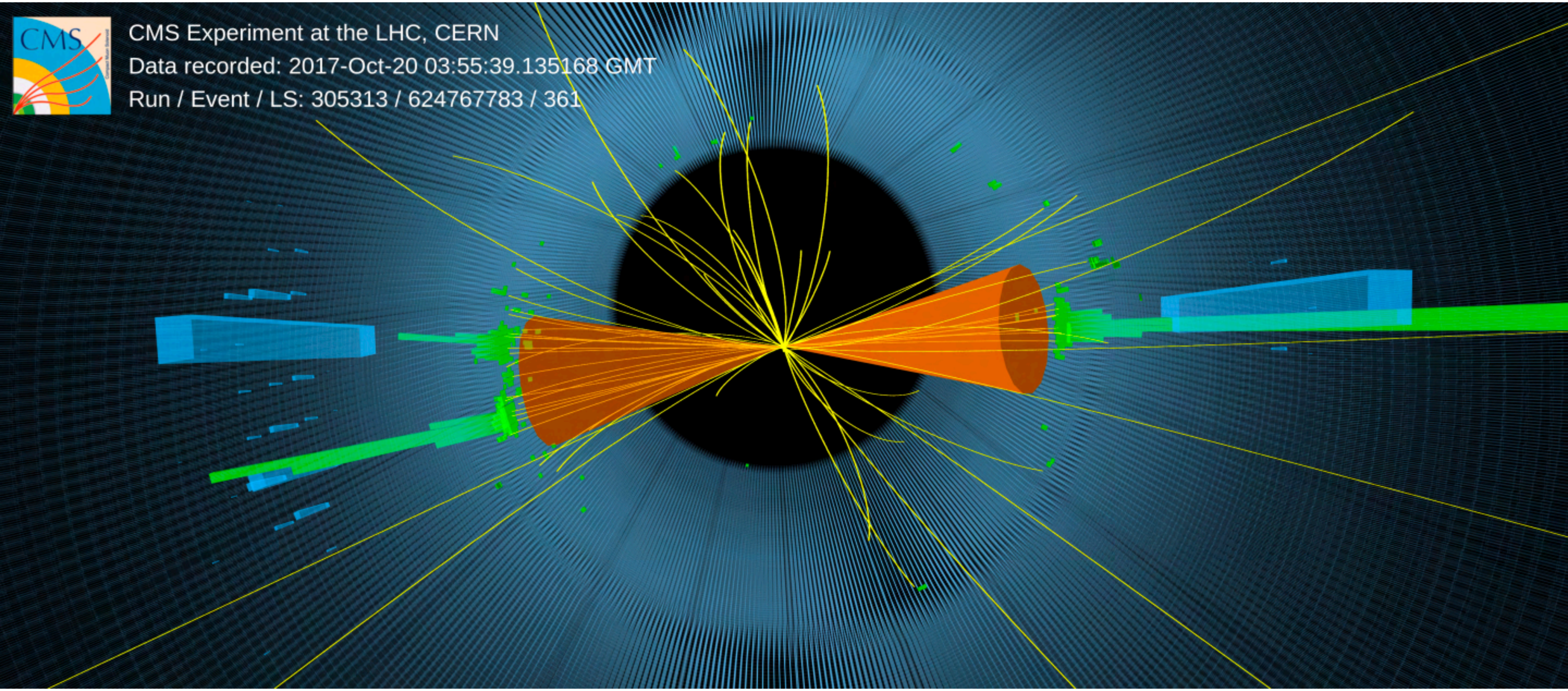
Computing For Big Data Experiments

Philip Harris MIT

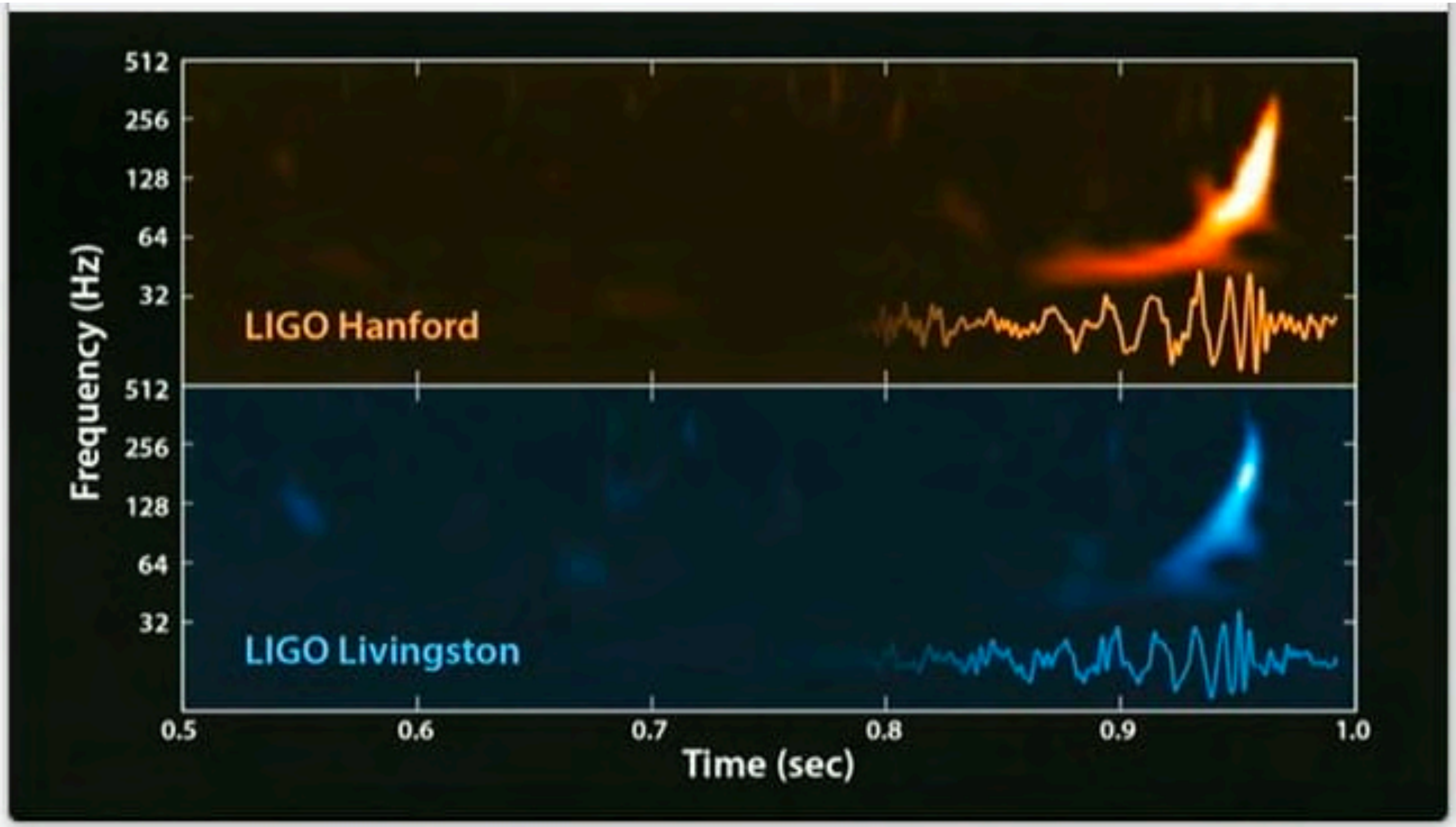
LHC Challenge: Can we process every collision?



- LHC collides 40 Million times per second
- Each collision is about 10 MB of data

400 Tera Bytes Per Second

LIGO Challenge: Can we find all mergers

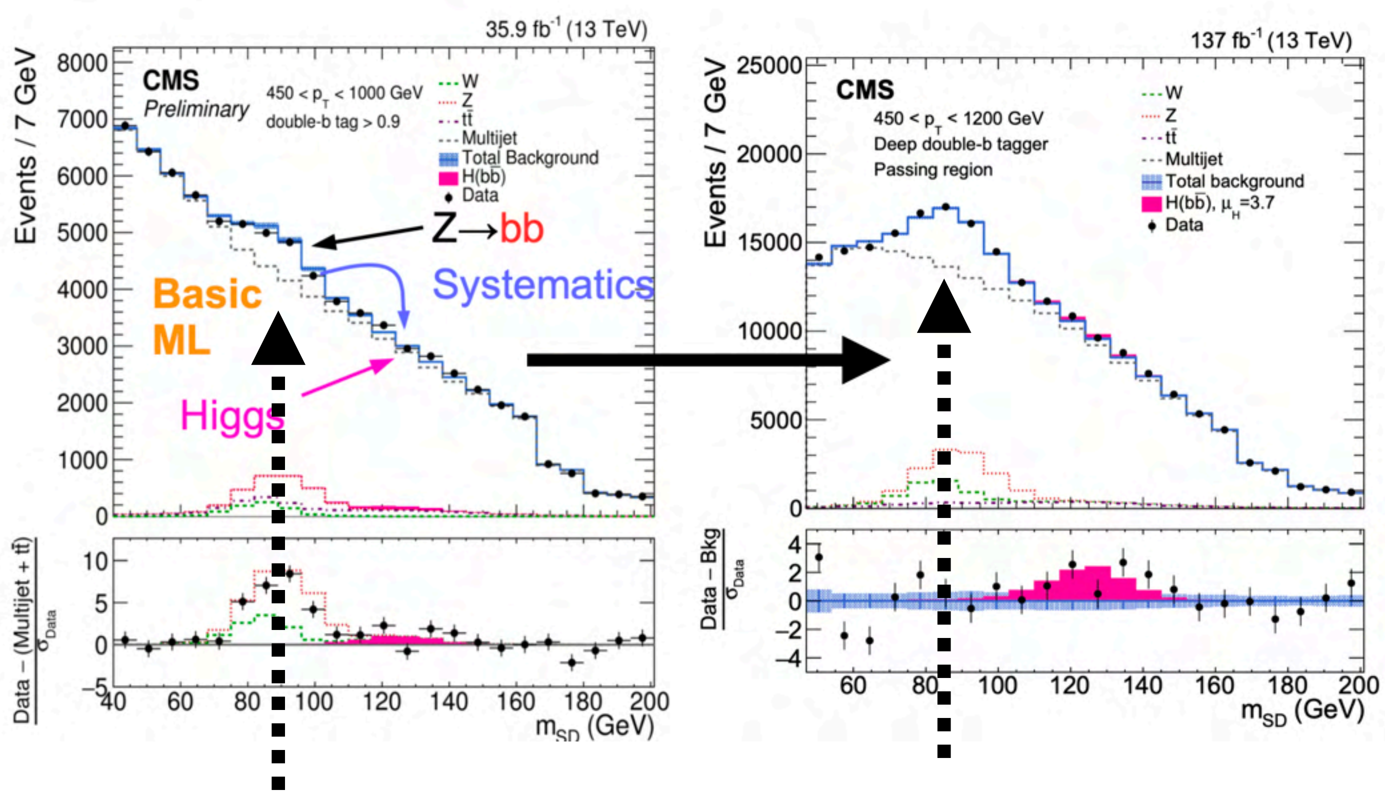


- LIGO has 10^5 channels at 1024 Hertz
- Looking for subtle signals hidden in the noise

Real-time Detailed (10k core) analysis every millisecond

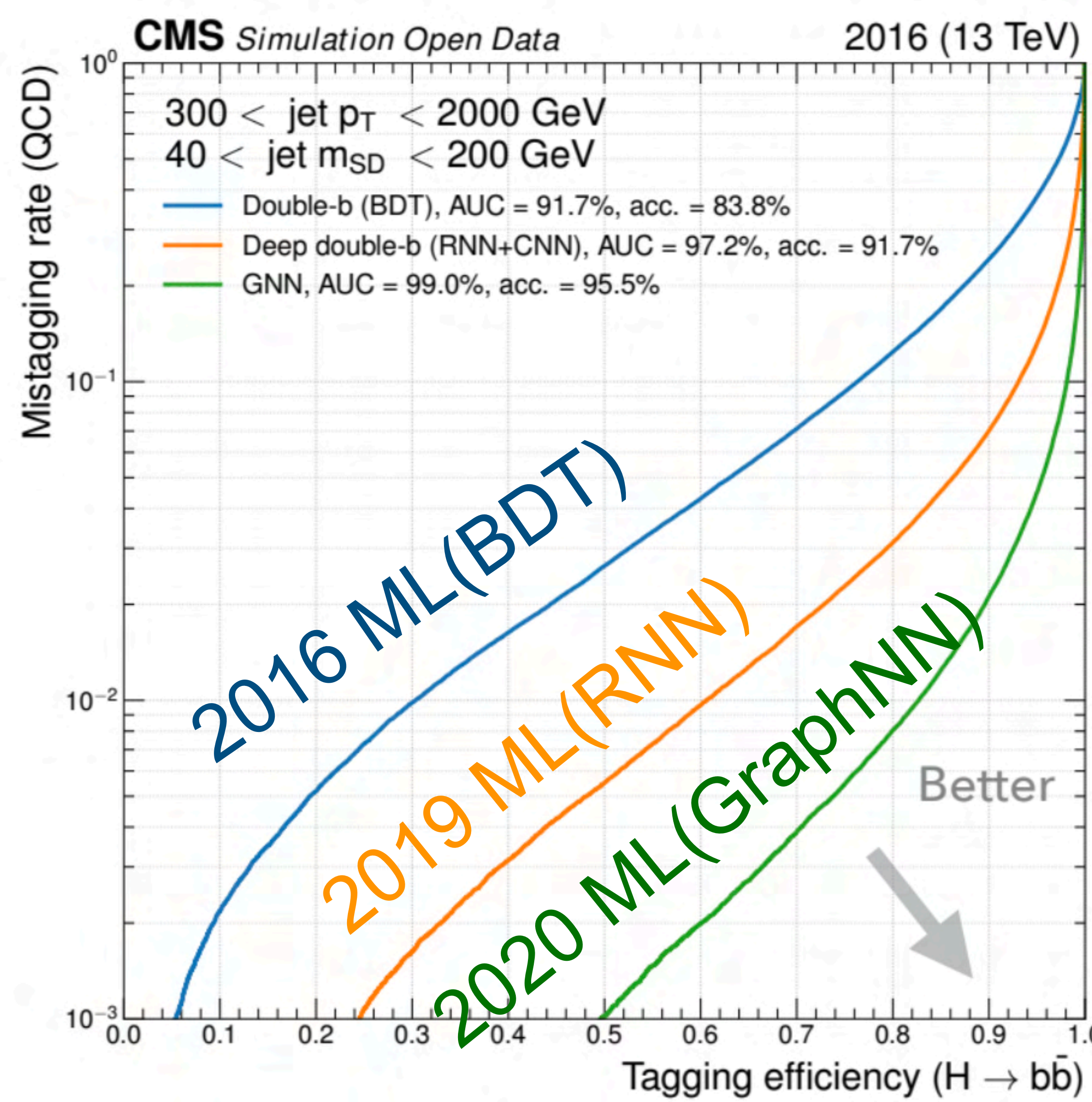
An Angle on AI revolution

- Things are starting to change in the way we compute
    - ML algorithms have the ability to go beyond algorithms
        - This is also b/c GPUs have helped to parallelize computation



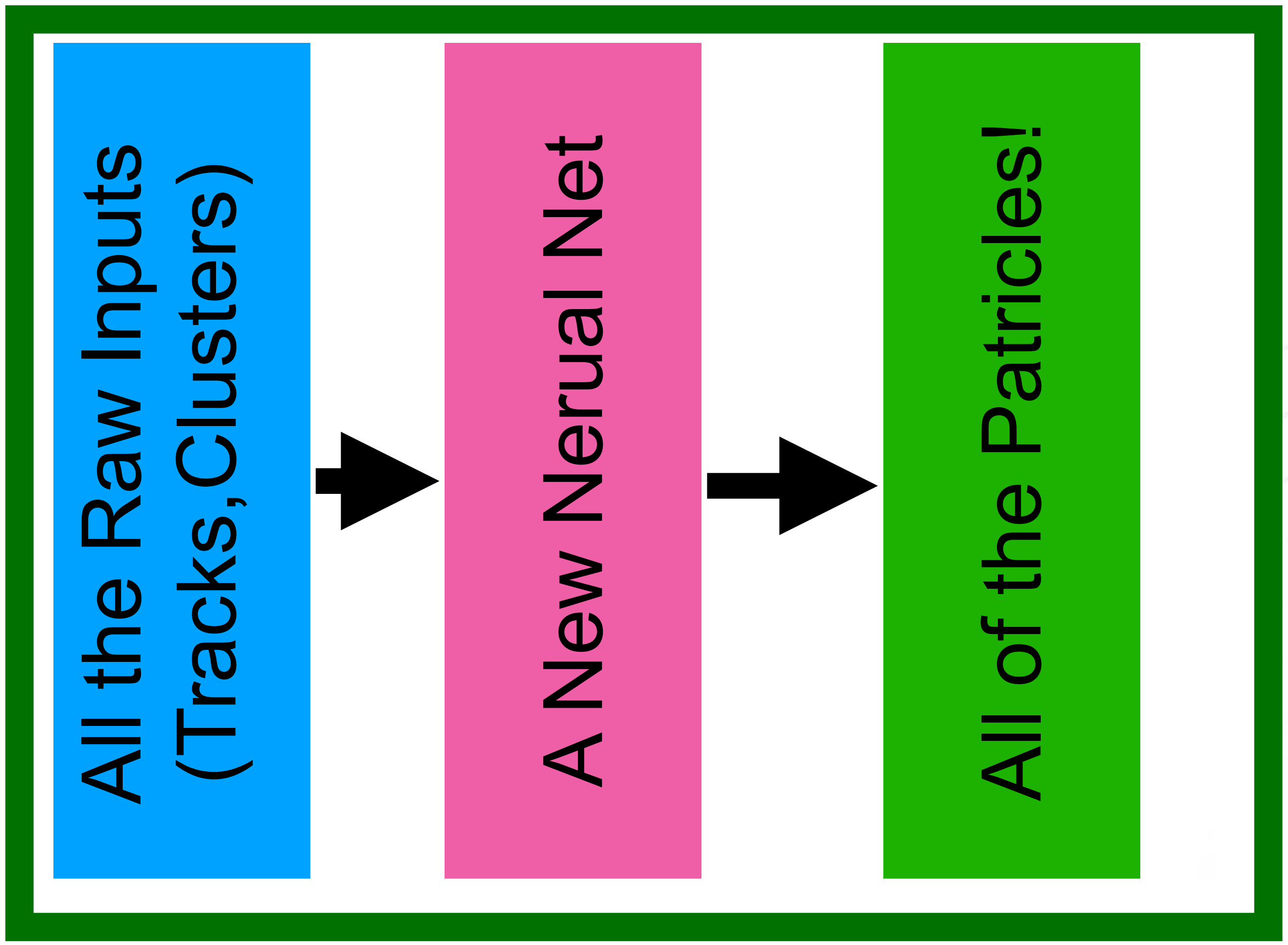
Small ML
Small Peak

Big ML
Big Peak

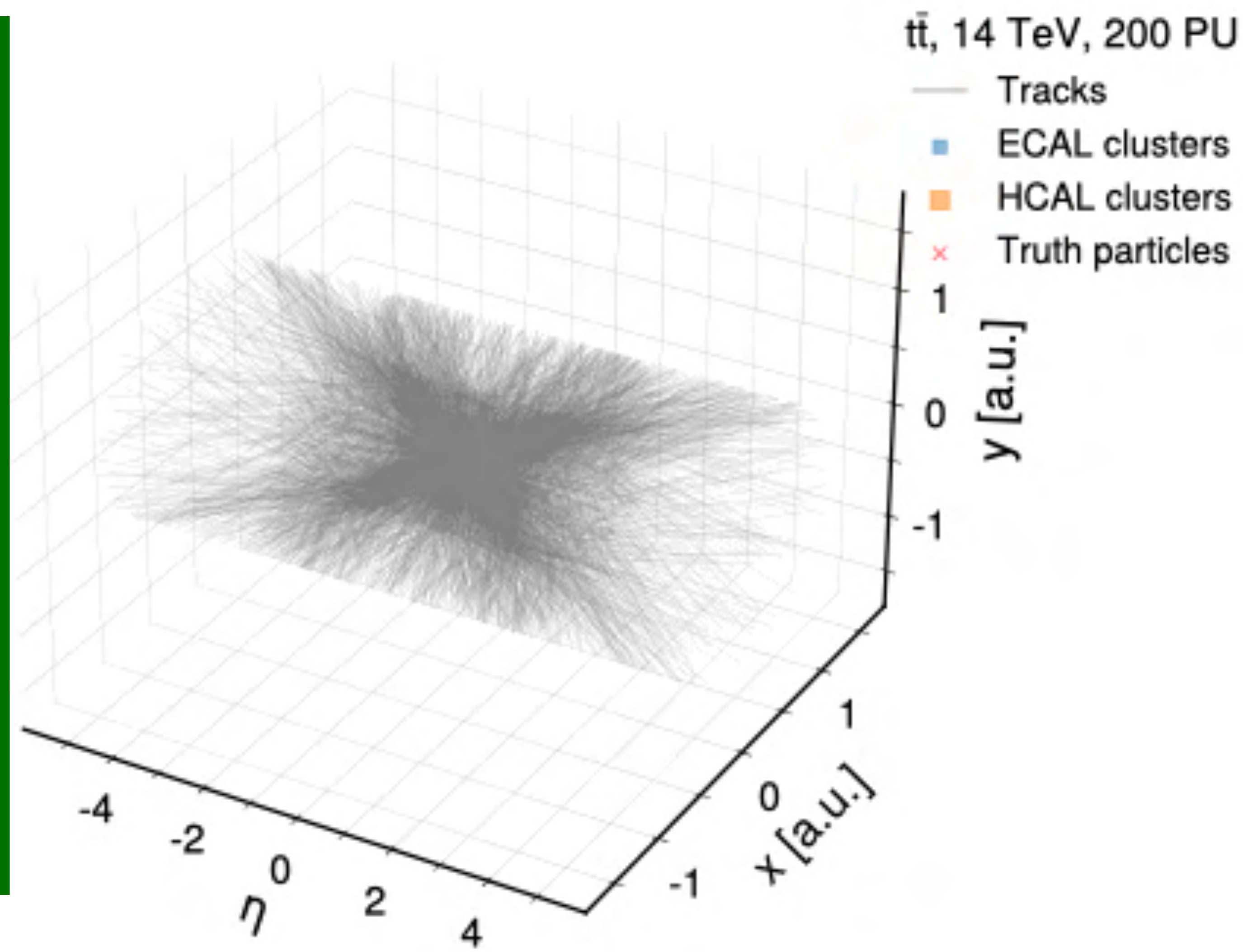


What does this mean?

- Inevitable that our algorithms will become progressively larger

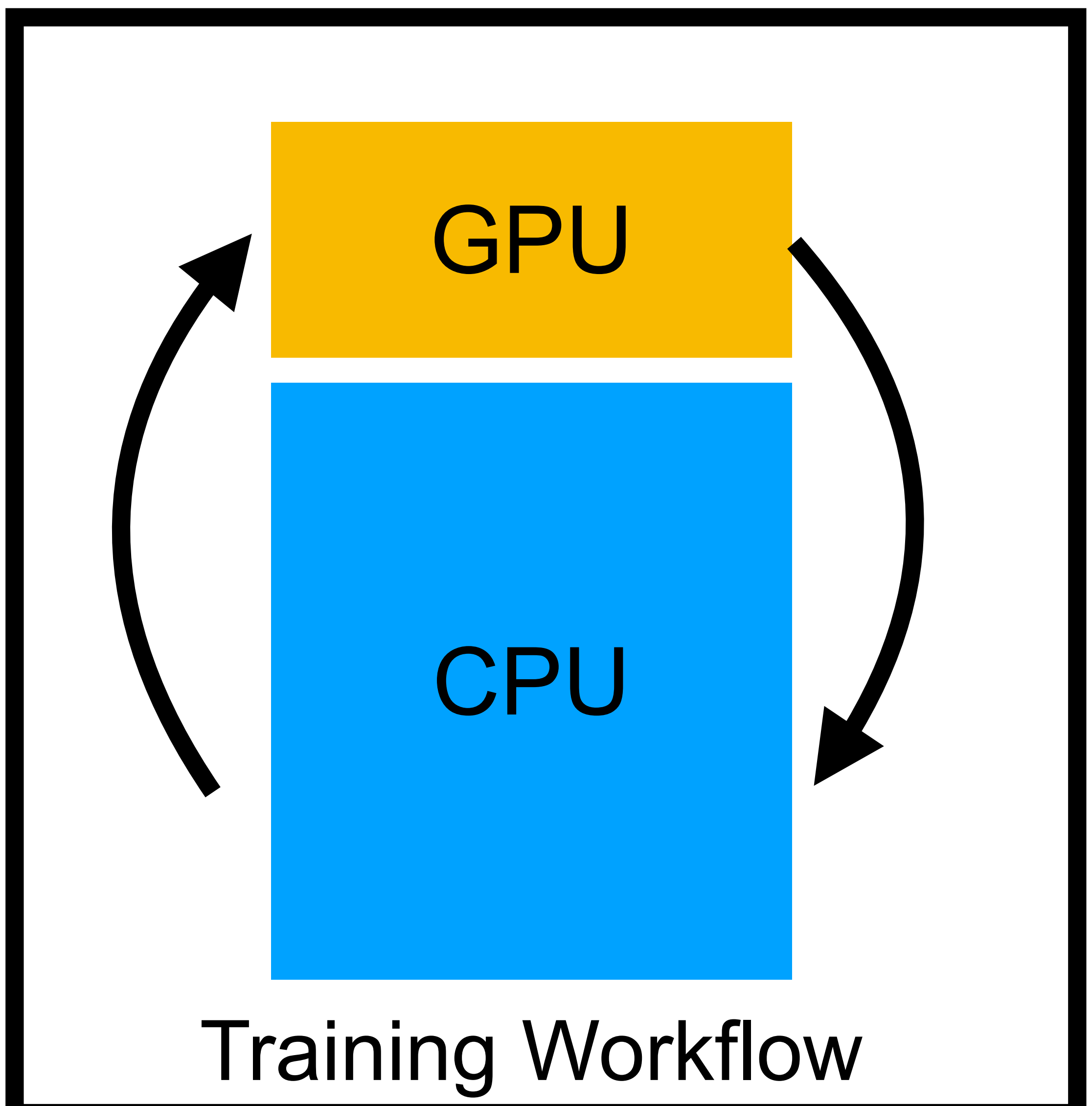


All particles in on fell swoop

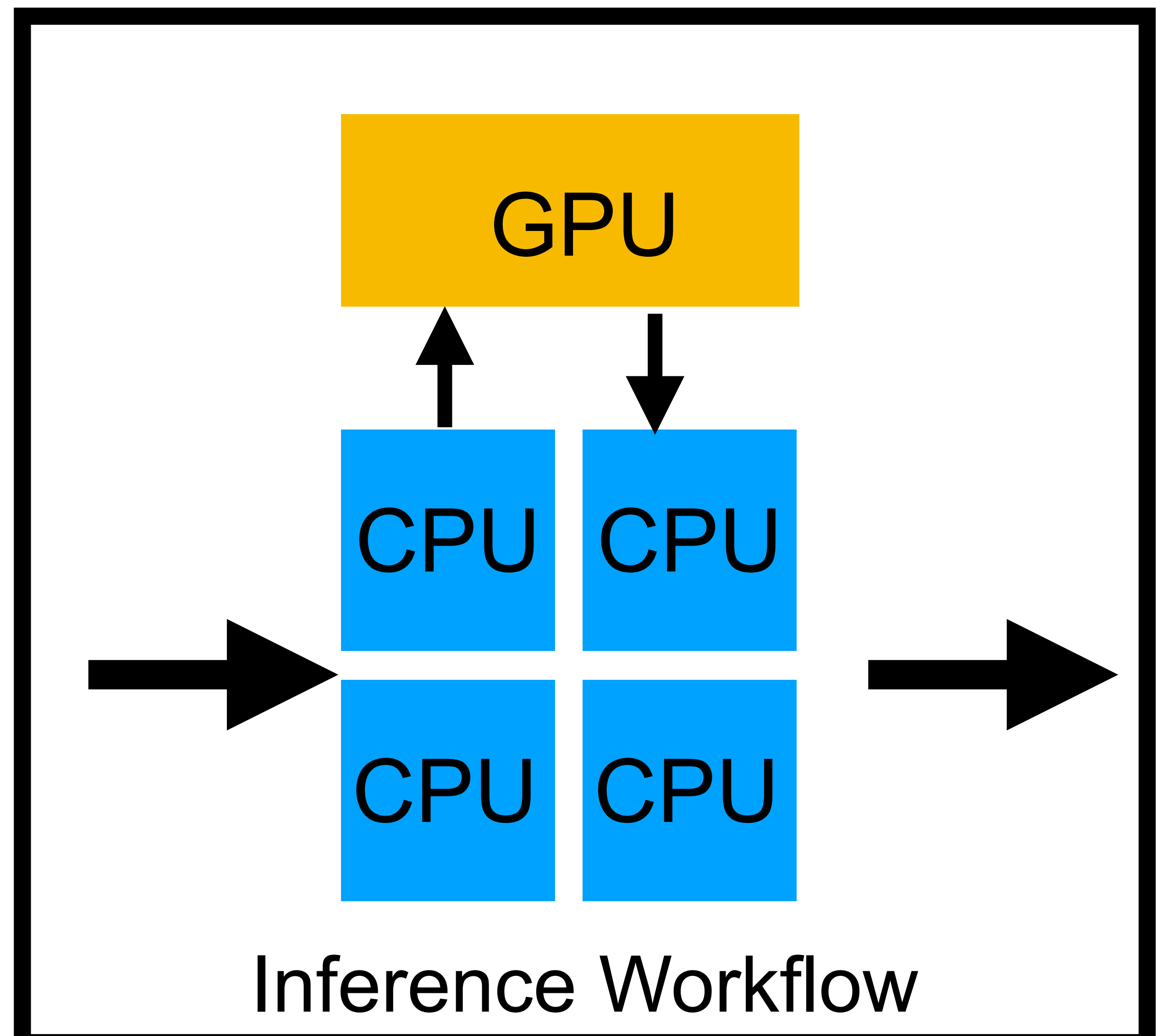


Algorithm Needs

- With the development of AI algorithms we need two things
  - Training and Testing
  - Processing power to run on the data



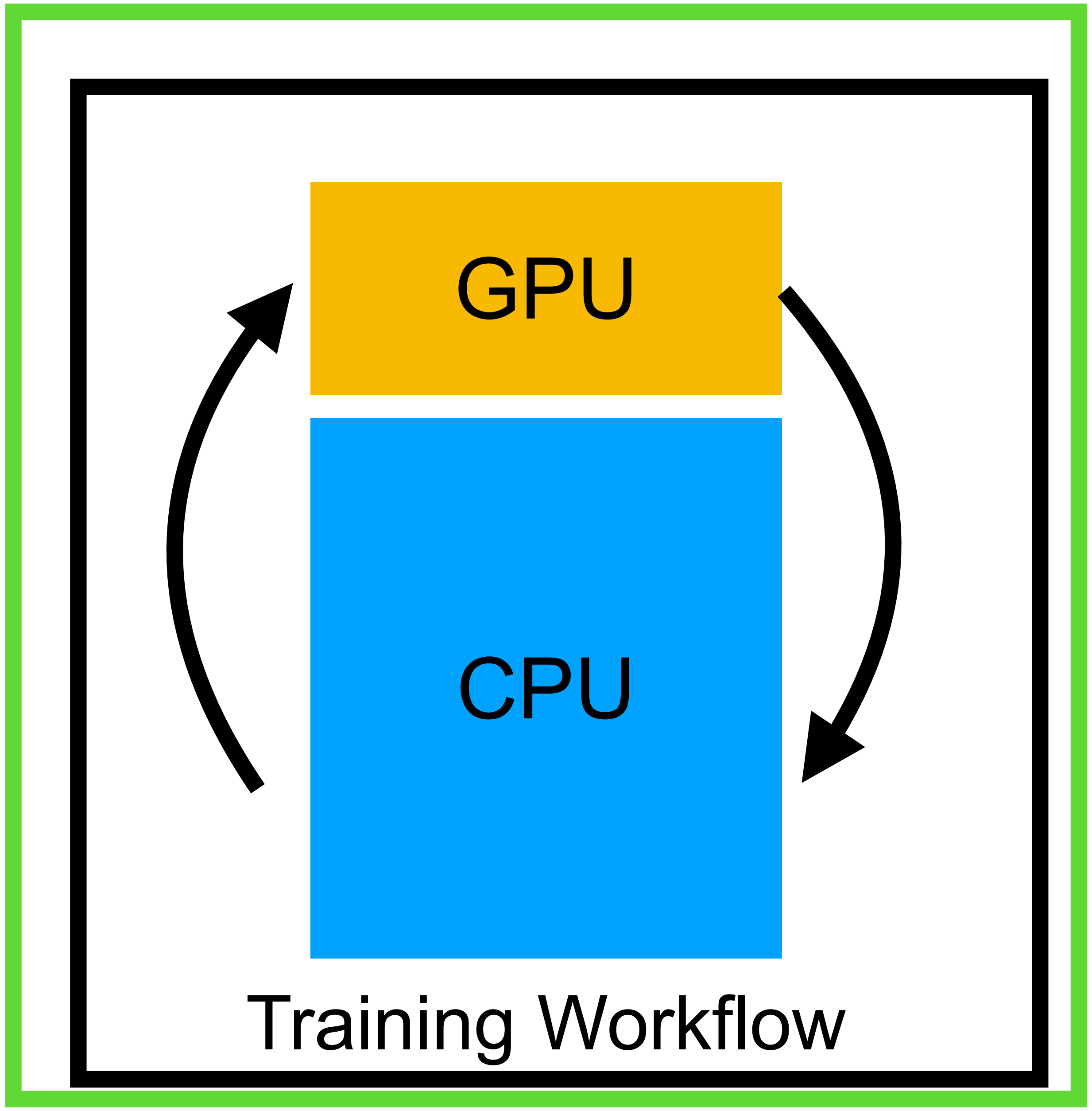
Training Workflow



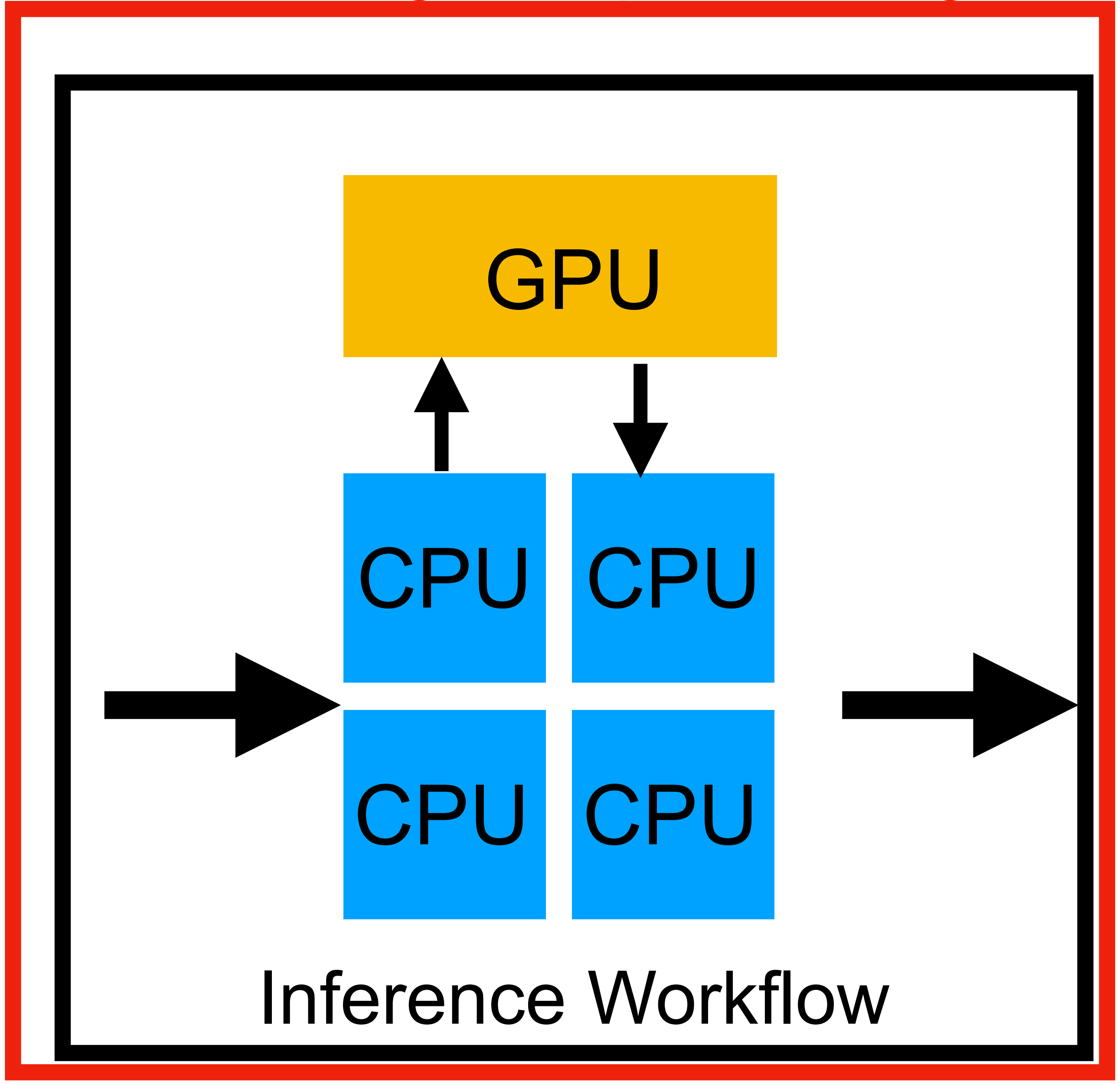
Inference Workflow

Algorithm Needs

Solved
Big HPCs dump as many GPUs as they possibly can in a room
Aim for the maximum compute



What we need
Requires Dynamic allocation to balance GPUs and CPUs focus is on dealing with processing



Algorithm Needs

Solved
Big HPCs dump as many GPUs as they possibly can in a room
Aim for the maximum compute

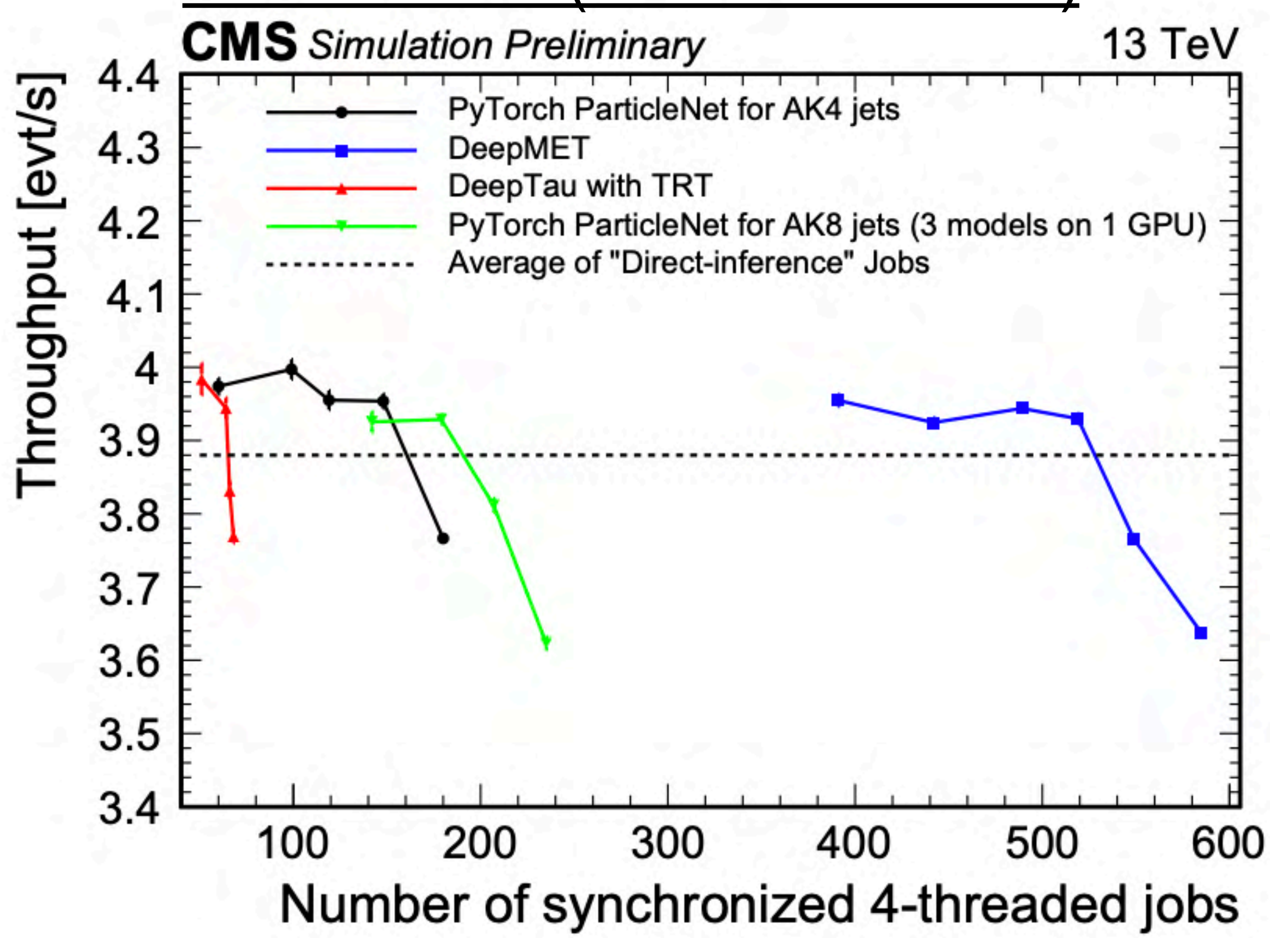
Training Workflow

What we need
Requires Dynamic allocation to balance GPUs and CPUs focus is on dealing with processing

Inference Workflow

Real World Examples

Public Slides (Publication soon)



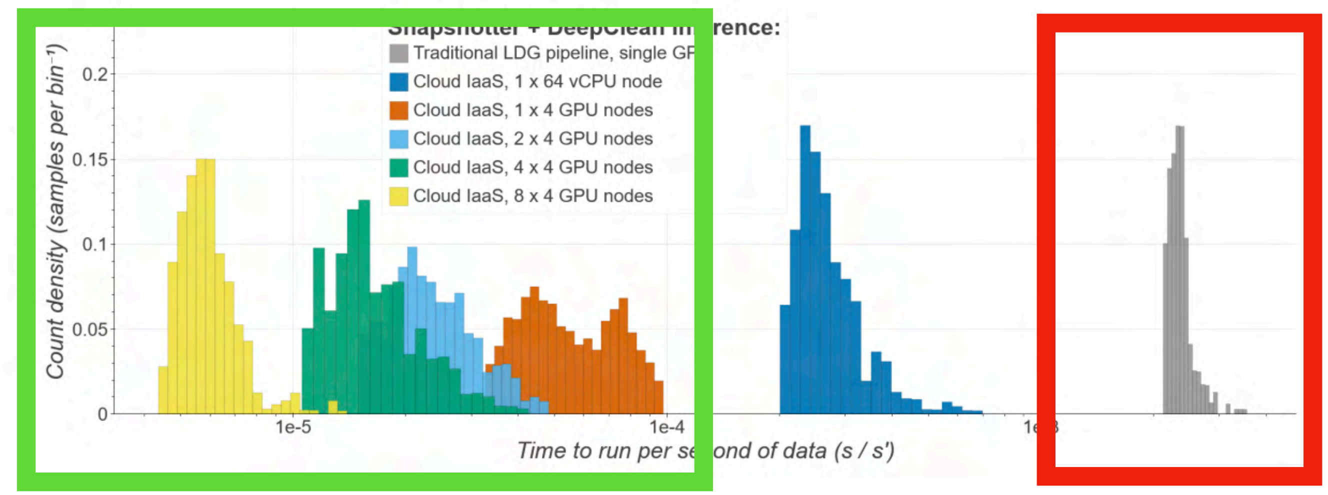
LHC

Saturation for a single GPU
W/ Many CPUs

Nature Astronomy 6 529-536

Updated Workflow

Current HPCs



Time to run per second of data

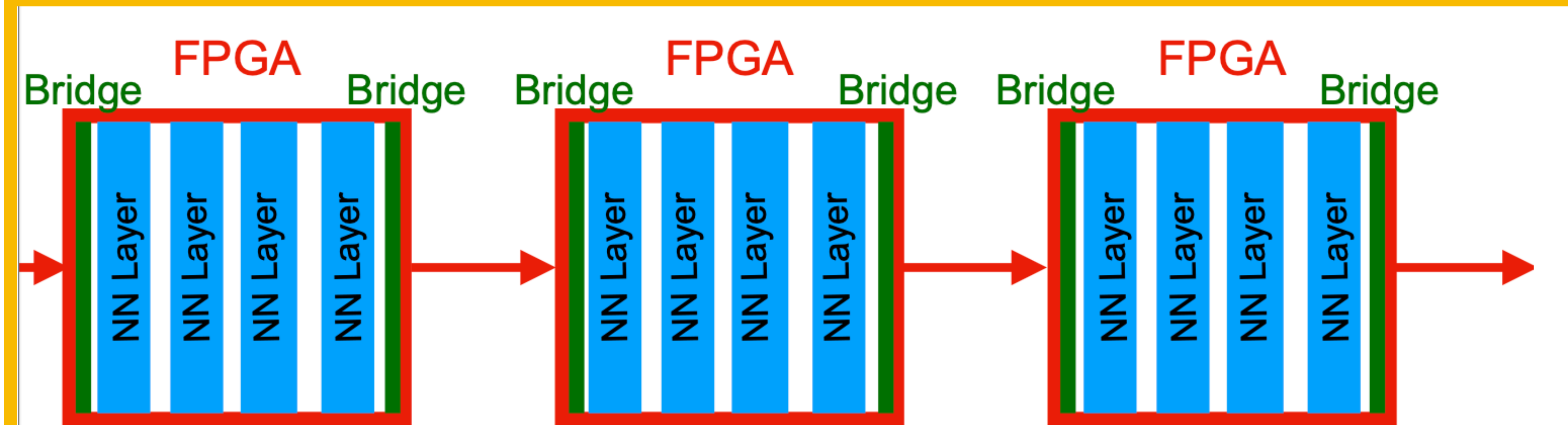
LIGO

Speed with Inference-as-a serivce

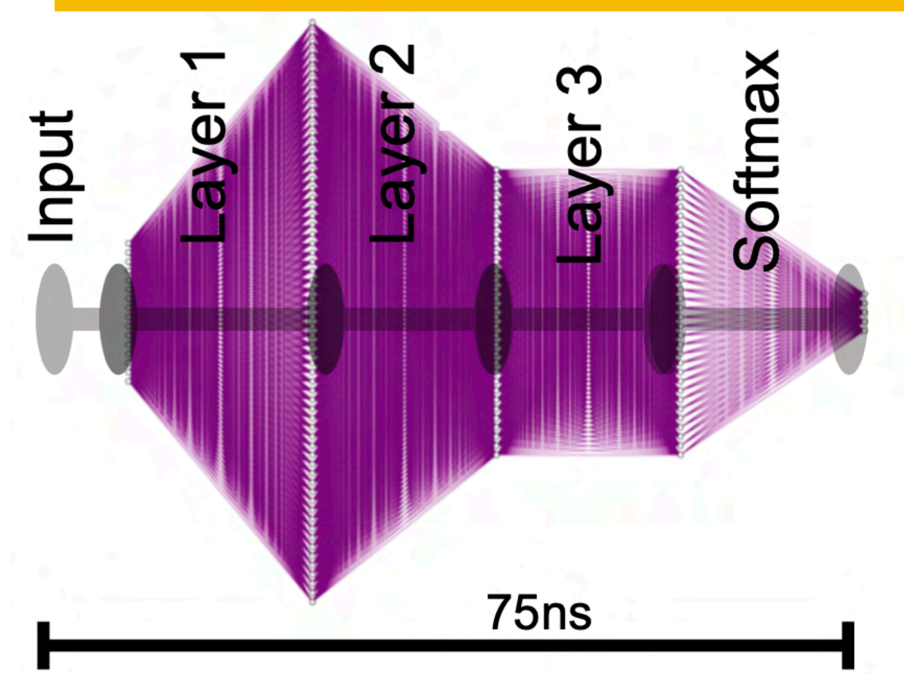
- Here is a glimpse of studies we have done to show this
- Run large scale studies demonstrating heightened throughput

Custom Computing

Ultra low latency Requires a fully custom solution
To achieve ultra high throughput at > 1 Pb/s we use FPGAs
This system doesn't look like an HPC/computer anymore



Compressed AI with Fixed Bit Precision



hls4ml

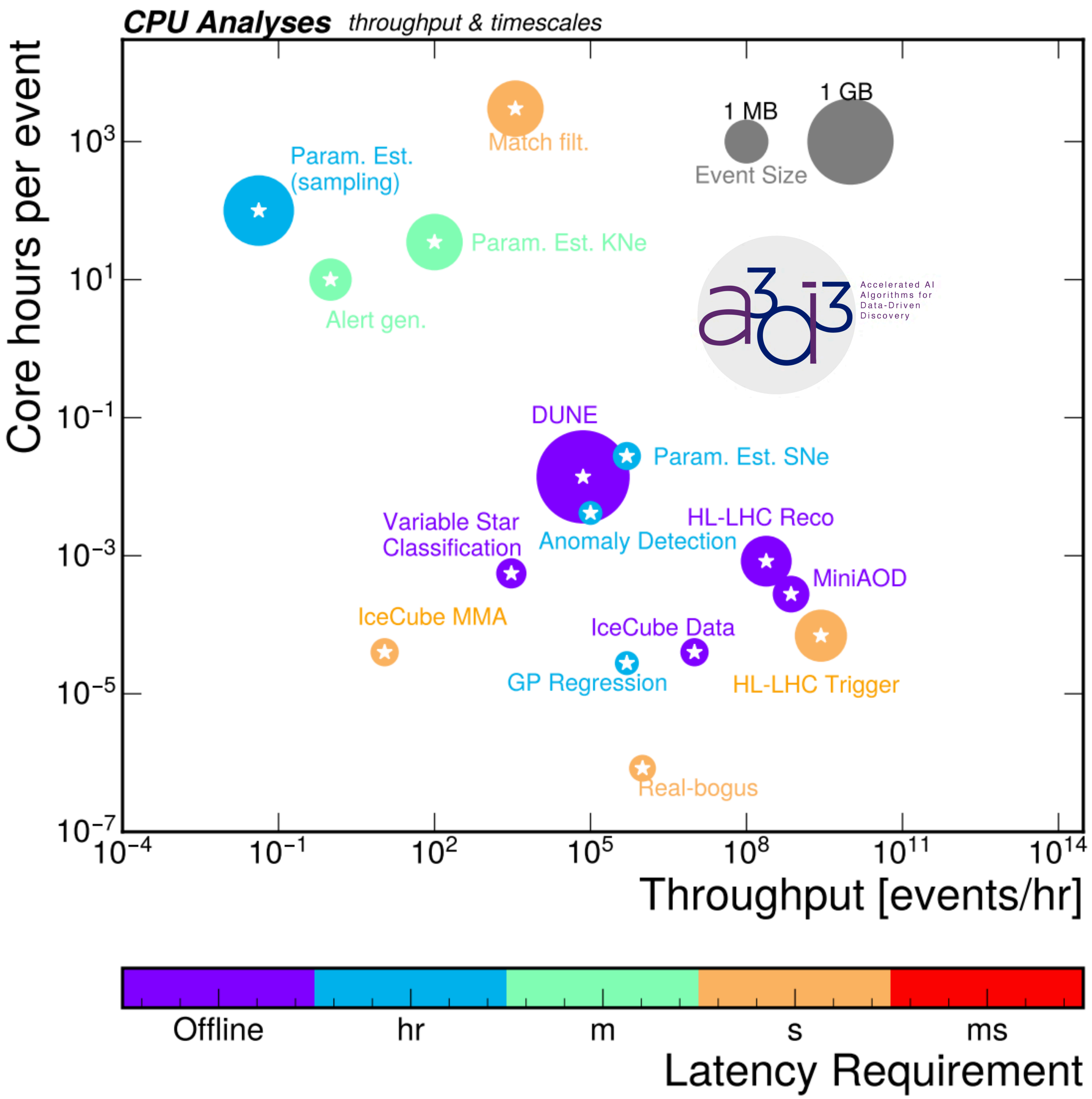
Applications:
LHC/Plasma Controls/Brain Controls/…

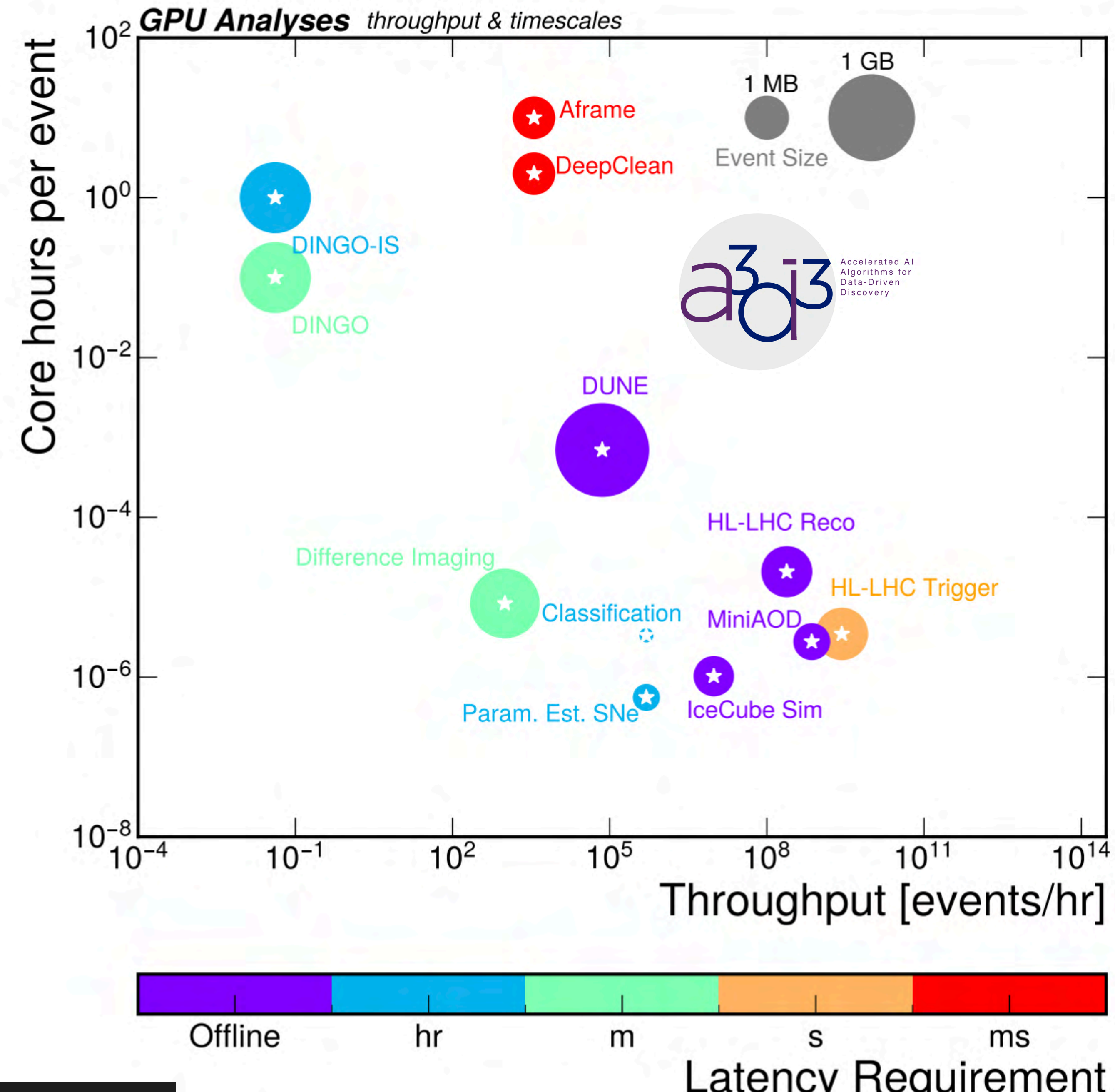

a3di3
Accelerated AI
Algorithms for
Data-Driven
Discovery

What is Critical?

- We would like to highlight commonalities across domains
  - **Computing demands**
    - Critically connected infrastructure for ML science deployment
    - Inference differs from training →Efficiency is Key
  - **Software Stack**
    - With all ML algorithms aim for a set of core software tools
    - Containerization: Docker/Singularity/Kubernetes/…
  - **ML Problems**
    - Awareness of the diversity of problems is critical (Not just LLM)
    - Highlighting the similarity across scientific domains is critical

Computing Demands





arXiv > hep-ex > arXiv:2306.08106

High Energy Physics – Experiment

[Submitted on 13 Jun 2023]

Applications of Deep Learning to physics workflows

Arxiv: 2306.08106

Have a whitepaper outlining Inference Workflows Demands

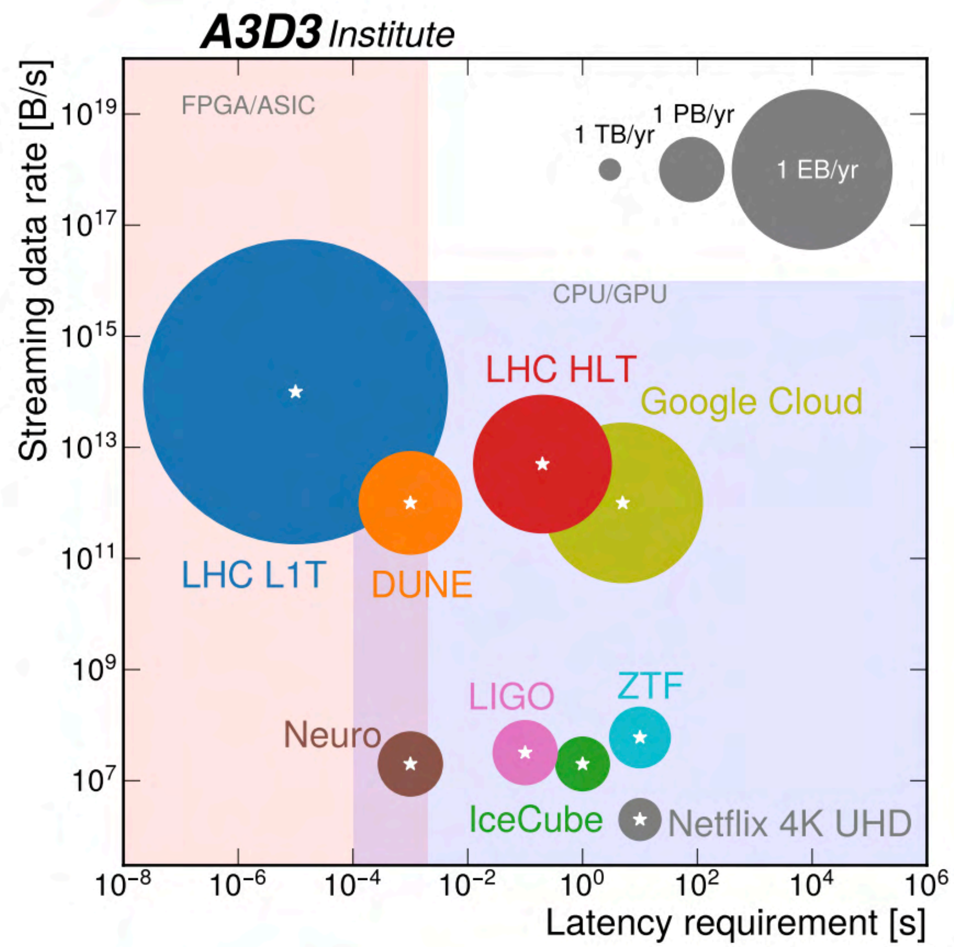
A Vision

- Can we align science across ML Challenges?
  - Details here following C. Herwig, N. Tran (Fermilab)

		Scientific Moonshots		
		Domain A	…	Domain N
AI thrusts	AI - 1: Real-time	Benchmark 1A		Benchmark 1N
	AI - 2: Control			
	AI - 3: Autonomous			
	AI - 4: Foundation			
	AI - 5: Generative	Benchmark 5A		Benchmark 5N

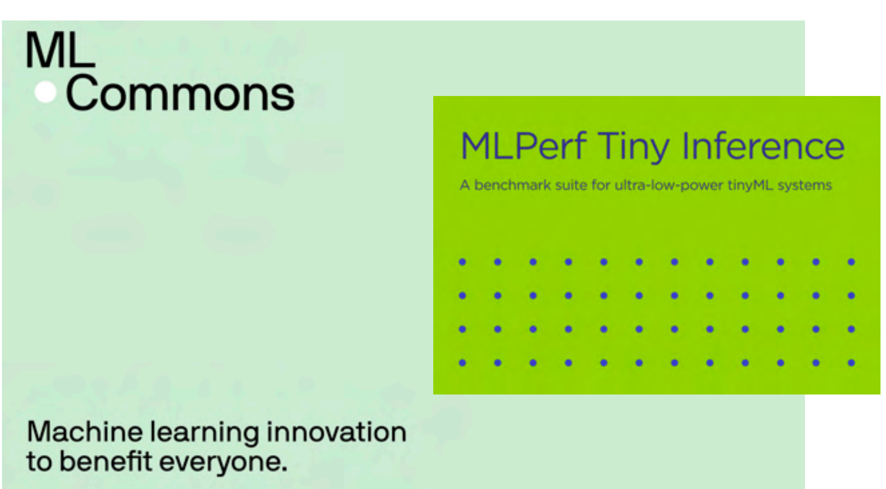
ML Challenges

- Aiming to build a website hosting Scientific ML Challenges



Accelerated AI Algorithms for Data-Driven Discovery

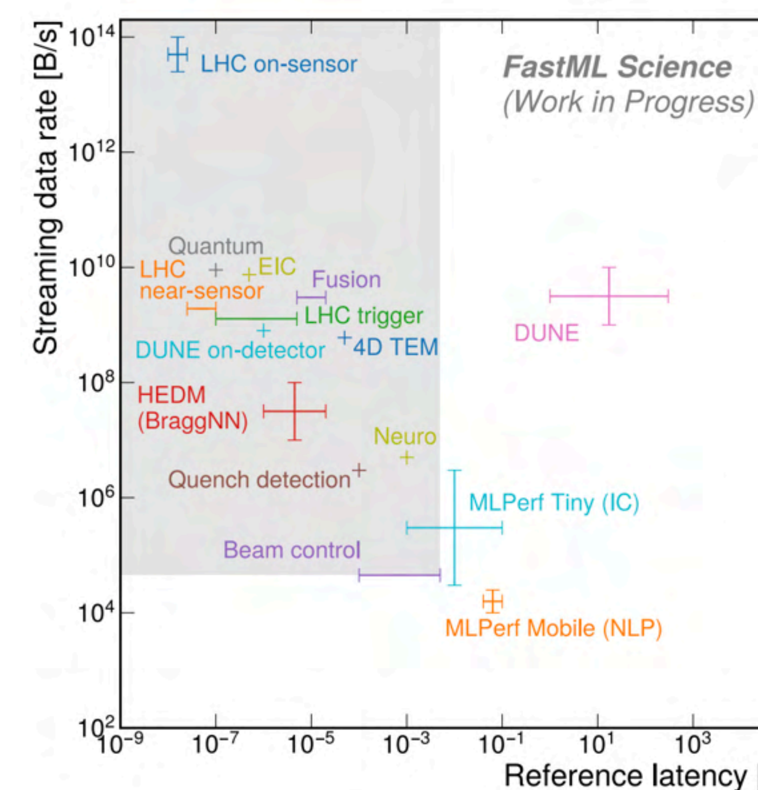
Connecting with ML Commons



Would like to highlight Criticality of Scientific Problems

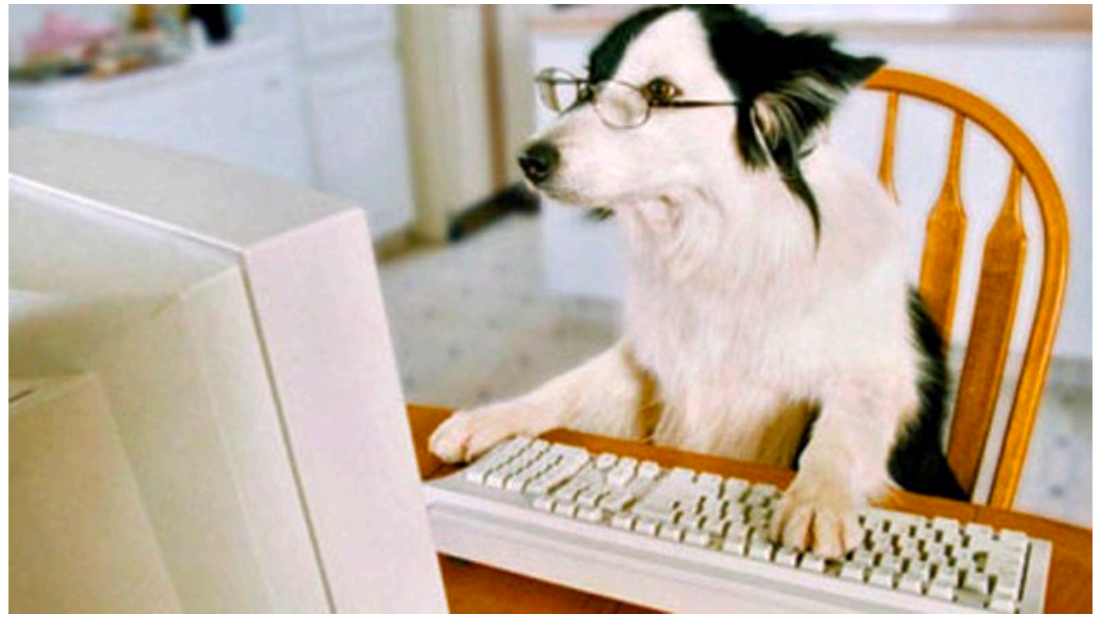
Support from NSERC <u>FAIRUniverse</u>

Connecting With Hardware

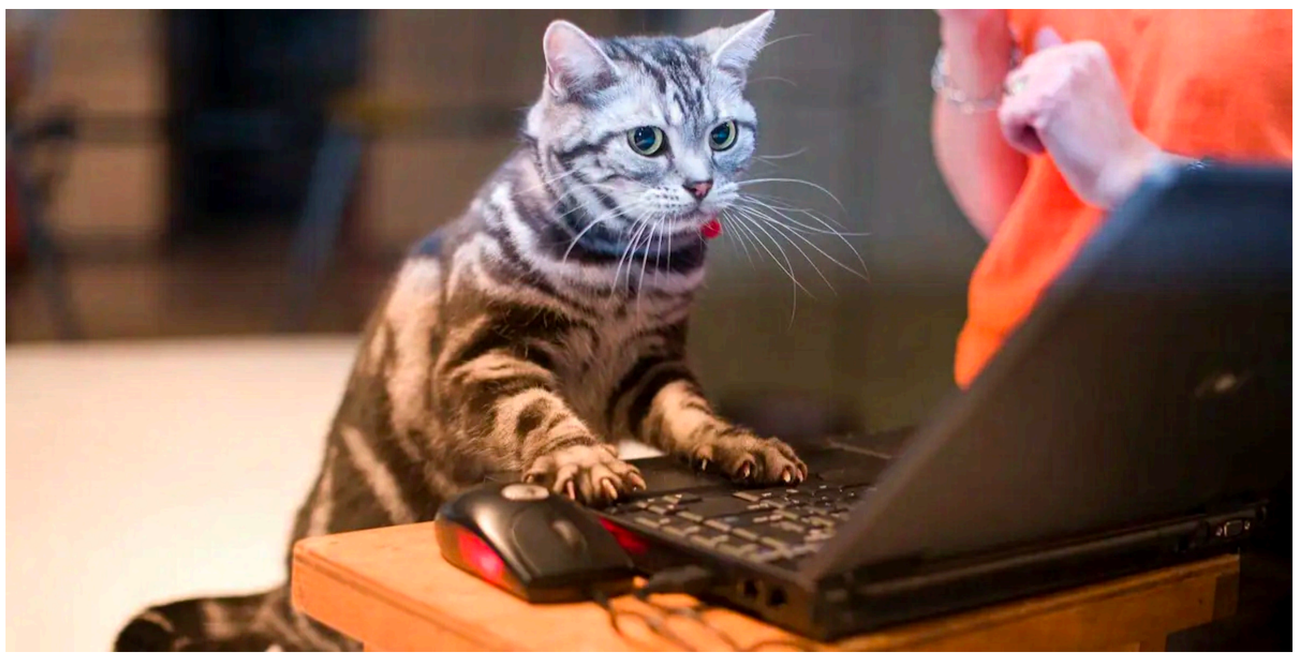


Recap

- There are a variety of large data experiments
  - Latency is often a critical element in the design
- HPCs & other computing sites are not necessarily the best
  - Coming up with a scheme/strategy to do this
- Have done a number of studies to show how this is possible
  - Requires new software stacks
  - Requires different approaches to building out the system
- Expect to have many more challenges coming soon
  - AI is quickly growing throughout the scientific community!



Accelerated AI Algorithms for Data-Driven Discovery

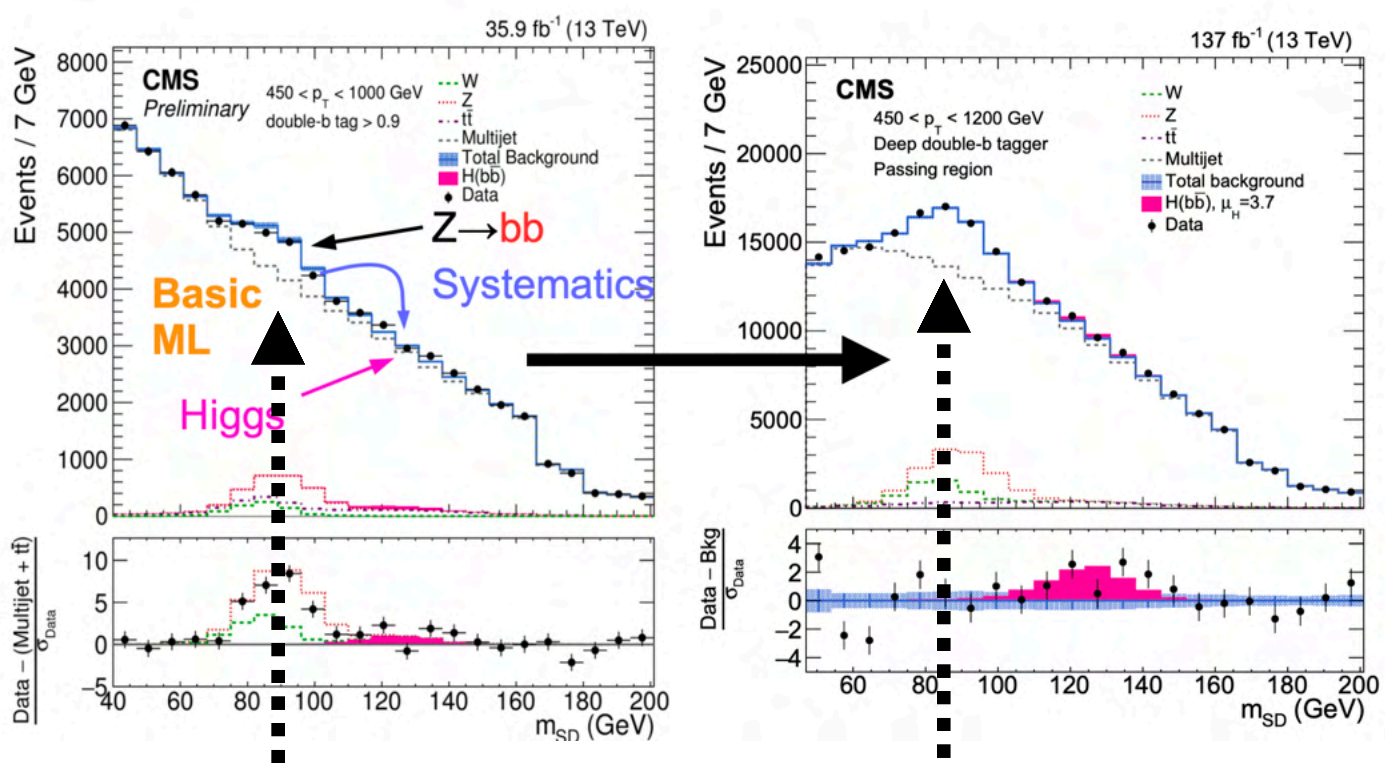


Despite differences in language, there is a common theme

Thanks

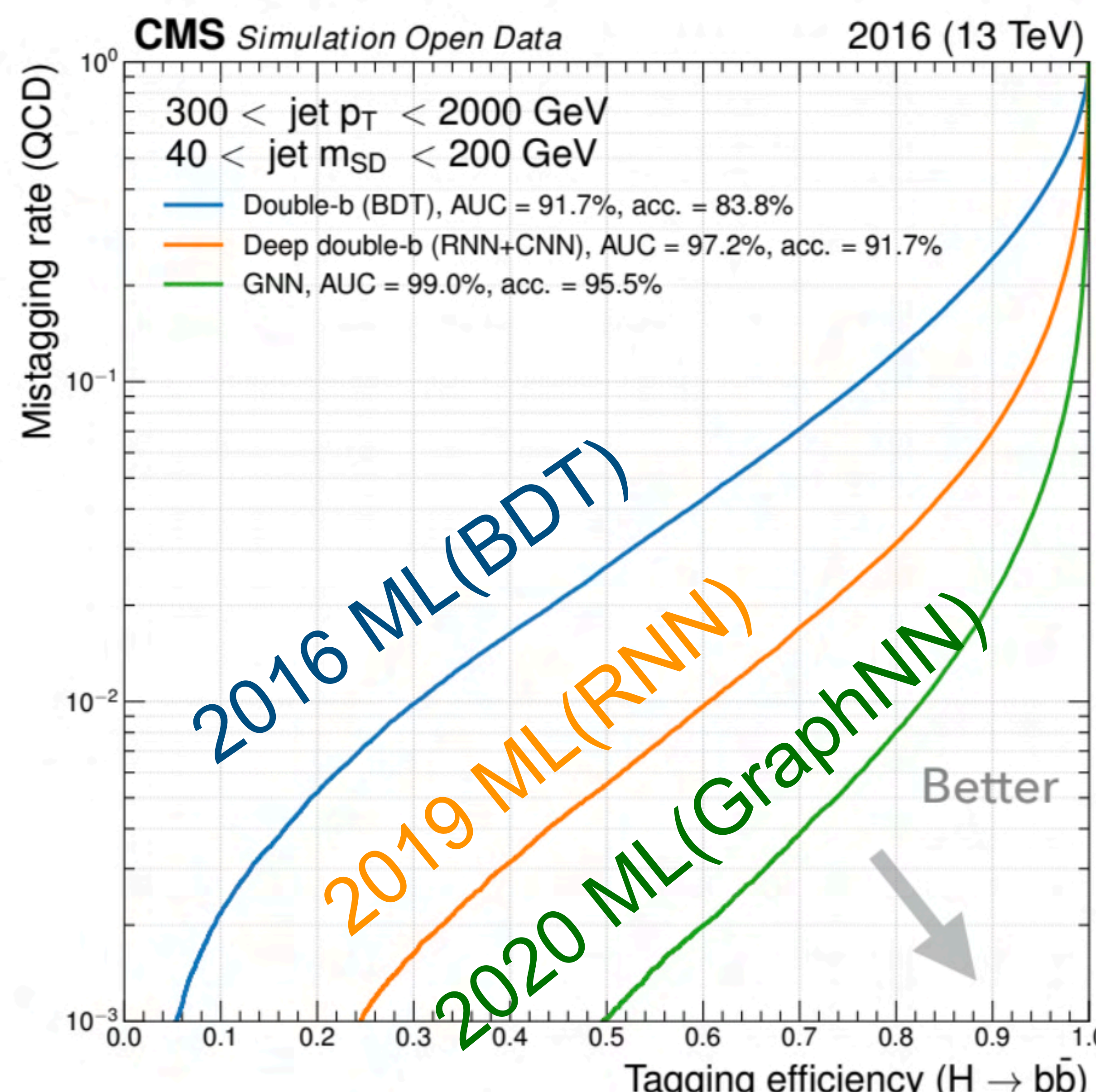
An Angle on AI revolution

- Things are starting to change in the way we compute
  - ML algorithms have the ability to go beyond algorithms
    - This is also b/c GPUs have helped to parallelize computation



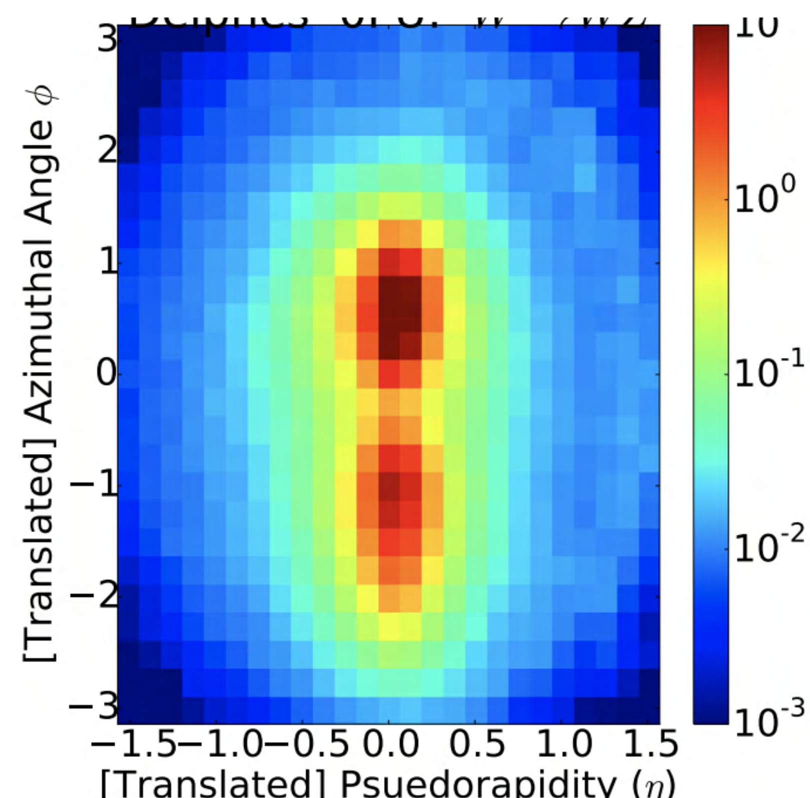
Small ML
Small Peak

Big ML
Big Peak



Deep Learning Progression

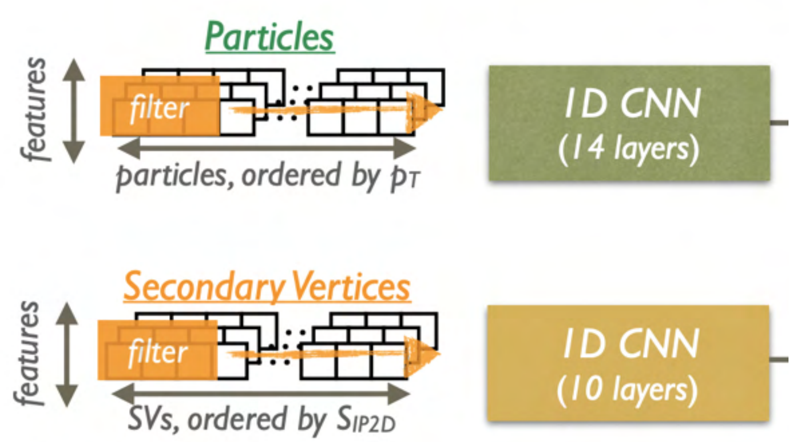
2016



Images

(not lorentz invariant)

2018

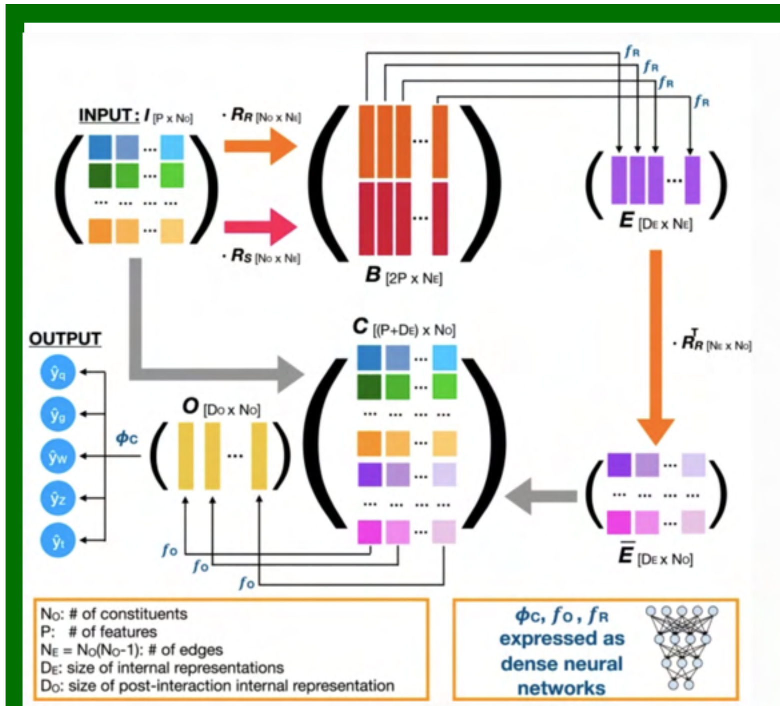


Particles and SVs
with 4-vectors+features

Particles

(limited correlations)

2020



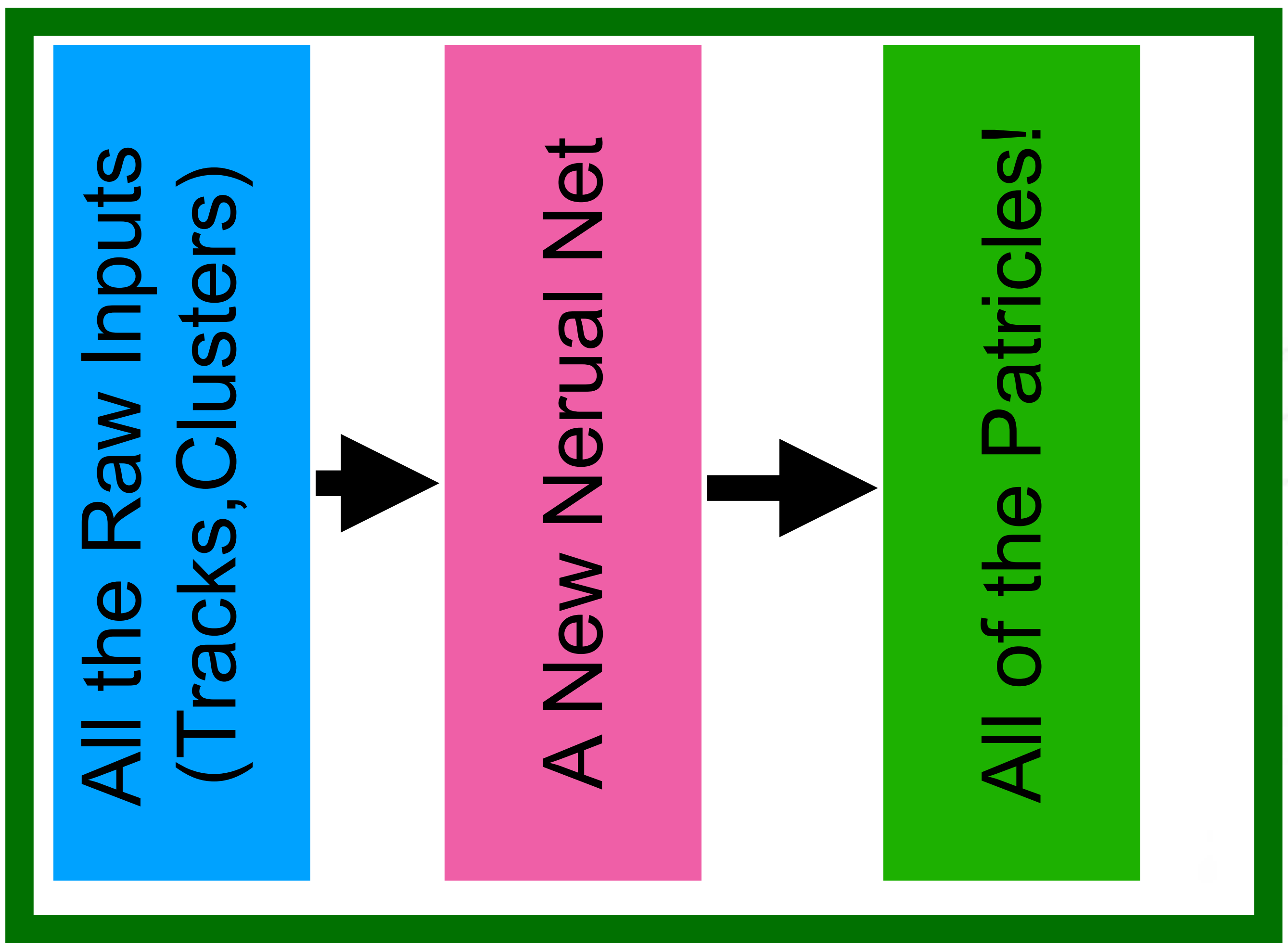
Graphs

(Particles+correlations)

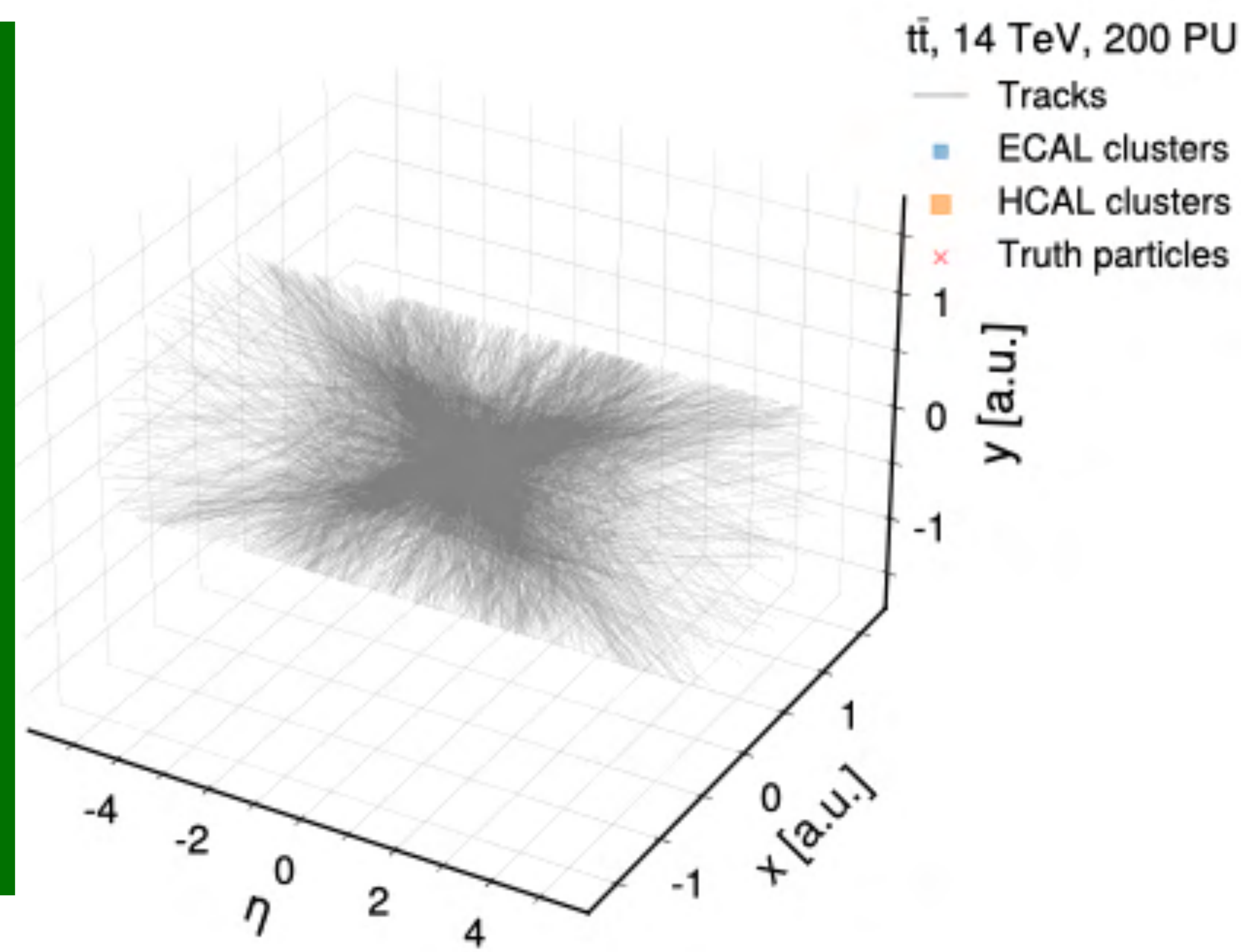
Progressively moving towards use of more info

What does this mean?

- Inevitable that our algorithms will become progressively larger

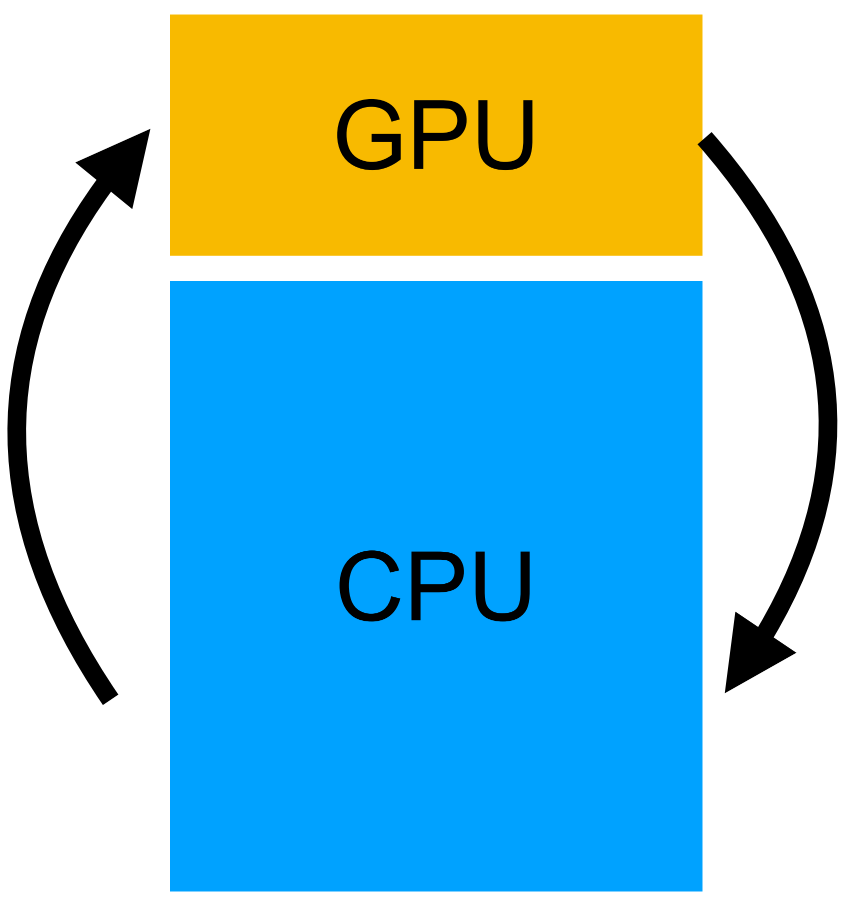


All particles in on fell swoop

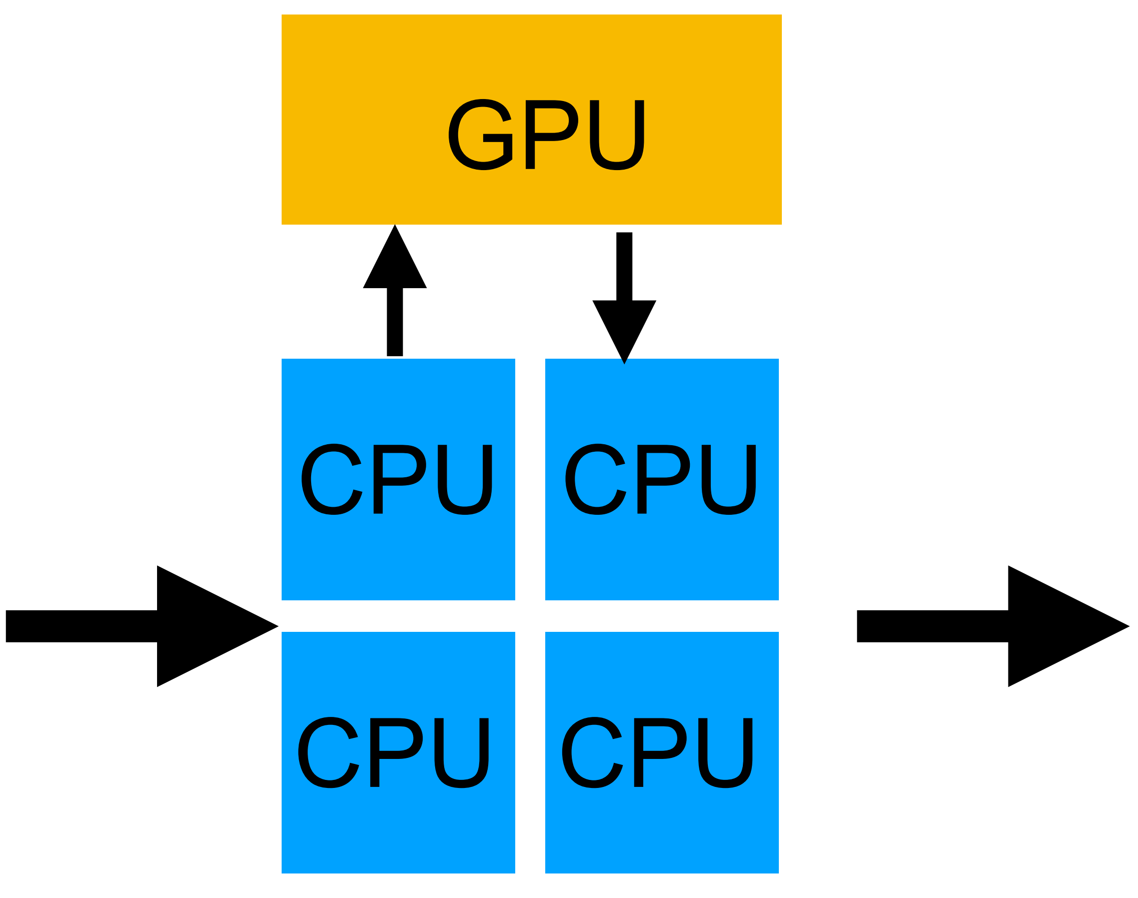


Algorithm Needs

- With the development of AI algorithms we need two things
  - Training and Testing
  - Processing power to run on the data



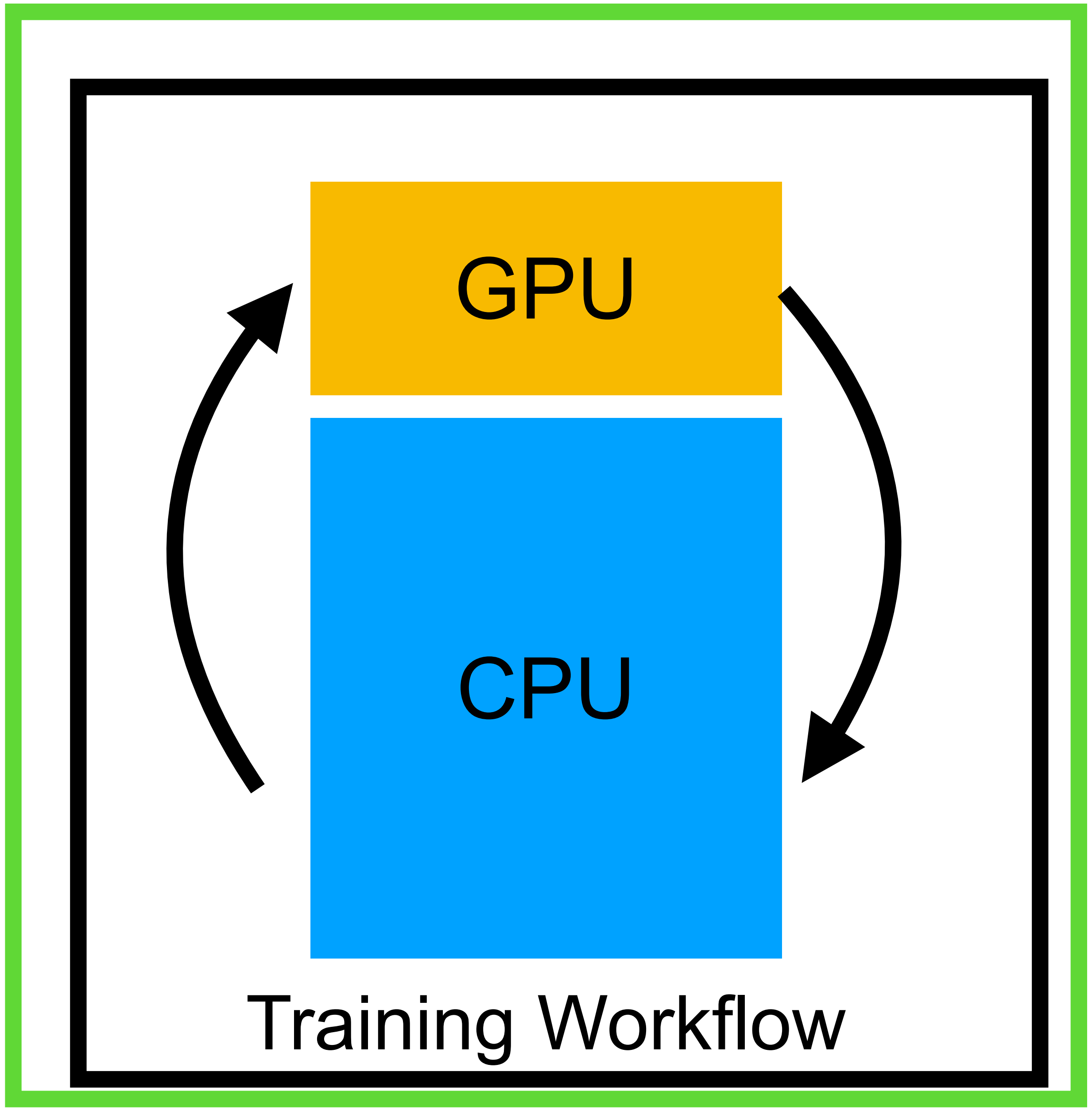
Training Workflow



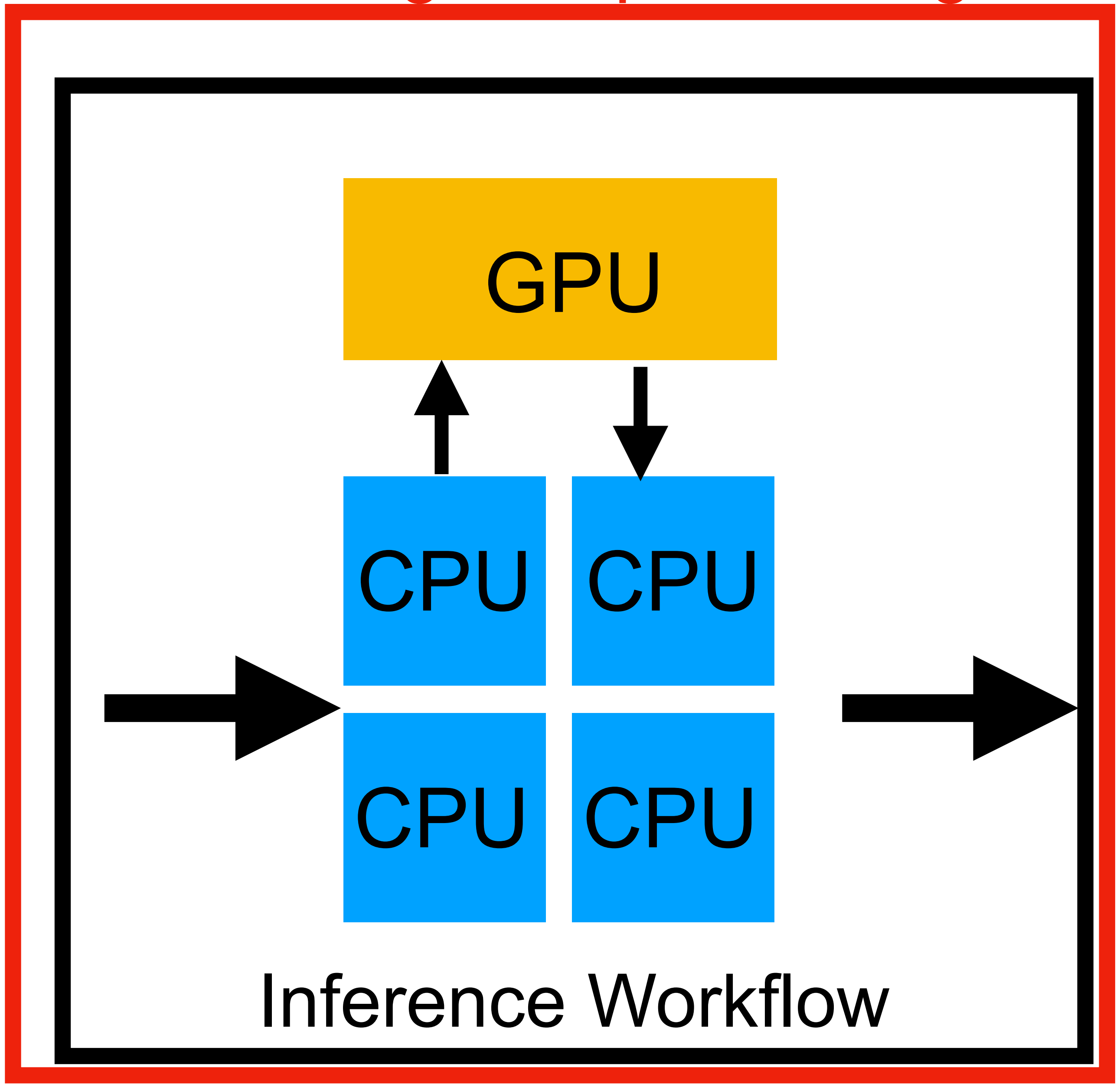
Inference Workflow

Algorithm Needs

Solved
Big HPCs dump as many GPUs as they possibly can in a room
Aim for the maximum compute



What we need
Requires Dynamic allocation to balance GPUs and CPUs focus is on dealing with processing



Algorithm Needs

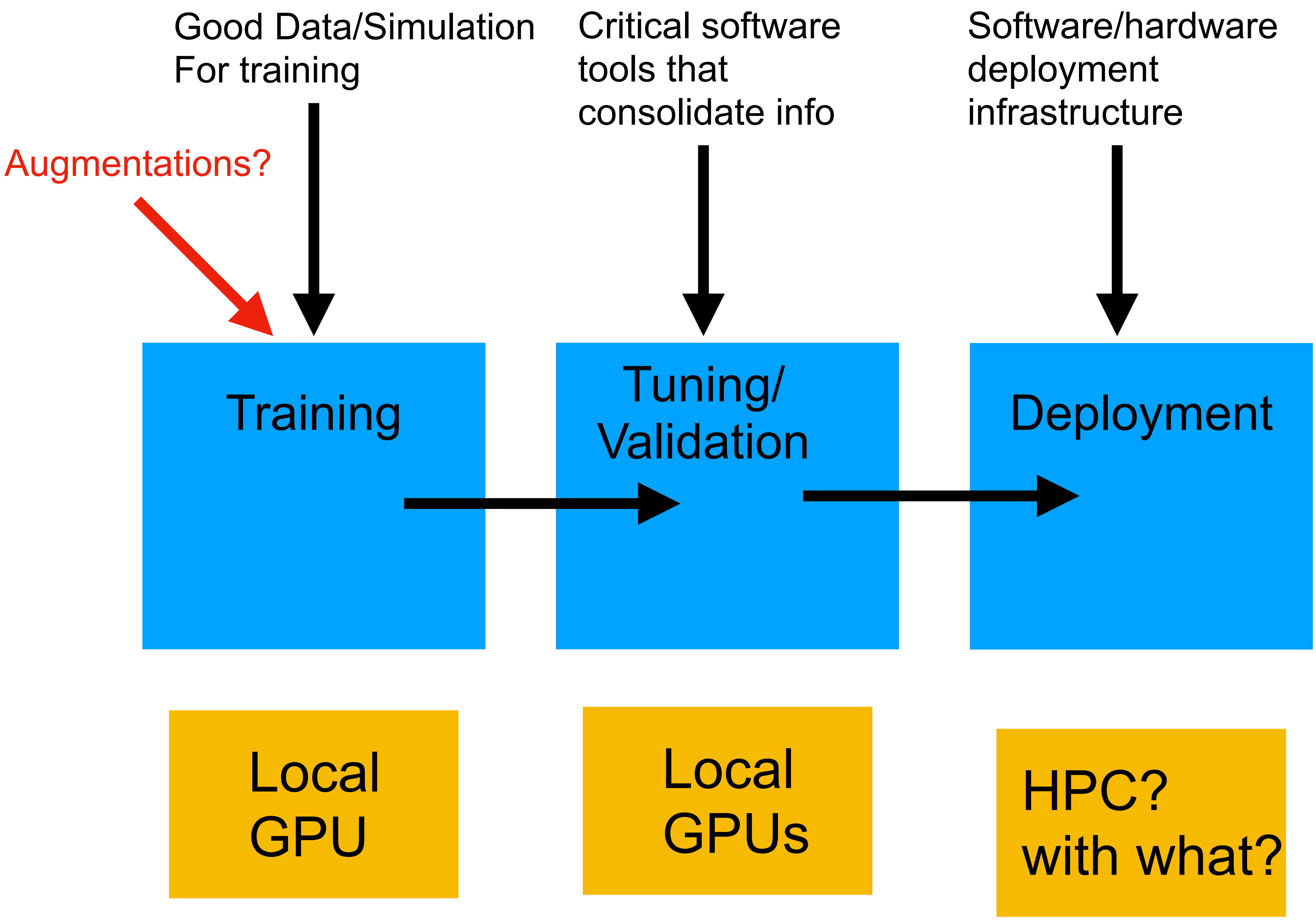
Solved
Big HPCs dump as many GPUs as they possibly can in a room
Aim for the maximum compute

Training Workflow

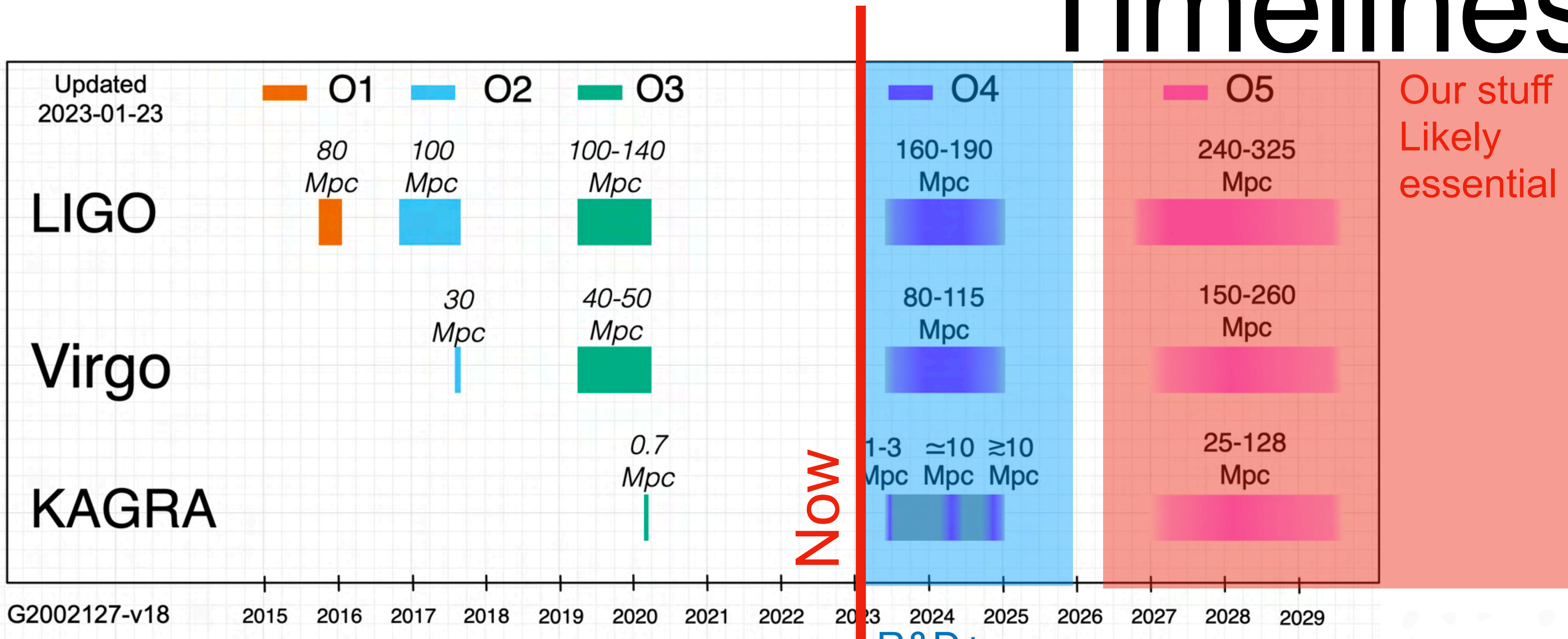
What we need
Requires Dynamic allocation to balance GPUs and CPUs focus is on dealing with processing

Inference Workflow

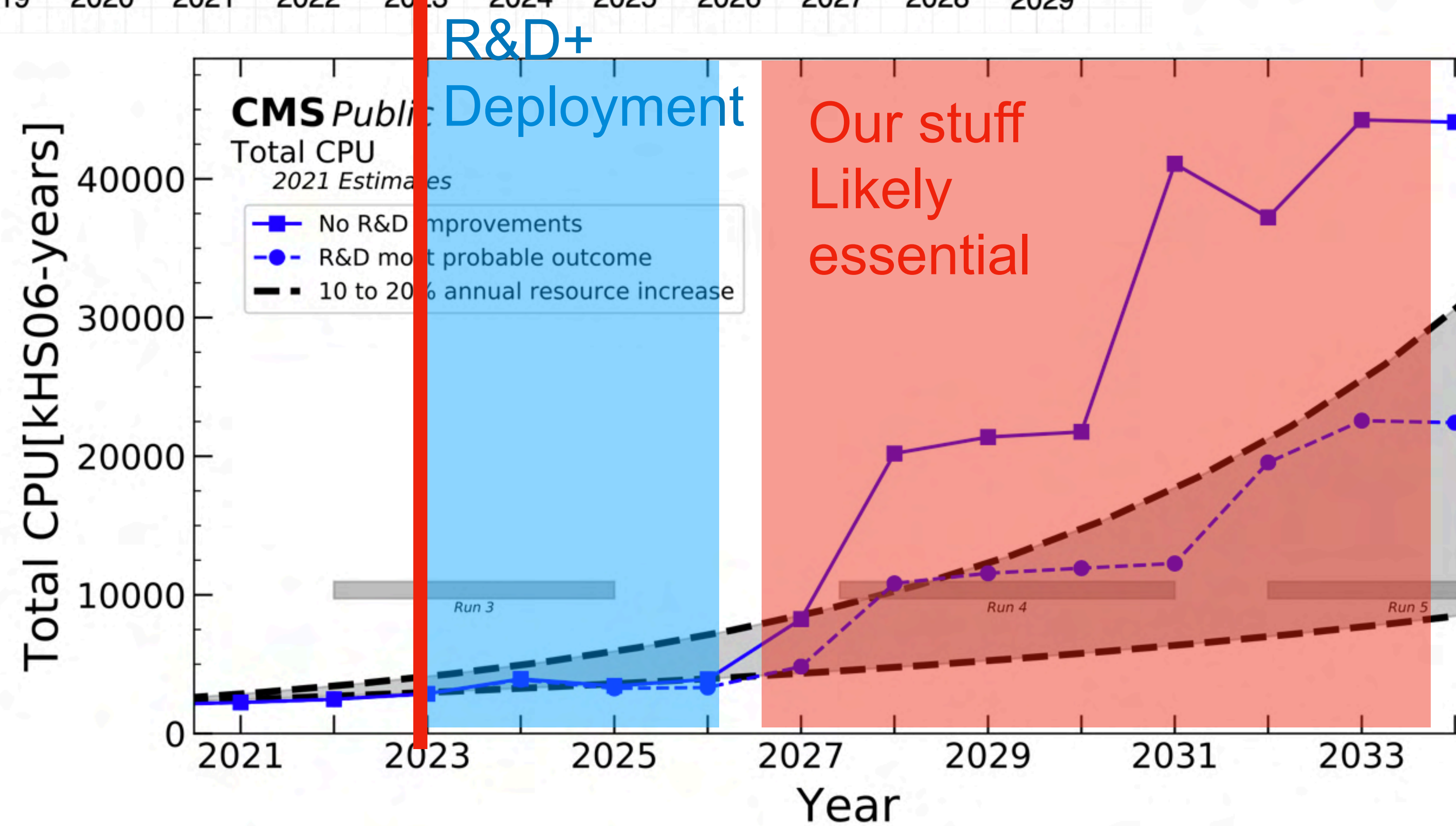
Anatomy of an Algo



Timelines

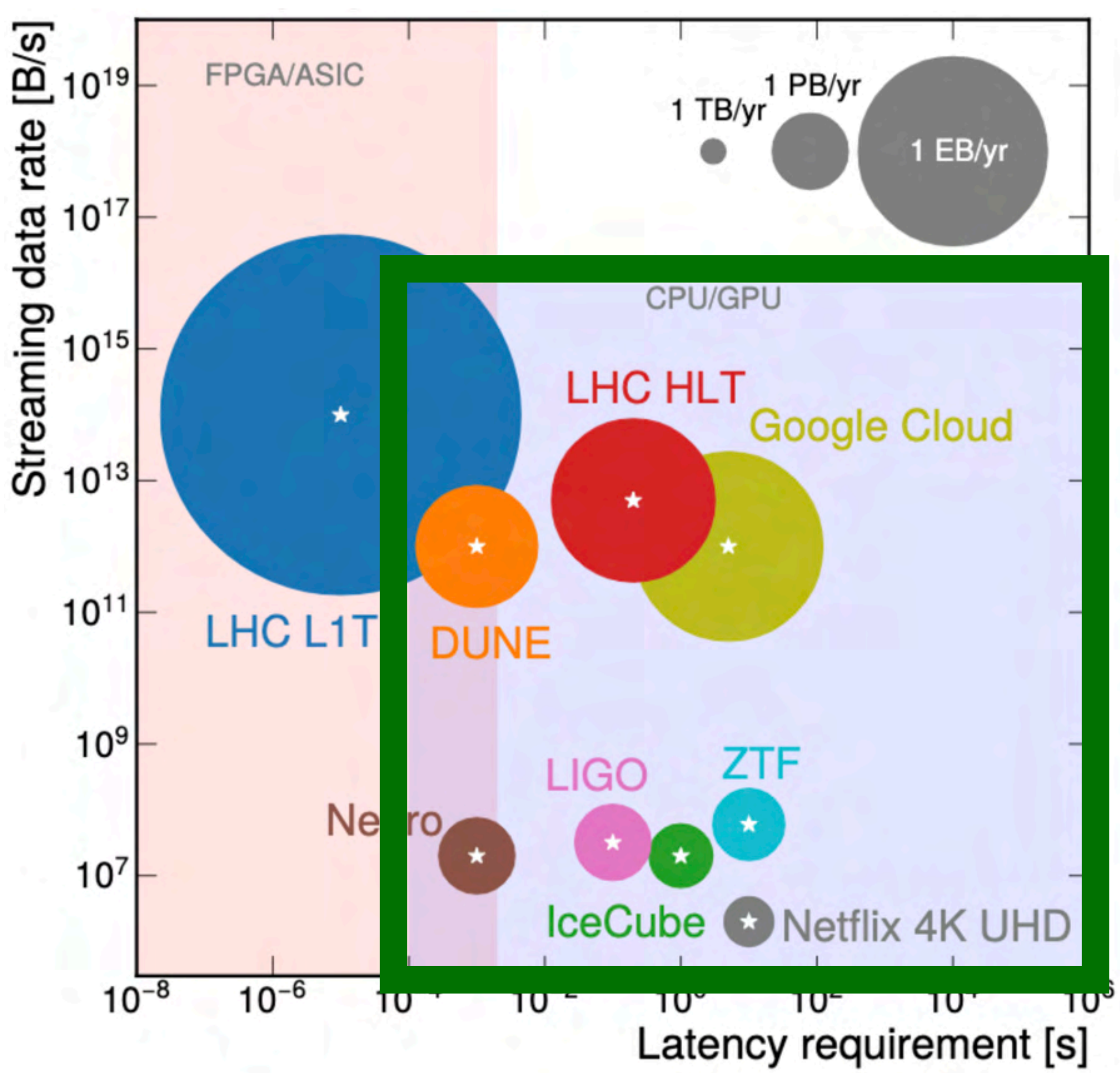


DUNE timeline and various astro timelines (Rubin/LSST)
Should also figure in our overall schedule



What computes are here?

- Within the FastML Community there is a broad range
  - We often try to characterize this range by customization
  - Low Latency and Low Power need more customization



This is our focus here

We want to understand the high throughput component

Visualizing Computing

- All of us in the room require at least one thing in common
  - Computers
  - Also, with GPUs/Coprocessors to accelerate things
- As part of this workshop we would like to create a graphic
  - This graph illustrates the computing demands
  - We hope this graphic can be used as a motivator
- The A3D3 graphic has gotten a lot of traction
  - Highlighting the specific challenges for this conference helps
  - Would like to share this with HPCs as a motivator

ML Challenges

- Through the HDR community
  - We are working to organize a set of ML Challenges
  - Aiming to align this work with two other communities
  - MLCommons scientific (through ML tiny)
  - FAIRUniverse grant aimed at supporting

White Paper
Really some reasonable source explaning

→

ML Challenges
Assemble a list from a few domains

→

Construction
Construct the FAIR dataset test this guy

→

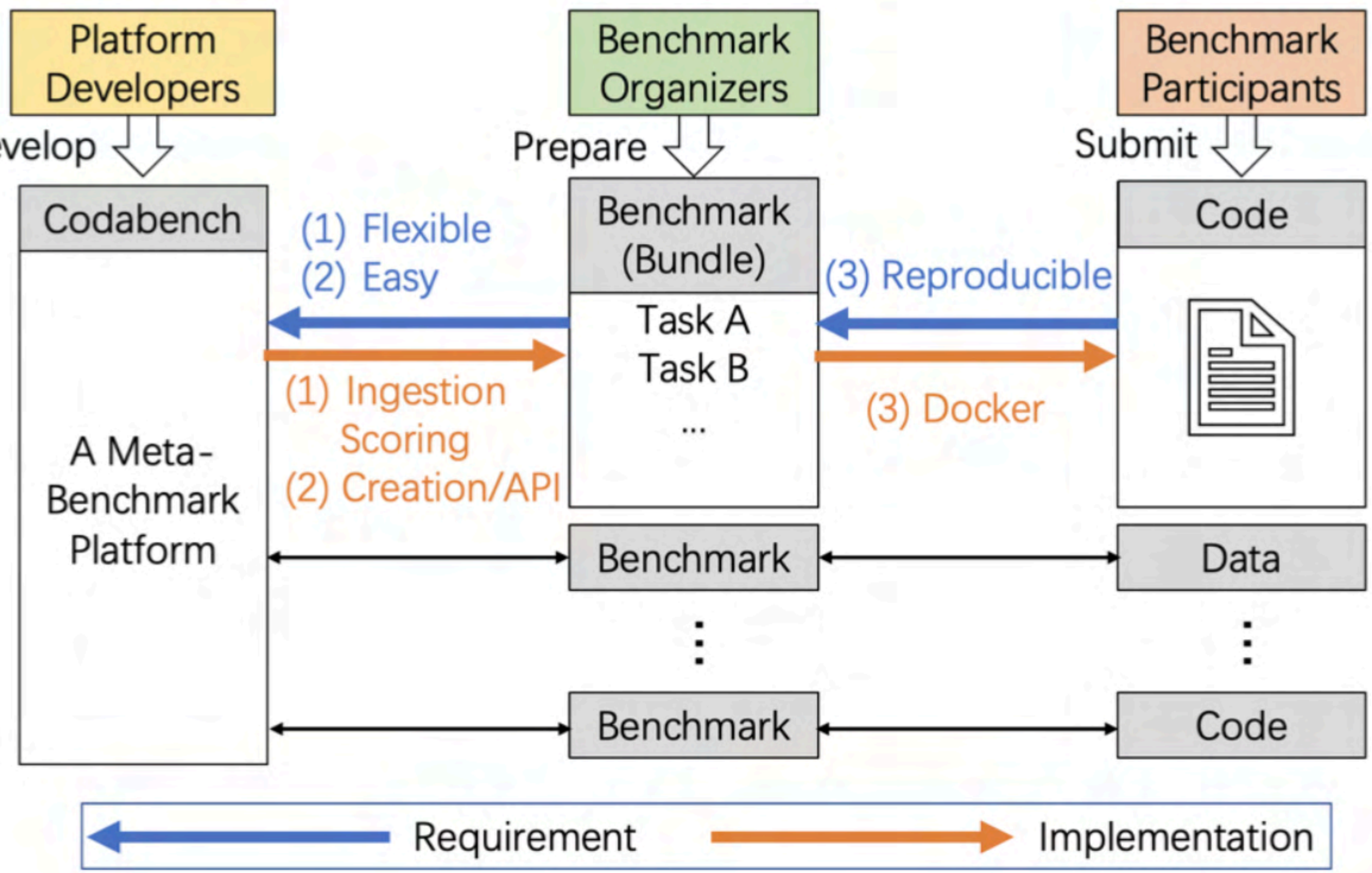
FAIRUniverse
Scheme to deploy models & challenges

- Annual Bootcamp at UW to award results & have a tutorial

FAIRUniverse has established Infrastructure

Codabench and "Fair Universe" Platform

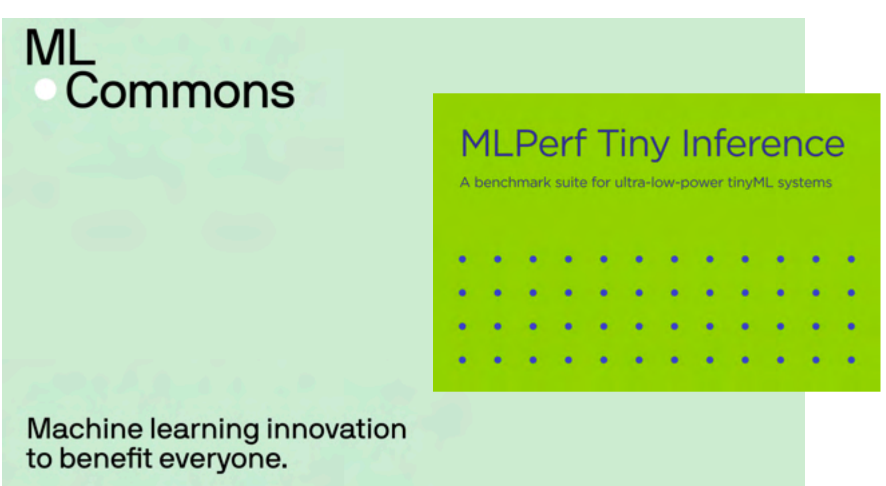
Based on https://www.codabench.org/

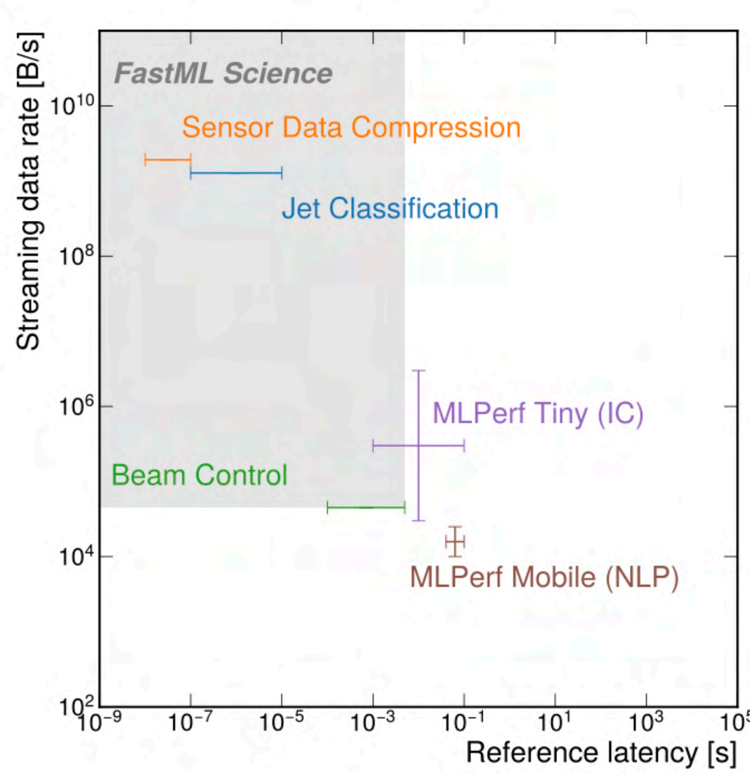


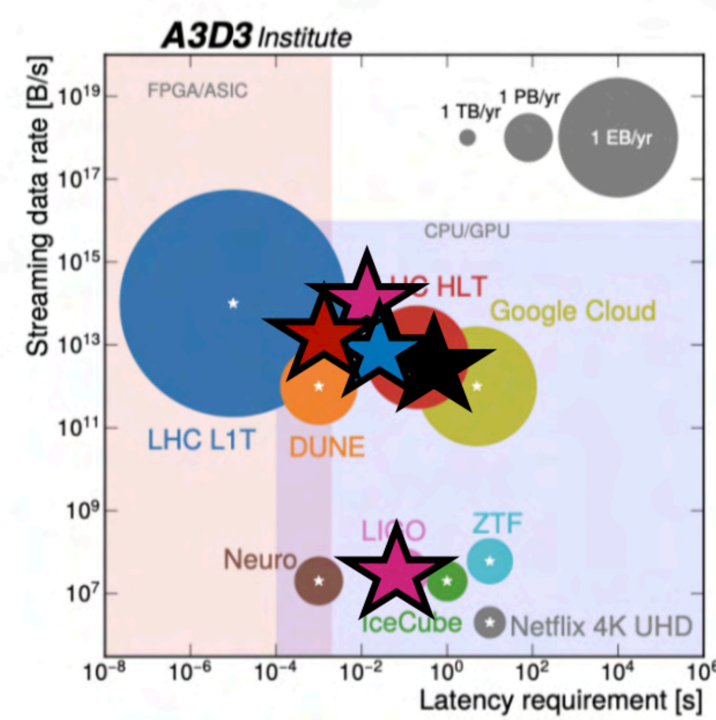
https://docs.google.com/presentation/d/1hqnlvmMgPgVfm7GzDjb6vJfgafl3PRInd9SX1H0GoFA/edit?usp=sharing

Idea for ML Challenges

- There is one underway Icecube Kaggle Challenge
- Dylan's talk from FastML lists some HEP benchmark motivations
  - LHC tracking as a new benchmark
  - LIGO DeepClean as another benchmark
- More complicated challenges
  - Can we make a data generation challenge, or scheduling







A Point to Highlight

- The best way for us to collaborate across domains
  - Making easy-to-use curated datasets or ML problems
  - We have the people in house to really test these datasets
- This is also a way to tie the different domains together
  - We can use this white paper to start testing out our challenges
    - Preparation of datsets
    - Release of models
- Can we get a dataset/model from each scientific domain
  - Also do we have the right benchmarks to do this?

Conclusions

- Welcome! Enjoy your time here in Cambridge
  - We would like to write a white paper
  - We have some discussion time at the end of the conference
- Outline for the White paper (Lets keep it short!)
  - **Discussion of computing tools and software**
    - Path to aligning these across domains
  - **List of critical models in the field**
    - What makes these models
  - **One plot to rule them all and bind these sections**
- A roadmap for future computing can helps us move this forward

Backup