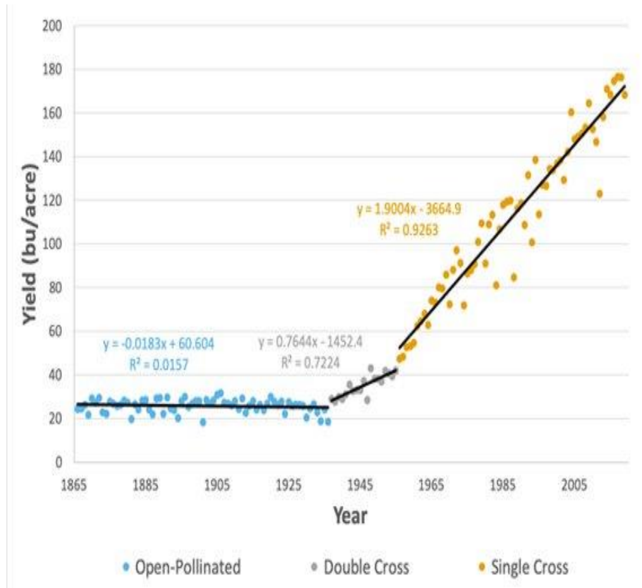
Using Genotype by Environment Interactions for Marker Selection in Maize

Ife Adgegbohungbe, Jenna Aguilar, Aziz Ebrahimi, Diana M. Escamilla Sanchez, Harrison Helmick, Dongqi Liu, Bilsha M. Mohan Das, Kobe Martin, Dhruv Narayanan, Amrish Nayak, Sven K. Nelson, George Ogilvie-Russell, Meenakshi Pavithran, Daniyaal Rasheed, Paula Sarmiento, Abhay Silina, Priyadarshini Subramaniam, Seth Tolley, Keishi Vannithamby

PURDUE UNIVERSITY

BAYER

Introduction

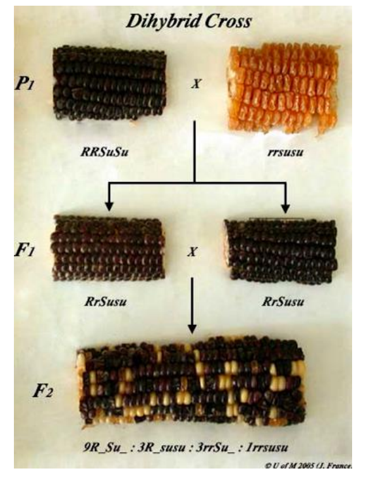


(USDA NASS: Ray *et al.*, 2013)

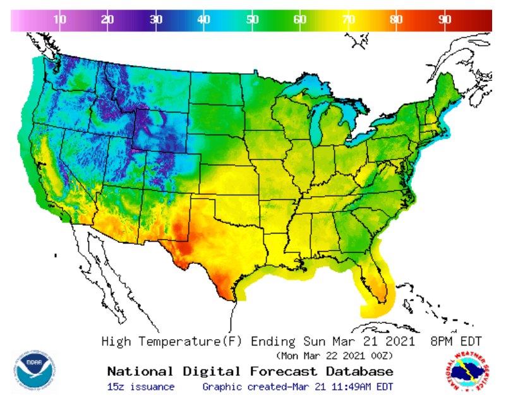
The plot to the left shows the trend of US Corn Yield over the past few hundred years. As we can see the yield has increased greatly over the past few hundred years due to improvements in breeding. ***However, it is estimated that to meet the demand of 2050, global corn production will need to double.***

Corn Yield is a product of

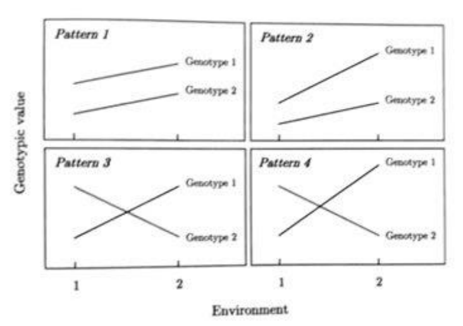
Genotype



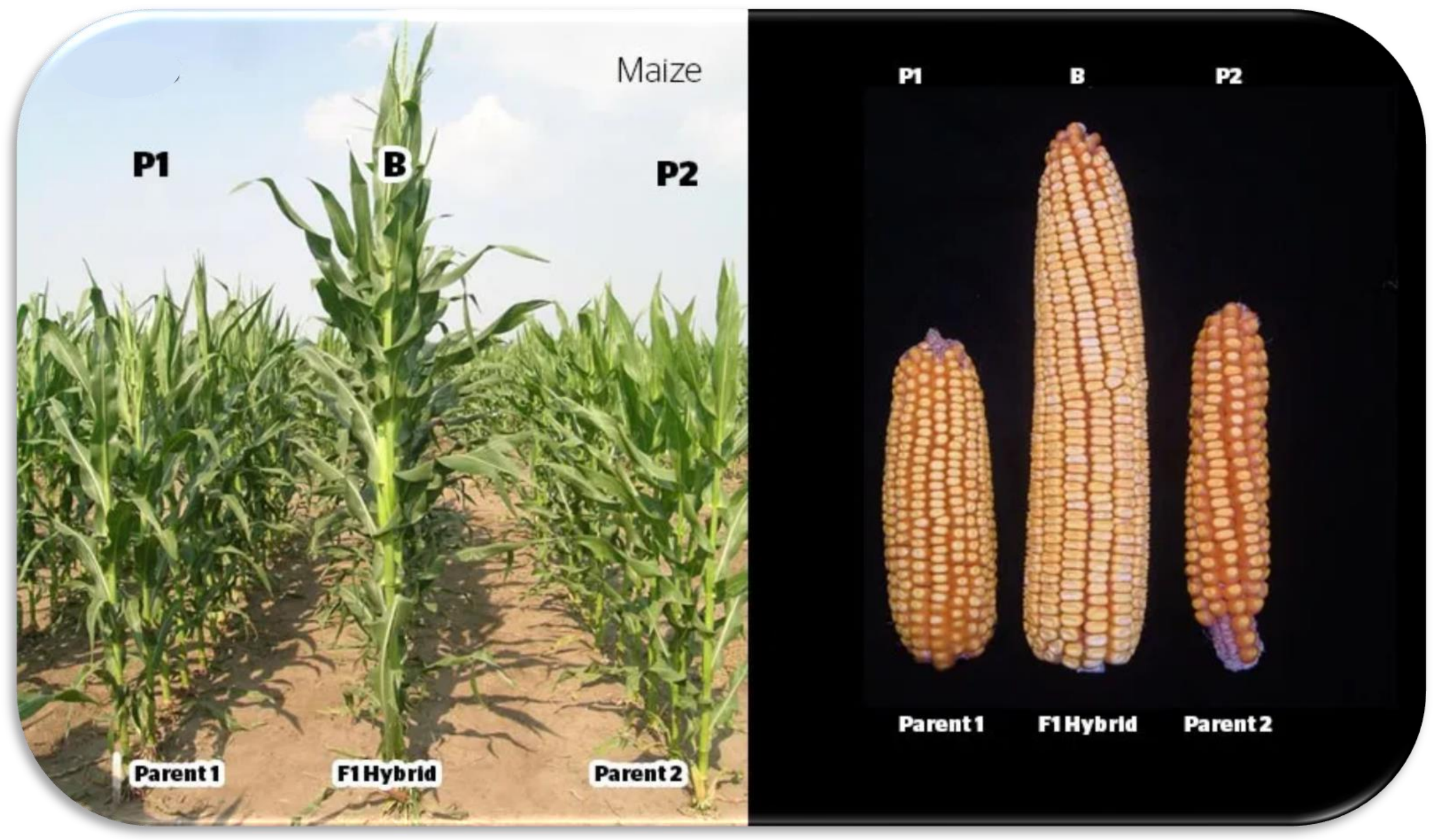
Environment


High Temperature(F) Ending Sun Mar 21 2021
National Digital Forecast Database

Genotype & Environment Interaction



The goal of our project was to **use machine learning to perform feature selection on genomic and environmental variables in maize hybrids**



Methodology

Dataset was divided in two clusters with inbred lines bred as either male or female, data provided included:

- Genetic information (Parents and progeny) of important markers.
- Phenotype data, including yield, height, weight maturity etc.
- And covariates as Location and Year.

All work was conducted on R4.0.0 and Python 3.6.11

Methodology

Data Wrangling

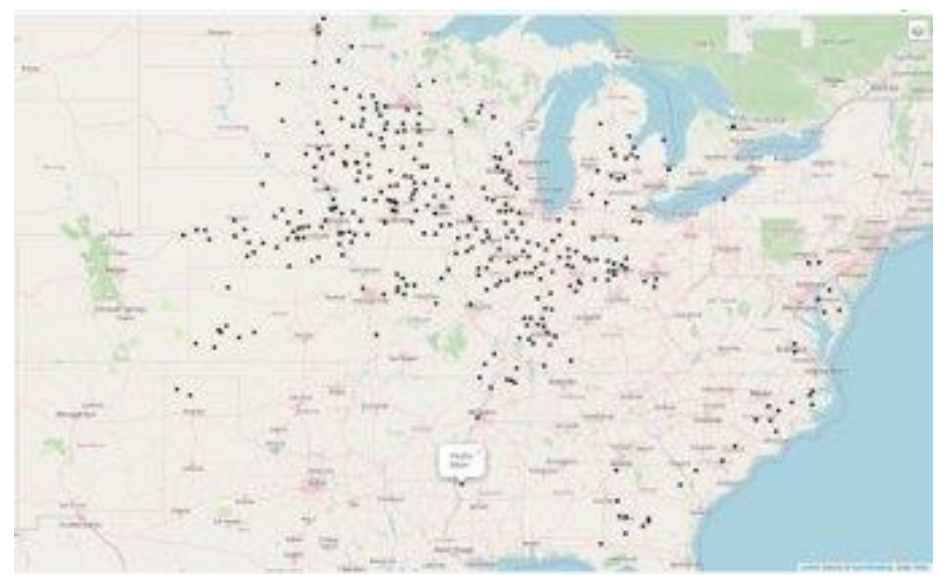
- Merging datasets
- Cleaning data
- Eliminated 317 of out-of-range values
- Outlier detection

Phenotypic metrics and ranges

Trait	Units	range
Ear Height (EHT)	inches	5-99
Estimated Relative Maturity (ERM)	days	60-180
Moisture (MST)	%	5-50
Plant Height (PHT)	inches	5-200
Root Lodging Percentage (RTLP)	%	0-100
Stalk Lodging Percentage (STLP)	%	0-100
Test Weight (TWT)	pounds/bushel	40-70
Yield Best Estimate (YLD_BE)	bushels/acre	1-400

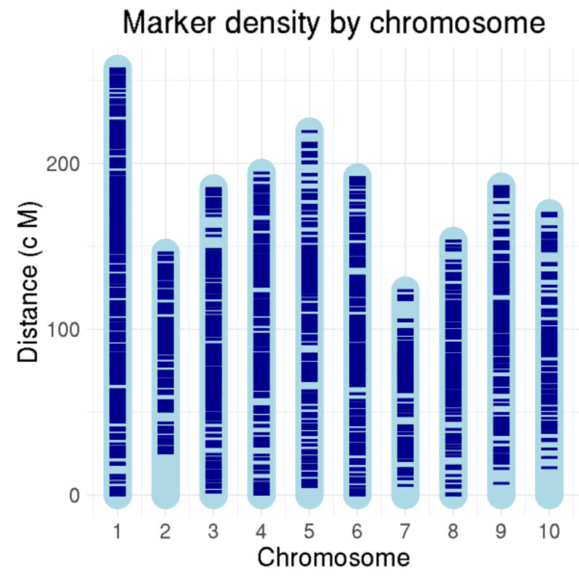
Assembling Environmental Datasets

- Weather stations close to field locations shown on map to the right
- In total, pulled data from 405 weather stations

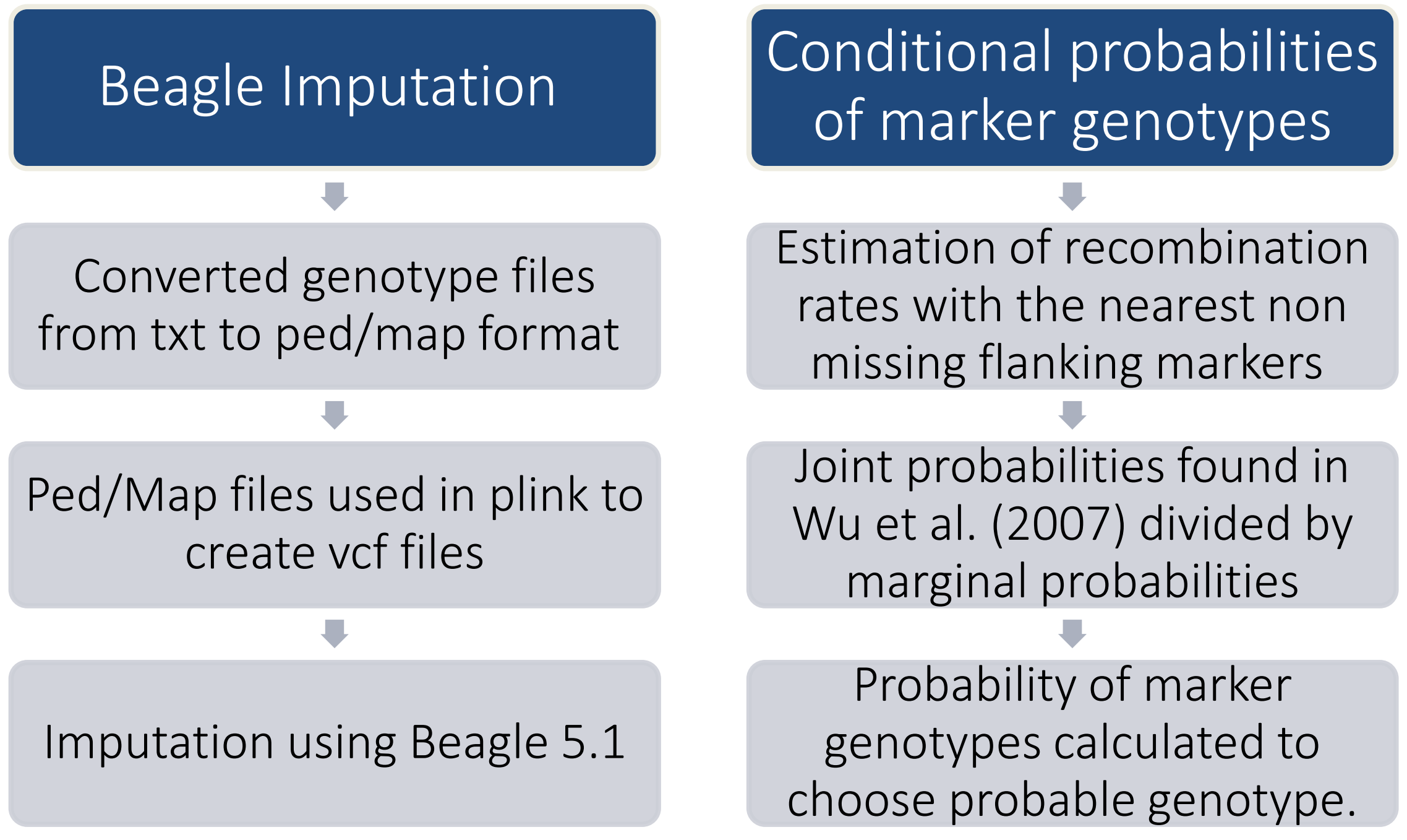


- Environmental Data Pulled from APIs:
  - NOAA (https://www.noaa.gov)
    - Precipitation, Temperature, Wind (16 metrics in 6 months)
  - SoilGrids (https://soilgrids.org)
    - Soil texture, nitrogen, organic carbon metrics, soil pH, and other soil metrics (16 metrics)
  - EPA Air Quality (https://www.epa.gov/outdoor-air-quality-data)
    - Air Quality Index (AQI) annual mean and maximums over daily averages (5 metrics)

Pipelines for Genomic Imputation



- Consensus linkage map used for imputation provided by Monsanto (shown to left)



Beagle Imputation

Converted genotype files from txt to ped/map format

Ped/Map files used in plink to create vcf files

Imputation using Beagle 5.1

Conditional probabilities of marker genotypes

Estimation of recombination rates with the nearest non missing flanking markers

Joint probabilities found in Wu et al. (2007) divided by marginal probabilities

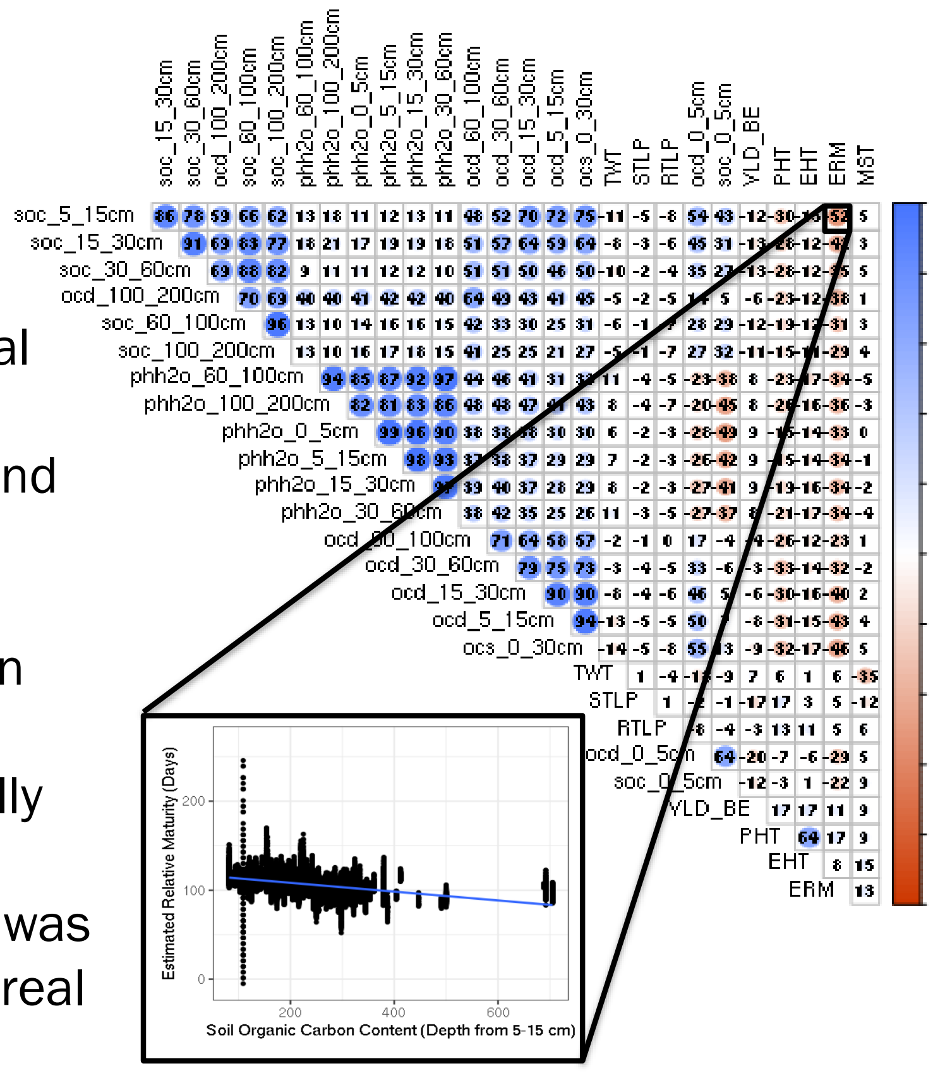
Probability of marker genotypes calculated to choose probable genotype.

Imputation of markers for male and female inbred populations

Results and Conclusions

Phenotype x Environment

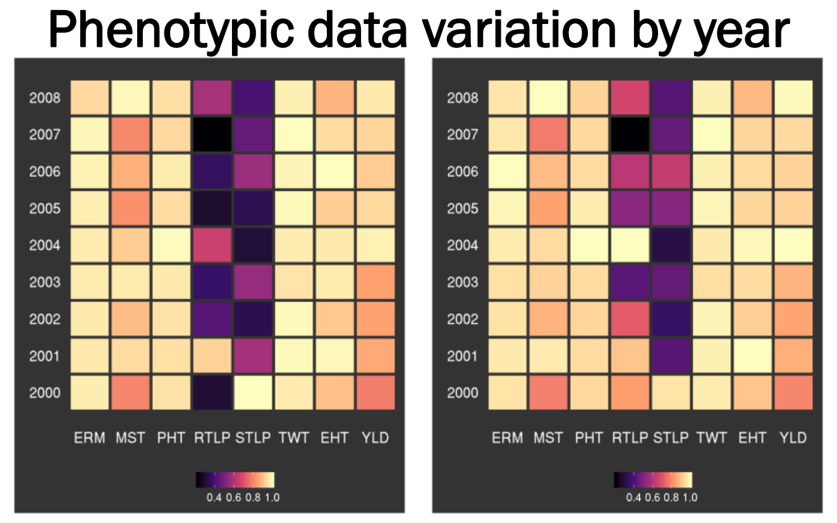
- Analysis of correlations between environmental metrics and phenotypic outputs (such as yield)
- To identify environmental factors impacting traits
- Such as nitrogen, AQI, and soil organic carbon
- For comparisons with a high absolute correlation (see correlation plot), scatter plots were visually inspected to examine whether the correlation was linear and likely to be a real relationship.



Year-to-year variance

- Heatmap at right shows variance for each phenotype by year.
- MST and EHT are variable across years
- Increasing trend in YLD is consistent with genetic improvements each year.

Phenotypic data variation by year



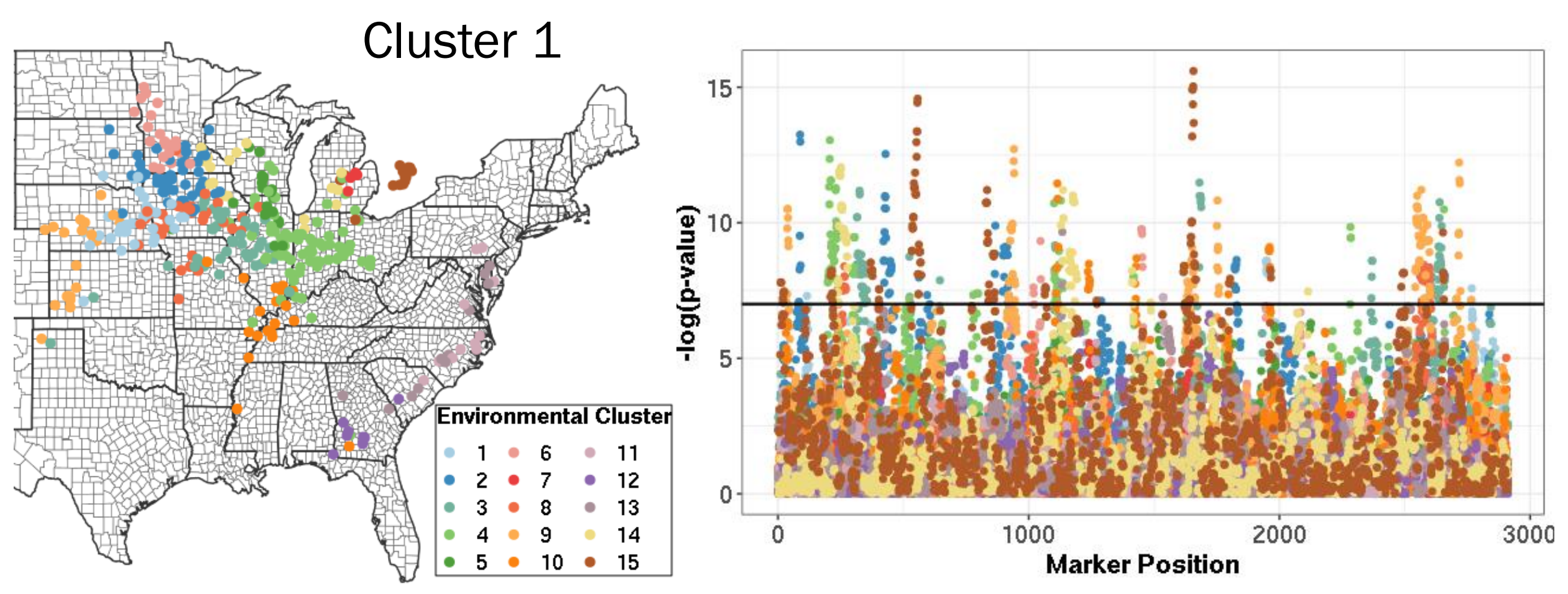
Cluster 1 Cluster 2

*Values normalized as a percentage of the max value.

Feature Selection

Selected 592 markers based on a significant association of markers and yield within environmental clusters

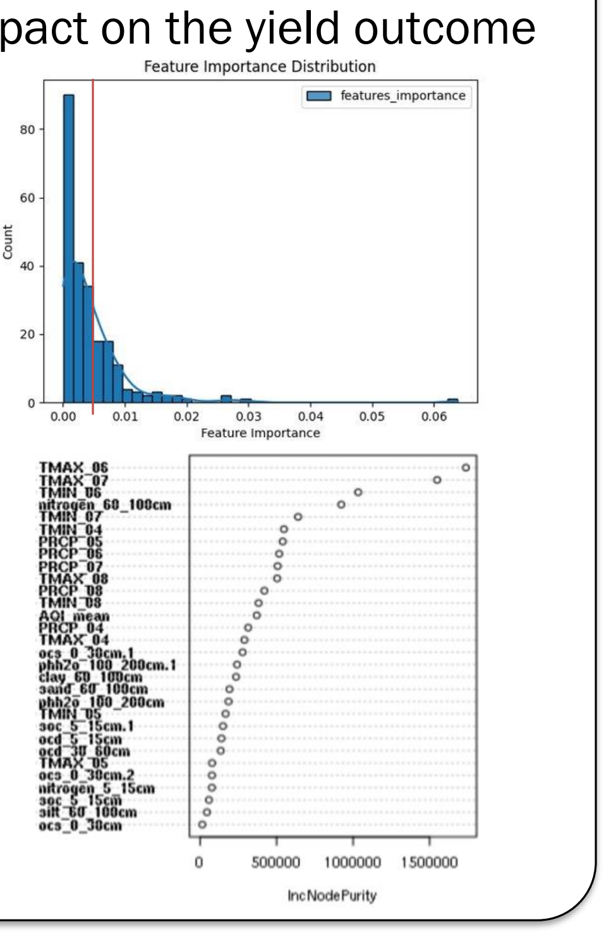
Cluster 1



Only a few environmental variables have a big impact on the yield outcome

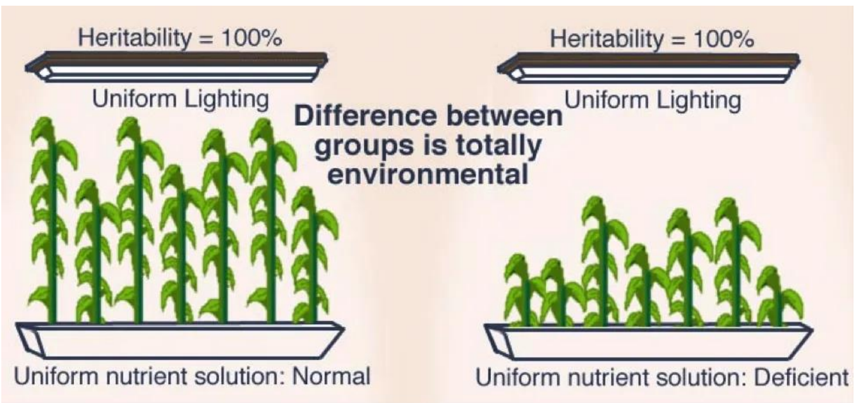
- Random forest for markers and environmental features
- Most important variables in yield predictions
  - Location
  - Temperature
  - Precipitation
  - Air Quality
  - Soil Type

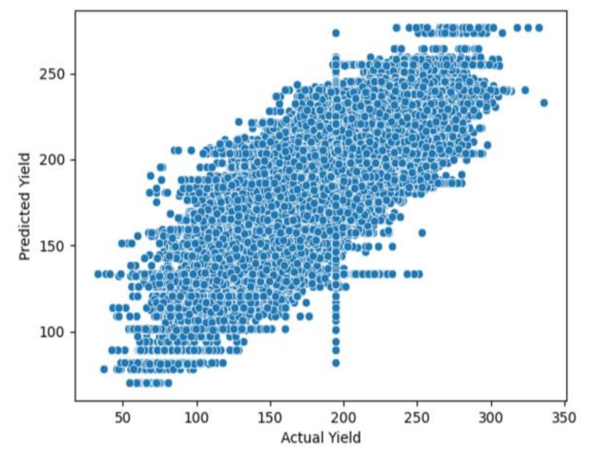
Traits with highest importance value used for predicting Yield.



Next steps

- Build prediction models for yield using Linear Regression and Random Forest
- Work towards developing more sophisticated Models such as Deep Neural Networks that could make more accurate predictions.
- Explore different sets of variables (genotype, environment) and alternate feature selection.
- Test how different imputation methods have effects on the model.
- Heritability Measures
  - Analysis and Marker Identification
  - Feature Selection and Prediction


Heritability = 100%
Uniform Lighting
Difference between groups is totally environmental
Heritability = 100%
Uniform Lighting
Uniform nutrient solution: Normal
Uniform nutrient solution: Deficient



Further reading

Barrera-Redondo J, Piñero D, Eguiarte LE (2020) Genomic, Transcriptomic and Epigenomic Tools to Study the Domestication of Plants and Animals: A Field Guide for Beginners. Front Genet **11**: 1

Bernardo, R. (2014). Essentials of Plant Breeding. Minnesota: Stemma Press.

Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, et al (2016) Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. Front Genet **7**: 1

Browning, S and Browning, B. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Jacobson A, Lian L, Zhong S, Bernardo R (2015) Marker Imputation Before Genomewide Selection in Biparental Maize Populations. The Plant Genome **8**: plantgenome 2014.10.0078

Jacobson A, Lian L, Zhong S, Bernardo R (2014) General Combining Ability Model for Genomewide Selection in a Biparental Cross. Crop Science **54**: 895–905

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics **9**: 166–177

Maydica DD, 2005 Genetic progress in yield of United States maize (Zea mays L.). ask-force.org

NOAA. API Web Service https://www.weather.gov/documentation/services-web-api.

Perez-Elizalde S, Cerón-Rojas JJ, Crossa J, Fleury D, Alvarado G (2014) Rindsel: An R Package for Phenotypic and Molecular Selection Indices Used in Plant Breeding. *In* Crop Breeding. Humana Press, New York, NY, New York, NY, pp 87–96

Ray D, Mueller ND, West PC, Foley, JA (2013). Yield Trends are Insufficient to Double Global Crop Production by 2050. PLoS ONE. 8(6): e66428. doi: 10.1371/journal.pone.0066428.

Robertsen CD, Hjortshøj RL, Agronomy LJ, 2019 Genomic selection in cereal breeding. mdpi.com

Shamshad M, Sharma A (2018) The Usage of Genomic Selection Strategy in Plant Breeding. Next Generation Plant Breeding. doi: 10.5772/intechopen.76247

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal **6**: 330–340

Acknowledgements

Nick Ames, PhD
Susana Diaz-Amaya, PhD
Fabiana Moreira, PhD
Mark Daniel Ward, PhD
Maggie Betz
Brian Dilkes, PhD
Pete Pascuzzi, PhD
Sofía Brandariz, PhD
Wei-Ting Liao, PhD
Adam Scott, PhD

The Data Mine Corporate Partners Symposium 2021