Machine Learning Competitions for G2P and End of Season Phenotype Prediction

Saint Louis University
Street Address: 1 N Grand Blvd, St. Louis, MO 63103
Postal Address: 3700 West Pine Mall, Fusz Hall, St. Louis, MO 63108
Lead PI: Abby Stylianou, Assistant Professor
(314) 977-2678, abby.stylianou@slu.edu

Administrative Point of Contact:
Matt Renaud
(314) 977-7742
grants@slu.edu

Agricultural Genome to Phenome Initiative (AG2PI)
2021 Seed Grants

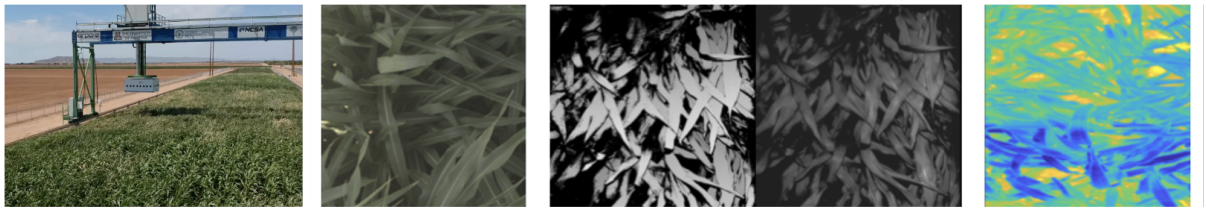

Figure 1: Data collected by the TERRA-REF Field Scanner System in June 2018. (a) An RGB image from the stereo RGB camera. (b) The depth (left) and reflectance (right) images from the laser 3D scanner. (c) A thermal image captured by the FLIR camera.

1 Objectives & Aims

In this project, we will organize competitions to engage with the broader machine learning community to produce models that can answer phenomic questions, using curated datasets from the Department of Energy ARPA-E Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform (TERRA-REF) program [Burnette et al., 2018]. The TERRA-REF program aimed to transform plant breeding by using remote sensing from a state of the art field scanner gantry system, seen in Figure 1 (top) to increase the speed at which plant traits can be measured. The field scanner includes a number of sensors, including a millimeter resolution laser 3D scanner, high resolution stereo-RGB cameras, multiple hyperspectral sensors, and a thermal camera, among others (example data products are shown at the bottom of Figure 1). Over the course of several seasons, this sensor collected over a petabyte of sensor data for bioenergy sorghum lines, their corresponding genetic data, a large volume of ground truth measurements of plant phenotypes and growing conditions, and a baseline set of algorithmic approaches for extracting phenotypic data.

We propose that structured machine learning contests are an excellent way to share data and ensure that the desired scientific questions are actually the ones that are answered. Contests, where a specific problem is shared, training and testing data are provided, and a specific evaluation protocol is defined, are a frequent and popular means of advancing results to specific questions within the machine learning community. For example, the Fine-Grained Visual Classification community hosts annual contests on difficult visual recognition problems, such as the iNaturalist contest to recognize different species in image data, which had 1,477 submissions last year from 249 different competitors [Van Horn et al., 2018]. These communities are hungry for well organized datasets with specific scientific questions.

In the following section, we define the three contests that will be hosted and promoted at relevant workshops at machine learning and computer vision conferences. In particular, we have worked with the organizers of the Fine-Grained Visual Classification workshop, hosted at the annual IEEE/CVF Conference on Computer Vision and Pattern Recognition, and the organizers of the Workshop on Computer Vision Problems in Plant Phenotyping and Agriculture (CVPPA), hosted this year at the annual International Conference on Computer Vision, to have support to host one of our competitions at each of the workshops. Below, we specify for each competition the question that the agricultural domain experts are seeking to have answered, the dataset that will be provided, and the contest and evaluation approach that defines the machine learning development that will be conducted. Unless otherwise specified, the training dataset for a contest will consist of data from one of the two plots grown for a particular cultivar in a season, while the testing set will consist of images from the other plot, in order to validate that models are not over-fitting to non-generalizable features (e.g., static ground features).

1.0.1 Contest 1: Cultivar Classification

Domain Expert Question: "What cultivar is shown in this image?" Predicting the cultivar in an image is an especially good challenge problem for familiarizing the machine learning community with the data that will be part of these ongoing contests. A high accuracy machine learning predictor of the species captured by the sensor data could be used to determine where errors in the planting process may have occurred. For example, seed may be mislabeled prior to planting, or planters may get jammed, depositing seeds non-uniformly in a field [Sharma et al., 2019]. Both types of errors are surprisingly

common and can cause major problems when processing data from large-scale field experiments with hundreds of cultivars and complex field planting layouts. **Dataset:** The dataset will include RGB imagery from the 2017 season labeled by cultivar. **Contest/Evaluation:** Contest participants will report cultivar classification accuracy for test sets that include: the full season; one week at the start of the season; one week mid-season; and one week at the end of the season.

1.0.2 Contest 2: Genetic Marker Prediction

Domain Expert Question: "Does this cultivar have a particular genetic mutation?" Figure 2 shows an example image of a particularly waxy sorghum leaf. From the crop breeding perspective, bioenergy sorghum cultivars that have high levels of this epicuticular wax have improved water use efficiency and greater yields [Uttam et al., 2017]. This is via a higher gas exchange ratio and regulation of water loss at the leaf surface. Rapid and automated identification of desirable waxy cultivars in a large diversity panel (such as the BAP) would help to accelerate the breeding of this trait into commercial varieties. This epicuticular wax phenotype is controlled by a (known) genetic marker, and is the sort of visual information that a machine learning algorithm might be able to model. It is interesting to understand how accurately different genetic markers like this can be predicted based on visual data. **Dataset:** The dataset will include RGB imagery from the 2017 season labeled by cultivar and genetic marker, for ten genetic markers known to affect leaf waxiness, dwarfing and drought tolerance. 80% of the cultivars will be in the training set. **Contest/Evaluation:** Contest participants will report accuracy on predicting genetic marker labels across the 20% of the cultivars held out to be the test set.

Figure 2: (An especially waxy sorghum plant.

1.0.3 Contest 3: End-of-Season Phenotype Prediction

Domain Expert Question: "How early in the season can we predict if a crop will have X trait?" Breeding programs in particular can be made more efficient by predicting end of season traits as early as possible. **Dataset:** The training data will consist of 3D-scanner data from early in the 2017 season from 80% of the cultivars, labeled by end-of-season phenotype measurements (e.g., plant height or leaf length). **Contest/Evaluation:** Participants will report on predicted plant height and leaf-length measurements for the remaining 20% of the cultivars based on early season measurements.

Planned Competitions. We will be running the End-of-Season Phenotype Prediction Competition at CVPPA 2021, with a goal of a second related competition in 2022, and will be running the Cultivar Classification and Genetic Marker Prediction Competitions at FGVC 2022.

Promotion. The goal is to initially focus on getting machine learning and computer vision communities involved in answering questions of interest for the plant phenotyping and G2P communities. To that end, promotion and advertisements are going to primarily be focused on those communities, through hosted competitions at relevant workshops. Additionally, PI Stylianou is currently the Social Media Chair for the IEEE/CVPR Computer Vision and Pattern Recognition (CVPR) conference and the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), and through those roles has become quite integrated in social media circles relating to computer vision and machine learning activities. PI Stylianou will focus on disseminating competition information (both competitions proposed in this project, and competitions being run simultaneously at the related workshops) through those relevant social media communities. Finally, PI Stylianou regularly gives talks both in the plant phenotyping and computer vision communities, and intends to spend time in such talks discussing these competitions, the findings that come out of them, and more generally the possibilities of leveraging engaged communities through curated competitions to answer specific scientific questions.

Applicability to Animal Sciences. While the focus of the competitions proposed in this project is on plant phenotyping and plant agricultural G2P, we are extremely excited about the general use of

curated machine learning competitions to answer scientific questions. It is our intent to develop best practices for machine learning competitions for the general AG2PI community; such best practices would be applicable for both plant and animal science questions. One of the communities that we will be working with, the Fine-Grained Visual Classification Community, has an extended history of working on animal classification, counting and behavior description in camera trap data. We are excited to work with the managers of those competitions to ensure that our proposed best practices are applicable to animal science and behavior questions.

2 Furthering the AG2PI Aims

The AG2PI aims to understand the interaction between genetics, the environment and plant phenotypes, and expresses that developments in these areas will require, among other things, better mechanisms for data sharing, the development and application of novel data science approaches, and better sharing of best practices. In this project, we will share curated agricultural datasets and specific questions in plant phenotyping and the relationship between genetics and observed traits through structured machine learning contests. The contests will engage a new community of machine learning practitioners, researchers and data scientists in developing novel algorithmic approaches for addressing the proposed questions. Developed models will provide interesting answers and insights for the posed questions, and will potentially be generalizable to other questions in agricultural phenomics and G2P. Moreover, the experience of running these contests and interacting with competitors will provide opportunities to understand how to best engage with the machine learning and data science communities in answering related questions. It is our aim that through iterating on these contests, soliciting feedback from competitors on their experience and from plant biologists on the quality of the developed models, we will be able to provide a set of best practices for others working in this space to organize data, pose questions and create similar competitions.

3 Evaluating Project Success

Project success will be determined both by the level engagement in the machine learning competitions and through the quality of the models provided by competitors. Well-performing models will be evaluated for quality (how well they answer the proposed scientific questions), generalizability (what other phenotyping and G2P questions the proposed models might be suitable for), reproducibility (whether the approaches could be repeated for both the proposed question or related questions), and explainability (whether the decisions or estimations made by the model are understandable by humans) by both the computer scientists on the proposing team, and biologists from the TERRA-REF and ARPA-E OPEN teams that have been working with this data. We have included a letter of support from the lead for those projects, Dr. Todd Mockler of the Donald Danforth Plant Science Center.

4 Expected Outcomes and Deliverables

The goals for this project are to produce:

- Engagement and interaction with the machine learning and data science communities through structured machine leanring competitions;
- Deployable machine learning models that address several interesting plant phenotyping and G2P questions; and
- A template or set of best practices for deploying additional machine learning challenges that relate to the AG2PI's goals.

Through running several competitions with specific plant phenotyping and G2P questions at their center, we will have multiple opportunities to produce useful models and answers to the important scientific questions posed in Section 1, to learn from competitors about challenges that arise in working with agricultural sensor and genetic data, and to make suggestions for others in the G2P space on how to best utilize structured machine learning competitions to answer their questions.

5 Qualifications

PI Abby Stylianou is an early career faculty member in Computer Science, and has worked for the last several years under the TERRA-REF and subsequent ARPA-E OPEN projects to use machine learning to answer a number of interesting questions about the bioenergy sorghum grown under the TERRA-REF gantry platform. These questions include how to best measure plant phenotypes from this system (e.g., "how tall is this sorghum plant?" or "how long are its leaves?"), how to predict end-of-season phenotypes from early-season measurements, how to generalize approaches to work for both greenhouse and field data, and how to predict phenotypes across environmental conditions. Stylianou publishes extensively in the areas of computer vision and machine learning and is leading the Hotel-ID 2021 competition at the Fine-Grained Visual Classification workshop at CVPR 2021 (the proposed venue for a publicized competition in 2022). Stylianou has also presented at NAPPN 2021 about explainable machine learning models trained to predict the presence or absence of genetic mutations in the Bioenergy Association Panel of sorghum grown under the TERRA-REF gantry.

Madison Pope, the Graduate Research Assistant who will be funded by this project, has worked extensively in the Baxter Lab at the Donald Danforth Plant Science Center as a lab tech, and is now pursuing her M.S. in Bioinformatics and Computational Biology, with a research focus on machine learning for plant phenotyping. Madison is excited to understand the ways in which curated machine learning competitions can be used to answer scientific questions. Madison will not simply be running the competitions, but will be interacting with competitors and scientific end-users to understand what best practices are both to encourage competitors to participate and engage at a non-superficial level, and to produce models that end-users can actually work with.

6 Timetable of Activities

The first competition (on end-of-season biomass prediction) will be launched within the first month of the project, aligned with the CVPPA Workshop. We will additionally launch first iterations of the Cultivar Classification and Genetic Marker competitions, which are being presented at FGVC 2021, and will be hosted FGVC competitions in 2022. Competitions will each run for four months. At the end of each contest, the top performing results will be evaluated for reproducibility. Top performing will be surveyed to understand their experiences with the datasets and scientific questions. It is also often the case that contestants in these sorts of contests will participate in not just one contest, but many contests from a contest provider if they have established that the provider provides clean datasets with clear questions and evaluation approaches. We intend to run each competition multiple times, incorporating feedback from both competitors and scientific end-users in the design of follow on competitions.

7 Engaging Scientific Communities & Under-represented Groups

This proposal is explicitly aimed at engaging a new scientific community – the machine learning community, and those in the plant phenotyping community who are interested in machine learning – in addressing questions relating to plant phenotyping and the genome to phenome relationship. Publicly accessible competitions which carefully curated data and clearly stated questions are an excellent way of engaging a broader community in developing models that are actually usable in the scientific domain of interest. In Kaggle competitions in particular, well-performing teams are often excited to both be engaged with the scientists posing the questions, and in providing publicly accessible write-ups of the ideas that led to superior performance in the competition.

PI Stylianou's lab has a record of mentorship for under-represented students, especially women in computer science. Women make up more than half of the undergraduate and graduate student researchers in the lab (in C.S., the average typically falls closer to 25%). PI Stylianou's lab has an emphasis leveraging engaged communities in addressing problems ranging from plant phenotyping to social justice and human trafficking, and the PI supports efforts at broadening the participation in computing, including by speaking at the National Center for Women & Information Technology and acting as a mentor at the recurring Computer Science Institute for Women at the Donald Danforth Plant Science Center.

References

[Burnette et al., 2018] Burnette, M., Kooper, R., Maloney, J. D., Rohde, G. S., Terstriep, J. A., Willis, C., Fahlgren, N., Mockler, T., Newcomb, M., Sagan, V., Andrade-Sanchez, P., Shakoor, N., Sidike, P., Ward, R., and LeBauer, D. (2018). TERRA-REF data processing infrastructure. In Sanielevici, S., editor, *Proceedings of the Practice and Experience on Advanced Research Computing, PEARC 2018, Pittsburgh, PA, USA, July 22-26, 2018*, pages 27:1–27:7. ACM.

[Sharma et al., 2019] Sharma, D., Rane, J., Singh, R., Gupta, V. K., and Tiwari, R. (2019). Comparison of different planting methods to determine the precision of phenotyping wheat in field experiments. In *Advances in Plant & Microbial Biotechnology*, pages 77–83. Springer.

[Uttam et al., 2017] Uttam, G. A., Praveen, M., Rao, Y. V., Tonapi, V. A., and Madhusudhana, R. (2017). Molecular mapping and candidate gene analysis of a new epicuticular wax locus in sorghum (sorghum bicolor l. moench). *Theoretical and Applied Genetics*, 130(10):2109–2125.

[Van Horn et al., 2018] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.