

Phil Harris MIT
Li zhing shi, Northwestern, Data Science
Yu Liang, U Tennessee Chattanooga, Computer Science
Harry Chao, OSU, Computer Science
Mark Neubauer, UIUC, Physics
Melissa Quinnan, UCSD, Physics
Josh Agar, Drexel, Mechanical Engineering

What can we make for a new ML Challenge for the future?

Is there a way to do ecological forecasting of the future?

=> Current challenge uses existing sensory data of lakes, temperatures, concentrations, etc...., the challenge is can predict the parameters of the lakes over time, given various sources

=> Another theme is predicting outbreaks, these are abundances of objects, such as ground beetles, the idea is to predict anomalous appearance of animals, bugs,

Eric: for imageomics out of sample challenge, out of sample could be having computer vision models identify specimens from new sites (train on data at one site, identify species at new site, e.g., for Beetles), which will inevitable have some divergence in their species list (the taxa that could be found there) and/or phenotypic variation that could vary with geography. Could we leverage work and recruit participants from the Imageomics Beeltepalooza to help with this? I think NEON would be excited to support this. Another application could be identifying invasions (e.g., new taxa) at NEON sites.

Anomaly detection problems:

Can we do additional problems, or do we do another theme?

Digital twin simulation?

Mapping simulation with reality?

Foundation model/LLM for domain knowledge?

Uncertainty quantification?

Quantum ML challenge?

More general or there good forecasting problems?

Embedded Long-tail problems?

Few-shot learning problems?

Symmetry based compression?

Out-of distribution extrapolation?

Are there strategies for integrating Meta-data, using Meta data for extrapolation and training

Can we deal with multi-scale/multi-modal/multi-granularity (Meta-data)

Mice-genomics, there is a link between genomics and features. However, there is a lack of benchmark for genomics?

One proposal would be to get a NAIRR pilot to pretrain an model, and then allow people to fine-tune this model

Plan of Action

1. We plan to focus year 2 of the ML challenge on **out-of-distribution problems**
 - a. Out-of-distribution is about forecasting trends
 - i. Extrapolation of expected distributions...
 - ii. Anomaly detection is about finding extremes from forecaste
 - b. We would like to make the data possibly multi-model
 - c. Provide both meta data, and the base data
 - d. We would like to include all institutes
2. In addition, we would like to increase the scope of this work
 - a. We would like to submit a NeurIPS workshop
 - b. We will (again) engage Geoffrey Fox and ML Commons Science
 - c. We plan to request a NAIRR pilot
3. In year 3 of the ML Challenge we are thinking to
 - a. Train foundation models with the NAIRR Pilot
 - b. Motivate a challenge that is fine-tuning on the foundation model
 - i. A proposal is to combine Anomaly detection with out of domain to do uncertainty in ML for scientific discovery
4. In the next 2-3 months, we will organize special meetings to solicit proposal problems
 - a. We want to engage all 5 institutes in the challenge this year
 - b. Since we have an existing challenge software framework and plan
 - i. The challenge prep will take less time
 - c. This will allow us to focus on point 2 about expanding the scope

Join ML Challenge Planning!

Name, email, Institute/affiliation

- Eric Sokol, esokol@battelleecology.org, NEON
- Sanjay Purushotham, psanjay@umbc.edu, University of Maryland Baltimore County,