Accelerated AI Algorithms for Data-Driven Discovery

A3D3 in NSF HDR

Scope

P. Harris

NSF HDR

- 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
  - This is large group of institutes from many different domains
  - A3D3 is an NSF HDR Institute
    - There are 4. other NSF HDR institutes

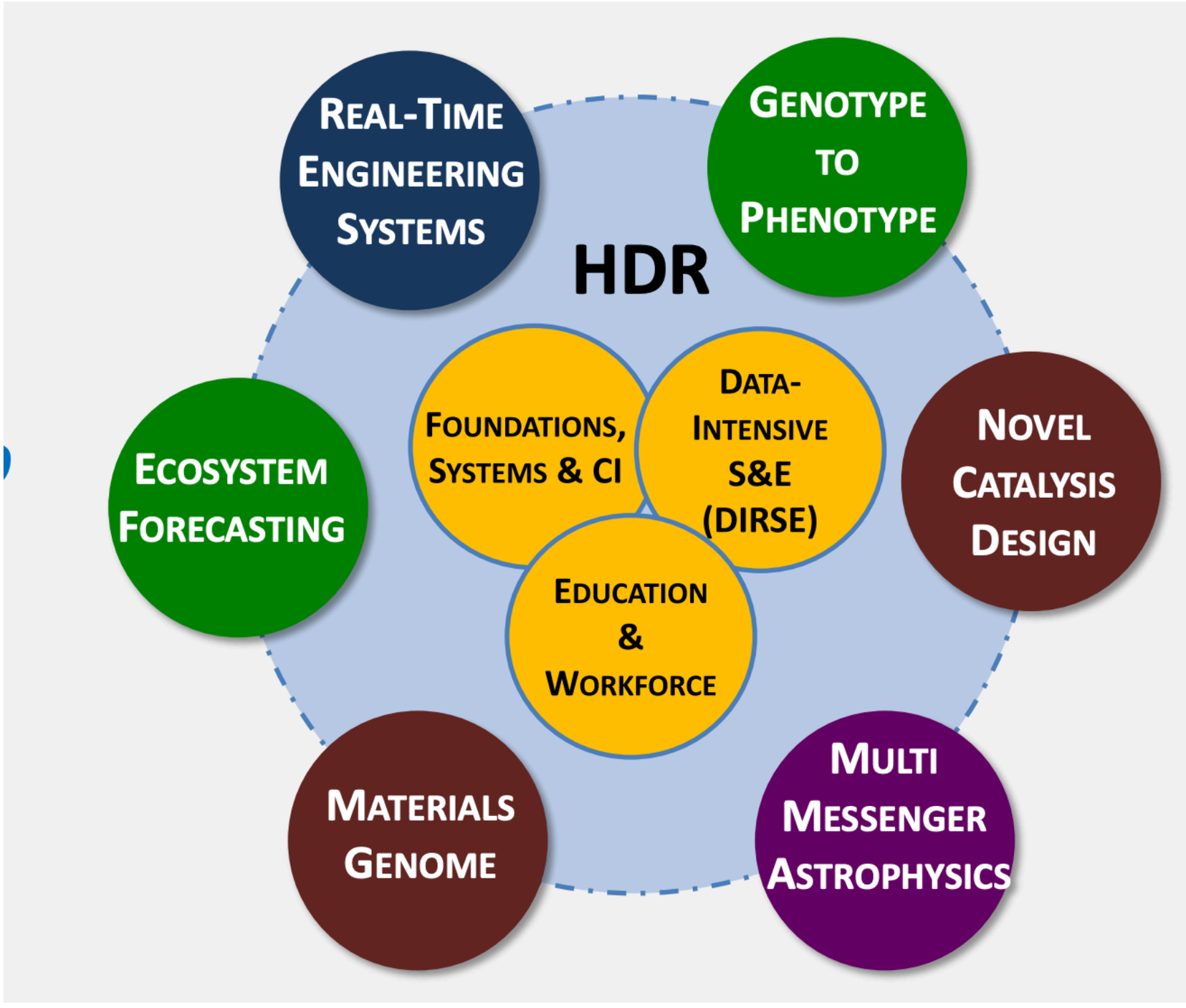
iHARP: NSF HDR Institute for Harnessing Data and Model Revolution in the Polar Regions

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

NSF HDR



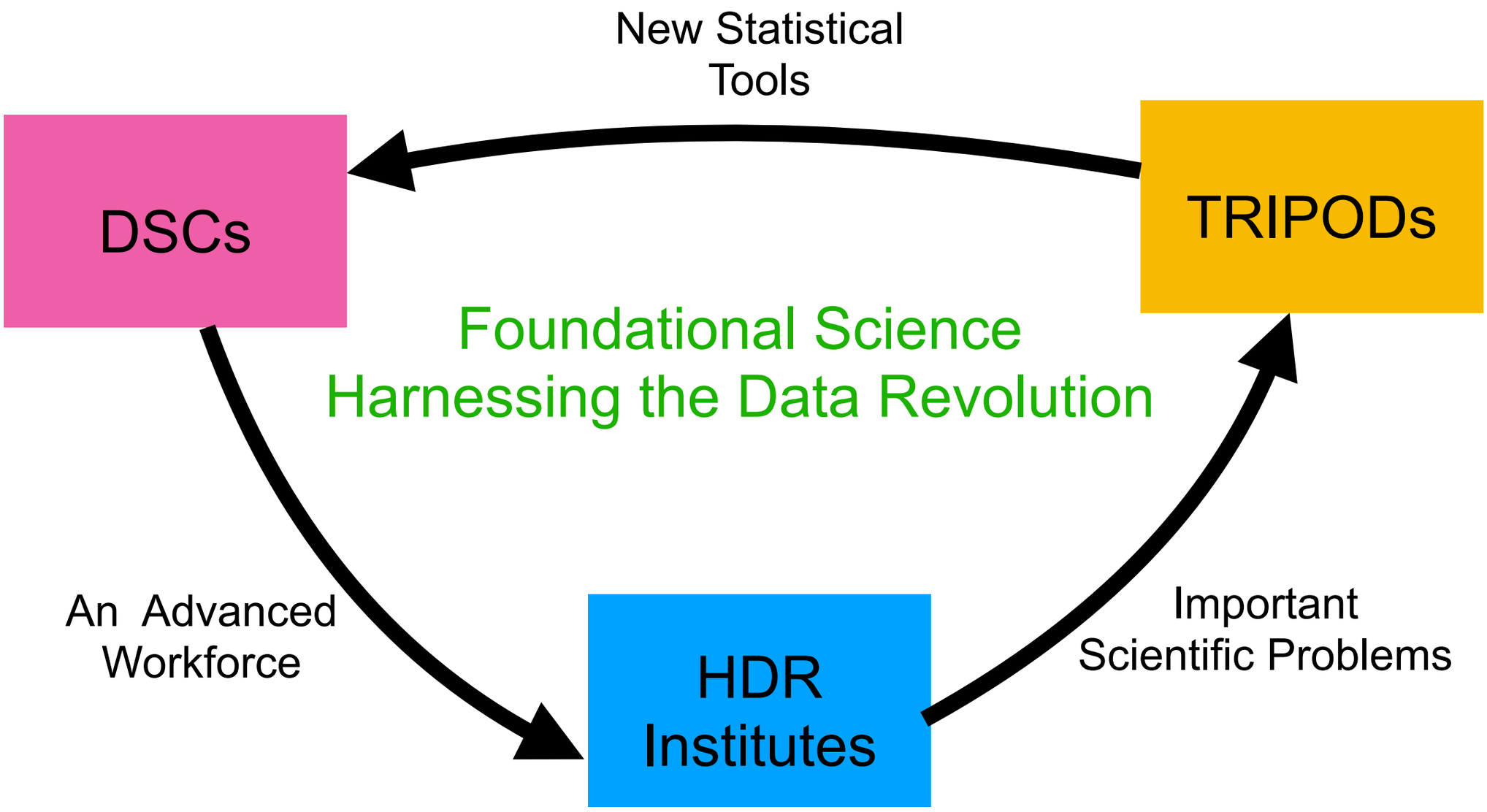
This is the NSF's Diagram for their view of the HDR

Other NSF HDR funds

- HDR Tripods:
  - Transdisciplinary Research in Principles of Data Science
  - Focuses on theoretical foundations of data science, core algorithmic, mathematical, and statistical principles
    - 2 Phase 2 institutes ($10M)
    - 15 Phase 1 smaller funds
- HDR DSCs:
  - *Data Science Corps* is one of the components of the HDR ecosystem enabling education and workforce development by focusing on building capacity for harnessing the data revolution at the local, state, and national levels
  - Roughly 15 Additional ones

Scope of HDR

- The goal is to establish convection between institutes



A3D3's Role Here

- In the spring
  - We were asked to submit a supplementary award
    - This award covers funding to connect HDR
  - As a result of this, we put together a short proposal
    - Aim of this proposal was to organize meetings
      - Supplemental funding twoward cross HDR efforts
    - We were the **only** institute awarded this money
      - However there might be other future opportunites
- Additional Money has been received this fall

What we proposed?

- 3 Core initiatives:
  - **HDR PI meetings** :
    - We are organizing the first HDR PI meeting in 3 weeks
  - **ML Challenges :**
    - Want to initiate a series of ML challenges
  - **Postbac enhancement** :
    - Would like to build on our existing program
    - Connect our work across the NSF HDR domains

Conferences

- In supplement, we proposed organizing HDR-wide

HDR Ecosystem Startup Meeting in DC Oct 2022

Regular HDR PI Meetings Annual Zoom 2023-

Big HDR Conference Meeting UIUC 2024

- We are organizing the first HDR conference
  - "From Harnessing the data revolution to Harvesting the data revolution" https://indico.cern.ch/event/1174814/
  - We will organize a big all HDR (students,PDs,PIs) conference
    - Mark (UIUC) is leading the big conference effort

ML Challenges

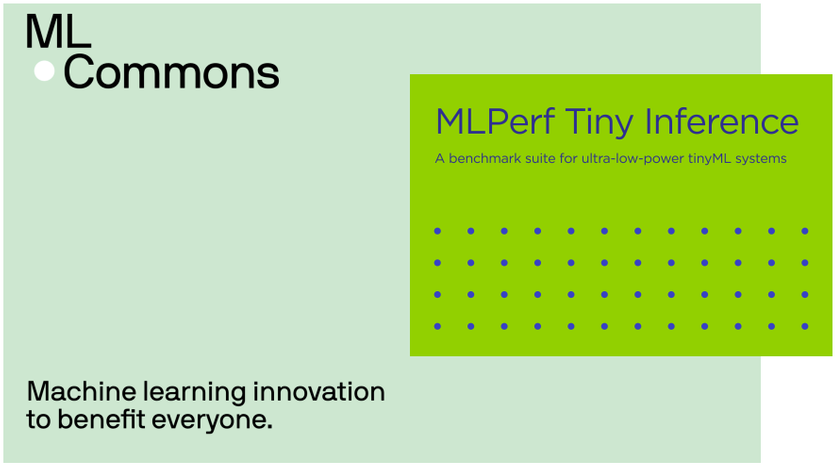
- Goals is to perform a yearly ML challenge
  - For each ML challenge, we prepare FAIR datasets
  - Also prepare a directed goal for each of the datasets
  - Plan: release challenge, then award ceremony/conference
  - Aim is to focus each challenge on the institute topic

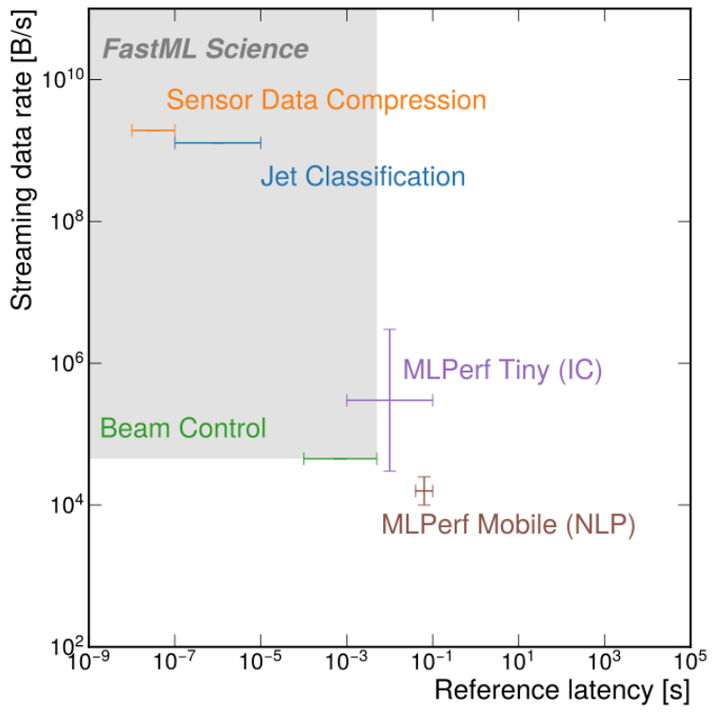
Year 1	**Year 2**	**Year 3**	**Year 4**
A3D3 Real-time ML Challenges	Imageomics Knowledge- Guided ML Development	ID4 Time-evolution & Complex Data types	iGUIDE Diverse Socio-economic environmental systems

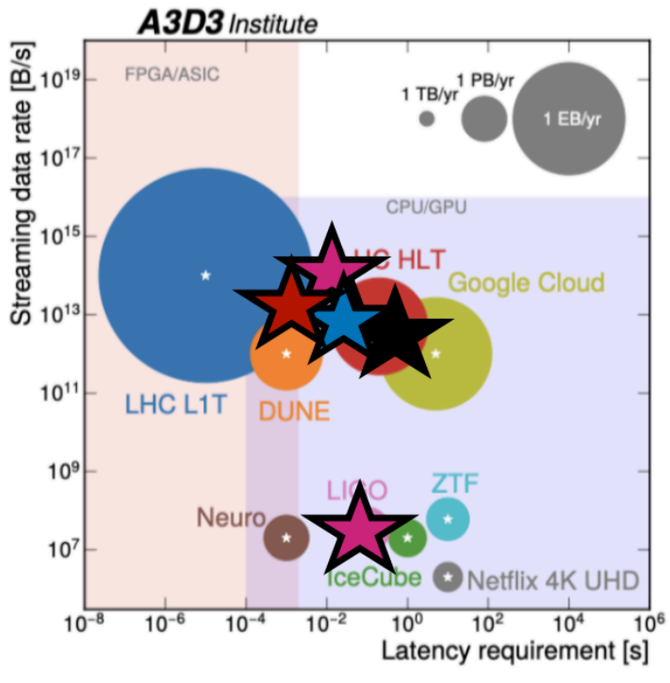
- Annual Bootcamp at UW to award results & have a tutorial

Idea for ML Challenges

- Following Dylan's talk, we have some idea of models
  - LHC tracking as a new benchmark
  - LIGO DeepClean as another benchmark
- From Jules' talk we have ideas for tiny models (resource constrainted)
  - ASIC based autoencoder/Jet dataset/Anomaly detection/RL Beam control
- Like to connect with larger MLPerf/MLCommons Science Community







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
  - Suggest that we do a first in house test in **Jan?**
    - 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?

Possible Idea

Neural Benchmarks

Public Challenge

MMA Bencharks

Public Challenge

HEP Bencharks

Public Challenge

FAIR4HEP

A3D3

A3D3 Internal Challenge

Public Challenge

To go from internal to public
We should seek Corporate
Sponsors

Have enough people in our
institute to build a pipeline

Postbac Improvements

- Our goal in extending the proposal was to:
  - Increase our recruitment of underrepresented minorities
    - Travel to conferences including Grace Hopper
    - Sponsor a booth at Grace Hopper
      - Join the recruitment database for various initiatives
  - Provide 2 day training session for incoming postbacs
    - Aim to cover AI/HPC/… relevant tools to HDR
    - Foster a community across the Postbacs
    - Potenitally collate these tutorials to online materials
- We have already initiated a multi-HDR effort towards this

A3D3 in Ecosystem

- Are the first institute to be awarded additional supplement
  - Perceive it as a good sign for future work from the institute
  - It would be good to leverage this position wisely
- Organizing the HDR meeting has been elucidating
  - Large community of data scientists
    - Lot of work on data science preparations/analysis
  - Our scientific problems are very compelling
    - Our computing work is equally compelling

Conclusions

- Big effort towards connecting HDRs
  - We have received funding to organize connecting events
  - Will help us to continue the effort over next few years
  - Hopefully will allow us to keep leadership in institutes
- Given we are leading this effort
  - We have the potential to guide problems and resources
  - We have compelling problems.
  - We can give more feedback following the PI meeting
- Would like to highlight ML challenge as a way of getting continuity