Minutes for Liaison Meeting 9/12

1. Liaisons discussed and asked whether we want to talk about waveform data or whether we should prioritize Statement of Work
2. Mehrezat introduced discussion on Statement of Work discussion
   1. Clarification of terms:
      1. Model: The machine learning program that will automatically identify waveforms.
      2. Interface: The way the program looks and is navigated by a user, e.g. an executable with visuals and buttons.
   2. What does the interface, or user side, look like?
      1. Dr. Backus and Dr. Reif prefer a usable program separate from the engineering team’s software.
   3. Defining success: what is an acceptable accuracy and for what species?
      1. Dr. Reif asks to hear about our perspective on what is possible, and specifically how we intend to approach the data (in terms of which species to start with): Does the species matter to the product or is a waveform a waveform?
      2. Dr. Backus states 90% is not usable: if this is the limit, the model must be very trainable.
      3. Aphid data is 20 years old, it is imperfect. There were no “families” or “subtypes” then.
   4. Goals, and primary goals vs. stretch goals
      1. Dr. Backus proposes a primary goal of 90% accuracy and a stretch goal of reaching 98%.
      2. Dr. Cooper proposes a primary goal of 95% accuracy at the family level. If we can proceed to identifying types, that would be a very useful stretch goal. Consistency is more important than accommodating all of the human “styles” of annotation.
      3. Dr. Backus responds to say that aphids and sharpshooters do not have such clear differences at the family and type levels, and those goals are very subjective.
      4. Professor Hope suggests that, due to the subjectivity, perhaps it is not useful to write a program that measures specific metrics, but perhaps it could be more of a suggestive model that can remedy weaker data portions.
      5. Dr. Cooper would love if the program would give an uncertainty level, if it could “tell us where it is uncertain, and we could tell it where we are uncertain as well”
      6. Milo reiterates that we may not be able to even reach 90% accuracy in our model. The accuracy that we expect will be the accuracy that we train on. We will have no idea what our success will be until we sink our teeth into it.
      7. Dr. Backus concludes that focusing on training might be more productive.
   5. Data
      1. Zach asks whether there is any data that has been annotated by more than one person; Dr. Cooper proposes the “golden standard” which is scored by many prospective annotators.
      2. Dr. Reif says that it is sufficient for the model to identify garbage data, it is useful to know where that is.
      3. ~~Dr. Backus~~
   6. On Dr. Reif’s question of what classifications matter to us, does species matter, is a waveform a waveform?
      1. Professor Hope answers: it is unlikely that we would be able to have one model that can handle multiple species.
      2. But it might be helpful to look at different species because we might learn about how to develop a model for this program on an “easier” set of data, and transfer those lessons towards building our final model.
      3. Dr. Reif responds: We need to try to balance everyone’s needs, aphid data is larger but is coarse. The mosquito data has a subtype aspect. She asks that we look at both species and come back with a more informed perspective on what the best approach is. Reminds of the obligation to funders to complete the mosquito model.
      4. Prof. Hope states that, if we determine that the aphid data is so different that it is not helpful to our goal of measuring blood-feeding arthropod data, then we will not pursue it. But we will need to explore, because the aphid data as a larger set could be useful for initial exploration of the problem.
   7. Milestones
      1. Tests, scheduled for our site visit (in March); we will need a very workable version. (Deadline for a usable program!)
      2. Dr. Reif bounces the question back, we will need to set our technical milestones once we know our data and can estimate.
      3. Report deadlines for funders are not until the summer.
      4. Dr. Hope suggests that we can still lay out very high-level milestones
   8. Receiving the data
      1. Resolved confusion on data size (misstated as 6.7 TB instead of GB) which caused concern about feasibility of sharing virtually.
      2. 6.7 GB of data
      3. We will receive data via OneDrive
   9. Explanation of WinDaq data by Dr. Backus
      1. Demonstration of data compression (how much data is shown on the screen)
      2. Description of waveform and labeling
         1. An initial spike in the waveform signifies an insertion of stylus
         2. Large spikes are the initiation of probing
         3. Shorter, compact spikes are various probing behaviors
         4. Baseline at 0 (= non-probing)
         5. WinDaq places the start of the recording in the middle of the screen the first time it is opened
         6. Left and right mouse buttons can amplify/deamplify the waveform on the y-axis
         7. Strange, thick, negative spike pointed out as “something strange that doesn’t have a name,” an error
         8. Description of specific wave shape labeling
      3. Clarification of column headings
         1. “Insect” heading = treatment, species description
         2. “Insect\_num” = date and channel
         3. “Waveform” = waveforms: families: P = pathway, Z = baseline, D = drop (see handout on aphid waveforms)
         4. “Waveform\_duration” = duration of classification (=an individual waveform event
         5. “TBF2” = time since beginning of file for the end of each waveform event