**Investigator (Name and Department):**

Dr. Dominic Reisig, Entomology and Plant Pathology

**Student Consultants:**

Autumn Biggie - Plan of Action

John Hinic

Brennan Clinch-Description

**Project Description:**

**Plan of Action:**

To answer the research questions of the study, we will perform an initial exploratory data analysis, but the bulk of our efforts will be focused on modeling abscission of cotton bolls as a function of the other variables in the data set. To facilitate proper model training and performance measurement, we will split the data into a training and test set, with 75% of the data in the training set and 25% in the test set. The training data will be used to train the models, while the test data will be used to examine their accuracy. It is important to separate the training and test data to avoid introducing bias, since the models should not be evaluated based on how well they predict on data they have “seen before.”

Exploratory Data Analysis (EDA) will be conducted on only the training set as well in order to avoid the bias mentioned above. Our purpose is to ensure that our findings are applicable to cotton fields outside of this study, and that the models will predict with accuracy given new data.

The following methods will be considered as possible candidates for modeling: Logistic Regression, Decision (Classification) Tree, Boosted Trees, and Random Forest. We will use cross-validation to select the best model parameters when applicable, train the models using the training set, and evaluate their performance with the test set. The goal is to provide models that predict abscission with the highest accuracy while maintaining interpretability/utility of results. We will also explore the possibility of an interaction between feeding type and variety, and produce any related plots.

The Logistic Regression and Decision Tree models are expected to be the most interpretable. A decision tree graphic will be provided as a visual representation of the relationships between variables. The Boosted Trees and Random Forest models are expected to produce the highest accuracy, but are less interpretable due to “averaging” over many generated trees. However, we can still produce a plot of variable importance measures from the Random Forest model and compare to the results from other models.

**Next Steps: Who is responsible for what and by when.**