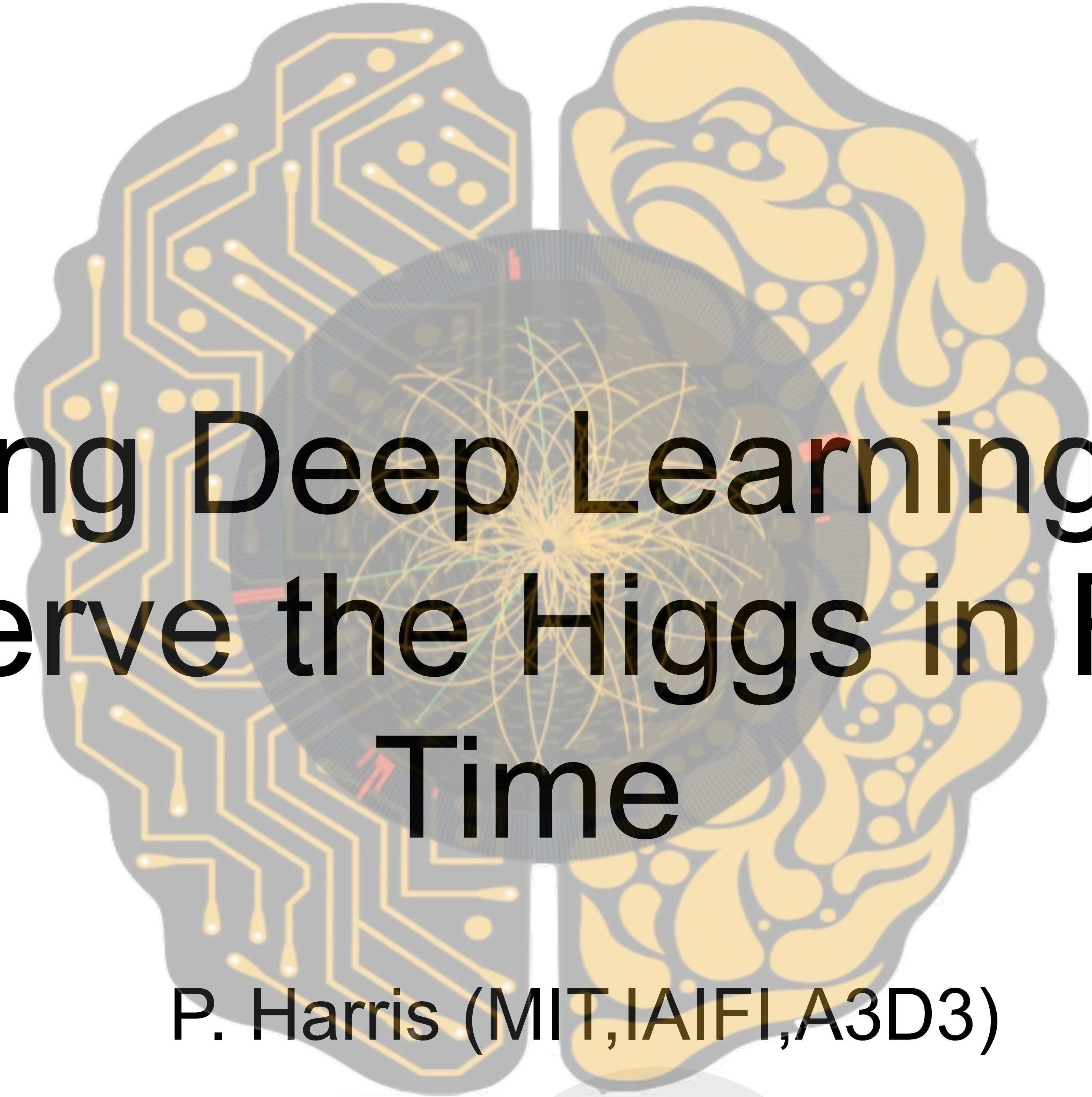


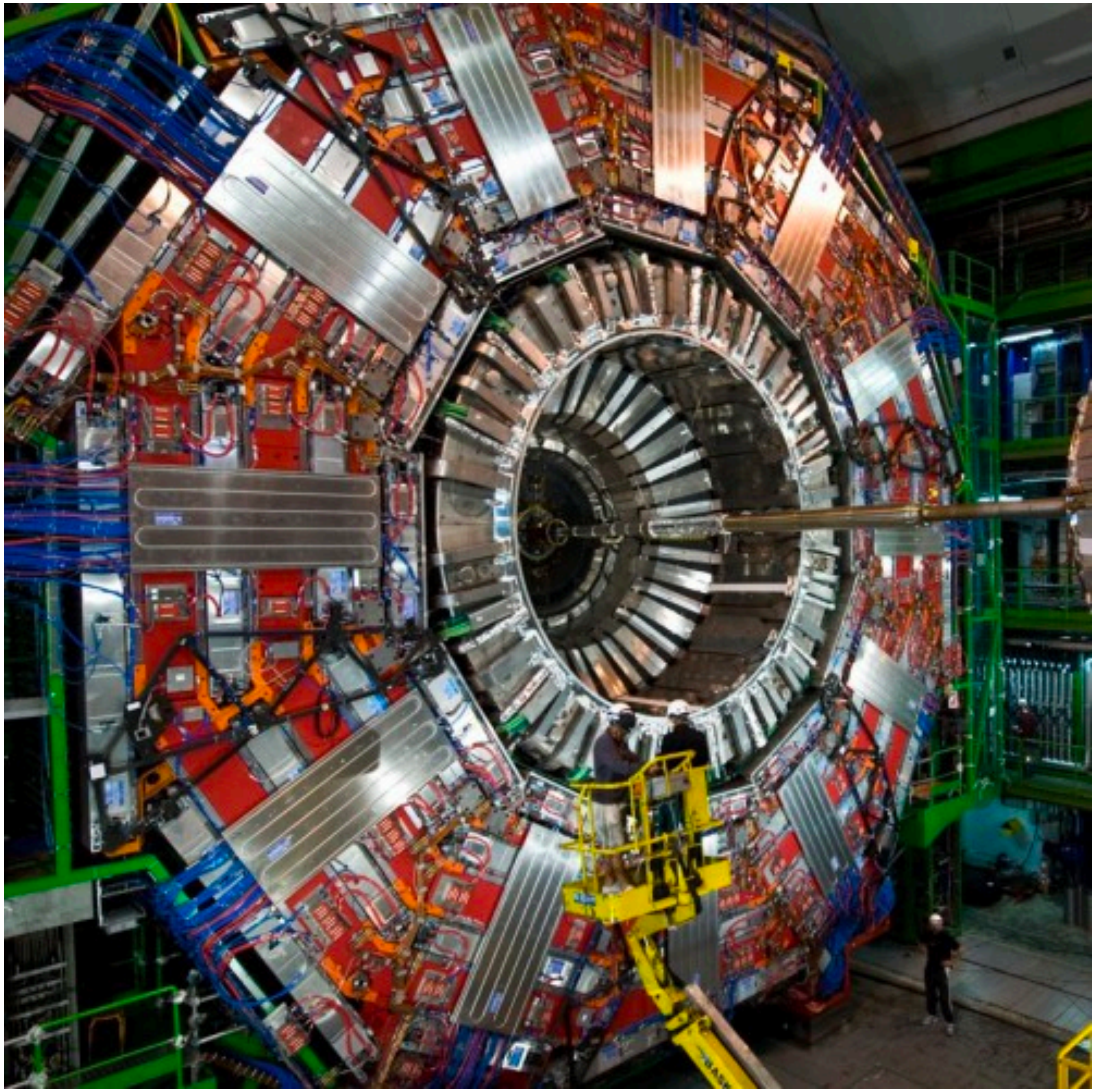
Using Deep Learning to Observe the Higgs in Real Time

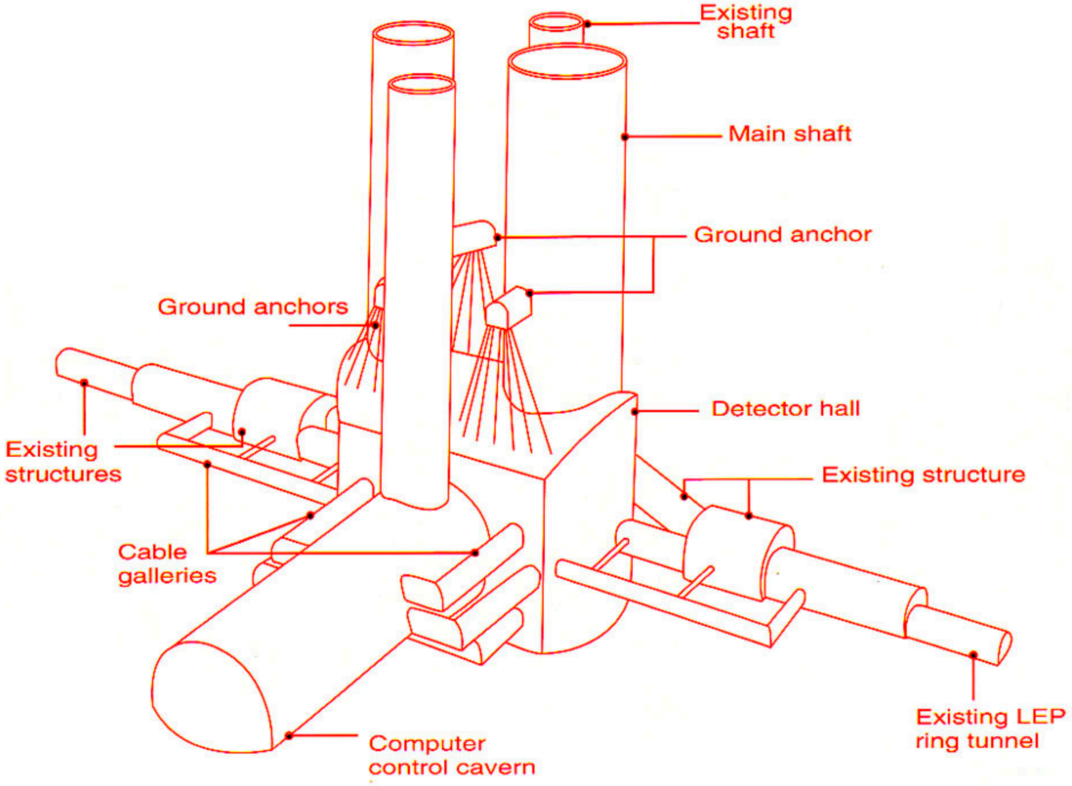
P. Harris (MIT,IAIFI,A3D3)

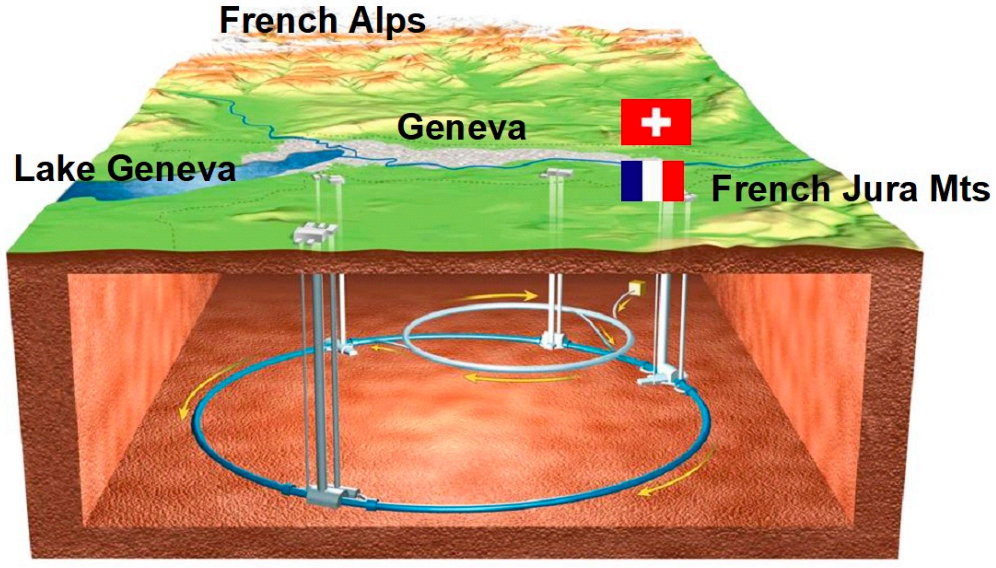
Accelerated AI Algorithms for Data-Driven Discovery

Large Hadron Collider

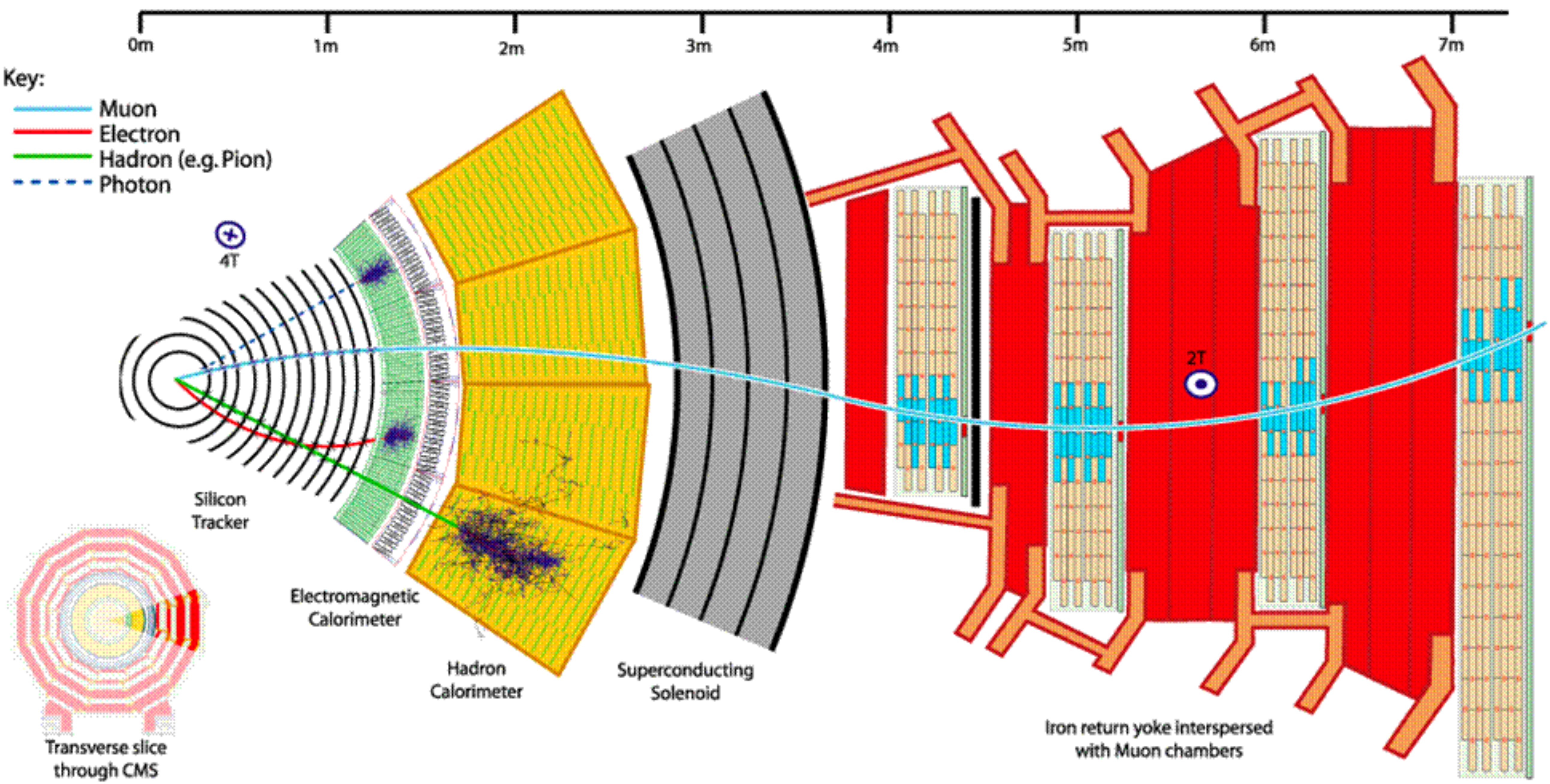
Detector at the LHC





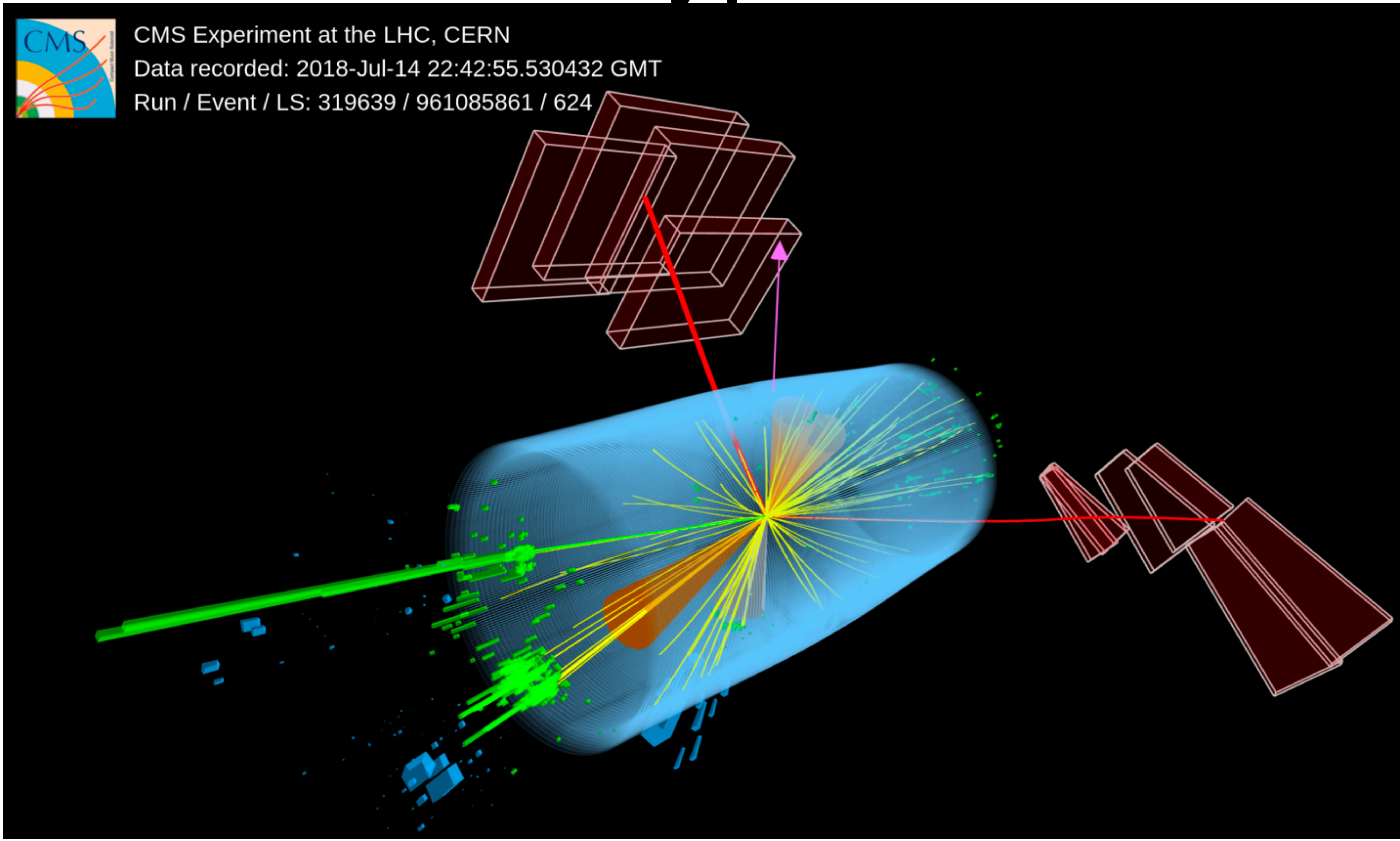


Particle Reconstruction

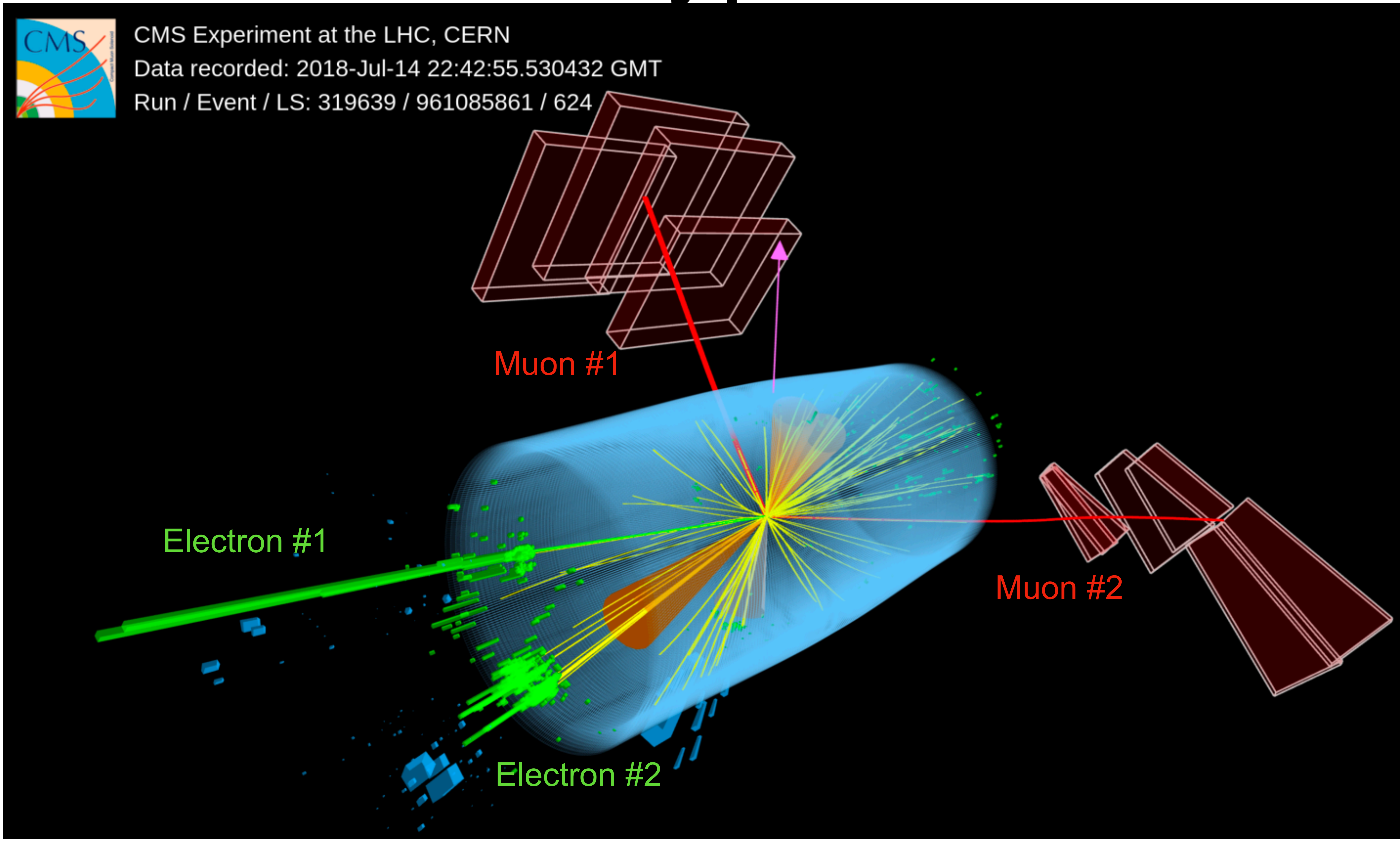


Go from detector deposits to particles

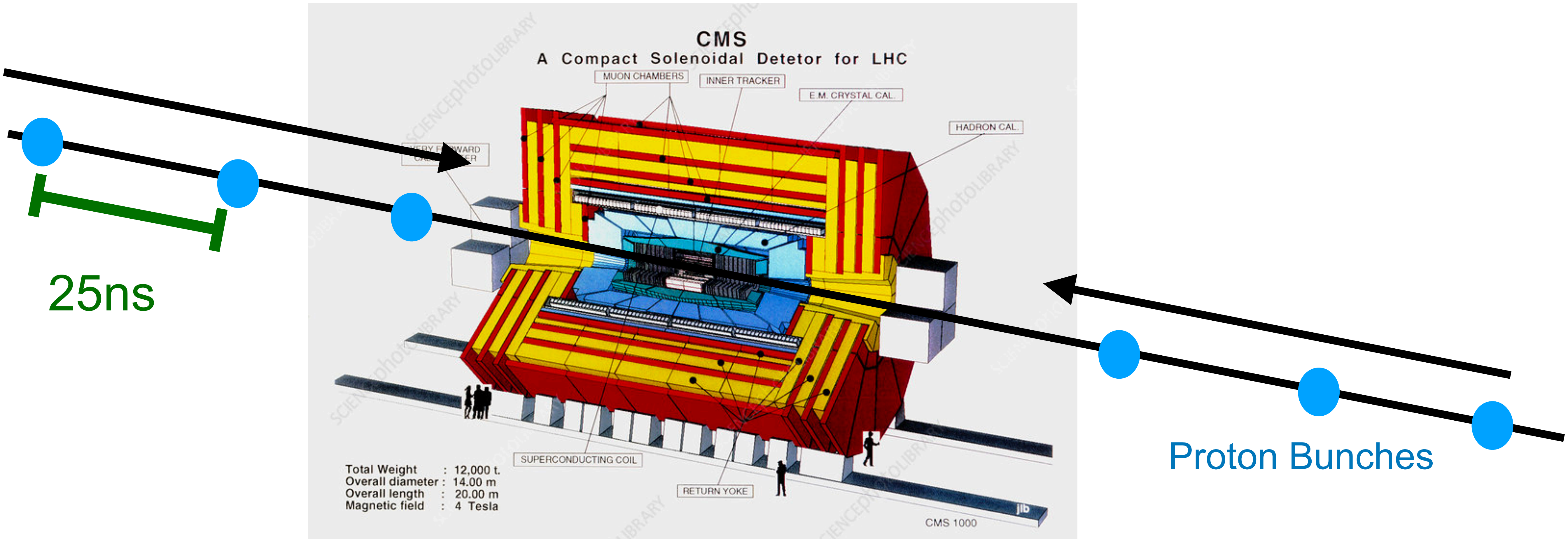
Typical Collision



Typical Collision

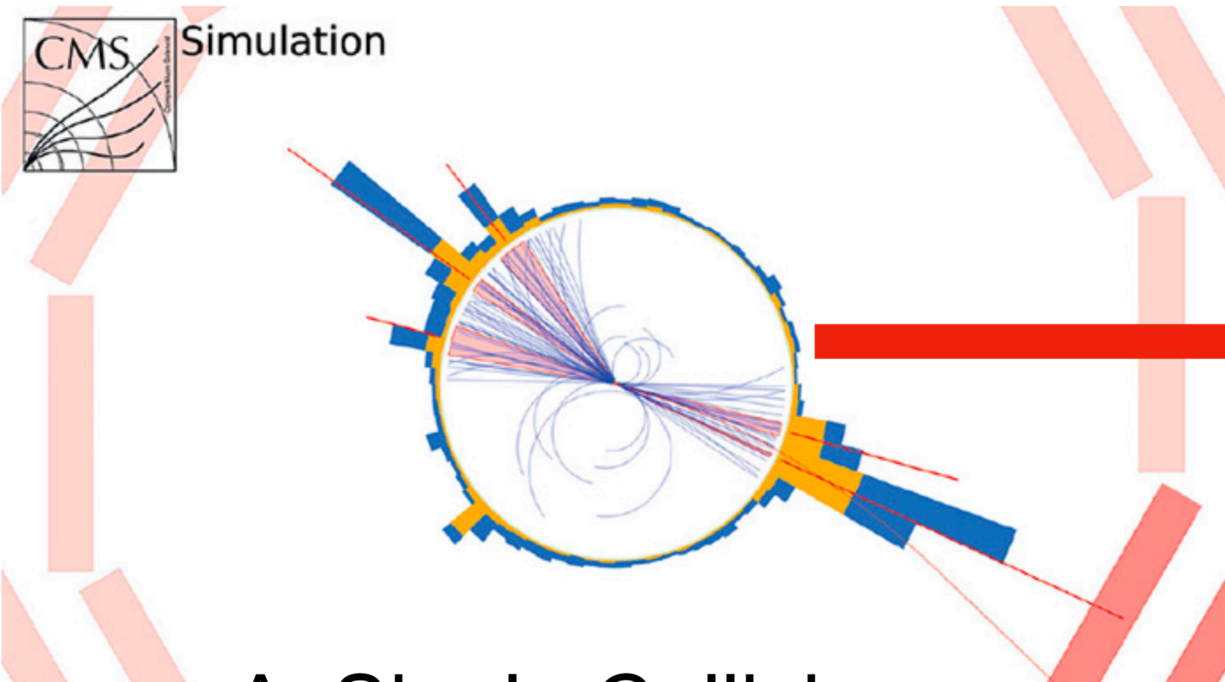


Finding something?

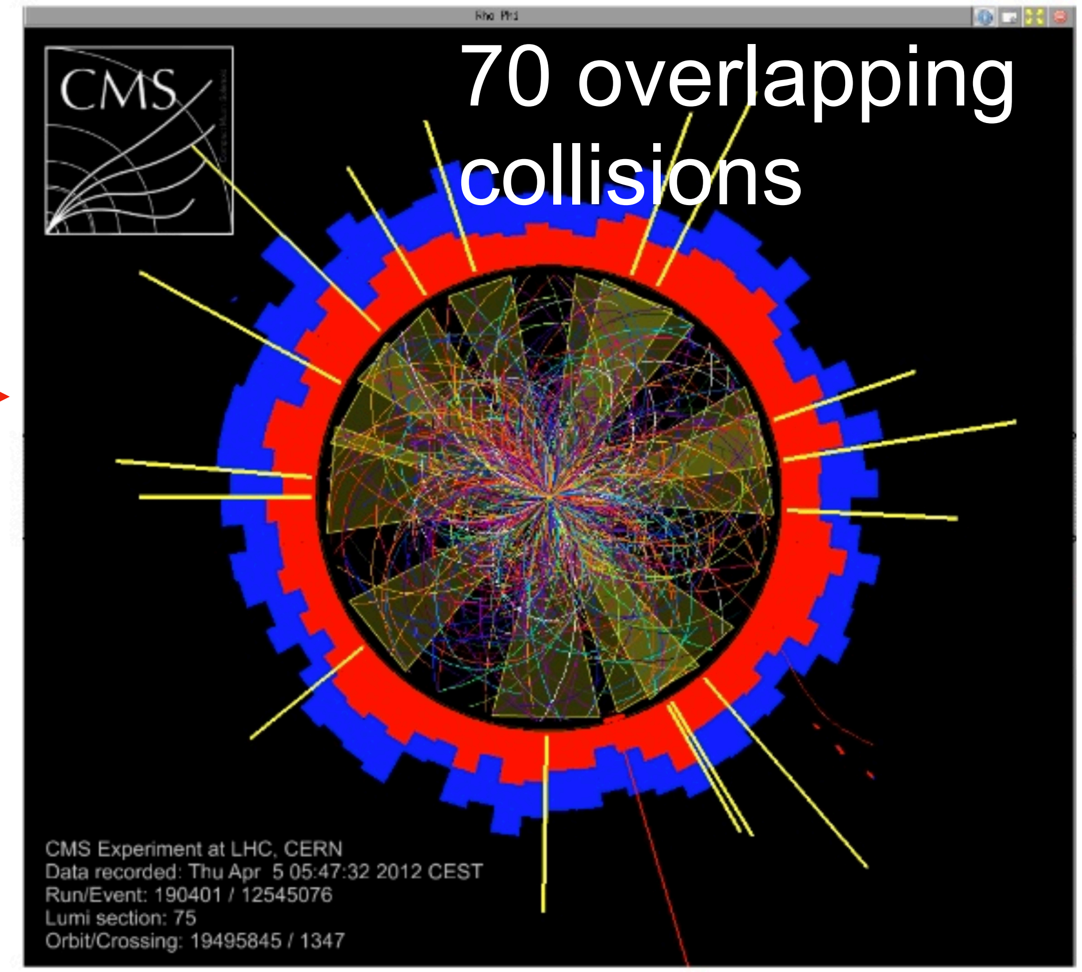


- To find something interesting we collide at a high rate
    - We collide collections of protons at 40 MHz
- This equates to a PIPELINE Initiation Interval of 25ns
- A single event is 8 Mb @ 40 MHz = **320 Tb/s**

Higher Rates



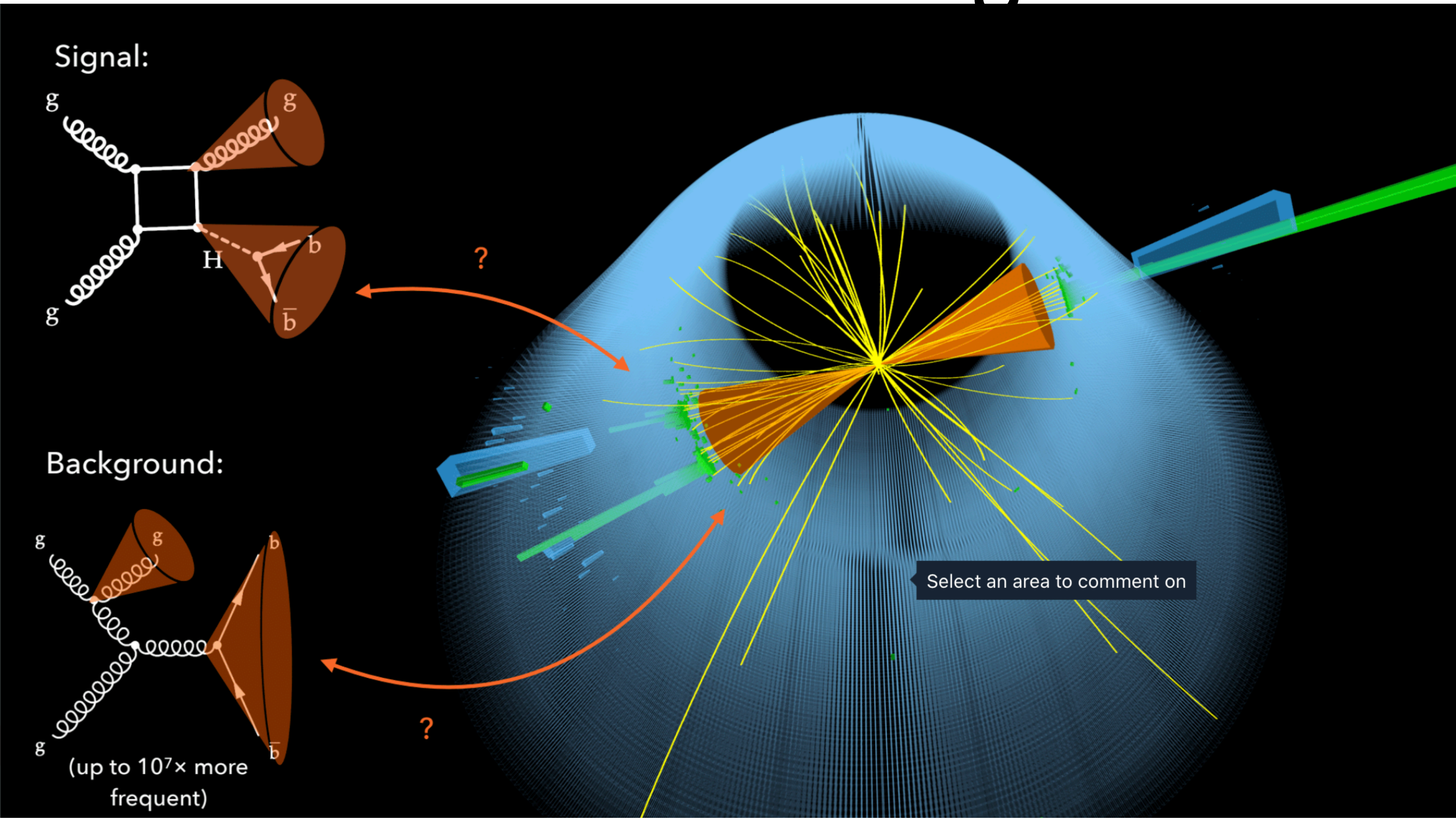
A Single Collision



200 overlapping collisions in future

- In addition to colliding at 40 MHz
  - We don't just collide one proton at a time
  - We (currently) collide about 70 protons at a time (Pileup Collisions)
- We have to pick out one collision on top of many overlapping collisions

What are we looking for now?

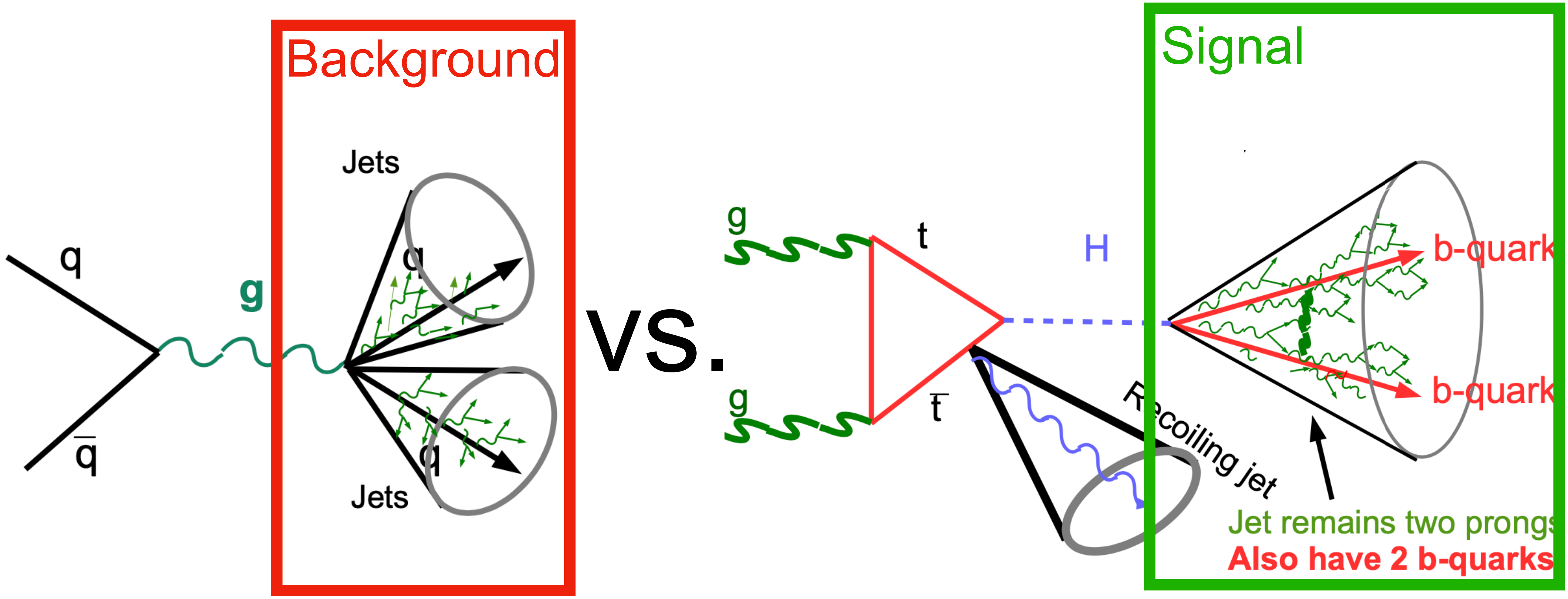


- Higgs boson at very high energies

arxiv:2006.13251

How to find the Higgs?

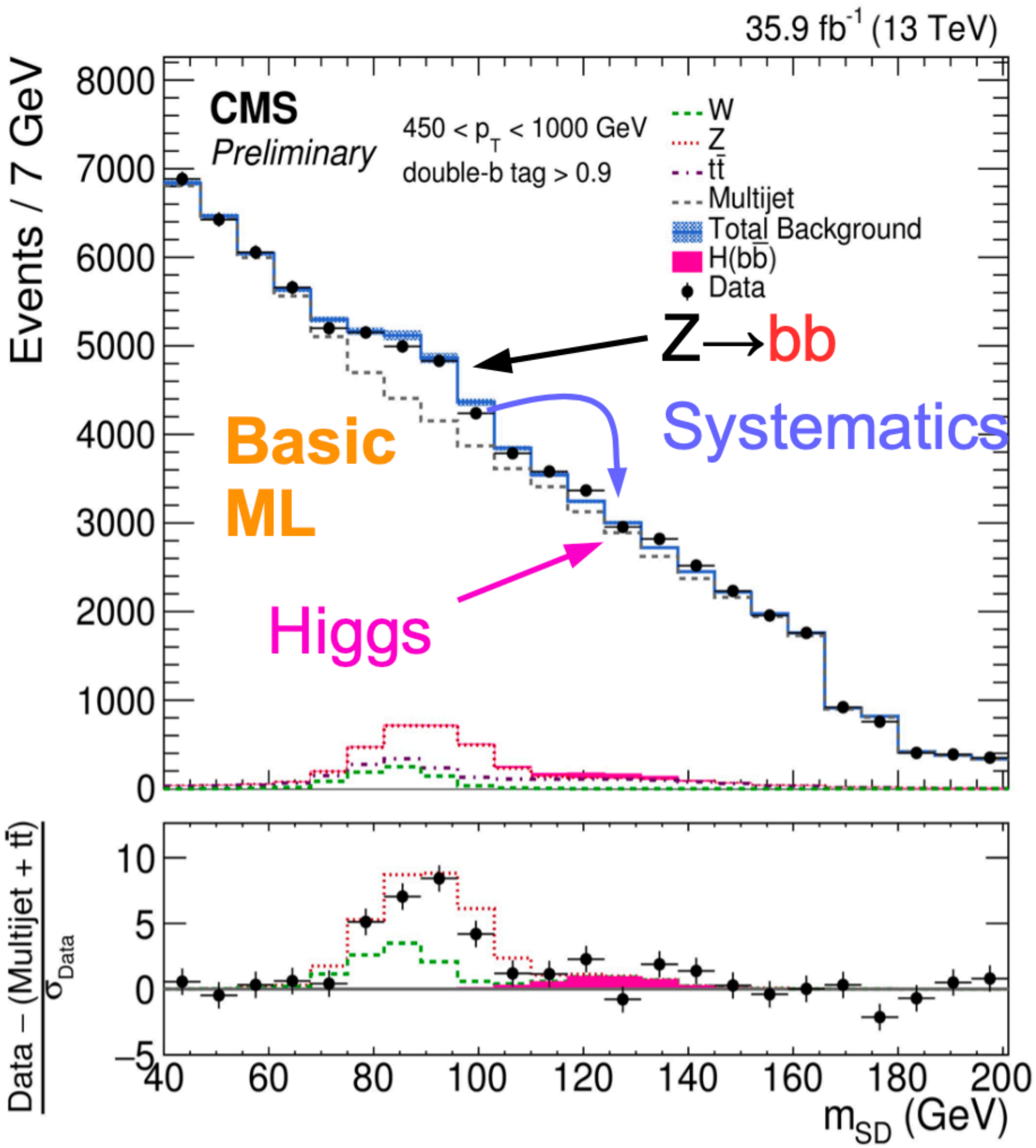
- This topology is very simple
  - But has a huge amount of background
  - Corr is to build a deep learning algorithm to separate



Jet final states consist of many particles (perfect for deep learning)

A first attempt for Higgs

- Sensitive to the Higgs at roughly 1 standard deviation



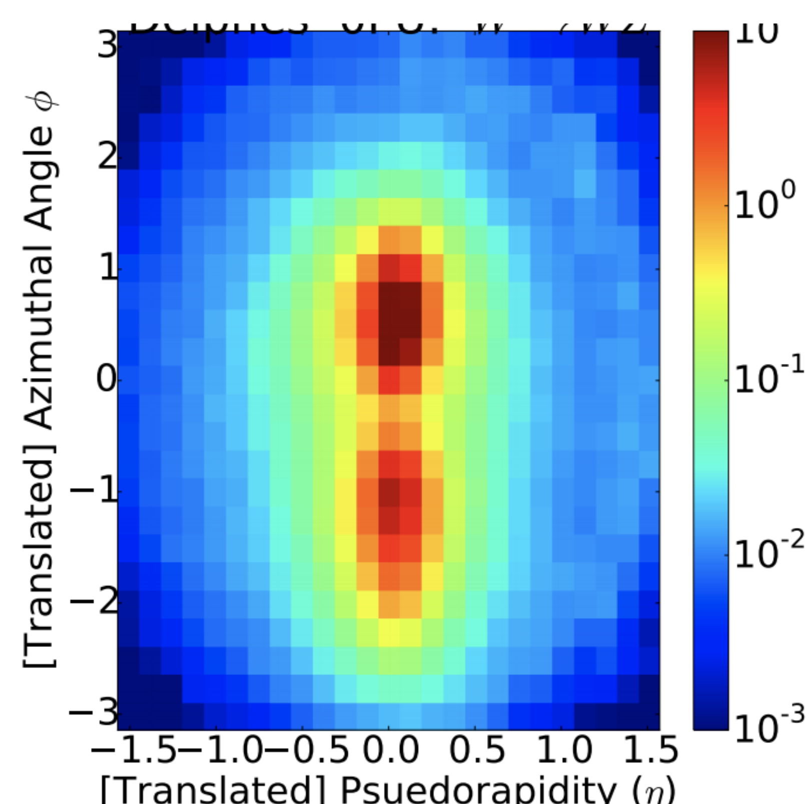
See a clear Z boson peak

Convinced us that this approach could be used to push identification much tighter

Measure tagger eff in-situ

Deep Learning Progression

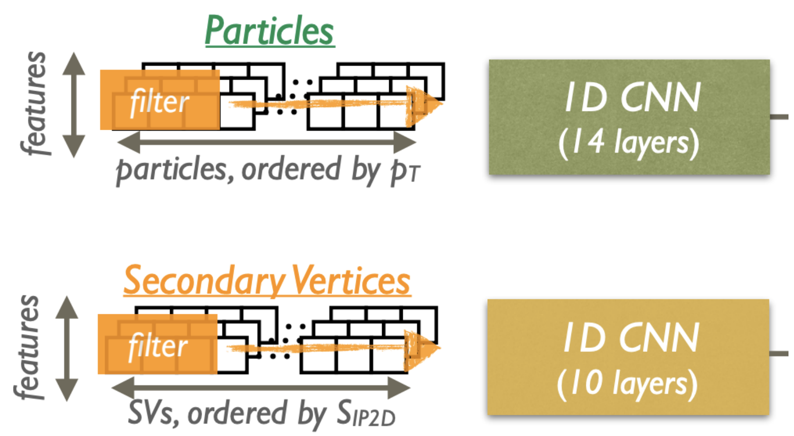
2016



Images

(not lorentz invariant)

2018

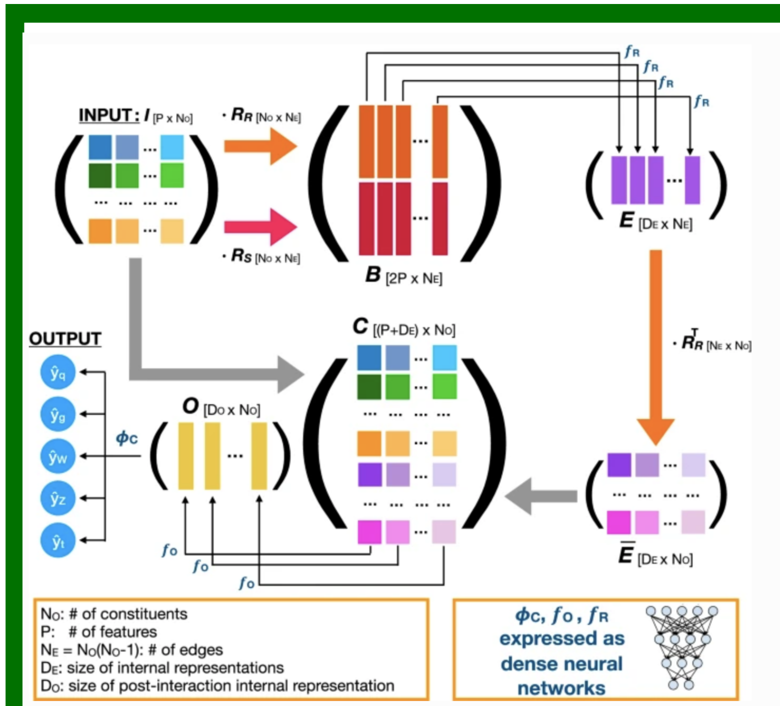


Particles and SVs
with 4-vectors+features

Particles

(limited correlations)

2020



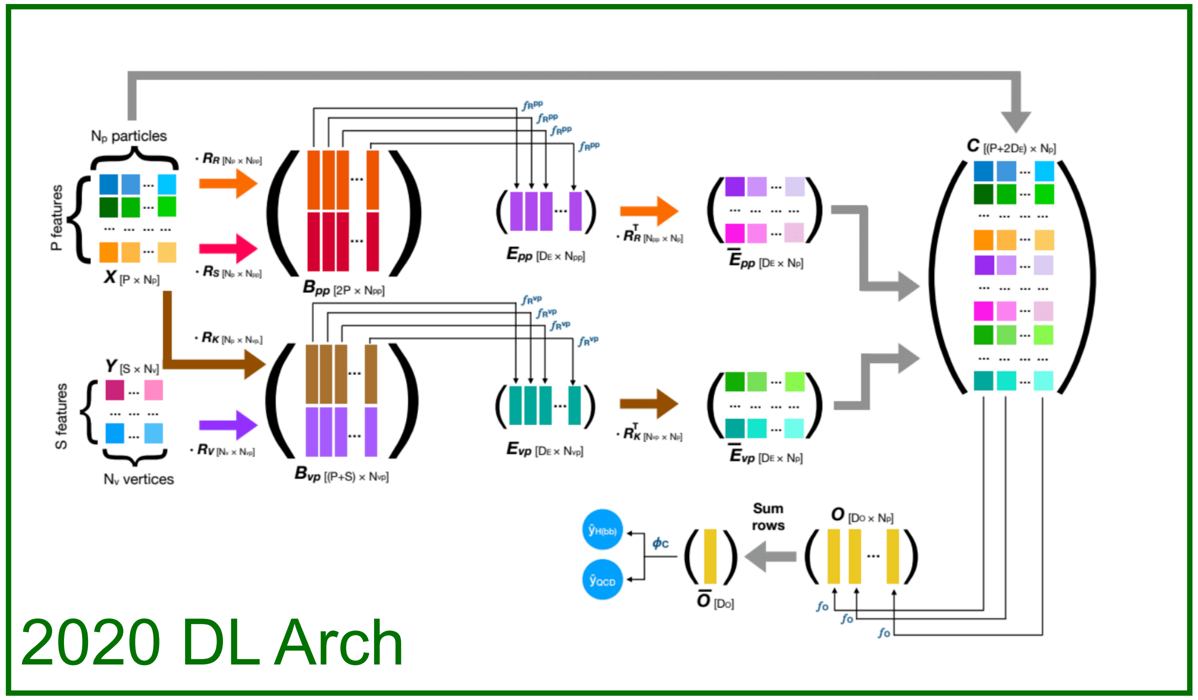
Graphs

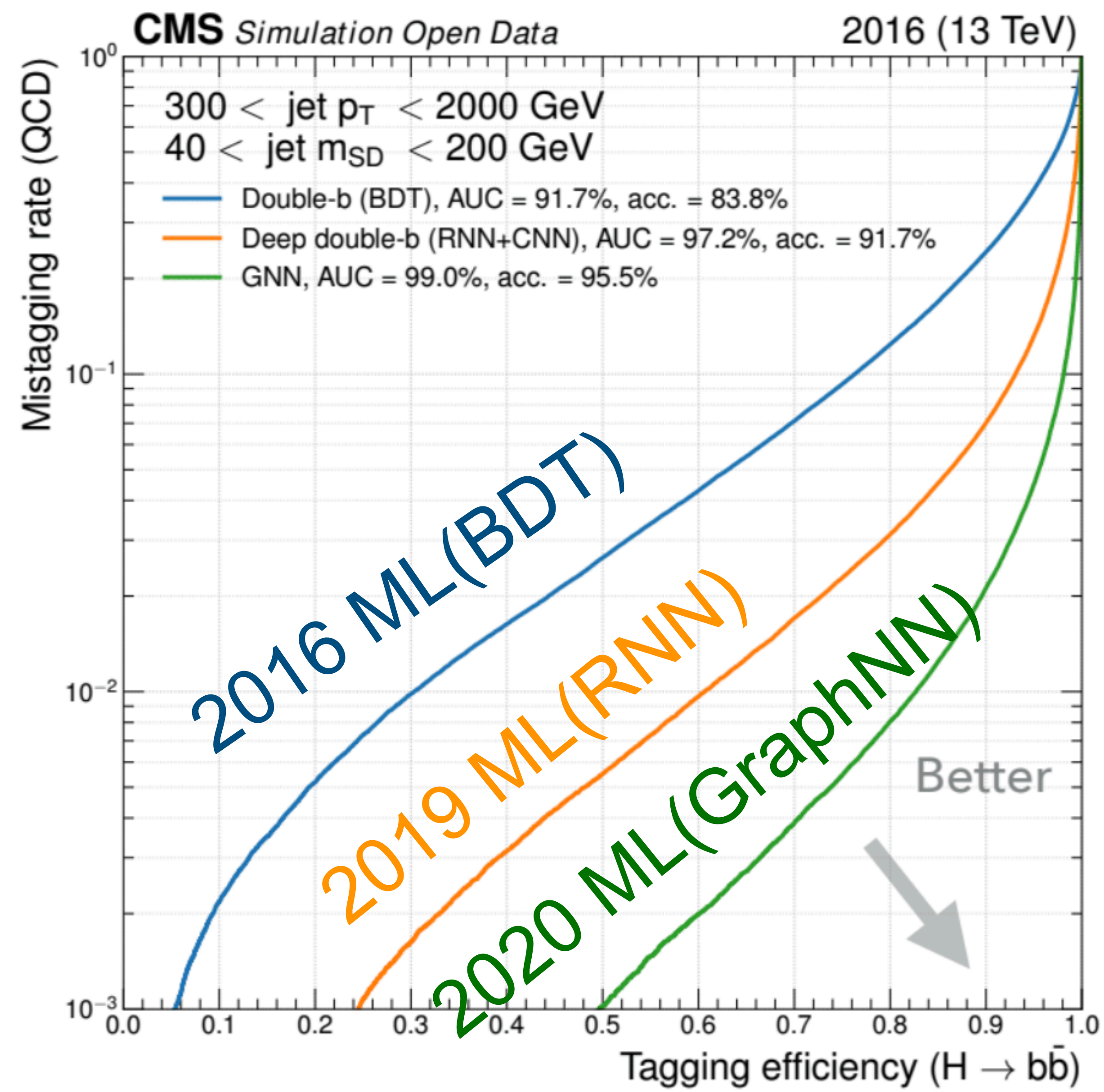
(Particles+correlations)

Current collaboration results

Progressively moving towards use of more info

Difficulty of finding Higgs



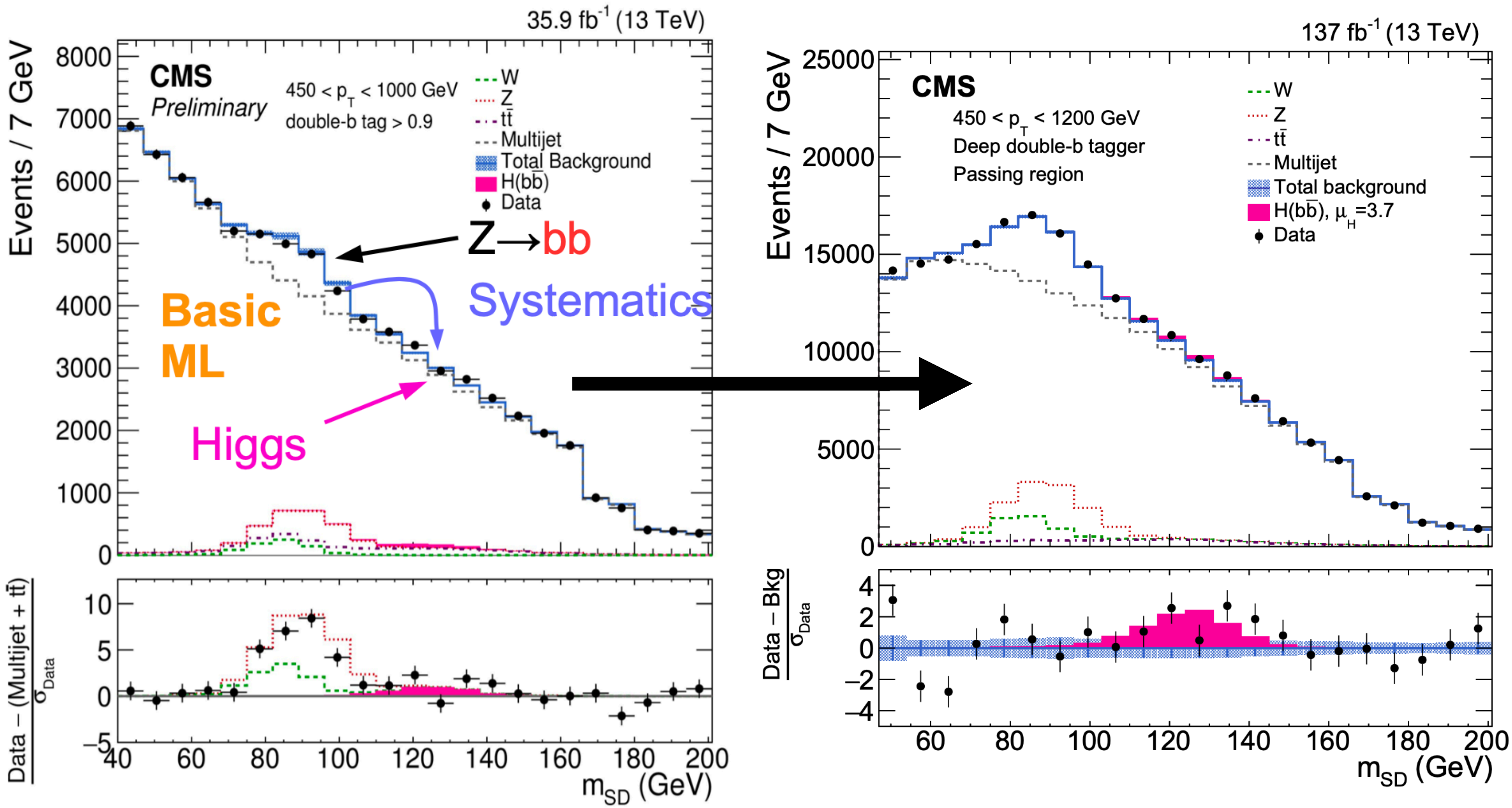


- For a Higgs boson at high energy
  - We have to rely on deep learning
- Deep learning is quickly leading to a major transformation
  - We can measure processes that we didn't think possible

arxiv:1909.12285

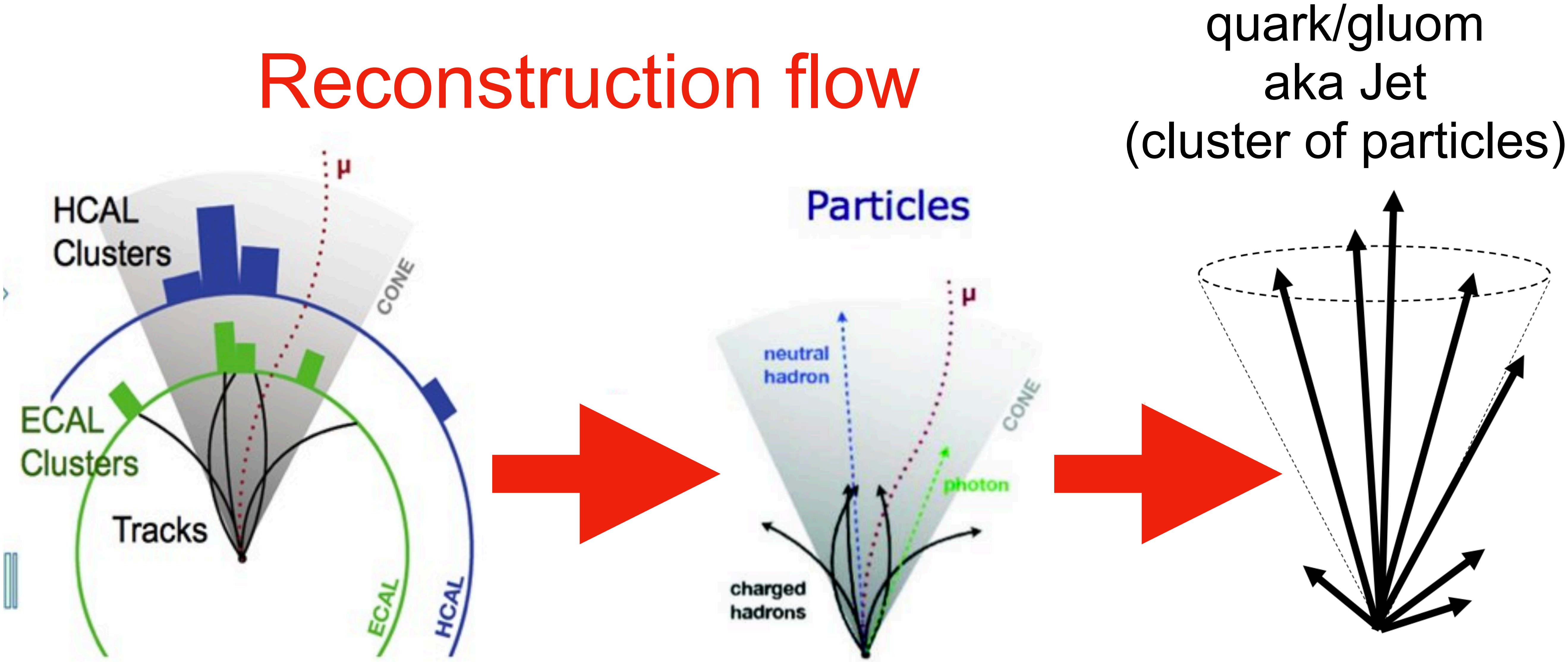
Higgs Boson Progression

- Z boson sensitivity is dramatically improved (thnkas DNN)



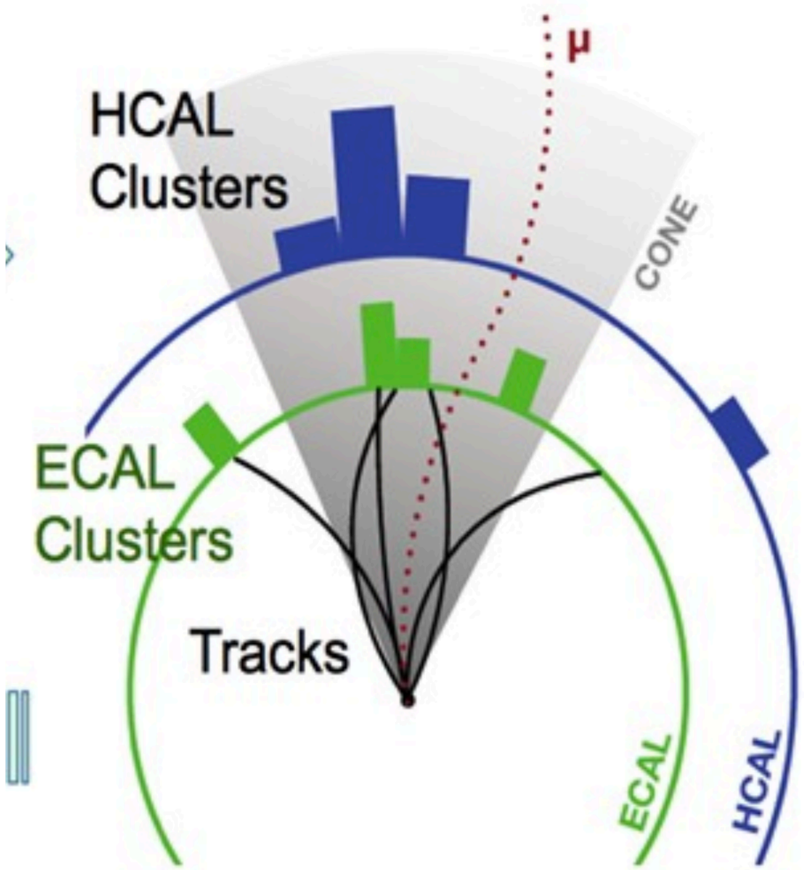
Deep Learning Evolution

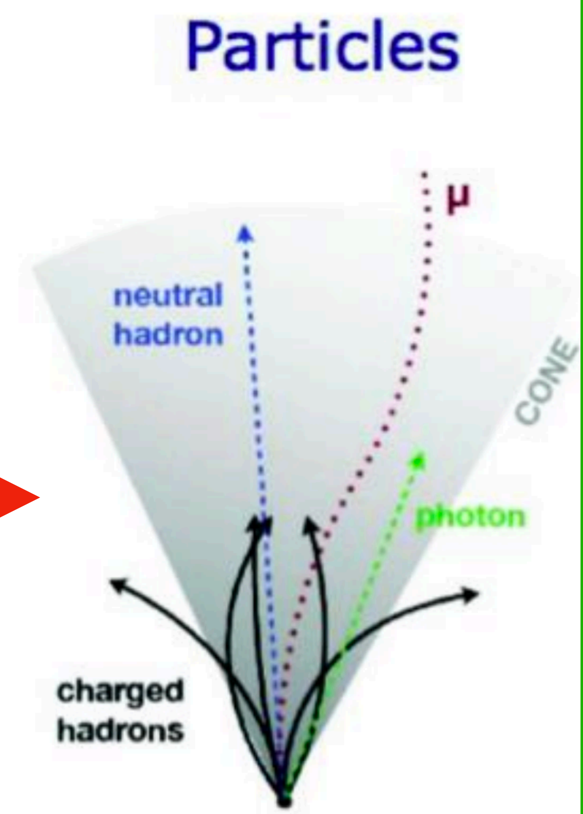
Reconstruction flow



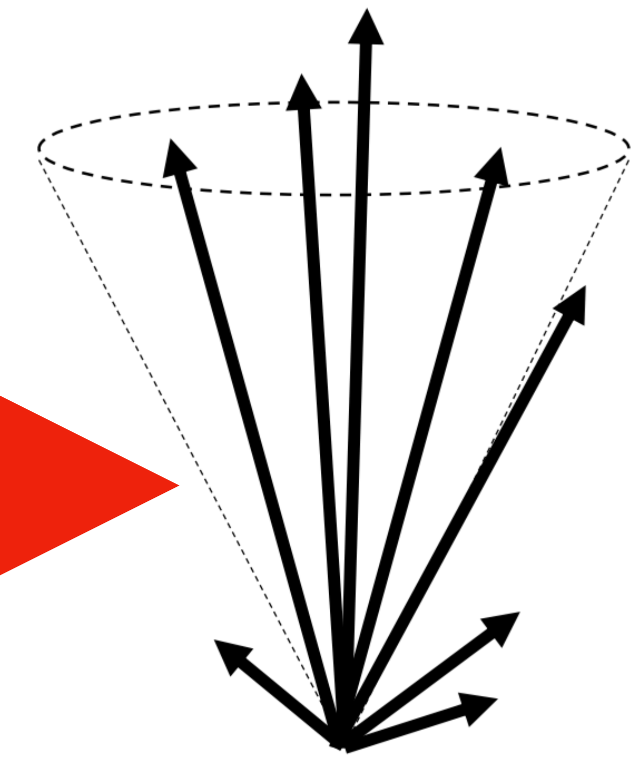
Deep Learning Evolution

Reconstruction flow



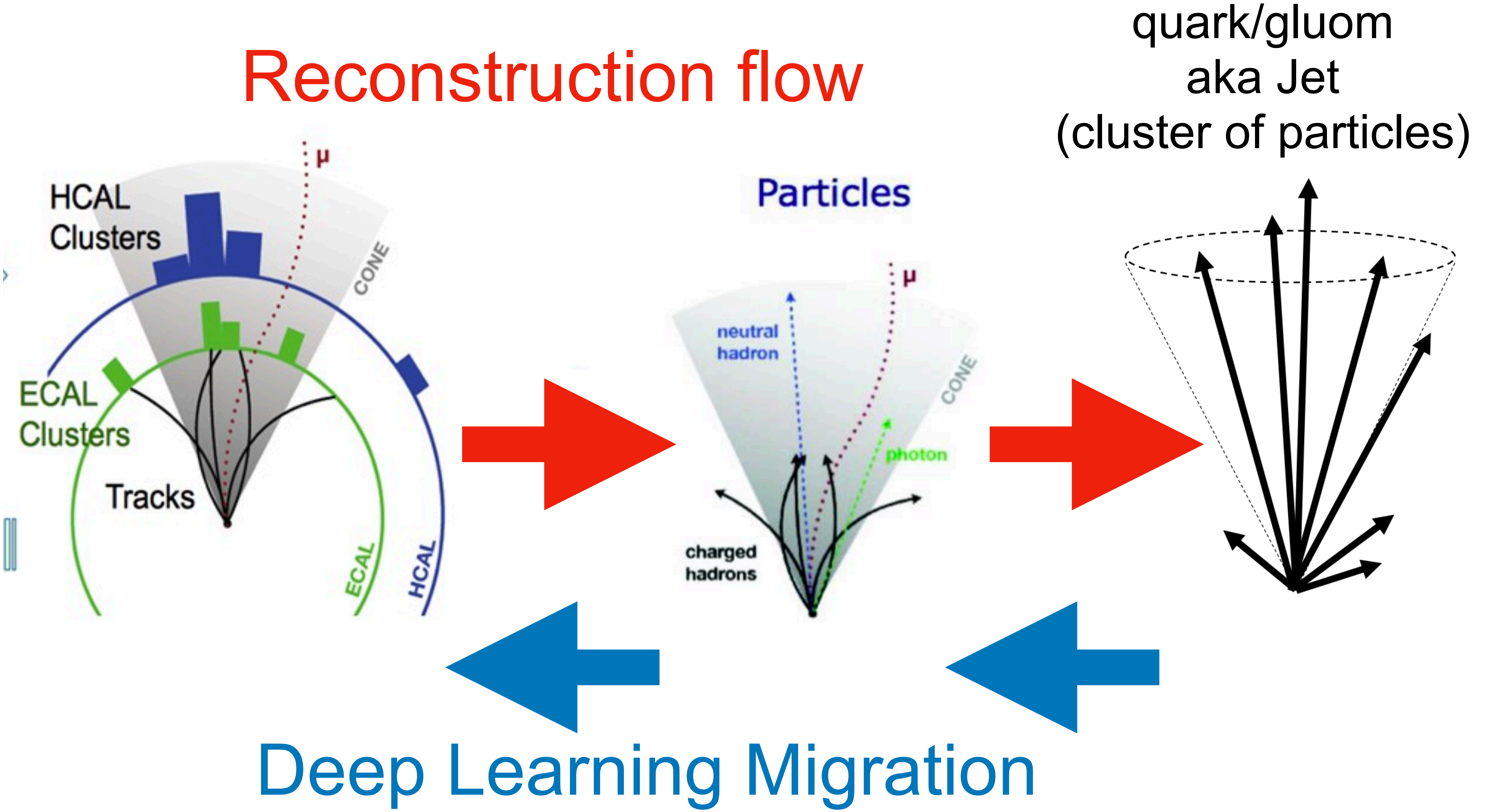


quark/gluom
aka Jet
(cluster of particles)



Where we do analysis

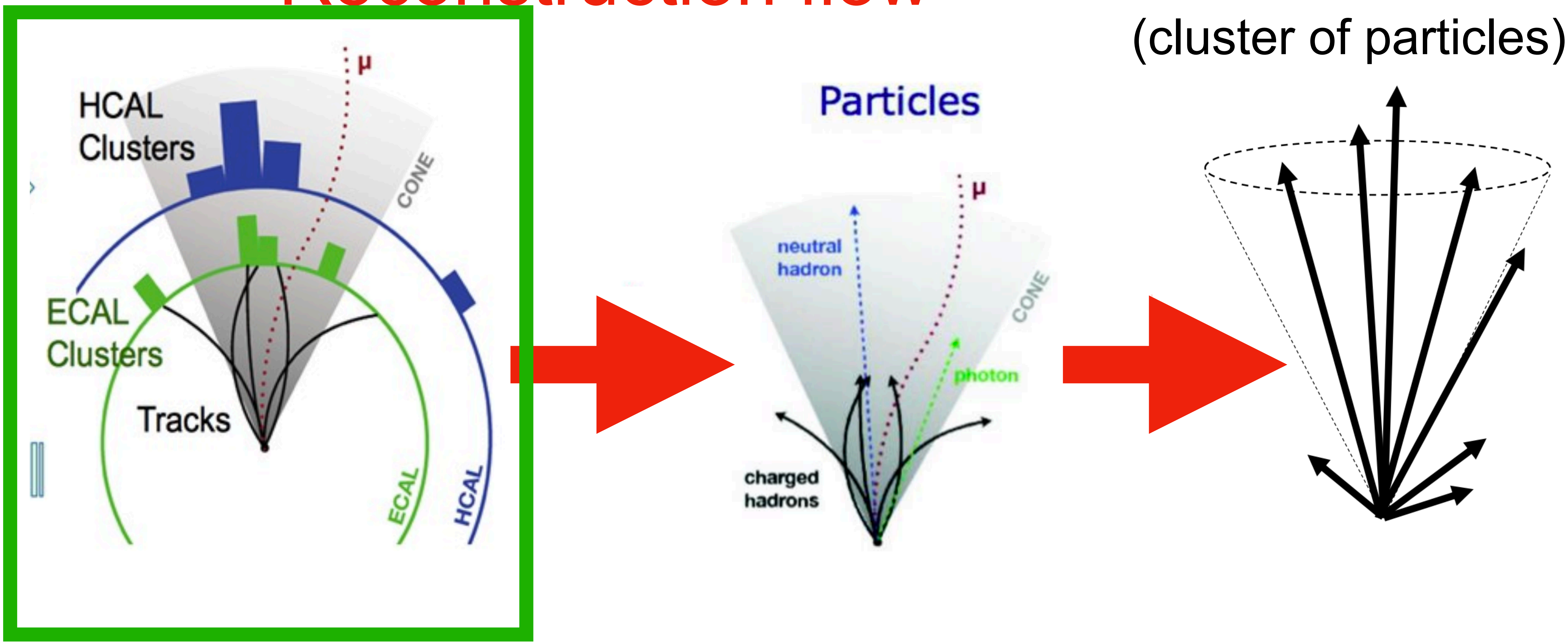
Deep Learning Evolution



Deep Learning Evolution

quark/gluom
aka Jet
(cluster of particles)

Reconstruction flow

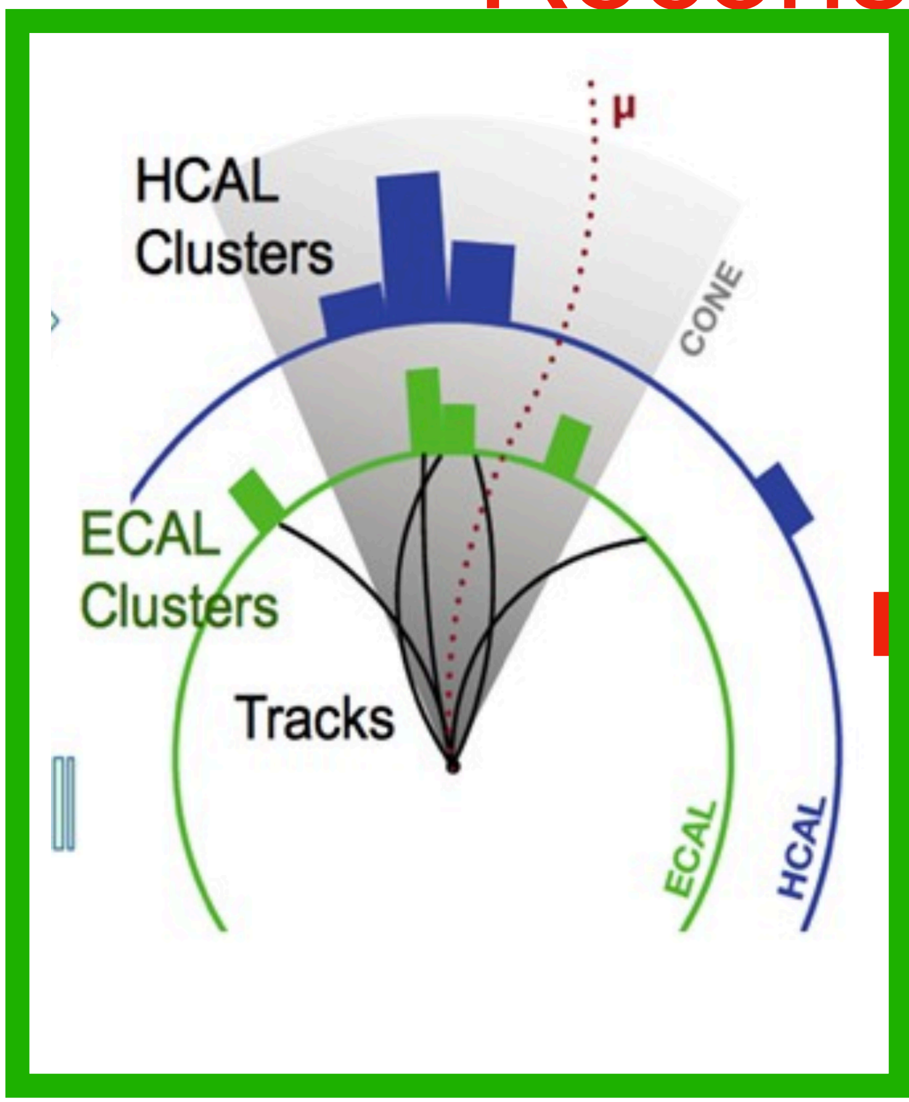


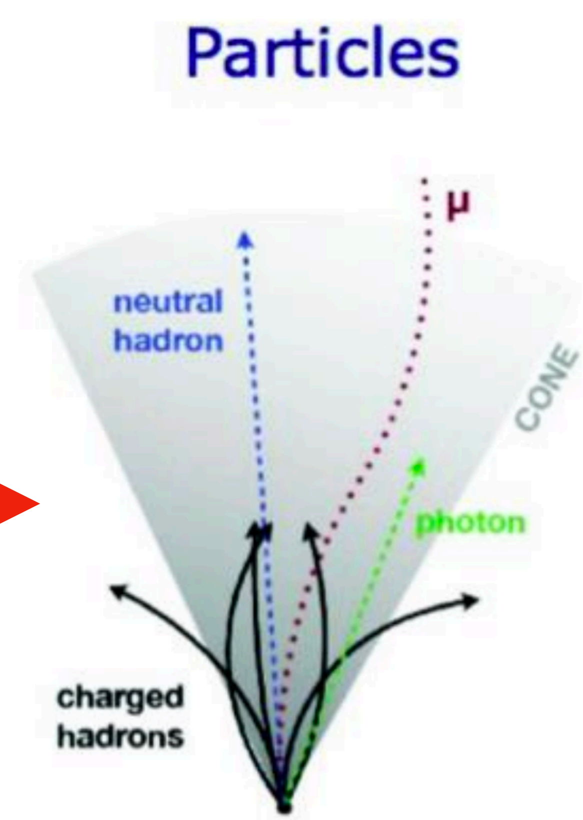
Challenge:
Can you go from Raw inputs to reco?

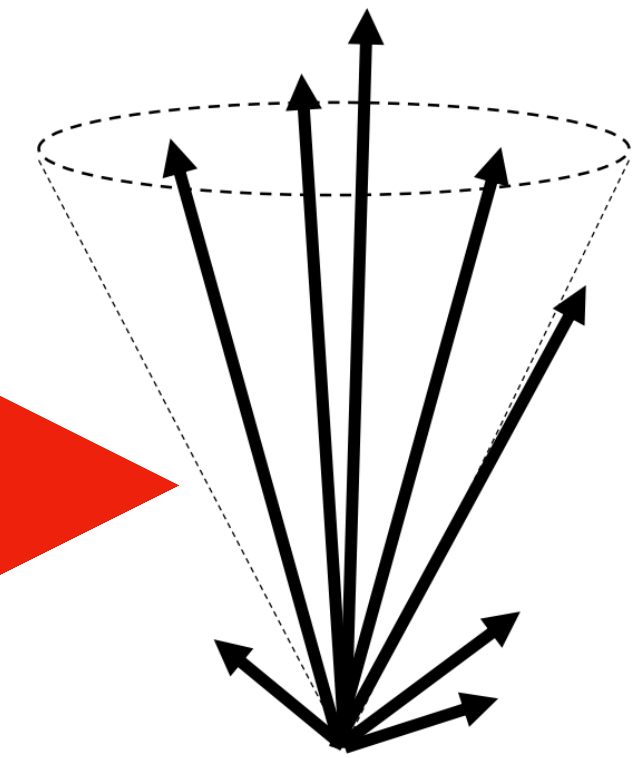
Deep Learning Evolution

Reconstruction flow

quark/gluom
aka Jet
(cluster of particles)



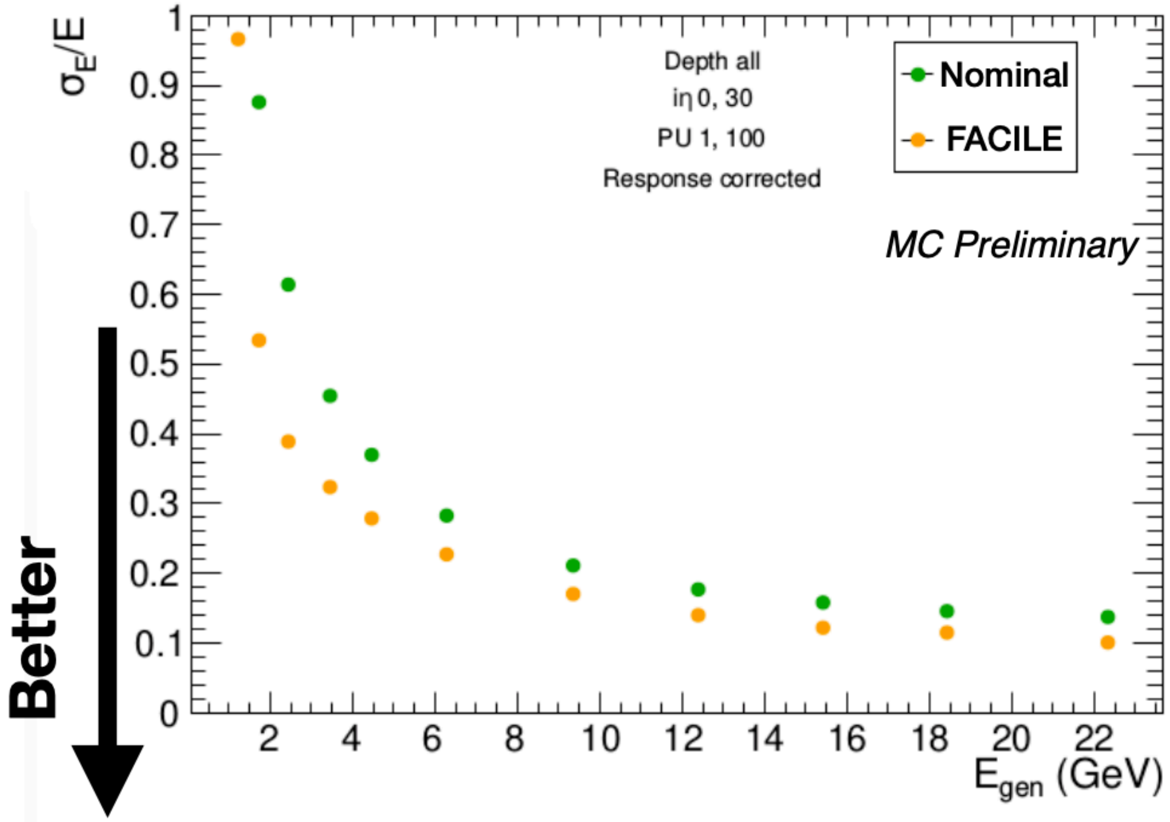


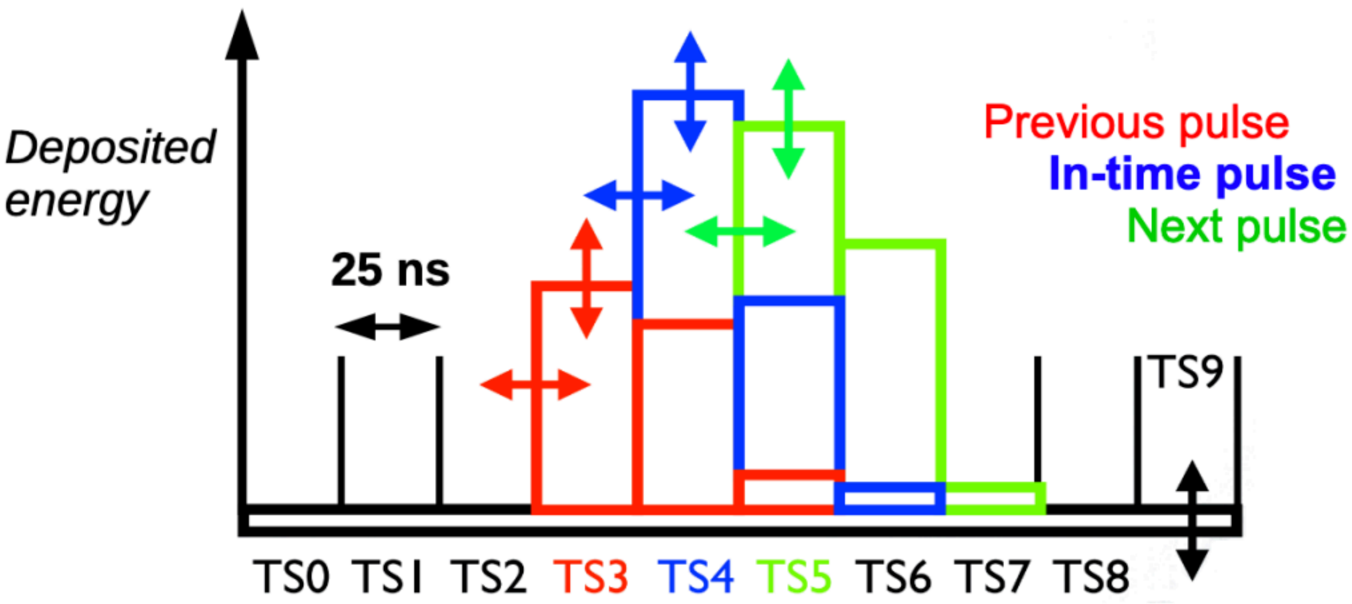


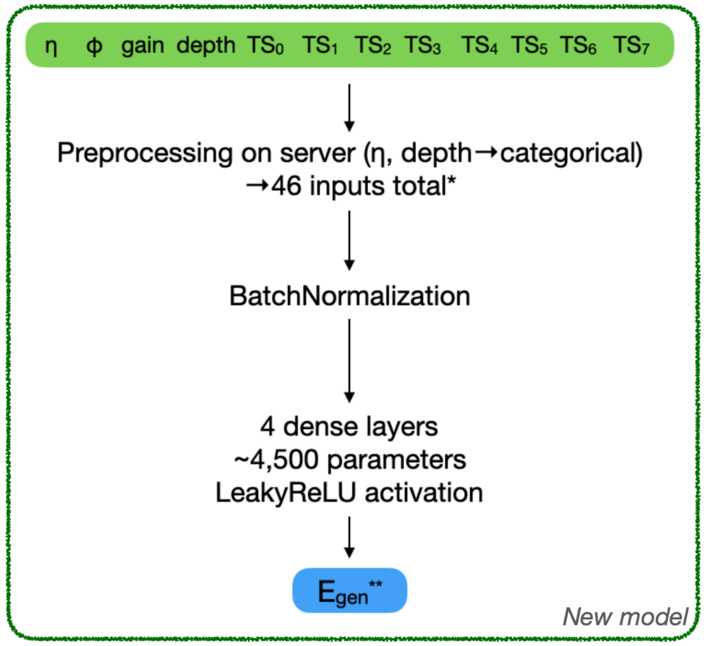
Challenge:
Can you go from Raw inputs to reco?

Simple Example

- Reconstructing a single calorimeter tower
  - FACILE Algorithm: Reconstruct integral of in-time pulse
  - Up to 5 overlapping pulse





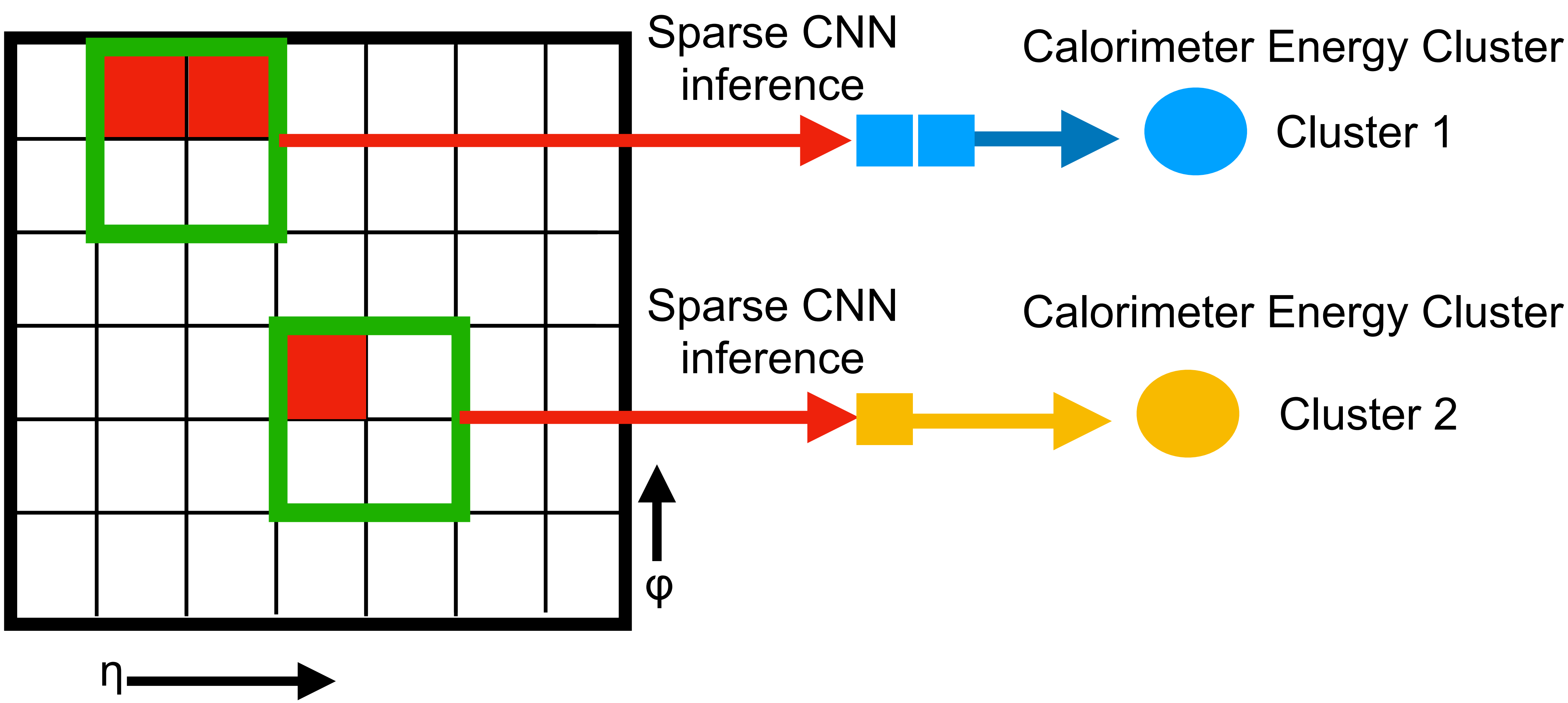


Simple NN
Can run fast

LeakyRelu
Critical to regression

From Single to Collection

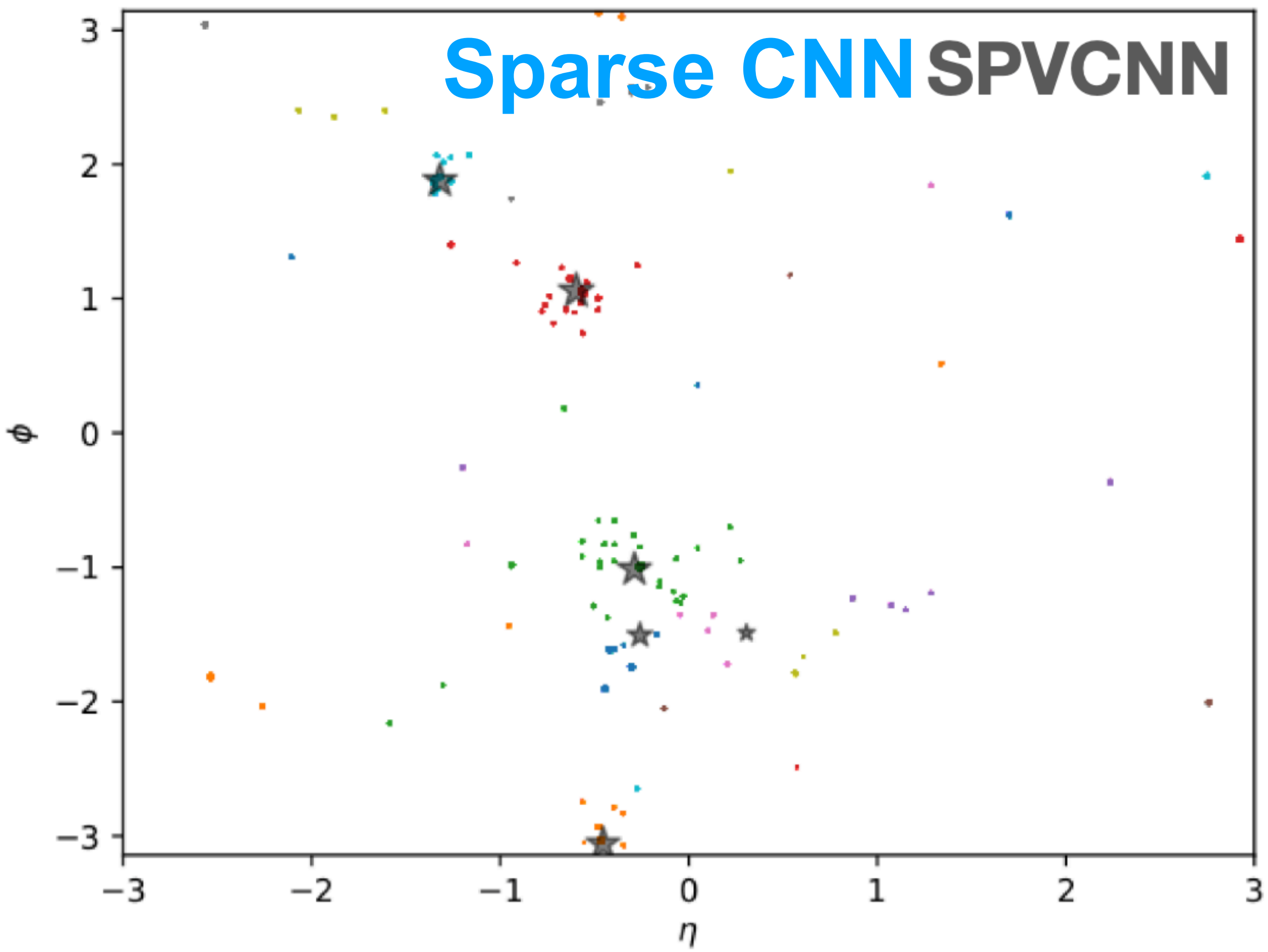
- Facile runs reconstruction on a single channel
  - We can envision an algorithm that takes in all channels
  - One way is to use a sparse CNN for graph-like inputs

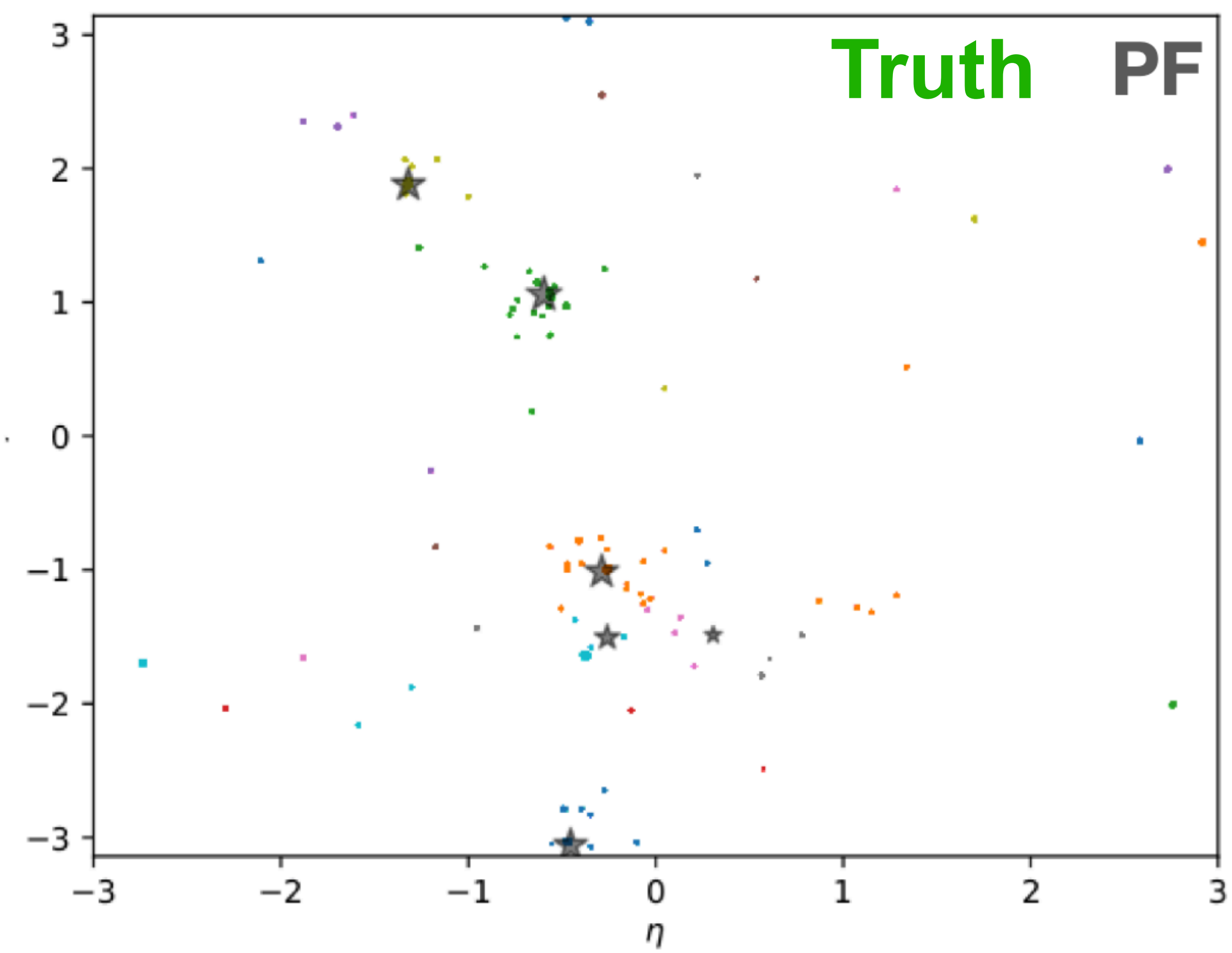


By taking the grid geometry of calorimeter can deploy Sparse CNN to Infer whole calo at once

Can compare to Reality

- A single algorithm is doing all of the clustering

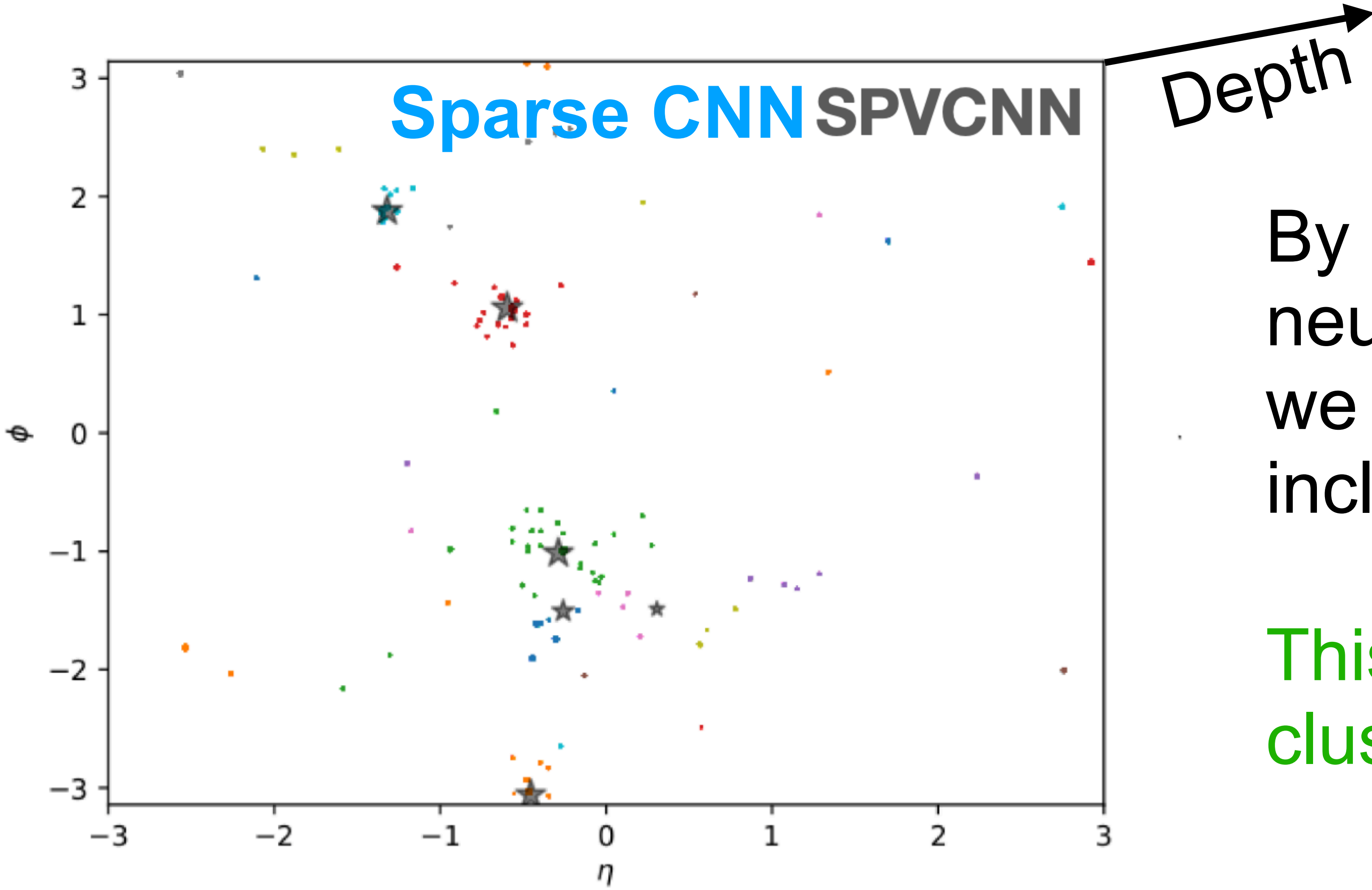




- Clustering algorithm produces very similar results to truth
  - Single algorithm that takes in whole detector at once

So what do we gain?

- A single algorithm is doing all of the clustering

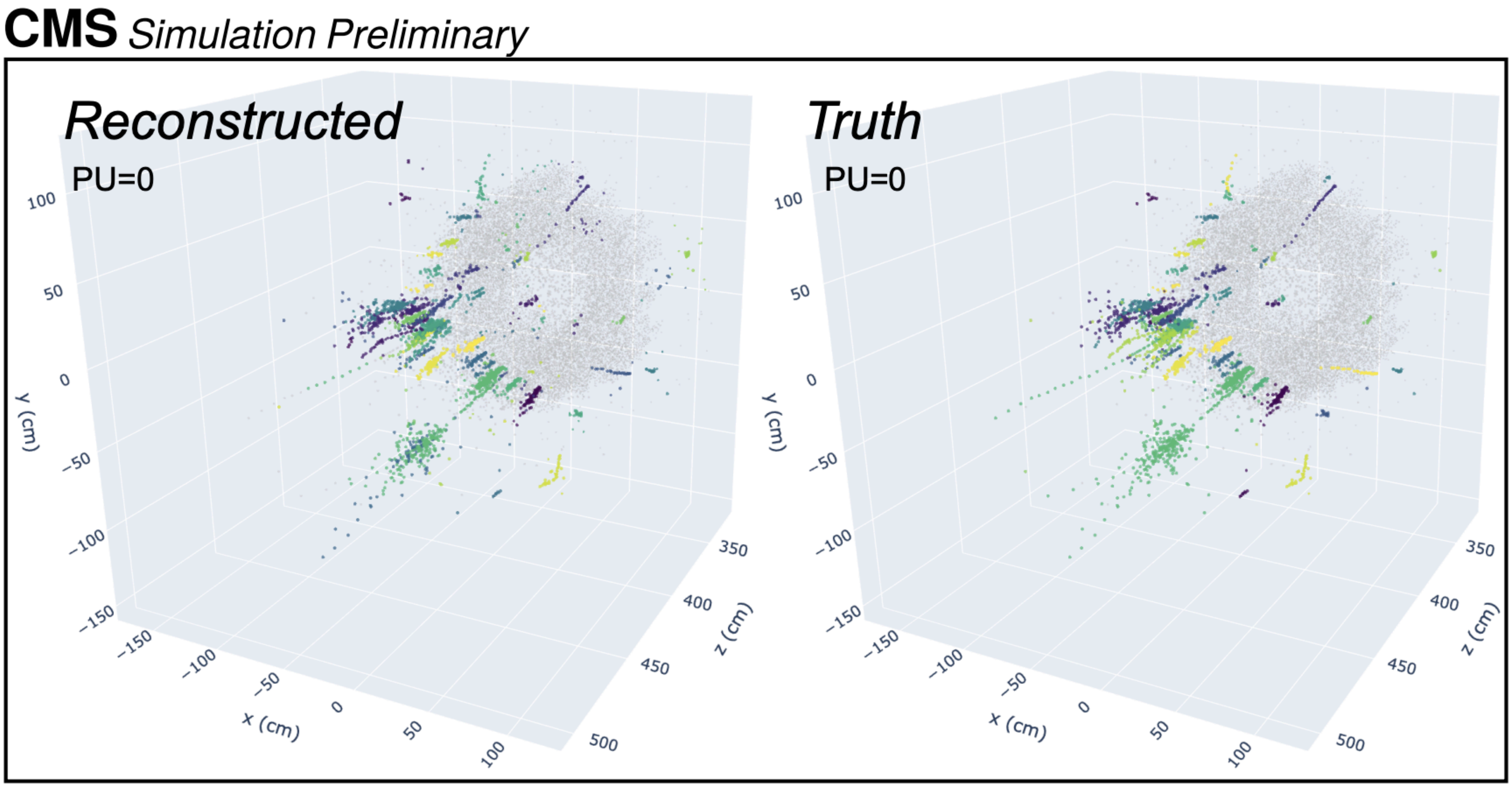


By embedding this in a neural network
we can extend algo to include more info

This is 1st algorithm to cluster with depth info

- Moreover this algorithm can now look at whole event to perform clustering
  - Awarness of the event can allow for dynamic thresholds/interpretations
- Finally, this algorithm is highly parallelize → Can Run it Fast!

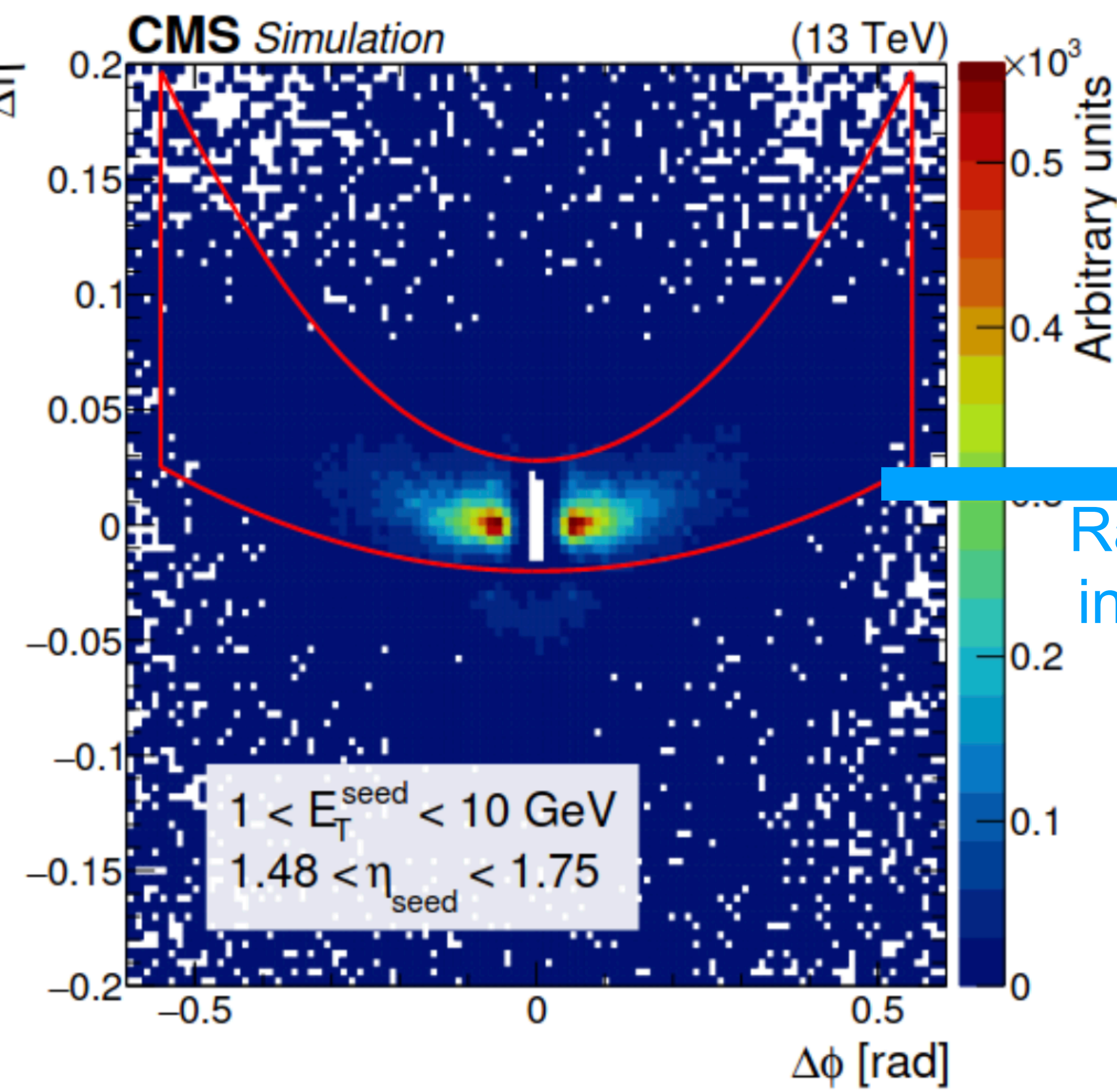
A more Extreme Example



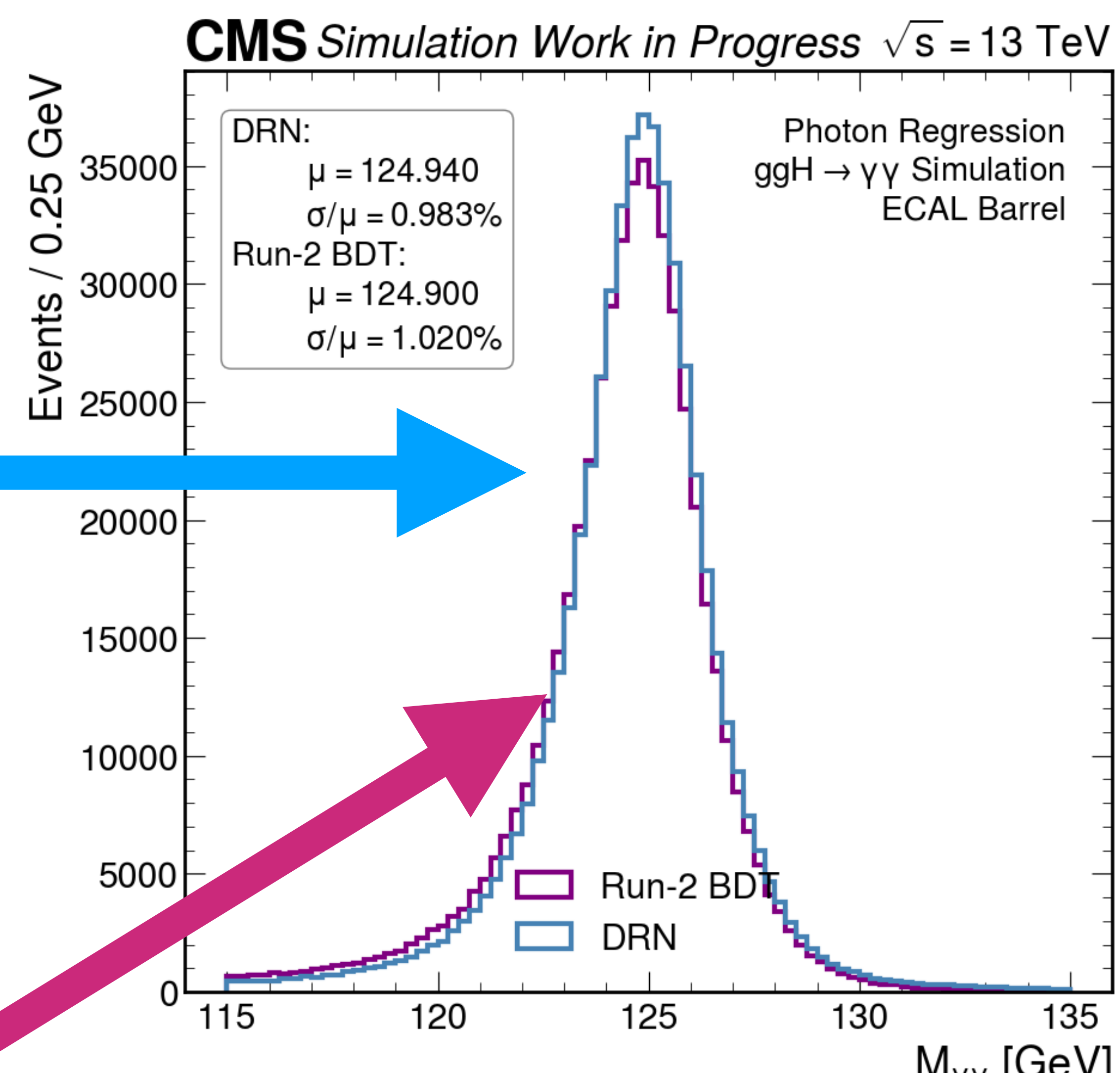
- Algorithm effective a reconstructing new complex topologies

Another Example

- Electron and Photon energy regression with an NN
  - Raw inputs to make an NN gives significant improvements



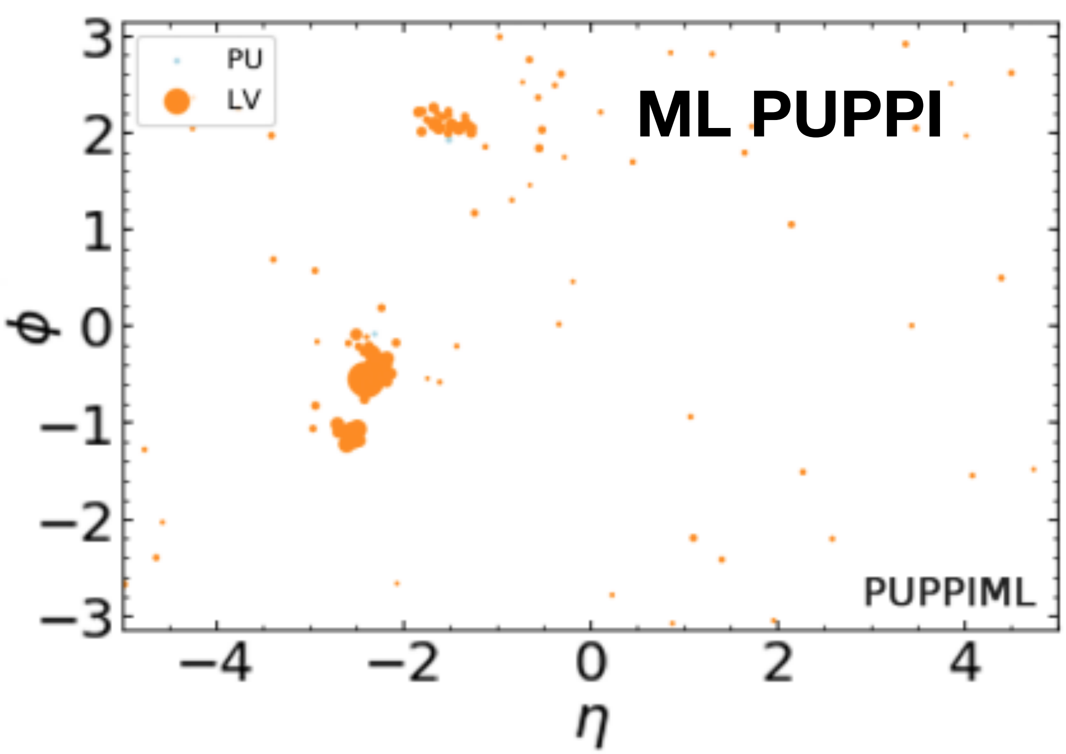
Raw Inputs into an NN



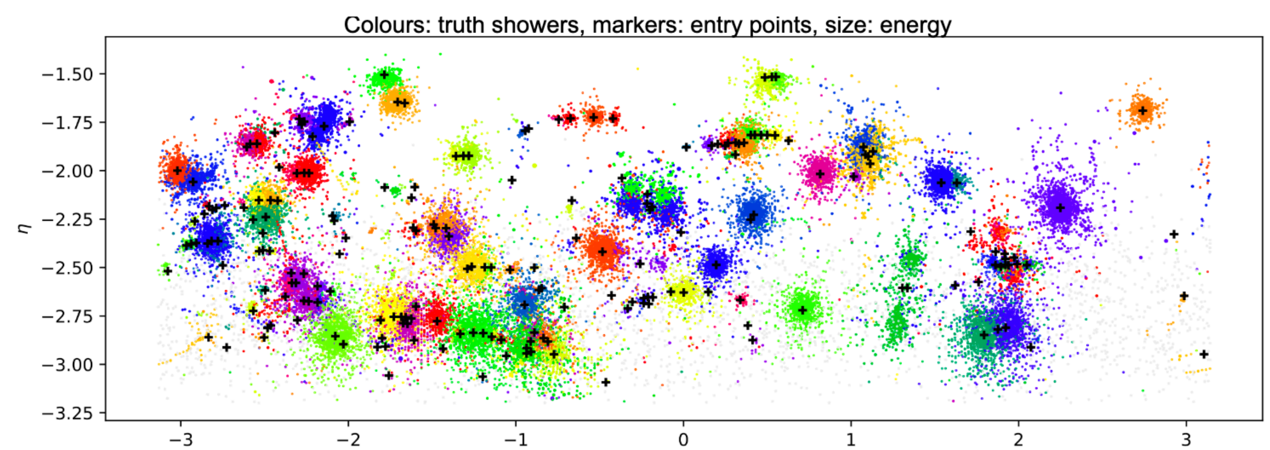
Previous Version used pre-reconstructed variables based on raw inputs

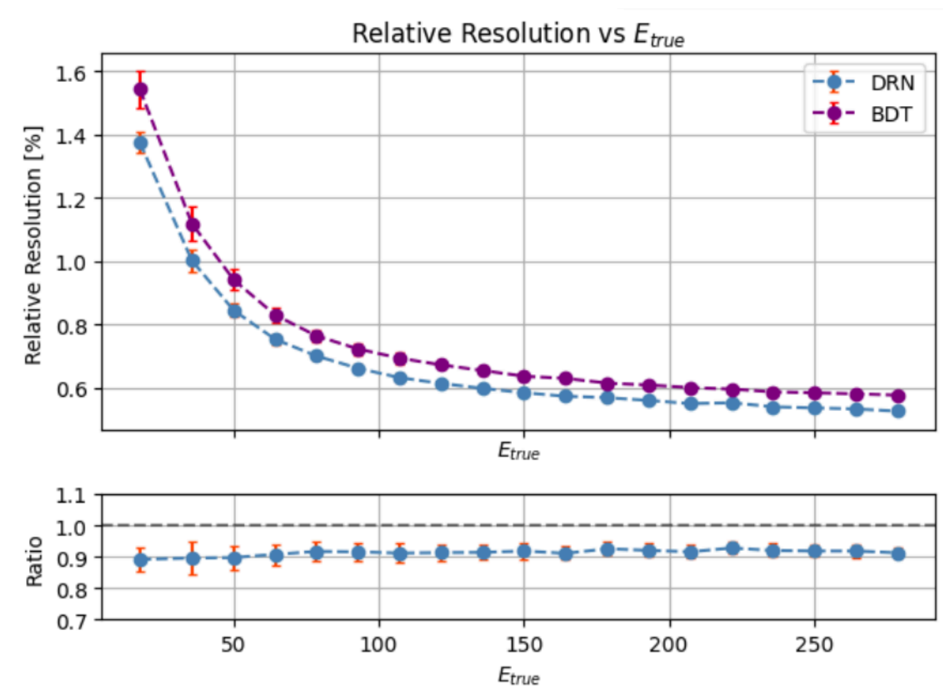
eg. $\langle \Delta\phi^2 \rangle_{crystals}, \langle \Delta\eta^2 \rangle_{crystals}$

Success of Deep Learning



Clustering: Graph NNs for HGCAL



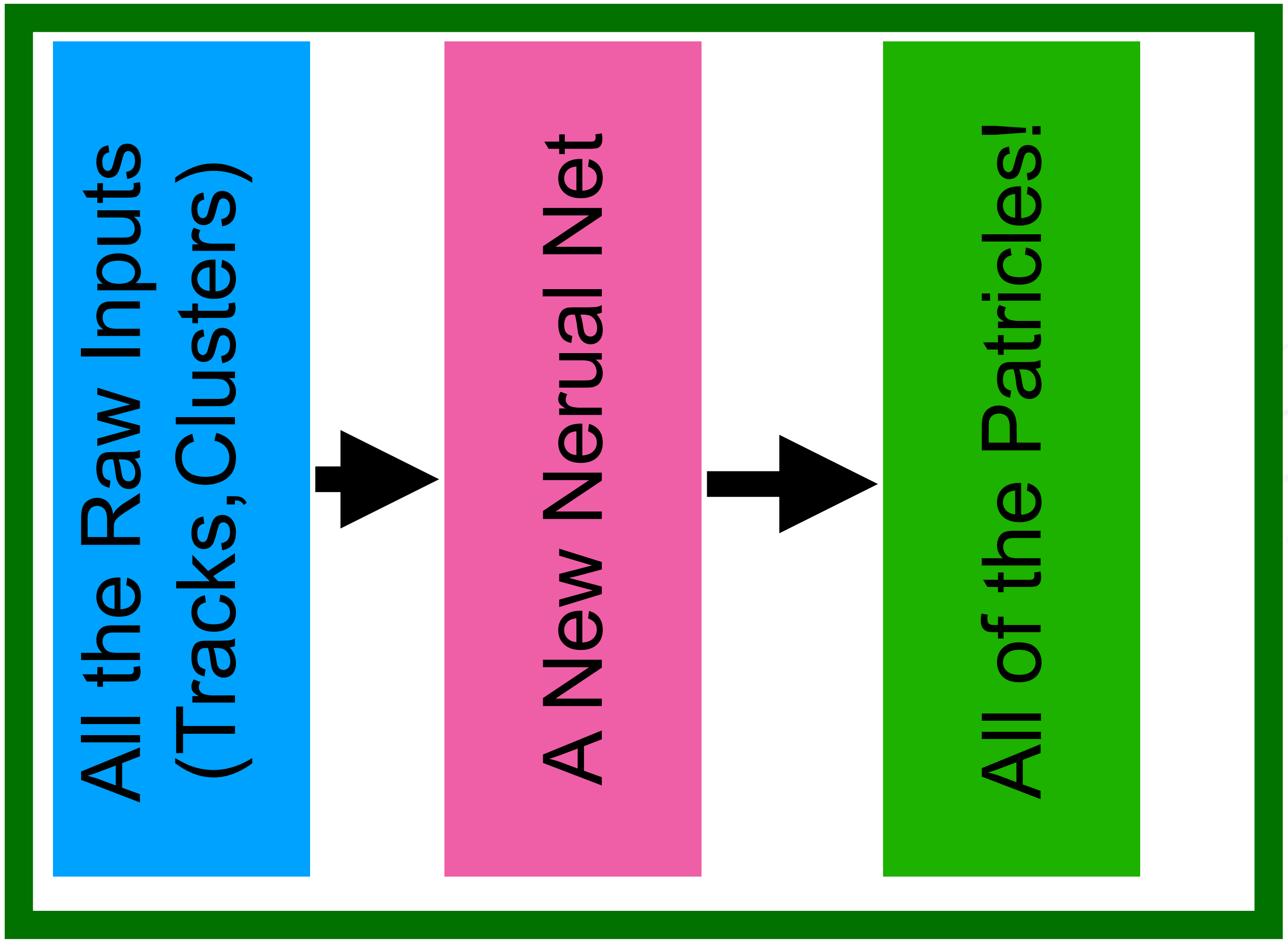


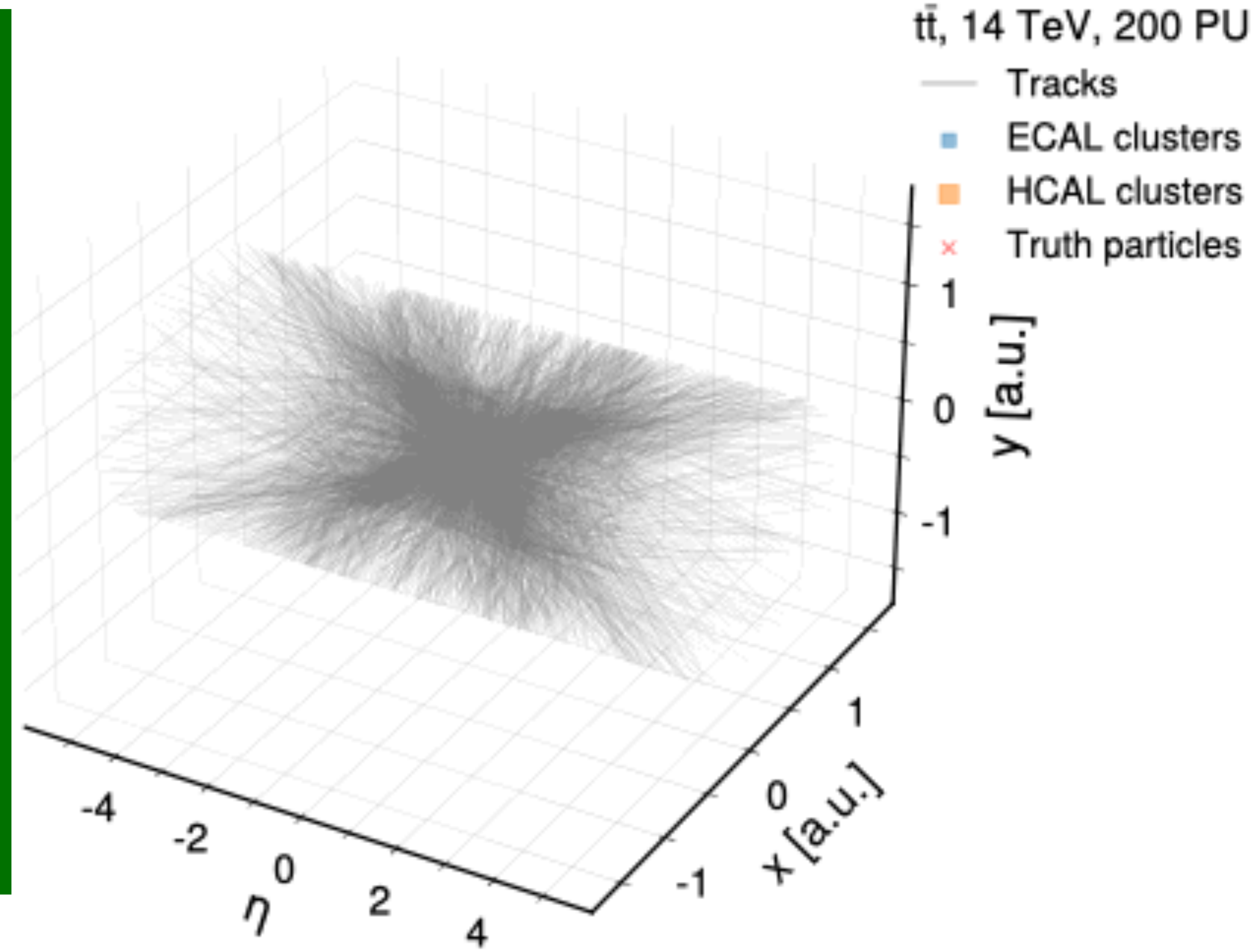
Dynamic reduction network for EGamma regression

S. Rothman

- Networks are emerging to do calorimeter clustering
- Additionally networks are emerging to identify all objects

Success of Deep Learning



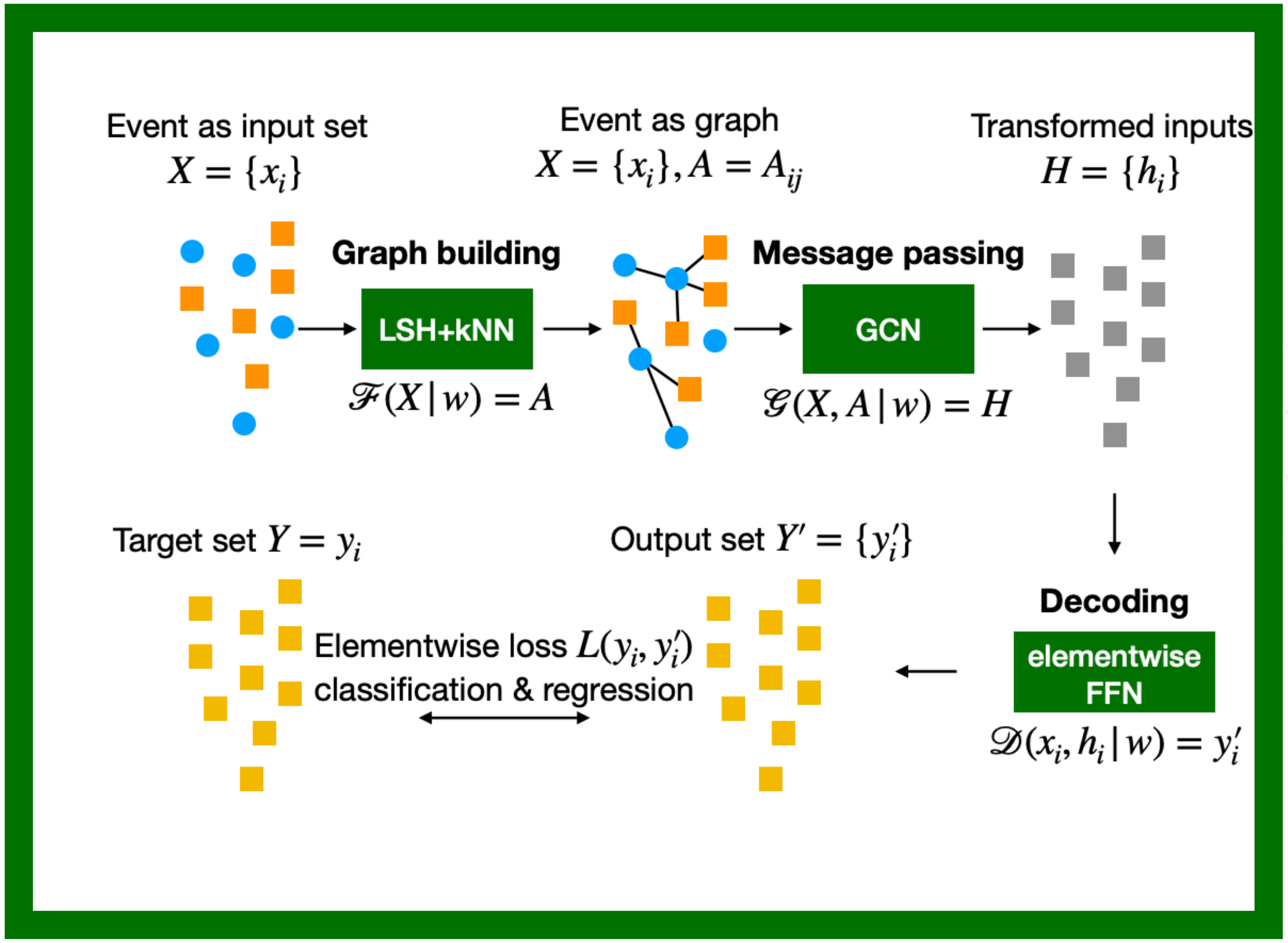


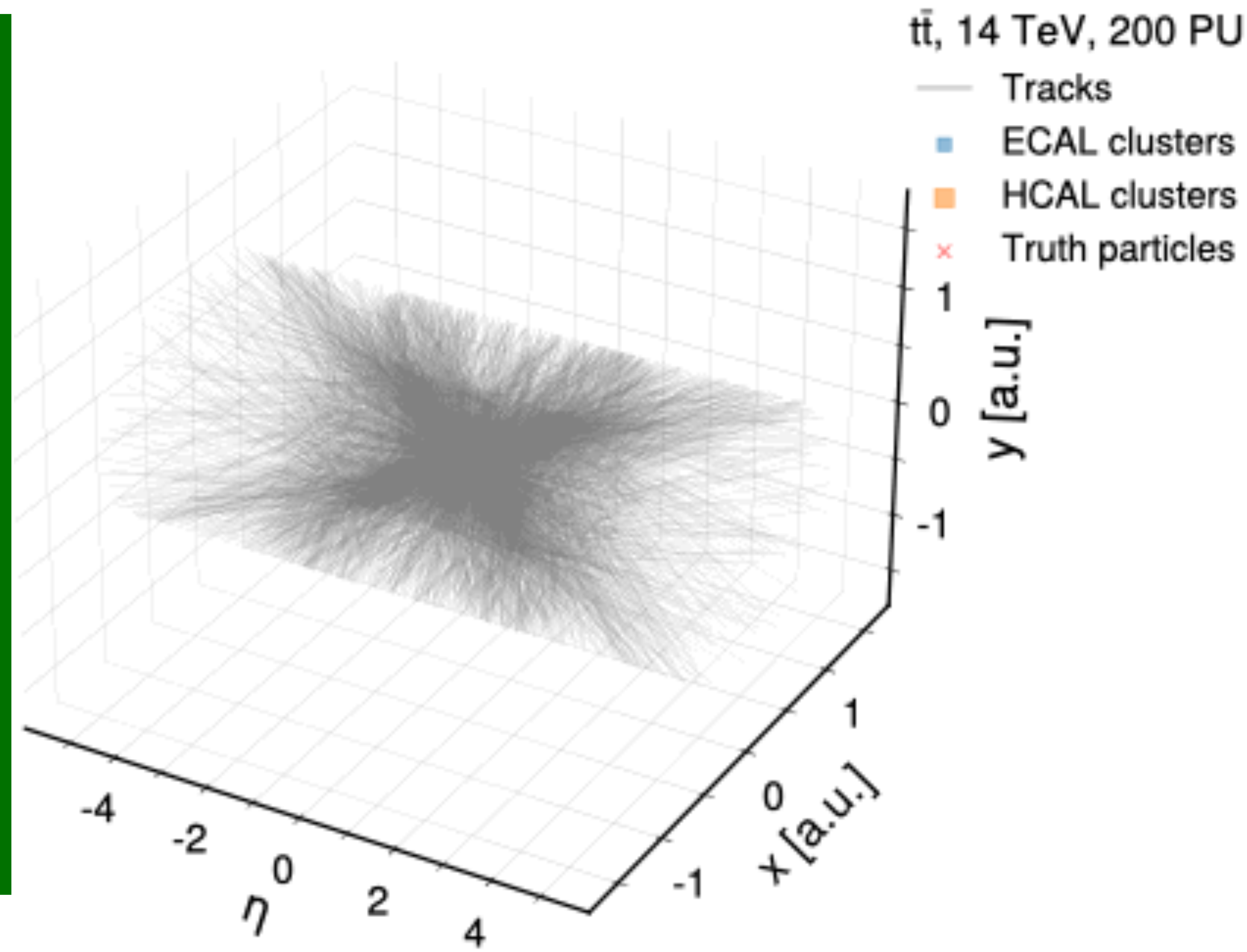
All particles in on fell swoop

- First ideas of full particle based reconstruction are emerging
- Tools are emerging to do particle reconstructeion in one go

arxiv:2101.08578

Success of Deep Learning

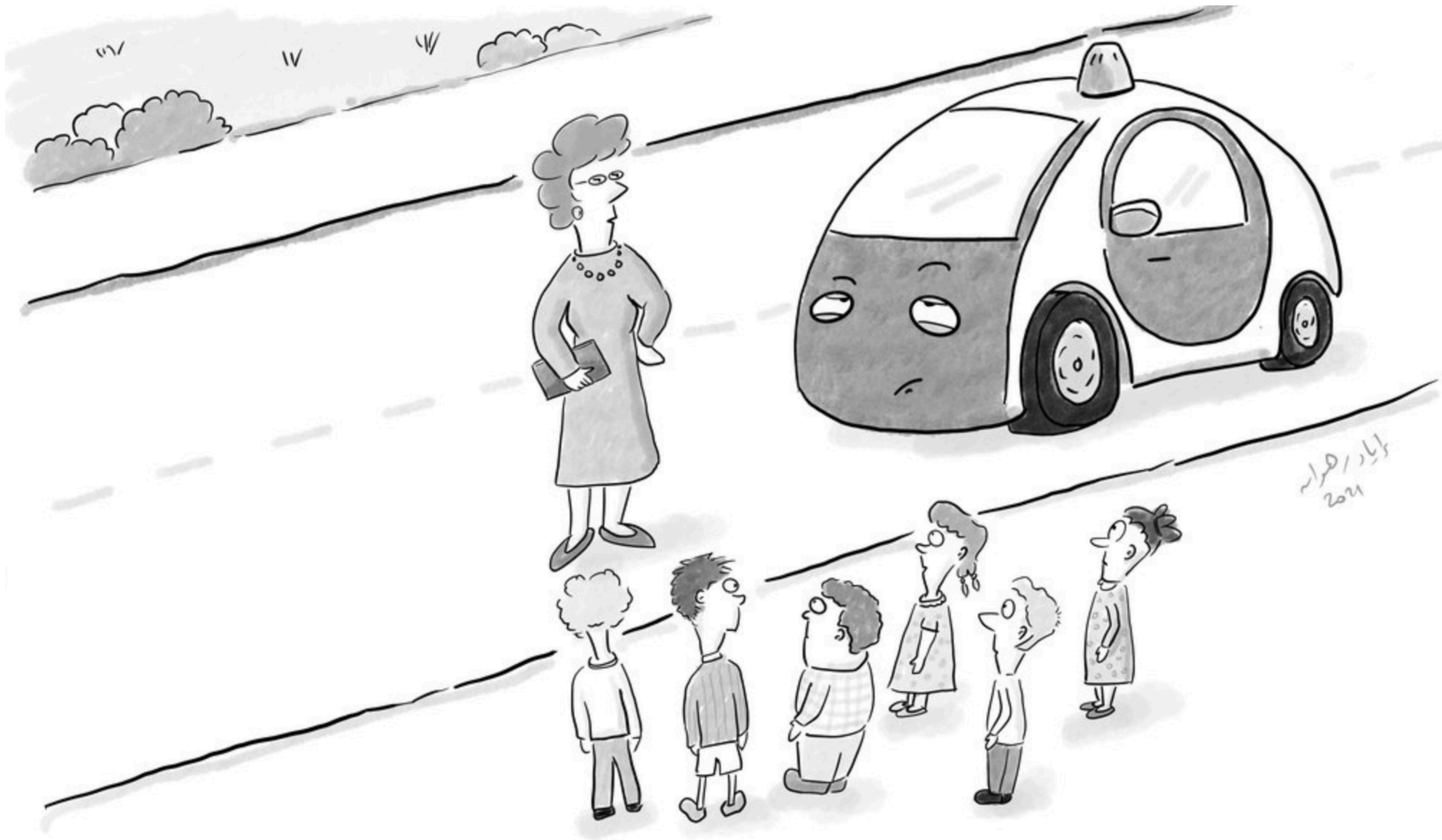




- First ideas of full particle based reconstruction are emerging
- LHC is a great place for DL because we have fantastic simulation

arxiv:2101.08578

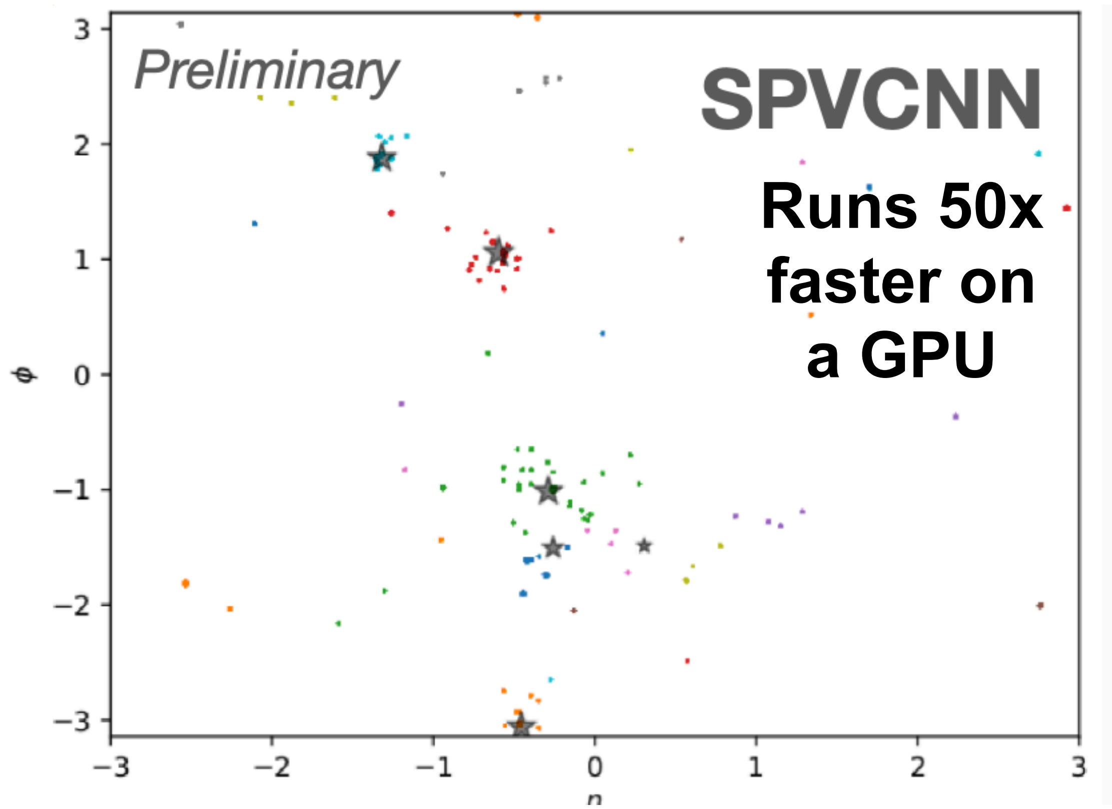
Taking a Leap of Faith

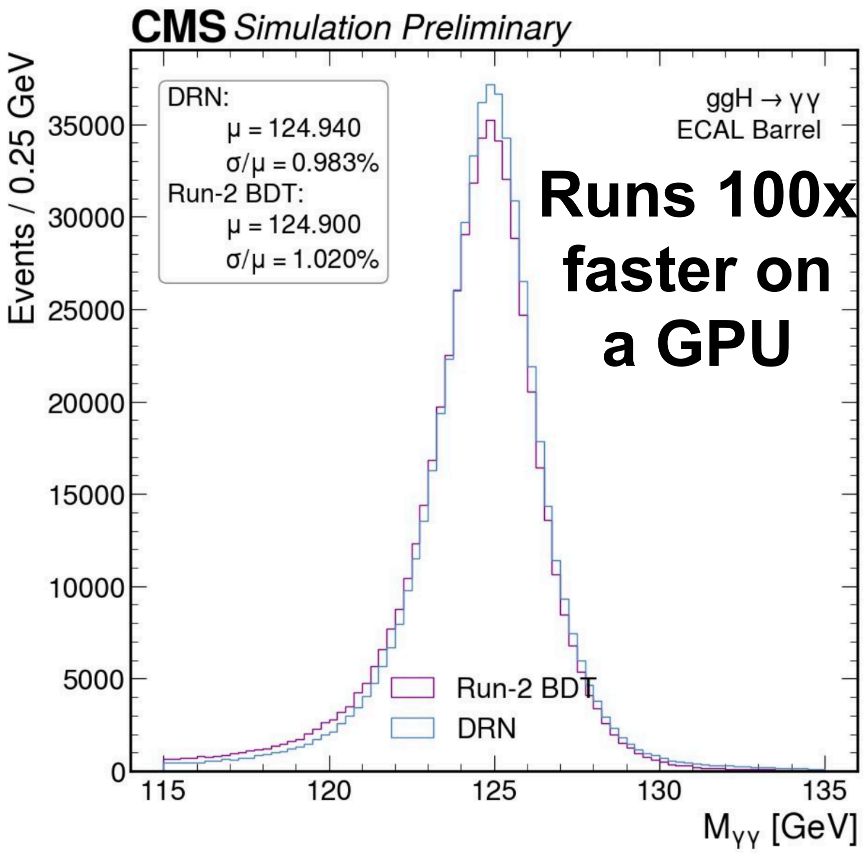


"Remember kids, you should never look before you cross, because driverless cars will always stop for you!"

Can we really trust AI to work from scratch well? always?

Deep Learning

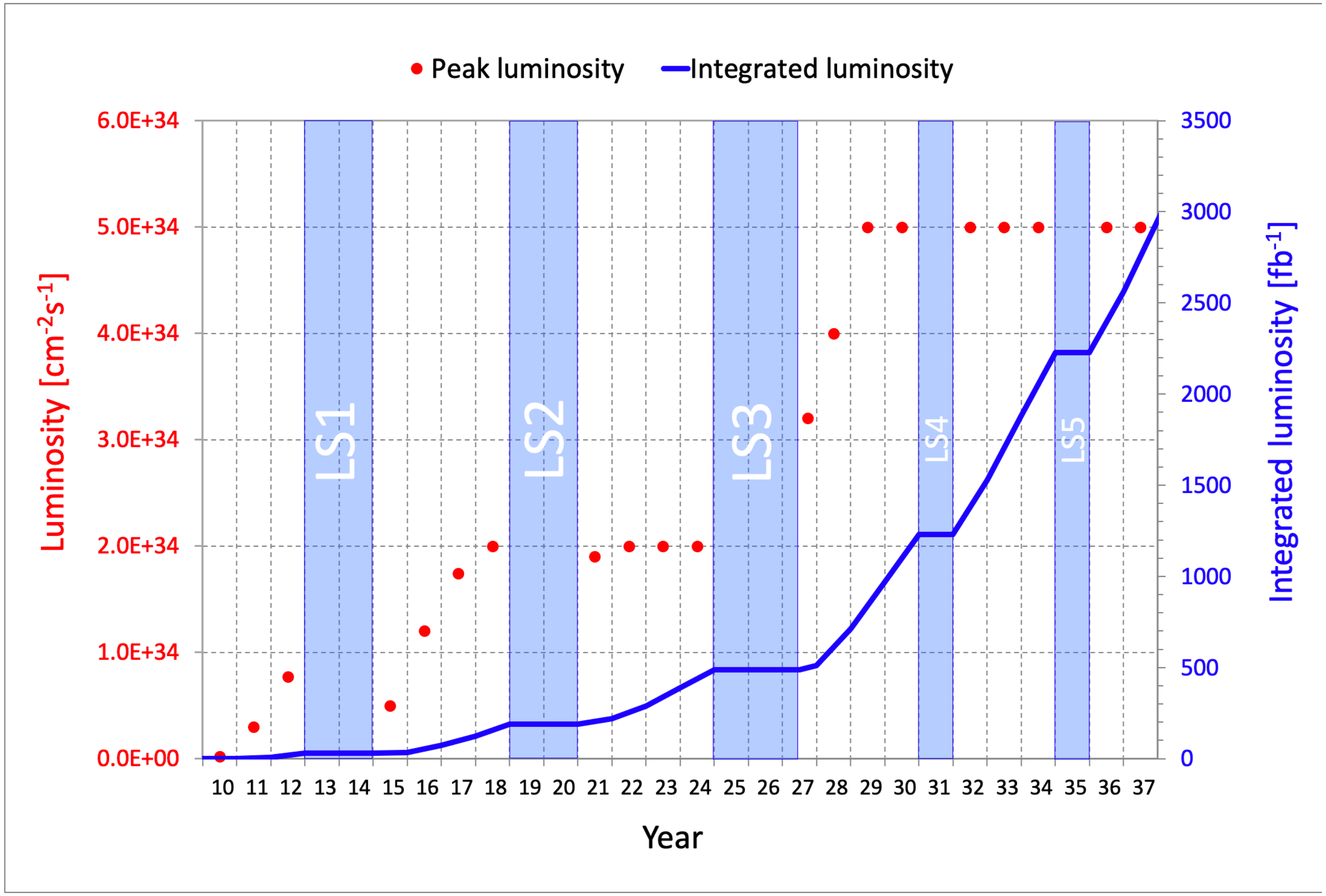




- We are building a number of algorithms with deep learning
  - These are quickly becoming part of LHC reconstruction algos
- Additionally these algorithms run dramatically faster on GPUs
  - Incorporate GPUs within our existing compute workflows

Where does this fit at LHC?

- The LHC has topped out in energy

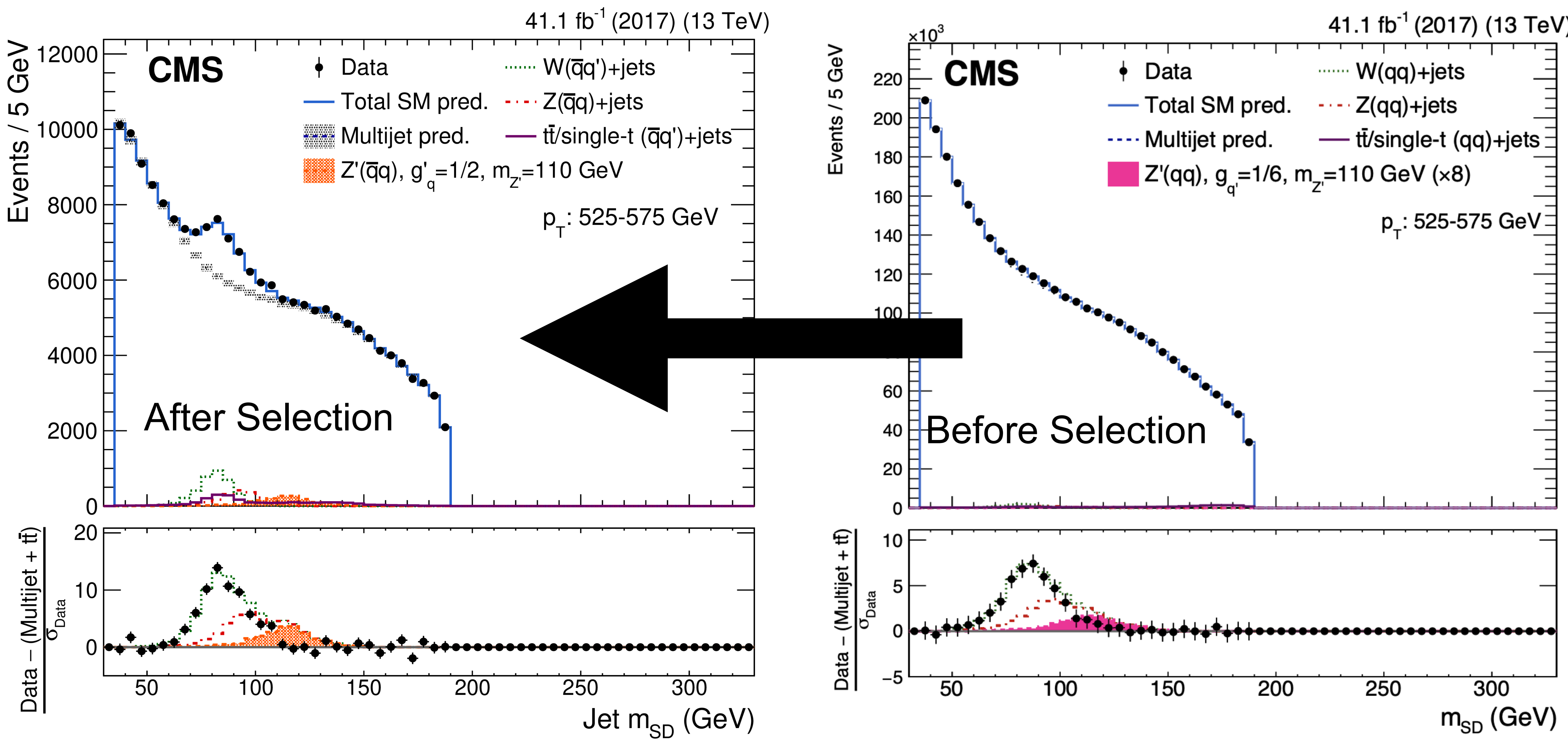


We still have 15 years of LHC running
20x more data to come

LHC Plan

- Lack of higher energy beams means
  - Analyses focus on measurements with lots of data
    - These are often hard and precise measurements
    - Long term analyses are focus (ie. W mass)
  - Creative final states we ignored in the past
    - Rethinking the strategy to search for new physics
    - Finding events that we couldn't in the past
- There are an incredibly diverse set things to explore

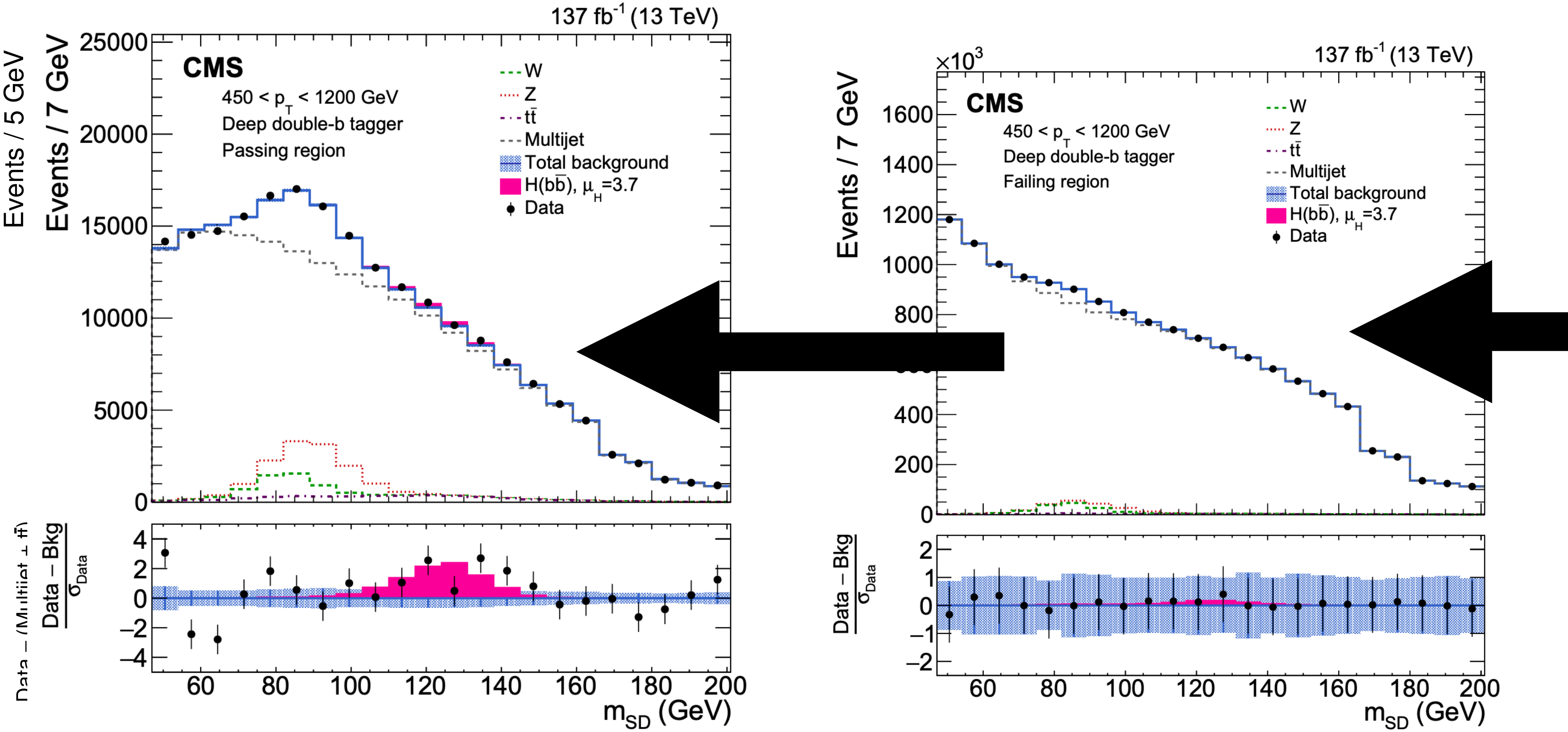
Looking for small signals



There is still a wealth of unexplored physics at the LHC

Its just a bit harder to find

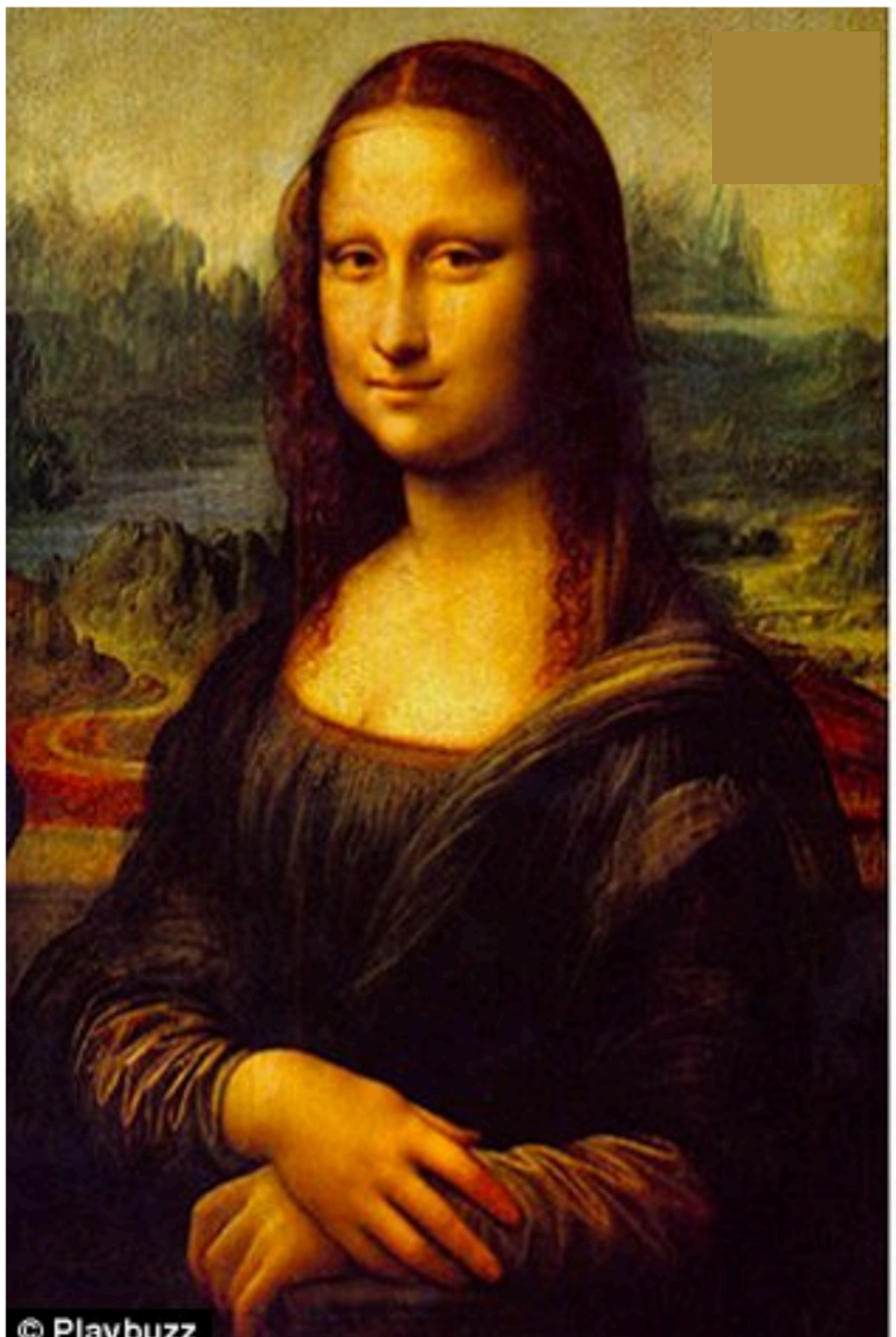
Looking for small signals



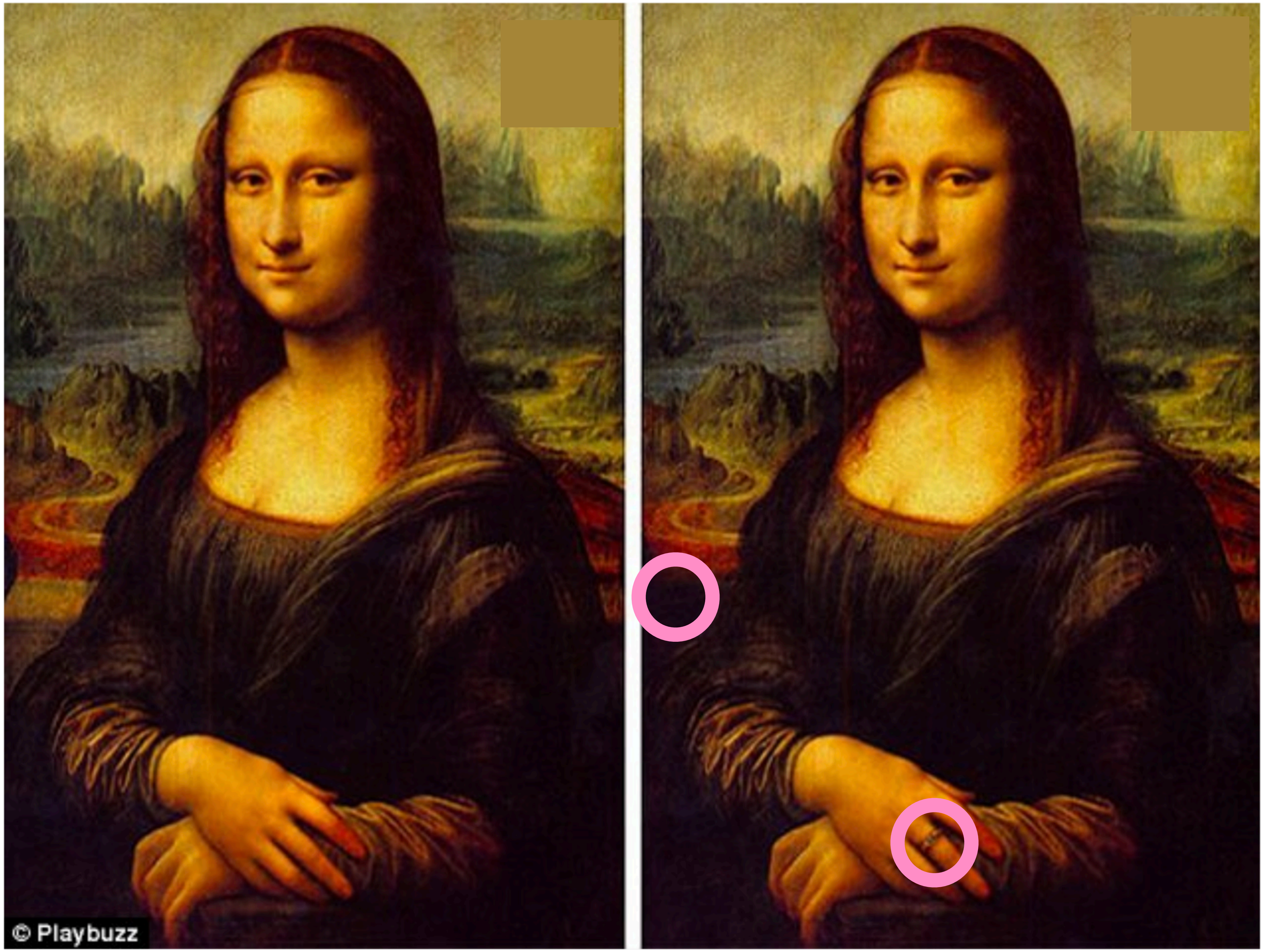
There is still a wealth of unexplored physics at the LHC

Its just a bit harder to find

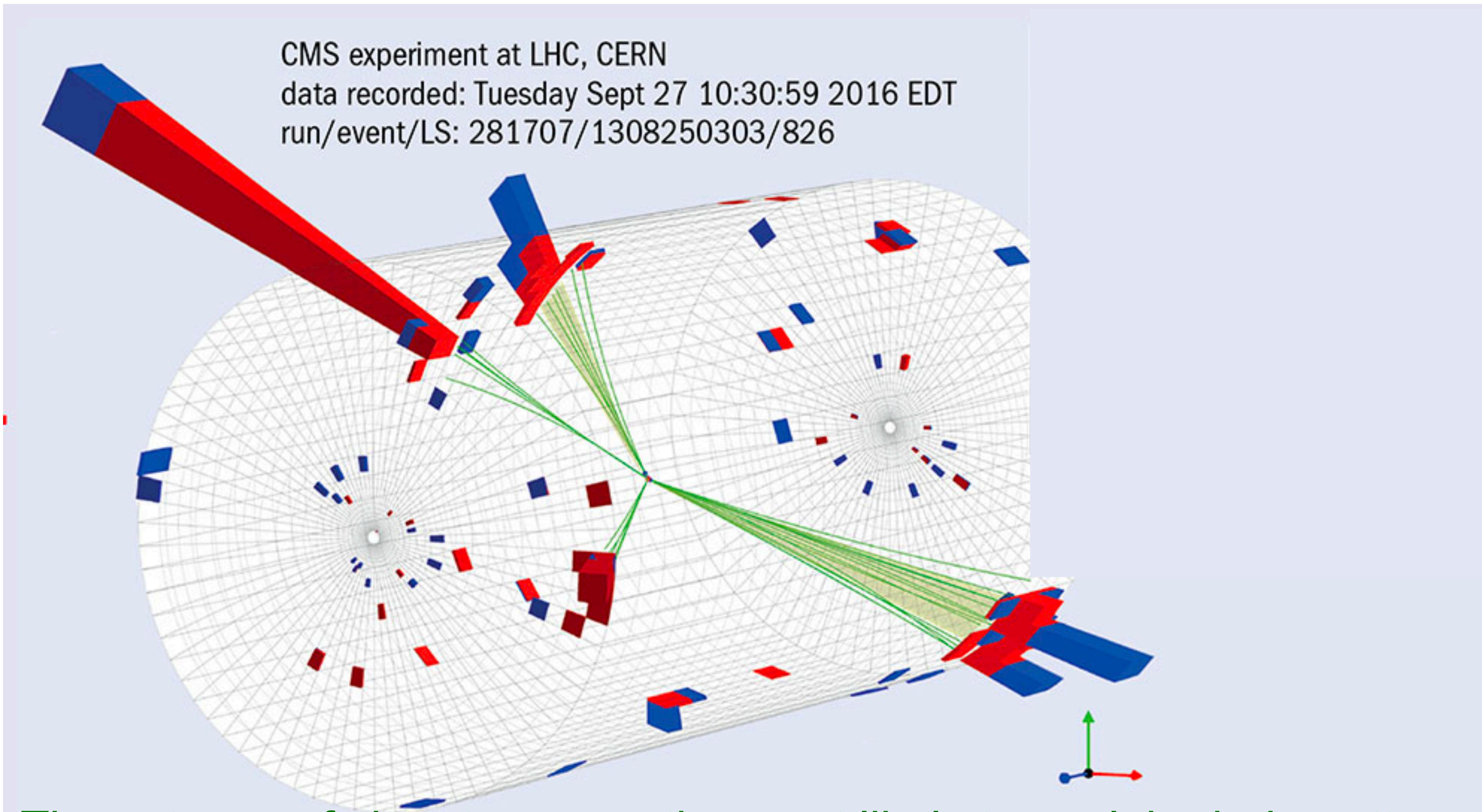
What is different w/Left and Right?



The Need for Subtlety

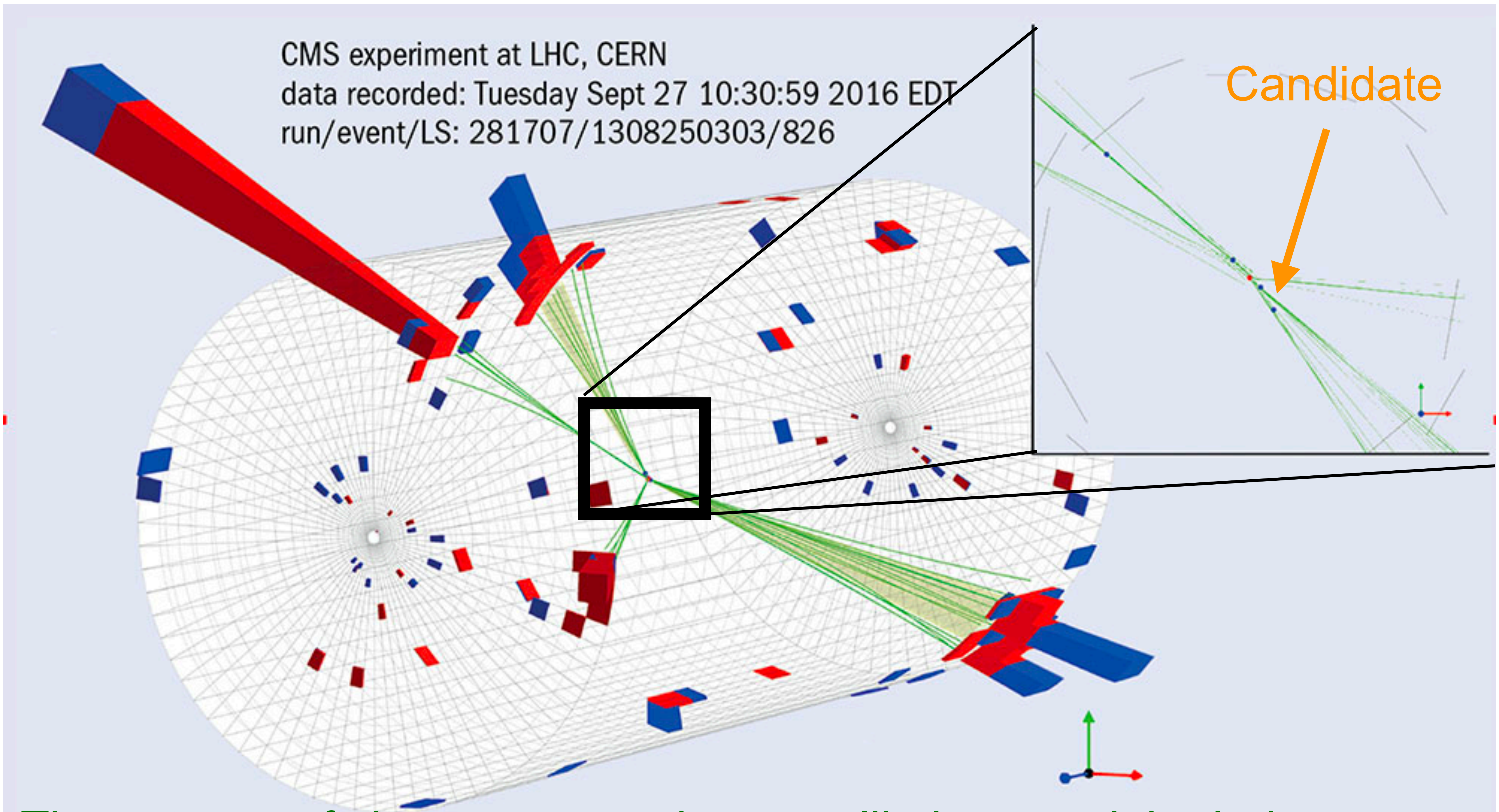


The Need for Subtlety



These types of signatures are the most likely to explain dark matter

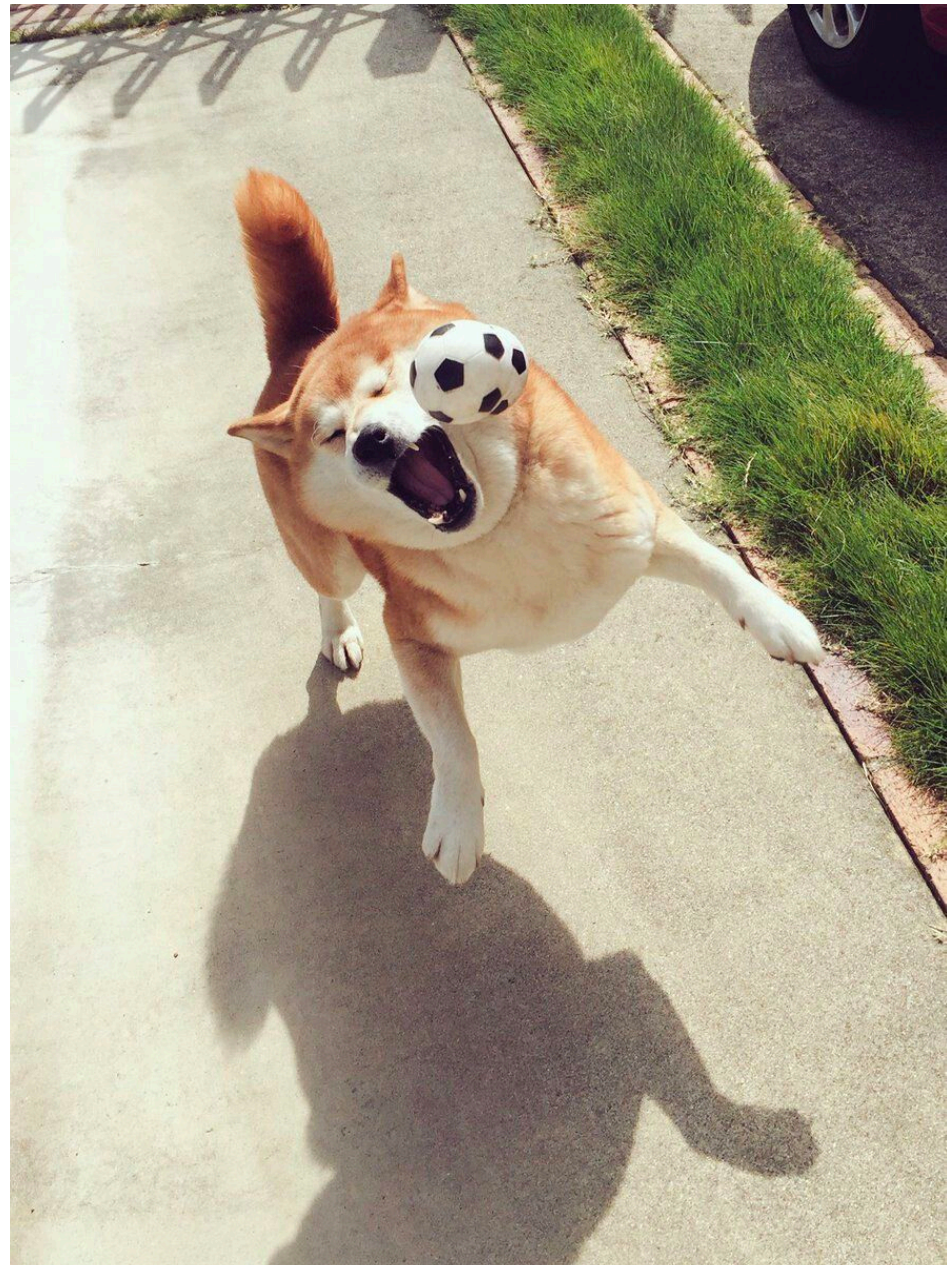
The Need for Subtlety



These types of signatures are the most likely to explain dark matter

Where are we now?

- The LHC has been running for the past 10 years
  - We have made some remarkable discoveries:
    - Higgs Boson
    - Measurements of top quarks, W, Z bosons…..
    - Strong constraints on Dark Matter and New Physics
- The times are changing:
  - We find ourselves doing more deep learning
  - We are also looking for harder to find signals



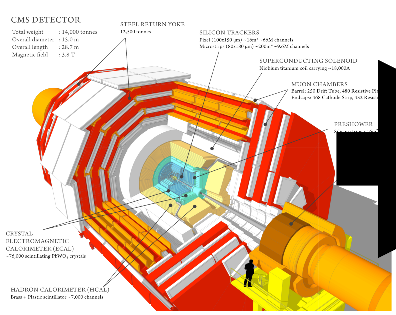
Think Fast
(NN Inference)

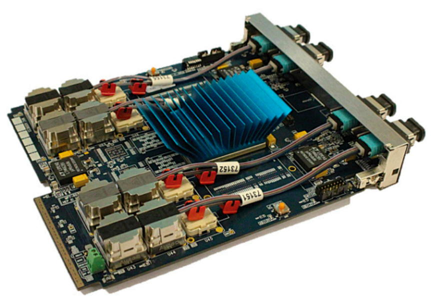
Spanning Frequencies

40 MHz **1 kHz**

Radiation Hard ASICs

FPGA Boards





Select 1 event in 400

The rest is thrown away **<u>Forever</u>**!

320 tb/s

Fast

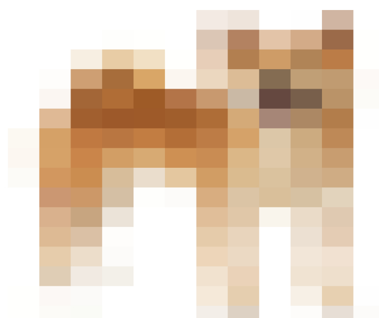
40 MHz Collisions

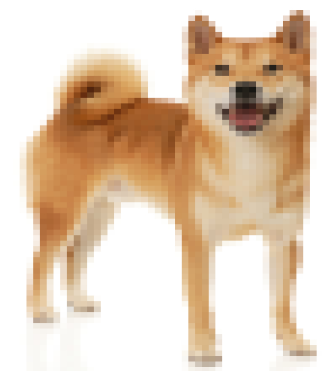
10 μs window

L1Trigger

Spanning Frequencies

40 MHz → **1 kHz**

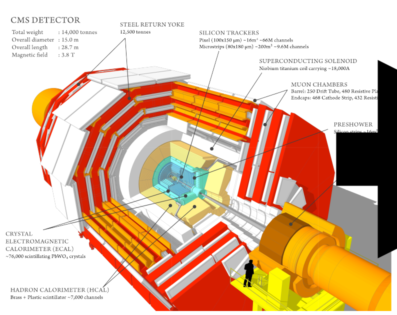


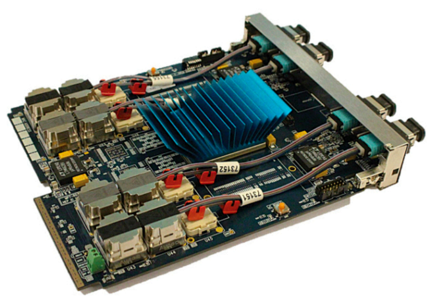


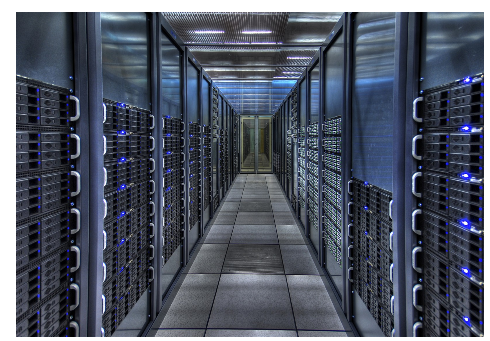
Radiation Hard ASICs

FPGA Boards

Local CPU Cluster







320 tb/s

1 tb/s

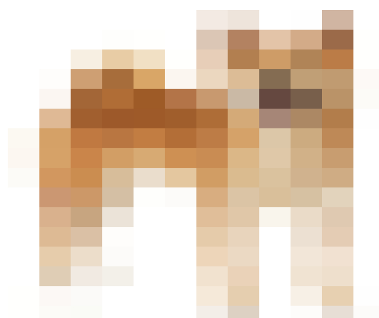
Fast
40 MHz Collisions
10 μs window
L1Trigger

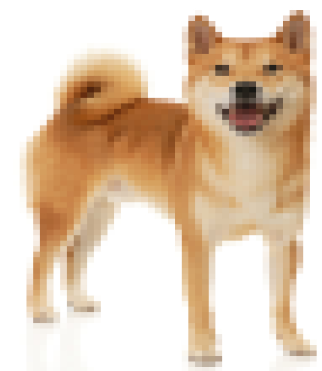
Intermediate
100 kHz Collisions
<500 ms window
High Level Trigger

Select 1 in 100

Spanning Frequencies

40 MHz → **1 kHz**



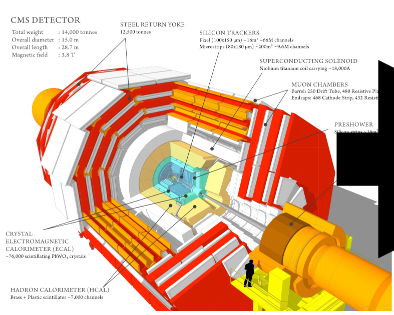


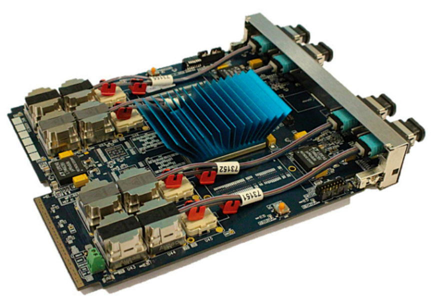
Radiation Hard ASICs

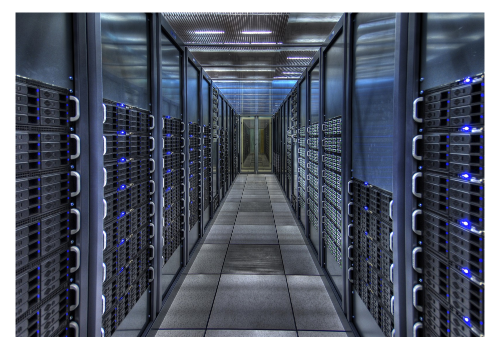
FPGA Boards

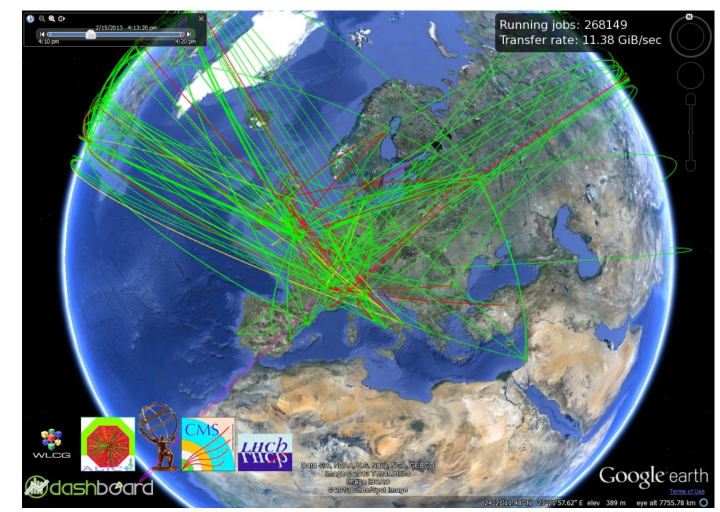
Local CPU Cluster

CPU Grid









320 tb/s

1 tb/s

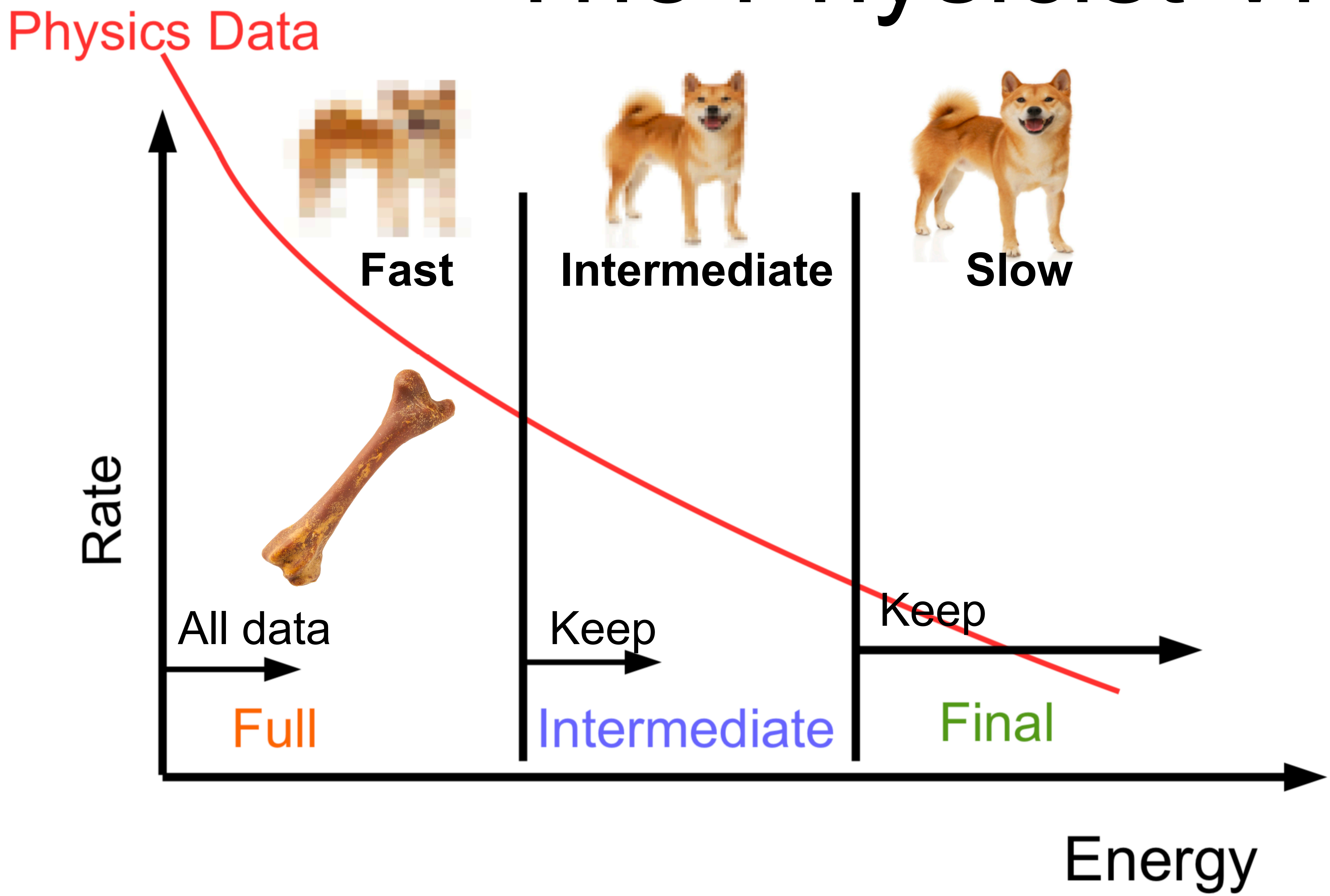
10 Gb/s

Fast
40 MHz Collisions
10 µs window
L1Trigger

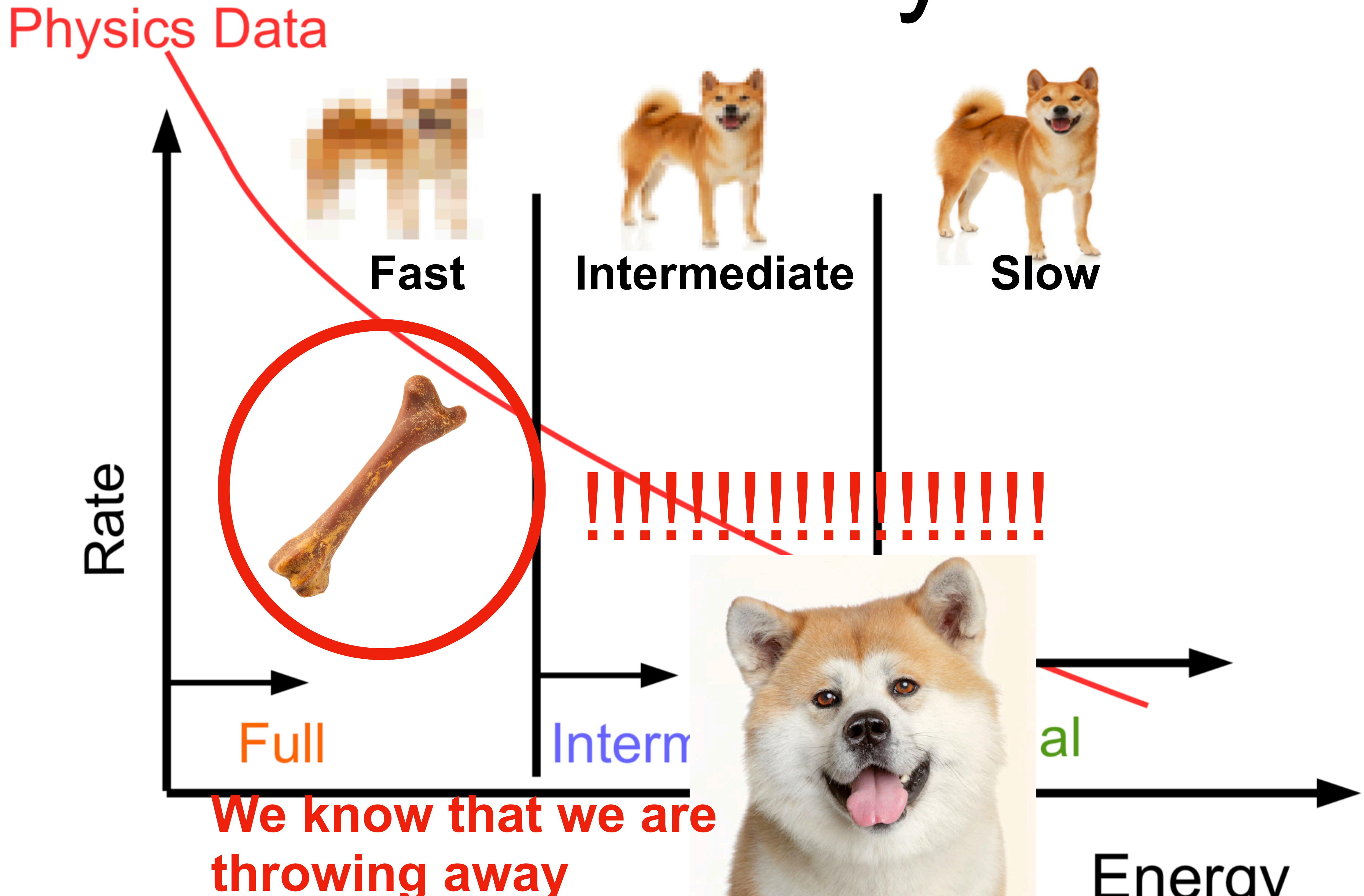
Intermediate
100 kHz Collisions
<500 ms window
High Level Trigger

Slow
1 kHz Collisions
10 s window
Offline Cluster

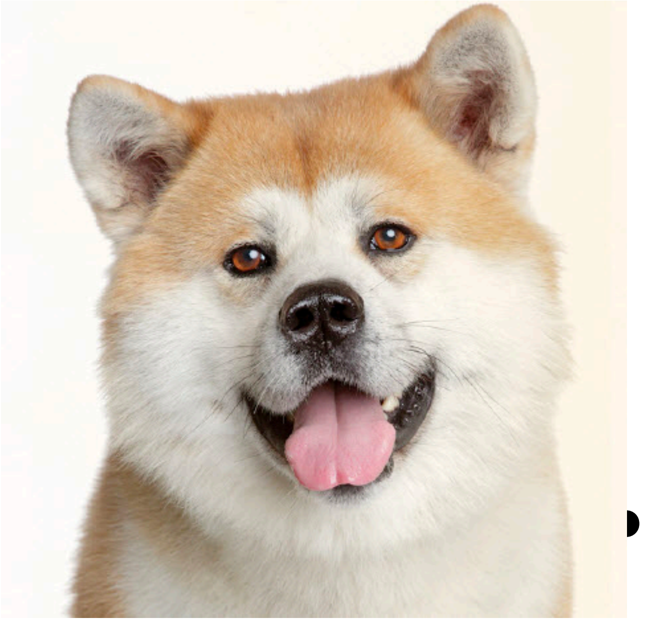
The Physicist View



The Physicist View

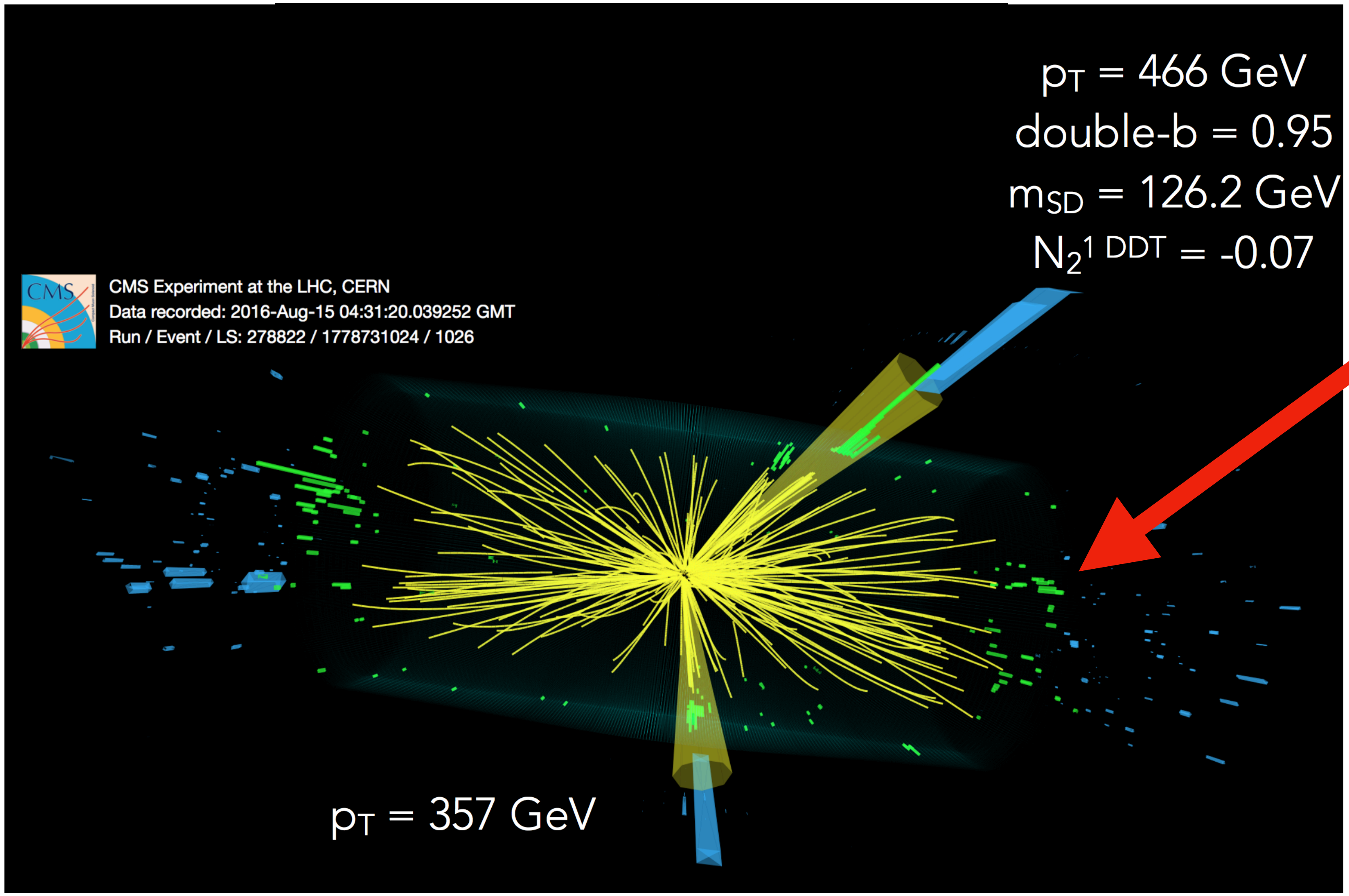


We know that we are throwing away a lot of good data



Hidden gems?

- There is a plethora of physics that we throw out



Higgs boson right on the cusp of being thrown out

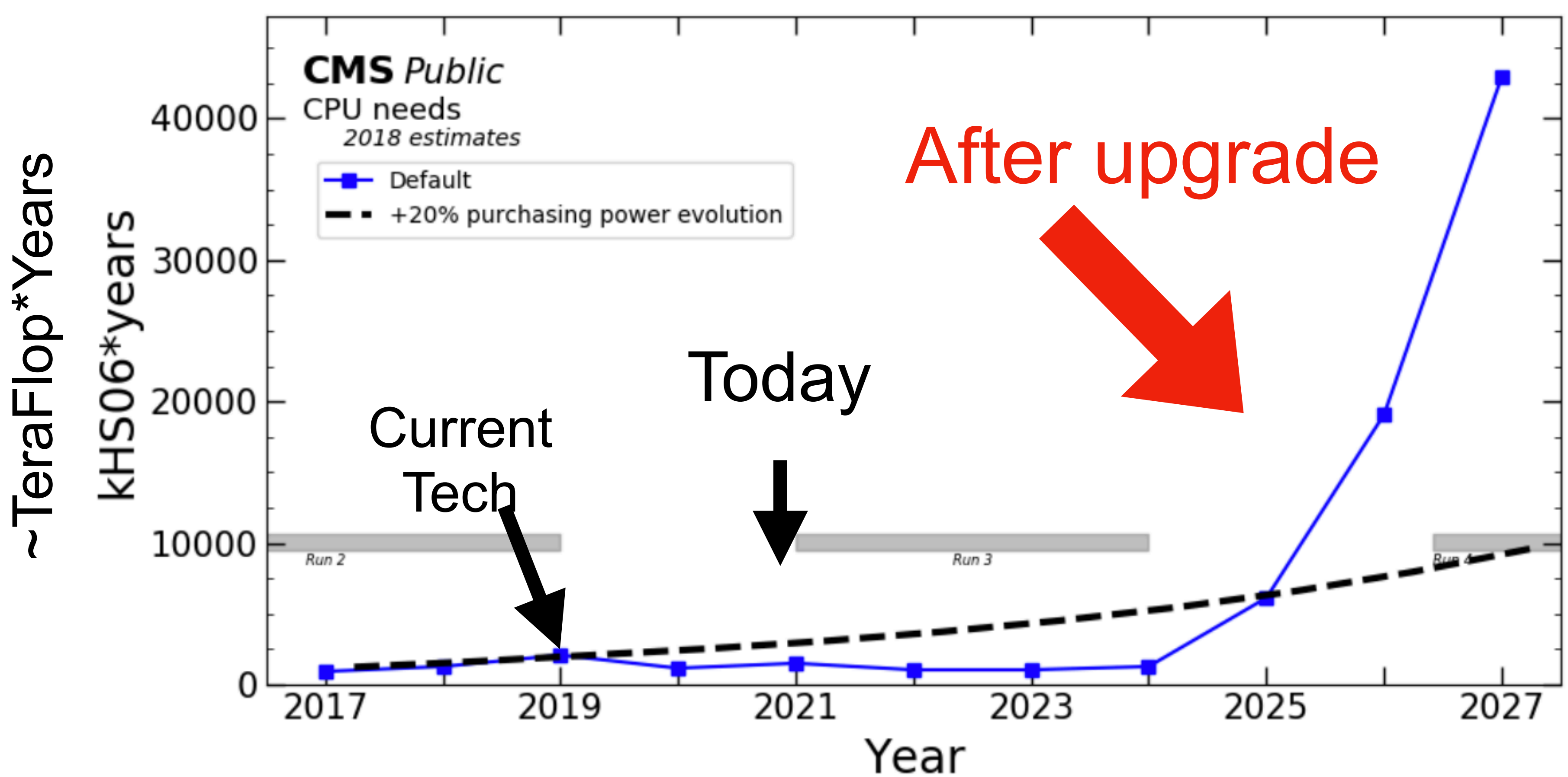
The dream

- At the moment:
  - We only get a full data of one in 40,000 collisions
  - There is interesting physics that we have to throw away

- We would like to analyze every collision at the LHC
  - To deal with this we need to increase our throughput
  - Ultimately this means going to 100s of Tb/s

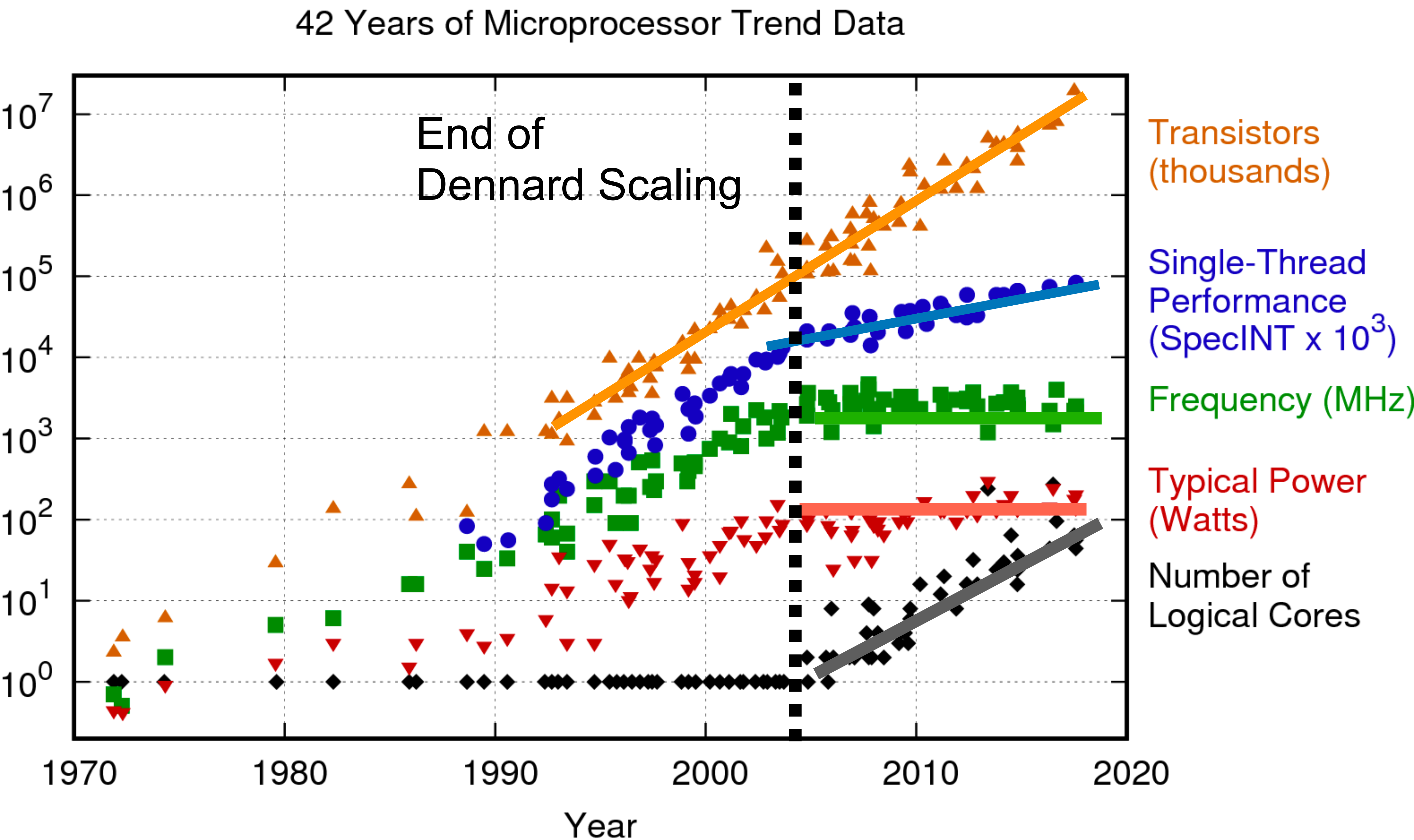
The Challenge

- To deal with the upgraded LHC intensity
- To preserve current physics we are upgrading the system
  - Our event size will have to be 10x larger
  - We will have to take data at 5 times the current rate



Results

The Crises



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

Processor Technology

Will we be able to handle the future upgrades?

Modern Processing

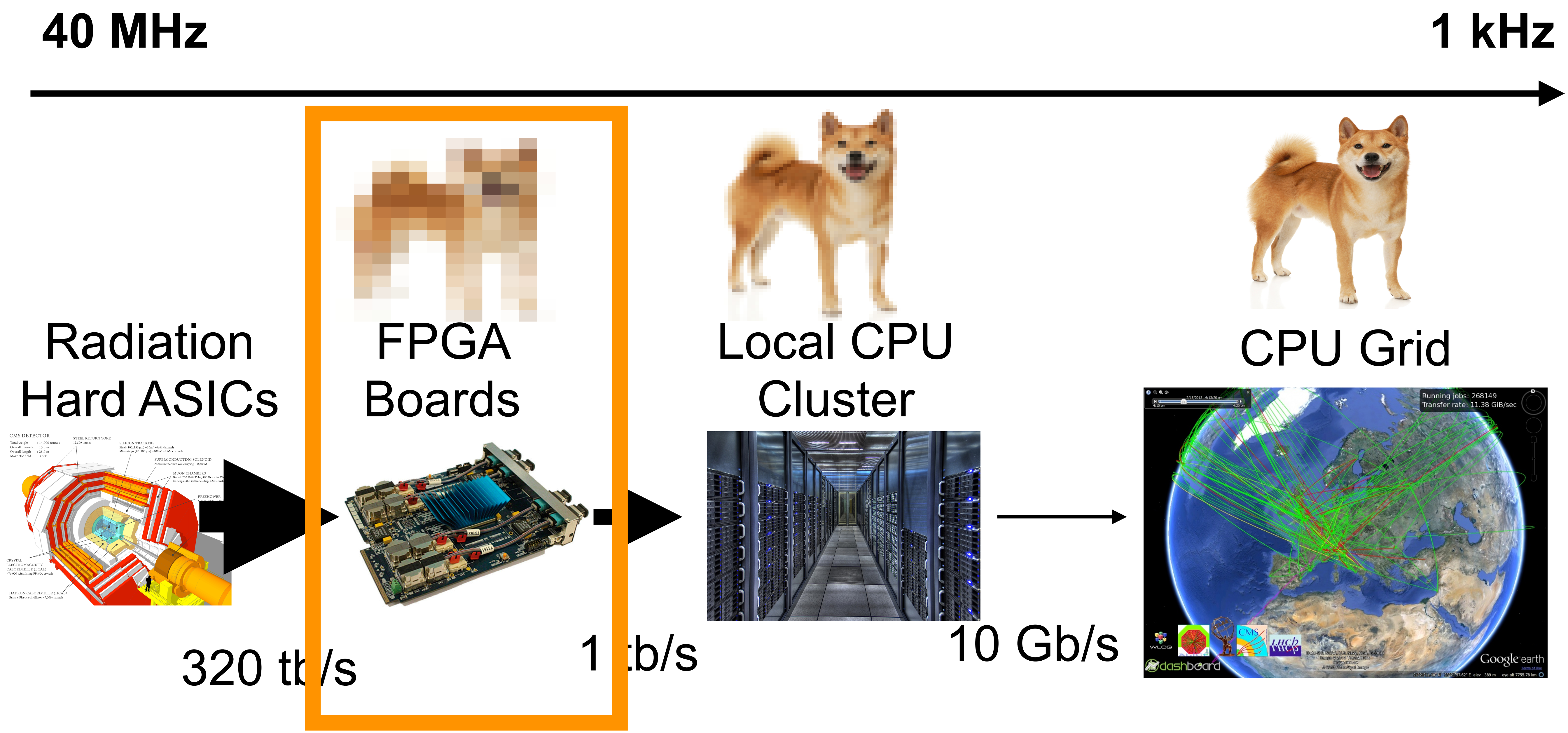
- Multi cores CPU:
  - Your standard CPU with split all up (you know this)
- GPU :
  - Effectively many multi-cores with simplified instructions
    - Many cores in parallel (O(500)) with addition and mult.
  - Power hungry (better, but not too different from a CPU)
- FPGA:
  - Pre-programed the chip to do the operation you want
  - Every switch and multiplier assigned to a fixed patter
  - Energy efficient and hyper parallel (5000 parallel)
- ASIC:
  - FPGA but with inability to be programmed

Processing Tech

CPU	GPU	FPGA
		
1 player	A few at same time	The whole chip
A soloist	A group	An orchestra
Whatever	Main theme and some freedom to improv	Score has to be known perfectly beforehand

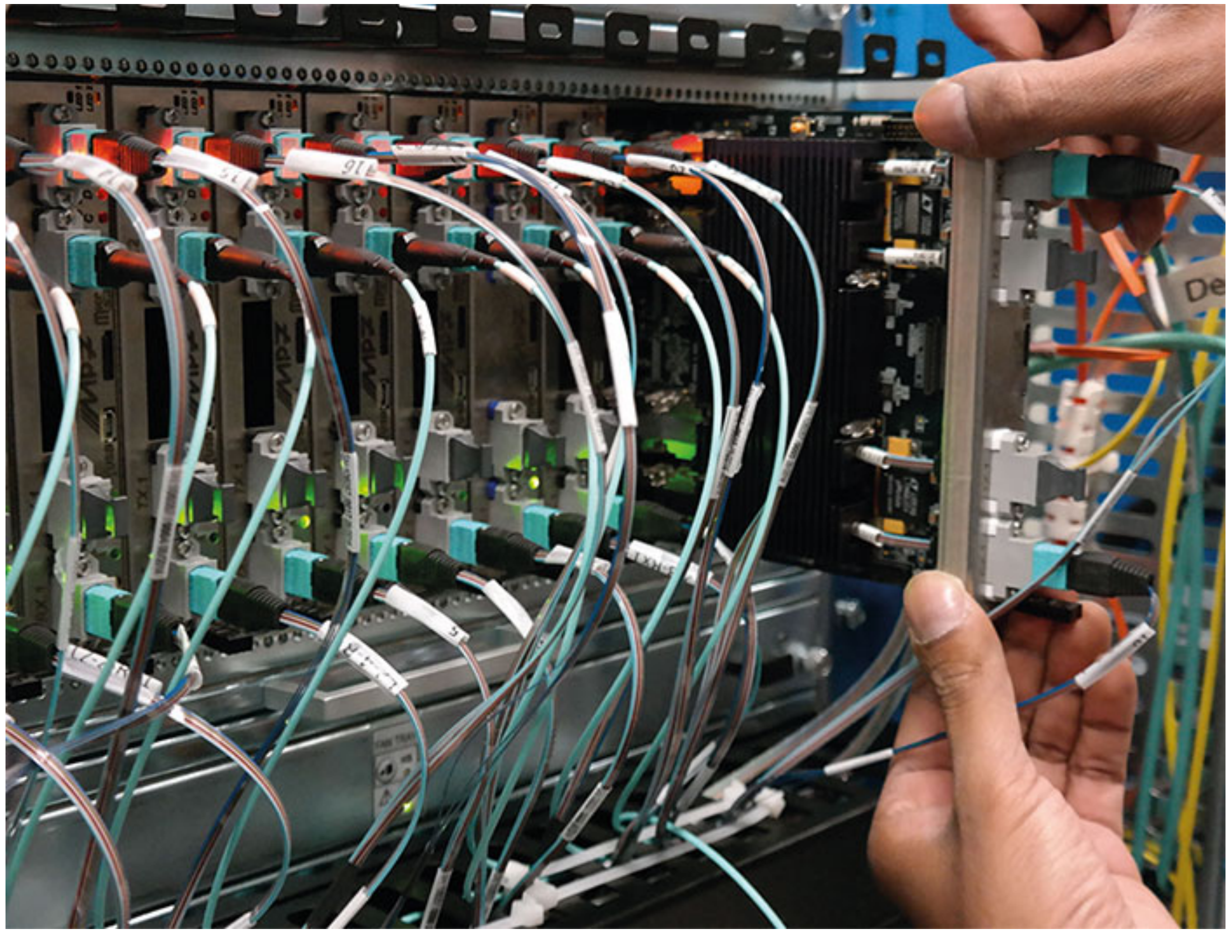
Processing Tech

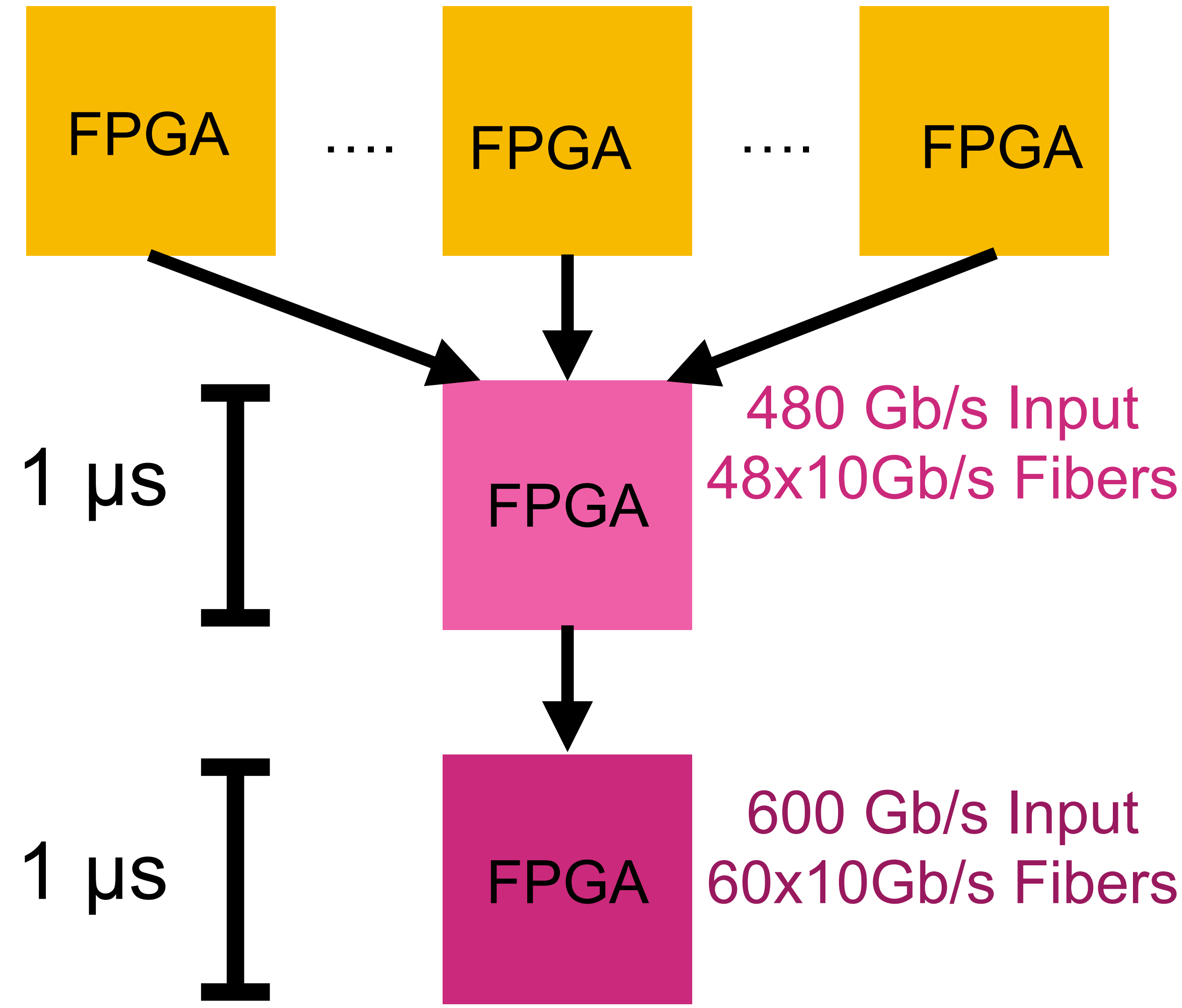
Past	Present	Future
CPU	GPU	FPGA
Past 		
Speed: 1	Speed: 20-50	Speed: 200-1000
Single complex operation per clock	Multiple simpler operations in parallel commonly available	Efficient packing of operations highly parallelized (low power)

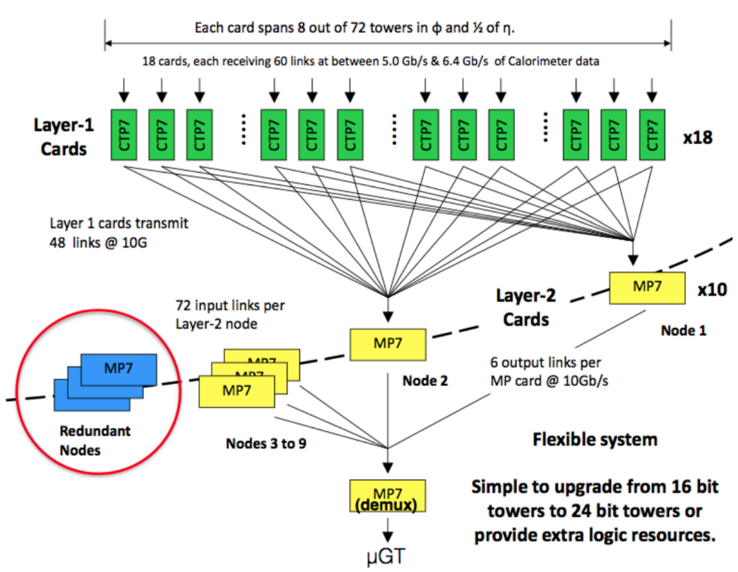


Real-time AI on every LHC Collisions

Current (Old) Tech



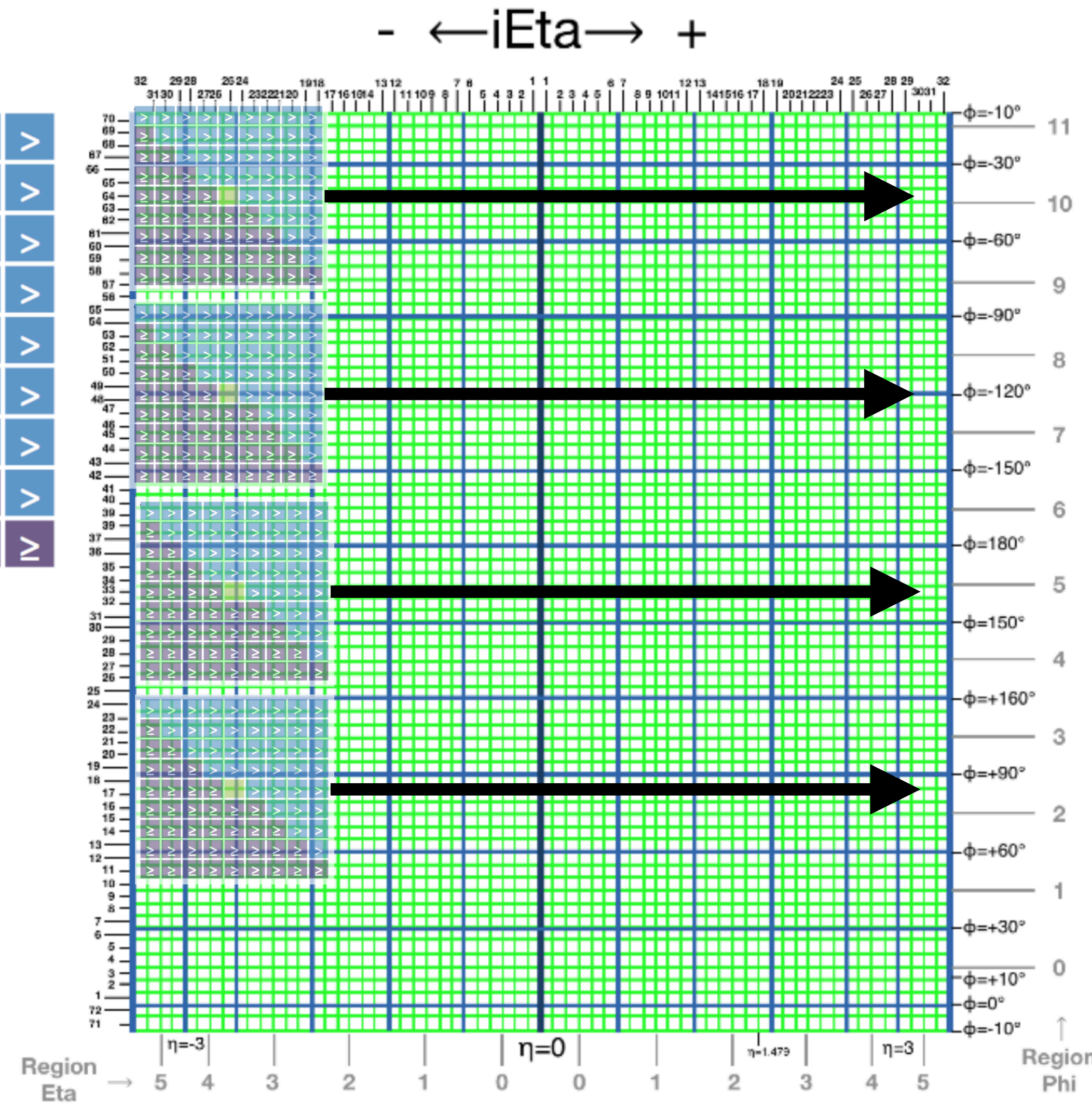




Current System is roughly 100 Virtex7 FPGAs interconnected with Fibers

Current Algos

Algo

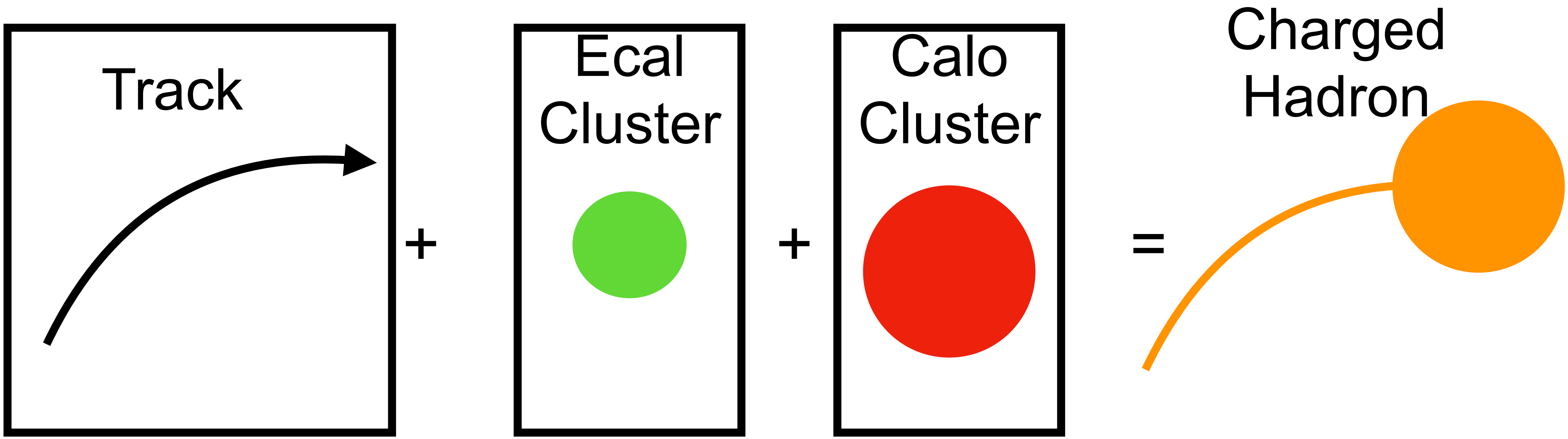


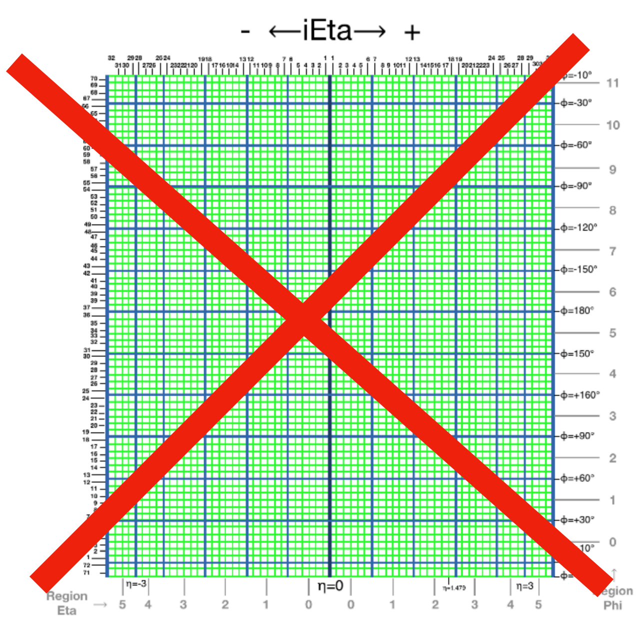
Simultaneously scan over calorimeter region a very simple algorithm

FPGA is essential to parallelize & deal w/ enormous bandwidth

Algorithms have traditionally been simple due to the size of the FPGAs + RTL code

Rethinking the Algos





No More Grid
of Information

Process information object by object

Reference

High Level Synthesis

Vivado™ HLS

Catapult

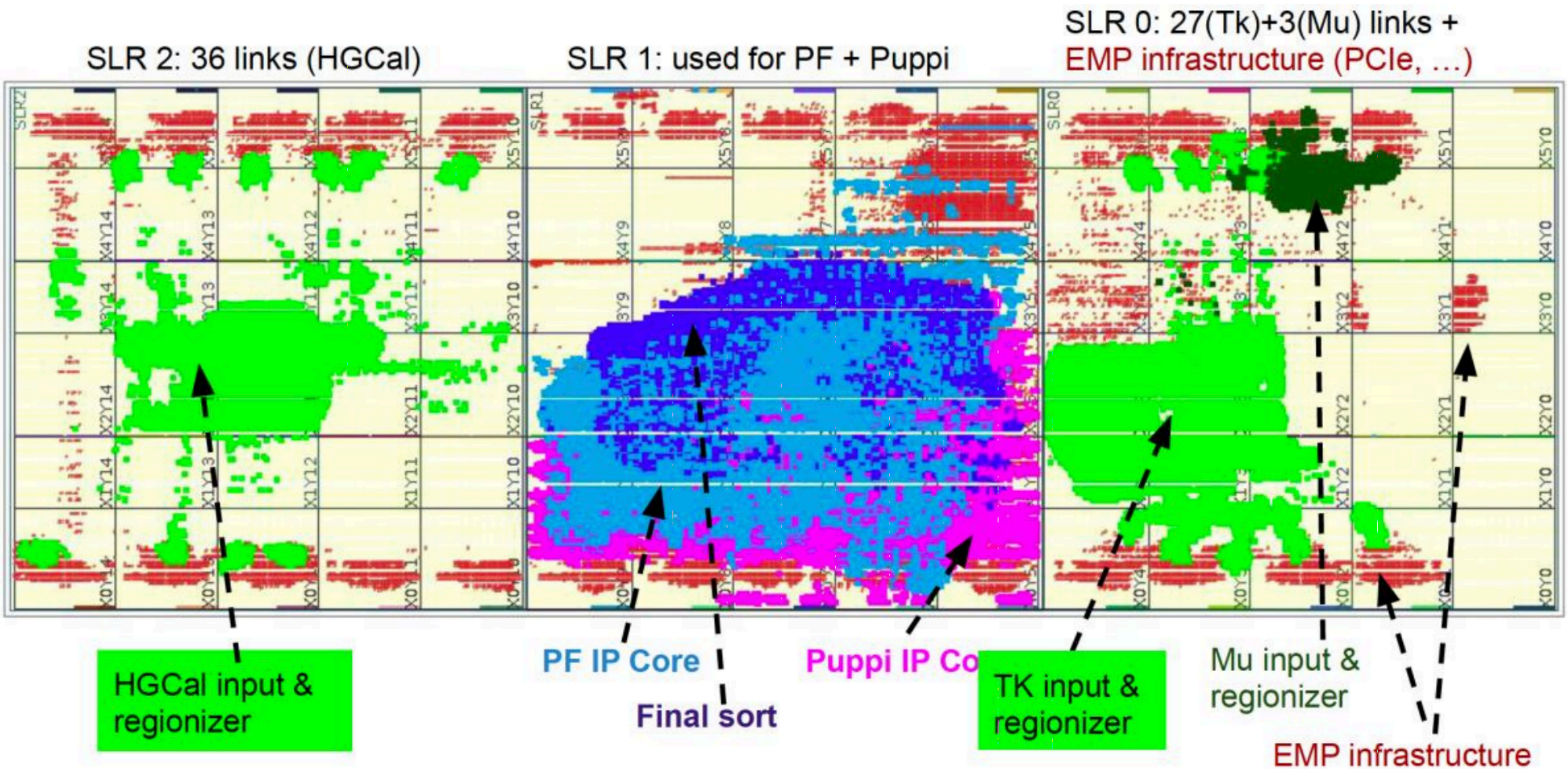
Mentor®

A Siemens Business

- Designing complex algorithms on FPGAs
  - Needed an approach to design/understand complex algos
  - We also wanted to be sure to capture the physics
  - As physicists, we prefer writing code in c++
- HLS has given us the possibility to develop algorithms quickly
  - Allows for fast turn around to deployment of algorithms

How does this fit?

An important element of the design flow
Make individual blocks small enough to fit on one die (SLR)
Crossing SLRs is slow



Real-Time Deep Learning

- We only have 1μs or less for the inference time
  - We need to run the networks at a rate > 40 MHz (II < 25ns)
  - Forced us to re-think DNN hardware implementations
- This work led us to the project:

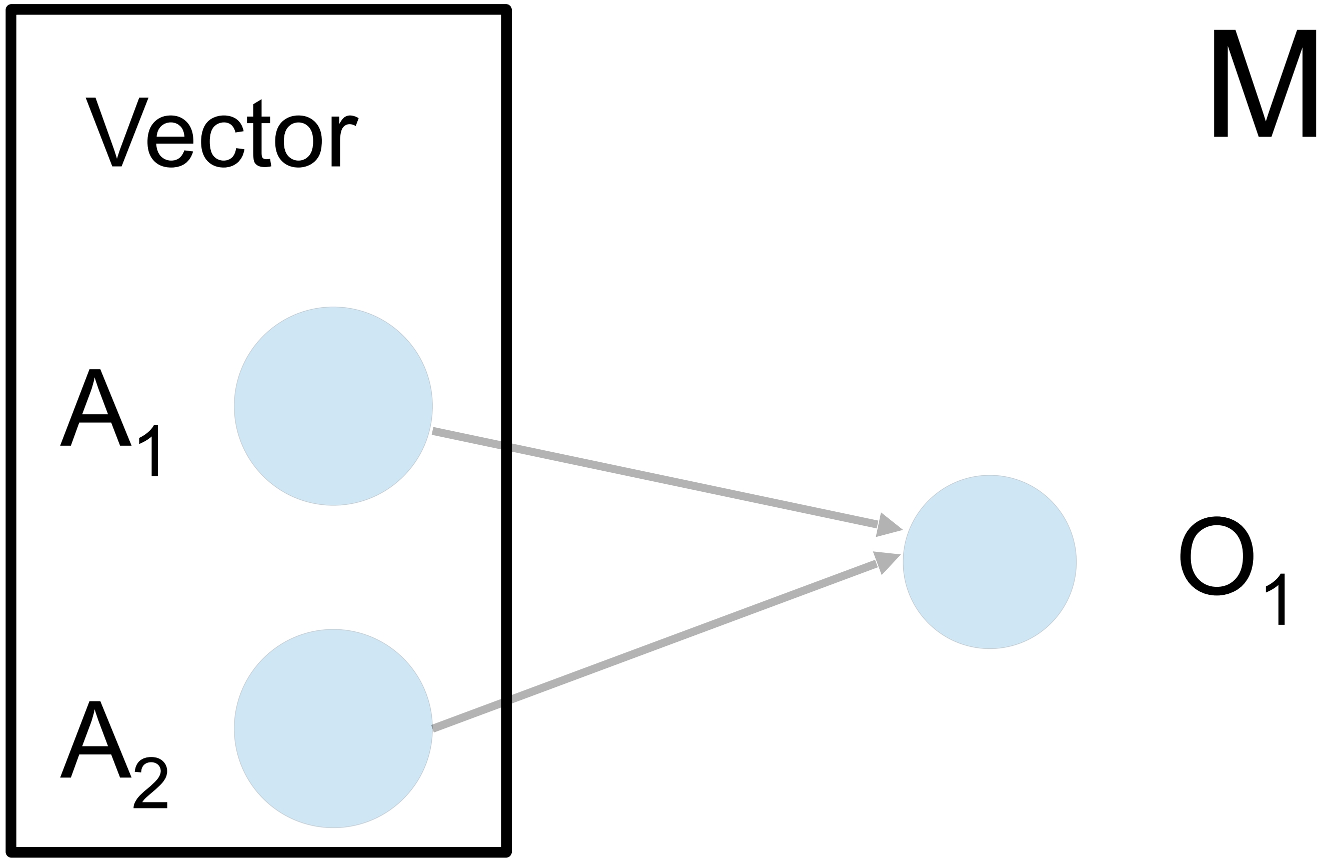
S. Han

D. Rankin

hls4ml

arxiv:1804.06913

Matrix Mult in Math



How can we parallelize this?

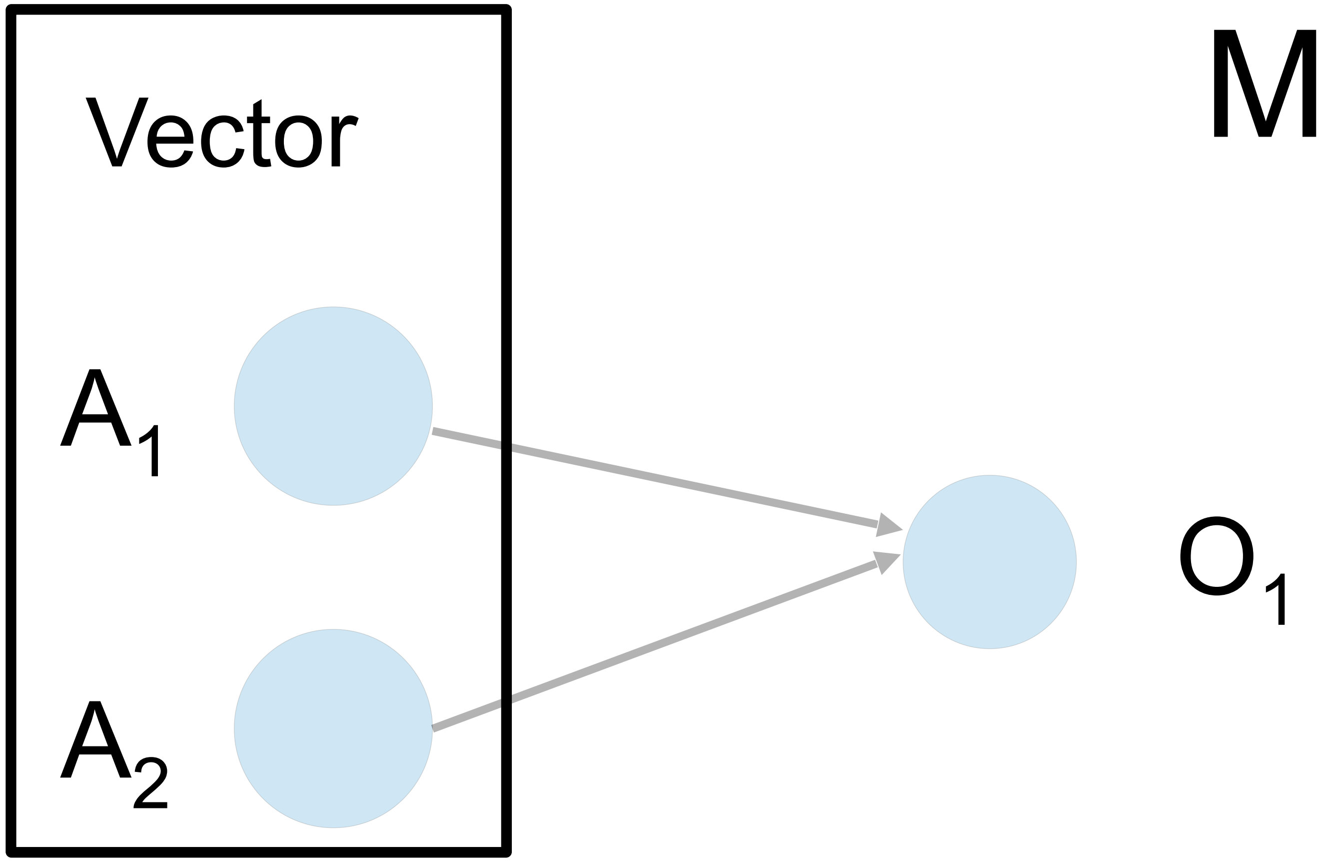
$$\varphi(A_1W_{11}+A_2W_{21}+B_1)=O_1$$

Vector Addition

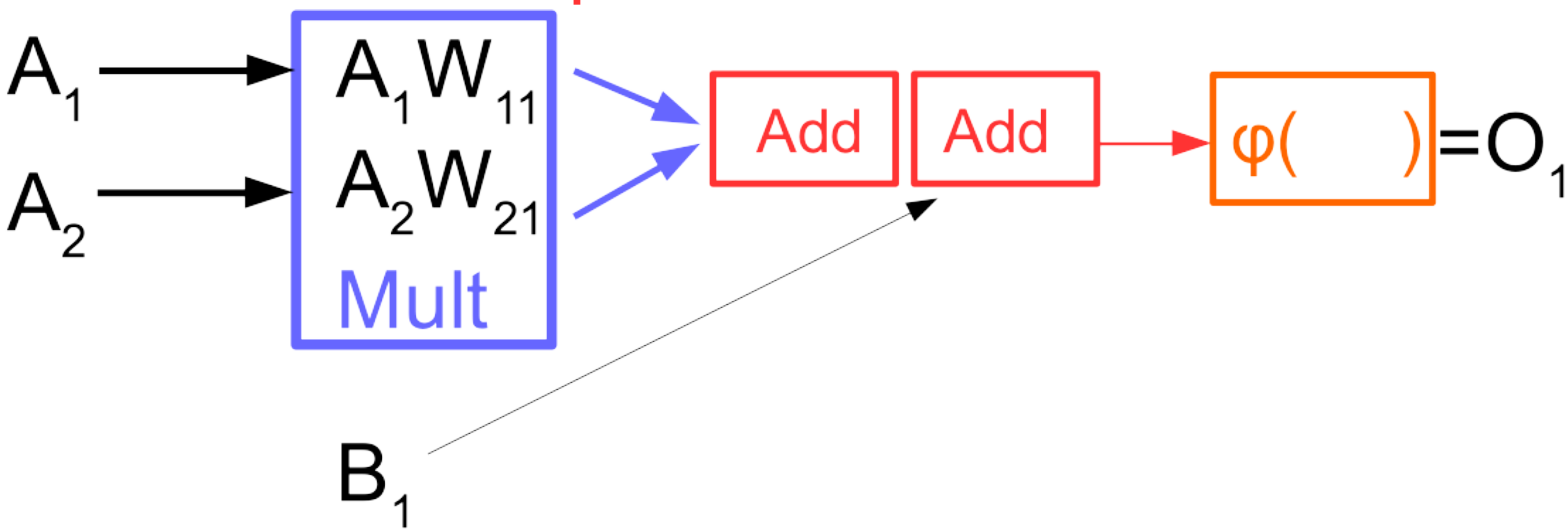
Matrix Multiplication

Activation function

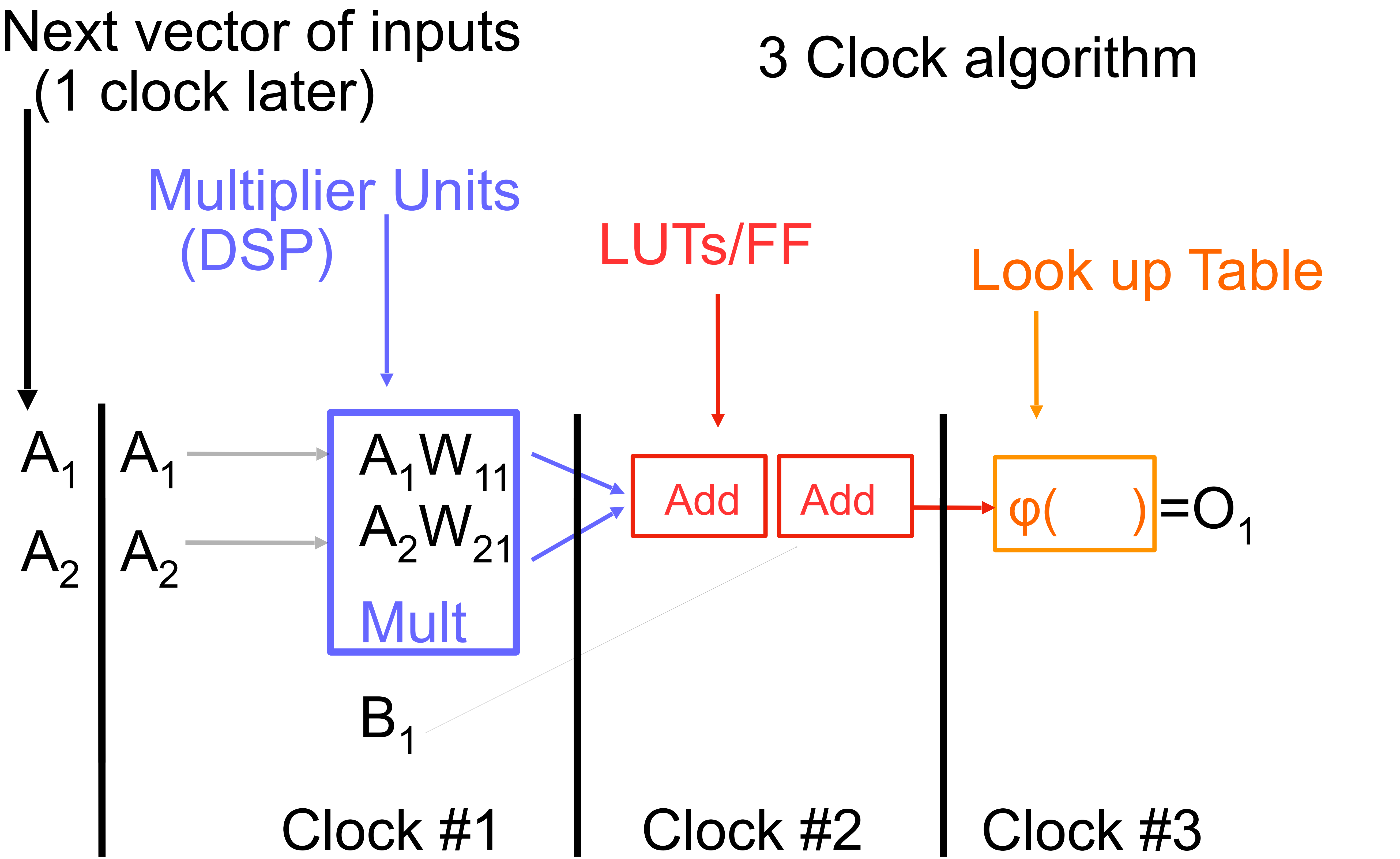
Matrix Mult in Math



How can we parallelize this?

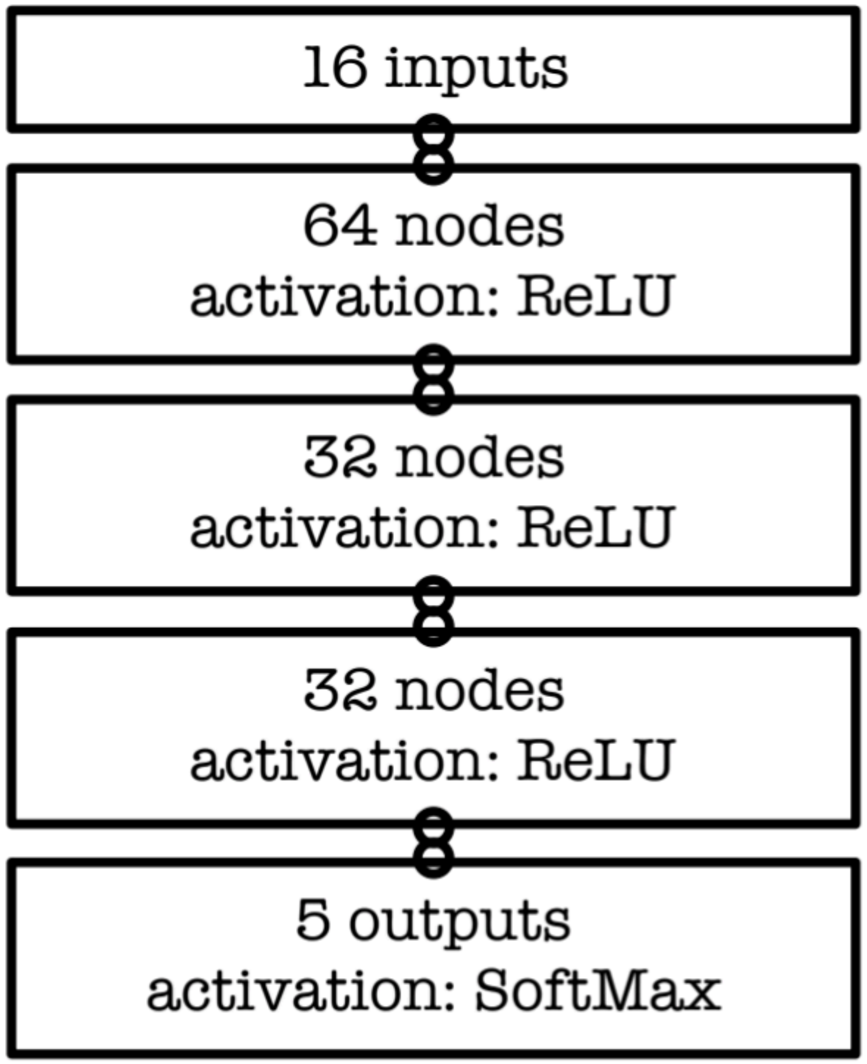


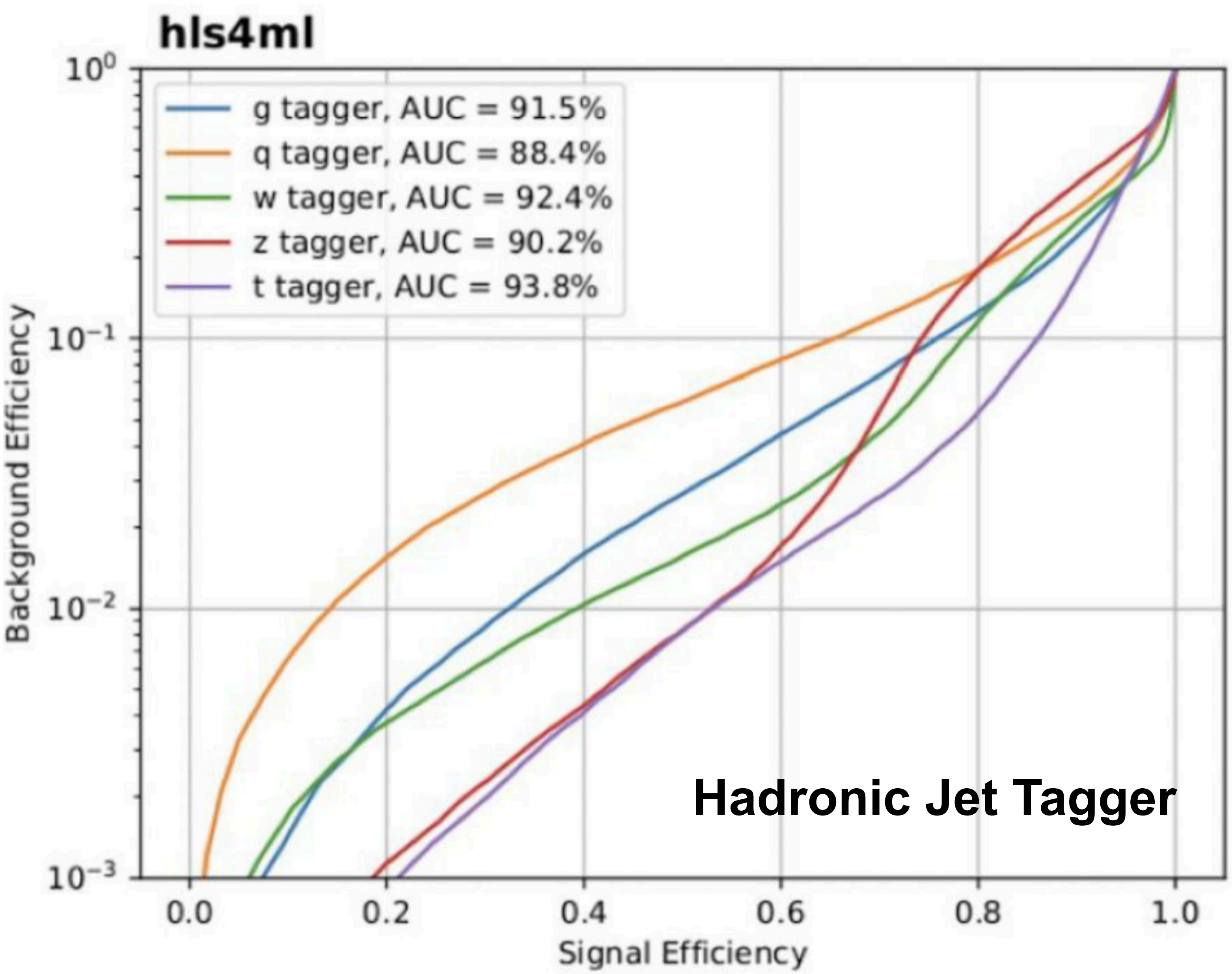
Matrix Mult in an FPGA



Results subject to precision outputs

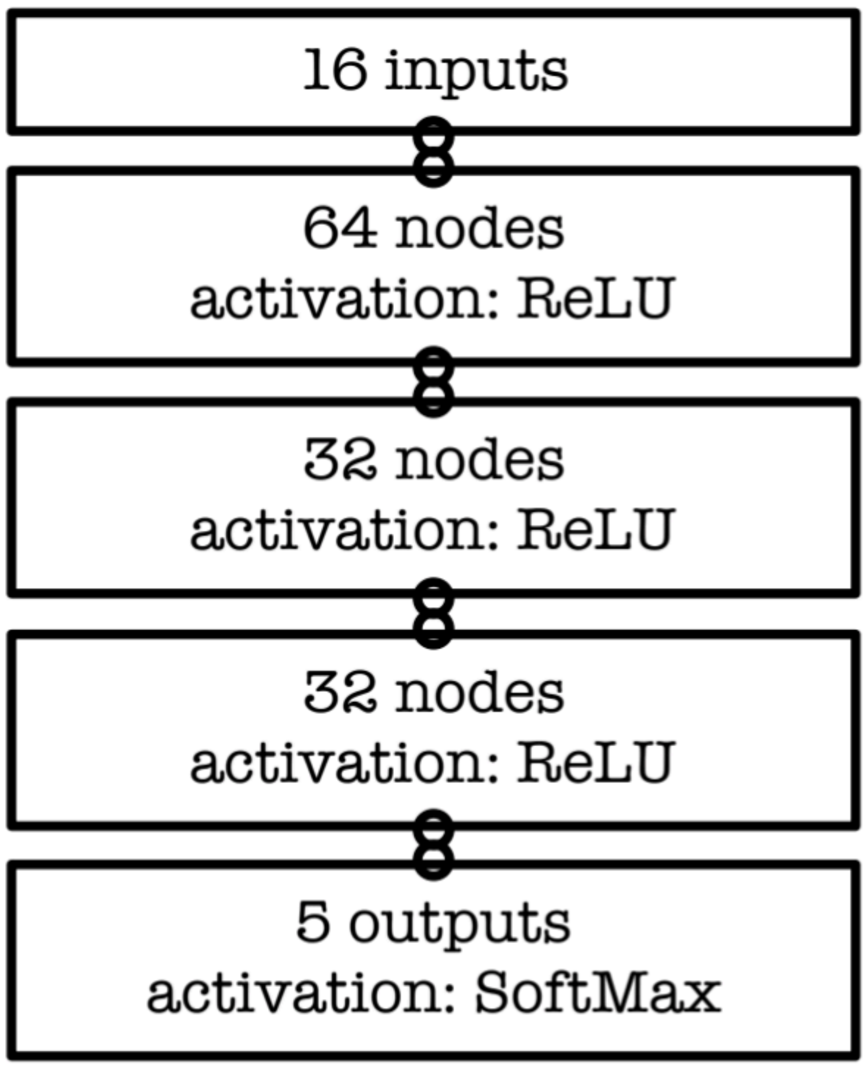
A full benchmark example

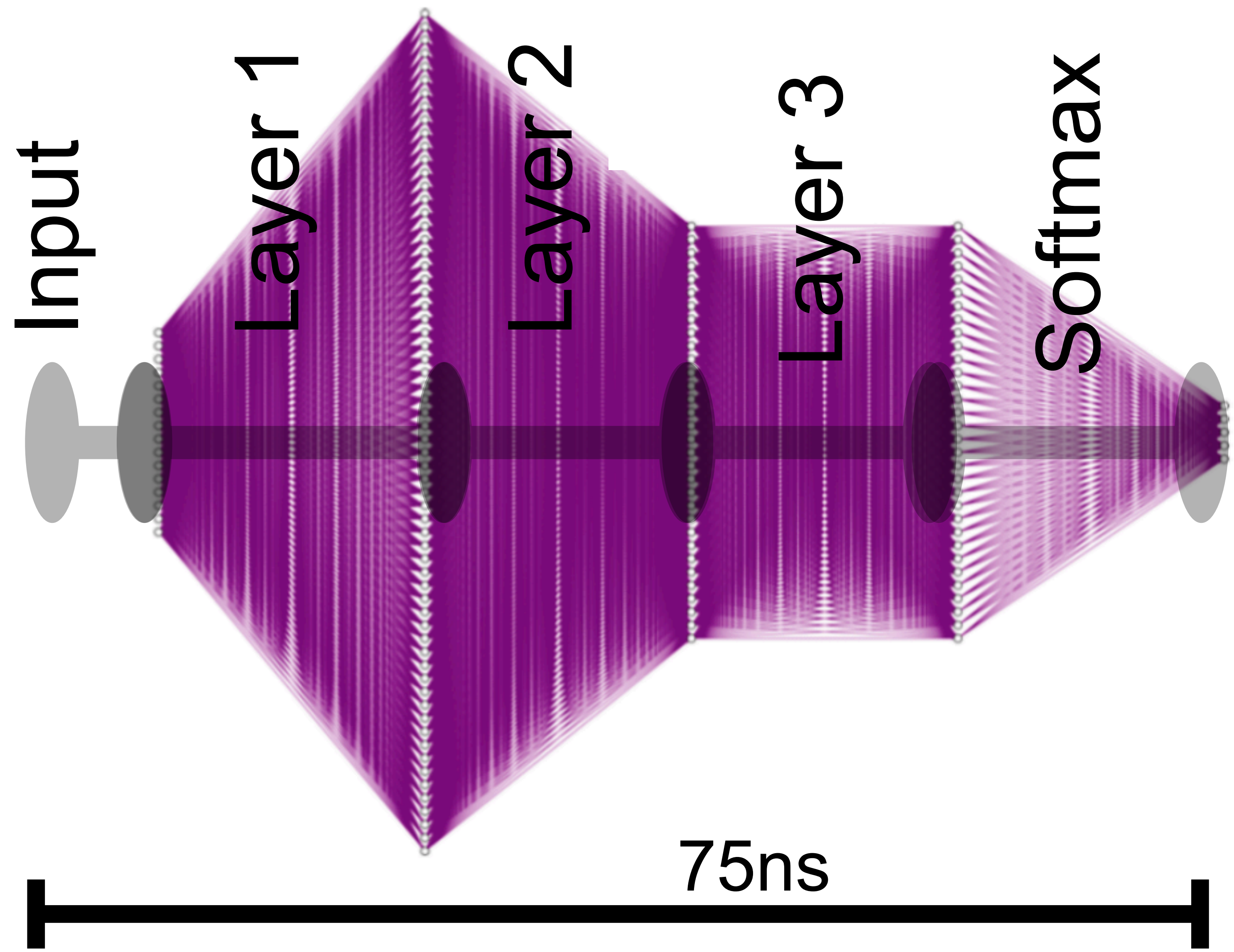




This network has an II of 1 clock, being run constantly
It has 4.3k weights and 4.3k DSPs at II=1

A full benchmark example





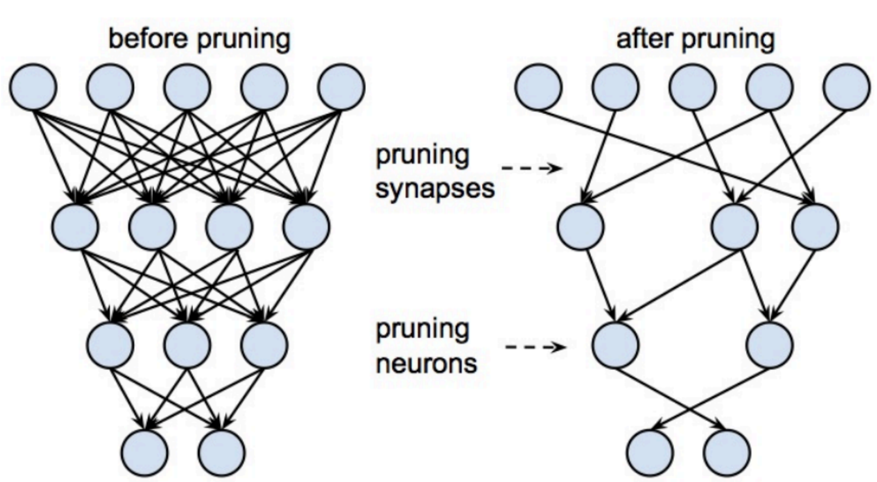
This network has an II of 1 clock, being run constantly
It has 4.3k weights and 4.3k DSPs at II=1

How can we reduce resources?

Focus on 3 ways to cut down resources

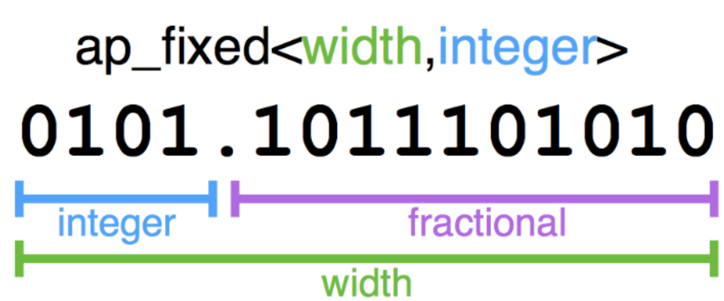
Is our algorithm overly complex?

Algorithmic Compression



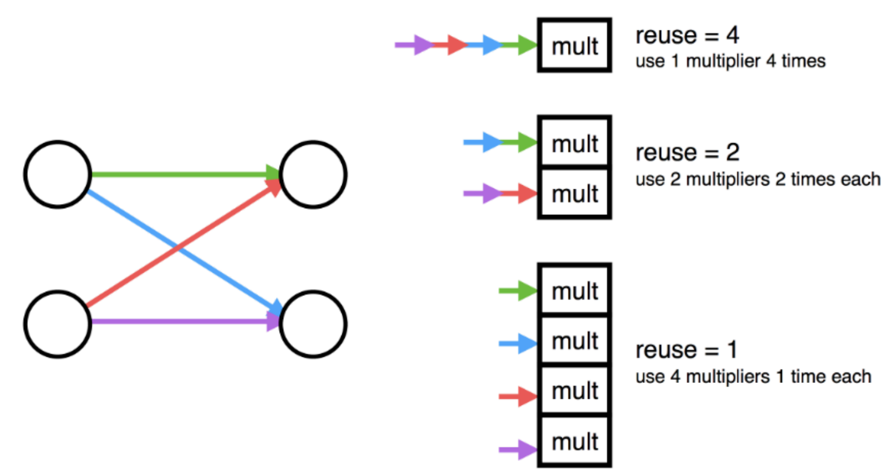
Are we too precise?

Quantization



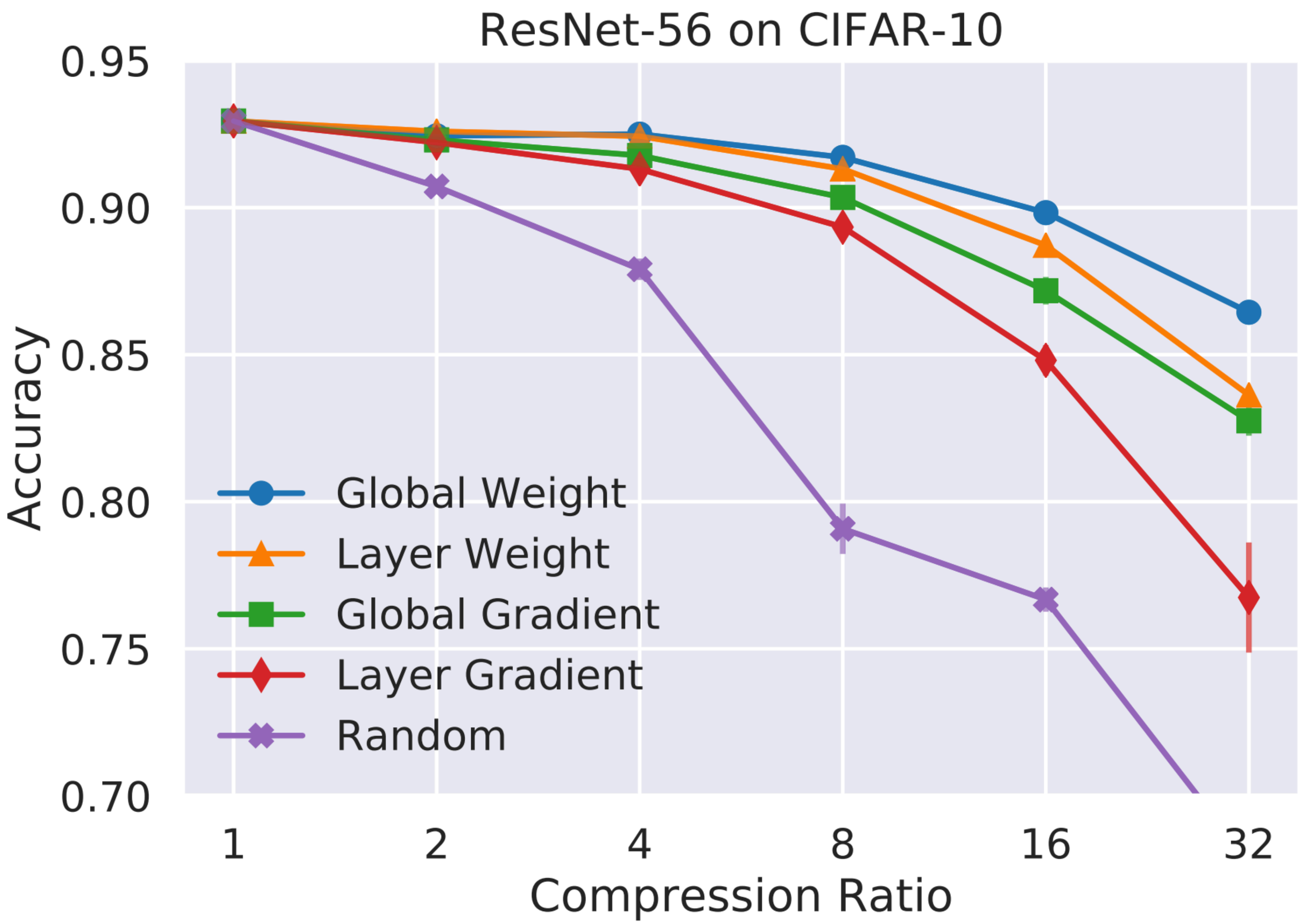
Does it really need to be this fast?

Reuse Factor



Algorithm Compression

- Compression is a critical aspect to reduce ML
- A suprising amount of weights in an NN are irrelevant

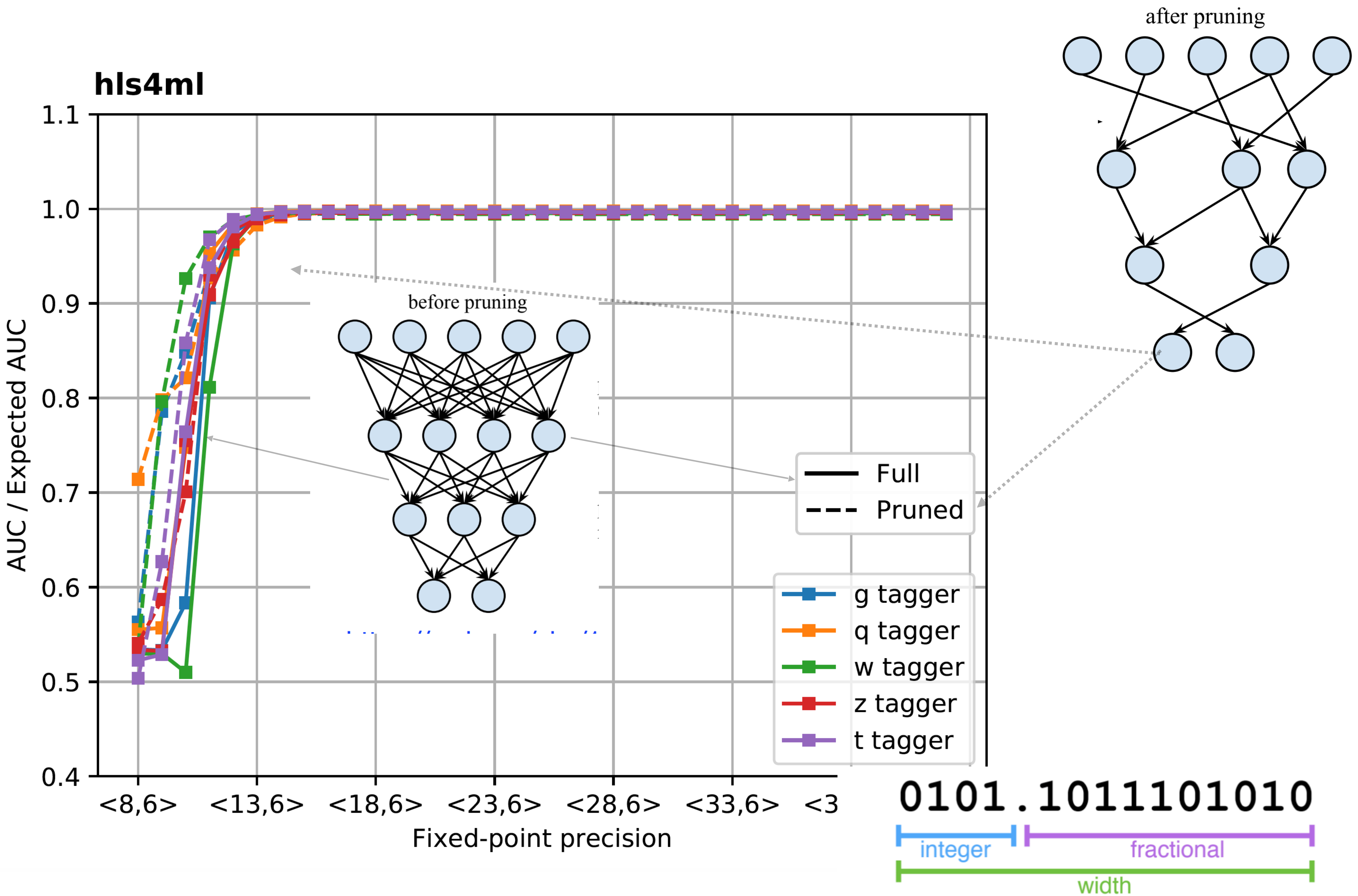


Model	Mult(DSP)	LUTs
Before	15%	13%
After	0%	1%

Same Performance
Smaller Latency (50→40ns)
Dramatic Compression

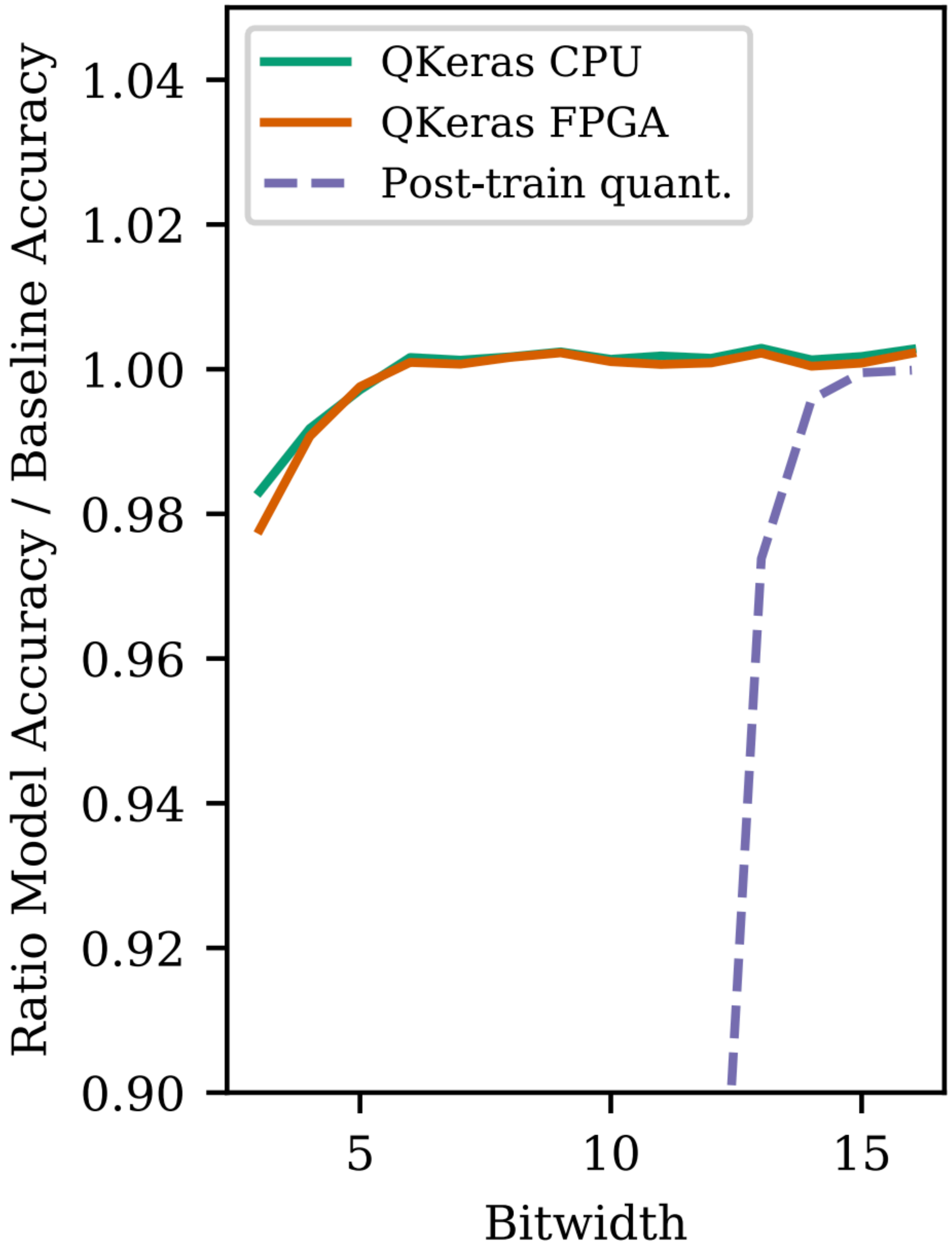
arxiv:1804.06913

Quantization



<Total bit width, integer bits above decimal>

Algorithm Compression



Fixed precision training
Weight pruning shrinks networks

arxiv:2103.05579

A Compiler than can do it

A Deep Neural Network Of Your chocie (mostly)!

→

Conversion tool

hls4ml

→

Specs for an FPGA

Latency/ Resources/ Precision/ Configure

→

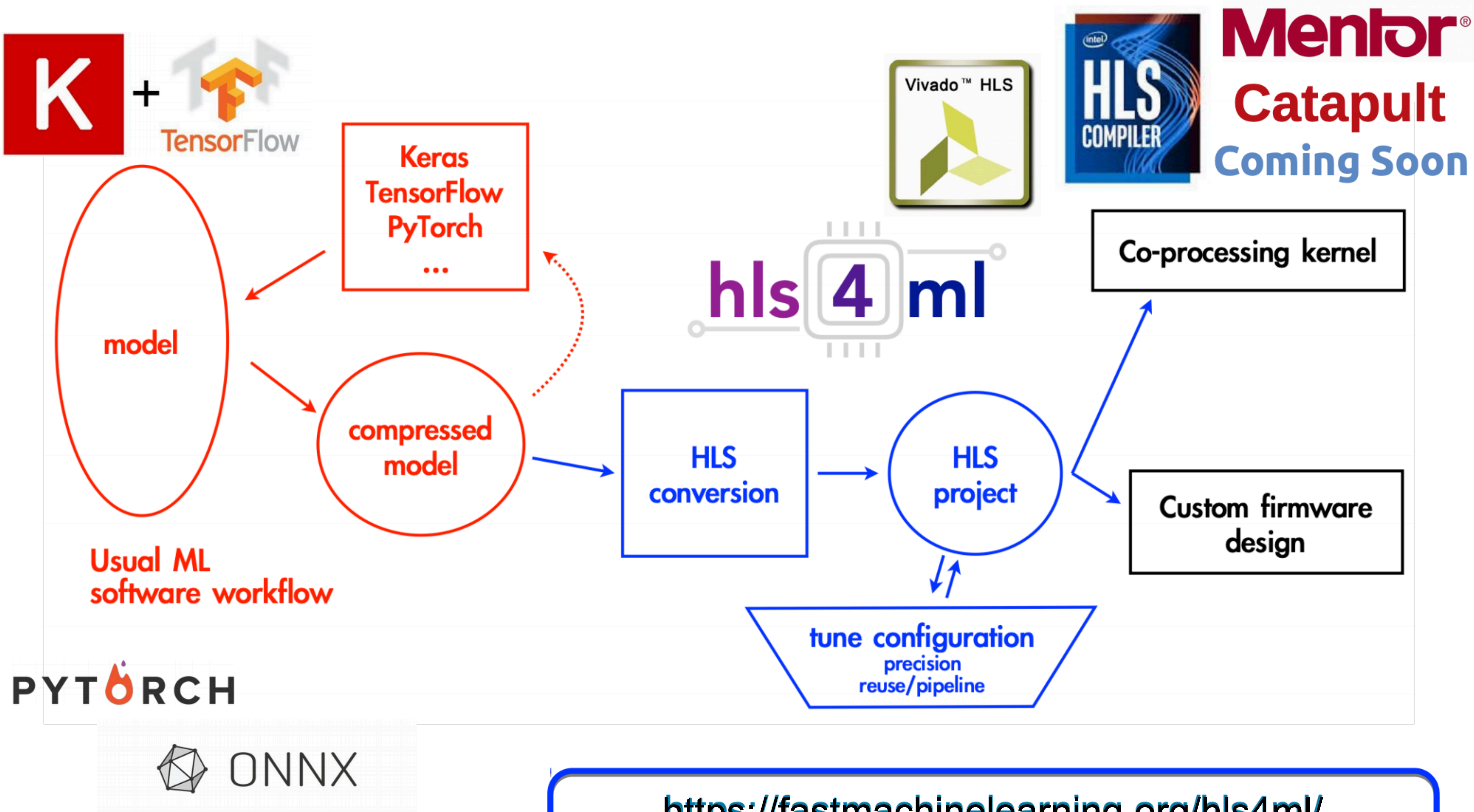
The Code to Run it

There are now a few tools
See Tae Min's Talk for another tool!

https://fastmachinelearning.org/hls4ml/

Summing Up the Data flow

```
python keras-to-hls.py -c keras-config.yml
```



https://fastmachinelearning.org/hls4ml/

Flexibility

- Many different types of collisions are analyzed at LHC
  - A diverse set of algorithms are required
  - There is no one size fits all NN that will solver our problems
- With HLS4ML we have continued to expand options
  - HLS has allowed for quick development

Algorithms

MLPs
CNNs
RNNs(LSTM/GRU)
Binary & Ternary NNs
Graph NNs(MPNN/GravNet/GarNet)
BDTs

arxiv:2003.06308
arxiv:2002.02534
arxiv:2008.03601
arxiv:2006.10159

Not yet in official release

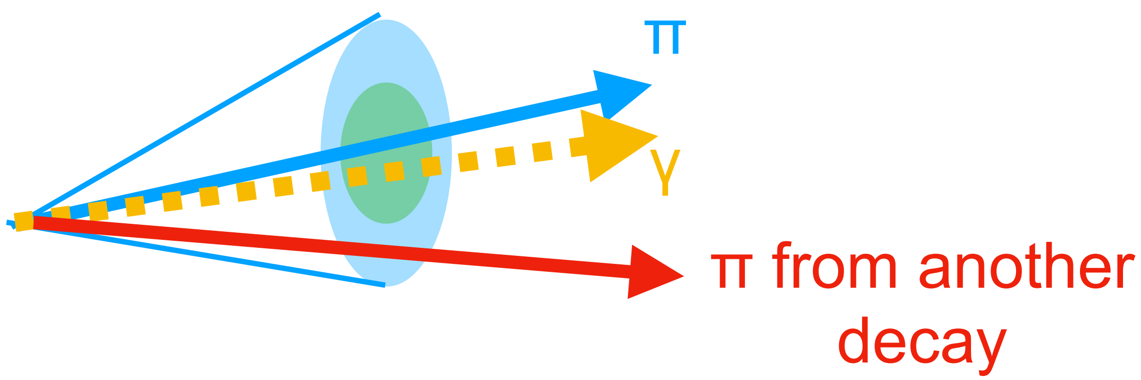
Backends

Xilinx Vitis HLS
Intel HLS Quertus
Mentor Catapult HLS
Intel OneAPI

Not yet in official release

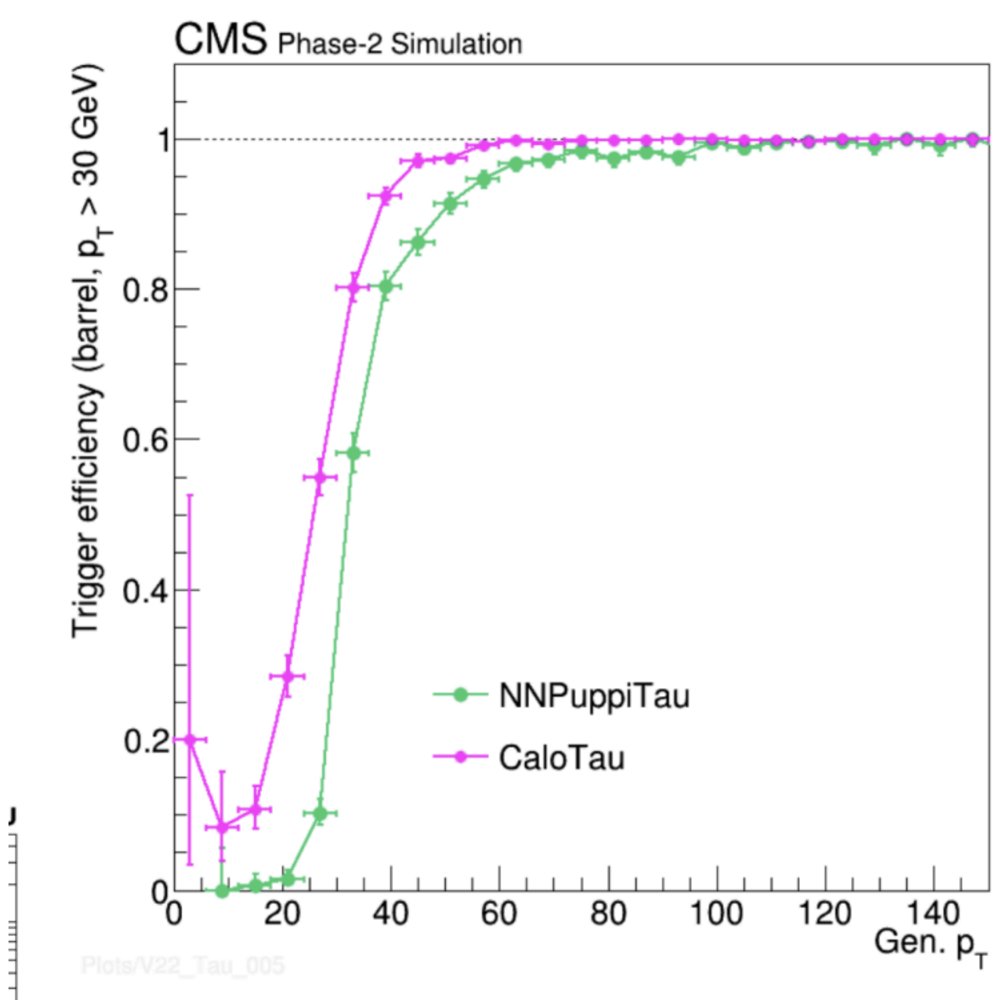
Example #1 Tau Tagging

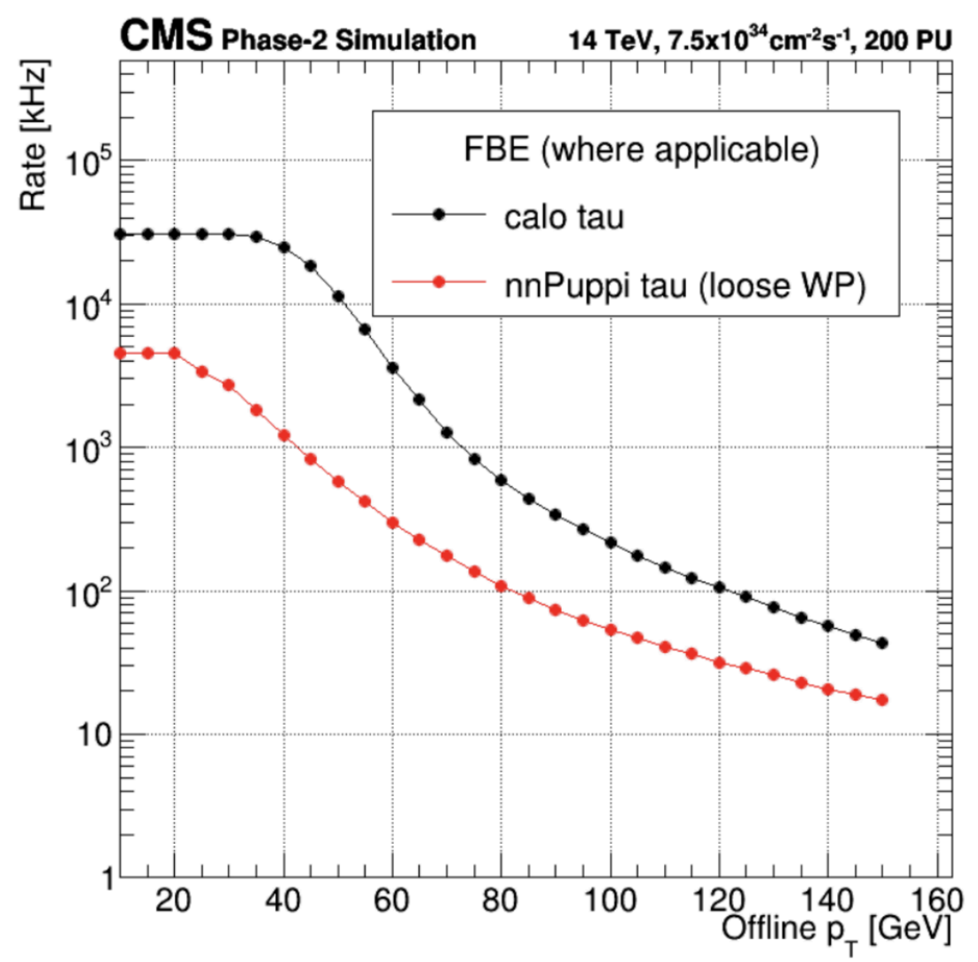
Tau Leptons have complex final states



Tau Lepton can decay to as many as 10 different particles
Background can decay to many more

Neural Network has long been the algorithm of choice to identify Taus

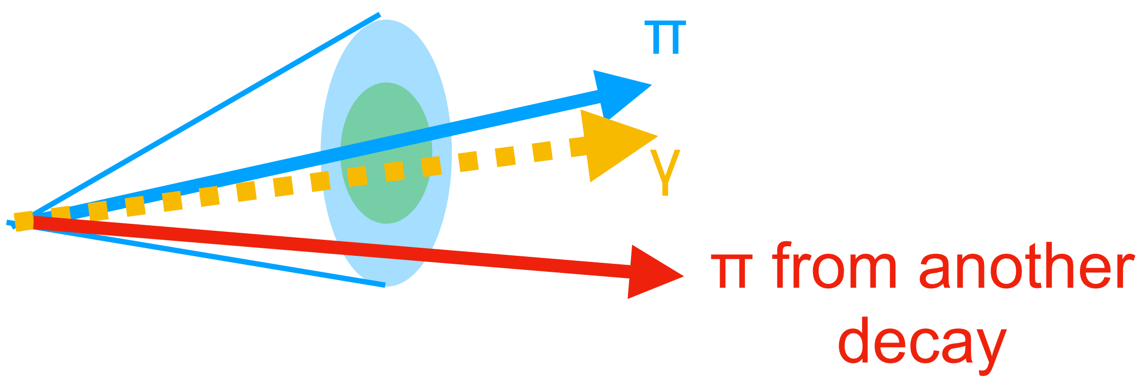




Example #1 Tau Tagging

Algorithm Takes 10 top particles in a cone and runs NN

With HLS4ML we can run this algorithm in 70ns

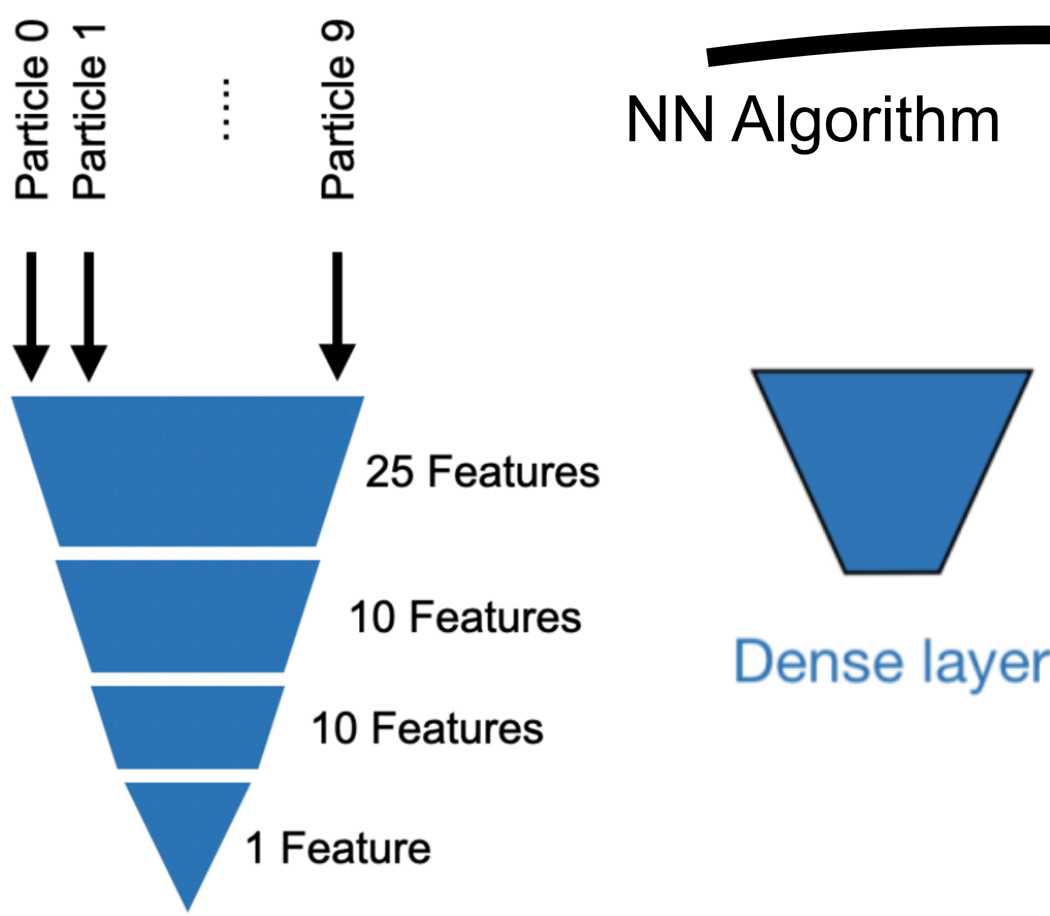


VU9P	DSP	FF	LUTs	BRAM
NNTau	11%	12%	18%	16%

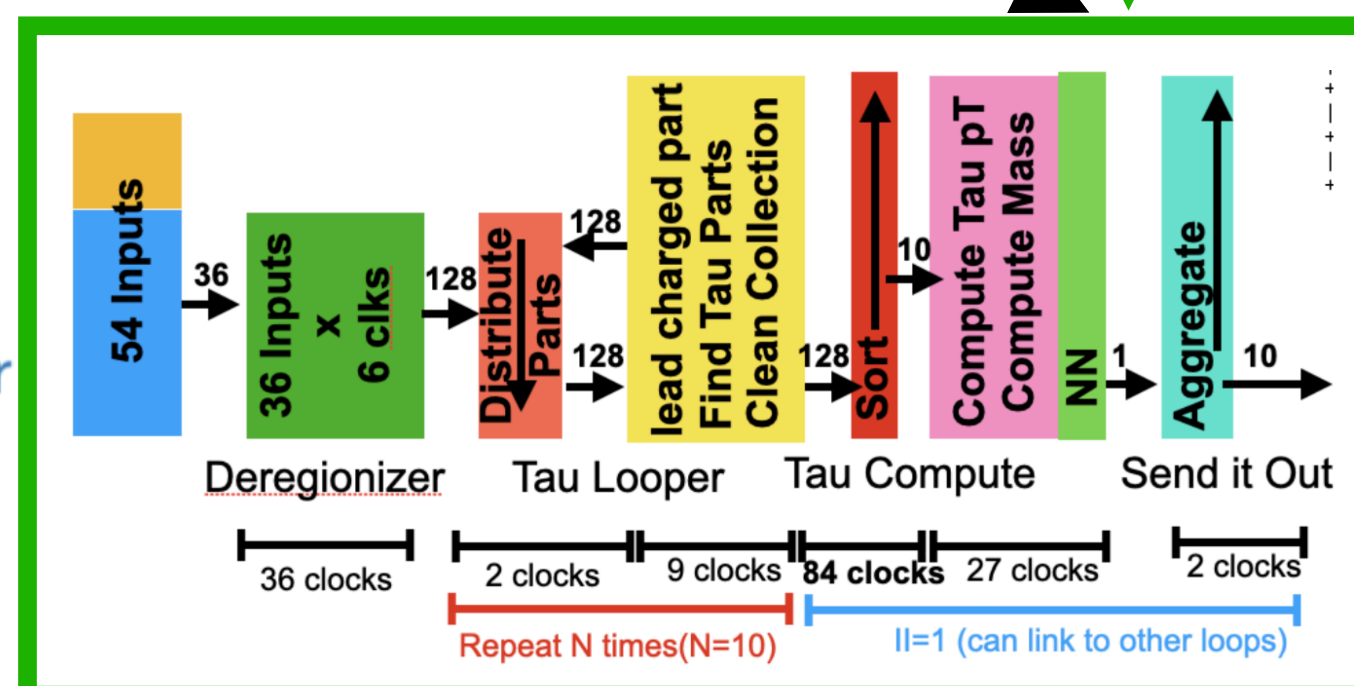
NN alone is <10% of the FPGA

Whole Algorithm Resources

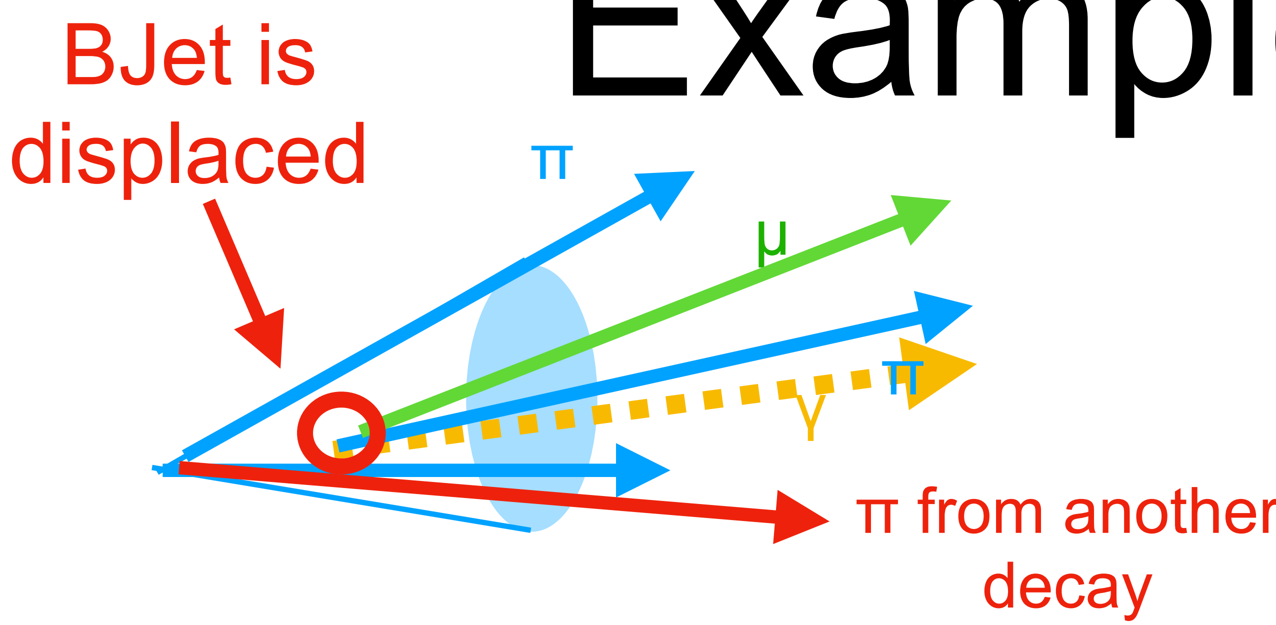
NN Algorithm



Whole Algorithm on Board

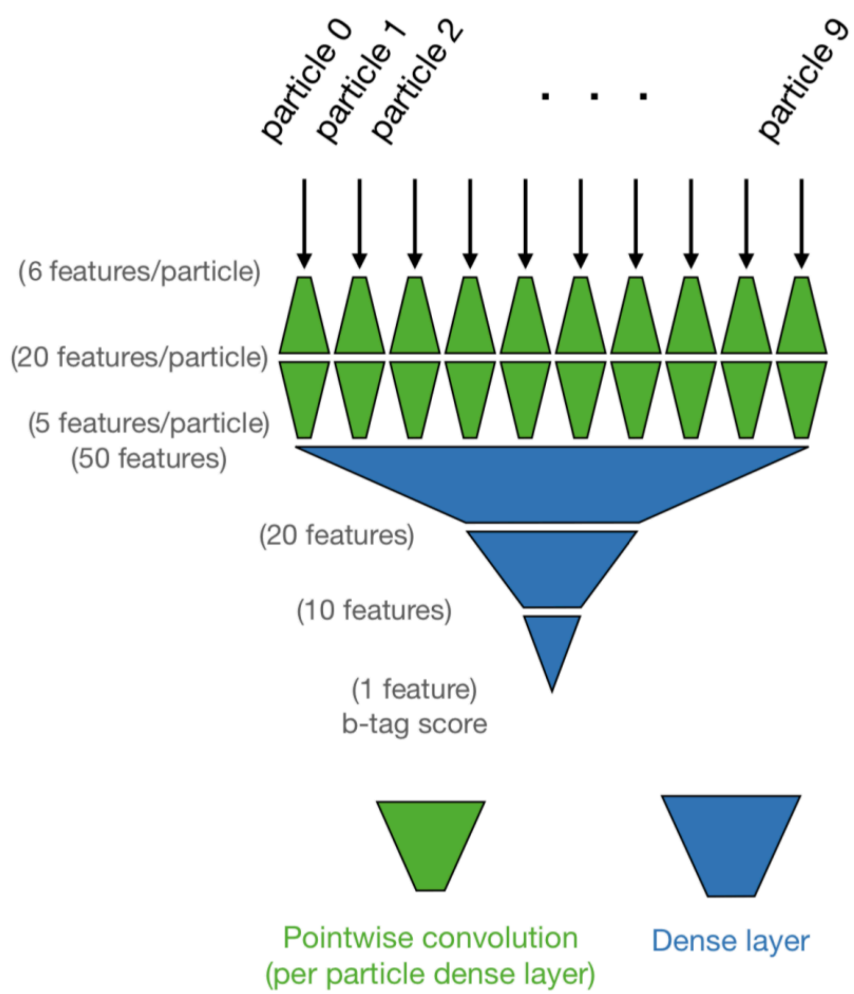


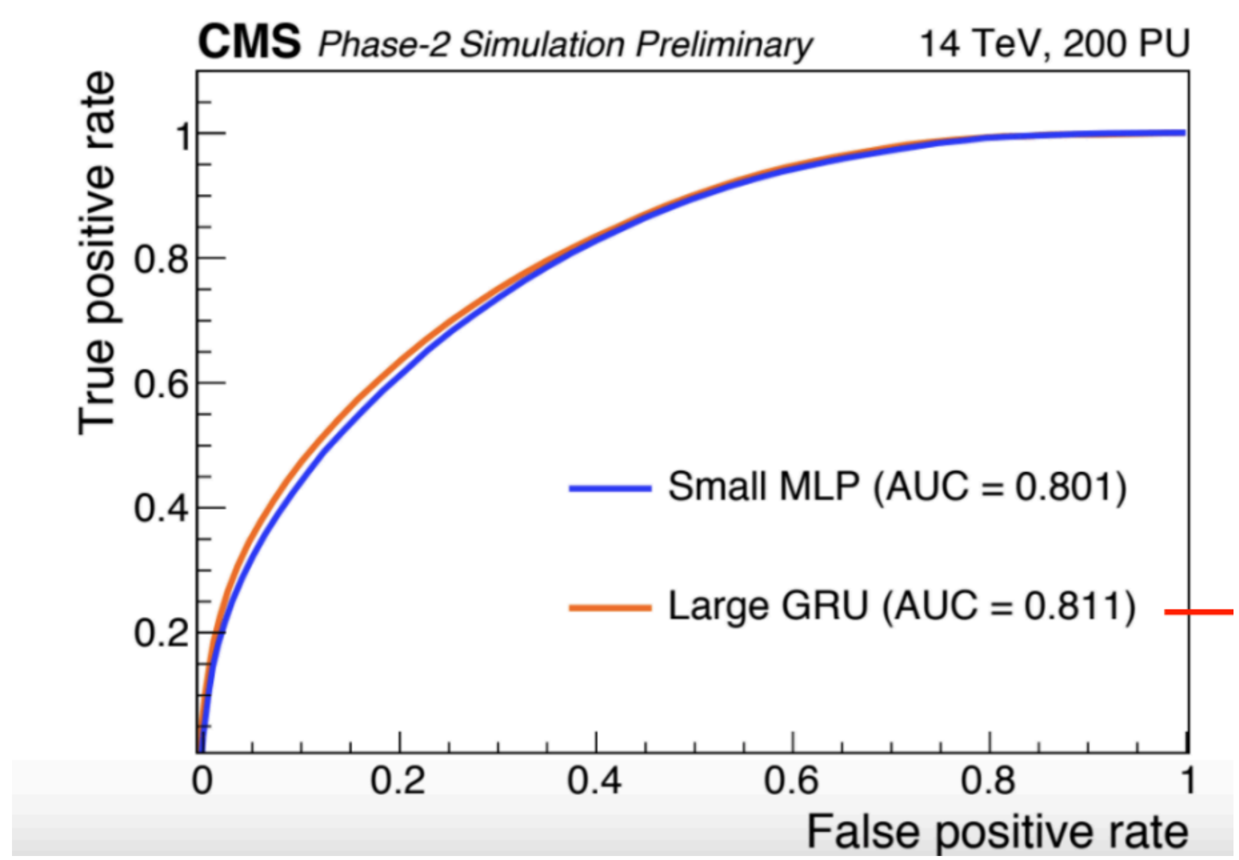
Example #2 BTagging



In addition to taus
B-tagging good ML candidate

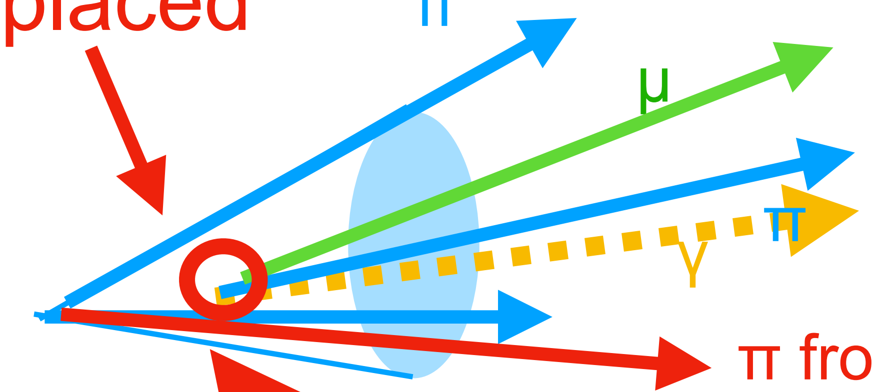
Not obvious CMS Trigger
vertex resolution is large





Example #2 BTagging

BJet is displaced

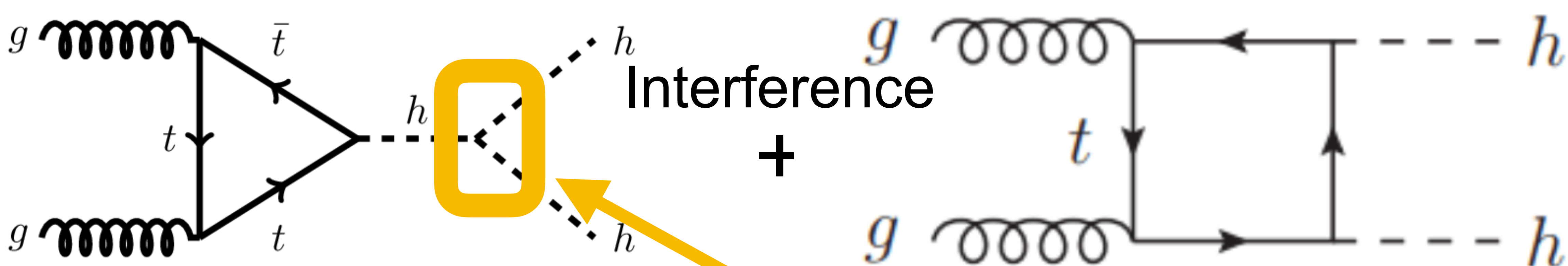


Hard

Resolution in Trigger is worse

In addition to taus
B-tagging good ML candidate

Not obvious CMS Trigger
vertex resolution is large

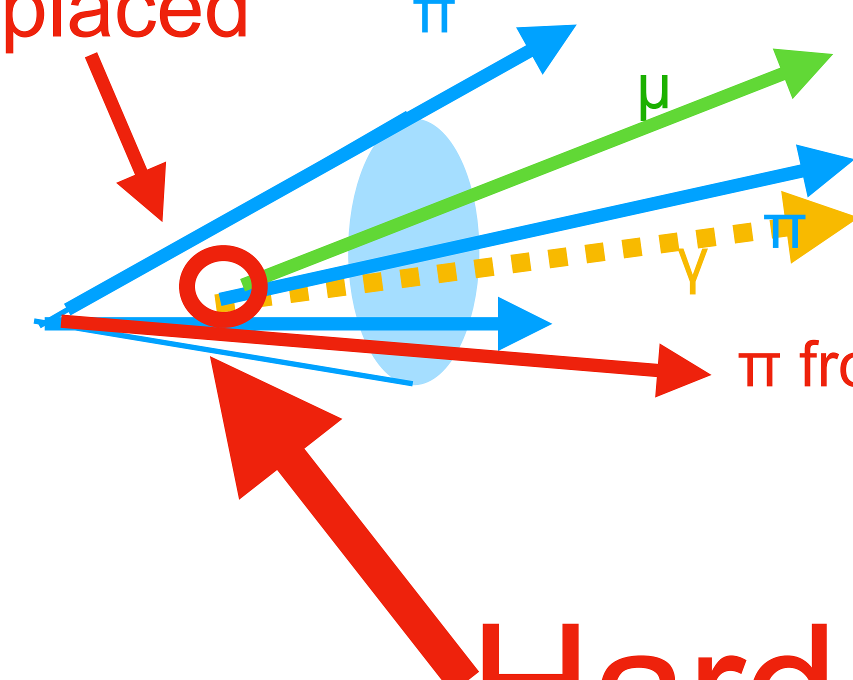


Di Higgs Boson Production

Higgs Self Coupling Term

Example #2 BTagging

BJet is displaced



π from another decay

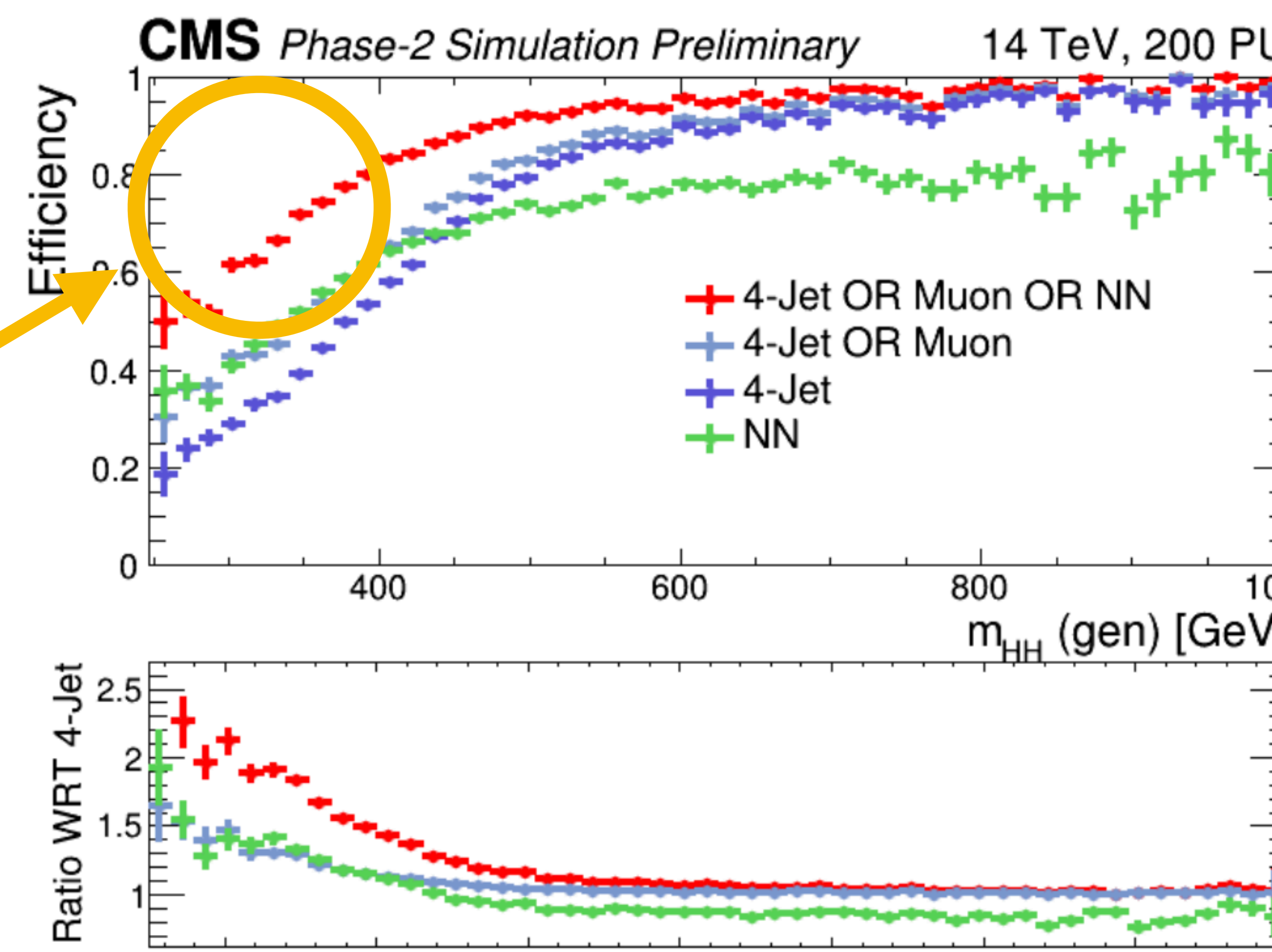
Hard

Resolution in Trigger is worse

In addition to taus
B-tagging good ML candidate

Not obvious CMS Trigger vertex resolution is large

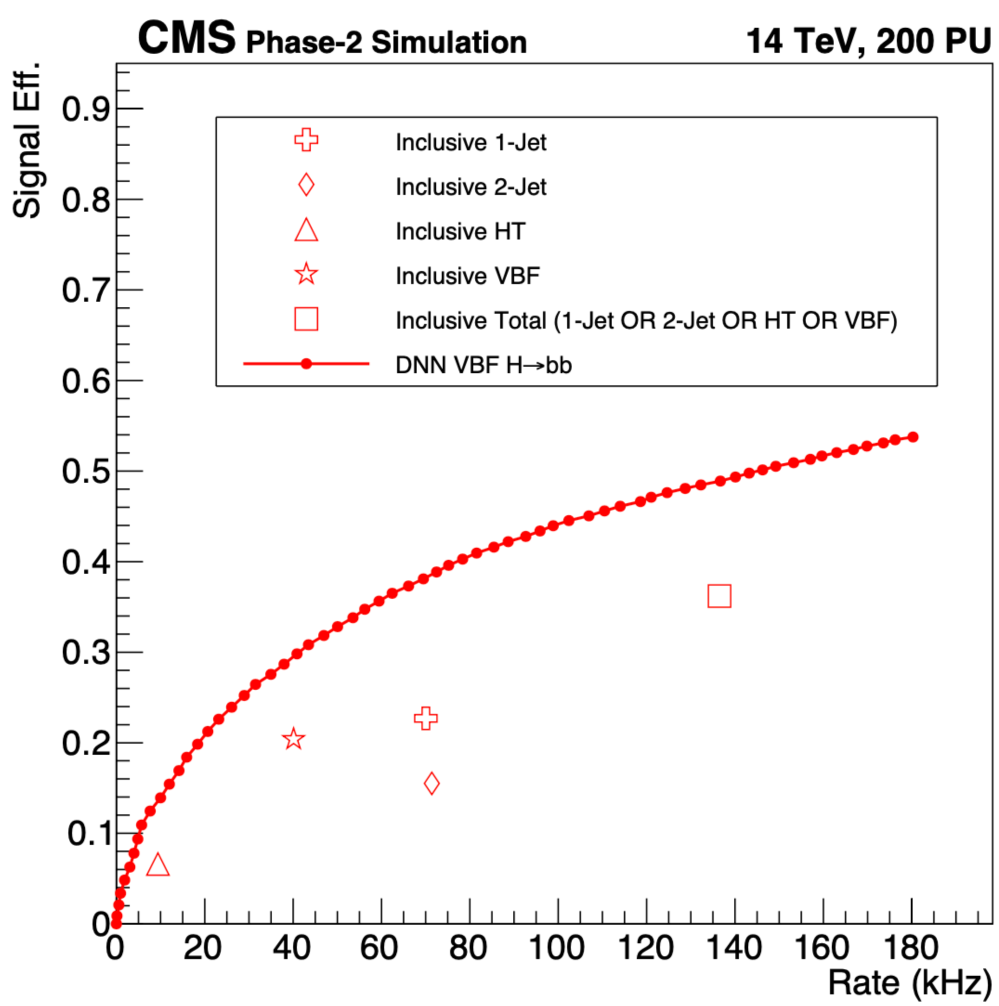
Critical Region
For Self Coupling

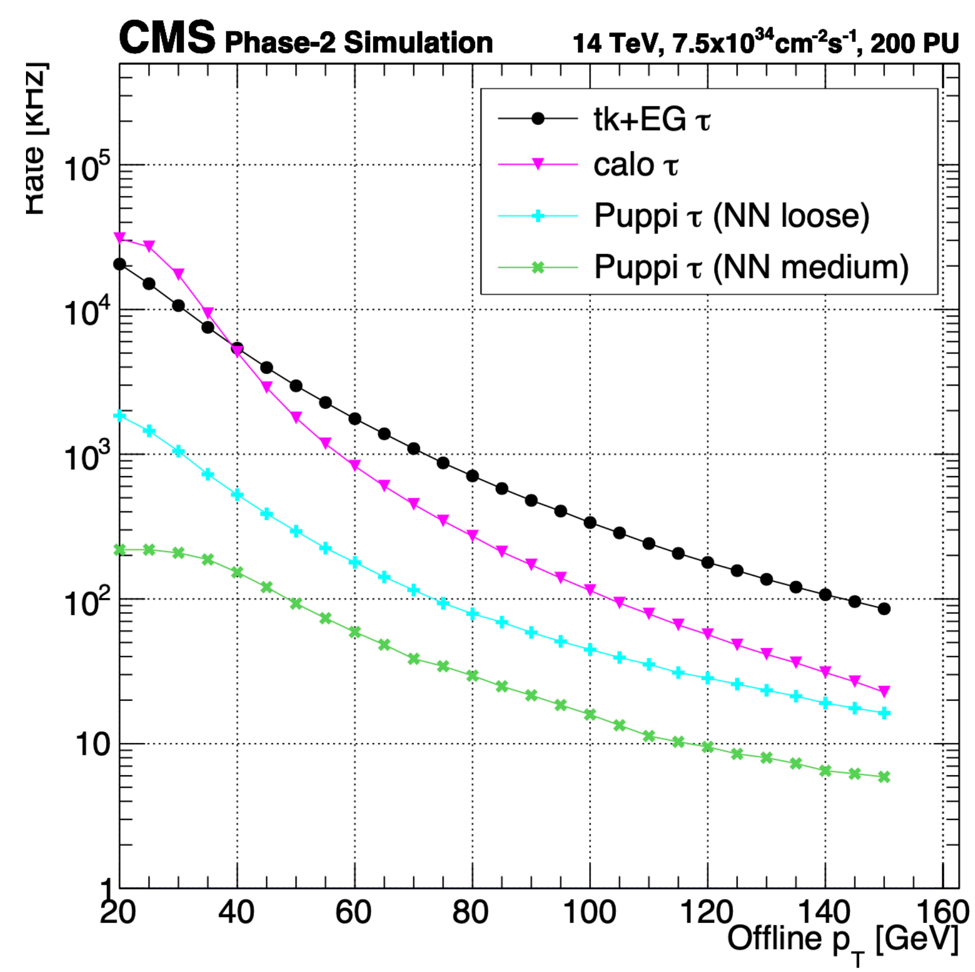


Accomplishments

- HLS4ML is rapidly being adopted in our trigger system
  - Will be used in the next running at the LHC
- We already see a number of substantial improvement

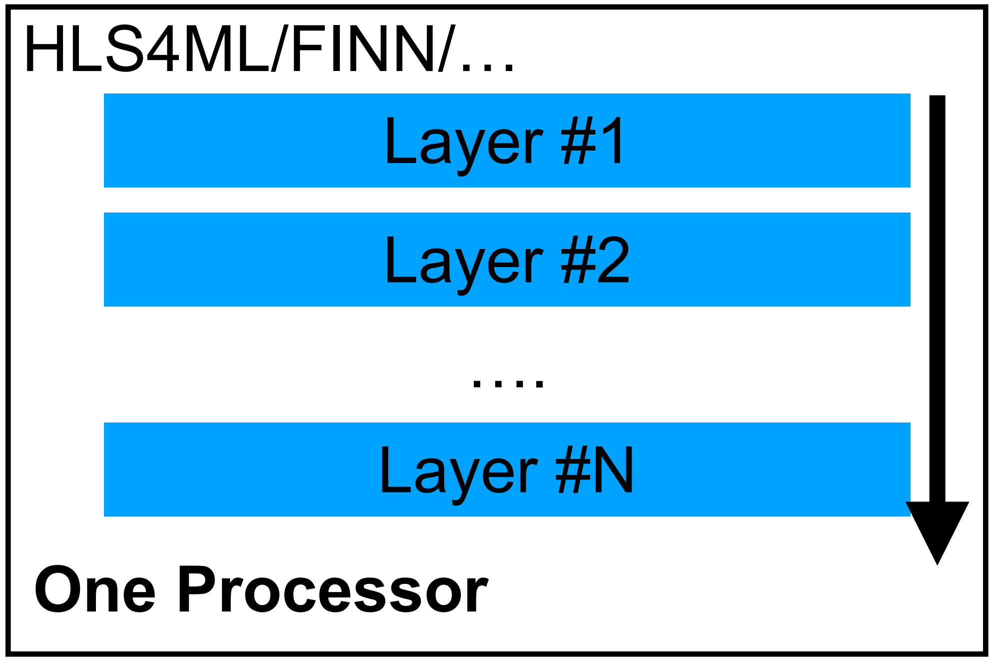
2-5 times More Higgs bosons with the same data rates



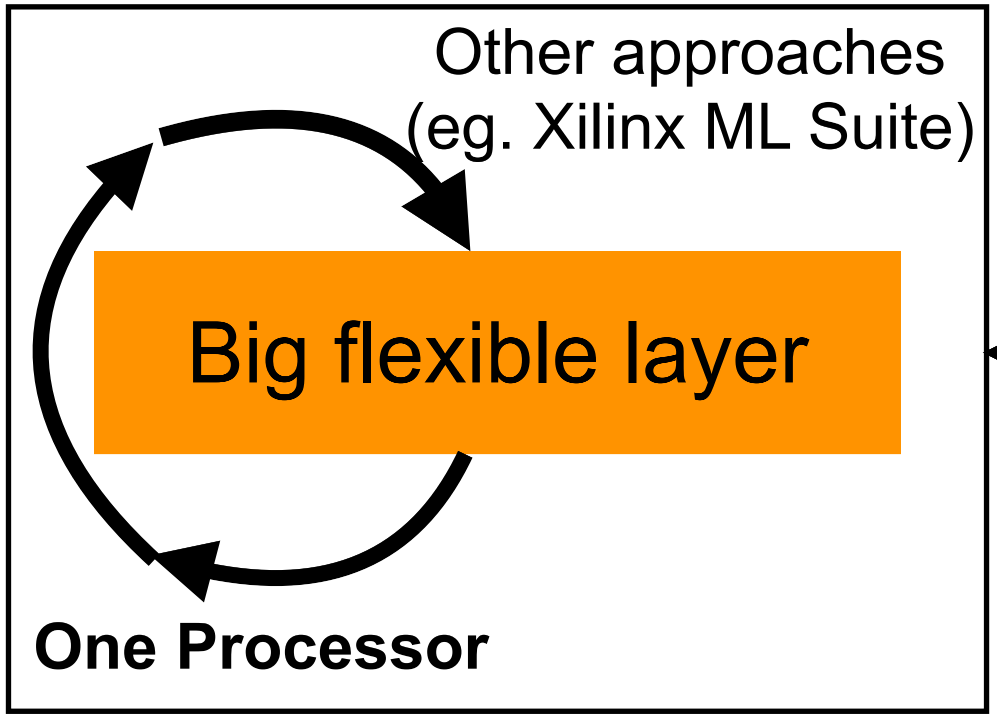


Other Deep Learning Models

- HLS4ML differs from other ML models



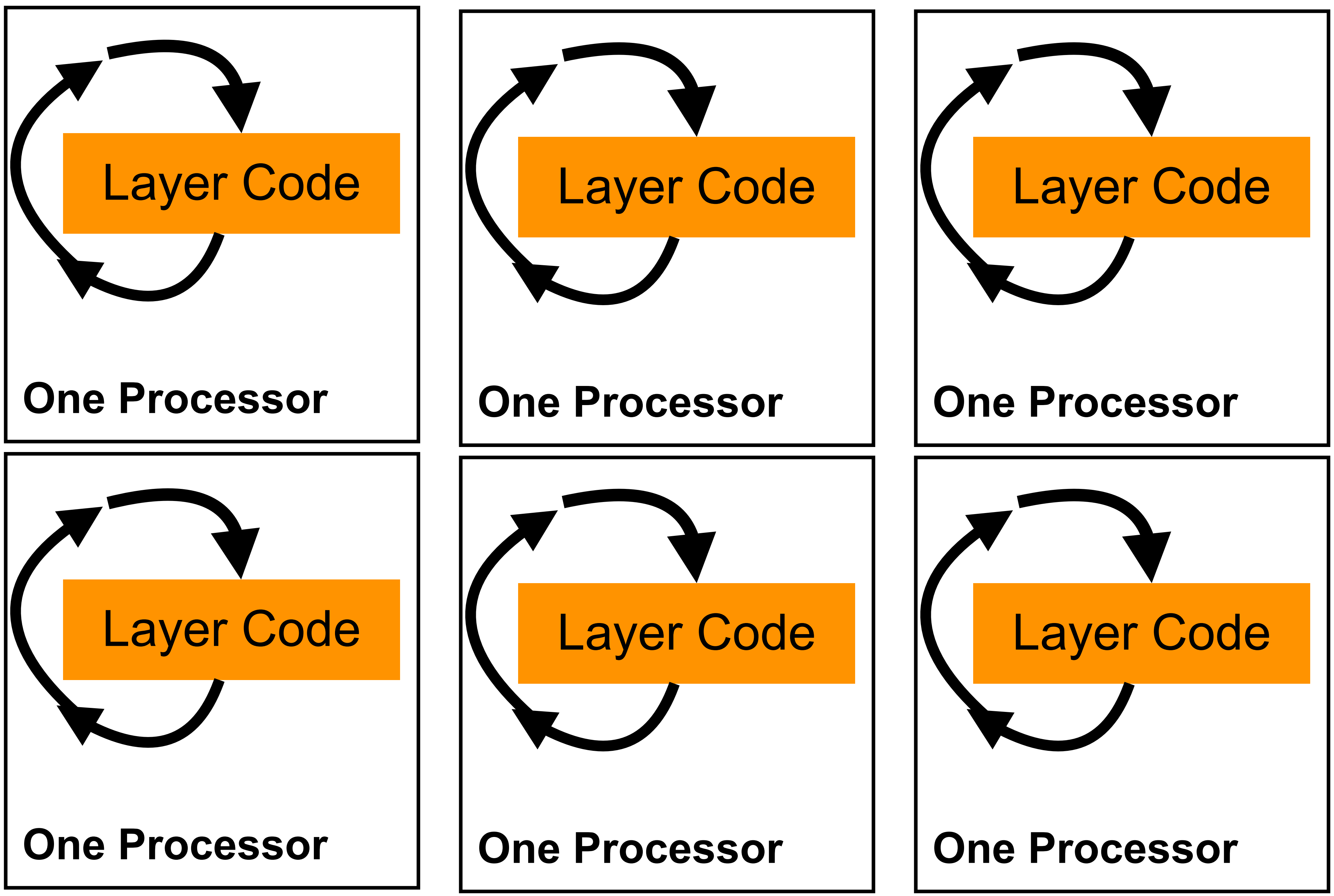
Good for small models where you need ultra low latency and ultra high throughput



Good for very large models where you can't fit the whole algorithm on the processor logic

How does a GPU do this?

- GPU is about even more standardization

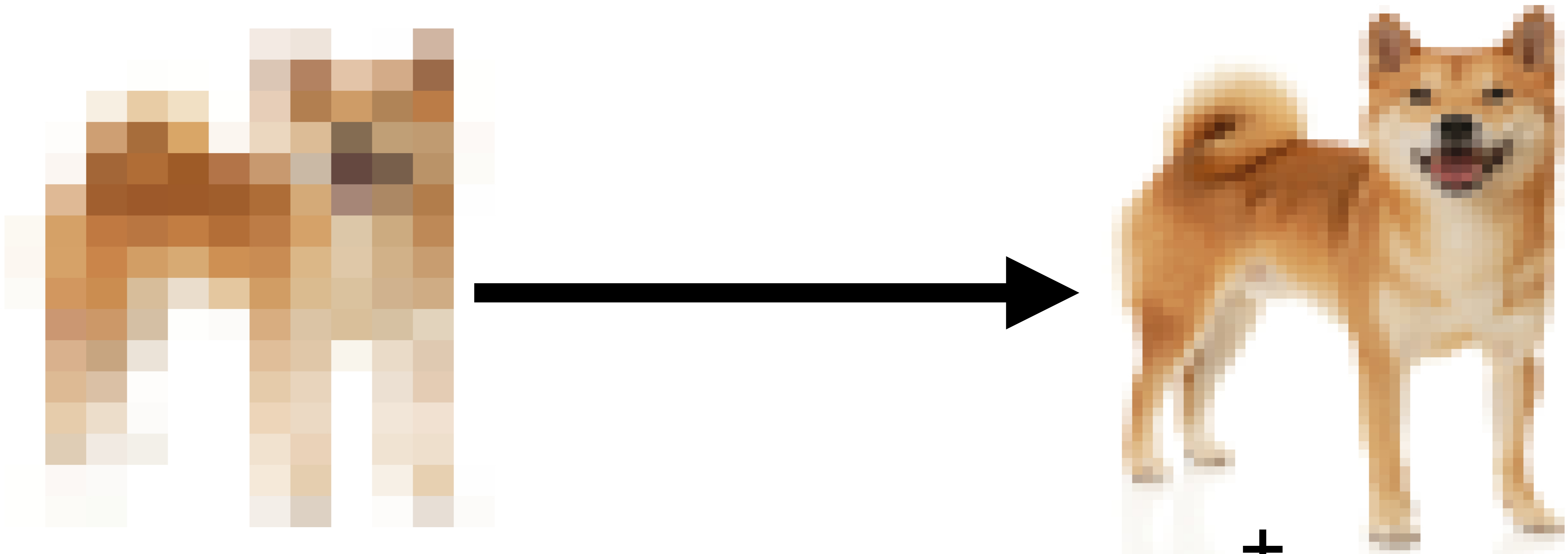


Great for many many evaluations of a big network

Not Great for a small network

.....

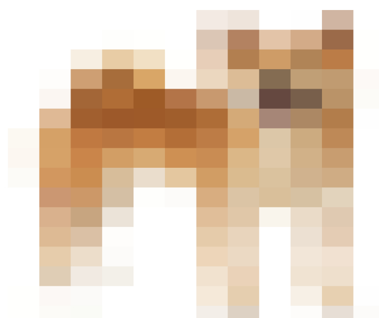
.....



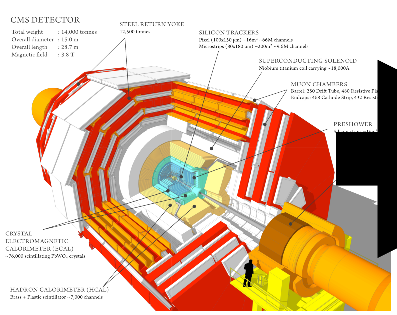
Running @
Longer latencies

HLT Trigger+Offline Reco

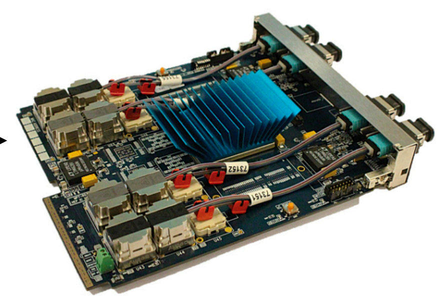
40 MHz **1 kHz**

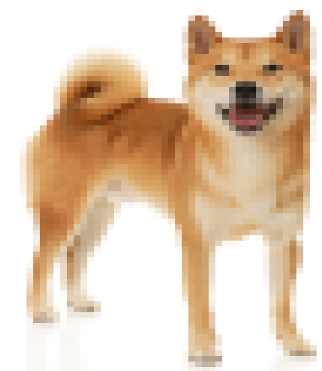


Radiation
Hard ASICs



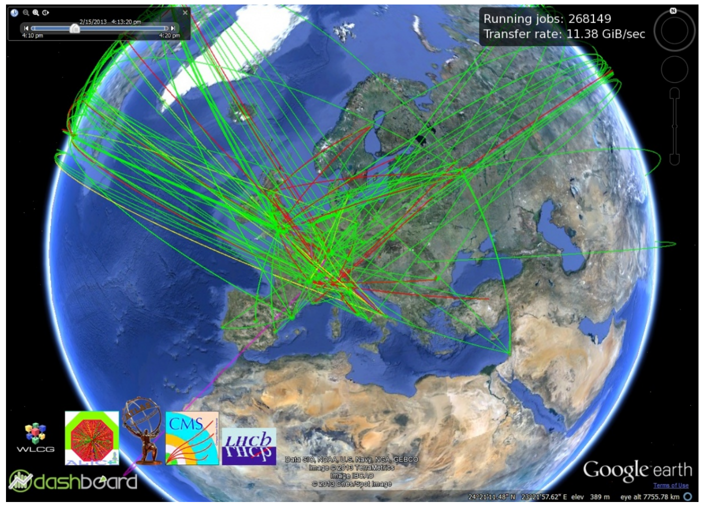
FPGA
Boards





Local CPU
Cluster

CPU Grid

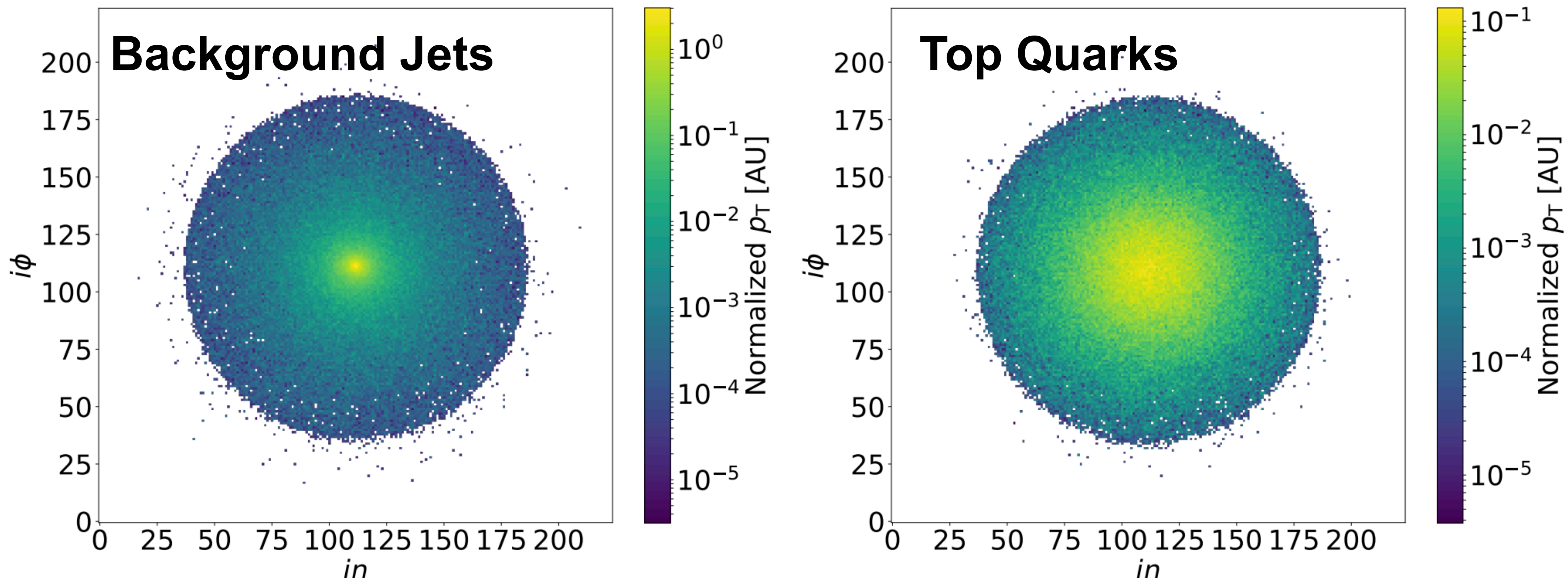


J. Krupa

D. Rankin

Both Tiers are CPU
milar algos(different scales)

What we learned?



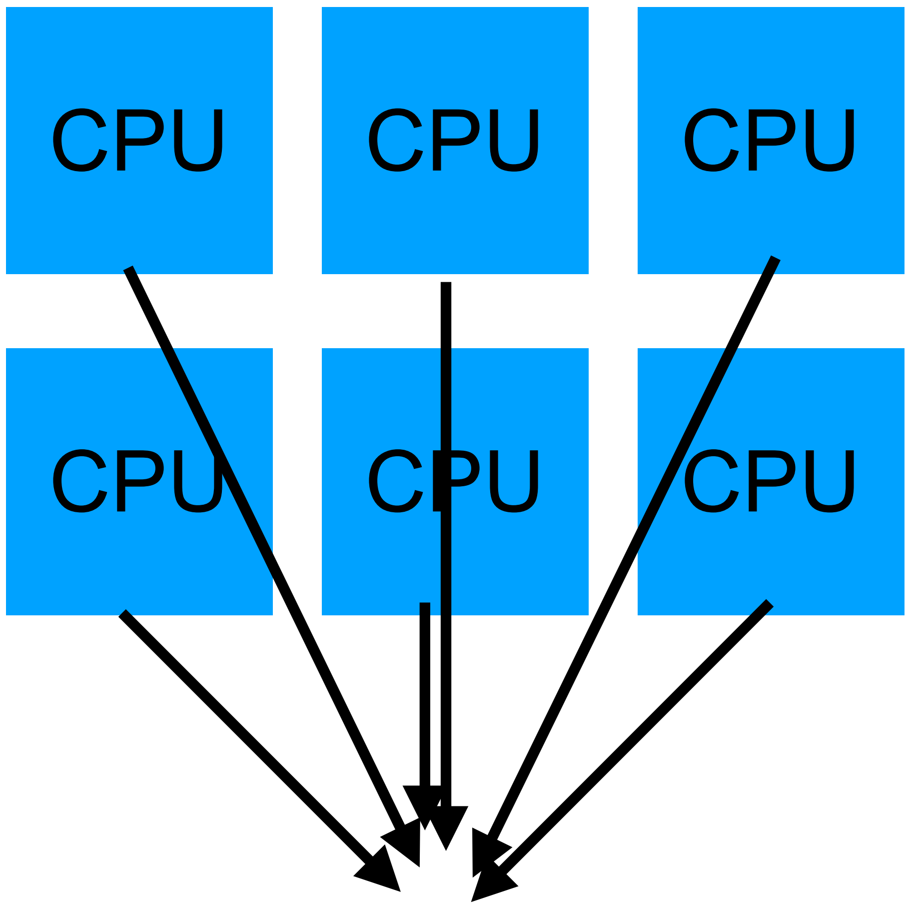
CNN using Resnet50 Arch

Algo	Per Event
CPU	1.75s
GPU Batch 1	7ms
GPU Batch 32	2ms
FPGA	1.7ms

1000x

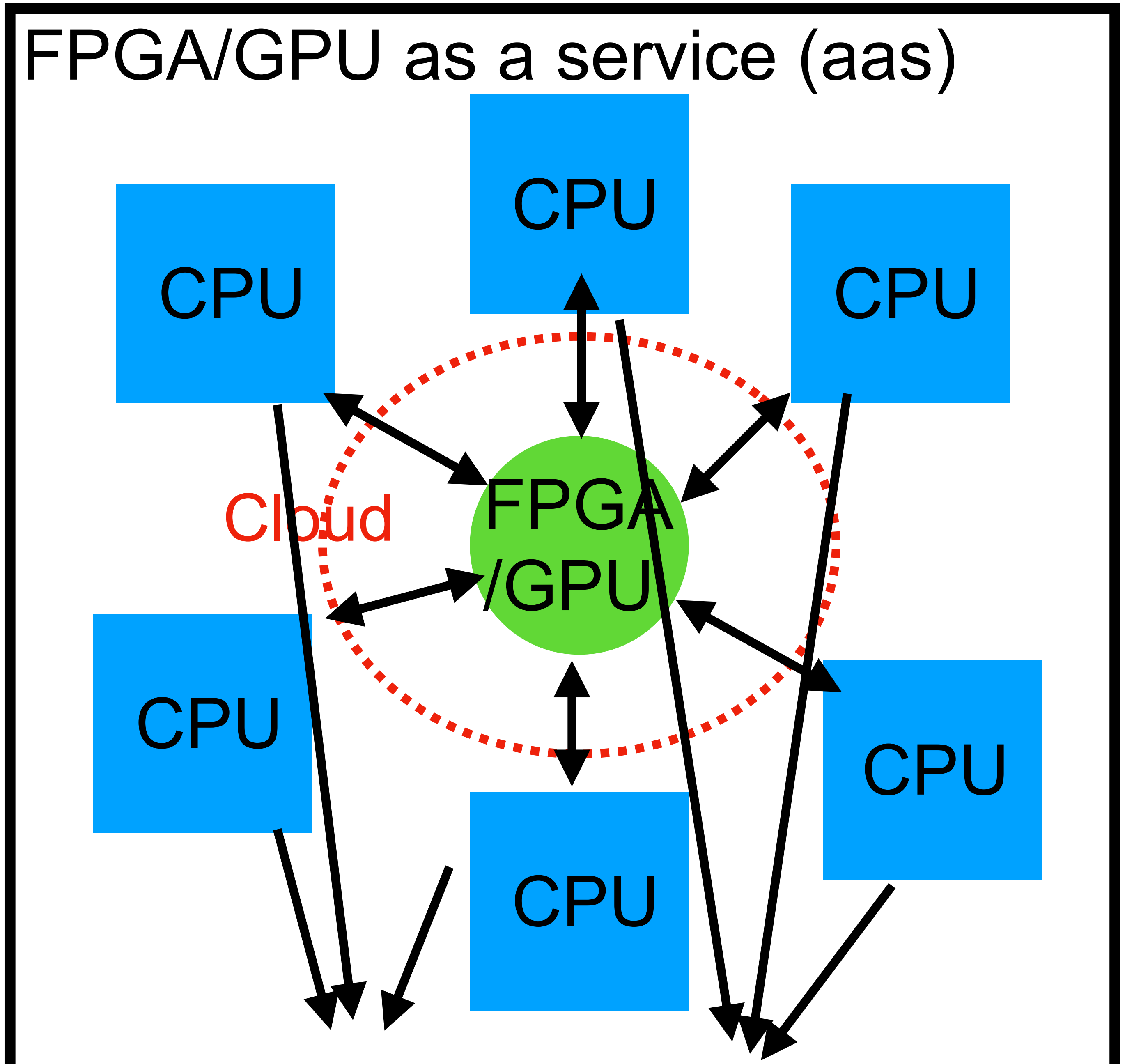
arxiv:1904.08986

What does this mean?



Process event by event

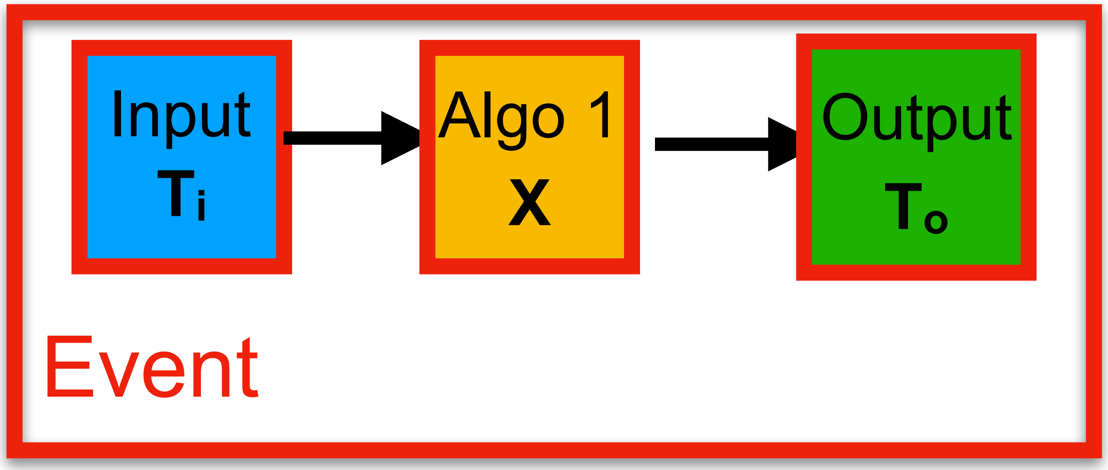
arxiv:1904.08986



Process (event by event)?
outsource specific algorithms

Deploying on a GPU

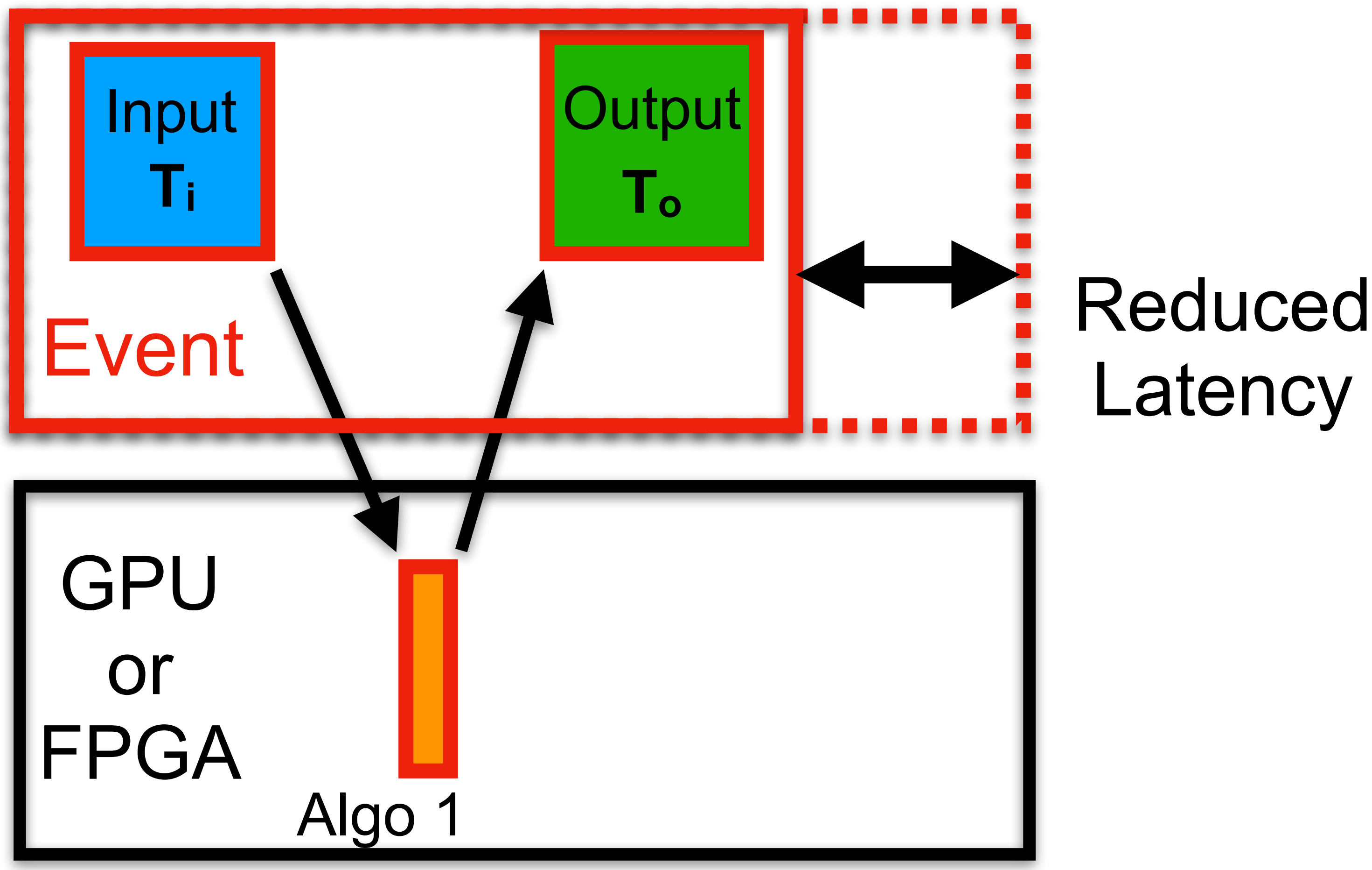
Process event by event



arxiv:1904.08986

Deploying on a GPU

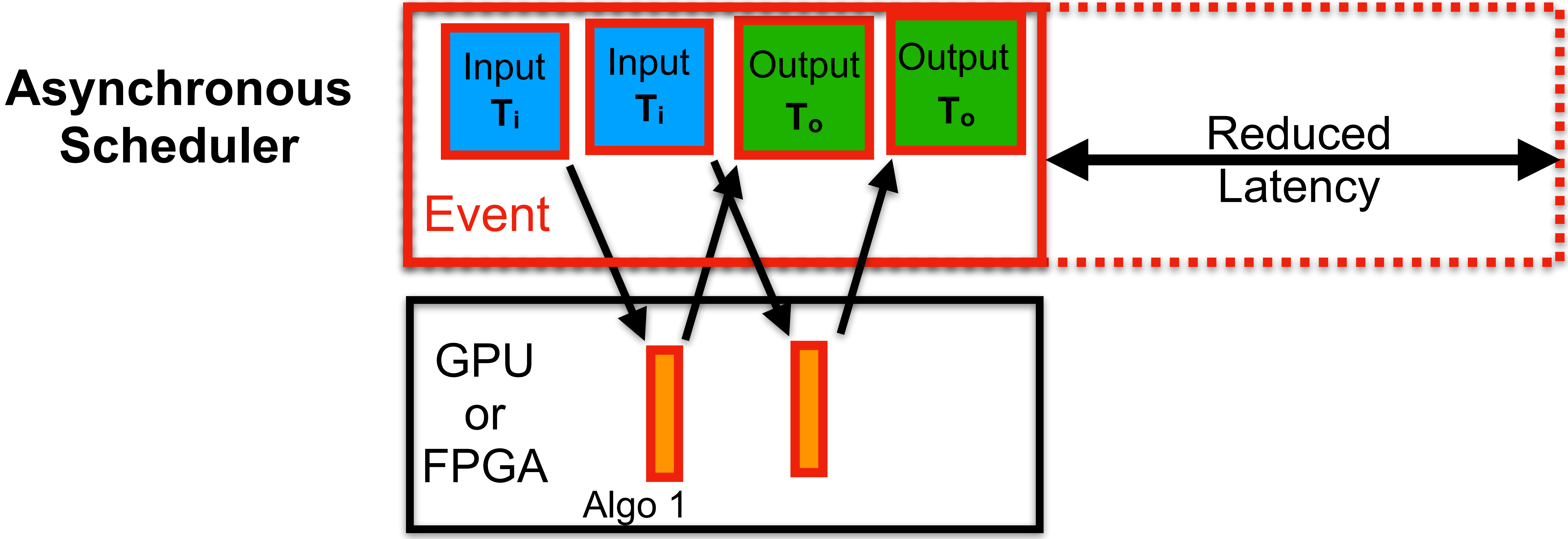
Process event by event



arxiv:1904.08986

Deploying on a GPU

Process event by event



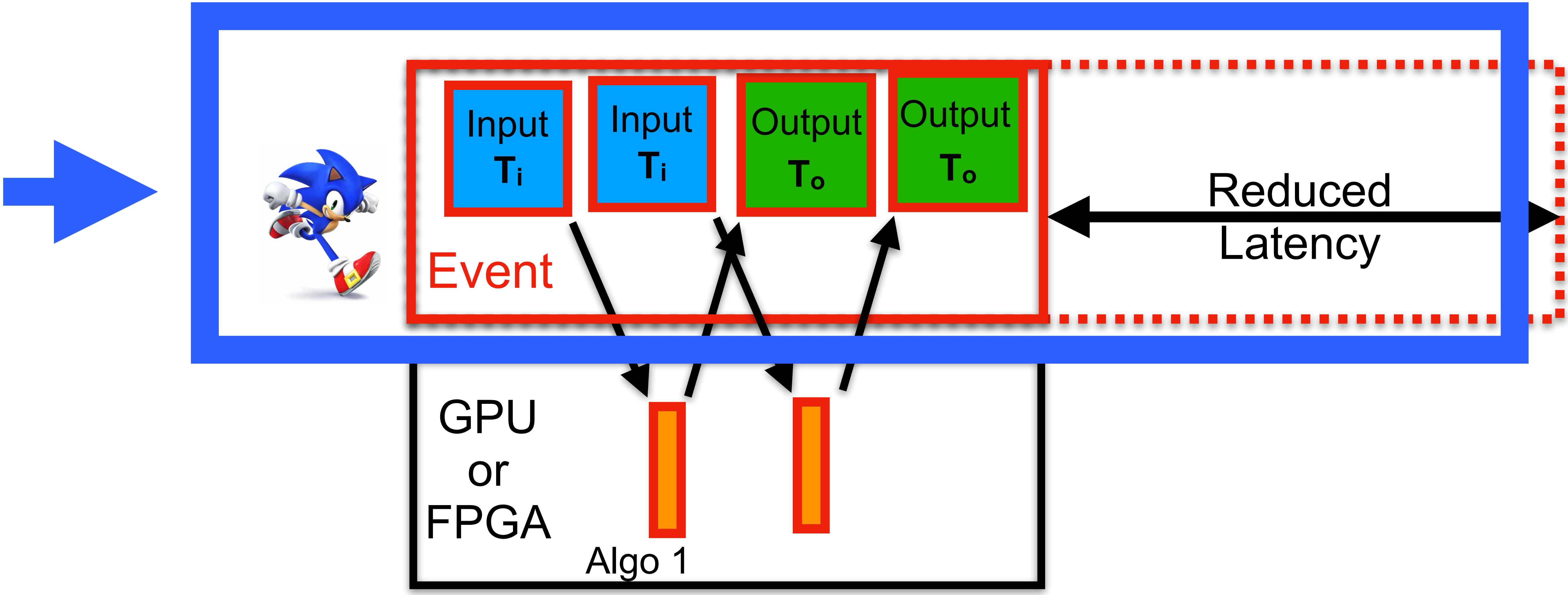
Asynchronicity allows for longer wait times

arxiv:1904.08986

Integrating with cloud

SONIC

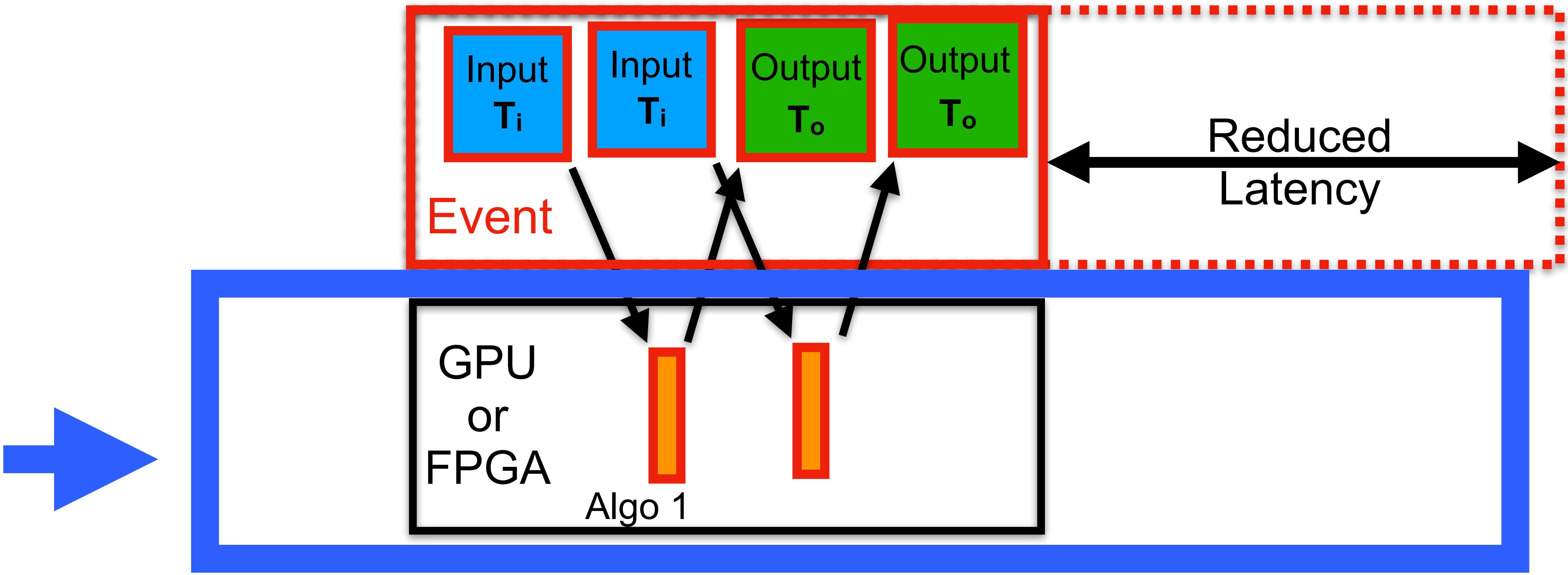
Services for Optimized Network Inference on Coprocessors



Integrating with cloud

SONIC

Services for Optimized Network Inference on Coprocessors



gRPC servers:

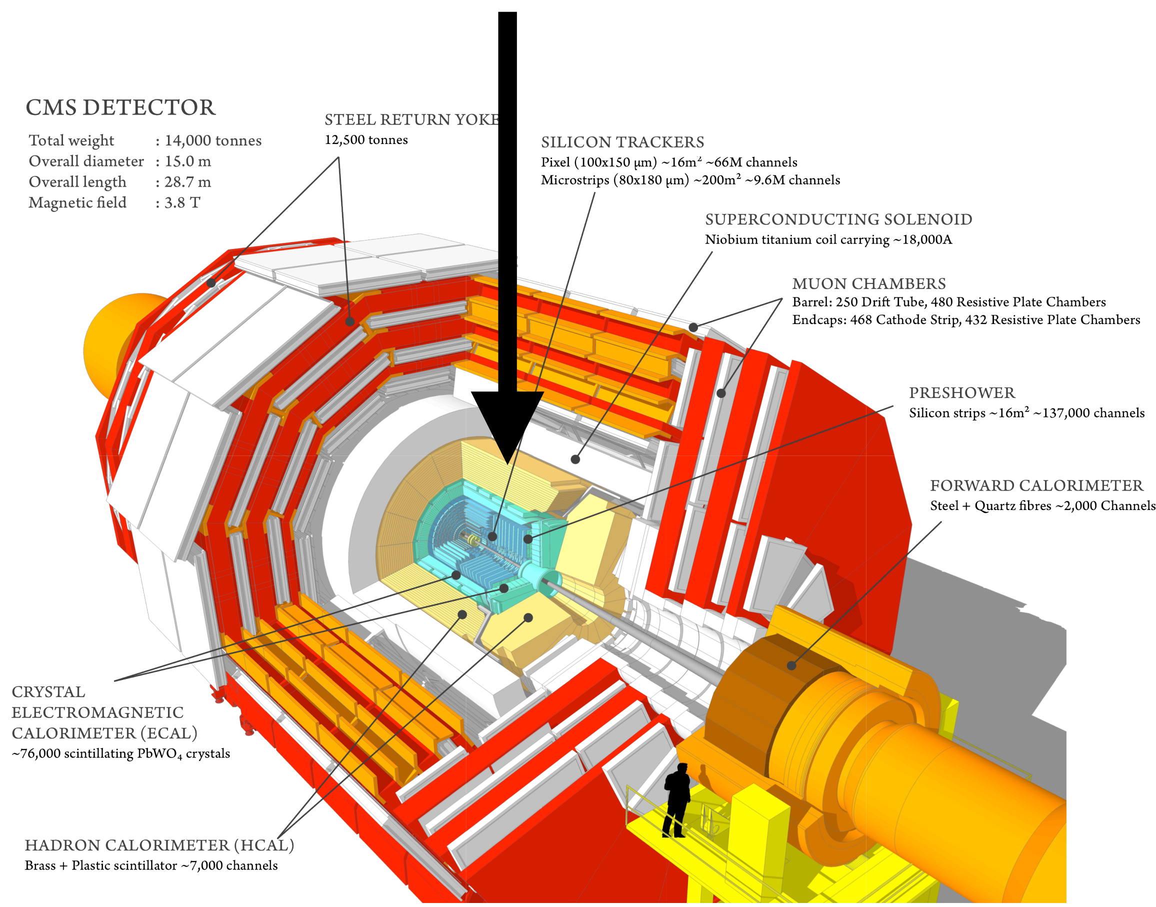
arxiv:1904.08986

FPGA-as-a-service Toolkit (FAAST) w/Xilinx ML Suite/HLS4ML/…

or NVIDIA TRITON INFERENCE SERVER

Case Study

Reconstructing this detector



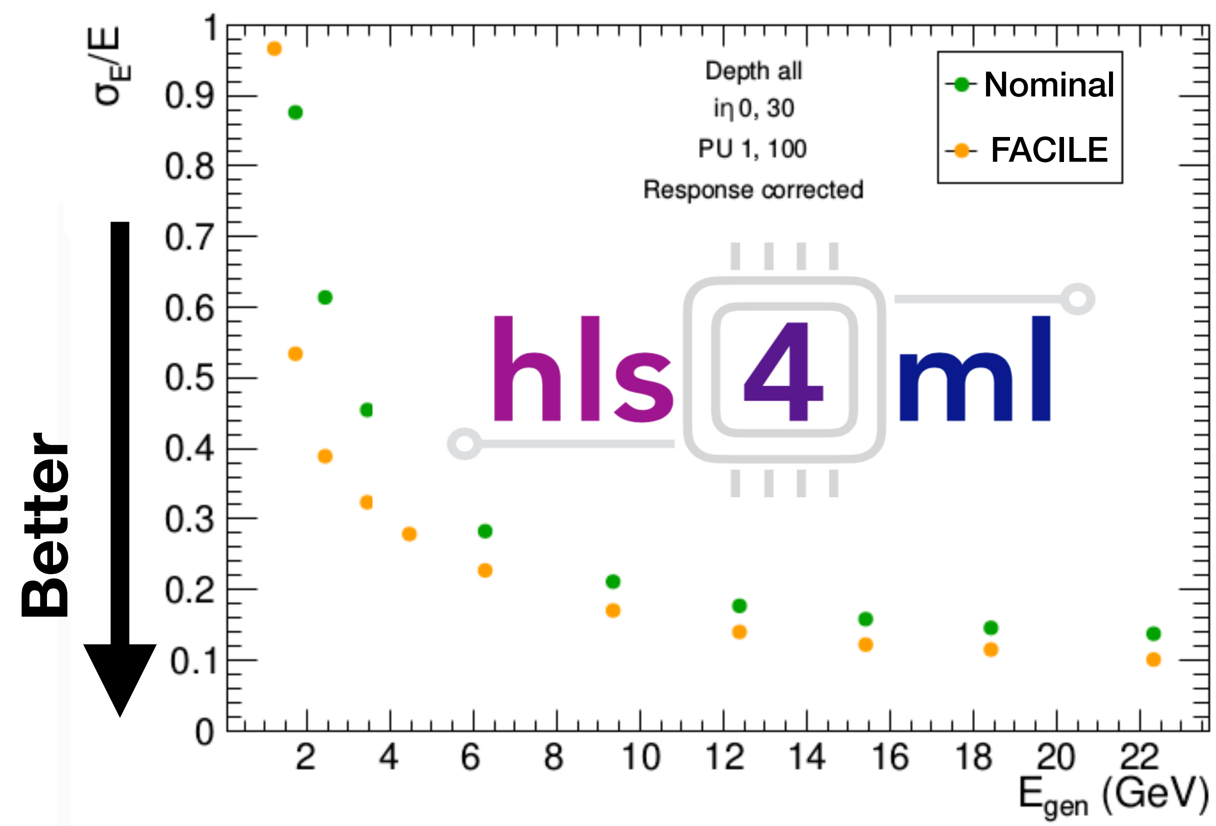
Deep Neural Network that reconstructs energy deposits

Applied to 16k (Batch) Channels

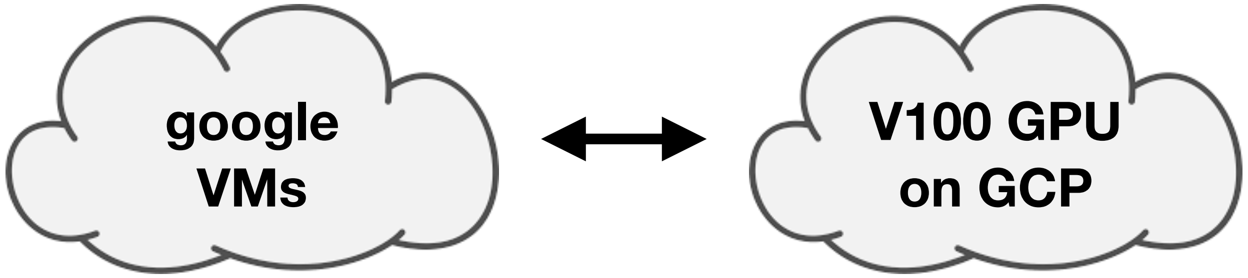
Run at batch 1 on FPGA

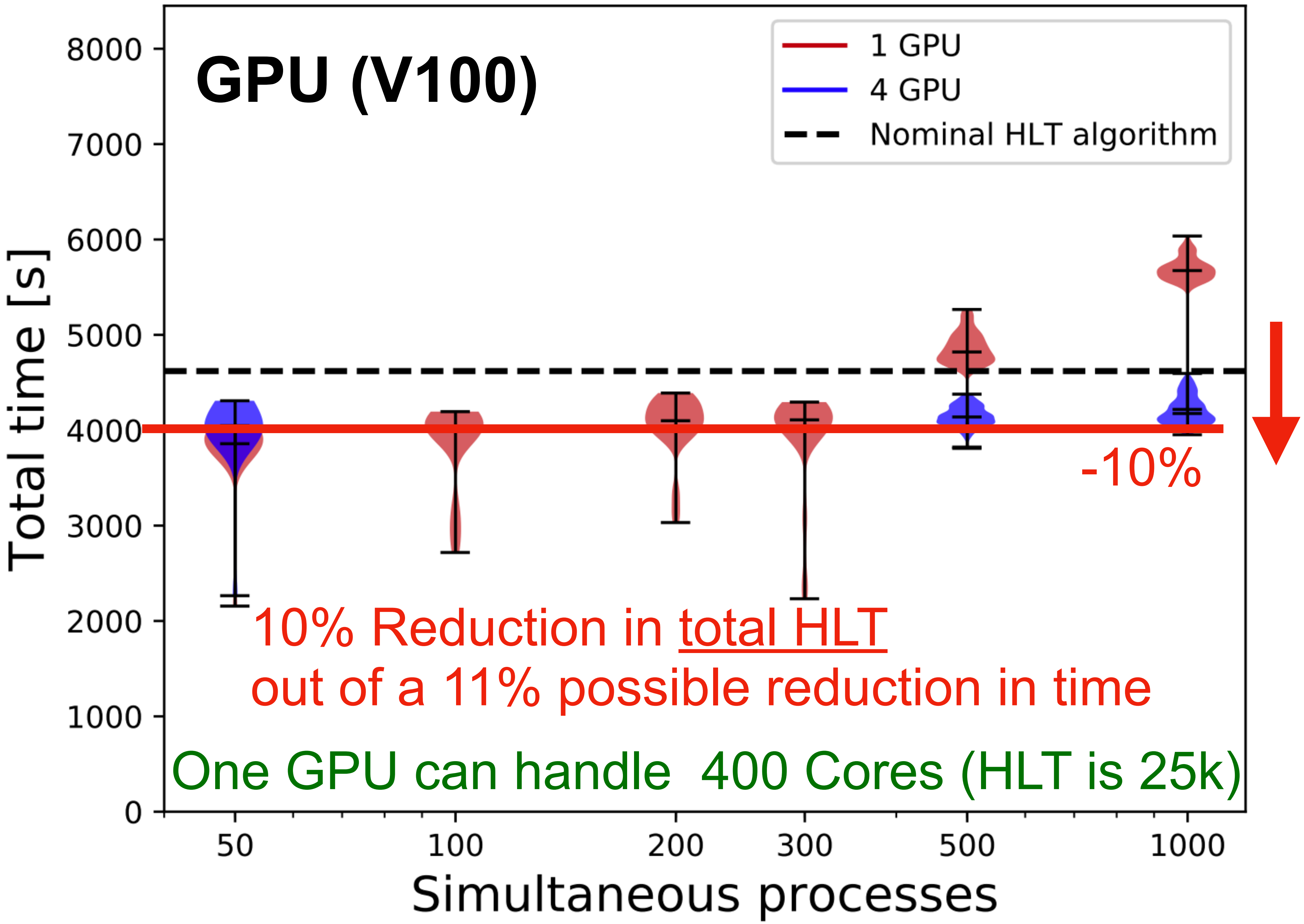
II=2 Clocks (8 ns)

Algorithm	Accelerator	Time
Nominal	None	60 ms
FACILE	GPU	2 ms*
FACILE	FPGA	0.1 ms*

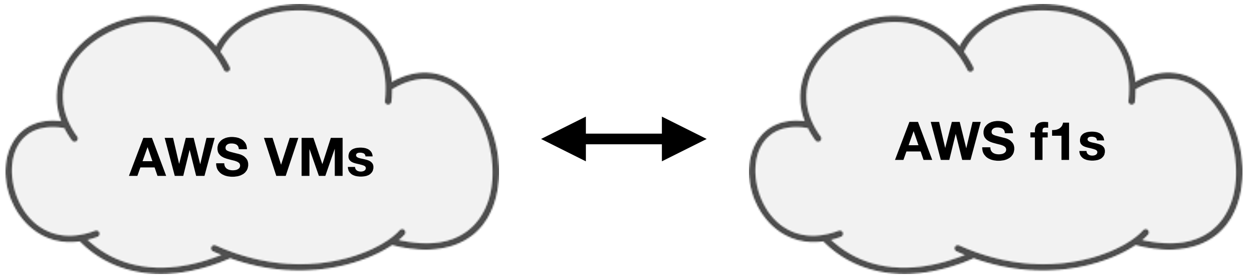


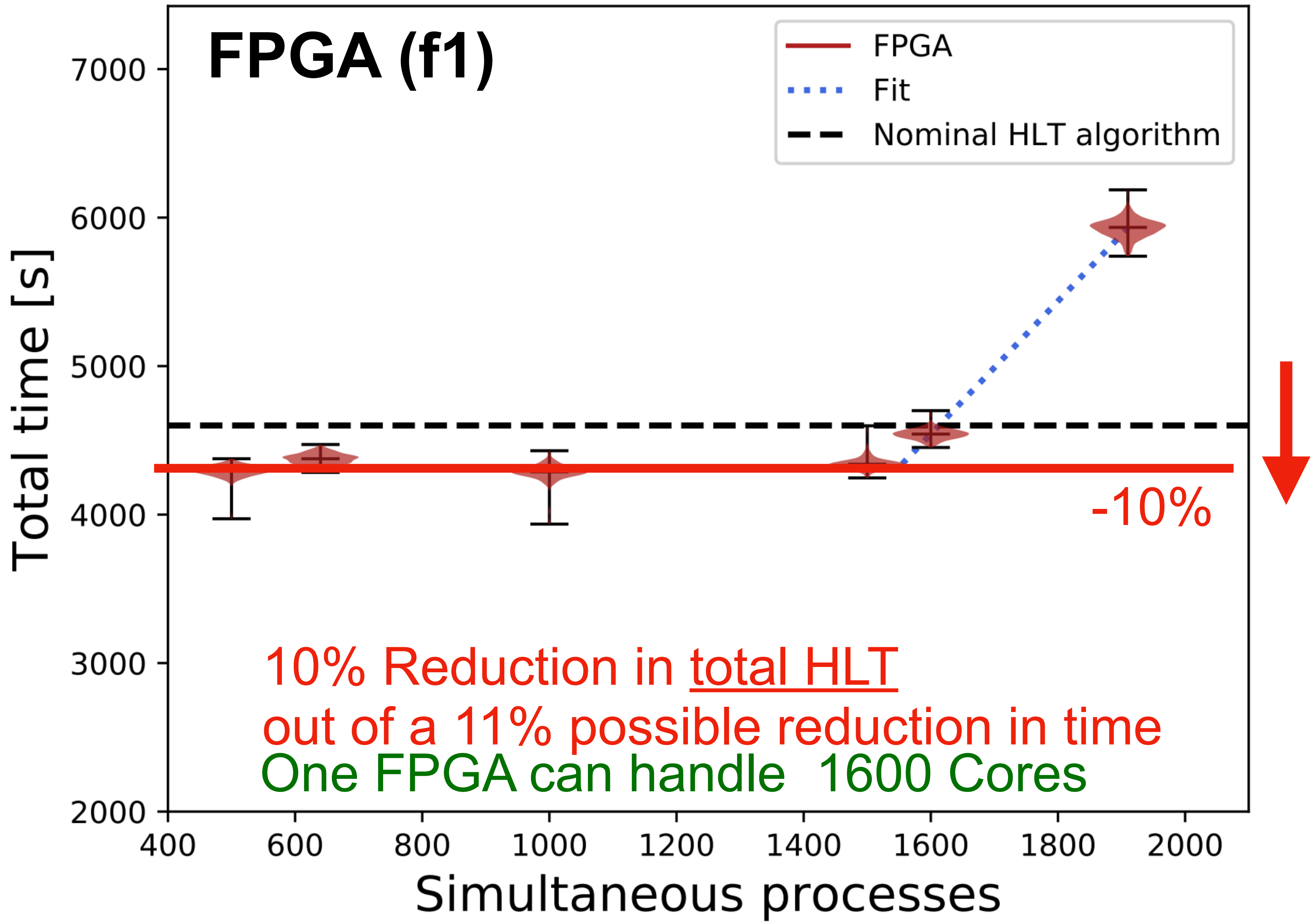
FPGA is on SLR of an Xilinx Alveo U250





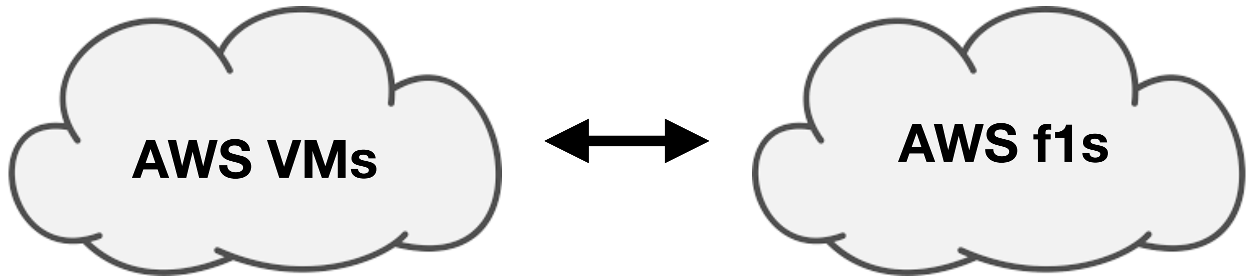
arxiv:2007.10359



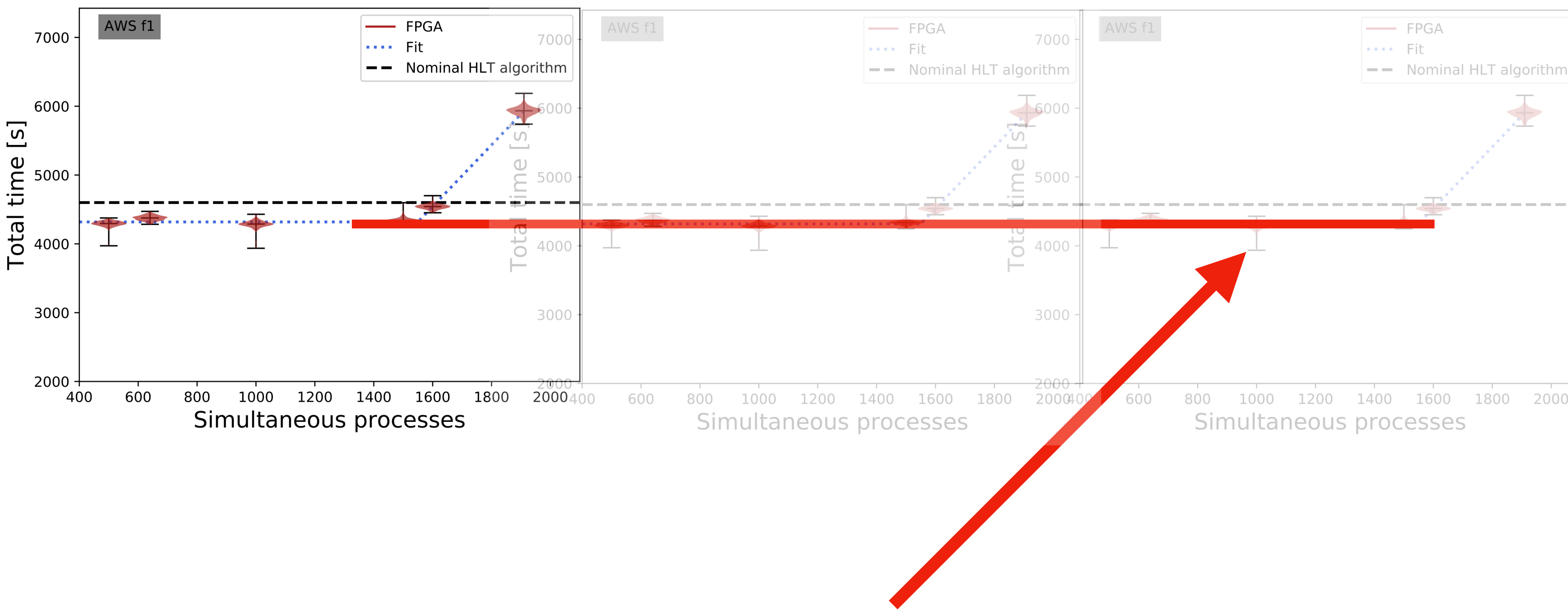


In fact the limit here is not from the FPGA its network (25 Gbps)

arxiv:2010.08556



Actual FPGA limit (f1)



Limit without 25 Gbps is actually at 5500 simultaneous processes

That means 6 FPGAs can reduce 30k core system by 10%!

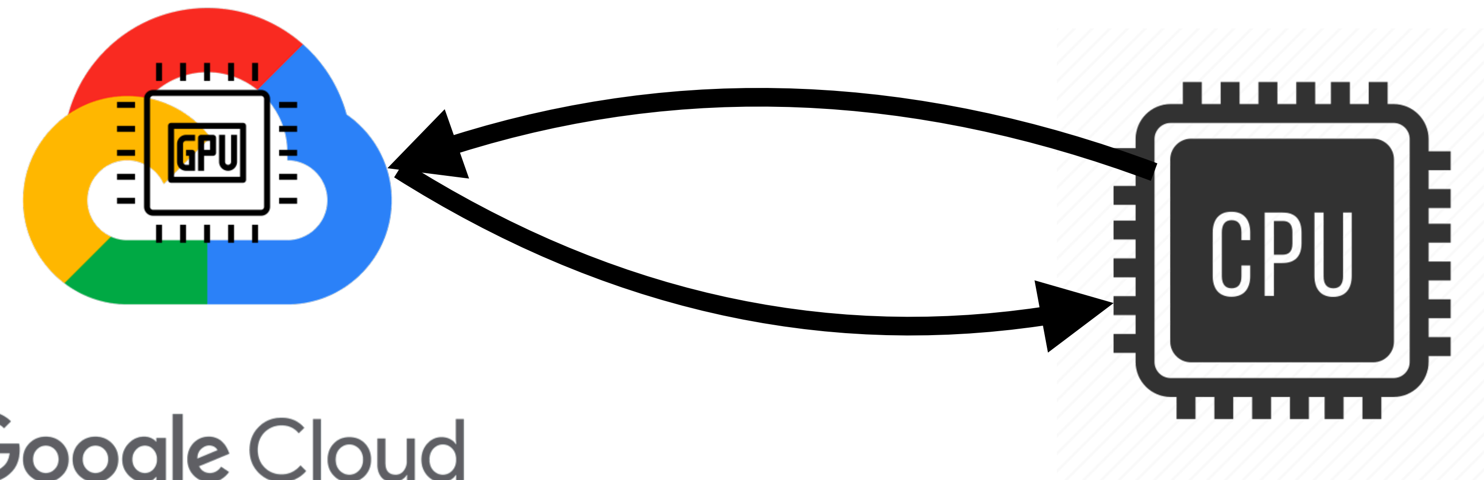
arxiv:2010.08556

Running To Scale

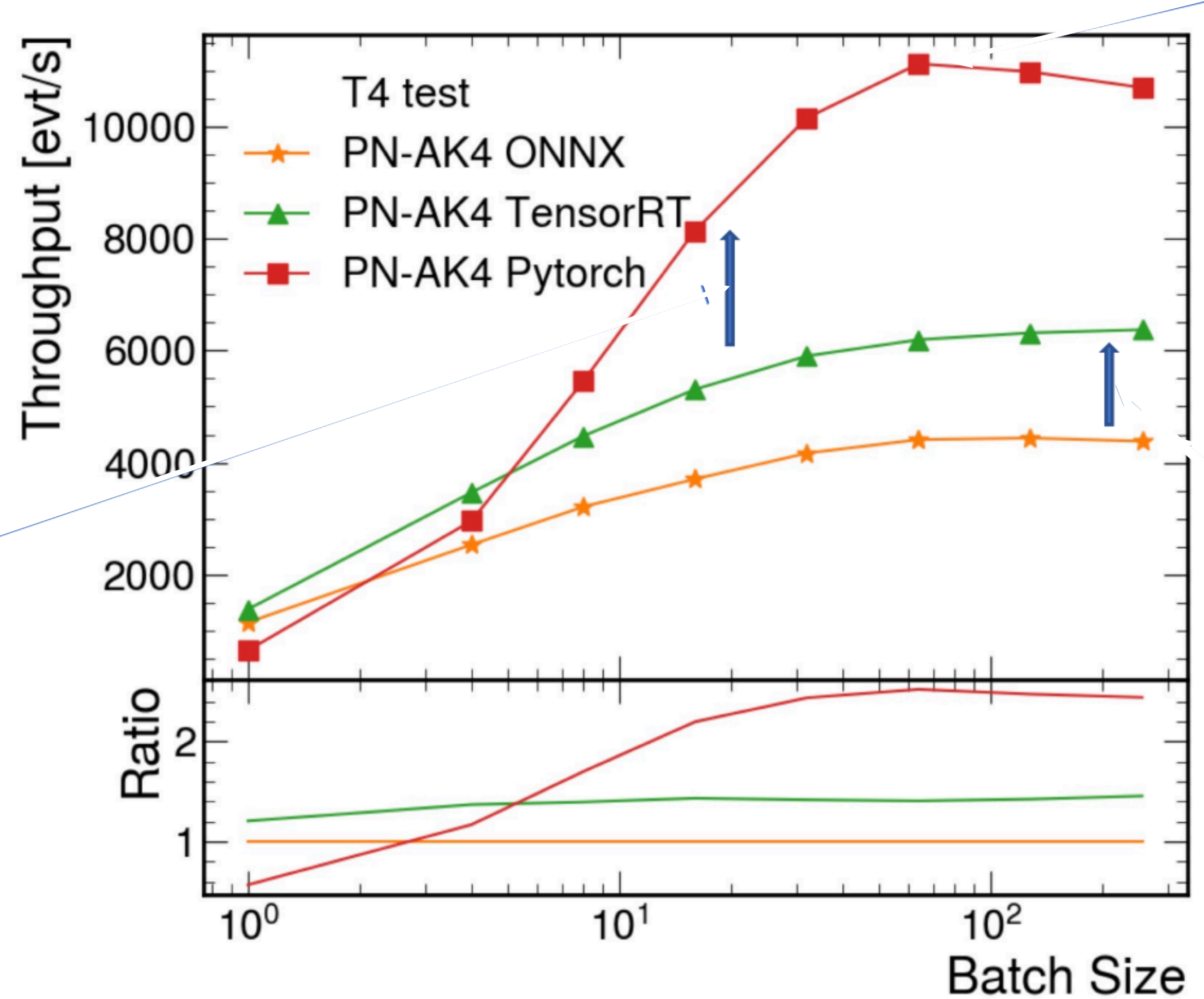
- In addition we have been able to run this work to scale

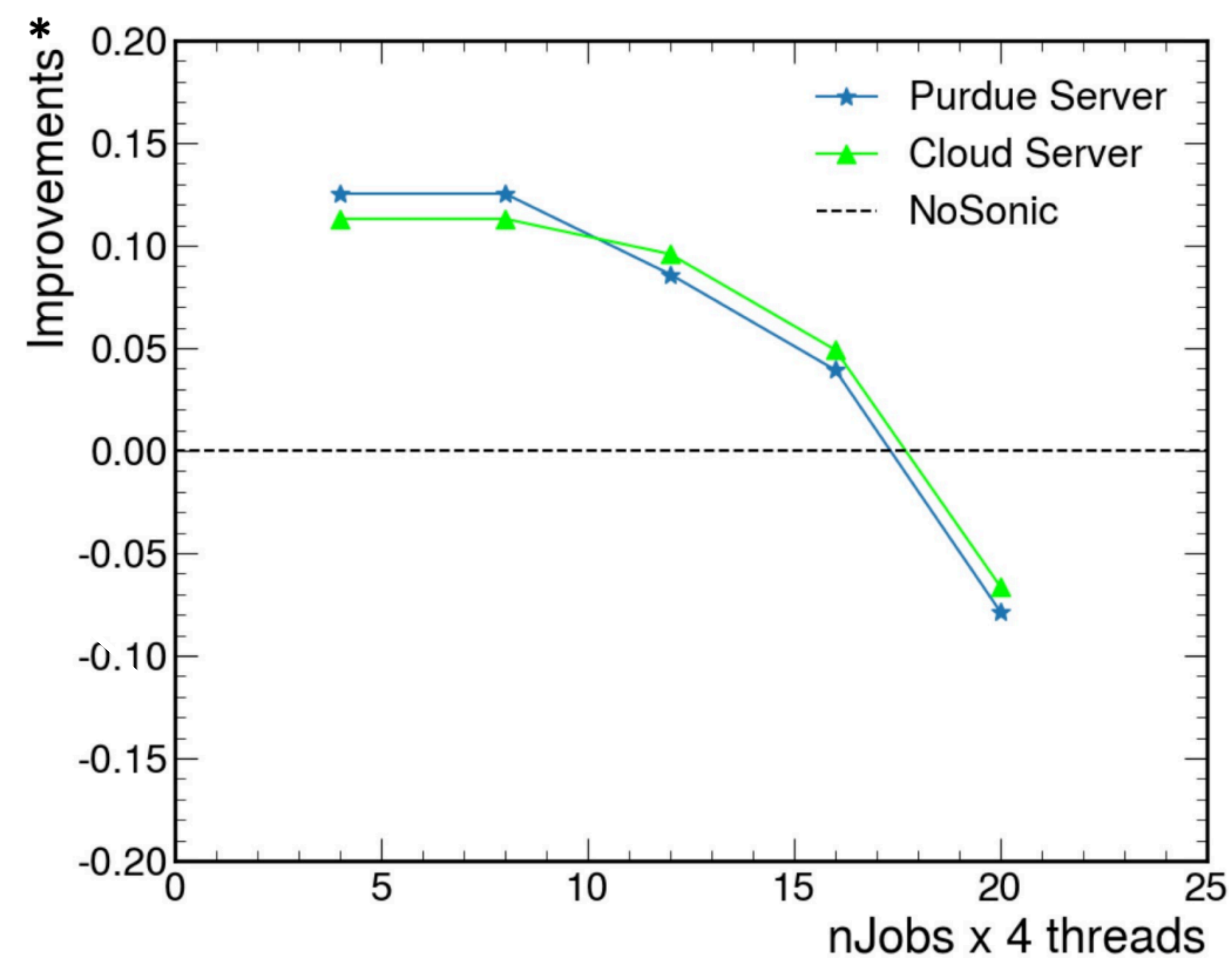
By Using Google Cloud

Sped up 3 algos currently in use gave 15% reco speedup



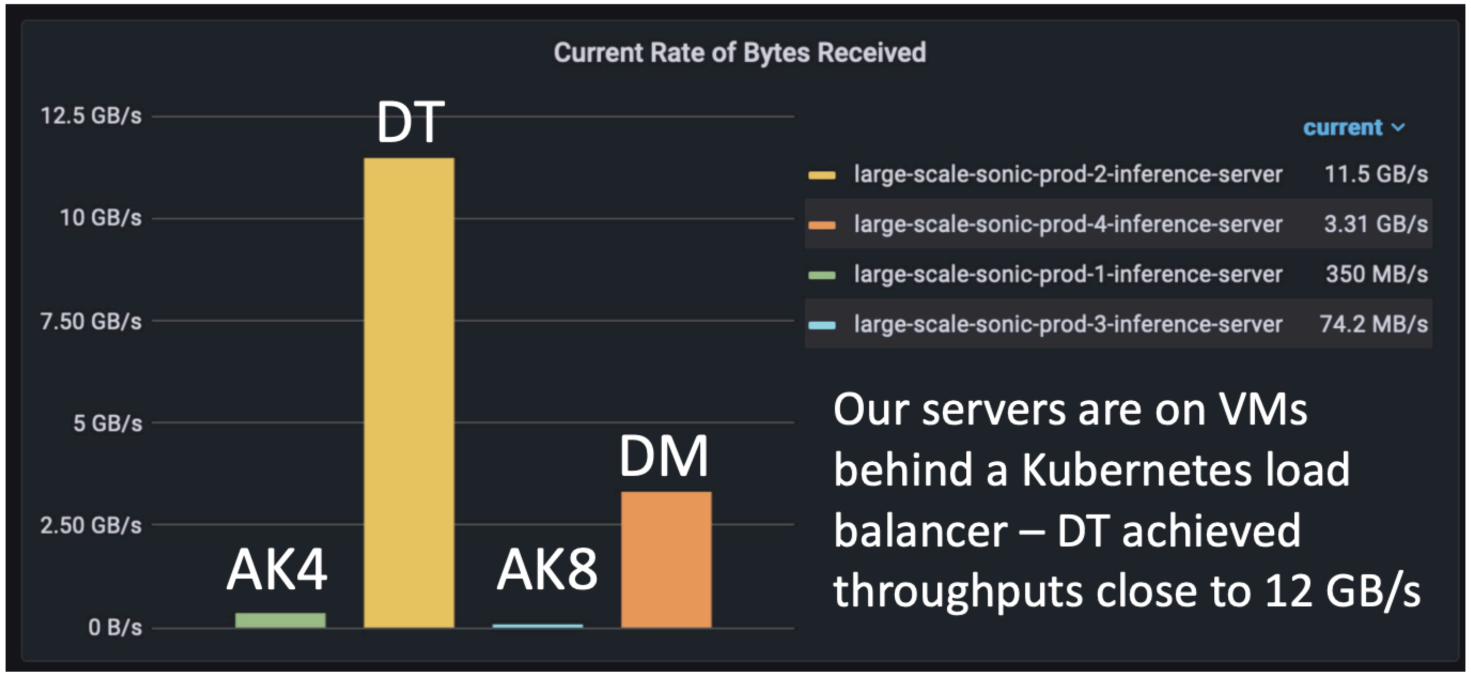
Google Cloud

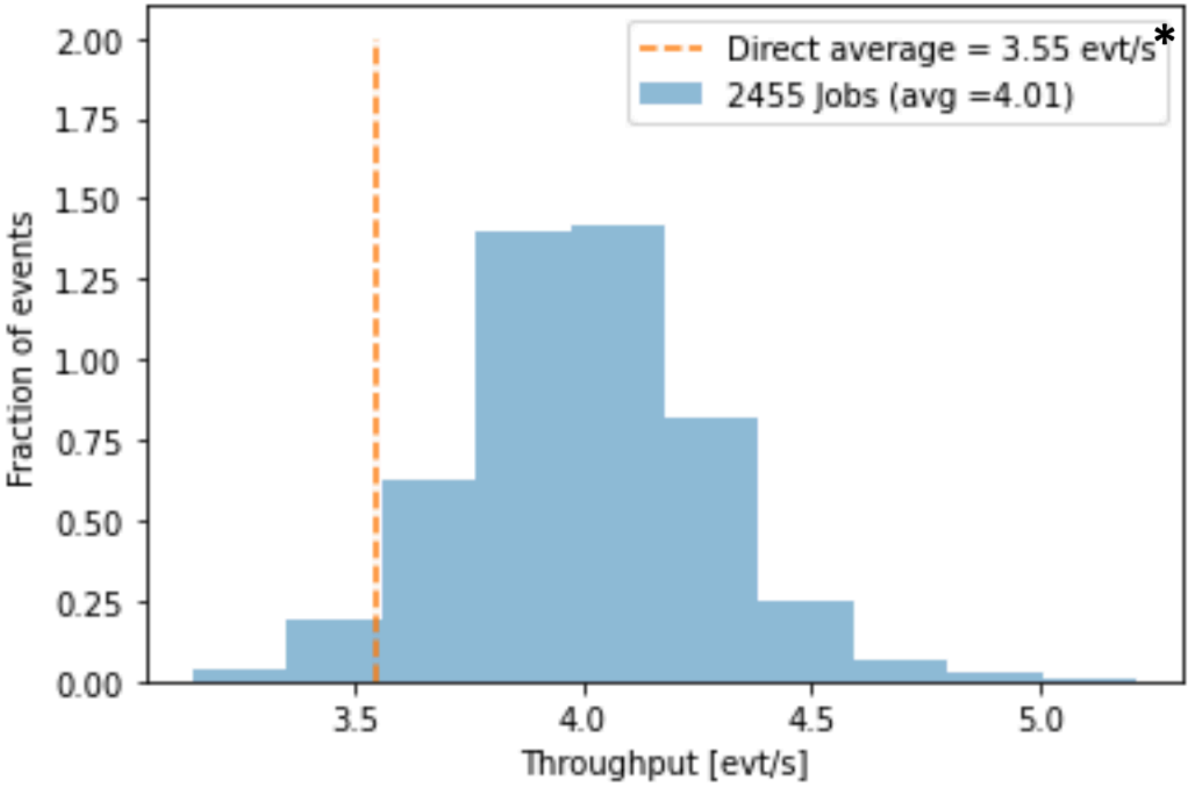




Running To Scale

- In addition we have been able to run this work to scale
  - Ran a test with 10000 CPU cores and 150 GPUs
  - Processes a realistic 150 TB sample
  - Demonstrated this paradigm works to scale!





Other Algos

We have considered a broad range of algorithms

Algo	Batch/Event	CPU	GPU	FPGA
Hcal (Prev Slides)	16000	60ms(16ms)	2ms	0.2ms
Electron Id	5	75ms	0.1ms	<1ms(tbd)
Top Quark(resnet50)	<1	1500ms	1.2ms	1.5ms

At Large batch(saturated)

Like the physics events: there is a wide variety of algorithms

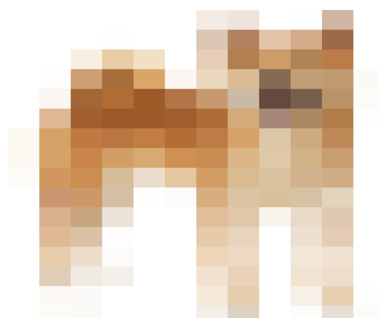
Small algorithms can benefit from optimizations on FPGA

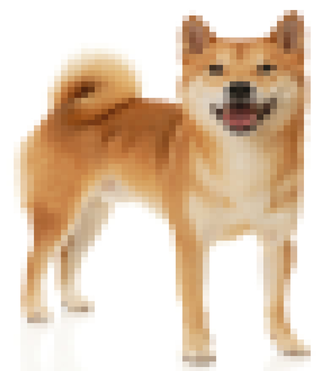
Larger algorithms+slower inference times GPUs start to work well

arxiv:2007.10359

A Broader Vision of DAQ

40 MHz **1 kHz**



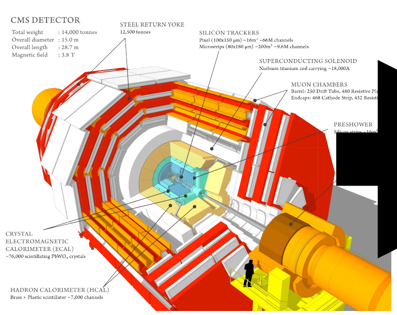


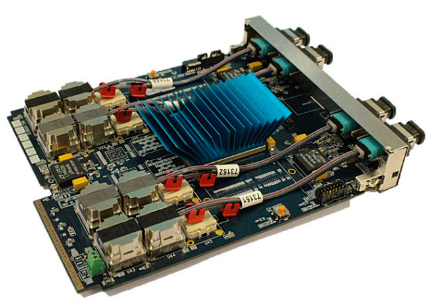
Radiation Hard ASICs

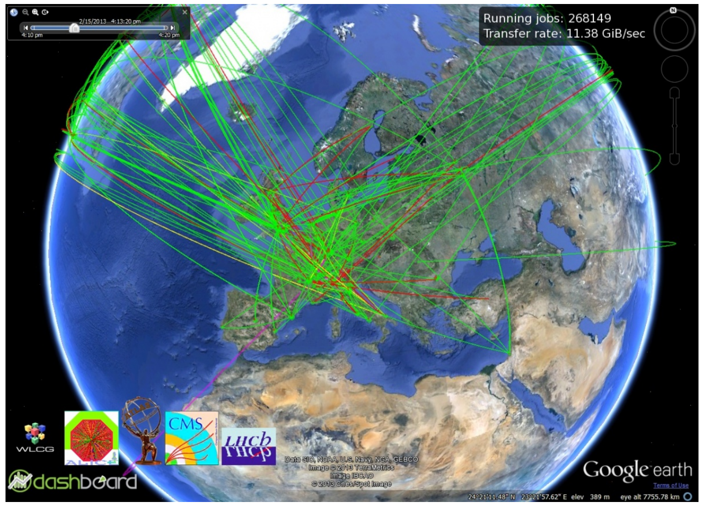
FPGA Boards

Local CPU Cluster

CPU Grid







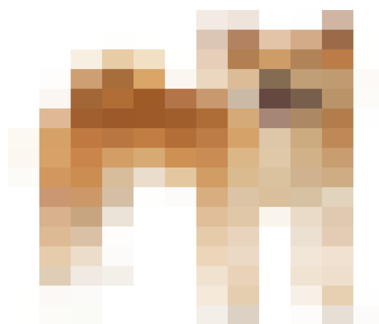
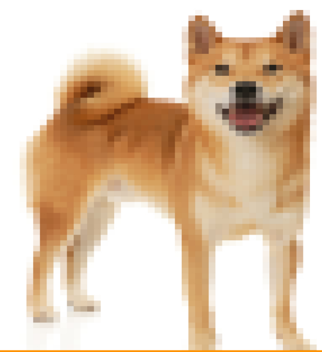
320 tb/s

1 tb/s

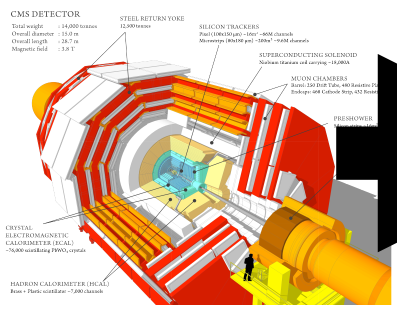
10 Gb/s

A Broader Vision of DAQ

40 MHz **1 kHz**

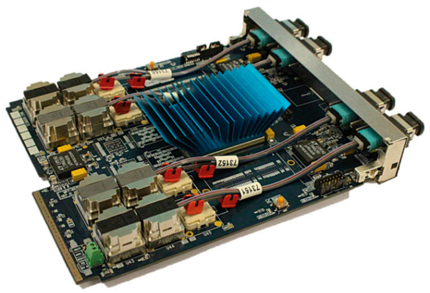




Radiation
Hard ASICs



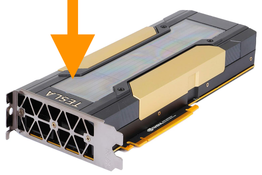
320 tb/s

FPGA
Boards



1 tb/s

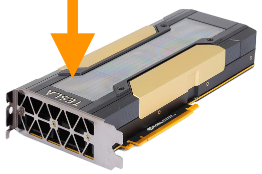
Local CPU
Cluster



Accelerator

10 Gb/s

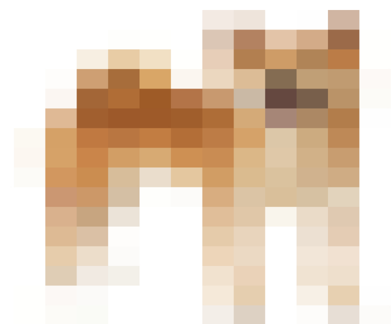
CPU Grid



Accelerator

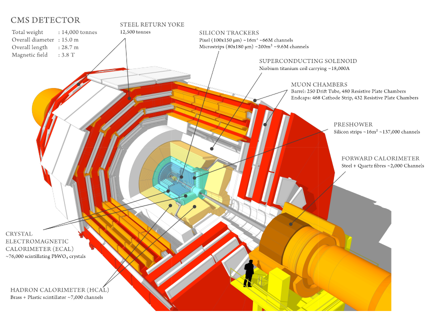
A Broader Vision of DAQ

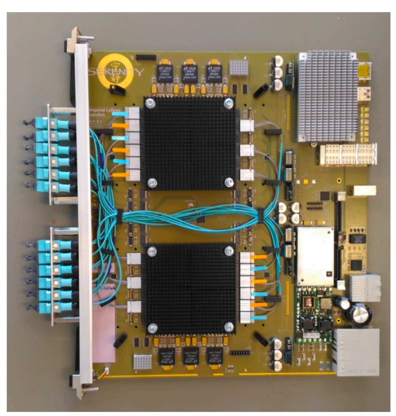
40 MHz **100 kHz**



Radiation Hard ASICs

FPGA Boards

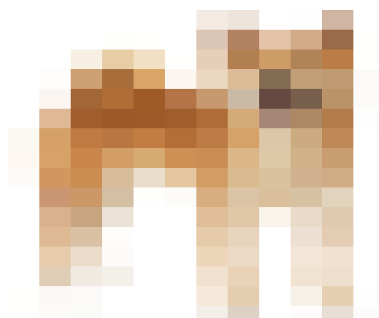


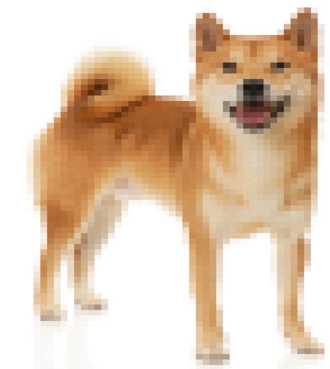


Now Lets Zoom In on our system

A Broader Vision of DAQ

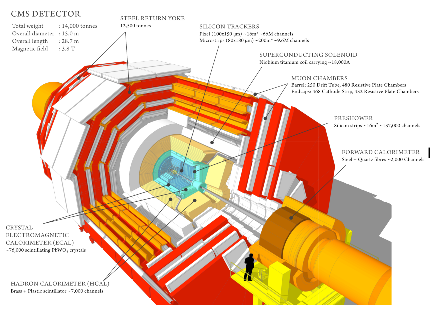
40 MHz **100 kHz**

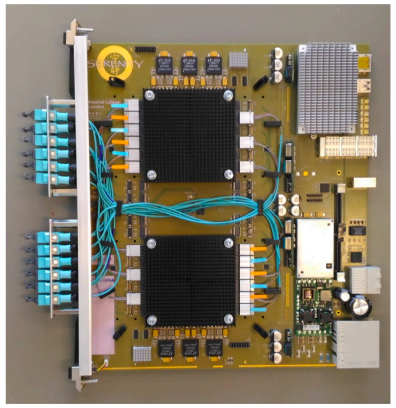


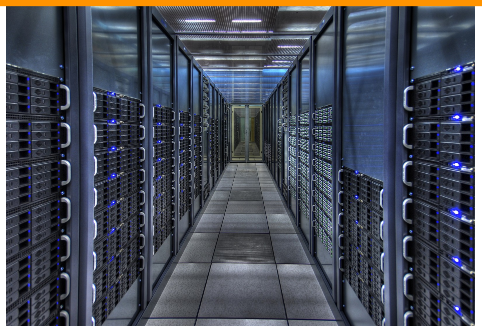


Radiation Hard ASICs

FPGA Boards



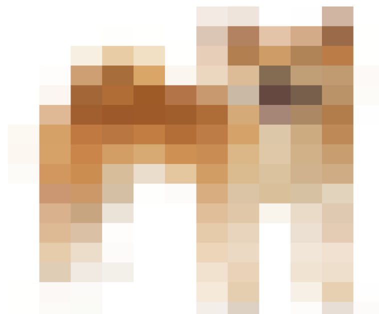


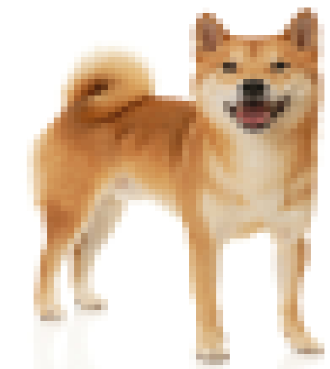


And Reconfigure it

A Broader Vision of DAQ

40 MHz **100 kHz**

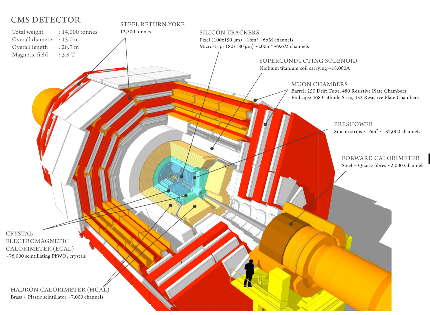


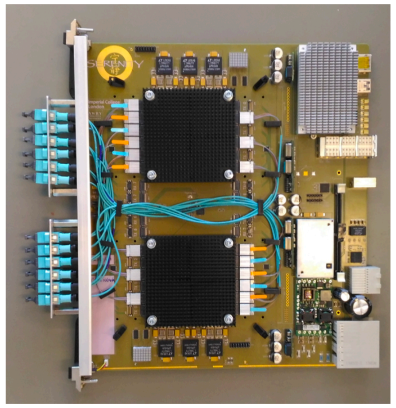


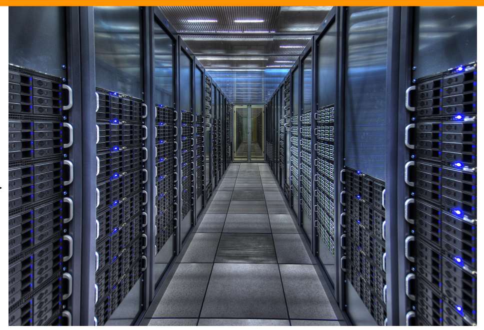
Radiation Hard ASICs

FPGA Boards

Throttle between 100kHz-40 MHz







What can we do if we go from Our FPGA system to accelerators?

Algean

P. Chow N. Tarafdar

UNIVERSITY OF TORONTO

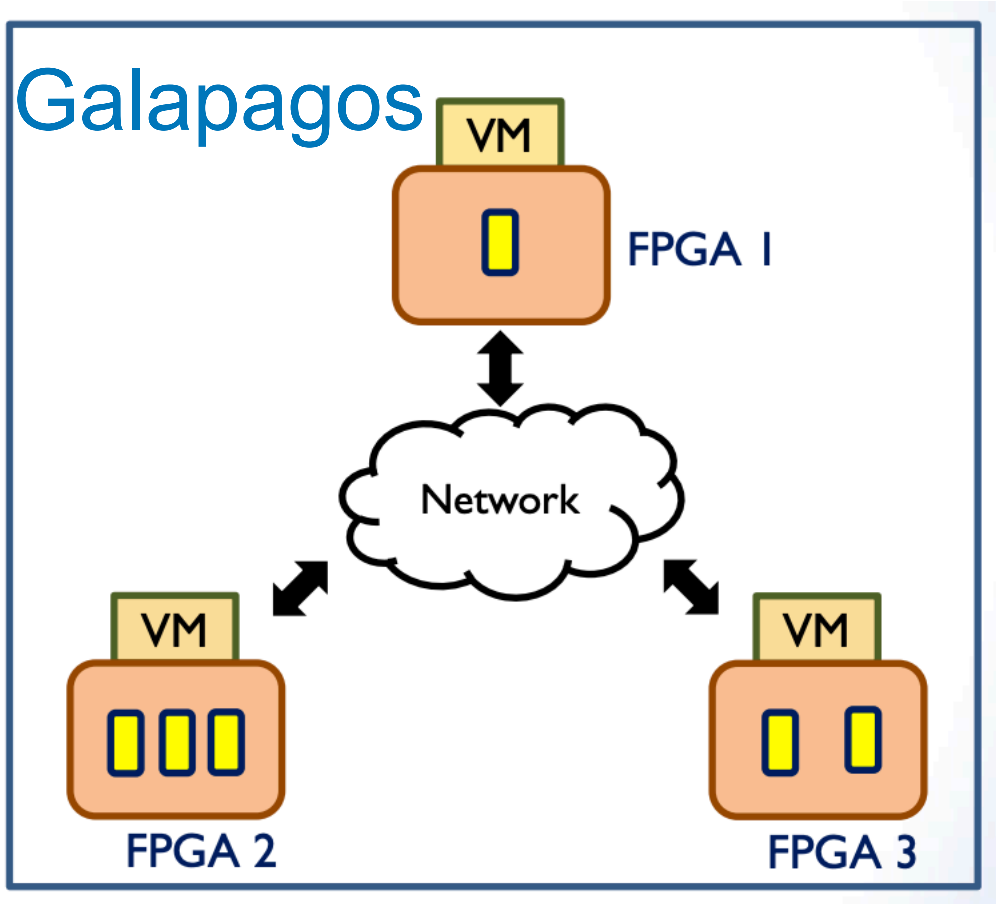
Combining Ideas

- What if we combine the two show concepts?

hls4ml

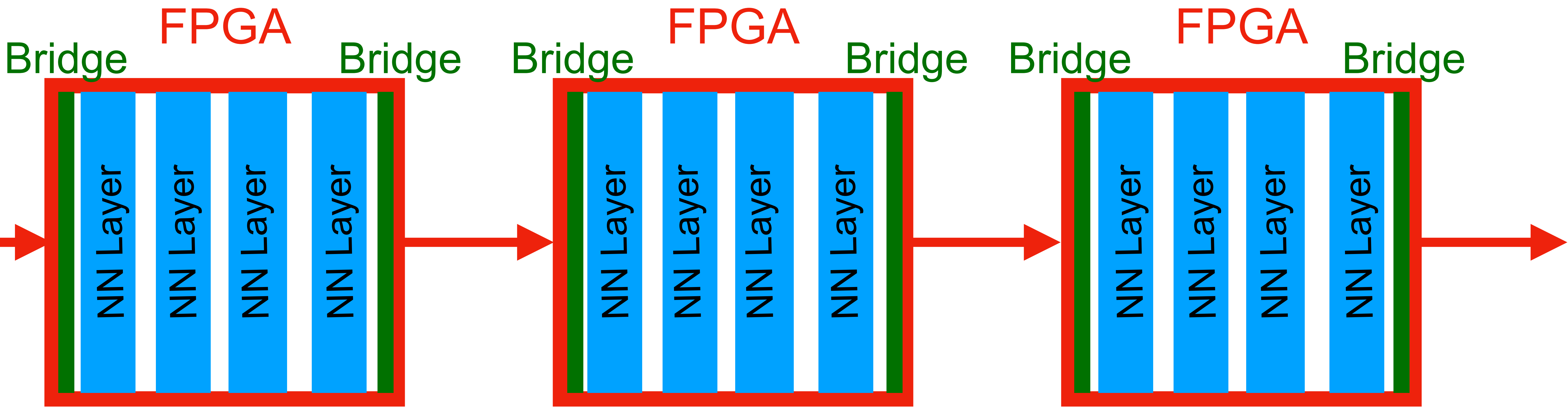
Fast Distributed Deep Learning Networks

+



Open Source Tool to talk to FPGAs Directly over Network

Algean



With Algean we can stretch out networks across many FPGAs
100 Gb/s protocol between FPGAs (can go to CPUs)

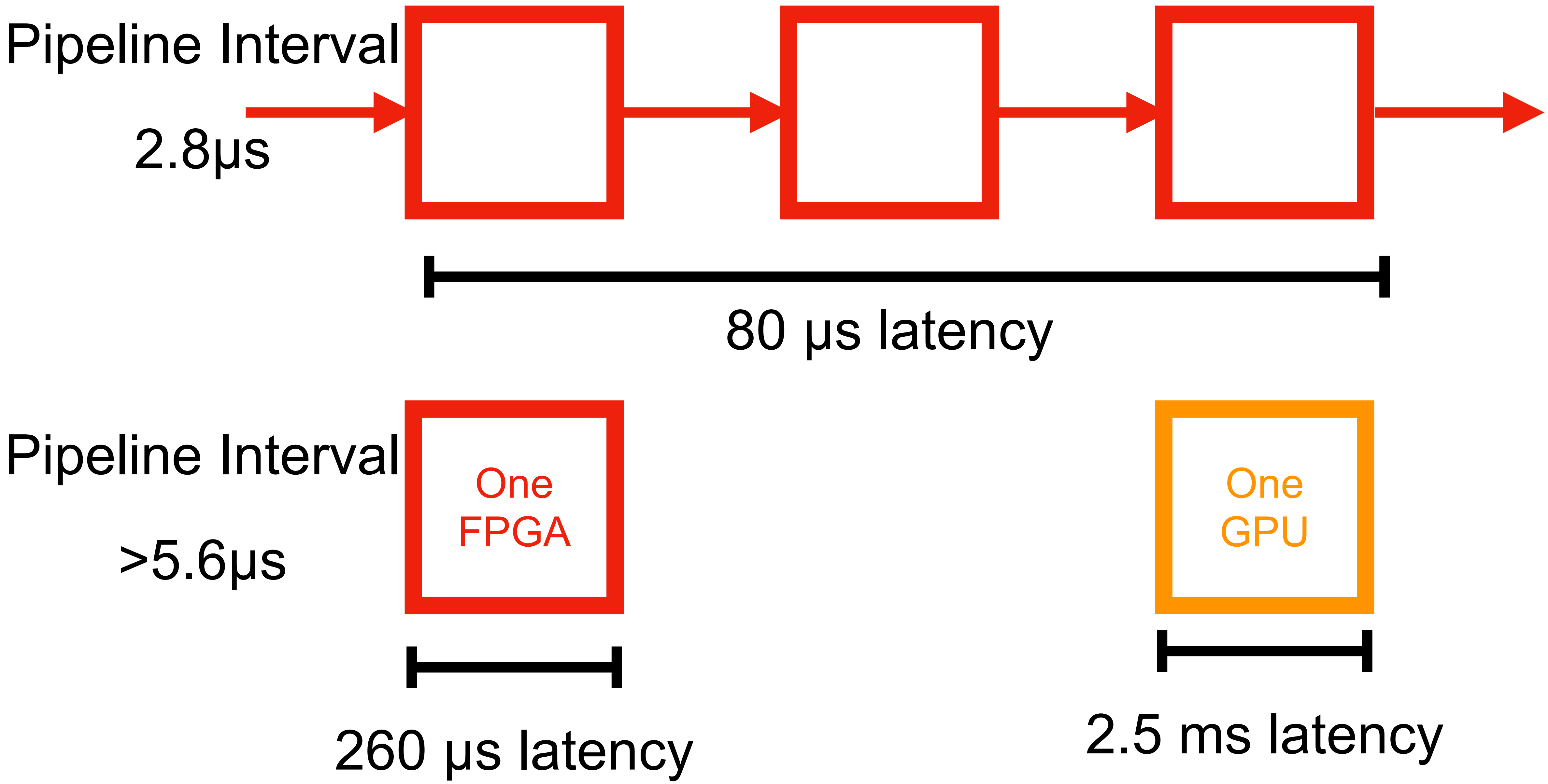
This allows us to run inference for very large networks

Very Fast

Tune our network to the resources we have

Example Autoencoder

Anomaly detection algorithm



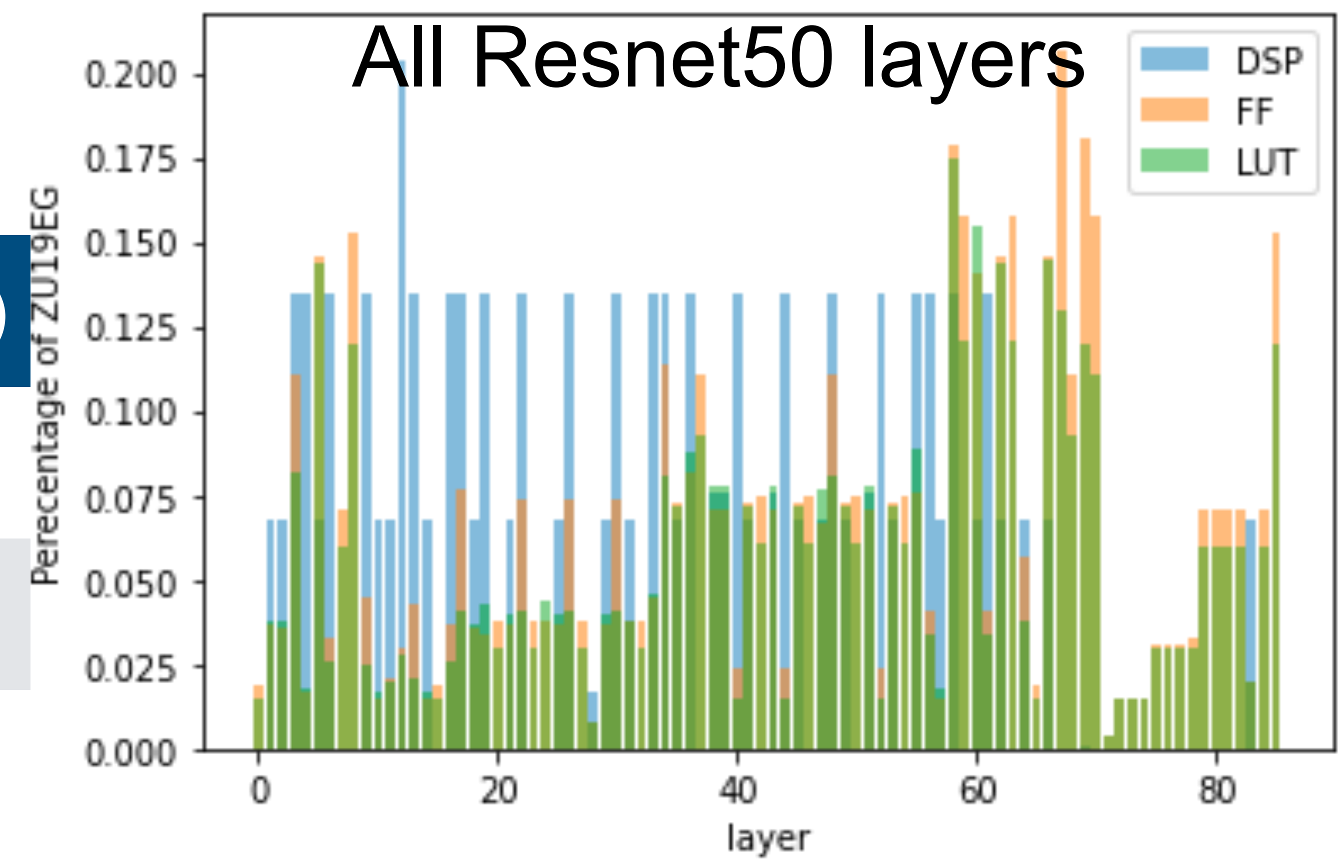
Resnet-50

8bit Resnet50 with a throughput of 1.5ms

Partitioned onto **9 ZU19EG FPGAs**
packed resources would fit 6

We can compile networks
over MANY FPGAs

Implementation	Result
Latency of Data Transfer of a Single Image from CPU to FPGA	2.5 ms
Projected AIgean Throughput of entire CPU/FPGA network	400 images/s
Projected AIgean Throughput on FPGA only	660 images/s
Microsoft Brainwave Batch-1 Throughput [38]	559 images/s



	Resources	Alveo U250(%)	ZU19EG(%)
DSP:	9475	0.77	4.99
LUT:	2895351	1.68	5.55
FF:	4952884	1.43	4.76

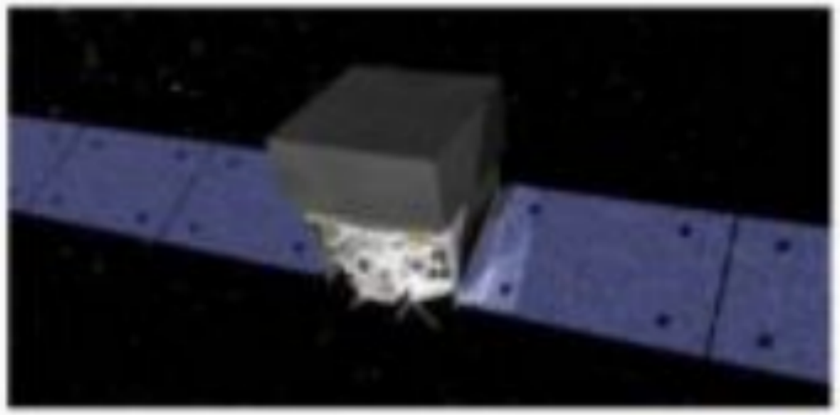
Use Cases?

- A new paradigm of computing
  - Unroll the whole network across many processors
  - Single inference (batch 1) latencies well beyond GPUs
  - Natural way to link CPUs and FPGAs together
  - Can start to envision a new paradigm of LHC Data Acquisition
- Lots of room to explore! OpenSource

Gravitational Waves

- Aiming to identify Gravitatoinal waves fast to do MMA
  - Correlating GW and Optical observations is powerful

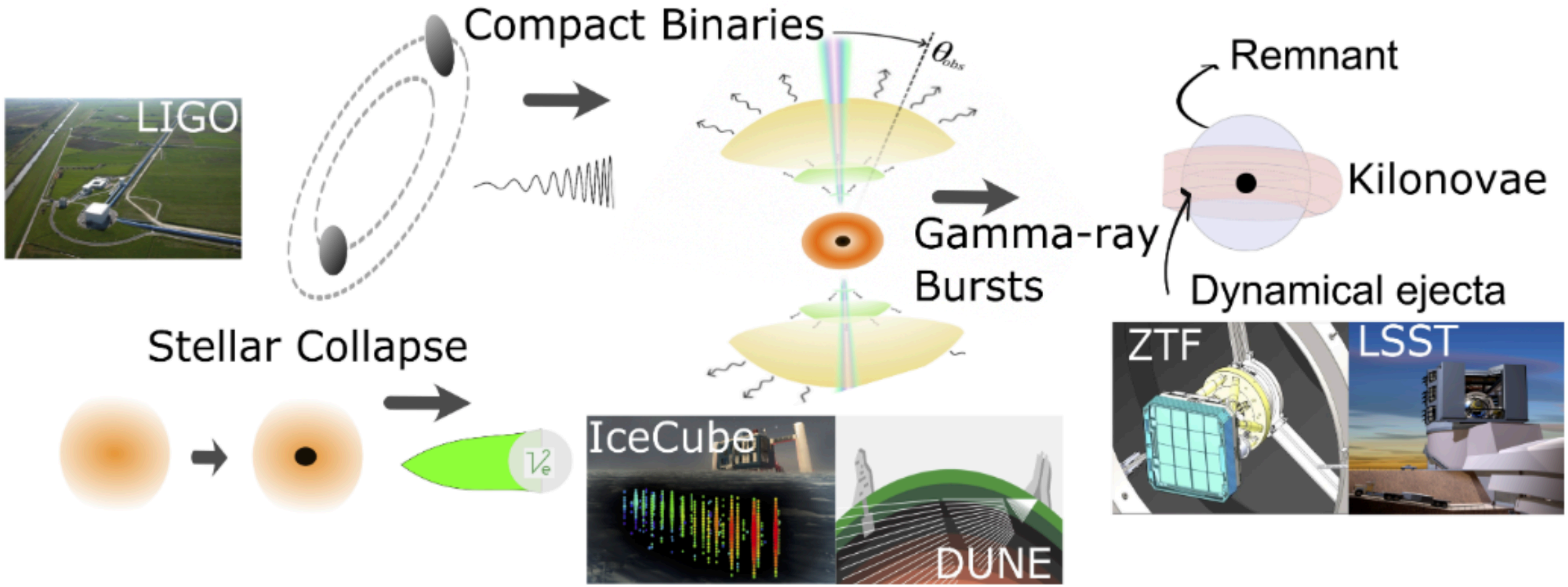
See a Gravitational Wave



Alert a Telescope

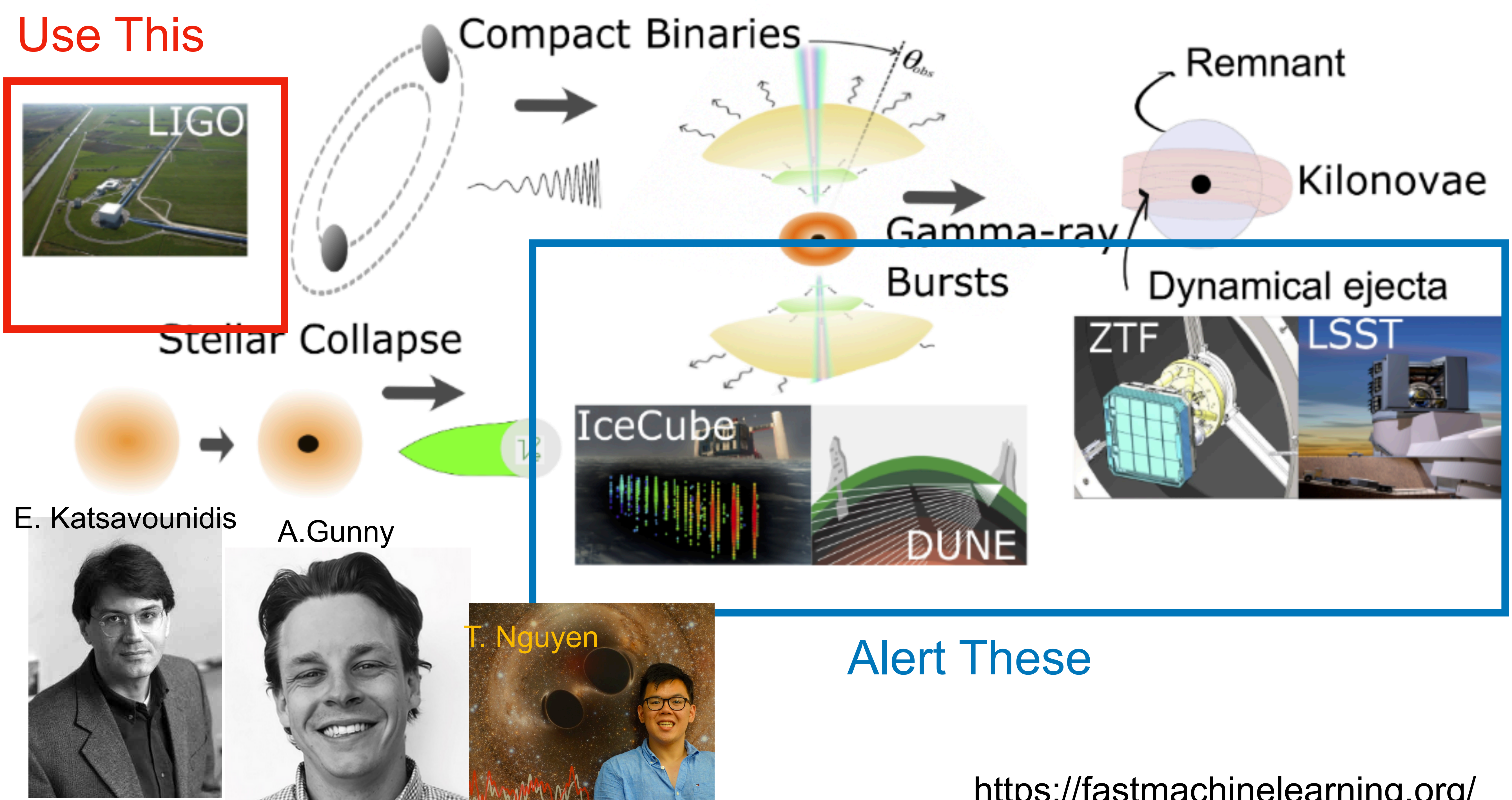
Can we make the GW reconstruction fast enough to be real-time?

Multi Messeng Astro



https://fastmachinelearning.org/

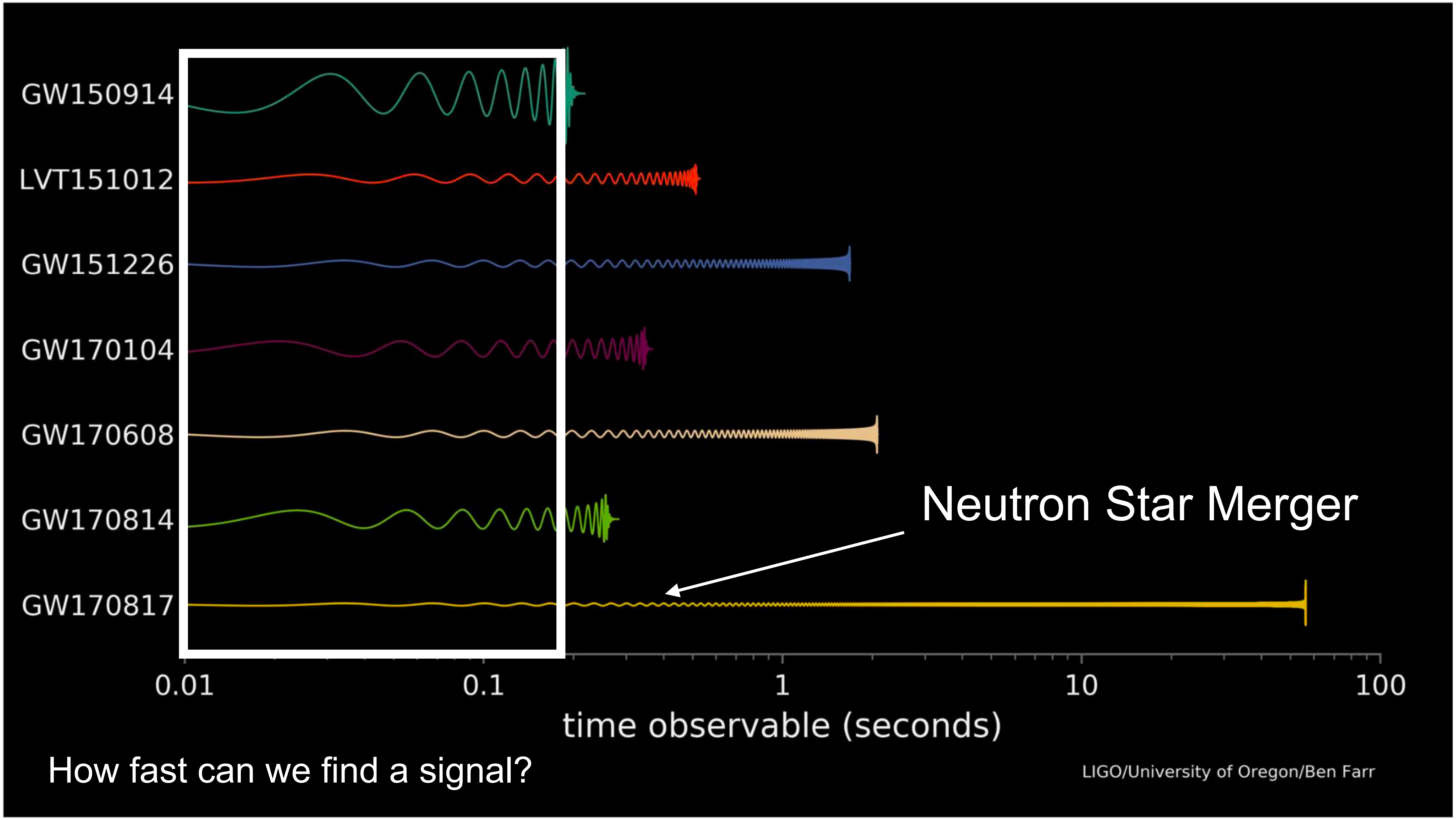
Multi Messenger Astro



https://fastmachinelearning.org/

Gravitational Waves

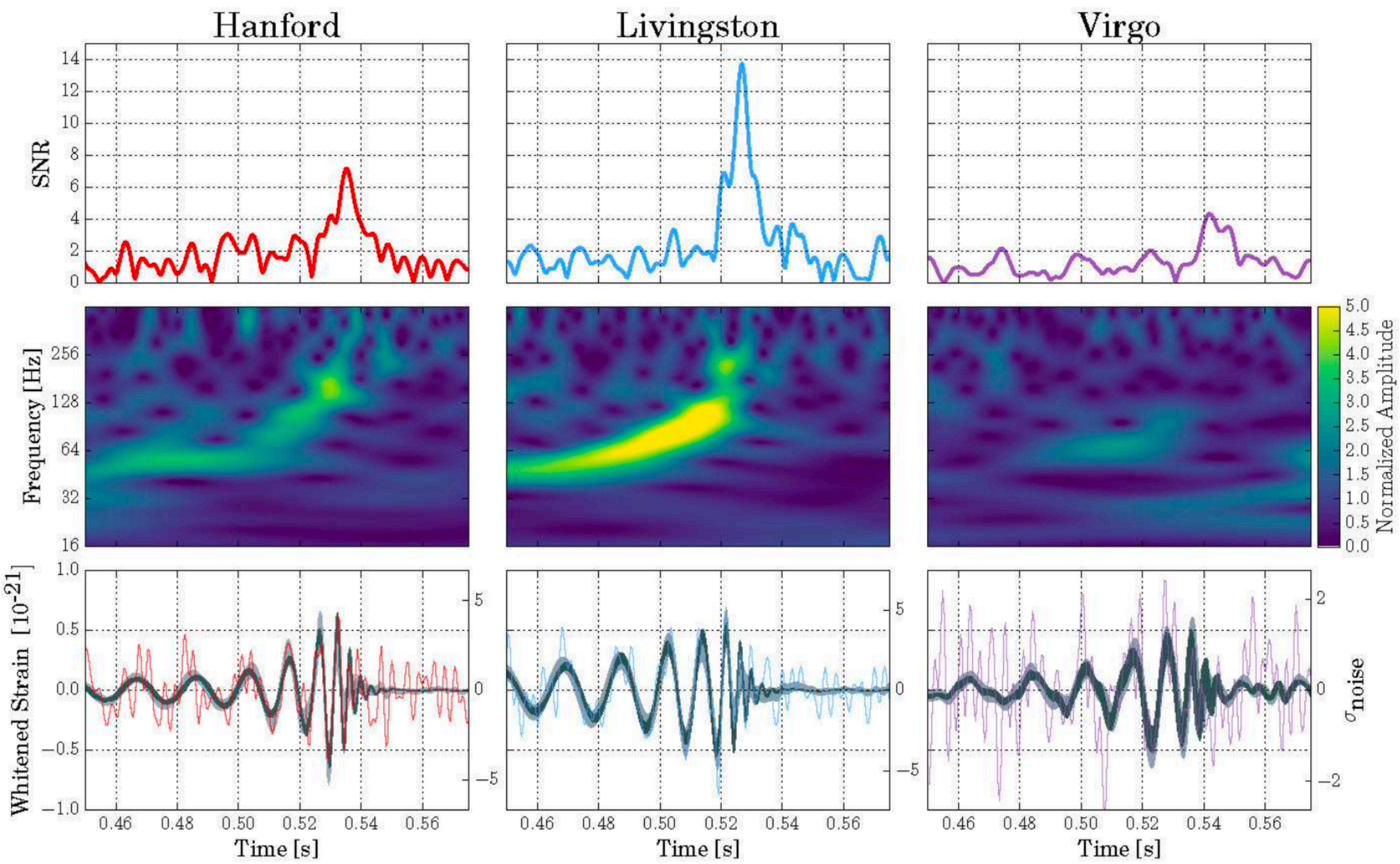
Observed signal durations (above ~30 Hz)



Arsenal of telescopes

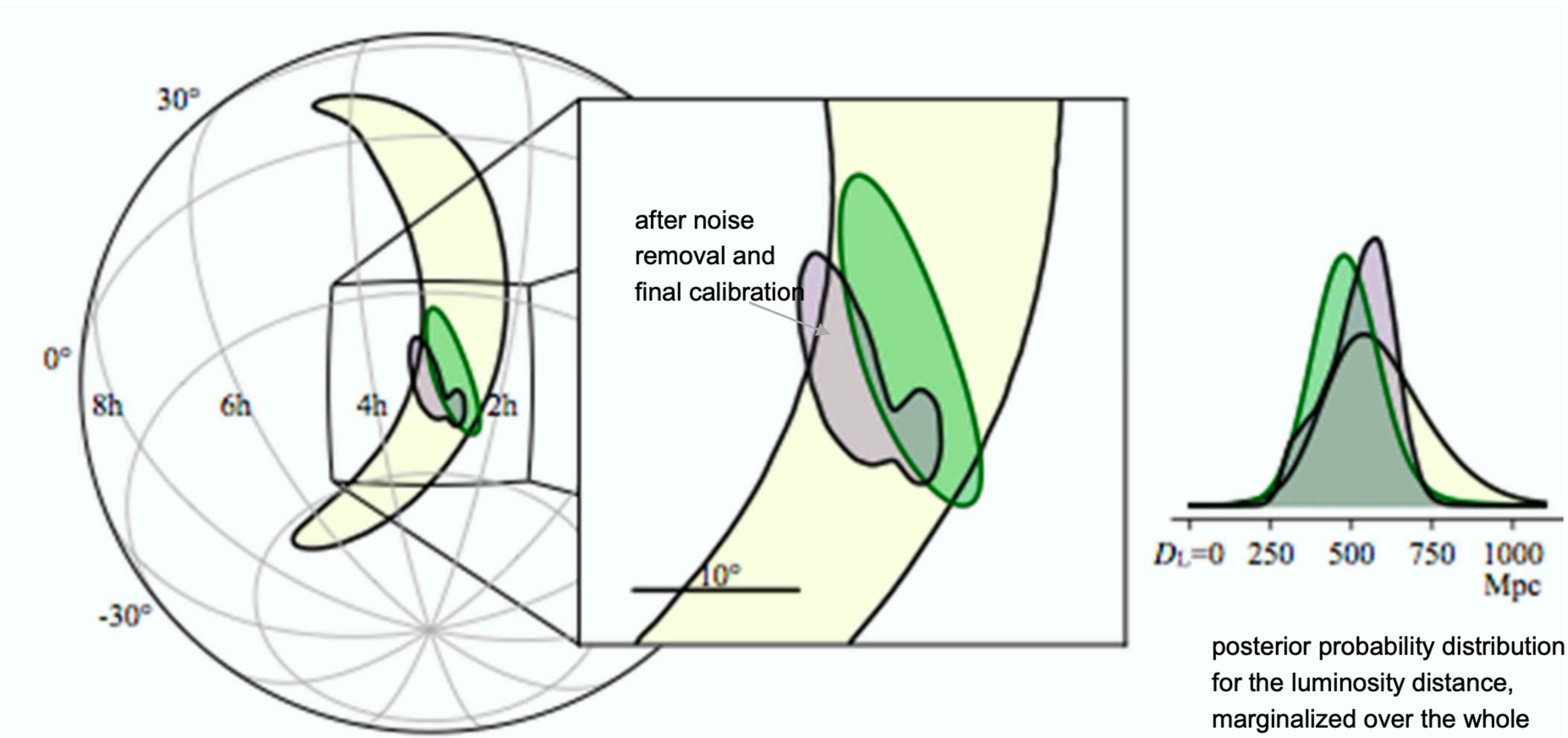
Once you have found the GW event
have to send the coordinates to a huge network

Three detectors: GW170814



A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence
Phys. Rev. Lett., 119:141101, 2017

GW170814 Sky Location



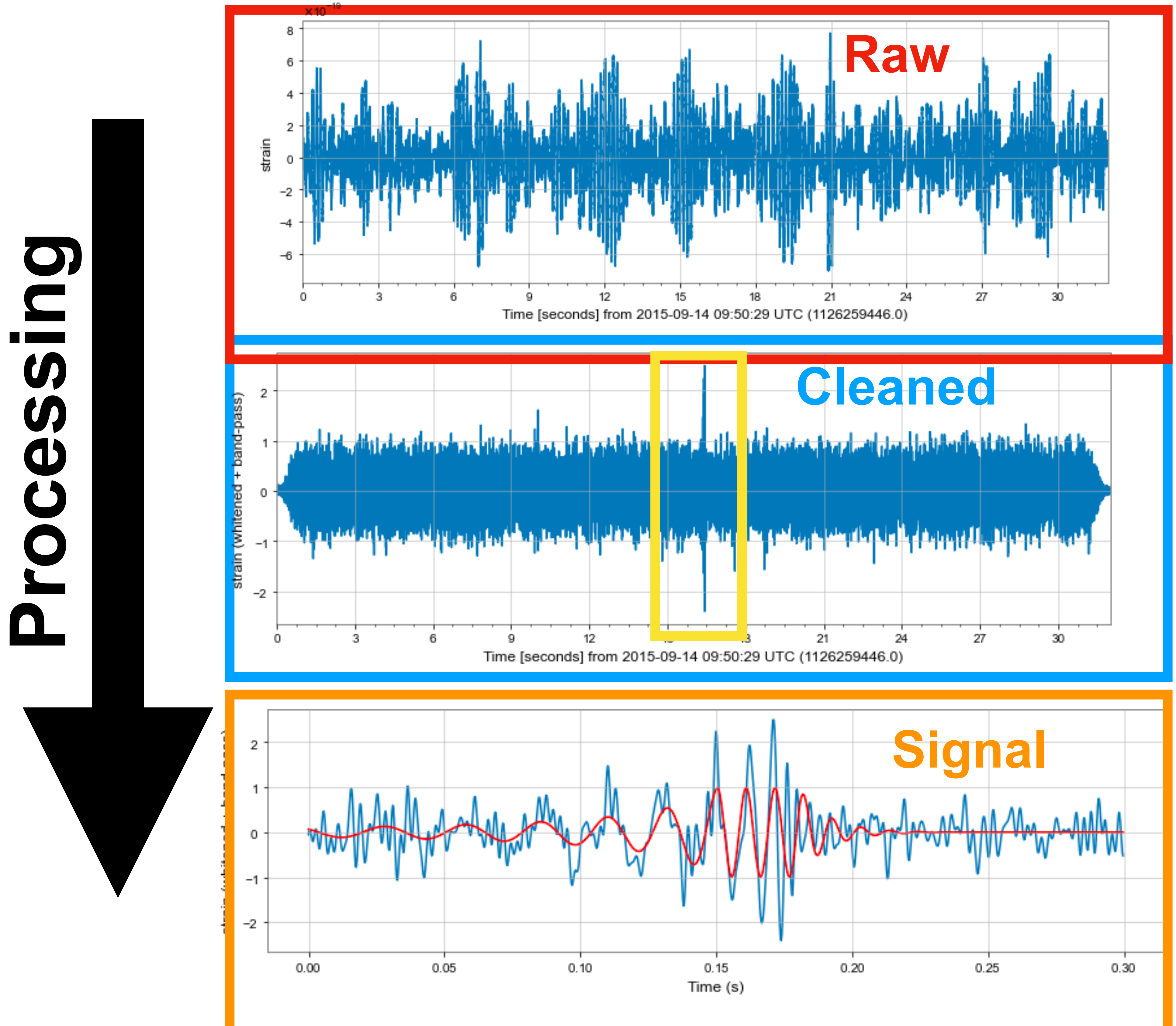
posterior probability distribution for the luminosity distance, marginalized over the whole sky.

LIGO and Virgo Collaborations
Phys. Rev. Lett., 119:141101, 2017

Currently it takes a while to get a good signature

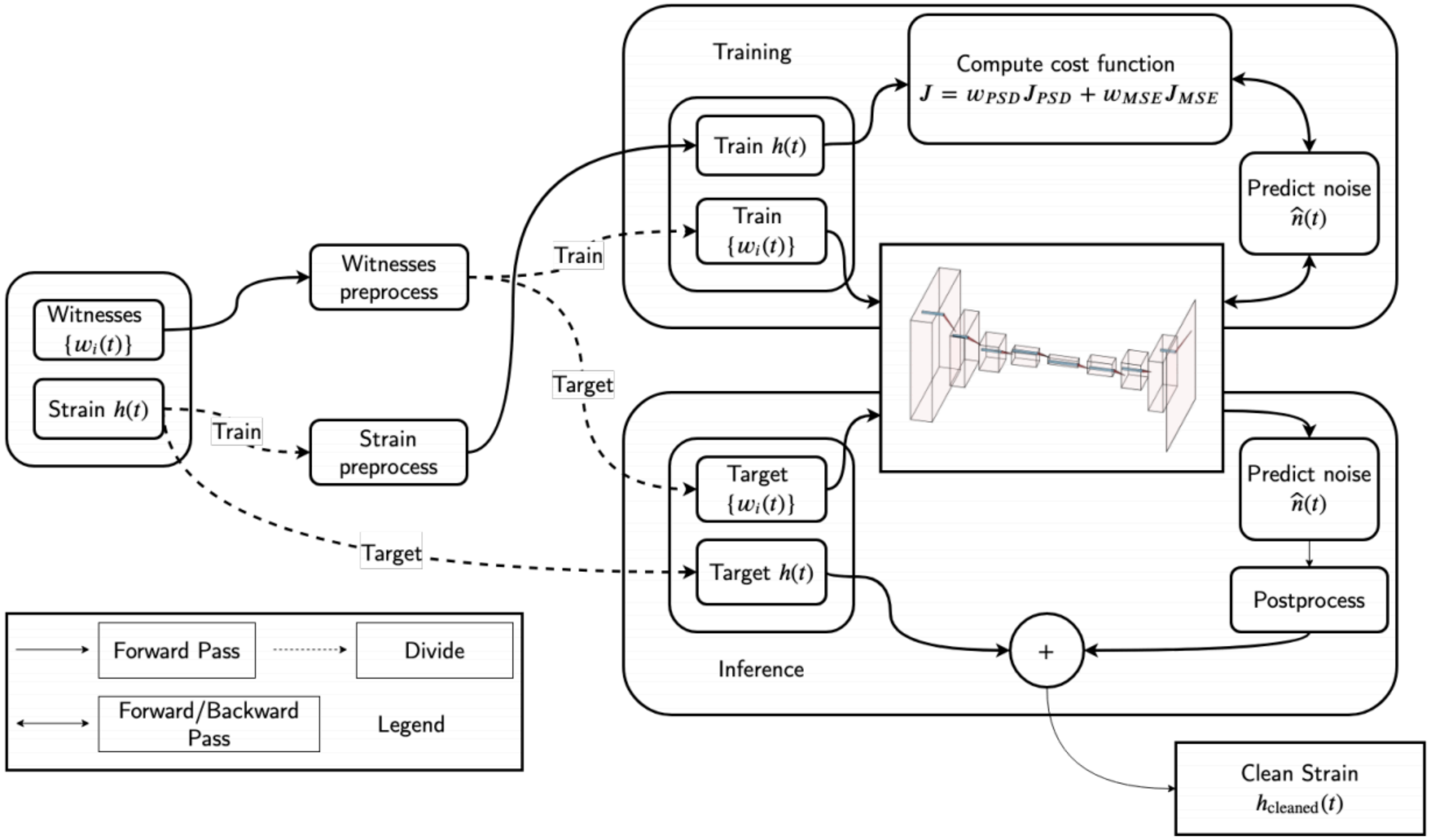
How do we do it Fast?

Preliminary



arxiv/2005.06534

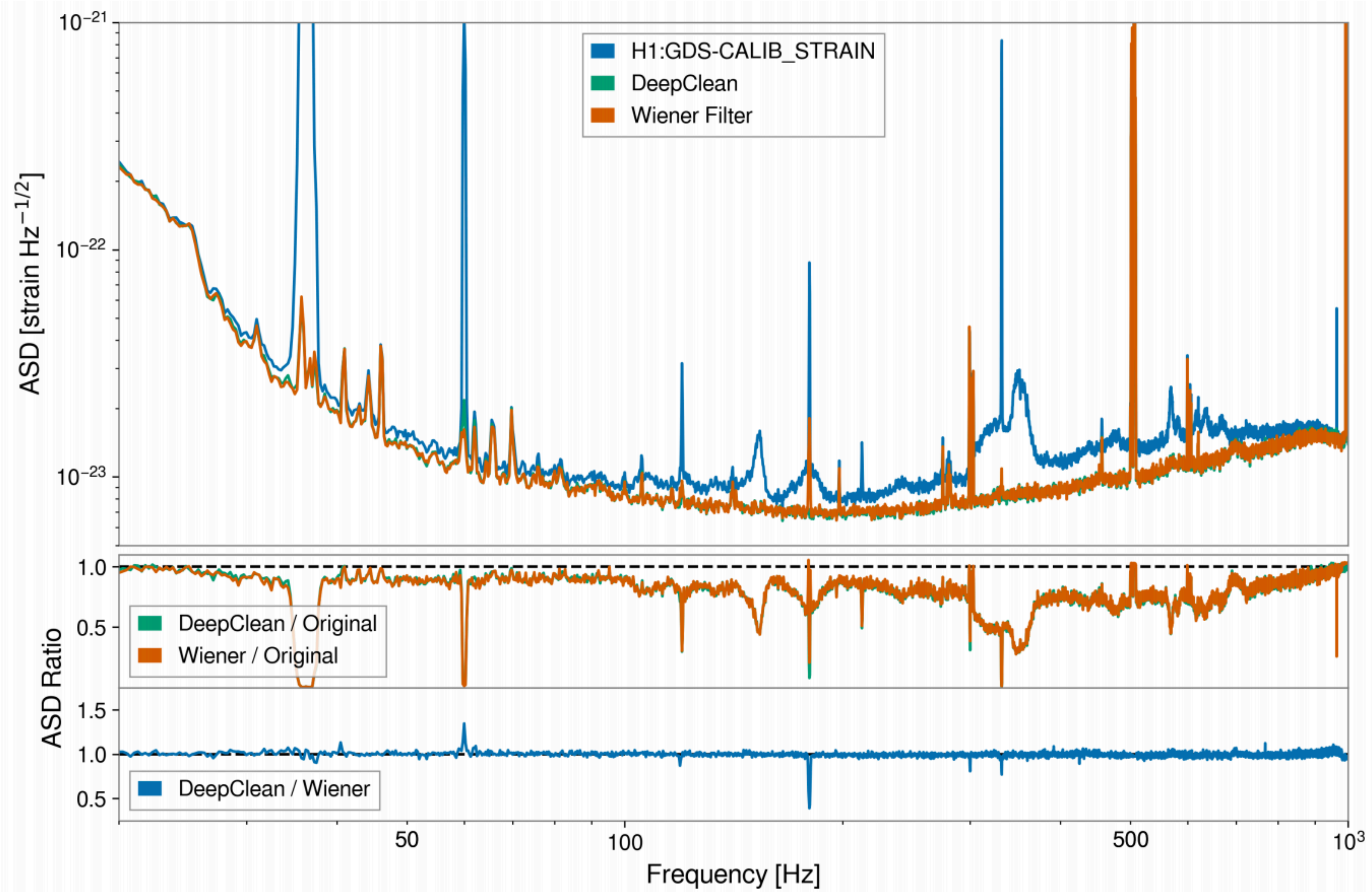
Cleaning the Data



- E. Katsavounids, T. Nguyen have developed a denoising DNN
- Algorithm is an effective AE with conv1d inputs (time series)
    - Lots of room for expansion of project

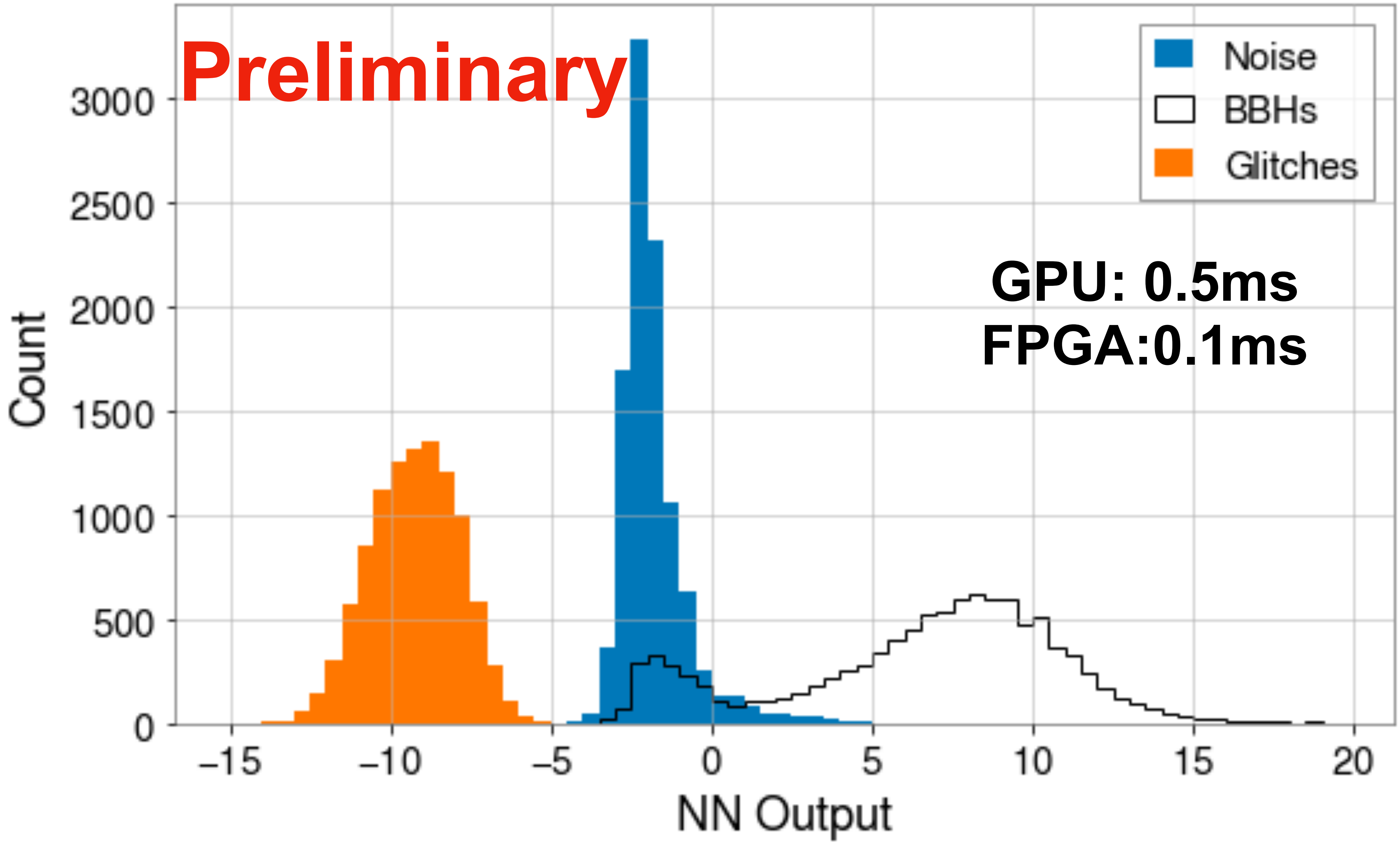
arxiv/2005.06534

Cleaning the Data



- DeepClean performs at the same level as Wiener Filter
- DeepClean can deal with non-linear correlations

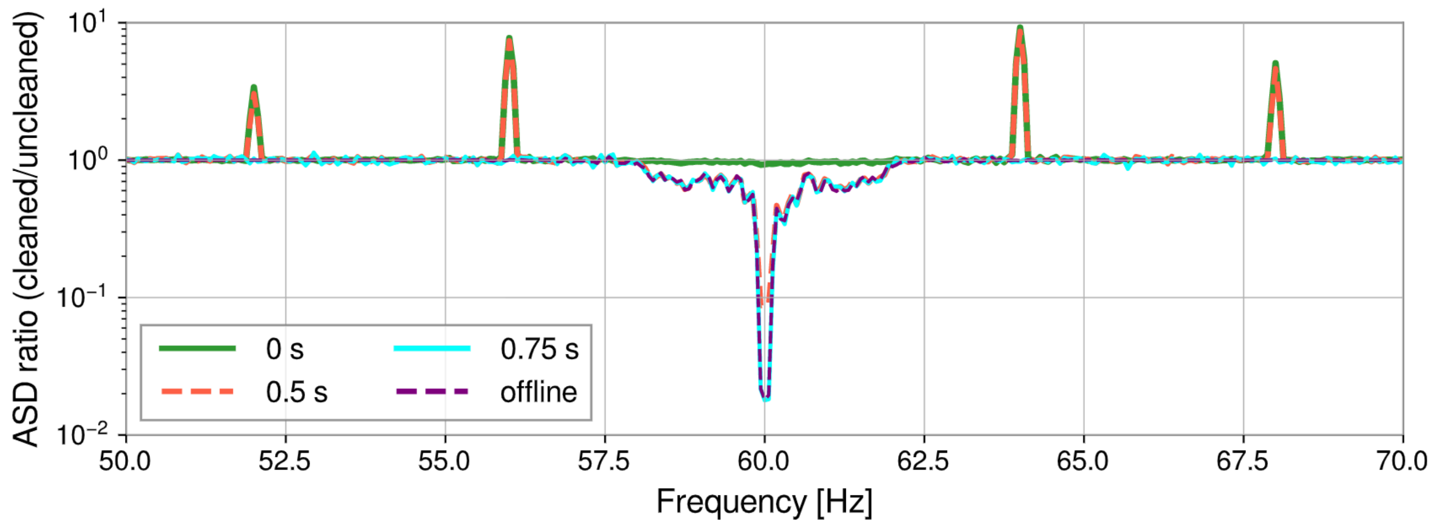
Identifying Gravitational Waves



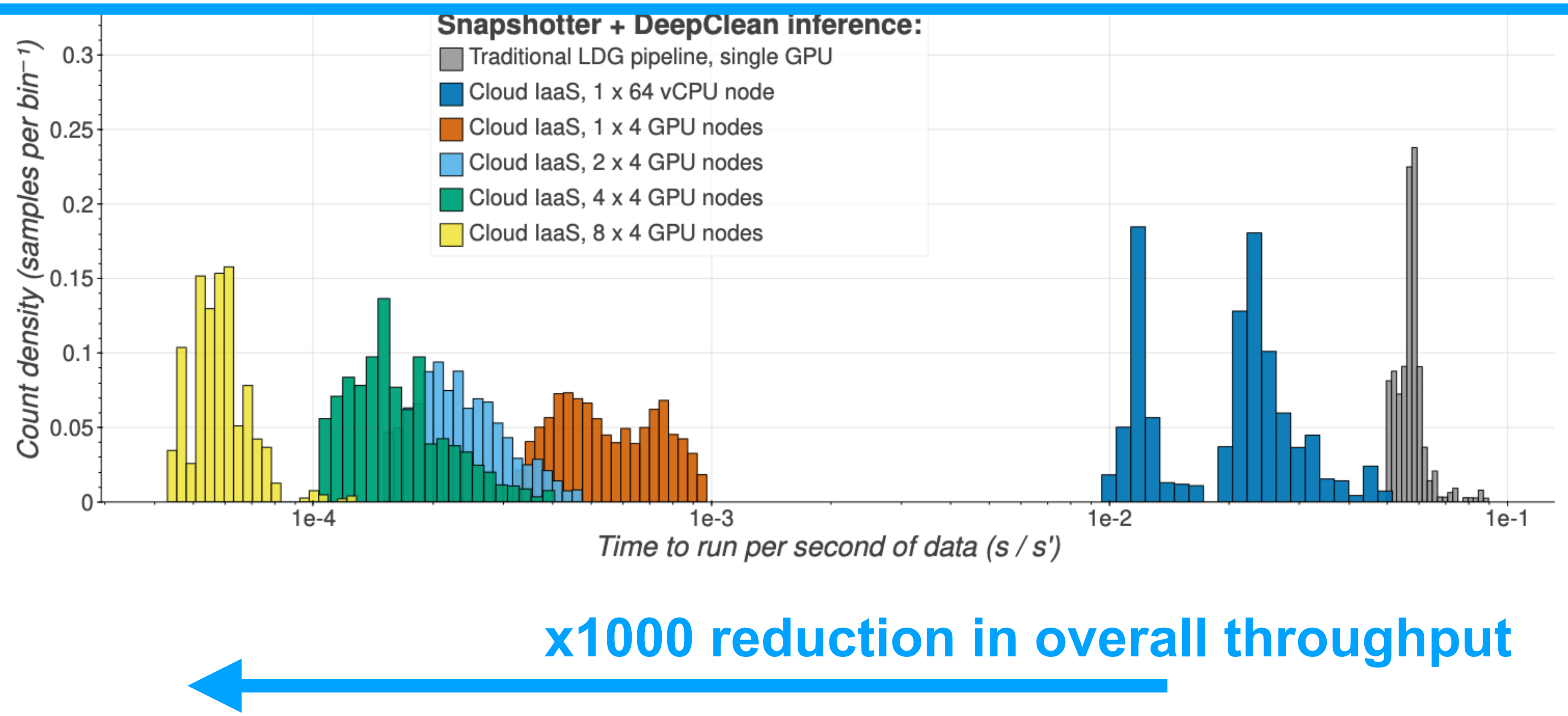
Currently have a preliminary result on fast BBH detection

Gravitational Waves

- Actively building an AI alert system to be deployed at LIGO



Developed AI-based Denoising and BBH detection



Constructed a GPU-as-a-service integration for GW low latency alerts

A3D3

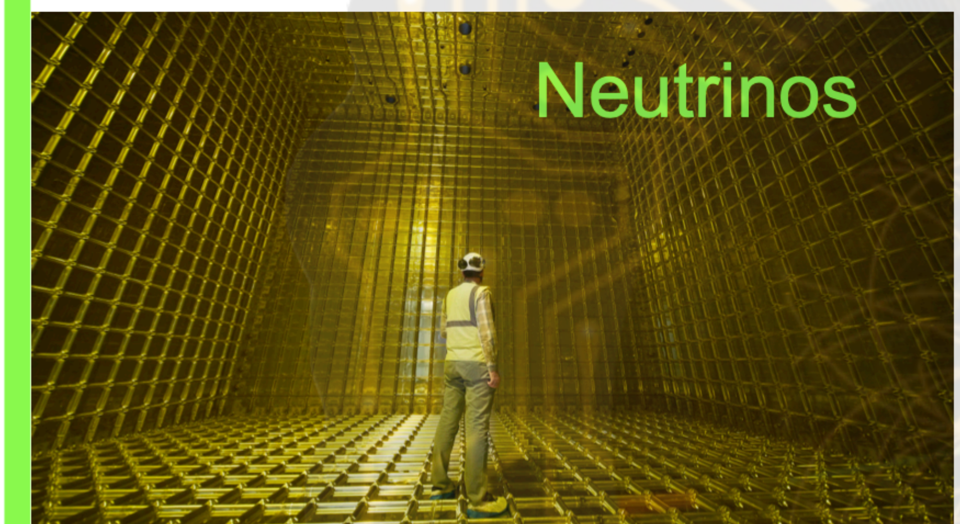
- An institute to unite real-time AI
  - Quickly looking for people to be part of extended team

Accelerated AI Algorithms for Data-Driven Discovery

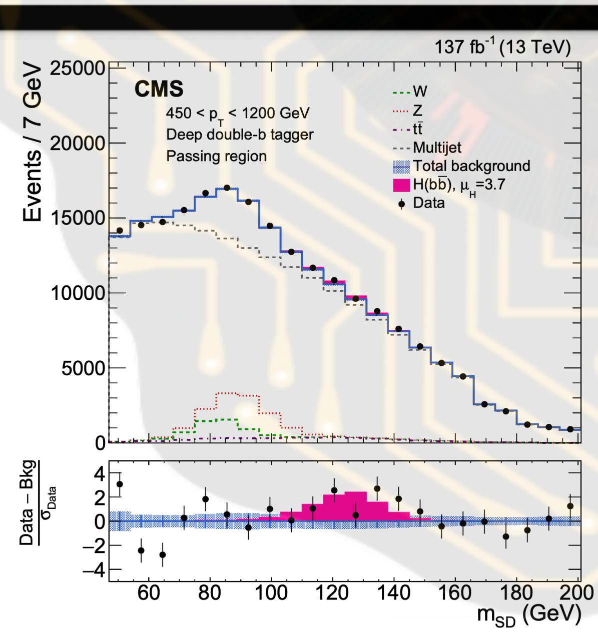
A New Institute: **A3D3**

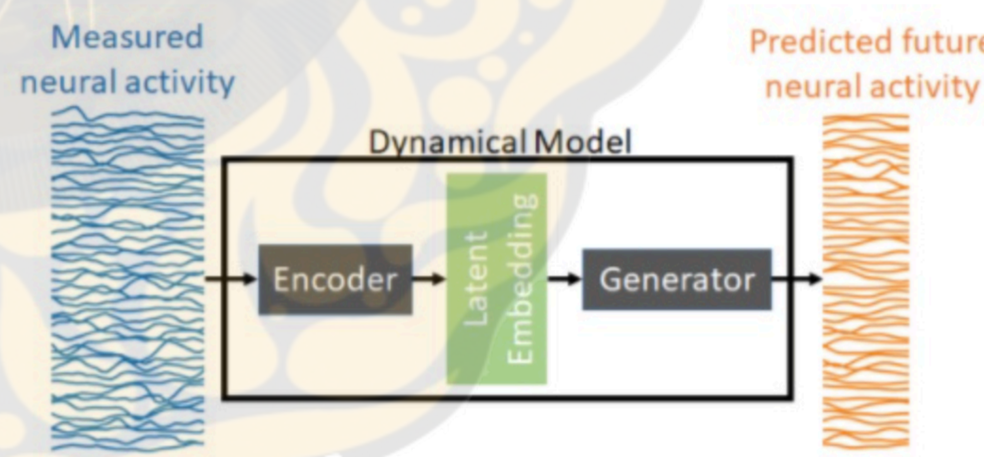
New Types of Computing

- We have been awarded a new institute to explore real-time AI
  - Accelerated AI Algorithms for Data Driven Discovery (A3D3)



LHC Physics

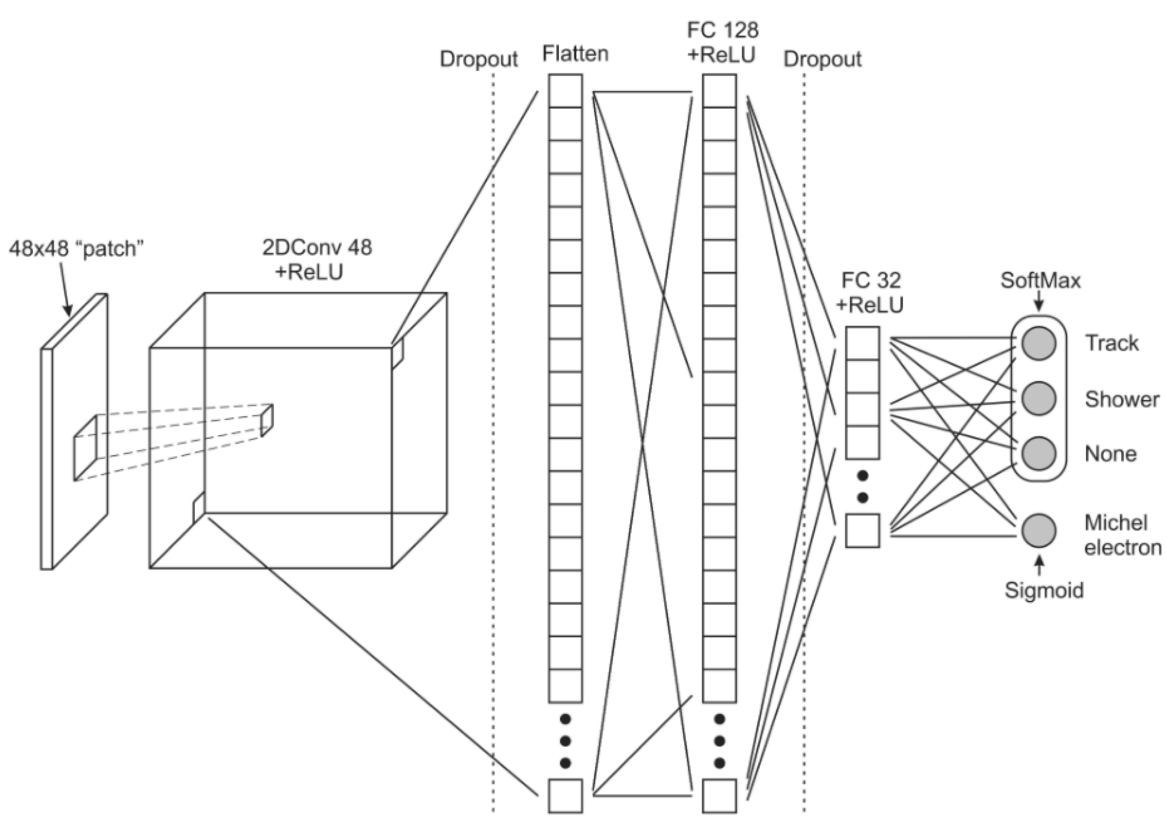




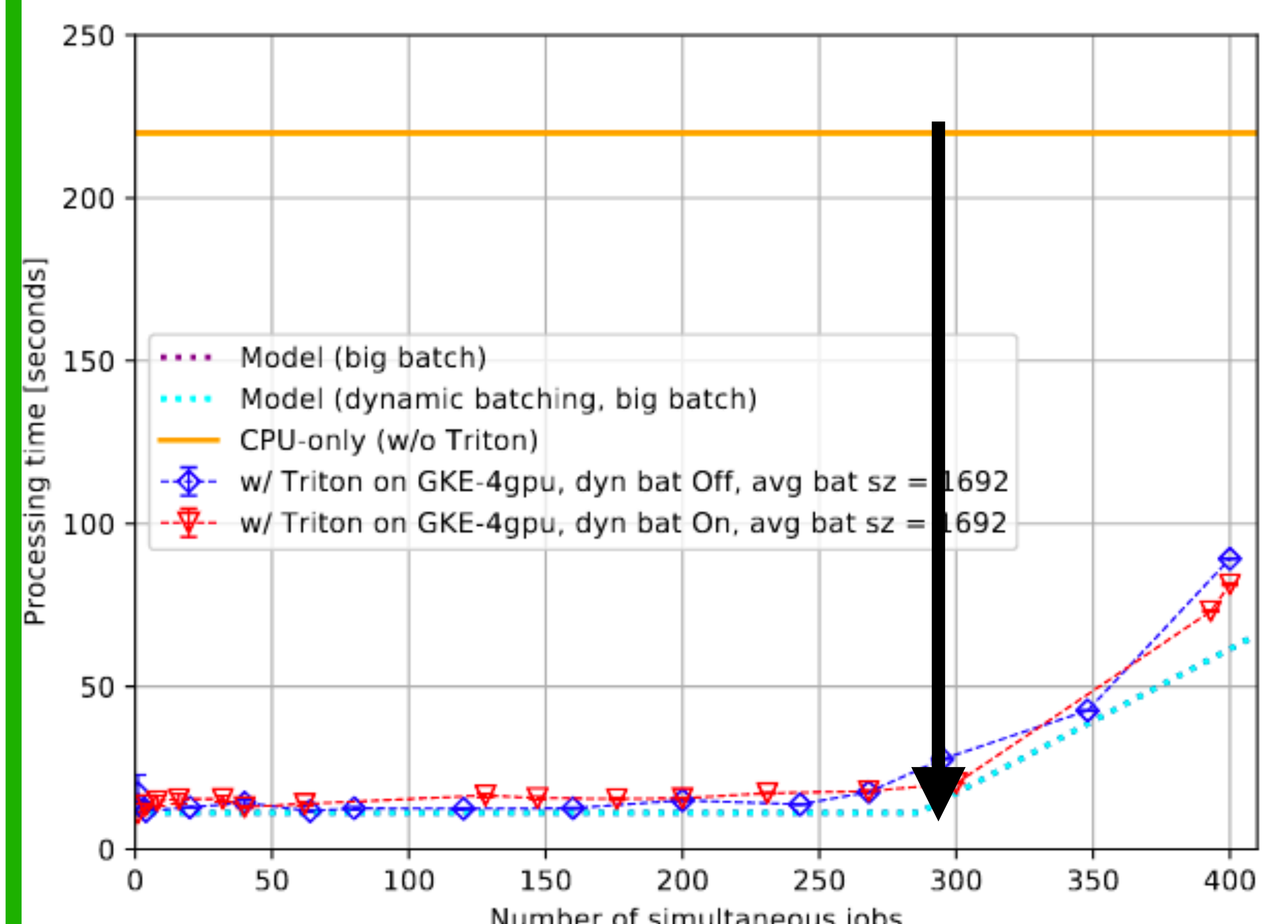
NeuroScience

Neutrino Physics

- We are pursuing the same idea in Neutrino physics



Michel Electron Id NN



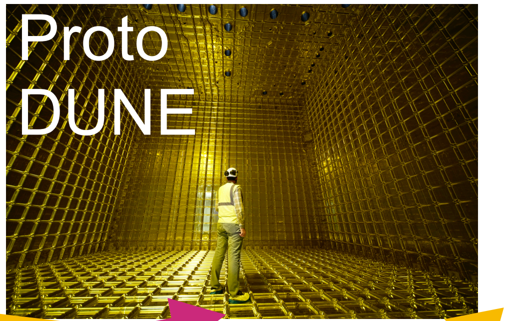
Large Factor in speed up

Overview Venn Diagram

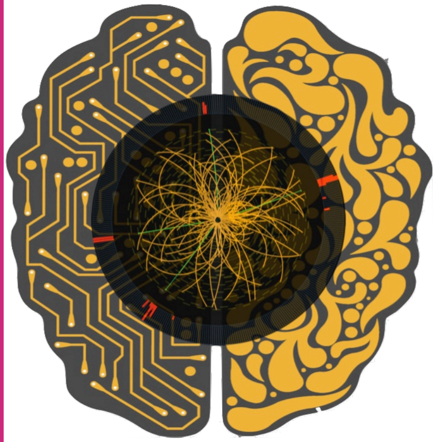
ATLAS EXPERIMENT

CMS

sPHENIX



LIGO VIRGO KAGRA



Fast Machine Learning Lab

Real-Time
Heavy Flavor
Tagging @ sPHENIX

a3d3 Accelerated AI Algorithms for Data-Driven Discovery

Real-time Multi-messenger
Alert

Exploring Clouds
to Accelerate Science

INTERNET2

AI based compression
For Silicon calorimeter
Readout (DOE ASCR)

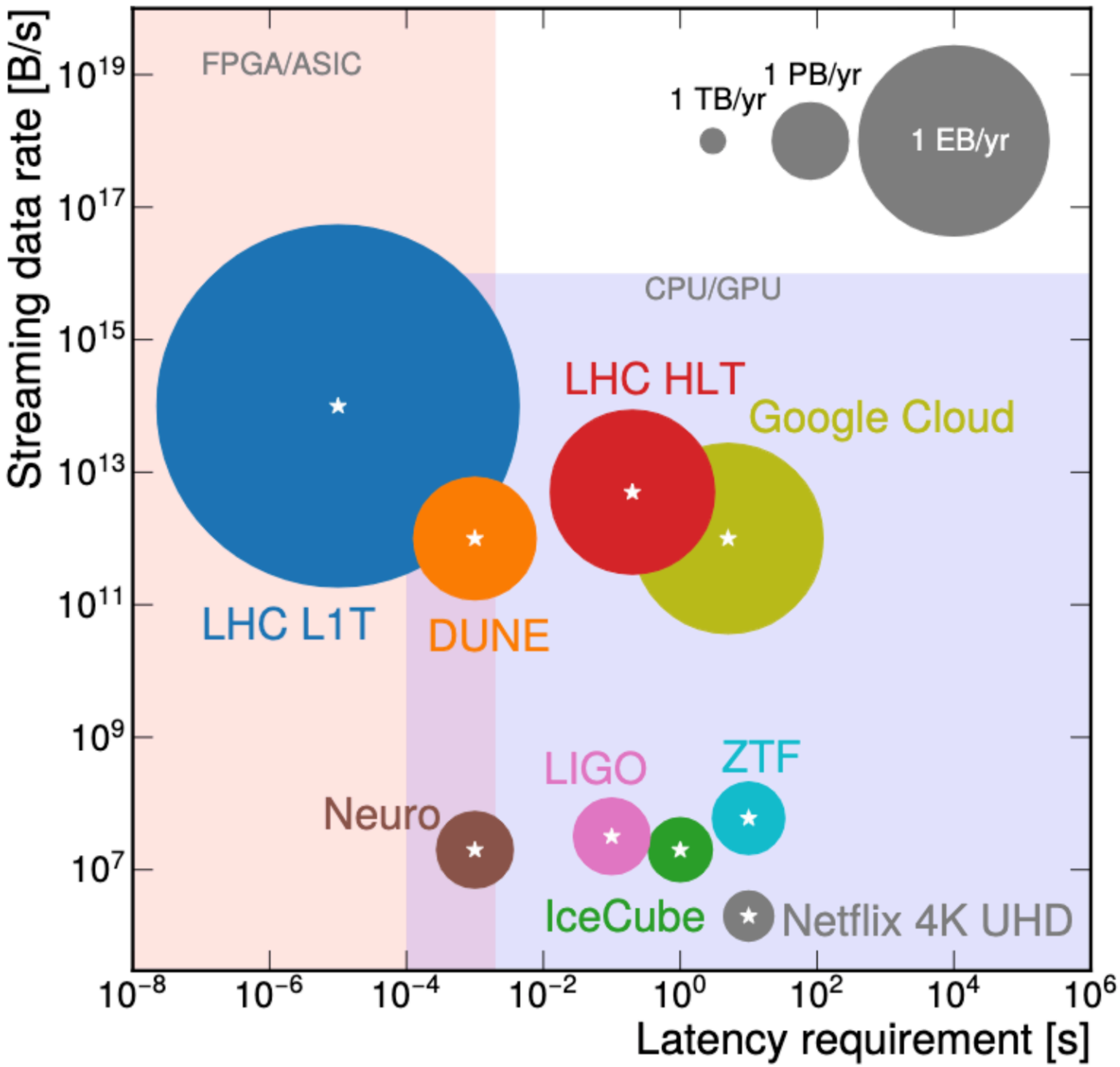
hls4ml

FAIR4HEP

AI Algorithms (AI²)

IAIFI

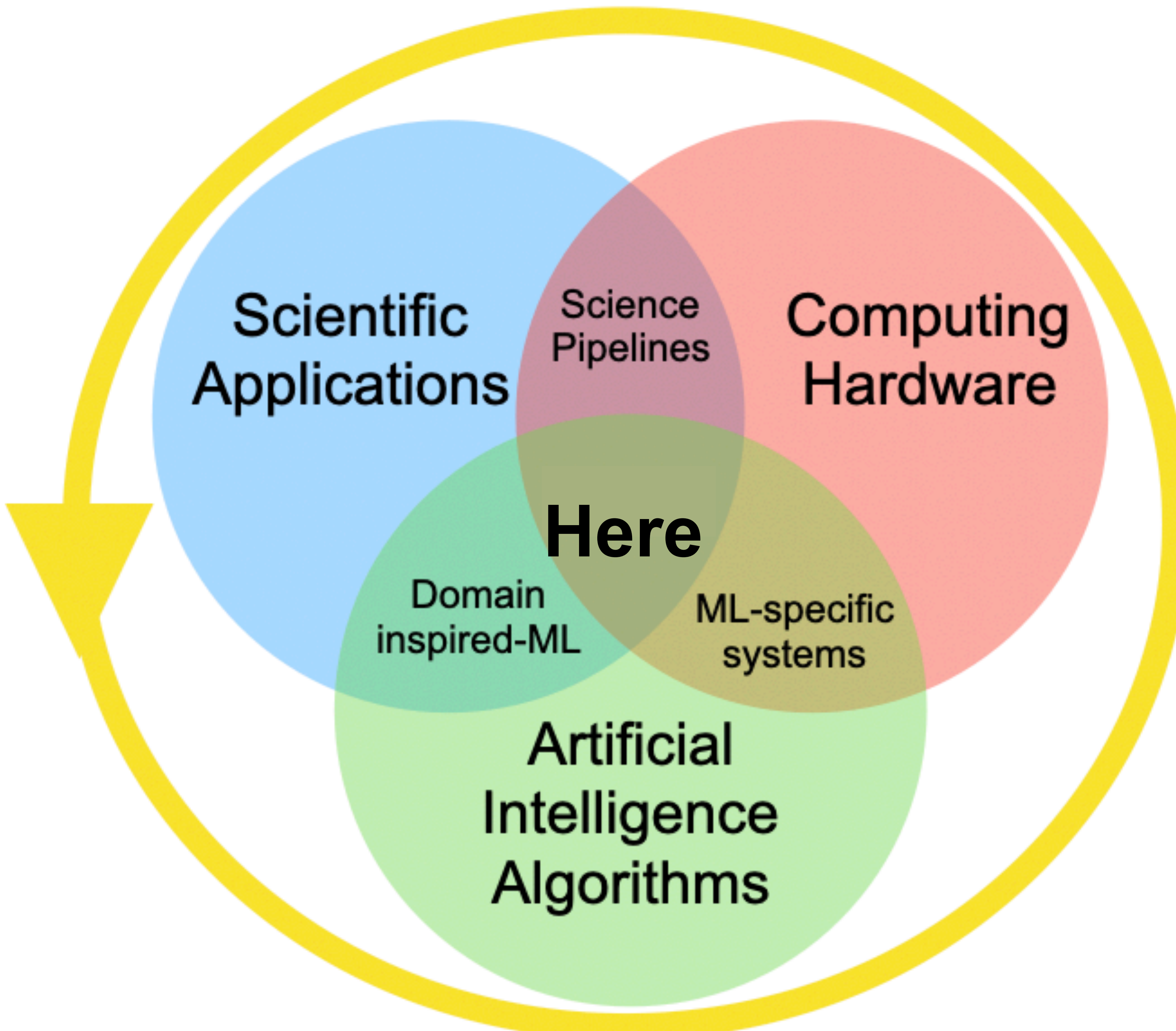
Preparing for the future



Who are we?

https://fastmachinelearning.org/

- Project started by adapting deep neural networks to LHC data flow
- Collaboration is now > 100 members at 10 institutes (2 years old)
- Our aim : bring the fastest machine learning to science


Scientific
Applications
Science
Pipelines
Computing
Hardware
Here
Domain
inspired-ML
ML-specific
systems
Artificial
Intelligence
Algorithms

Conclusions

Real time deep learning

In science has the potential to open new doors

Thanks!

INTERNET 2

NSF

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

XILINX

ALL PROGRAMMABLE™

Microsoft

MIT Quest for Intelligence

Google Cloud Platform

aws

Fast ML Team

Fermilab

Massachusetts Institute of Technology

UNIVERSITY of WASHINGTON

UNIVERSITY OF TORONTO

PURDUE

Microsoft

XILINX

UCSD

CERN

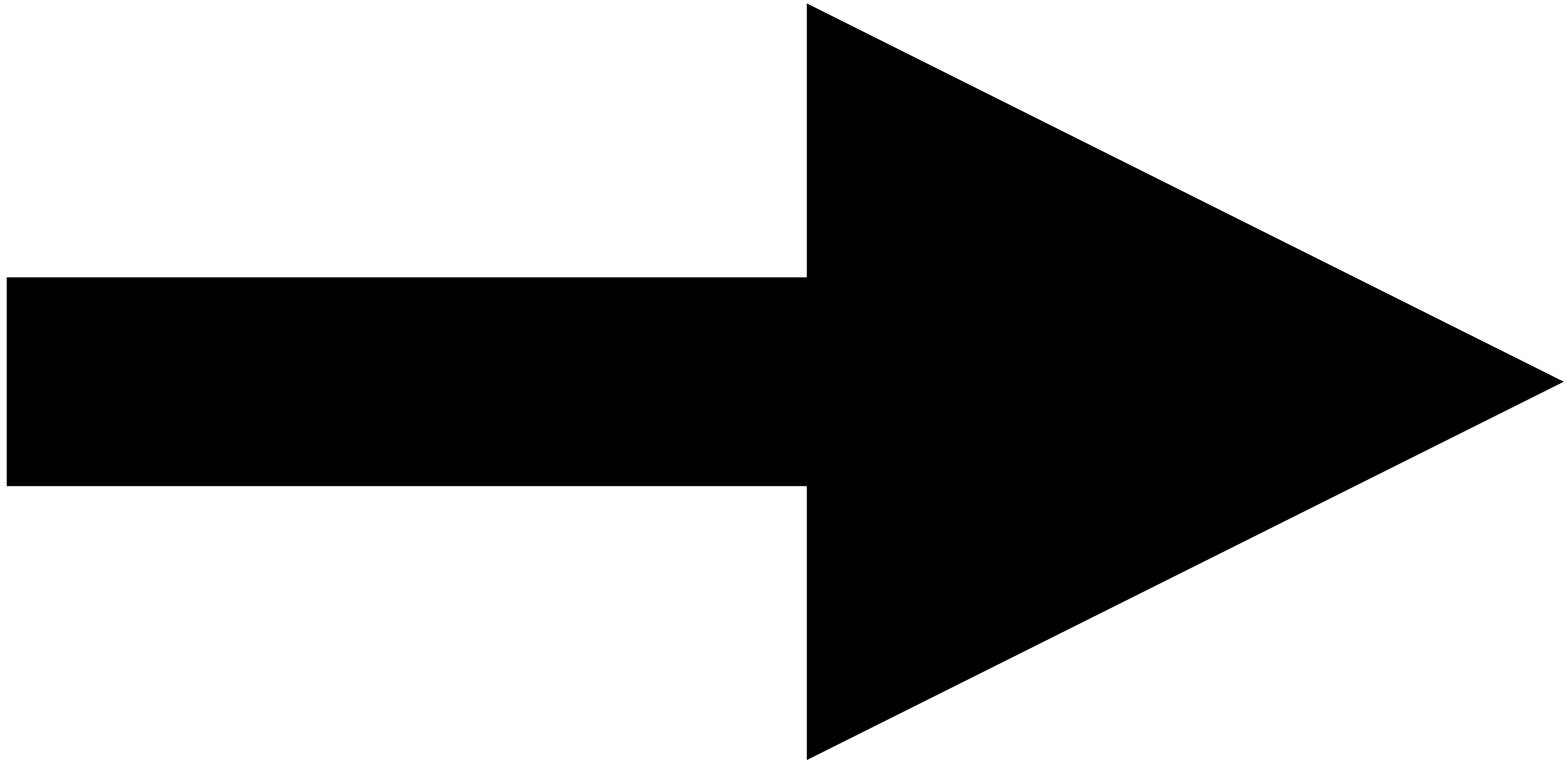
ILLINOIS

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

HawkEye 360

NORTHWESTERN UNIVERSITY 1851

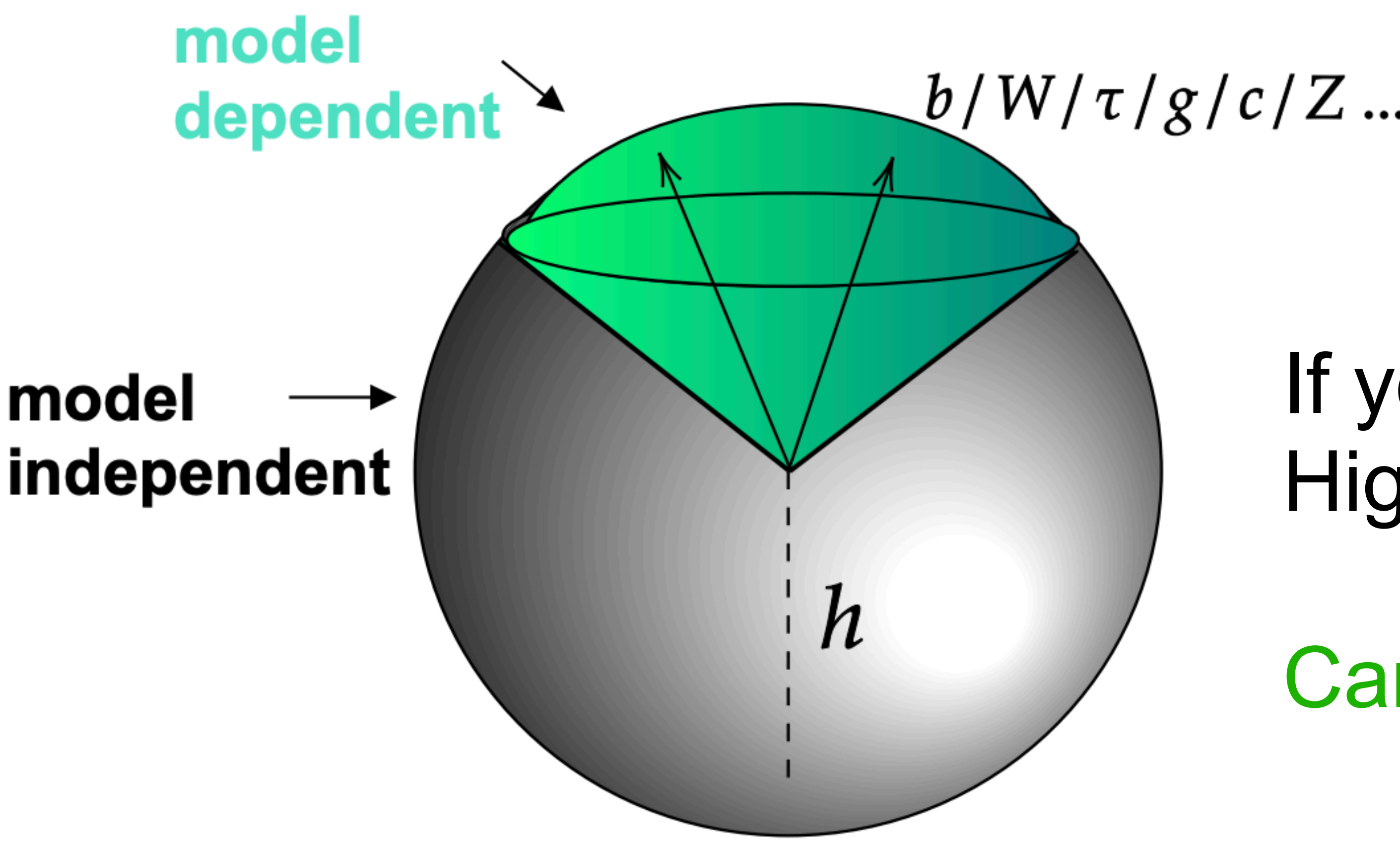
Right Brain



This is a story of
an IAIFI Collaboration


CTP 2019
IAIFI Prep Meetings

Stuck on a Problem



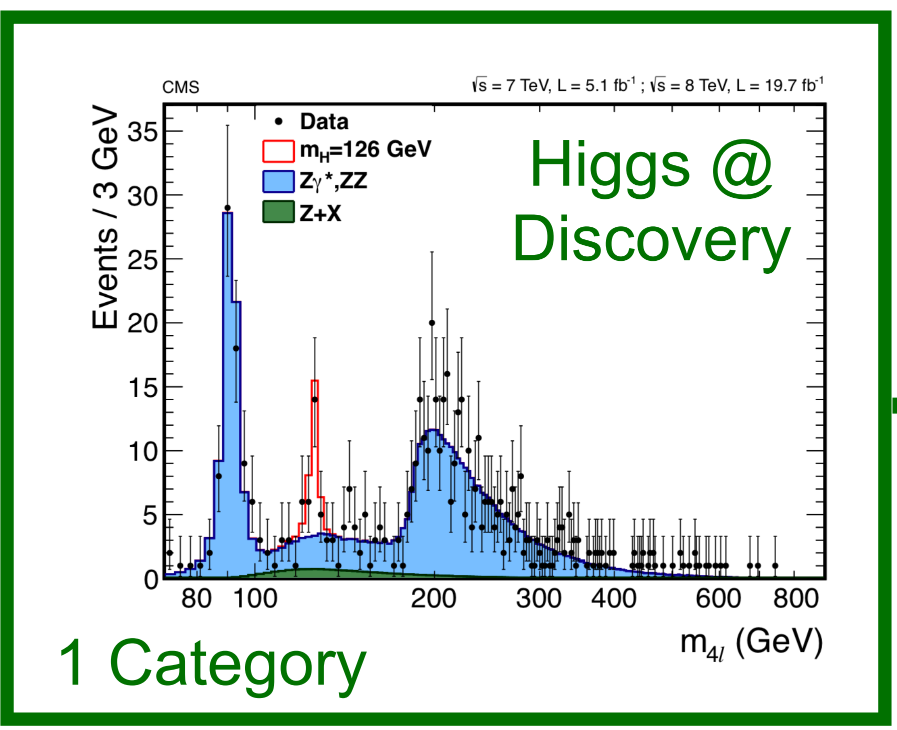
If you can make an inclusive Higgs boson measurement

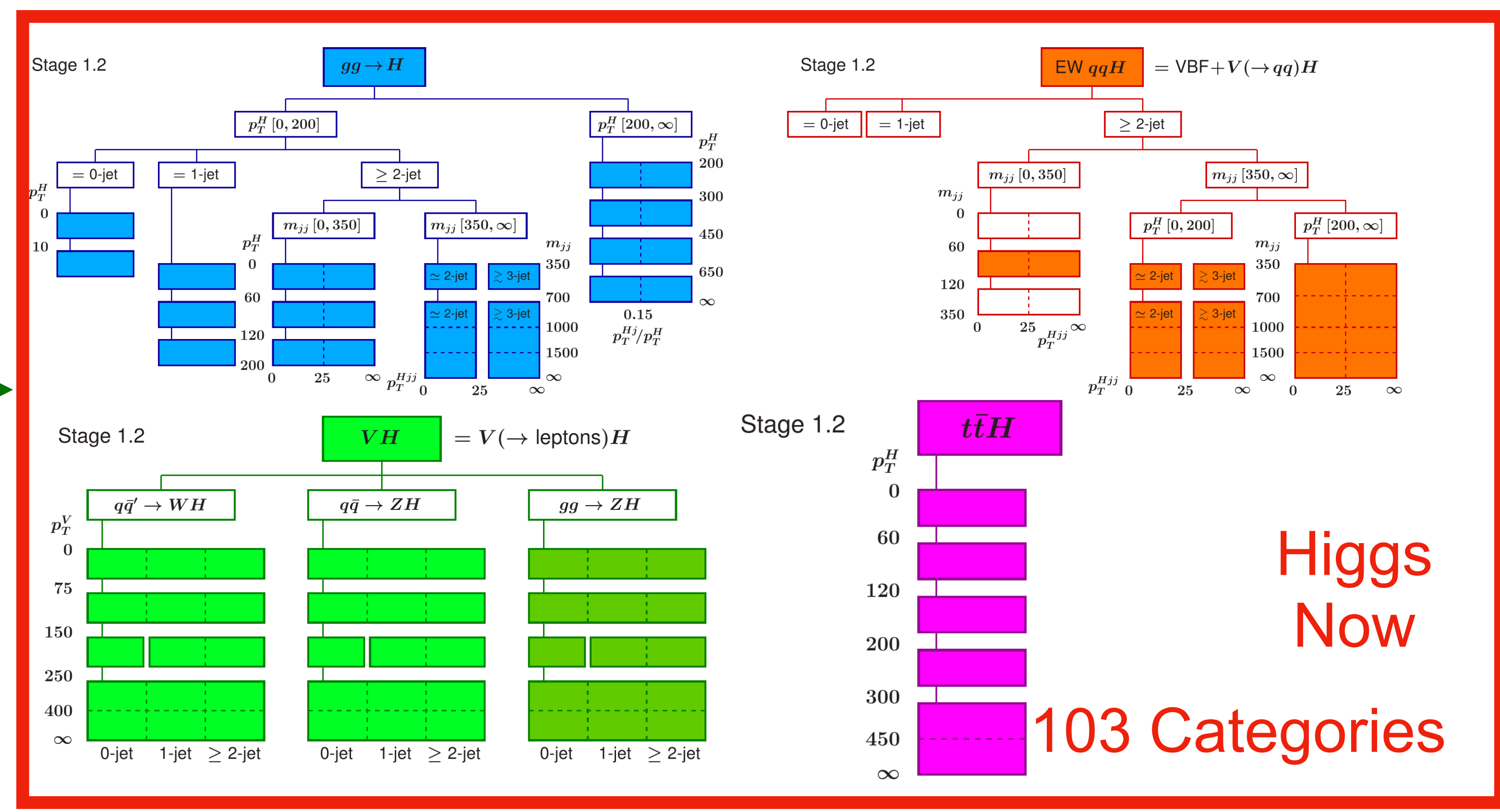
Can measure the total width

How do you search for every final state at once?

Ageing Analyses @LHC

- Data analyses at the LHC are changing
  - Analyses are becoming much more complex
    - Many categories and many final states
- General trend towards more complicated analyses



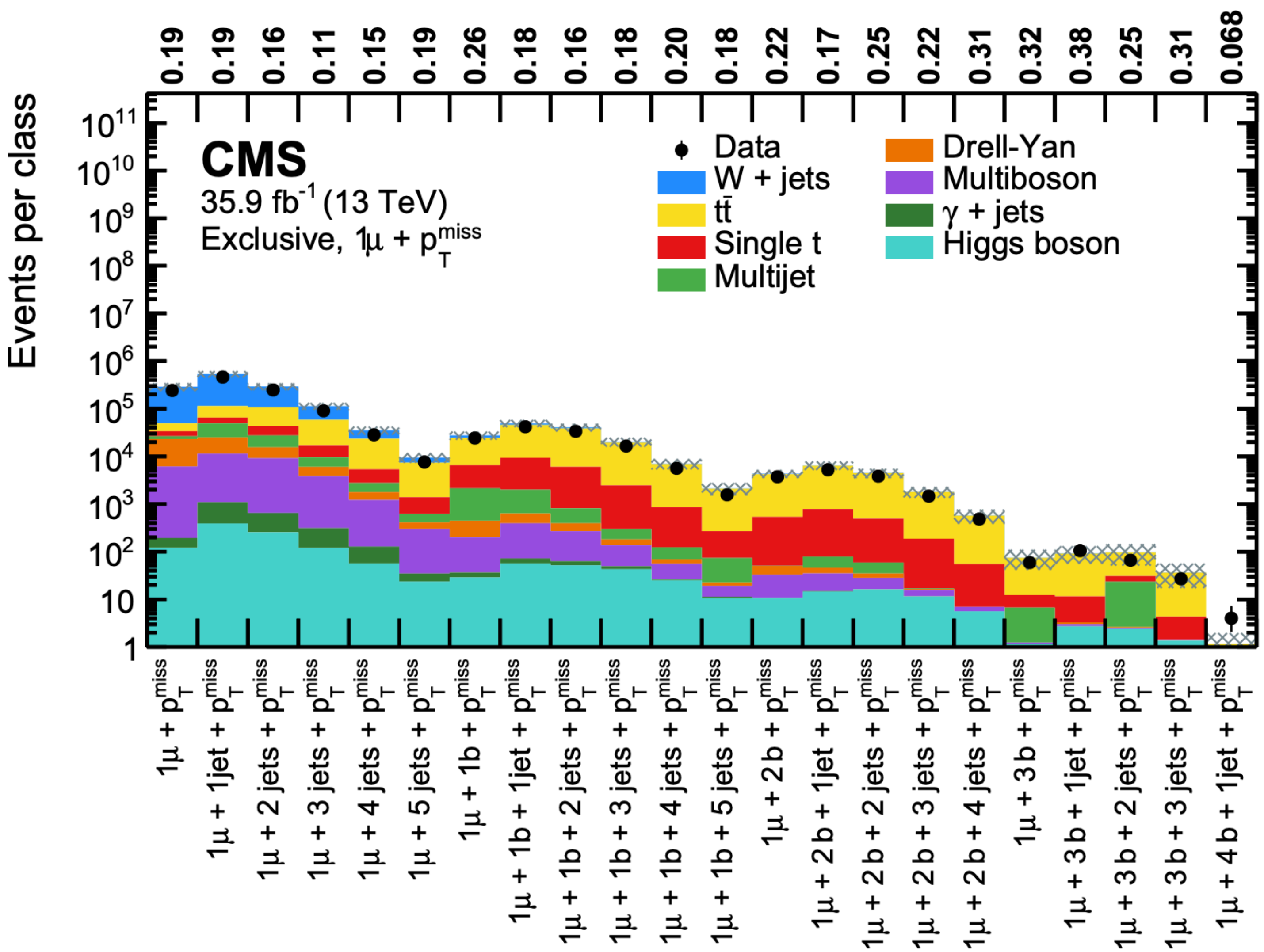


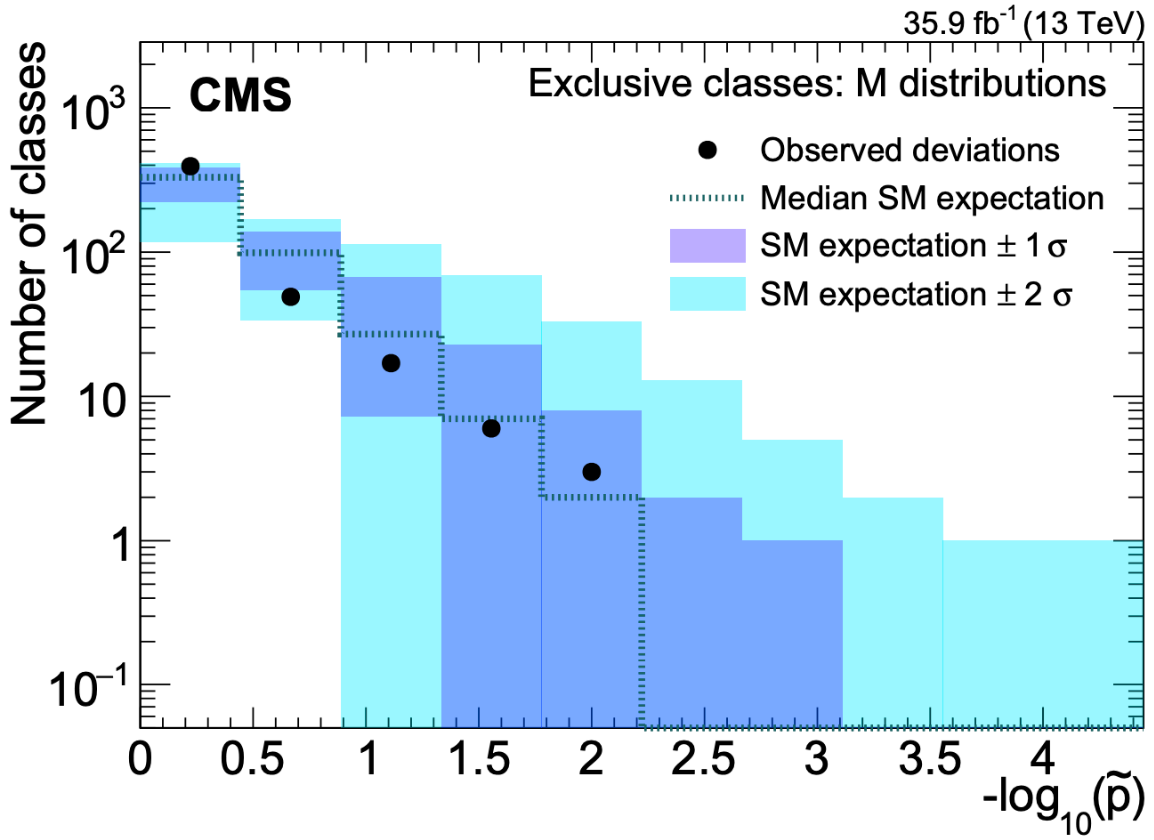
What has caused trend?

- The power of computing
  - Complex many parameter fits run much faster these days
  - Newer optimization strategies that are proven to be robust
  - Along with the ease of use of complex fitting tools
    - Many tools now auto build likelihood and minimize
- A better understanding of our simulation
  - Many processes are understood
  - Steps to making categories has become progressively simpler
- Encroaching on a general philosophy to do more in one swoop

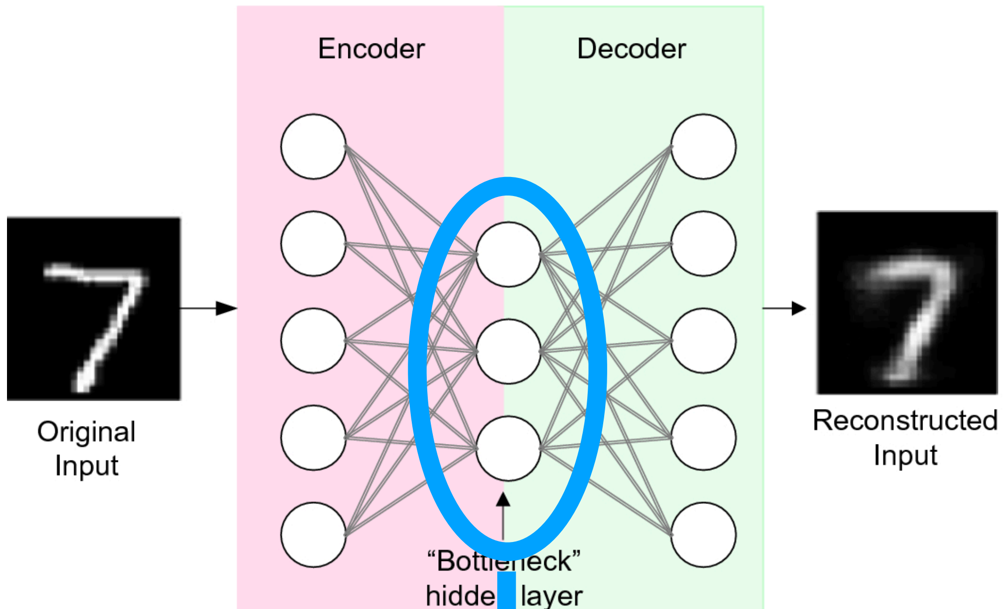
From this trend

- Some old ideas are starting to be taken more seriously
    - Can we perform analyses on a broad range of data at once



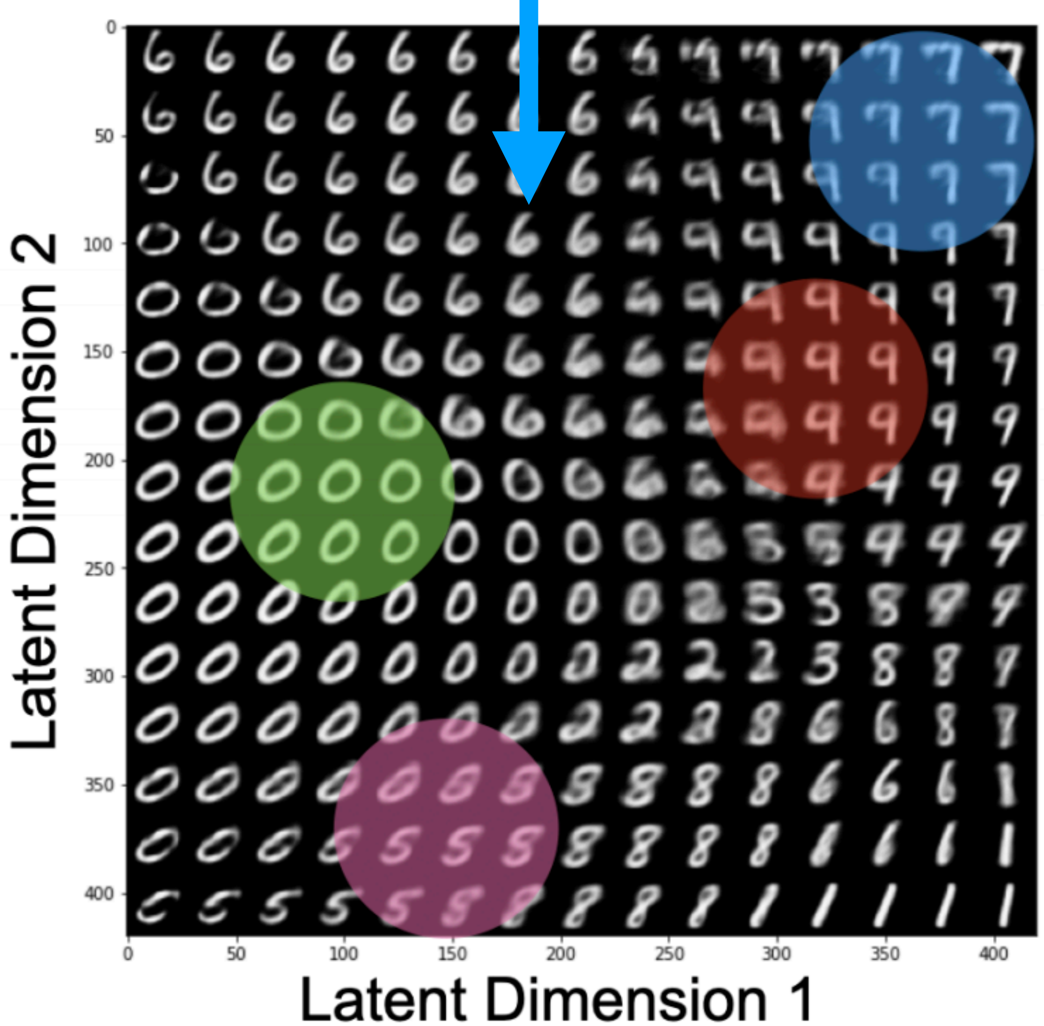


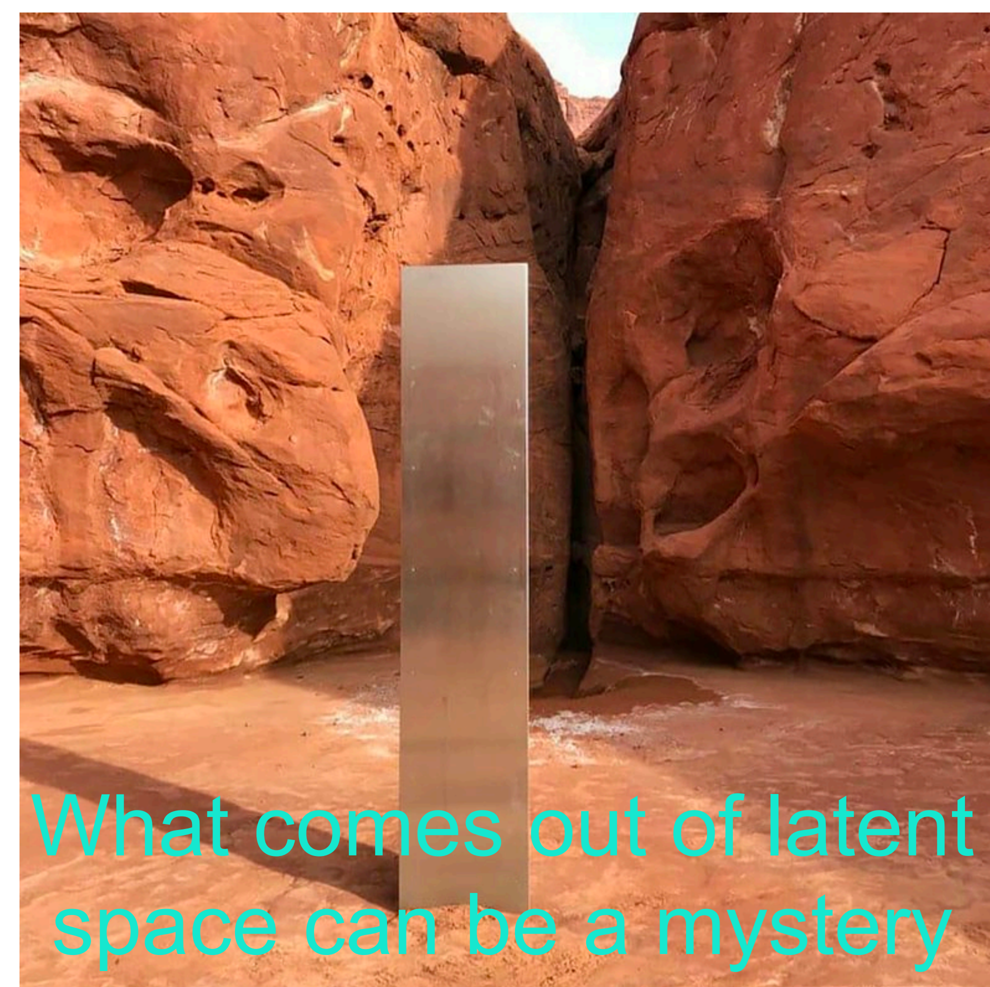
The Latent Space



Latent space aims to organize the information

Normalizing Flow allow for adaptive capture of physics





One-Shot Learning

One-shot learning aims to build a space of similar objects

Normalizing Flow

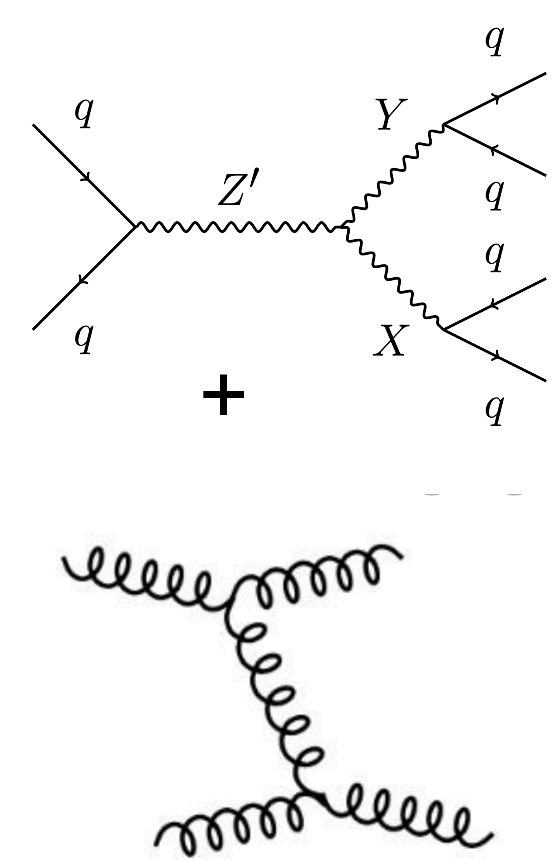
Similar

Our idea:
Normalizing Flow to build
a latent space of physics objects

Towards Having it all

- Can we search for an arbitrary signal and find it?
- There was a recent challenge to look at this:
  - LHC Olympics 2020

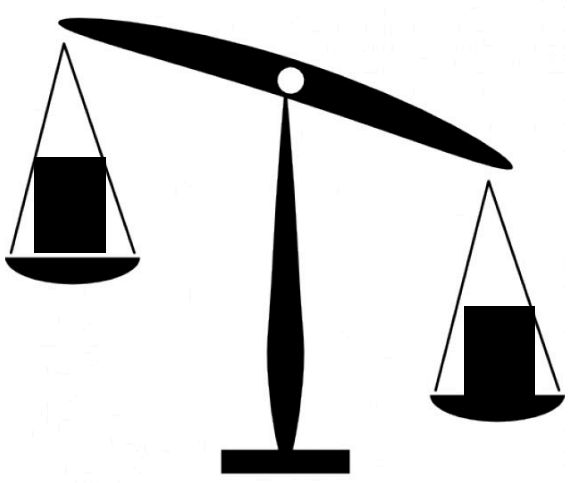
Black Box



Anomaly Strategies@LHC

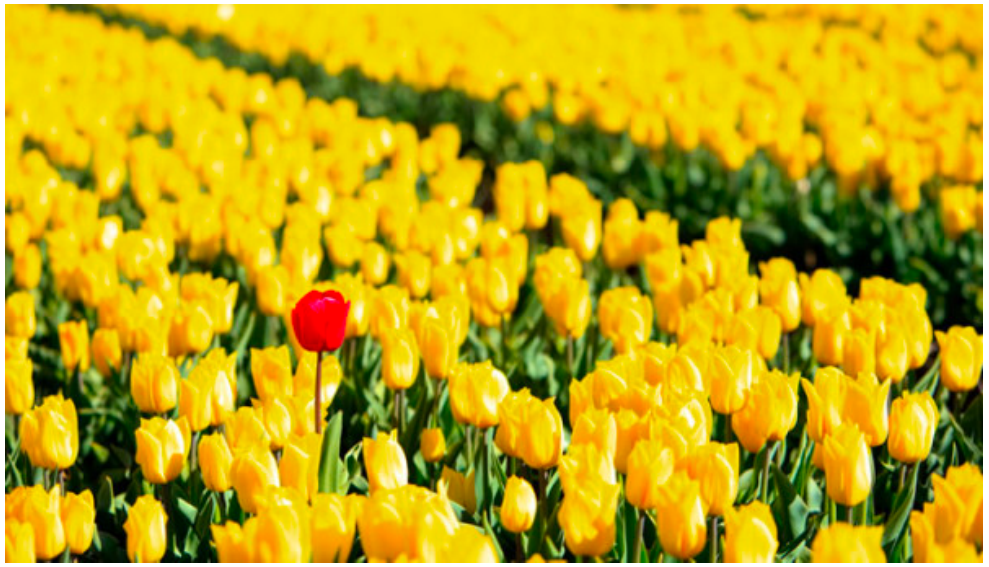
- Anomaly Strategies at LHC fall into two categories

I know regions where new physics does not exist



I want to leverage those regions against other parts of the data to find differences

I know how to predict all collisions



Are there any collisions that I cannot predict?

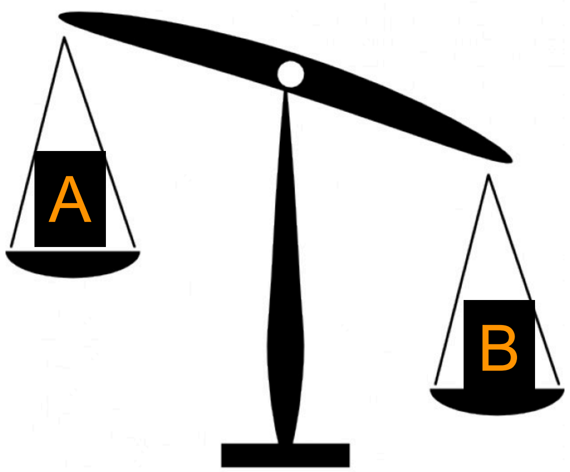
Anomaly Strategies@LHC

- Anomaly Strategies at LHC fall into two categories

Weakly-Supervised

I know regions where new physics does not exist

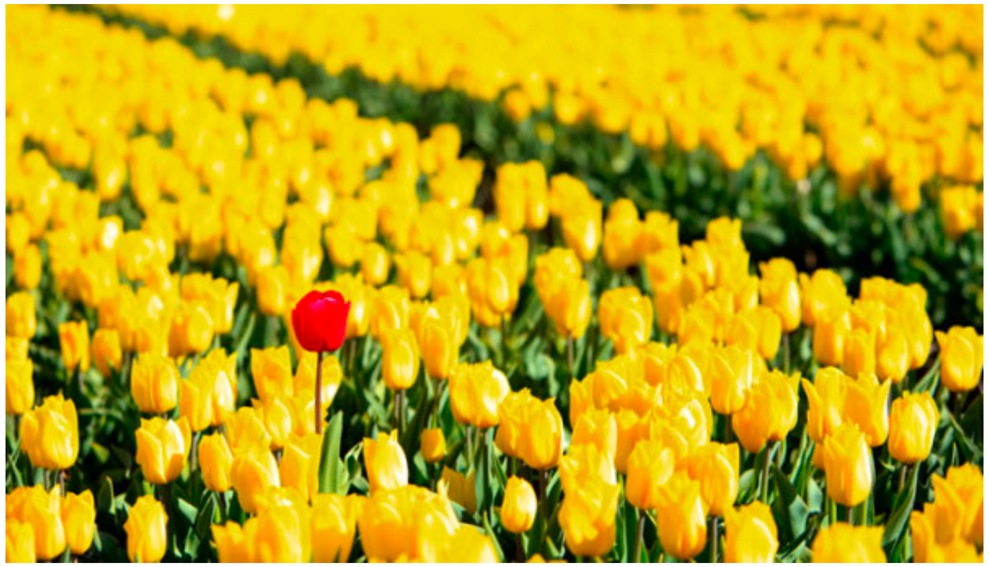
Classification W/O LAbels



I want to leverage those regions against other parts of the data to find differences

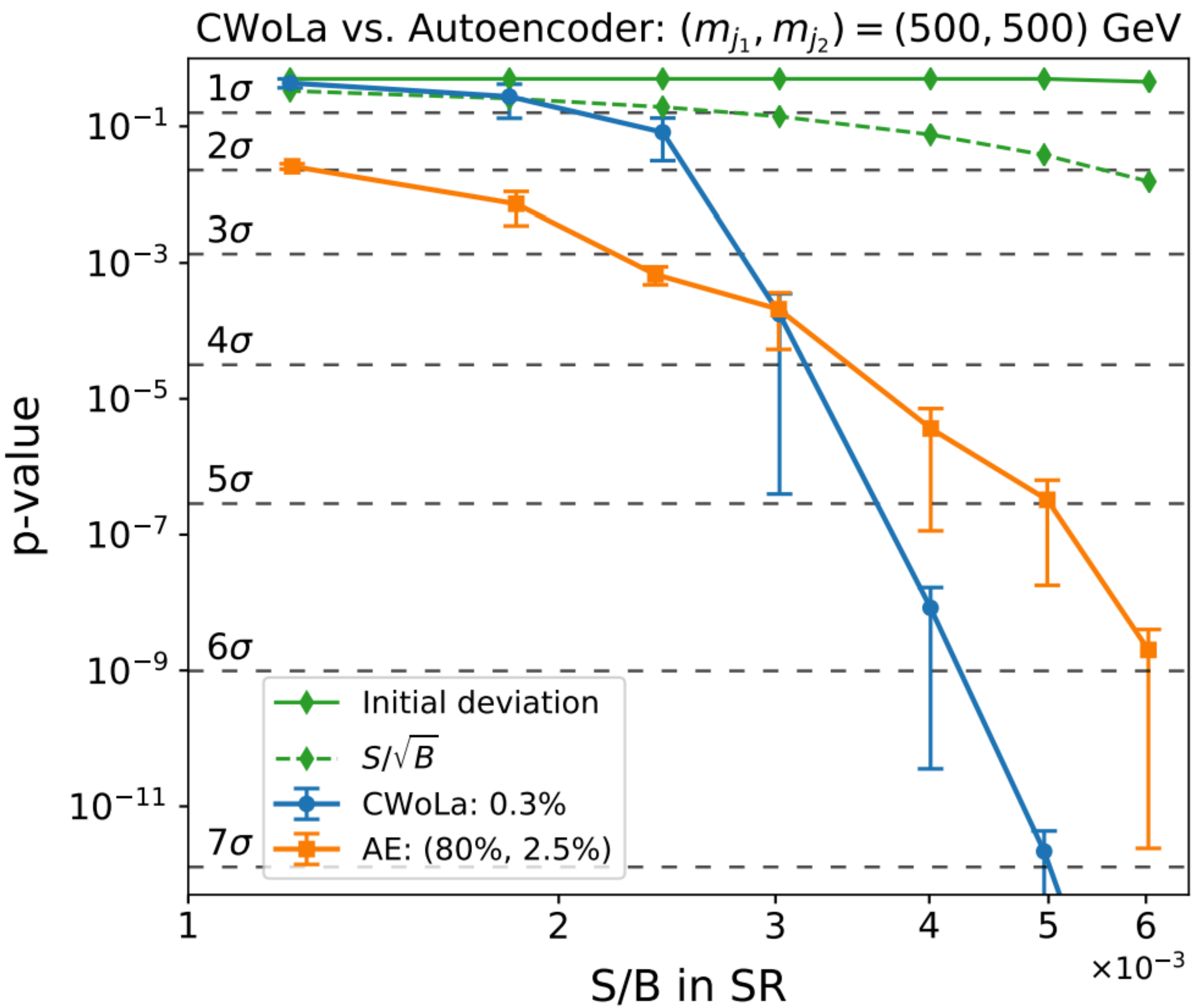
Autoencoders

I know how to predict all collisions

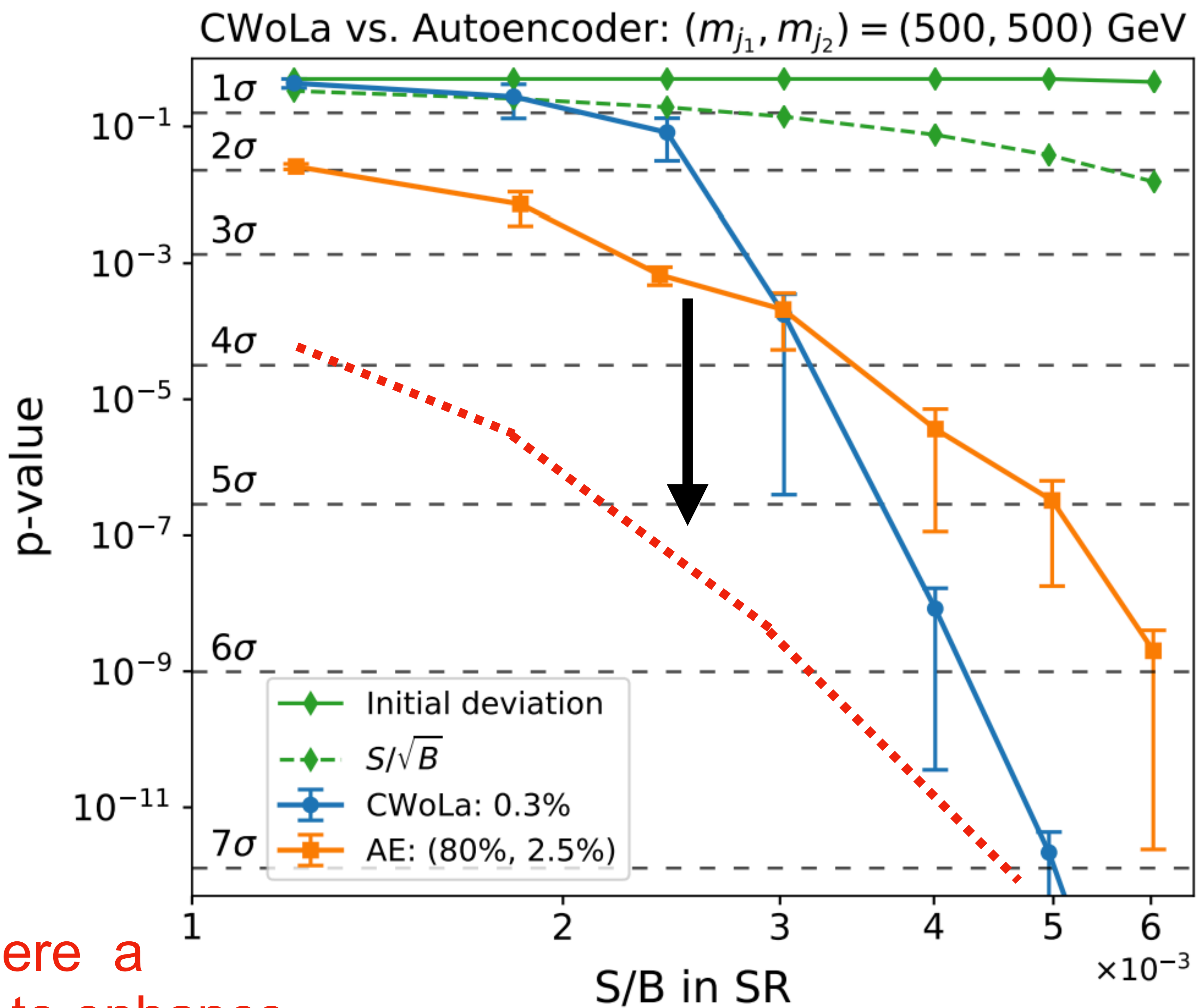


Are there any collisions that I cannot predict?

Performance Observations



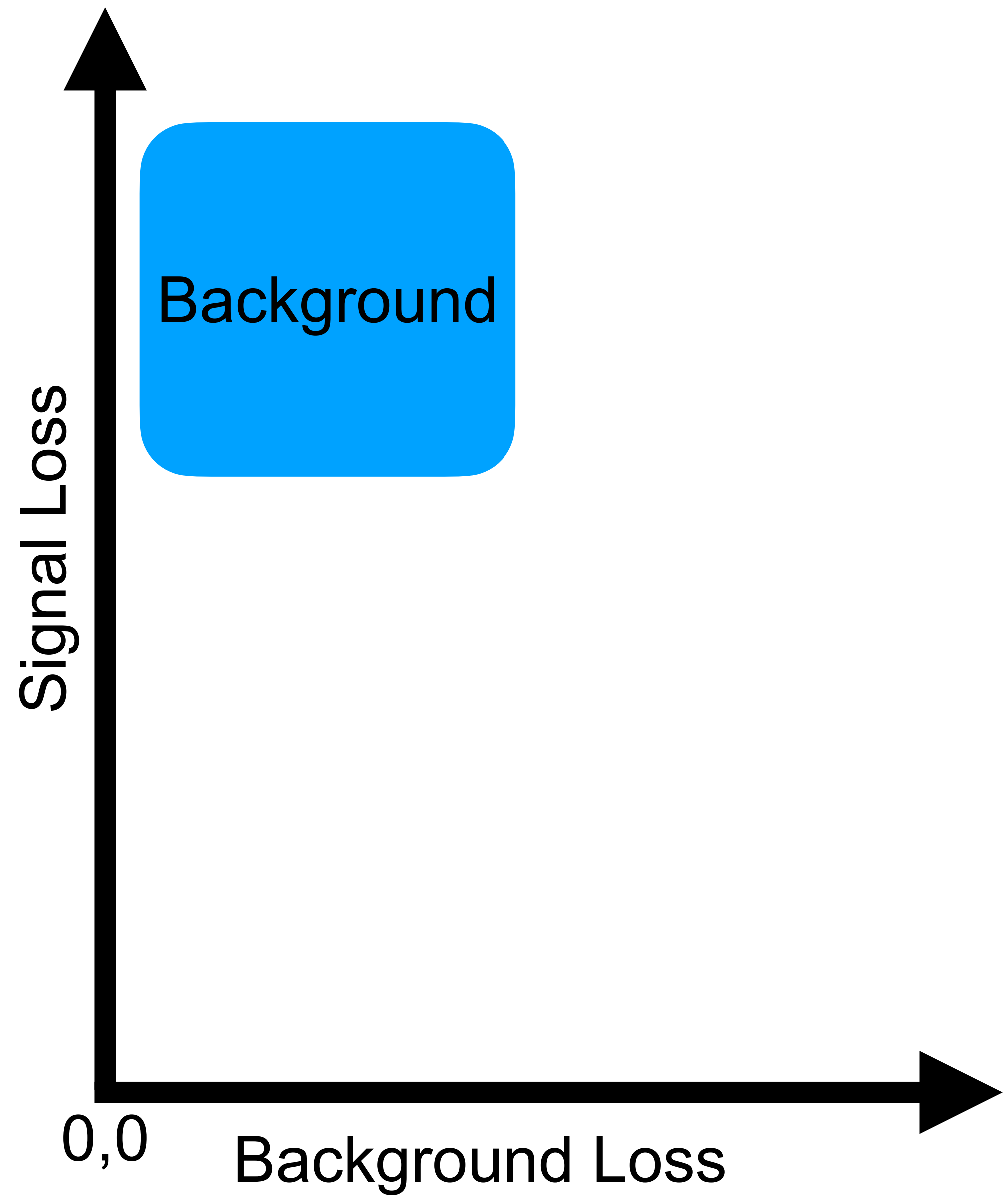
Performance Observations



Is there a way to enhance signal at low S/B?

QUasi Anomalous Knowledge

Normalizing
Flow
Trained
On signals

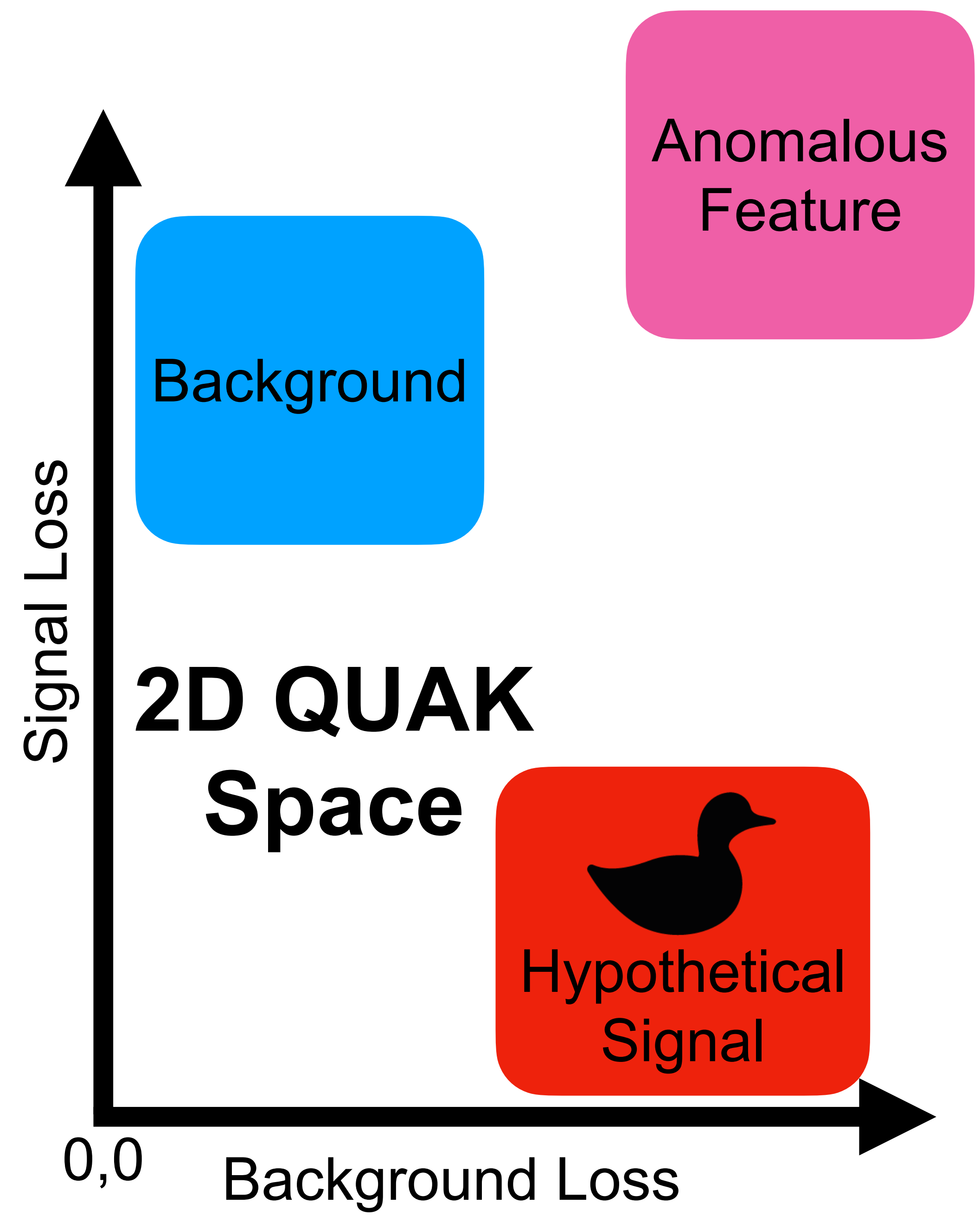


Normalizing Flow Trained On Backgrounds

S. Park

QUasi Anomalous Knowledge

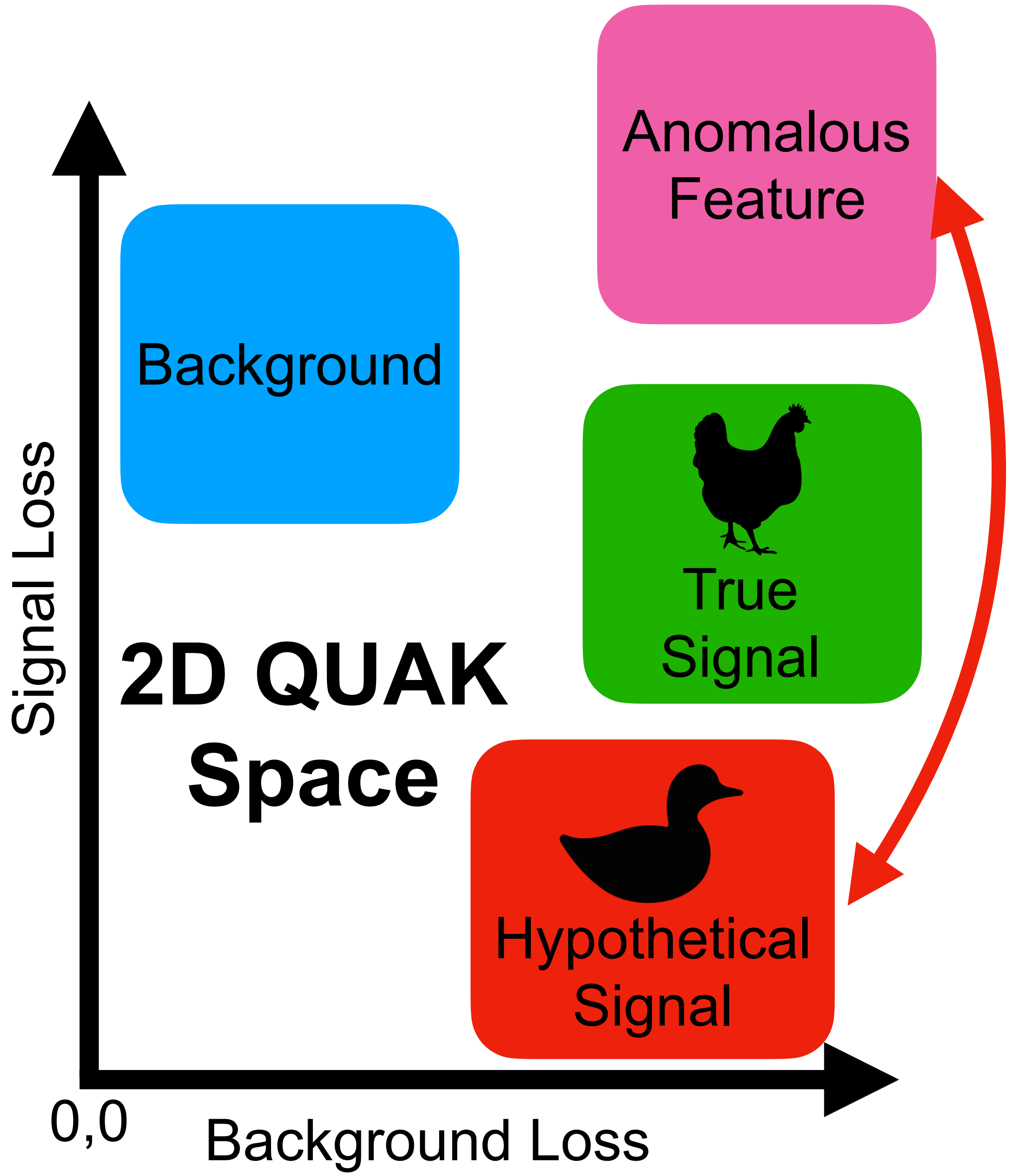
Normalizing
Flow
Trained
On signals



Normalizing Flow Trained On Backgrounds

QUasi Anomalous Knowledge

Normalizing Flow Trained On signals

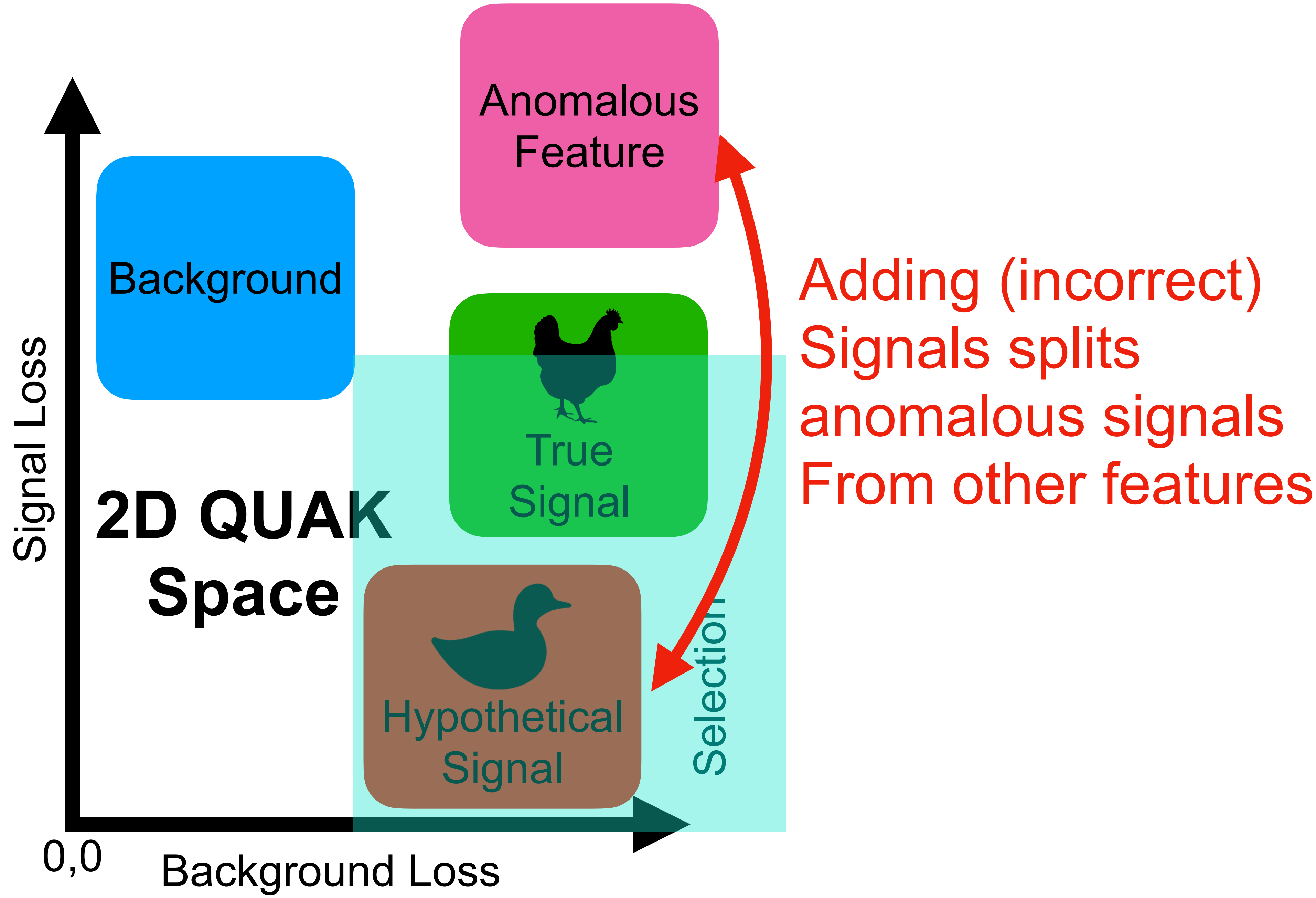


Adding (incorrect) Signals splits anomalous signals From other features

Normalizing Flow Trained On Backgrounds

QUasi Anomalous Knowledge

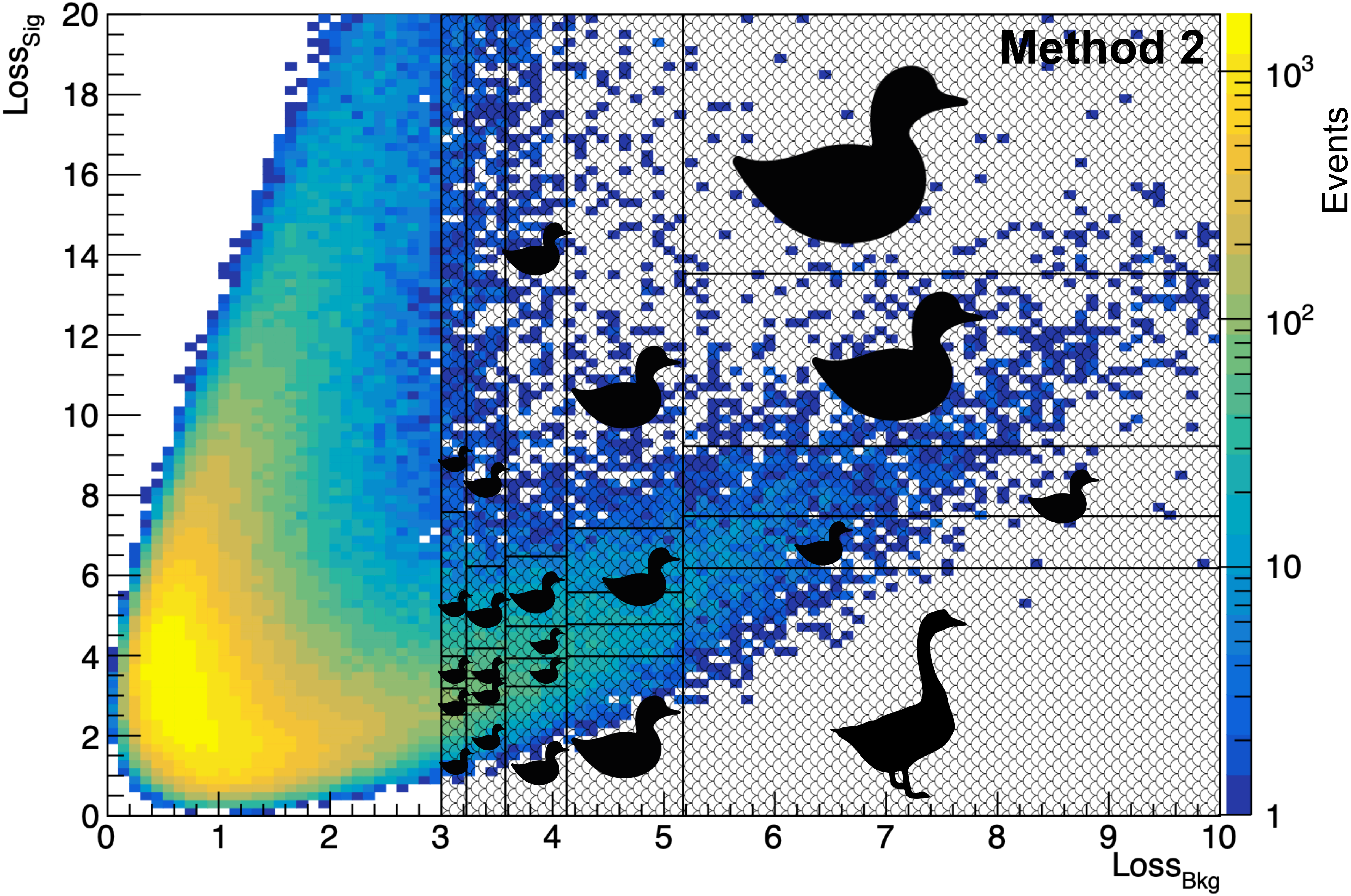
Normalizing
Flow
Trained
On signals



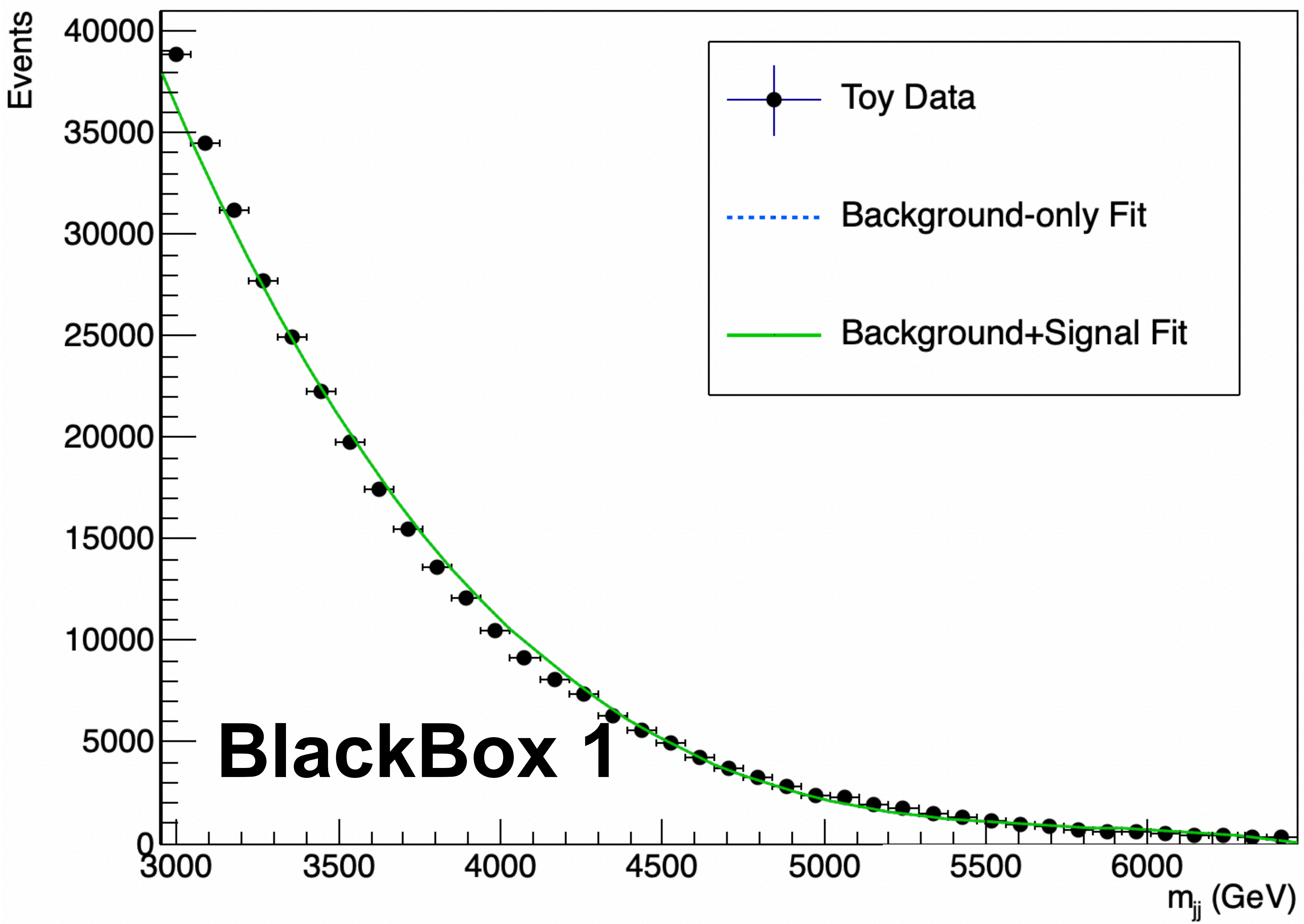
Normalizing Flow Trained On Backgrounds

Duck Duck Goose!

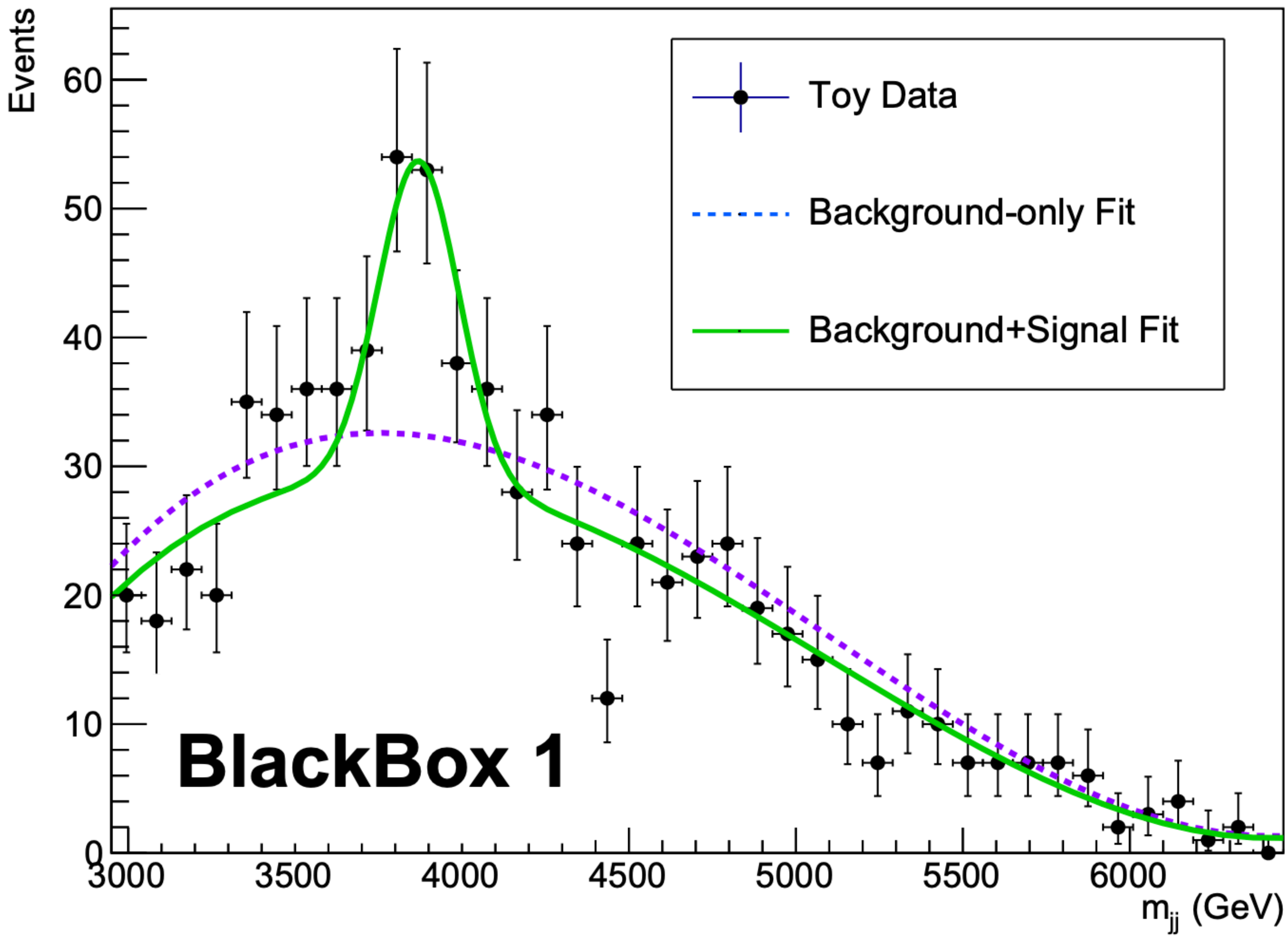
Search all of the regions **one big simultaneous fit**



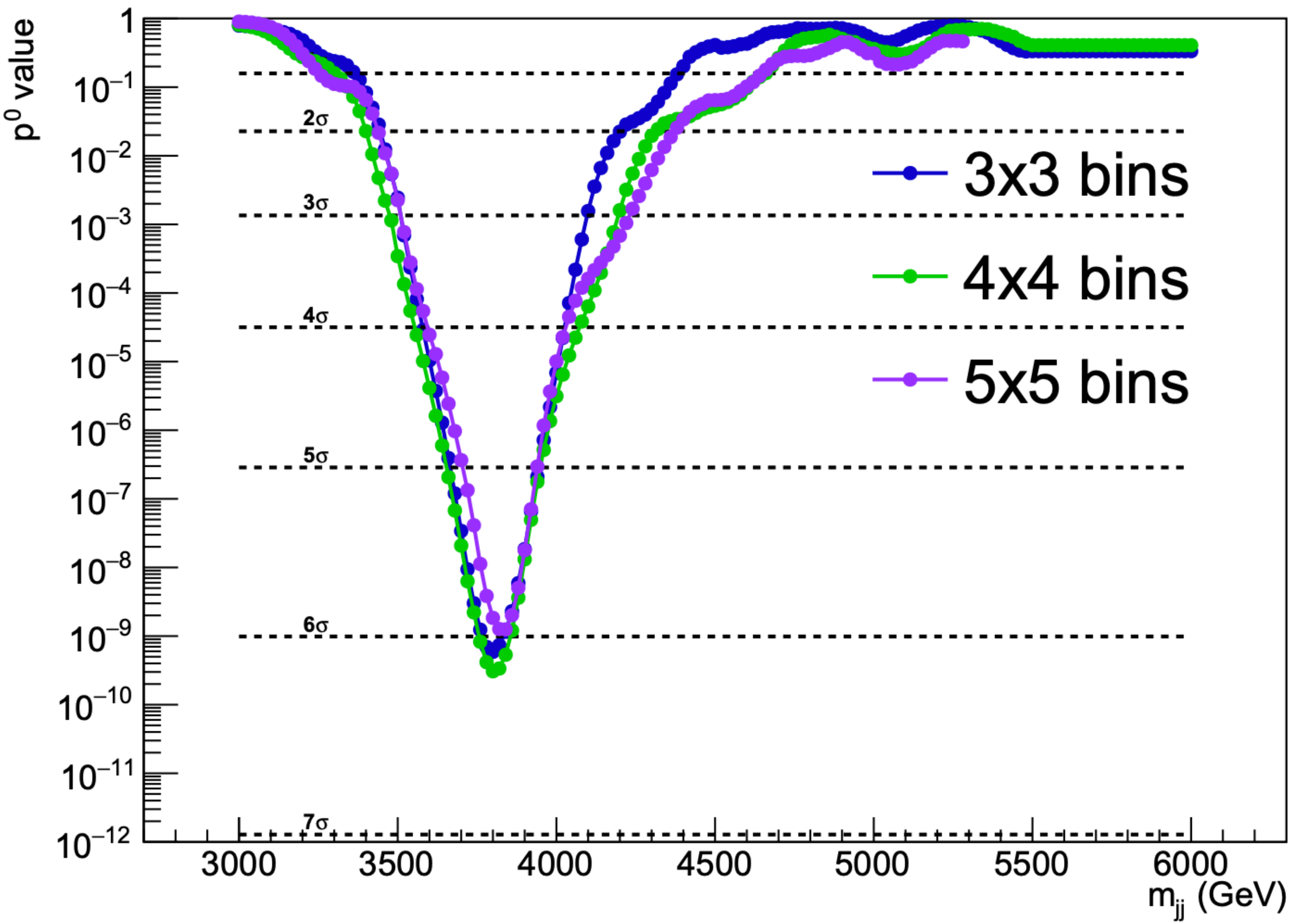
Seeing a Signal



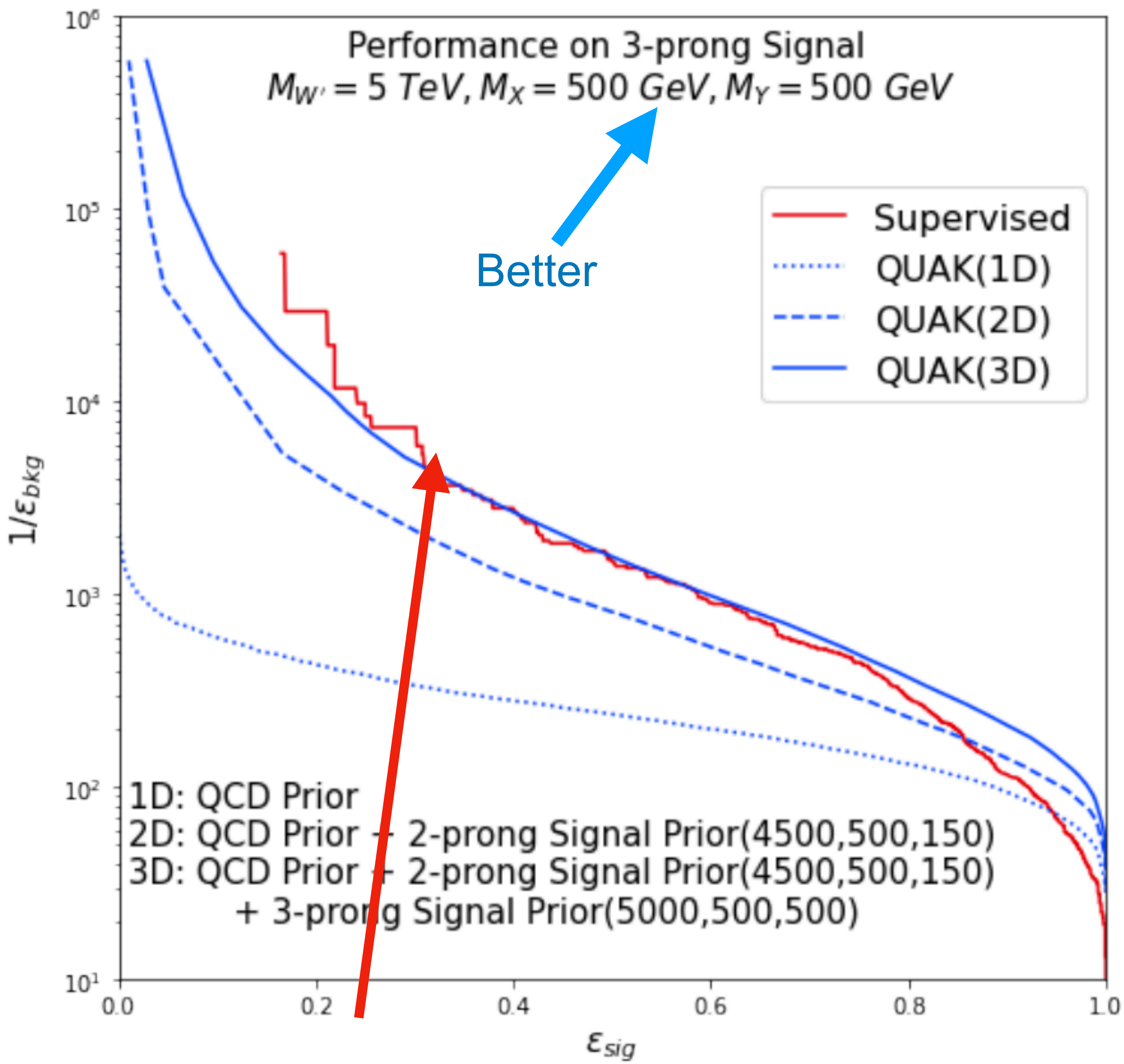
Seeing a Signal



Applying to Anomaly

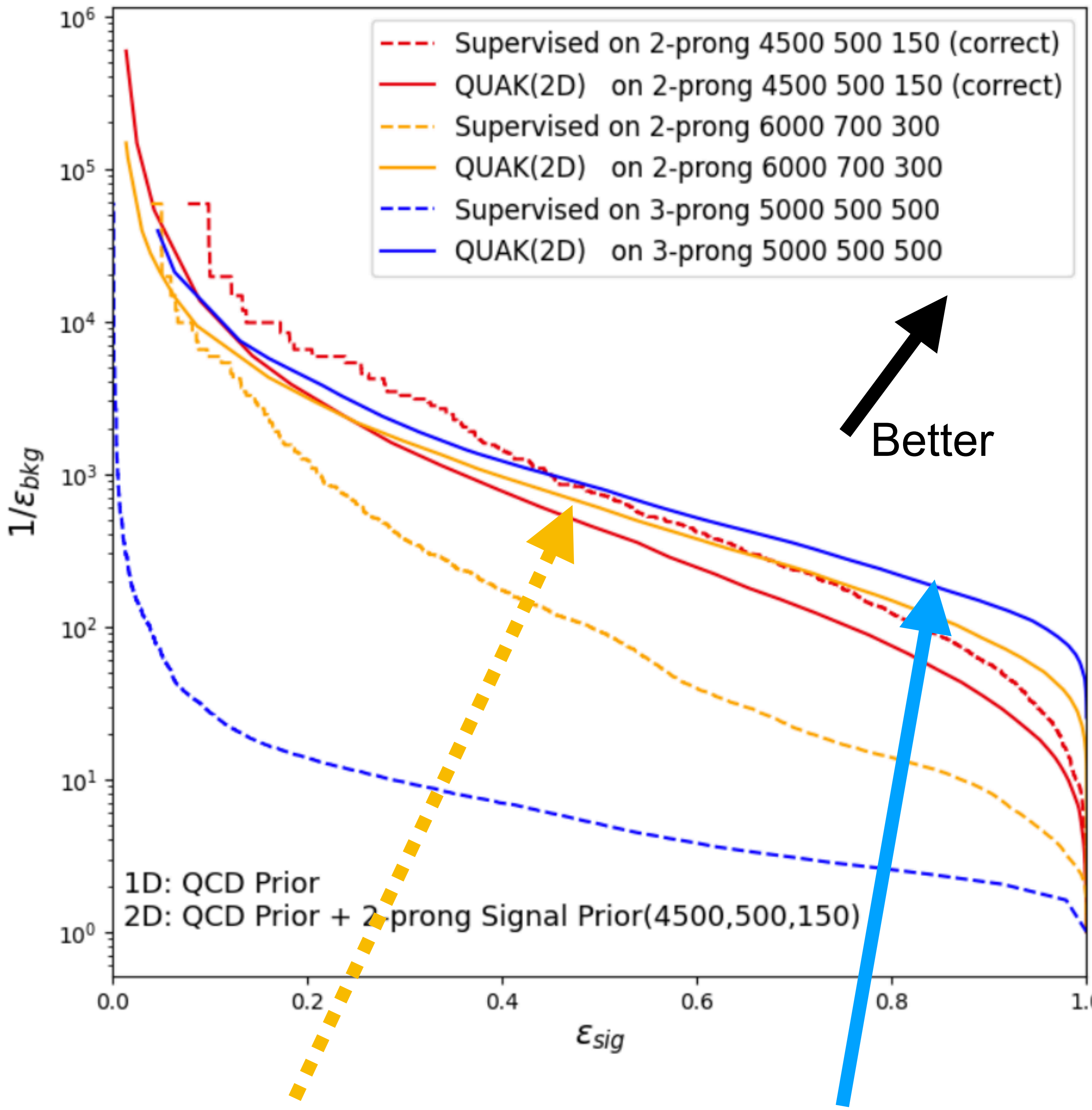


How Close to Optimal?



QUAK can outperform a supervised network
When signals are the same

How Close to Optimal?



One Supervised Network

One QUAK Network

What will the future be?

- Like to think this is a harbinger for things to come

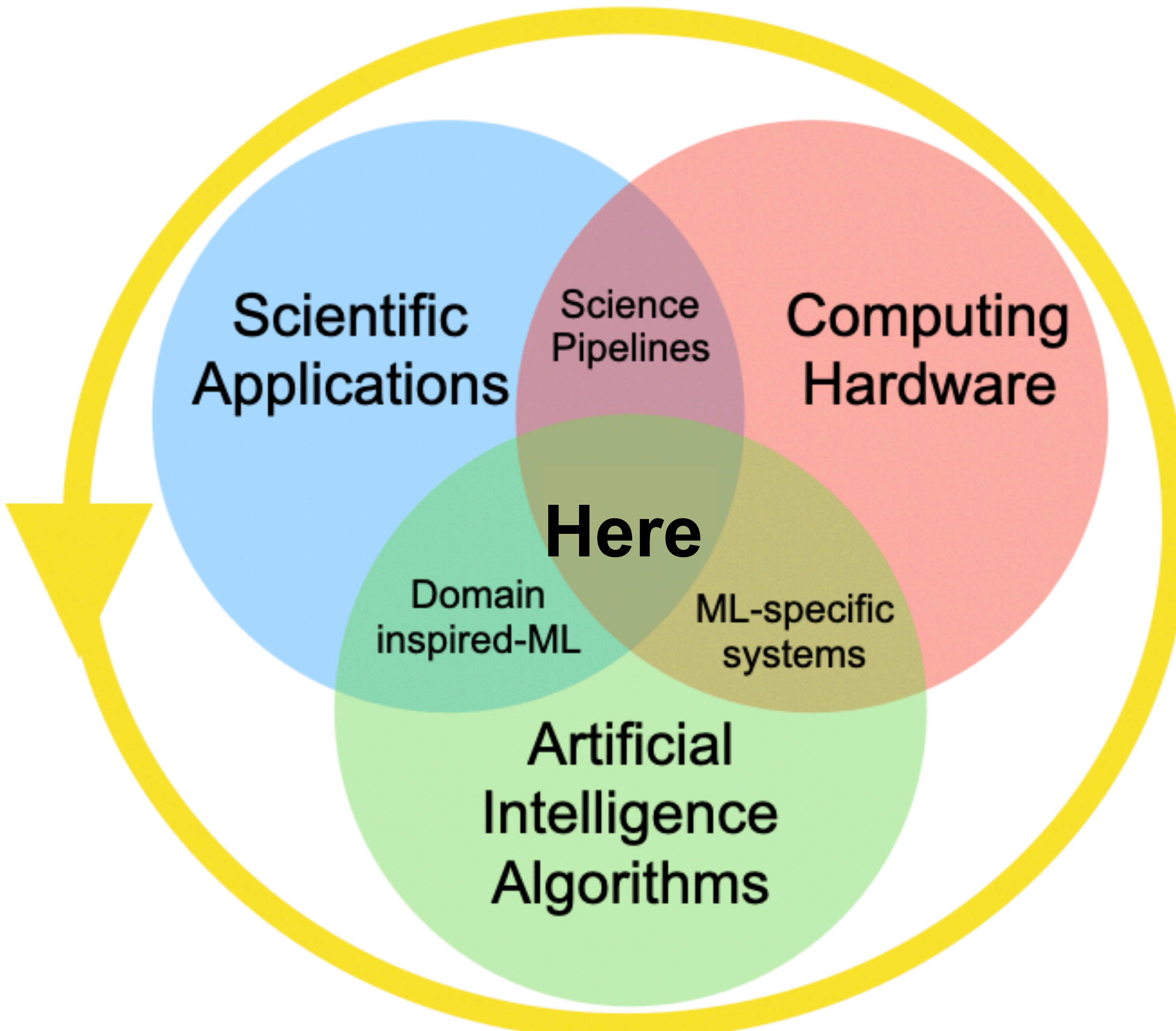
Did we find all the Higgs bosons in there?

Towards The Future

What are all the hidden signals in there?

and Can we do it Real-time?

Can we see it all? When its coming?


Scientific Applications
Science Pipelines
Computing Hardware
Here
Domain inspired-ML
ML-specific systems
Artificial Intelligence Algorithms

Conclusions

Real time deep learning

In science has the potential to open new doors

Thanks!

Fast ML Team

Fermilab

Massachusetts Institute of Technology

University of Washington

University of Toronto

Purdue

Microsoft

XILINX

UCSD

CERN

ILLINOIS

Columbia University in the City of New York

HawkEye 360

Northwestern University