GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN
Division of Plant Breeding Methodology
Department of Crop Sciences

CiBreed
Center for Integrated Breeding Research

Understanding maize phenotypic traits variability using a machine learning method that incorporates genomic and environmental covariates

Cathy Jubin
University of Göttingen, CiBreed

Genomes to Fields Workshop
Phenome2020 01/24/2020

Poster 209

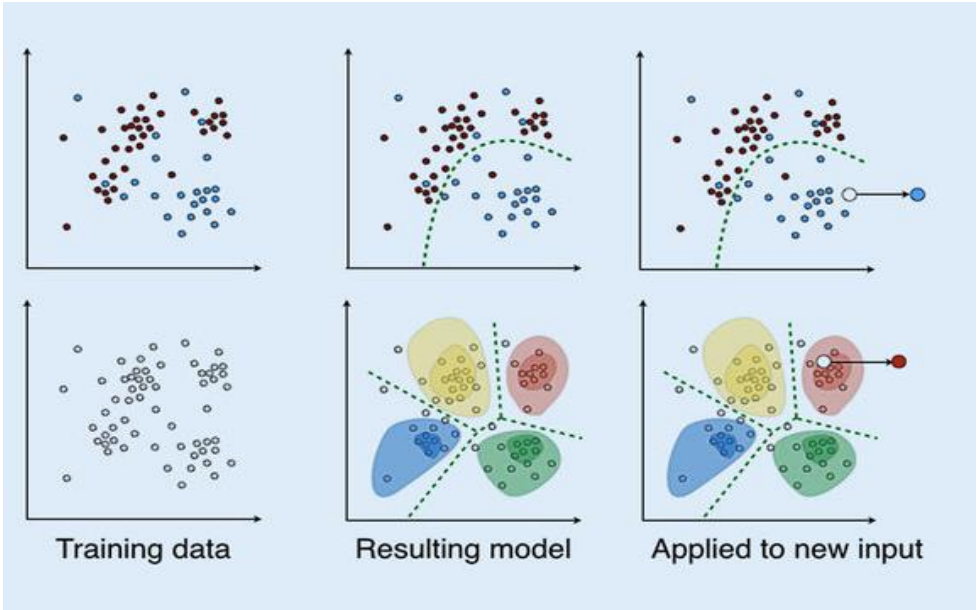
Machine Learning: a data-driven science

Branch of artificial intelligence concerned with the development of algorithms able to recognize patterns, or to predict categorical/quantitative targets on new data, after learning from examples.

Two main learning problems categories:

Supervised

Unsupervised



→ **Classification problems**
→ **Regression problems**

Langs, G., et al. (2018)

Why applying Machine Learning on multi-environment trials datasets?

More and more field data generated… how to use it ?

Environment
- Weather variables
- Soil characteristics
- Agronomic practices
- Biological environment

Various data types
High dimensional datasets

Genome and multi-omics data

Phenotype

‚Typical' breeding data: flowering time

Weather station

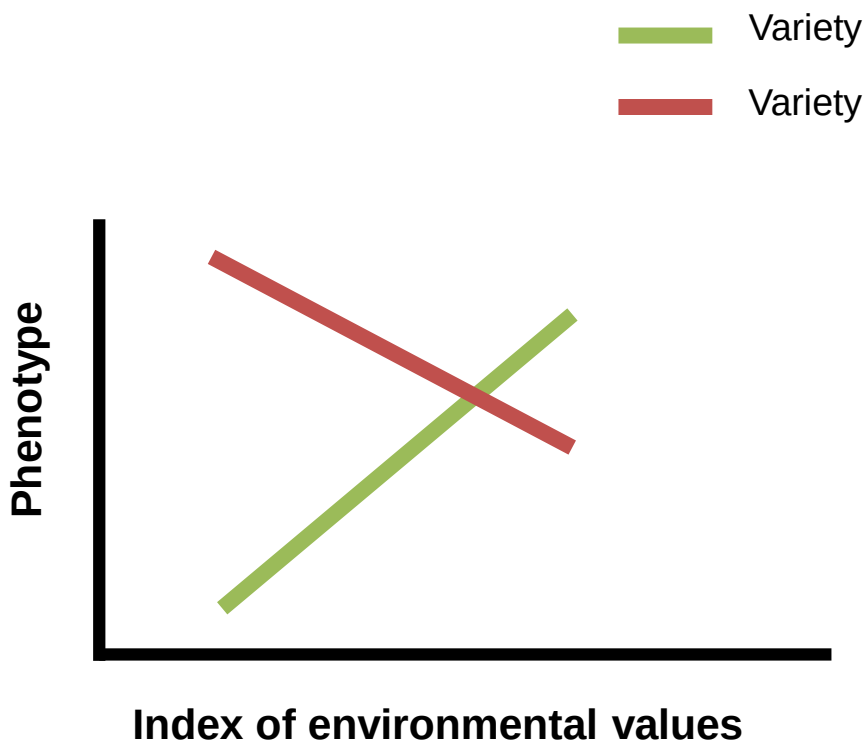
Picture source: Barış Alaca, Göttingen

+ drones, soil moisture sensors, high-throughput phenotyping platforms

Why applying Machine Learning on multi-environment trials datasets?

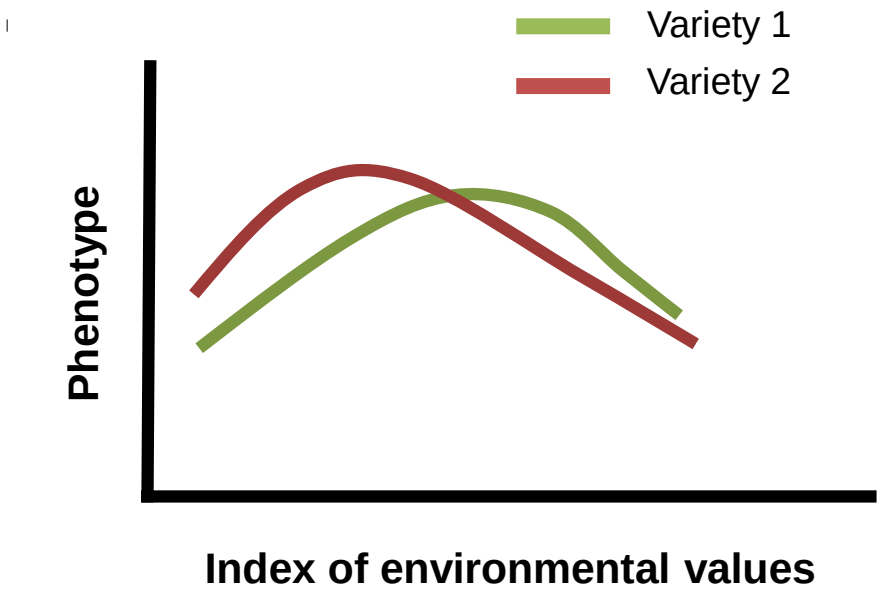
GxE, cross-over interactions

Various statistical models accounting for GxE interactions (Heslot et al. 2014, Jarquin et al. 2014)



But often more complexity:

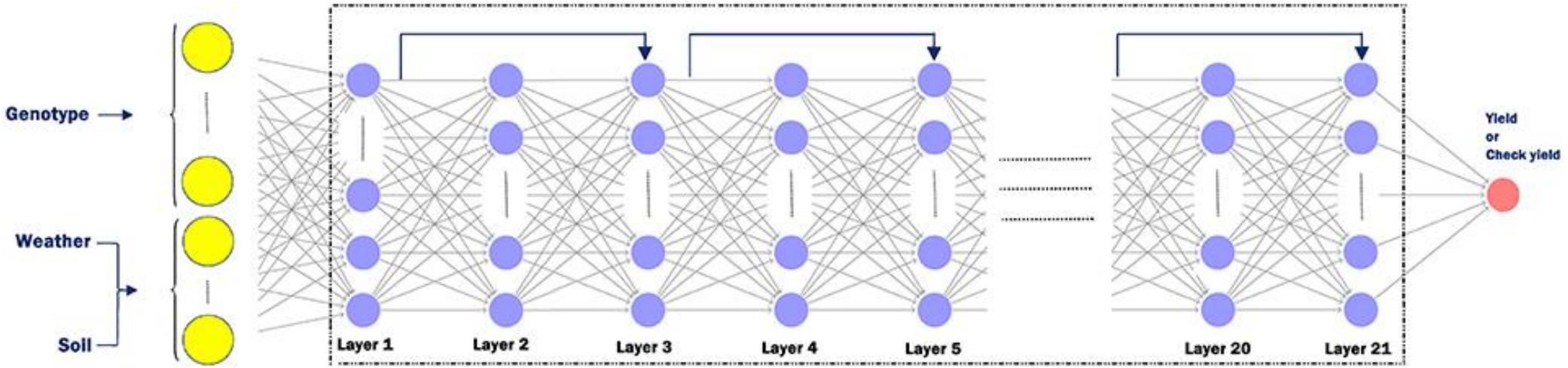
- Non-linear responses of genotypes to environmental stresses
- Variability among genotypes responses (GxE)



Why applying Machine Learning on multi-environment trials datasets?

Machine Learning → model flexibility
to link phenotypes to genomic and environmental features

→ Data-driven predictions…



Khaki and Wang (2019)

… at the cost of biological interpretability (black-box models) ?

Today's presentation

- Proposed approach to integrate genomic and environmental covariates in a pooled dataset

- First evaluation of the most determinant factors affecting maize grain yield and plant height with one machine learning algorithm

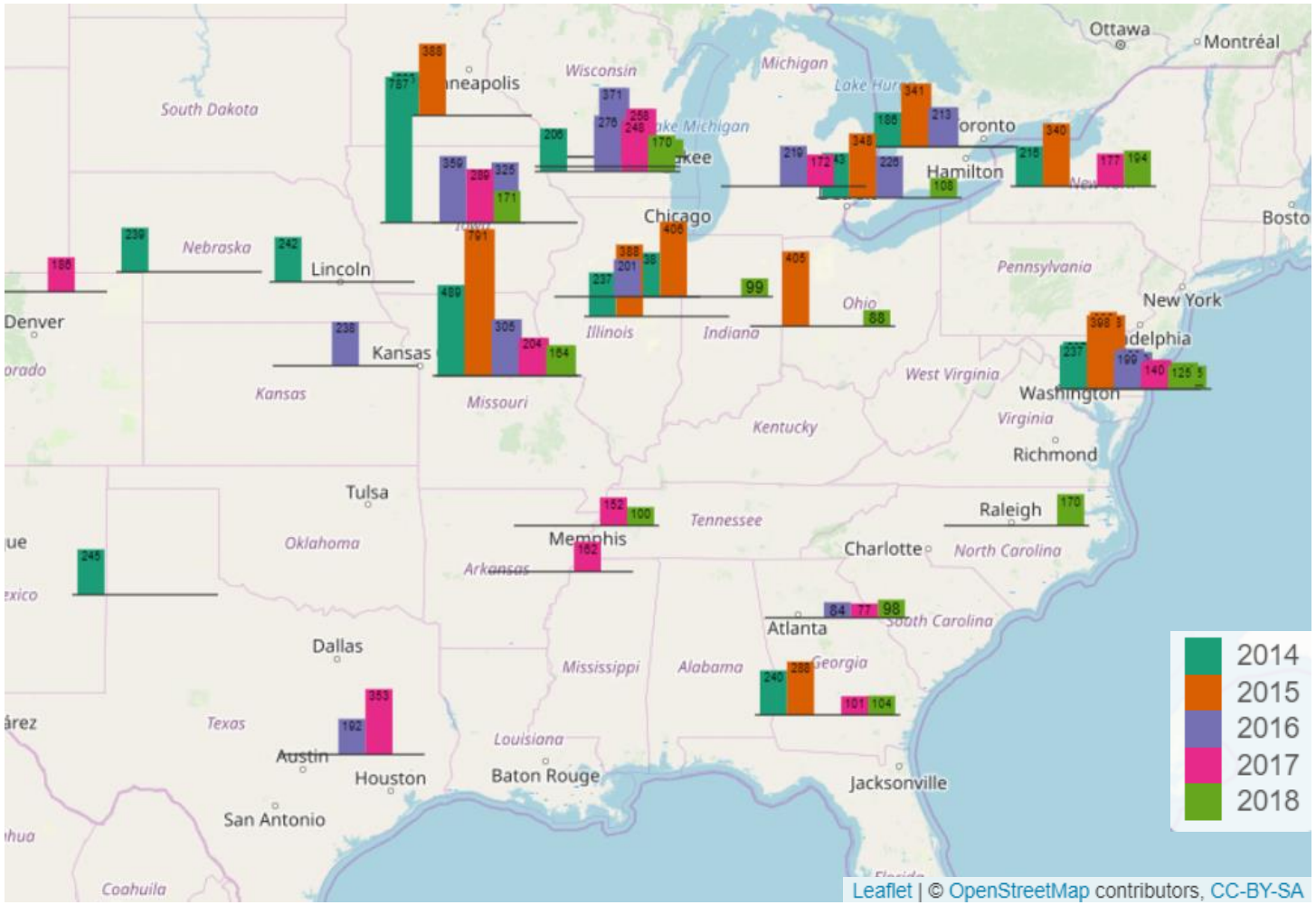
- First evaluation of the model performance (on the whole dataset) with the input features

Data preparation

Hybrid phenotypic observations (2014-2018) & hybrid genotype data

- About 15,500 phenotypic observations (5 years, 26 counties) remaining after quality control, and matching with genotype data
- Plots(trials removed (disease, missing information)

- GBS genotypic data from inbred lines used
  - > 900,000 SNPs initially
  - Filtering: ~ 255,000 SNPs on resulting synthetic hybrid genotype matrix



Processing of weather data (2014-2018)

Step 1: obtain a daily weather dataset for each station

Missing values often need to be imputed for ML algorithms (R Package '*nasapower*', Sparks 2018)

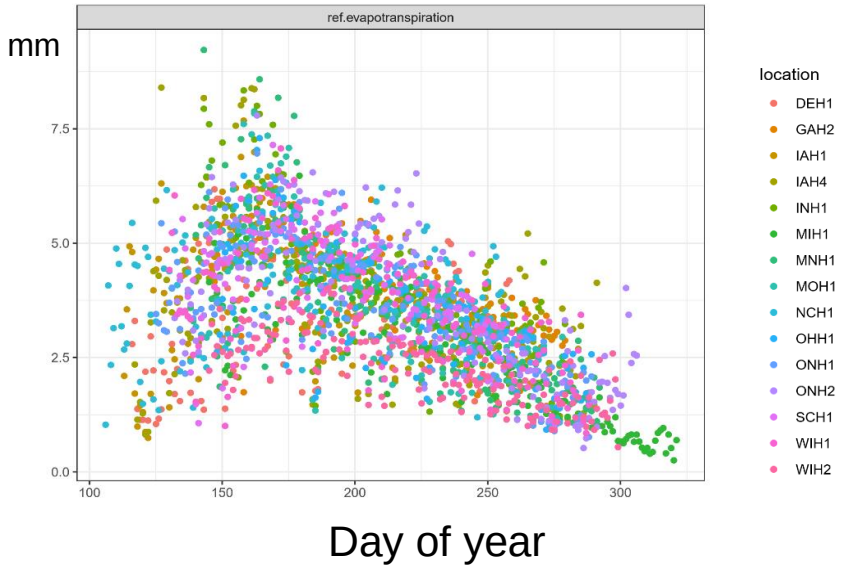
Weather station in Goettingen

Step 2: add irrigation data when available

Step 3: Feature engineering: creating new input predictors from existing ones

→ Low-level sensory data (e.g. relative humidity) might be more meaningful after transformation

Daily reference evapotranspiration (ETo) calculated according to standard FAO-56 Penman and Monteith definition



Environmental covariates derived for each hybrid per estimated growth stage

Hybrid-dependent environmental covariates: 3 growth stages estimated using recorded flowering date and sowing date (non-overlapping)

Vegetative stage (V): 1 week after sowing date to FD1

Flowering and pollination growth stage (F): From FD1 to FD2

Grain fill growth stage (G): From FD2 to ~ 55 days after

Picture source: Cathy Jubin, Barış Alaca

FD1: 5 days before 50% recorded silking date (reference basis of a 130 days growth period)
FD2: 10 days after 50% recorded silking date (reference basis of a 130 days growth period)

Weather information restricted to the growing season (pre-sowing information absent)

Environmental covariates derived per hybrid per growth stage

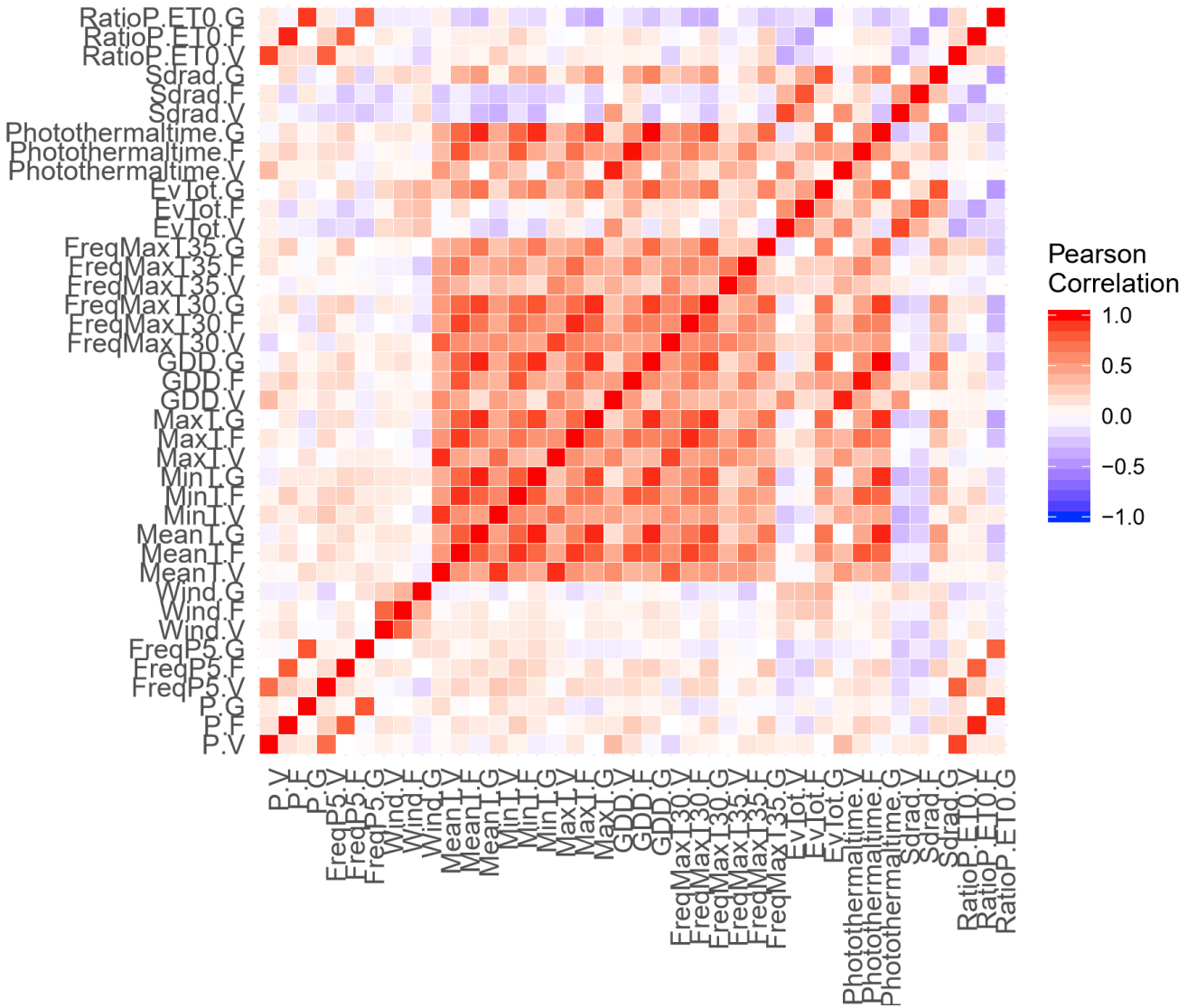
Acronym	General description
P	Accumulated precipitation (mm)
FreqP5	Frequency of days with more than 5 mm precipitation
MeanT	Average of daily mean temperature (°C)
MinT	Average of minimum daily temperature (°C)
MaxT	Average of maximum daily temperature (°C)
GDD	Cumulative growing degree days Base 10 (°C)
FreqMaxT30	Frequency of days with maximum temperature above 30°C
FreqMaxT35	Frequency of days with maximum temperature above 35°C
EvTot	Reference evapotranspiration ET_0 according to Penman-Monteith (total amount of evaporated water from the hypothetical grass reference surface. (mm)
RatioP.ET0	Ratio of total rainfall to total Penman-Monteith reference evapotranspiration (P/ET_0) in the growing stage - Aridity index.
Photothermal.Time	Sum of [daily hours of photoperiod (daylength) * daily GDD]
WindS	Average wind speed at 2 meters (m/s)
Sdrad	Accumulated daily solar radiation (MJ m^{-2} day^{-1})

Vegetative Stage
Flowering Stage
Grain Filling Stage

Growing season length included

→ Characterize each environment more accurately
→ Account for the plant growth stage

Correlations among computed environmental covariates



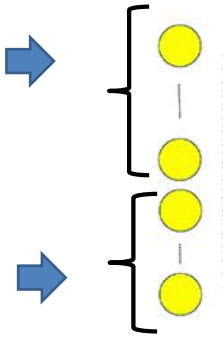
Pearson's coefficients of correlation within and across environmental covariates derived per growth stage.

→ Covariates derived from temperature information correlated

→ Collinearity: issue for predictions with machine learning ?

Input features for Machine Learning- based regression

Hybrid genotype matrix



Acronym	General description
P	Accumulated precipitation (mm)
FreqP5	Frequency of days with more than 5 mm precipitation
MeanT	Average of daily mean temperature (°C)
MinT	Average of minimum daily temperature (°C)
MaxT	Average of maximum daily temperature (°C)
GDD	Cumulative growing degree days Base 10 (°C)
FreqMaxT30	Frequency of days with maximum temperature above 30°C
FreqMaxT35	Frequency of days with maximum temperature above 35°C
EvTot	Reference evapotranspiration ET_0 according to Penman-Monteith (total amount of evaporated water from the hypothetical grass reference surface. (mm)
RatioP.ET0	Ratio of total rainfall to total Penman-Monteith reference evapotranspiration (P/ET_0) in the growing stage - Aridity index.
Photothermal.Time	Sum of [daily hours of photoperiod (daylength) * daily GDD]
WindS	Average wind speed at 2 meters (m/s)
Sdrad	Accumulated daily solar radiation (MJ m^{-2} day^{-1})

Machine Learning algorithm

Phenotypic observations
(grain yield, plant height)

Machine learning to assess determinant predictors (preliminary results)

Handling predictors before implementing prediction scenarios

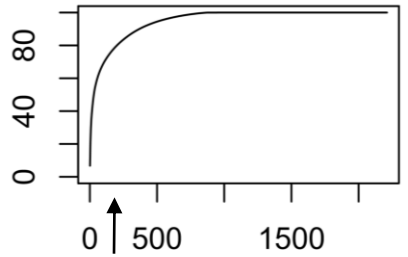
Overfitting problem: large number of predictors, with high collinearity

'large p, small n'

Feature extraction on hybrid marker matrix

- Singular Value Decomposition of the standardized hybrid genotype matrix (same approach as Ehret *et al*. 2014): $\boldsymbol{X = UDV'}$

Cumulative variability explained by each column



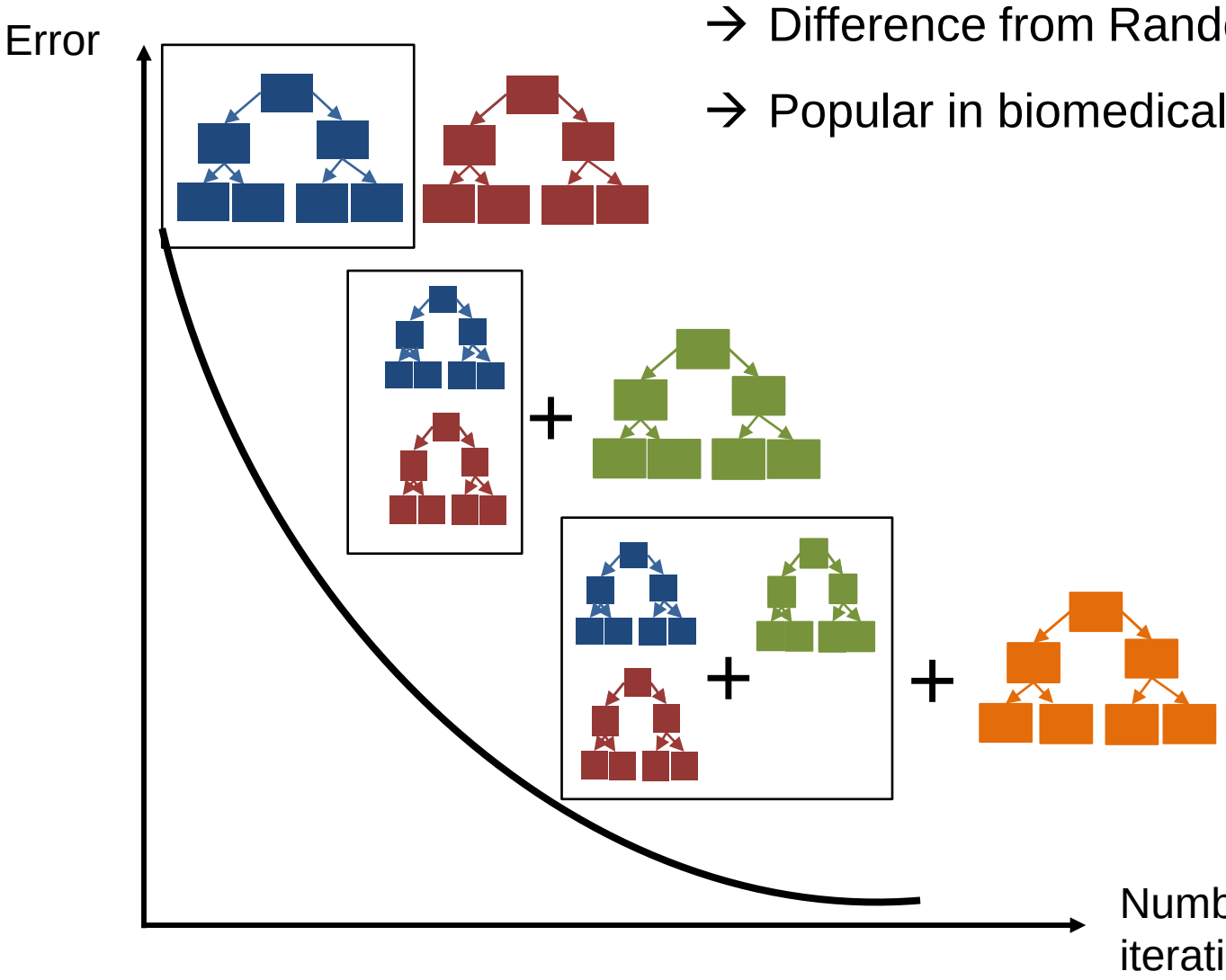
800 first scores accounting for 94.5% of the variability

→ Top 800 principal component scores used as input features

Feature selection (FS) for environmental predictors

- Assessment of variable importance on the whole dataset

Gradient Boosted Trees to evaluate variable importance



→ Difference from Random Forest: all trees are built independently

→ Popular in biomedical studies

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 August 2019
doi: 10.3389/fgene.2019.00735

Identification of Cancer-Related Long Non-Coding RNAs Using XGBoost With High Accuracy

Xuan Zhang[1,2†], Tianjun Li[3†], Jun Wang[4†], Jing Li[1], Long Chen[3] and Changning Liu[1*]*

[1] CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming, China, [2] University of Chinese Academy of Sciences, Beijing, China, [3] Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China, [4] Institute of Medical Sciences, Xiangya Hospital, Central South University, Changsha, China

Dinh *et al. BMC Medical Informatics and Decision Making* (2019) 19:211
https://doi.org/s12911-019-0918-5

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

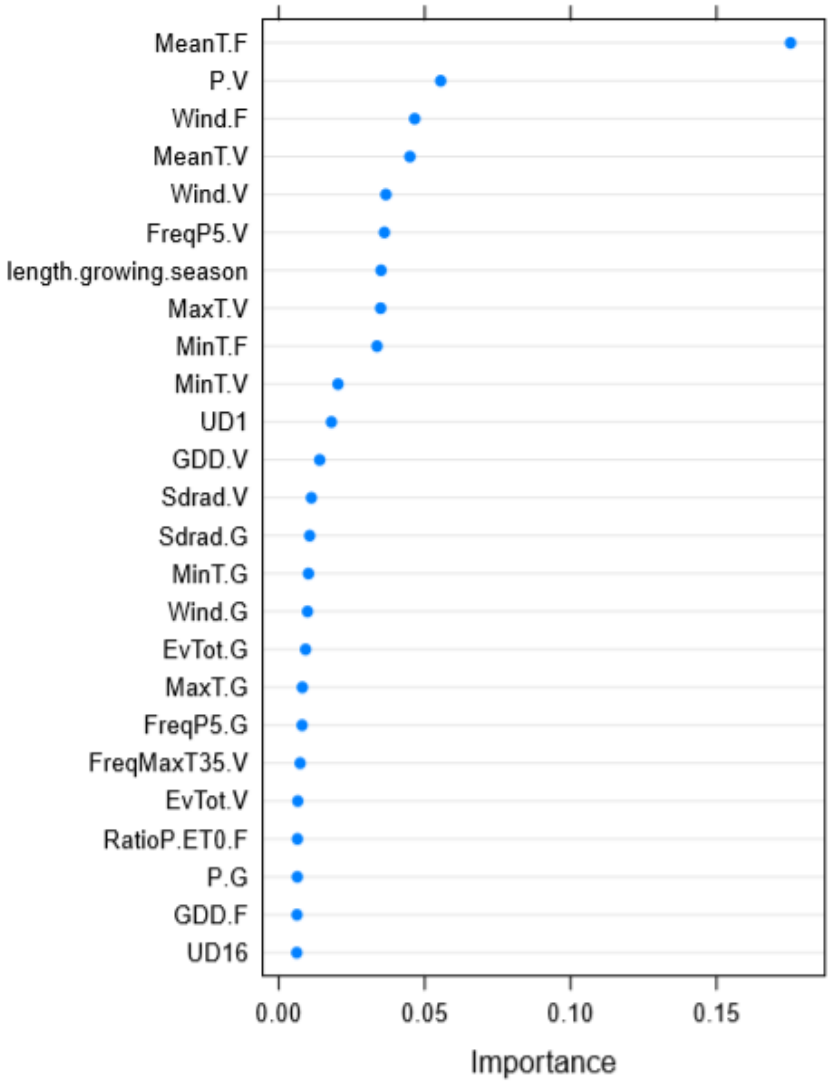
A data-driven approach to predicting diabetes and cardiovascular disease with machine learning

An Dinh[1†], Stacey Miertschin[2†], Amber Young[3†] and Somya D. Mohanty[4*]

Quantification of the variable importance (VI) – Grain yield

Top 25 variables shown for model including all covariates.
Suffixes refer to: V: vegetative stage ; F: flowering stage ; G : grain fill stage

UD: principal component scores

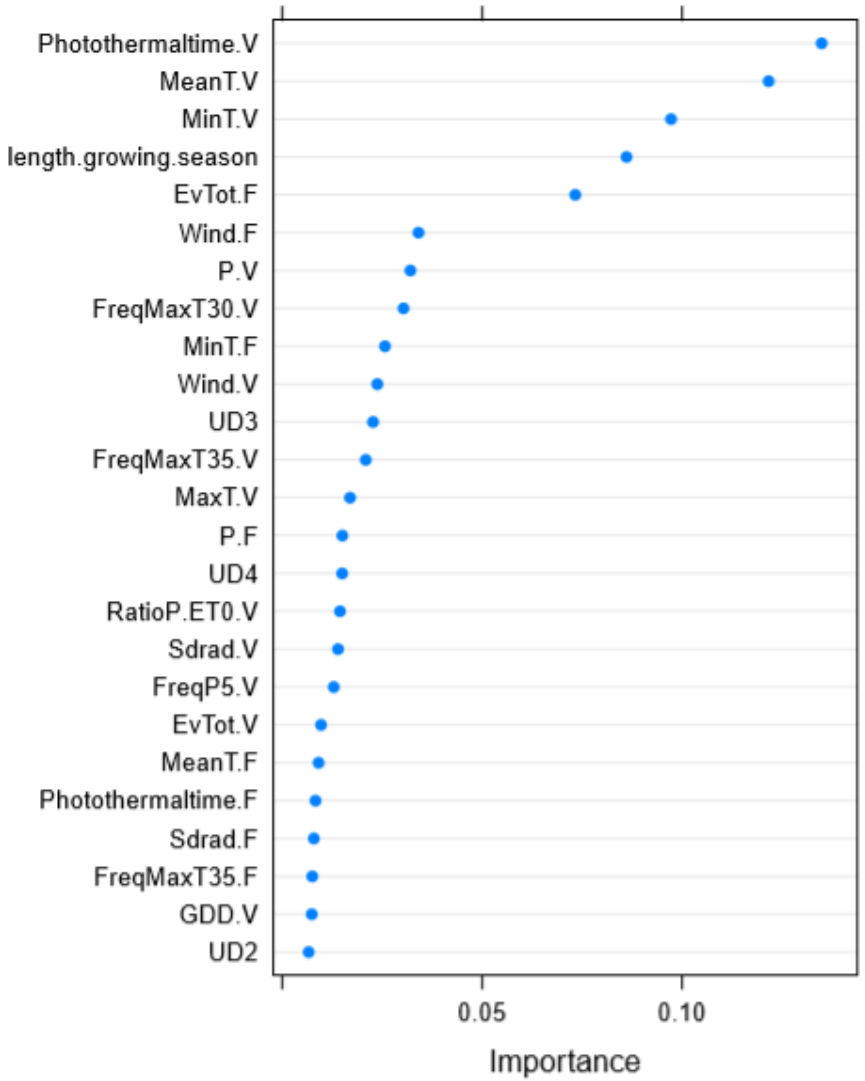


Based on gain (improvement in accuracy brought by a variable to the branches it is on)

1000 boosting iterations with xgbTree method, with 5-fold cross-validation on the whole dataset

Number of predictors: 840
Root mean square error: 23.31
R^2: 0.69
Mean absolute error: 17.72

Quantification of the variable importance (VI) – Plant height



Top 25 variables shown for model including all covariates (grain fill stage excluded)

Suffixes refer to: V: vegetative stage ; F: flowering stage ; G : grain fill stage

UD: principal component scores

Number of predictors: 827
Root mean square error: 13.16
R^2: 0.881
Mean absolute error: 9.85

Model performance obtained with different sets of input variables

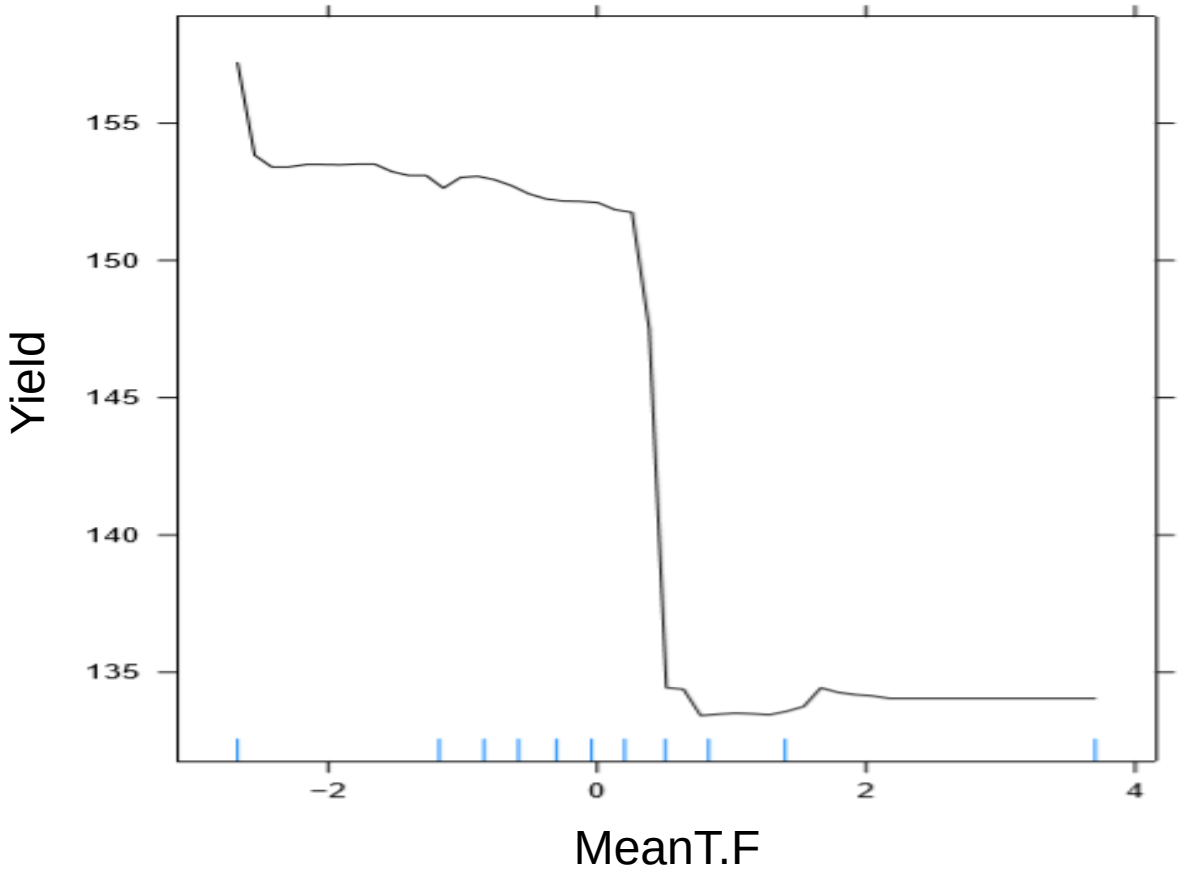
Trait	Input variables	Root mean square error	R^2	Mean absolute error
Grain yield (1)	ECs, PCs	23.31	0.692	17.72
Grain yield (2)	Y, L, PCs	26.54	0.603	19.90
Grain yield (3)	ECs	27.32	0.578	20.90
Grain yield (4)	PCs	43.30	0.0465	34.42
Plant height (5)	ECs, PCs	13.16	0.881	9.85
Plant height (6)	Y, L, PCs	13.66	0.872	10.20
Plant height (7)	ECs	17.02	0.802	13.03
Plant height (8)	PCs	37.59	0.091	27.84

EC: all environmental covariates ; L: location (indicated as county) ; Y: year
PCs: principal component scores based on SNPs matrix
Model assessed with 5-fold cross-validation

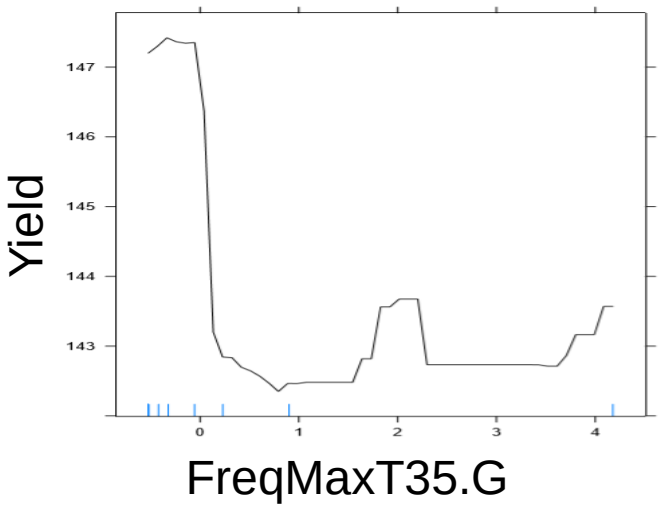
Partial dependence plots: partial relationships of the predictors to grain yield

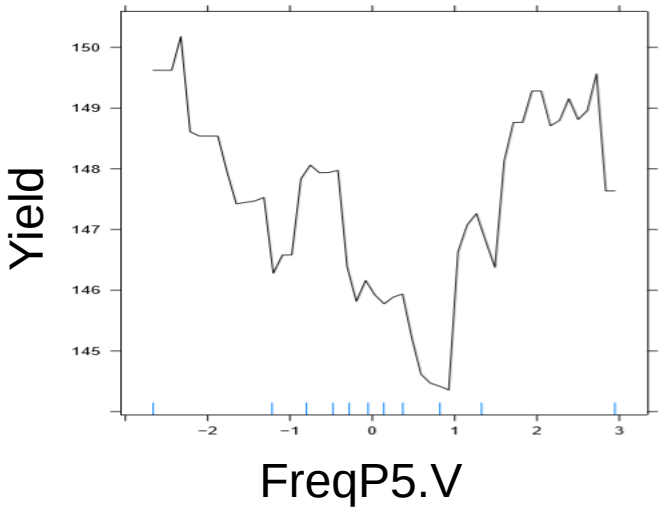
Shows the marginal effect a feature has on the predicted outcome

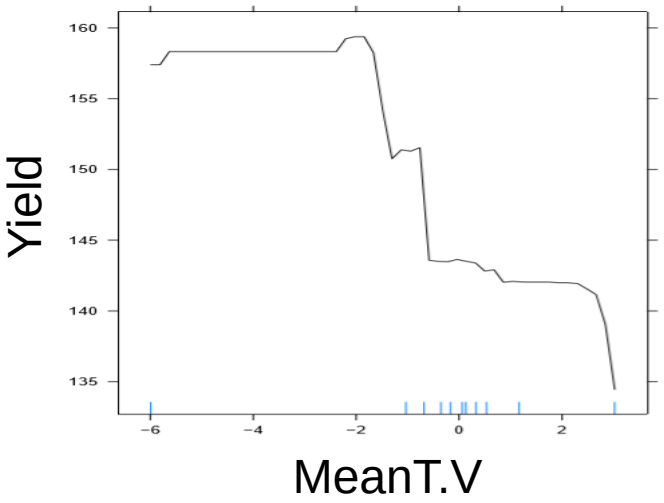
Negative influence of temperature on the output variable



Other predictors relationships with the target variable





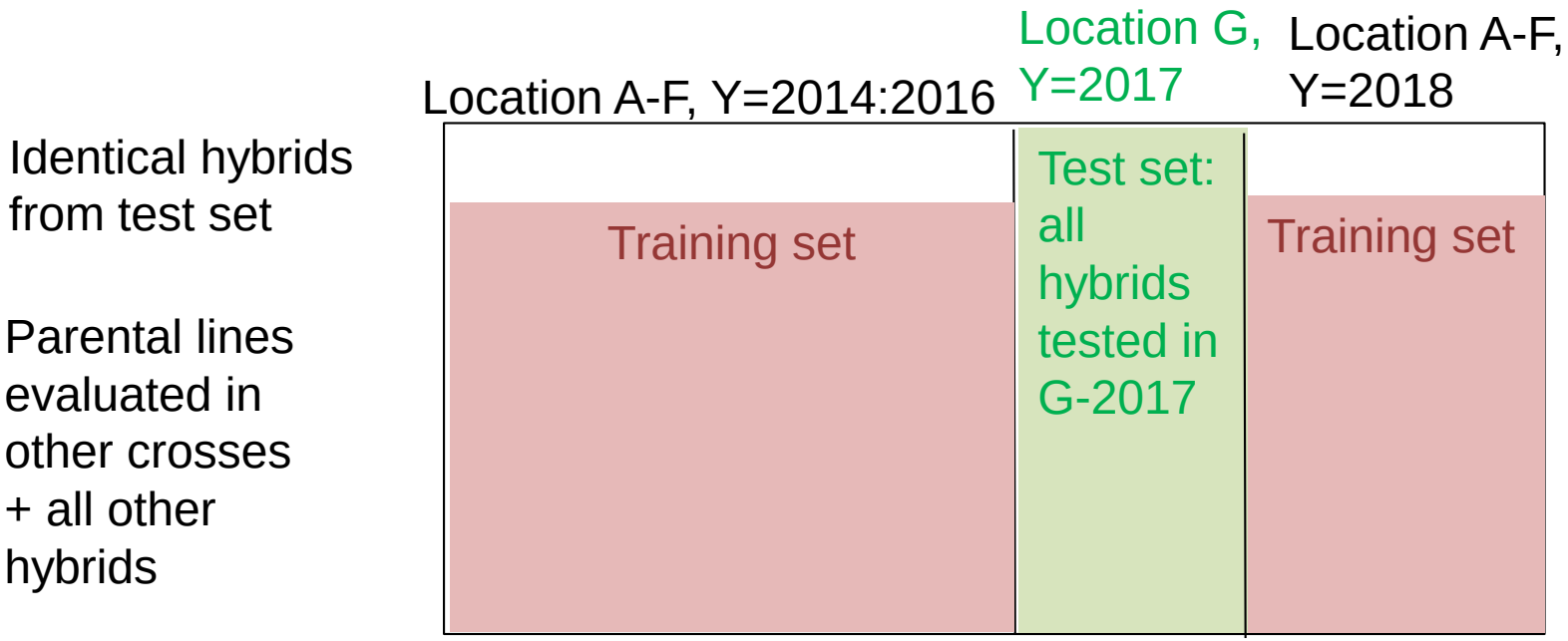


Weather covariates standardized

Outlook: What type of plant breeding scenarios to predict ?

Predicting new hybrids (never assessed in any other environment): T0,T1 and T2 hybrids in a new year and in a new location (county).

Environments



Identical hybrids present both in training and testing set removed from training set.

Discussion

- **Environmental covariates restricted here to weather information (soil, agronomic practices, pests or pre-sowing information not included)**
- **Quality of weather-based variables (weather station, imputation of missing values)**
- **Precision of definition of growth stages**
- **Model performance dependent on hyperparameter optimization and on the type of algorithm**
- **Pooling all data together: advantageous or not for prediction of specific environments ?**

Take-home messages

- **Here, climatic factors alone explained more of the phenotypic trait variability than the principal component scores derived from SNPs data**
- **Weather-based covariates related to temperature and heat stresses were the most critical for grain yield determination**
- **Machine learning algorithms can help us understand the relationships (linear and non-linear) between predictors and the target variables**

Acknowledgements

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

PhD supervisors:

Timothy M. Beissinger

Henner Simianer

Wolfgang Link
Stefanie Griebel
Medhat Mahmoud
Barış Alaca
Reimund Rötter
Patrick Thorwarth
Jan-Christof Richter
Gregory Mahone

the Genomes to Fields initiative

We thank the G2F Consortium for making data publicly available, sharing them among collaborators and for their help with this study.

KWS

We are grateful to KWS SAAT SE & Co. KGaA for financial support of the PhD thesis.

GWDG Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Thank you for your attention !

Poster 209