FarmView: Regression Analysis of 2016 Sorghum Composition

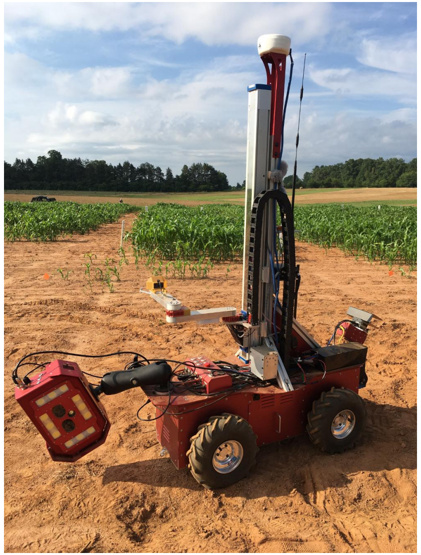
Ben Parr
bparr@cmu.edu

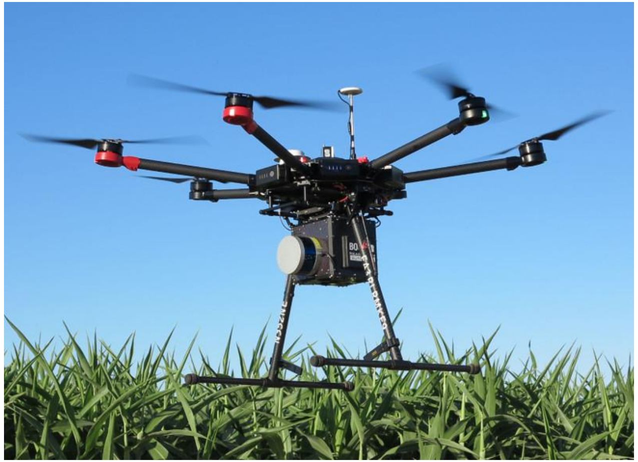
Sorghum bicolor

- Drought-tolerant and highly productive grass.
- Diverse gene pool containing over 40,000 genetic varieties.
- A preferred bioenergy candidate.
- Worldwide production is increasing.

Ground Robot and Aerial Drone

- Ground robot traverses rows of sorghum.
- Aerial drone flies above field of sorghum.



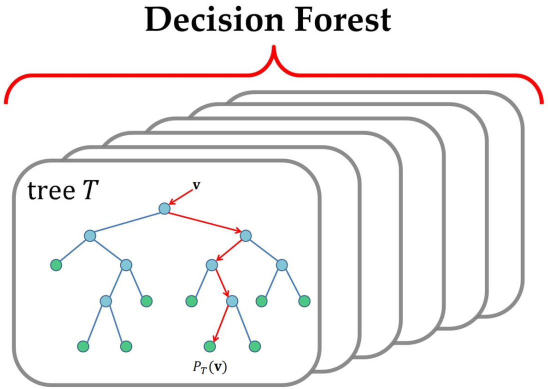


2016 Sorghum Dataset

- Field cultivated in Pendleton, South Carolina in 2016.
- 698 subplots (samples).
- 29 input features: accession (e.g. country of origin), GPS field location, ground robot, aerial drone, harvest phenotypes (e.g. harvested plant weight).
- 21 composition features as the output features.

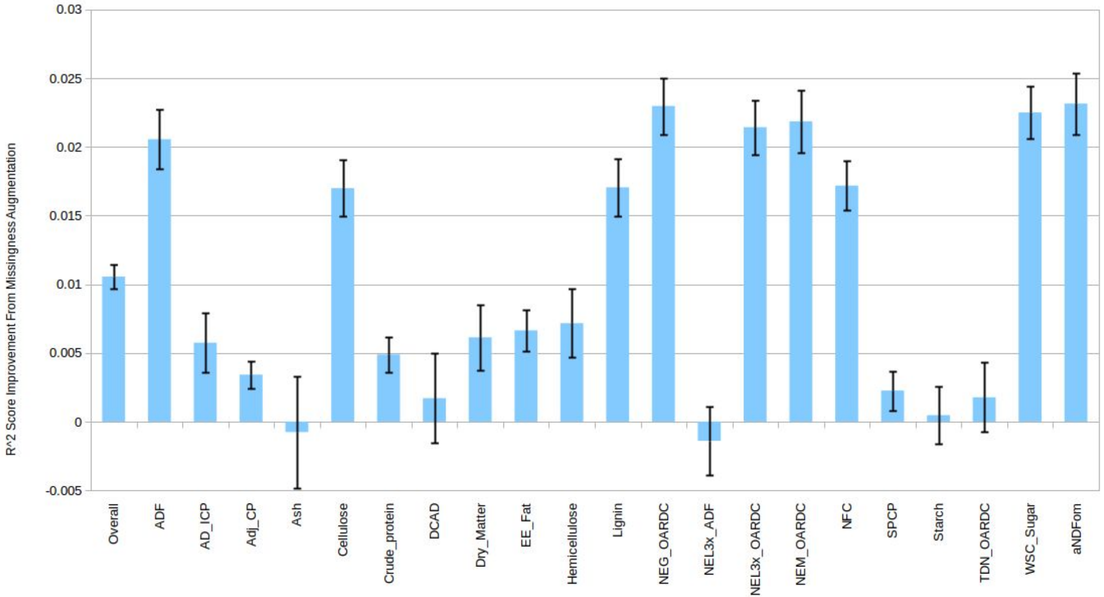
Random Forest Regressor

- First formally defined by Breiman in 2001.
- Final prediction is average of the outputs of its constructed regression decision trees.
- Fast and perform well in practice.



Missingness Augmentation

- Over half of the samples are missing >24% of the input features.
- 2.5% increase in r^2 score by augmenting the training set with training samples that had selected values removed and replaced with missing values.



Best Regressor: Random Forest

- Overall r^2 score of 0.436
  - Multi-dimensional r^2 score across all 21 output features.
- Random forest with 100 regression trees, max depth of 10.
- 10-fold cross validation.
- Full results, source code and input files available at https://github.com/bparr/dap/.