AFRL-RI-RS-TR-2021-158

MODSQUAD, THE PURDUE DATA-DRIVEN DISCOVERY OF MODELS TEAM

PURDUE UNIVERSITY

SEPTEMBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-158 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
PETER A. JEDRYSIK	JULIE BRICHACEK
Work Unit Manager	Chief, Information Systems Division
	Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government’s approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* SEPTEMBER 2021	**2. REPORT TYPE** FINAL TECHNICAL REPORT	**3. DATES COVERED** *(From - To)* OCT 2016 – MAR 2021

4. TITLE AND SUBTITLE MODSQUAD, THE PURDUE DATA-DRIVEN DISCOVERY OF MODELS TEAM	**5a. CONTRACT NUMBER** FA8750-17-2-0111
	5b. GRANT NUMBER N/A
	5c. PROGRAM ELEMENT NUMBER 62702E
6. AUTHOR(S) William Cleveland, Wen-wen Tung, Jeffrey Baumes, Roni Choudhury, Ryan Hafen, and Curtis Lisle	**5d. PROJECT NUMBER** D3MP
	5e. TASK NUMBER S0
	5f. WORK UNIT NUMBER 04
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Purdue University 250 N University Street West Lafayette Indiana 47907	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RISB 525 Brooks Road Rome NY 13441-4505 DARPA/I20 675 N. Randolph St. Arlington VA 22203-2114	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/RI
	11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2021-158

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our team consisted of both Statistics and Software Design experts. Our involvement in the DARPA Data-Driven Discovery of Models (D3M) program included developing a TA3 system, which we entitled ModSquad, participating in joint standards design meetings, presenting at the project workshops on a variety of topics, and investigating the TERRA-REF datasets and challenge problems, resulting in an interactive data exploration and model fitting interactive interface for several TERRA-REF datasets, and finally culminating with an analysis of state-of-the-art atmospheric research data. We feel that we accomplished the tasks assigned to our team to the benefit of the overall program and provided a high return on investment for DARPA, AFRL, and the D3M program leadership.

15. SUBJECT TERMS

Machine Learning, Statistics, TERRA-REF

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** **PETER A. JEDRYSIK**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	57	**19b. TELEPHONE NUMBER** (*Include area code)* **N/A**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

The DARPA Data-Driven Discovery of Models (D3M) program facilitated the development of Automatic Machine Learning model systems and corresponding User Interfaces to train and invoke these models. Our team focused on User Interface development and team membership consisted of both Statistics and Software Design experts. Our involvement in the D3M program included developing a user interface, which we entitled ModSquad, participating in joint standards design meetings, presenting at the project workshops on a variety of topics, investigating the TERRA-REF datasets and challenge problems, resulting in an interactive data exploration and model fitting interactive interface for several TERRA-REF datasets, and finally culminating with an analysis of state-of-the-art atmospheric research data. We feel that we accomplished the tasks assigned to our team to the benefit the overall program and provided a high return on investment for DARPA and the D3M program leadership.

2.0 INTRODUCTION

The Data-Driven Discovery program was initiated to refine the state-of-the-art in automatic machine learning model development. Prior to this program, the majority of machine learning models were developed by data scientists who had to accept data from a domain scientist, clean-up or transform the data by hand or using custom programming scripts, and then fit one or more statistical models to the data. This fit process was generally done using custom programming in the R, python, or Java languages. The primary goal of the D3M program was to develop software systems capable of automating the procedures described (preparing data, fitting models, measuring the accuracy of model fits). The program goals were divided into several different Technical Areas, which were given numbers and abbreviations, such as TA1, TA2, TA3. Subdividing a problem into several technical areas and assigning teams to each task area is a common technique used in DARPA-led programs. For this program, Technical Area One (TA1) referred to the development and packaging of specific algorithms using a common interface, so they could be called from an automated system. Example TA1 algorithms include outlier detection, data normalization, data type determination, and statistical models. Technical Area Two (TA2) comprises development of the integrated machine learning platforms that receive data and fit models, using the algorithmic components from TA1. Finally, Technical Area Three (TA3) refers to the User Interface development, which allows a human user to interact with the overall system.

After the project summary information, this report is divided into sections. Section 3 covers the design methods for our ModSquad user interface and the TERRA-REF datasets, Section 4 describes the results of using ModSquad for live data modeling, the results of our analysis of TERRA-REF data, and finally covers our use of D3M technology to analyze a state-of-the-art research dataset in Statistical Climate Science -- research on Atmospheric Rivers performed at Purdue.

2.1 TEAM RESEARCH GOALS

Our original goals were to innovate through the development of novel data science interactive interfaces based on our team's strong prior experience developing interfaces and running data analyses on a wide variety of problems (remote sensing, time sequence volume analysis, geospatial dataset exploration, and others.). We accomplished a lot, but weren't able to strongly address these research goals during the course of our participation in D3M largely because of the emphasis by DARPA on early software integration and our limited budget. We feel the standardized TA2/TA3 protocols are helpful in the final architecture but required integrating with a constantly changing execution environment early in the program. The rigorous testing that was applied early in the program before stable standards were in place required a lot of engineering integration hours.

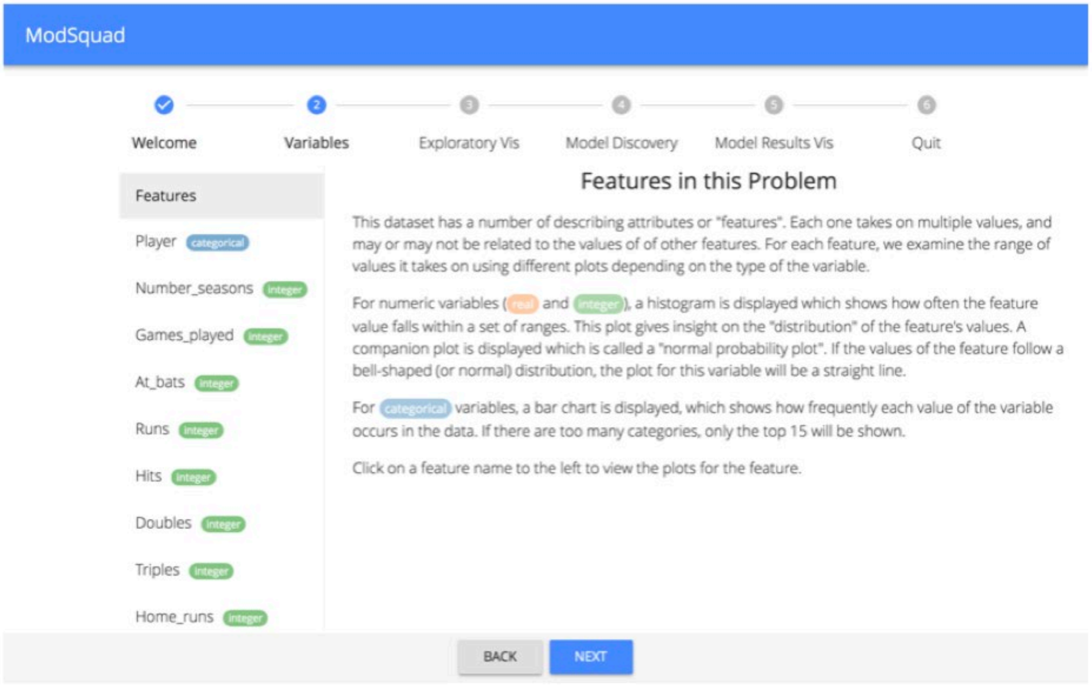
That said, we understood and appreciated the need for standardization of the software process and environment to yield an integrated final product. The D3M project appears on target to produce an integrated primitive library and a set of user interface and model development applications that will hopefully become a popular addition to the open source data analysis tools currently available. In hindsight, we do feel that if the strict integration standards had been enforced a bit later in the program, some Performer teams would have been able to insert more research goals into their program deliverables and further increase the novelty of the overall final D3M product. We recommend that DARPA consider this suggestion in the execution of later software development programs.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

In this section, we cover the methods, developments, and procedures the Purdue team performed during the course of this effort. All members of our team, but particularly Dr. Cleveland and Dr. Hafen, have experience partnering with domain scientists from different communities, listening to their scientific goals, and performing supportive data science activities to enable progress towards the stated scientific goals. Because of this prior expertise, our Purdue team developed a user interface that was focused on engaging with a domain-scientist who understood her or his problem, without requiring much understanding of statistics or being even familiar with the machine learning community vocabulary. We implemented the first steps of a visual diagnostics process that Dr. Cleveland uses when first meeting a new dataset and trying to understand what is in the dataset. Most datasets contain multiple independent variables and one or more dependent variables. To increase the engagement of users, our interface implements a multi-step process and shows the user constantly where they are in the process through a navigation bar always displayed along the top of our system interface. The navigation bar is visible along the top of Figure 1, showing that they have completed the Welcome step and are currently in the “Variables” step.

3.1 EXPLORING FEATURES ONE AT A TIME.

It makes sense to first show the user what variables are present in their dataset and allow the user to then explore each variable by itself. The goal of this first step is to let the user become familiar with the values each individual variable takes on throughout the dataset. A subject matter expert, who knows what values should be present, may be able to identify outliers or data collection errors in this initial step. This part of our User Interface is illustrated in Figure 1. Along the left column, each dataset variable is listed along with its datatype. We used color coding to visually indicate the data type — providing a visual cue to speed understanding. The user simply selects one of the variables and is then presented with either a histogram and an all-values plot (if the variable is continuous) or a bar chart of the value counts (if the variable is categorical). Figure 2 shows the display after a continuous variable has been selected. In this example, it is the number of “At bats” for a dataset of baseball players.



Features in this Problem

This dataset has a number of describing attributes or "features". Each one takes on multiple values, and may or may not be related to the values of of other features. For each feature, we examine the range of values it takes on using different plots depending on the type of the variable.

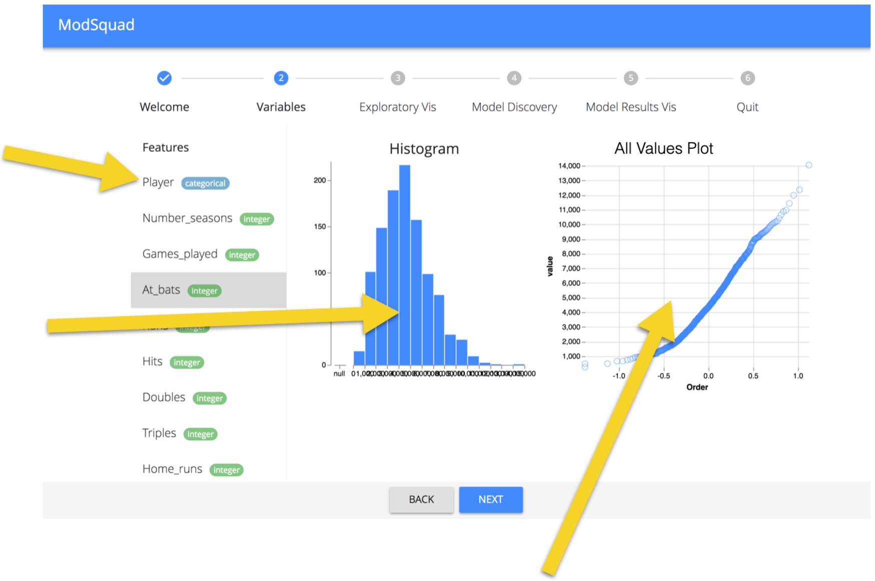
For numeric variables (real and integer), a histogram is displayed which shows how often the feature value falls within a set of ranges. This plot gives insight on the "distribution" of the feature's values. A companion plot is displayed which is called a "normal probability plot". If the values of the feature follow a bell-shaped (or normal) distribution, the plot for this variable will be a straight line.

For categorical variables, a bar chart is displayed, which shows how frequently each value of the variable occurs in the data. If there are too many categories, only the top 15 will be shown.

Click on a feature name to the left to view the plots for the feature.

Figure 1 Variable Exploration Interface

The “all values” plot is not among the most used plots among young and upcoming data science and visualization scientists. However, it provides unique insight to both a domain scientist, who understands what the variables values should look like and to a novice who is trying to understand a dataset for the first time. In plotting all the data ordered by increasing value, the user is quickly shown if there are gaps in the values taken on by a variable in question, and if there are outliers (substantially different values from what the variable usually assumes).   Someone unfamiliar with a dataset can begin forming an effective conceptual understanding of each variable using this plot as an initial exploration tool.



Click on a feature along the left column

The histogram counts the number of values for a particular feature (or variable) that land in each "bin". This shows an approximate distribution of how the feature values vary across the entire dataset

The All-values plot shows all the values this feature takes on in the dataset. It shows whether the values vary: discretely, continuously, or with some unusual range or pattern.

Figure 2 ModSquad Single Variable Exploration

It was mentioned previously that categorical variables are summarized in the exploration interface by generating a bar graph showing the number of times the variable under observation takes on each of its different values. An example of this, again using the baseball dataset, is shown in Figure 3. Here the "position" variable has been selected (see the highlighted grey region which is positioned when the user selects a variable). The separate values ('outfield', 'catcher', etc.) form the categorical (x) axis of the bar chart. This method allows the user to understand the distribution of values taken on by any categorical variable.

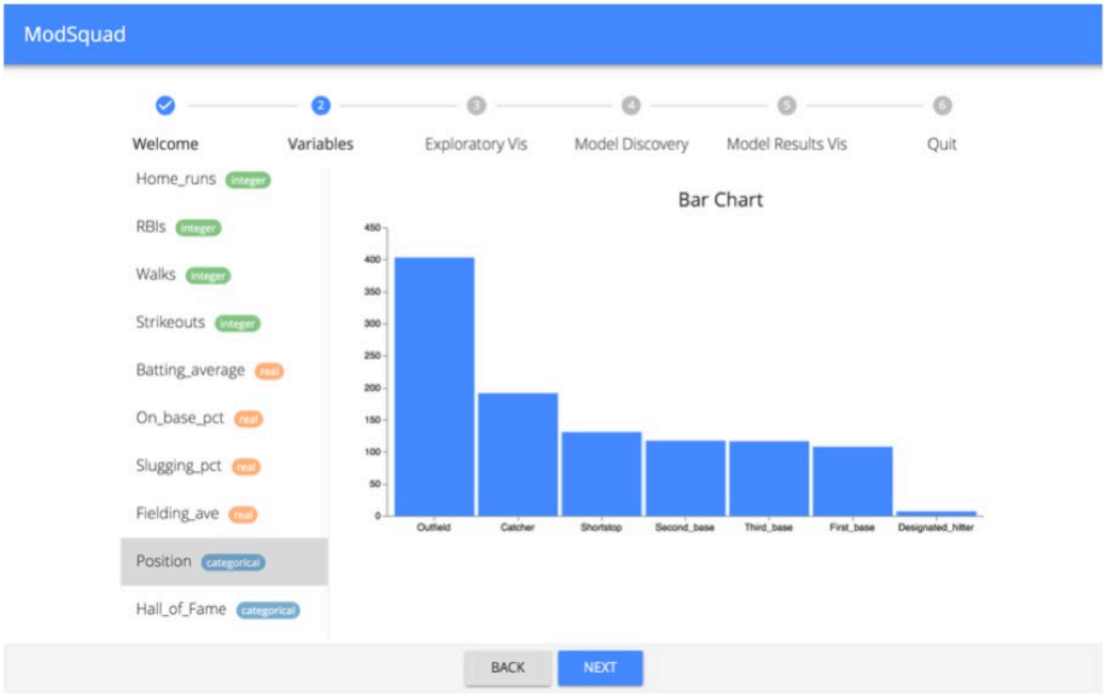


Figure 3 Exploration of a Categorical Variable

3.2 EXPLORING THE INTERACTIONS OF FEATURES TWO AT A TIME

After a user is comfortable with exploring the values each variable takes on, our team feels is best to next continue building the user's understanding and look at how the values of variables compare with each other in a "pairwise" fashion. Variables are, therefore, taken two at a time and placed on the same axis, so the user can explore how each pair of variables are related. Our interface uses either a scatterplot, a box plot, or a heat map, depending on the types of the variables being plotted together. We will show examples of each type below.

If the variables plotted together are continuous, then a scatterplot is utilized. For each instance in the dataset, the values of the two variables under study are used as indices on a Cartesian plot. An example taken from our interface is shown as Figure 4. The value of a bivariate scatterplot is to provide the user an easy visual understanding if the values of the two observed variables are positively correlated. In this context, positive correlation means that an increased value of one variable tends to indicate there will be a proportional increase in the second variable. In the inverse case, negatively correlated variables will tend to vary inversely with each other. The X axis of Figure 4 shows the number of hits a player had up to this point in their career. The vertical axis lists the number of those hits which were triples. In this case there is a positive correlation because there can be observed to be an upward trend of the points as the eye moves from the left to the right on the plot. A red line is drawn by hand to show the average of values taken by the variables in the dataset. The upward slope of the red line indicates that these variables are positively correlated.

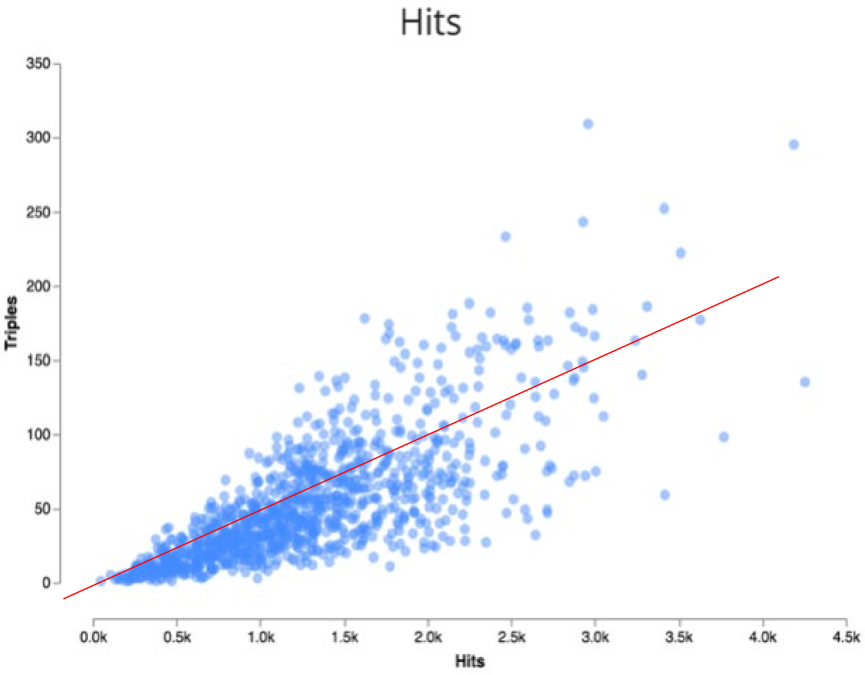


Figure 4 Scatterplot of Two Variables

3.3 EXPLORING A SET OF VARIABLES

If all pairs of two variables from a dataset were taken simultaneously, this would result in an NxN matrix of plots (given that N is the number of variables in the dataset.) This rendering is often done in the form of a scatterplot matrix, where each individual scatterplot shows the bivariate relationship between two variables. An example scatterplot of three leaf characteristics of sorghum during a growth season, from the TERRA-REF dataset, is shown in Figure 5. The scatterplot matrix is a popular way for an experienced data scientist to look at how a set of variables are inter-related. However, when a large number of plots are displayed together, this can become overwhelming for even experienced data scientists. In the Statistics literature, it is known that the scatterplot matrix visualization technique is best used to visualize relationships across variables in smaller datasets.

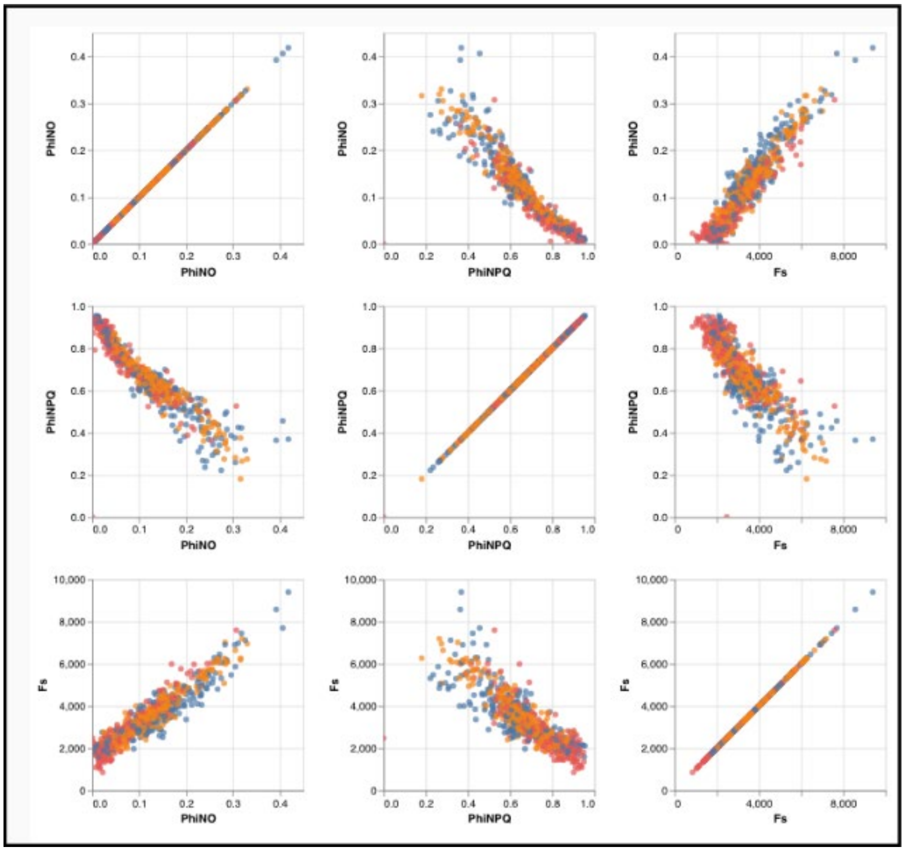


Figure 5 Scatterplot Matrix

Since our User Interface was designed to expose non-data scientists to dataset exploration, we chose to offer only bivariate exploration using a set of single scatterplots. Our interface asks the user to select a target variable from the dataset which will be plotted along the Y axes of each of a set of N-1 plots, representing the interaction between the selected variable and all the others in the dataset. The interface then generates all the plots and allows the user to scroll through them, observing which variables (if any) appear to be correlated or inversely-correlated with the selected target variable. If the target variable is a categorical variable, then our interface shows the bivariate relationship using a box plot instead. An example box plot from our interface is shown in Figure 6. The categorical value the target variables takes on are listed along the Y-axis. The most common values of the X-axis variable are contained within the size of the box and the size of the box tells the user about the distribution of values the variable plotted along the X-axis will take for each target variable value. This box plot example shows that designated hitters have a generally higher number of home runs when compared to the other playing positions. In the terminology of the D3M program and much of the current data science literature, the target variable means the variable whose value we want to learn from and then predict using machine learning methods. In the statistics literature, this is generally called the dependent variable, assuming that its value “depends” on the value of the other variables, called the independent variables.

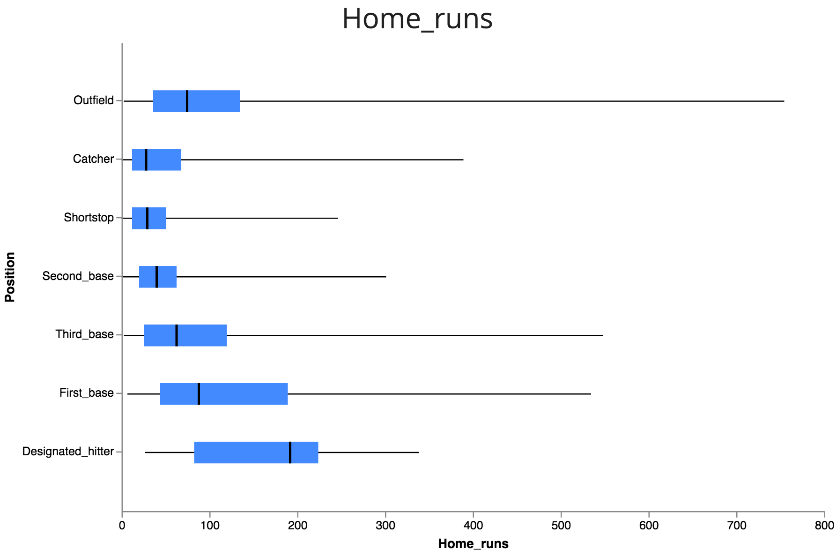


Figure 6 A Box Plot for Categorical Data

Finally, we have the case where both the target variable and the independent variable being compared are categorical in nature. For this case, our interface uses a heatmap, and an example is shown in Figure 7. A heatmap is a two-dimensional matrix where each element corresponds to a particular value for each of the variables in the bivariate relationship being explored. The colors assigned to each location correspond to the number of times this value combination appears in the dataset. In the example here, the heatmap shows that the largest subset of the dataset consists of outfielders who are not inducted in the Hall of Fame.

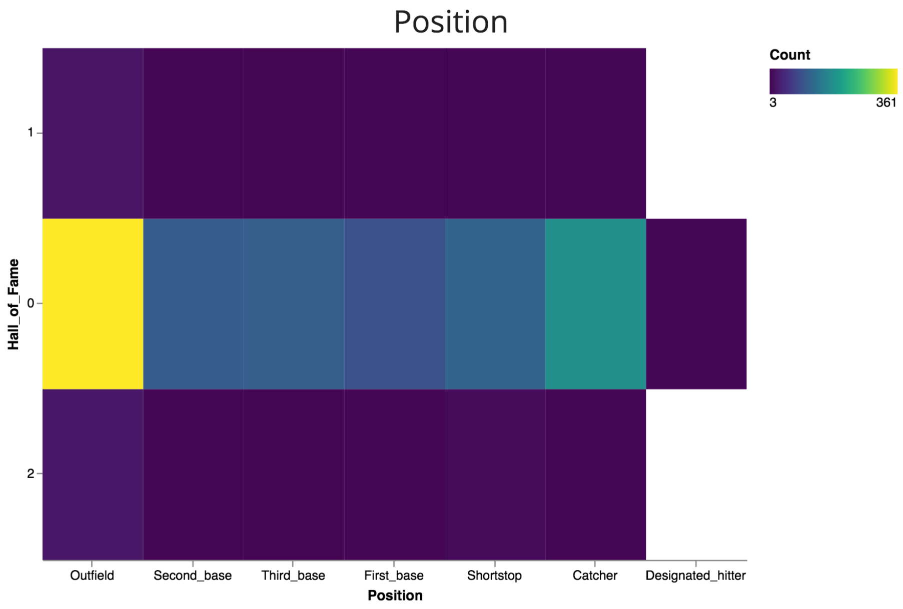
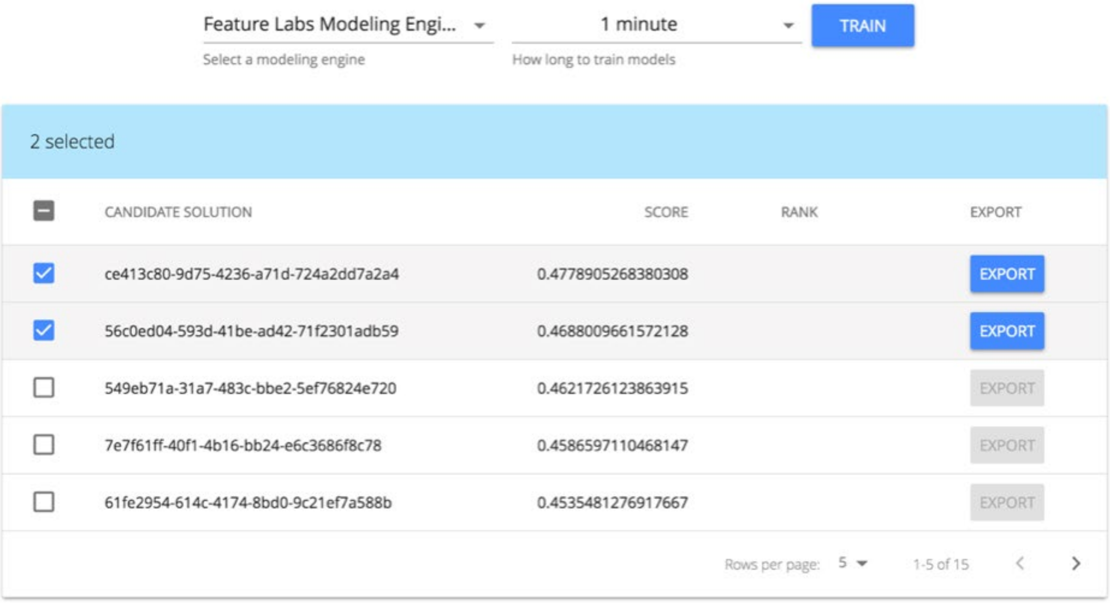


Figure 7 A heatmap showing how players were inducted into the Hall of Fame

3.4 FITTING A MODEL IN THE INTERFACE

Since the purpose of the D3M user interface is to allow a user to train and evaluate a model without much prior experience, the next screen of our interface allows the user to fit a model to the dataset and chosen target variable they have just explored through bivariate exploratory visualization. Our next screen allowed the user to pick a modeling engine to try and predict the dependent or target variable, given the values of the dependent variable. Our system was designed to be connected to more than one autoML solution engine, a feature that some other teams adopted later. The user was also allowed to specify an amount of time allowed for candidate solutions to be fit and presented. When several candidate solutions to the problem have been discovered by the modeling engine, a table of those results along with preliminary fitness scores is displayed on the interface. The user may investigate whichever of the candidate solutions she or he feels appropriate by selecting them and moving to the next step. Figure 8 shows the case where the user chose the Modeling Engine developed by MIT Feature Labs, selected 1 minute for candidate solutions to be proposed, and then selected the top two scoring solutions for comparison.



2 selected

⊟	CANDIDATE SOLUTION	SCORE	RANK	EXPORT
☑	ce413c80-9d75-4236-a71d-724a2dd7a2a4	0.4778905268380308		EXPORT
☑	56c0ed04-593d-41be-ad42-71f2301adb59	0.4688009661572128		EXPORT
□	549eb71a-31a7-483c-bbe2-5ef76824e720	0.4621726123863915		EXPORT
□	7e7f61ff-40f1-4b16-bb24-e6c3686f8c78	0.4586597110468147		EXPORT
□	61fe2954-614c-4174-8bd0-9c21ef7a588b	0.4535481276917667		EXPORT

Figure 8 Candidate Solution Interface

When model fit results are presented to the user for review, the type of visualization provided is dependent on whether the target variable was numerical/continuous or categorical in nature. For categorical variables, a heatmap is used with the predicted values listed along the Y-axes and the actual values listed along the X-axis. In this case, we are looking for a diagonal across the matrix showing correlation between the actual and the predicted values. Off-diagonal squares indicate instances where the prediction did not match the ground truth. An example for predictions from a regional disturbance dataset are shown in Figure 9. For purposes of explanation, we have superimposed a dotted line over the diagonal where the model has predicted the correct result based on its analysis of the independent variables. Each value the target variable takes on (a ground truth value) is represented as a column in this visualization. We opted to tally all of the predicted values for each actual ground truth and apply a separate color mapping (from black up to yellow) to each column. The brighter color indicates a higher percentage of the predicted values fall in this location. So the brighter the squares are at the diagonal, the better the model is fitting the ground truth. The yellow square on the diagonal in the “Riots/Protests” category indicates the model predicted almost all of these instances correctly. However, there are other values for the target variable where the model was less successful. White areas of the heat map correspond to values the model never predicted.

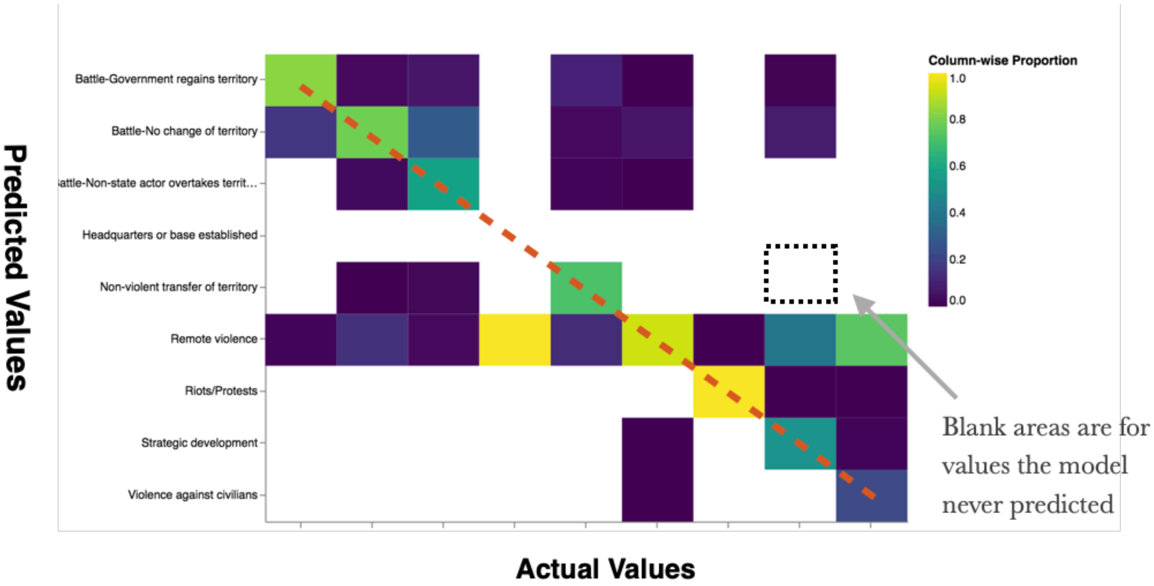


Figure 9 Model Results for a Categorical Target Variable

When the target variable being predicted is continuous, our TA3 system generates residual plots to illustrate the resulting model fit. Errors between the model prediction and the actual value are called residuals. The residual plot is a scatterplot where the points are plotted in locations according to the error between the model predictions and the actual target variable values. An example residual plot result from the baseball dataset is shown in Figure 10. In this case, the model is predicting if the player is inducted in the Hall of Fame and the possible value the target variable takes is [0, 1, or 2]. The residuals will have values in the interval [-2,2] depending on the model prediction and the actual Hall of Fame value. This plot shows the residuals when the number of games played by the individual is plotted along the X-axis.

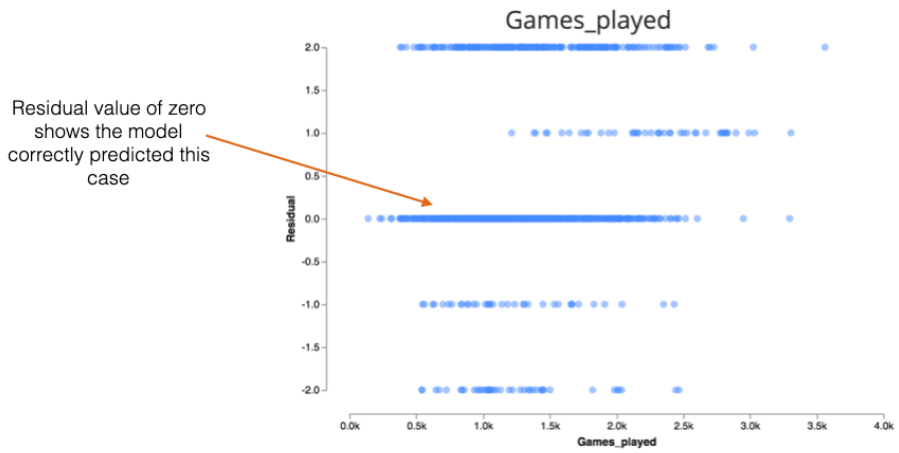


Figure 10 Model Results for a Continuous Target Variable

To help the user relate the model accuracy to the way the user explored the dataset, our user interface presents the model results as a series of bivariate charts — one for each independent variable. Each plot shows the model's target variable prediction, given the values a particular independent variable takes on. The user can look for correlation or inverse correlation between the target predictions and the other independent variables.

Our interface allows the user to iterate between different candidate solutions, exploring the residual plots and ultimately deciding which model they want to select and export for later use. When the user completes the decision, an Export button is available on the interface (see Figure 8) to enable them to select this model. The user's decision is communicated to the AutoML system to cause the AutoML save out enough information to re-run this trained model later.

3.6 USER INTERFACE IMPLEMENTATION DETAILS

Our User Interface system was implemented using a client/server architecture. Our backend server uses the Python computer language and was built as an extension to the open-source Girder platform released by Kitware, Inc. Girder provides a capability similar to dropbox, in that it allows the uploading, storage, and controlled access to any type of digital media asset, data table, image or a file, or JSON object. Our project created a github.com repository to hold our software development prototypes created under this program. The source code for our Girder extension is available on this repository. The python server is the portion of our system which connects directly to the standardized interface implemented by the D3M program (using the GRPC protocol) that connects between User Interface systems and the AutoML modeling engines. As the D3M program released updates to the protocol, we built the python wrappers and included these files in our source tree.

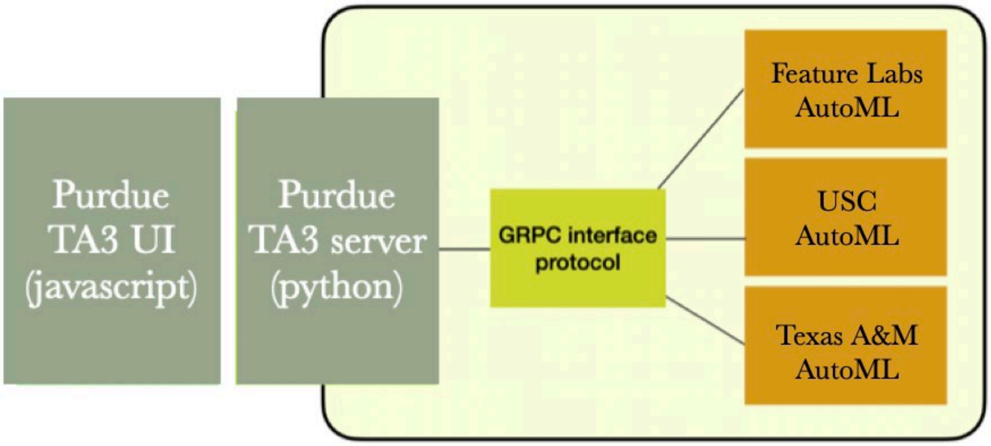


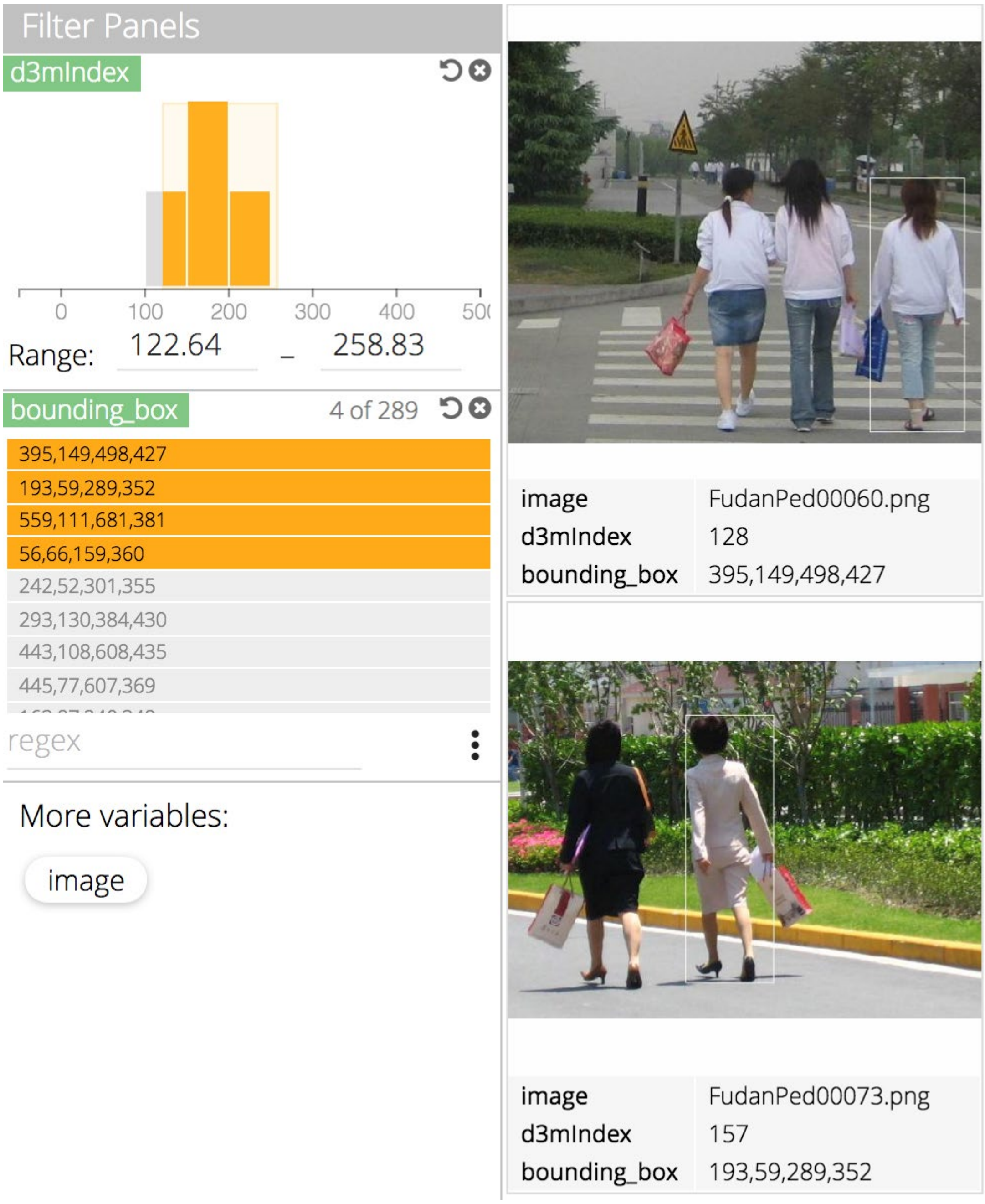
Figure 11 The Architecture of the Purdue TA3 (the user interface task)

The Purdue user interface client was developed in the JavaScript language and uses Facebook's React component framework. As described earlier in this report, our interface consists of several pages which lead the user first through the understanding of their dataset and then through the training and evaluation of a model. We selected React because of both the visually pleasing appearance of websites using it with the Material Design guidelines but also because React's extension Redux made it easier to debug state transition issues as we integrated with a number of different AutoML systems. The diagram shows that different AutoML systems could be running on the other side of the GRPC protocol. During the course of the D3M program, we tested our user interface against the MIT Feature Lab, the USC, and Texas A&M AutoML systems.

3.7 EARLY IMAGE MANAGEMENT INTERFACE

At the D3M program's request, we spent a small amount of engineering effort to develop an imagery-specific version of our User Interface and conducted a demo for the government team. This was before other TA3 systems had developed capabilities in the imaging area. Our team worked with USC and MIT/LL to formulate the standard formats for the object detection problems, which was the first imaging problem implemented for the D3M program. The datasets consisted of a series of images that contained or did not contain an object of interest. The ground truth for the datasets was provided as a set of bounding boxes covering the extents of the target objects in image coordinates.

For our interface, we adopted *Trelliscope*, developed by Dr. Ryan Hafen, one of our team members. Trelliscope supports browsing through the images and displays the ground truth or model prediction bounding boxes over the images, so the user can review the dataset. A screenshot of this interface is provided in Figure 12 running on a people-detection dataset. Trelliscope allows the user to browse several instances from a dataset and iteratively sort and visualize the instances according to filters applied to their attributes. Ultimately, our funding level in the program did not allow us to continue this development, but we were able to assist the government team in defining the problem and worked with the other TA3 teams to explain the problem and potential interfaces that could be used. We still feel that this use of Trelliscope offers unique capabilities not available in the imaging interfaces from other TA3 performers on the D3M program. This is a capability that might be useful to build on at a later time.



Range: 122.64 – 258.83

395,149,498,427
193,59,289,352
559,111,681,381
56,66,159,360
242,52,301,355
293,130,384,430
443,108,608,435
445,77,607,369

image	FudanPed00060.png
d3mIndex	128
bounding_box	395,149,498,427

image	FudanPed00073.png
d3mIndex	157
bounding_box	193,59,289,352

Figure 12 TA3 Prototype for Imaging Problems

3.8 TERRA-REF DATA EXPLORATION

During the second year of the D3M program, DARPA directed our team to study the TERRA-REF datasets. This was a result of the collaboration between DARPA and ARPA-E (the research wing of the Department of Energy) to better understand the growth and development of agriculture. A tremendous amount of data was collected by the DOE sponsored TERRA-REF program over the past ten years but relatively little post-collection analysis had been performed on this data at the time we were assigned by DARPA leadership.

MIT-LL had previously developed auto machine learning problems suitable for other D3M training performers to use, but these problems were not that helpful to the domain scientists (biologists observing how to maximize crop yield). Since our team includes leading statisticians with substantial previous experience and data analysis across many disciplines, our team was asked to hand-explore this data first in order to better understand what is available in the datasets and to assist in the creation of additional machine learning problems.

The University of Illinois at Champlain was a subawardee on the original TERRA-REF program and was responsible for archiving and analyzing the data coming from the instrumented growing fields. A picture of the robot used to take measurements during agricultural growth is provided in Figure 13. This robot, located in Mericopa, AZ is a one-of-a-kind system able to record imagery, height, and thermal profiles for agriculture under observation.

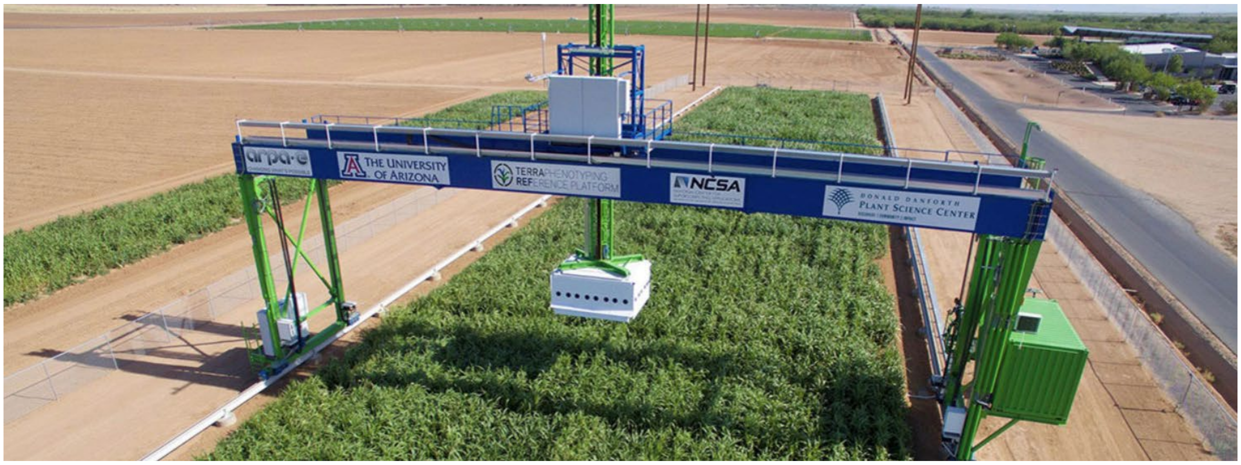

arpa-e
THE UNIVERSITY OF ARIZONA
TERRA PHENOTYPING REFERENCE PLATFORM
NCSA
DONALD DANFORTH PLANT SCIENCE CENTER

Figure 13 Mericopa Data Collection Robot

The TERRA-REF program had hand delivered some datasets to the D3M team, but these datasets did not seem very complete to our group, after inspection. We elected to go to the online repositories and directly extract data recorded by the robot and compiled by the TERRA-REF team. The University of Illinois created an online repository called BetyDB, the TERRA REF Phenotype Database. The archive is an SQL compatible database system containing multiple years of recorded phenotype attribute measurements. After direction from the D3M program leadership, we chose to focus on two growing seasons: Season Four and Season Six. These seasons both focused on the growth of various cultivars of sorghum, a large grass used for biofuel production. One of the scientific goals was how to optimize the growth of this grass in order to produce the largest yield of biofuel at the end of a growing season. Furthermore, how soon into a season could we tell whether the watering and treatment protocol being used would be successful? Armed with this information, we began an initial exploration of the TERRA data led by Dr. Ryan Hafen. In the following paragraphs, we will discuss each of the datasets that we analyzed, with particular focus on Season Four. Season Six yielded similar results to Season Four and is not discussed in detail in this report because of its similarity.

We didn't receive detailed explanations about the scientific goals of each of the TERRA-REF seasons, but we believe that Season Four is associated with tracking the growth of sorghum when faced with environmental factors such as water deprivation. There was both an automated data collection process (by the moving robot) and a manual data collection effort performed during this season. Please refer to Figure 14 below to review a plot of which variables were recorded and when during the season. 114 different features were measured at some time or other during the season (which lasted from late April through late August). The majority of these features (drawn in pink) were measured by hand only during August.

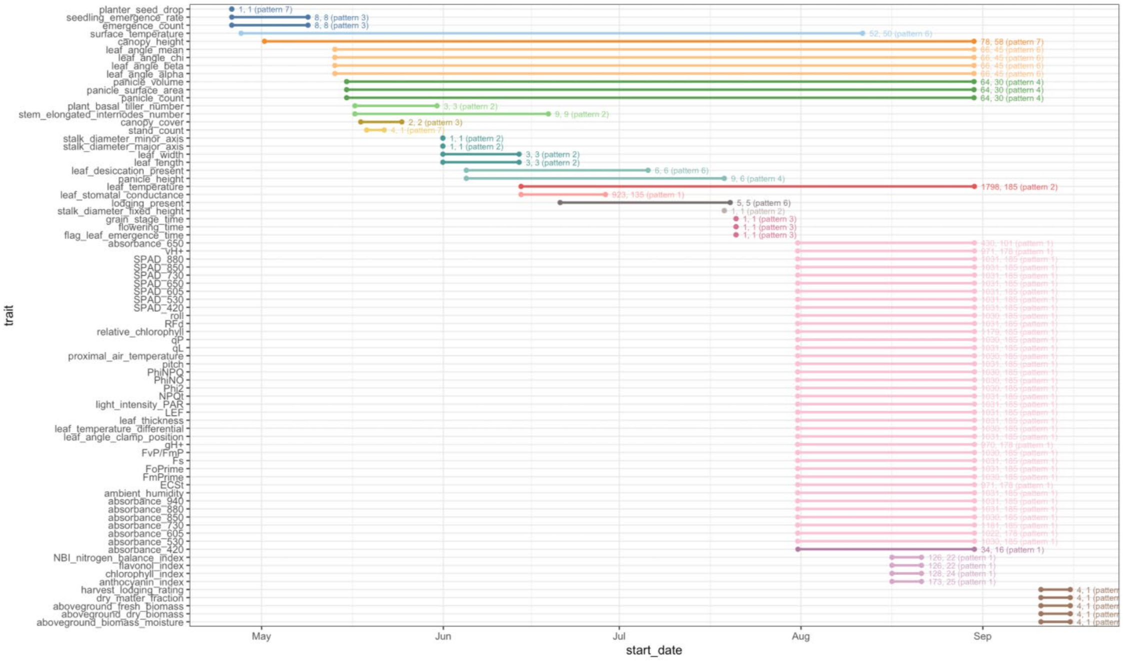


Figure 14 TERRA-REF Season Four

An initial review of the figure shows that, even though this dataset was produced by a very careful scientific effort, this dataset is still not a square table with all future values filled in for every measurement time. Data sampling was not consistent across the time domain. Figure 15 shows that sampling was not consistent in the spatial domain either. We estimate approximately eleven different patterns of spatial sampling. The sparsest being where hand measurements were done during August on only a few cultivars. The dense patterns were from the robot recorded once per day during the season. Most of the common distributions are shown below for different recorded features. Each panel corresponds to a single measured feature and the black dots represent where in the Maricopa field the measurement was taken.

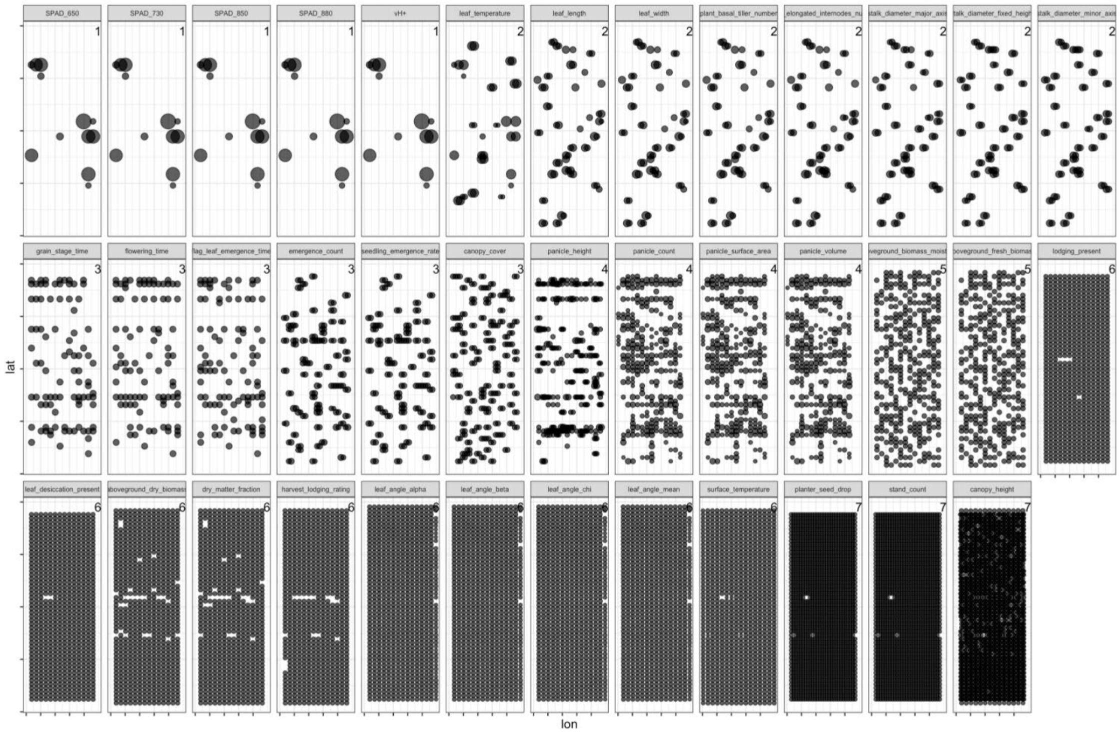


Figure 15 Spatial Distribution of Season Four Measurements

Some of the complexity of the TERRA-REF problem comes from the fact that it combines several different types of statistical learning problems. This data includes *time sequence* data since the sorghum plants are growing and being periodically measured in height during the season. However, this data also includes feature engineering (which data is important for the scientific goal) and a regression problem: how to predict the final produced biomass for each cultivar as early as possible.

4.0 RESULTS AND DISCUSSION

The first of the following sections cover the results we achieved through the development and testing of ModSquad, our domain-scientist focused user interface. After this, we discuss results achieved during our study of the TERRA-REF datasets.

4.1 MODSQUAD SYSTEM RESULTS

Our team supported the ModSquad user interface to AutoML engines for the first two years of the D3M program. After this, our budget allocated from DARPA did not support continuing the development effort. During the time we were supporting our user interface, our team was proud of achieving the following results.

- Rated Easiest Interface – During the first user studies led by Parenthetic, our user interface was rated as the easiest to understand and use by the Parenthetic team and by the testing users. Furthermore, this result was achieved under a compressed time schedule.
- Tied for 2nd in User Scores – During the first graded User Interface testing (Summer 2018), our system tied for second place results with several other teams who had undertaken substantially larger, more expensive software development efforts. Our UI system design was simple and effective, requiring only a small “learning curve”, which was the stated goal from DARPA for the user interface task. During the same testing event, we were the only team that received a user vote indicating users felt they understood the dataset better after using our interface. Again, this addressed one of the D3M’s major program goals.
- Enhanced Relationship with D3M Government partner (TERRA) – During the last year of our participation in D3M, we focused on the datasets of the TERRA program, an element of the Agricultural Research Portfolio for ARPA-E. Our team presented at an ARPA-E yearly program review briefing on behalf of D3M and developed a dedicated interface for the TERRA domain scientists, which was favorably reviewed by the domain scientists. This occurred after other D3M Performers’ demonstrations had initially failed to engage and excite the same domain scientists. After reviewing results from our team’s data analysis efforts, TERRA scientists now have a positive impression of the D3M program technology and were planning to continue a level of inter-program collaboration with D3M technology.

- First Team to Ingest Raw Datasets – During the early years of the D3M program, the government provided a library of pre-organized, pre-described problems to "jump start" the user interface and model generation systems. This helped simplify earlier engineering, but reduced the usability of D3M in deployed environments, because the TA2/3 systems could only process pre-prepared datasets running within a specific execution environment. Our ModSquad User Interface demonstrated live ingest and model fitting of raw datasets in April 2019, in advance of a directive from the D3M leadership that all systems should pivot to raw data ingest.

4.2 SELECTED MODSQUAD DEMONSTRATION VIDEOS AVAILABLE

Throughout the course of the program, our team authored several demonstration videos which are uploaded to YouTube. Several selected videos are referenced here. A URL to reach each video is listed here along with a short description of the video content:

- ModSquad Interface Walkthrough: https://www.youtube.com/watch?v=zvL1UzGj6Qw. In this video, the screens of the interface described previously in this report are demonstrated and explained for a first time user. The demonstration dataset is the popularity dataset (about children's interests in school, their popularity, and careers). This dataset is from the D3M dataset archive.
- Raw Data Ingest and Augmentation Demo: https://www.youtube.com/watch?v=mp-dKP98XOk&t. This demonstration shows ModSquad being used as a step in an actual analytical process. Raw data is curated from another data product, the data is uploaded to ModSquad and analyzed. Finally, the data is augmented through the ModSquad interface to demonstrate improvement in the model training scores. This demonstration is described in more detail in a later section of this report.

4.3 TERRA-REF ANALYSIS RESULTS

As we were just starting to understand the data, we fit several different model types to the canopy-height measurement (the height of the top covering leaf canopy) across the entire field. Substantial variation was observed across the cultivars and field locations, which can't be sufficiently learned from only using the location, leaf measurements, the cultivar, and the measurement date to determine predictions. All models tended to follow the average performance across the set of all cultivars, but this doesn't generalize to either "high achiever" cultivars or low achieving (smaller) cultivars. Figure 16 plots the results of a decision-tree model, a Gradient Boosting model, and a multi-layer perceptron neural network (MLP), when compared to a "high achiever" cultivar. The vertical axis corresponds to the actual plan canopy height in centimeters.

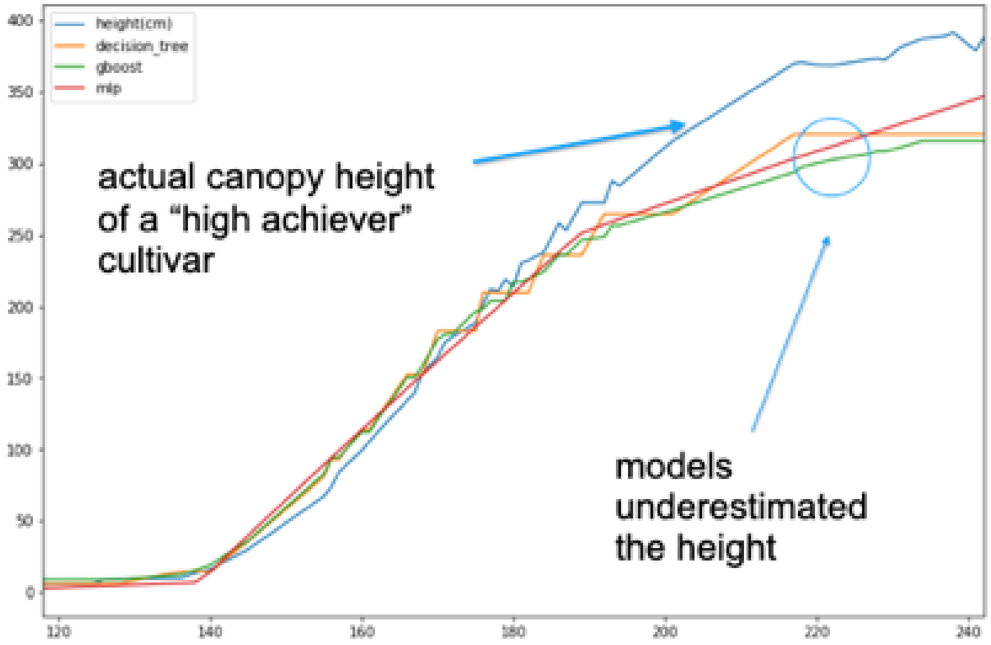


Figure 16 Early Model Fits to Cultivar Height

So it was determined that a single model predicting the height of any cultivar instance located anywhere in the field was not an acceptable solution, given the model performance in the source data available. We tried instead to train two additional sets of models: (1) a set of models with one model per cultivar, and (2) a set of models with a complete model for each location in the field. The error in canopy height measurements was substantially reduced for the cases where a set of models were trained and the correct model is used, depending on the field location. These early results were presented by our team member, Dr. Lisle from KnowledgeVis, to the ARPA-E leadership during the TERRA portfolio review in San Antonio, TX on November 12, 2019.

Histogram plots of the prediction error sizes are shown for the single model and the per-location case for the Gradient Boosting model in Figure 17. At first look, it appears each distribution approximates a Gaussian (bell) curve with a tail to the right. However, the error rates for the multiple model case (presented at right) are over 1000x more accurate. The single model predicted correct height to within 5.2% of the observed values. However, the multiple model fit was within 0.003%. A reason the canopy height is of particular interest, is that canopy height seems to be the closest proxy to the overall size of the plant and therefore the amount of the resulting biomass at the end of the season.

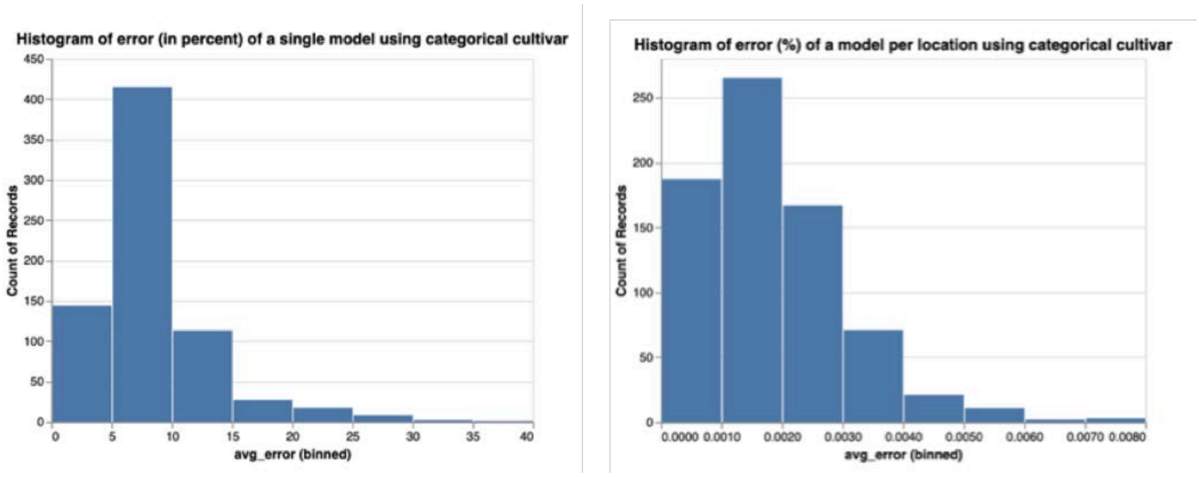
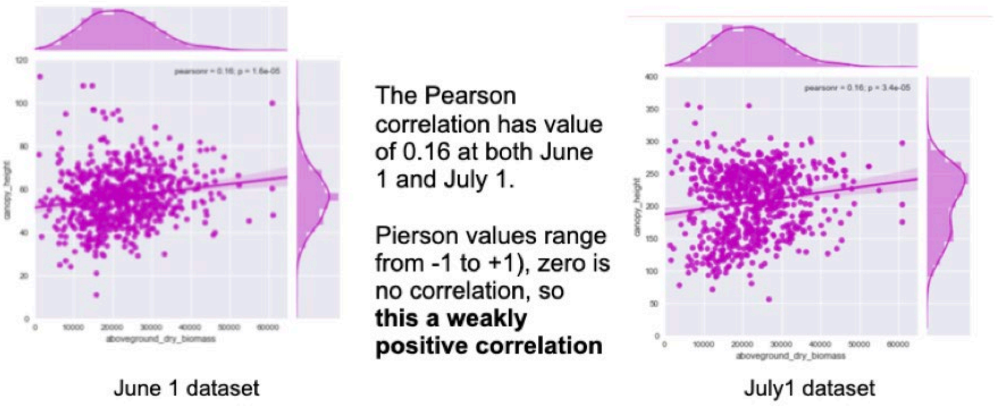


Figure 17 Accuracies for Single and Multiple Models

But how good is canopy height actually as an indicator of final biomass? To make this determination, we measured all the canopy height values on June 1st and again on July 1st to see how these values correlated with the measured final biomass at the end of the growing season. Figure 18 shows all observations plotted along with the distributions of their values plotted as histograms along the axes. A slightly positive correlation was observed between canopy height and final biomass, and this correlation had the same strength for both dates.



June 1 dataset

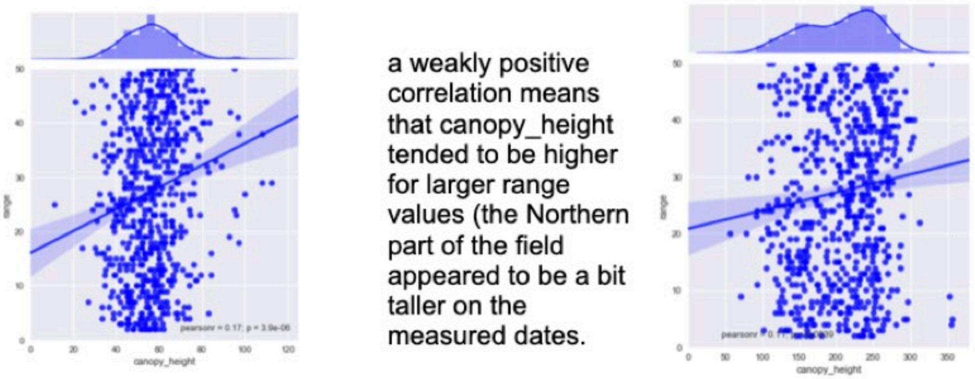
The Pearson correlation has value of 0.16 at both June 1 and July 1.

Pierson values range from -1 to +1), zero is no correlation, so **this a weakly positive correlation**

July1 dataset

Figure 18 Correlation Between Canopy Height and Biomass

We also explored if each plant's relative position in the growing field had any noticeable correlation with being final biomass or final canopy height. We found no correlation with biomass, but we did find a correlation indicating that plants along the Southern end of the field tended to be taller at the end of the season than the corresponding instances at the Northern end of the field. This correlation is shown in Figure 19 for both June 1st and July 1st dates. These charts are drawn with the range (the row) of the field going from the bottom row, which is the Northernmost, up to the highest row, which is the Southernmost.



a weakly positive correlation means that canopy_height tended to be higher for larger range values (the Northern part of the field appeared to be a bit taller on the measured dates.

June 1 dataset

July1 dataset

Figure 19 Correlation Between Canopy Height and Field Position

Since there were many different measurements taken during the season, we employed a feature engineering approach to evaluate each of the measured features and decide which were most predictive of the final biomass. To evaluate how the influence of different variables changes throughout the growing season, we performed a predictive feature analysis for June 1st and again at July 1st. By this, we mean that we took a snapshot of measured data from June 1st and used it to predict the final biomass. We repeated this exercise for available measurements on July 1st. In Table 1, we list the most important features (and their correlation weights) used in developing the prediction model that takes data from June 1 and predicts final biomass for all cultivars. Table 2 provides the same for the prediction model based on July 1 values.

Table 1 Feature Importance for June 1 Predictive Model

Feature	Importance
planter_seed_drop	0.305125
canopy_height	0.297484
stand_count	0.236735
canopy_cover	0.053942
seedling_emergence_rate	0.032933
emergence_count	0.030095
canopy_height_diff	0.020092
stalk_diameter_minor_axis	0.010246
leaf_width	0.004026
stalk_diameter_major_axis	0.003005
plant_basal_tiller_number	0.002756
leaf_length	0.002569
stem_elongated_internodes_number_slope	0.000992

Reviewing features that are present in each table, we call attention to the emergence of the correlated features as the season progresses. We will call these features the *primary plant size indicators* and we include canopy cover, stalk dimensions, and canopy height as the important features. Table 1 consists of these primary size indicators along with a number of other measurements which may exhibit spurious correlations just because it is early in the season. At the start of the season, any number of measurements may appear to be trending together early in the growth cycle. We did not have the opportunity to further investigate other features that appeared to be important to the early model only. During June, between our two feature-based predictive model dates, stalk diameters (major and minor) and the canopy measurements (coverage and height) become the best predictors of final biomass at the end of the season. We do call attention that the correlation between canopy height and biomass is reduced between the start of June and the start of July. We surmise this is due to the increased diversity resulting from different phenotypes as each has had further time to develop. In general, our results reinforce what was already surmised: the canopy height measured is the overall best in-season predictor of the final biomass of a particular cultivar instance at the end of the growing season. The overall results of the feature-based models were reasonably accurate, predicting end of the year biomass within 8% (for the June model) and within 5% (July model) after a few outliers were removed. The results for the single July 1 model (at about 5% accuracy) are comparable to the results obtained with the single random forest model trained on the season of data. In this case, by-hand feature engineering achieved a comparable accuracy using less data compared to automated model fitting, but the hand-engineered solution required effort by an expert data scientist (our team-member Dr. Ryan Hafen from Hafen Consulting).

Table 2 July 1st Dataset Feature Correlation Values

Feature	correlation to final biomass
canopy_height	0.155752
stalk_diameter_minor_axis	0.121776
stalk_diameter_major_axis	0.120870
canopy_cover	0.112504
stem_elongated_internodes_number_slope	0.074044
leaf_angle_beta_slope	0.070265

To explore variations across cultivars in detail, we inserted the Season 4 data into the interactive visualization system *Trelliscope*, developed by Dr. Hafen. Trelliscope allows a user to apply sorting and filtering operations dynamically to handle large datasets yet be able to zoom into any part of the dataset and compare the detailed interactions of the independent and dependent variables. Figure 20 shows Trelliscope operating on TERRA-REF Season Four data. The black lines in the top charts are the actual recorded measurements for two different cultivars while the colored lines (light blue, orange, red, and teal) show the predictions of the different models for those cultivars. In addition, the performance of each model is shown using a bar chart in the smaller charts on the lower part of the interface. For these charts, each prediction is shown along with the median and the quartile lines (this is the usual definition of a bar chart) to show how clustered (or how diverse) were the predictions from each of the models throughout the growing year for this cultivar. As mentioned previously, the per-cultivar and per-location models were substantially more accurate. This is observed because the vertical axis of the left charts is four orders of magnitude finer in order to adjust to the low error measurements of these models compared with the models in the right bar charts (single decision tree and single XGBoost model), which exhibited higher error rates. Trelliscope further lets a user subset the portions of a dataset according to dependent variables (such as the location in the field).

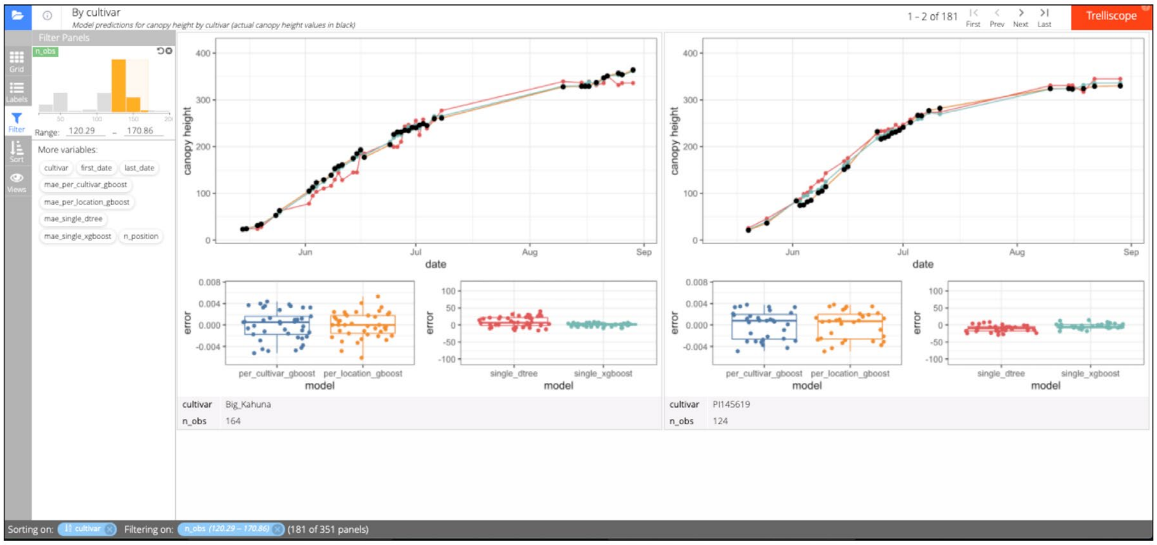


Figure 20 TERRA-REF Season Four in Trelliscope

One of the questions the biologists in the TERRA-REF program wanted to know was how much of the observed variation in the phenotypes of the cultivars could be attributed to changes in their genetic code. One way to determine this utilizes a special type of hierarchical clustering called a phylogenetic tree. The phylogenetic tree places each observed instance in the hierarchy clustered with its most closely related instances. A precise DNA sequencing of the Sorghum cultivars was completed by the TERRA-REF program, but processing this into an actual phylogenetic tree requires a lot of computation and detailed biological knowledge outside of our team's expertise area, so we didn't undertake this full conversion. To illustrate the value of this type of analysis, Dr. Lisle constructed an interim phylogenetic tree based only a subset of the cultivar's genetic differences. This isn't a biologically-exact phylogenetic tree, but is an approximation of the true tree.   Given the interim tree we constructed and the daily automatic readings from the Sorghum instances, we placed the data together in an interactive rendering showing how attribute values varied across the Sorghum cultivars. This visualization, shown in Figure 21, show some observable clusters of instances that appear to have similar phenotypic measurements, indicating the genetic makeup of the instance is related to its presenting phenotype (agreeing with the biologists' hypothesis). Note the example regions highlighted by light red boxes where phylogenetically clustered instances have similar measured features.

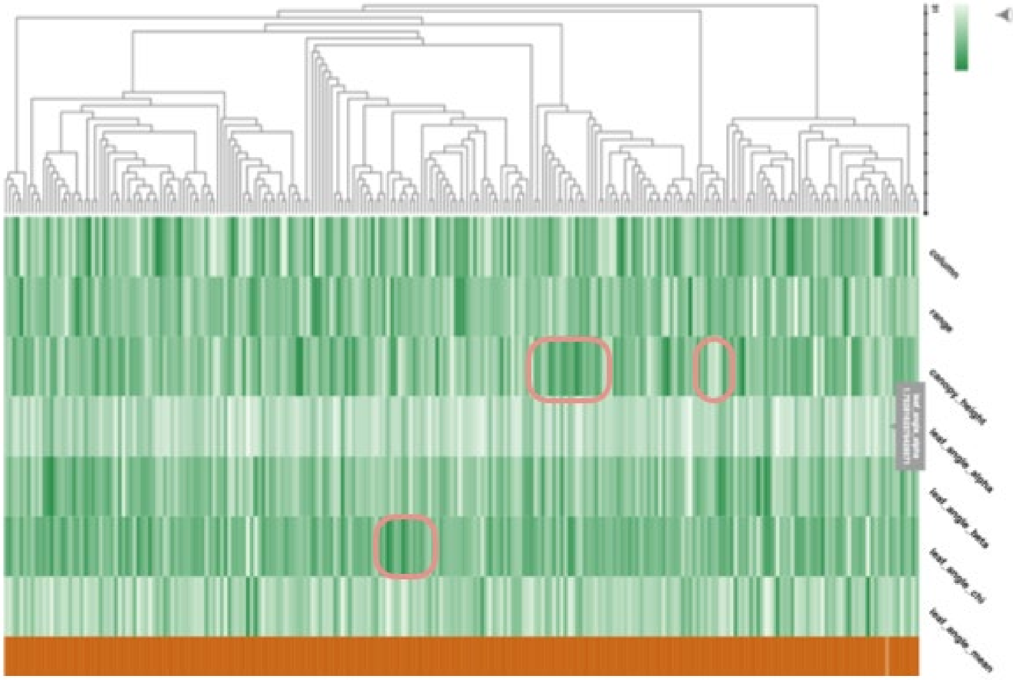


Figure 21 Sorghum Instances Arranged in a Phylogenetic Tree

4.4 HOSTING INTERACTIVE TERRA-REF MODELS AND VISUALIZATIONS ON THE WEB

When the above results were presented to the biologists, the response was very positive about how visualization illuminated relationships they supposed but hadn't previously been able to observe. They further requested that an interactive interface be built for exploring the TERRA-REF datasets, which was consistent with a request from the D3M program to develop a TERRA-specific version of our user interface.

A screenshot of the resulting system, which consists of several "mini-applications" that each explore a different aspect of the TERRA dataset, is provided in Figure 22. This system was publicly hosted on the Amazon cloud during the Spring and Summer of 2020 to collect further feedback from the domain science community. A video walkthrough of our team's demo system is published on YouTube here (https://youtu.be/o6H7rpJ_Wwk). The biologists responded that they would like to have systems like this available to review data as it comes in during future seasons of research for the Maricopa field. This could be an opportunity for later installation of this and other D3M developed technologies.

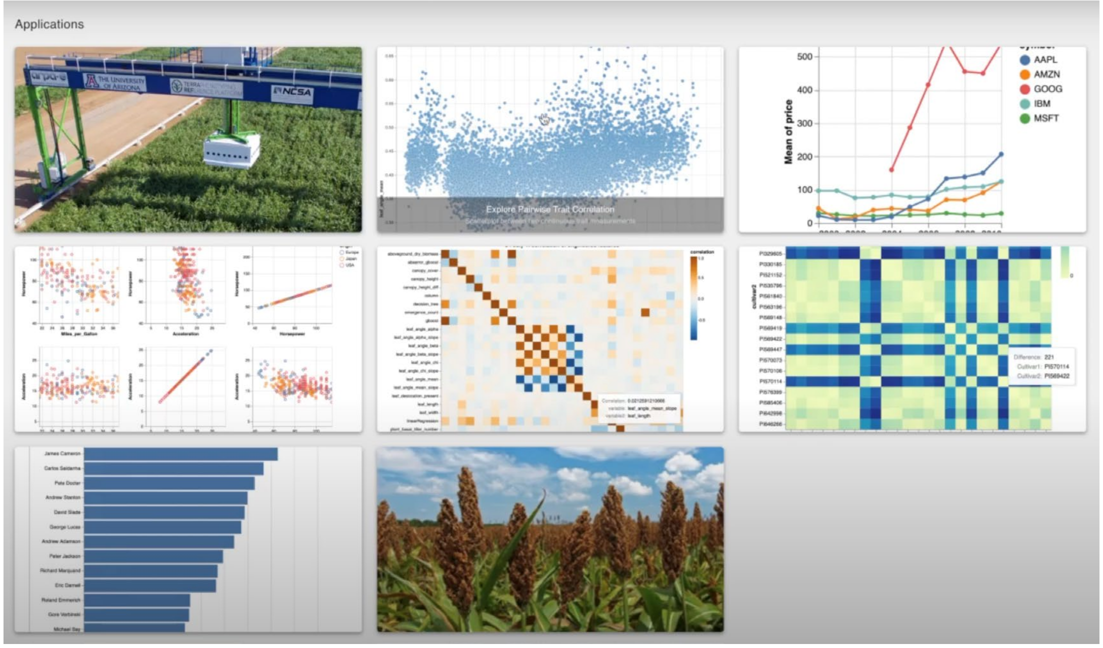


Figure 22 Applets Shown in the Purdue/D3M TERRA-REF Interface

Season Six - As previously mentioned, Season Six growth and result data is somewhat similar to Season Four. The interactive demonstration application supports visualization and model fitting on both Season Four and Season Six data, but the results of Season Six are not included separately in this report because of their similarity to Season Four.

4.5 LESSONS LEARNED AND BROUGHT BACK TO THE D3M PROGRAM

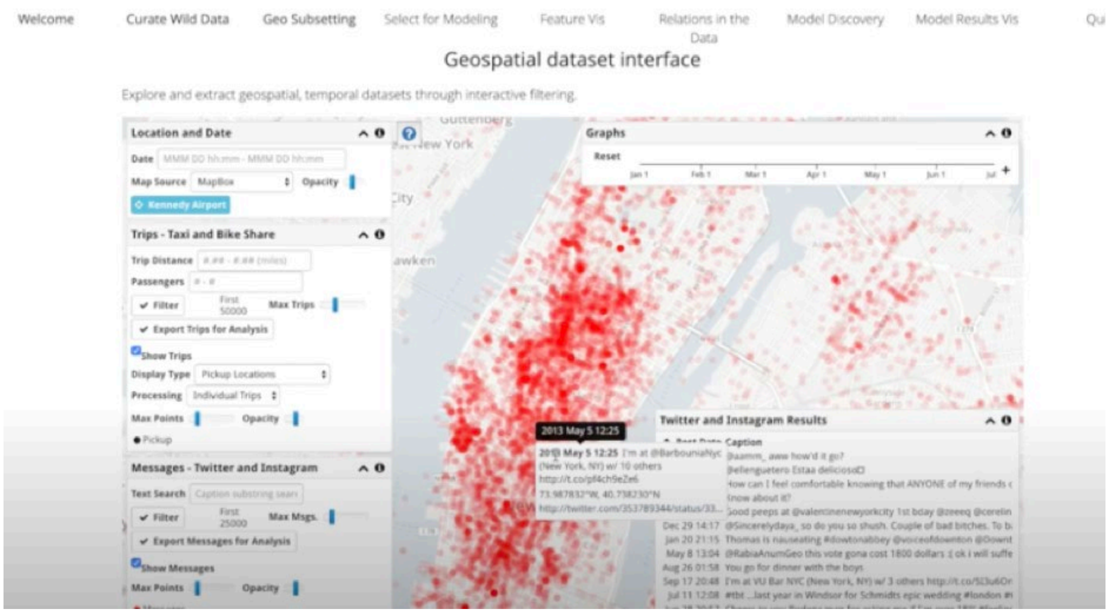
Our team gave a briefing during June 2020 to the D3M Performer teams during a special virtual project meeting. The presentation described the interfaces we developed for the TERRA analysis demonstration and mentioned several issues that we noticed during our processing of the TERRA-REF datasets, such as temporal data gaps. Even though this was a well-curated dataset, it still took our team a noticeable time to process and prepare the data for analysis and visualization. We feel this is a motivation for AutoML systems like D3M and others to embrace some or all of the “ETL” process (extract, transform, and load) for dataset preparation as practical. These “lessons learned” are applicable for teams to apply to their AutoML interfaces and system developments. The presentation we gave is included in this report as Appendix A.

Summary of the TERRA-REF Engagement - The Purdue D3M team appreciated the opportunity to engage with the TERRA-REF dataset, the scientific team behind the data collection effort, and their archived data for several growing seasons of the Maricopa, Arizona agricultural research field. We feel that the analyses and visualization products we developed and shared were helpful to the domain scientists and also provided examples to inspire the further development of D3M and other AutoML systems. The value of our analysis and products was confirmed to us by multiple members of the TERRA-REF scientific team.

4.6 TRAINING MODELS ON RAW DATA WITH MODSQUAD

During early 2019, our TA3 was the first in the D3M program to embrace raw, real-world data. Up to that point in the program, all other D3M Performers had focused on processing pre-curated datasets, where the problem description had been already explicitly created and the data had been preprocessed to fit a standard specification. It was right for D3M to do this in the beginning to reduce the burden of data ETL on its AutoML solutions. However, we felt as we began to transition to an operational system, D3M pipelines should be able to process raw datasets.

As a representative test case, we chose to analyze taxi and instagram data for New York City. During the DARPA XDATA project, our team member, Kitware, Inc. created an ingest and exploration interface for geospatial datasets. During early 2019, we developed an integration connector between the geospatial exploration system and ModSquad, our user interface, and we demonstrated model fitting and data augmentation on raw datasets as they were created through the exploration interface. We believe our effort was the first demonstration of using D3M technology to build models from actual data feeds. The geospatial interface is shown in Figure 23 displaying taxi and instagram data interactively. The analyst can use the interface to explore the raw data and subset temporally and geospatially — ultimately creating a data subset for further analysis through fitting a D3M-generated model to the data.


Welcome
Curate Wild Data
Geo Subsetting
Select for Modeling
Feature Vis
Relations in the Data
Model Discovery
Model Results Vis
Qu
Geospatial dataset interface
Explore and extract geospatial, temporal datasets through interactive filtering.
Location and Date
Date MMM DD hh:mm - MMM DD hh:mm
Map Source MapBox
Opacity
Kennedy Airport
Trips - Taxi and Bike Share
Trip Distance #.## - #.## (miles)
Passengers # - #
✓ Filter
First 50000
Max Trips
✓ Export Trips for Analysis
☑ Show Trips
Display Type Pickup Locations
Processing Individual Trips
Max Points
Opacity
● Pickup
Messages - Twitter and Instagram
Text Search Caption substring search
✓ Filter
First 25000
Max Msgs.
✓ Export Messages for Analysis
☑ Show Messages
Max Points
Opacity
Graphs
Reset
Jan 1
Feb 1
Mar 1
Apr 1
May 1
Jun 1
Jul
New York
City
awken
Guttenberg
Twitter and Instagram Results
Post Date
Caption

2013 May 5 12:25

2013 May 5 12:25 I'm at @BarbouniaNyc (New York, NY) w/ 10 others http://t.co/pl4ch9eZe6
73.987832°W, 40.738230°N
http://twitter.com/353789344/status/33...

Post Date	Caption
	@aamm_ aww how'd it go?
	@ellenguetero Estaa deliciosoD
	How can I feel comfortable knowing that ANYONE of my friends c know about it?
	Good peeps at @valentinenewyorkcity 1st bday @zeeeq @corelin
Dec 29 14:17	@Sincerelydaya_ so do you so shush. Couple of bad bitches. To b
Jan 20 21:15	Thomas is nauseating #dowtonabbey @voiceofdownton @Downt
May 8 13:04	@RabiaAnumGeo this vote gona cost 1800 dollars :(ok i will suffe
Aug 26 01:58	You go for dinner with the boys
Sep 17 20:48	I'm at VU Bar NYC (New York, NY) w/ 3 others http://t.co/5l3u6Or
Jul 11 12:08	#tbt ...last year in Windsor for Schmidts epic wedding #london #t

Figure 23 The ModSquad Geospatial Interface

D3M curated a library of state-of-the-art algorithms for datatype identification, outlier detection, value imputation, and other operations helpful during data ETL (extract transform and load) operations. These primitives are used by the AutoML systems during the process of model fitting. However, there is a danger that an AutoML system could apply these primitives in a way that is semantically incorrect according to the context in which the incoming data was generated. To address this problem, ModSquad empowers the user to run selected primitives interactively since the subject matter expert can be expected to understand the appropriate context for data cleaning, aggregation, or other semantically-sensitive operations.

In our example analysis, the user was predicting taxi pickups around the NYC (New York City) Kennedy Airport and the number of instagram messages authored in the immediate vicinity around the airport was used as a data augmentation technique to improve the model fit results. In the left panel of Figure 24, a user is visually browsing taxi pickup data that they used ModSquad to aggregate into hourly totals. On the right panel of the figure is the output plot showing the error between the actual taxi pickup totals and the totals predicted by the AutoML model trained on the pickup data. This analysis then continued by augmenting the hourly taxi pickup totals with hourly-aggregated Instagram messages and demonstrated that the additional column of relevant data improved model fit results.

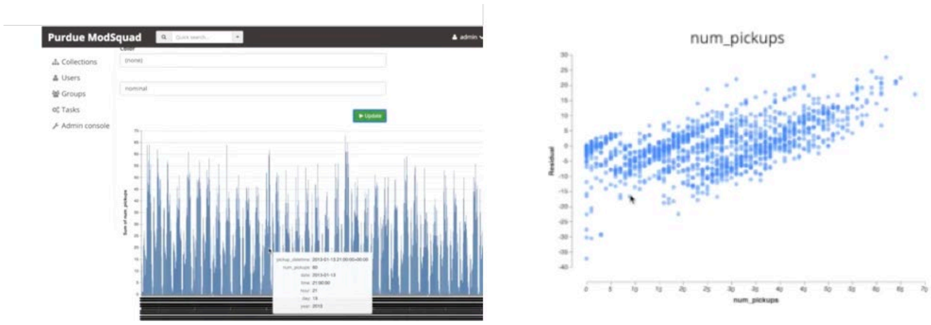


Figure 24 Input and Resulting ModSquad Model Residual Plot

As far as we are aware, this exercise demonstrated ModSquad was the first TA3 team to ingest real-world ground truth data, enable a domain expert to augment the original data using a semantically-related dataset and then use a TA2 AutoML system to train models on the original and augmented dataset. We view this as an important demonstration of how D3M technology can be used to solve real-world problems. A video of this demonstration is available on YouTube at https://www.youtube.com/watch?v=mp-dKP98XOk&t=696s.

4.7 D3M IN ATMOSPHERIC RIVERS RESEARCH

As part of our evaluation of the potential impact and applicability of D3M technology to real-world problems, our team decided to exercise tuning AutoML models to state-of-the-art statistical research datasets. Since one of the goals of D3M was to directly help domain scientists remain focused on their problem without having to spend a lot of time learning mathematical modeling, our team analyzed climate data and explored using D3M technology to directly fit the data from this scientific domain.

Atmospheric rivers (AR) are long narrow filaments of enhanced water vapor transport in the lower troposphere and they are known to accompany extreme rain and winds. They are important weather systems for US water resources on the West Coast and in the Midwest. The Purdue Statistics department lead research on atmospheric rivers by performing large-scale data analysis on a Hadoop cluster over climate data from the US West Coast and US Midwest regions.

For the research focus, our team asked which impacts, in which region, and in what time scale and period were Atmospheric River activities of concern. We then used an approach, combining climate significant-event or extreme-event criteria, image processing, and statistical analysis to create eighty-one (81) West Coast AR indices and the same number of Midwest indices from January 1980 to June 2017 for answering the questions using detailed visualization. We found that an optimal AR index for precipitation depends on the defined precipitation impacts, regional physical mechanisms of precipitation, season, and duration. One of the AR measurements, Integrated water vapor (IWV) can represent the broad-stroke presence and accumulation of precipitation in regions studied. Longer duration thresholds also led to higher accumulated precipitation. Combined moisture with wind fields using another AR metric, integrated water vapor transport (IVT), is necessary to get extreme West Coast AR orographic precipitation. IWV well represents moderate to extreme Midwest AR precipitation events for all seasons. The combination of IVT and IWV is useful to get snapshots of extreme precipitation events. The complete analysis was recently published as a journal paper in the issue 10.1029/2020JD033667 of the ”Journal of Geophysical Research: Atmospheres”.

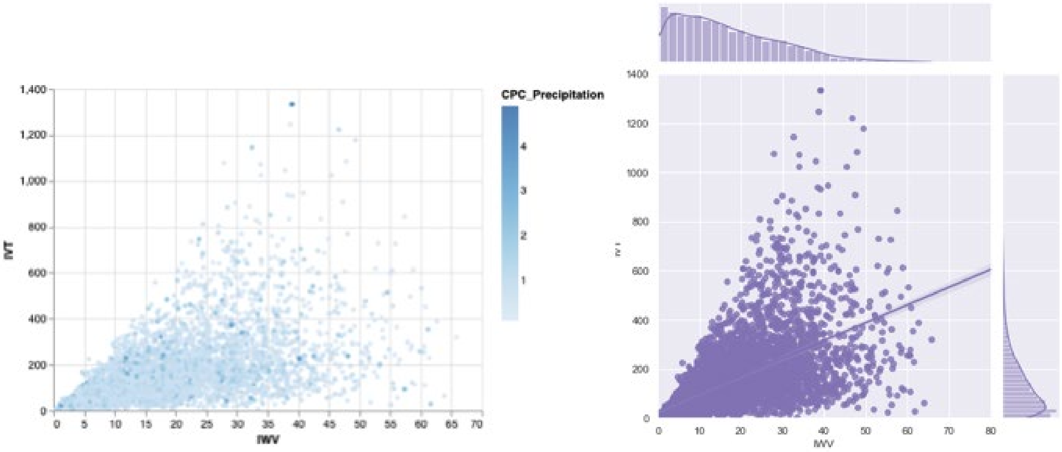


Figure 25 IVT and IWV Plots for the Subset Area

To evaluate the use of D3M technology on this application area, we experimented with a subset of the compiled AR and precipitation data since the D3M program was focused on AutoML methods instead of processing very large datasets. We chose a section of West Cost 2017 AR data and first generated some exploratory visualizations using ModSquad interface technology. We were searching to assess what can be learned about Atmospheric River metrics through quick visual examination at first. We examined the covariation of IVT against IWV to see if there was any easily viewable correlation evident with measured precipitation in this geospatial and temporal subset. The left panel of Figure 25 plots IVT against IWV with the color saturation of the dots tied to the precipitation levels. A general correlation of the AR metrics is observed without being able to make any observations on precipitation levels.   The right panel shows the same scatterplot data with the addition of the plot trend line and a rendering of the distributions of IVT and IWV. It is noted that for a particular region and time viewed, the measured AR activity had distributions scaled toward lower values but with a long tail, indicating the presence of a few events of higher atmospheric moisture content.

To fit a simple model to this dataset subset, an XGBoost classifier was used to predict the occurrence of precipitation using only IVT and IWV as independent variables. The classifier was trained on 75000 point from the subset. 60000 points were randomly selected for training with the remaining 15000 points held back for model evaluation. The confusion matrix from the classifier (shown in Figure 26) indicates an overall 58% correctness when using only Atmospheric River metrics for predicting precipitation in the selected West Coast region. This demonstrates that "somewhat helpful" models can be trained for precipitation prediction using only a subset of the AR metrics alone. However, these results also demonstrate that the interaction between Atmospheric Rivers and observed participation is more complex than just a simple classification can detect and understand. This reinforces the results described in our full paper.

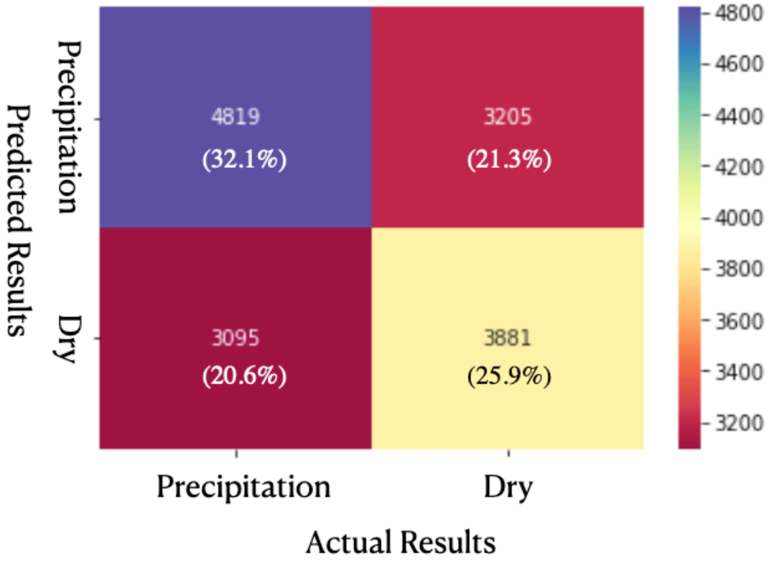


Figure 26 Accuracy of ML Model

5.0 CONCLUSIONS

The Data-Driven Discovery of Models Program focused on the development of AutoML software, methods, and interfaces to make data science easier for domain scientists to perform. Throughout the course of the program, our team, which includes well-known practicing data scientists, offered our experience of how to meet new datasets, what visualizations are the most helpful for domain scientists, and we implemented some of these practices in our ModSquad TA3 interface.

Throughout the course of this program, our team engaged as much as our budget allocation allowed and our team contributed at several key moments during the arc of the research performed on the D3M project. Contributions included early user testing, pioneering the move to kubernetes for automated system testing, analyzing, and augmenting real-world data through our ModSquad interface, engaging with outside scientists in the TERRA-REF program to understand and illustrate their data, and finally applying data science techniques to a brand-new statistics research area in the field of the characterization of Atmospheric Rivers. Our team is proud of the accomplishments we achieved. We are also grateful for the opportunity offered by DARPA to join with other expert Performers throughout the span of this research program.

APPENDIX A - LESSONS LEARNED PRESENTATION (TERRA-REF DATA ANALYSIS)

TERRA-REF Datasets and Problems and Collaboration with D3M

D3M Purdue ModSquad TA3 team
Presented by: Curtis Lisle, KnowledgeVis, LLC

Problem Overview

- The TERRA-REF program has collected a large amount of genetic and growth data related to the sorghum plant over several growing seasons. More background about the program have been provided in previous presentations
- The phonemic data (observed growth and descriptive measurements) is the first category of data available
- There are also several image datasets currently coming online (whole field overhead, multispectral, infrared imagery, etc.).
- D3M could help them explore this new image data modality (more about imagery later)

TERRA-REF Seasons

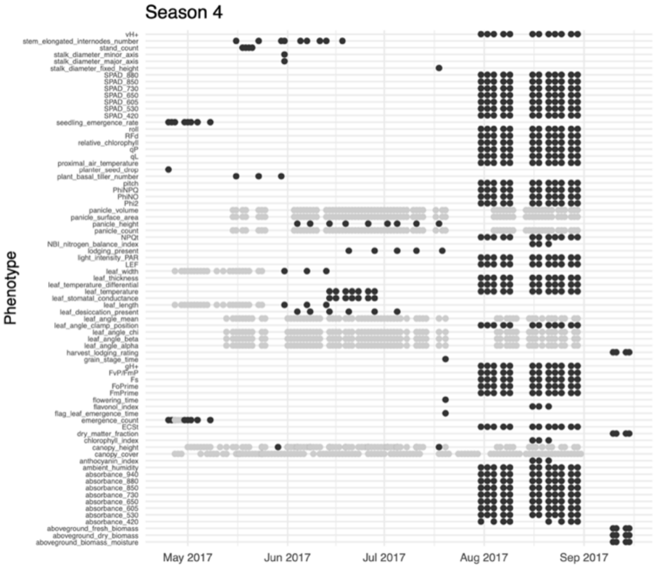
- There were multiple Seasons, each one has its own biological motivation. Here is a simplified summary of three seasons with overlapping cultivars
  - Season 4 - Cultivars selected for wide variety of growth patterns, but all from a set of cultivars known for relatively large biomass. These cultivars are "lines" from the BAP - from the Biomass Association Panel. Drought tolerance test was performed in August.
  - Season 6 - Same cultivars as Season 4; cultivar arrangement seems relevant: "Experiment was arranged in two replicates (blocks) as a row-column design with a further constraint that lines (accessions) are blocked by height class. This blocking is applied across ranges. In specifying a model for analysis, lines (accessions) are nested within height/stature classes."
  - Season 9 - Overlap of a few of the BAP cultivars. Complex watering, pre-growth, and fertilizer regimen is listed in the season description.

Season 4

- This is the data D3M has seen so far (canopy_height and leaf_angle problems)
- A mixture of automatic and manual measurements
- Manual measurements were all taken in last month of the season
- canopy_cover is available
- drought tolerance study during August

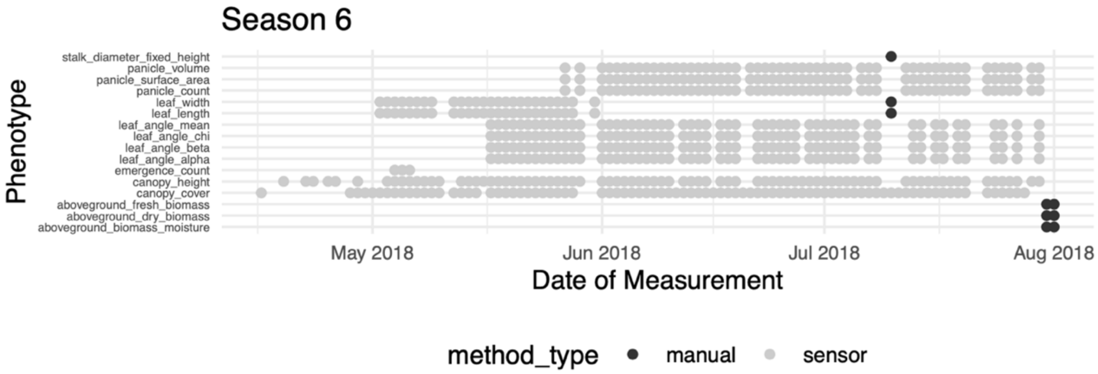
The table at right is from a manuscript pre-print available here:

https://github.com/terraref/data-publication/blob/master/terraref-dryad.pdf



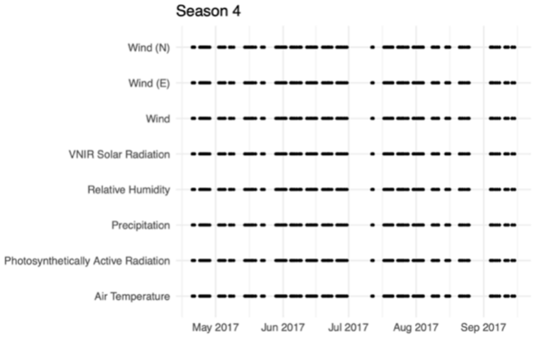
Season 6

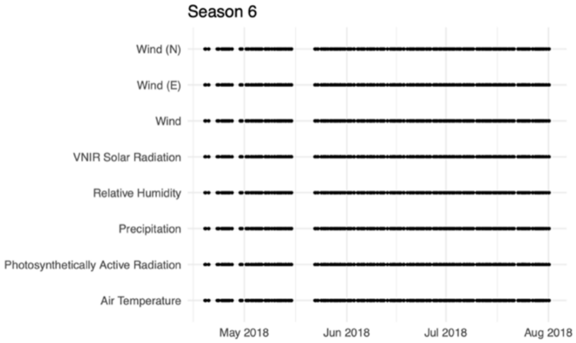
- Large overlap in cultivars with season 4 (324 matches)
- Similar sensor measurements (leaf properties, canopy height, cover)
- Panicle measurements



Weather

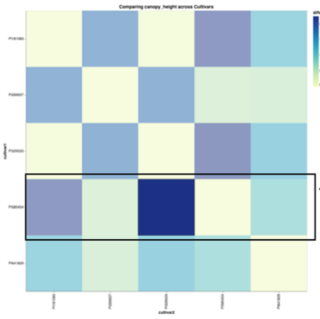
- Taken from local field sensors; missing dates could be filled-in from external sources
- A DataMart opportunity here?



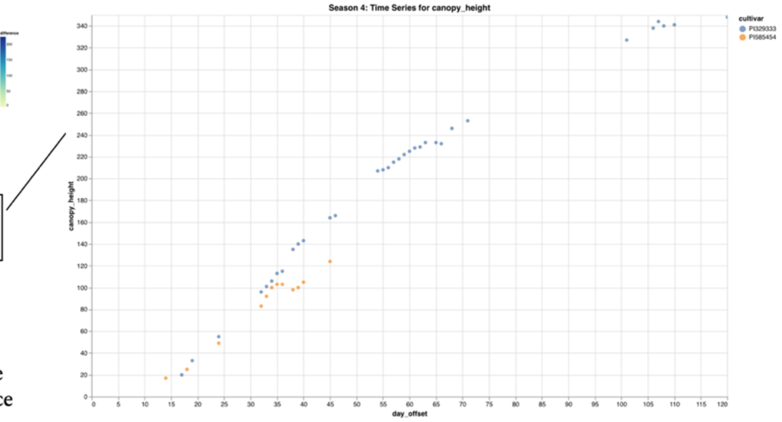


Temporal Sampling Issues

- Visualizations should expose irregular sampling so the scientists don't overlook it
- Here is a case where cultivars had different height primarily because of sampling

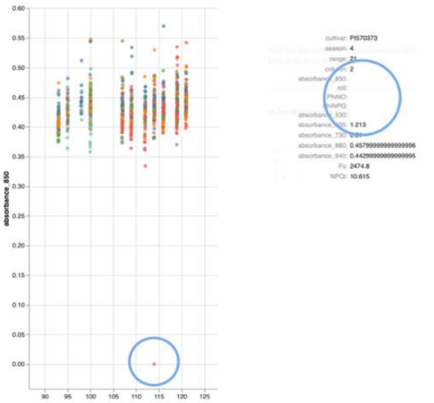


Side-by-side height comparison. We drill-down into the biggest difference



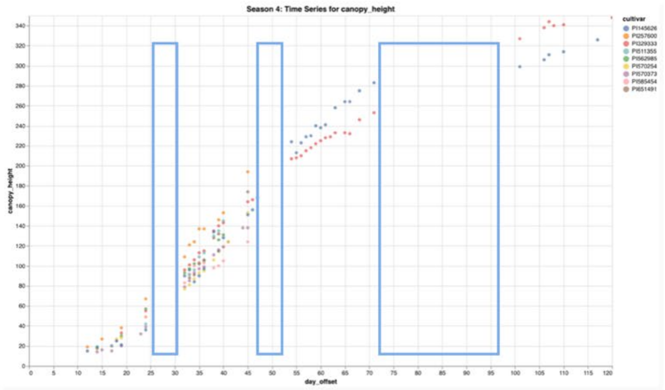
Quality Control / Curation?

- This has been largely out of the scope of D3M, but this dataset could benefit from algorithmic curation
- Outlier Detection/Removal



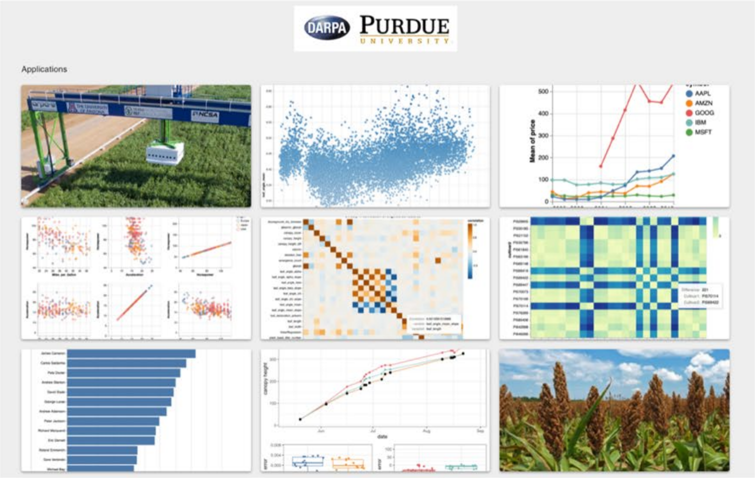
cultivar: PI570373
season: 4
range: 21
column: 2
absorbance_850:
roll:
PhiNO:
PhiNPQ:
absorbance_530:
absorbance_605: 1.213
absorbance_730:
absorbance_880: 0.45799999999999996
absorbance_940: 0.44299999999999995
Fs: 2474.8
NPQt: 10.615

- Temporal Imputing / Resampling



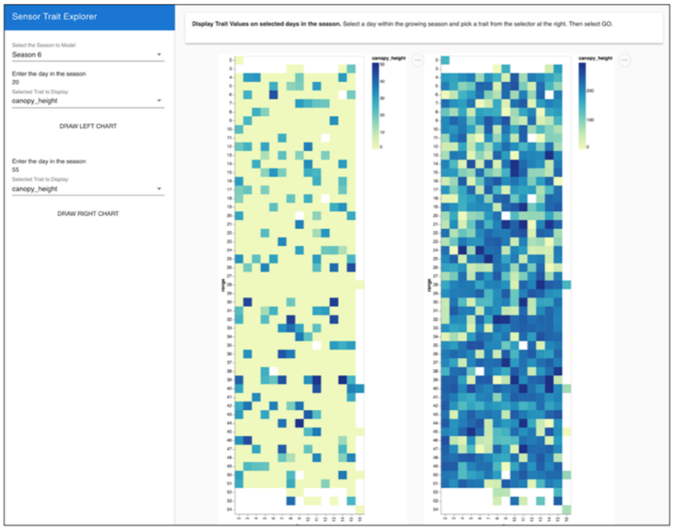
An Interactive UI for TERRA datasets

- Available at http://terraref.knowledgevis.com:8080
- Implements a number of data exploration "mini apps" including a per-cultivar XGBoost model
- YouTube demo walkthrough at https://youtu.be/o6H7rpJ_Wwk



Sensor Trait Explorer

- Allow exploration of any auto collected sensor trait
- Can be used to compare the same trait at different times in the season or different traits at same or different times
- Select season, day-into-season, and trait to display
- Choose values independently for left and right chart



Sensor Trait Explorer

Select the Season to Model
Season 6

Enter the day in the season
20
Selected Trait to Display
canopy_height

DRAW LEFT CHART

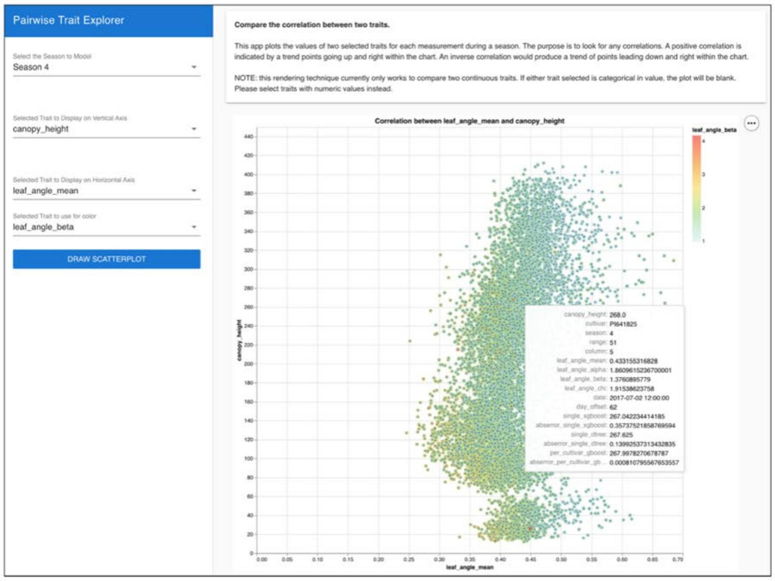
Enter the day in the season
55
Selected Trait to Display
canopy_height

DRAW RIGHT CHART

Display Trait Values on selected days in the season. Select a day within the growing season and pick a trait from the selector at the right. Then select GO.

Trait Scatterplot

- Allow exploration of an entire season's measurements
- Discover any easily observable trends, correlations, or outliers



Pairwise Trait Explorer

Select the Season to Model: Season 4

Selected Trait to Display on Vertical Axis: canopy_height

Selected Trait to Display on Horizontal Axis: leaf_angle_mean

Selected Trait to use for color: leaf_angle_beta

Compare the correlation between two traits.

This app plots the values of two selected traits for each measurement during a season. The purpose is to look for any correlations. A positive correlation is indicated by a trend points going up and right within the chart. An inverse correlation would produce a trend of points leading down and right within the chart.

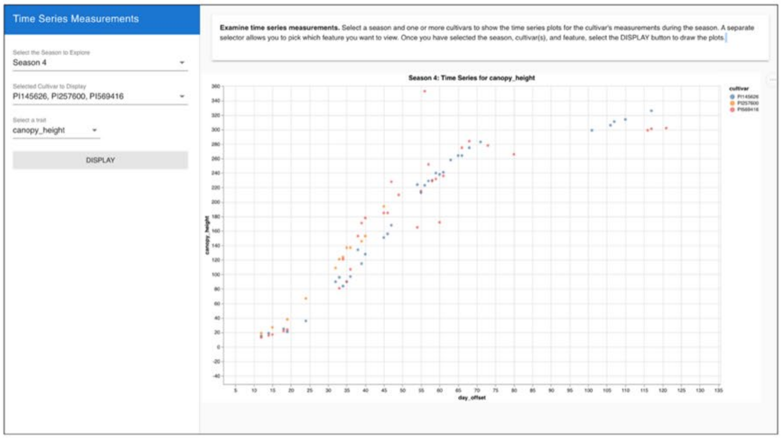
NOTE: this rendering technique currently only works to compare two continuous traits. If either trait selected is categorical in value, the plot will be blank. Please select traits with numeric values instead.

Correlation between leaf_angle_mean and canopy_height

canopy_height: 268.0
cultivar: PI641825
season: 4
range: 51
column: 5
leaf_angle_mean: 0.433155316828
leaf_angle_alpha: 1.8609615236700001
leaf_angle_beta: 1.3760895779
leaf_angle_chi: 1.91538623758
date: 2017-07-02 12:00:00
day_offset: 62
single_xgboost: 267.042234414185
abserror_single_xgboost: 0.35737521858769594
single_dtree: 267.625
abserror_single_dtree: 0.13992537313432835
per_cultivar_gboost: 267.9978270678787
abserror_per_cultivar_gb... 0.00081079556765355

Time Sequence across Cultivars

- Pick a season and one or more cultivars to explore
- Plot any trait of the cultivars against the days of the season (x-axis)
- See how a trait varies across the season and across a subset of cultivars



Time Series Measurements

Select the Season to Explore: Season 4

Selected Cultivar to Display: PI145626, PI257600, PI569416

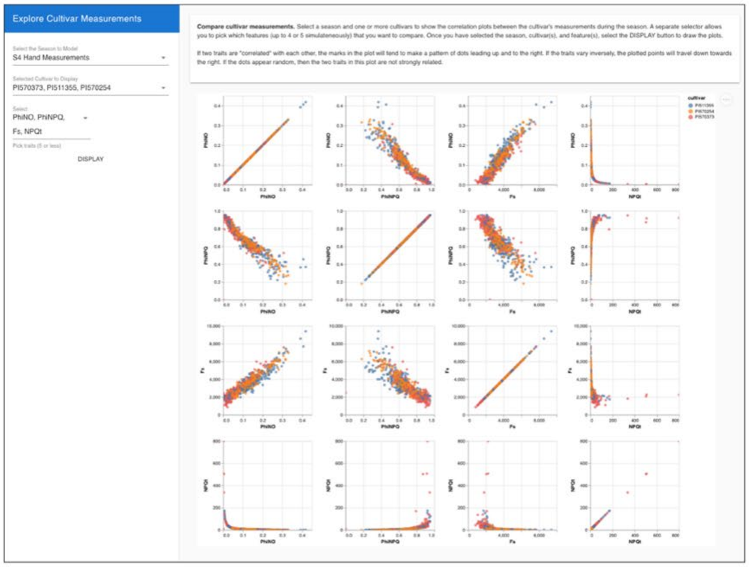
Select a trait: canopy_height

Examine time series measurements. Select a season and one or more cultivars to show the time series plots for the cultivar's measurements during the season. A separate selector allows you to pick which feature you want to view. Once you have selected the season, cultivar(s), and feature, select the DISPLAY button to draw the plots.

Season 4: Time Series for canopy_height

Scatterplot Matrix

- Pick a season and one or more cultivars to explore
- Pick several traits to explore their joint interactions
- bivariate, trivariate, etc. relationships



Explore Cultivar Measurements

Select the Season to Model: S4 Hand Measurements

Selected Cultivar to Display: PI570373, PI511355, PI570254

Select: PhiNO, PhiNPQ, Fs, NPQt

Pick traits (5 or less)

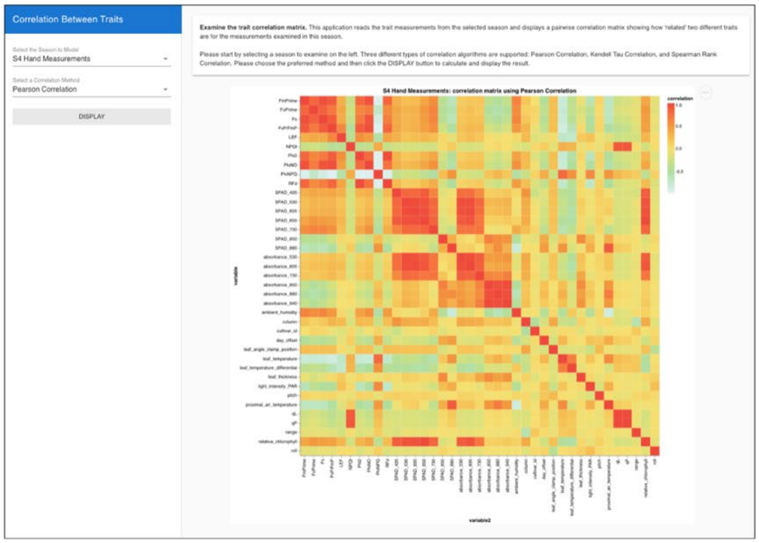
DISPLAY

Compare cultivar measurements. Select a season and one or more cultivars to show the correlation plots between the cultivar's measurements during the season. A separate selector allows you to pick which features (up to 4 or 5 simulateneously) that you want to compare. Once you have selected the season, cultivar(s), and feature(s), select the DISPLAY button to draw the plots.

If two traits are "correlated" with each other, the marks in the plot will tend to make a pattern of dots leading up and to the right. If the traits vary inversely, the plotted points will travel down towards the right. If the dots appear random, then the two traits in this plot are not strongly related.

Trait Correlation Matrix

- Pick a season and look at the correlation of trait values across all cultivars in the season
- Explore which pairs of traits vary in a coordinated fashion



Correlation Between Traits

Select the Season to Model: S4 Hand Measurements

Select a Correlation Method: Pearson Correlation

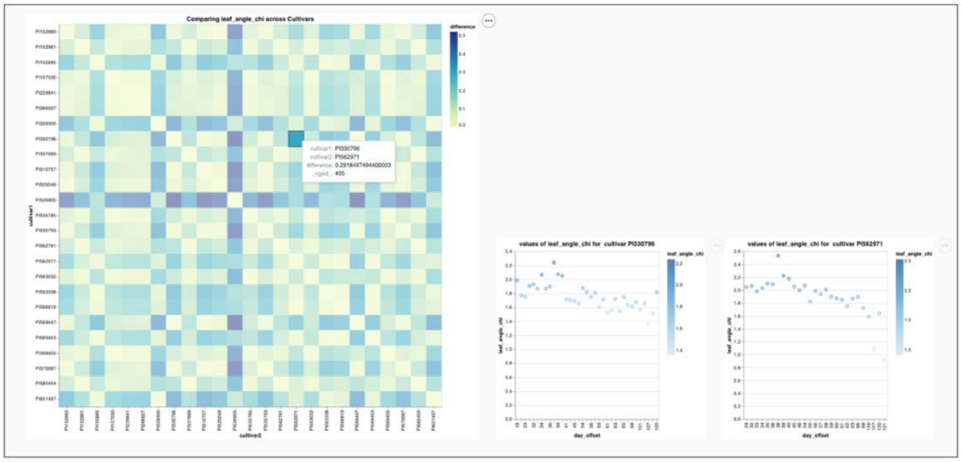
DISPLAY

Examine the trait correlation matrix. This application reads the trait measurements from the selected season and displays a pairwise correlation matrix showing how 'related' two different traits are for the measurements examined in this season.

Please start by selecting a season to examine on the left. Three different types of correlation algorithms are supported: Pearson Correlation, Kendell Tau Correlation, and Spearman Rank Correlation. Please choose the preferred method and then click the DISPLAY button to calculate and display the result.

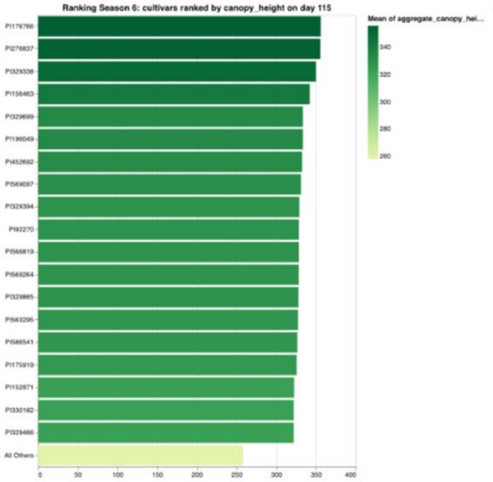
Compare Cultivar Pairs

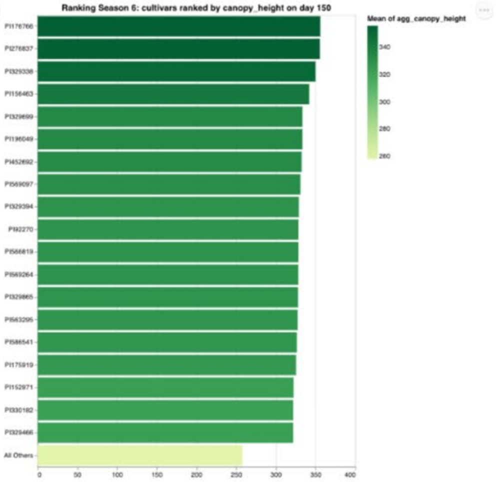
- Compare how a trait across the season between two different cultivars
- After a trait and a pair of cultivars is chosen, the scientist can observe trends in trait value during the season


Comparing leaf_angle_chi across Cultivars
difference
0.5
0.4
0.3
0.2
0.1
0.0
cultivar1
cultivar2
cultivar1: PI330796
cultivar2: PI562971
difference: 0.2918497494400003
vgsid: 405
values of leaf_angle_chi for cultivar PI330796
values of leaf_angle_chi for cultivar PI562971
leaf_angle_chi
day_offset

Top-10 Ranking

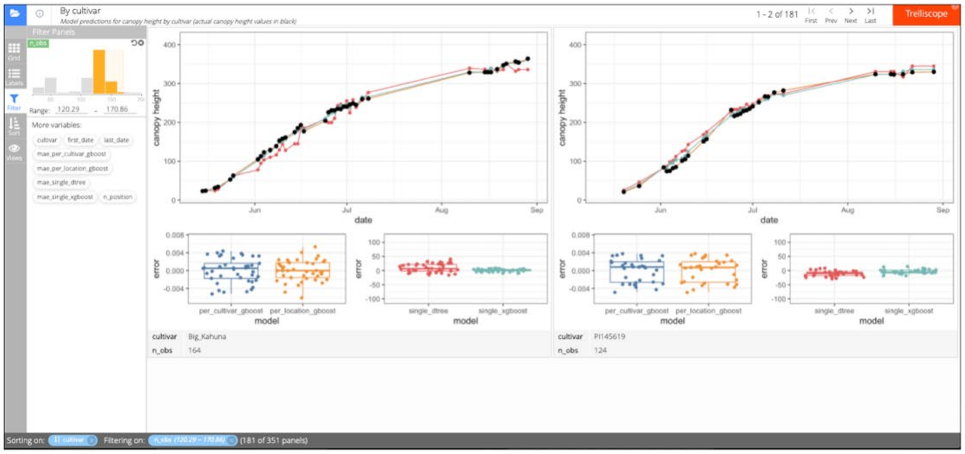
- Given any selected trait and a day of the growing season, rank the cultivars in order of the selected trait's average measured value





Trelliscope - Compare Model Fit Accuracy

- Flexible filtering and ranking of plots to explore model fitness against ground truth data



By cultivar

Model predictions for canopy height by cultivar (actual canopy height values in black)

Range: 120.29 - 170.86

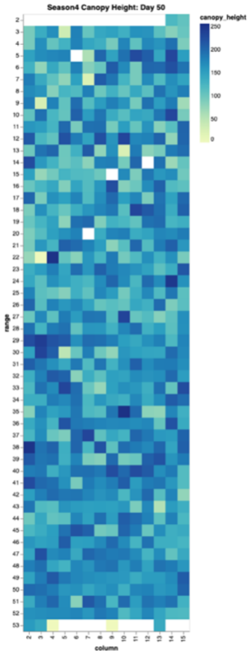
More variables:

cultivar, first_date, last_date, mae_per_cultivar_gboost, mae_per_location_gboost, mae_single_dtree, mae_single_xgboost, n_position

cultivar	Big_Kahuna
n_obs	164

cultivar	PI145619
n_obs	124

Sorting on: cultivar Filtering on: n_obs (120.29 - 170.86) (181 of 351 panels)



Closest Daily Measurement

- Closest Daily Measurement is an example custom display algorithm for traversing TERRA measurement data
- **We wanted scientists to be able to recreate the state of the observed field any day during the growing season.**
- Hypothesis: this would improve their ability to observe trends (showing example of day #15 vs day #50)
- The algorithm returns the most recent measurement available for each location (which might not be the same time as its neighbors)
- This retrieval algorithm is used in several of the interactive apps

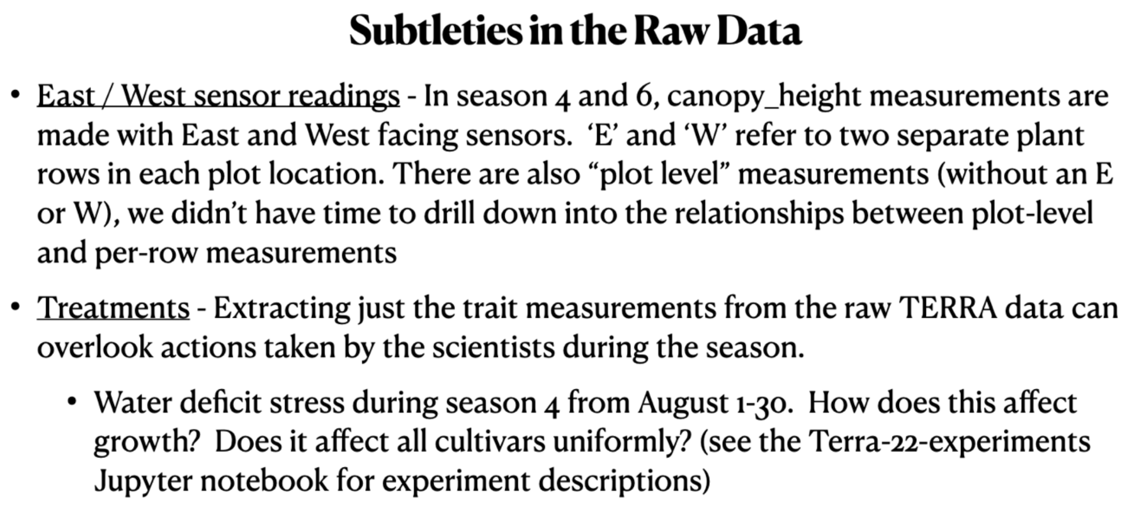
Potential New "problems" using TERRA-REF data

- Predict final biomass from early season canopy_height
- Explore the role of canopy_cover (% field covered from above) in the prediction of growth curves and biomass result
- Train on Season 4 and test on Season 6 or vise versa. Which BAP cultivars are consistent across seasons? Does this yield a "reliable cultivar" which is more dependable under different weather conditions?
- Augment TERRA-REF weather with NOAA or other weather measurements where there is a gap in the TERRA sensor data
- Use panicle measurements to refine final biomass results
- Does panicle data determine probability for each cultivar to "flower" (grow a panicle) as this stunts the plant growth. Do panicle-count have any effect or just how far into the season before panicles appear? (Curt's idea; not validated yet)

Potential New "problems" using TERRA-REF data

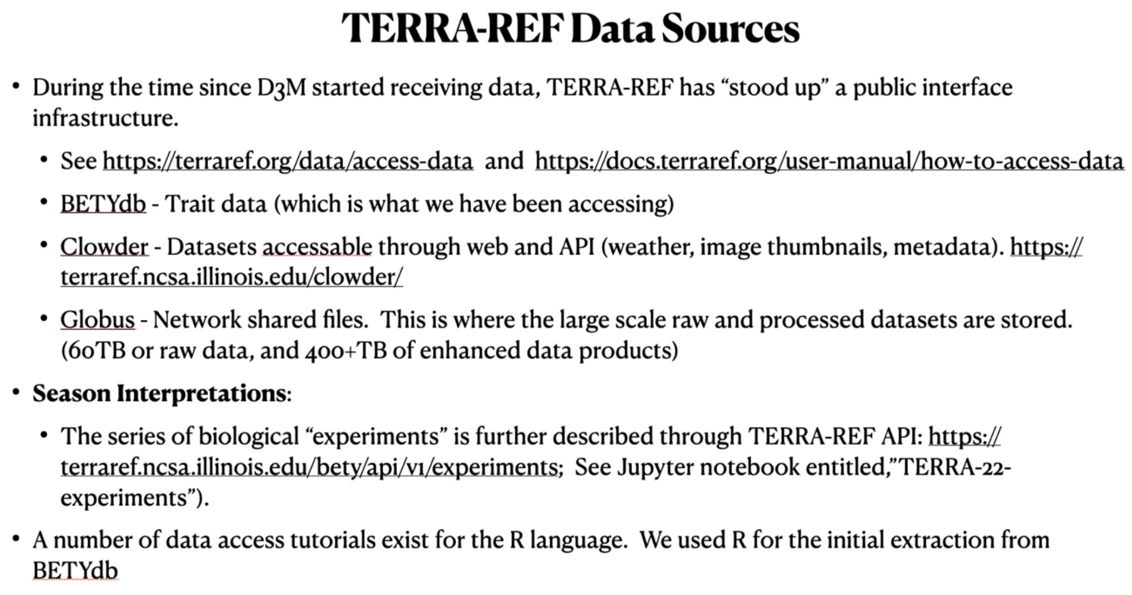
- Image Single or Multiple Regression - Train an image regressor using trait values and hyperspectral imagery. Predict on imagery from other cultivars. Which bands are most predictive? Do these bands correspond to certain physiological characteristics (e.g. chlorophyl level?)
- Image regression - Given annotated pictures of plants from different cultivars, can we predict the genetic distance between them? We believe time series RGB and 3D scanner data is available for many or all cultivars during season 4 and 6

Season 4 RGB Image from 1.5m overhead



Subtleties in the Raw Data

- East / West sensor readings - In season 4 and 6, canopy_height measurements are made with East and West facing sensors. 'E' and 'W' refer to two separate plant rows in each plot location. There are also "plot level" measurements (without an E or W), we didn't have time to drill down into the relationships between plot-level and per-row measurements
- Treatments - Extracting just the trait measurements from the raw TERRA data can overlook actions taken by the scientists during the season.
  - Water deficit stress during season 4 from August 1-30. How does this affect growth? Does it affect all cultivars uniformly? (see the Terra-22-experiments Jupyter notebook for experiment descriptions)



TERRA-REF Data Sources

- During the time since D3M started receiving data, TERRA-REF has "stood up" a public interface infrastructure.
  - See https://terraref.org/data/access-data and https://docs.terraref.org/user-manual/how-to-access-data
  - BETYdb - Trait data (which is what we have been accessing)
  - Clowder - Datasets accessable through web and API (weather, image thumbnails, metadata). https://terraref.ncsa.illinois.edu/clowder/
  - Globus - Network shared files. This is where the large scale raw and processed datasets are stored. (60TB or raw data, and 400+TB of enhanced data products)
- **Season Interpretations**:
  - The series of biological "experiments" is further described through TERRA-REF API: https://terraref.ncsa.illinois.edu/bety/api/v1/experiments; See Jupyter notebook entitled,"TERRA-22-experiments").
- A number of data access tutorials exist for the R language. We used R for the initial extraction from BETYdb

TERRA-Ref Relationship

- TERRA-REF scientists remain interested in models D3M teams can build using existing data
- Regarding the TERRA-REF challenge problem document: Some data mentioned is not available, but it seems some available problems listed haven't been attempted yet
- Hyperspectral imagery is very interesting for them, but this data is just becoming available. If D3M provides a date when data needs to be ready (e.g. for Summer workshop, etc.), they indicated they would try to provide suitable imagery for testing by that date
    - Schedule a follow-up data-focused meeting with TERRA data analysis team to identify a few problems and particular datasets to collaborate on further

LIST OF ACRONYMS

AR – Atmospheric Rivers; long, narrow filaments of enhanced water vapor carried in the upper atmosphere

ARPA-E – The Advanced Research Projects Agency of the Department of Energy, an Agency in the United States Government

AutoML – Automatic Machine Learning; This refers to any computational system designed that tries to select and fit mathematical, statistical, or deep learning models to incoming data with little or no intervention required from the user

BetyDB – Biofuel Ecophysiological Traits and Yields DataBase; BetyDB is an online resource through which early results from the TERRA-REF program were made publicly available for download. This is where our team acquired the TERRA-REF Season 4 and Season 6 data for processing.

DARPA – the Defense Advanced Research Projects Agency, a unit of the Department of Defense, an Agency of the United States Government. DARPA funded the D3M program among other research programs

D3M – The Data-Driven Discovery of Models program funded and lead by DARPA

ETL – Extract, Transform, and Load; this refers to the overall process of cleaning and adapting raw collected data before it is ready to be used to train mathematical models.

GRPC – Google Remote Procedure Call; a technology for creating well-defined interfaces between computer systems and exchanging messages between those systems. GRPC technology is used in the D3M program to communicate between a user interface and an AutoML system

IWV – integrated water vapor; this represents the presence of an accumulation of precipitation in an atmospheric region; it is a measurement of Atmospheric Rivers (AR) in the atmosphere

IVT – Integrated Vapor Transport; Mathematically different than IWV, but similar in definition, IVT is a metric used to characterize the amount of moisture carried and its relationship to wind velocity in Atmospheric Rivers

ML – Machine Learning; the technology focused on computational solutions for solving classification, regression, time-series, image analysis and other problems. Machine Learning is a body of scientific research focused on developing these methods for use in any application area

MLP – Multi-Layer Perceptron; An MLP is the simplest, canonical example of an artificial neuron that is a building block of neural networks commonly used in deep learning applications. Deep learning is a subspecialty within ML dealing exclusively with neural networks of different architectures.

JSON – The JavaScript Object Notation; JSON is a human-readable format for representing arbitrary hierarchical data. JSON is used extensively for exchanging datasets between systems across the internet. Many database systems now offer storage and indexing of data stored in JSON format.

MIT/LL – Massachusetts Institute of Technology / Lincoln Labs; a technology-focused unit of MIT consisting of staff available to be applied to externally funded research programs

SQL – the Structured Query Language; SQL is a syntax for posing data retrieval questions to relational database systems. MySQL and Postgres are examples of database systems that support SQL queries to retrieve data stored in a managed database

TA – Technical Area; this terminology is used in DARPA programs to delineate part of the problem being solved in the program. Each Task Area is a focus area of one or more teams working on the DARPA program

TA1 – Technical Area One; in the D3M program, this referred to the development and packaging of specific algorithms using a common interface, so they could be called from an automated system. Example TA1 algorithms include outlier detection, data normalization, data type determination, and statistical models.

TA2 – Technical Area Two (TA2) comprises development of the integrated machine learning platforms that receive data and fit models, using the algorithmic components from Task Area One.

TA3 - Technical Area Three (TA3) refers to the User Interface development, which allows a human user to interact with the overall system.

TERRA - Transportation Energy Resources from Renewable Agriculture; This is a portfolio (a group of related research programs) funded by ARPA-E to explore how to increase agricultural production to combat possible food shortages in the future

TERRA-REF – the TERRAphenoytping REFerence platform; This is a program funded by the ARPA-E TERRA Portfolio; the TERRA-REF program developed and deployed technology to carefully record information about the growth process of several types of agriculture to enhance the production of Biofuels and other agricultural products

USC – University of Southern California

XGBoost – the eXtreme Gradient Boosting model – a statistical model that extends decision tree technology and acts as an effective algorithmic ML method. XGBoost instances can trained from input data and then used to predict outputs that mimic the training data it has previously seen