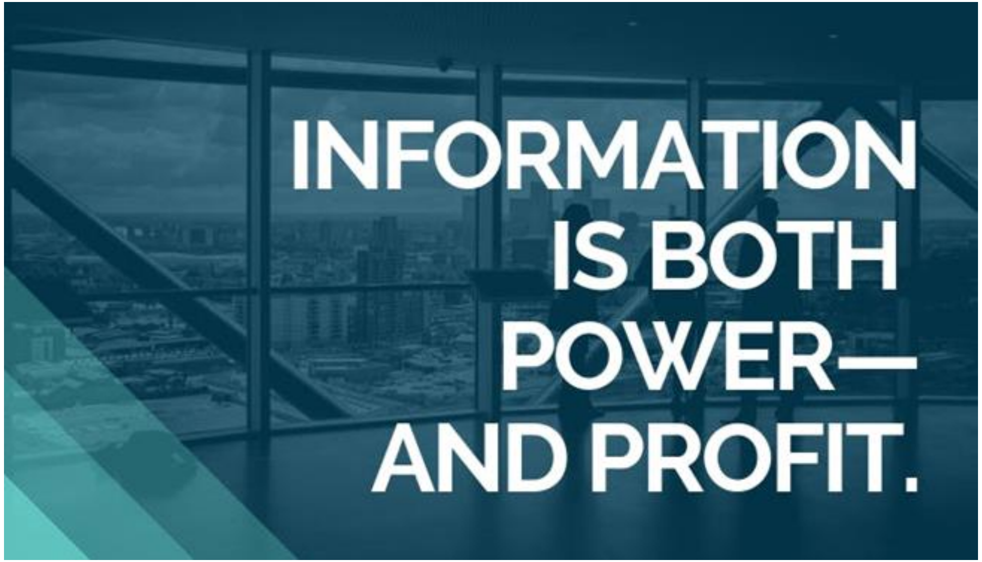
CORTEVA™
agriscience

Machine learning-based AI applied to breeding

Alencar Xavier
Breeding Analyst at Corteva
Adjunct professor at Purdue

Adequate use of

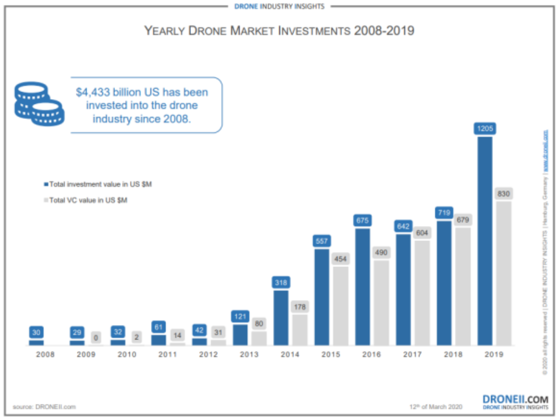


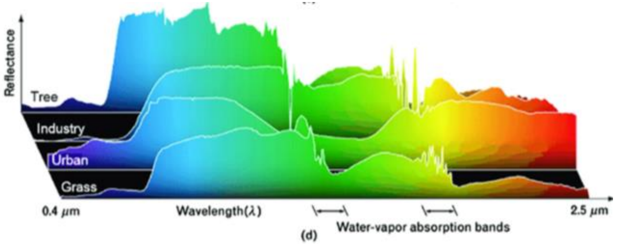
Outline

1. **Introduction**
   - More data
   - Branching ML
2. **Machines**
   - Filters
   - Engines
3. **Analytics**
   - Target G x E x M
   - Validation
   - Cases of study
4. **Conclusion**

1. **Introduction**
   - More data
   - Branching ML
2. **Machines**
   - Filters
   - Engines
3. **Analytics**
   - Target G x E x M
   - Validation
   - Cases of study
4. **Conclusion**

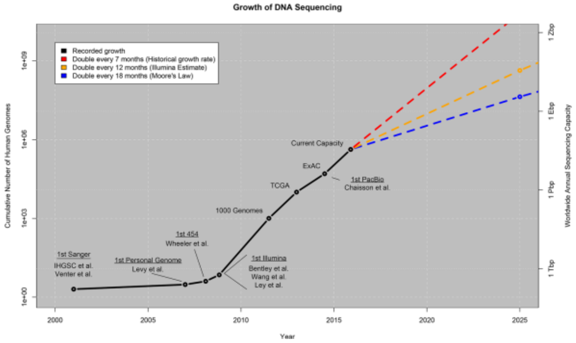
More Pheno



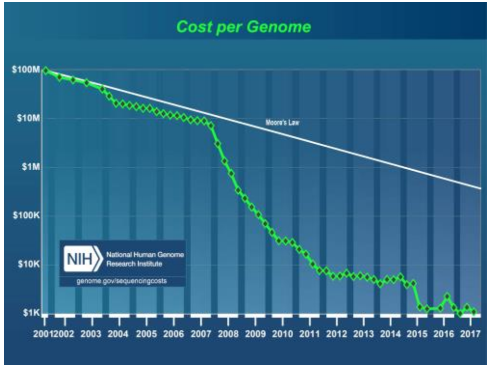


https://www.mdpi.com/2076-3417/12/5/2570

More Geno



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

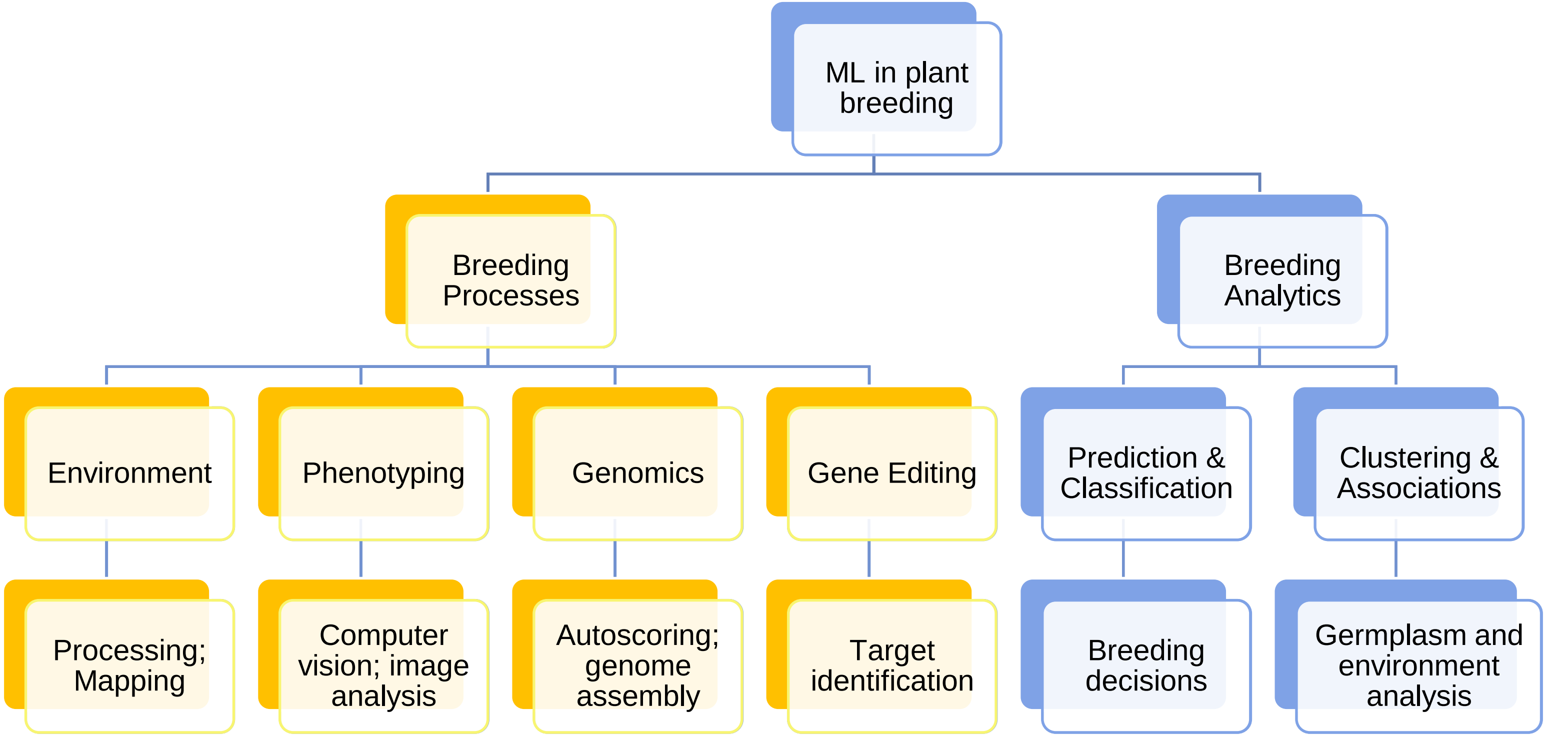


Stephens, Z. D.et al. (2015). Big data: astronomical or genomical? *PLoS biology*, *13*(7), e1002195.

More Env

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

More Computing


ML in plant breeding

Breeding Processes

Breeding Analytics

Environment

Phenotyping

Genomics

Gene Editing

Prediction & Classification

Clustering & Associations

Processing; Mapping

Computer vision; image analysis

Autoscoring; genome assembly

Target identification

Breeding decisions

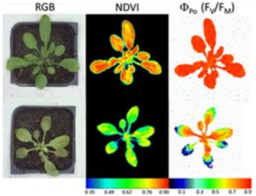
Germplasm and environment analysis

ML in breeding processes

Enhancing databases, automating lab tasks field work

phenotyping

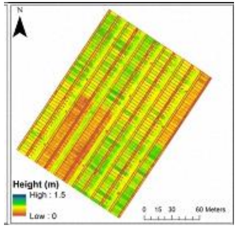
Disease, stress scoring

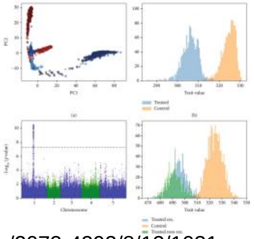


https://www.mdpi.com/2673-2688/2/3/26
https://www.biomedcentral.com/collections/phenomics

Phenotype automation

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

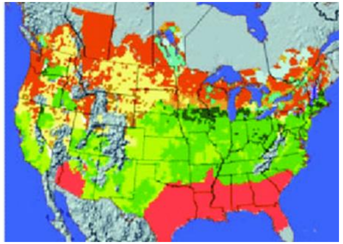




https://www.mdpi.com/2072-4292/8/12/1031
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7706325/

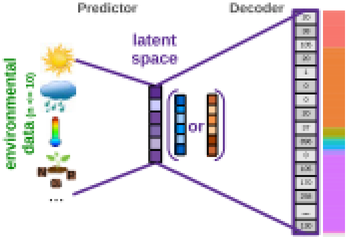
environment

Mapping / zoning



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

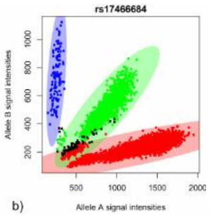
Latent weather, soil

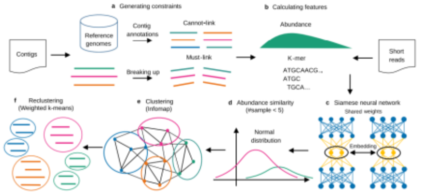


https://doi.org/10.1093/bioinformatics/btaa971

biotech

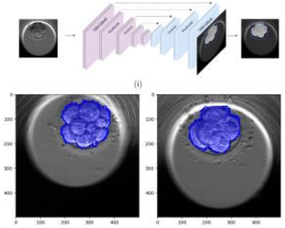
SNP calls, genome assembly





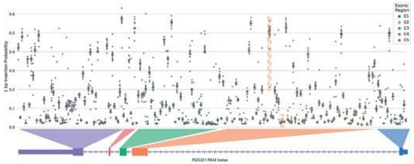
https://doi.org/10.1186/1753-6561-3-s7-s58
https://www.nature.com/articles/s41467-022-29843-y

Embryo rescue DH production

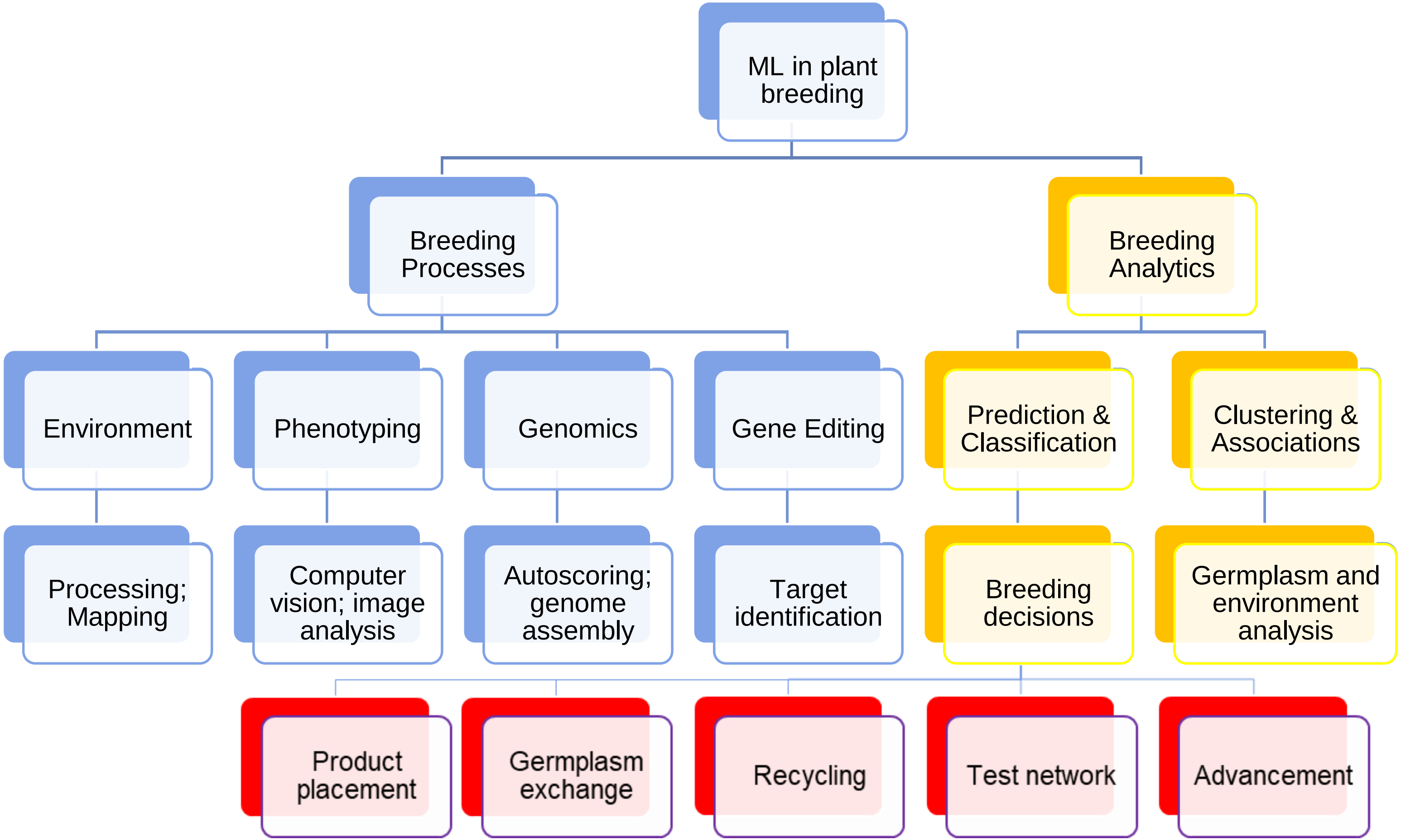


https://www.nature.com/articles/s41598-022-06336-y

Gene editing targets



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



ML in plant breeding

Breeding Processes

Breeding Analytics

Environment

Phenotyping

Genomics

Gene Editing

Prediction & Classification

Clustering & Associations

Processing; Mapping

Computer vision; image analysis

Autoscoring; genome assembly

Target identification

Breeding decisions

Germplasm and environment analysis

Product placement

Germplasm exchange

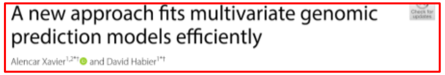
Recycling

Test network

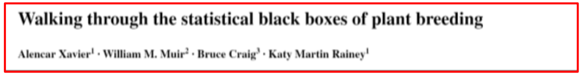
Advancement

1. **Introduction**
    - More data
    - Branching ML
2. **Machines**
    - Filters
    - Engines
3. **Analytics**
    - Target G x E x M
    - Validation
    - Cases of study
4. **Conclusion**

Machine Learning Engines


A new approach fits multivariate genomic prediction models efficiently

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


Walking through the statistical black boxes of plant breeding

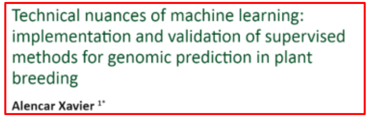
Alencar Xavier[1] · William M. Muir[2] · Bruce Craig[3] · Katy Martin Rainey[1]

Efficient Estimation of Marker Effects in Plant Breeding

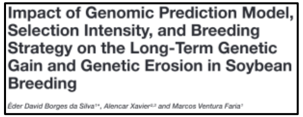
Alencar Xavier[*,†,1]

[*]Corteva Agrisciences, 8305 NW 62nd Ave. Johnston IA, and [†]Purdue University, 915 W State St. West Lafayette IN

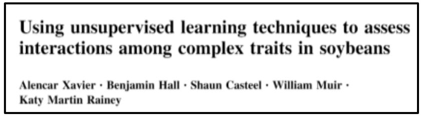
ORCID ID: 0000-0001-5034-9954 (A.X.)


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

Alencar Xavier [1*]


Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding

Éder David Borges da Silva[1*], Alencar Xavier[2,3] and Marcos Ventura Faria[1]

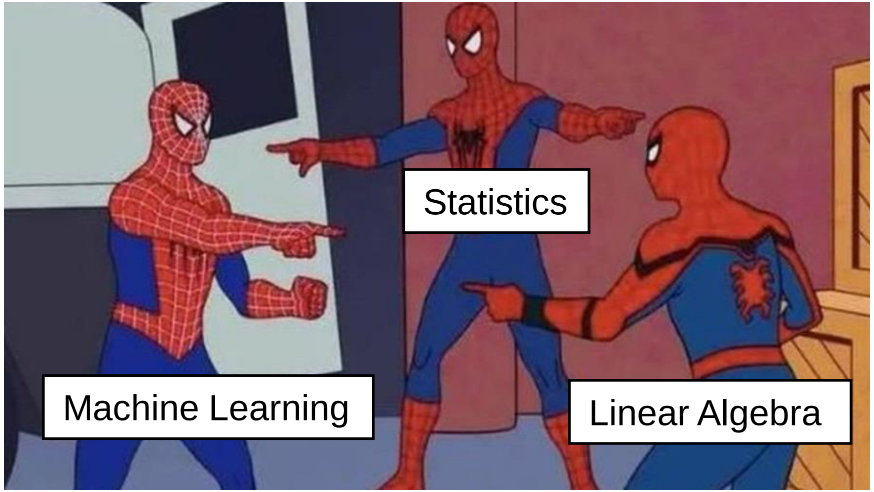

Using unsupervised learning techniques to assess interactions among complex traits in soybeans

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

Article

Joint Modeling of Genetics and Field Variation in Plant Breeding Trials Using Relationship and Different Spatial Methods: A Simulation Study of Accuracy and Bias

Éder David Borges da Silva [1,2,*], Alencar Xavier [3,4] and Marcos Ventura Faria [2]

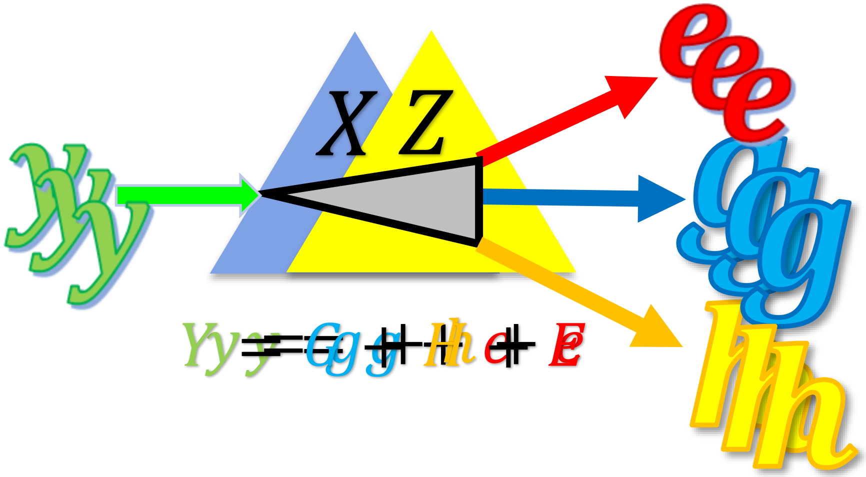

Statistics
Machine Learning
Linear Algebra

Key idea of supervised learning: FILTERING

Multi-trait filter

Multiple filters

Single filter

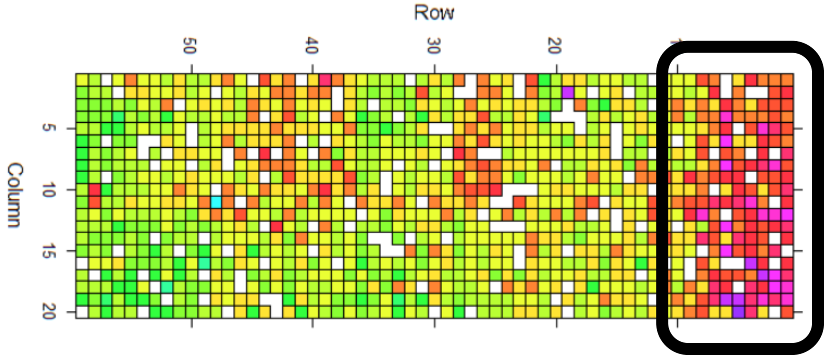


$Y = G + H + E$

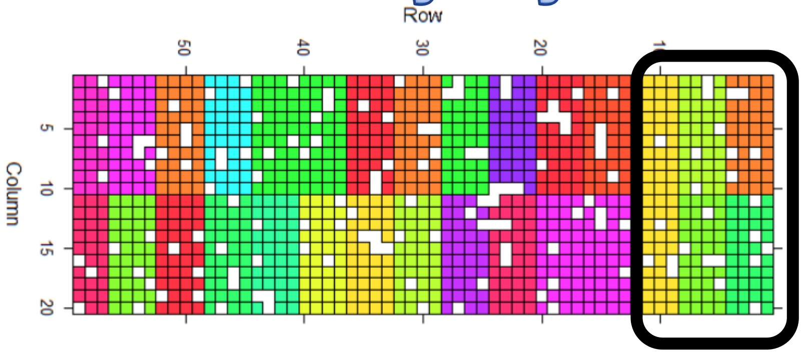
$y = g + h + e$

Why bother with multiple filters?

Field variation

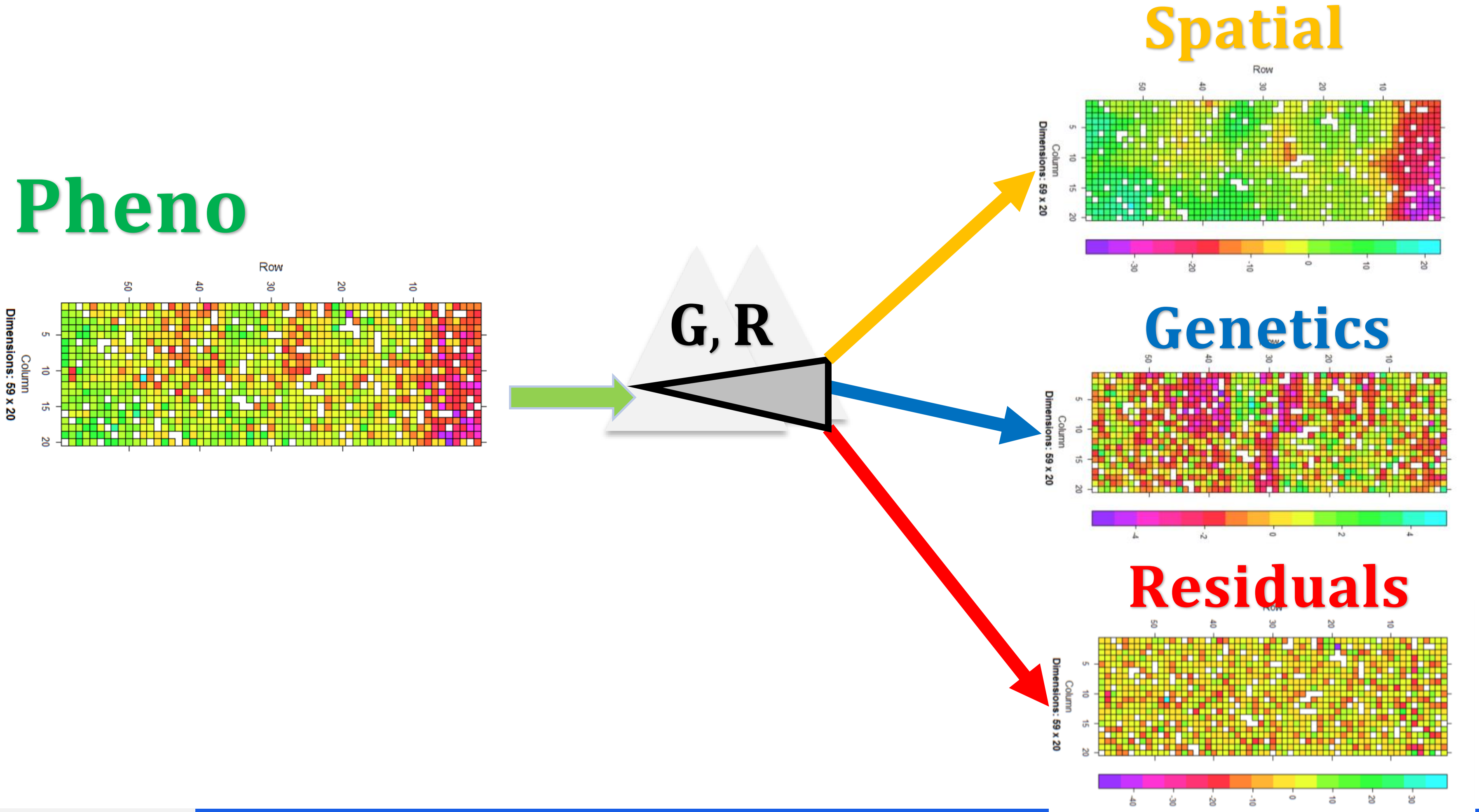


Family layout



Some families were placed on unfavorable side of the field…

SoyNAM field,
Indiana 2014


Pheno
Row
50
40
30
20
10
Column
5
10
15
20
Dimensions: 59 x 20
G, R
Spatial
-30
-20
-10
0
10
20
Genetics
-4
-2
0
2
4
Residuals
-40
-30
-20
-10
0
10
20
30

Why bother with multi-response filters?

Simple (bivariate) model:

INFORMATION GAIN

$$\mathbf{y} = \mathbf{g} + \mathbf{e}$$

$$Var\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

Why bother with multi-response filters?

$$y = Zg + e, \qquad y \sim N(0, V)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

- Covariance structure

$$V = G \otimes \Sigma_a + I \otimes \Sigma_e = G \otimes \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix} + I \otimes \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

- Model equation

$$\begin{bmatrix} Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11} & Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12} \\ Z_2'\Sigma_e^{12}Z_1 + G^{-1}\Sigma_a^{12} & Z_2'\Sigma_e^{11}Z_2 + G^{-1}\Sigma_a^{22} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} Z_1'(\Sigma_e^{11}y_1 + \Sigma_e^{12}y_2) \\ Z_2'(\Sigma_e^{22}y_2 + \Sigma_e^{12}y_1) \end{bmatrix}$$

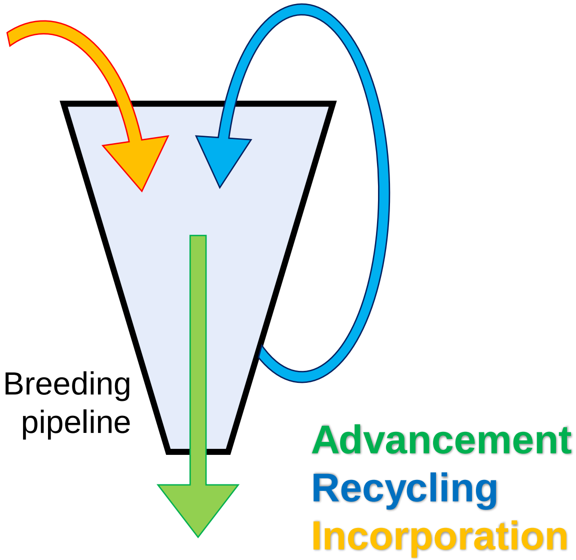
- **<u>Univariate vs bivariate</u>**

$$\begin{aligned} g_1 \quad &= (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'\Sigma_e^{11}y_1) \\ g_1|g_2 &= (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'(\Sigma_e^{11}y_1 + \overbrace{\Sigma_e^{12}y_2) - (Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12})g_2}^{\text{INFORMATION GAIN}}) \end{aligned}$$

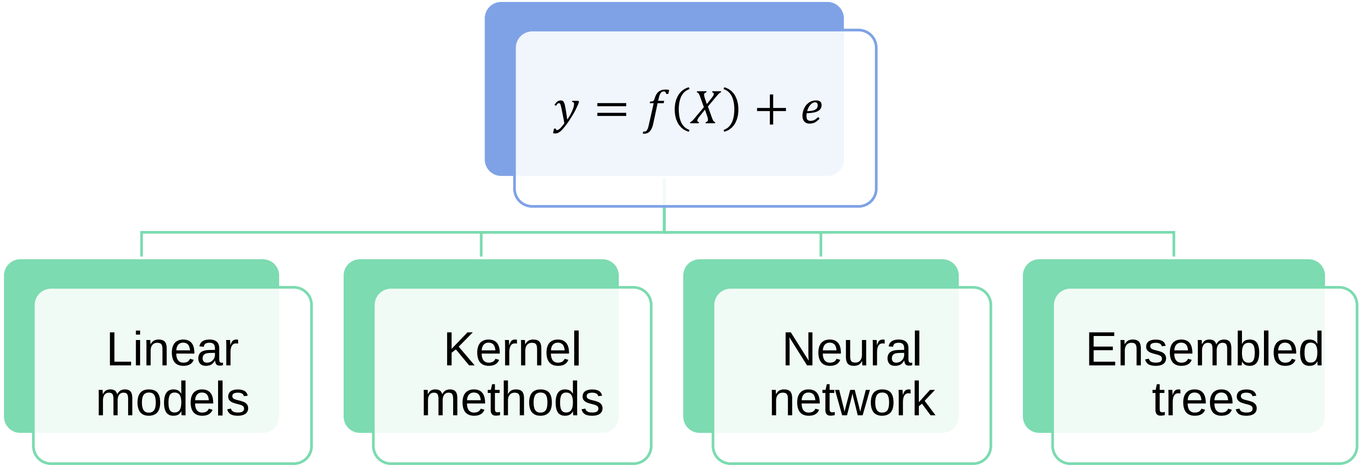
Does the choice of filter matter?

- **ADDITIVE LINEAR FILTERS** **(GEBV)**
  - *Pattern*: ADDITIVE GENETICS - ***heritable***
  - *Method*: GBLUP, RIDGE, LASSO
  - *Suits*: **RECYCLING, ADVANCEMENT**

- **NON-LINEAR FILTERS** **(EGV)**
  - *Pattern*: ANY GENETIC SIGNAL
  - *Method*: RKHS, DNN, Random Forest
  - *Suits*: **ADVANCEMENT, PRODUCT PLACEMENT**



Main classes of learners



$$y = f(X) + e$$

Linear models

Kernel methods

Neural network

Ensembled trees

Solving: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$

$$Finding \rightarrow argmin(e'e + \lambda b'b)$$

- Coordinate descent

  (Use diagonals of LHS)

$$\hat{b}_j^{t+1} = \frac{x_j'(y - X_{-j}\hat{b}_{-j})}{x_j'x_j + \lambda}$$

Used for p>>n solvers

glmnet, BGLR, bWGR, GS3

- Gradient descent

  (Does not build LHS)

$$\hat{b}^{t+1} = b^t - \frac{2r}{n}\left[X'(y - X\hat{b}^t) + \lambda\hat{b}^t\right]$$

Used for Deep Neural Nets

TensorFlow Keras, PyTorch, MXNet

- Second order

  (Builds entire LHS)

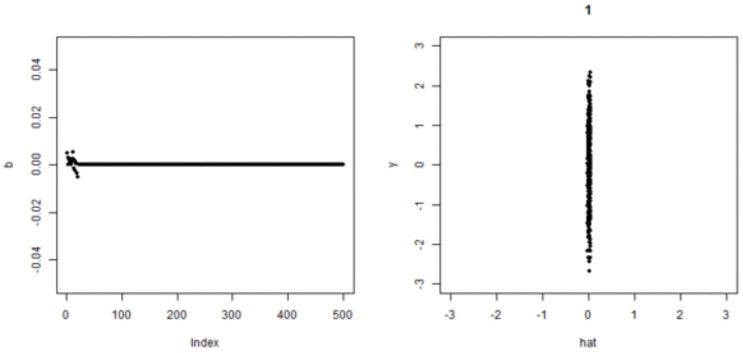
$$\hat{b} = (X'X + \lambda)^{-1}(X'y)$$

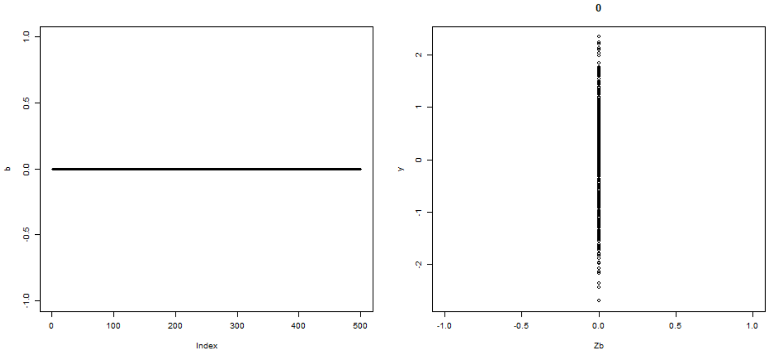
Used for everything else

ASREML, lme4, SAS

Coordinate descent

Gradient descent





$$\hat{b}_j^{t+1} = \frac{x_j'(y - X_{-j}\hat{b}_{-j})}{x_j'x_j + \lambda}$$

$$\hat{b}^{t+1} = b^t - \frac{2r}{n}\left[X'(y - X\hat{b}^t) + \lambda\hat{b}^t\right]$$

What about the deep learning? 🤔

$$y = \alpha(\alpha(XB_1)B_2)b_3 + e$$

i.e., just a “stack of solvers”

Data > Method

Unnecessarily complex analysis should not be used as a foil to disguise lower quality datasets

Kruuk (2004 *apud* Walsh and Lynch 2018)

1. **Introduction**
    - More data
    - Branching ML
2. **Machines**
    - Filters
    - Engines
3. **Analytics**
    - Target G x E x M
    - Validation
    - Cases of study
4. **Conclusion**

Analytics

“Breeding objective”

- Set of traits of interest (**TOI**)

bred into a

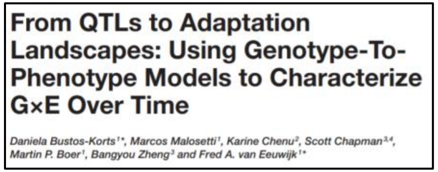
- Target population of genotypes (**TPG**)

for a given

- Target population of environments (**TPE**)

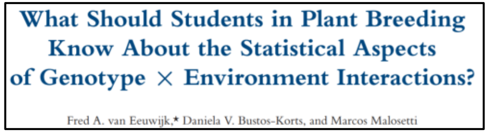
TPE, TPG, TPM

- **Target population of environments (TPE)**
  - Influences accuracies via GxE correlation
  - Which environments should I be able to predict?
- **Target population of genotypes (TPG)**
  - Influences accuracies via genetic relationship
  - Which genetics should I be able to predict?
- **Target population of management (TPM)**
  - Herein nested in TPE



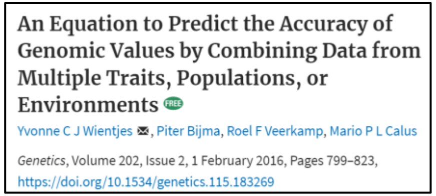
From QTLs to Adaptation Landscapes: Using Genotype-To-Phenotype Models to Characterize G×E Over Time

Daniela Bustos-Korts[1], Marcos Malosetti[1], Karine Chenu[2], Scott Chapman[3,4], Martin P. Boer[1], Bangyou Zheng[3] and Fred A. van Eeuwijk[1]**



What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype × Environment Interactions?

Fred A. van Eeuwijk,* Daniela V. Bustos-Korts, and Marcos Malosetti



An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments FREE

Yvonne C J Wientjes ✉, Piter Bijma, Roel F Veerkamp, Mario P L Calus

Genetics, Volume 202, Issue 2, 1 February 2016, Pages 799–823, https://doi.org/10.1534/genetics.115.183269

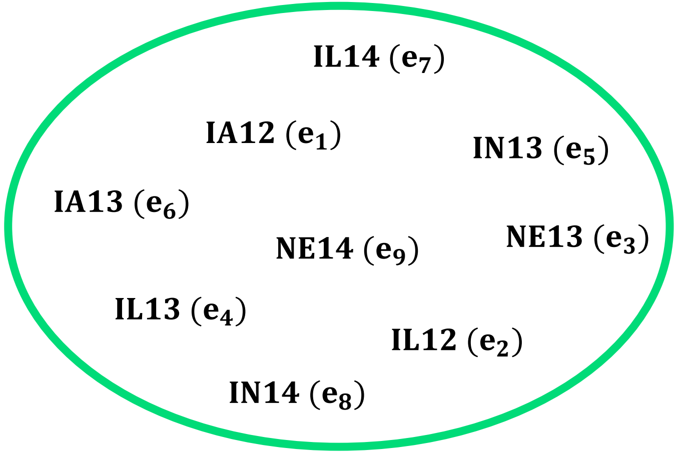
TPE

- Any given trial happens in each environment-management combination, that is sample of much larger population:

$$e_i \in E$$

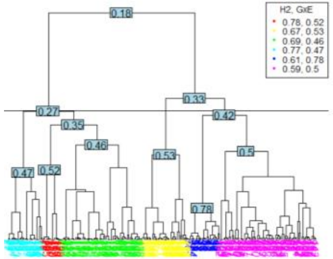
That is:

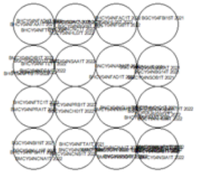
$$\mathbf{TPE\ (E) =}$$

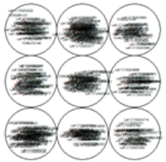


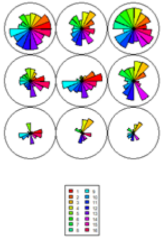
$$\begin{bmatrix} y_{e_i} \\ y_{e_j} \\ g_E \end{bmatrix} = \begin{bmatrix} \sigma^2_{g(e_i)} + \sigma^2_{\epsilon(e_i)} & \sigma_{g(e_i,e_j)} & \sigma_{g(e_i,E)} \\ \sigma_{g(e_j,e_i)} & \sigma^2_{g(e_j)} + \sigma^2_{\epsilon(e_j)} & \sigma_{g(e_j,E)} \\ \sigma_{g(E,e_i)} & \sigma_{g(E,e_j)} & \sigma^2_{g(E)} \end{bmatrix}$$

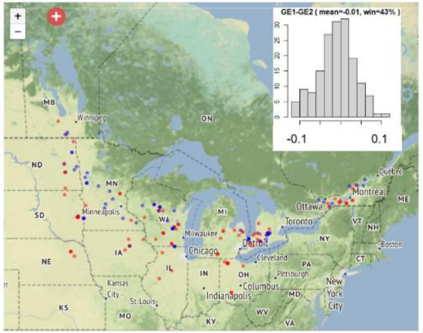
NOTE: GxExM patterns within TPE are largely assessed using different methods of ML

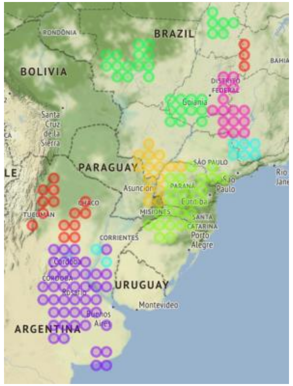


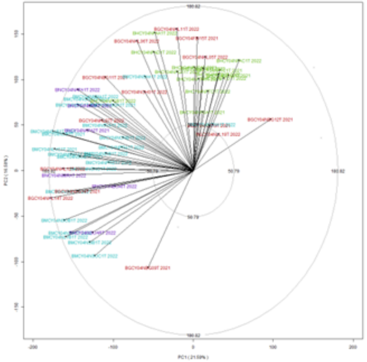


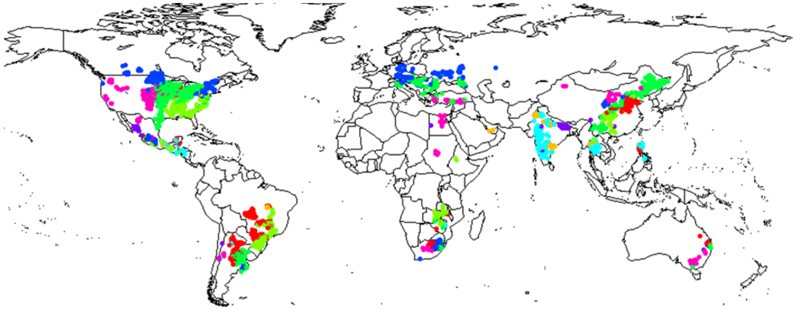












TPG + TPG

- Accuracy (Wientjes et al 2016) = correlation(true signal, estimated signal),
- It is a function of heritability, GxE, representativeness of the calibration set
- For:

$$y = g + e,$$
$$\mathrm{var}(y) = V, \qquad \mathrm{var}(g) = G$$

Then accuracy is

$$a_i = \mathrm{cor}(g_i, \hat{g}_i) = \frac{\mathrm{cov}(g_i, \hat{g}_i)}{\mathrm{var}(g_i)\mathrm{var}(\hat{g}_i)} = \frac{\mathrm{var}(\hat{g}_i)\ r^2_{GxE}}{\mathrm{var}(g_i)\mathrm{var}(\hat{g}_i)} = r^2_{GxE}\sqrt{\frac{G_{i,y}V^{-1}G_{y,i}}{G_{i,i}}}$$

Thus, we **know** how much signal to expect in any given prediction

Validation schemes

1) CV type – Test intent

- **Random CV** = Upper-bound predictive potential
- **Leave-one-out** = Assess structured scenarios (e.g., geography-out, year-out)
- **Holdout** = Reproduce true applications (e.g., predict individuals from upcoming)

2) TPE/TPG relation

	Genotype	Environment	Difficulty
CV00	New	New	*****
CV0	Observed	New	***
CV1	New	Observed	***
CV2	Observed	Observed	*

Adapted from Crossa et al. (2017) doi.org/10.1016/j.tplants.2017.08.011

3) Signal availability

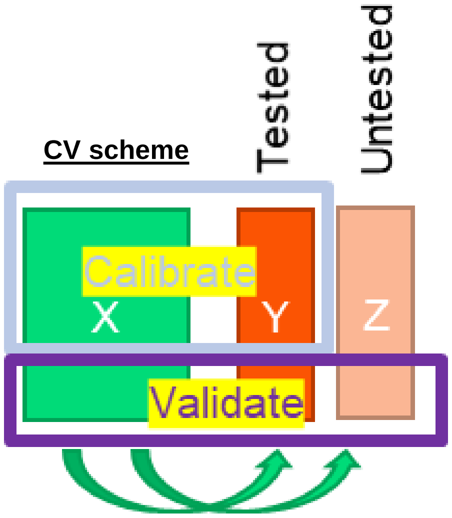
Genetic information available in different cross-validation setups

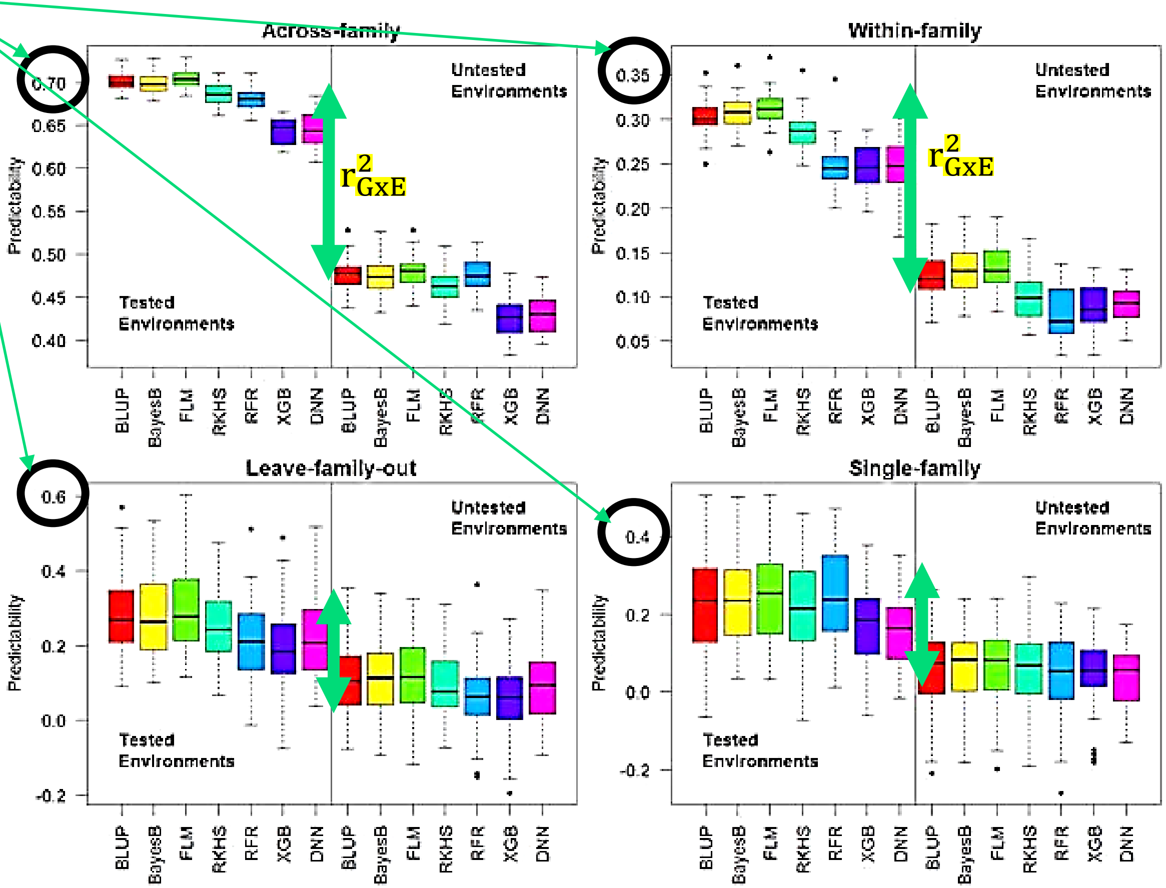
- ***Intra-family***: Linkage*
- ***Within-family***: Linkage and LD
- ***Across-family***: Relationships**, Linkage and LD
- ***Leave-family-out***: Relationships and LD
- ***Untested environments***: Same as above x (GxE)

Validation metrics

- **Correlations**
  - Most common metrics in breeding (e.g., predictability)
  - Pertinent to ranking and selection of complex traits
- **Prediction error**
  - Utilized when the predicted values must be as close as possible to original scale
  - Pertinent to risk prediction (e.g., disease risk)
- **Success**
  - Accommodate complex or subjective criteria, independent or otherwise
  - Pertinent to decision involving data from multiple sources (e.g., advancement)

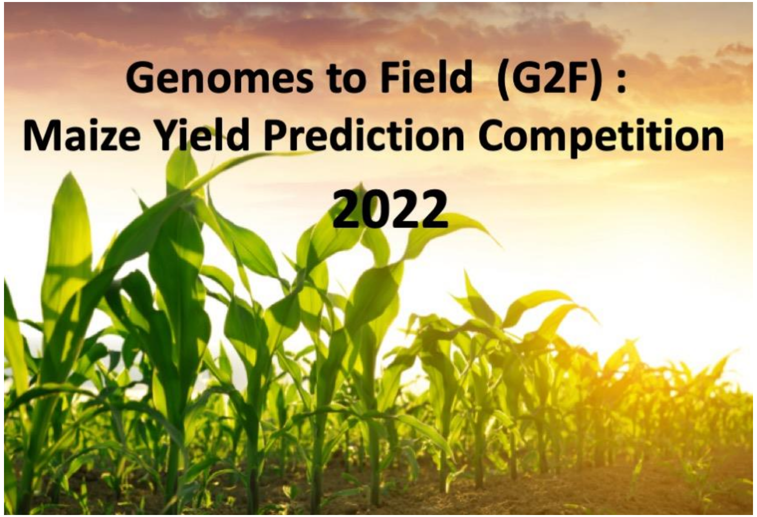
Amount of signal that can be captured in different structures





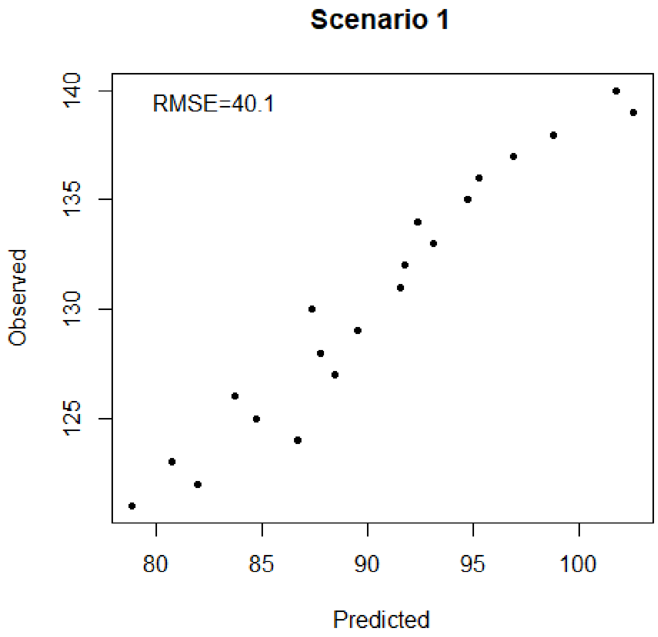
SoyNAM data
ES: 2012 (7 loc)
PS: 2013 (4 loc)
#Fam = 40
Genos = 5600
SNPs = 4300
Obs: 3k-5k obs/loc

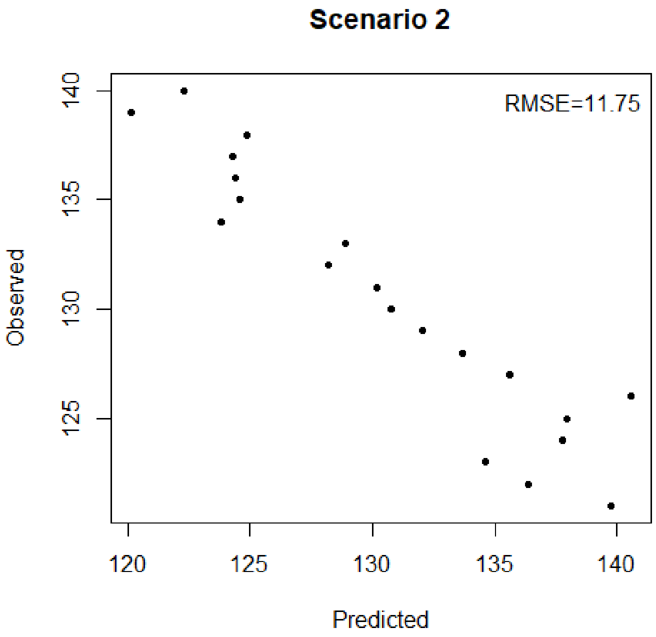
Case of study



Genomes to Field (G2F) :
Maize Yield Prediction Competition
2022

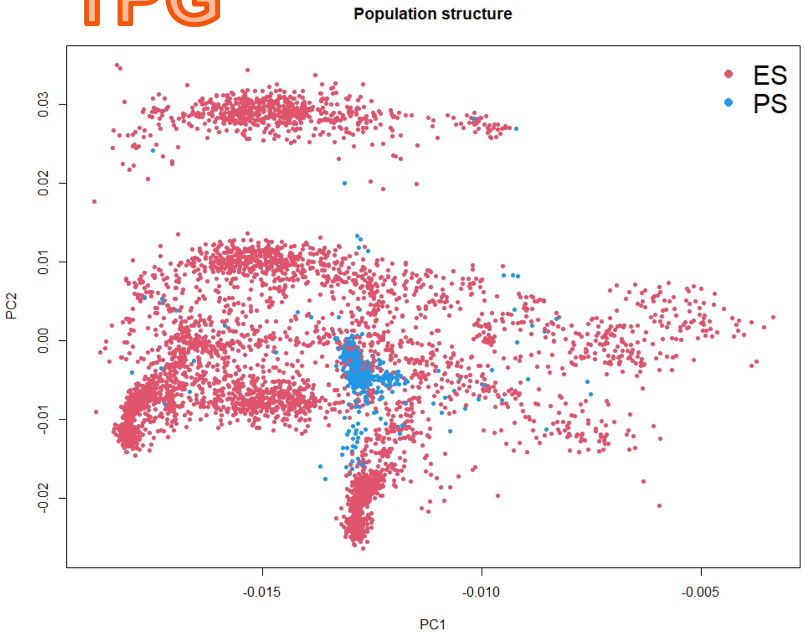
Evaluation criterion



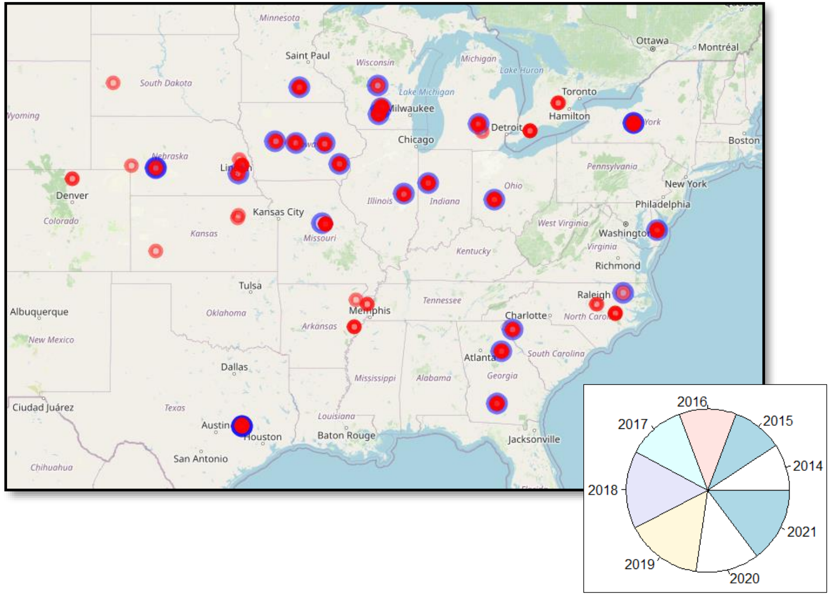


2022 G2F GxE prediction competition

TPG



TPE



Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

What was modeled?

$$y|E_i = \mu_i + g|E_i$$ (Two FILTERS)

$$\text{Phenotype @ i}^{\text{th}}\text{ Loc} = \text{i}^{\text{th}}\text{ Loc Mean} + \text{Genetic effect @ i}^{\text{th}}\text{ Loc}$$

- The winning approach:
  - Predict location means using mixed model and random forest
  - Predict genetic performance with index from multi-response based on **TPE/TPG**

2022 G2F GxE prediction competition

Realized results

Team Name	Within RMSE
CLAC	2.329
igorkf	2.345
phenomaize	2.374
UCD_MegaLMM	2.387
CGM	2.391
breedingteam	2.398
Purdue	2.402
SmAL	2.425
ML_APT	2.472
MPB_Group	2.544

Ranking with alternative metrics

Team Name	Cor Within Loc
CLAC	0.357
CGM	0.353
MPB_Group	0.342
UCD_MegaLMM	0.338
SmAL	0.285
DeepCropVision	0.281
CropEnthusiast	0.279
AllModelsAreWrong	0.272
DataJanitors	0.256
supermanwasd	0.243

Team Name	Cor Across Loc
breedingteam	0.650
DataJanitors	0.644
CLAC	0.631
Purdue	0.631
UCD_MegaLMM	0.628
phenomaize	0.617
igorkf	0.600
CGM	0.587
SmAL	0.586
AllModelsAreWrong	0.575

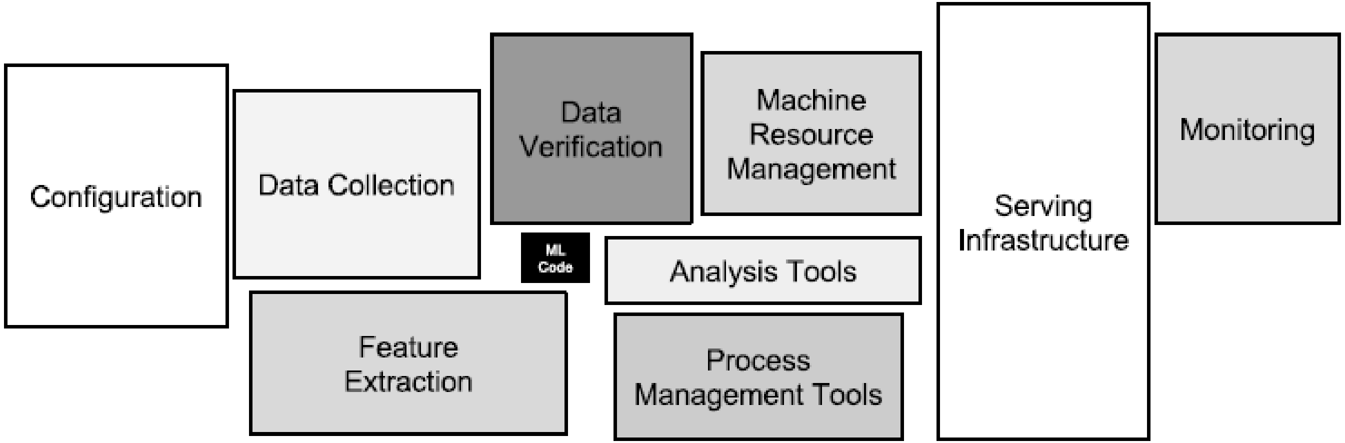
Source: Jacob Washburn, Jose Ignacio Varela, Alencar Xavier

1. **Introduction**
    - More data
    - Branching ML
2. **Machines**
    - Filters
    - Engines
3. **Analytics**
    - Target G x E x M
    - Validation
    - Cases of study
4. **Conclusion**

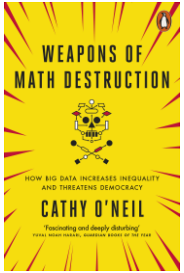
There is more to ML than proof of concepts using cross-validations

Hidden Technical Debt in Machine Learning Systems

doi/10.5555/2969442.2969519



- How easily can an entirely new algorithmic approach be tested at full scale?
- What is the transitive closure of all data dependencies?
- How precisely can the impact of a new change to the system be measured?
- Does improving one model or signal degrade others?
- How quickly can new members of the team be brought up to speed?


WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
'Fascinating and deeply disturbing'

Thank you for your attention!

Final remarks:

1) Plant breeding uses machine learning for multiple purposes in processes and analytics
2) Filter settings are important to maximize signal, but it is less important than data
3) Validation metrics and validation schemes matter to design meaningful models

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com