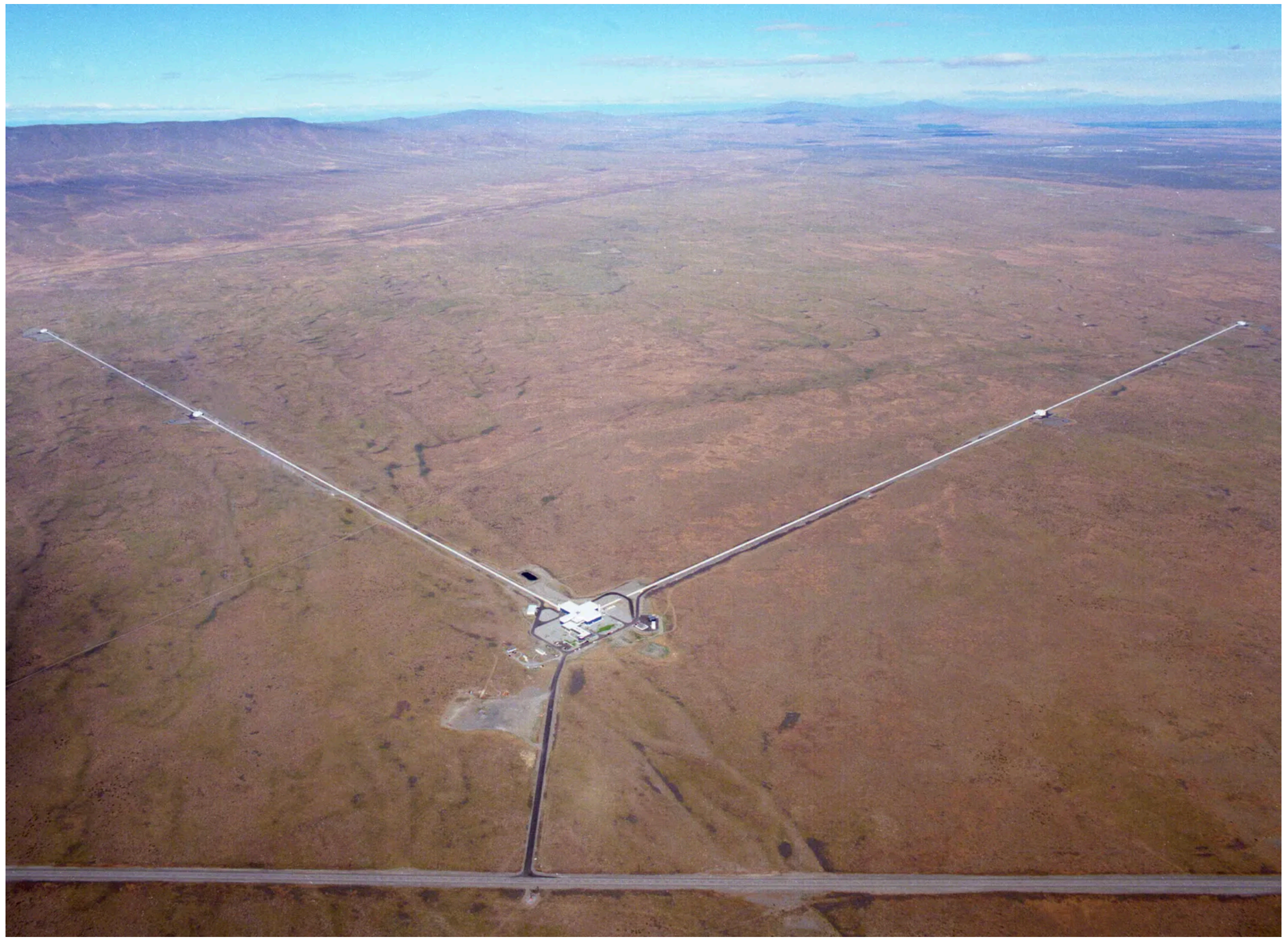
AI/ML Computing for Gravitational Waves

Accelerated AI Algorithms for Data-Driven Discovery



Phil Harris(MIT)
A3D3 deputy director

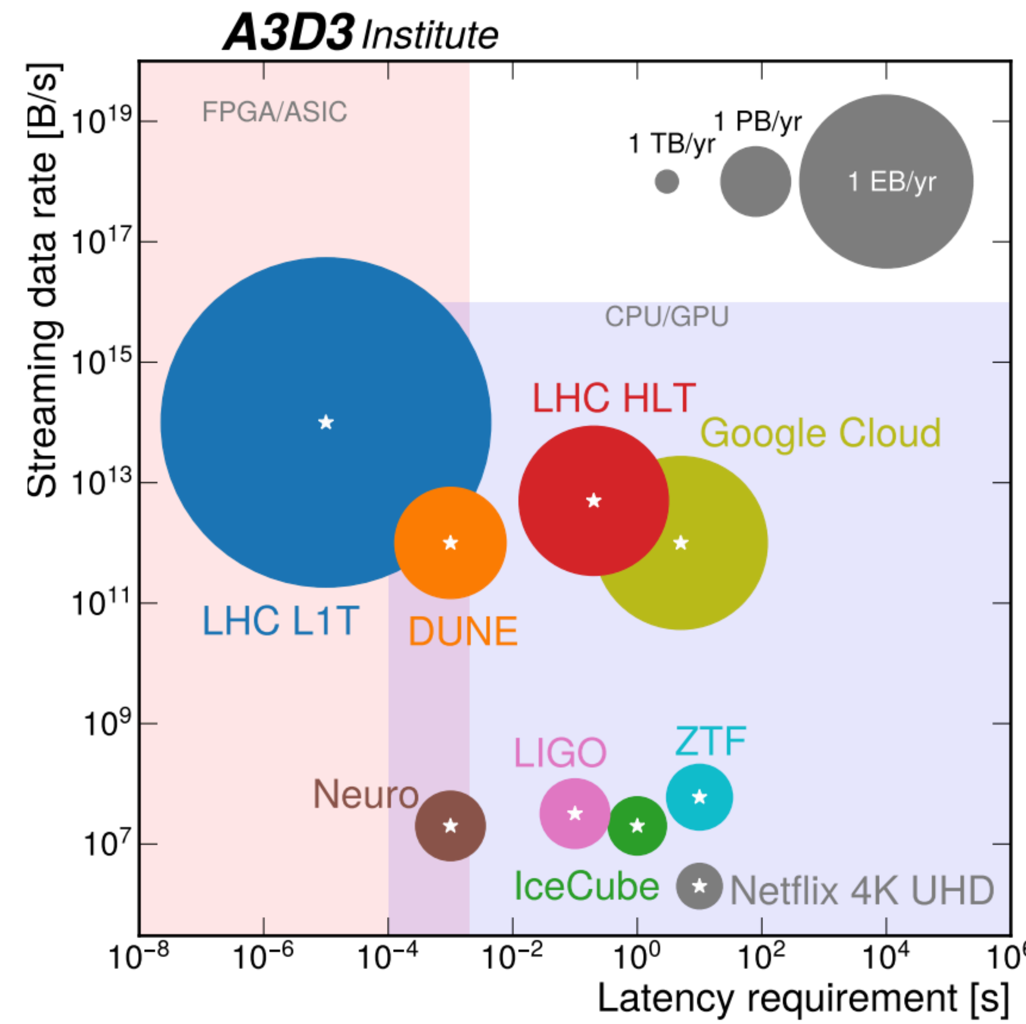
Alec Gunny[1], Ethan Marx[1], Will Benoit[2], Deep Chatterjee[1], Michael Coughlin[2], Katya Govorkova[1], Erik Katsavounidis[1], Eric Moreno[1], Rafia Omer[2], Ryan Raikman[1], Muhammed Saleem[2]

1 - Massachusetts Institute of Technology
2 - University of Minnesota

Phil Harris Lab@MIT

Understanding this Problem

FastML For Science

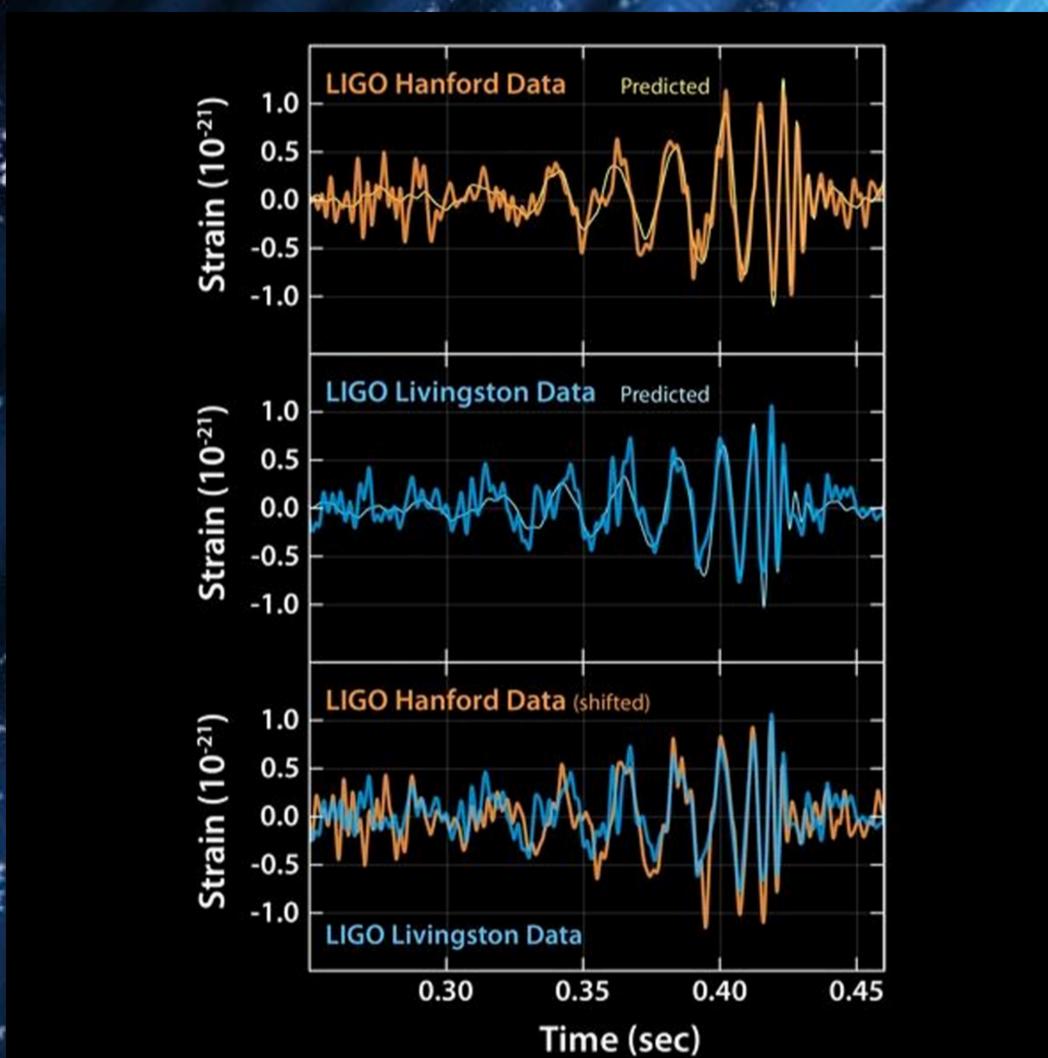


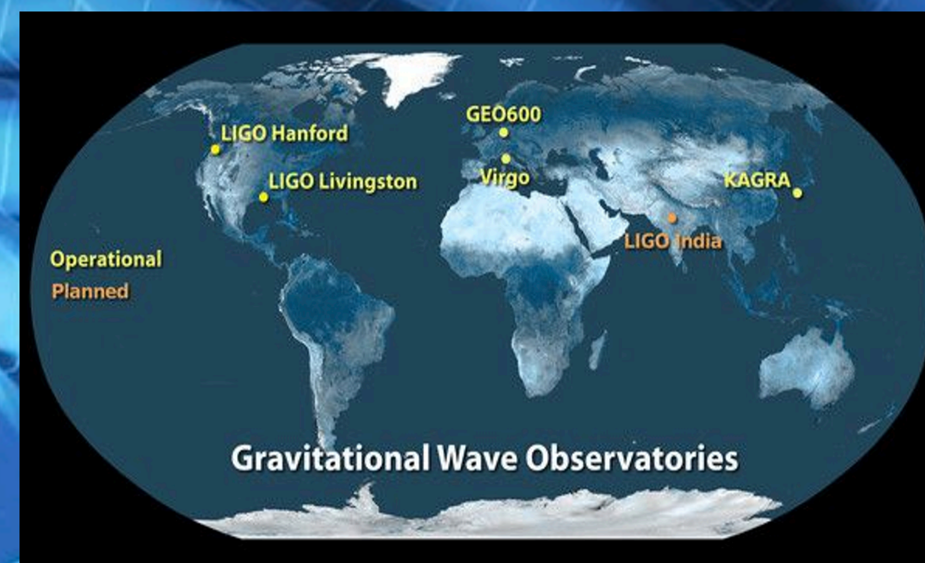
- FastML/A3D3 created to address real-time AI for science
  - Developing ML + GPU integration for large throughput computing
  - Developing ML+ FPGA/ASIC for low latency computing
- Science benchmarks are competitve with the rest of AI world

Gravitational Waves

Ripples in Spacetime

International Gravitational Wave Network to characterize Gravitational Wave

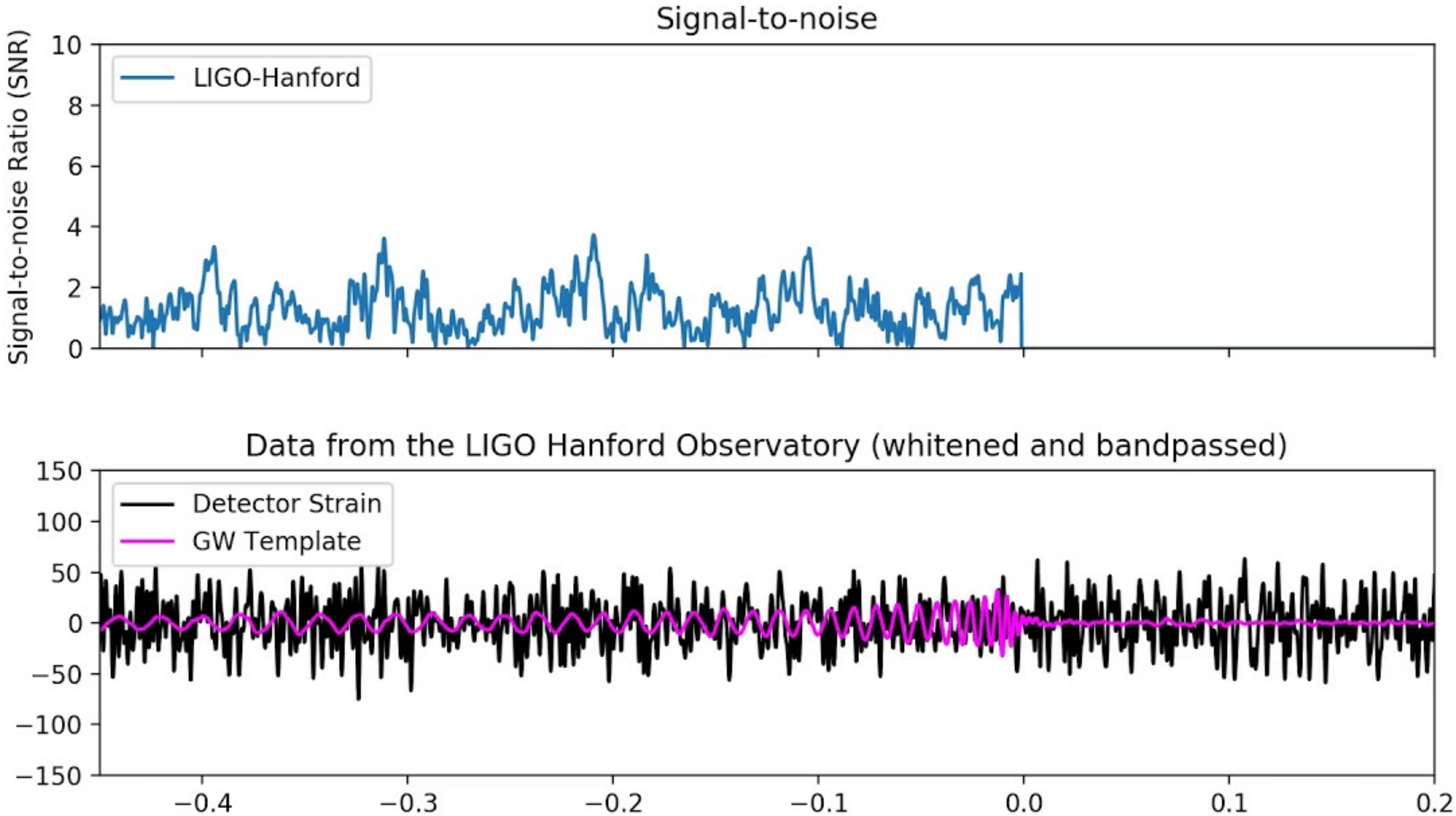




Grand Challenge:

Can we identify GWs fast for downstream telescopes?

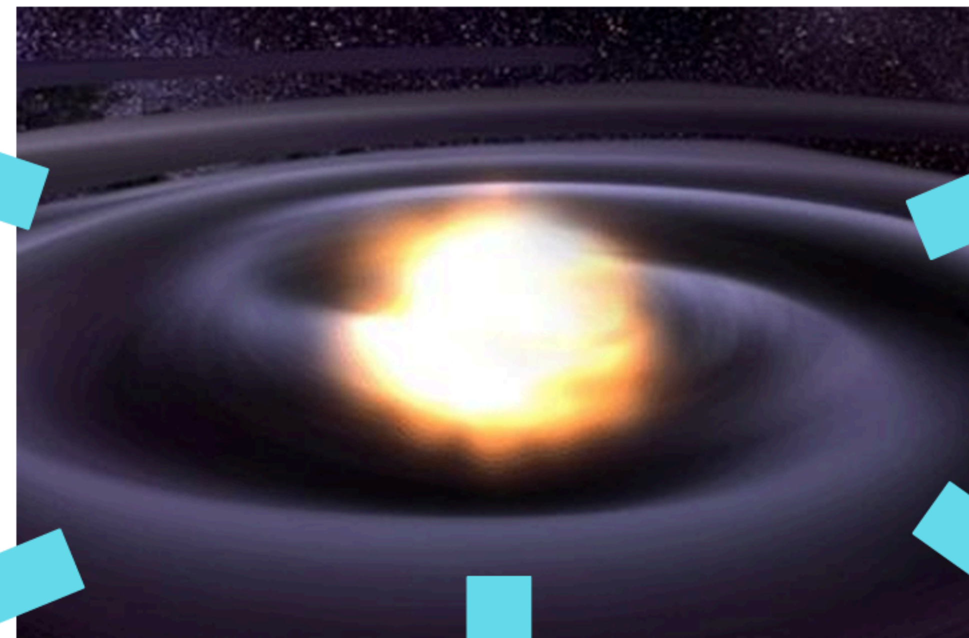
Typical LIGO Signal

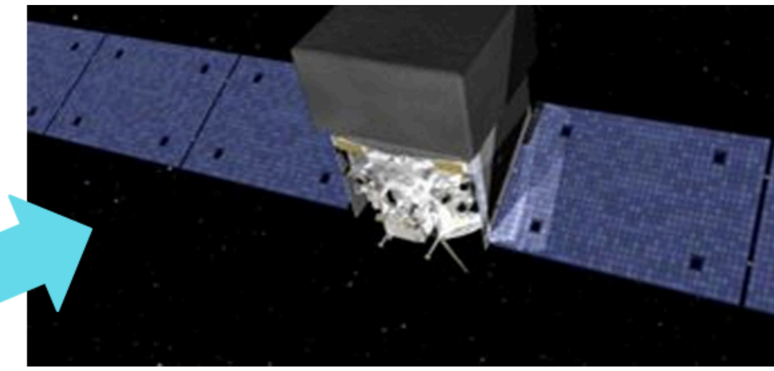


To ensure we aren’t seeing a glitch we use at multiple detectors
2 LIGO detectors in US + Virgo detector in Europe + Kagra in Japan

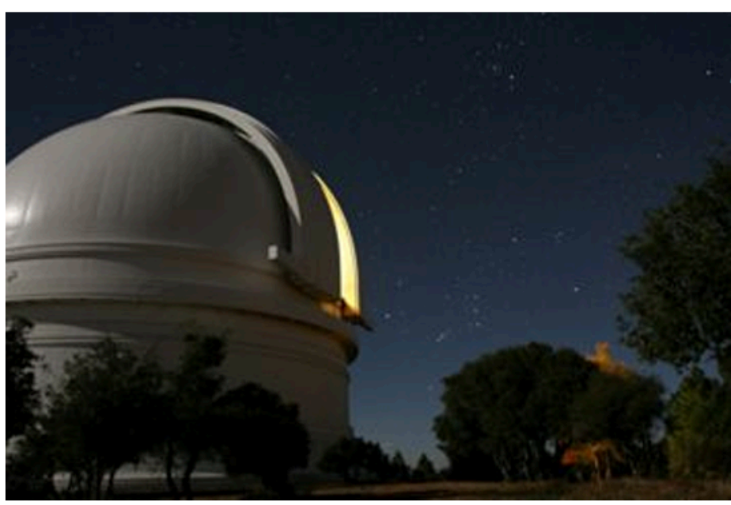
Multi-Messenger Astronomy

Gravitational waves



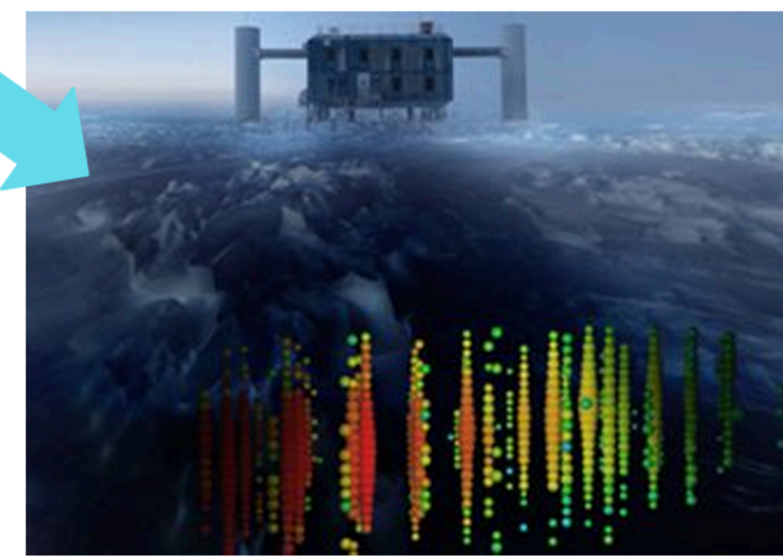


X-rays/Gamma-rays



Visible/infrared light

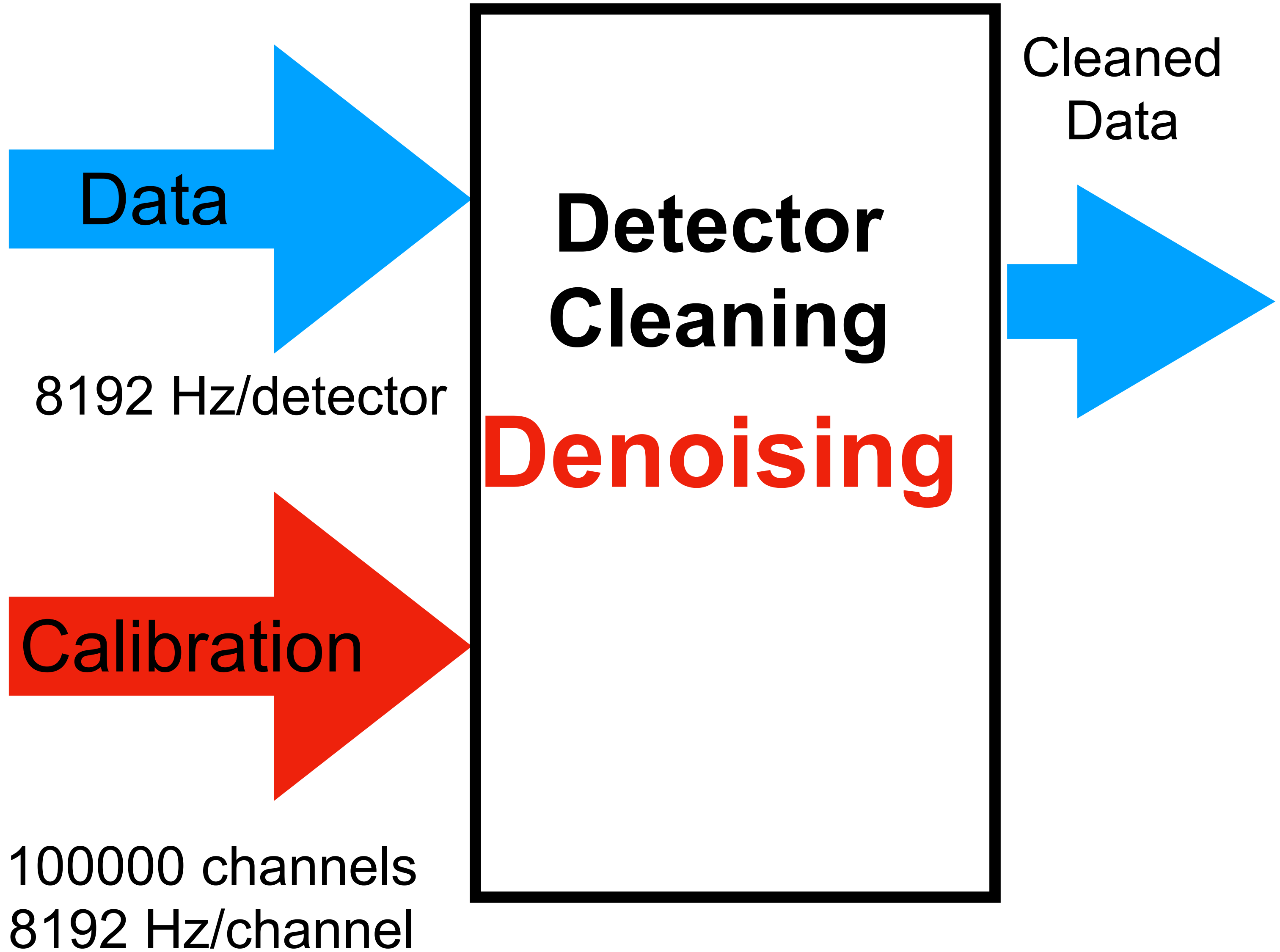
Radio waves



Neutrinos

Performing fast identification of GWs critical to alerting world!

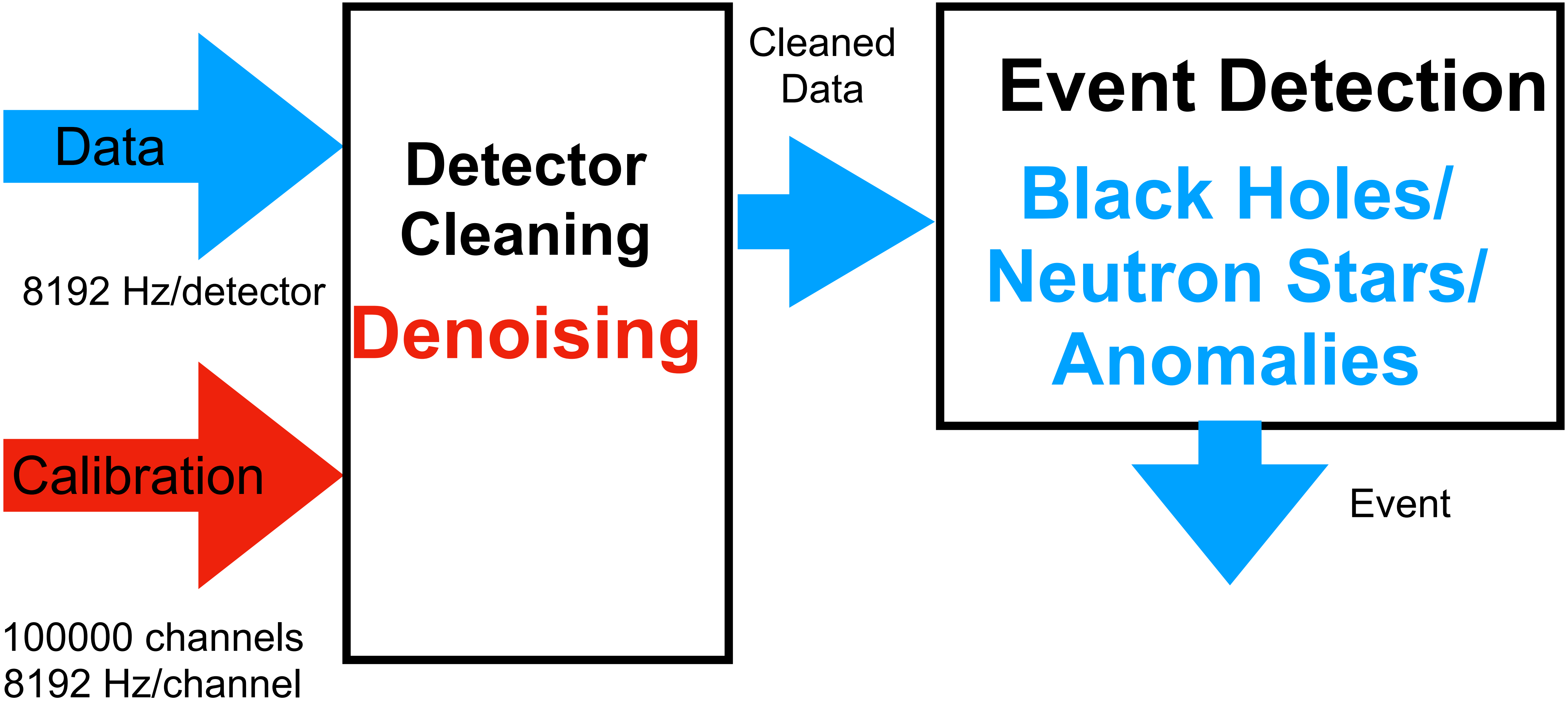
LIGO Data Workflow



Challenge is to run this in real-time

Roughly 1 PB per year

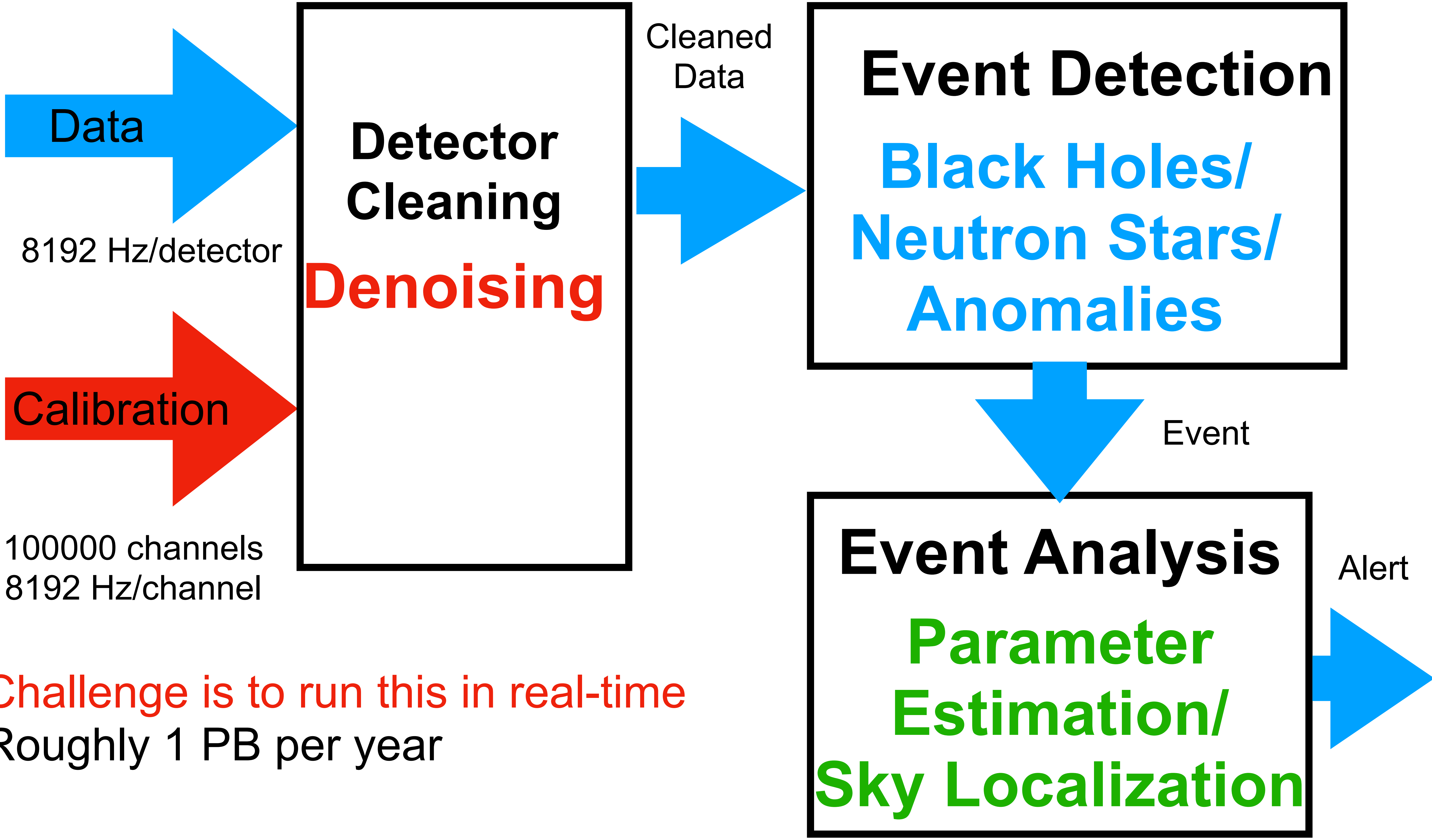
LIGO Data Workflow



Challenge is to run this in real-time

Roughly 1 PB per year

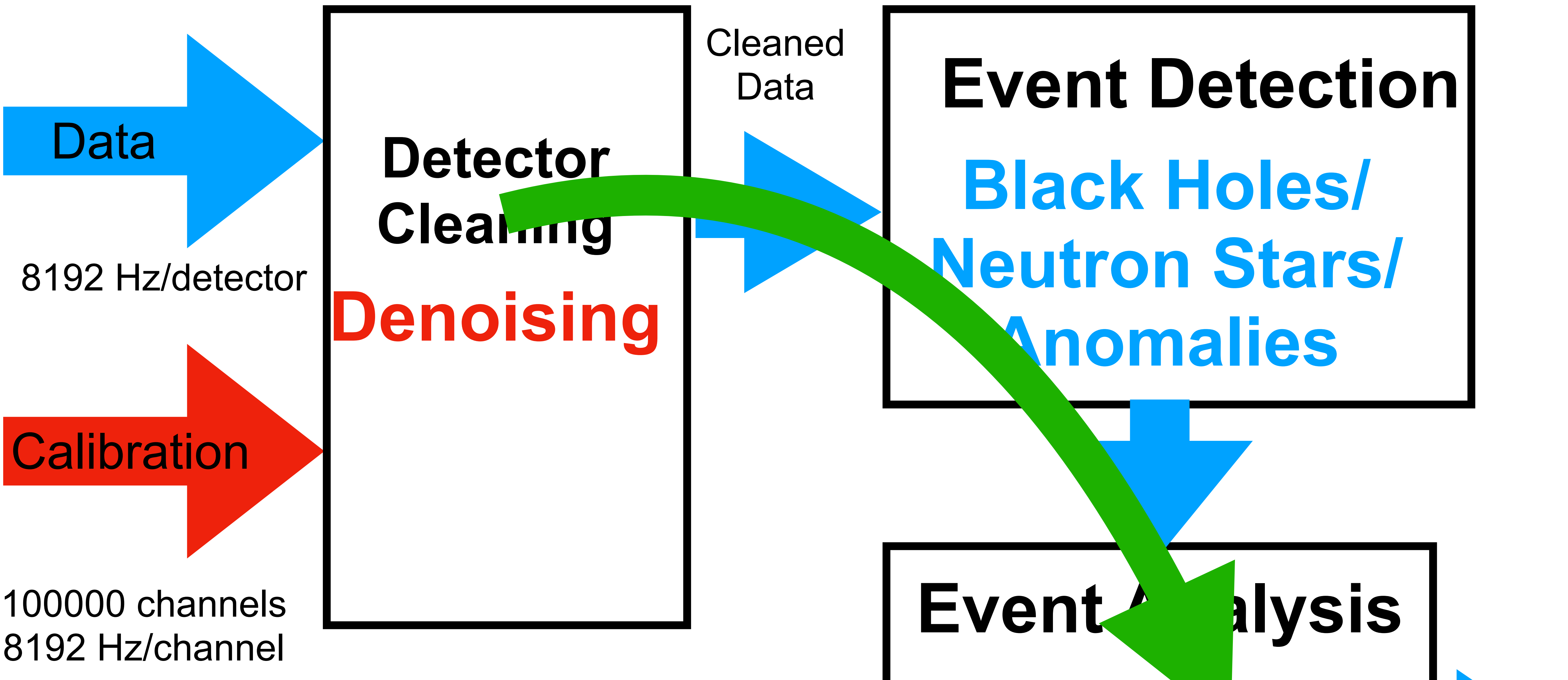
LIGO Data Workflow



Challenge is to run this in real-time

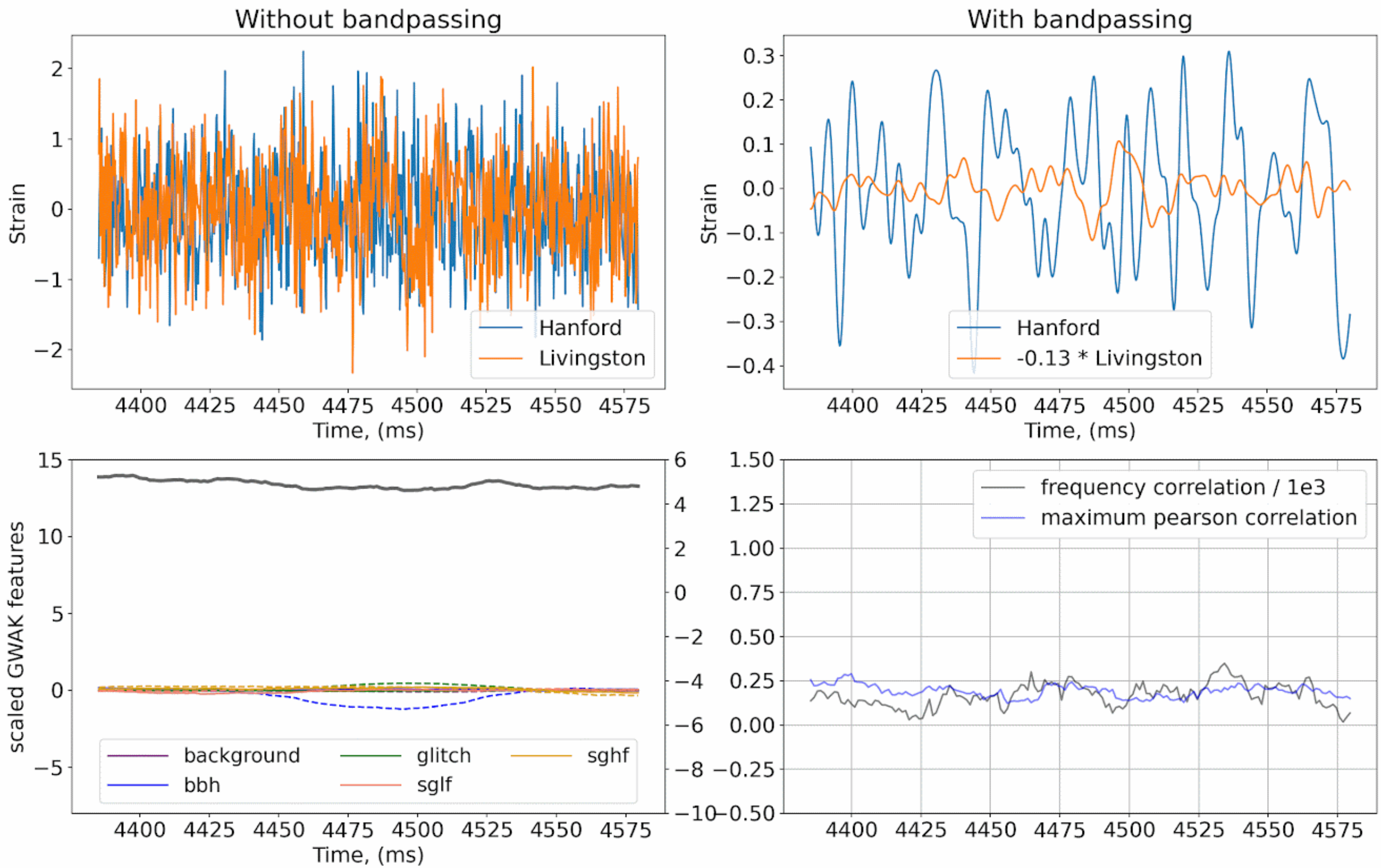
Roughly 1 PB per year

Our Upgraded Workflow



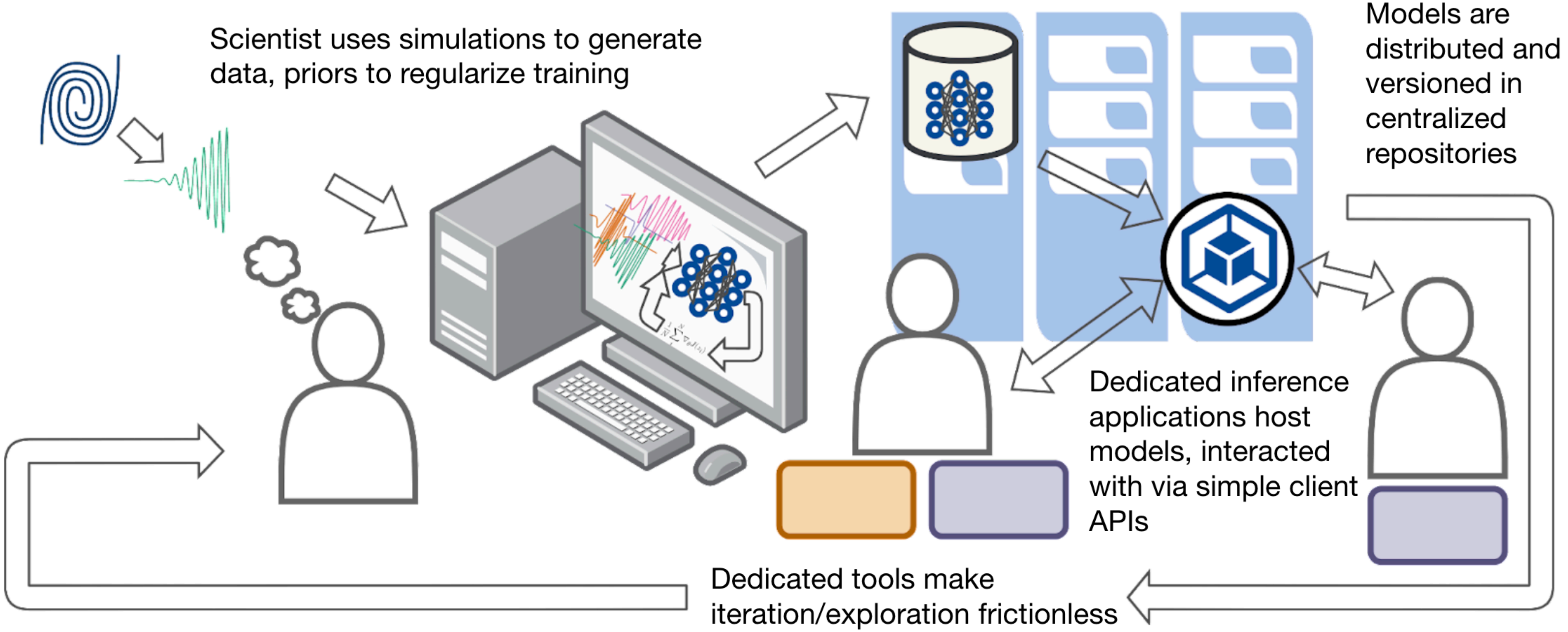
We replaced this pipeline with a chain of deep learning algos

LIGO Data



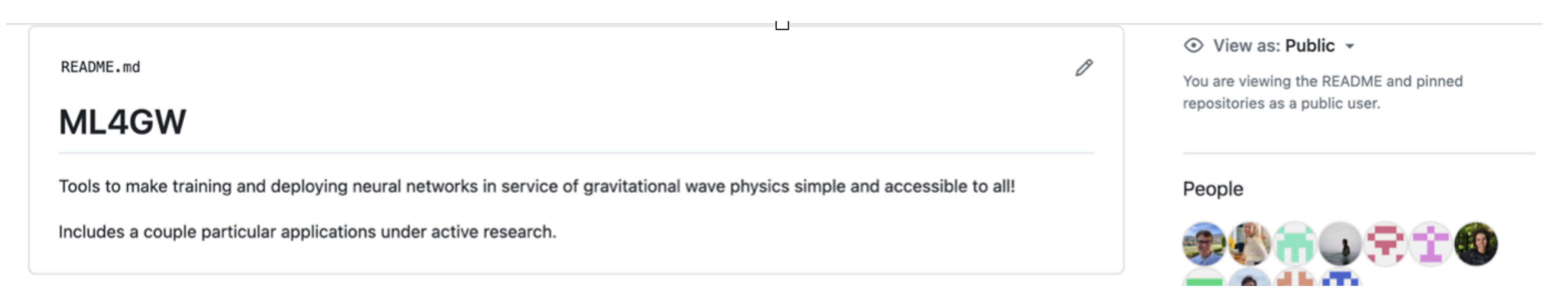
Without bandpassing
Strain
Hanford
Livingston
Time, (ms)
With bandpassing
Hanford
-0.13 * Livingston
scaled GWAK features
background
glitch
sghf
bbh
sglf
frequency correlation / 1e3
maximum pearson correlation

MLOps Toolkit



ML4GW Toolkit

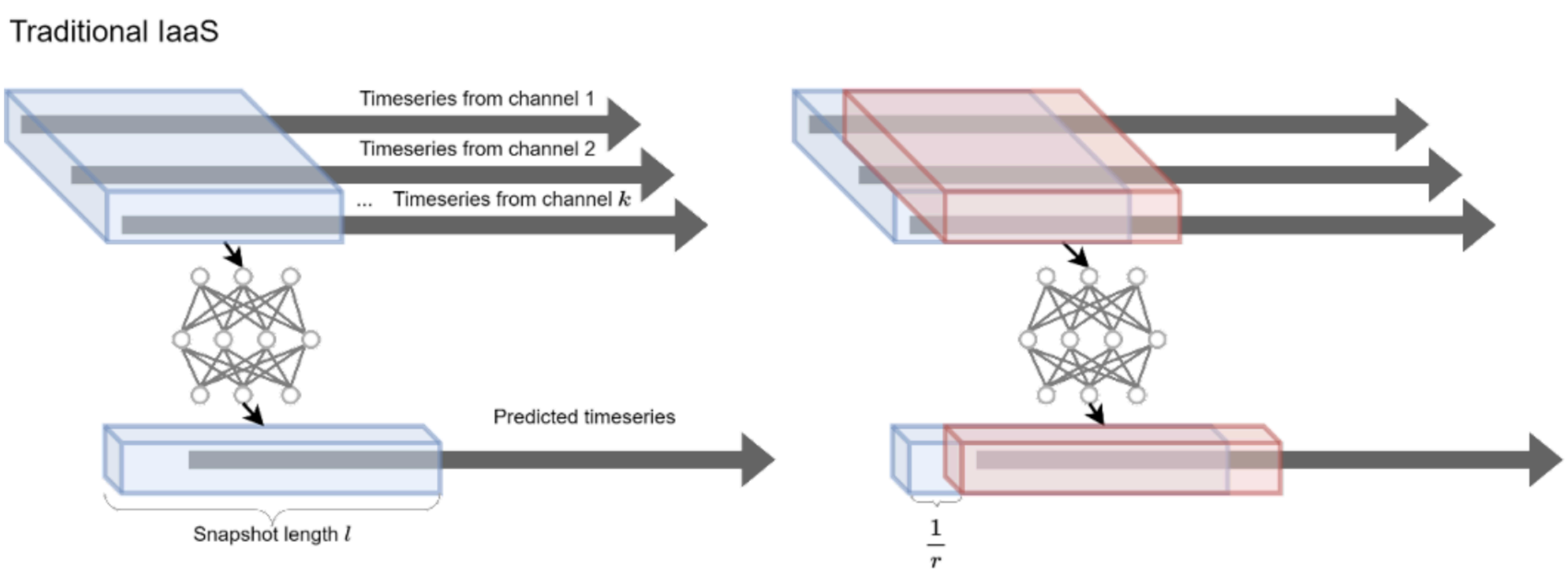
To enable fast deployment + optimized heterogeneity



Example: Stateful Caching

Conventional Processing of Time series data

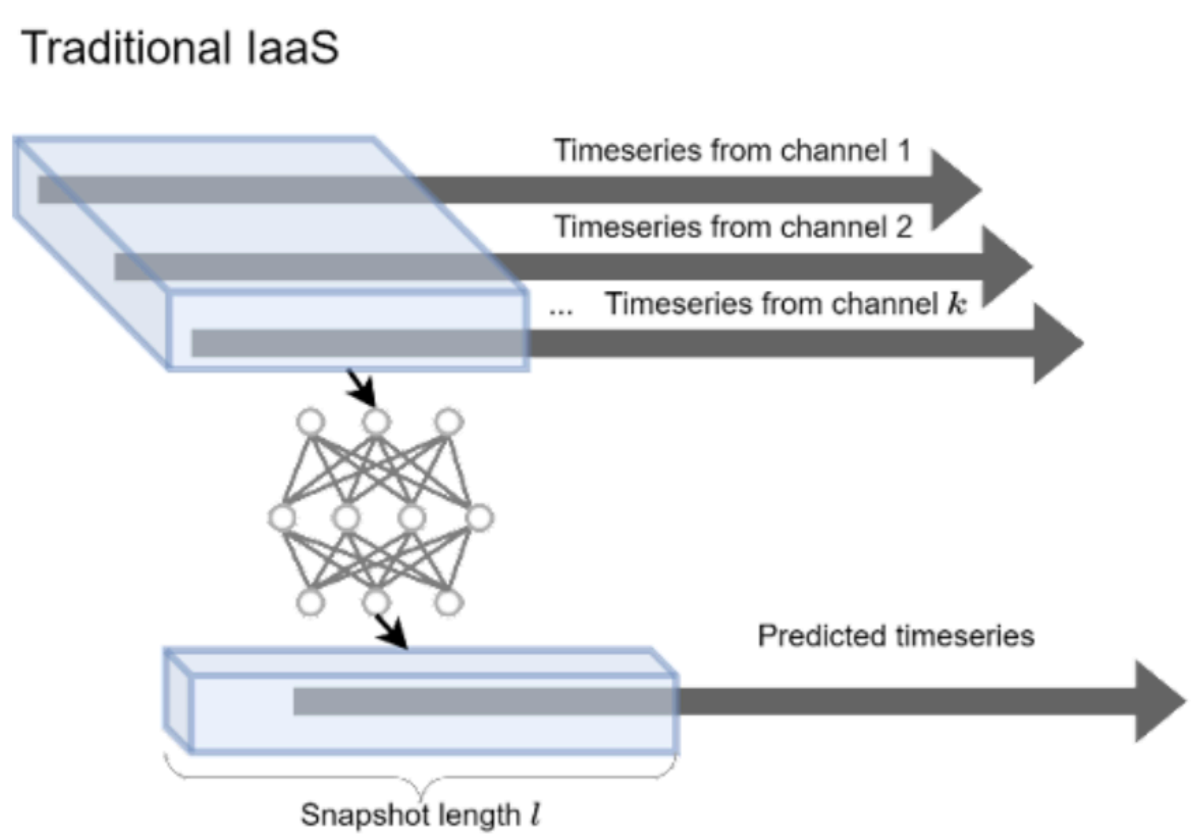
First ML Inference on data   Second ML Algorithm for data



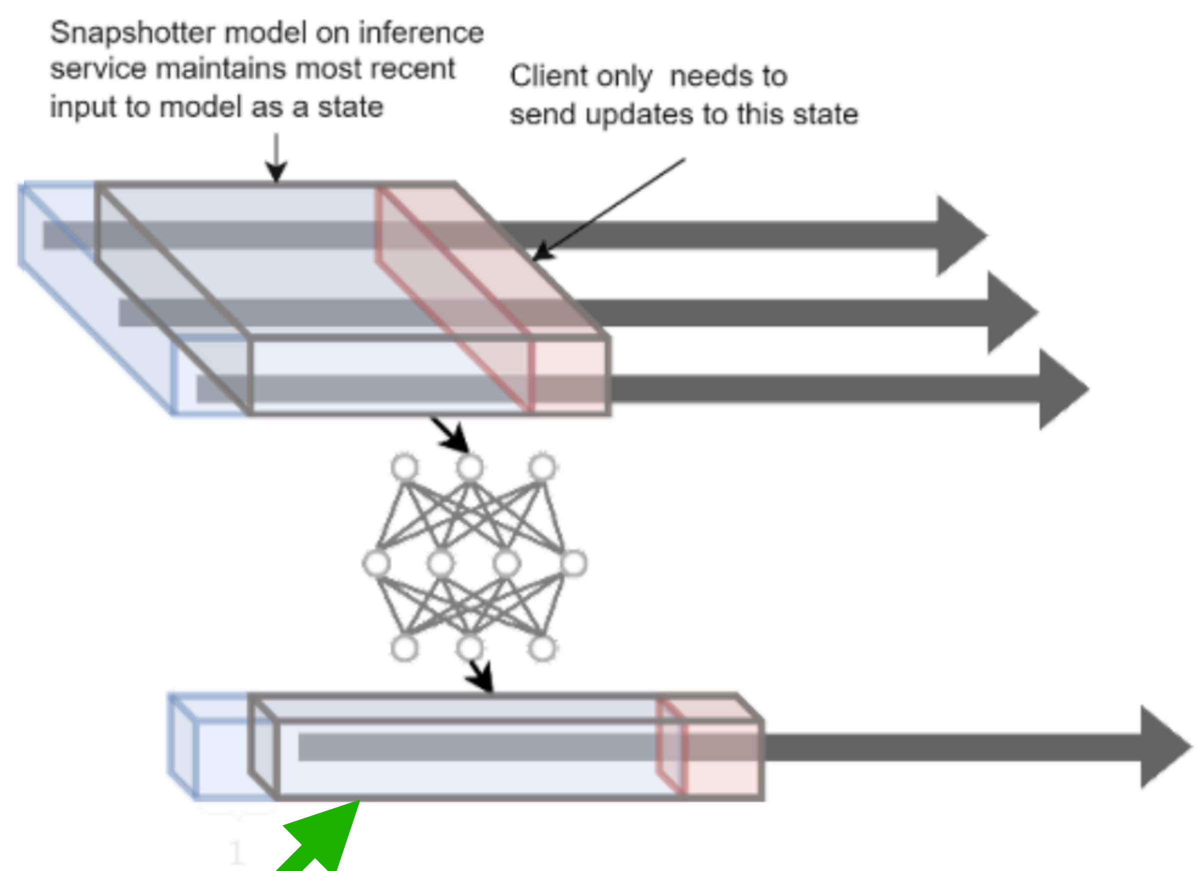
Example: Stateful Caching

Optimzied Processing of Time series data

First ML Inference on data



Second ML Algorithm for data



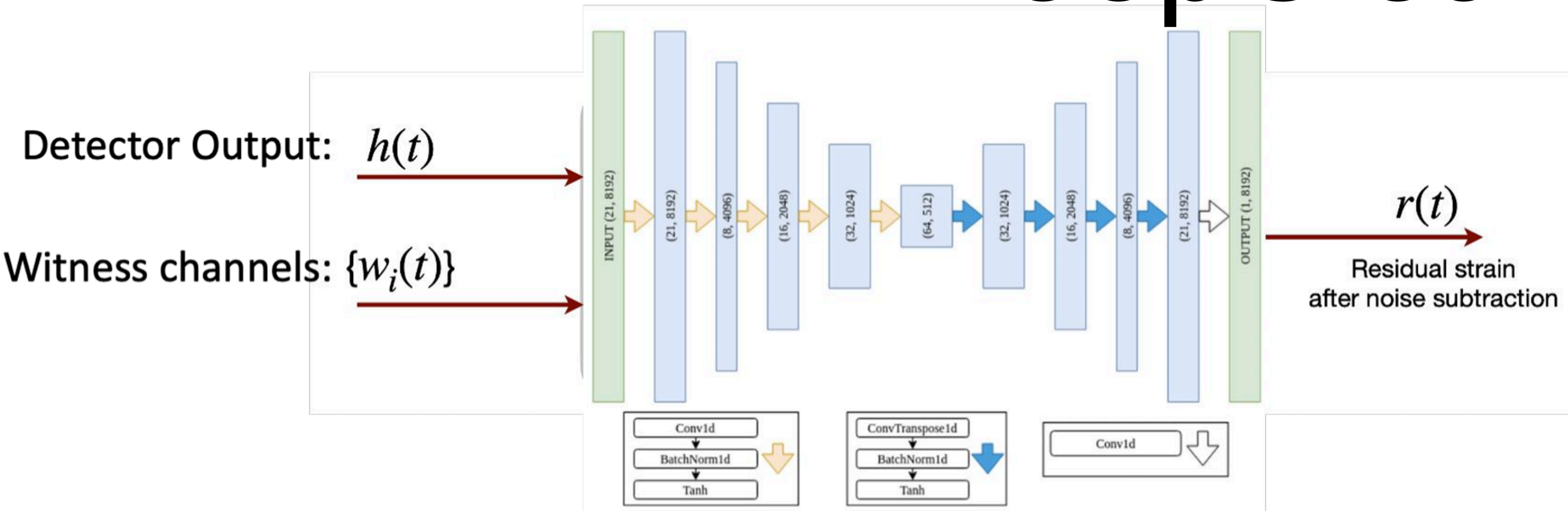
Cache on the GPU

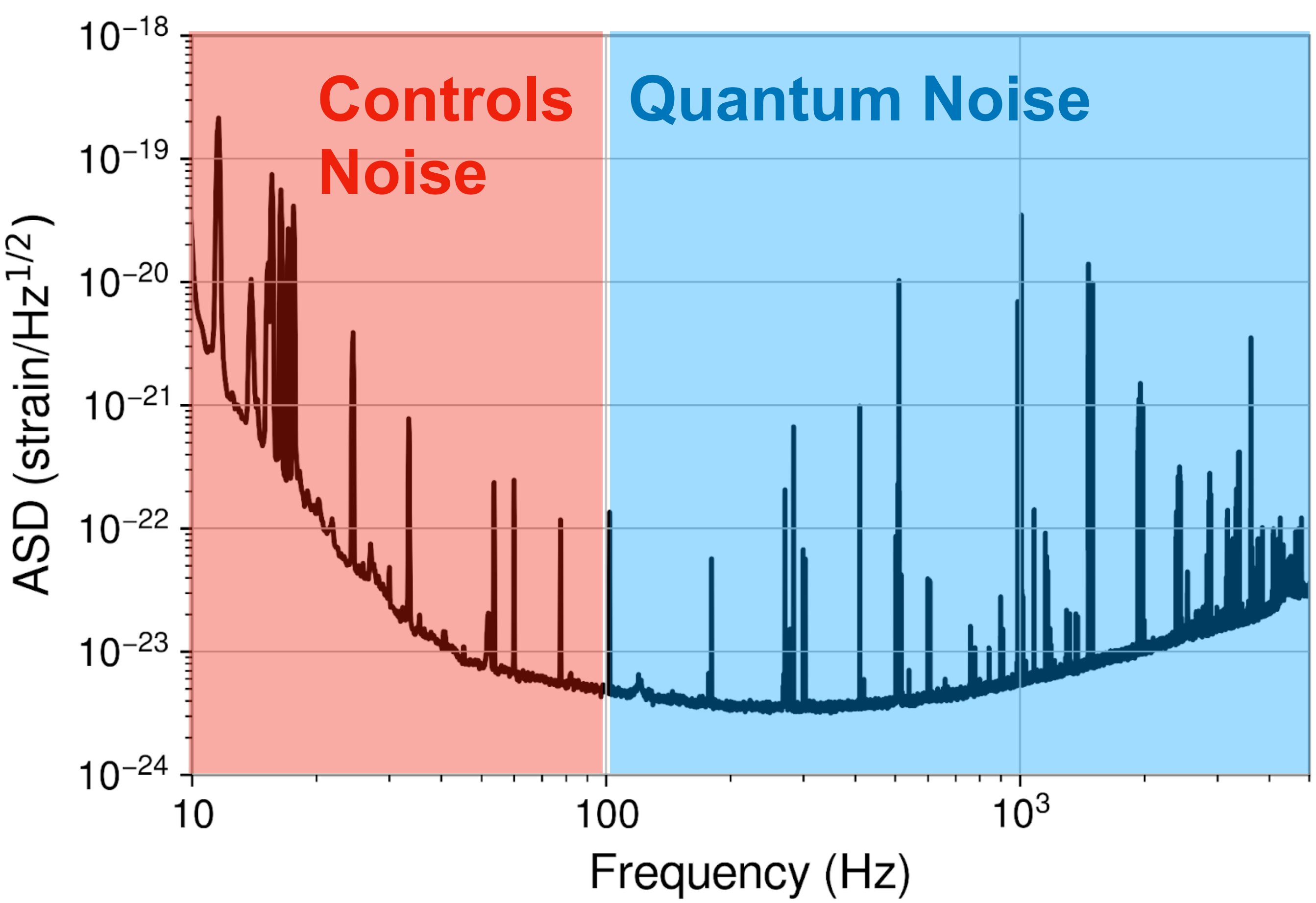
Cacheing+optimizations speeds up ML by 10x over vanilla pytorch

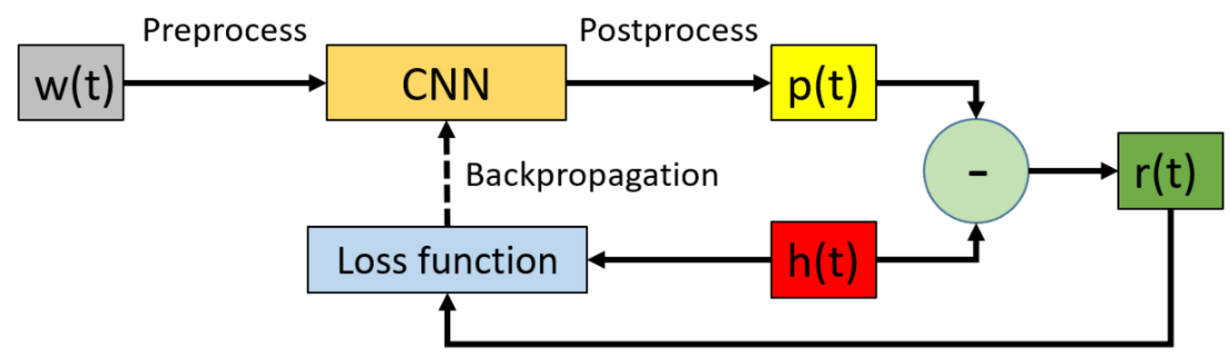
The Algos

- Step 1: Denoising
  - Deepclean: Cleaning up detector in real-time
- Step 2: Transient detection
  - A-frame: black hole merger detection in real-time
  - GWAK : unmodeled (anomalous) GW detection
- Step 3:
  - AMPLIFI: real-time parameter estimation using LFI
- Step 4: Alert

DeepClean







1000 Monitored Witness channels

100k total Witness chanels

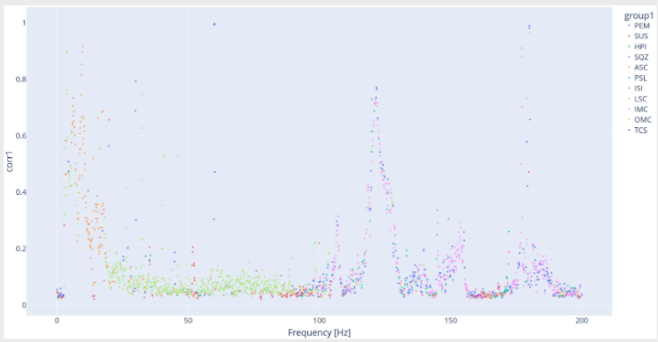
Noise reduced through cross correlation w/witness channels

Real time Monitoring

- Daily monitor noise and its correlation w/1000 channels
  - Find channels on a daily+frequency basis w/highest correltion
  - Feed them into the NN for training and decorrelation

Real time Monitoring

Coherence Monitor



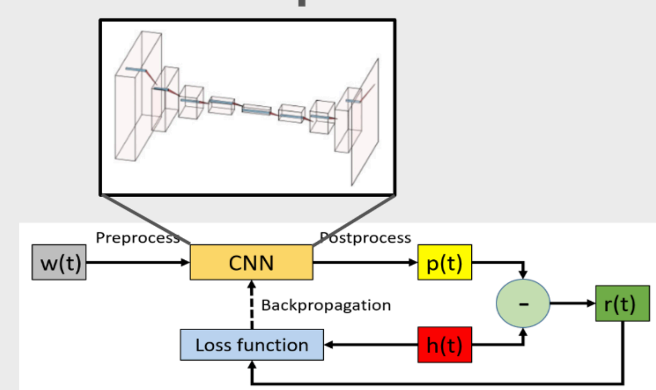
Requirements:
- 1st/2nd highest coherence
- coherence > 0.2

Witness channels

H1:ASC-CSOFT_P_OUT_DQ
H1:ASC-DHARD_Y_OUT_DQ
H1:PEM-CS_MAG_LVEA_OUTPUTOPTICS_X_DQ
H1:ASC-CHARD_Y_OUT_DQ
H1:ASC-DSOFT_P_OUT_DQ
H1:PEM-CS_MAG_EBAY_LSCRACK_X_DQ
H1:ASC-REFL_A_RF45_I_YAW_OUT_DQ
H1:LSC-POP_A_RF45_I_ERR_DQ
H1:ASC-DHARD_P_OUT_DQ
H1:SUS-ETMX_L3_OPLEV_PIT_OUT_DQ
H1:ASC-AS_A_DC_NSUM_OUT_DQ

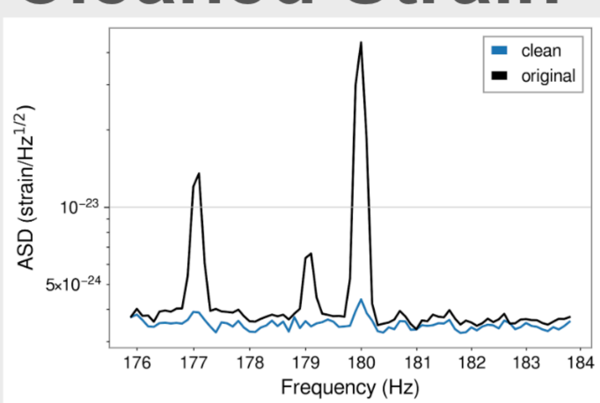
Fully automated training (hyperparameters, …)

DeepClean



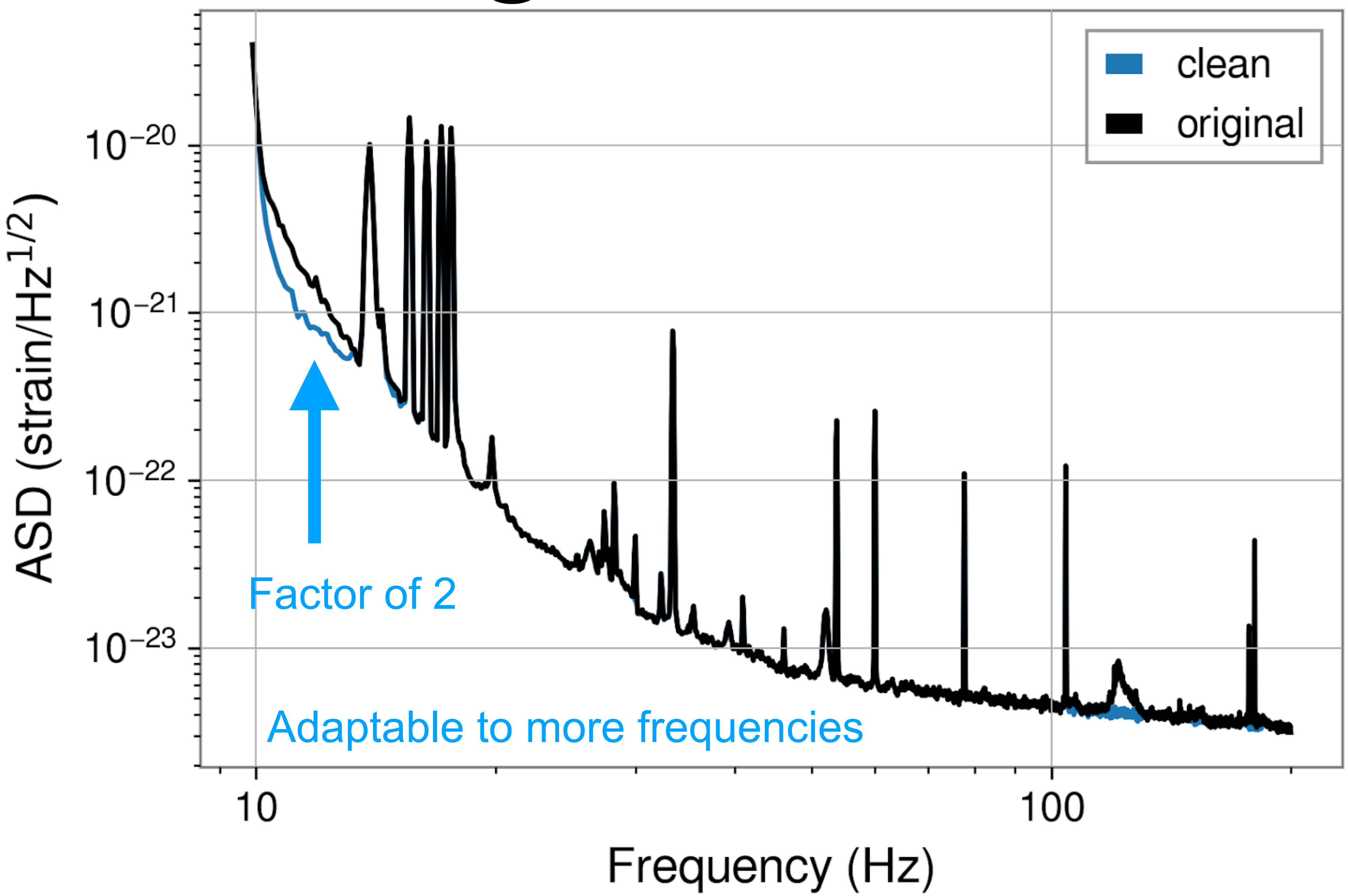
Apply cleaning on independent data

Cleaned Strain



- Aim to run end-to-end turnaround in < 1h of arrival

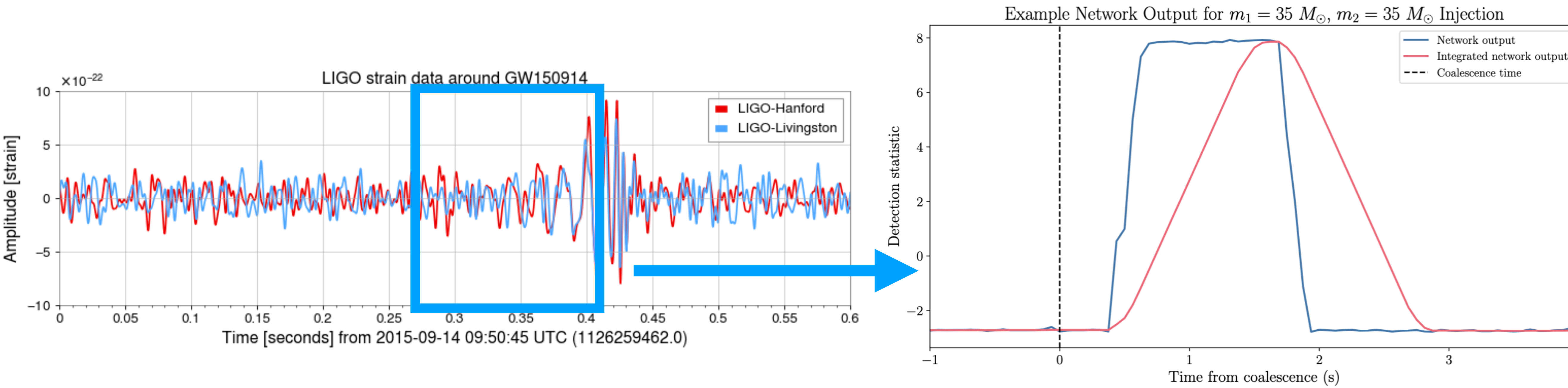
Significant Reduction



- Toolkit is adaptable for many different frequency problems
  - Denoising in frequency space is a common problem

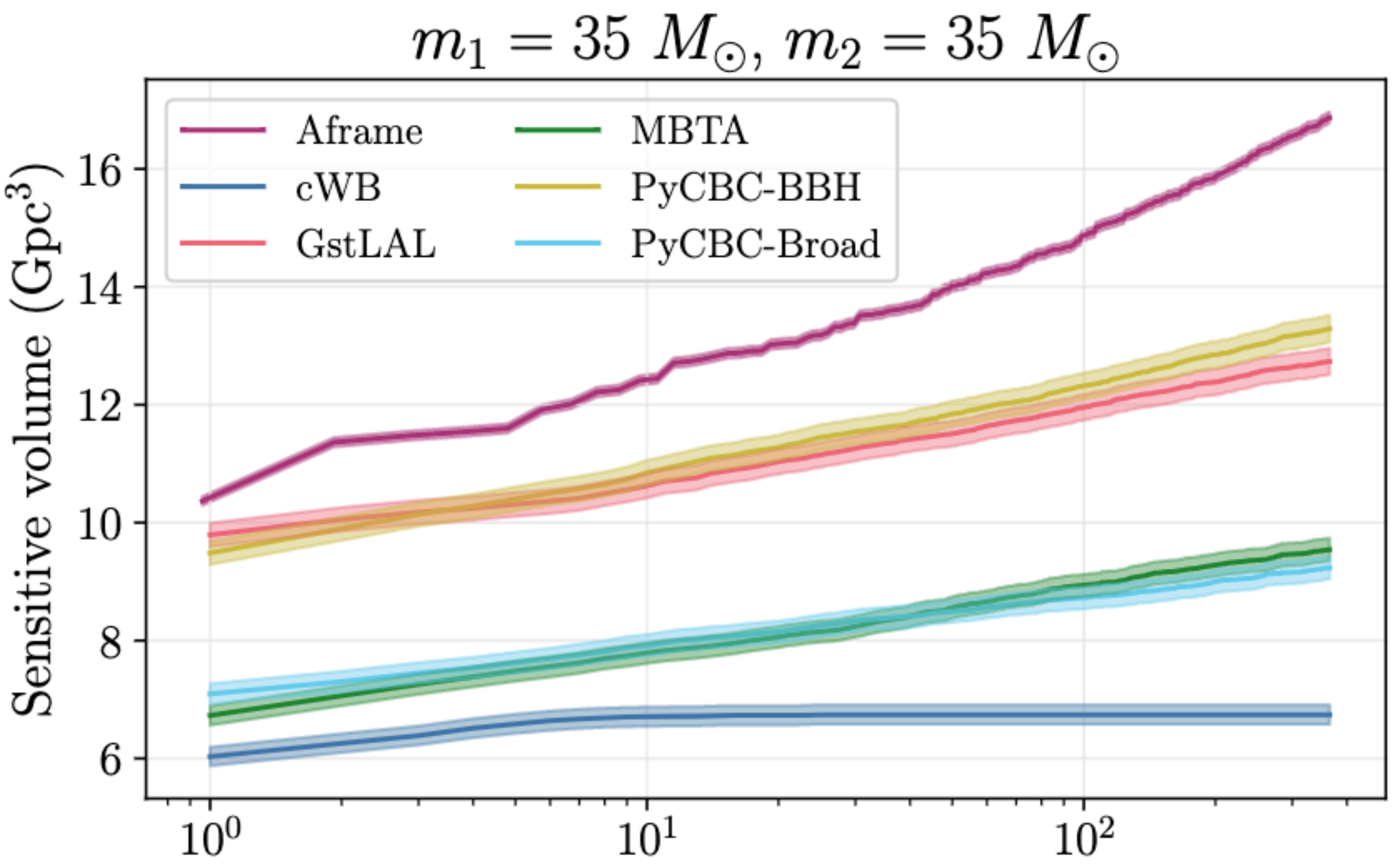
A-Frame

Sliding NN Score



- Neural network targeting Compact Binary Coalescence
  - Aka Black Hole Mergers
  - Working to adapt this to other signatures
    - Neutron star mergers
  - Curriculum learning scheme and Glitch mitigation essential
    - Sophisticated on GPU mixing tools for this

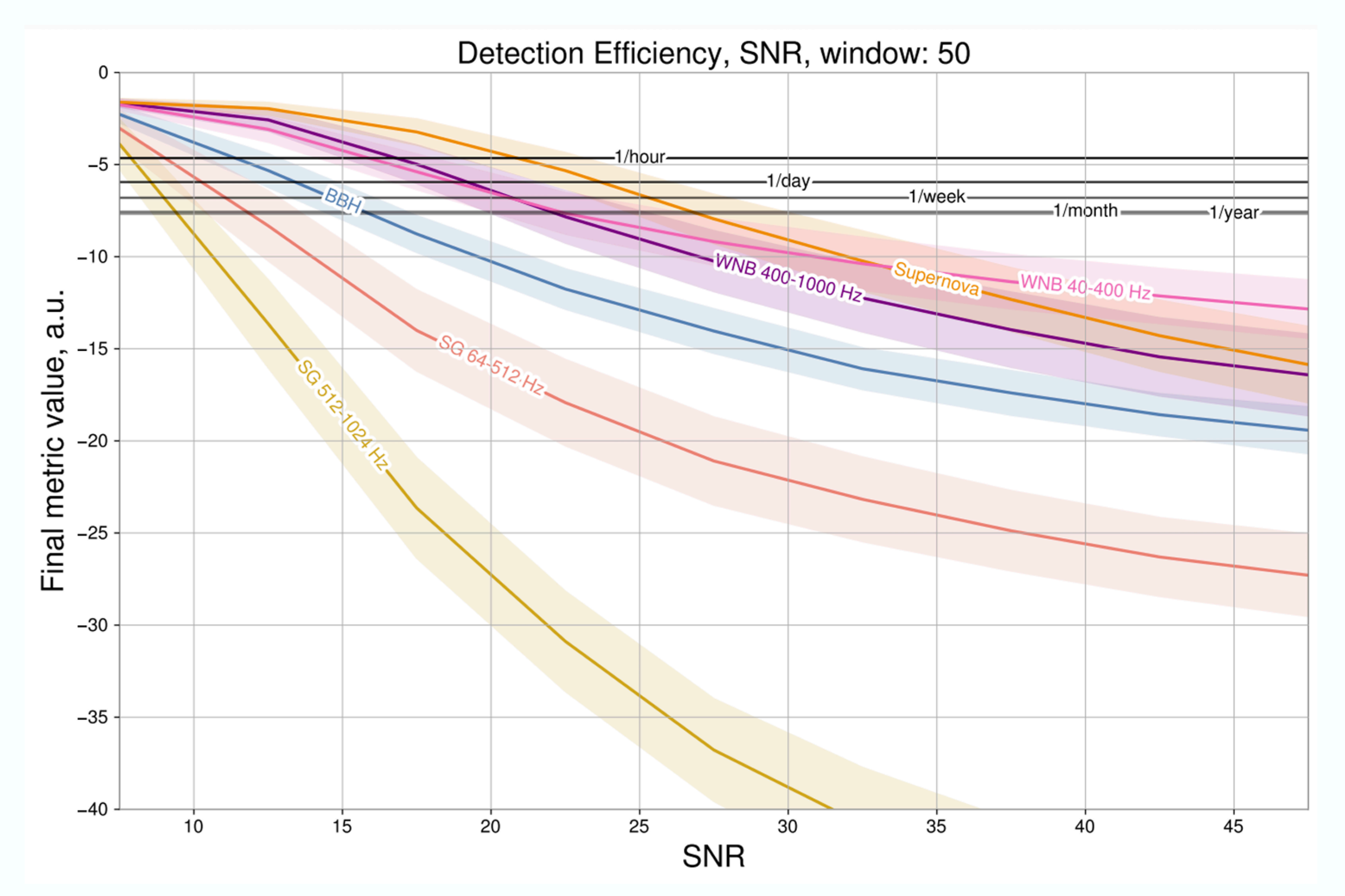
A-Frame



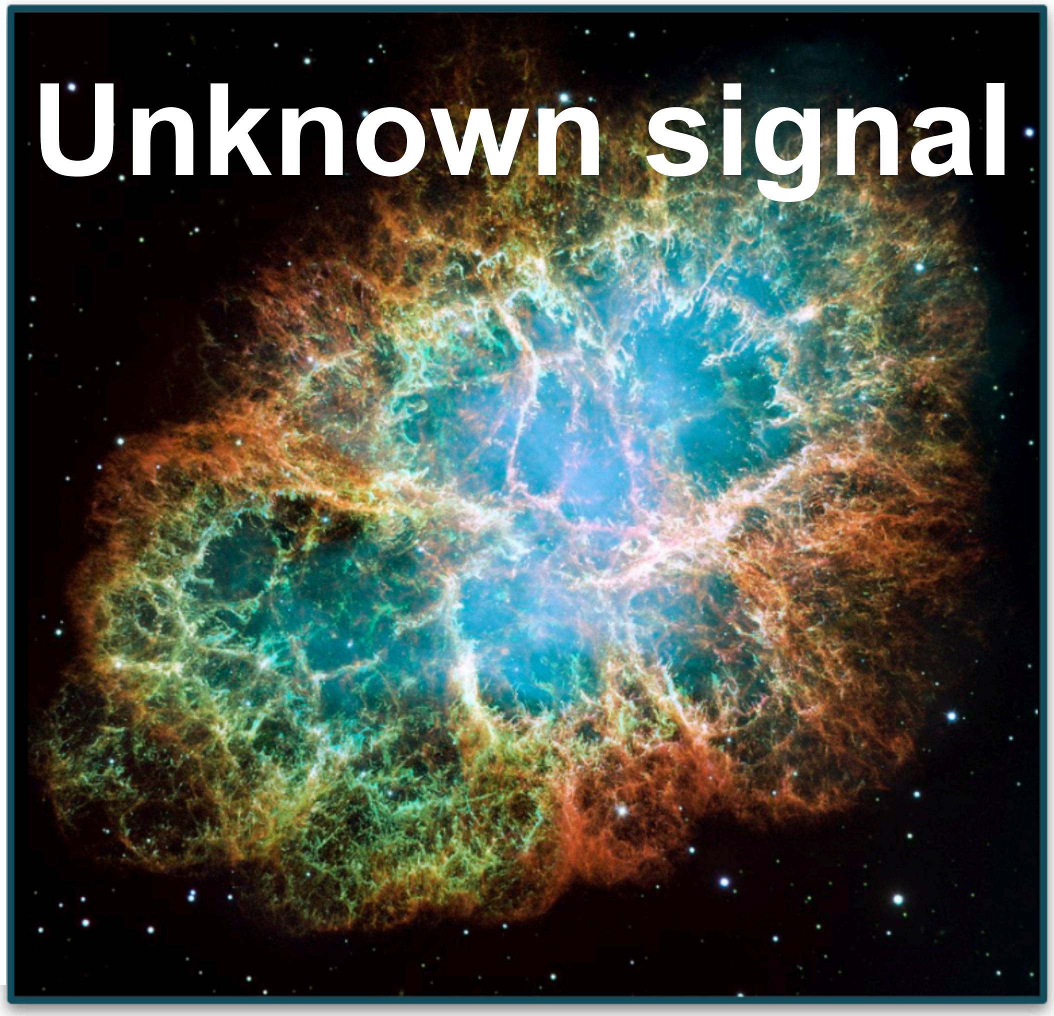
- Best sensitivity over the highest mass mergers
  - Consistent observations with running data
  - In the final stages of review to become a public merger

GWAK: Anomaly Detection

- Gravitational wave signatures can have unknown shapes
  - We can use AI based anomaly detection to identify these
  - AI based algorithm can runs on the fly (real-time)

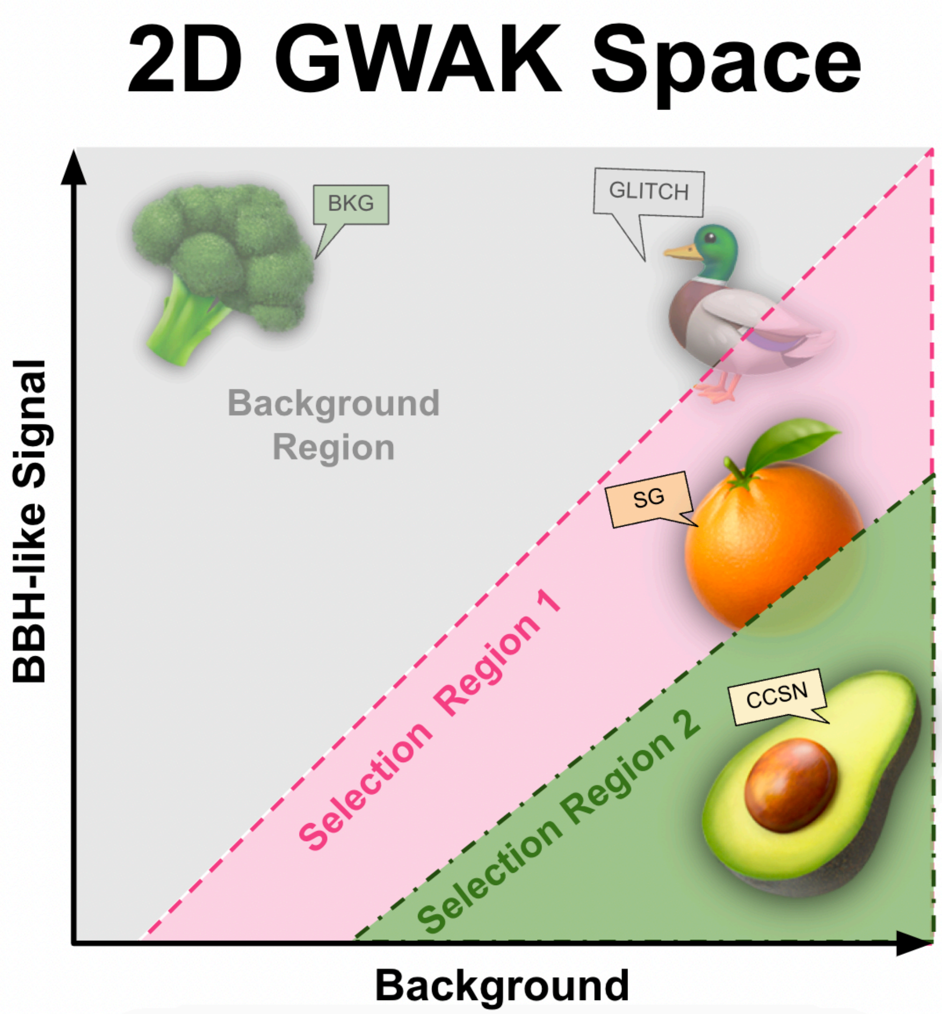


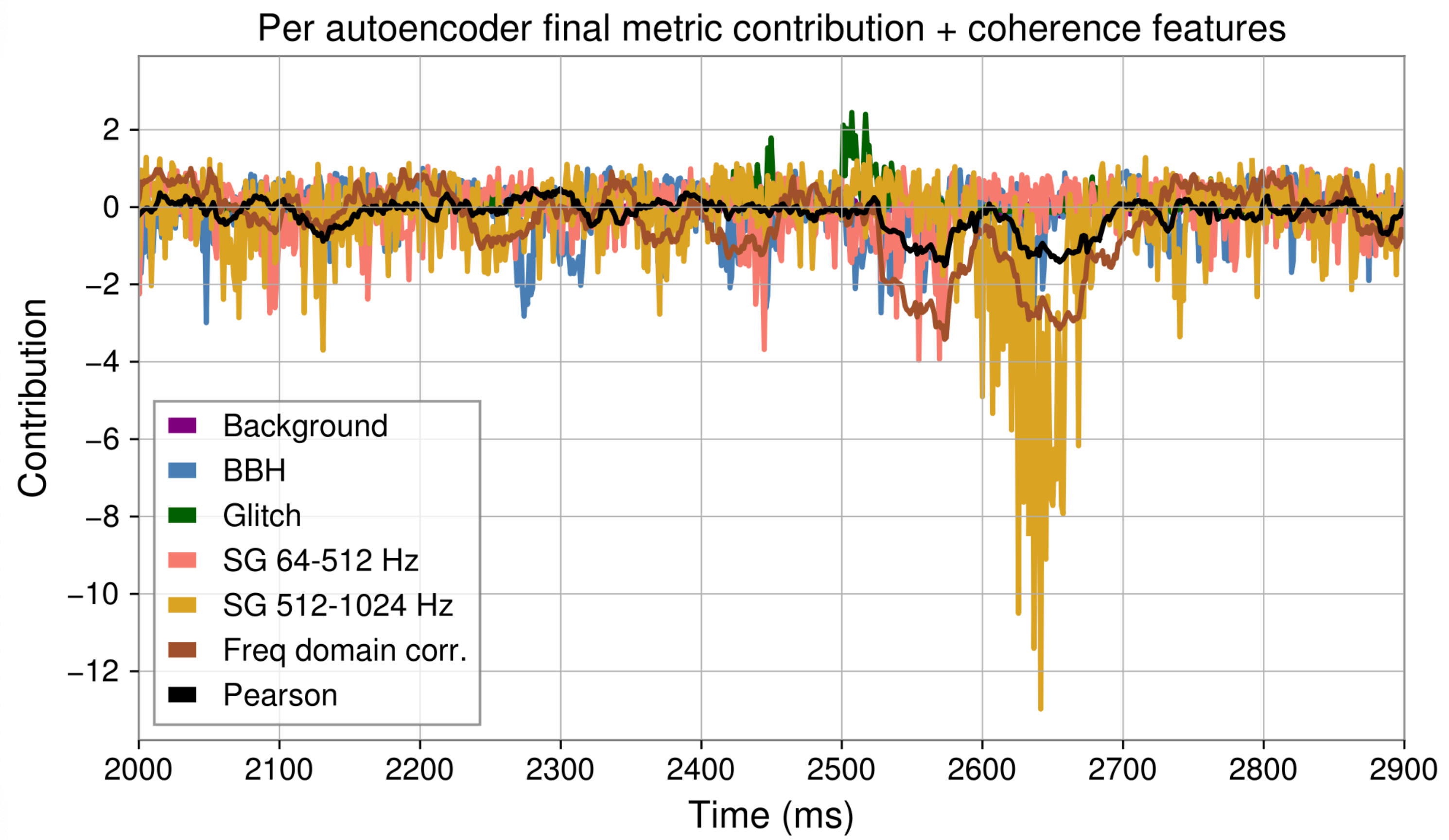
Core-collapse supernova (CCSN)



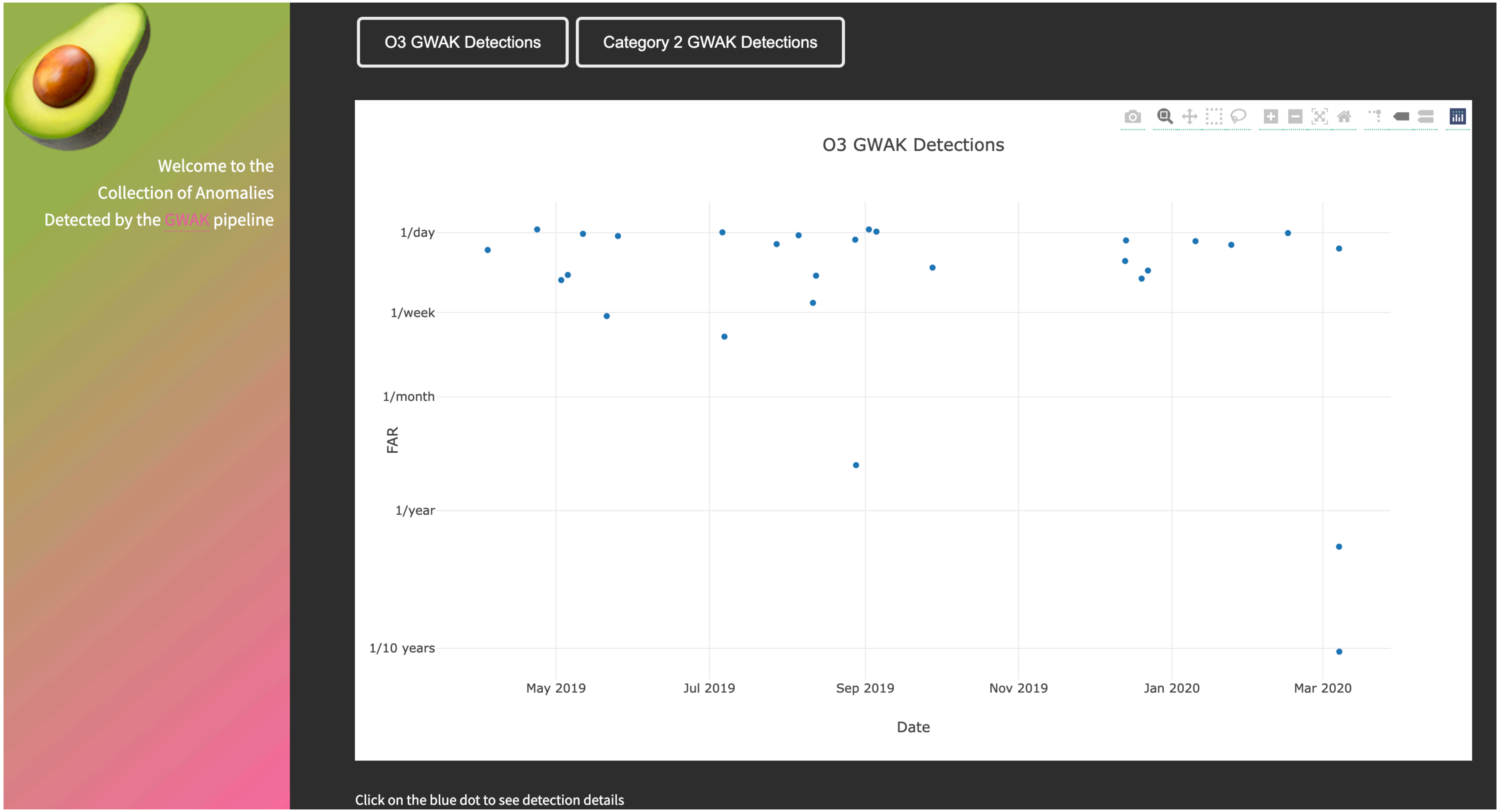
Embedded Space

- GWAK works by constructing a 11 dimensional space
  - Space presents a likelihood of a signal in a specific region
    - Likelihood on typical signals (bbh/sine gaussian)
  - Metric is constructed by a hyperplane in the space



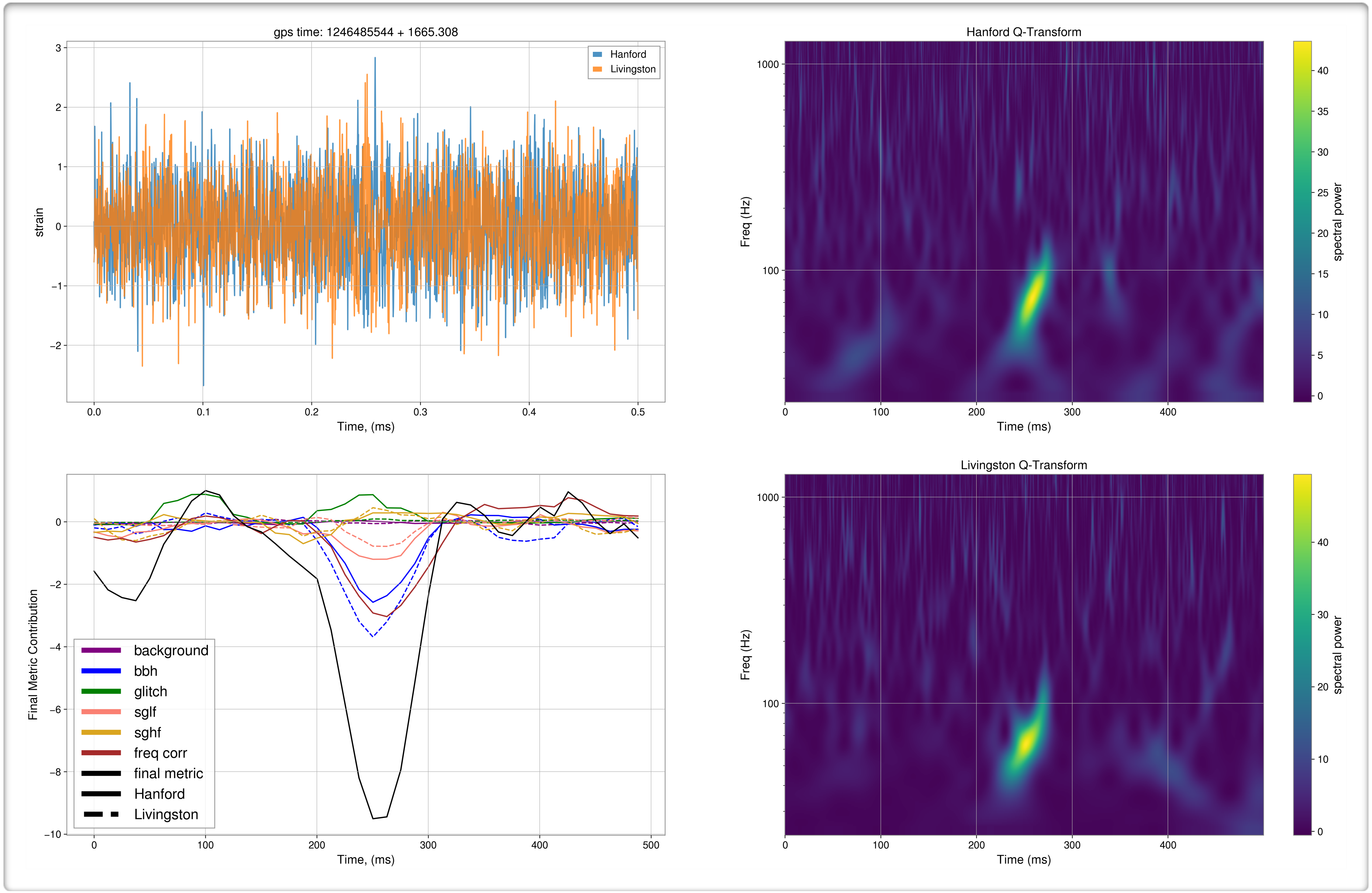


GWAK on Data



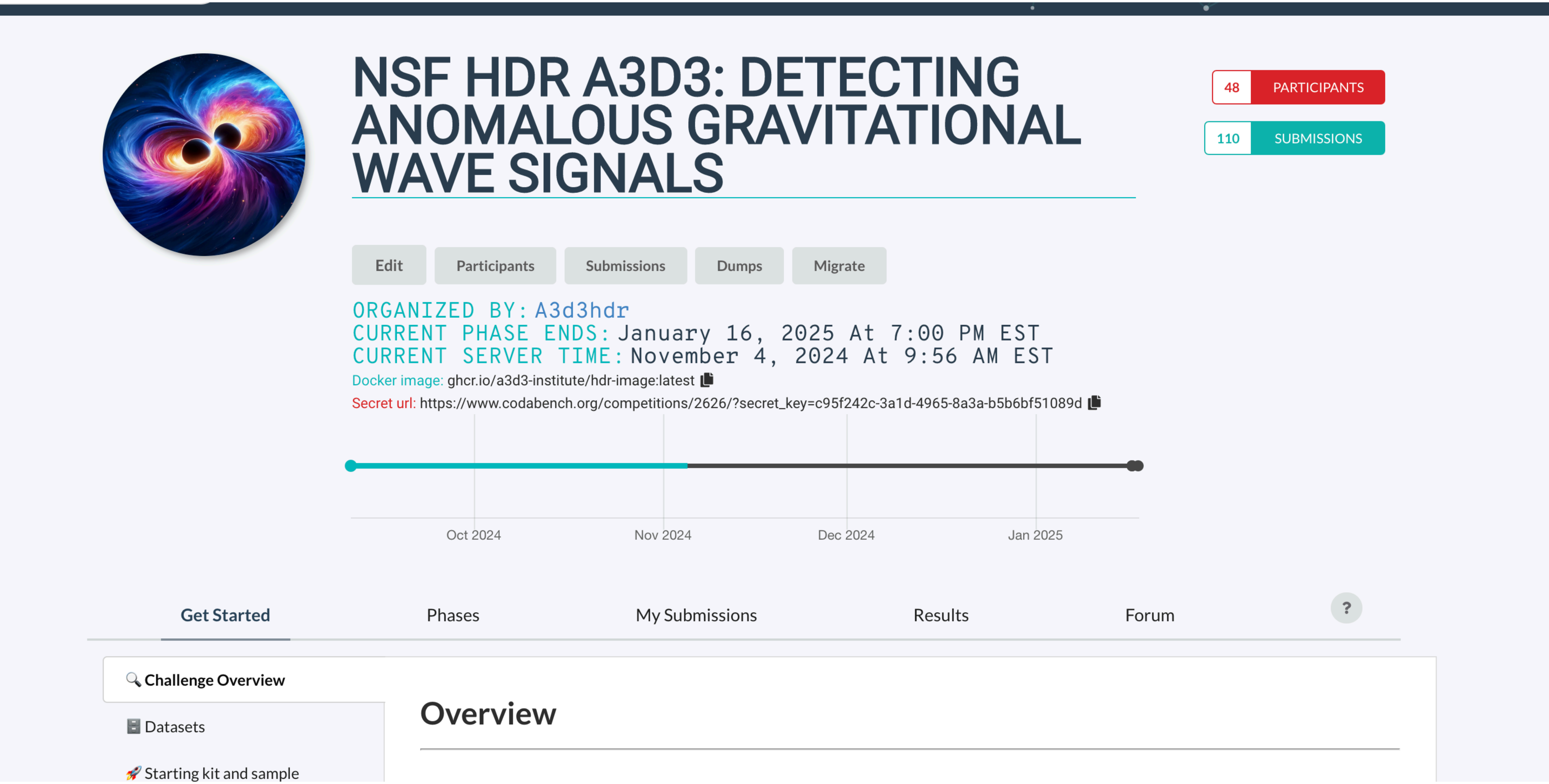
- Working to add this to real-time alerts
  - Already running internally in LIGO

GWAK on Data



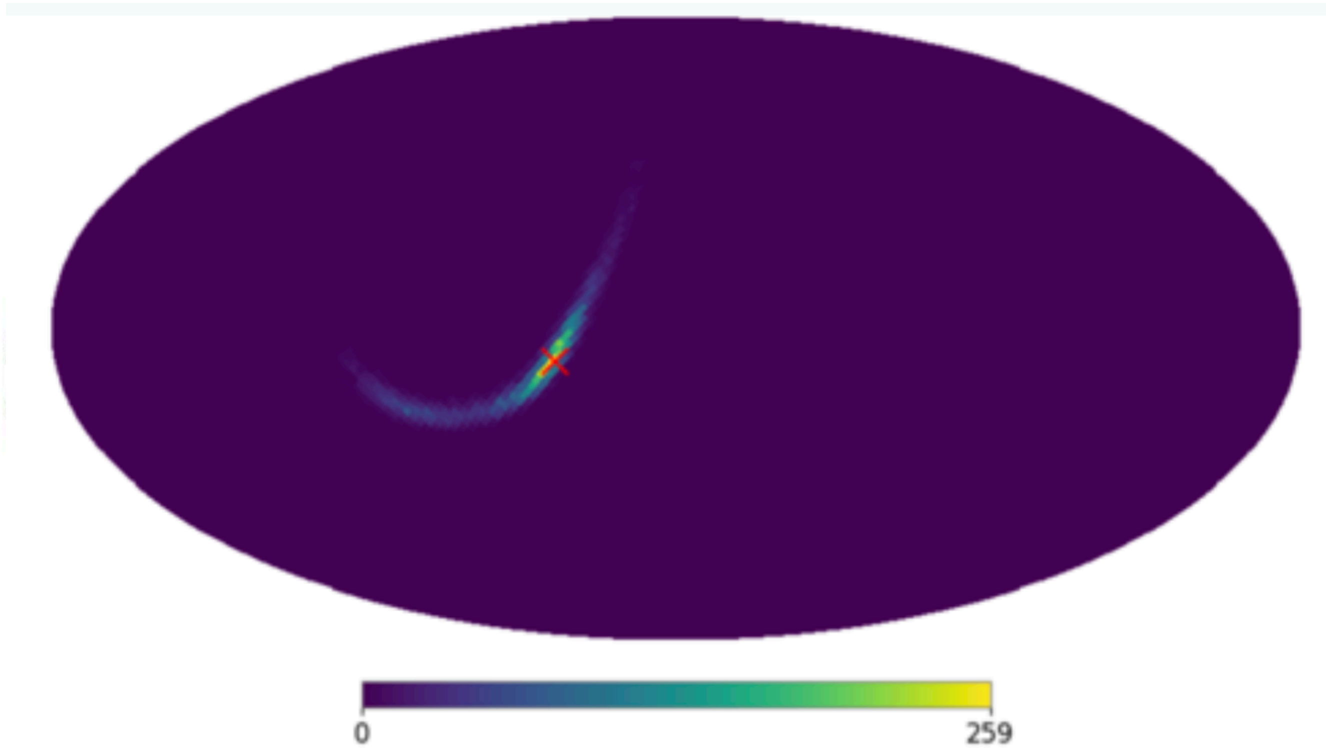
- Working to add this to real-time alerts
  - Already running internally in LIGO

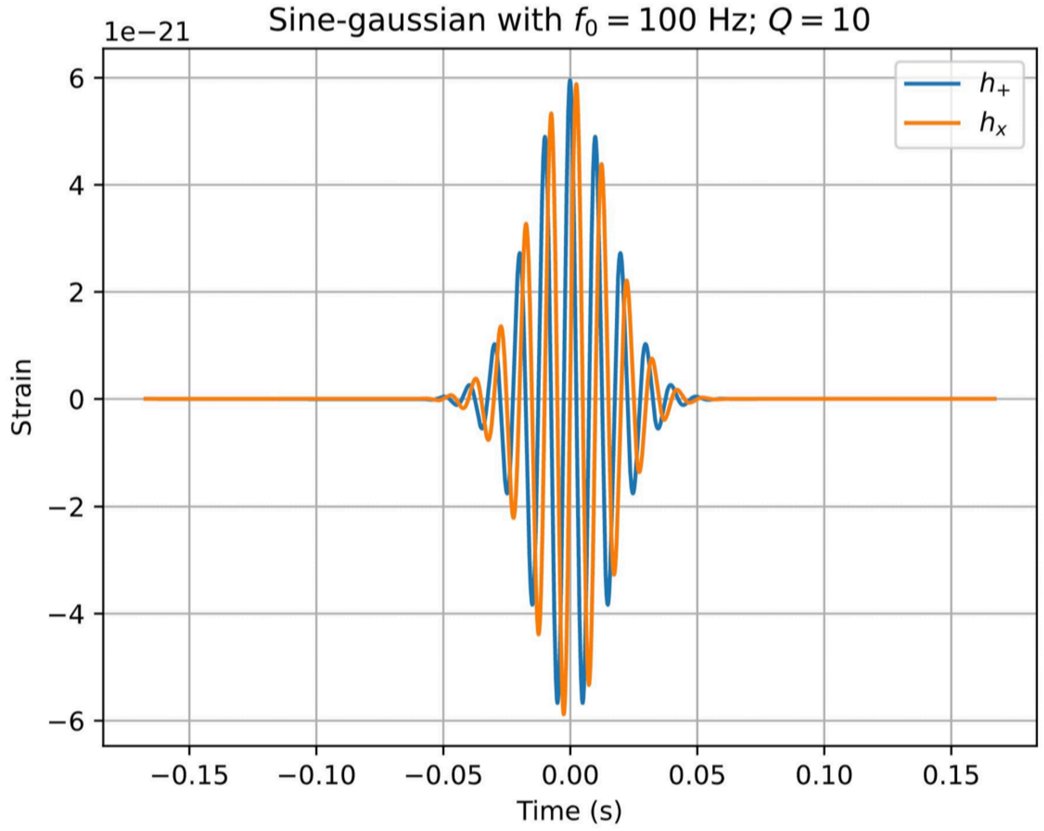
Anomaly detetion ML challenge



https://www.nsfhdr.org/mlchallenge

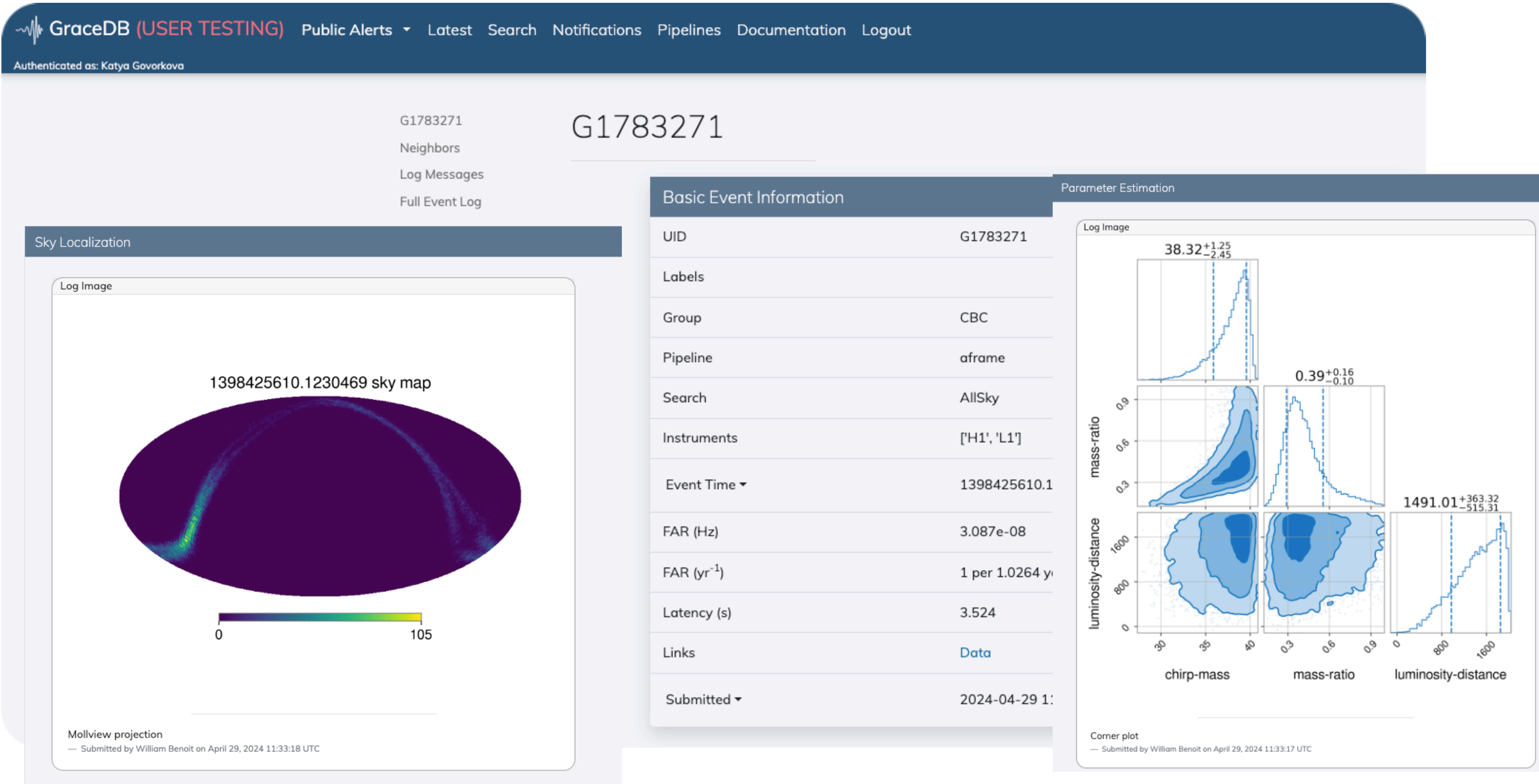
Parameter Estimation





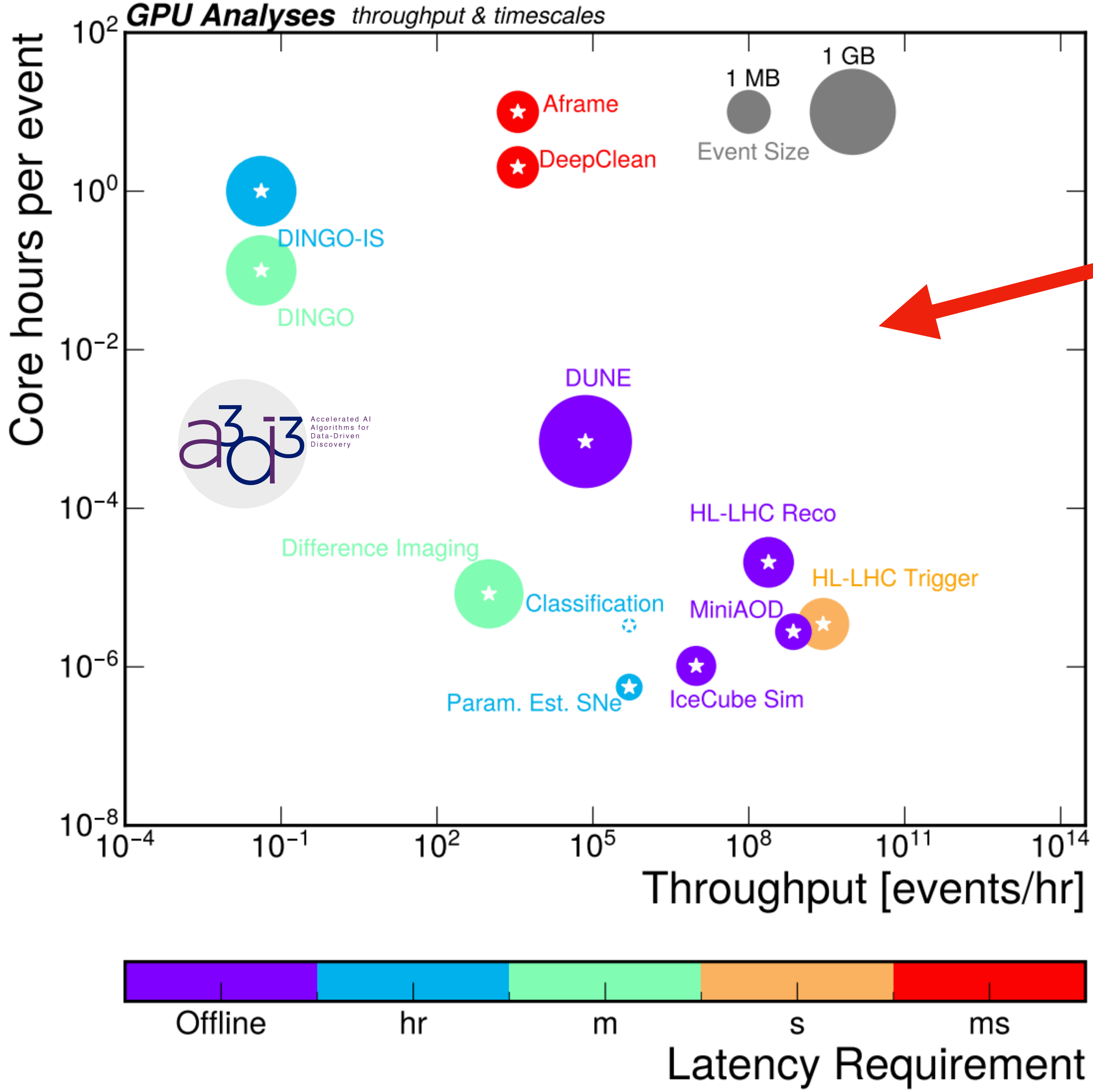
- Perform sky localization and gravitational wave parameters
- Perform fast parameter estimation using likelihood free inference
  - Normalizing flows embed broad knowledge of waveforms
    - Customized embedding to ensure compressed info
  - Parameter estimation done within seconds (or potentially less!)

Building a Pipeline



Pipeline is up and running

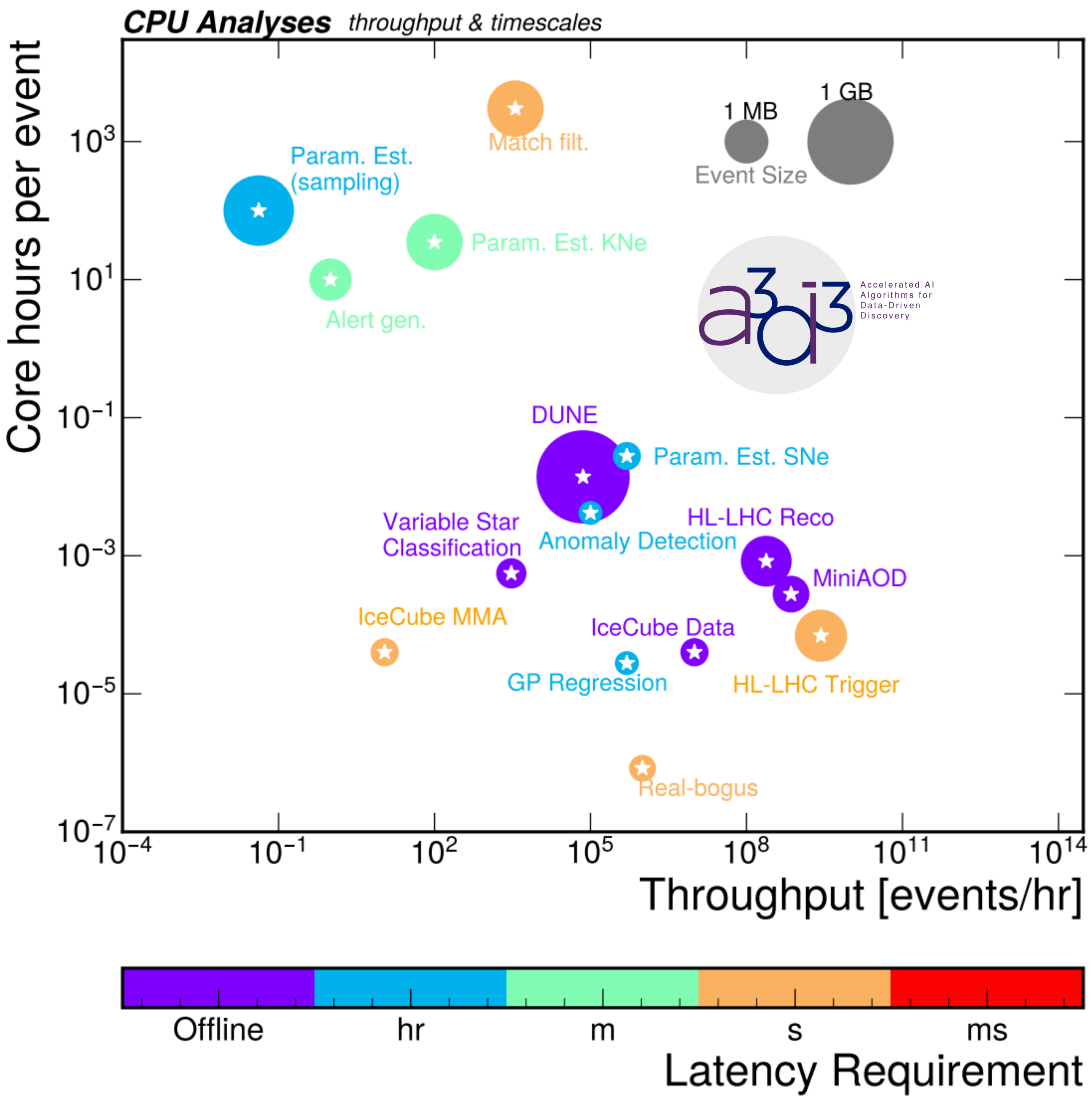
GPUs for ML

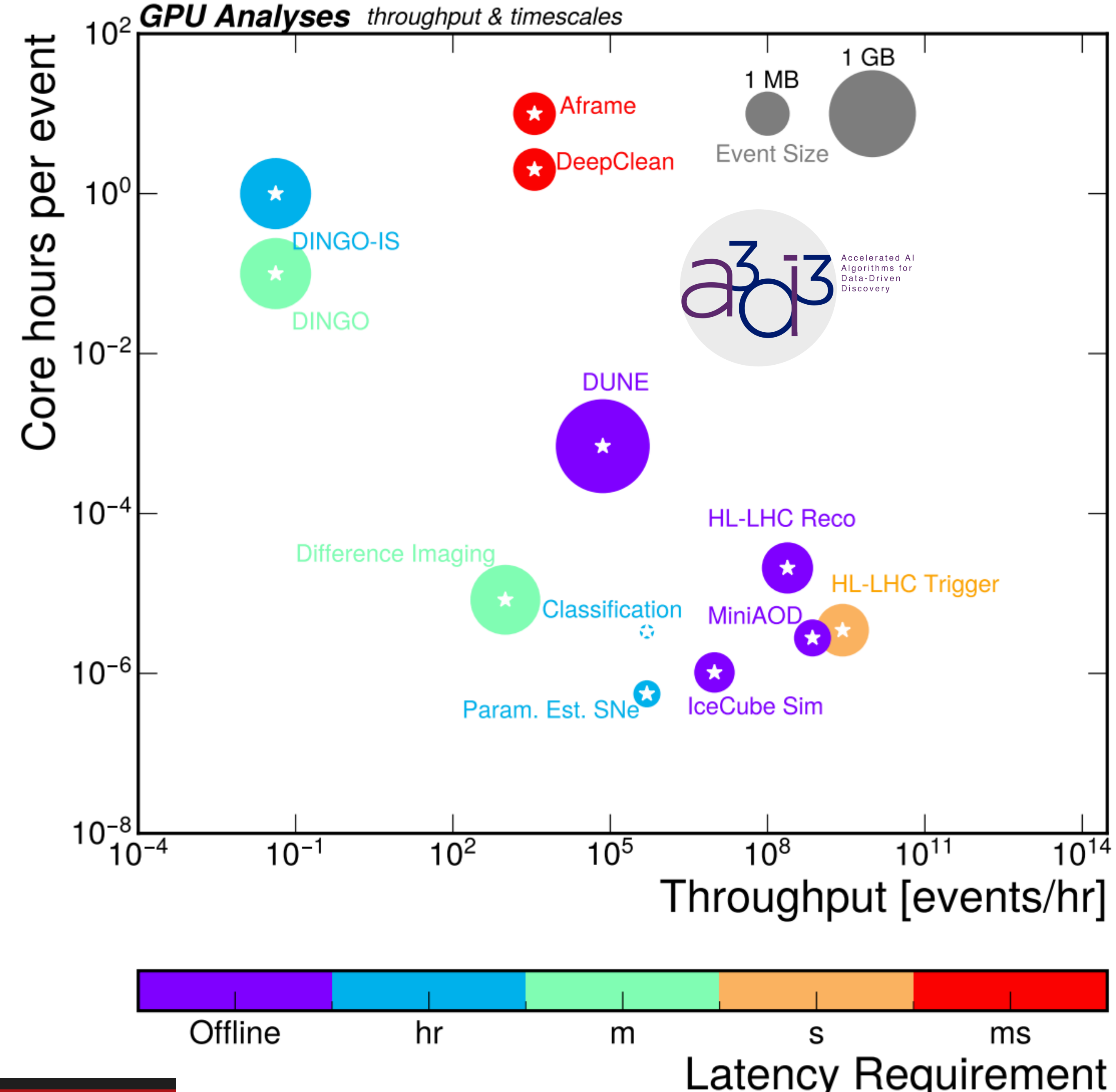


For large algorithms parallelizability and shear compute is unprecedented

Major demand emerging in many domains

Computing Demands





arXiv > hep-ex > arXiv:2306.08106

High Energy Physics – Experiment

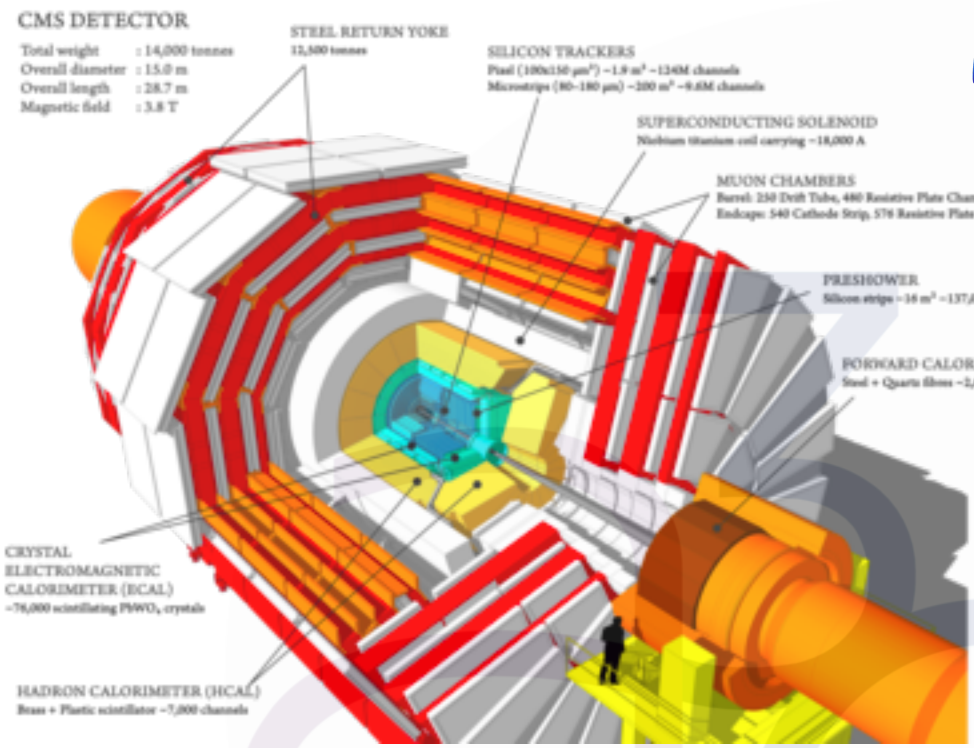
[Submitted on 13 Jun 2023]

Applications of Deep Learning to physics workflows

Arxiv: 2306.08106

Have a whitepaper outlining Inference Workflows Demands

Building an Ecosystem



DEEP UNDERGROUND NEUTRINO EXPERIMENT

DUNE infernece as-aserivce

arxiv:2301.04633

CMS Inference-as-a-service
(arxiv:2402.15366v1)

Tracking-as-a-service

LIGO inference-as-a service
arxiv:2108.12430

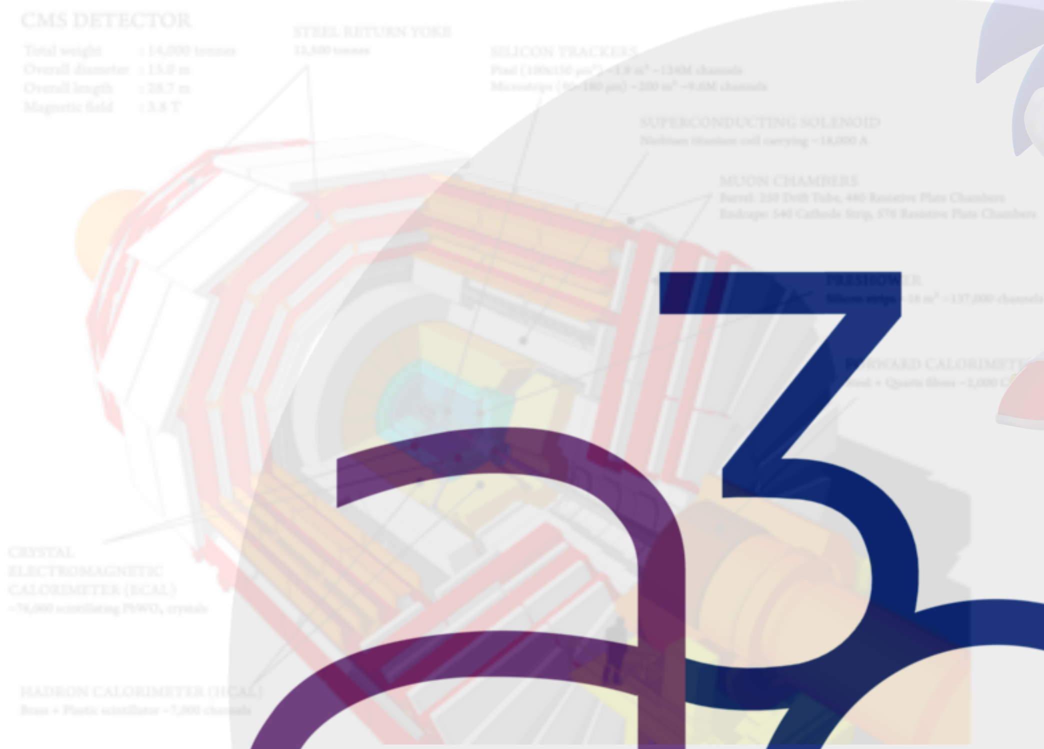
ICECUBE

Starting to Investigate

- Super-SONIC
  - Landing page of SW for Iaas for Physics

shared toolkit across many experiments

Building an Ecosystem



Accelerated AI Algorithms for Data-Driven Discovery

DUNE inference-as-a-service
arxiv:2301.04633

CMS Inference-as-a-service
(arxiv:2402.15366v1)

Tracking-as-a-service

LIGO inference-as-a service
arxiv:2108.12430

- Super-SONIC
  - Landing page of SW for Iaas for Physics

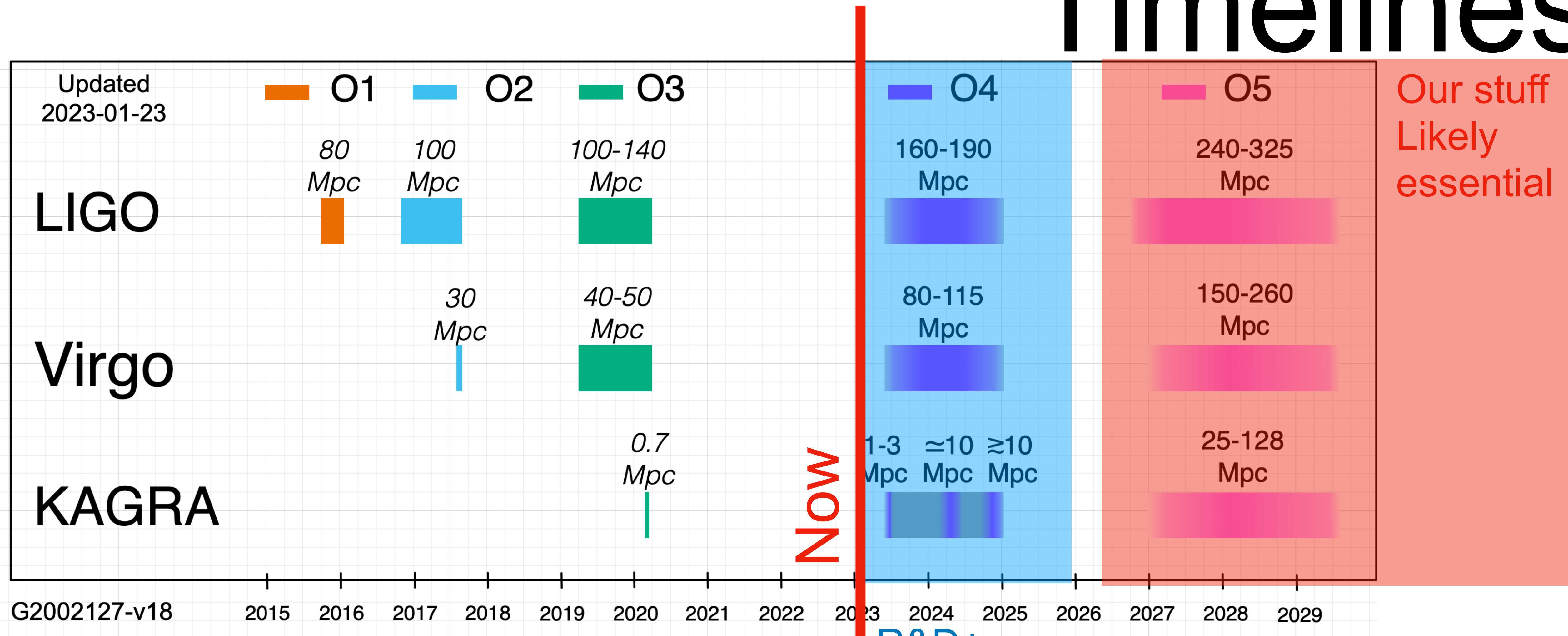
ICECUBE
Starting to Investigate

shared toolkit across many experiments

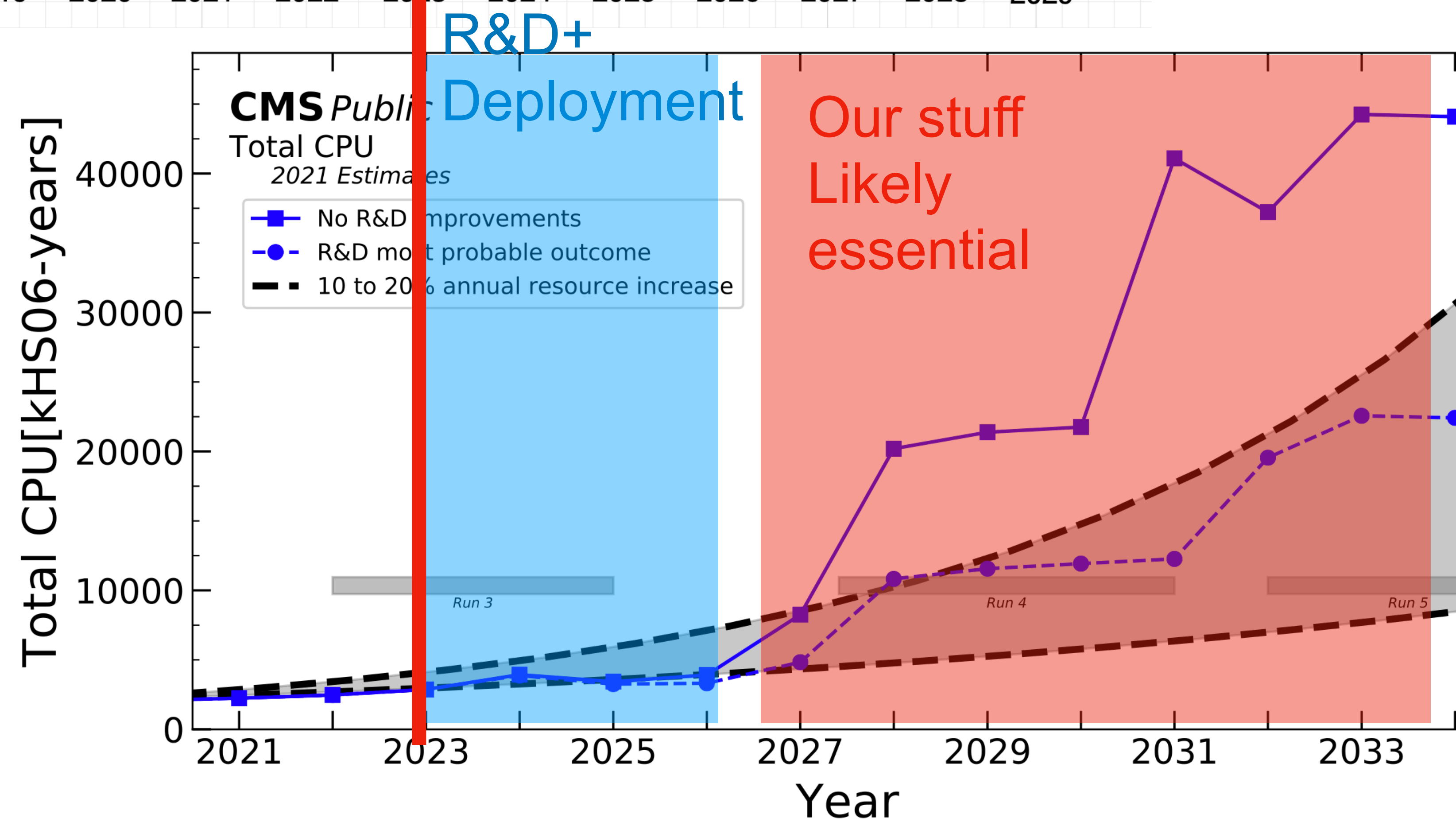
What is Critical?

- 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: Apptainer/Kubernetes/…
  - **ML Problems**
    - Awareness of the diversity of problems is critical (Not just LLM)
    - Highlighting the similarity across scientific domains is critical

Timelines



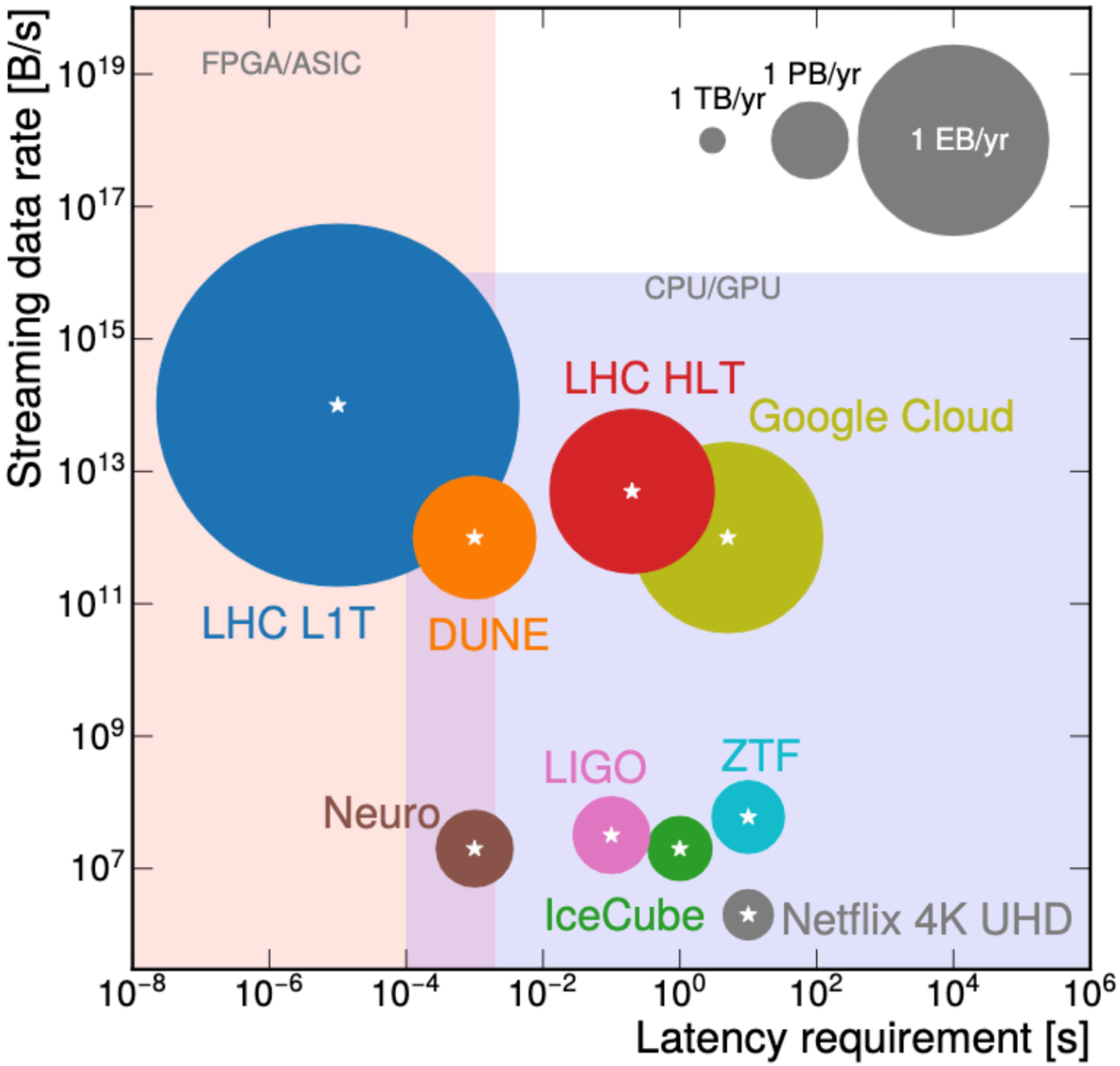
DUNE timeline and various astro timelines (Rubin/LSST)
Should also figure in our overall schedule



Recap

- We are building a real-time GW pipeline using ML
  - Latency and throughput are critical element in the design
  - Effective integration of heterogeneous compute is critical
- Our pipelines are up and running internal and soon externally
  - We see this as a path to be a main pipeline for future (O4) running
  - We encourage many others to build on our toolkit
- While our focus is on GW the tools here are broadly applicable
  - SBI toolkit has already been adapted for Kilonovae
  - In discussions for adapting parts of the toolkit for other time series
- Looking to expand the scope of our work under ML4GW toolkit
  - More collaborations within GW and byeond

Thanks!



Accelerated AI Algorithms for Data-Driven Discovery

LIGO Data Workflow

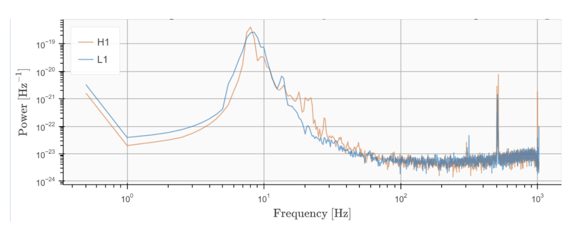
Data

8192 Hz/detector

Calibration

100000 channels
8192 Hz/channel

Detector Characterization



Use Witness Channels to denoise

Cleaned Data

Challenge is to run this in real-time

Roughly 1 PB per year

LIGO Data Workflow

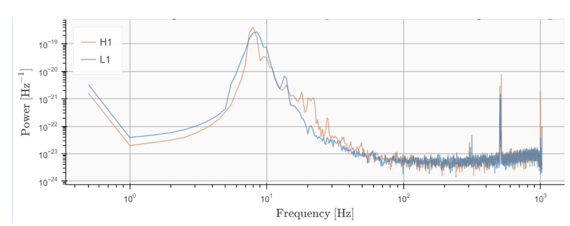
Data

8192 Hz/detector

Calibration

100000 channels
8192 Hz/channel

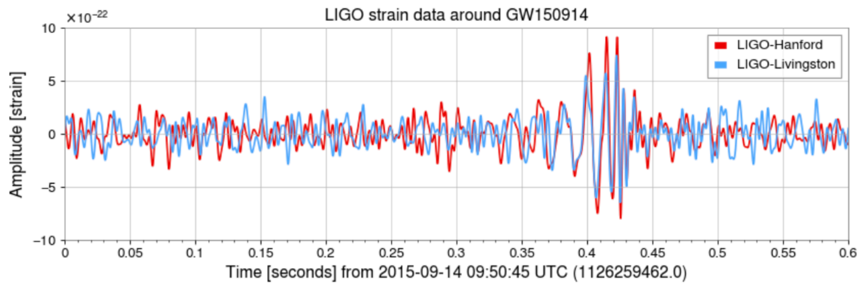
Detector Characterization



Use Witness Channels to denoise

Cleaned Data

Event Detection



Black Holes/Neutron Stars/???

Event

Challenge is to run this in real-time

Roughly 1 PB per year

LIGO Data Workflow

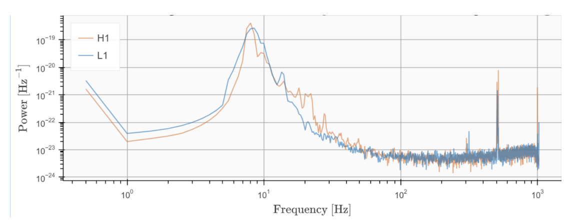
Data

8192 Hz/detector

Calibration

100000 channels
8192 Hz/channel

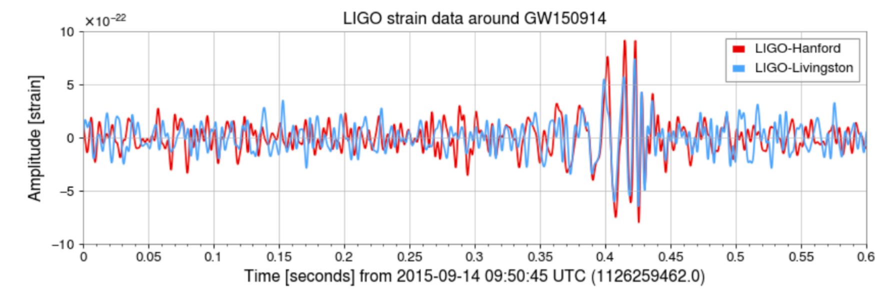
Detector Characterization



Use Witness Channels to denoise

Cleaned Data

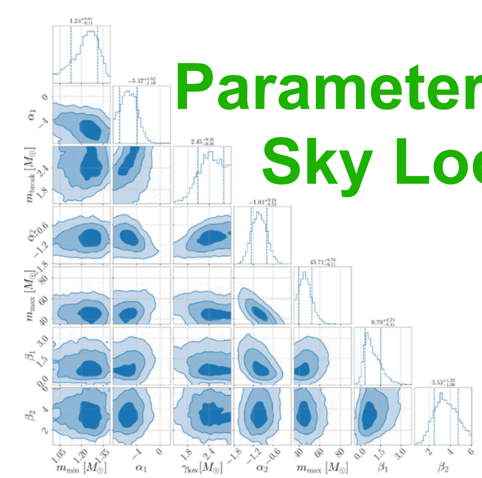
Event Detection



Black Holes/Neutron Stars/???

Event

Event Characterization



Parameter Estimation/ Sky Localization

Alert

Challenge is to run this in real-time

Roughly 1 PB per year

Current Workflow

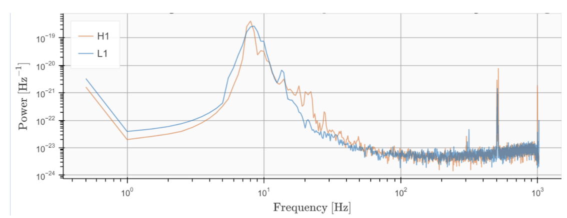
Data

8192 Hz/detector

Calibration

100000 channels
8192 Hz/channel

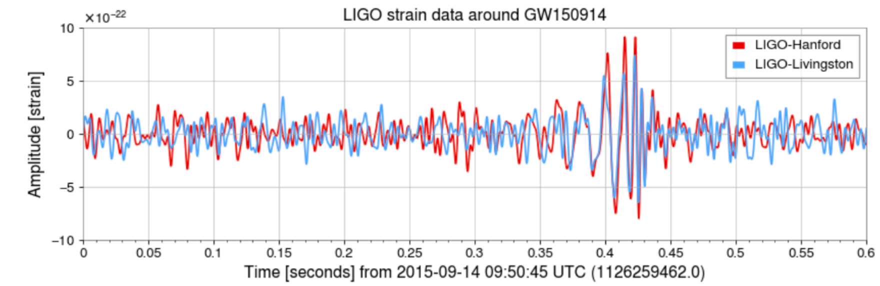
Detector Characterization



Correlation
On previous data

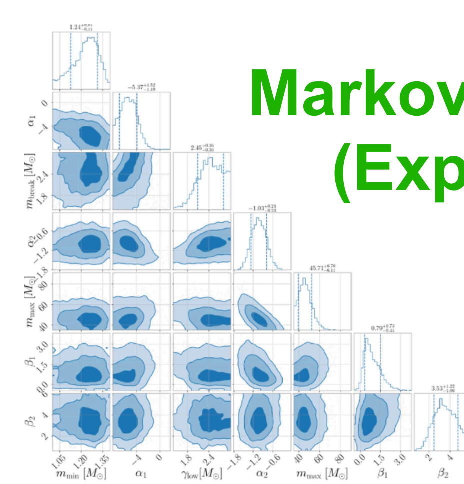
Cleaned
Data

Event Detection



Fast Template Matching

Event Characterization



Markov chain MC
(Expensive)

Current workflow
Uses (CPU) LIGO data grid
With rule based algorithms

Our Upgraded Workflow

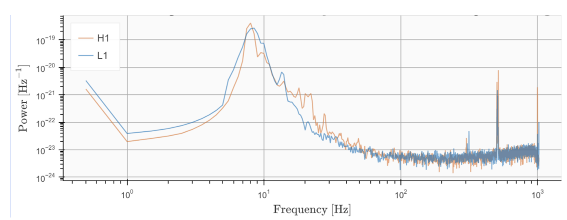
Data

8192 Hz/detector

Calibration

100000 channels
8192 Hz/channel

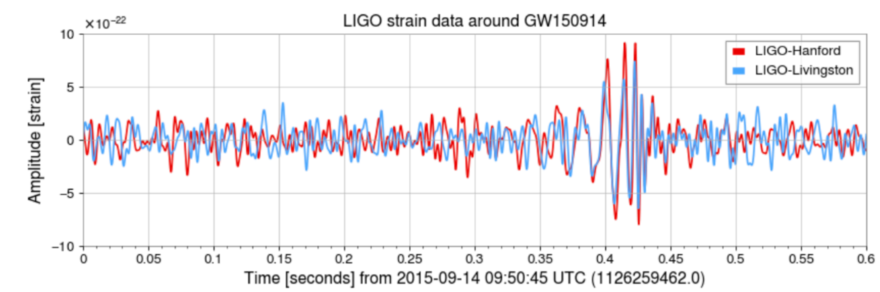
Detector Characterization



DeepClean
NN based AE

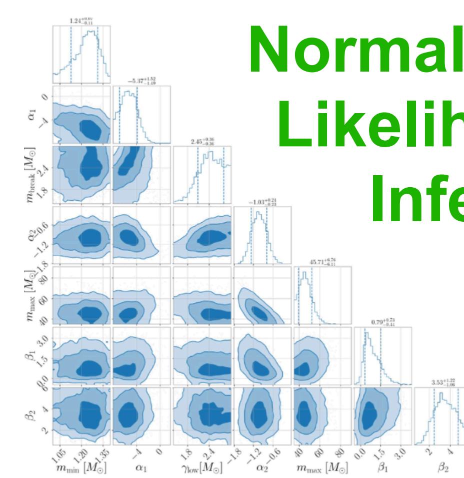
Cleaned Data

Event Detection



NN-based Algos
A-Frame/Neutron Star/GWAK

Event Characterization



Normalizing Flow
Likelihood Free Inference

ML Based Workflow
Almost 100% GPU based
Only a handful of GPUs for real-time

Computing Resources

LIGO Data Grid

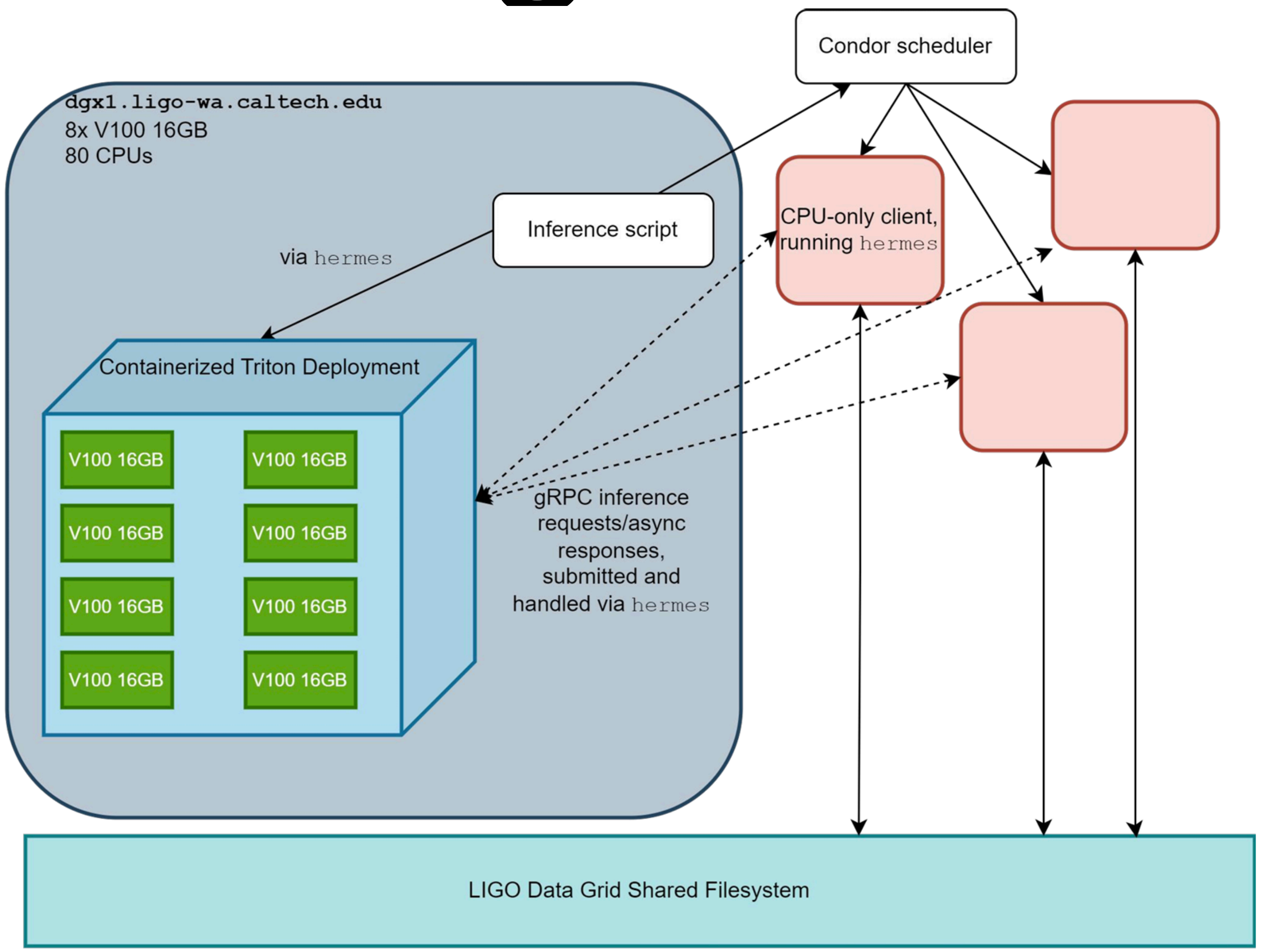
- LIGOs computing ecosystem of mostly CPU resources
- Limited GPUs, workloads not scalable
- GPUs resources are not sufficient to support large ML workflows

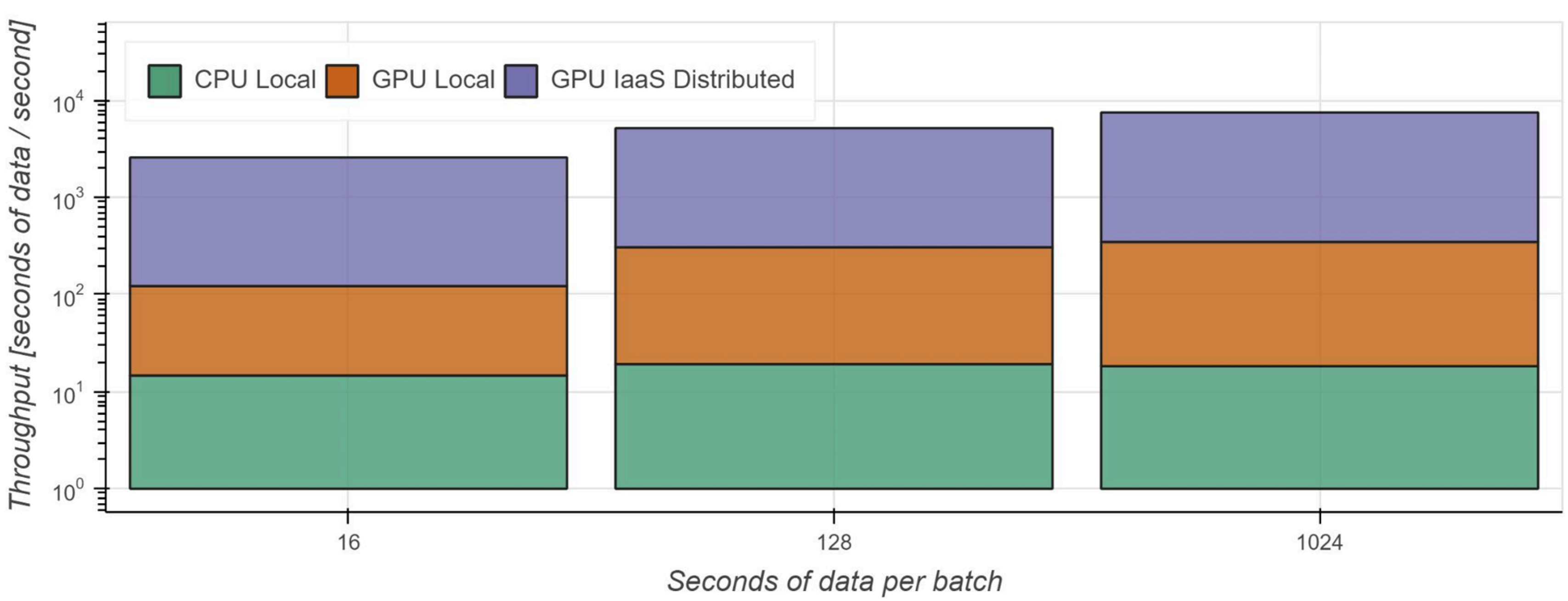
Nautilus HyperCluster

- Collection of computing clusters containing 1000s of GPUs
- Containerized workloads
- Easily scalable with Kubernetes
- With Kubernetes infra, can easily migrate to other cloud resources

NRPs GPU resources makes it possible for us to scale to full analyses

Large Scale deployment





Throughput of 3800 s'/s achieved

100 years of data in 2 weeks!

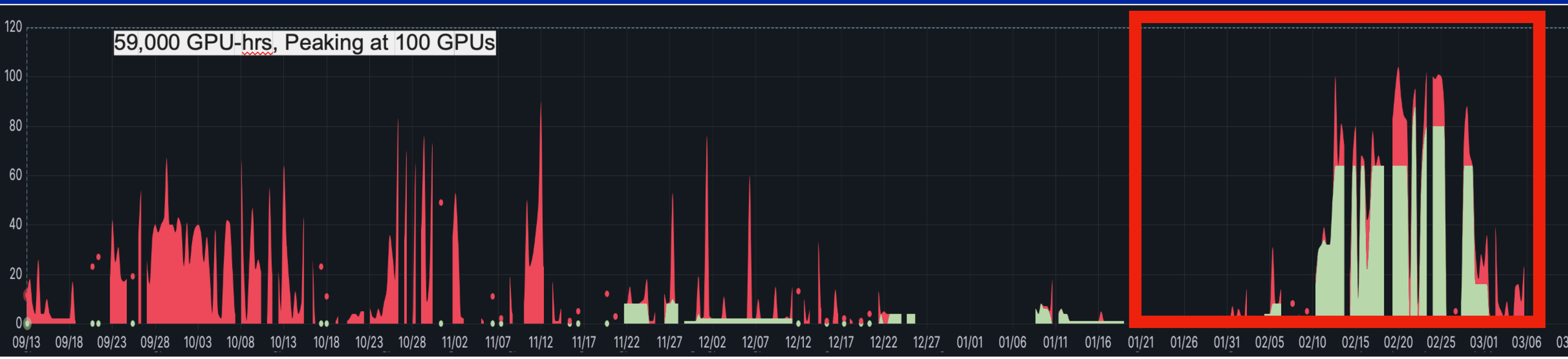
10x speed up from conventional GPU + More possible (FP16)

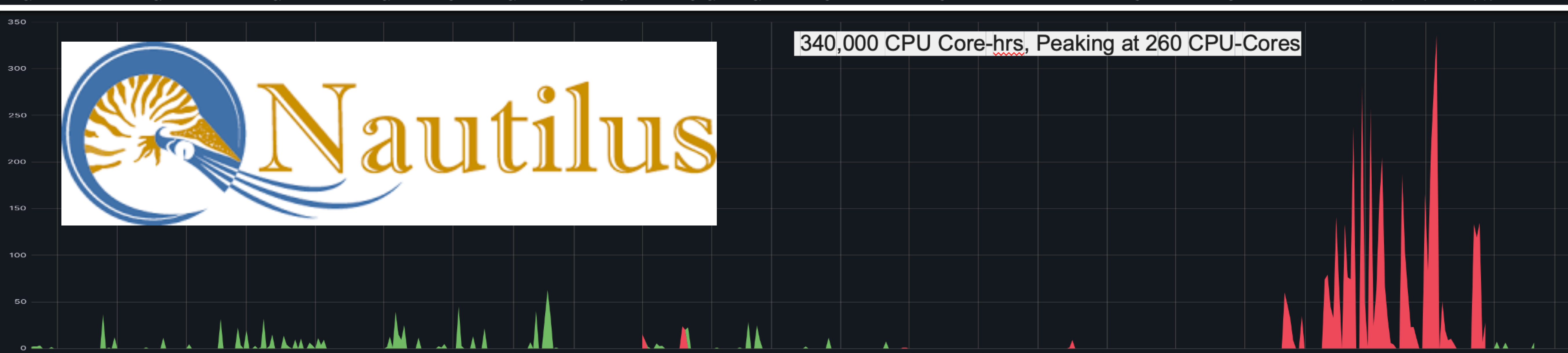
Ability to scale processing to large clusters with k8s

Current Usage

Phil Harris, MIT

LIGO GPU/CPU Usage Per Day, Last 6 Months





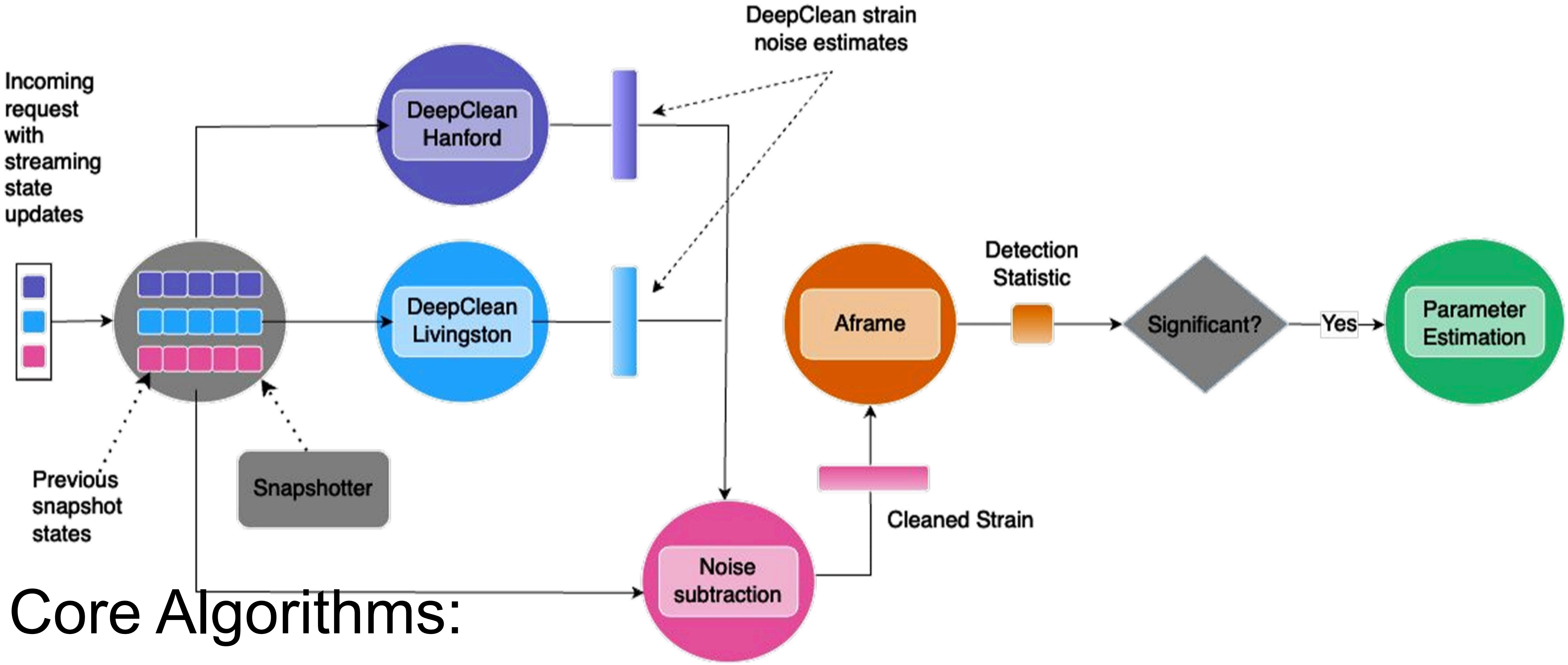
NATIONAL RESEARCH PLATFORM

Namespace osg-ligo, bbhnet

NRP Nautilus

Algorithm Training of a black hole merger algorithm

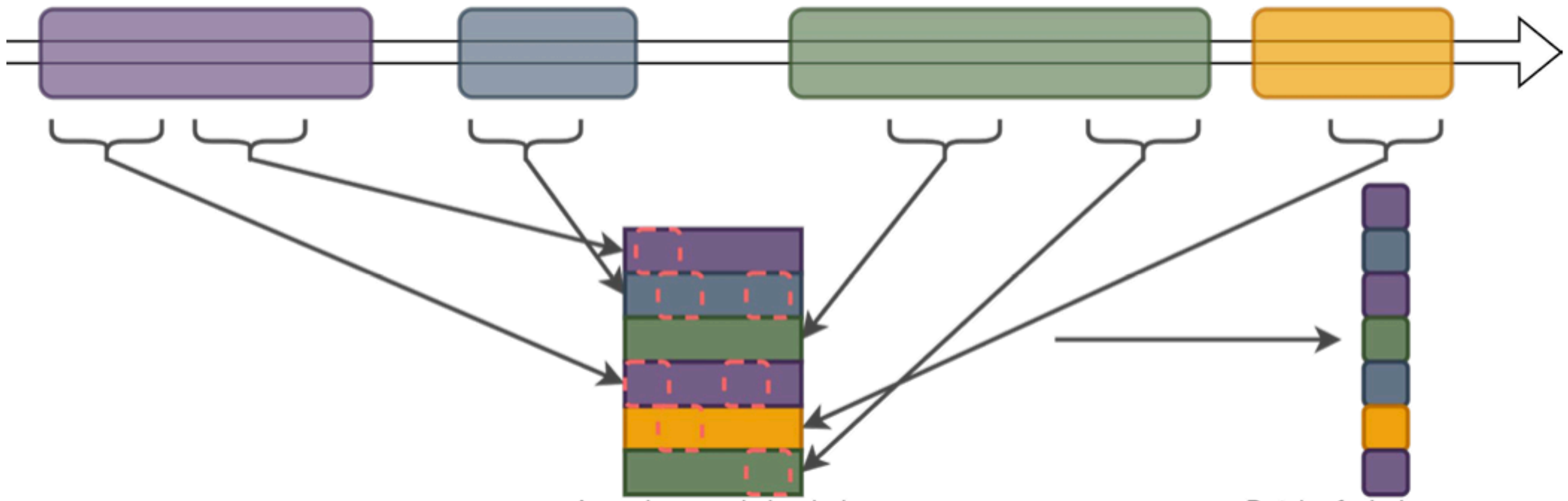
What Algorithms exist?



- Core Algorithms:
  - Detector Denoising => Deep Celan
  - Black Hole Merger detection => A-frame
  - AI based anomaly detection => GWAK
  - Neutron Star Merger detection
  - Parameter Estimation

Time Series Caching

Transitioning to larger datasets



Chunked loading of background data

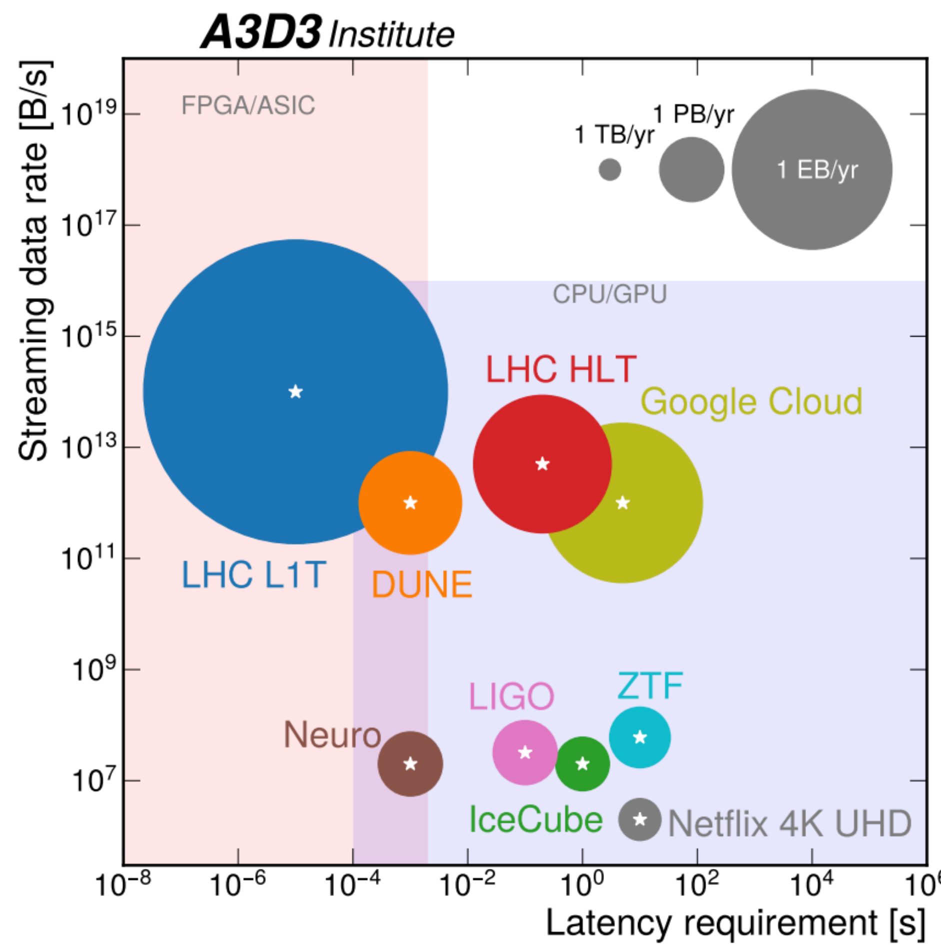
GPU batching can be enabled through chunked data loading
Parallel time series processing allows fast training

Computing Challenge

- GW data is time series data
  - Our toolkit targets critical time series setup
- Vanilla ML processing workflow
  - Load ML model and shove data through it :
    - 512s of data per second (s’/s) on 16 GB V100 GPU
    - 1 year of data is 17hrs of computing
    - 100 years (needed for analysis) is 70 days
      - Too slow to iterate on ideas
- Our workflow utilzes optimized schemes to avoid these issues

ML Challenges

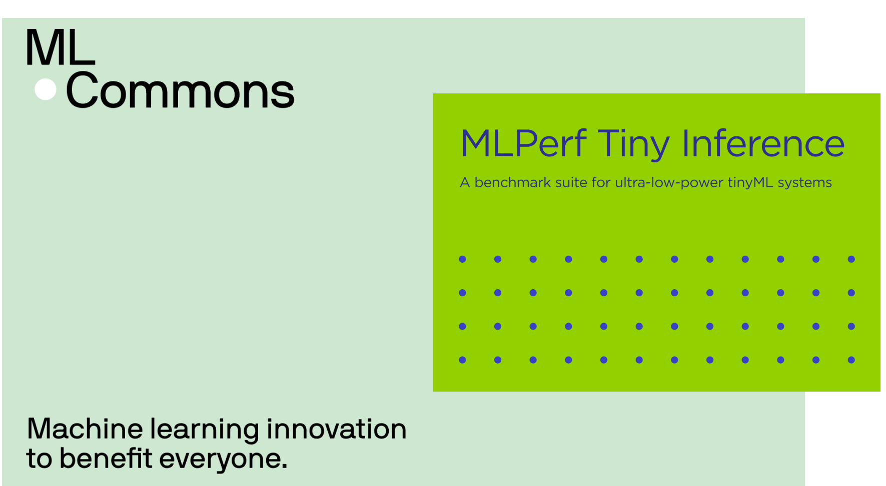
- Aiming to build a website hosting Scientific ML Challenges



Accelerated AI Algorithms for Data-Driven Discovery

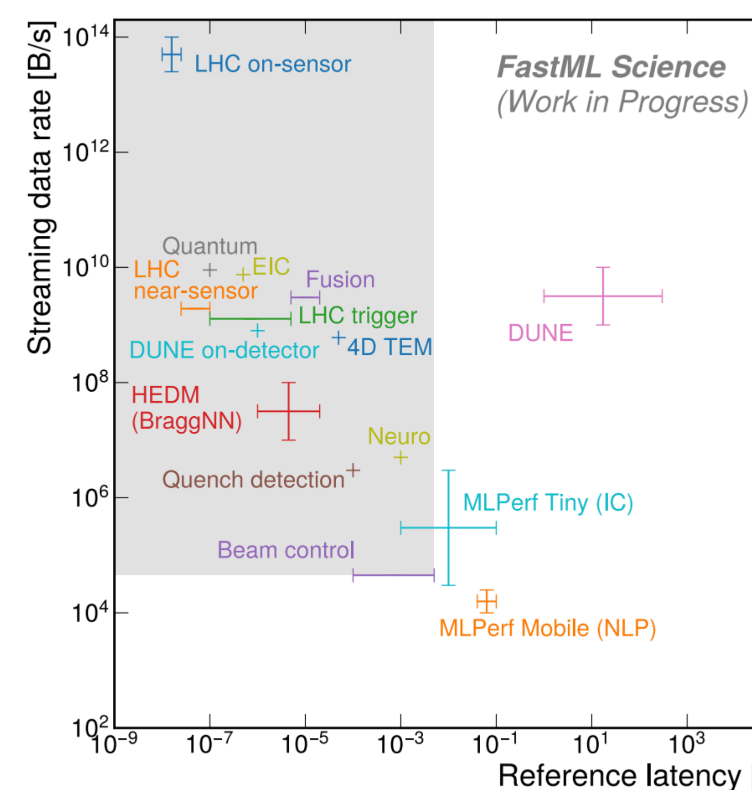
Connecting with ML Commons

Connecting With Hardware



Would like to highlight Criticality of Scientific Problems

Support from NSERC <u>FAIRUniverse</u>

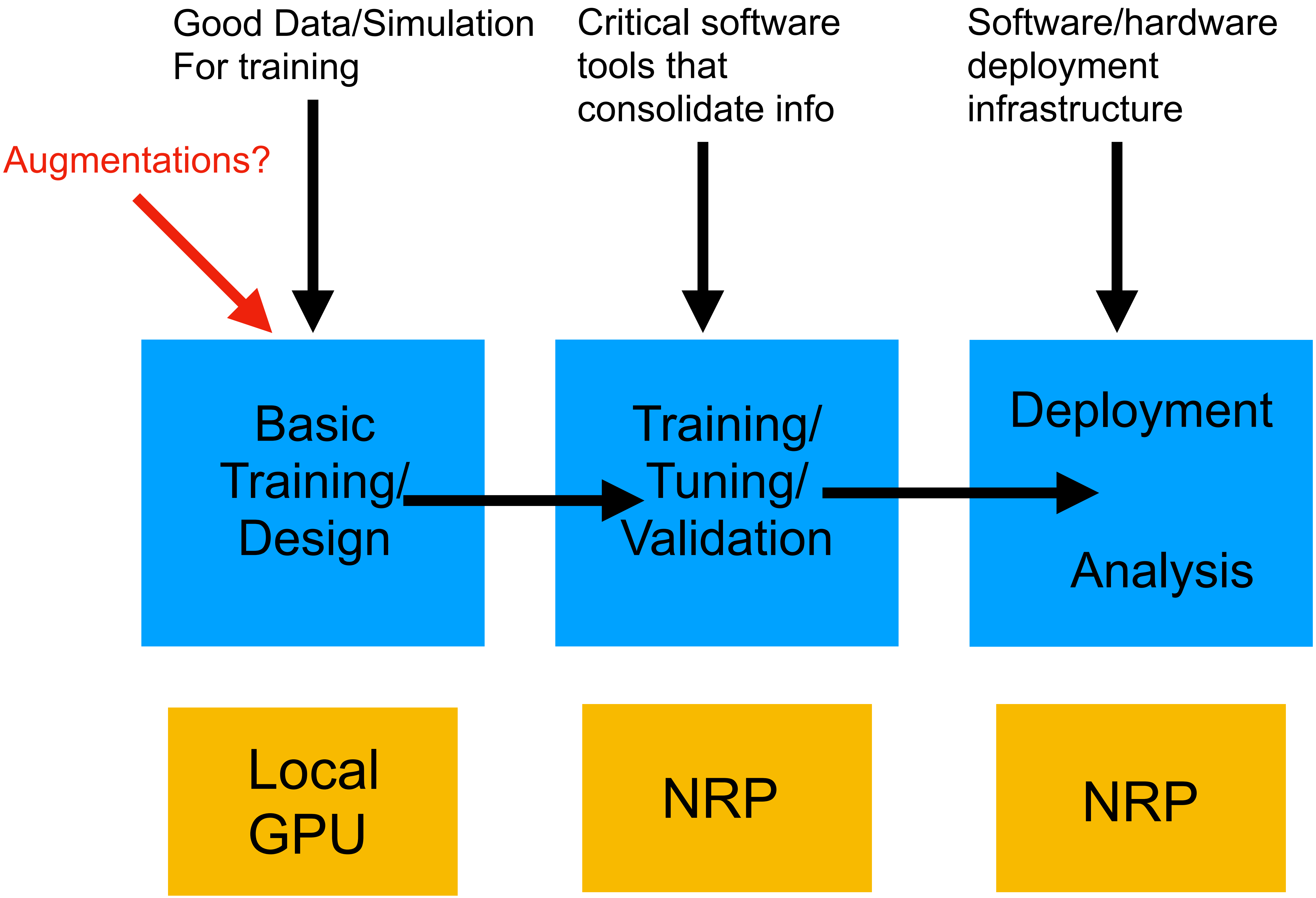


A Vision

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

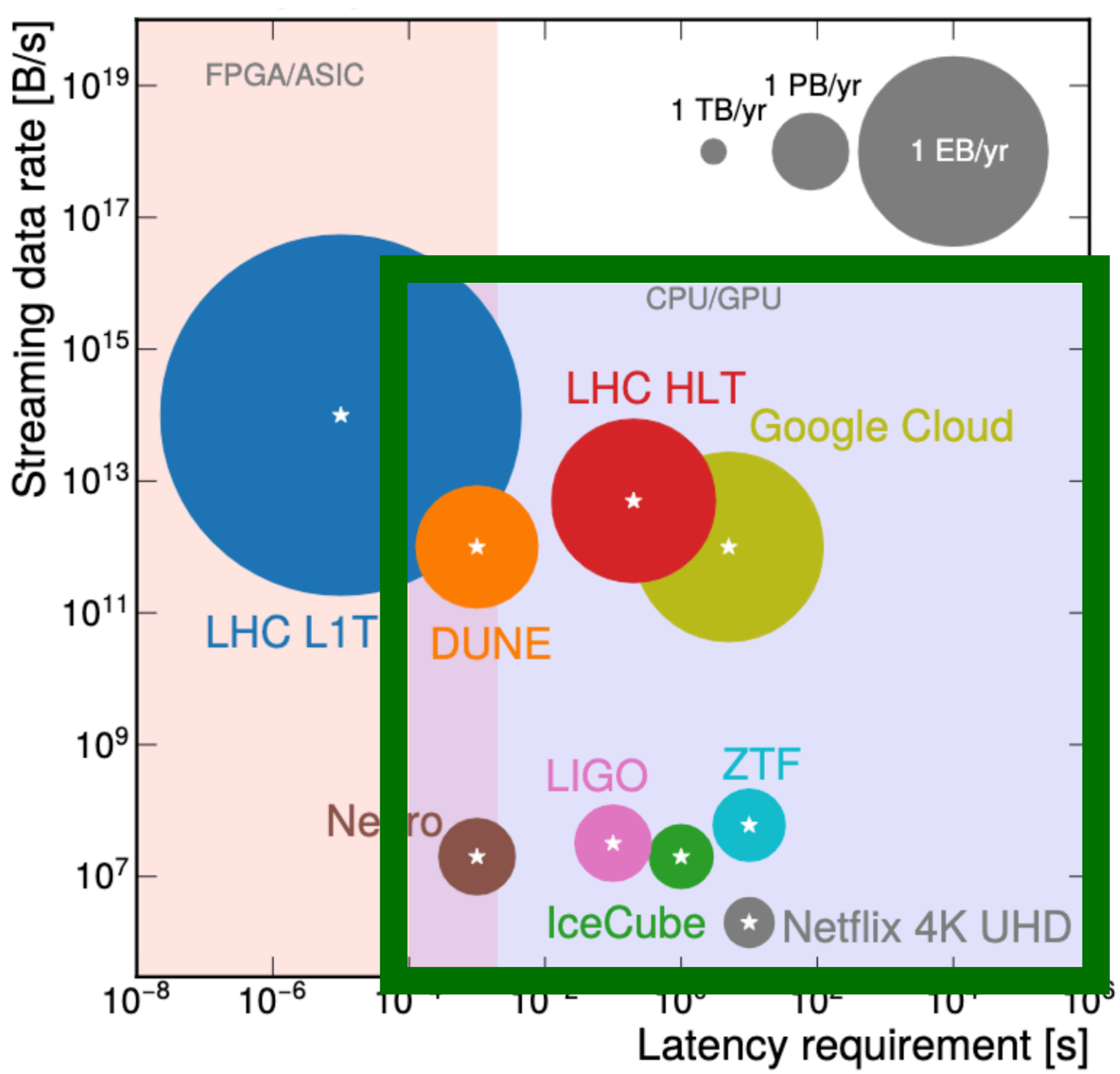
		Scientific Moonshots		
		Domain A	…	Domain N
AI thrusts	AI - 1: Real-time	Benchmark 1A		Benchmark 1N
	AI - 2: Control			
	AI - 3: Autonomous			
	AI - 4: Foundation			
	AI - 5: Generative	Benchmark 5A		Benchmark 5N

Anatomy of an Algo



What computes are here?

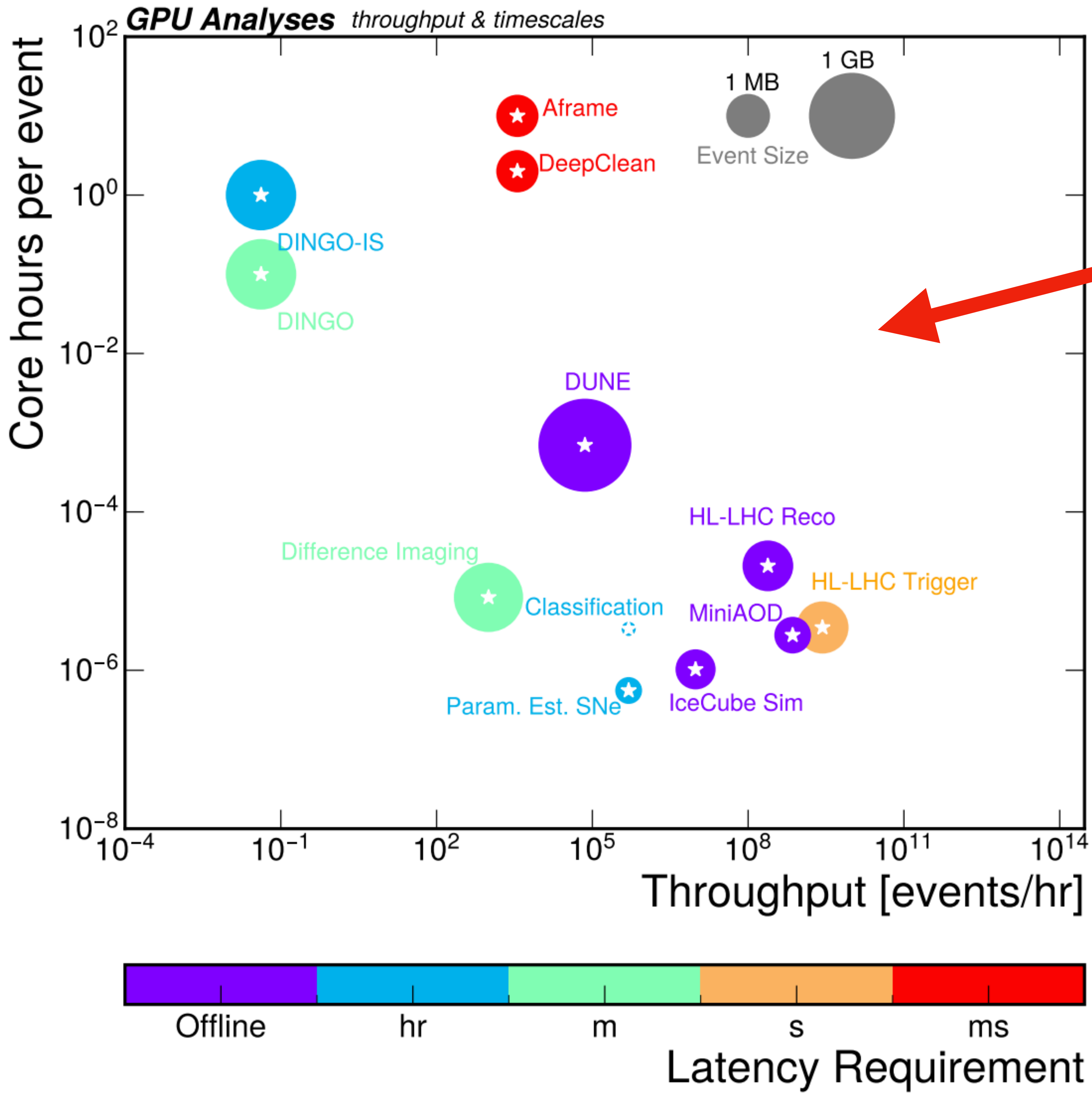
- Within the FastML Community there is a broad range
  - We often try to characterize this range by customization
  - Low Latency and Low Power need more customization



This is our focus here

We want to understand the high throughput component

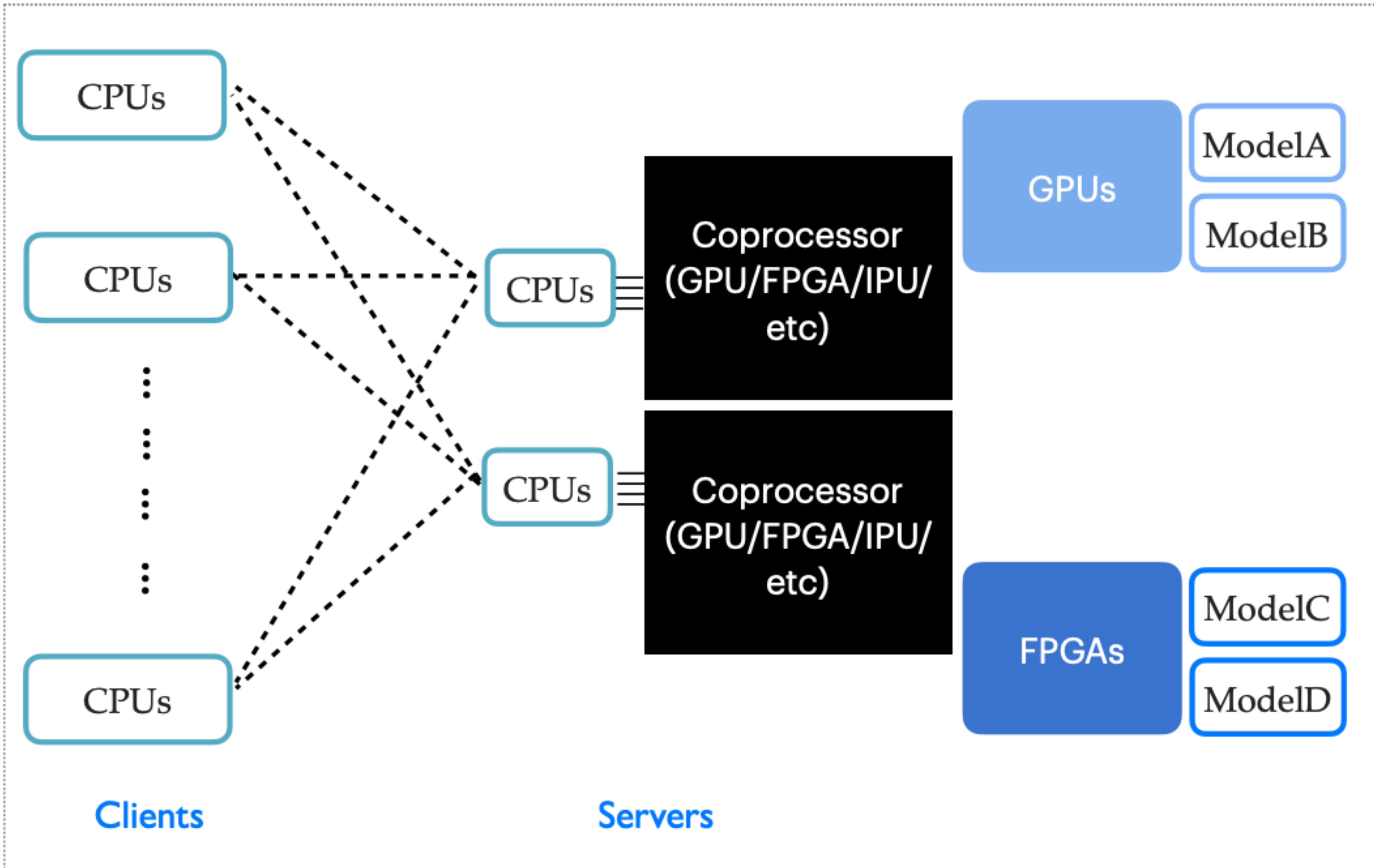
GPUs for ML



For large algorithms parallelizability and shear comput is unprecedented

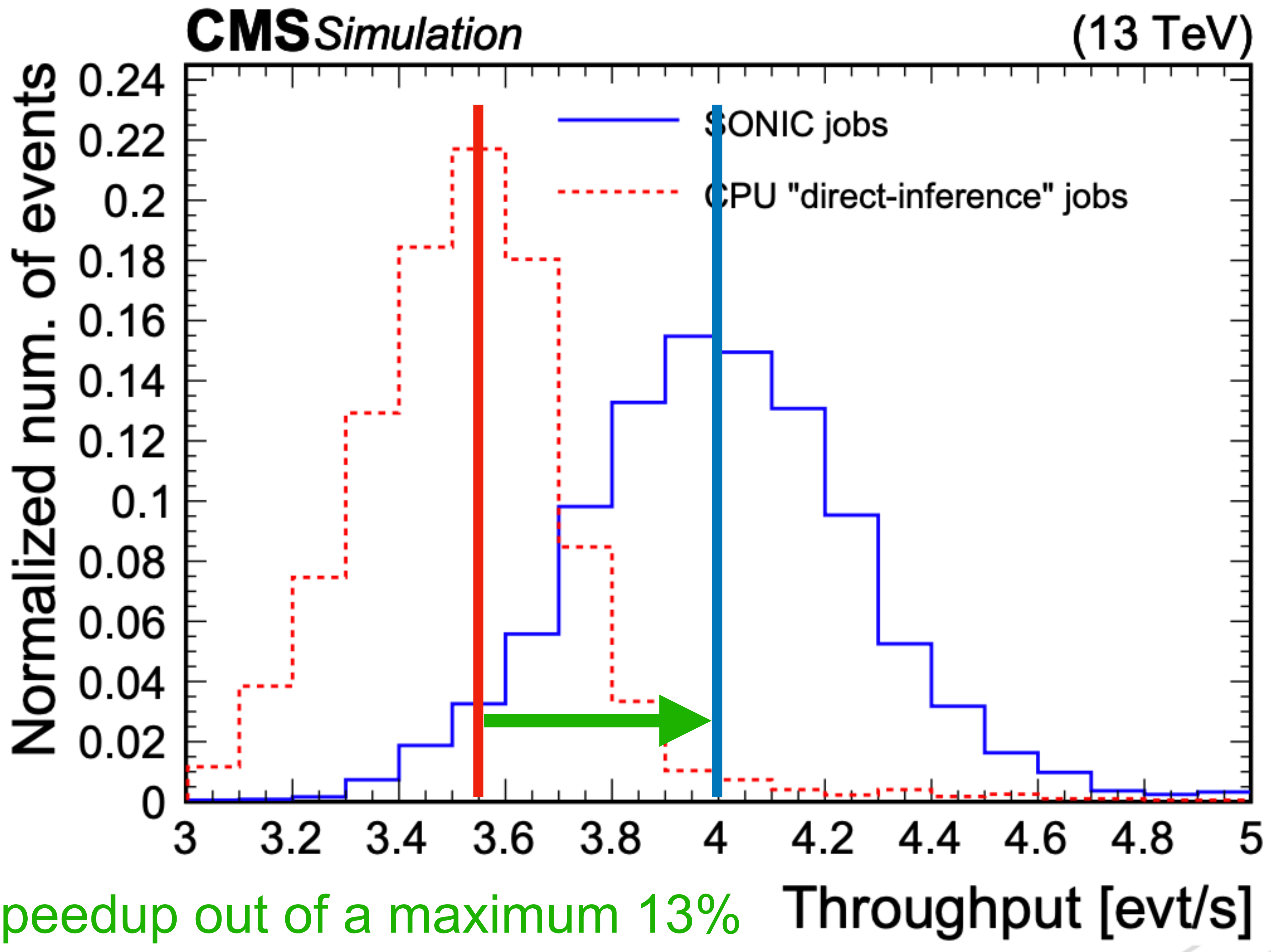
Major demand emerging in many domains

Deploying to Scale



Proof we can

10k CPUs & 200 GPUs



12% speedup out of a maximum 13%

A3D3

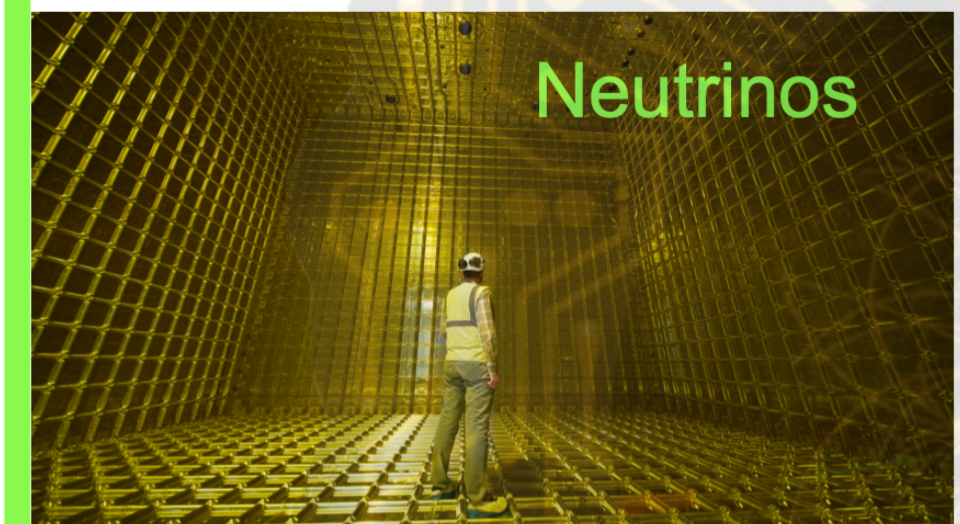
- An institute to unite real-time AI
  - Quickly looking for people to be part of extended team

Accelerated AI
Algorithms for
Data-Driven
Discovery

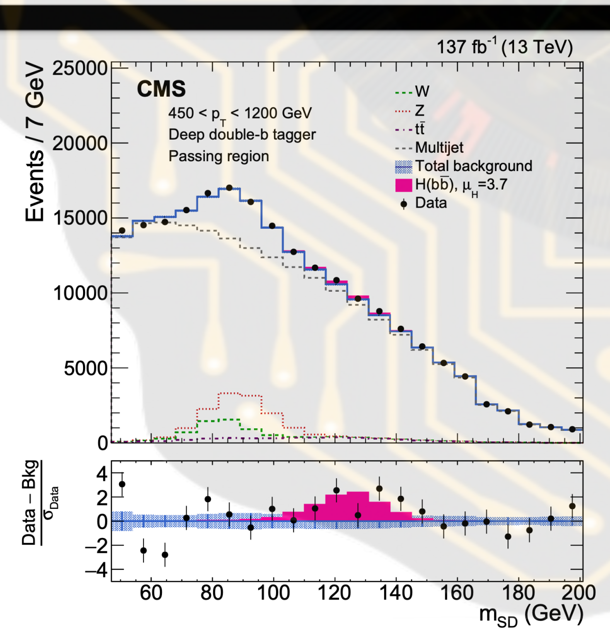
An Institute: **A3D3**

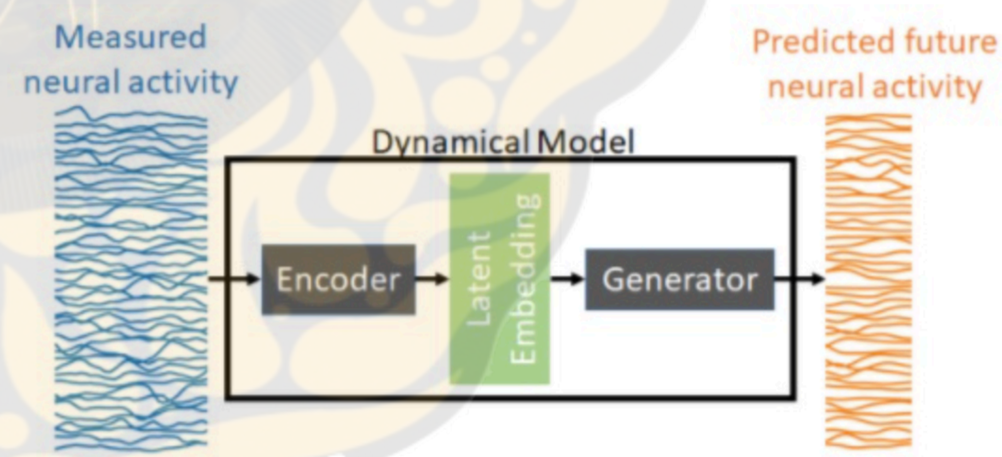
- We have been awarded a new institute to explore real-time AI
  - Accelerated AI Algorithms for Data Driven Discovery (A3D3)

New Types of Computing



LHC Physics

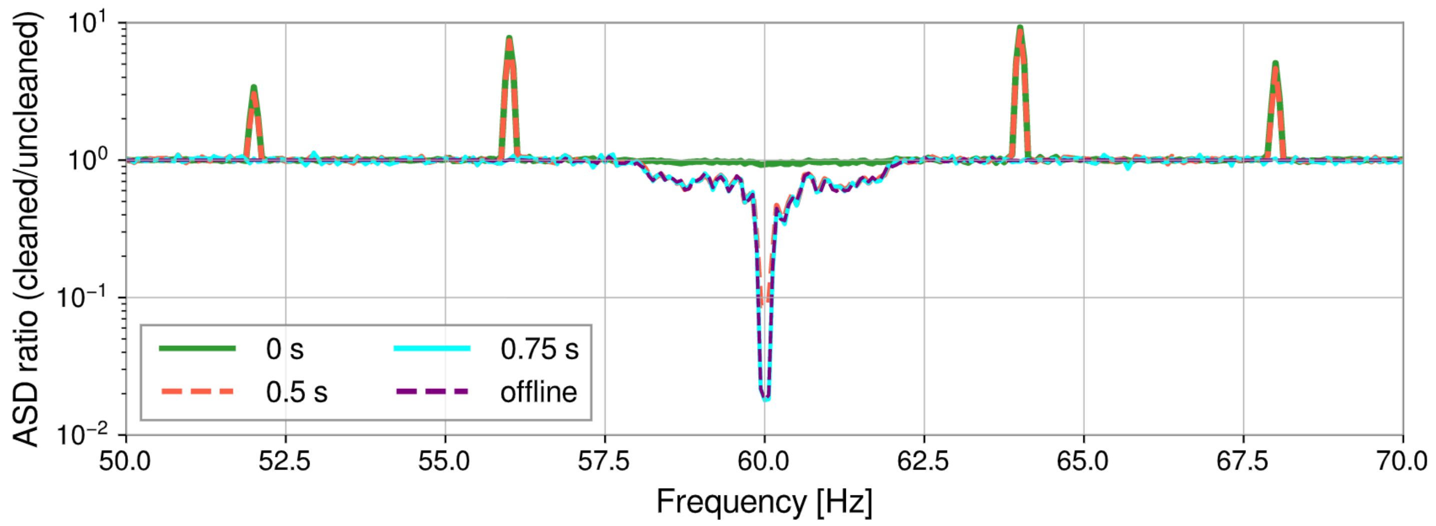




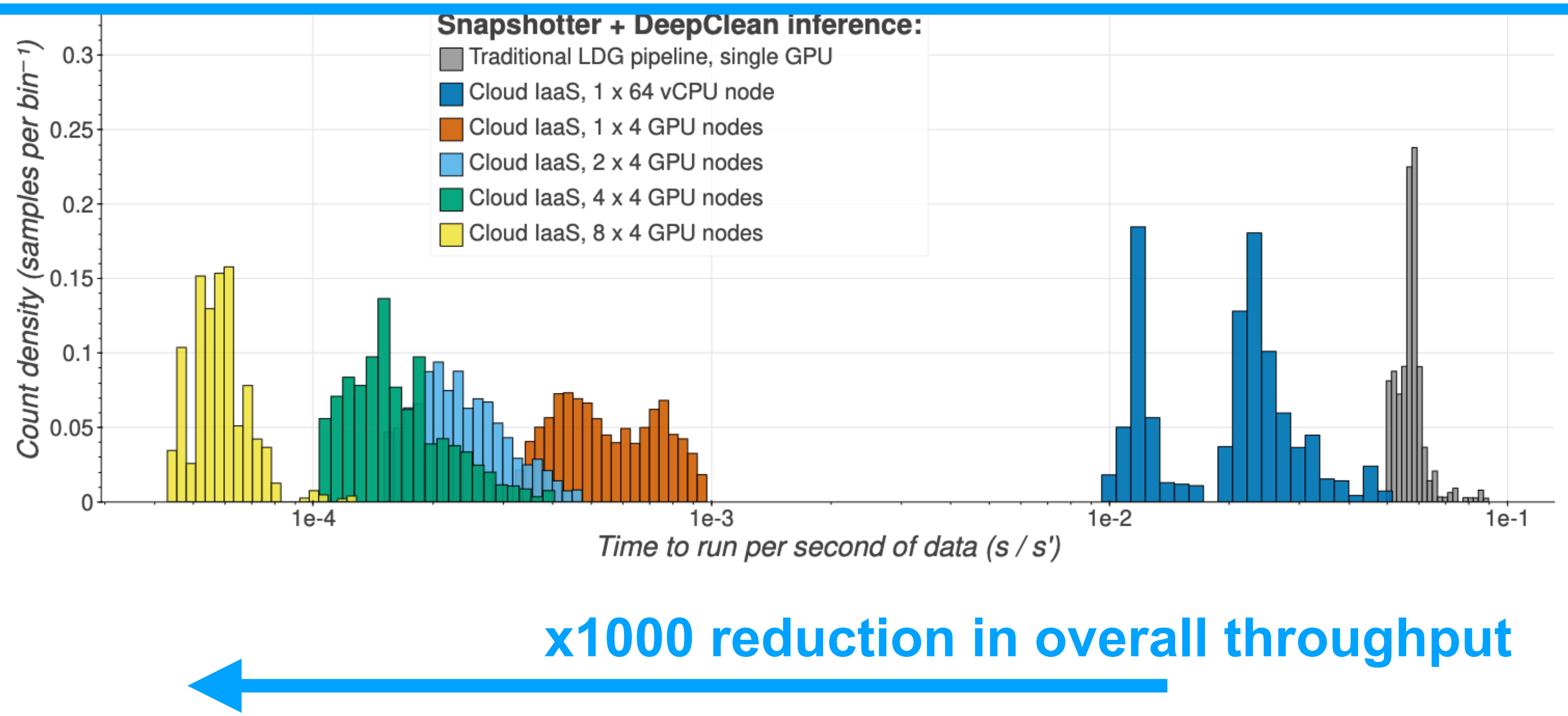
NeuroScience

Gravitational Waves

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



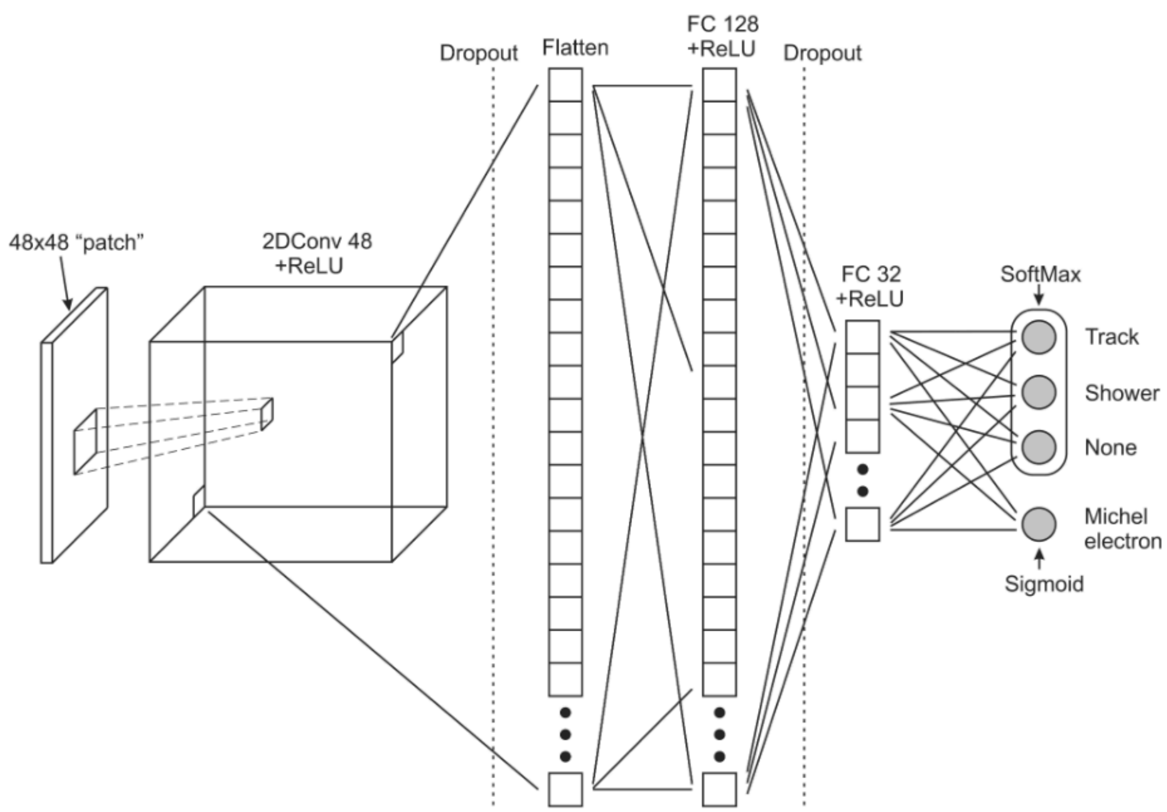
Developed AI-based Denoising and BBH detection



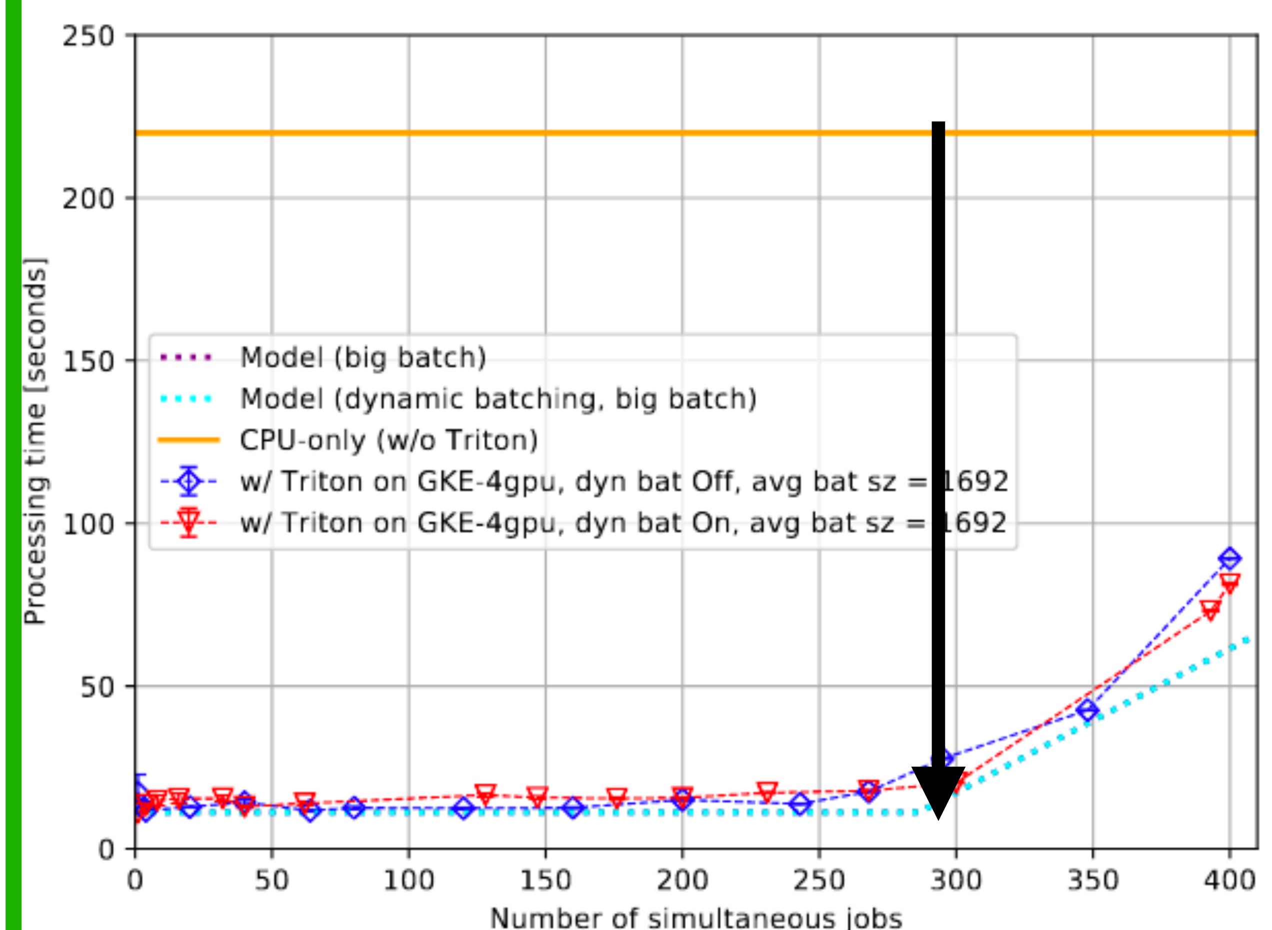
Constructed a GPU-as-a-service integration for GW low latency alerts

Neutrino Physics

- We are pursuing the same idea in Neutrino physics

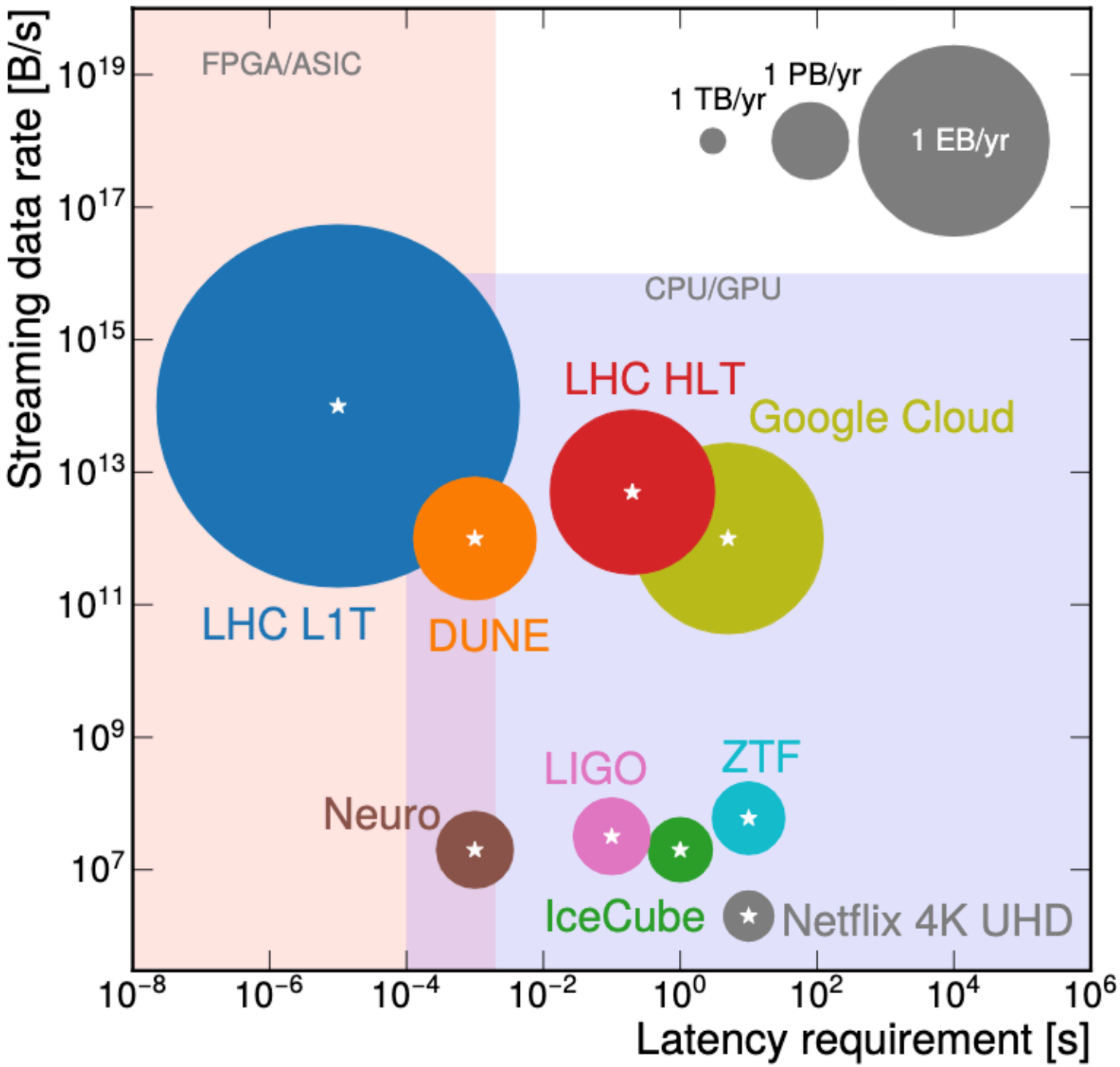


Michel Electron Id NN

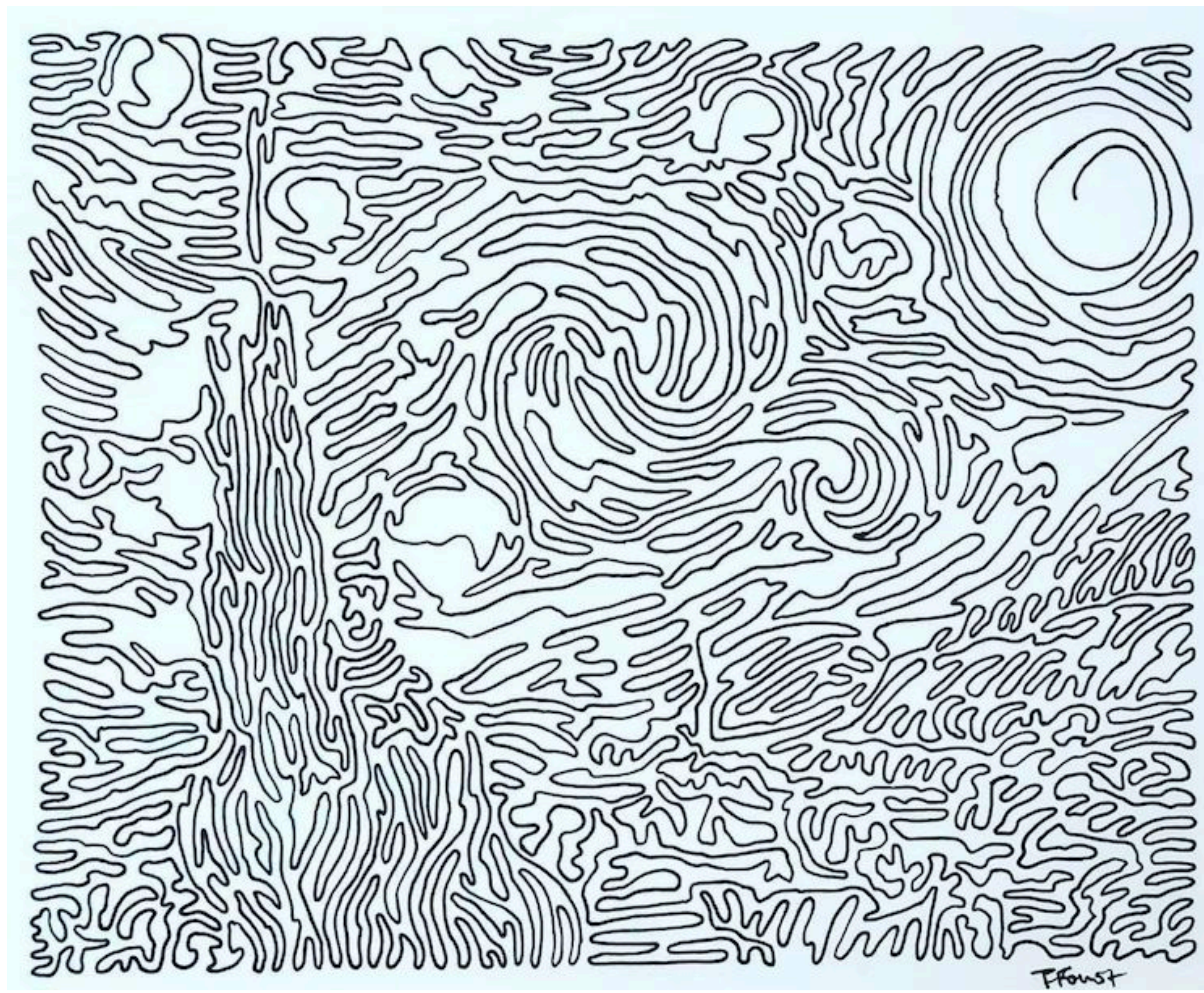


Large Factor in speed up

Preparing for the future

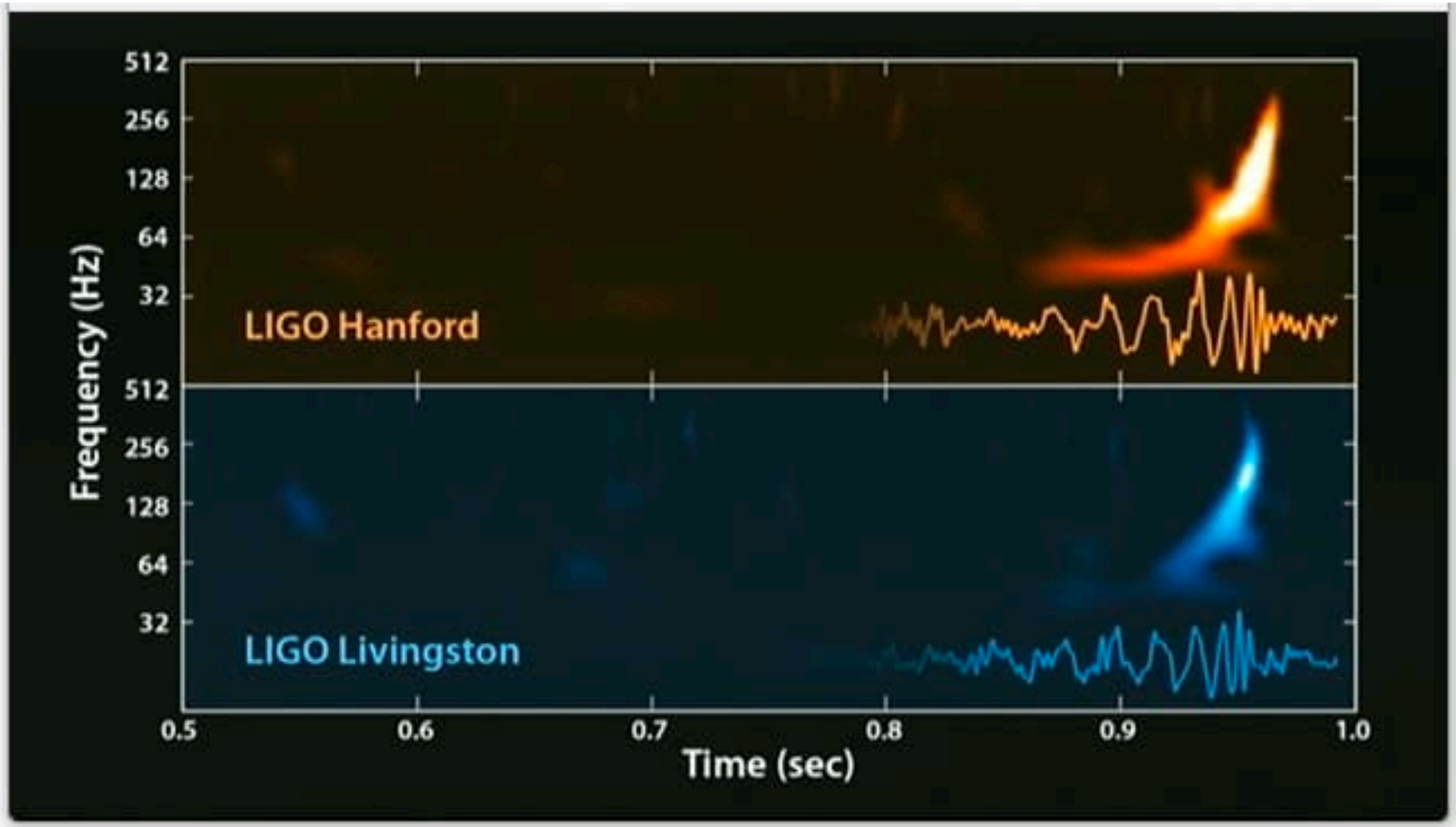


Thanks!



Accelerated AI
Algorithms for
Data-Driven
Discovery

LIGO Challenge: Can we find all mergers

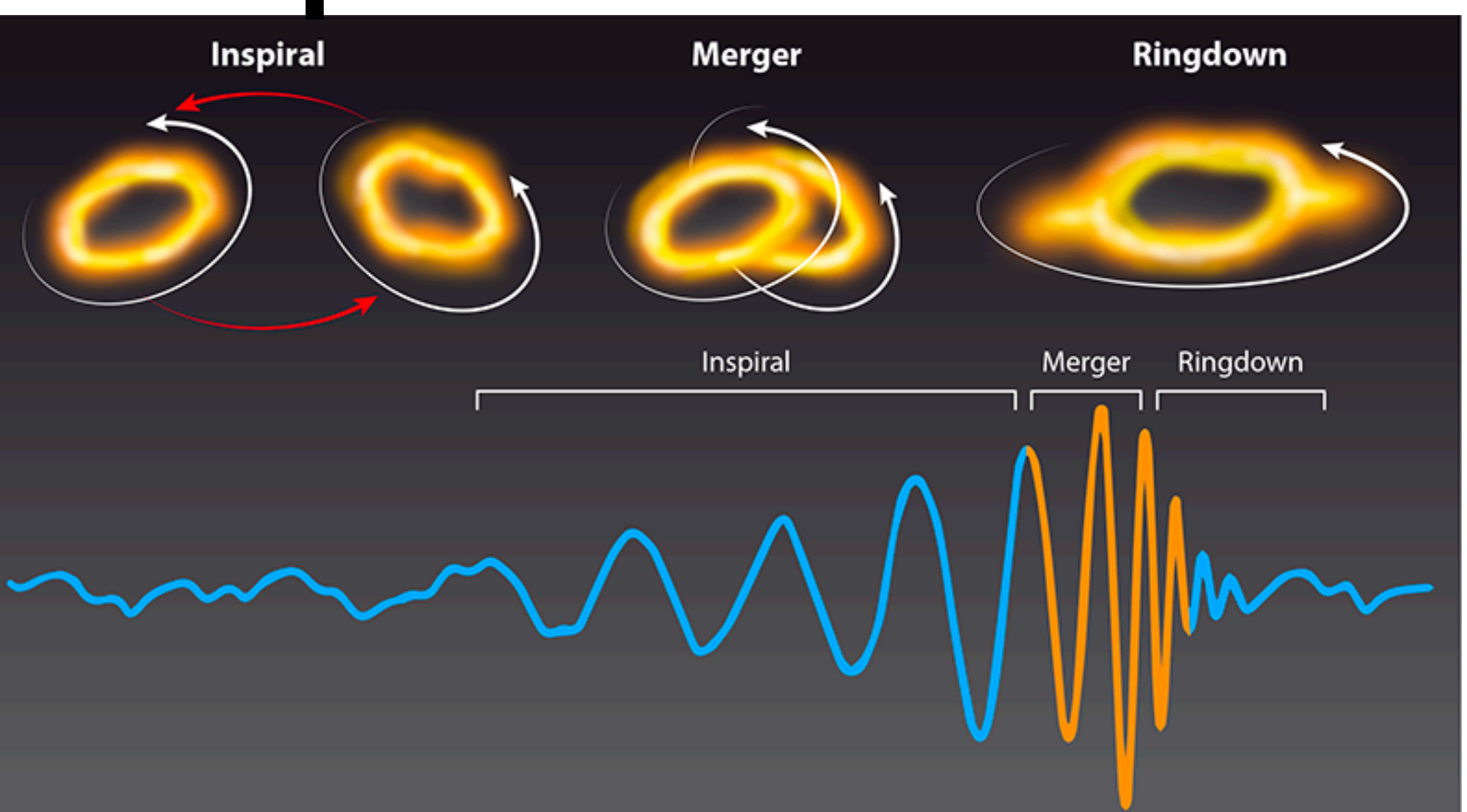


- LIGO has 10^5 channels at 1024 Hertz
- Looking for subtle signals hidden in the noise

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

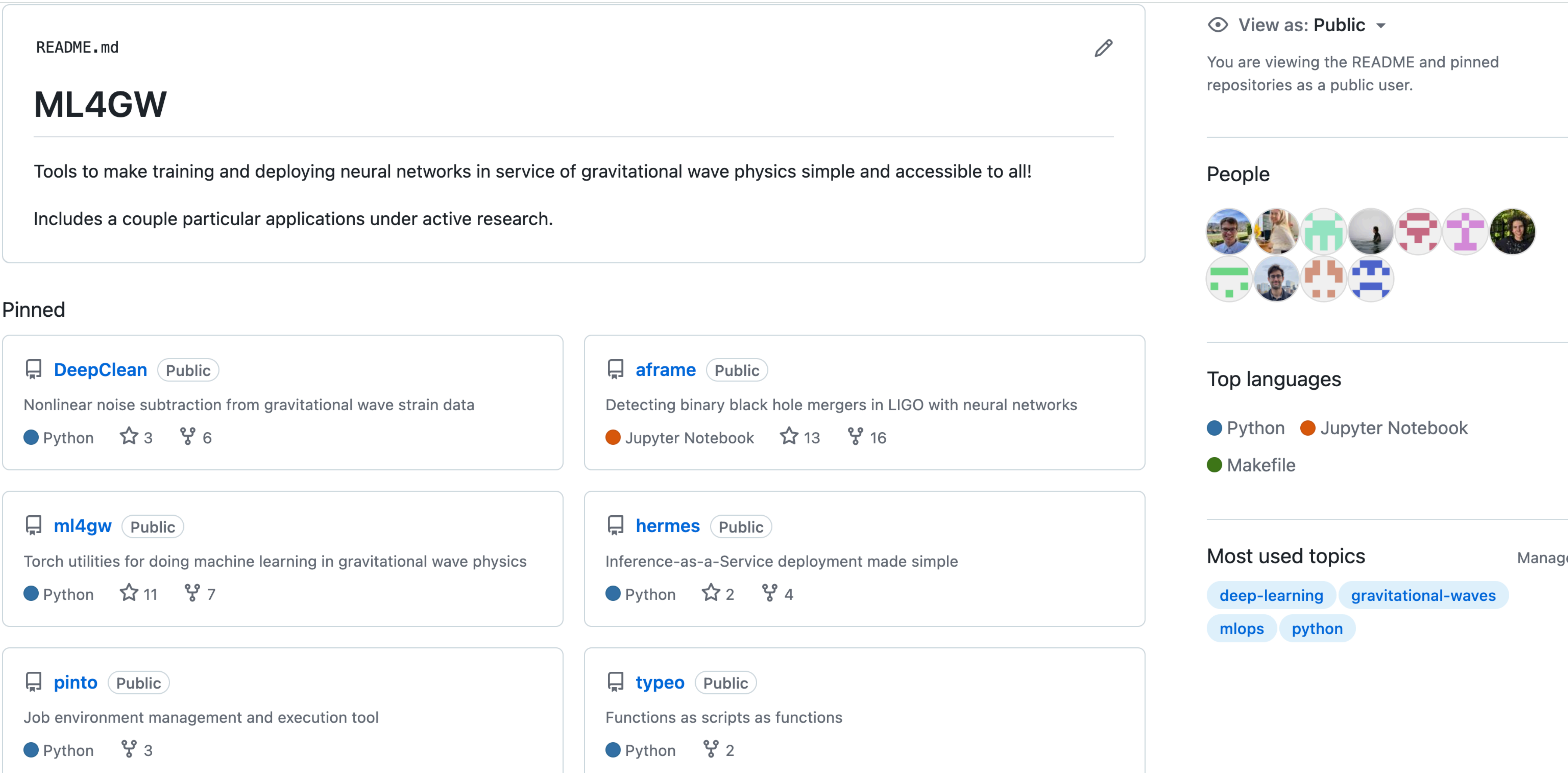
LIGO Path to success

Data Taking → Signal Processing (Denoising) → Core Algorithm → Parameter Estimation → MMA Alert → Sky localization Merger Type

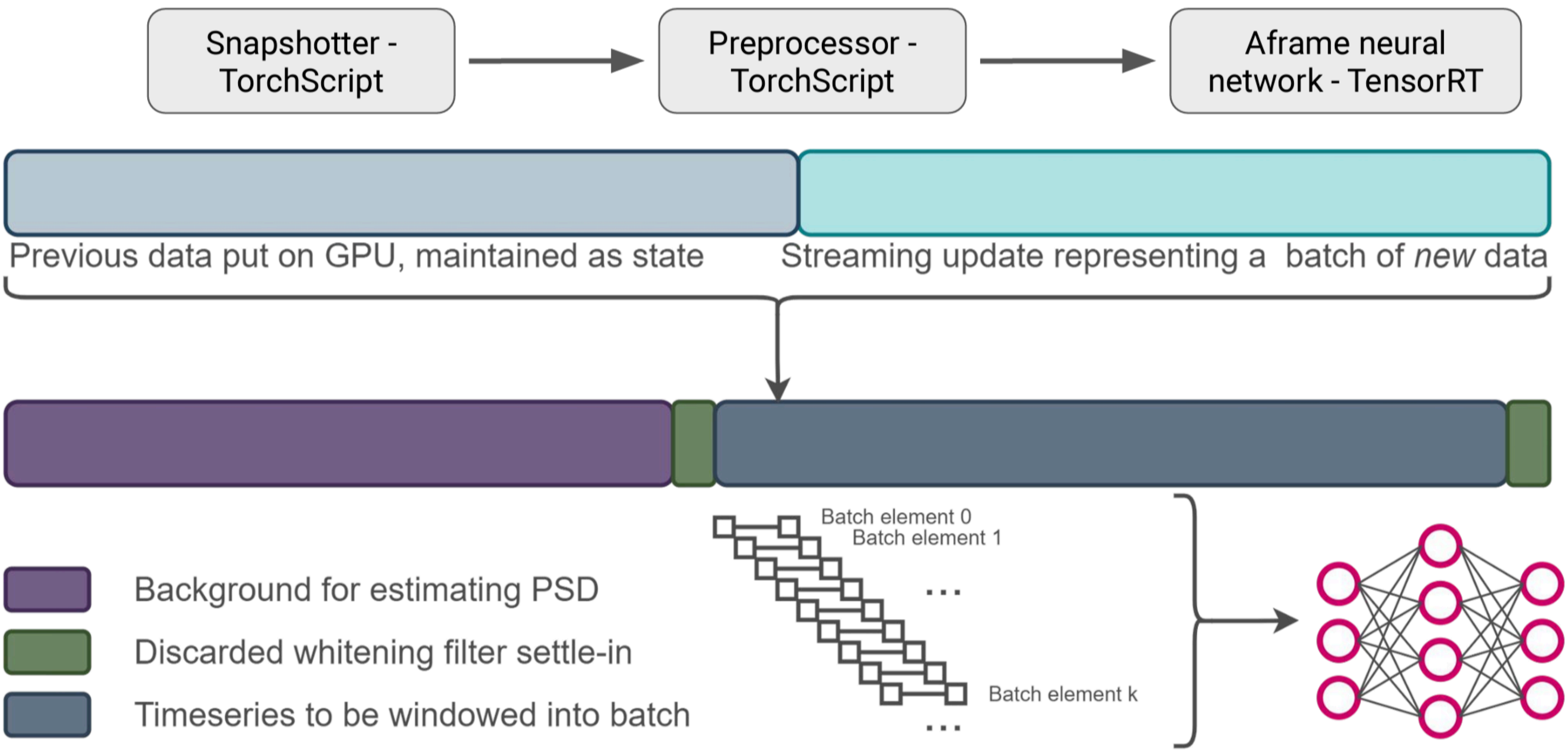


ML4GW toolkit

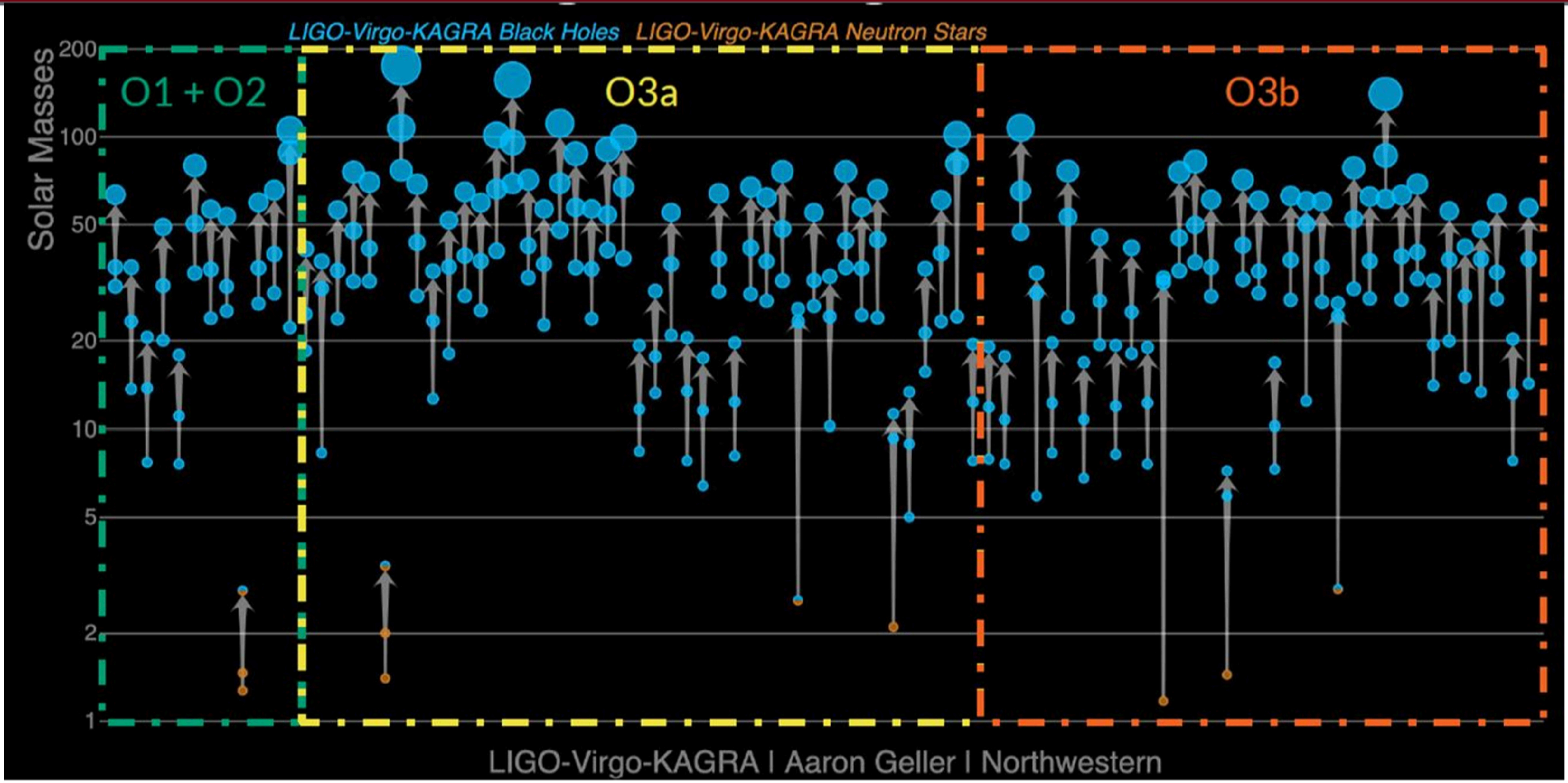
- Enable a complete AI pipeline we have developed ML4GW
  - Comprehensive toolkit for ML pipeline in Gravitational Waves



Hermes: Inference -as-a service deployment



Third transient event catalog: GWTC-3



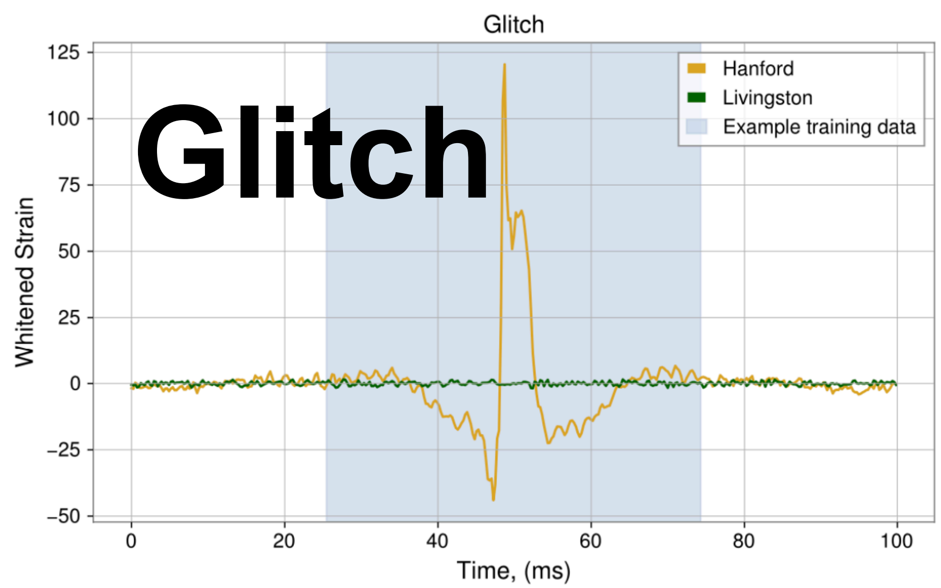
11 events from O1+O2

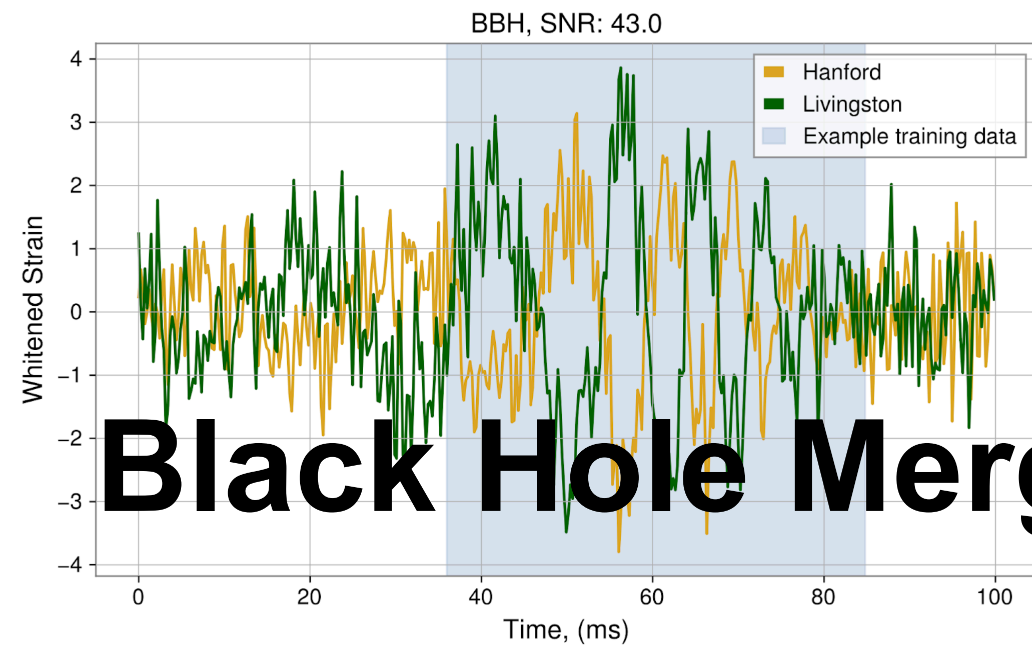
44 events in O3a, 55 total
1041 “subthreshold” events in O1,O2,O3a

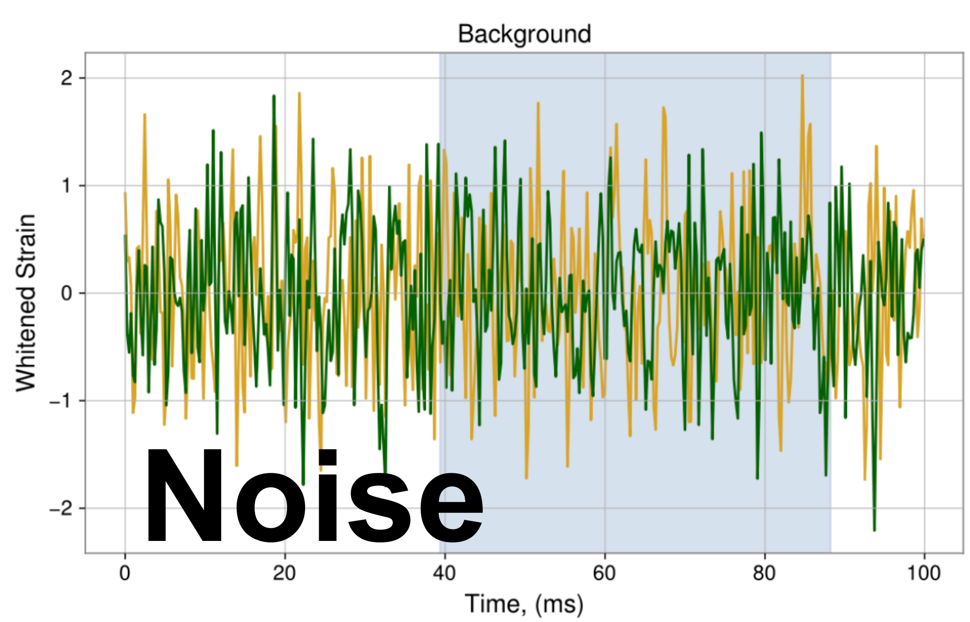
35 events in O3b, 90 total (catalogs are cumulative)

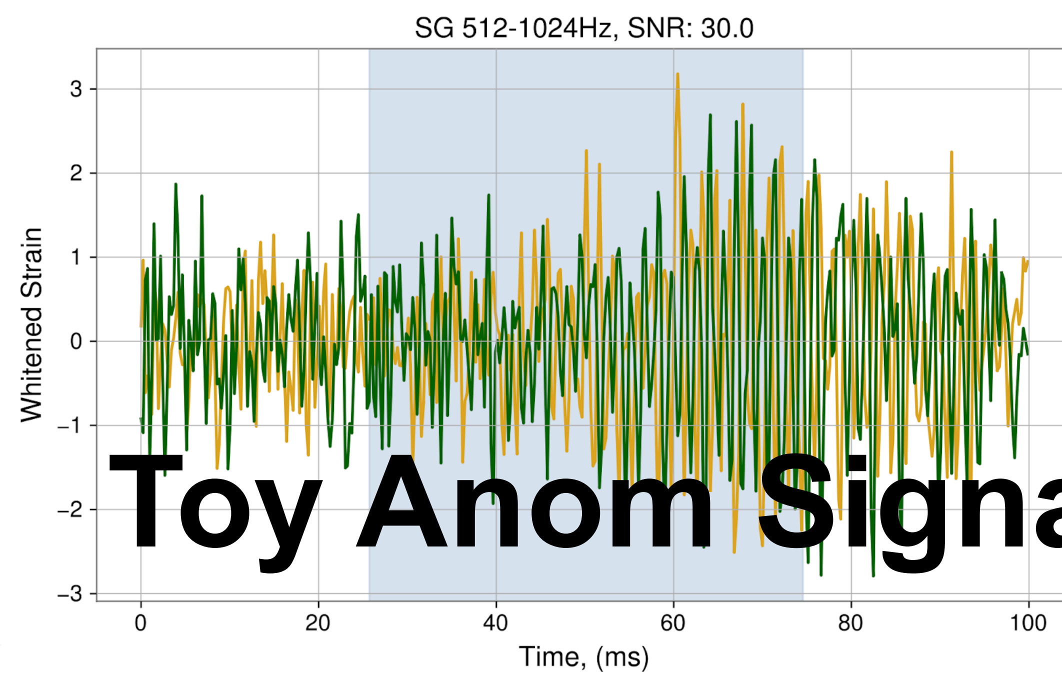
GWAK Space

- GWAK stands for GW (QU)AK like guacamole



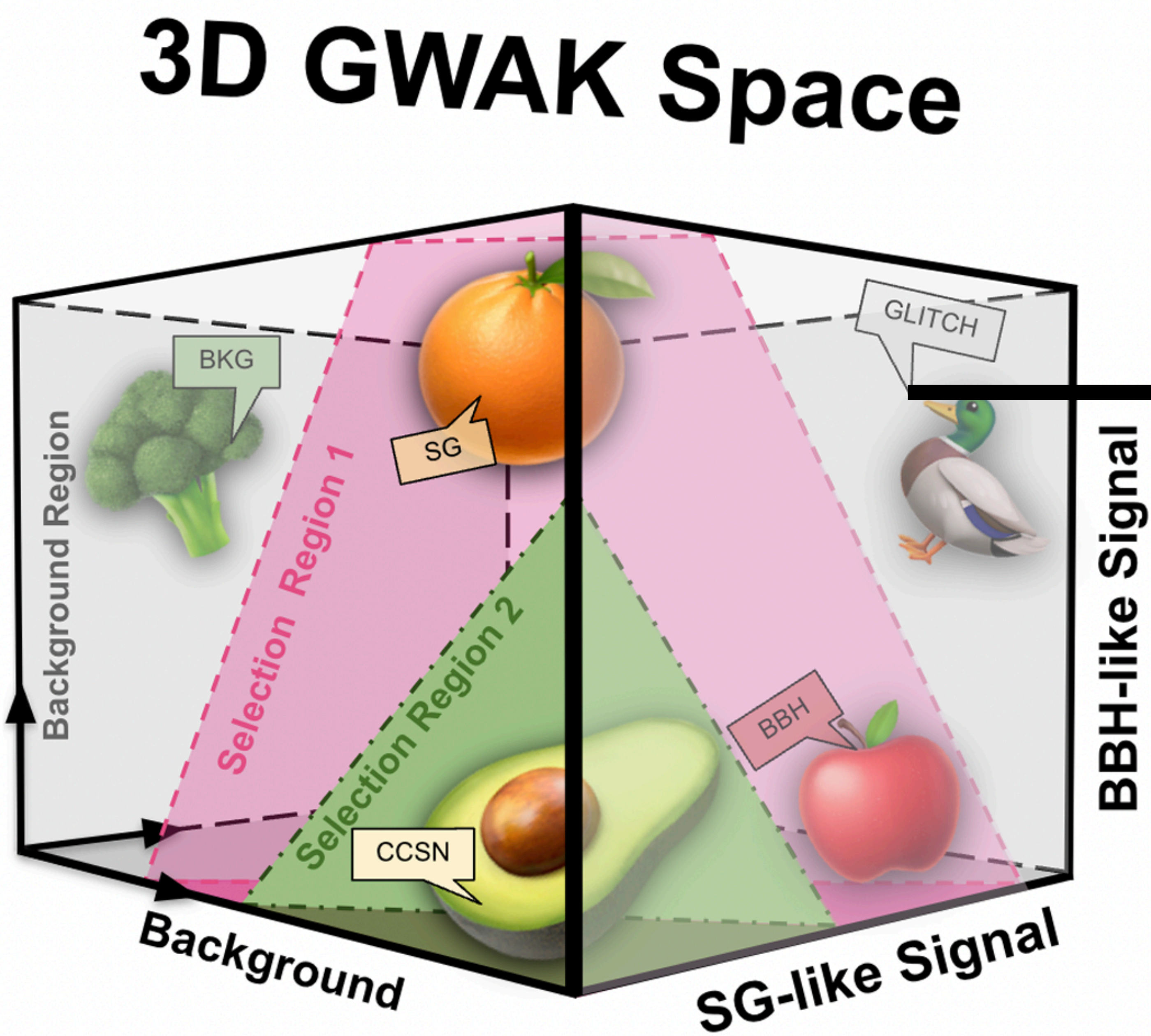






GWAK Space

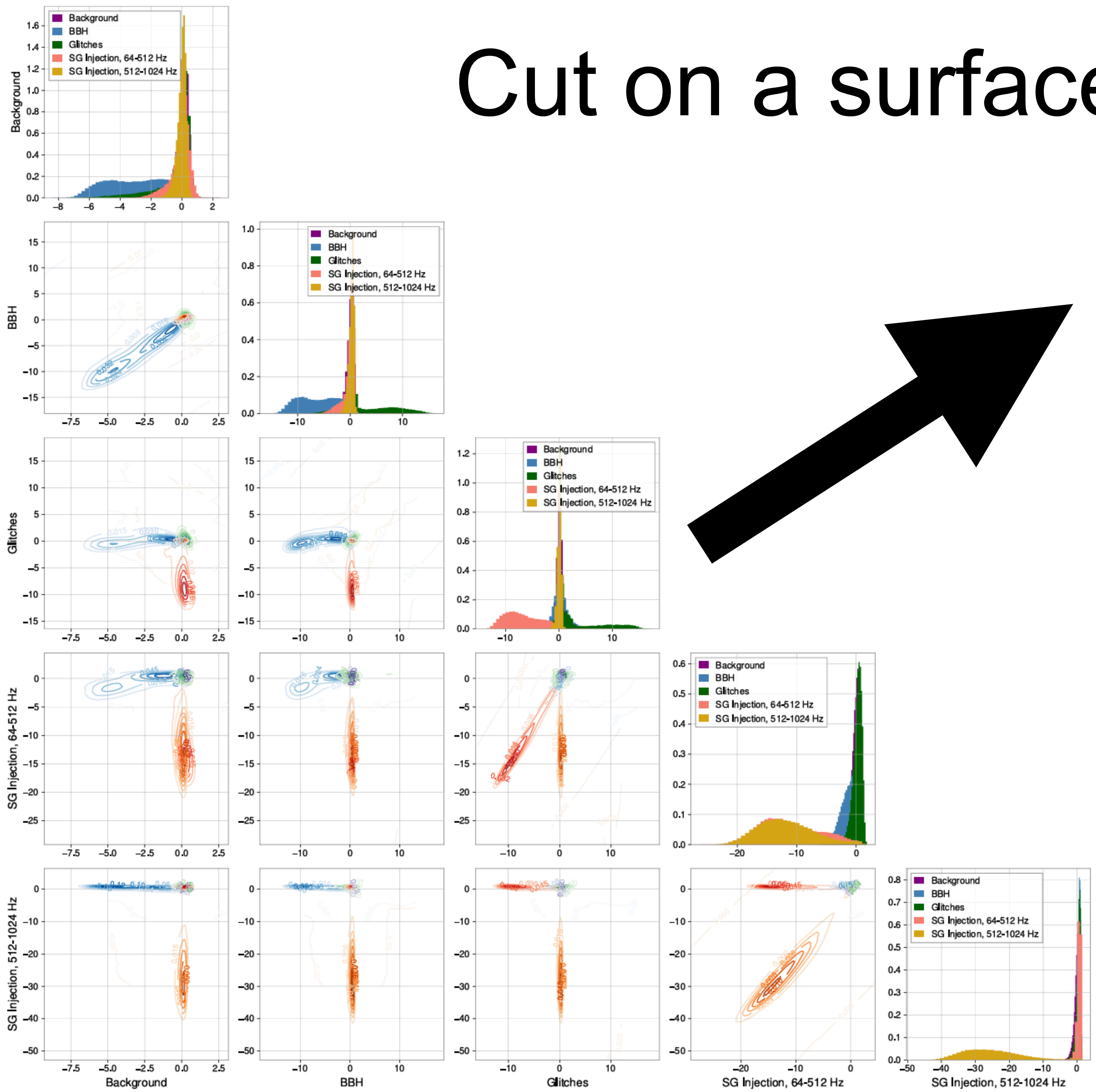
- Using autoencoders we cosntruct a similar space



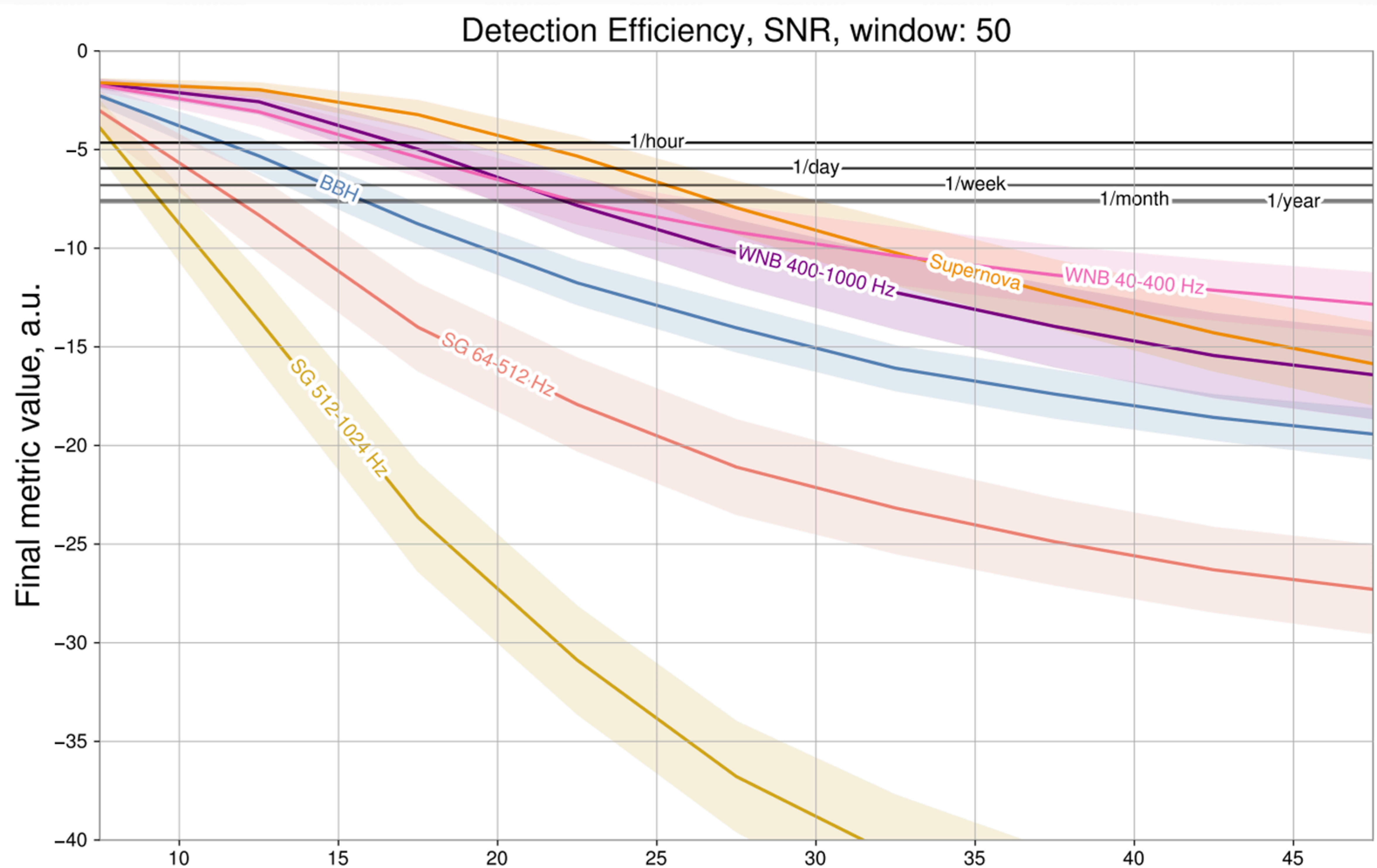
We Make a 5D Space

GWAK Space

Cut on a surface in the 5D Space

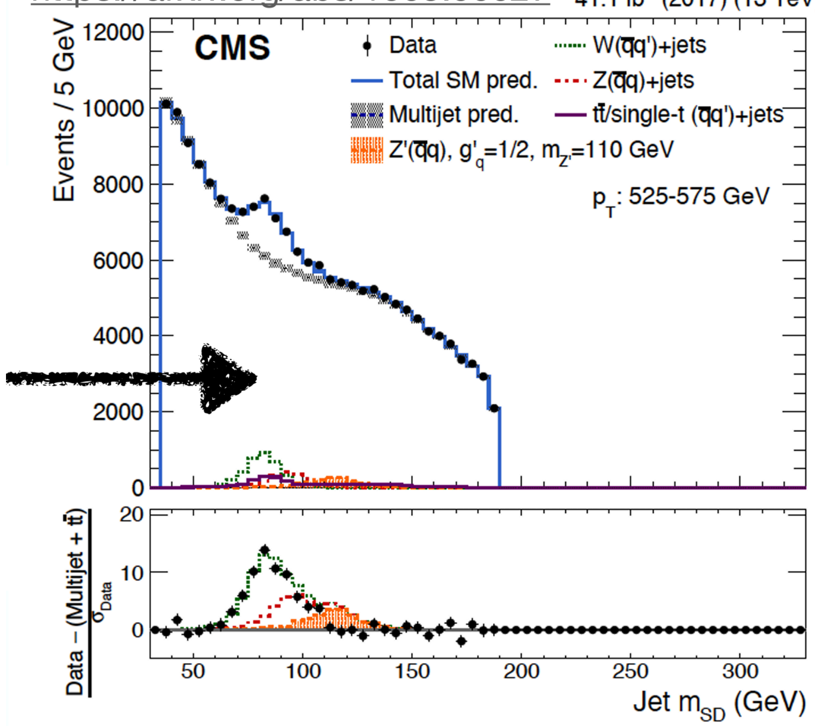


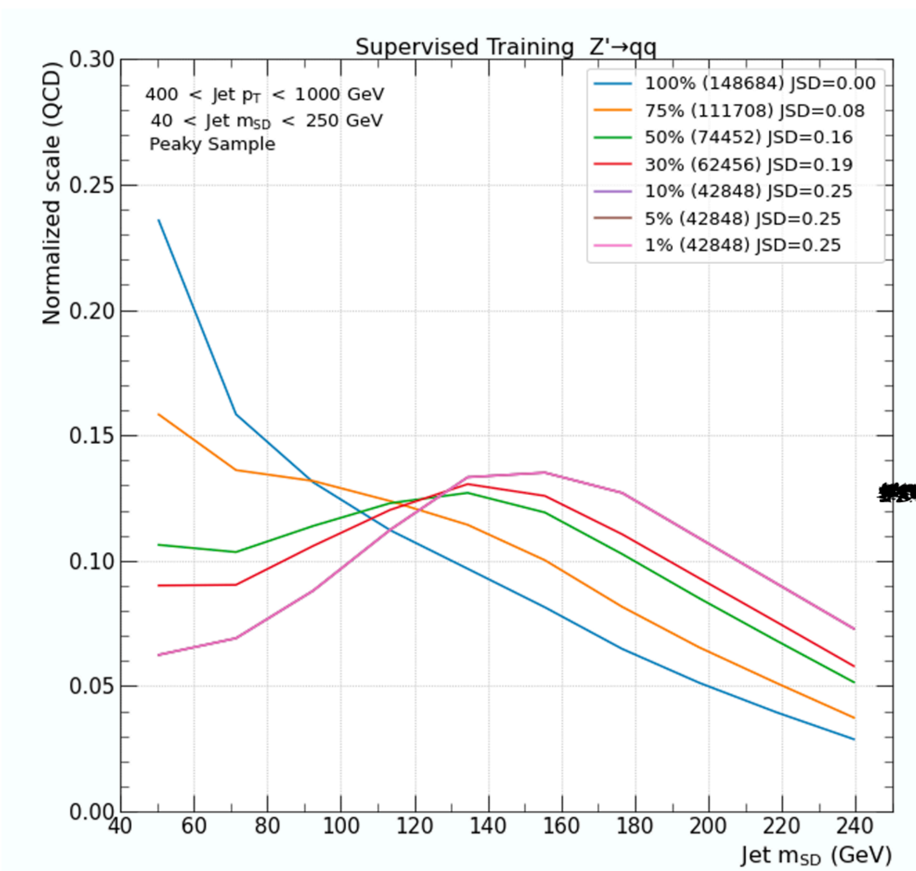
GWAK Algorithm

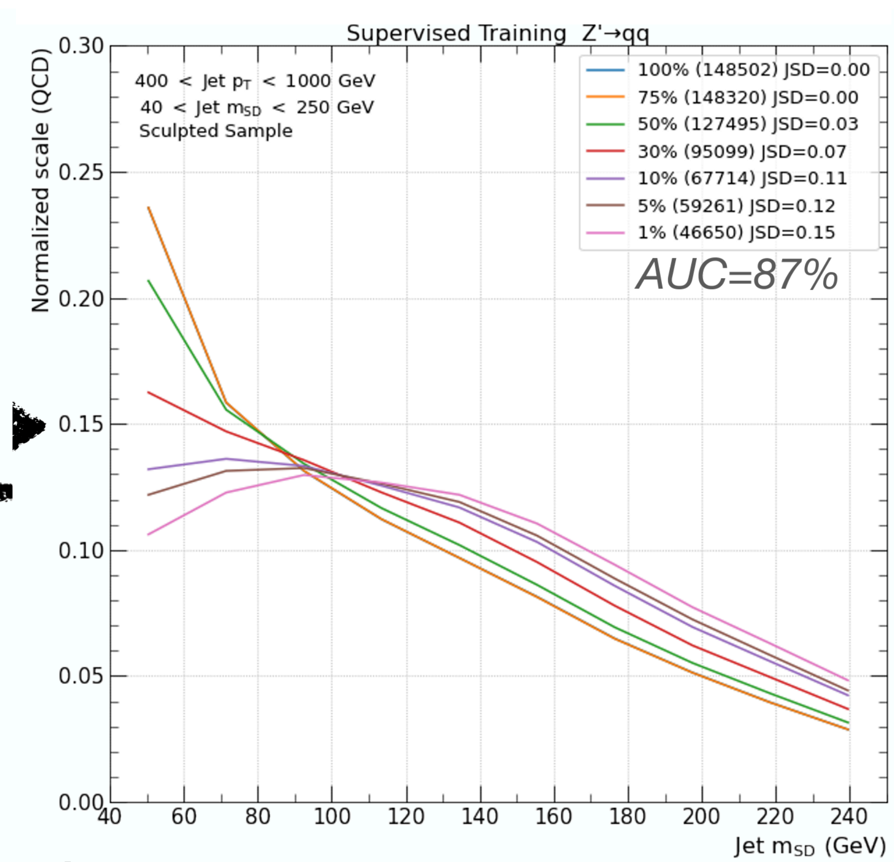


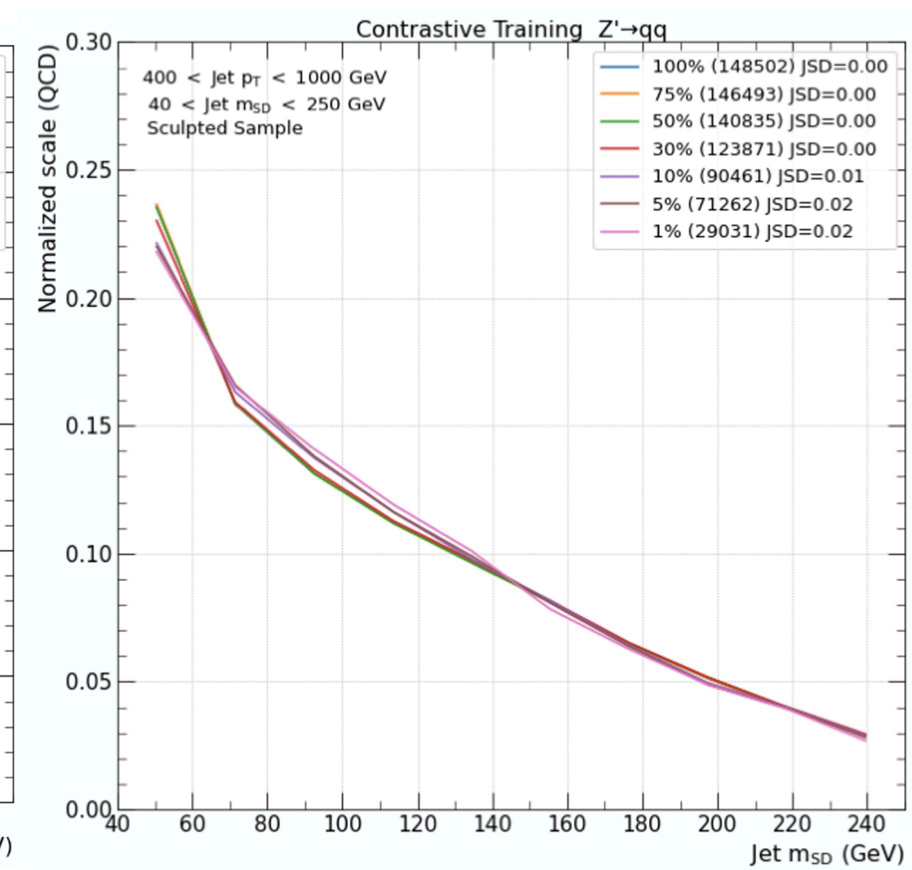
Signal to Noise Ratio (SNR)

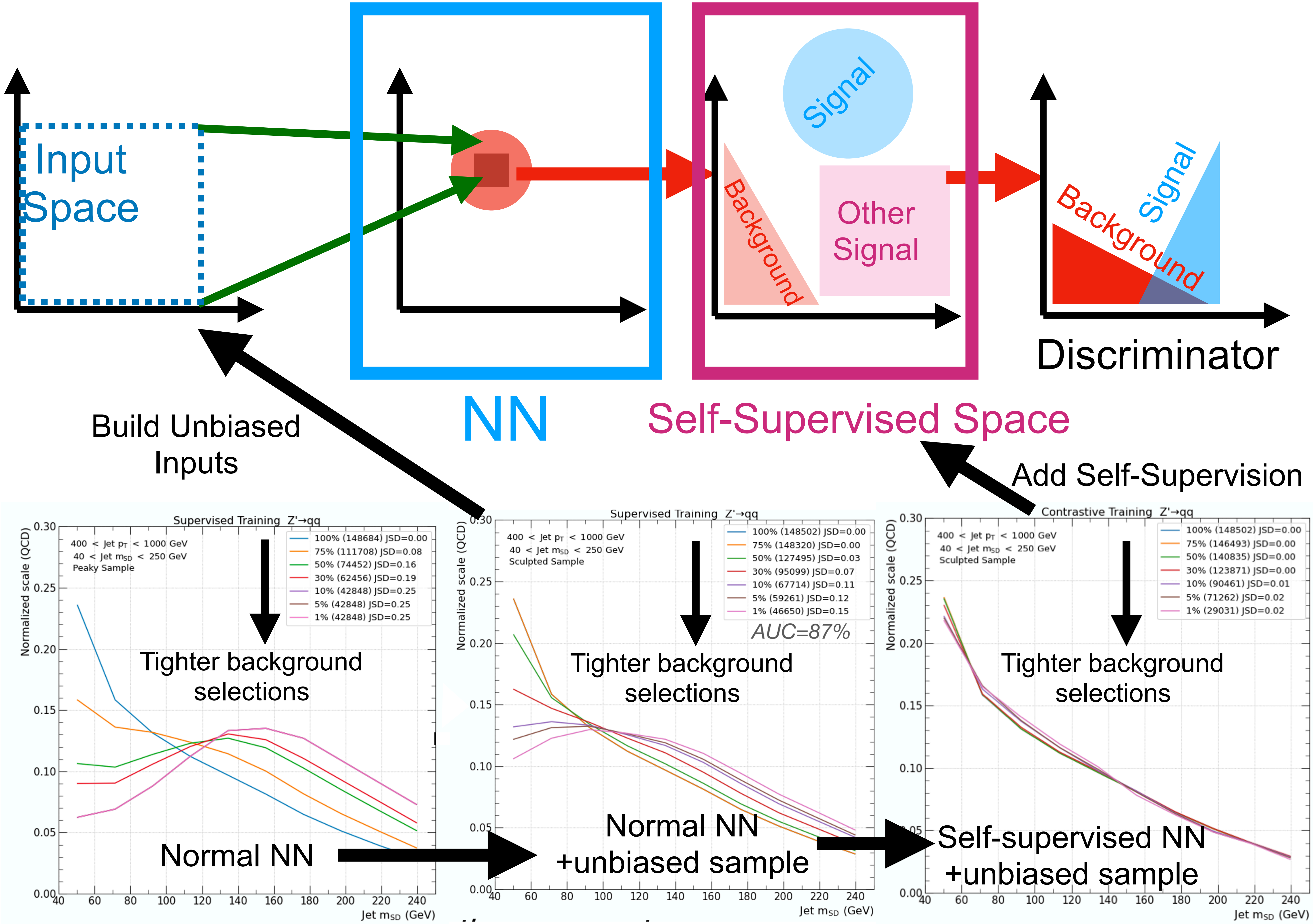
https://arxiv.org/abs/1603.00027



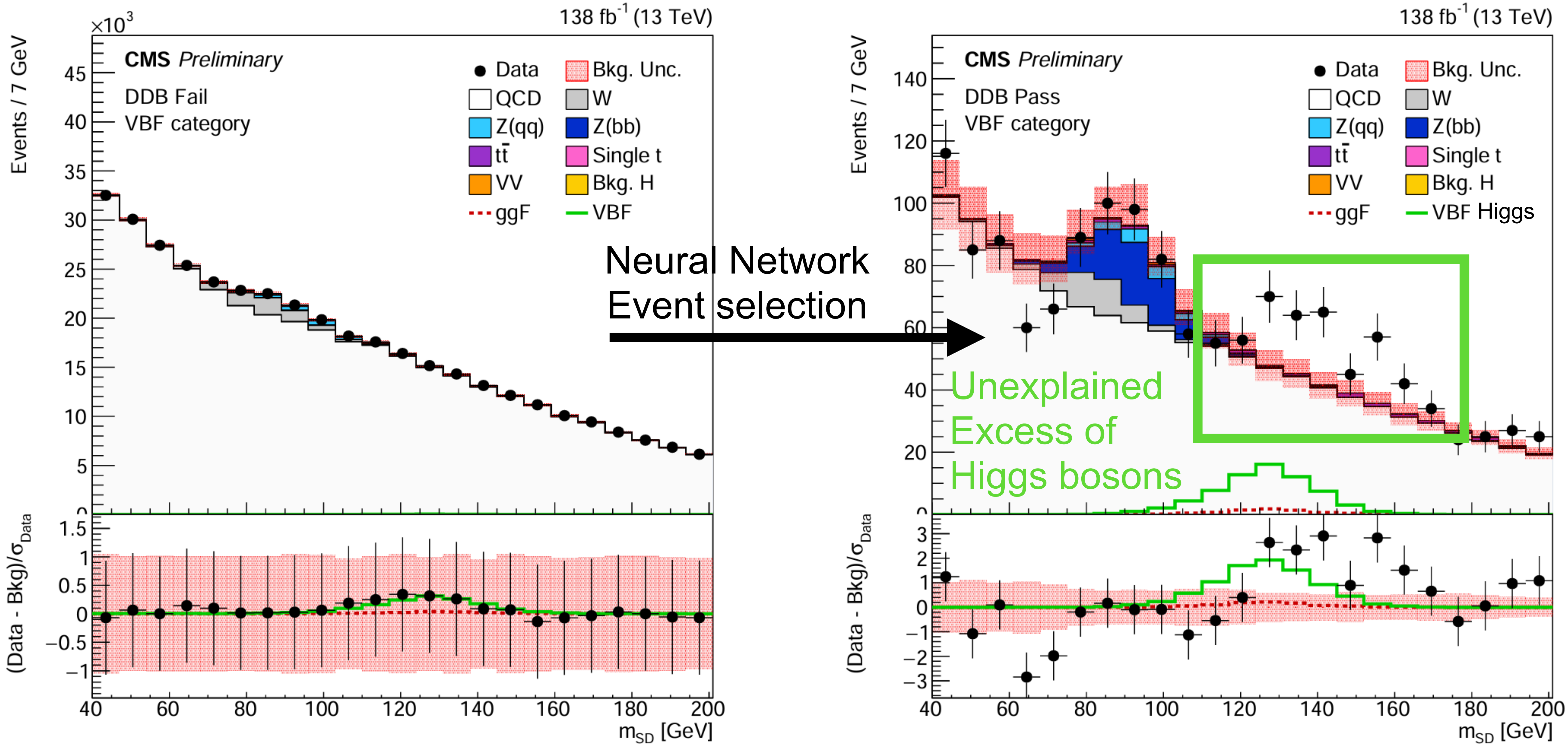








Input Space
NN
Signal
Background
Other Signal
Self-Supervised Space
Background
Signal
Discriminator
Build Unbiased Inputs
Add Self-Supervision
Supervised Training Z'→qq
400 < Jet p_T < 1000 GeV
40 < Jet m_{SD} < 250 GeV
Peaky Sample
100% (148684) JSD=0.00
75% (111708) JSD=0.08
50% (74452) JSD=0.16
30% (62456) JSD=0.19
10% (42848) JSD=0.25
5% (42848) JSD=0.25
1% (42848) JSD=0.25
Tighter background selections
Normal NN
Normalized scale (QCD)
Jet m_{SD} (GeV)
Supervised Training Z'→qq
400 < Jet p_T < 1000 GeV
40 < Jet m_{SD} < 250 GeV
Sculpted Sample
100% (148502) JSD=0.00
75% (148320) JSD=0.00
50% (127495) JSD=0.03
30% (95099) JSD=0.07
10% (67714) JSD=0.11
5% (59261) JSD=0.12
1% (46650) JSD=0.15
AUC=87%
Tighter background selections
Normal NN +unbiased sample
Normalized scale (QCD)
Jet m_{SD} (GeV)
Contrastive Training Z'→qq
400 < Jet p_T < 1000 GeV
40 < Jet m_{SD} < 250 GeV
Sculpted Sample
100% (148502) JSD=0.00
75% (146493) JSD=0.00
50% (140835) JSD=0.00
30% (123871) JSD=0.00
10% (90461) JSD=0.01
5% (71262) JSD=0.02
1% (29031) JSD=0.02
Tighter background selections
Self-supervised NN +unbiased sample
Normalized scale (QCD)
Jet m_{SD} (GeV)



138 fb^{-1} (13 TeV)
$\times 10^3$
CMS Preliminary
DDB Fail
VBF category
Data
Bkg. Unc.
QCD
W
Z(qq)
Z(bb)
$t\bar{t}$
Single t
VV
Bkg. H
ggF
VBF
Events / 7 GeV
(Data - Bkg)/σ_{Data}
m_{SD} [GeV]
Neural Network
Event selection
138 fb^{-1} (13 TeV)
CMS Preliminary
DDB Pass
VBF category
Data
Bkg. Unc.
QCD
W
Z(qq)
Z(bb)
$t\bar{t}$
Single t
VV
Bkg. H
ggF
VBF Higgs
Unexplained
Excess of
Higgs bosons
Events / 7 GeV
(Data - Bkg)/σ_{Data}
m_{SD} [GeV]

Hermes: Inference -as-a service deployment

https://github.com/ML4GW/hermes

Export → Acceleration → Deployment → Inference

Export
- Managing model repository
- Pythonic interfaces to protobuf configs
- Simple support for stateful streaming models
- Supports Torch and TensorFlow export

Acceleration
- Conversion of Torch models to ONNX
- ONNX → TensorRT conversion with FP16 support

Deployment
- Python contexts for deploying a local inference service
- Throughput and latency metrics monitoring service

Inference
- Asynchronous inference request submission and response handling
- Input/output shape/dtype inference

Hermes is our python package for end to end inference-as-a service

Computing Ecosystem

- Ray: Handle distribution of CPU brokers for data
- PyTorch Lighting: Optimized training and model development
- Apptainer: containerization to enable maximum flexibility
- Luigi: lightweight task execution, Kubernetes support contrib (Spotify)
- LAW: Wrapper for Luigi to enable Condor/Slurm support (HEP)
- Kubernetes for distribued server balance

Open source tools some from industry

ML in GW Processing

Online

Real-time analysis with goal of alerting electromagnetic astronomers (MMA) of significant events

Detect events → Localize on Sky → Send public alerts

Main consideration is ***latency***

Potential Future Use Case

Offline

Large scale analysis of archival data for

- End to end searches
- Validating new methods, performing new research

Main consideration is ***throughput***

What we want now w/Nautilius

Nautilus