NSF HDR ML Hackathon 2024

Max Cohen

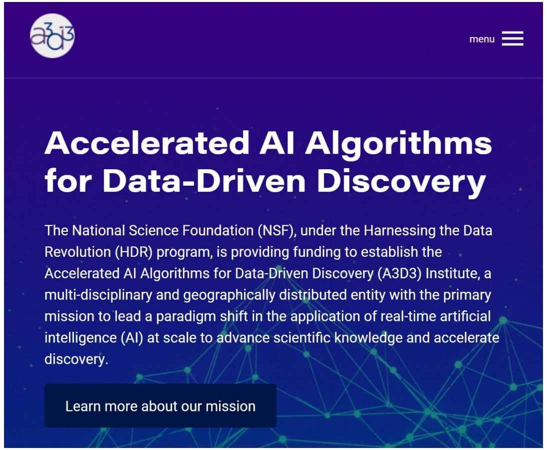
Support by

Introduction

Welcome to the NSF HDR ML Hackathon!

- This event is hosted by A3D3 (Accelerated AI Algorithms for Data-Driven Discovery), a national organization dedicated to the development of AI for science

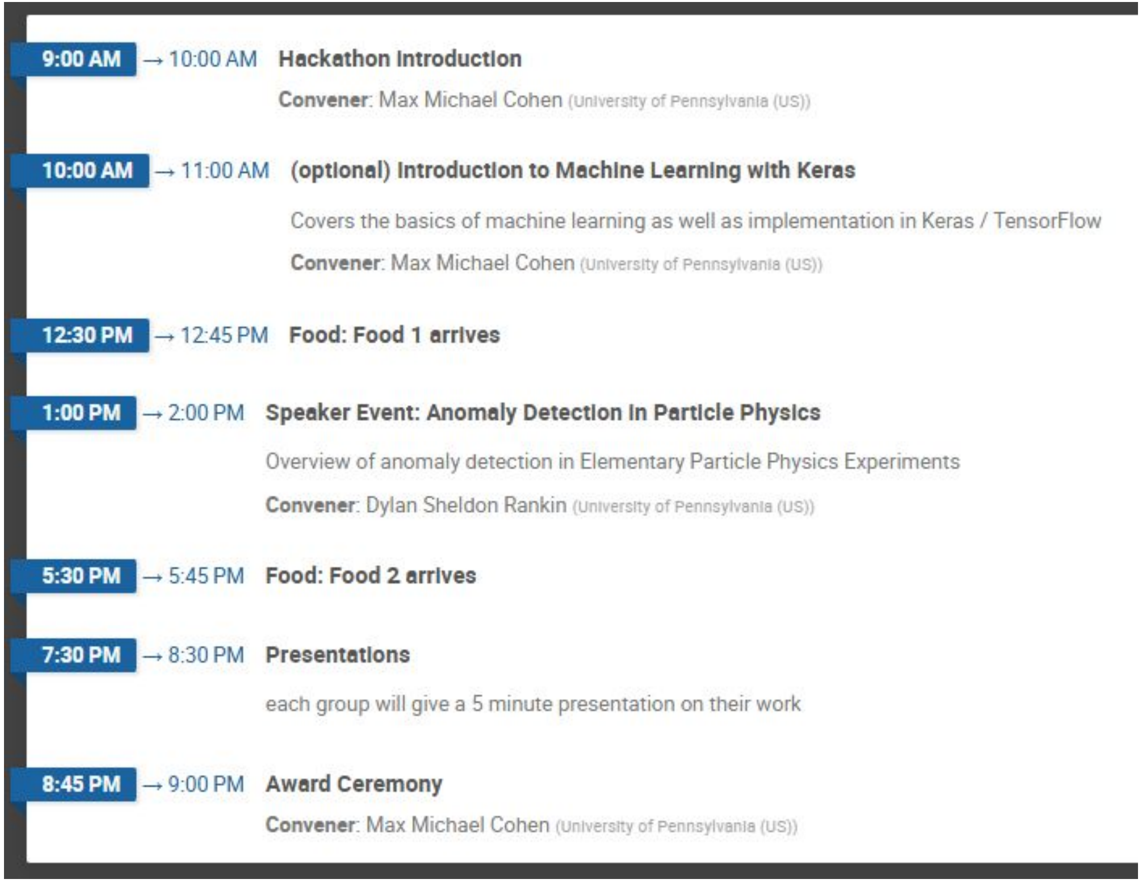
- I'm **Max Cohen**, an A3D3 affiliate and PhD student here at Penn. My current work is developing anomaly detection algorithms for use in **elementary particle physics**



Accelerated AI Algorithms for Data-Driven Discovery

The National Science Foundation (NSF), under the Harnessing the Data Revolution (HDR) program, is providing funding to establish the Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute, a multi-disciplinary and geographically distributed entity with the primary mission to lead a paradigm shift in the application of real-time artificial intelligence (AI) at scale to advance scientific knowledge and accelerate discovery.

Event Schedule



9:00 AM → 10:00 AM **Hackathon Introduction**

Convener: Max Michael Cohen (University of Pennsylvania (US))

10:00 AM → 11:00 AM **(optional) Introduction to Machine Learning with Keras**

Covers the basics of machine learning as well as implementation in Keras / TensorFlow

Convener: Max Michael Cohen (University of Pennsylvania (US))

12:30 PM → 12:45 PM **Food: Food 1 arrives**

1:00 PM → 2:00 PM **Speaker Event: Anomaly Detection in Particle Physics**

Overview of anomaly detection in Elementary Particle Physics Experiments

Convener: Dylan Sheldon Rankin (University of Pennsylvania (US))

5:30 PM → 5:45 PM **Food: Food 2 arrives**

7:30 PM → 8:30 PM **Presentations**

each group will give a 5 minute presentation on their work

8:45 PM → 9:00 PM **Award Ceremony**

Convener: Max Michael Cohen (University of Pennsylvania (US))

Today: Choose one of three AD challenges

Detect anomalous gravitational waves from real LIGO data

Find hybrid butterfly species from only pictures

Predict flood events along the east coast from satellite data

Hackathon Goals and Awards

During the challenge, you'll work on two goals:

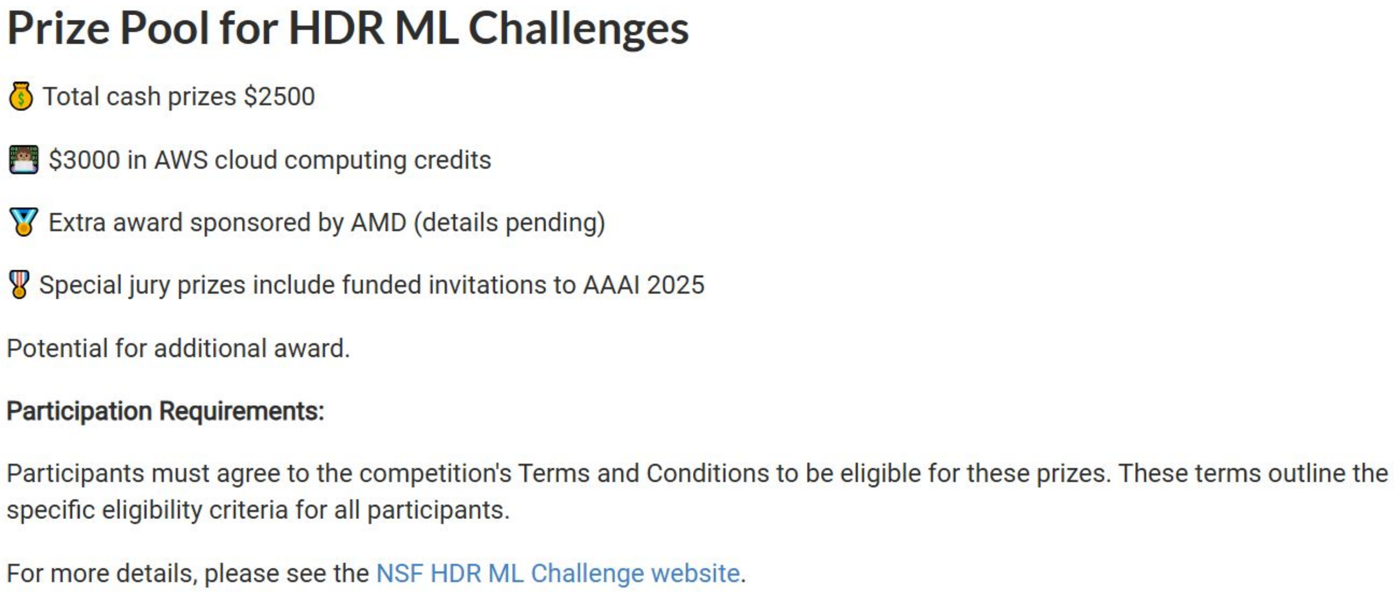
1) Train Smaller Models
   a) Submit these models to receive feedback and test scores in real-time.
2) Plan a Larger Network
   a) Design and write the code for a larger model you don't have time to train today.
   b) Motivate why it should work, and test it on a small number of events to debug.

Awards Categories:

- **Most Performant Model:** Achieves the best performance in real-time tests.
- **Most Exciting Plan:** For the most promising design of a larger network.
- **Most Creative Approach:** Celebrates unique and creative strategies.

Additionally: ML Challenge Awards

Anyone participating in the ML challenge at any institution, whether during a hackathon or otherwise, is eligible for the following awards:



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.

Participation Requirements:

Participants must agree to the competition's Terms and Conditions to be eligible for these prizes. These terms outline the specific eligibility criteria for all participants.

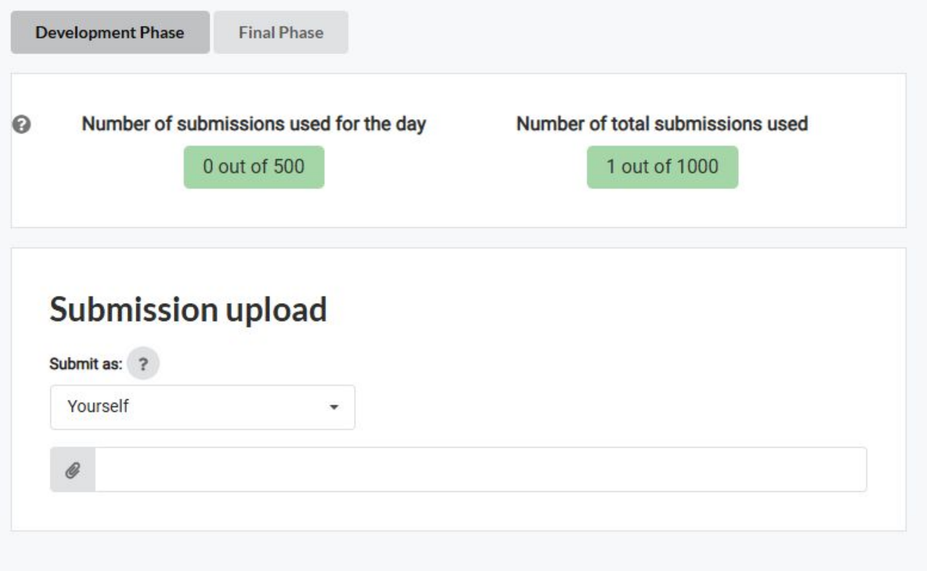
For more details, please see the NSF HDR ML Challenge website.

Most Performant Model Award

For each of the three challenges, you'll have the opportunity to submit a trained network to be run on a testbench and given a performance score.

The Most Performant Model Award will go to a team which has exceptional performance on this submitted testbench!

IMPORTANT: Submitted Networks will only be evaluated between 1:00pm and 6:00pm



Number of submissions used for the day: 0 out of 500

Number of total submissions used: 1 out of 1000

Submission upload

Submit as: Yourself

Most Exciting Plan Award

With only 12 hours, there may not be enough time to train larger networks.

We encourage you to:

- **Design a Network:** Propose an architecture you believe will excel at your challenge
- **Implement a Pipeline:** Program the full training and testing pipeline. Test it on a small subset of events to ensure it's bug-free
- **Plan for Training:** Outline when and how you will train the network after the hackathon

The Most Exciting Plan Award will go to the team with an especially notable network design and a ready-to-run pipeline!

Most Creative Approach Award

This award celebrates **innovative thinking and unique solutions** to the challenges.

It recognizes a team that:

- Explores unconventional ideas or techniques
- Thinks outside the box in designing networks, features, or strategies
- Develops novel approaches to anomaly detection that go beyond baseline methods

This could be demonstrated through:

- Clever data preprocessing or feature engineering
- Innovative use of architecture or algorithms
- Applying concepts from other fields or surprising approaches to problem-solving

The **Most Creative Approach Award** is about showcasing ingenuity and pushing boundaries of what's possible in this hackathon setting!

Group Presentations

- At the end of the day, your group will give a short, 5 minute presentation about your work
  - You'll report the test scores you obtain, as well as your plan for training the larger network

- Additionally (not required), groups can submit github repositories containing codes, which will make it easier for us to judge the quality of your work

Based on these presentations and github repos, we'll determine the winners of the prizes

Winners will get certificates printed with their names!

Anomaly Detection

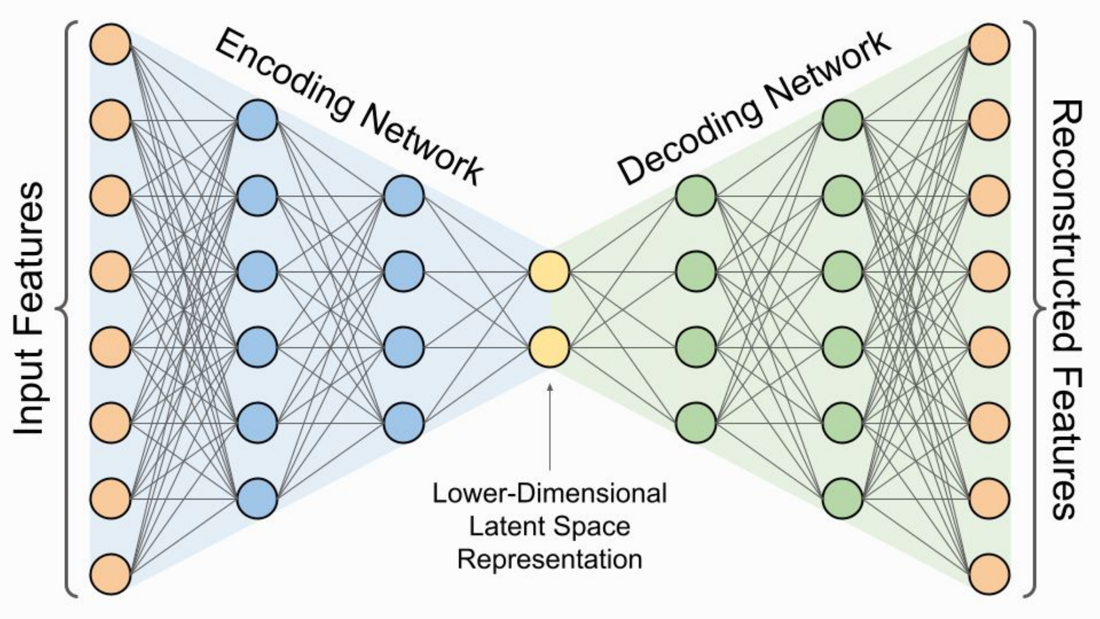
Anomaly detection (AD) seeks to identify outliers of a dataset, often done with unsupervised machine learning.

Anomaly detection **does not require any knowledge of the signal** itself, but rather just that the **signal should look different from background.**

Anomaly Detection

Common network architecture for AD: **Autoencoders**

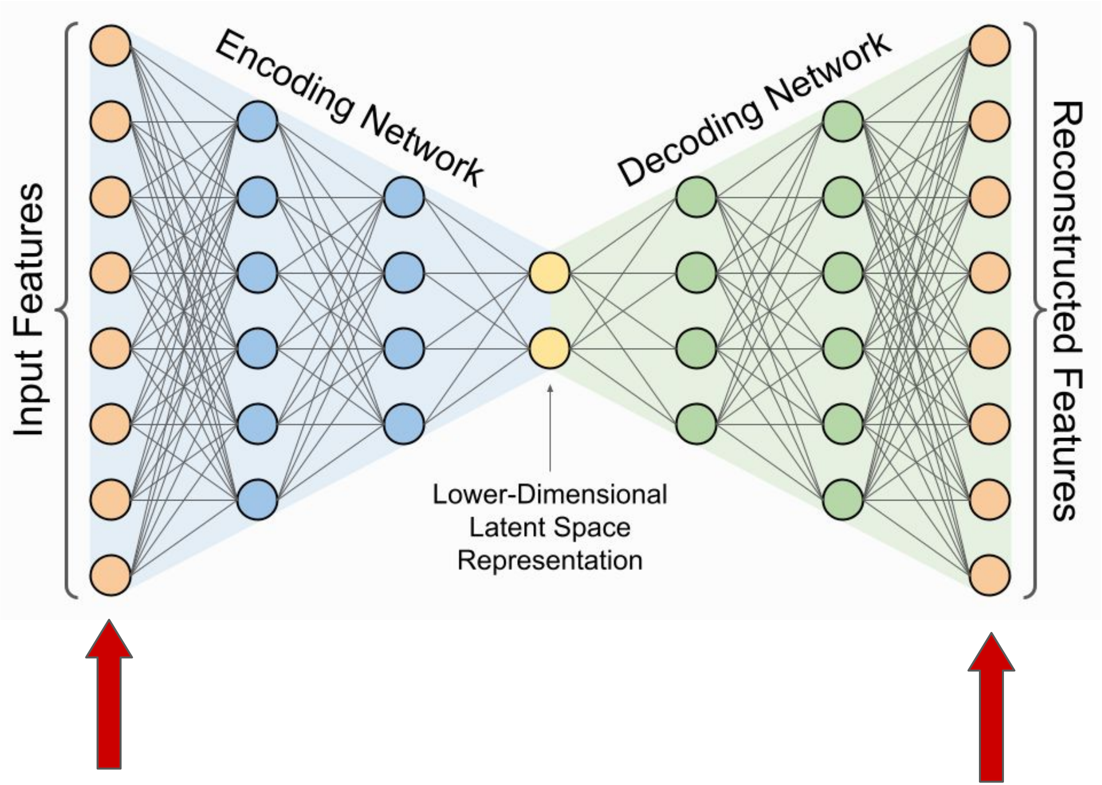
- Encoder compresses data into a lower dimensional representation
- Decoder takes this representation and attempts to recreate the input



Anomaly Detection

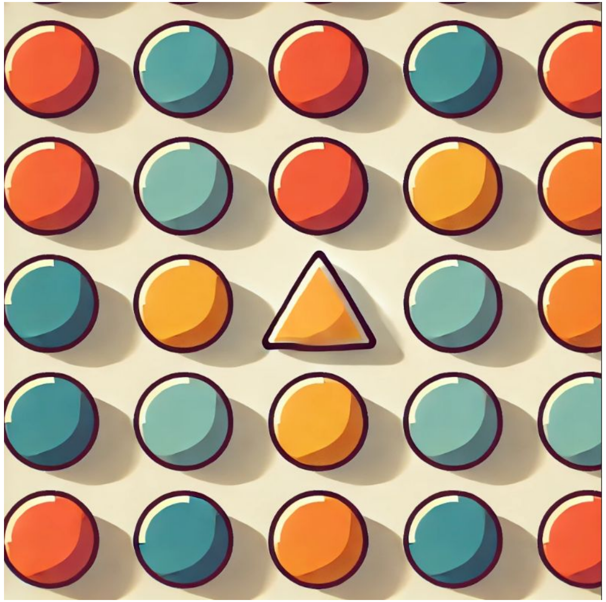
Common network architecture for AD: **Autoencoders**

- The loss is calculated by computing the MSE between the input and output of the autoencoder



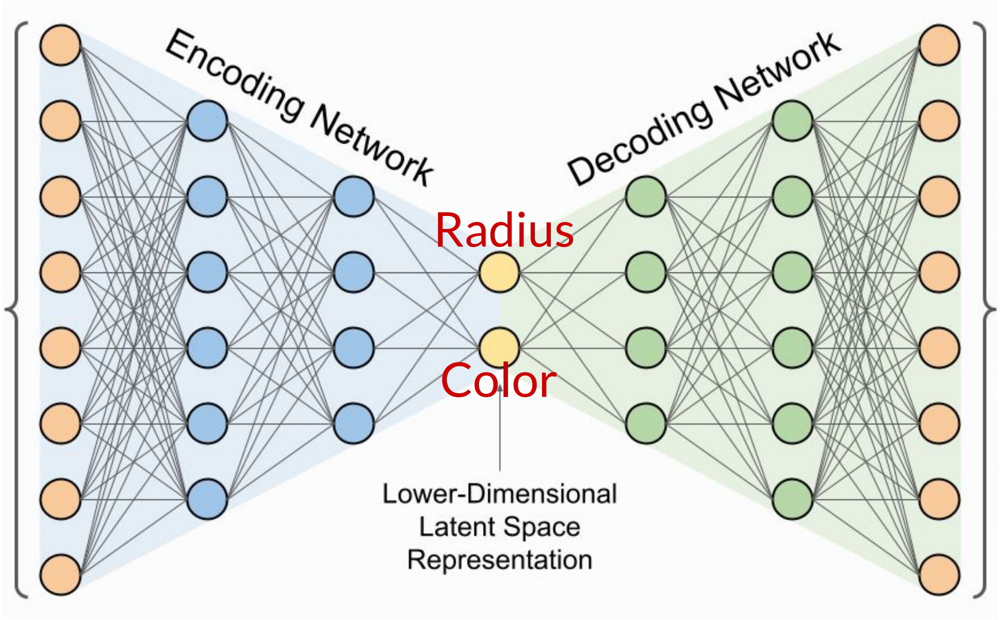
Autoencoder Example

Imagine the network is trying to pick out anomalous pictures of shapes:



Autoencoder Example

During training, the network will learn features and patterns of the data:

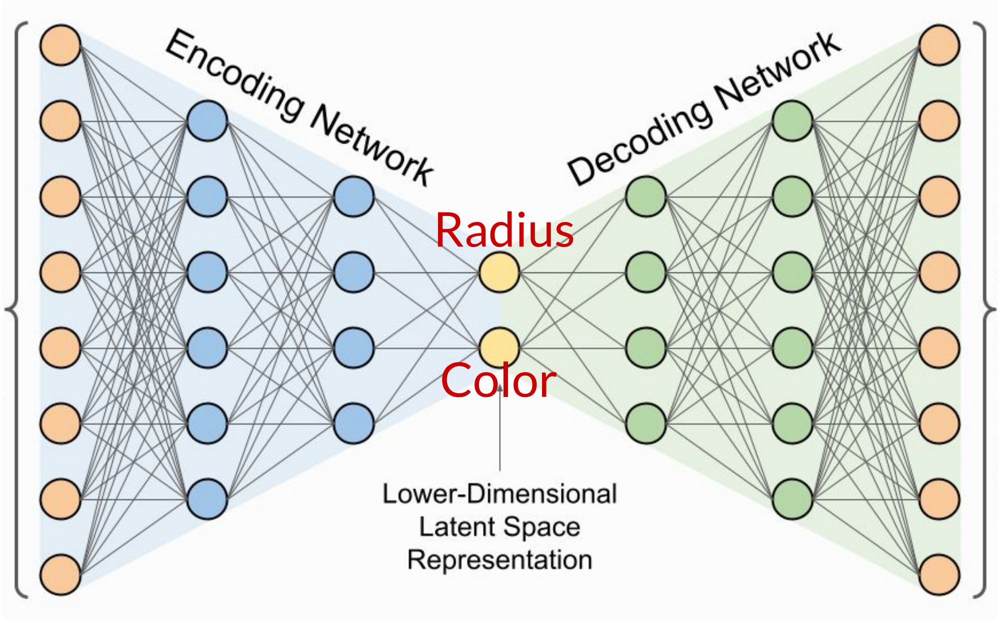


Radius

Color

Autoencoder Example

But these features fail to describe anomalous samples, yielding a large MSE loss!



Today: Choose one of three AD challenges

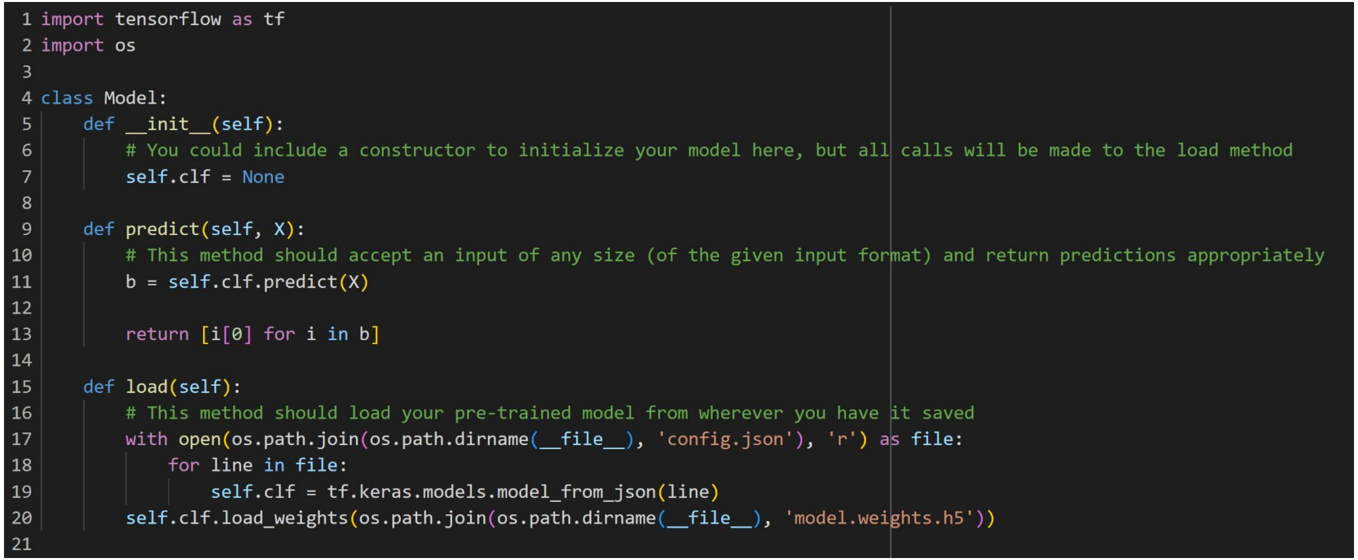
Detect anomalous gravitational waves from real LIGO data

Find hybrid butterfly species from only pictures

Predict flood events along the east coast from satellite data

Submissions: Gravitational Waves and Butterfly Hybrids

IMPORTANT: Submissions must be in the following format (except flood detection):

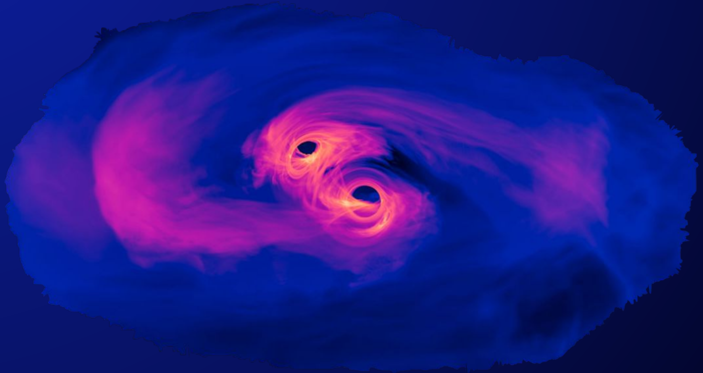


```python
import tensorflow as tf
import os

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

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

        return [i[0] for i in b]

    def load(self):
        # This method should load your pre-trained 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'))

```

- The submission must be a zipped file containing this model.py as well as any auxiliary files (model weights, config.json, etc)
- Keras, TensorFlow, Pytorch, Ski-kit learn are all allowed
- List of allowed packages found here

Submissions: Flood Detection

For the iHARP Sea level rise challenge, you will not submit a model, but rather a CSV file with your results. This will be further explained later.



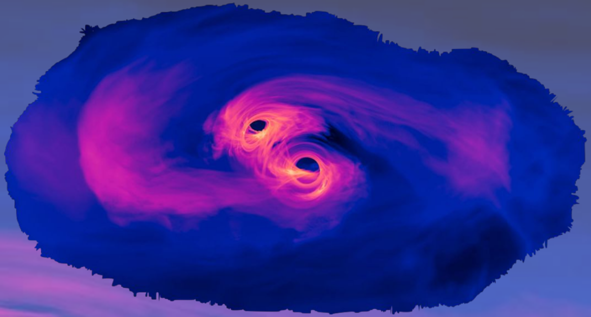
NSF HDR A3D3: Detecting Anomalous Gravitational Wave Signals

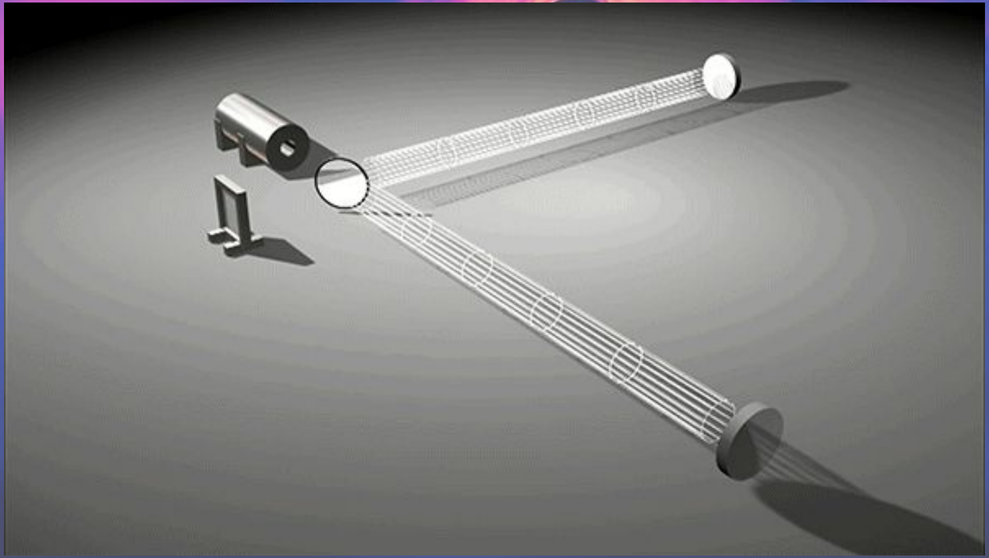
Slides created by:

Katya Govorkova, Yuan-Tang Chou, Phil Harris

Gravitational Waves and Their Detection

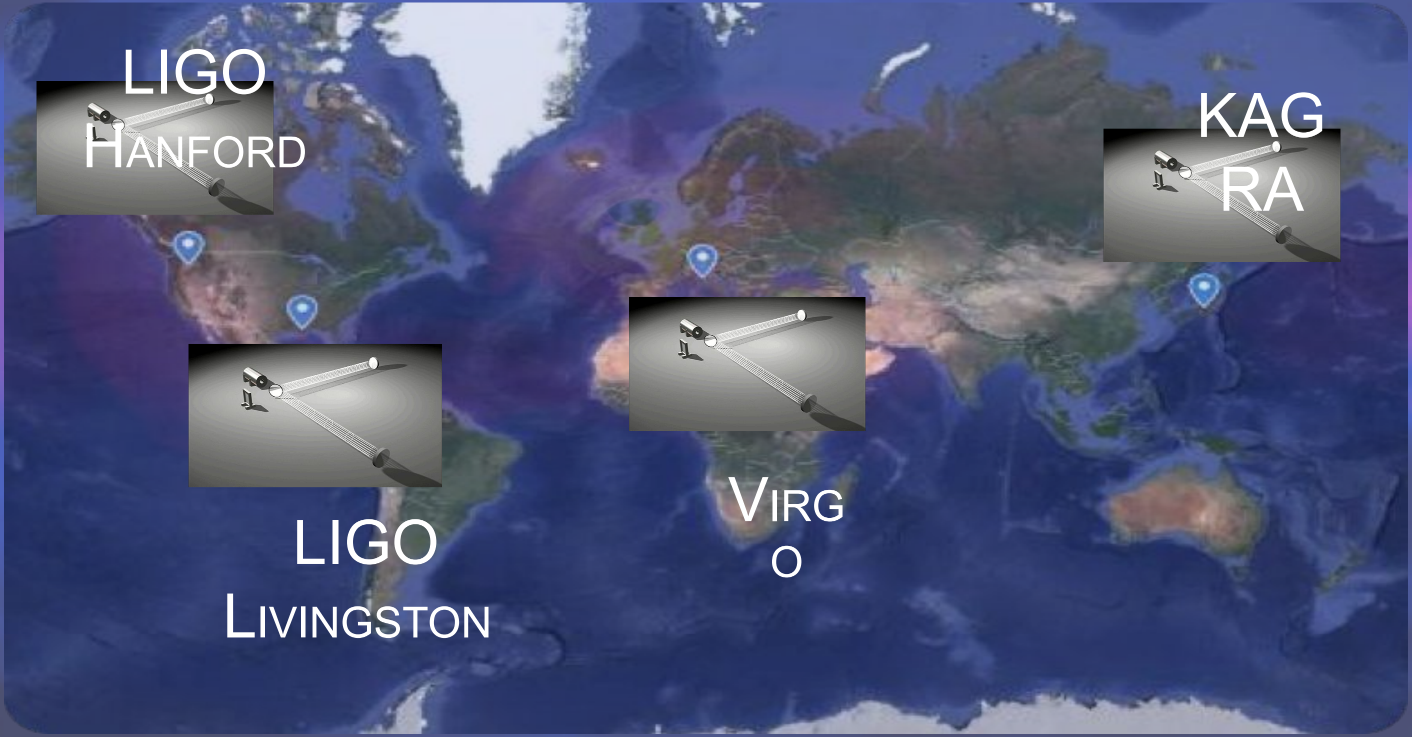
Accelerating masses produce deformations in space time that we can detect via interferometry





The LIGO-Virgo-KAGRA Collaboration

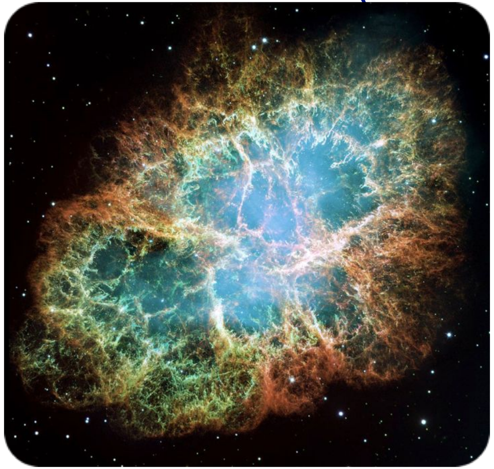
A signal will appear in at least two interferometers, with the time delay because of the distance between the detectors



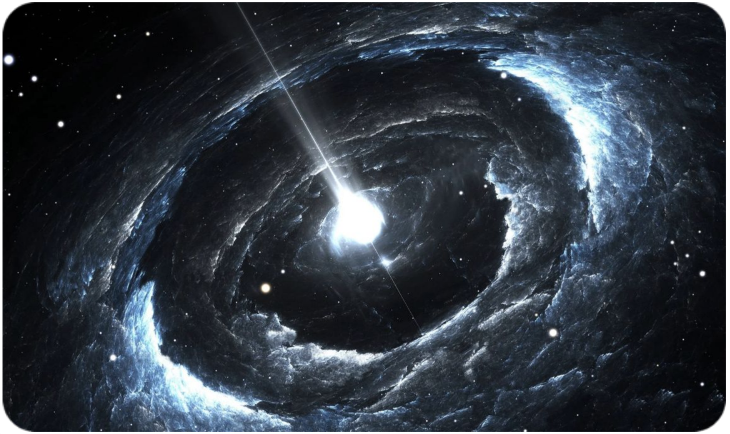
GWAK Anomalous Gravitational Wave Sources

Known “Unknowns” possible signal sources that are poorly modelled and therefore cannot be easily detected using the match filtering pipeline

Core-collapse supernova (CCSN)



Neutron Star Glitches



GWAK Anomalous Gravitational Wave Sources

Unknown “Unknowns” new, unexpected GW sources

We refer to them as anomalous and aim to develop a semi-supervised approach which would let us to discover anomalous signals without explicit modelling

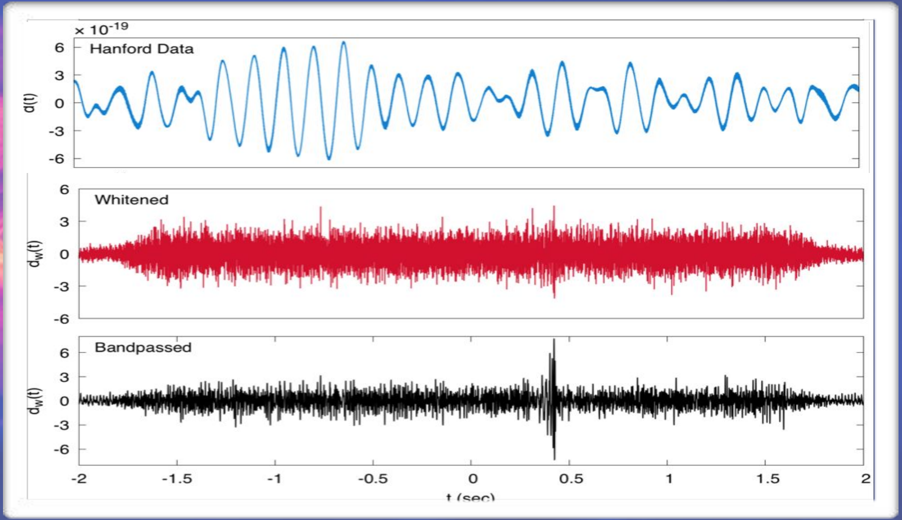
?

Gravitational-wave data preprocessing

Continuous time series 4096 Hz

Whitening

IS TRANSFORMING THE DATA SO THAT IT HAS A FLAT (UNIFORM) POWER SPECTRAL DENSITY, MAKING DIFFERENT FREQUENCY COMPONENTS COMPARABLY SCALED FOR MORE EFFECTIVE SIGNAL DETECTION



Bandpassing 30 Hz < x < 1500 Hz

IS A FILTERING TECHNIQUE THAT ISOLATES THE FREQUENCY RANGE WHERE GRAVITATIONAL WAVE SIGNALS ARE EXPECTED, REMOVING BOTH LOW-FREQUENCY NOISE AND HIGH-FREQUENCY

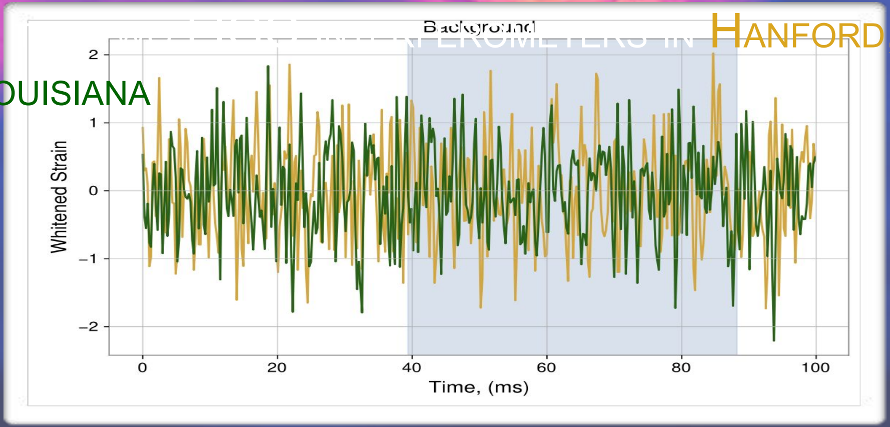
Background dataset

K.Govorkova et al
MLST
10.1088/2632-2153/a
3a31

Sampling rate is 4096 Hz, meaning there are 4096 data points recorded every second

The data is divided into segments of 50 milliseconds each, which contains 200 data points (50 milliseconds * 4096 samples/second = 200 samples)

The dimension of the input data is (N, 200, 2), where N represents the number of data segments. The last dimension of 2 corresponds to the data streams from [illegible] interferometers in Hanford, Washington, and Livingston, Louisiana



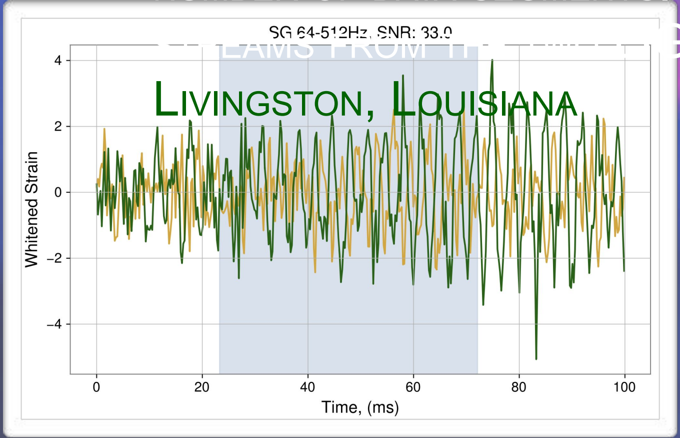
Signal datasets

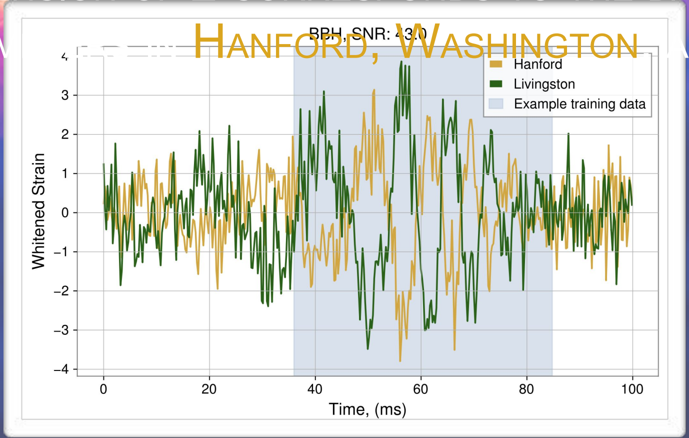
K.Govorkova et al MLST 10.1088/2632-2153/a 2 1

Sampling rate is 4096 Hz, meaning there are 4096 data points recorded every second

The data is divided into segments of 50 milliseconds each, which contains 200 data points (50 milliseconds * 4096 samples/second = 200 samples)

The dimension of the input data is (N, 200, 2), where N represents the number of data segments. The last dimension of 2 corresponds to the data streams from th GO interferom Hanford, Washington and Livingston, Louisiana





Example submission

```
import tensorflow as tf
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 method
    self.clf = None

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

    return [i[0] for i in b]

  def load(self):
   # This method should load your pre-trained 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'))
```

Resources

- The notebook with example https://colab.research.google.com/drive/1hatkYT5Xq6qauDXY6xFrfnGzB66QPsV8?usp=sharing
- The paper with more details and our algorithm MLST 10.1088/2632-2153/ad3a31
- Challenge page with details about the dataset https://www.codabench.org/competitions/2626/
- Any questions should be submitted as a GitHub issue https://github.com/a3d3-institute/HDRchallenge/issues

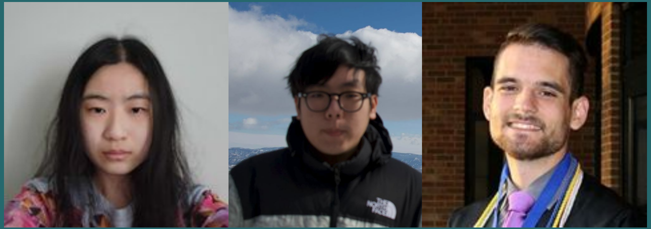
Anomaly Detection: Hybrid Butterflies

Slides created by:

Elizabeth G. Campolongo

and

The Imageomics ML Challenge Team

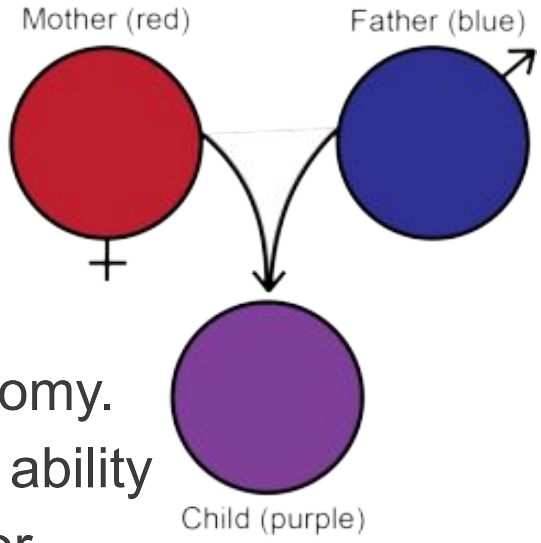


Special thanks to Lisa Wu, Ziheng Zhang, David Carlyn

Hybrid Detection

A brief history

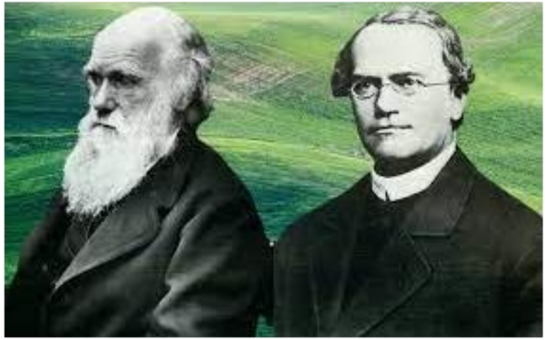
Hybrid Detection



- Researchers have sought a means to detect hybrids since the creation of the field of taxonomy.
- Detecting hybrids would give taxonomists the ability to determine what constitutes a true *species* or *subspecies.*
- The question is ***how****?*
  - *How* do we recognize a hybrid?
  - What does a hybrid look like?

Hybrid Detection: History

- Darwin first posed this question of “What does a hybrid look like?”
- Mendel answered with his pea plant experiment.



Hybrid Detection: History

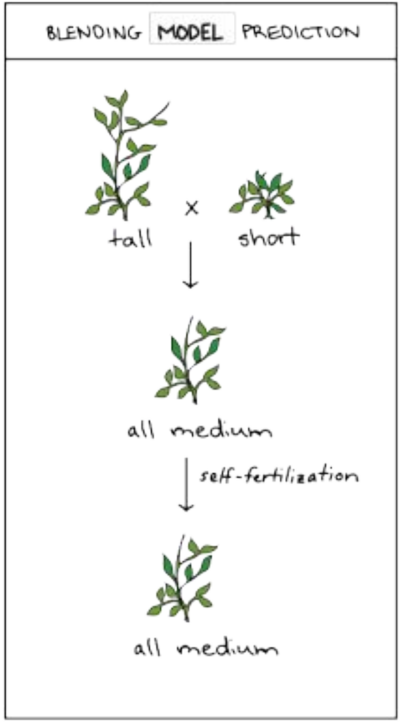
Mendel’s *Hypothesis*:

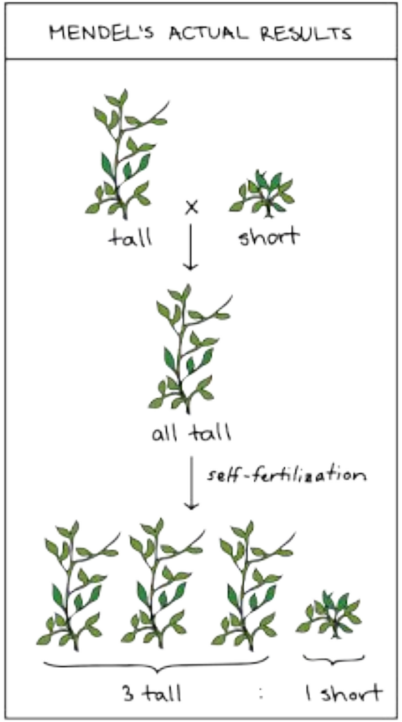
Blending Inheritance

- Inheritance of traits is ***continuous***.

Mendel’s *Results*:

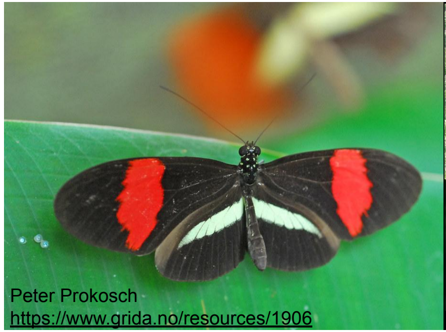
Inheritance is often ***discrete***.



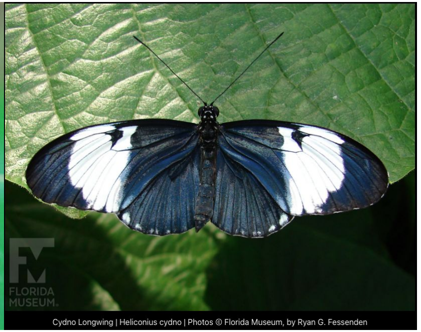


Hybrid Detection: Butterflies

- Consider these two species:
- Hybridization may lead to a variety of resulting patterns.
- There are several [dominant] genes that control color pattern on wings.
  - Ex: red on hindwings is a dominant trait.
- Dominance: hybrids may look like one parent.
- In practice, identifying hybrids requires knowledge of their parent species/subspecies.

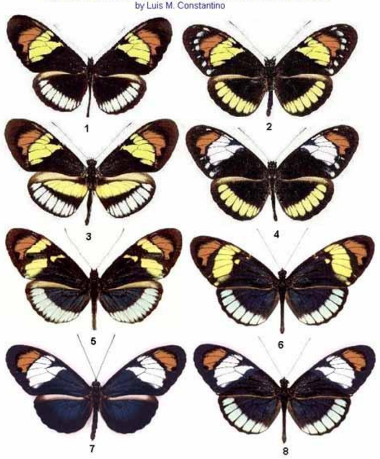


Peter Prokosch
https://www.grida.no/resources/1906



Cydno Longwing | Heliconius cydno | Photos © Florida Museum, by Ryan G. Fessenden

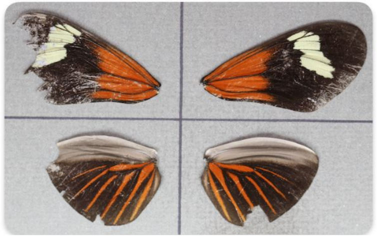
Hybrids between H. cydno and H. melpomene from Colombia
by Luis M. Constantino



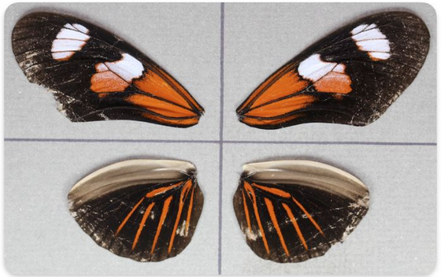
Our Challenge

How ***you*** can contribute to answering this important biological question

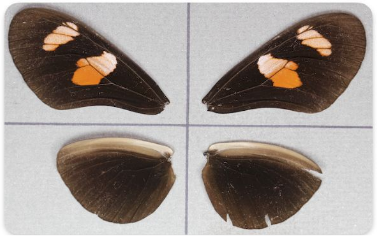
Species A subspecies I



Hybrid



Species A subspecies II



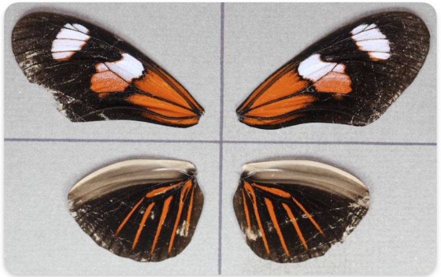
Images are from Zenodo records 2714333, 3082688, 2677821, 2686762, and 2684241, and are licensed under CC-BY 4.0.

Images are from Zenodo records 2714333, 3082688, 2677821, and 2686762, and are licensed under CC-BY 4.0.

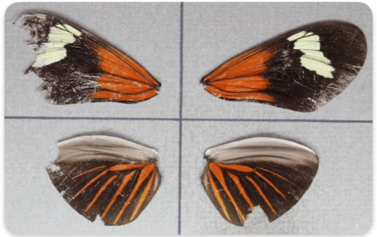
Images are from Zenodo records 2714333 and 3082688, licensed under CC-BY 4.0.

Hybrid graphic generated using Canva Magic Media, manually edited.

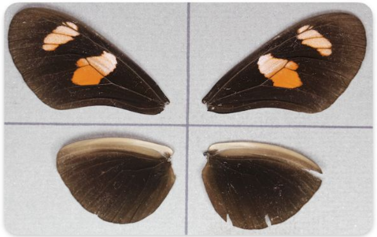
Hybrid



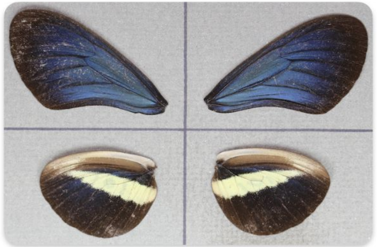
Species A subspecies I



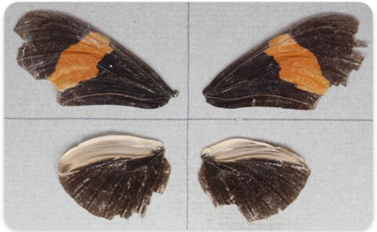
Species A subspecies II



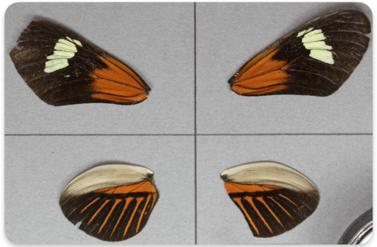
Species A subspecies III



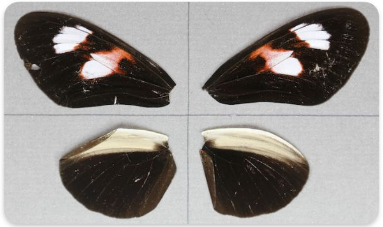
Species A subspecies IV



Species B subspecies I



Species B subspecies II



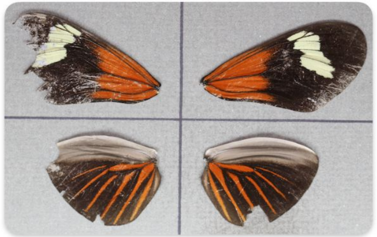
Images are from Zenodo records 2714333, 3082688, 2677821, 2686762, and 2549524, and are licensed under CC-BY 4.0.
Hybrid graphic generated using Canva Magic Media AI, then manually edited.

Our Challenge: Training Data

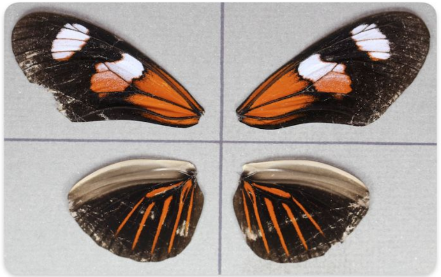
Training Data

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

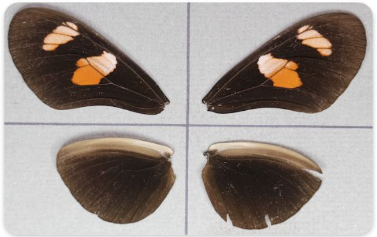
Species A subspecies I



Signal Hybrid



Species A subspecies II



Images are from Zenodo records 2714333 and 3082688, licensed under CC-BY 4.0.

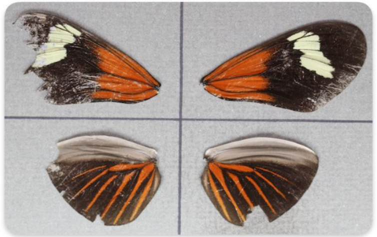
Our Challenge: Dev & Test Data

- 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).
- The numbers:
  - Validation Data (Dev): ~1100 images
  - Test Data: ~2200 images

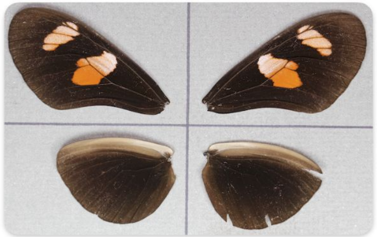
The Challenge: Find the Hybrids

- 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)?

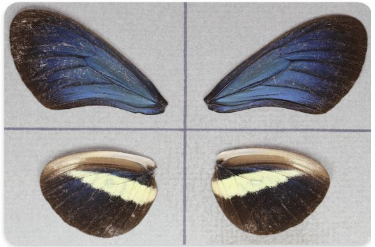
Species A subspecies I



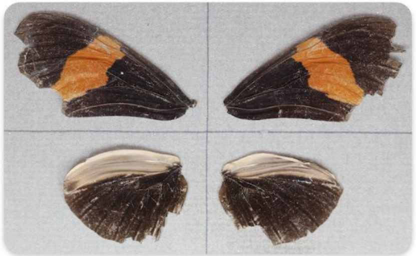
Species A subspecies II



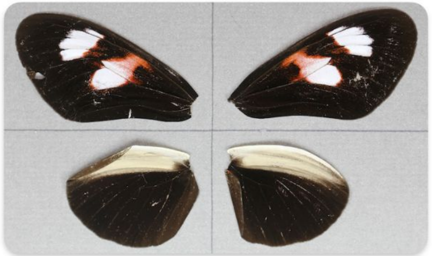
Species A subspecies III



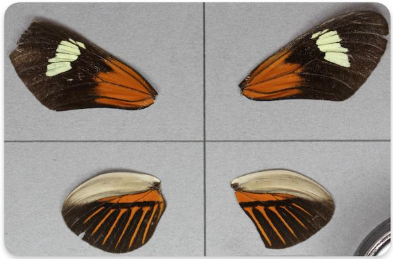
Species A subspecies IV



Species B subspecies II

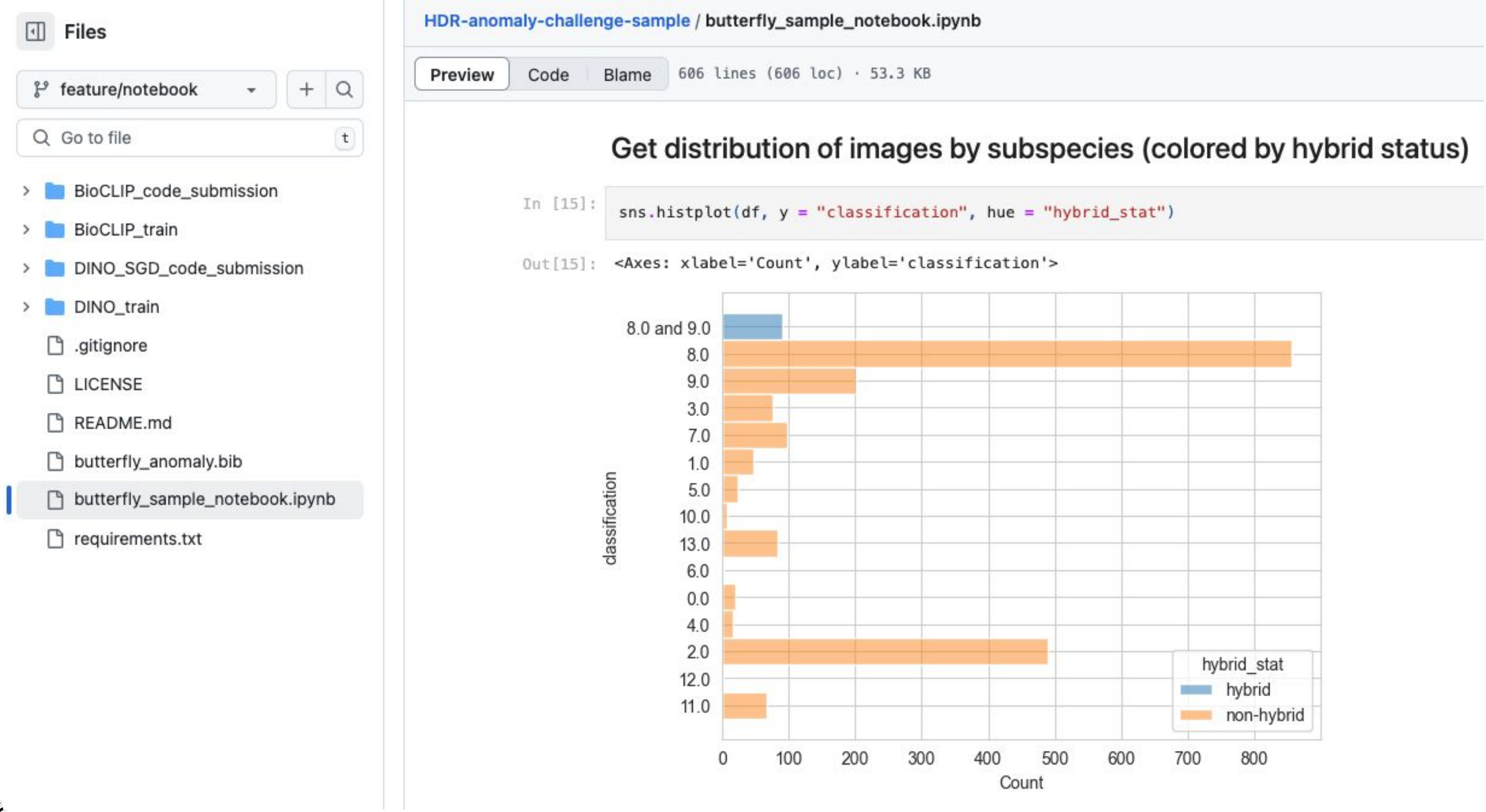


Species B subspecies I



Images are from Zenodo records 2714333, 3082688, 2677821, 2686762, and 2549524, and are licensed under CC-BY 4.0.

Sample Submissions Repository



Files

feature/notebook

Go to file

- BioCLIP_code_submission
- BioCLIP_train
- DINO_SGD_code_submission
- DINO_train
- .gitignore
- LICENSE
- README.md
- butterfly_anomaly.bib
- butterfly_sample_notebook.ipynb
- requirements.txt

HDR-anomaly-challenge-sample / butterfly_sample_notebook.ipynb

Preview Code Blame 606 lines (606 loc) · 53.3 KB

Get distribution of images by subspecies (colored by hybrid status)

```
In [15]: sns.histplot(df, y = "classification", hue = "hybrid_stat")
```

```
Out[15]: <Axes: xlabel='Count', ylabel='classification'>
```

Join the Challenge!

Thank you!

Questions?

Detecting anomalous sea level rise events

Slides created by:

Subhankar Ghosh & Aneesh Subramanian

w/ Shashi Shekhar, Vandana Janeja, Josephine Namayanja

iHARP
NSF HDR INSTITUTE

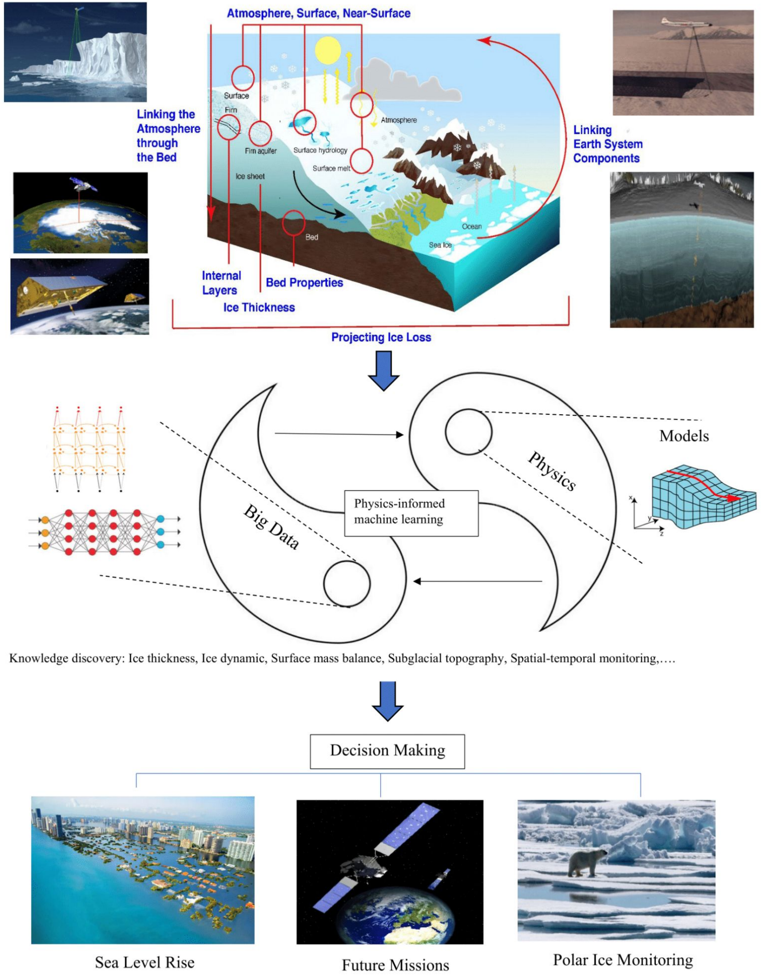
NSF HDR INSTITUTE

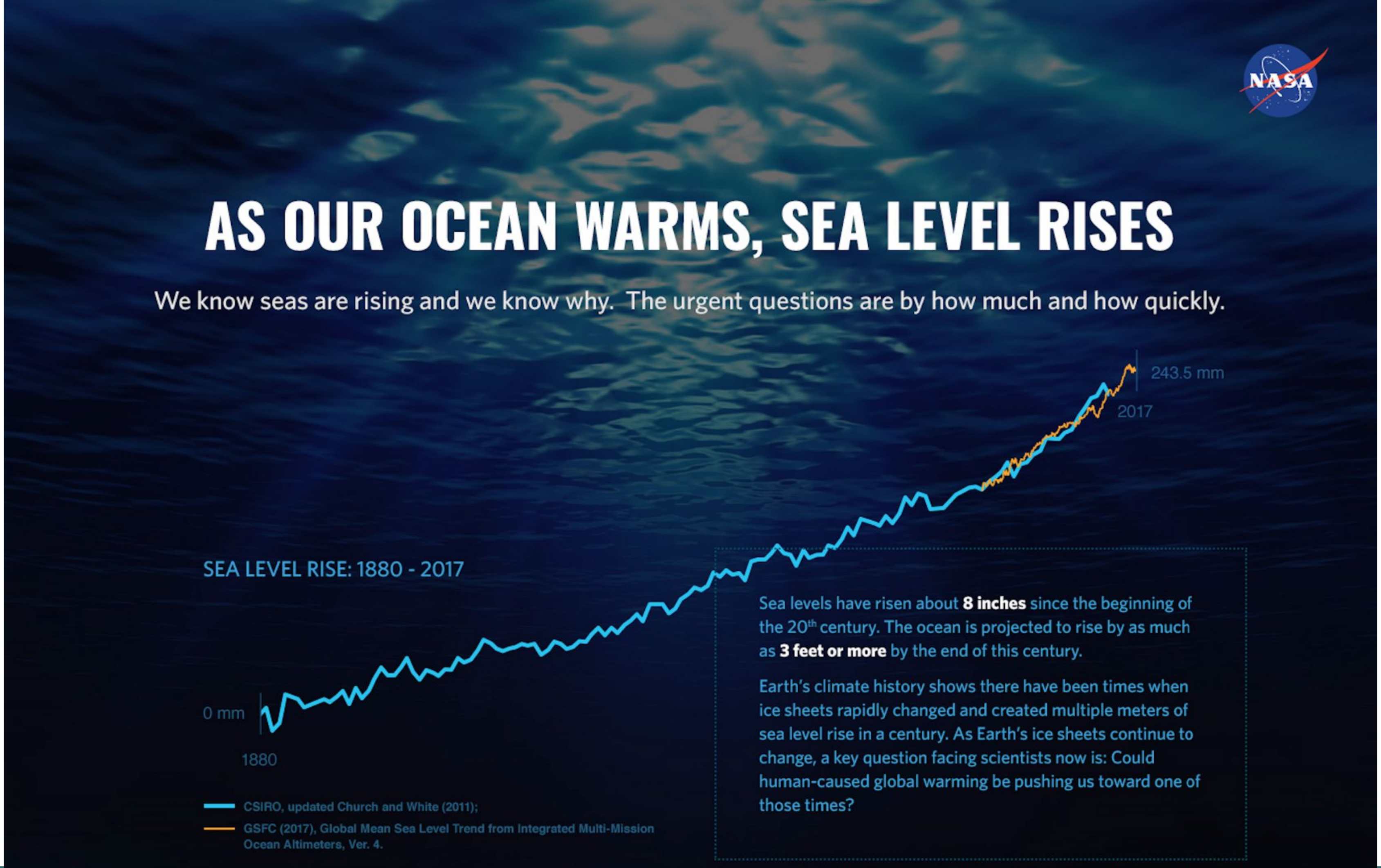
iHARP Vision

iHARP advances our understanding of the response of polar regions to climate change and its global impacts by deeply integrating data science and polar science to spur physics-informed, data-driven discoveries.

iHARP Mission

iHARP conducts data intensive research, education, outreach, and cyberinfrastructure development that will transform understanding of the effects of climate change in polar regions. This institute brings together stakeholders and leading scholars in data science and polar science to reduce uncertainties in projecting Greenland and Antarctica's future mass balance, associated sea-level rise, and impacts on global communities.


Atmosphere, Surface, Near-Surface
Linking the Atmosphere through the Bed
Surface
Firn
Firn aquifer
Surface hydrology
Surface melt
Atmosphere
Ice sheet
Bed
Ocean
Sea Ice
Linking Earth System Components
Internal Layers
Bed Properties
Ice Thickness
Projecting Ice Loss
Models
Physics
Big Data
Physics-informed machine learning
Knowledge discovery: Ice thickness, Ice dynamic, Surface mass balance, Subglacial topography, Spatial-temporal monitoring,….
Decision Making
Sea Level Rise
Future Missions
Polar Ice Monitoring



AS OUR OCEAN WARMS, SEA LEVEL RISES

We know seas are rising and we know why. The urgent questions are by how much and how quickly.

Sea levels have risen about **8 inches** since the beginning of the 20th century. The ocean is projected to rise by as much as **3 feet or more** by the end of this century.

Earth's climate history shows there have been times when ice sheets rapidly changed and created multiple meters of sea level rise in a century. As Earth's ice sheets continue to change, a key question facing scientists now is: Could human-caused global warming be pushing us toward one of those times?

CSIRO, updated Church and White (2011);

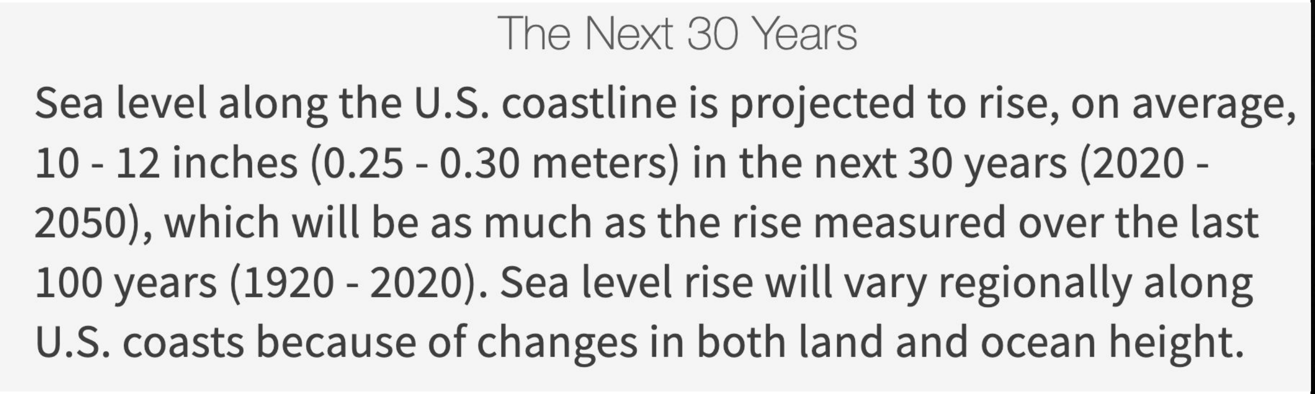
GSFC (2017), Global Mean Sea Level Trend from Integrated Multi-Mission Ocean Altimeters, Ver. 4.

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.

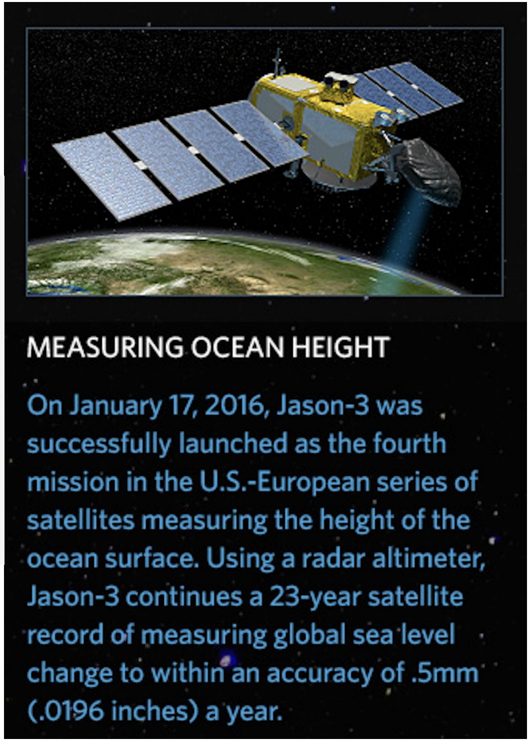
Making Better Predictions of Sea Level Rise

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



The Next 30 Years

Sea level along the U.S. coastline is projected to rise, on average, 10 - 12 inches (0.25 - 0.30 meters) in the next 30 years (2020 - 2050), which will be as much as the rise measured over the last 100 years (1920 - 2020). Sea level rise will vary regionally along U.S. coasts because of changes in both land and ocean height.

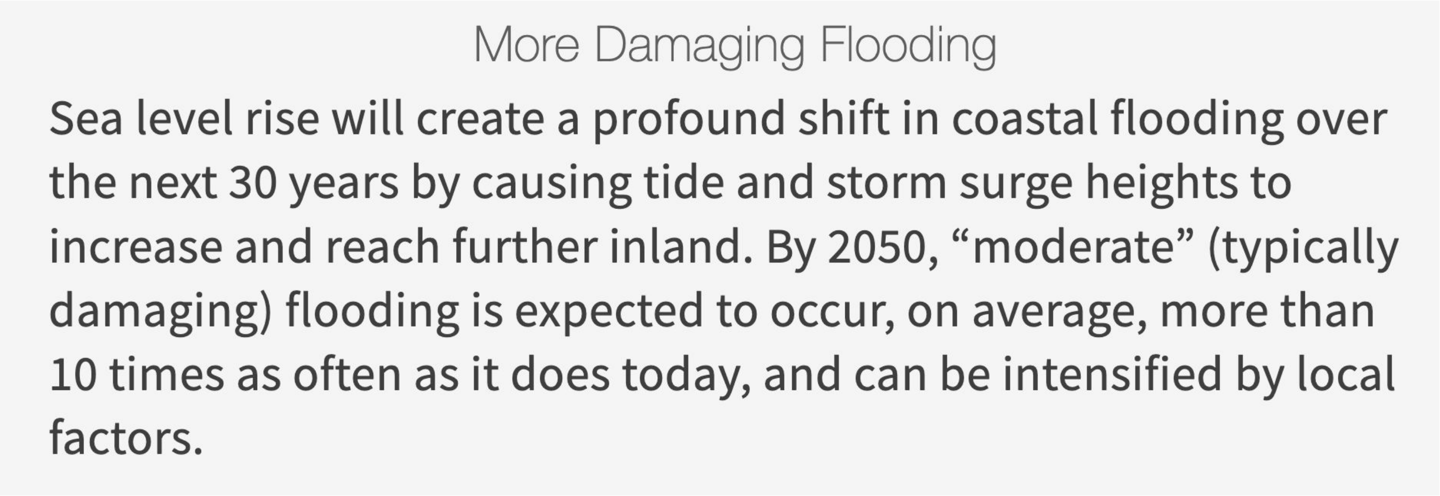


MEASURING OCEAN HEIGHT

On January 17, 2016, Jason-3 was successfully launched as the fourth mission in the U.S.-European series of satellites measuring the height of the ocean surface. Using a radar altimeter, Jason-3 continues a 23-year satellite record of measuring global sea level change to within an accuracy of .5mm (.0196 inches) a year.

Making Better Predictions of Sea Level Rise

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

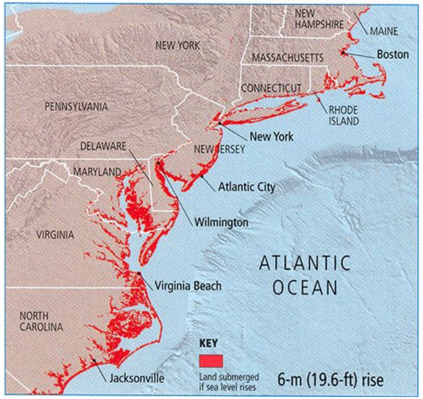


More Damaging Flooding

Sea level rise will create a profound shift in coastal flooding over the next 30 years by causing tide and storm surge heights to increase and reach further inland. By 2050, “moderate” (typically damaging) flooding is expected to occur, on average, more than 10 times as often as it does today, and can be intensified by local factors.

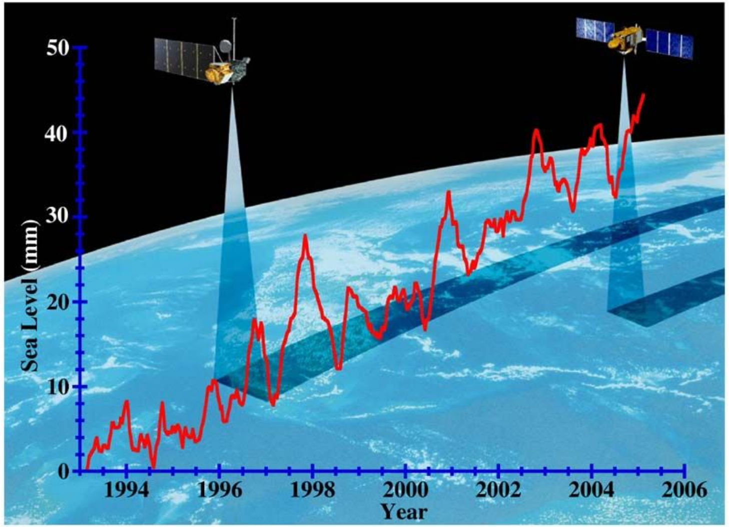
NORTHEAST COASTLINE

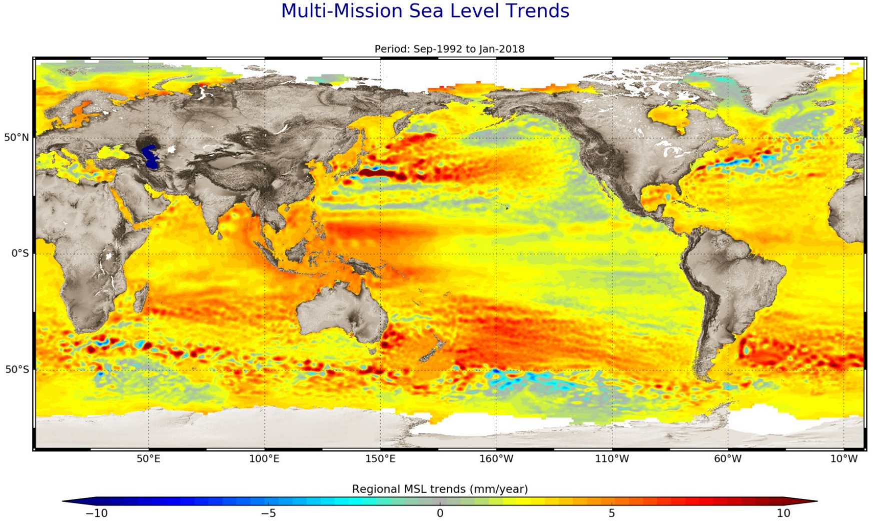
Most of New York City and Boston would be submerged if sea level were to rise by 6 m (19.6 ft).



Continual Tracking

Continuously tracking how and why sea level is changing is an important part of informing plans for adaptation. Our ability to monitor and understand the individual factors that contribute to sea level rise allows us to track sea level changes in a way that has never before been possible (e.g., using satellites to track global ocean levels and ice sheet thickness). Ongoing and expanded monitoring will be critical as sea levels continue to rise.

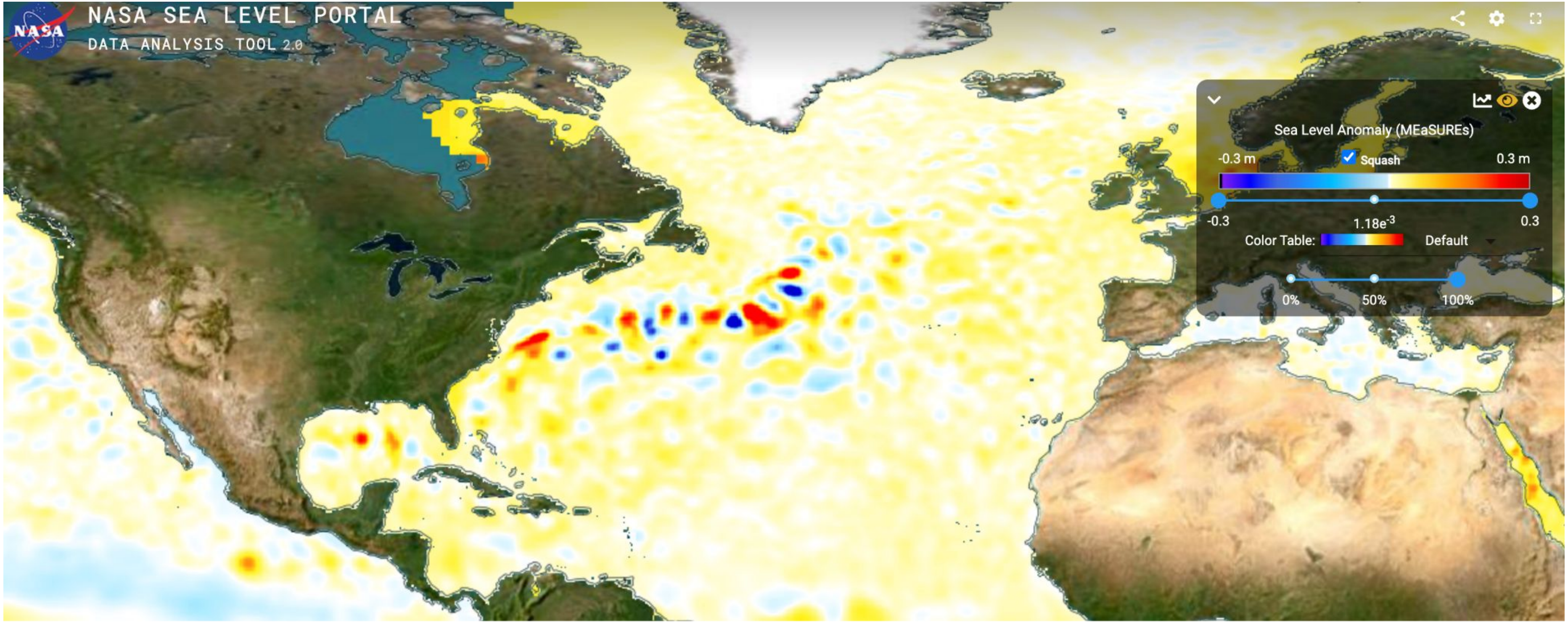




Multi-Mission Sea Level Trends
Period: Sep-1992 to Jan-2018
50°N
0°S
50°S
50°E
100°E
150°E
160°W
110°W
60°W
10°W
Regional MSL trends (mm/year)
−10
−5
0
5
10

© EU Copernicus Marine Service/CNES/LEGOS/CLS, 2018

Machine Learning Challenge: Detect anomalous flooding events from satellite sea level maps



☑ Squash

Color Table: Default

Machine Learning Challenge: Detect anomalous flooding events from satellite sea level maps

- We provide daily satellite sea level anomaly data over the North Atlantic for the past 30 years
- We provide dates of anomalous flooding along 12 US East coast stations for the past 30 years
- Challenge is to detect anomalous flooding events at each station along the US East Coast with the maps of sea level over the North Atlantic

Summary

Gravitational Waves:

Example Submission on Codabench

Butterfly Hybrids:

Example Submission repo

Example model

- Can use this to ensure your submission has the correct format
  - Network reads inputs correctly, returns outputs correctly

iHarp Sea Level Rise Detection:

Instead of submitting the model directly, you will submit a csv file:

"The submission should be a single .csv file where each row should represent a day from 2014 to 2023. The columns should include binary values (0 for false, 1 for true) to predict the presence of an anomaly in each of the stations."

Thank you for attending the hackathon!

In order to match you with others interested in the same challenge, please fill out this survey: