NSF HDR ML Challenges

P. Harris (MIT), S.C. Hsu (UW), M. Neubauer(UIUC)
S.Chakrabaty(JPL),A. Gangopadhyay(UMBC),D. Matteson (Cornell),C. Stewart(RPI), S. Wang(UIUC)
HDR Community

What is the NSF HDR program ?

- NSF's Harnessing the Data Revolution (HDR) Big Idea is a national-scale activity to enable new modes of data-driven discovery that will address fundamental questions at the frontiers of science and engineering
- There are 5 NSF HDR institutes

A3D3: NSF HDR aimed at deploying real-time machine learning into scientific domains

iHARP: NSF HDR Institute for Harnessing Data and Model Revolution in the Polar Regions

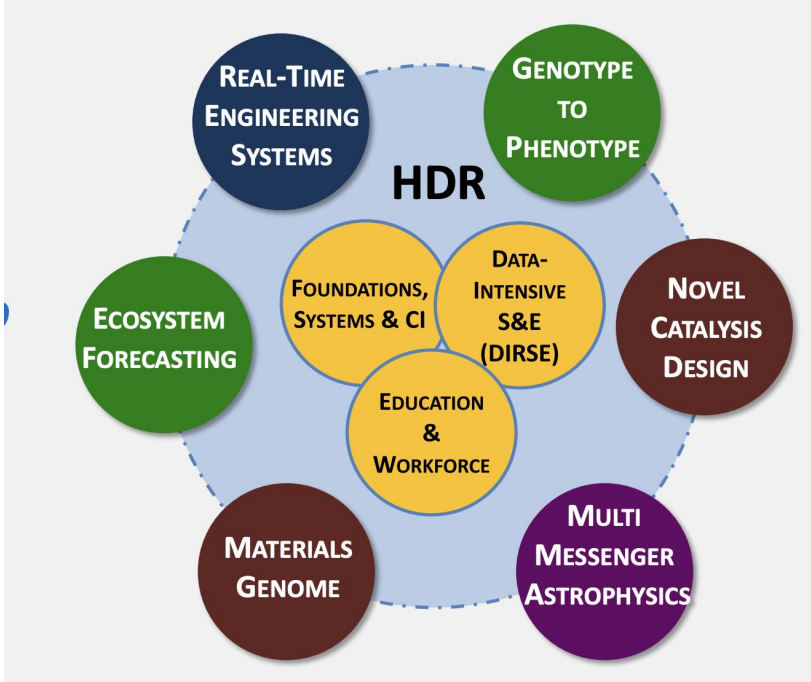
ID4: Institute for Data Driven Dynamical Design

iGuide: Institute for Geospatial Understanding through an Integrative Discovery Environment

A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning

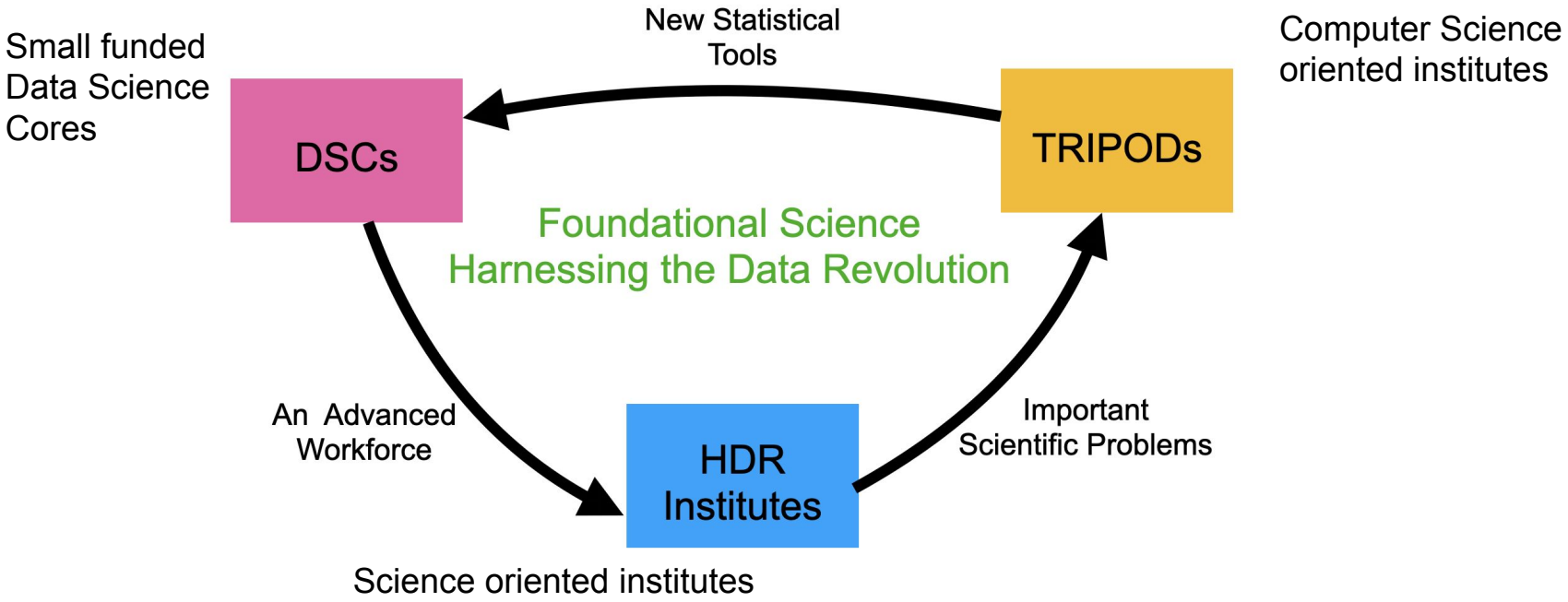
How HDR operation is envisioned

- HDR Aims to build a strategy around data science research



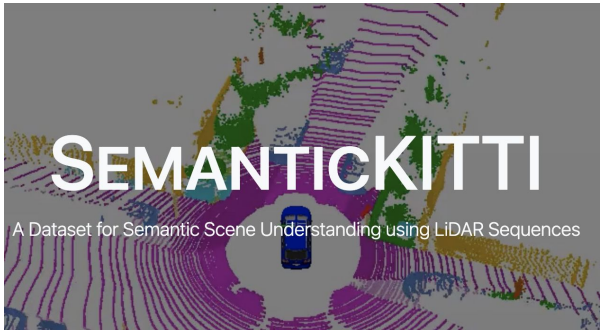
Scope of HDR

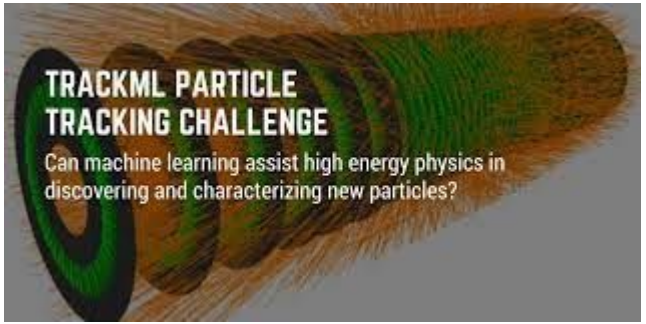
- The goal is to establish convection between institutes



What are ML Challenges?

- Aim is to make a series of datasets released to public
  - Use this dataset to make an ML Challenge
- Many datasets exist within the common use
  - This is our opportunity to align datasets with Science
  - Need to make it clear datasets have a different & special input
- Push community to think beyond the standards (below)





IMAGENET

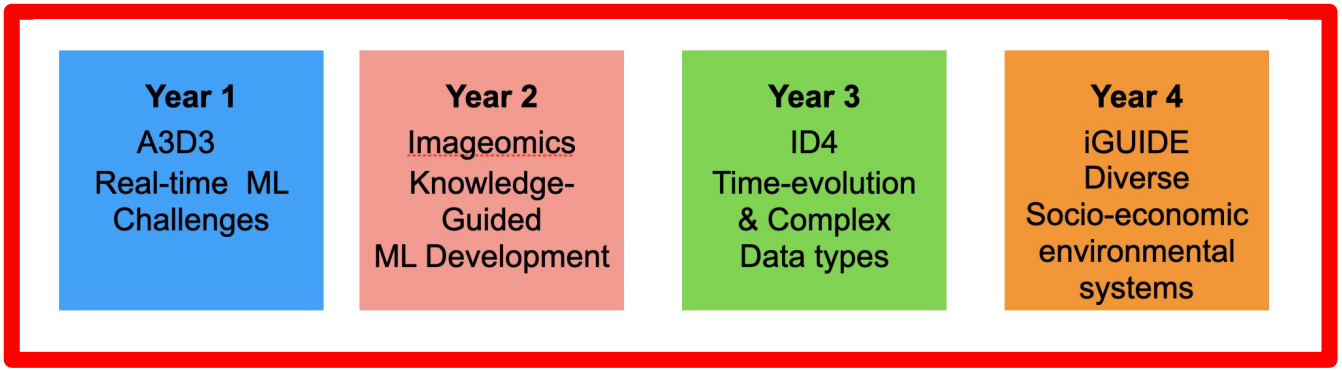
Advantages of ML Challenges

- A good way to advertise important scientific problem
  - Bring awareness to challenges within a field
  - Extends scope of problems beyond conventional ones
- A good way to expand FAIRness in community
  - ML Challenge datasets should aim to be FAIR
- A way to engage members outside of field/academia
  - Members of industry
  - Other scientific domains
- New and creative solutions to important scientific problems

HDR ML Challenges

- Goals is to perform a yearly ML challenge
  - For each ML challenge, we prepare FAIR datasets
  - Also prepare a directed goal for each of the datasets
  - Plan: release challenge, then award ceremony/conference
  - Aim is to focus each challenge on the institute topic
- Annual Bootcamp at UW to award results & have a tutorial

Suggested Plan

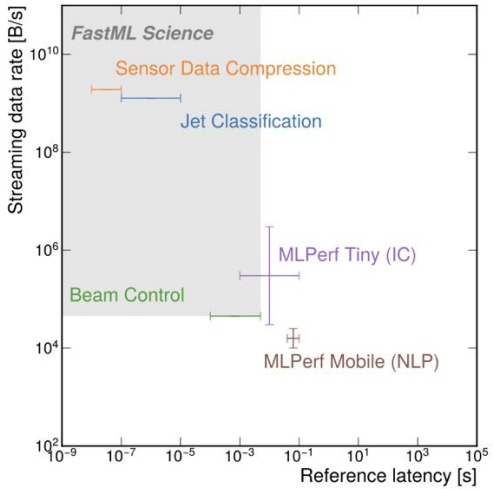


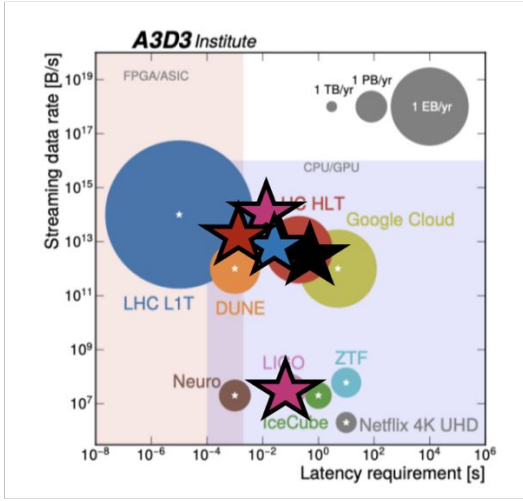
Cross HDR discussion

- We would like to connect with the work done within MLCommons Science
- These models can be potentially be integrated to metrics (such as MLPerf)
  - We will have a talk from MLPerf Tiny community later on
- More importantly : **This is a way to strengthen the HDR Community!**

ML Commons

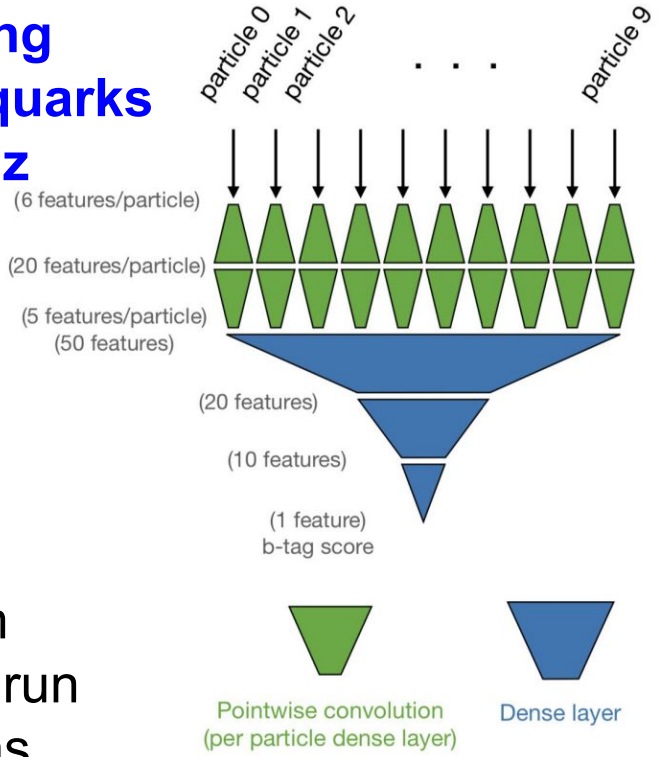
Machine learning innovation to benefit everyone.



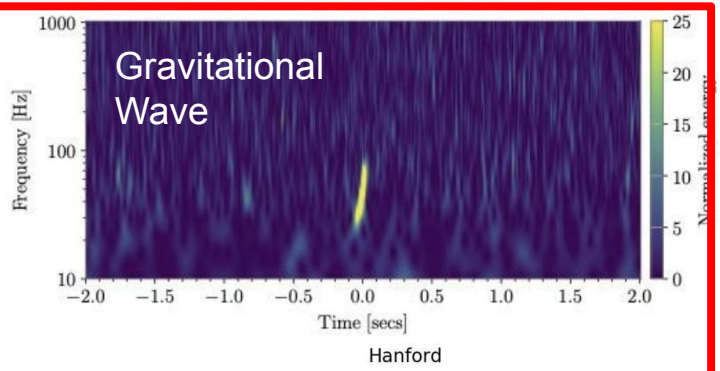


List of Possible Models

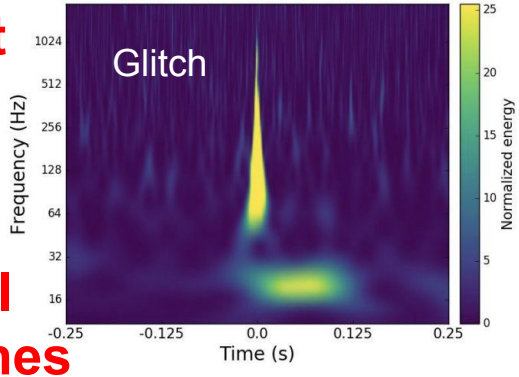
Identifying bottom quarks at 40 MHz



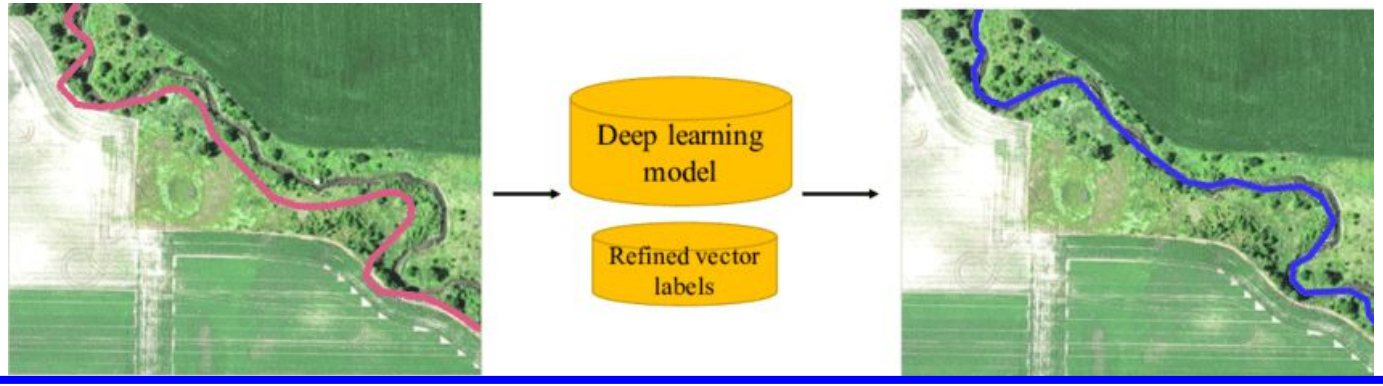
Algorithm needs to run in < 500ns



Building out dataset to generate+ identify gravitational waves glitches



List of Possible Models



Mapping out water resources from USGS surveys

Discovering power outages from atmospheric sensors.

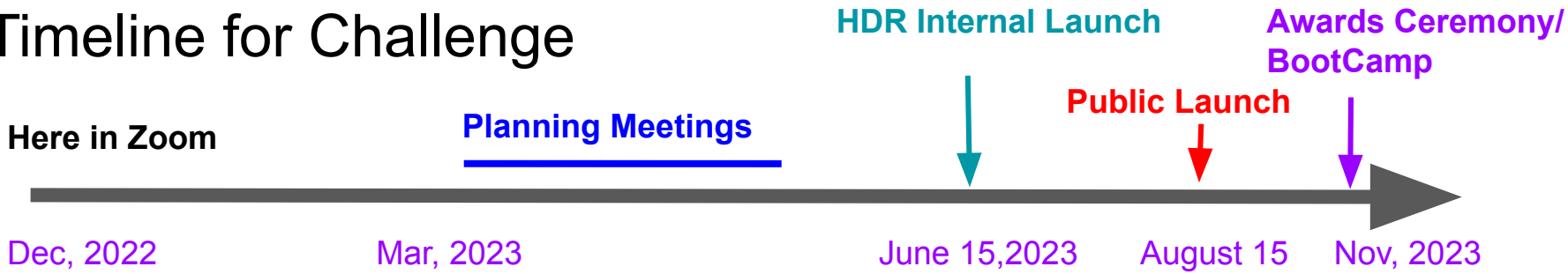
Where many of the sensors are propetary
So we can’t release challenge data

Except a small validation set

What needs to be done?

- We are lacking a clear framework for testing and validation
  - There are potentially a few options:
    - Hugging Face
    - https://www.modelshare.org/
    - Perhaps will hear more from the next talk
- Go for a future framework should have capability for many metrics
  - Basic research performance for minimizing loss
  - Low latency performance for checking optimized timing
  - Potentially these can evolve over time
- Aiming to have regular working group meetings
  - Would like as a goal for first ML Challenge meeting of before December
- Plan to have an internal ML challenge in summer (June 1st)
- Ultimate goal is to release ML Challenge by Fall (September 1st)

Timeline for Challenge



We aim for a one year timeline to setup the challenge to ceremony

We would like to engage i**ndustry partners to help sponsor our challenge**

- We can potentially release this on Kaggle or something else
- The other speakers here can really help push our plans forward

We support Awards ceremony/Bootcamp for the Challenge winners & for training

Summary

- Through the HDR community we are looking to make ML challenges
  - We see ML challenges as a way to promote science
  - This is also a way to curate a bunch of good data sets
    - New and interesting problems
- We have support to host an awards ceremony and develop challenge
  - Would need sponsor for monetary award of challenge
- Would like to engage several related communities
  - MLCommons science seems like a good way
  - Will hear from FAIRUniverse about hosting these