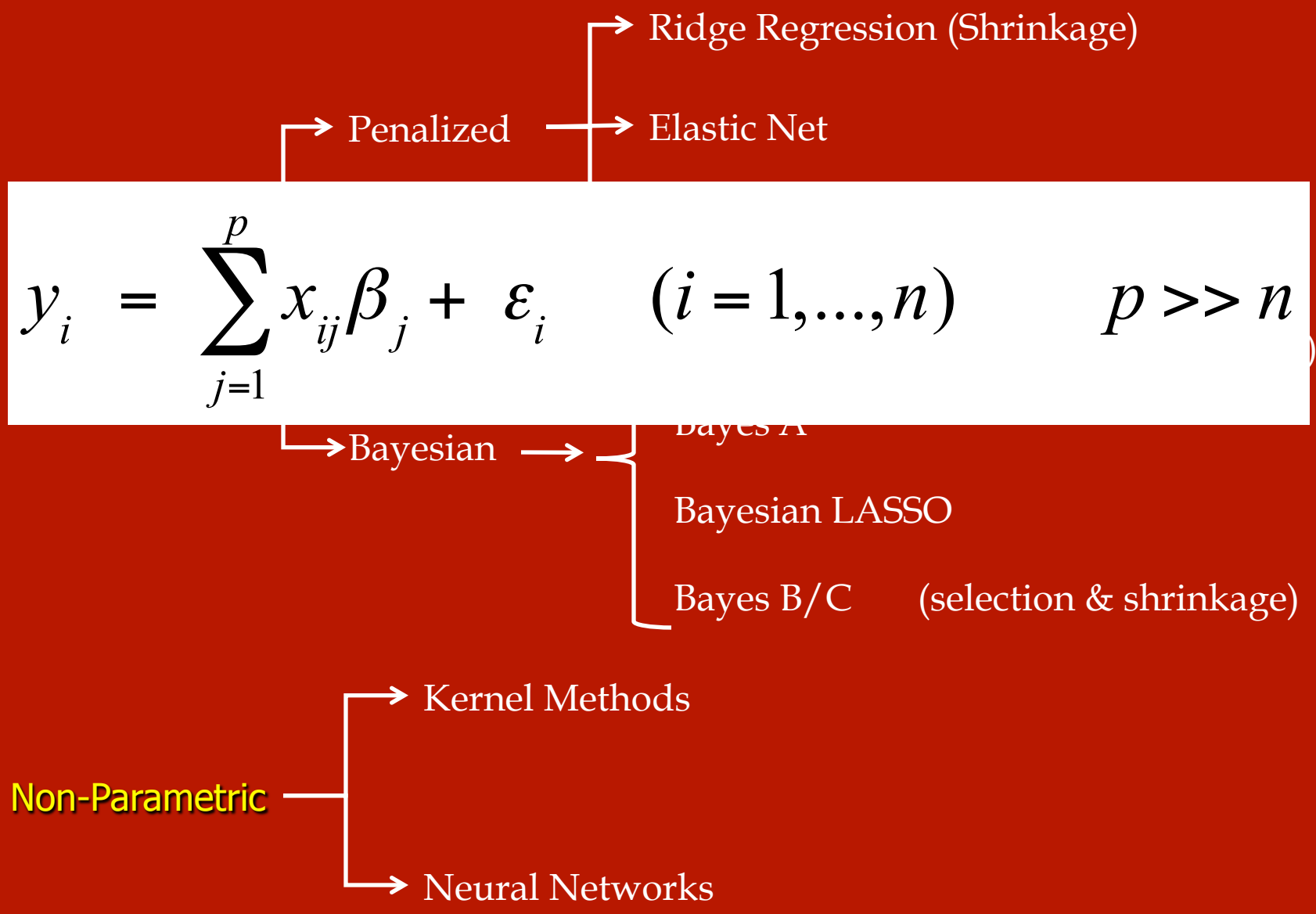
Genomic Prediction in the era of Big Data

- Overview of methods

- & Selected computational and statistical challenges

Gustavo de los Campos

Whole-Genome Regression/ Prediction Methods [1]



[1]: Meuwissen Hayes & Goddard (2001)

Penalized

Penalty

Ridge Regression (RR) $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$

Elastic Net (EN) $J(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \sum_{j=1}^{p} \beta_j^2 + (1-\alpha)\alpha \sum_{j=1}^{p} |\beta_j|$

LASSO $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$

Bayesian

Prior Density

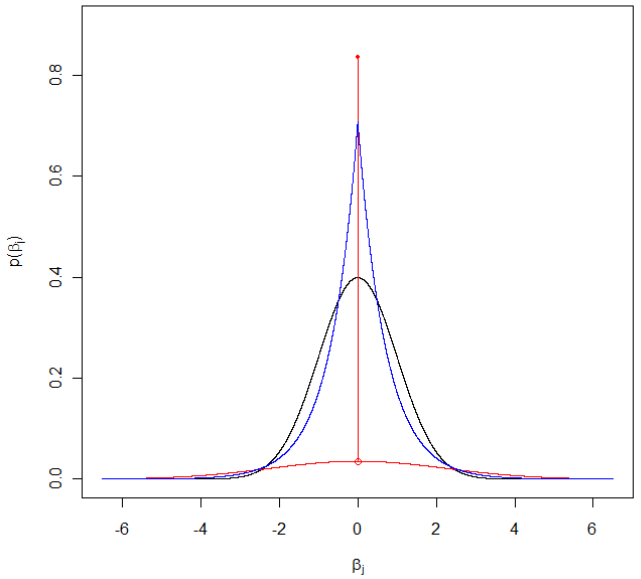
Gaussian (RR-BLUP/ G-BLUP)

Scaled-t (Bayes A)

Double-Exponential (Bayesian Lasso)

Spike-Slab (SSVS)

Point of mass & Slab (Bayes B, Bayes C)



Prior densities commonly assigned to marker effects

Software (selected examples)

	Language	Penalized	Bayesian	Shrinkage	Variable Selection	Kernel Regression	Nueral Networks	Deep Lerning
glmnnet	R	x		x	x			
rrBLUP	R	x		x		x		
ASREML		x		x				
BGLR	R		x	x	x	x		
BLUP-f90		x		x				
GenSel			x	x	x			
Tensorflow	Python	x		x	x		x	x

- We have a diverse array of software for genomic regression
- Most of the available packages scale to very large problems
- Most packages are user friendly and open-source

Selected Computational & Statistical Challenges

1. Genomic prediction with big data
2. Dealing with imperfect LD and highly heterogeneous data sets (sometimes small is better…)
3. Modeling and leveraging GxE
4. Integrating high-dimensional phenotypes from high-throughput phenotyping

1: Genomic Prediction with Big Data

- Data sets are becoming increasingly large (hundreds of thousands of genotypes linked to phenotypic records)

- Handling these data sets requires becoming familiar with a few important concepts (e.g., memory mapping, distributed arrays, distributed computing, etc)

- But there is already software that can handle extremely large data sets

- The main challenge is how to train ourselves and our students to become proficient on big data analyses

2: Dealing with imperfect LD (sometimes small is beautiful!)

- SNPs are in imperfect LD with the alleles at causal loci.
- The models we use are at best good local approximations to highly complex problems (epistasis is pervasive)
- Therefore, across generations and meiosis, LD breaks, allele frequencies change and therefore additive effects change.
- For this reason, sometimes combining very large data sets with distantly related genotypes my harm prediction accuracy.
- How do we adequately balance the benefits and potential problems of Big Data?

How far should we go to train genomic prediction models?

	SET 1	SET 2	SET 3	SET 4	SET 5
Scenario A_4518	TRN	TRN	TRN	TRN	TST 4
Scenario A_4515	TRN	TRN	TRN	TST 3	TST 4
Scenario B	TRN	TRN	TST 2	TST 3	TST 4
Scenario C	TRN	TST 1	TST 2	TST 3	TST 4

Source: Wolc et al., Animal Frontiers, 2016

Predictive Correlation

Scenario	SET 1 (N=8,144)	SET 2 (1,655)	SET 3 (1,758)	SET 4 (3,400)	SET 5 (3,492)
I	N=16,794				0.451
II	N=15,036			0.553	0.412
III	N=11,636		0.514	0.429	0.348
IV	N=8,144	0.436	0.392	0.367	0.308

3: Modeling and leveraging GxE

- Genetic-by-environmental interaction is very important in plants

- In principle we can use genomic information, linked to environmental data, to breed for target environments.

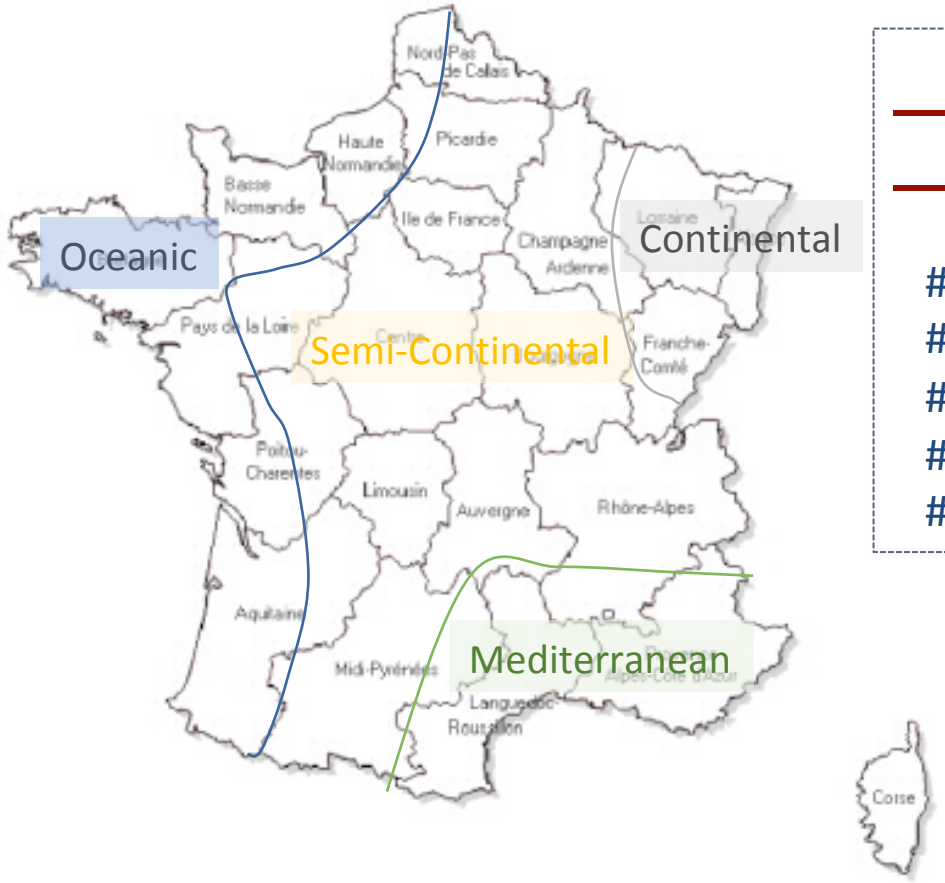
- There are several (old and new) approaches for dealing with GxE in genomic predictions, e.g.,

    - Marker-by-environment interactions (e.g., Lopez-Cruz et al., G3, 2015)
    - Reaction norm models for SNPs and environmental covariates (e.g., Jarquin et al., TAG, 2014)
    - Crop models (e.g., Cooper et al., Crop Sci, 2016)

- An important problem is how to integrate GxE when designing breeding strategies (an old problem).

- Predict the future from the past… how to leverage historical whether records with trial data?

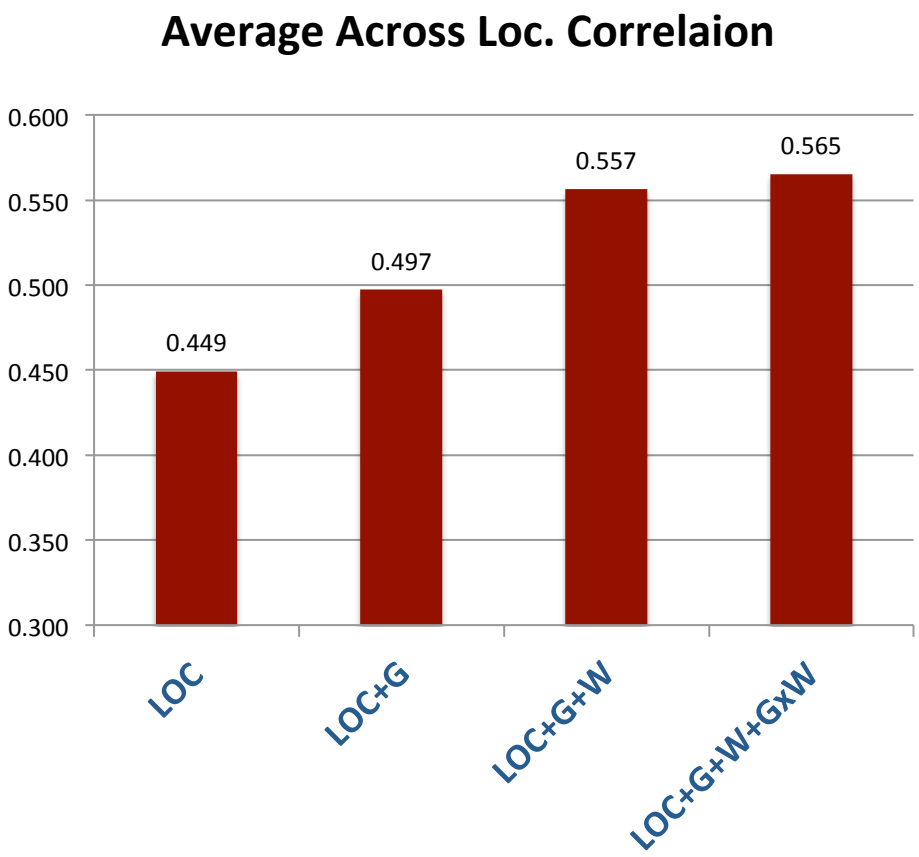
DATA



DATA

of records: 28,554 (yield adjusted-by design)
# of wheat lines: 601
# of markers: 213,339
# of year-locations: 875
# of environmental covariates: 125.

CV-Correlation



Remarks

- Adding markers and environmental covariates increased the correlation by ~25%

Conceptual Framework/Pipeline

Experimental Data

- Phenotypes (from many varieties, evaluated at several locations, over many years)

- Genotypes (SNPs)

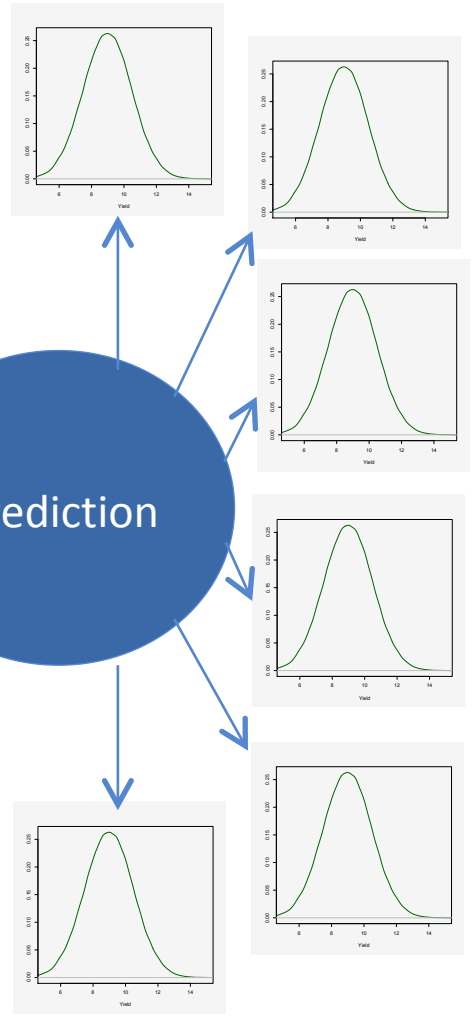
- Environmental Covariates

Historical Weather Data

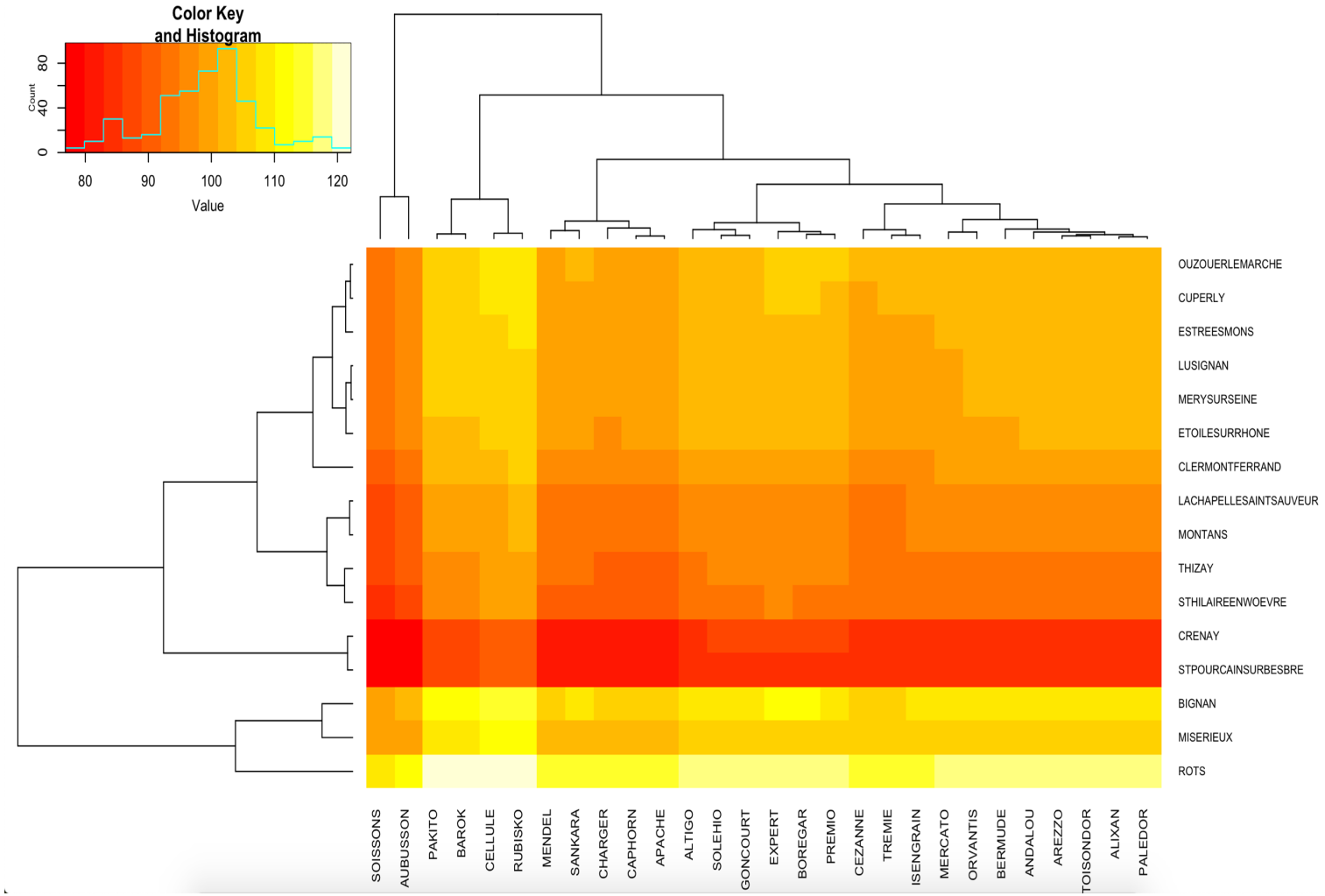
Fitted Model

Compute environmental covariates for a set of lines and locations over many years

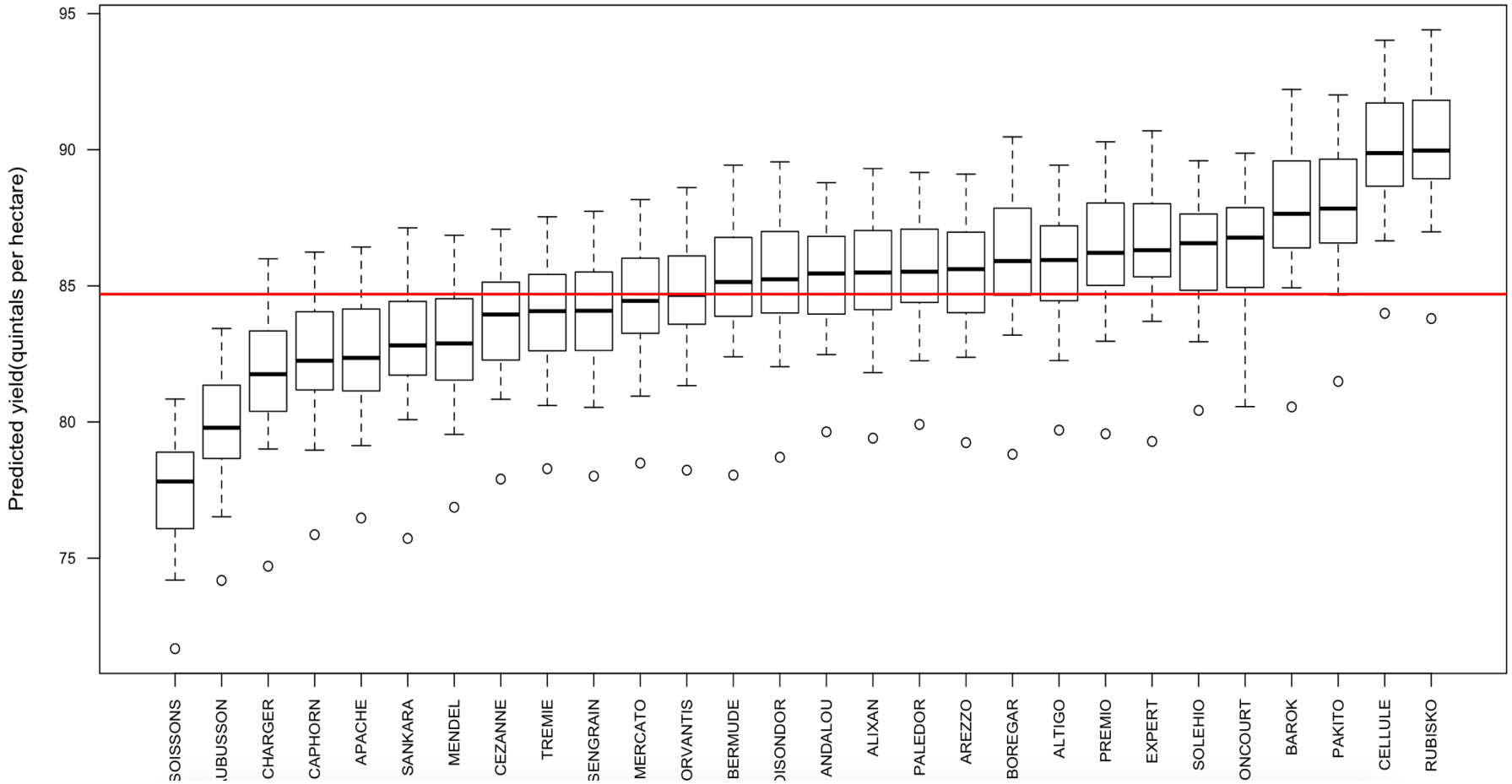
Prediction



3: Modeling and leveraging GxE



3: Modeling and leveraging GxE



4: Leveraging High throughput phenotyping

4: Leveraging High throughput phenotyping

- HTP has been adopted in agricultural research and commercial production.
- There are large volumes of research on how to use HTP to predict phenotypes an to optimize management practices.
- However, there is much less research on how to incorporate HTP data on breeding schemes.
- HTP platforms generate high-dimensional phenotypes (hundreds or thousands of traits per unit being monitored)
- For example, hyper-spectral cameras (reflectance at hundreds of wavenumbers over many time-points).
- How do we integrate this information in breeding schemes?
- Two challenges:
    - **Breeding strategy**: at what steps of the breeding process and with what objective we integrate HTP?
    - **Statistical**: what methods can be used to integrate high dimensional phenotypes into genomic prediction models?

Integrating Hyper-Spectral Crop Imaging Into Breeding Using Penalized Selection Indices

Marco Lopez-Cruz

Penalized Selection Indices

Ridge Regression

LASSO

Elastic Net

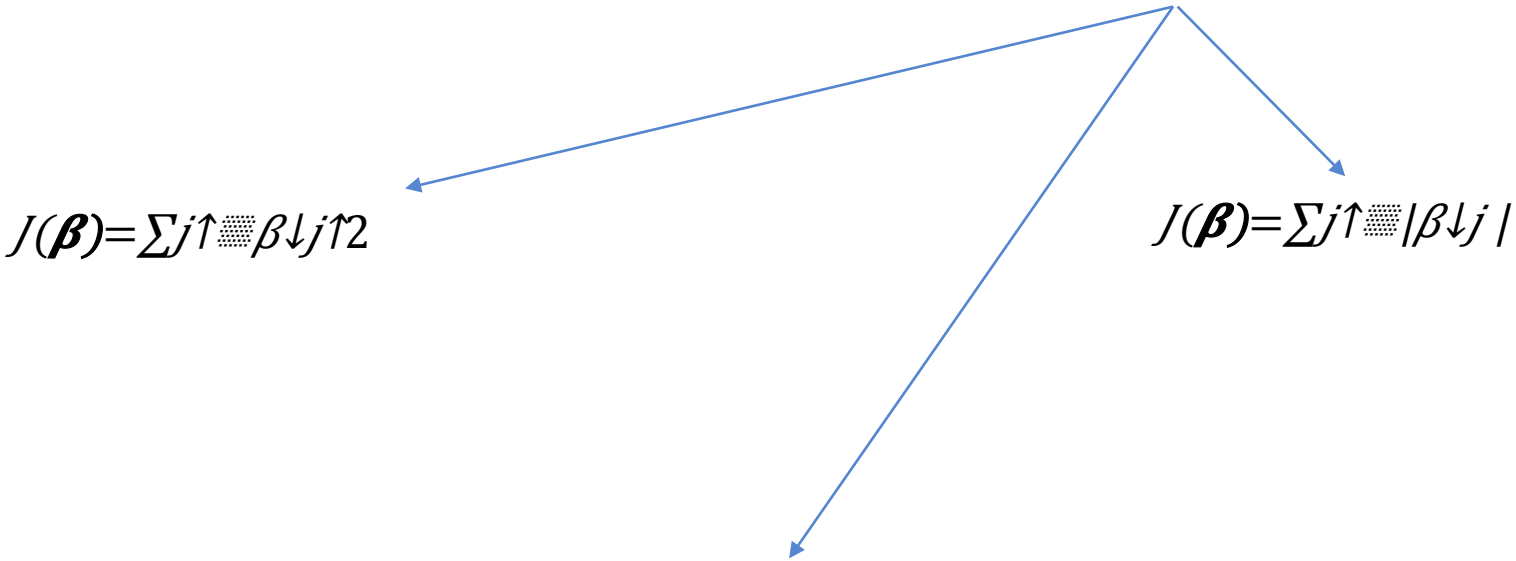
Compressed Sensing

Graphical Lasso

Support Vector Machine

Penalized Selection Index

$$\boldsymbol{\beta}\hat{} = \downarrow E(g\downarrow y - \boldsymbol{x}\uparrow' \boldsymbol{\beta})\uparrow 2 + \lambda J(\boldsymbol{\beta})$$



Accuracy of Indirect Selection of Canonical and Penalized SI for wheat yield