August 19, 2024 Theory Working Group Call

Attendees: Abby Lewis, Caleb Robbins, Mike Dietze, Jon Borrelli, Shubhi Sharma, Saeed Shafiei Sabet, Jody Peters, Alyssa Willson, Cole Brookson

Agenda:

1. Poll to schedule monthly calls for September to December. Mark your general availability and make sure you timezone is selected.

2. Please submit nominations for EFI2025 Conference session topics and potential speakers by Aug 30.
    a. You do not need to lead any sessions that you propose
    b. There will be time for working groups to get together on May 22
    c. Want to have a presentation about the Theory working group work at the conference

3. Check in with Caleb - Using the NEON Forecasting Challenge to explore predictability across variables and scales
    a. Made connection with Steve Munch about EDM models
    b. GitHub Repo: https://github.com/robbinscalebj/NeonPredictability/issues
    c. Still working on resubmitting forecasts for the chlorophyll targets.
        i. Just need to work with Quinn to get the scores for those resubmitted models
    d. Working on updating code to Abby's forecasts and making things modular
    e. Caleb is working on storing large files and using GitHub Actions - trained ML models that are >100 MB. There is a 100 MB limit for GitHub
    f. If you change 1 line in a 20,000 line. It just changes that 1 line
    g. With binary options if you change a line then you have to restore entire file
    h. Image files, pdf files, netcdf, R data binary files - avoid putting these files in GitHub
    i. Has anyone used the large file storage system?
        i. Cole has played around with it. Didn't like it. Doesn't let you treat the files like other smaller files. The large files are not available to others in the repo like smaller files. If you want to share files with people through GitHub, it doesn't really work. Others can access it, but it doesn't treat it like a file in storage. It is a pointer to the file. When you have the pointer not sure how to access it.
        ii. Think it is worth going in a different direction
    j. Mike suggests sticking the large file in a bucket somewhere. Could do it in something like Google drive, but harder to read/write to compared to a generic bucket.
    k. Lots of Challenge materials are stored in S3 buckets, but the group isn't sure how to use them. Are there tutorials for it?

l. When submitting forecasts you are using buckets.
m. There is a tutorial on cloud basics on EFI2024 conference. Mike can put the slides.
    i. Jody will check with Istem to see if that was recorded.
n. Do need your own bucket to upload to. Is it free?
    i. Probably. This is something to check in with Quinn and Carl to see what the options are
    ii. Probably the easiest thing is to create a Cyverse account
    iii. Cyverse has a lot of material on their website on how to do stuff - they have alot of instructions. The instructions are in bookdown so can be hard to find, but once you find them they are really helpful.
        1. Tyson Swetnam at the U of Arizona is at Cyverse and can help
o. Caleb will take the lead on checking out buckets
p. Collaboration safety on the repo - have been pushing and pulling without forking/branching.
    i. Caleb/Abby had been working on it with separate subdirectories
    ii. More than using branches - suggest using pull requests even for your own files.
    iii. You need to fork the repo, pull to your machine, push back to
    iv. Have your own fork. Then have to GitHub website to create pull request then merge it yourself or set up someone
    v. Still pulling/pushing with your own remote fork, but then submit pull request to merge with main repo
    vi. Think forking is more useful than branches for joint projects

4. ESIIL Idea - would be great to have a form of funding to get some dedicated time and space for the group together to work on these projects
    a. They have funds for working group
    b. There is a call for working group proposals in November
    c. Think it would be good for this group to apply for a working group to work on projects about predictability.
    d. Would give time to sit down and write code together which doesn't work as well for monthly calls
    e. Proposals are 5 pages, due Nov 22.
    f. Have ESIIL team that helps people formulate proposals and another team that evaluates the working groups
    g. 15 Working Groups get funded. Supposed to have ~15 people in the meeting
    h. 2 in-person meetings over 2 years
    i. 3 things for the proposal
        i. Want a big question
        ii. What people work on needs to be of interest to computer science - do you need really big compute? If so, then will help use Cyverse big compute
            1. Goal is to support projects that leverage big compute and move big compute forward

        2. Push science beyond what you can do on a laptop
        3. Don't think this will be a hard box to check
    iii. Are you including new people in the proposal and working group?
j. Is there a limit to the number of people can attend the meetings?
    i. Yes. 15 is the max number of participants to get everyone to Boulder, CO for the 2 meetings. You don't need all 15 people, just need a quorum.
    ii. Don't want groups smaller than 12
    iii. Expect it will be the same group of people instead of rotating off some people from the first meeting and rotating on a few new people
k. Short answer there are funds to bring international people, but it is complicated.
l. There is ~$32,500 available for in-person meetings
m. RFP: 2024 WG Docs to Download
n. This seems like a bit of a deviation from the working group structure of EFI since it locks in a team of 15 people for 2 years that is very Theory group adjacent. Would take capacity away from people in the group to attend the monthly working group calls
o. Write up what you are proposing and send to the Steering Committee
    i. We are working on finalizing the rules on how entities submit proposals
    ii. But keep the EFI Steering Committee group in the loop
p. What you are doing is locking in the 15 people to participate in the in-person meetings.
q. Will need an internal process of deciding if we have 15 people who are interested - do we have more than 15 people and need to finalize. Or do we have fewer than 15 people and need to recruit additional people
r. Don't think this will offend too many people who participate on the working group
s. For participant table, can leave placeholders for types of people to bring in
    i. If you are subbing people in and out you will hopefully bring in people who bring in similar skills
t. Expect there will be 2-6 project leads
u. Get vibe check from the group
    i. Caleb - yes, Mike thumbs up, Shubhi - yes, Saeed thumbs up, Alyssa happy to contribute - can't take on an extra thing in addition to monthly meetings but happy to stay involved and contribute in smaller ways, Jon - would like to contribute
    ii. The 2 projects the group has been working on may be too narrow and may need to broaden to fill in the 2 years. Will probably take more than 2.5 years to figure out the nature of predictability
v. What timeline for sending to Steering Committee
    i. Expect SC will meet during the second week of October
w. Cole to ask - if you get a proposal with 4 student/postdoc co-leads is that a competitive proposal or do you need a permanent faculty position
    i. Cole interested in leading if this works out
    ii. Don't want this to jeopardize what the Theory group is formed to do

   iii. But if it is possible to do it as an EFI supported activity, it would be really cool
   iv. From Mike: diversity of career stages is a great question for ESIIL. Personally I'd love to see this led by early career folk but with the clear blessing and support of more senior EFI folk
x. Start with the core team (Cole, Abby, Caleb, Shubhi) - meet next week

5. Check in with Cole and Shubhi - simulations, weighted permutation entropy and handling data gaps.
   a. Using simulations and NEON data - can we pull info from time series using weighted permutation entropy and realized permutation entropy - use it to look at how predictable the time series are
   b. 2 things -
      i. Simulate from models and permutation entropy
      ii. Looking at the NEON data
   c. How does data gaps affect understanding of predictability

6. Marcus - machine learning forecasts and NEON Forecasting Challenge.

7. Blog post idea for code review materials (Jody)
   a. On January call the group talked about Jody drafting the blog post and running it by the group. There is no definite timeline for this, but hopefully within the next month or two
      i. SORTEE group has a lot of energy between code review and seems like there is overlap with EFI
      ii. SORTEE: https://www.sortee.org/
      iii. Has Slack channel - here is the link to join
      iv. Library of code mistakes:
         1. Can add anonymously issues that people have found when their code has been reviewed
         2. E.g., day/time errors, misunderstanding of function
         3. It is structured with the same headings that were in the paper on code review