Predicting Phenotype from Multi-Scale Genomic and Environment Data using Neural Networks and Knowledge Graphs: An Introduction to the NSF GenoPhenoEnvo Project

Anne E Thessen, Michael Behrisch, Emily J Cain, Remco Chang, Bryan Heidorn, Pankaj Jaiswal, David LeBauer, Ab Mosca, Monica C Munoz-Torres, Arun Ross, Tyson Swetnam

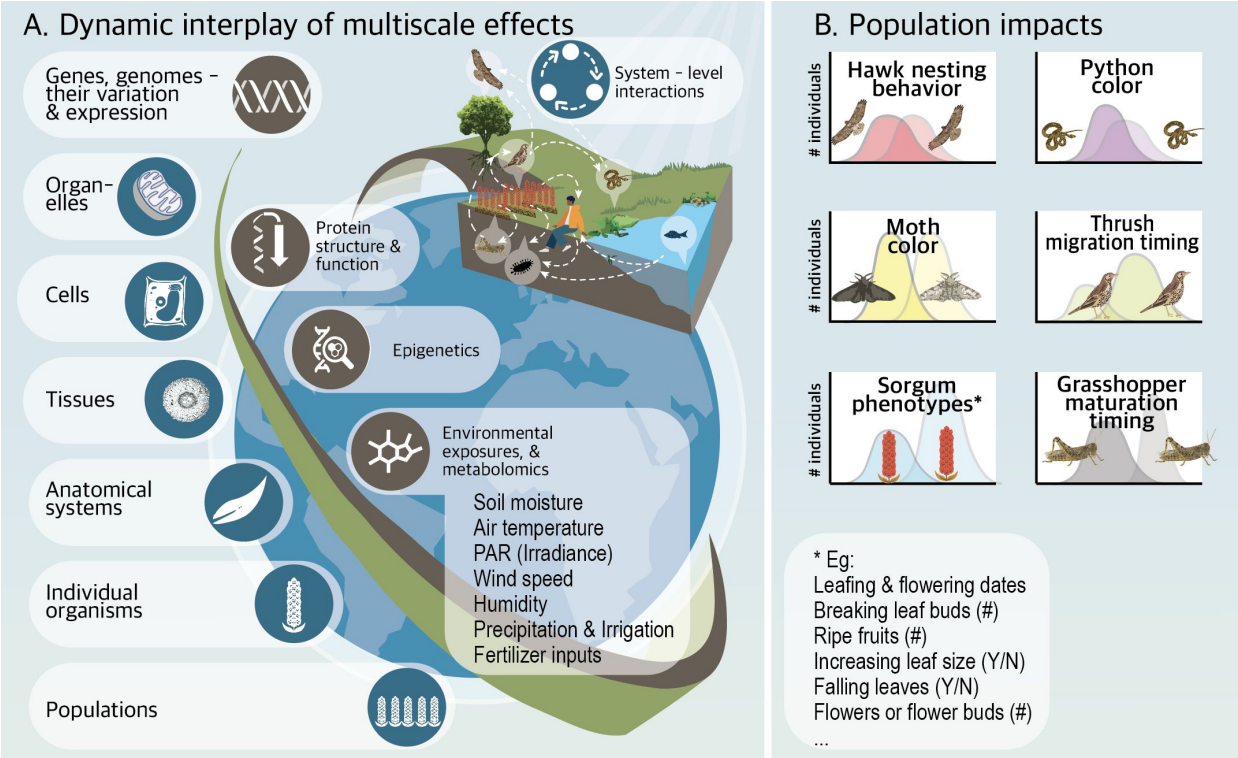
Acknowledgements

- Translational and Integrative Sciences Lab OSU (tislab.org)
- Two new members: Ishita Debnath (MSU) and Ryan Bartelme (UA)
- NSF Ideas Lab
- NSF Award 1940330 Harnessing the Data Revolution

TERRAPHENOTYPING
REFERENCE PLATFORM

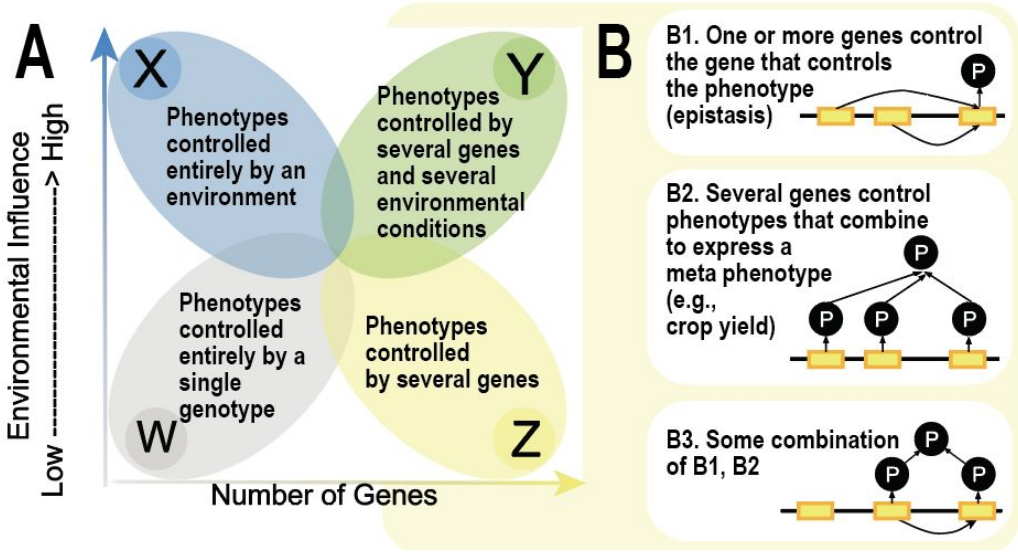
Predicting Phenotype from Genes and Environments

- G + E = P only works in the simplest systems, if at all
- There's a lot we don't know about how genes are translated into phenotypes
- How do phenotypes affect the ecosystem?



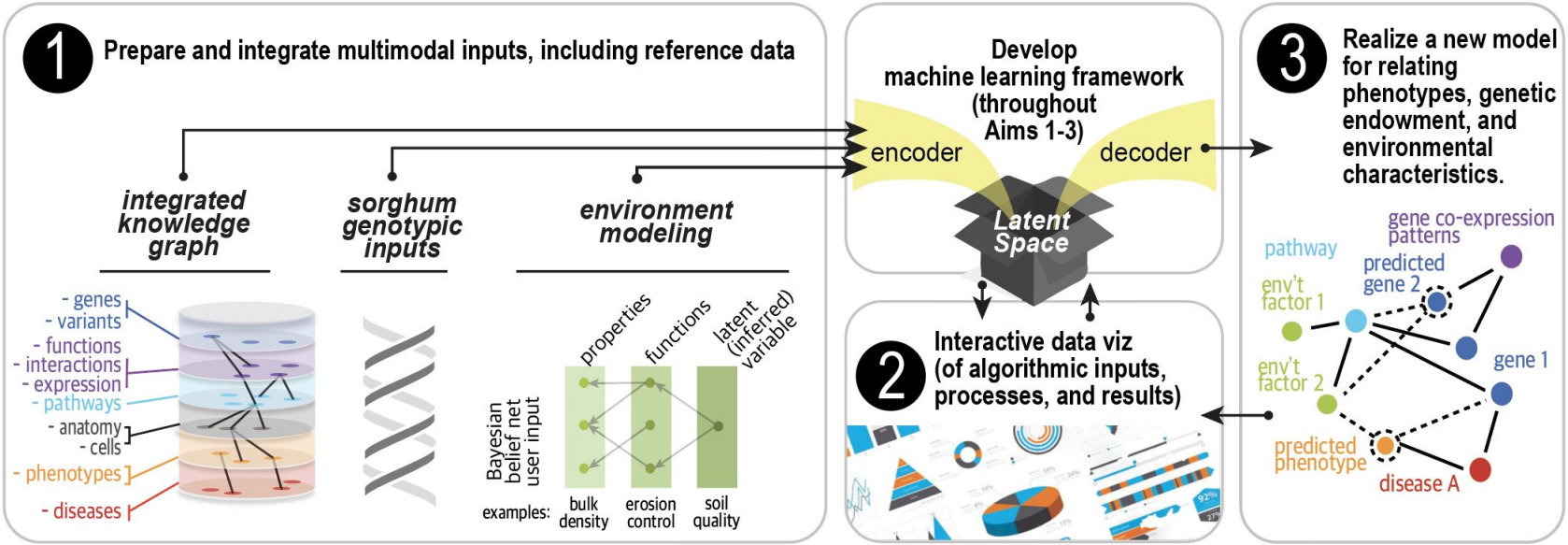
How can we predict phenotype given an organism's environmental conditions and genomic endowment?

- State-of-the-art statistical modeling has led to many insights, but has been applied to very controlled systems.
- Getting the phenotype is only part of the answer.
- Can we use the predictive model to reveal hidden processes? Critical variables?



Machine Learning for Results and Process

- Pros
  - Capable of coping with non-linearity in biological systems
  - Find hidden relationships
- Cons
  - Output is opaque
  - Not enough curated data not available
  - Data that are available not “ML ready”



GenoPhenoEnvo Project

- Goal: Develop a machine learning framework capable of predicting phenotypes based on multi-scale data about genes and environments.
  - Leverage existing, well-structured, cross-species reference data about genes and phenotypes
  - Provide interactive data visualizations for examining and interpreting the “black-box” behavior of ML models and their results
  - Realize a new model for relating phenotypes, genetic endowment, and environmental characteristics
- Just started Oct 1
- Now in Year 1 Q2

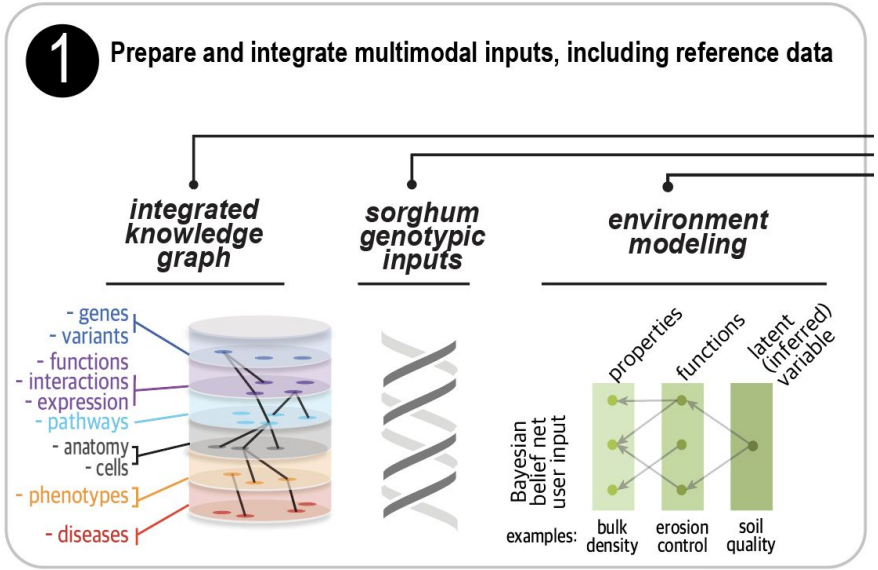
Training Data

Year 1

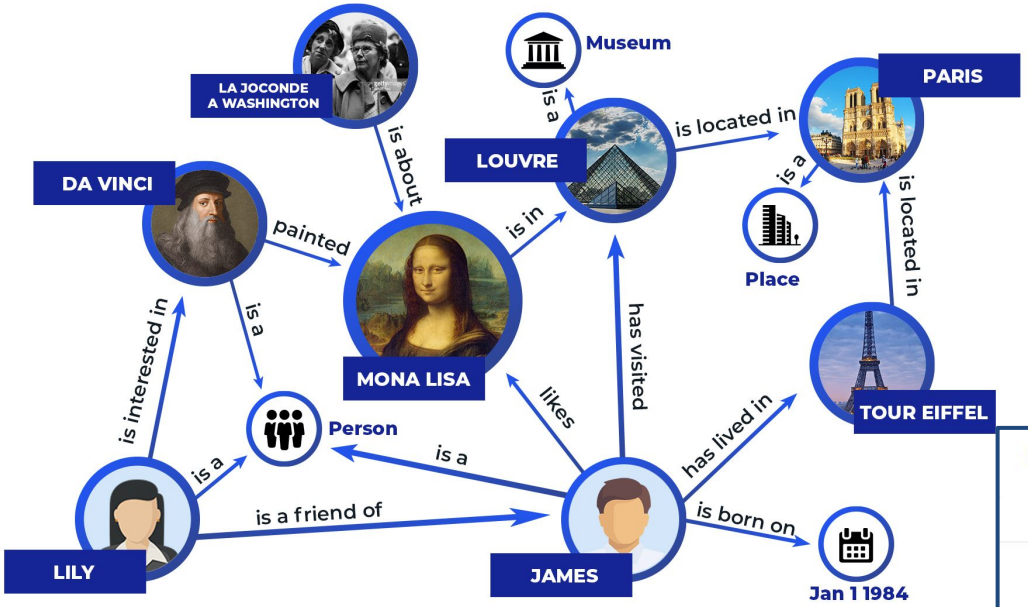
- TERRA-REF sorghum
- Heavily controlled and measured environment data
- Thorough genotyping and phenotyping
- Knowledge graph links
- **How do we prepare these data to be ML ready?**

Year 2

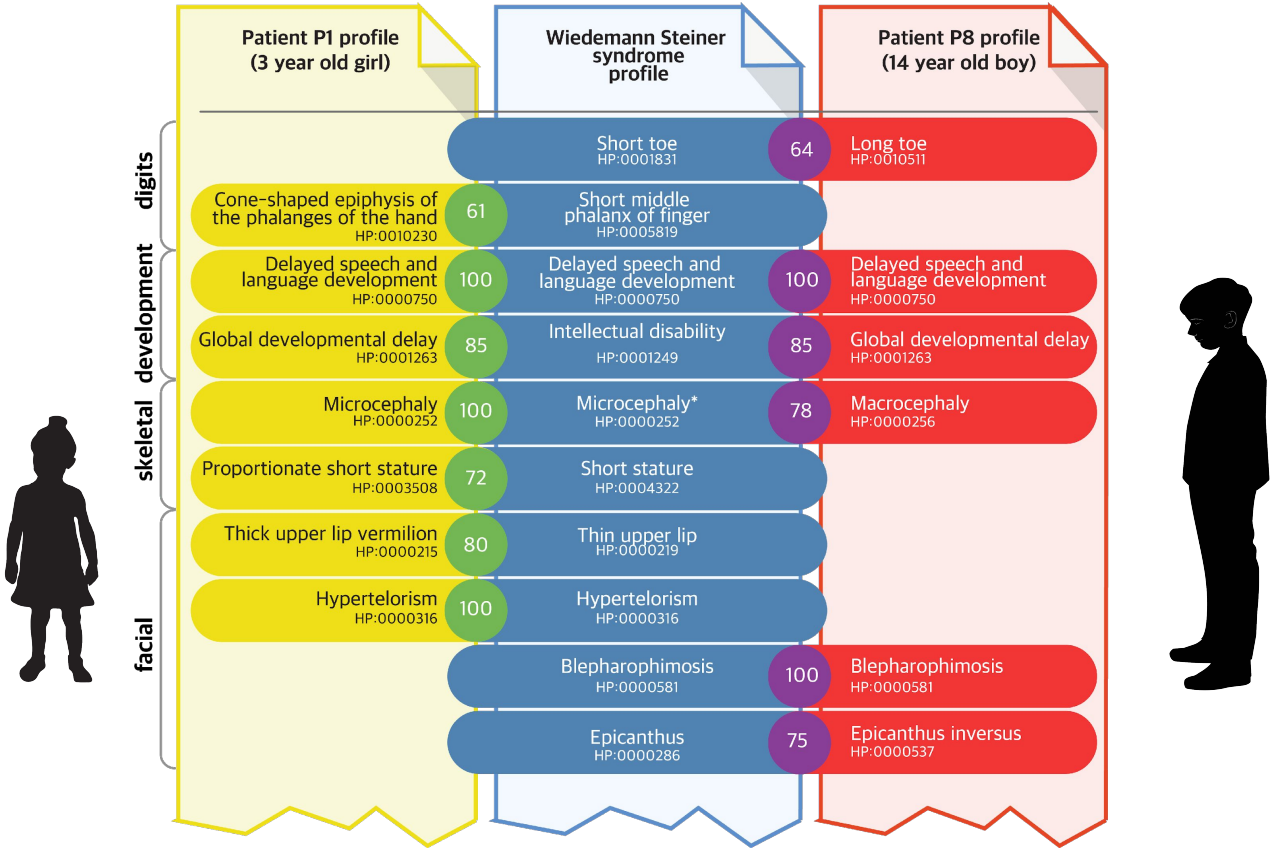
- TERRA-REF wheat
- NEON, EOS
- Citizen science phenology



What is a Knowledge Graph?

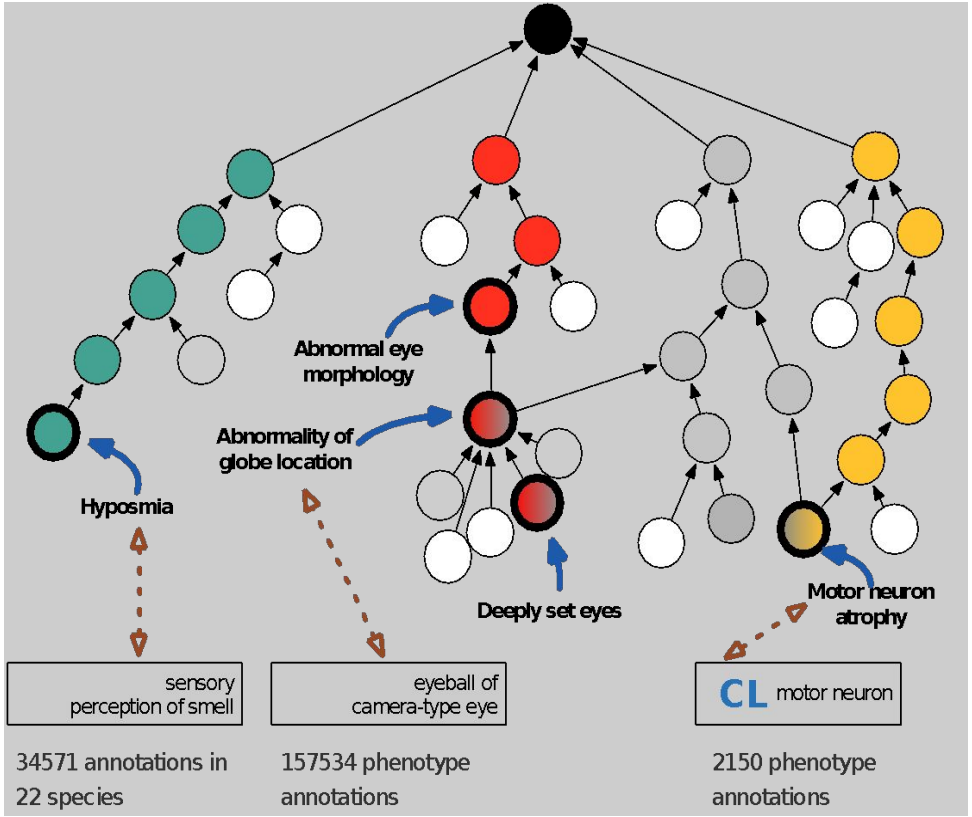


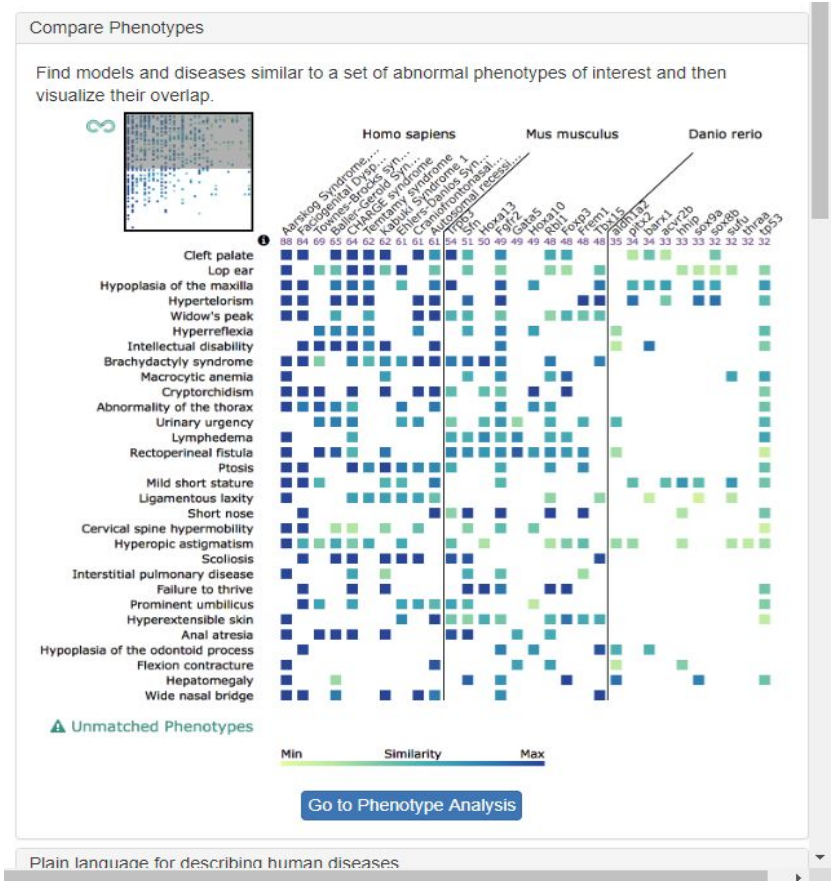
How Can Knowledge Graphs Help?



DOI: 10.1126/scitranslmed.3009262

How Can Knowledge Graphs Help?





How Can Knowledge Graphs Help?

- Constrain ML and prioritize results
- Quality control - sanity check
- Integrate heterogeneous data
  - Manage terminology
  - Manage scale and granularity
- Find new relationships
- Fill in data gaps with inferencing

Training Data - Genomic

List 1: TERRA-REF data for Y1

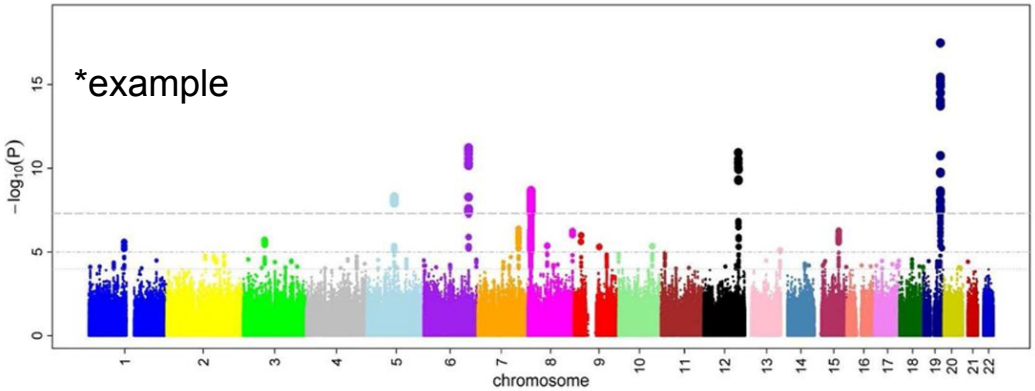
Gene Information (G)
- Sorghum whole genome
- Sorghum genotypes[219]

Phenotype Information (P)
- Emergence Date
- End of Season Biomass
- End of Season Height
- Flowering Date

Environment Information (E)
- Soil moisture
- Air temperature
- PAR (Irradiance)
- Wind speed
- Humidity
- Precipitation and Irrigation
- Fertilizer inputs

GWAS results can be combined with the knowledge graph results to reduce input variables for ML

- Use VCF files and phenotype data from TERRA-REF
- SNP to phenotype associations with p values and effect sizes (GWAS)
- Manhattan plot for all phenotype data



By M. Kamran Ikram et al - Ikram MK et al (2010) Four Novel Loci (19q13, 6q24, 12q24, and 5q14) Influence the Microcirculation In Vivo. PLoS Genet. 2010 Oct 28;6(10):e1001184. doi:10.1371/journal.pgen.1001184.g001, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=18056138

Training Data - Phenomic

TERRAPHENOTYPING REFERENCE PLATFORM

List 1: TERRA-REF data for Y1

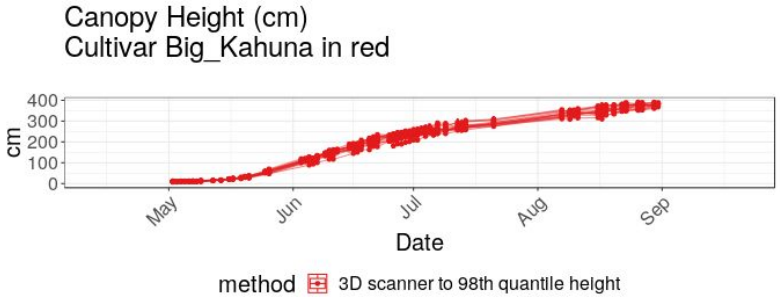
Gene Information (G)
- Sorghum whole genome
- Sorghum genotypes[219]

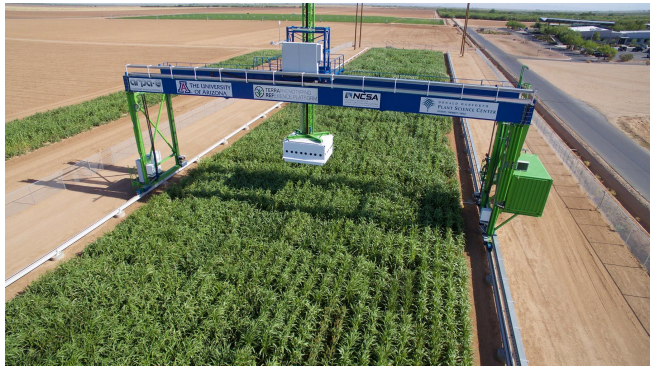
Phenotype Information (P)
- Emergence Date
- End of Season Biomass
- End of Season Height
- Flowering Date

Environment Information (E)
- Soil moisture
- Air temperature
- PAR (Irradiance)
- Wind speed
- Humidity
- Precipitation and Irrigation
- Fertilizer inputs

- Days to flowering
- Growing Degree Days to flowering
- Days to flag leaf emergence
- Growing Degree Days to flag leaf emergence
- Canopy height
- End of season canopy height
- Above ground dry biomass at harvest

Phenotypes can be represented as categories or integers (10 cm or decreased height?)





Training Data - Environmental

List 1: TERRA-REF data for Y1
Gene Information (G)
● Sorghum whole genome
● Sorghum genotypes[219]
Phenotype Information (P)
● Emergence Date
● End of Season Biomass
● End of Season Height
● Flowering Date
Environment Information (E)
● Soil moisture
● Air temperature
● PAR (Irradiance)
● Wind speed
● Humidity
● Precipitation and Irrigation
● Fertilizer inputs

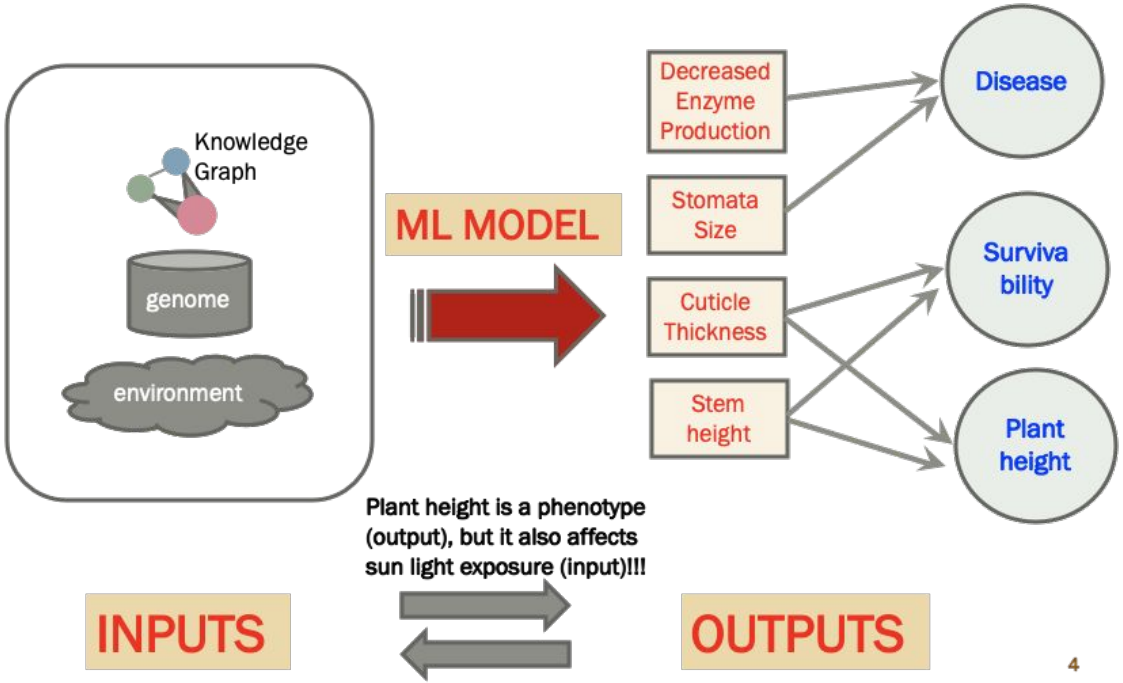
- Air temperature
- Relative humidity
- Precipitation
- Wind speed and direction
- Growing degree days
- Cumulative precipitation

Data from weather station and gantry

Abstracted to daily average, min, and max

Machine Learning - Preliminary

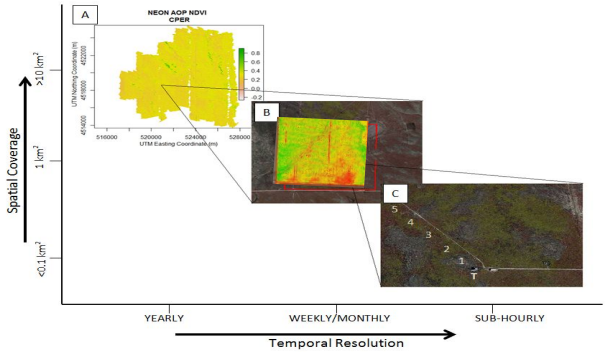
1. Regression Models
2. Simple Neural Networks
3. Deep Neural Networks



Year 2 - Expanding to Ecosystems (*Preliminary*)

Leverage data and resources from multiple NSF supported programs:

- Analyze genetic and remote sensing data from NEON
- Utilize the CyVerse Data Science Workbench
- XSEDE for ML computations



List 2: Observational & EOS data for Y2	
Gene Information (G) • Cottonwood whole genome • Cottonwood genotypes • Sorghum whole genome • Sorghum genotypes • Wheat whole genome • Wheat genotypes	**Phenotype Information (P)** • Leafing date • Flowering date • Breaking leaf buds (#) • Ripe fruits (#) • Increasing leaf size (Y/N) • Falling leaves (Y/N) • Colored leaves (Y/N) • Flowers or buds (#) • Open flowers (#) • Pollen release (Y/N) • Recent fruit/seed drop (Y/N) • Fruits (#) • NDVI (EOS)
Environment Information (E) • Soil moisture (e.g., SMAP) • Precipitation (Daymet) • Air temperature (Daymet) • PAR (NARR) • Soil Type (USDA)	

National Ecological Observatory Network

XSEDE
Extreme Science and Engineering Discovery Environment

GenoPhenoEnvo Project Information

- Join our Google Group
- Watch our GitHub Repo
  github.com/genophenoenvo
- Search Twitter hashtag
  #GenoPhenoEnvo
- Visit the project web page

- Anne E Thessen
- annethessen@gmail.com

Questions?