CORTEVA™
agriscience

Agriculture Division of DowDuPont

Good learners, faster learning

Overview on machine learning, stating problems, coordinate descent

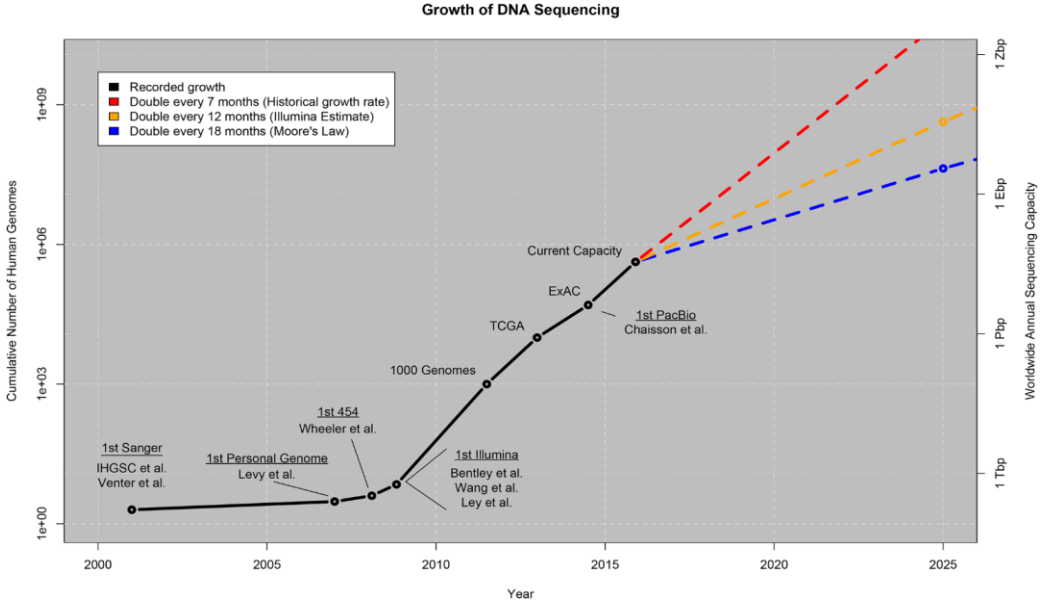
Alencar Xavier, Plant and Animal Genome 2019

Outline

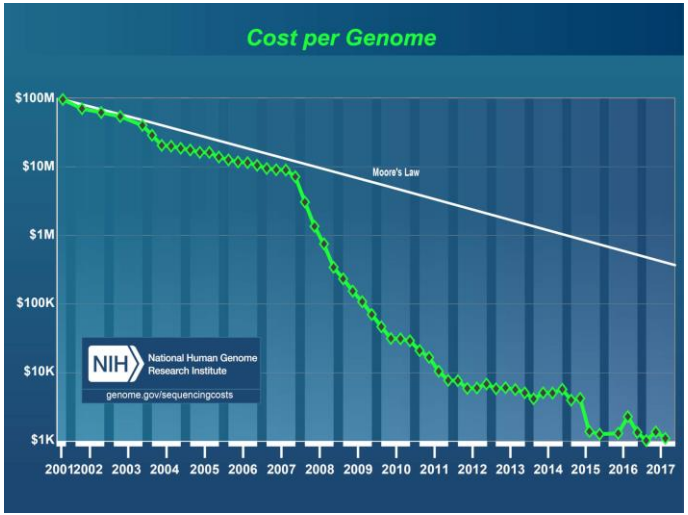
1. **Overview of Machine Learning**
   - Machine learning
   - Context on breeding
2. **Good Learners**
   - Metrics of success
   - Case of genomics
   - Case of phenomics
3. **Fast Learning**
   - Coordinate descent
   - Test on REML for RR & GWAS
   - Test on a Laplacian model

1. **Overview of Machine Learning**
   - Machine learning
   - Context on breeding
2. **Good Learners**
   - Metrics of success
   - Case of genomics
   - Case of phenomics
3. **Fast Learning**
   - Coordinate descent
   - Test on REML for RR & GWAS
   - Test on a Laplacian model

Part 1 – Overview on ML



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

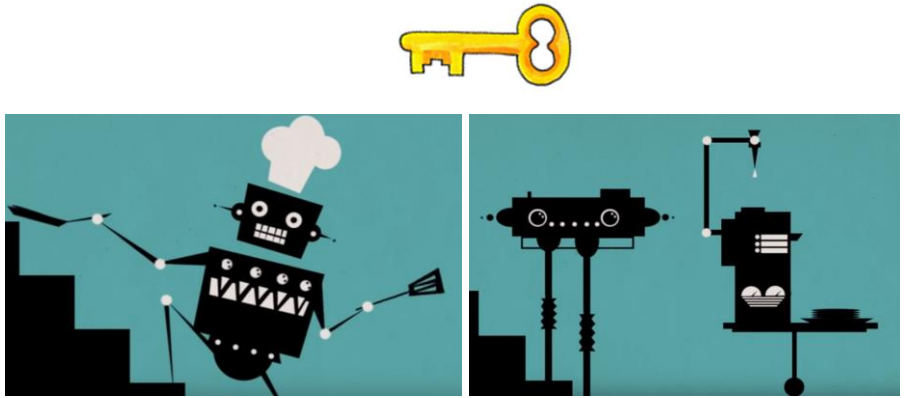


The Cost of Sequencing a Human Genome. NIH. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

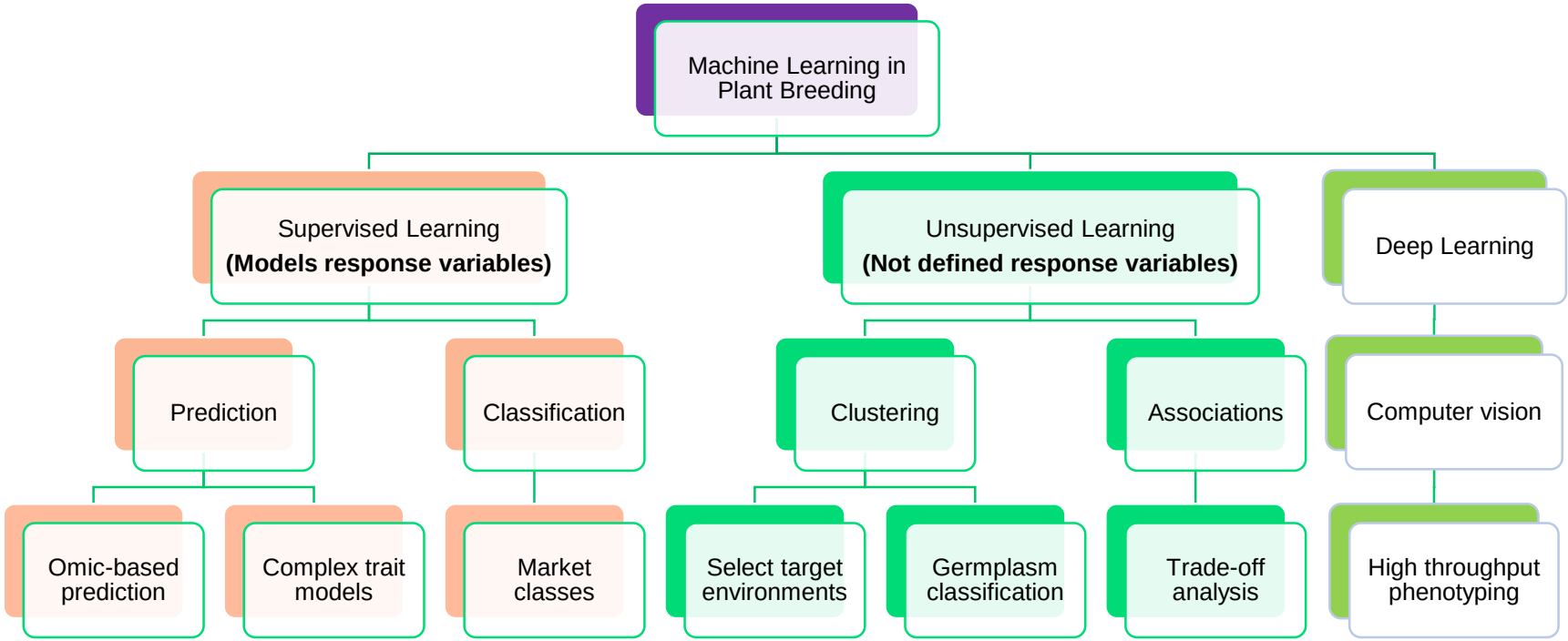
Machine Learning

- Good for solving **well-defined problem**

- Genomics
  - Prediction and selection
  - Germplasm analysis
- Phenomics
  - Automate existing traits
  - Identify new traits

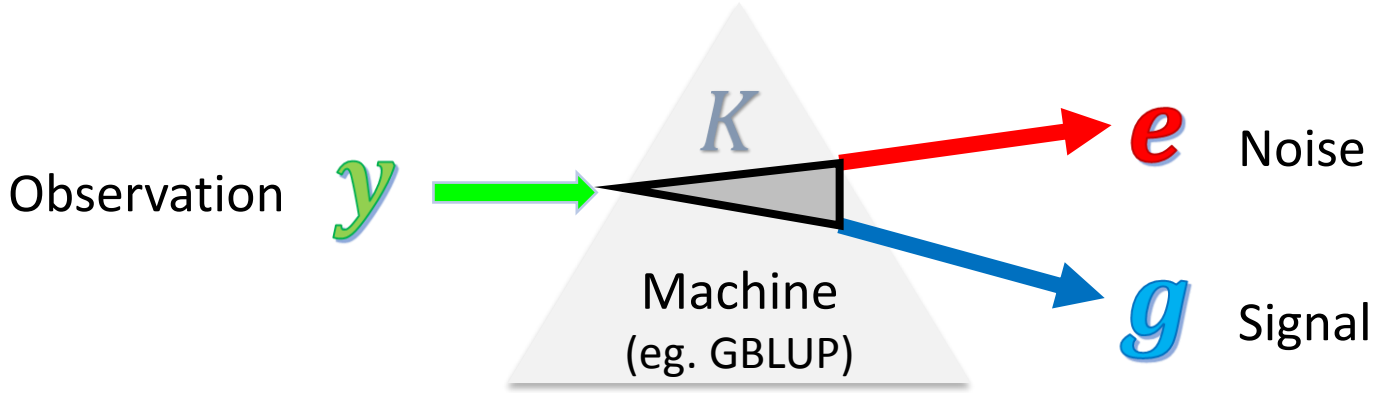


Source: https://www.youtube.com/watch?v=MPR3o6Hnf2g


Machine Learning in Plant Breeding
Supervised Learning (Models response variables)
Unsupervised Learning (Not defined response variables)
Deep Learning
Prediction
Classification
Clustering
Associations
Computer vision
Omic-based prediction
Complex trait models
Market classes
Select target environments
Germplasm classification
Trade-off analysis
High throughput phenotyping

Supervised machine: use to distinguish signal from noise

$$y = \mu + g + e$$



1. Overview of Machine Learning
   - Machine learning
   - Context on breeding
2. **Good Learners**
   - Metrics of success
   - Case of genomics
   - Case of phenomics
3. Fast Learning
   - Coordinate descent
   - Test on REML for RR & GWAS
   - Test on a Laplacian model

Part 2 – Good Learners

Defining the problem: Metrics for success

1. Scientist (**why**): to define the problem mathematically (easy to get it wrong)

2. Metric (**what**): Correlations, MSPE, Jaccard index (top X %)

3. Testing (**how**):
    1. Simulation or cross-validation (CV) on real data?
    2. Predicting phenotype or GEBV?

Testing machines for different scenarios of genomic prediction

	Genotype	Environment	Prediction 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

Case 1: Test machines to recover ***additive signal***

1. **Prediction target**: Estimate GEBVs with the entire dataset
2. **How (Testing)**: Omit a subset (e.g. family) and train model
3. **What (Metric)**: Predict GEBV subset; correlate to target GEBV

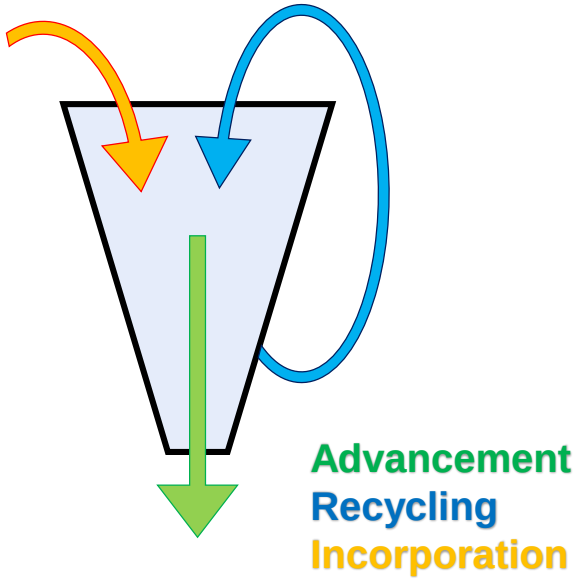
Predicted signal → **Observed signal**

GEBV GEBV

Why is additive genetics  ??

- Breeding value (**GEBV**)
  - *Pattern*: ADDITIVE GENETICS
  - *Method*: GBLUP, BayesABC, LASSO
  - *Suits*: **RECYCLING**, **ADVANCEMENT**

- Genomic value (**EGV**)
  - *Pattern*: ANY GENETICS
  - *Method*: RKHS, DNN, Random Forest
  - *Suits*: **ADVANCEMENT**



Case 2: HTP traits for yield improvement

- **Task**: Search for HTP trait that is ***additive-genetically correlated*** (r_A) to yield

  - $y_1 \rightarrow$ Trait of interest
  - $y_2 \rightarrow$ Secondary trait

$$\text{Indirect selection} = \frac{\textbf{Correlate Response}}{\textbf{Direct Response}} = \frac{h^2_{y_2} \times r_A(y_1, y_2)}{h^2_{y_1}}$$

Where does r_A come from?

Where does additive genetic correlation (r_A) come from?

- For a given model used to fit two traits (y_1, y_2):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{u} + \mathbf{e}, \qquad \mathbf{u} \sim N(0, \mathbf{G}\sigma_a^2)$$

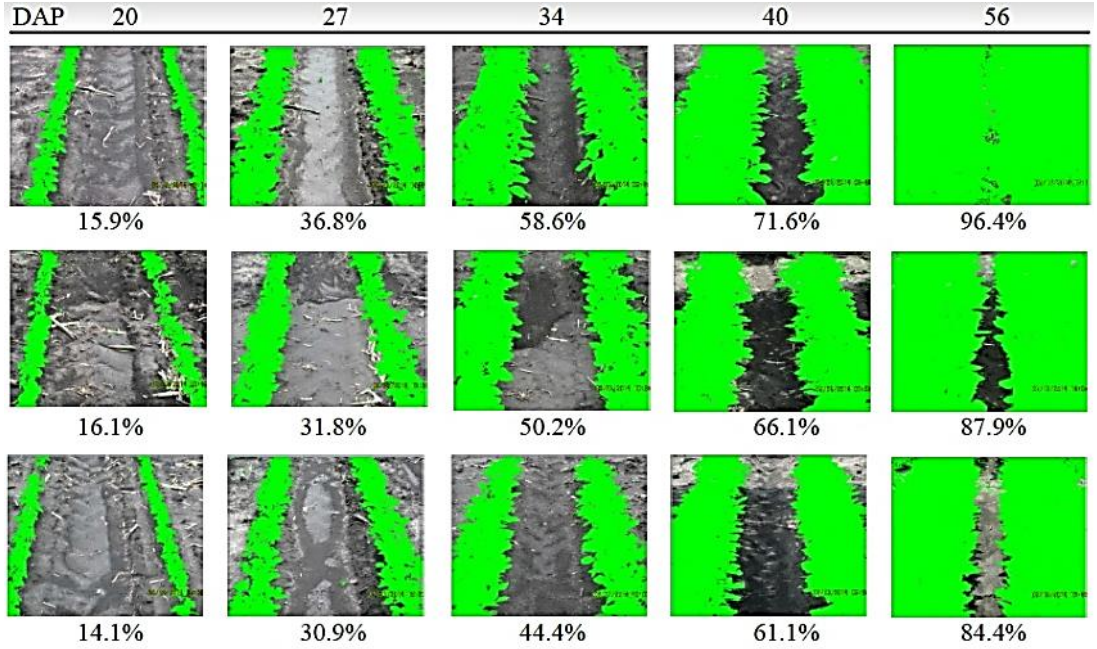
- r_A

$$r_A = \frac{\hat{u}_1' G^{-1} \hat{u}_2}{\sqrt{\hat{u}_1' G^{-1} \hat{u}_1 \times \hat{u}_2' G^{-1} \hat{u}_2}}$$

- Not r_A

$$r = \frac{\hat{u}_1' \hat{u}_2}{\sqrt{\hat{u}_1' \hat{u}_1 \times \hat{u}_2' \hat{u}_2}}$$

The case of canopy coverage in soy



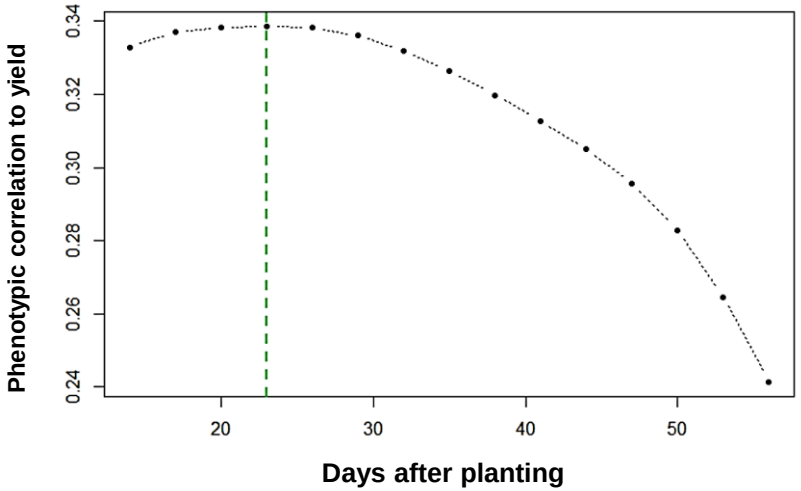
Drone HTP: Canopy coverage in soybeans

Multivariate mixed model

$$\rho_{cc,y} = \frac{CR}{R} = \frac{h^2_{cc} \times r_{cc,y}}{h^2_y} = \frac{0.76 \times 0.88}{0.58} = 1.14$$

Xavier et al. (2017)
https://doi.org/10.1534/genetics.116.198713

Phenotypic analysis



1. Overview of Machine Learning
   - Machine learning
   - Context on breeding
2. Good Learners
   - Metrics of success
   - Case of genomics
   - Case of phenomics

Part 3 – Fast Learning

3. **Fast Learning**
   - Coordinate descent
   - Conditioning and approximation for RR & GWAS
   - Test on more complicated machines

Coordinate Descent (CD)

(see Friedman et al. 2010, dx.doi.org/10.18637/jss.v033.i01)

- ***Key concept***: Reduce the dimensionality of the problem to univariate through "*conditioning*"

- *How does CD work?*
  1. Update a coefficient
  2. Update residuals

- *Other popular names for CD*
  - **Gauss-Seidel residual update** (see Legarra & Misztal 2008, 10.3168/jds.2007-0403)
  - **Full-conditional expectation** – The backbone of *Gibbs sampling* methods

CD works on Variance Components

$$\sigma_i^2 = \frac{\mathbf{u}_i' \mathbf{K}_i^{-1} \mathbf{u}_i + \mathrm{tr}\left(\mathbf{K}_i^{-1} \mathbf{C}^{ii}\right)\sigma_e^2}{q_i} = \textit{HARD TO GET } (\textit{for large marker datasets})$$

$$\sigma_e^2 = \frac{\mathbf{y}'\mathbf{e}}{n - r_X} = \textit{EASY TO GET}$$

- **Consider the model**

$$\mathbf{y} = \mu + \mathbf{M}\mathbf{a} + \mathbf{e}$$

- If you **condition** the response variable to a **single marker**

$$\tilde{\mathbf{y}} = \mathbf{y} - \left(\mu + \mathbf{M}_{-j}\mathbf{a}_{-j}\right) = \mathbf{m}_j a_j + \mathbf{e}$$

- The genetic variance becomes

$$\sigma_{SNP}^2 = \frac{\mathbf{a}'\mathbf{K}^{-1}\mathbf{a} + \mathrm{tr}(\mathbf{K}^{-1}\mathbf{C}^{ii})\sigma_e^2}{q} = \mathbf{a^2} + \frac{\boldsymbol{\sigma_e^2}}{\mathbf{m}'\mathbf{m} + \boldsymbol{\lambda}}$$

Study 1A) conditioning and approximation of RR-BLUP

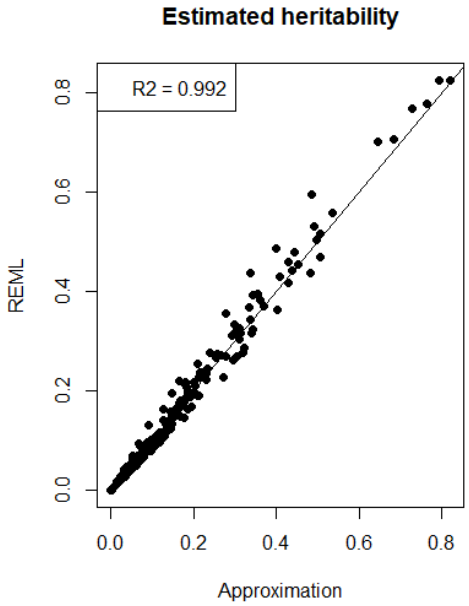
$$\mathbf{y} = \mu + \mathbf{Ma} + \mathbf{e}, \qquad \mathbf{a} \sim N(0, \sigma^2_{SNP})$$

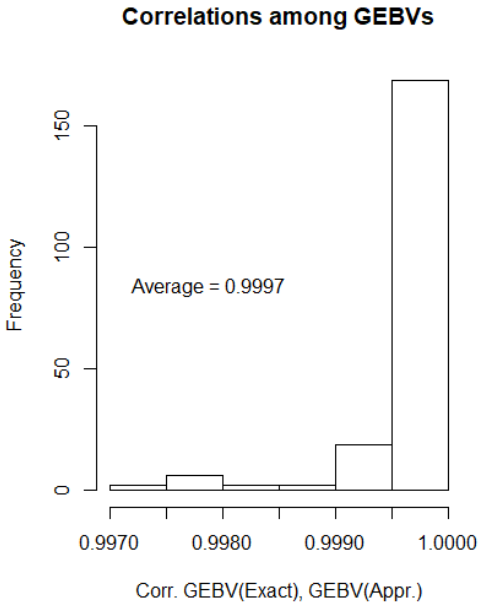
- A simple speed up using
  1. Conditioning: Use **CD** to marker effects ($\mathbf{a}$) 
  2. Approximation: To get genetic variance (σ^2_{SNP}) 

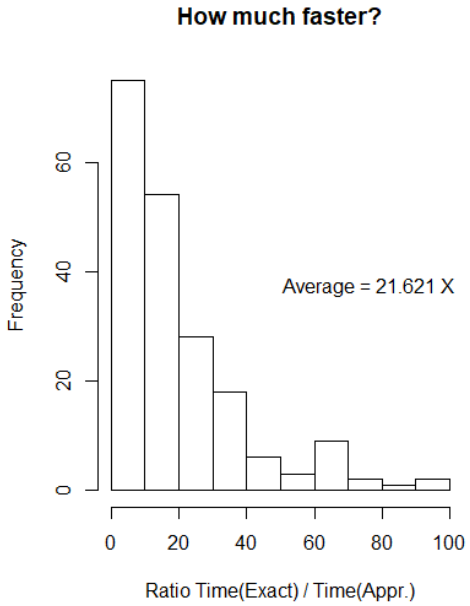
$$\boldsymbol{\sigma^2_{SNP} \cong \frac{\sigma^2_y - \sigma^2_e}{\sum \sigma^2_{M_j}}}$$

- Compare to EMMA REML

Variance components and breeding values







200 simulated datasets, varying N, P and h2

Some contribution from Vishnu Ramasubramanian (Beavis Lab, Iowa State)

Study 1B) conditioning and approximation of GWAS

- Test SNPs via likelihood ratio costs only 1 extra round of CD

  - Reduced model *(does not contain the marker of interest)*

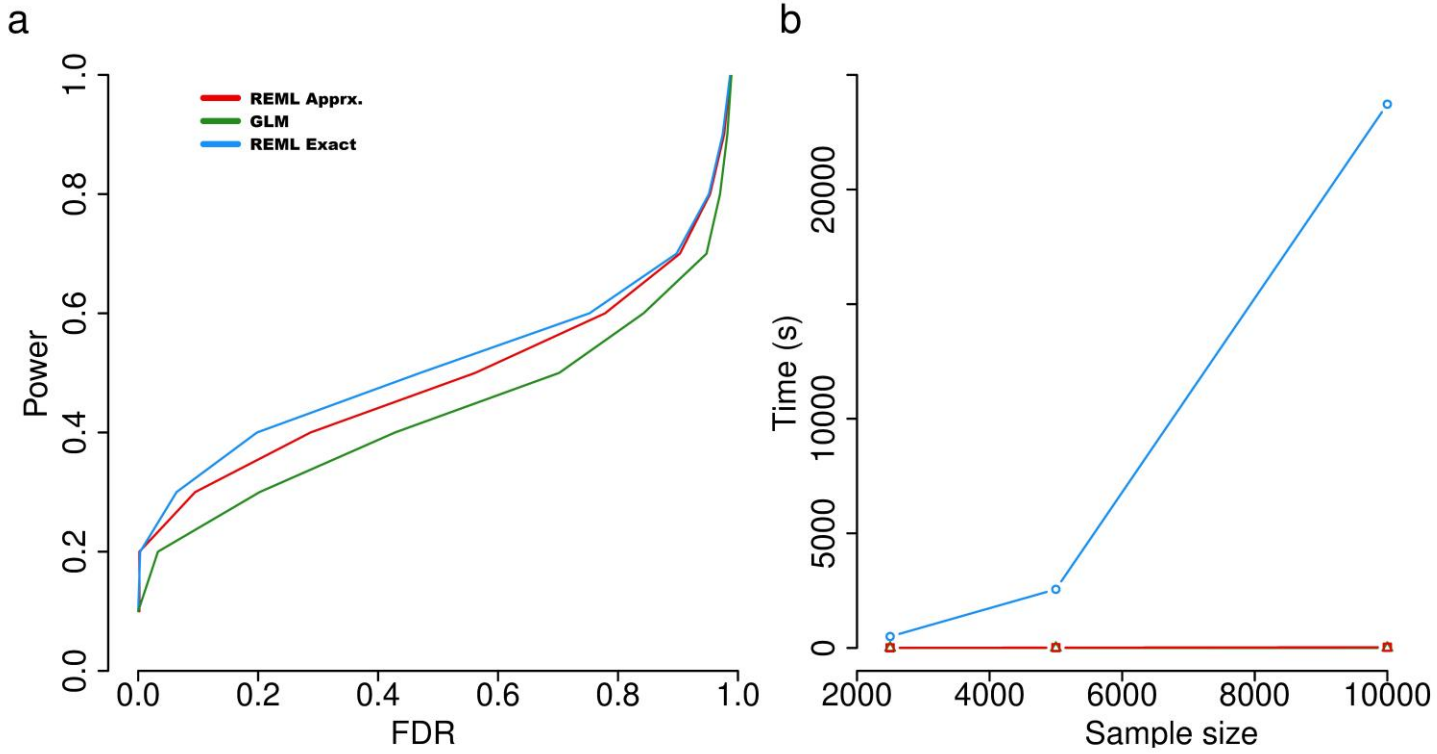
$$\mathbf{y} = \mu + \mathbf{M}_{-j}\mathbf{a}_{-j} + \mathbf{e}$$

  - Full model *(contains the marker of interest)*

$$\mathbf{y} = \mu + \mathbf{m}_j a_j + \mathbf{M}_{-j}\mathbf{a}_{-j} + \mathbf{e}$$

Legend
Fixed effects
Random effects

Power, false positives, computing time



Analysis provided by Meng Huang (Rainey Lab, Purdue)

Study 2) Conditioning a Laplacian machine

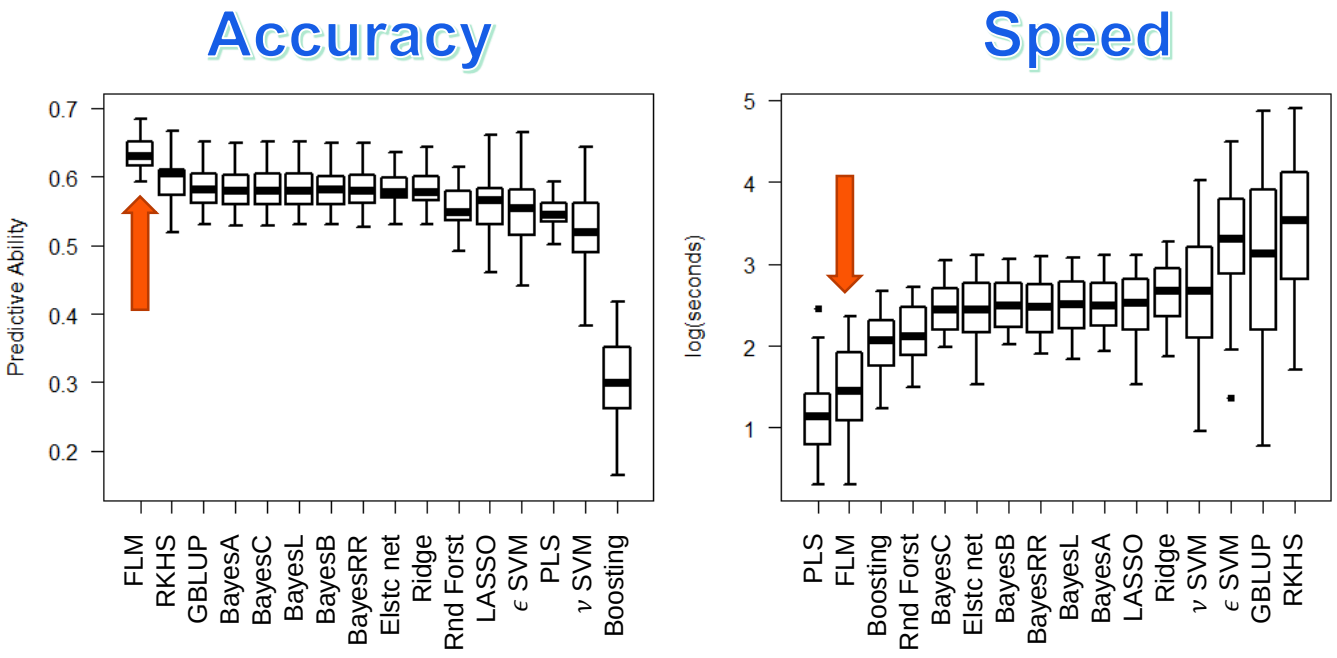
A. First, we test as Whole-Genome Regression (WGR)

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Ma} + \mathbf{e}, \qquad \mathbf{a} \sim \mathrm{N}(0, \mathrm{T}^2\sigma_e^2)$$

B. Test a single-step: Plugging WGR into a mixed model via conditioning

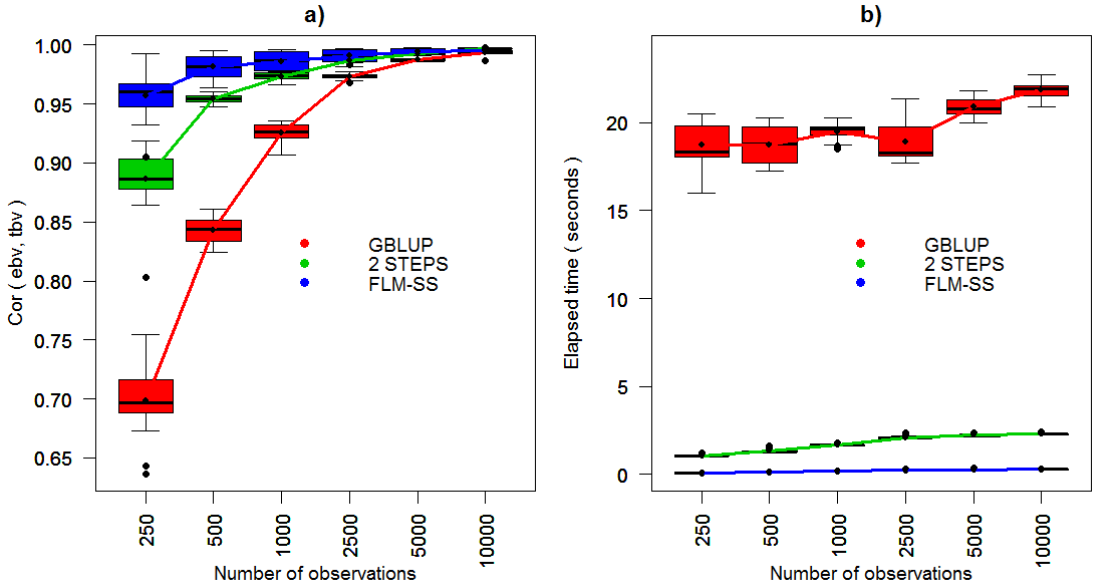
$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \qquad \textbf{where} \quad \mathbf{u} = \mathbf{Ma} + \boldsymbol{\varepsilon}$$

Study 2A) Laplacian machine on genomic prediction



Maize datasets (from 2 heterotic groups and 10 geographies, 5-fold cross-validations 20x): Box plot of prediction accuracy (left) and computing time (right) across methods. Models ordered based on the average performance.

Study 2B) Laplacian machine on single-step model



This comparison was performed across a variable number of individuals (n = 250, 500, 1000, 2500, 5000 and 10000) and trait heritability (h^2 = 0.25, 0.50, 0.75), each scenario was repeated 40x with different seeds to sample the individuals.

Concluding Remarks ( 's)

1. Machines respond to well defined problems

2. Metrics of success are critical (e.g. r_A and $gebv$'s)

3. Conditioning and approximations can make good machines

Concluding Remarks ( 's)

1. Machines respond to well defined problems

2. Metrics of success are critical (e.g. r_A and $gebv$'s)

3. Conditioning and approximations can make good machines

Acknowledgments

- **Students (CD Study 1)**
  - Meng Huang (Purdue)
  - Vishnu Ramasubramanian (ISU)
- **Insight**
  - Mak Geha (Corteva)
  - David Habier (Corteva)
  - Bill Muir (Purdue)
  - Shizhong Xu (UC Riverside)
- **Funding source**
  - Corteva Agriscience
  - United Soybean Board
- **Canopy study**
  - Katy Rainey (Purdue)
  - Ben Hall (Purdue)
  - Keith Cherkauer (Purdue)
  - Anthony Hearst (Purdue)
- **Additional support and review**
  - Radu Totir (Corteva)
  - Keith Boldman (Corteva)
  - Jing Wang (Corteva)

Concluding Remarks

1. Machines respond to well defined problems
2. Metrics of success are critical (e.g. r_A and $gebv$'s)
3. Conditioning and approximations can make good machines

Thank you for your attention!

Questions??

Alencar Xavier

alencar.xavier@Corteva.com