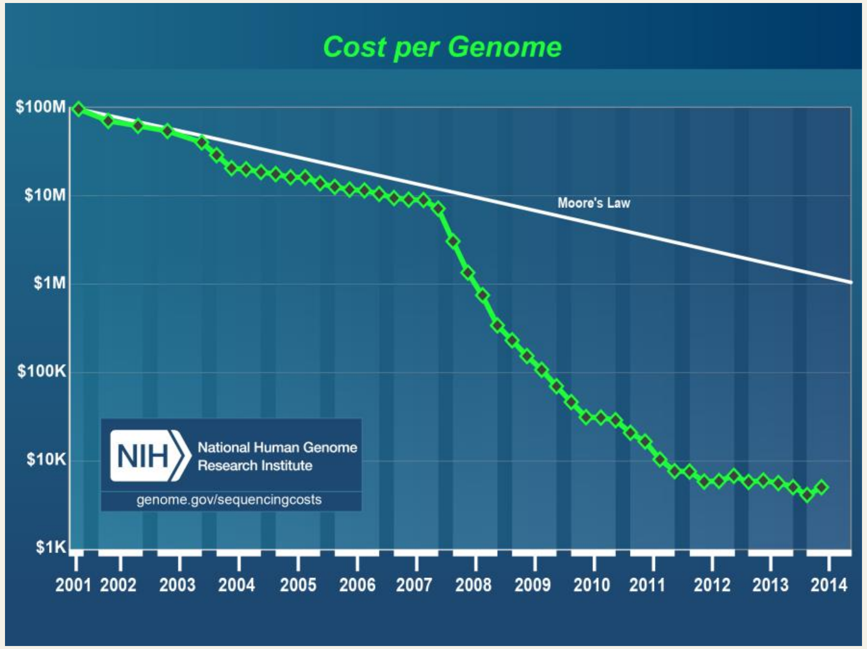
informs ANNUAL MEETING
2021 ANAHEIM, CALIFORNIA

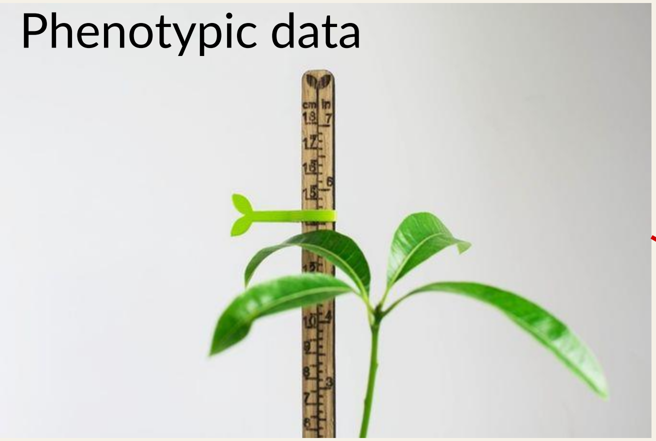
Combining Multi-Type Data to Improve Categorical Trait Prediction

Vamsi Manthena, Diego Jarquin, Reka Howard

UNIVERSITY OF Nebraska Lincoln

Genomic Prediction (GP)





Phenotypic data

$$$

Marker data

$

G-BLUP model

$y = g + e$

Prediction and Selection

Test candidates

Multi-type Data and Classification

- Modern plant breeding programs collect several types of data
  - Phenotypic trait data, genomic data
  - HTP, weather, image data

- Categorical phenotypic traits
  - Susceptibility to disease
  - Resistance to drought/salinity
  - Number of tassels

- How do we leverage other data types to improve prediction?

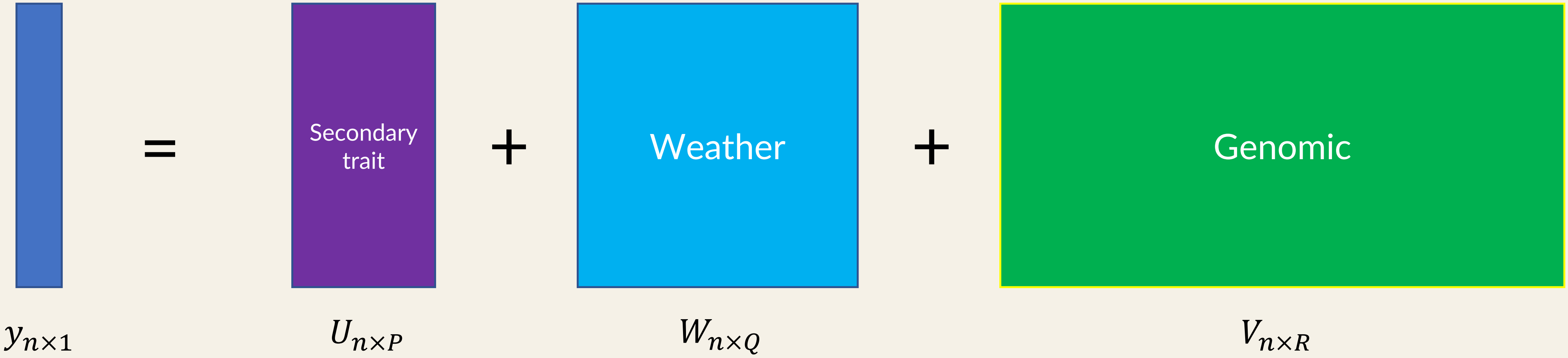
Developing Methods for Classification using Multi-Type Data

- Combine three types of data:
  - Secondary traits
  - Weather
  - Genomic data

- Predict multi-class categorical response

Key Challenges

- High-dimensionality of genomic data
- Genomic covariates (R) >> weather covariates (Q) > secondary trait (P)
- Confounding: Genomic & weather covariates affect main trait and secondary traits.
- Threshold optimization for classification



= Secondary trait + Weather + Genomic

$y_{n\times 1}$ $U_{n\times P}$ $W_{n\times Q}$ $V_{n\times R}$

Step 1:
Intrinsic Effect of Secondary Traits

- Regress each secondary trait on the set of weather variables and obtain residuals:

$$\hat{u}_{ip} = u_{ip} - w_i^T \, \hat{b}_p$$

where $\hat{b}_p = (\hat{b}_{p1}, \hat{b}_{p2}, \dots, \hat{b}_{pQ})$.

- The regression coefficients are estimated by minimizing:

$$\sum_{i=1}^{n} \left(u_{ip} - w_i^T \, b_p\right)^2 + \lambda \sum_{q=1}^{Q} pen\left(\left|b_{pq}\right|\right)$$

- Penalty functions: LASSO, aLASSO, RR, SCAD.

Step 1: Contd.

- Regress each residual on the set of genomic variables and obtain double residuals:

$$\hat{\hat{u}}_{ip} = \hat{u}_{ip} - v_i^T \hat{d}_p$$

where $\hat{d}_p = (\hat{d}_{p1}, \hat{d}_{p2}, \dots, \hat{d}_{pR})$.

- The regression coefficients are estimated by minimizing:

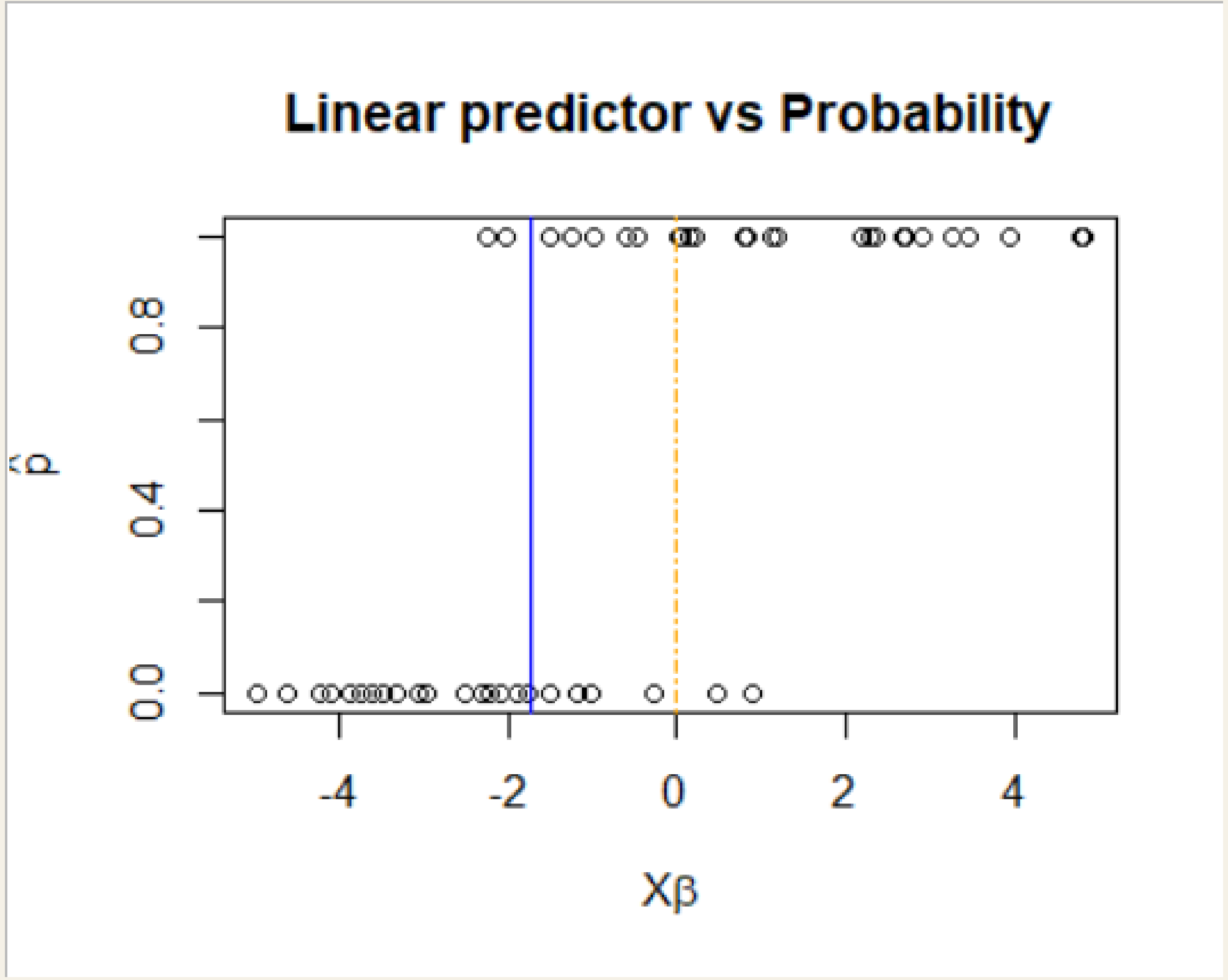
$$\sum_{i=1}^{n} \left(\hat{u}_{ip} - v_i^T d_p\right)^2 + \lambda \sum_{r=1}^{R} pen\left(\left|d_{pr}\right|\right)$$

- Penalty functions: LASSO, aLASSO, RR, SCAD.

Step 2:
Logistic Regression with Penalized Forward Selection

- Forward selection
  - Differing sizes of data types
- Penalized logistic regression
- Advantages:
  - Sparse Models – Interpretability
  - Control on data type entering model

Step 3:
Threshold Optimization for Classification



Step 4:
Model Evaluation

- Classify test observations
  - Coefficients from step 2
  - Threshold from step 3

- Compute classification errors for evaluation
  - Accuracy
  - TPR
  - TNR

Data Description

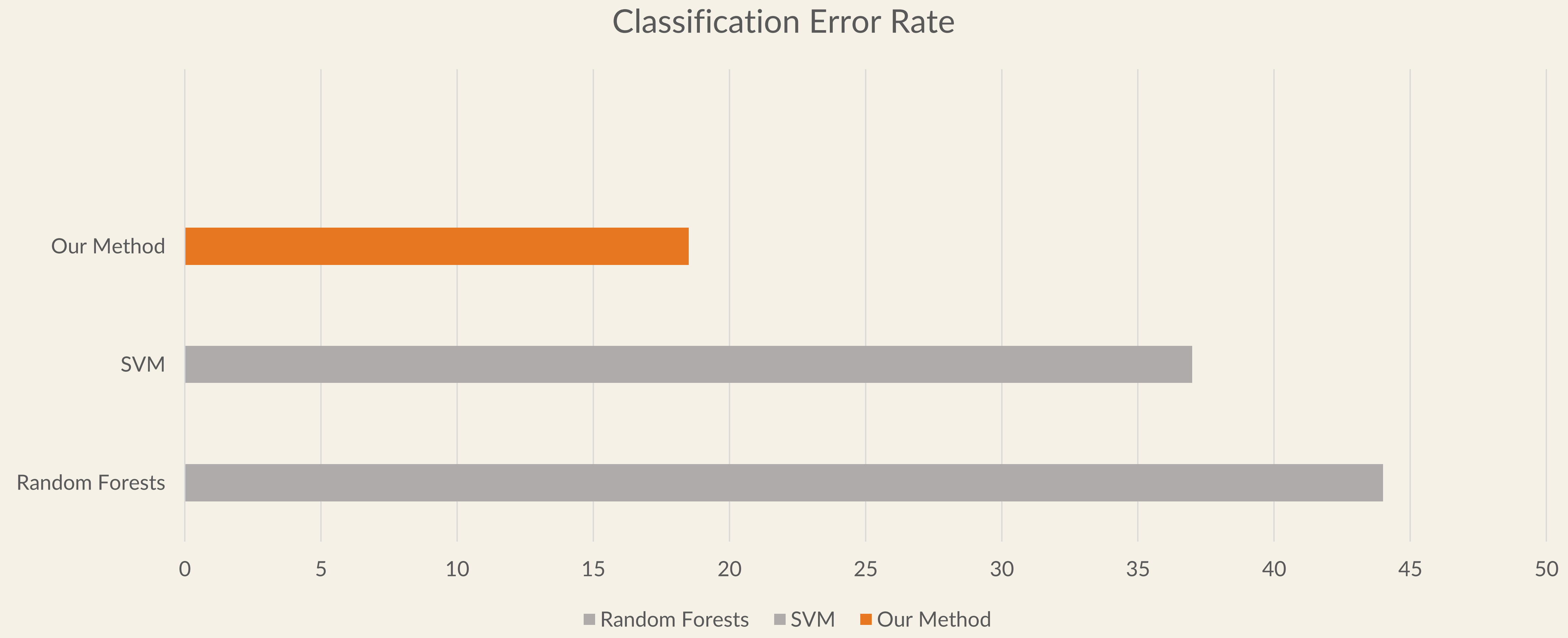
- Chickpea Data set with 278 lines
- Main Trait: Days to Maturity – Low/High

Secondary Traits – 6 traits

Weather – 4 × 100 weather covariates

Marker Data – 10000 markers

Preliminary Results



Conclusions

- Promising preliminary results
  - Outperforming ML methods
- Future work:
  - Imbalanced classification
  - Multi-class classification
  - Weather window optimization
- Applications:
  - Genomic Prediction
  - Biomedical precision medicine

informs ANNUAL MEETING
2021 ANAHEIM, CALIFORNIA

Combining Multi-Type Data to Improve Categorical Trait Prediction

Contact:

vamsi.manthena@gmail.com

www.vamsimanthena.com

N | UNIVERSITY OF Nebraska Lincoln

Select References

- Ghosal S, Hwang WY, Zhang HH. FIRST: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. Statistics and Its Interface. 2009;2(3):341–348. doi:10.4310/SII.2009.v2.n3.a7
- Turnbull B, Ghosal S, Zhang HH. Iterative selection using orthogonal regression techniques. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2013;6(6):557–564. doi:https://doi.org/10.1002/sam.11212
- Ghosal S, Turnbull B, Zhang HH, Hwang WY. Sparse Penalized Forward Selection for Support Vector Classification. Journal of Computational and Graphical Statistics. 2016;25(2):493–514. doi:10.1080/10618600.2015.1023395
- Crain J, Mondal S, Rutkoski J, Singh RP, Poland J. Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding. The Plant Genome. 2018;11(1):170043. doi:https://doi.org/10.3835/plantgenome2017.05.0043
- Jarquin, D., Roy, A., Clarke, B. and Ghosal, S., 2015. Combining phenotypic and genomic data to improve prediction of binary traits. https://www4.stat.ncsu.edu/~sghosal/papers/JRCG_multidata.pdf