HDR ML Challenge

Detect Anomalies in Science with Machine Learning!



Carnegie Mellon University

imageomics BIO ML

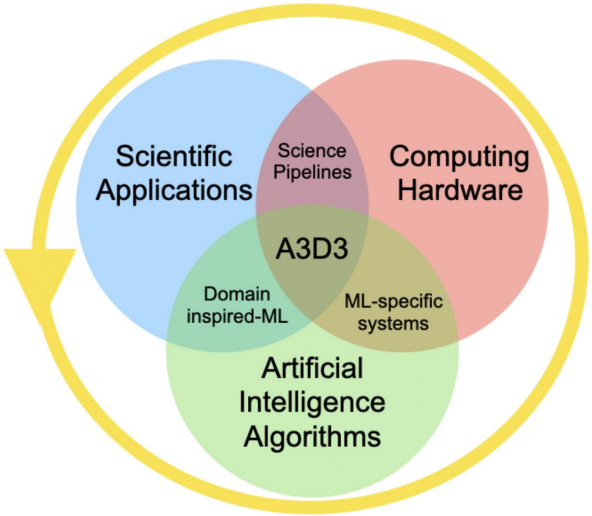
a3d3

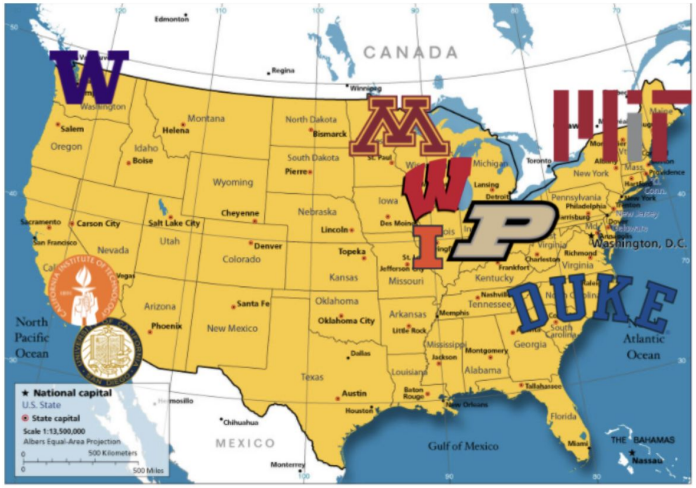
December 2nd, 2024

Harnessing the Data Revolution

- The HDR program aims to develop **AI solutions** to accelerate science & discovery
  - Funded by National Science Foundation (NSF)
  - Active, multi-disciplinary collaboration via the:

    Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute





CMU only recently joined A3D3!

Harnessing the Data Revolution

- The HDR program aims to develop **AI solutions** to accelerate science & discovery
  - Funded by National Science Foundation (NSF)
  - Active, multi-disciplinary collaboration via the:

    Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute

- **A3D3 challenges YOU** to develop Machine Learning models for **anomaly detection!**
  - Detect hybrid butterfly species
  - Find unmodeled gravitational waves
  - Detect unusual fluctuations in water levels

- A solid prize pool, including funded invitations to the AAAI 2025 conference!
  (Challenge last until 2025/01/17)

Prize Pool for HDR ML Challenges

- 💰 Total cash prizes $2500
- 👩‍💻 $3000 in AWS cloud computing credits
- 🎖️ Extra award sponsored by AMD (details pending)
- 🎖️ Special jury prizes include funded invitations to AAAI 2025

Potential for additional award.

Harnessing the Data Revolution

Agenda

- Harnessing the Data Revolution

- Introduction to the challenges
  - Butterfly Hybrid Detection
  - Sea Level Anomaly detection
  - Detecting Anomalous Gravitational Wave

- Model Submission Platform

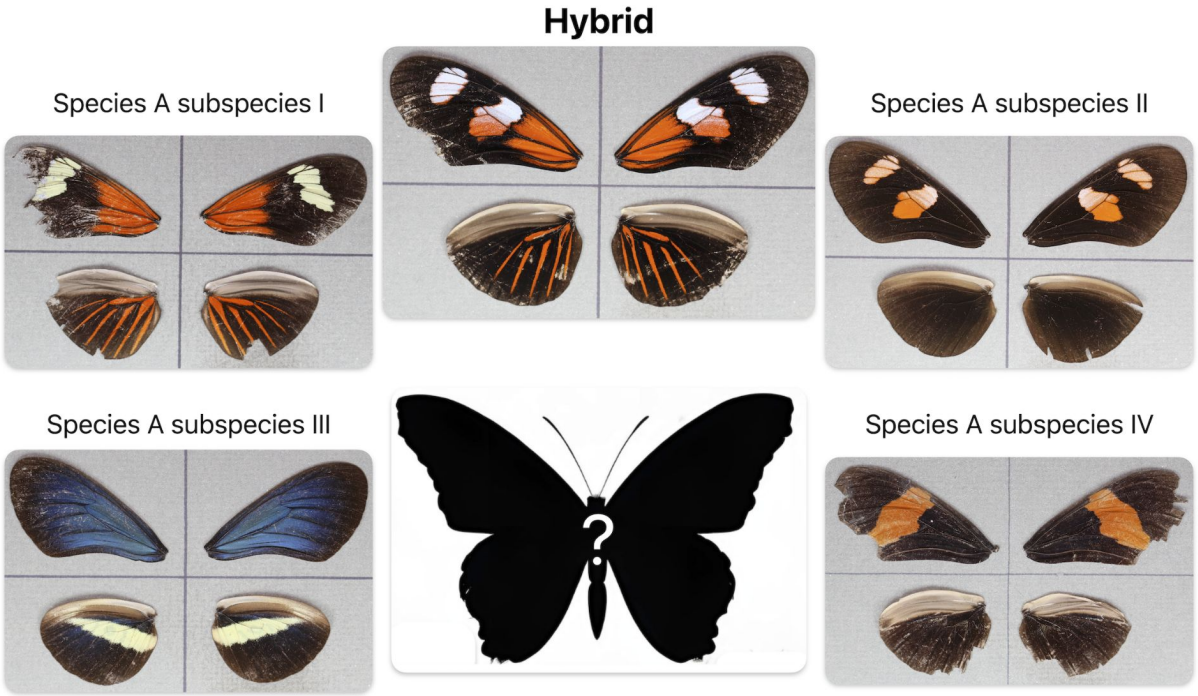
- References & Practicalities

The Challenges

Butterfly Hybrid Detection

- Hybridization may lead to a variety of resulting butterfly wing patterns
  - Identifying hybrids requires knowledge of their parent species/subspecies

- Can ML models automatically identify (unseen) hybrid cases?



Butterfly Hybrid Detection

- All information + a starter kit given in codabench

- **Training data** = ~2200 images of Species A
  - Includes:
    - Multiple ***sub***species.
    - Selected signal hybrids of two ***sub***species

- **Test/Dev data** = ~1100 images
  - Includes:
    - All Species A subspecies.
    - Signal hybrids from training data.
  - Further introduces:
    - Other Species A hybrids (non-signal).
    - Species B: Mimics of Species A signal hybrid parents (& their hybrids).

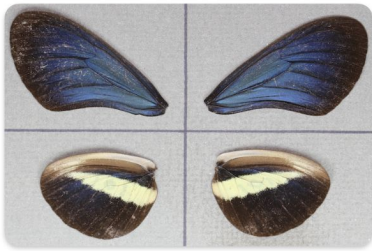
Butterfly Hybrid Detection

- Among Species A & B, can your algorithm find…
  - Species A signal hybrids?
  - Species A non signal hybrids?
  - Species B hybrids (mimics of Species A signal hybrids)?

Species A subspecies I

Species A subspecies II

Species A subspecies III

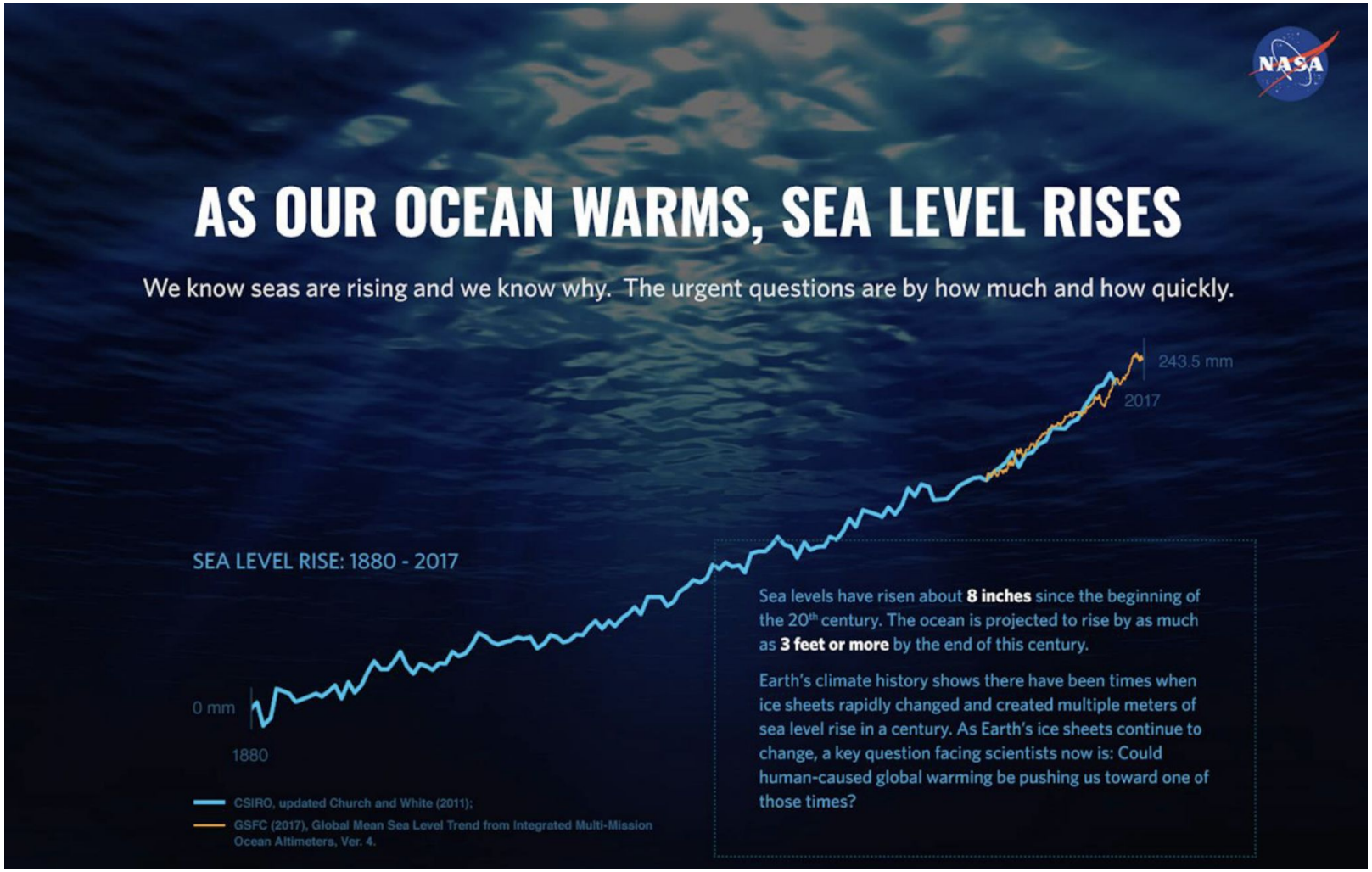


Species A subspecies IV

Species B subspecies II

Species B subspecies I

Sea-Level Anomaly Detection



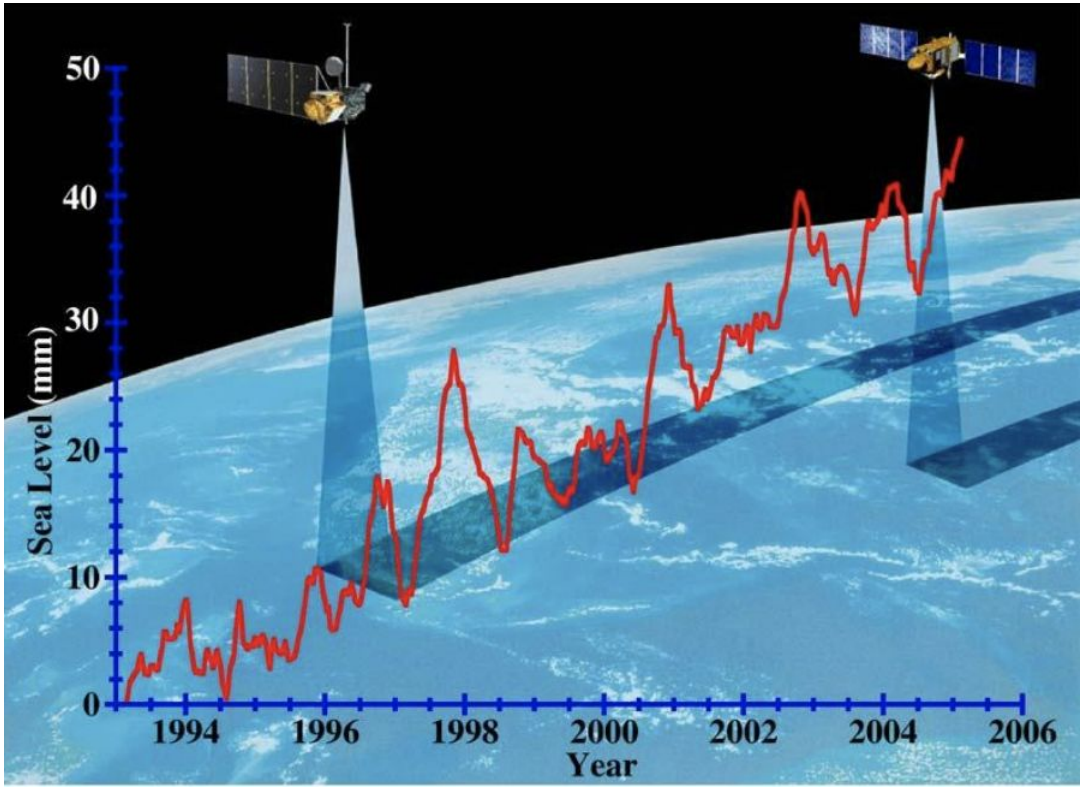
Sea-Level Anomaly Detection

SEA LEVEL RISE AFFECTS US ALL

More than **160 million people** live along coasts in the U.S., about half the nation's population. **Eleven of the world's 15 largest cities** lie along shores, including New York City. Sea level rise means the ocean will gradually inundate low-lying areas, and storms like hurricanes, bolstered by even higher seas, will extend their reach inland. All of society bears the burden for storm damage and those costs are expected to rise: Annual losses from flooding in the world's biggest coastal cities could rise from about **$6 billion a year** today to **$1 trillion a year** by 2050.

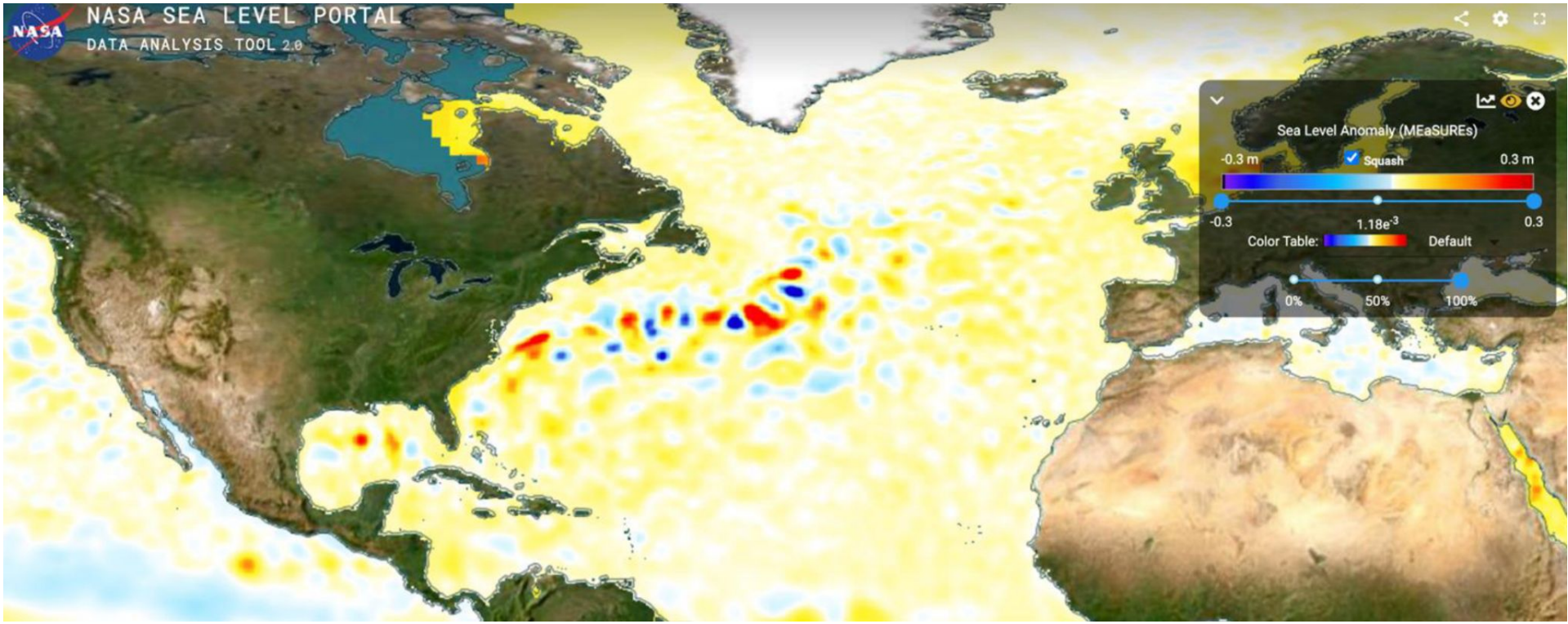
Sea-Level Anomaly Detection

- As the ocean rises, the ability to provide even more precise information about coastal sea level rise is crucial



Sea-Level Anomaly Detection

- **Goal:** Detect anomalous flooding events along the US East Coast with the maps of sea-level over the North Atlantic



Sea-Level Anomaly Detection

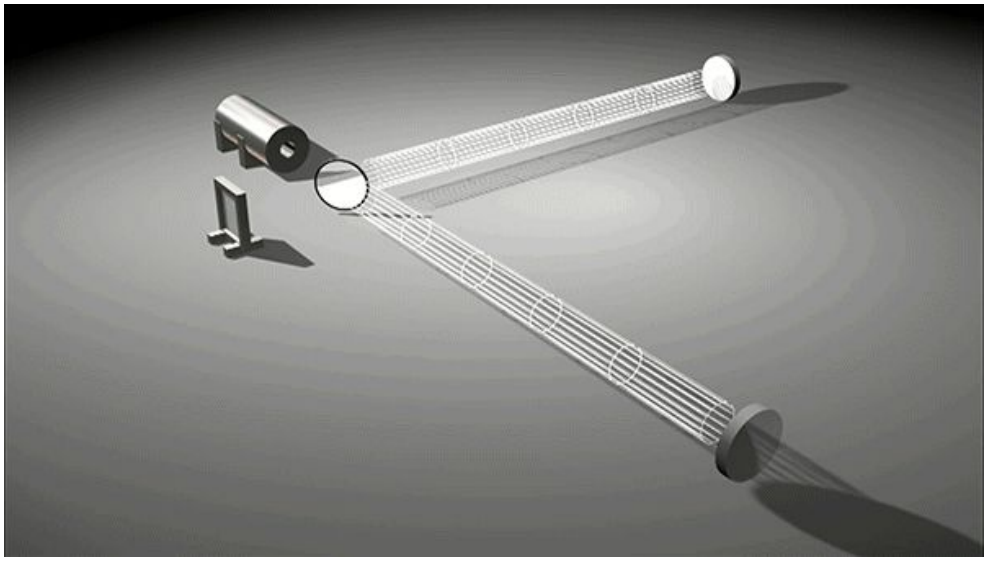
- **Goal:** Detect anomalous flooding events along the US East Coast with the maps of sea-level over the North Atlantic

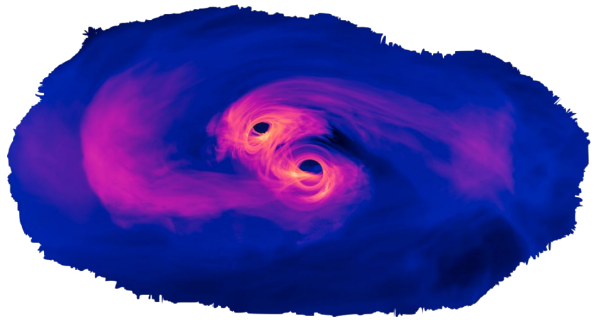
- All information + a starter kit given in codabench

- Data provided:
  - Satellite maps of sea level data over the North Atlantic for the past 30 years
  - Labeled Anomalous Flood: Dates of anomalous flooding along the US East Coast

Anomalous Gravitational Wave Detection

- Accelerating masses produce deformations in space-time that we can detect via interferometers
  - Multiple large-scale interferometers throughout the world (LIGO, VIRGO, …)
  - A signal will appear in at least two interferometers, with small time delay



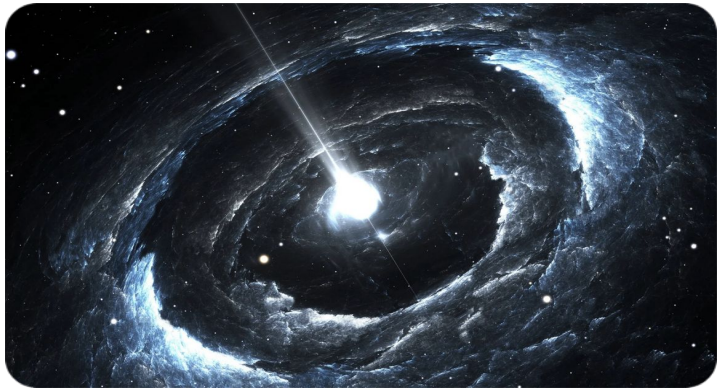


Anomalous Gravitational Wave Detection

- Accelerating masses produce deformations in space-time that we can detect via interferometers
  - Multiple large-scale interferometers throughout the world (LIGO, VIRGO, …)
  - A signal will appear in at least two interferometers, with small time delay

CORE-COLLAPSE SUPERNOVA

NEUTRON STAR GLITCHES

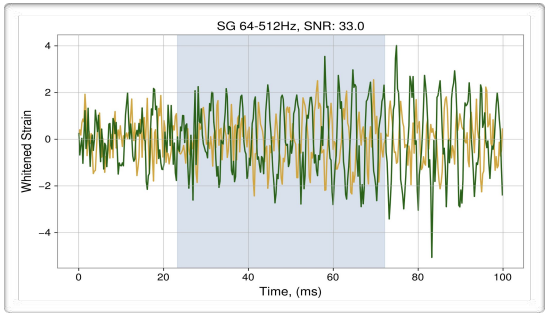


Known unknowns: We known they exist, but they are poorly modeled, so hard to detect!

Unknown unknowns: New anomalous GW sources we haven't thought off…

Anomalous Gravitational Wave Detection

- **Goal:** Design a ML model that identifies anomalous gravitational waves from (un)known unknowns

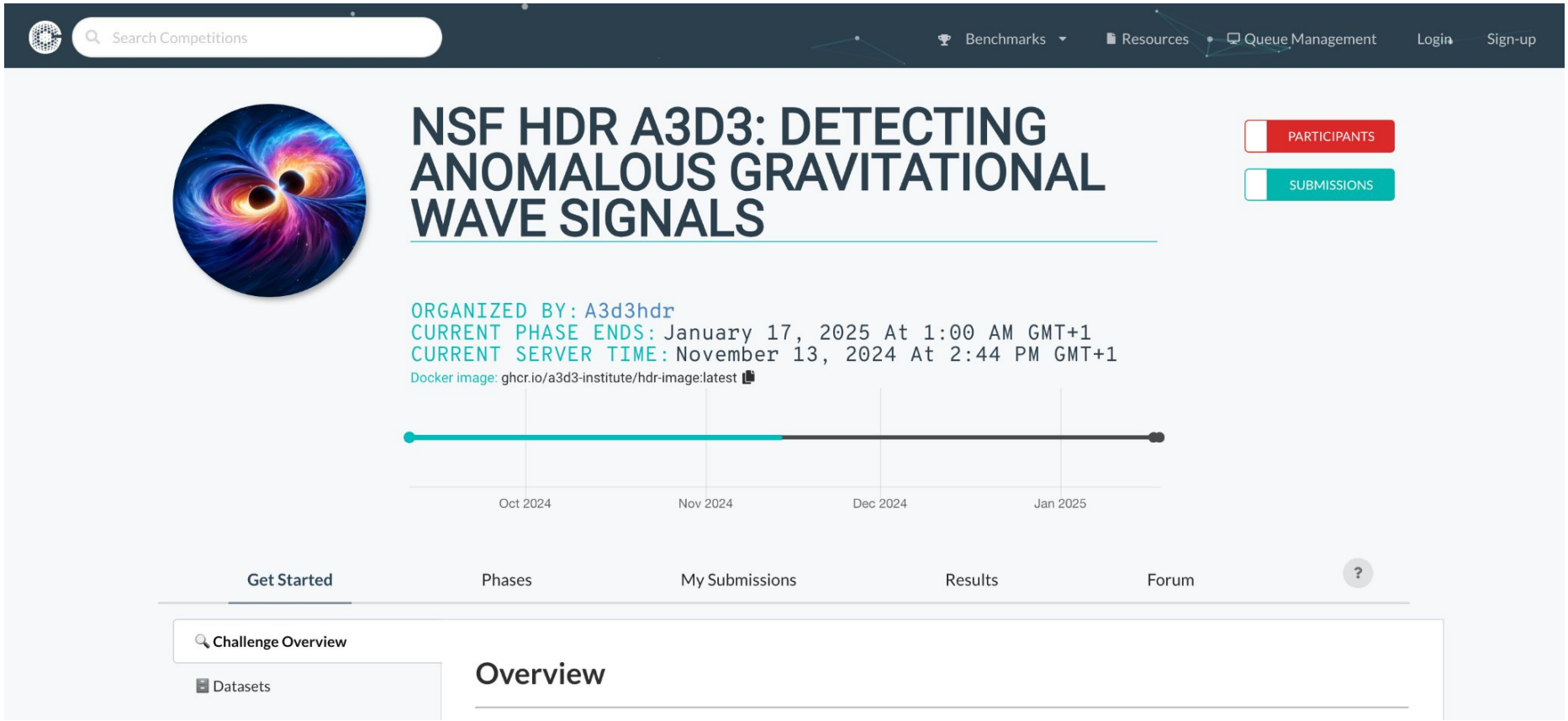


- All information + a starter kit given in codabench

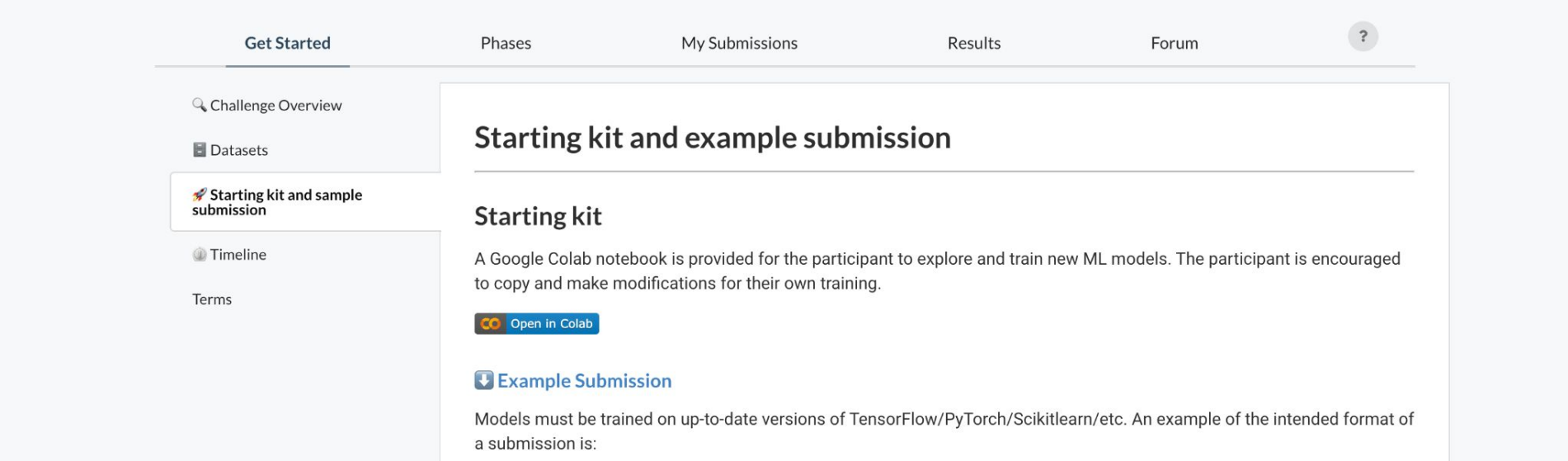
- Data provided = Continuous time series of interferometer data at 4096 Hz
  - Already pre-processed: whitened & band-passed (30<f<1500 Hz)
  - Divided into 50 ms segments, each containing 200 data points (50 ms * 4096 samples/second = 200 samples)
  - Dimension of input data is (N, 200, 2), with N representing the number of data segments. The last dimension of 2 corresponds to the datastreams from the two LIGO interferometers in **Hanford (WA)** & **Livingston (LA)**

Submission Platform

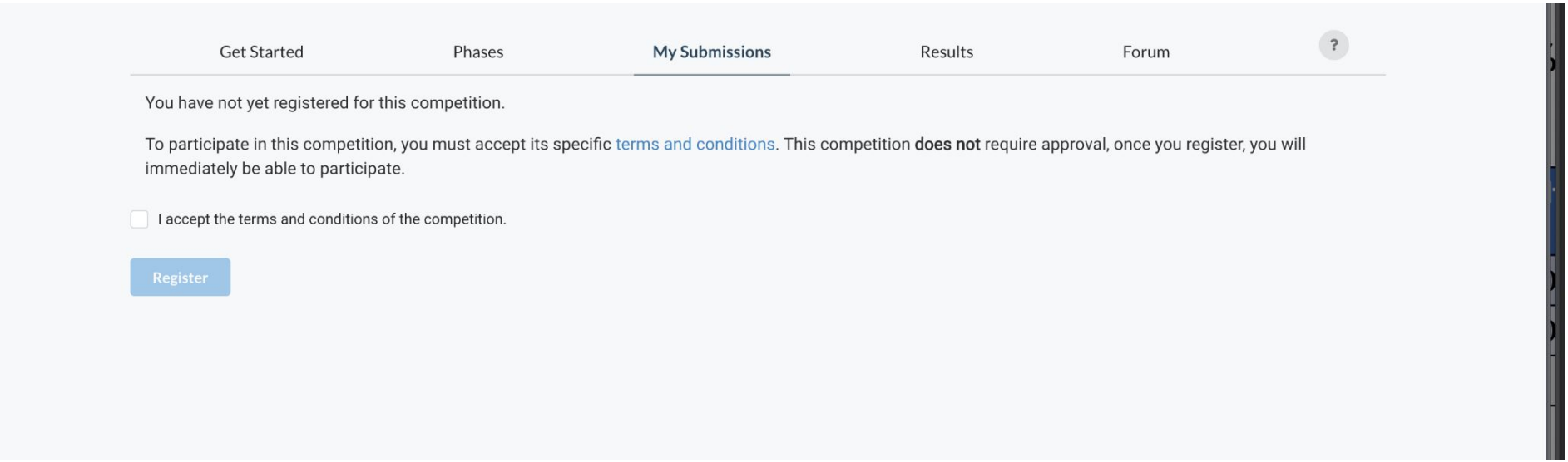
1. Login or Create Account on Codabench



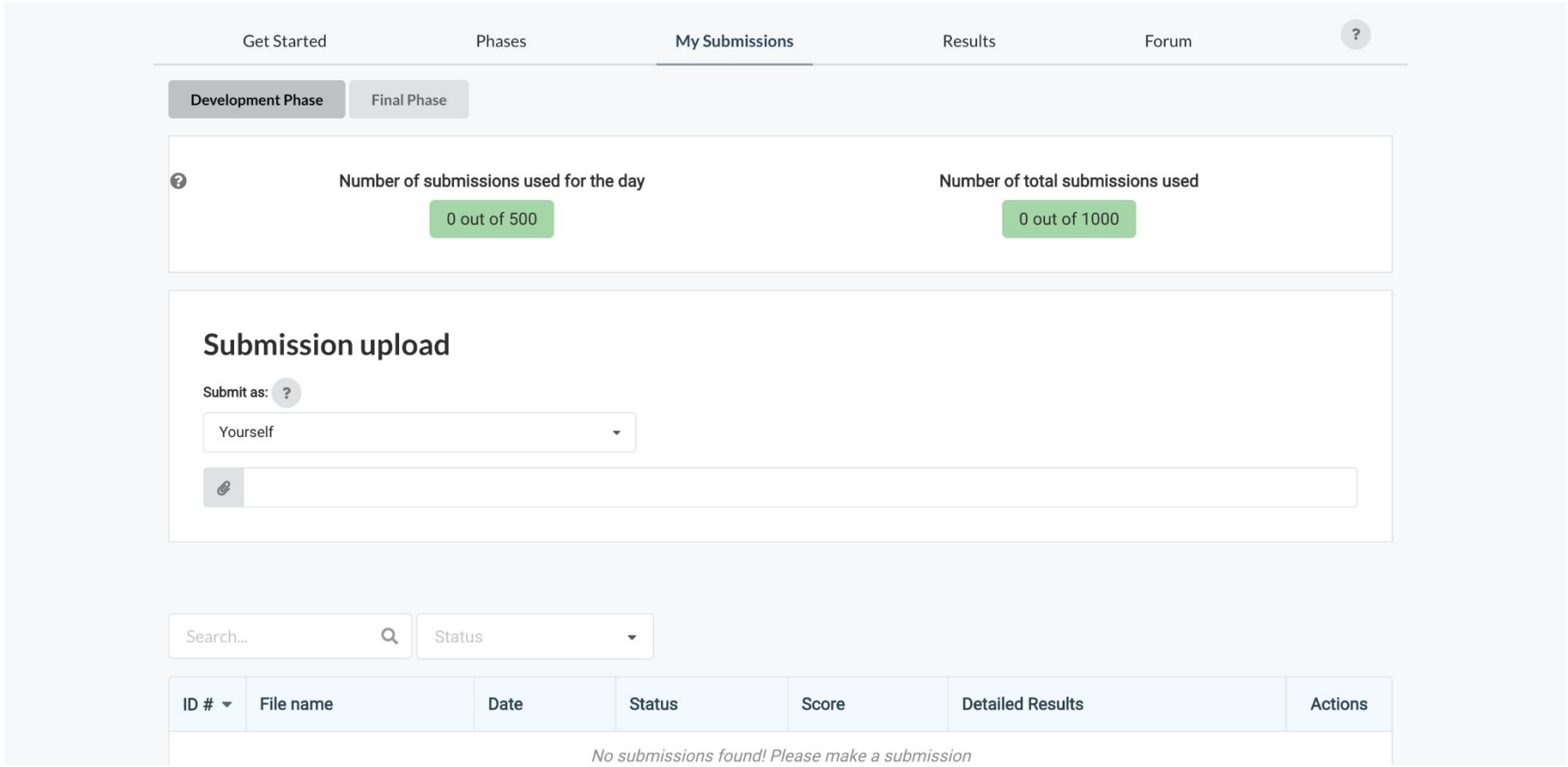
2. Download Dummy Submission



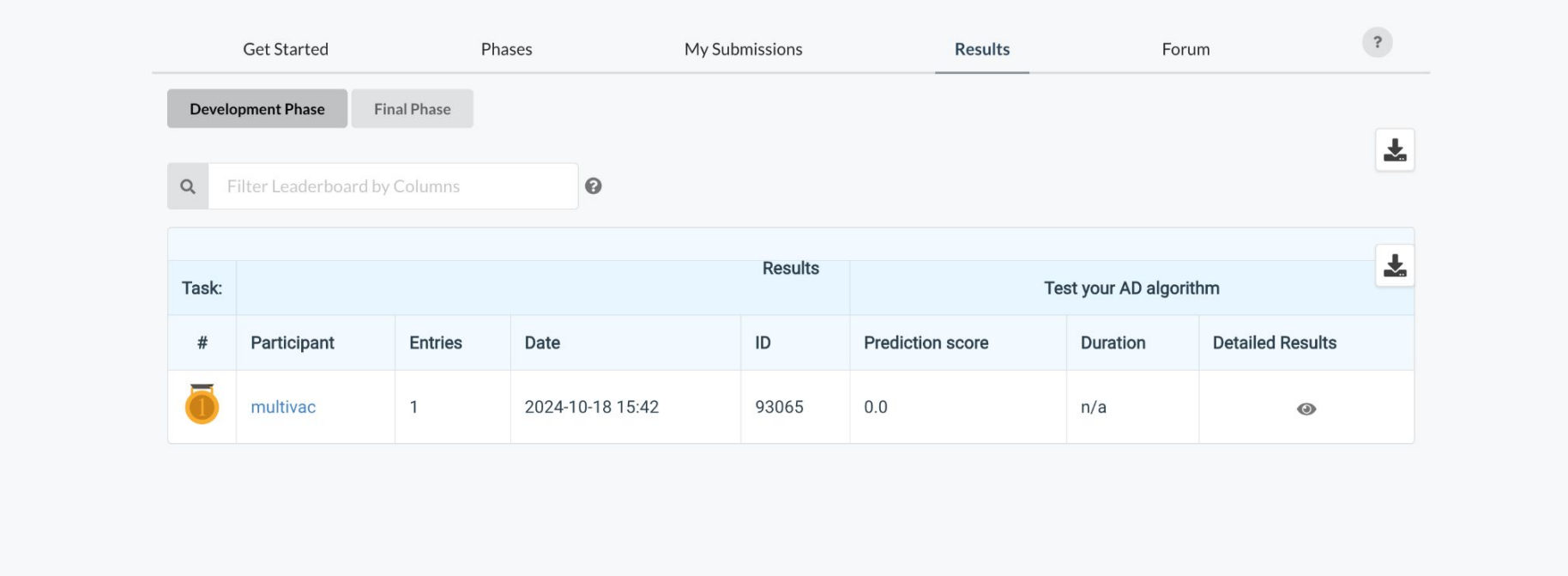
3. Register in the Competition



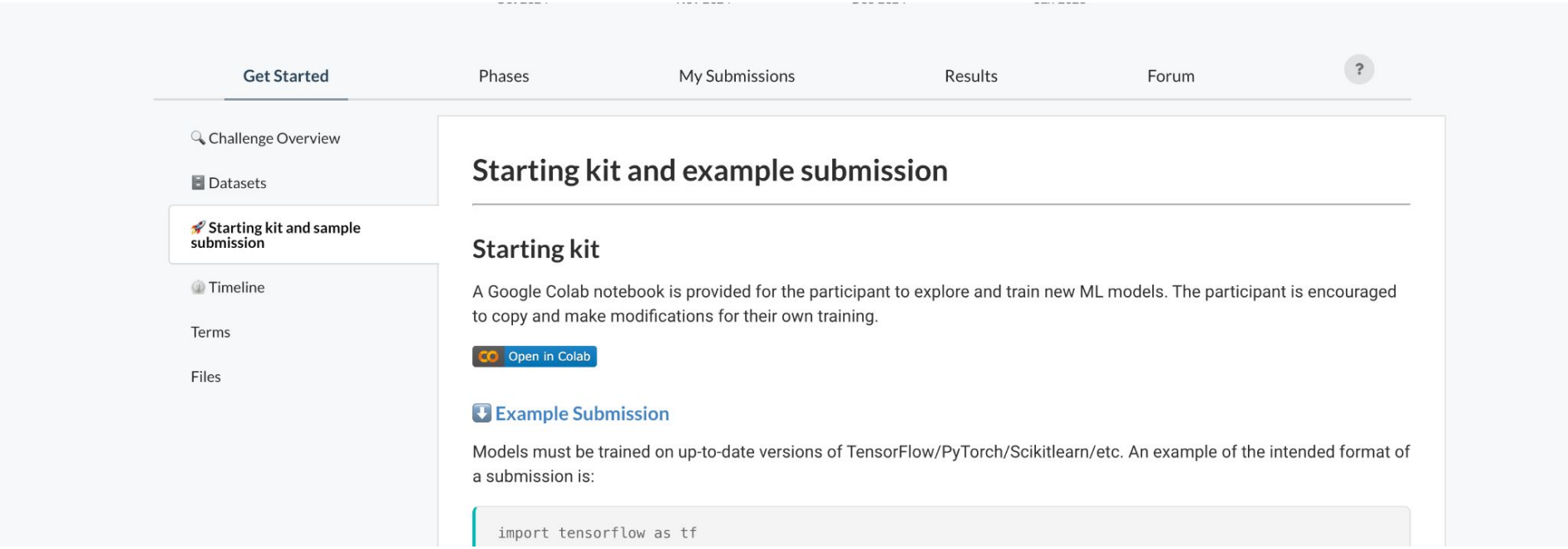
4. Submit Dummy Submission



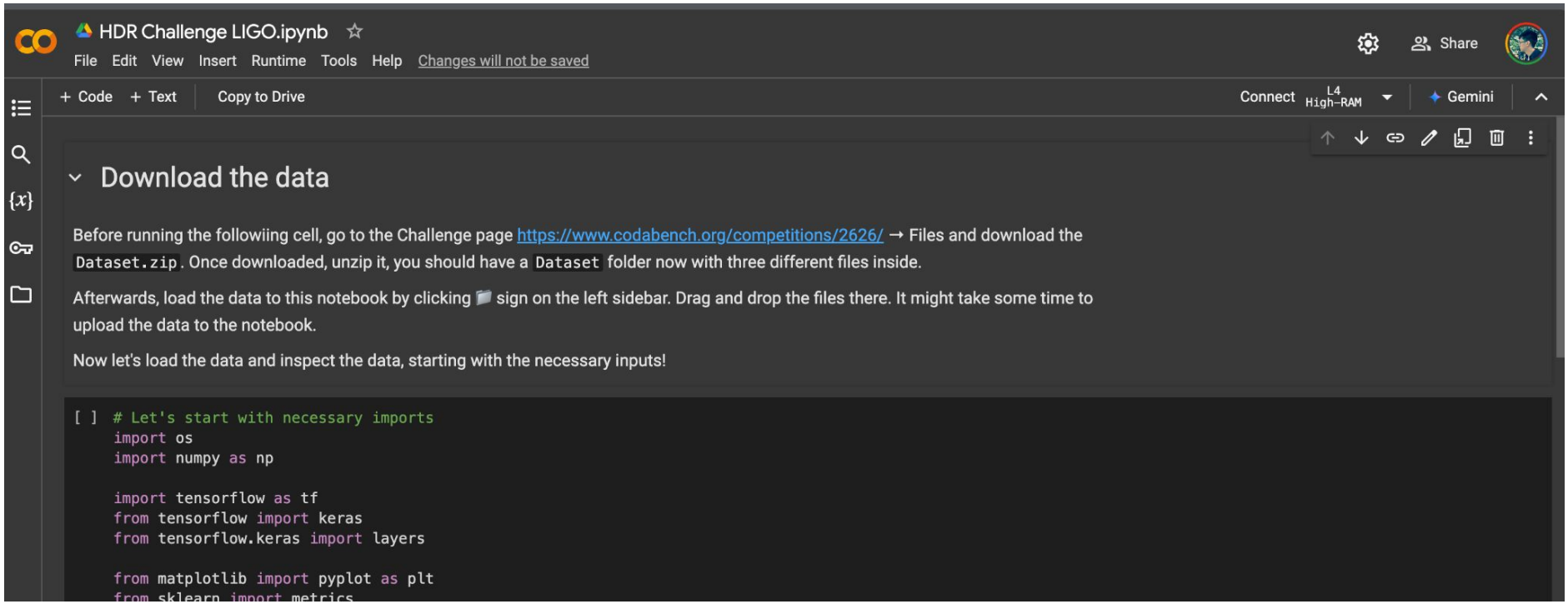
5. Check results in the leaderboard



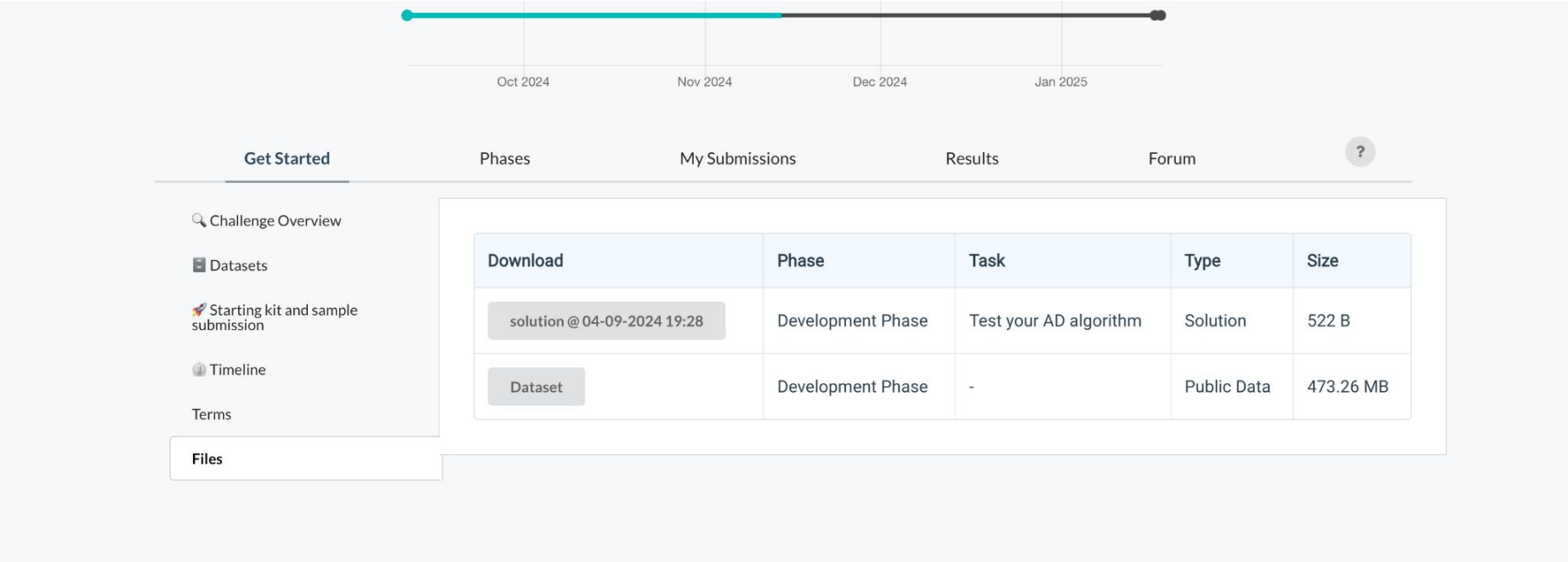
6. Check out the starting kit



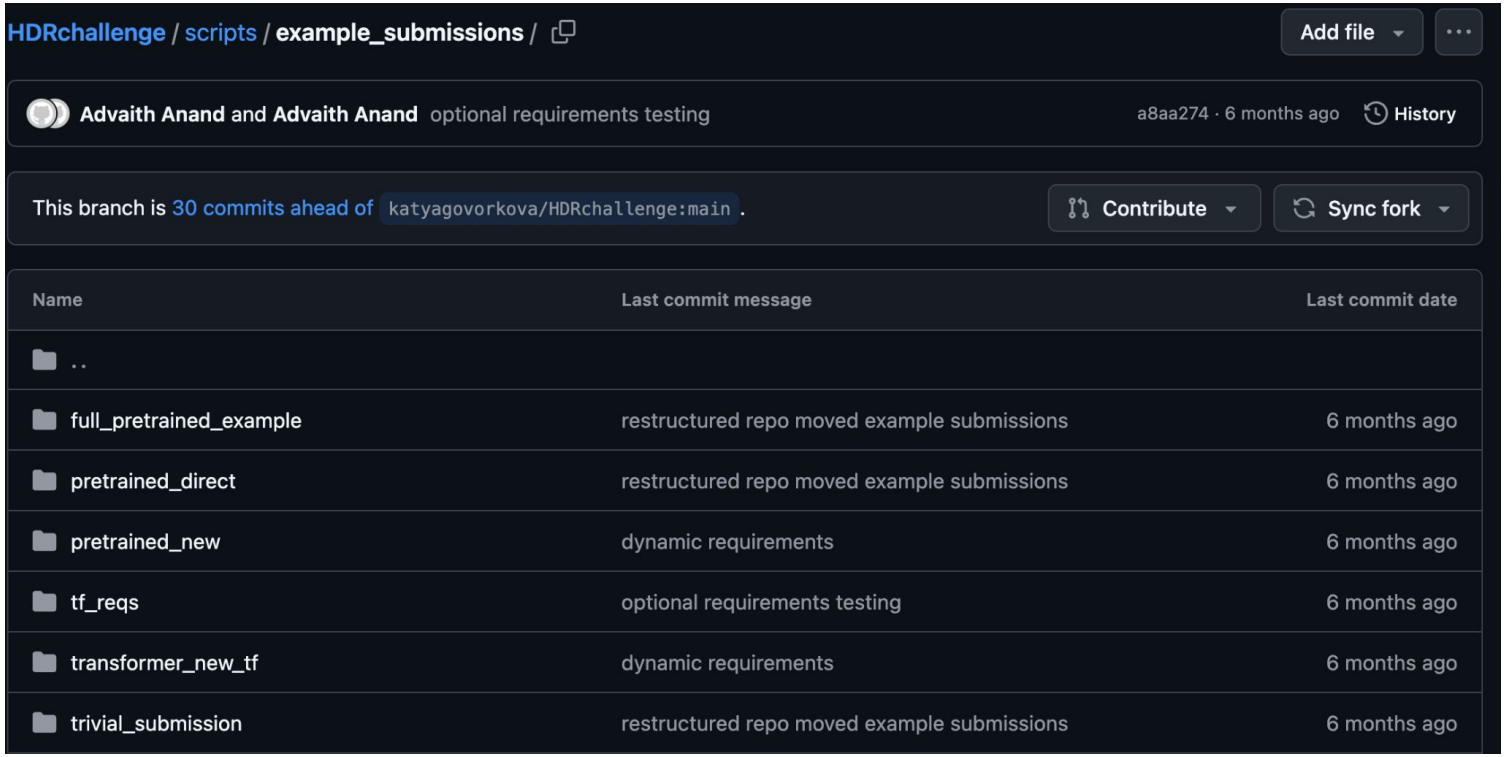
7. Starting kit as a Google Colab Notebook



8. Get Public Data



9. Checkout example submissions



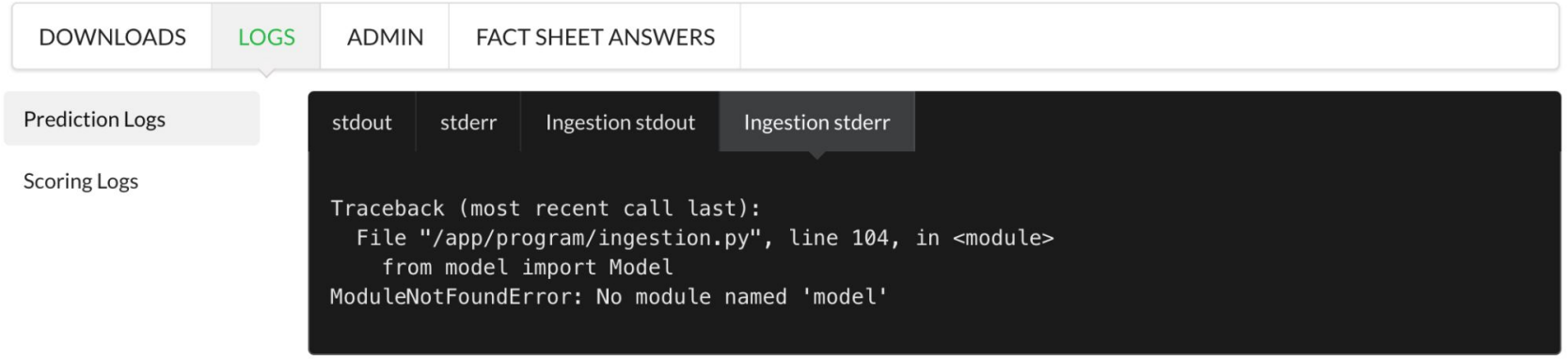
10. Code submission structure [example]

```python
import tensorflow as tf
import json
import os

class Model:
    def __init__(self):
        # You could include a constructor to initialize your model here, but all calls will be made to the load meth
        self.clf = None

    def predict(self, X):
        # This method should accept an input of any size (of the given input format) and return predictions appropri
        preds = self.clf.predict(X)
        print(preds)
        return preds

    def load(self):
        # This method should load your pretrained model from wherever you have it saved

        with open(os.path.join(os.path.dirname(__file__), 'config.json'), 'r') as file:
            for line in file:
                self.clf = tf.keras.models.model_from_json(line)
        self.clf.load_weights(os.path.join(os.path.dirname(__file__), 'model.weights.h5'))
```

[*] Follow the example to load your model. Avoid hard-coded path to model weight

Common issue

[!!] Do not zip the whole folder. ONLY select the model.py and relevant weight files to make the tarball



If you see the above error, mostly likely you zip the whole folder when making the tarball

References & Practicalities

References & Practicalities

- HDR ML Hackathon page: https://www.nsfhdr.org/mlchallenge
  - Todays event page: https://indico.cern.ch/event/1482320/
  - **Model submissions are accepted until January 17th, 2025**

- Details and starter-kits for each challenge:
  - Butterfly hybrid detection
  - Sea-level anomaly detection
  - Gravitational wave detection

- GPU resources are being organised through the Pittsburgh Supercomputing Center

HUGE thank you to them!

PSC

PSC Resources Overview

Allocations → PSC Bridges-2 RM | 300 Core Hours

PSC Bridges-2 GPU | 1k GPU Hours

- Set up an ACCESS account: https://operations.access-ci.org/identity/new-user
  - Email Zach with your ACCESS ID once it is set up: zbaldwin@cmu.edu
- The best resource is the PSC Bridges-2 User's Guide which provides all the necessary information on things like
  - Password Management/Acceptable Use Policy
  - Support/Reporting a problem (help@psc.edu)
  - Usage (environments, software, etc.)

- If any issues arise with PSC resources, email Zach!

or cc him in messages with support

PSC Quick Look

All information provided here (and more) can be found in extensive detail in the User's guide

- Connect to Bridges-2 login node (web browser or command line)
  - OnDemand
  - Command line interface | *ssh -Y <username>@bridges2.psc.edu*

$HOME - User's home directory

$LOCAL - Local file system (only visible to the node the current system is attached to)

$PROJECT - File storage

Transfering files - rsync, scp, sftp or Globus

Running Jobs - interactive, batch, or OnDemand

OpenMP & OpenACC available!

Backup