CORTEVA™ agriscience

Trends of predictive breeding

Changes in the plant breeding landscape driven by more, better and different data

Alencar Xavier
Breeding Analyst at Corteva Agrisciences
Adjunct professor at Purdue University

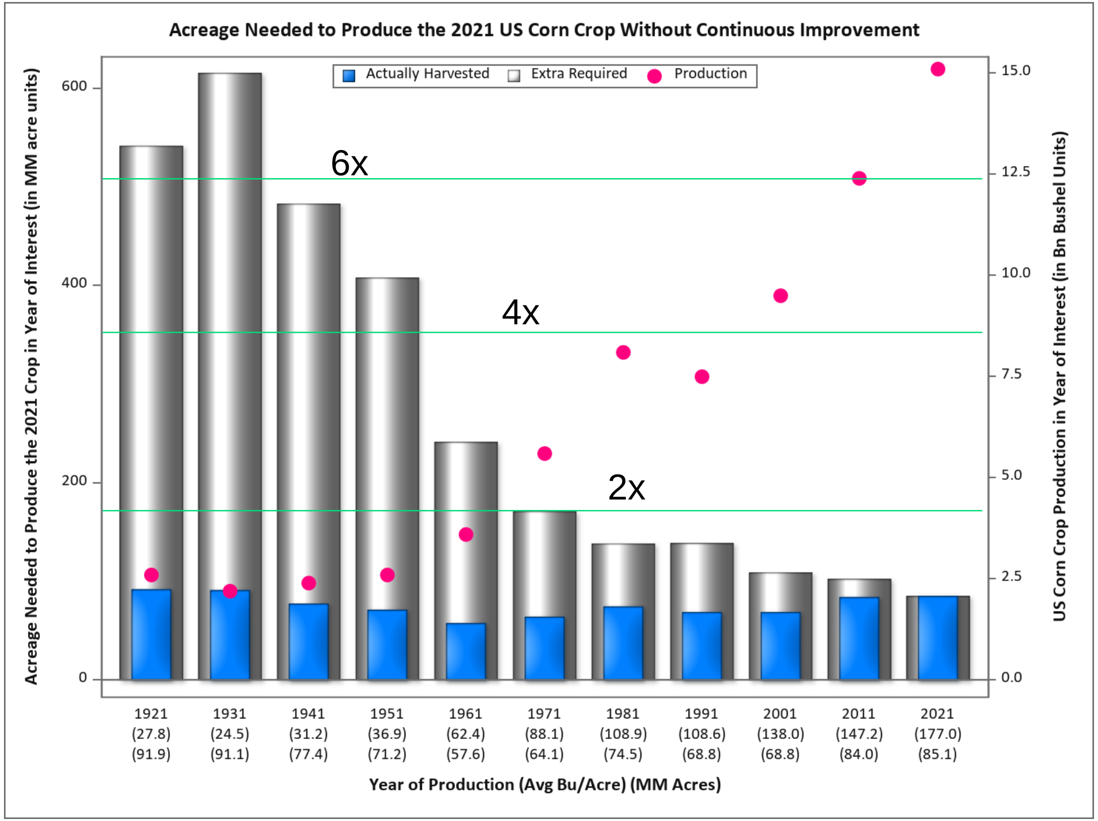
02/2024

Contributions from Radu Totir, Frank Technow, David Habier, David Bubeck, Abelardo de la Vega

1. **Introduction**
    - Rationale
    - Breeding with ML
2. **Modeling**
    - Approaches
    - Correlated information
3. **Analytics**
    - Breeding objectives
    - Target G x E x M
    - Validation
4. **Conclusion**

1. **Introduction**
   - Rationale
   - Breeding with ML
2. **Modeling**
   - Approaches
   - Correlated information
3. **Analytics**
   - Breeding objectives
   - Target G x E x M
   - Validation
4. **Conclusion**

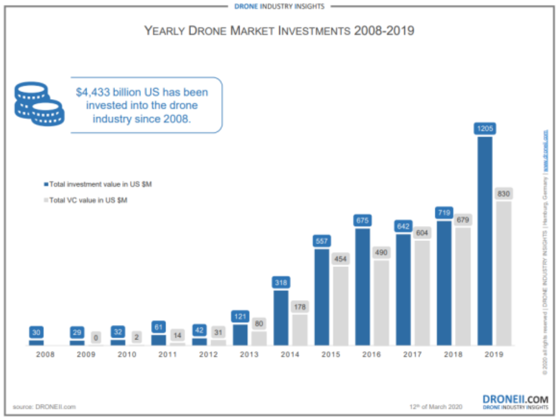
What are some implications of continuous Corn Improvement?

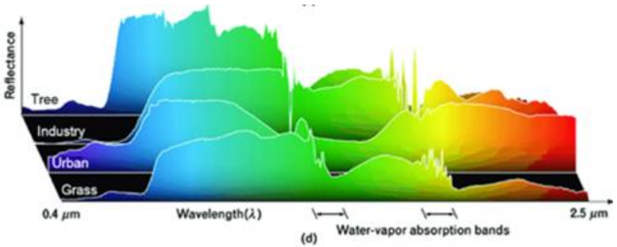


Source: Totir 2021, ASTA

*Based on 2021 USDA NASS data

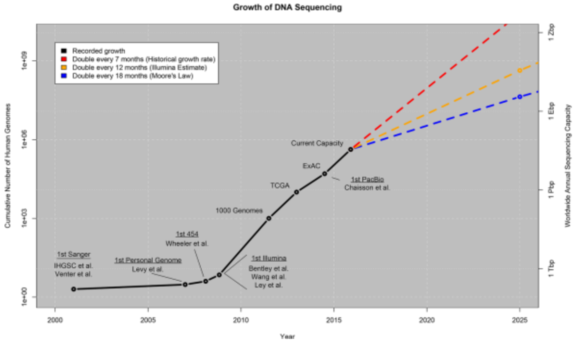
More Pheno



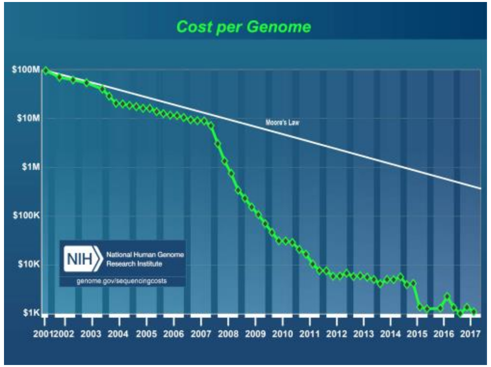


https://www.mdpi.com/2076-3417/12/5/2570

More Geno



The Cost of Sequencing a Human Genome. NIH.
https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

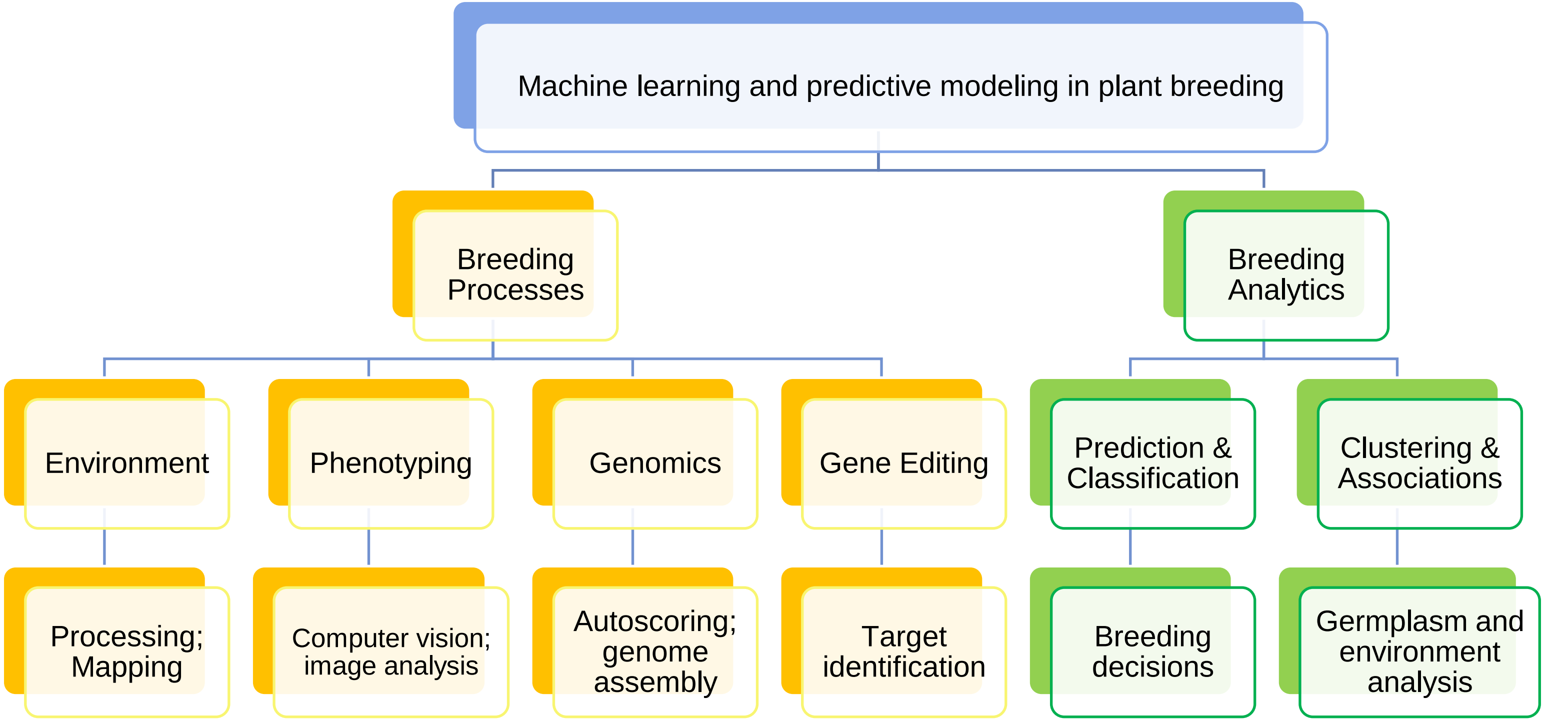


Stephens, Z. D.et al. (2015). Big data: astronomical or genomical? *PLoS biology*, *13*(7), e1002195.

More Env

- UC Merced GridMET
- NWS NOAA
- NASA GISS, NASA power
- Harmonized SoilDB
- USDA SSURGO

More Computing



Machine learning and predictive modeling in plant breeding

Breeding Processes

Environment

Processing; Mapping

Phenotyping

Computer vision; image analysis

Genomics

Autoscoring; genome assembly

Gene Editing

Target identification

Breeding Analytics

Prediction & Classification

Breeding decisions

Clustering & Associations

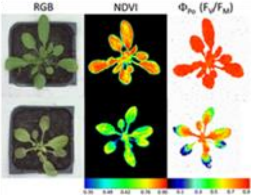
Germplasm and environment analysis

Machine learning in breeding processes

Enhancing databases, automating lab tasks field work

phenotyping

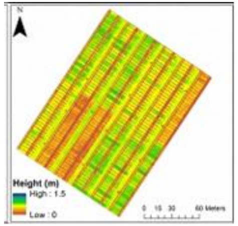
Disease, stress scoring

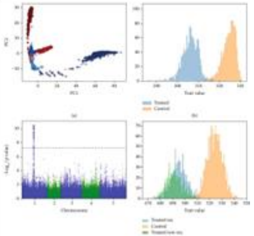


https://www.mdpi.com/2673-2688/2/3/26
https://www.biomedcentral.com/collections/phenomics

Phenotype automation

(e.g., plant height, identify new traits)

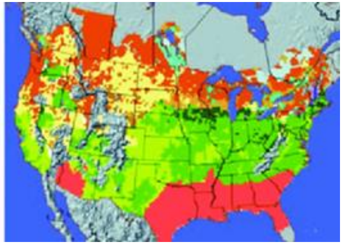




https://www.mdpi.com/2072-4292/8/12/1031
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7706325/

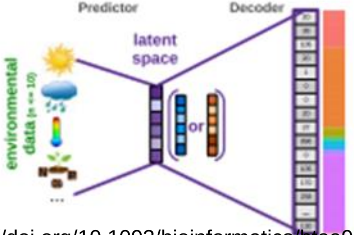
environment

Mapping / zoning



https://www.publish.csiro.au/cp/CP14007

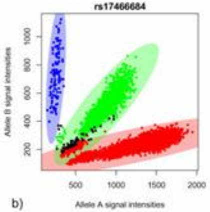
Latent weather, soil

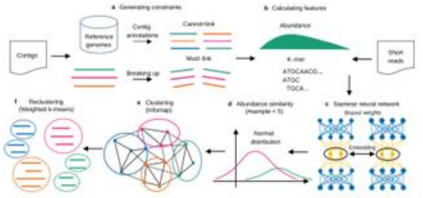


https://doi.org/10.1093/bioinformatics/btaa971

biotech

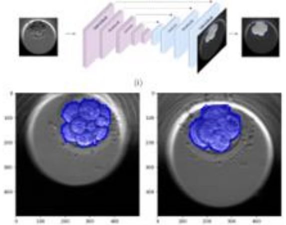
SNP calls, genome assembly





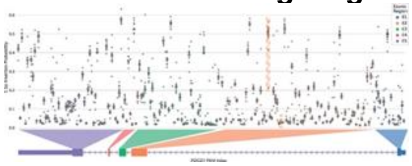
https://doi.org/10.1186/1753-6561-3-s7-s58
https://www.nature.com/articles/s41467-022-29843-y

Embryo rescue DH production



https://www.nature.com/articles/s41598-022-06336-y

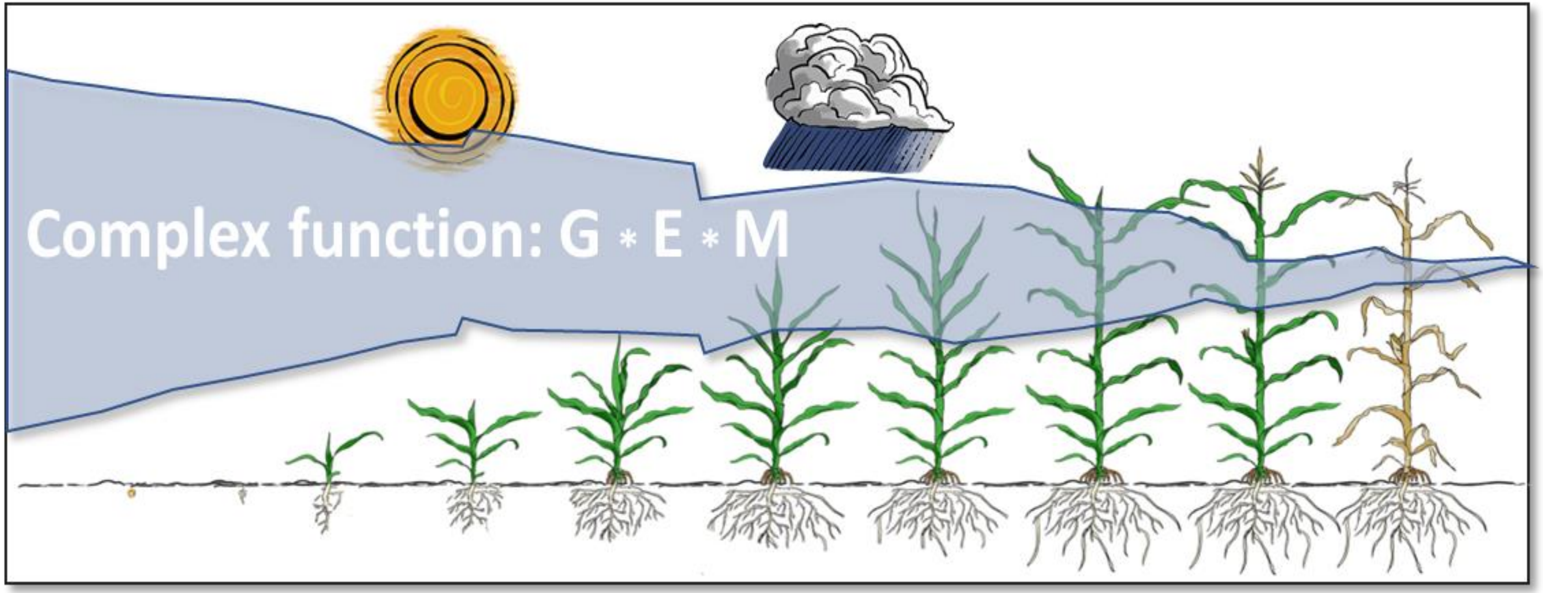
Gene editing targets



https://doi.org/10.1093/bioinformatics/btab268

1. **Introduction**
    - Rationale
    - Breeding with ML
2. **Modeling**
    - Challenges and opportunities
    - Correlated information
3. **Analytics**
    - Breeding objectives
    - Target G x E x M
    - Validation
4. **Conclusion**

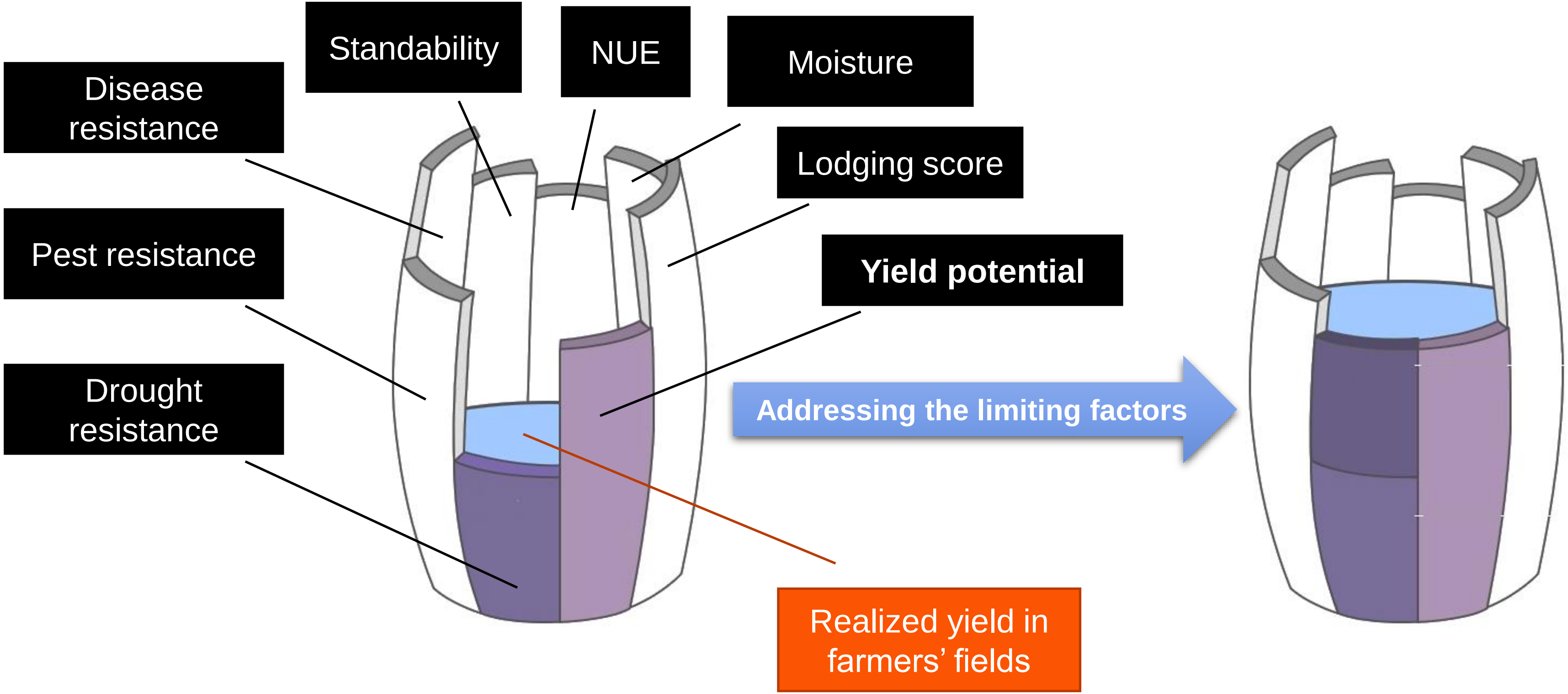
Challenges in Corn Improvement



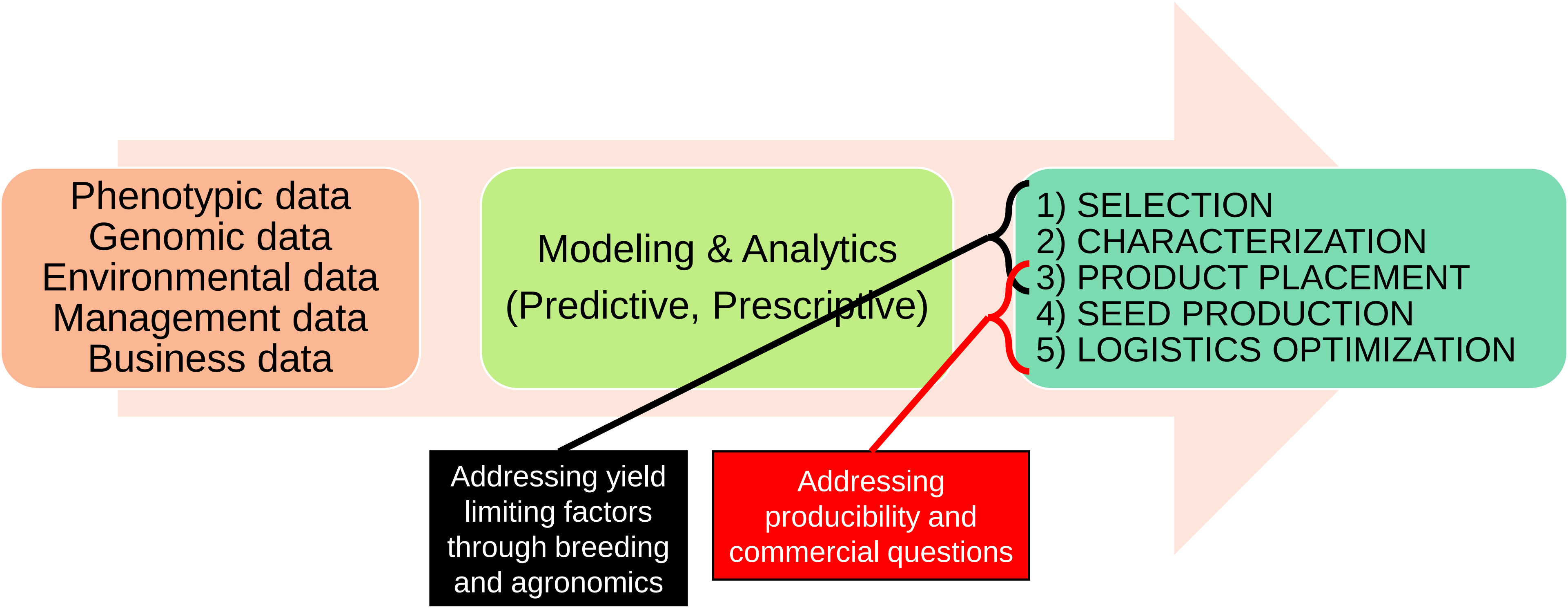
G = Genetics; E = Environment; M = Management

Source: Totir 2021, ASTA

Law of the minimum



End goal of GxExM characterization



Phenotypic data
Genomic data
Environmental data
Management data
Business data

Modeling & Analytics
(Predictive, Prescriptive)

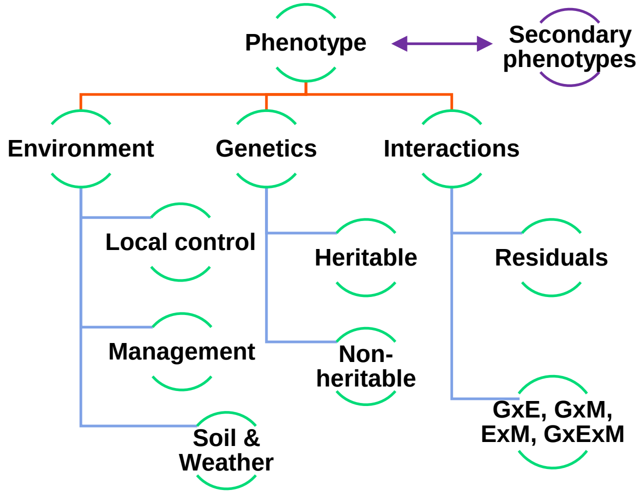
1) SELECTION
2) CHARACTERIZATION
3) PRODUCT PLACEMENT
4) SEED PRODUCTION
5) LOGISTICS OPTIMIZATION

Addressing yield limiting factors through breeding and agronomics

Addressing producibility and commercial questions

Linear models

Source: https://evolution.berkeley.edu/teach-resources/genes-environment-phenotype/


Phenotype ↔ Secondary phenotypes

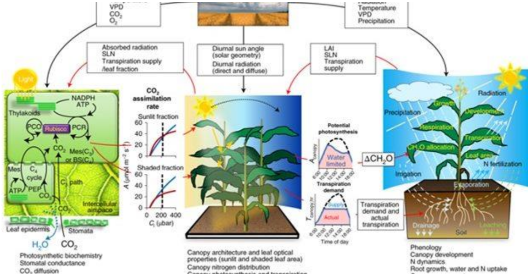
- Environment
  - Local control
  - Management
  - Soil & Weather
- Genetics
  - Heritable
  - Non-heritable
- Interactions
  - Residuals
  - GxE, GxM, ExM, GxExM

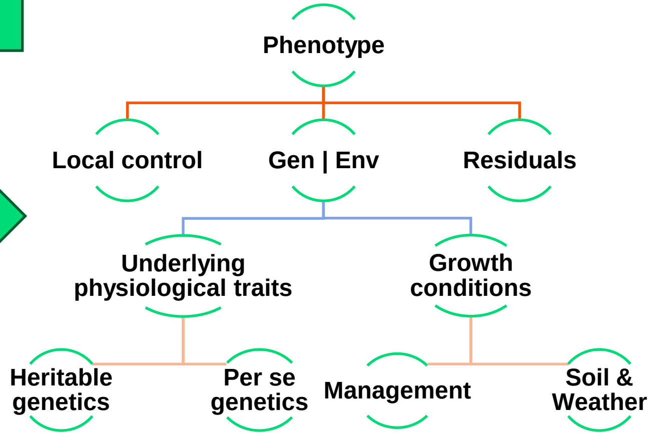
General framework, trait-crop agnostic

Biological knowledge, trait-crop specificity

Crop models

https://www.nature.com/articles/s41477-019-0398-8




Phenotype

- Local control
- Gen | Env
  - Underlying physiological traits
    - Heritable genetics
    - Per se genetics
  - Growth conditions
    - Management
    - Soil & Weather
- Residuals

Walking through the statistical black boxes of plant breeding

Alencar Xavier[1] · William M. Muir[2] · Bruce Craig[3] · Katy Martin Rainey[1]

Integrating Crop Growth Models with Whole Genome Prediction through Approximate Bayesian Computation

Frank Technow[1]*, Carlos D. Messina[2], L. Radu Totir[1], Mark Cooper[2]

Linearly correlated phenotypes

Using machine learning to infer connections from data

$$\text{Phenotype (y)} = \text{Gen|Trials(g)} + \text{Plot level noise(e)}$$

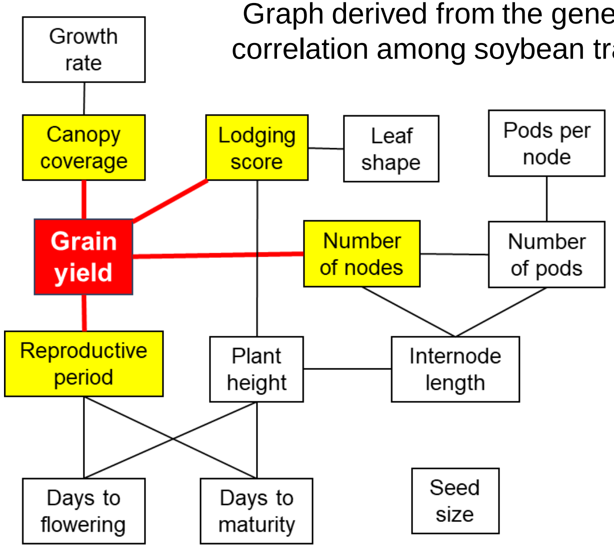
$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

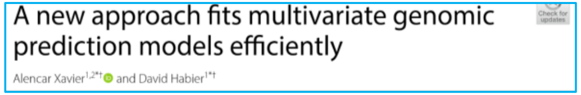
$$\text{Variance} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} \\ \sigma_{g_{12}} & \sigma^2_{g_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

CORRELATED INFORMATION

Phenotypic network

Graph derived from the genetic correlation among soybean traits

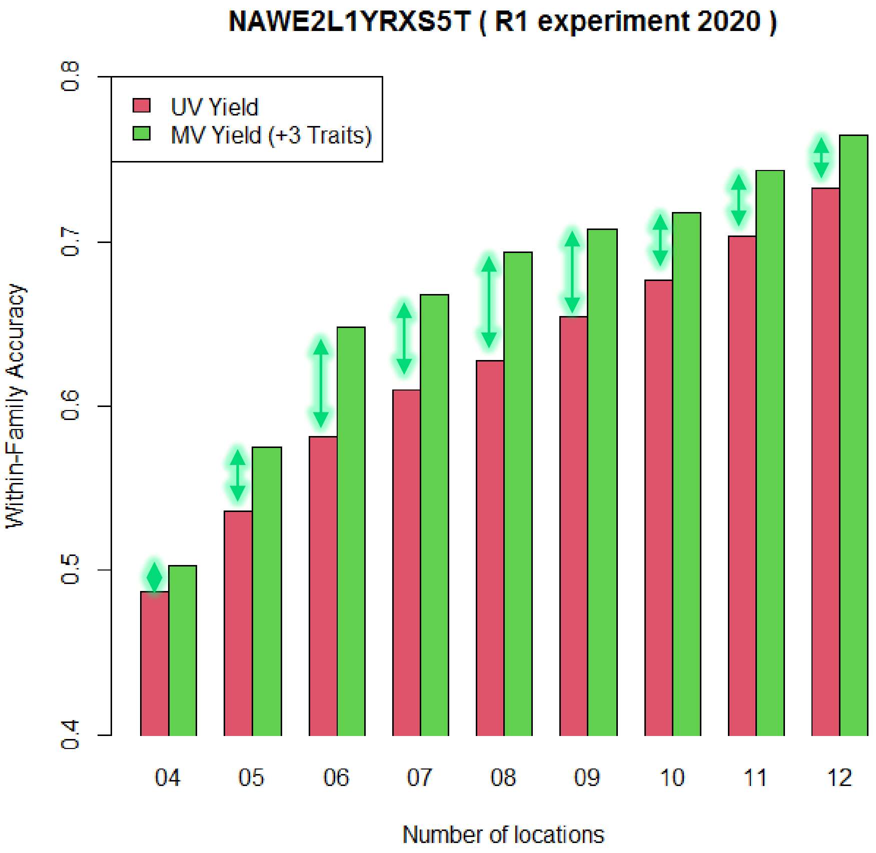



A new approach fits multivariate genomic prediction models efficiently

Alencar Xavier[1,2*†] and David Habier[1*†]

Using unsupervised learning techniques to assess interactions among complex traits in soybeans

Alencar Xavier · Benjamin Hall · Shaun Casteel · William Muir · Katy Martin Rainey



Leveraging information from secondary traits

Multi-trait analysis provided an average increase in accuracy of **0.03**

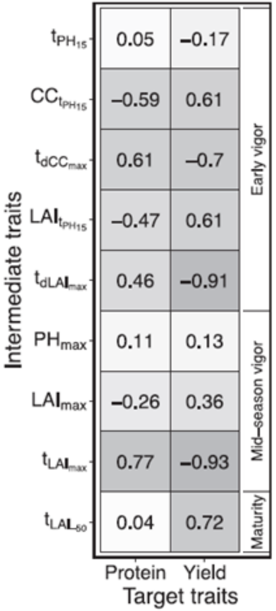
Equivalent to adding **~1.7 locations**

High-throughput phenotyping

More data, more information, more complexity, more opportunities…

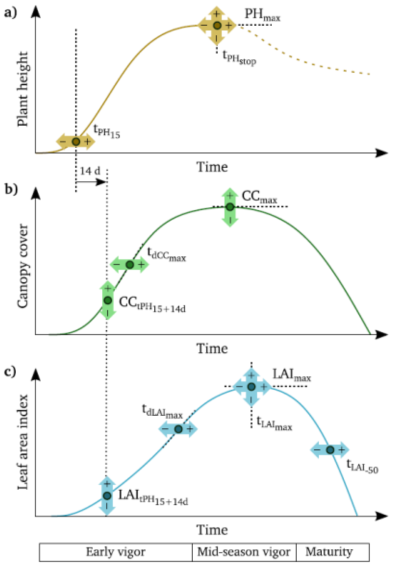
Singh A.K. et al. (2021)
High-Throughput Phenotyping in Soybean.
https://doi-org/10.1007/978-3-030-73734-4_7

Correlation(HTP, main trait)



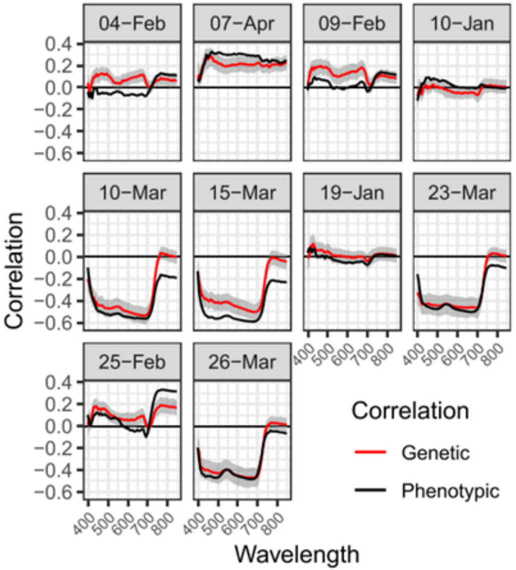
Intermediate traits	Protein	Yield	
$t_{PH_{15}}$	0.05	−0.17	Early vigor
$CC_{t_{PH_{15}}}$	−0.59	0.61	
$t_{dCC_{max}}$	0.61	−0.7	
$LAI_{t_{PH_{15}}}$	−0.47	0.61	
$t_{dLAI_{max}}$	0.46	−0.91	
PH_{max}	0.11	0.13	Mid–season vigor
LAI_{max}	−0.26	0.36	
$t_{LAI_{max}}$	0.77	−0.93	
$t_{LAI_{50}}$	0.04	0.72	Maturity

Target traits



Roth et al. (2022) High-throughput field phenotyping of soybean: Spotting an ideotype. https://doi.org/10.1016/j.rse.2021.112797

Correlation(HTP, main trait)
with *large number of traits*



Runcie et al. (2021) Mega-scale linear mixed models for genomic predictions with thousands of traits. https://doi.org/10.1186/s13059-021-02416-w

1. Introduction
    - Rationale
    - Breeding with ML
2. Modeling
    - Approaches
    - Correlated information
3. **Analytics**
    - Breeding objectives
    - Target G x E x M
    - Validation
4. Conclusion

“Breeding objective”

$$f\,(\text{market segment, farming systems})$$

- Set of traits of interest (**TOI**) bred into a

WHAT

Yield, moisture, relative maturity, disease resistance, stability, **trait package, producibility**

- Target population of genetics (**TPG**) for a given

WHO

Corn 111-121, corn 122-130
white corn 118-123

- Target population of environments (**TPE**) and management (**TPM**) practices

WHERE

HOW, WHEN

Drought, irrigation, early planting, varying levels of disease pressure, different soil types

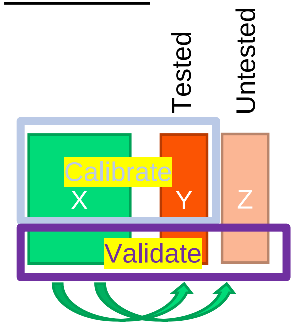
$$\rho_{\text{GxExM}} = \rho_{\text{TPG}} \times \rho_{\text{TPE}} \times \rho_{\text{TPM}}$$

Model testing and validation schemes

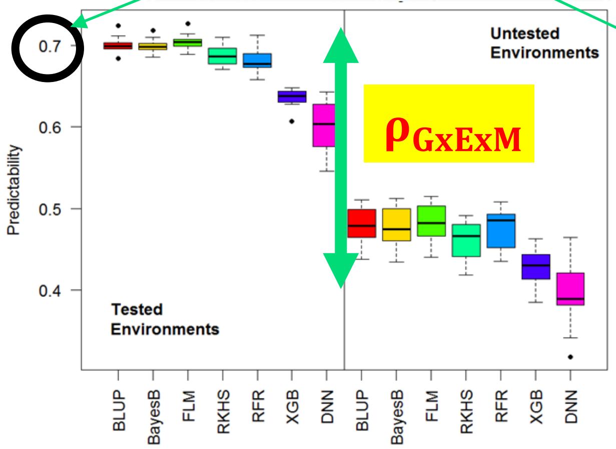
$$\textbf{Prediction accuracy} \propto \sqrt{H^2} \times \rho_{\text{GxExM}}$$

Upper limit is determinate by the application $\cong \sqrt{\text{H}^2}$

CV scheme

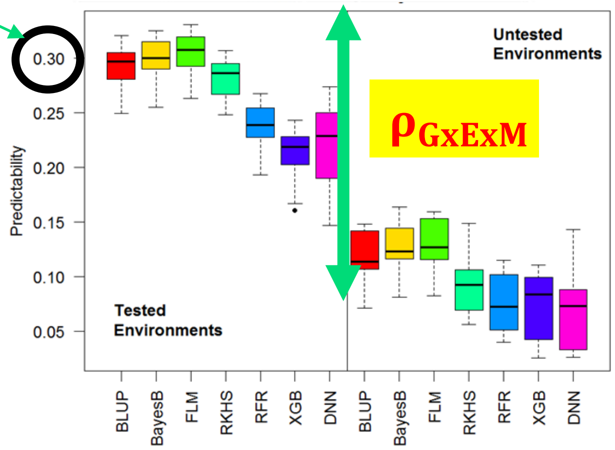


Scenario 1



Different prediction methods

Scenario 2

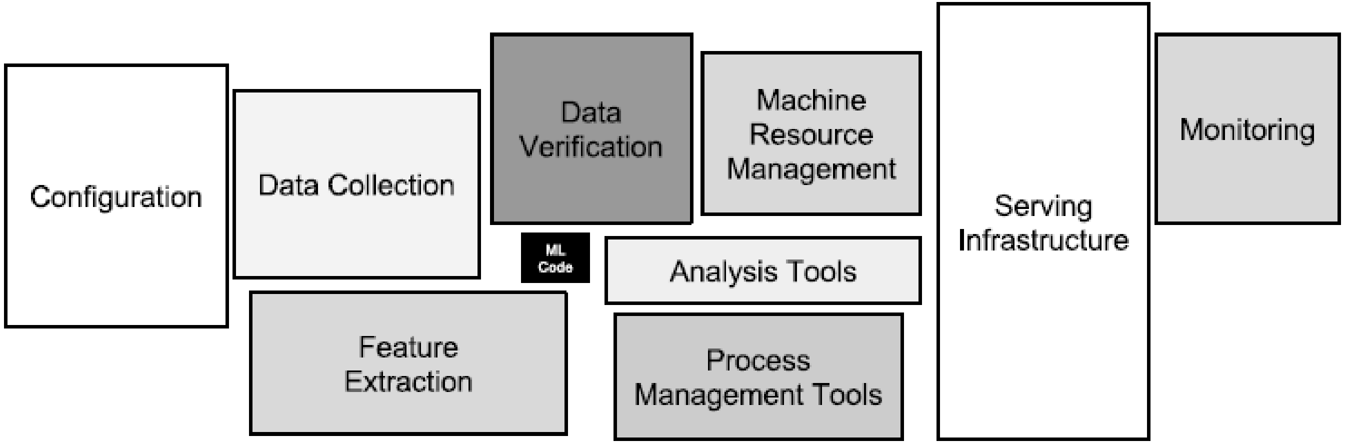


Different prediction methods

Technical nuances of machine learning: implementation and validation of supervised methods for genomic prediction in plant breeding

Alencar Xavier [1*]

Multiple factors play a role on the implementation of predictive breeding and automated systems beyond proof of concepts using cross-validations



Hidden Technical Debt in Machine Learning Systems

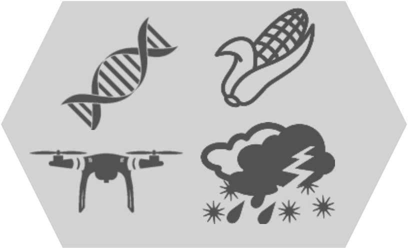
doi/10.5555/2969442.2969519

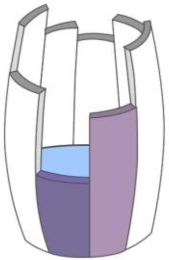
1. **Introduction**
    - Rationale
    - Breeding with ML
2. **Modeling**
    - Approaches
    - Correlated information
3. **Analytics**
    - Breeding objectives
    - Target G x E x M
    - Validation
4. **Conclusion**

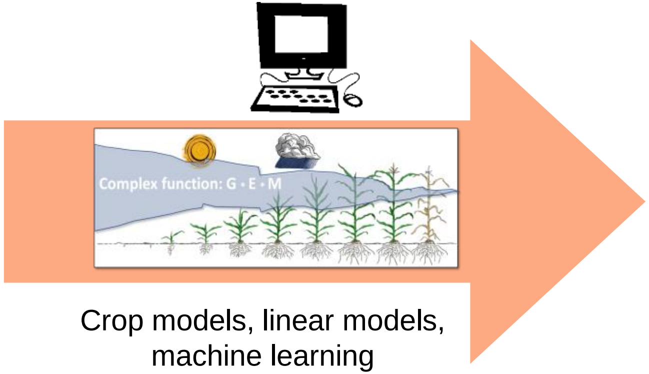
With more data and better analytics, breeding can respond faster the new farming challenges and trends

- New & better management
- Changing environment
- New pests and diseases

Predictive breeding (TPE, TPG, TPM)

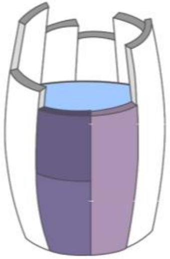






Crop models, linear models, machine learning

$$\mathrm{Acc} \propto \mathbf{GxExM} \times \mathbf{h}^2$$



Performance

Producibility

Robustness

Thank you for your attention!

Final remarks:

1) Plant breeding relies on analytics for multiple processes and analytics
2) Harnessing GxExM information benefits accuracy, business impact
3) Modeling is contingent to the target genetics, environments and management

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com