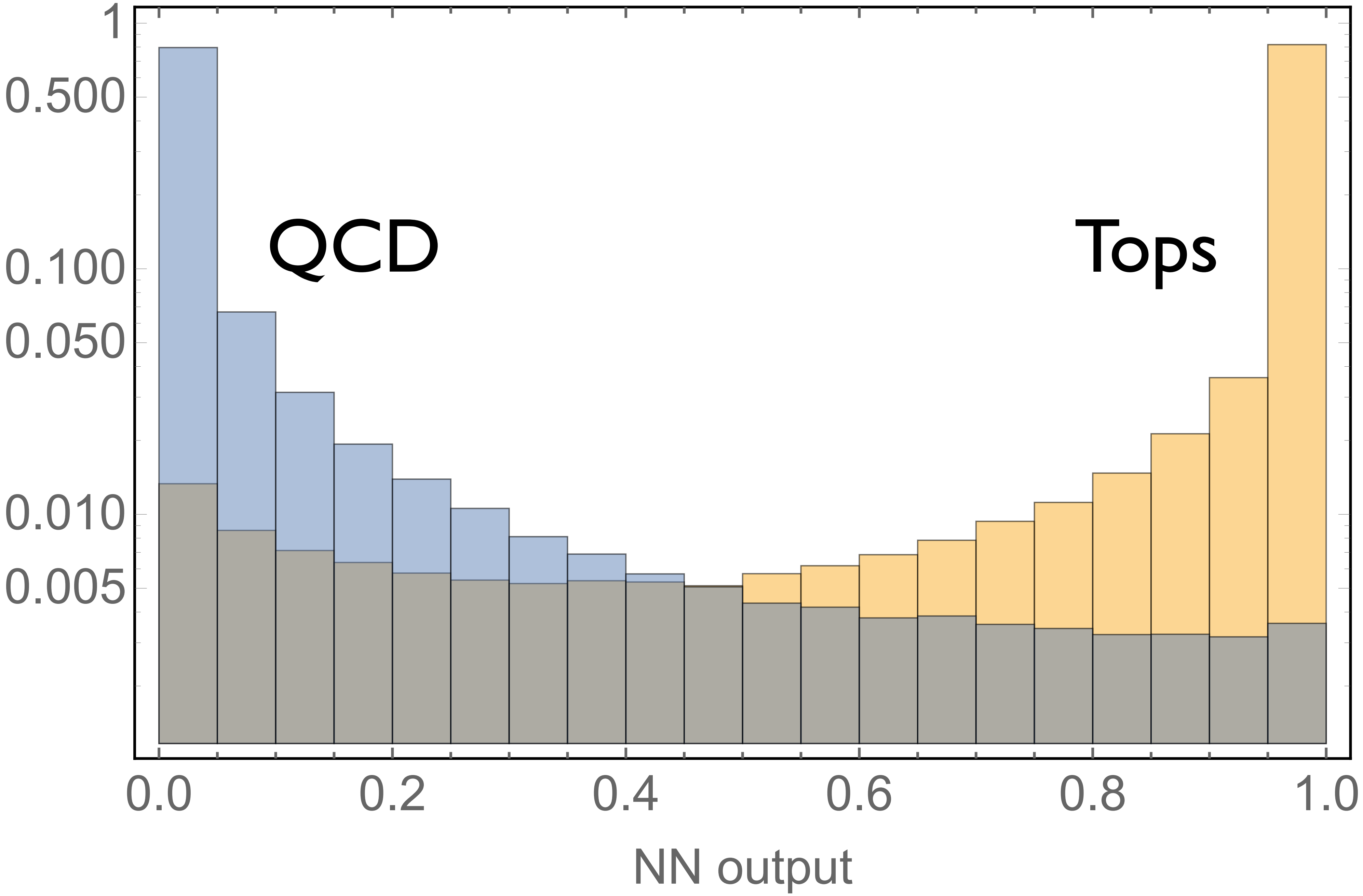
$\eta = 0.3$ with annealing schedule

minibatch size=128

cross entropy loss

Top Tagging with CNNs

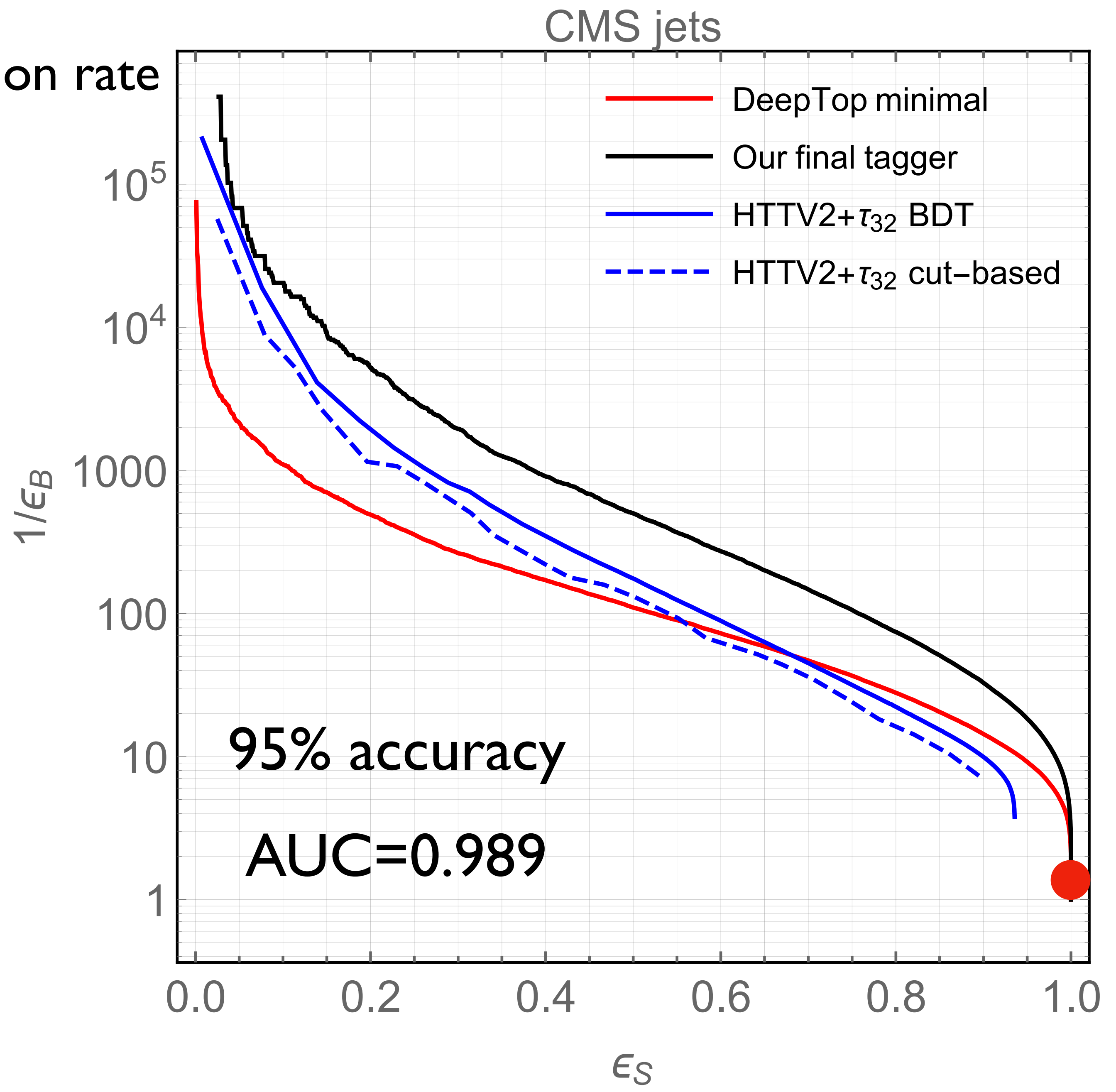
Macaluso & DS 1803.00107



Top Tagging with CNNs

Macaluso & DS 1803.00107

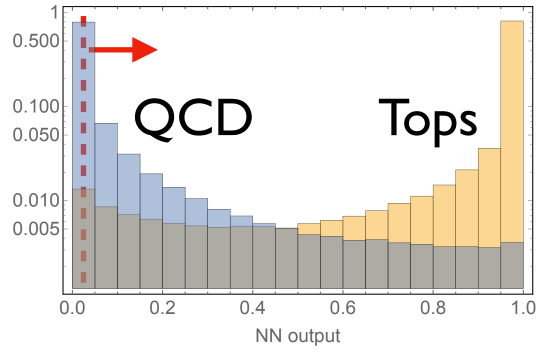
QCD rejection rate



95% accuracy

AUC=0.989

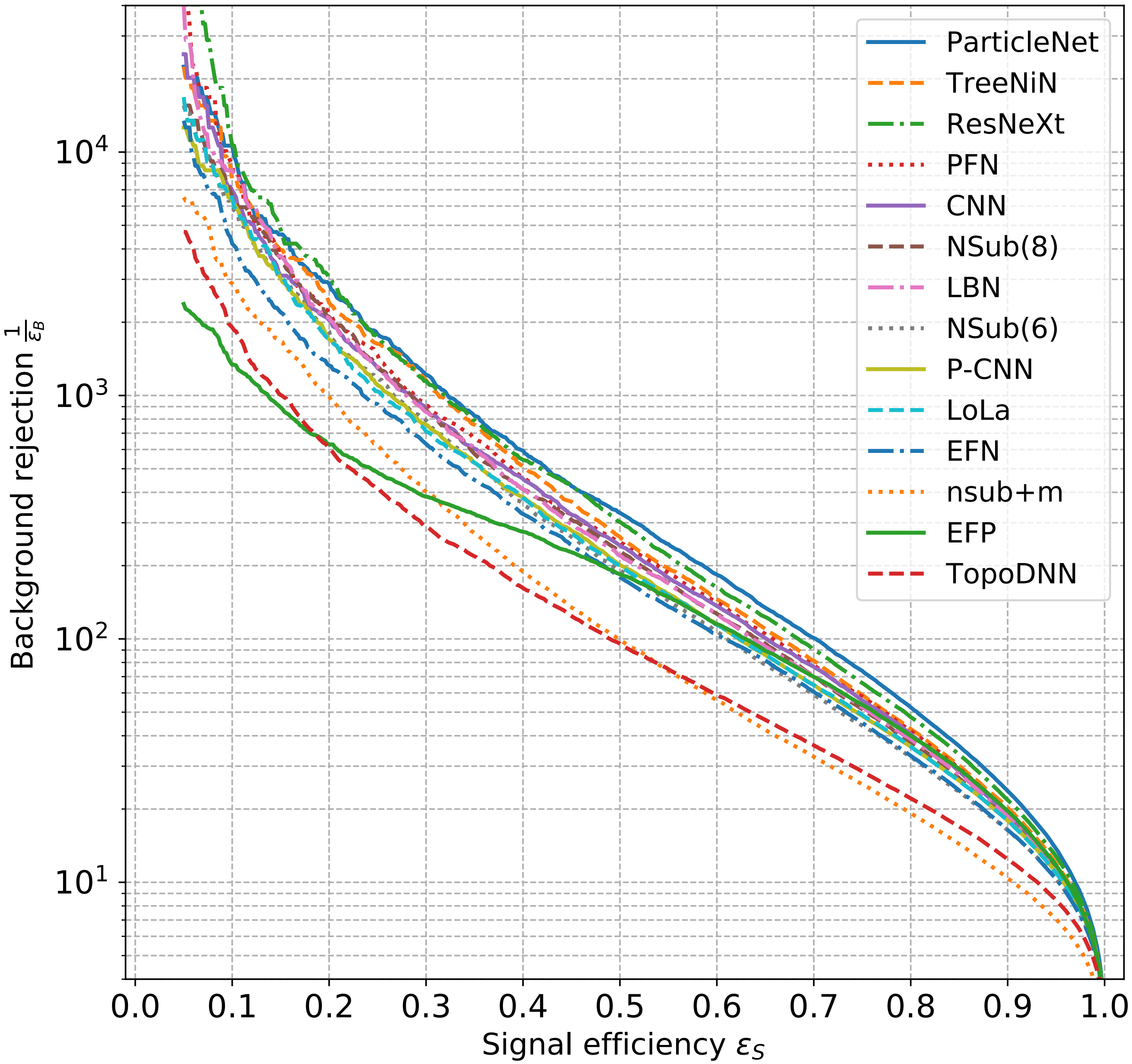
Top tagging efficiency



Can achieve factor of ~3 improvement over cut-based approaches and BDTs!

Community top tagging comparison

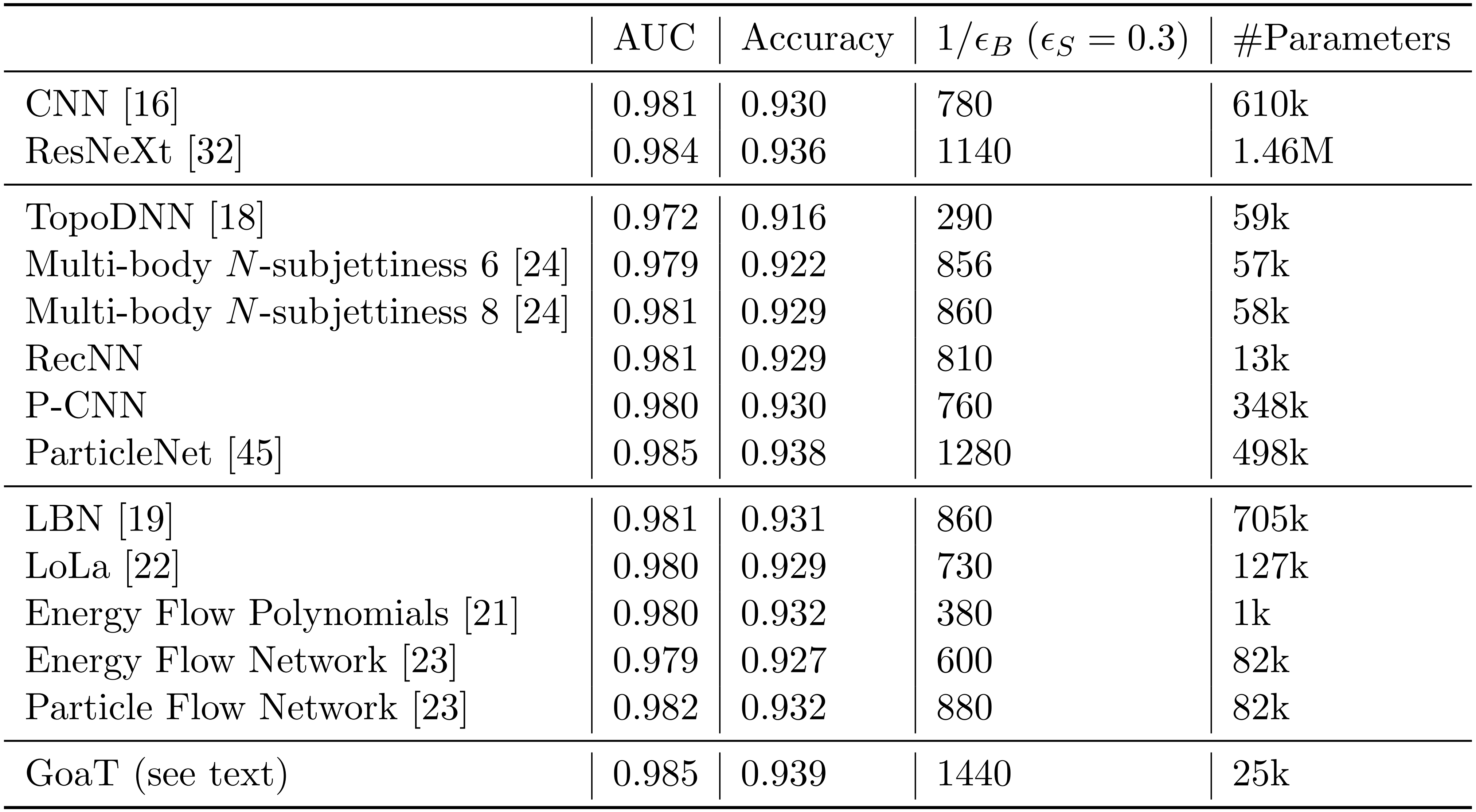
Kasieczka, Plehn et al 1902.09914



Apples-to-apples comparison of various deep learning top taggers on a common dataset.

Community top tagging comparison

Kasieczka, Plehn et al 1902.09914



	AUC	Accuracy	$1/\epsilon_B$ ($\epsilon_S = 0.3$)	#Parameters
CNN [16]	0.981	0.930	780	610k
ResNeXt [32]	0.984	0.936	1140	1.46M
TopoDNN [18]	0.972	0.916	290	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	856	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	860	58k
RecNN	0.981	0.929	810	13k
P-CNN	0.980	0.930	760	348k
ParticleNet [45]	0.985	0.938	1280	498k
LBN [19]	0.981	0.931	860	705k
LoLa [22]	0.980	0.929	730	127k
Energy Flow Polynomials [21]	0.980	0.932	380	1k
Energy Flow Network [23]	0.979	0.927	600	82k
Particle Flow Network [23]	0.982	0.932	880	82k
GoaT (see text)	0.985	0.939	1440	25k

Further improvements to our CNN are possible.
Have we found the optimal tagger??

Supervised vs Unsupervised ML

Top tagging is a prime example of supervised machine learning.

It is a straightforward classification task with fully-labeled (QCD or top) datasets.

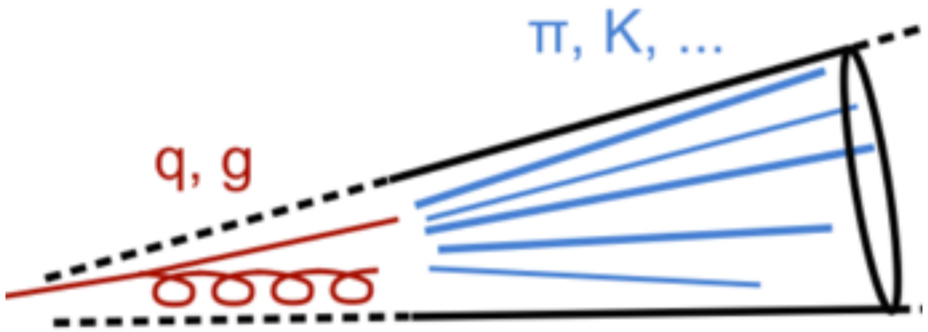
What if the data is not labeled — e.g. it is the actual LHC data and not simulation?

Can we apply ideas from unsupervised ML to discover patterns and features in the data?

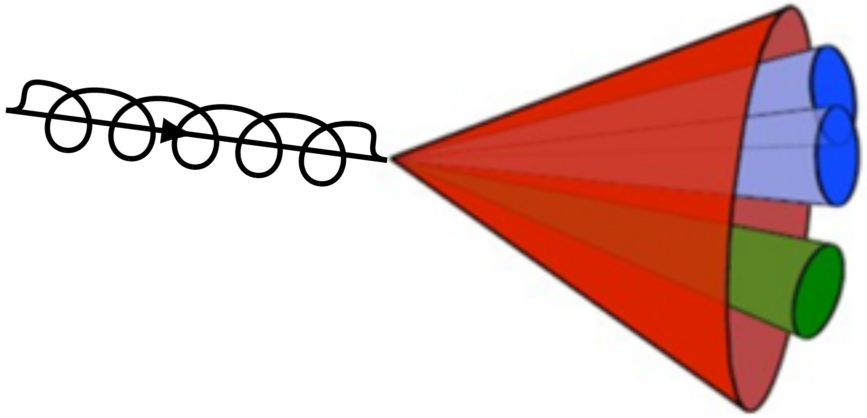
Can we discover unexpected new physics this way?

Statement of the problem

Consider a collection of jets at the LHC. [See Jesse and Anders talks for more on jets.]



Most will be from SM processes (quark/gluon showering and hadronization).



But a small fraction could be from an unknown (heavier) new physics particle with exotic properties.

How can we use ML algorithms to discover the exotic new particle without knowing what it looks like?

Statement of the problem

This is a standard anomaly detection problem in data science!

Train directly on the data

→ unsupervised anomaly detection (clustering)

Train on background-only (would probably need simulations for this)

→ weakly-supervised anomaly detection ("one-class classification")

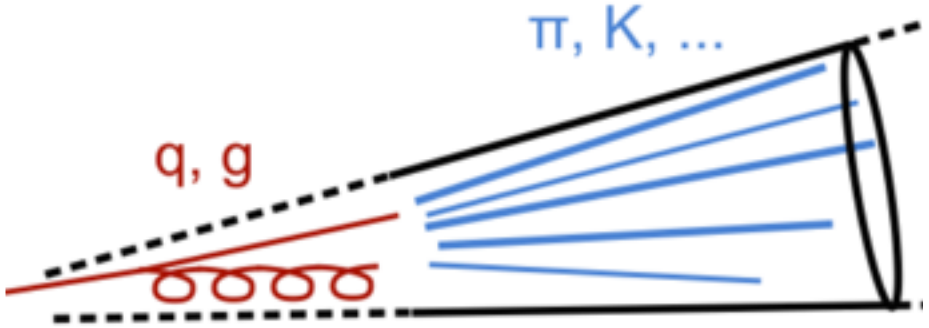
Sample definitions

Same jet specifications as for top tagging study. We used:

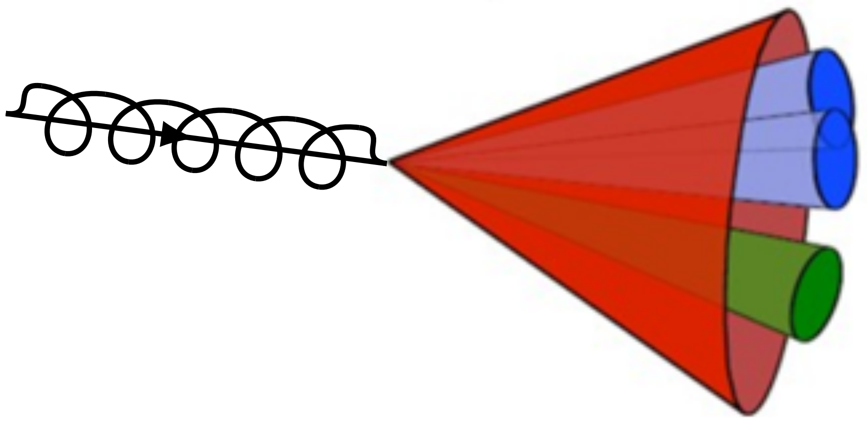
- q/g jets as background, and
- boosted tops and 400 GeV gluinos (decaying via RPV) as signal

We simulated

- 1.2M jets of each type
- using standard, open-source particle physics tools [Pythia8 and Delphes]
- and turned them into 37x37 grayscale images.



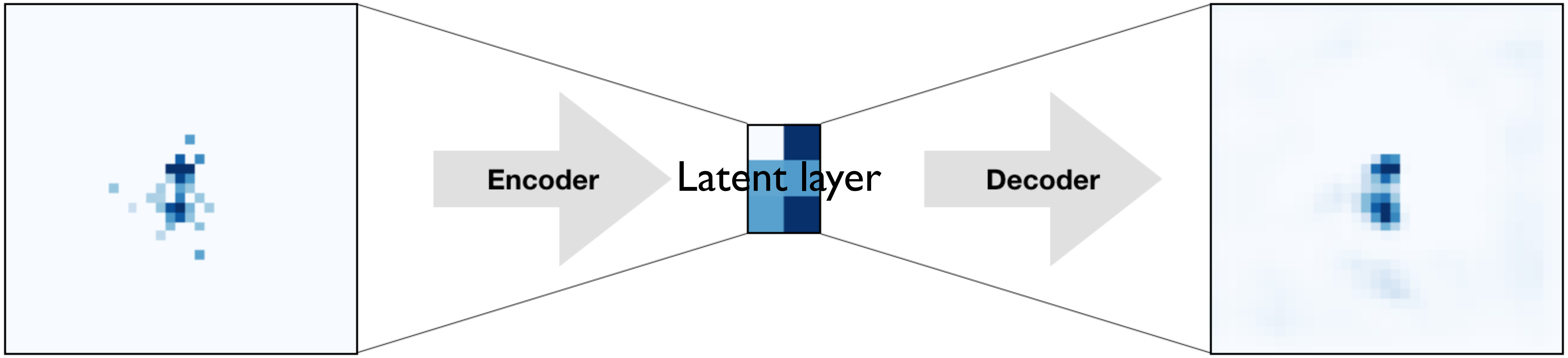
q/g jets



boosted tops and RPV gluinos

A promising idea: deep autoencoders

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992



An autoencoder maps an input into a "latent representation" and then attempts to reconstruct the original input from it.

The encoding is lossy, so the decoding cannot be perfect.

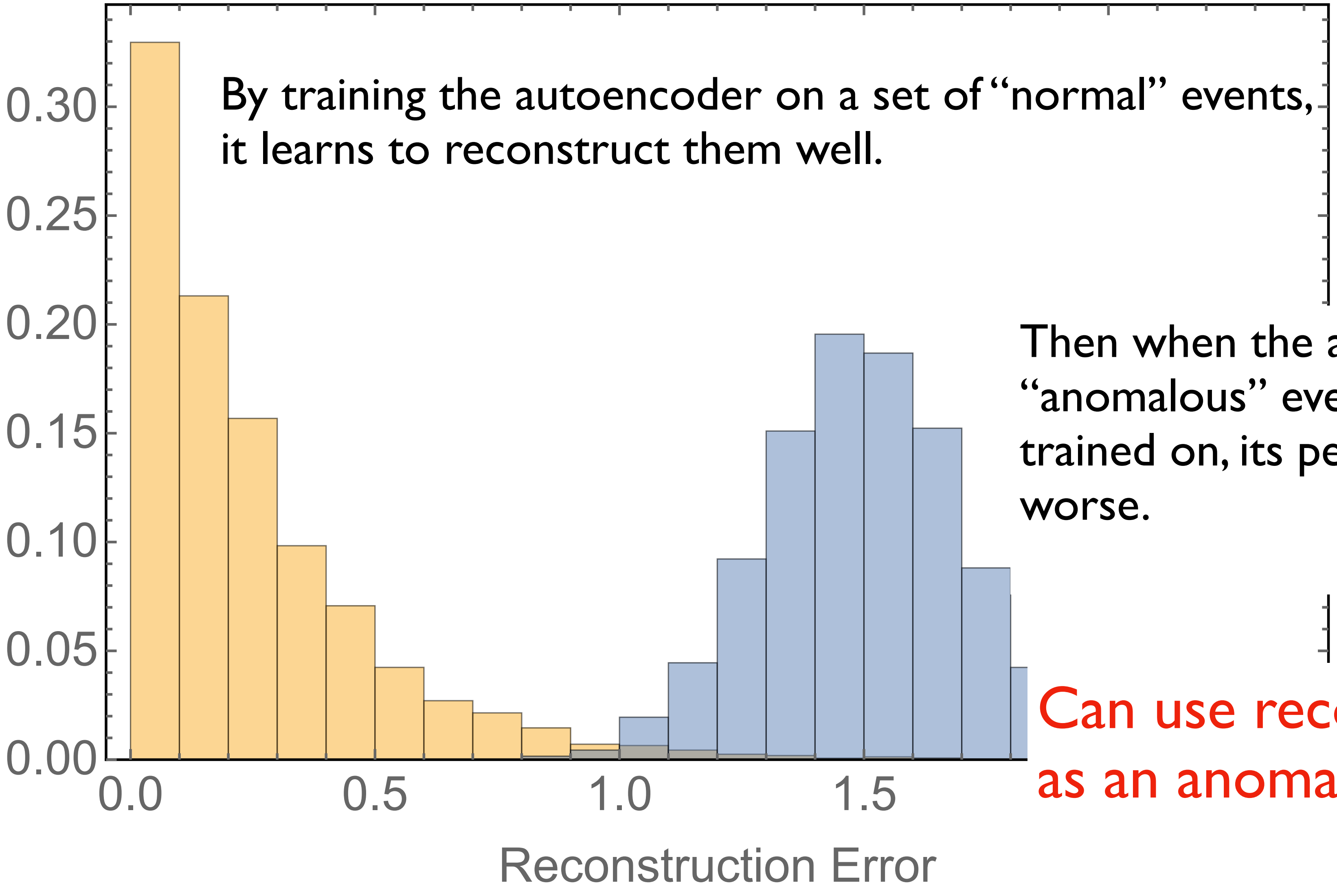
Some previous approaches:
Aguilar-Saavedra et al, "A generic anti-QCD jet tagger" 1709.01087
Collins et al, "CWoLa Hunting" 1805.02664
Hajer et al "Novelty Detection Meets Collider Physics" 1807.10261

Deep autoencoders for anomaly detection

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992

Quantify AE performance using reconstruction error:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

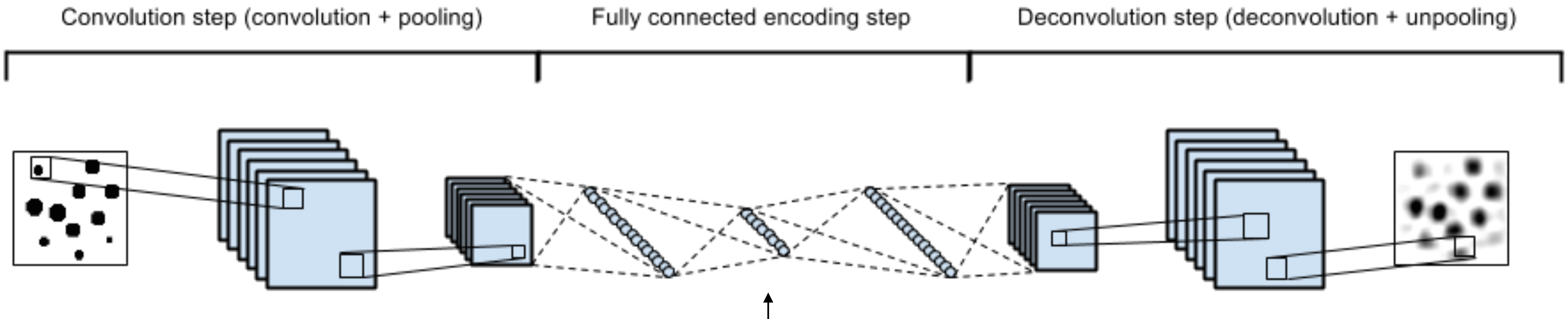


By training the autoencoder on a set of "normal" events, it learns to reconstruct them well.

Then when the autoencoder encounters "anomalous" events that it was not trained on, its performance should be worse.

Can use reconstruction error as an anomaly threshold!

Autoencoder architecture

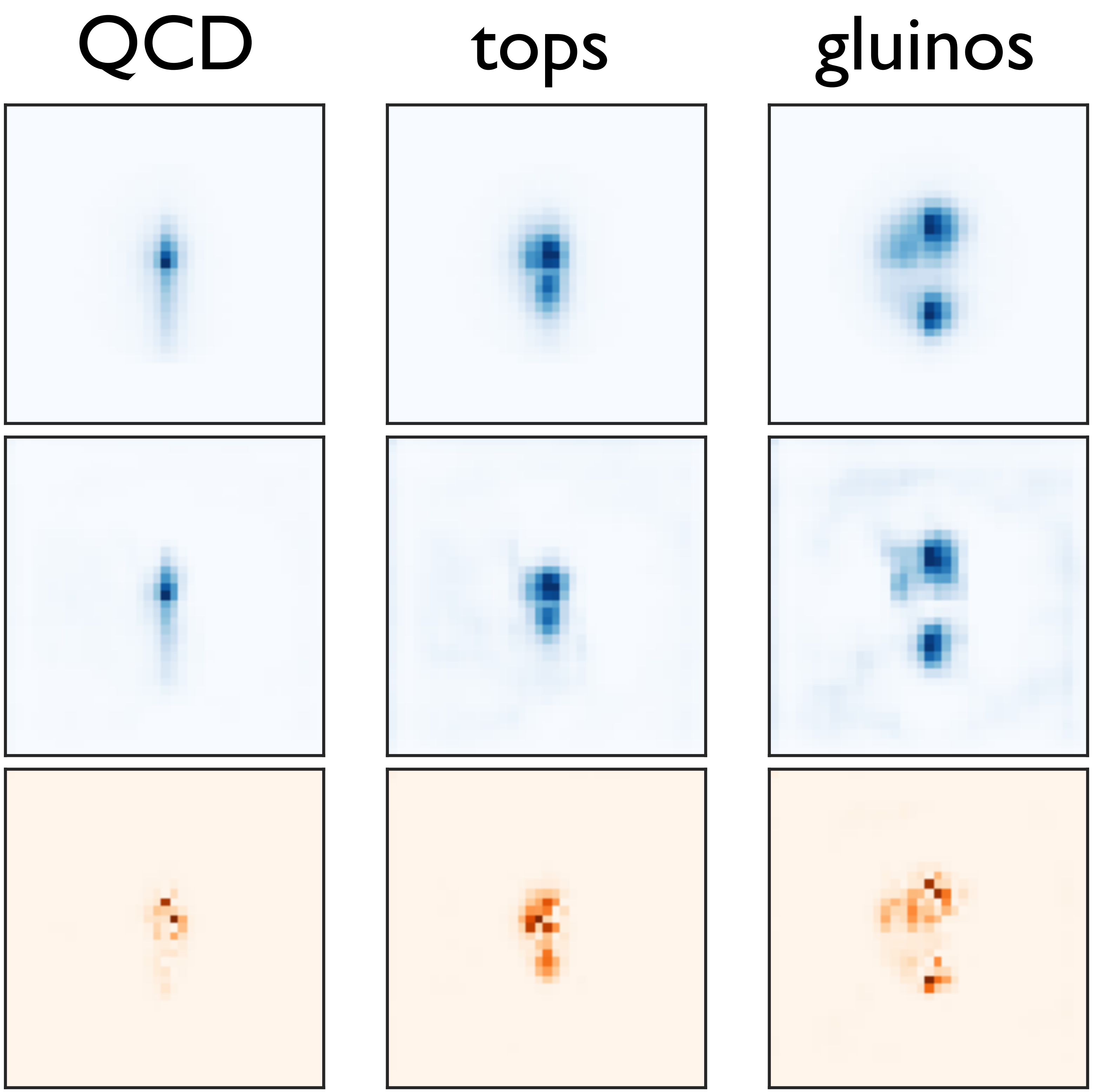

Convolution step (convolution + pooling)
Fully connected encoding step
Deconvolution step (deconvolution + unpooling)

M. Ke, C. Lin, Q. Huang (2017)

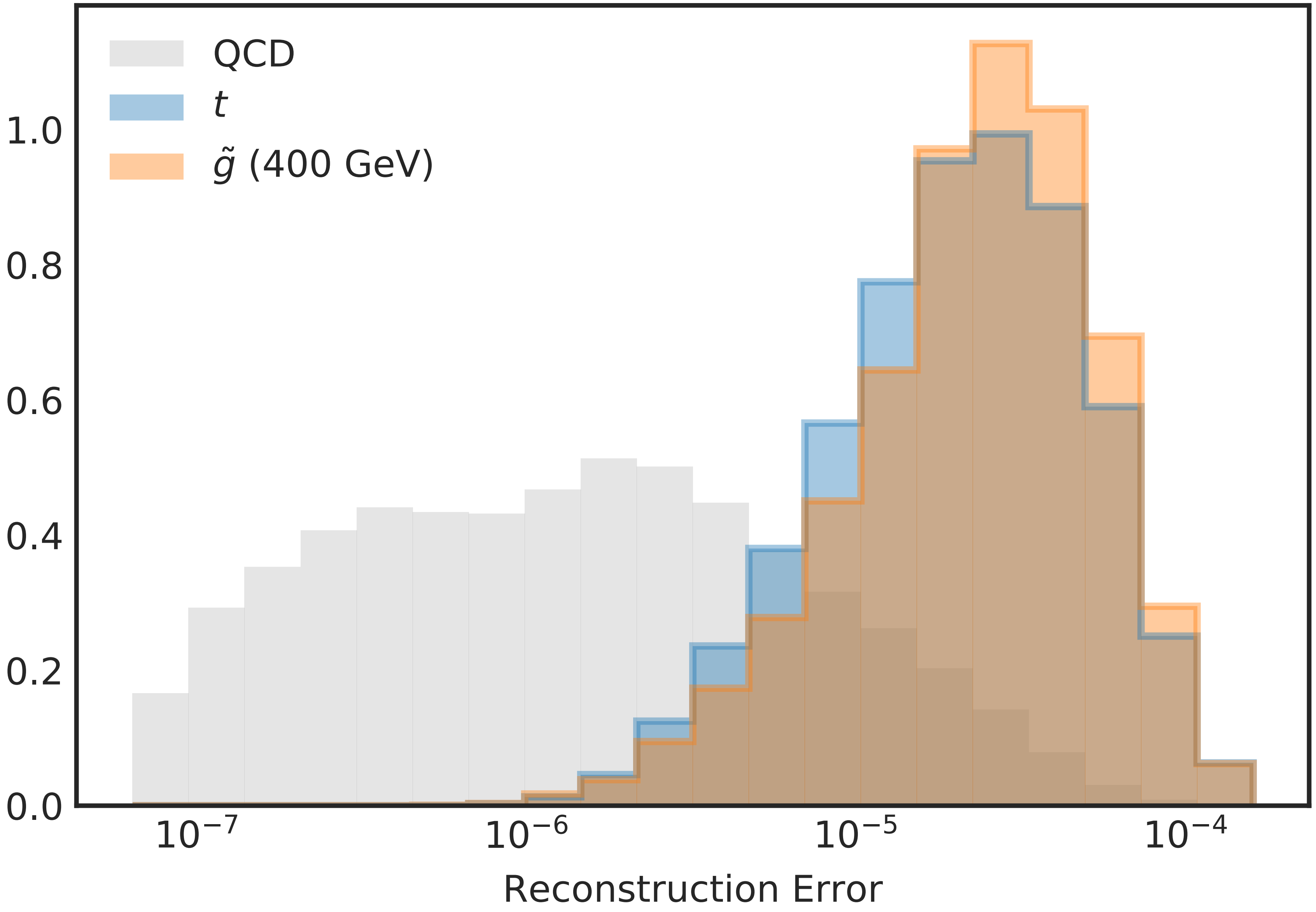
Encoder **Latent space** **Decoder**

128C3-MP2-128C3-MP2-128C3-32N-6N-32N-12800N-128C3-US2-128C3-US2-1C3

Performance should be worse on “anomalous” events that autoencoder was not trained on.



Can use reconstruction error as an anomaly threshold.

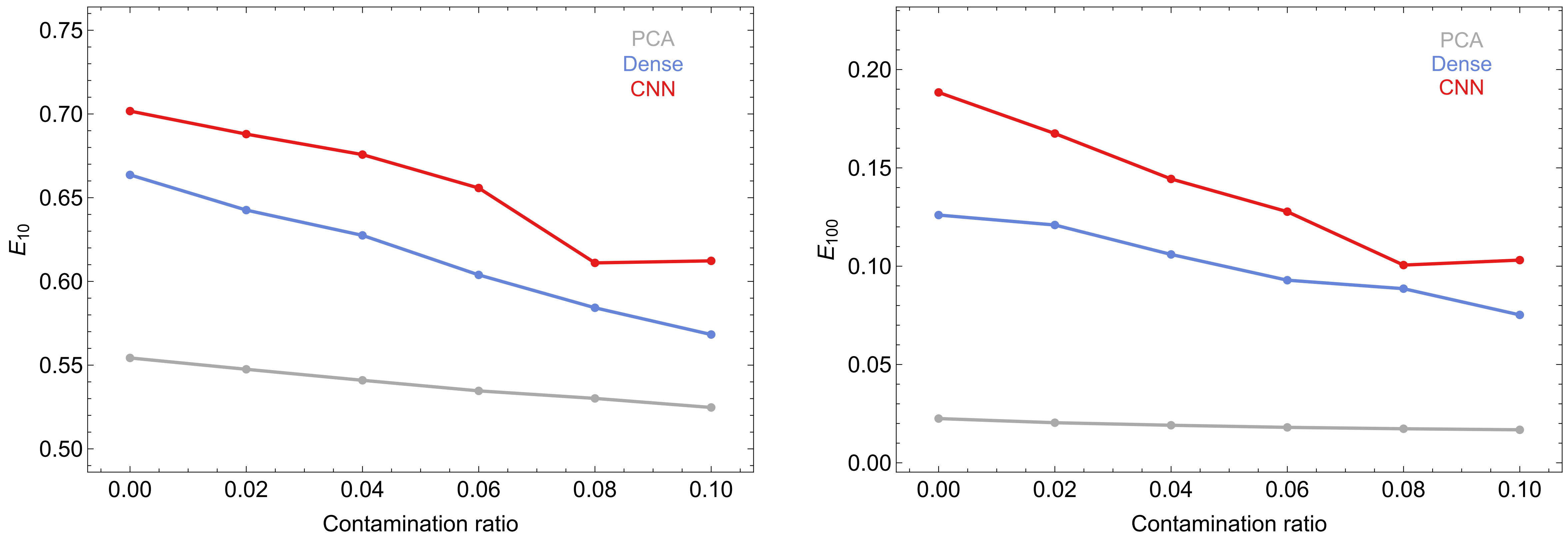


The algorithm works when trained on QCD backgrounds!

Fully unsupervised learning

Train on sample of QCD background "contaminated" with a small fraction of signal.

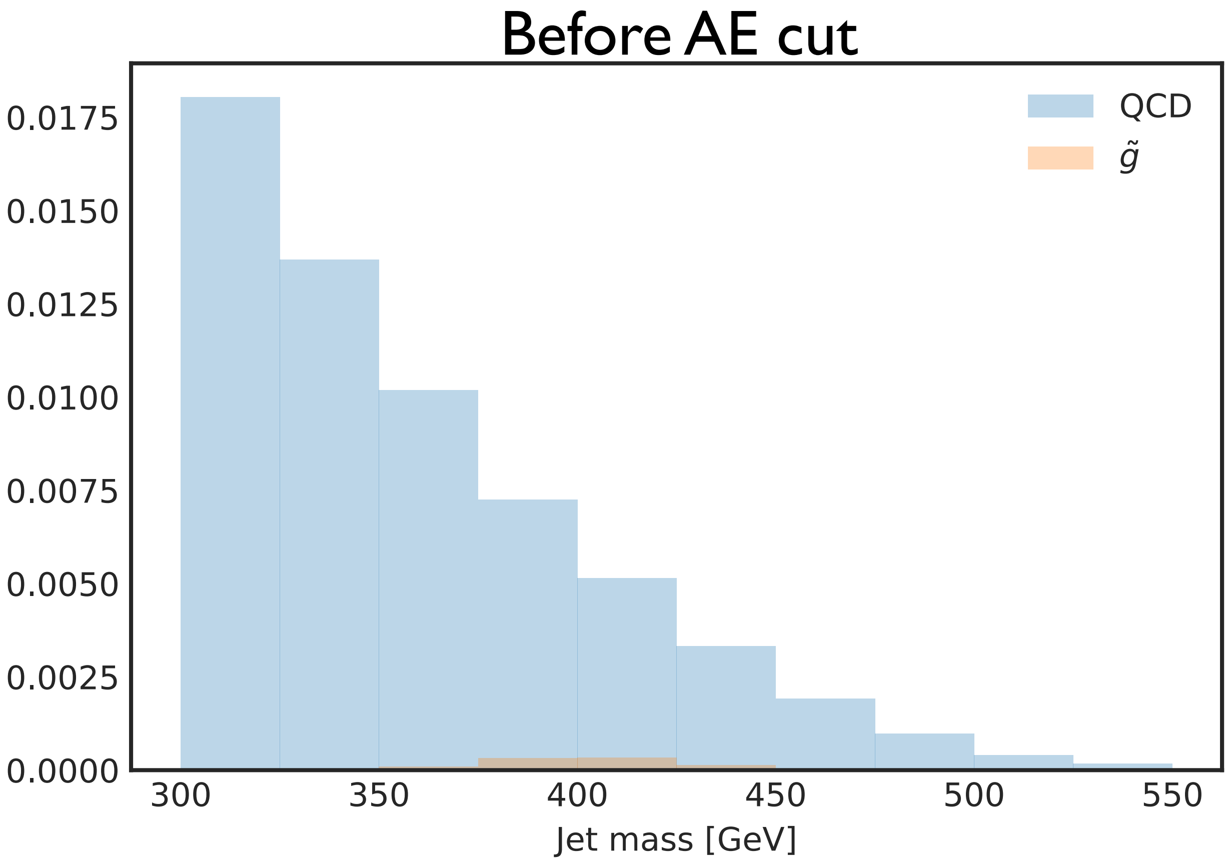
Representative of actual data.

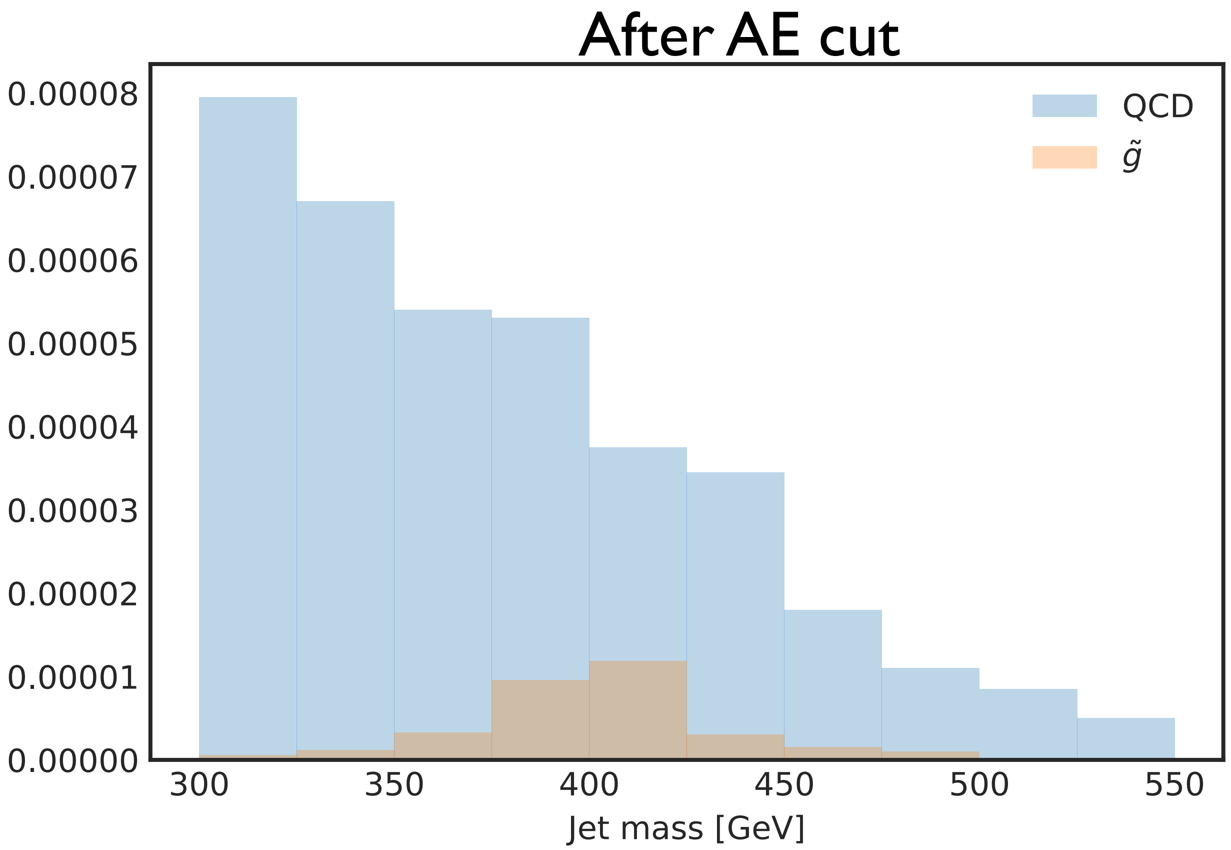


(E_x = signal efficiency at bg rejection = x)

Performance of AE surprisingly robust even up to 10% contamination!

Bump hunt with deep autoencoder


Before AE cut


After AE cut

Can train directly on data that contains 400 GeV gluinos,
use the AE to clean away "boring" SM events,
and improve S/N by a lot

Could really discover new physics this way!

Summary

Deep learning has revolutionized the field of artificial intelligence and has given birth to a number of stunning real-world applications.

The revolution is coming to high-energy physics!

In this talk, we gave an overview of deep learning applications to the LHC.

Then we focused on two promising applications:

- Top tagging with jet images and CNNs (supervised learning)
  - ➡ Enormous gains in performance over cut-based and shallow ML methods.
- Deep autoencoders for open-ended anomaly detection (unsupervised learning)
  - ➡ Novel proposal for searching for new physics in the data without prejudice.

Summary

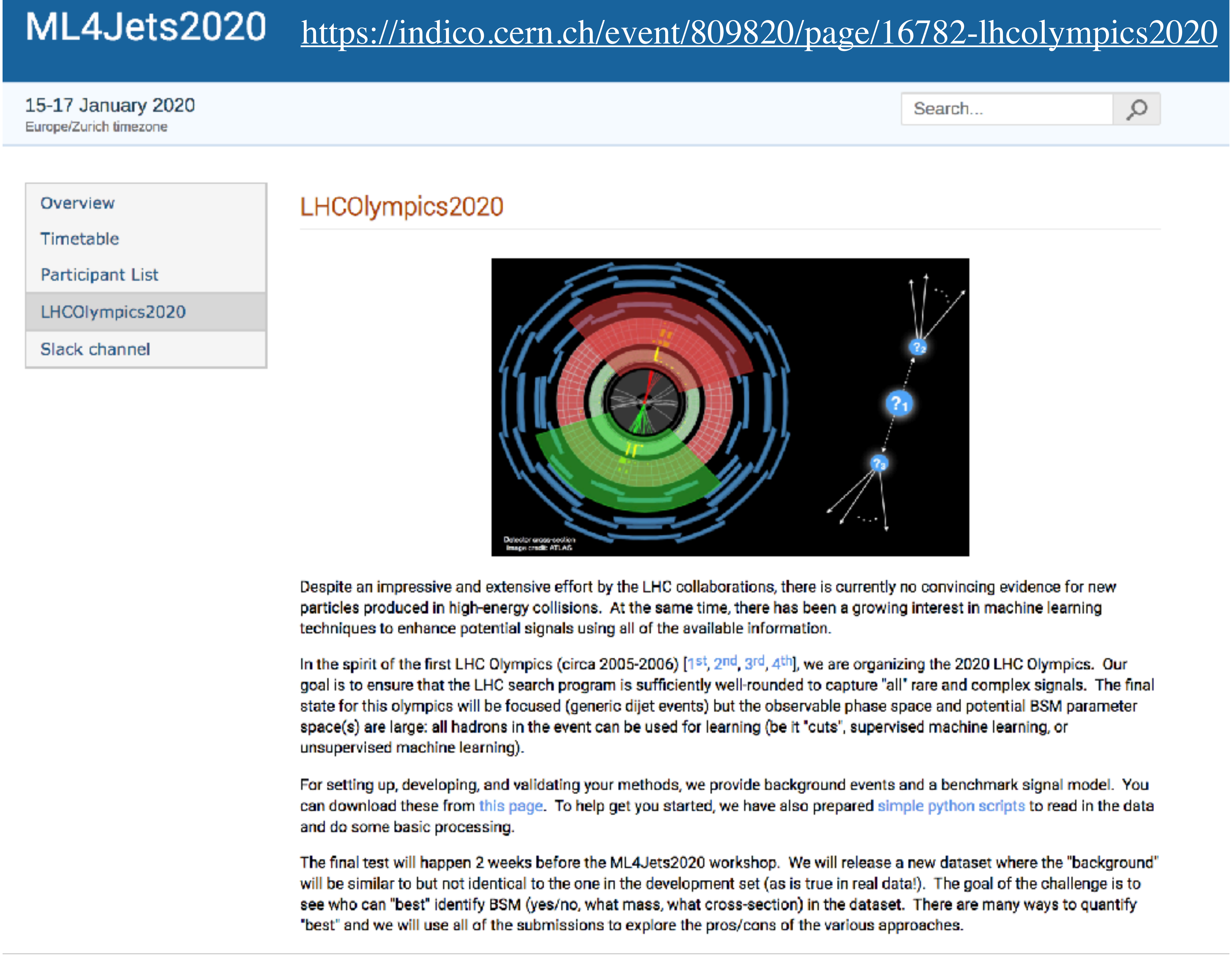
The Standard Model has withstood the test of time for over 40 years.

Despite knowing that new physics beyond the SM is out there, we have yet to see any evidence for it at the LHC.

We need more ideas for how to search for the unexpected at the LHC.

- Autoencoders for anomaly detection are a promising direction but there are surely many more!

Input from the ML experts in the audience would be most appreciated!



ML4Jets2020

https://indico.cern.ch/event/809820/page/16782-lhcolympics2020

15-17 January 2020
Europe/Zurich timezone

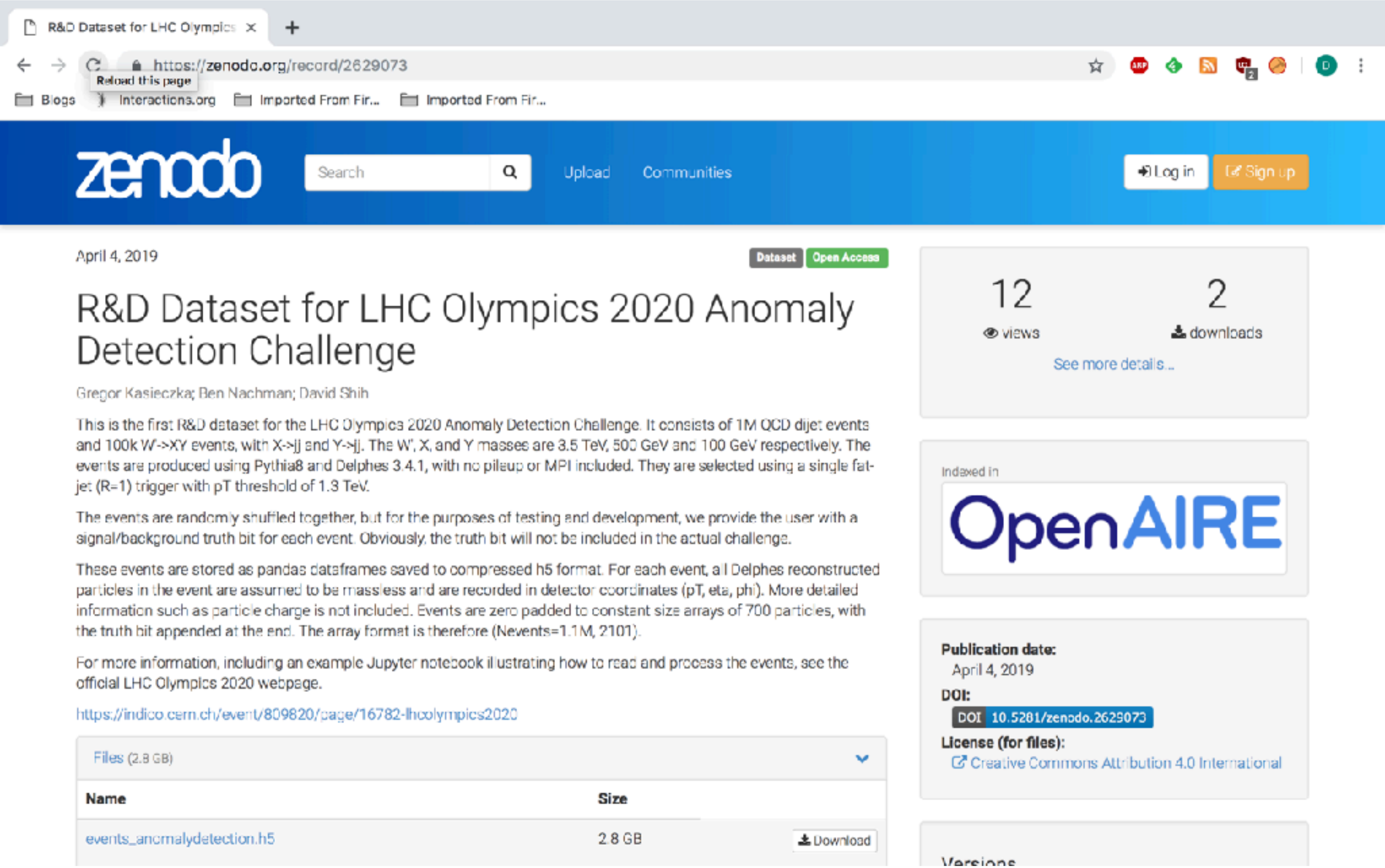
LHCOlympics2020

Despite an impressive and extensive effort by the LHC collaborations, there is currently no convincing evidence for new particles produced in high-energy collisions. At the same time, there has been a growing interest in machine learning techniques to enhance potential signals using all of the available information.

In the spirit of the first LHC Olympics (circa 2005-2006) [1st, 2nd, 3rd, 4th], we are organizing the 2020 LHC Olympics. Our goal is to ensure that the LHC search program is sufficiently well-rounded to capture "all" rare and complex signals. The final state for this olympics will be focused (generic dijet events) but the observable phase space and potential BSM parameter space(s) are large: all hadrons in the event can be used for learning (be it "cuts", supervised machine learning, or unsupervised machine learning).

For setting up, developing, and validating your methods, we provide background events and a benchmark signal model. You can download these from this page. To help get you started, we have also prepared simple python scripts to read in the data and do some basic processing.

The final test will happen 2 weeks before the ML4Jets2020 workshop. We will release a new dataset where the "background" will be similar to but not identical to the one in the development set (as is true in real data!). The goal of the challenge is to see who can "best" identify BSM (yes/no, what mass, what cross-section) in the dataset. There are many ways to quantify "best" and we will use all of the submissions to explore the pros/cons of the various approaches.



April 4, 2019

R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge

Gregor Kasieczka; Ben Nachman; David Shih

This is the first R&D dataset for the LHC Olympics 2020 Anomaly Detection Challenge. It consists of 1M QCD dijet events and 100k W'->XY events, with X->jj and Y->jj. The W', X, and Y masses are 3.5 TeV, 500 GeV and 100 GeV respectively. The events are produced using Pythia8 and Delphes 3.4.1, with no pileup or MPI included. They are selected using a single fat-jet (R=1) trigger with pT threshold of 1.3 TeV.

The events are randomly shuffled together, but for the purposes of testing and development, we provide the user with a signal/background truth bit for each event. Obviously, the truth bit will not be included in the actual challenge.

These events are stored as pandas dataframes saved to compressed h5 format. For each event, all Delphes reconstructed particles in the event are assumed to be massless and are recorded in detector coordinates (pT, eta, phi). More detailed information such as particle charge is not included. Events are zero padded to constant size arrays of 700 particles, with the truth bit appended at the end. The array format is therefore (Nevents=1.1M, 2101).

For more information, including an example Jupyter notebook illustrating how to read and process the events, see the official LHC Olympics 2020 webpage.

https://indico.cern.ch/event/809820/page/16782-lhcolympics2020

Files (2.8 GB)

Name	Size	
events_anomalydetection.h5	2.8 GB	Download

12 views

2 downloads

See more details...

Publication date:
April 4, 2019

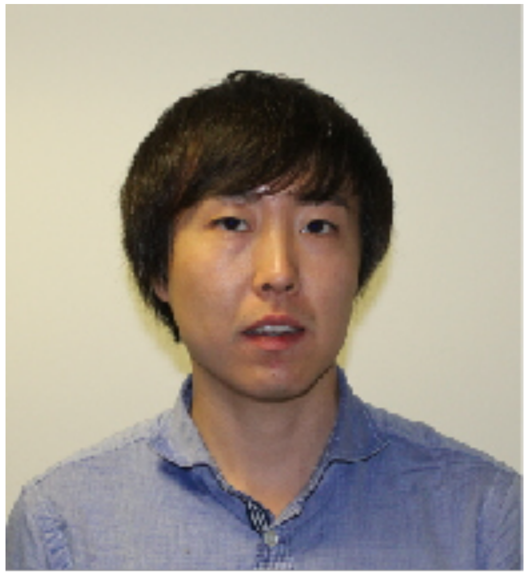
DOI:
DOI 10.5281/zenodo.2629073

License (for files):
Creative Commons Attribution 4.0 International

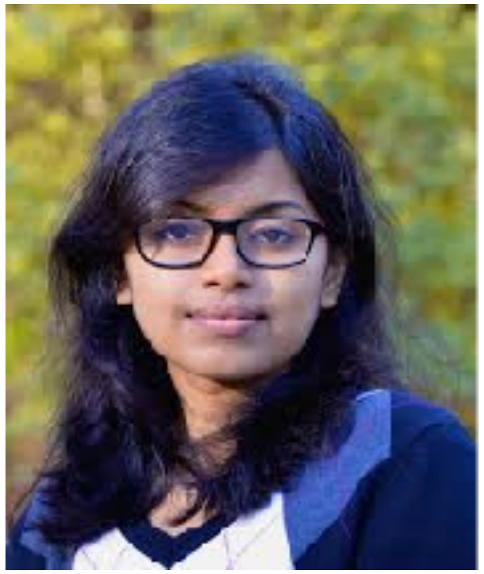
Thanks for your attention!

Sebastian Macaluso

Marco Farina



Yuichiro Nakai



Dipsikha Debnath

Matt Buckley

Scott Thomas

Backup material

Autoencoder architectures

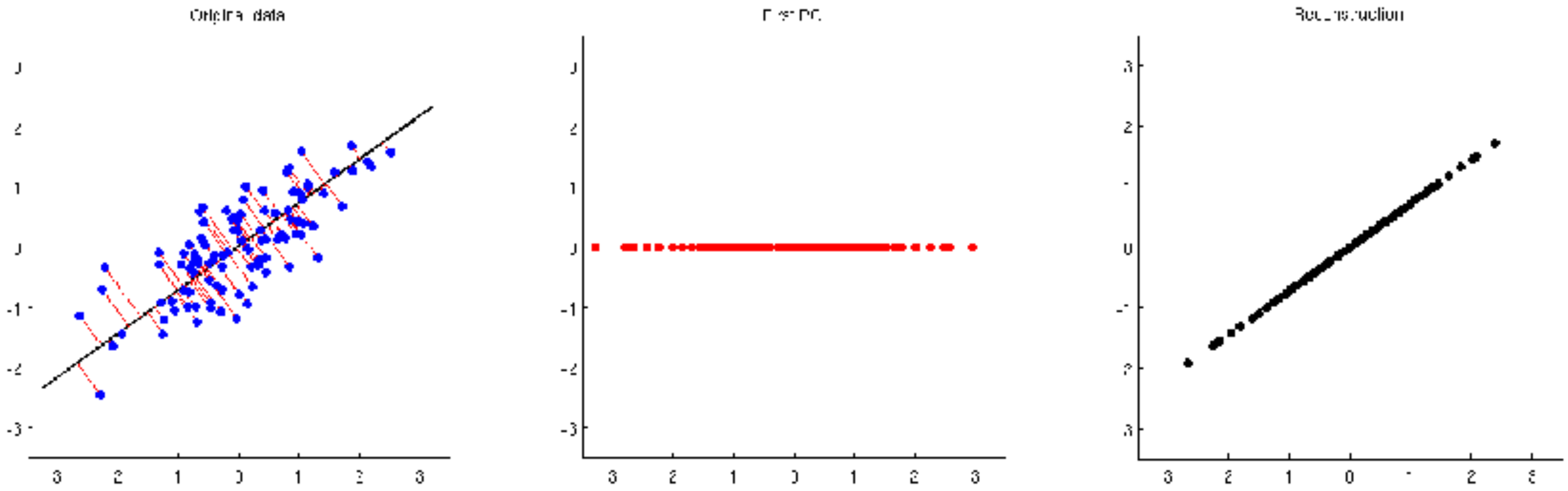
We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convolutional NN

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)



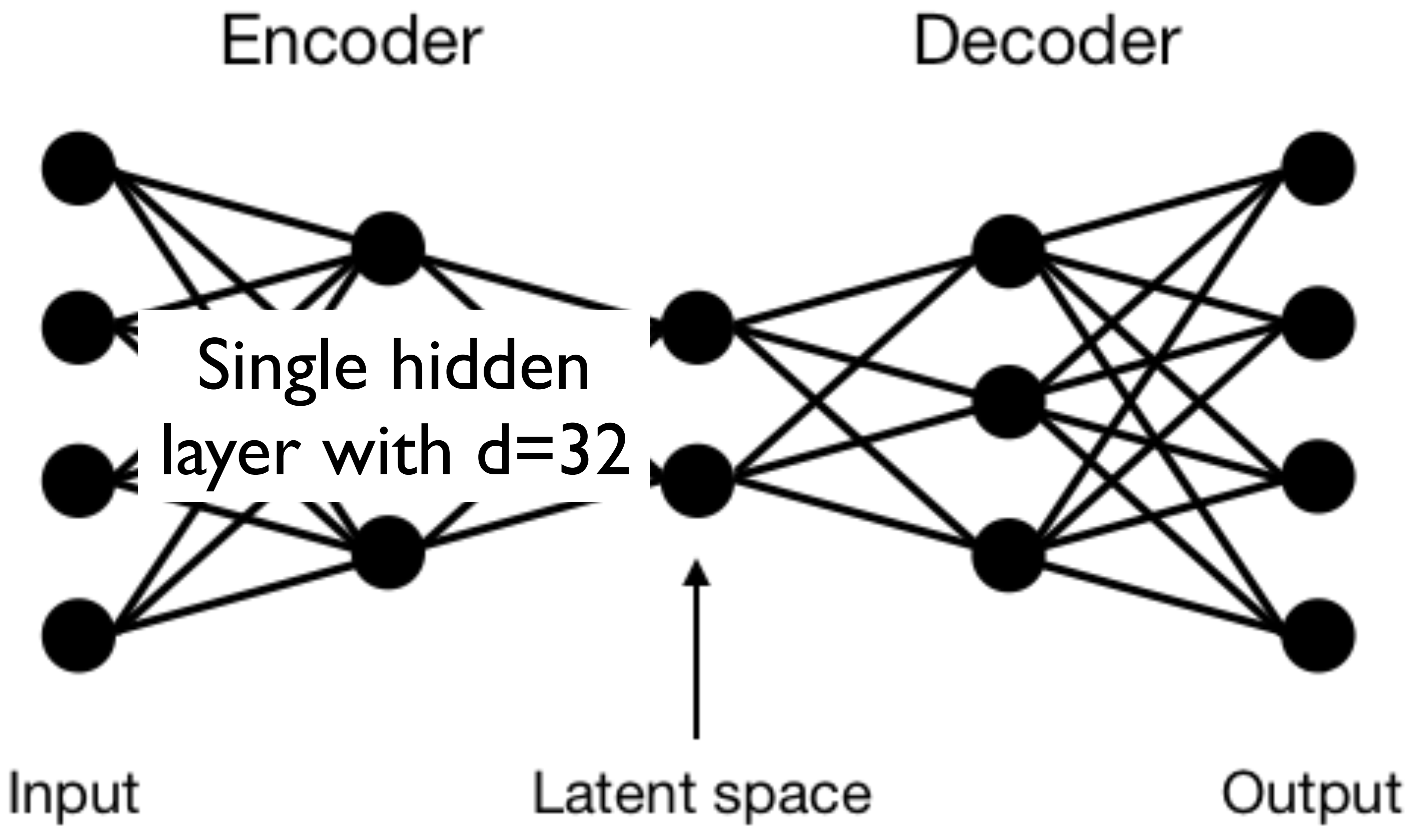
Project onto the first d PCA eigenvectors $z = \mathcal{P}_d x_{in}$

Inverse transform to reconstruct original input $x_{out} = \mathcal{P}_d^T z = \mathcal{P}_d^T \mathcal{P}_d x_{in}$

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convoluti



Flatten input into column vector

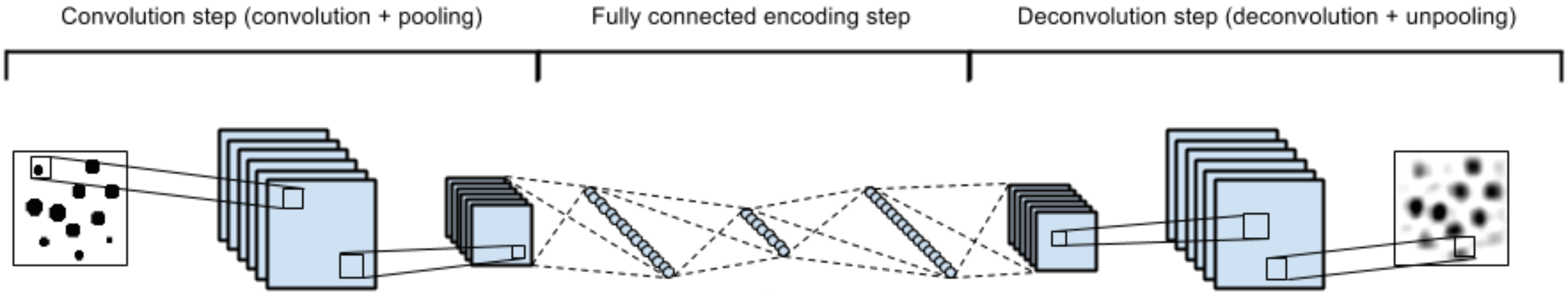
Single hidden layer with d=32

(dimension d)

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convolutional NN


Convolution step (convolution + pooling)
Fully connected encoding step
Deconvolution step (deconvolution + unpooling)

M. Ke, C. Lin, Q. Huang (2017)

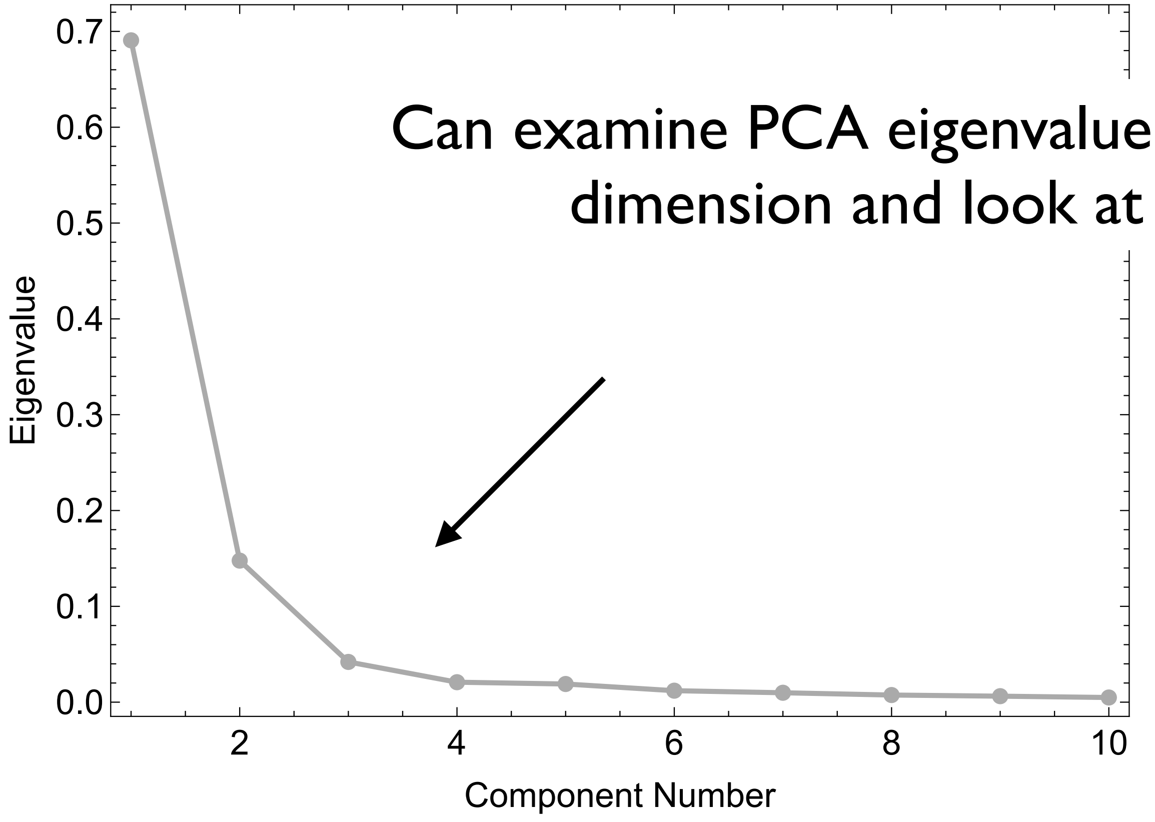
Encoder **Latent space** **Decoder**

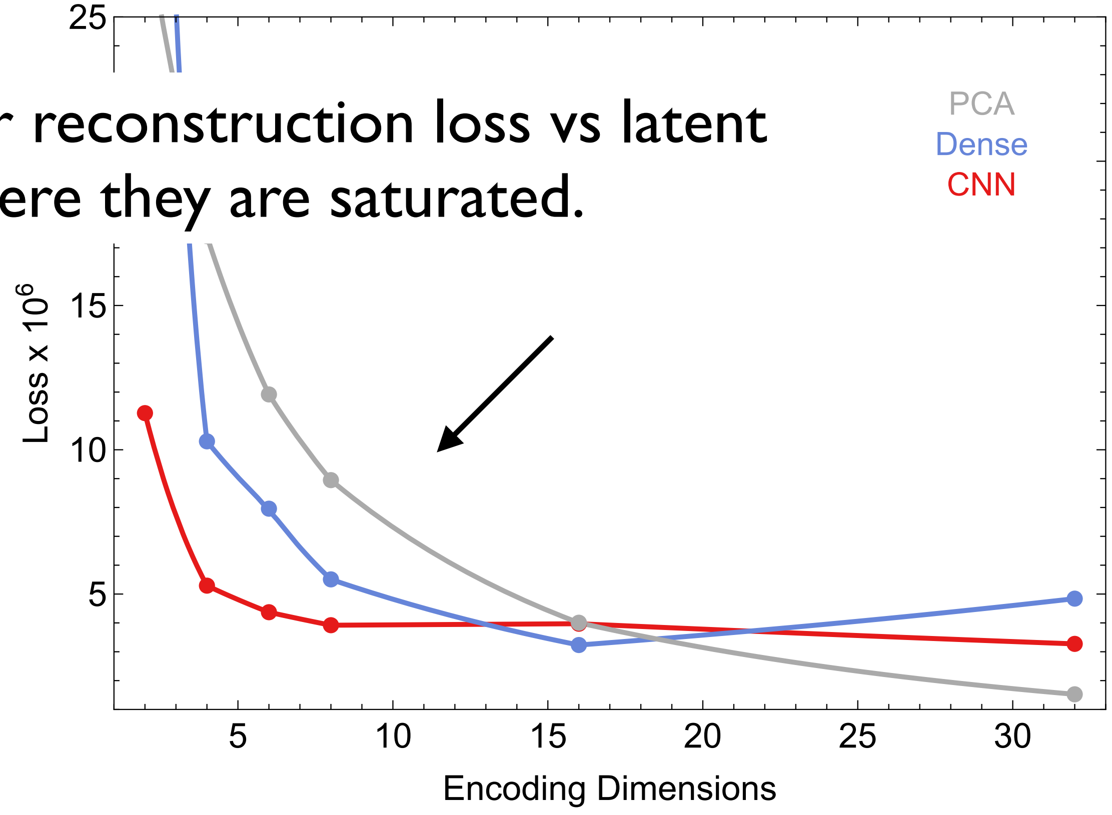
128C3-MP2-128C3-MP2-128C3-32N-6N-32N-12800N-128C3-US2-128C3-US2-1C3

Choosing the latent dimension

d too large → autoencoder becomes identity transform
d too small → autoencoder cannot learn all the features

Should choose the latent dimension in an unsupervised manner
(ie without optimizing on a specific signal)

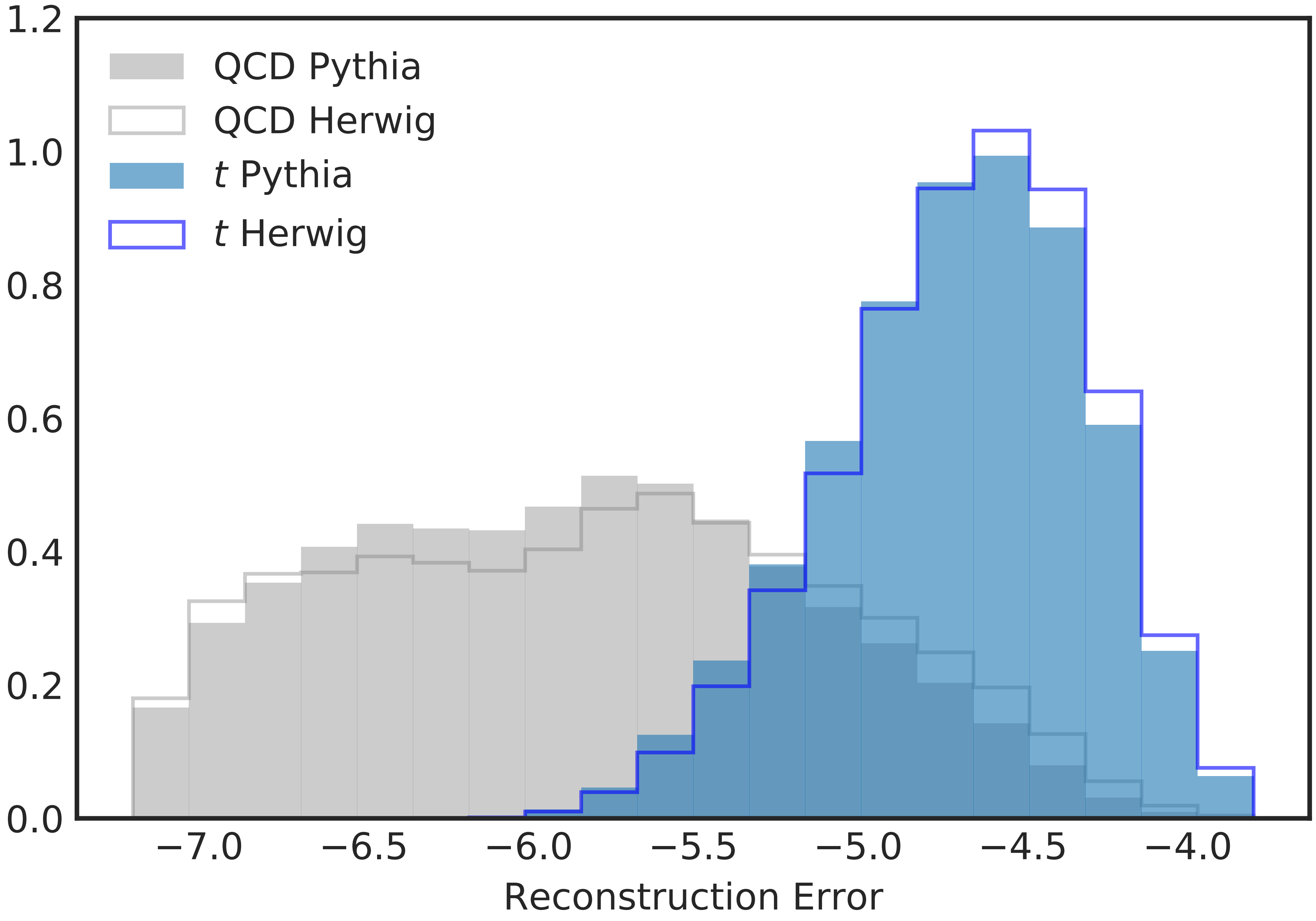




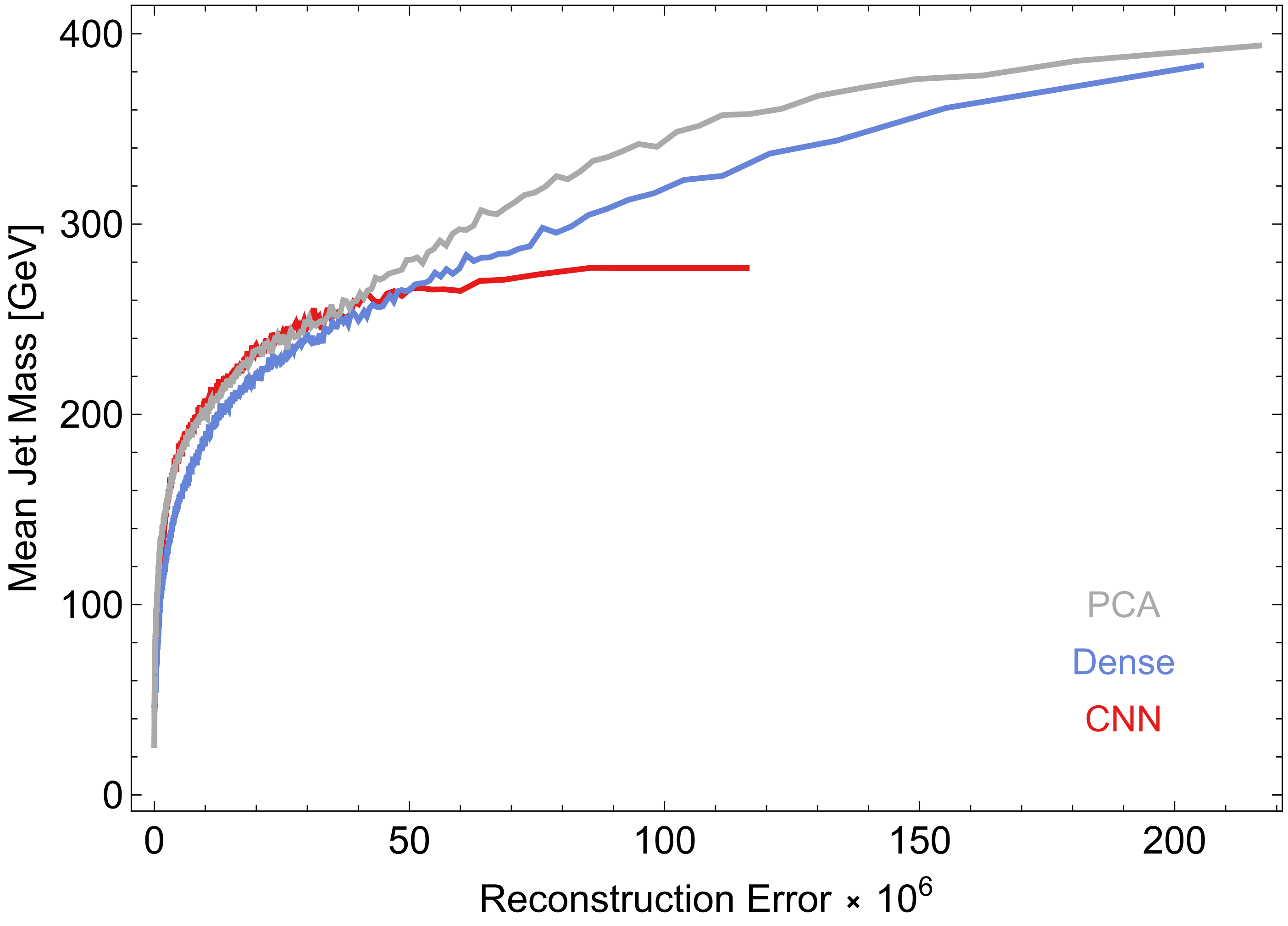
Can examine PCA eigenvalues or reconstruction loss vs latent dimension and look at where they are saturated.

We chose d=6

Robustness with other Monte Carlo



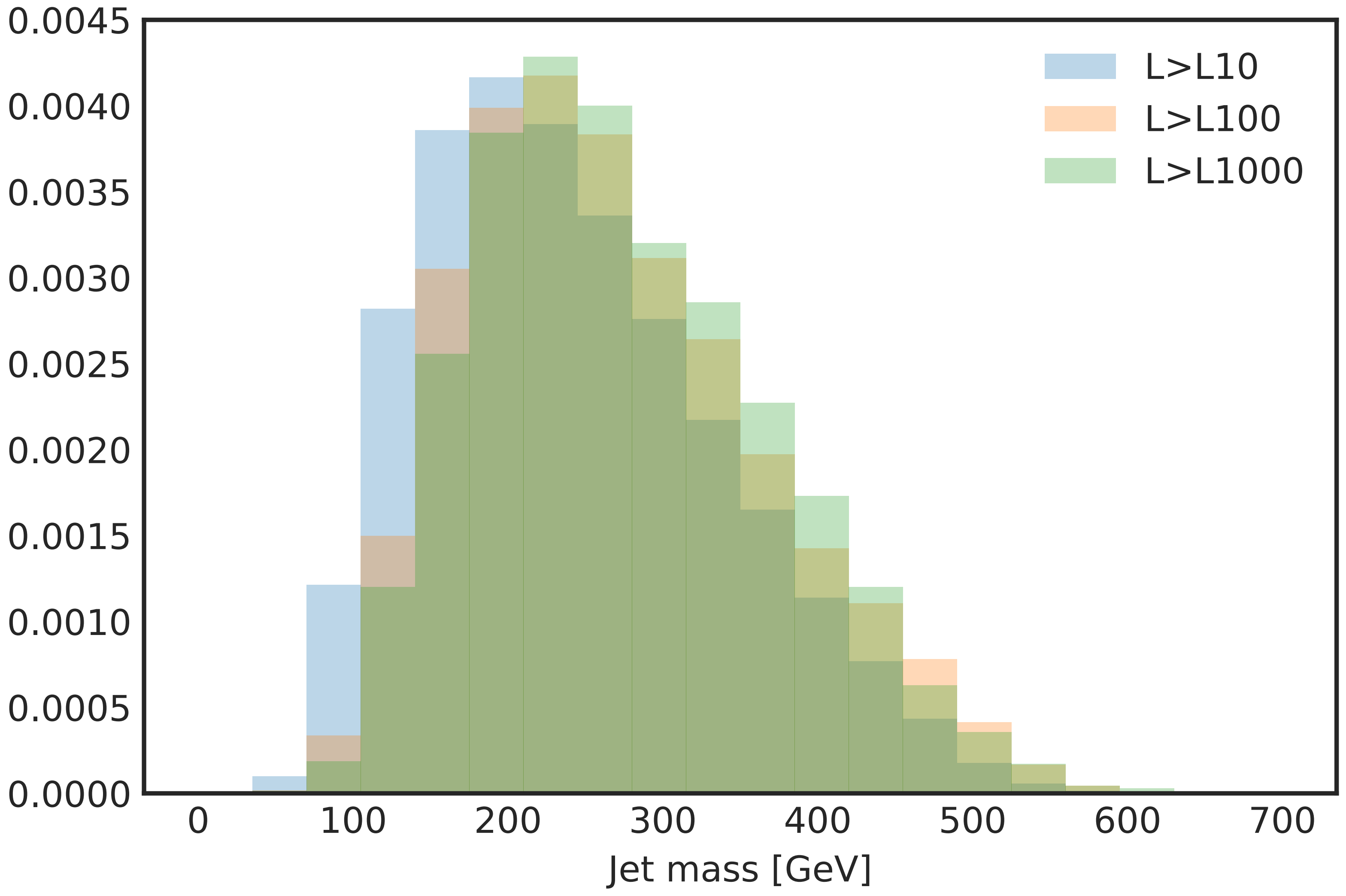
Correlation with jet mass



Indeed, this is confirmed by looking at mean jet mass in bins of reconstruction error for the QCD background.

CNN is no longer correlated with jet mass for m≳250 GeV

Correlation with jet mass



The QCD jet mass distribution is stable against harder cuts on the reconstruction error, for the CNN autoencoder.