



National Science Foundation (NSF)  
Harnessing the Data Revolution (HDR)

# Machine Learning Challenges: Anomaly Detections

Local Hosts: Seungbin Park, Megan H. Lipton, Maria C. Dadarlat-Makin,  
Arghya Ranjan Das, Yao Yao, Mia Liu

*Nov 26, 2024  
MJIS 2001, Purdue University*

Collaborate with



Support by



# HDR ML Challenges

Entrance:

<https://www.nsfhdr.org/mlchallenge>



# Introduction



PURDUE  
UNIVERSITY®

- Anomaly detection
- Three challenges:
  - [Butterfly Hybrid Detection](#)
  - [Detecting Anomalous Gravitational Wave\(GW\) Signals](#)
  - [iHARP HDR Anomaly Challenge: Detecting East Coast Sea-level Anomalies](#)
- Dataset and starting kit.
- Submit your model to Codabench
  - Dummy submission today.
- Discussion and team up!

# A few things...

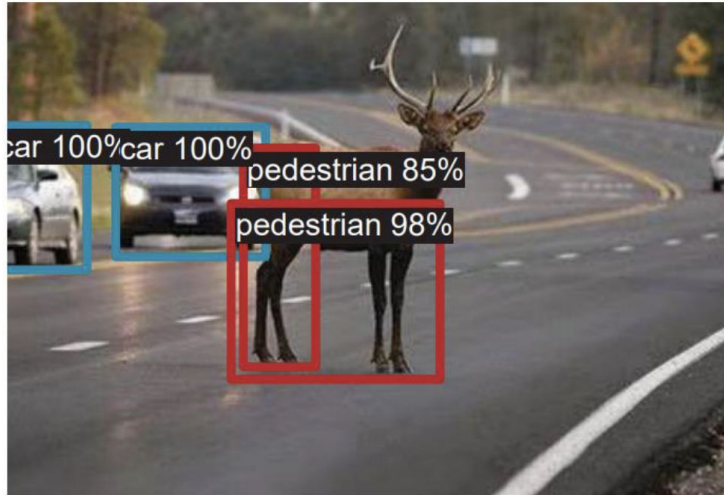
- <https://indico.cern.ch/event/1477186/>  
Winner Award:
  - 💰 Total cash prizes \$2500
  - 👤 \$3000 in AWS cloud computing credits
  - 🏆 Extra award sponsored by AMD (Details pending)
  - 🌟 Special jury prizes include funded invitations to AAAI 2025
- Training resource is not provided, but **FREE** GPU resources do exist.
- For questions regarding the challenges, please contact **challenge creators**.
  - Usually through “issues” on their github repo.
- Your cooperation in following the [Code of Conduct](#) is appreciated!

# What is Anomaly?



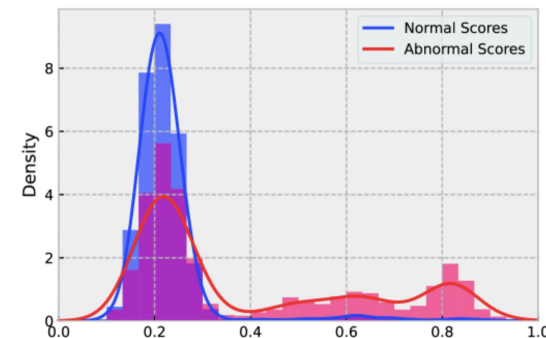
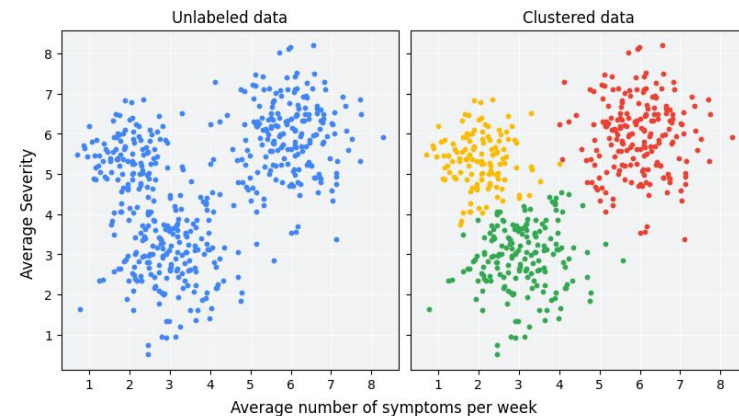
PURDUE  
UNIVERSITY

- Unexpected, unusual, **Rare** behavior:
  - Banking: Detecting fraudulent or irregular spending or requests.
  - Weather: Experiencing a snowstorm in summer or unusually hot temperatures in winter.
  - Automate Driving: a pedestrian ran into the road?



# Anomaly Detection with ML

- Choose algorithm based on if the historical anomalies are well-documented and labeled, or anomaly is well-simulated.
  - **Supervised learning:** when all anomalies are included in the dataset.
  - **Unsupervised learning:** identifies deviations from data, relying on clustering or density-based methods.
  - **Semi-supervised learning:** Trains models on mostly normal data to detect anything outside the learned pattern.
- Feature engineering:
  - Time-series data.
  - Select and transforming input features.
- Evaluation Metrics:
  - Precision, recall, F1-score, AUC...
- Challenges: imbalance dataset, false positives too high, ...

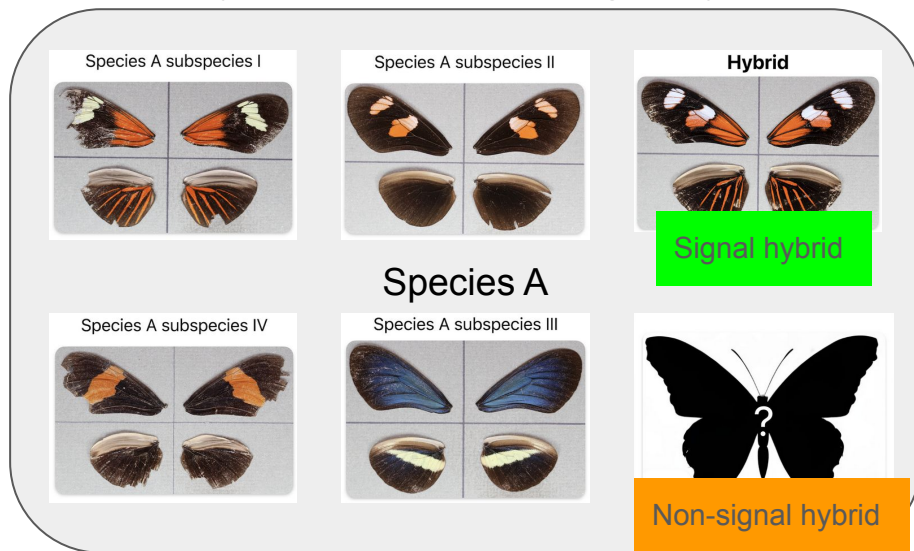


# Butterfly Hybrid



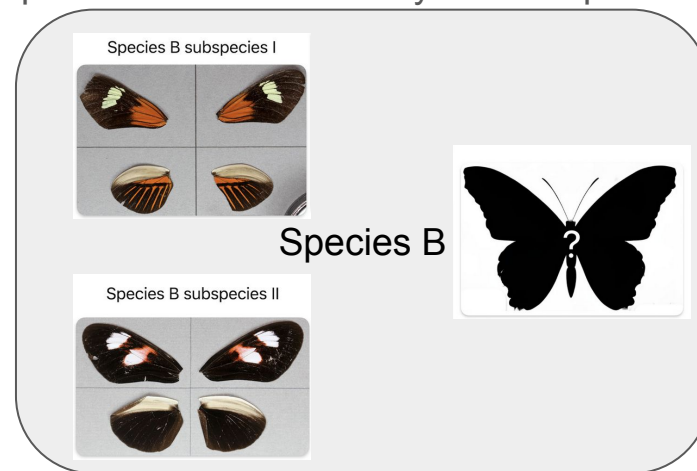
PURDUE  
UNIVERSITY

- All Hybrids are anomalies!
- The signal hybrid is the most common anomalies.
- Other hybrids are called non-signal hybrid.



## Goal of the algorithm:

- Detect signal hybrid in Species A.
- Also can detect non-signal hybrid in Species A.
- This algorithm should be able to be applied to Species B and detector hybrid for Species B.



Note: Species B's certain subspecies mimic Species A, so their subspecies have similar appearance. The purpose is to not break your trained algorithm when applying to a different species.

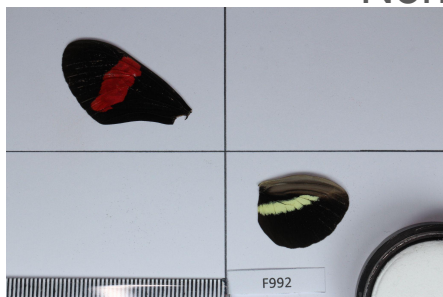
# Butterfly Hybrid



PURDUE  
UNIVERSITY

- <https://www.codabench.org/competitions/3764/>
- Train data: Species A subspecies and Species A signal hybrid. (that is what you are provided).
- Test data (test score feedback from codabench):
  - Species A: subspecies, signal hybrid and non-signal hybrid.
  - Species B: subspecies, hybrid.
- Data:
  - The Photos (7MB each, so make sure to down-size the photos when downloading)
- Starting kit: baseline algorithms (**DinoV2** and **BioCLIP** based) written in pytorch.  
<https://github.com/Imageomics/HDR-anomaly-challenge-sample>

Non-hybrid



Hybrid





# Anomalous GW with LIGOs

- Accelerating masses produce deformations in space-time that we can detect via *interferometry*.
- A signal will appear in at least two interferometers, with the time delay because of the distance between the detectors. In this challenge, we use the processed data from the LIGO (Laser Interferometer Gravitational-wave Observatory) in Washington and the LIGO in Louisiana.



- Known sources: signal sources that well understood and modeled. The mergers.
  - i) Binary black hole (BBH); ii) Binary Neutron Star (BNS) iii) Black hole-Neutron Star (BHNS)
- **Unknown source** (Anomalies).
  - i) Supernovae explosion; ii) Continuous Wave from Spinning Neutron Stars; iii) Stochastic Background...
- **Goal of the algorithm**: develop a semi-supervised approach to discover anomalous signals without explicit modeling.

# Anomalous GW



PURDUE  
UNIVERSITY

- <https://www.codabench.org/competitions/2626/>

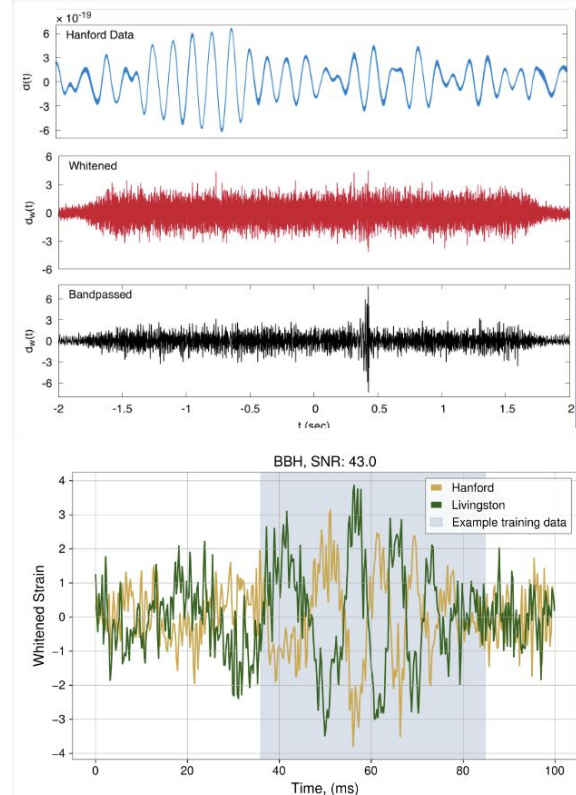
## Dataset

- Has been pre-processed. Whitening and bandpassing.
- Sample rate 4096 Hz => 4096 data point per second
- Every 200 data points is called a segment. So 1 segment ~ 50 milliseconds.
- Dimension of the input data (N, 200, 2). N (number of segment), 200 data points per segment, and 2 laser interferometers.
- **Tips:** Data points are time-ordered, and compare the waves from the two interferometers to reduce local noise.

- background.npz, bbh.npy (known source), sglf.npy (sine-gaussian low frequency, that fakes unknown source). *Note: you model will not be scored based on the provided sglf sample.*

## Starting kit

- A mini transformer written in tensorflow, on Google Colab. The link is provided on codabench.

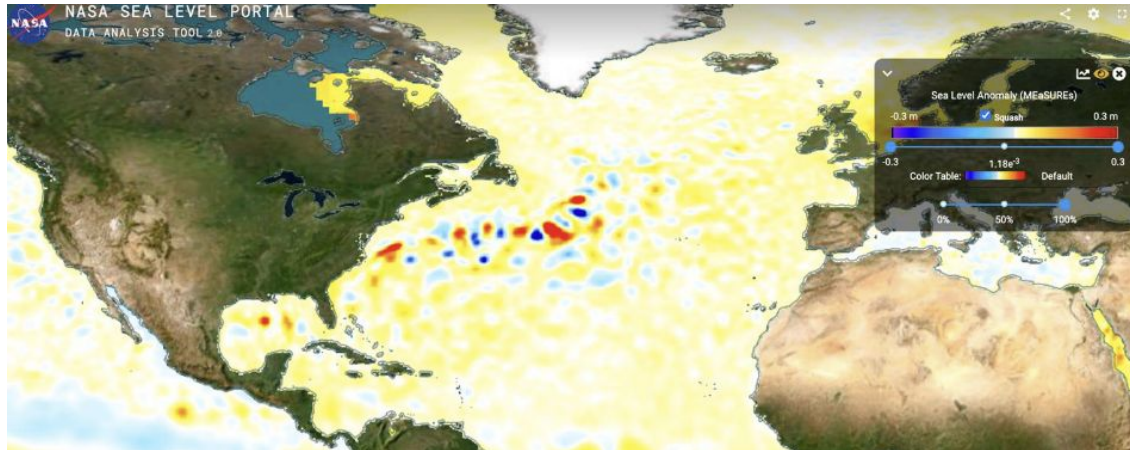


# Coastal Hazard



PURDUE  
UNIVERSITY

- Predict anomalous sea-level observations from daily tide gauge data along the US East Coast affected by changes in the sea-level elevation values on the Atlantic Ocean.



- The training dataset spans 20 years (1993 - 2013), consisting of daily sea level measurements from 12 coastal stations along the US East Coast and sea-level elevation values in the North Atlantic.
- **Goal of the algorithm:**
  - predict the sea-level anomalies across the 12 stations for each day from 2014 to 2023.

# Coastal Hazard



PURDUE  
UNIVERSITY

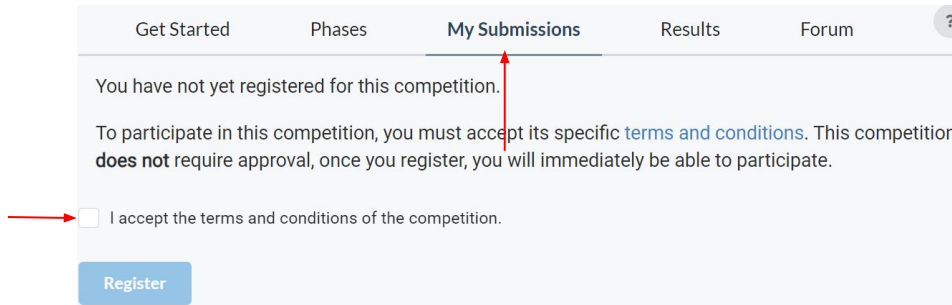
- <https://www.codabench.org/competitions/3223/>
- Data:
  - Each Buoy station has a .csv file, listing its location, anomaly (0 or 1) for every day. -> 12 files.
  - Atlantic data in .nc format (NetCDF files). Ask chatGPT or Google about how to extract data.
  - Example of data extraction:  
[https://colab.research.google.com/drive/1lwQpJID4xScUpCCxhgYkv6F\\_YjHZwx5-?usp=sharing](https://colab.research.google.com/drive/1lwQpJID4xScUpCCxhgYkv6F_YjHZwx5-?usp=sharing)
- Please split the data into training and testing.
- Private testing dataset: A subset of the time series data from each station with hidden anomalies. The challenge organizers will use this dataset to evaluate the submitted models and determine the final scores.
- Evaluation metric: F1 score.
  - Precision = True Positives / (True Positives + False Positives)
  - Recall = True Positives / (True Positives + False Negatives)
  - F1 Score =  $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$

# Submission Workflow with Codabench

**Account creation:** For submitting the models and other associated files participants must have an account in Codabench, hence [signup](#) and [login](#) to the accounts.

**Download the Dummy submission:** Every challenge has its own dummy submission formats and requirement. Navigate to the Sample Submission section to download.

**Register in the competition to begin:**



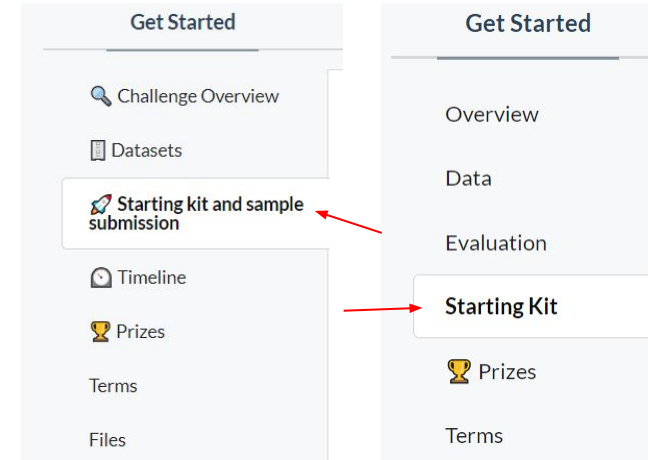
Get Started   Phases   **My Submissions**   Results   Forum   ?

You have not yet registered for this competition.

To participate in this competition, you must accept its specific [terms and conditions](#). This competition **does not** require approval, once you register, you will immediately be able to participate.

I accept the terms and conditions of the competition.

Register



**Explore the Starting Kits:** Various starting Kits are provided for the challenges. Participants are encouraged to look at them to familiarize with the competition's structure.

# Submission Workflow with Codabench



PURDUE  
UNIVERSITY

## Make Dummy Submission:

- After the registration into the competition, participants will be able to make dummy/real submissions.
- To ensure a smooth start, participants are required to make at least one submission, starting with the Dummy files.
- The participants can **later** train their own models and submit the **real submissions** in the same format as the dummy one.

**Leaderboard:** Can be accessed from the **Results** tab.

## Accessing Public Data:

- All the publicly available data and files can be found by navigating to the **Files** section.
- Note: that the format and type of data and files available are different for the different challenges.

Submission upload

Submit as: ?  
Yourself

Submission

Leaderboard

Search... Status

ID #	File name	Date	Status	Score	Detailed Results	Actions
No submissions found! Please make a submission						

Get Started Phases My Submissions

Challenge Overview

Datasets

Starting kit and sample submission

Timeline

Prizes

Terms

Files

Download

solution @ 04-09-2024 19:28

Dataset

Public Data

# Submission Workflow with Codabench



PURDUE  
UNIVERSITY

## Train your models

- The publicly available data and the associated files can be used or modified as needed for the model training and development.
- Training resources are not provided. Participants must use their own resources or free platforms like Google Colab for model training.

Note: Some files may be labeled as testing data. These are intended for dummy testing purposes and may differ from the data used during backend evaluation.

## Example Submissions: [Link](#)

- The final files should be zipped together and submitted.
- Only submit the relevant files and not the entire workplace folder. Look at the Github link for some example submissions

## Common Issues

1. Do not zip the entire folder. Instead, create a tarball with only the required files (model.py and weight files).  
For Example:

```
tar -cvf submission.tar model.py weights/
```

Or else you may face errors like:

Traceback (most recent call last):  
File "/app/program/ingestion.py", line 104, in <module>  
from model import Model  
ModuleNotFoundError: No module named 'model'

2. Avoid hard-coded path to the model weights in model.py for reference look at the *Example Submissions*

# Discussion

- Feel free to chat, ask questions, share experience and team up.
- Enjoy coffee and cookies!