

Sommertagung 2026 der Arbeitsgruppe Landwirtschaftliches Versuchswesen

# Balancing Stability and Efficiency in Random Forests for Genomic Selection

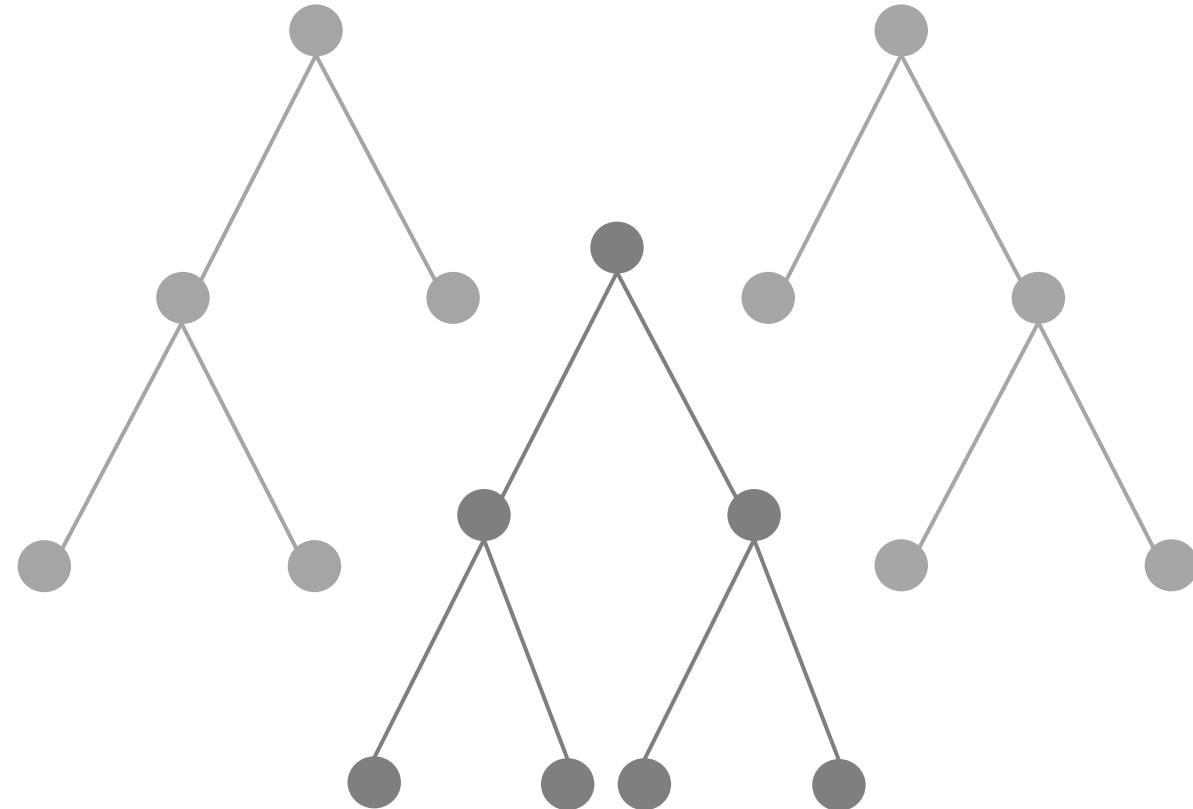
Dr. Thomas Martin Lange  
Postdoctoral researcher  
Breeding Informatics  
Georg-August University of Göttingen

## Genomic Selection and Random Forest

- Genomic selection uses genotypic data to predict phenotypes and select top-performing individuals
- Random forest is well-suited for genomic selection
  - Non-parametric (Montesinos López et al., 2022) and able to handle the large  $p$ , small  $n$  problem (Jannink et al., 2010)
  - Integrates SNP interactions into the prediction model (Lunetta et al., 2004)
- Particularly promising for predicting disease resistance
  - Often non-normal data distribution (Lange et al., 2021)
  - Phenotype might be affected by a combination of major and minor genes (Merrick et al., 2021) and by gene interactions (González et al., 2012)

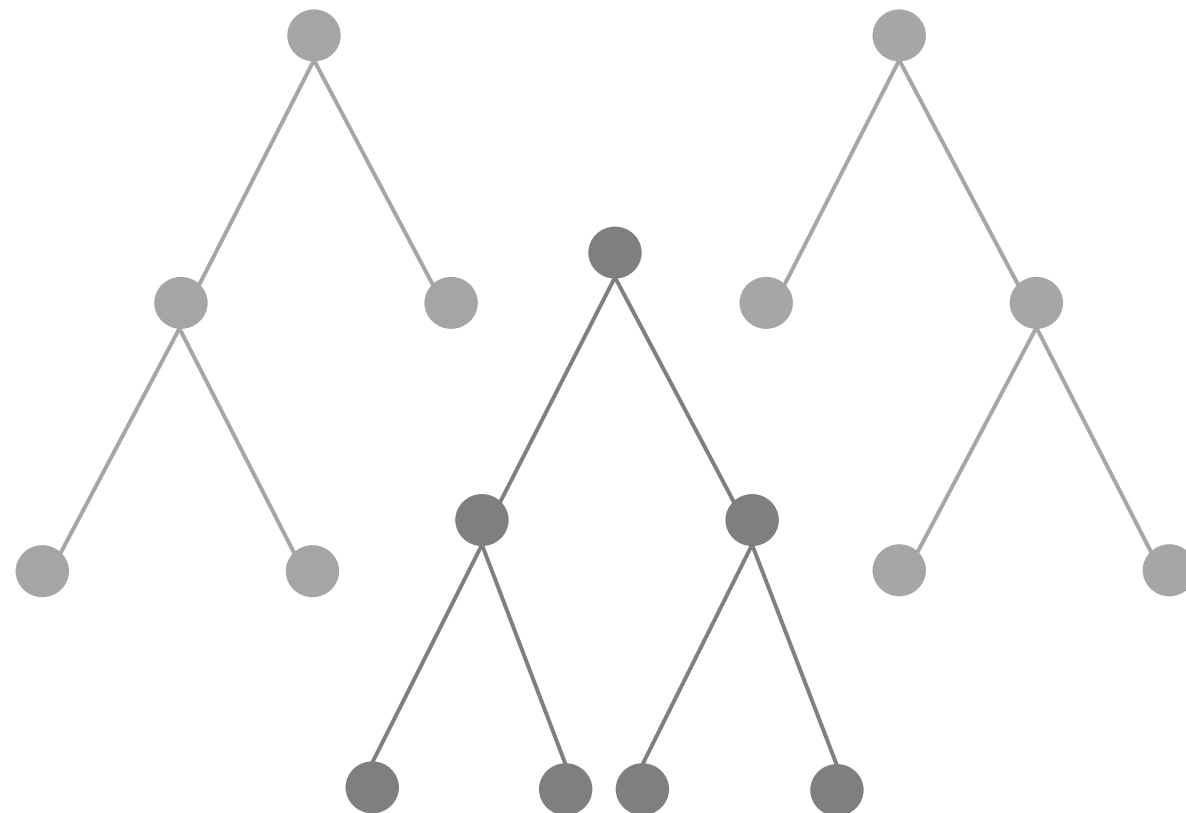
## How Does Random Forest Work?

- Random forest works by growing multiple decision trees
- The final random forest prediction is the average of the decision tree predictions
- To increase accuracy, randomness is introduced in the decision trees by variable selection and bootstrapping
- This randomness makes random forests naturally **non-deterministic**
- Predictions and decisions can change even with the **same input data** just by chance



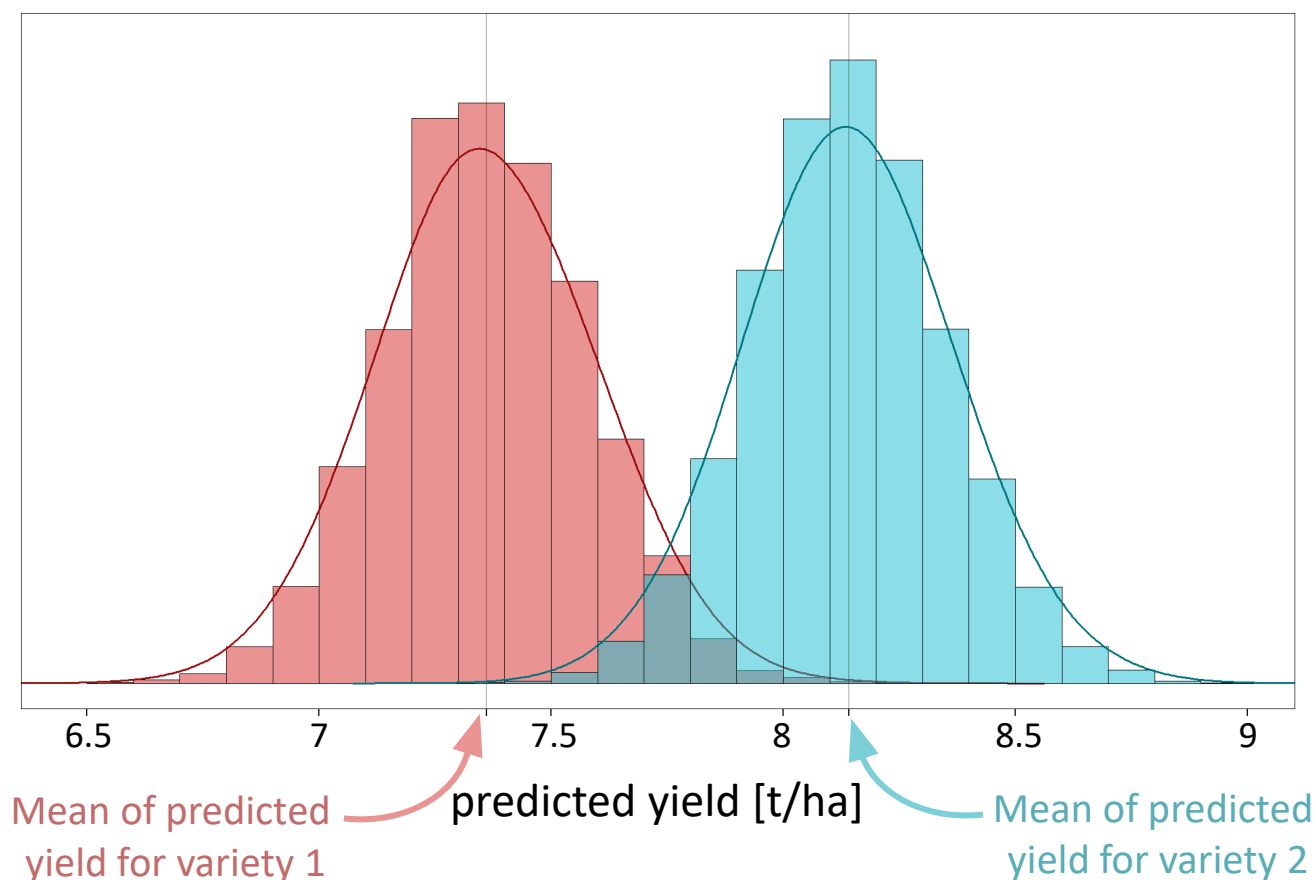
## How Does Random Forest Work?

- Random forest works by growing multiple decision trees
- The final random forest prediction is the average of the decision tree predictions
- To increase accuracy, randomness is introduced in the decision trees by variable selection and bootstrapping
- This randomness makes random forests naturally **non-deterministic**
  - Predictions and decisions can change even with the **same input data** just by chance



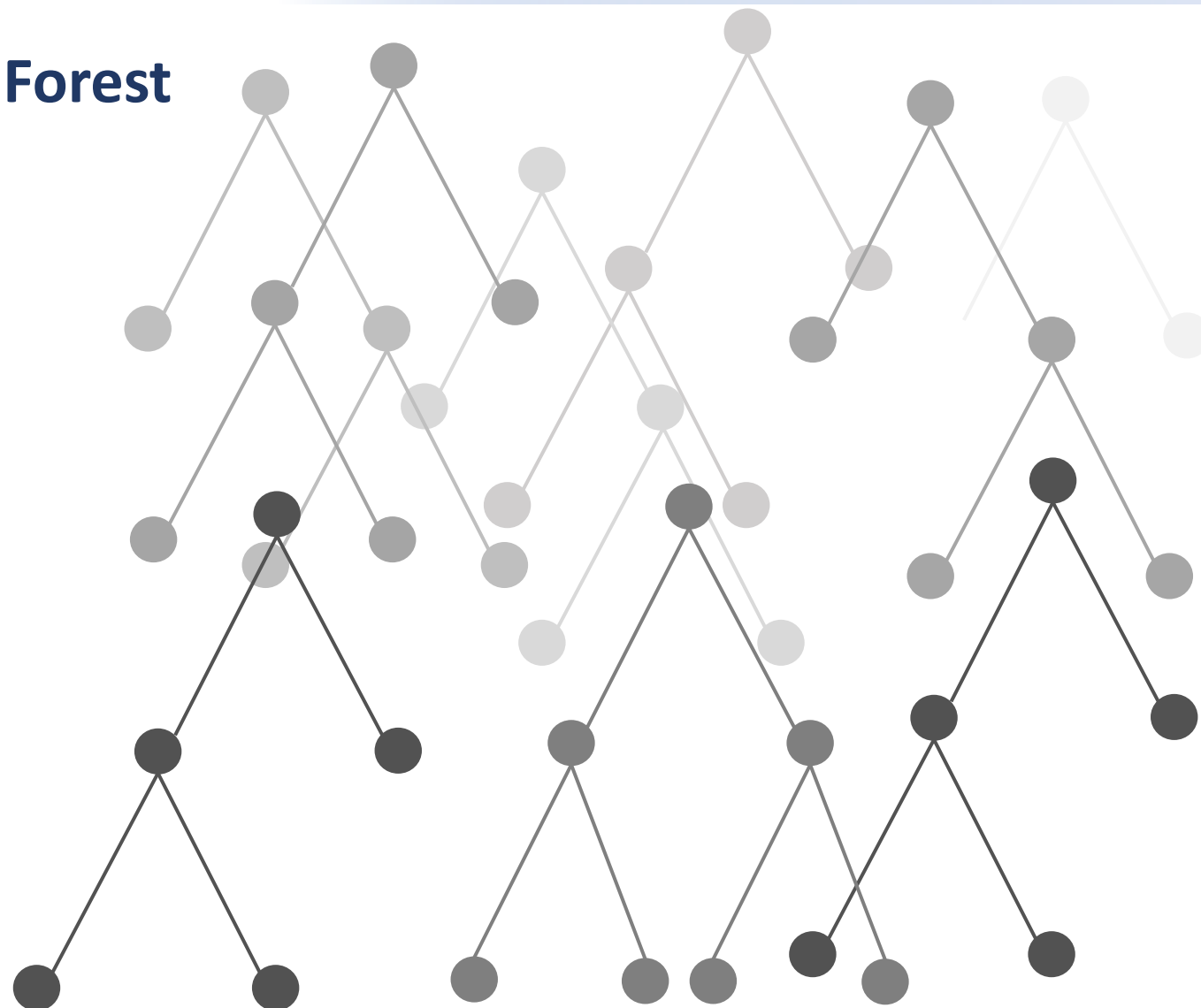
## How Non-Determinism Can Affect Selection Decisions

- For example: The predicted yield of two varieties using random forest 10,000 times
- The predictions are normally distributed around a mean prediction per variety
- The predictions show some overlap: It could be possible to make a wrong decision just by chance



## Increasing the Stability of Random Forest

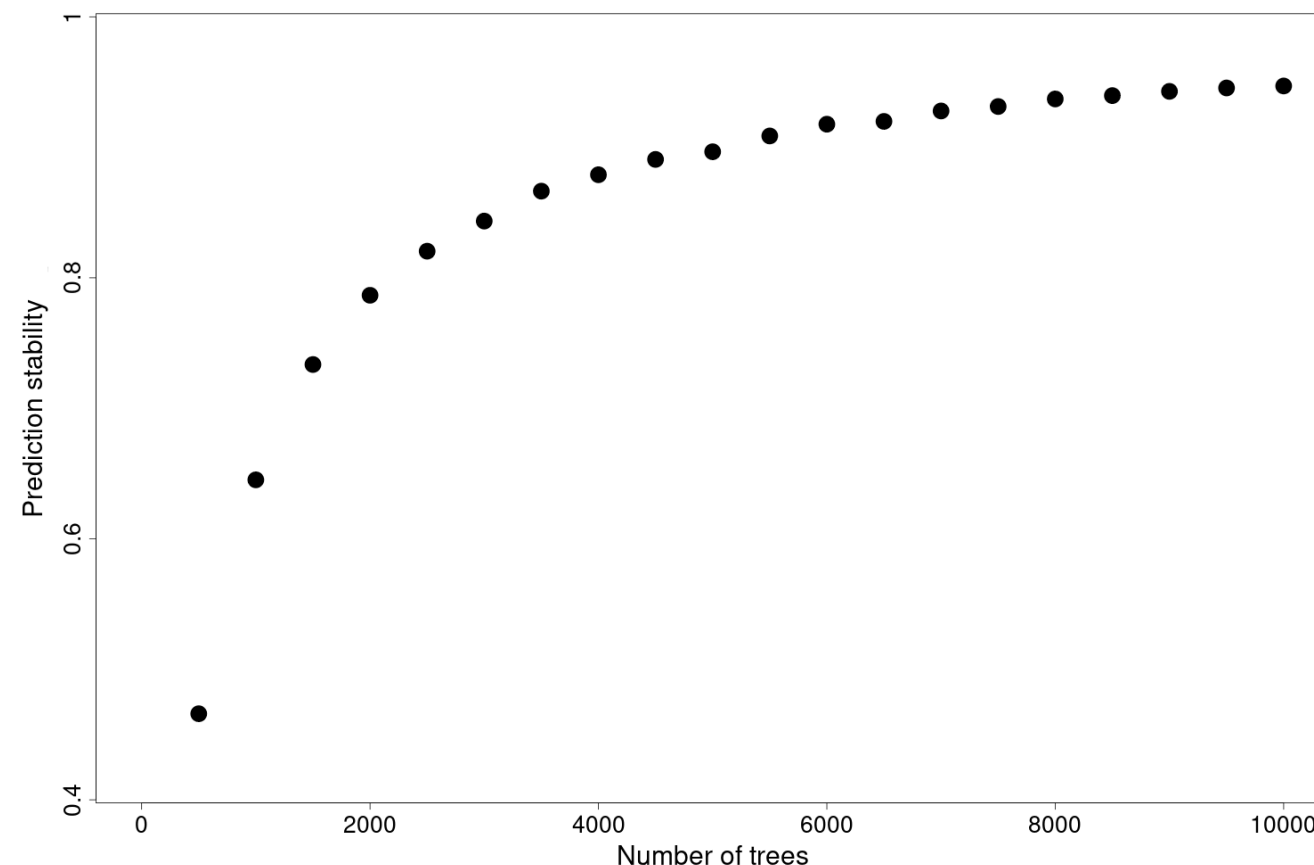
- More decision trees will lead to more stable predictions
- Disadvantage:
  - Growing a decision tree requires computation time
  - Growing 1,000 decision trees takes twice as long as growing 500 decision trees



## Relationship Between Stability and the Number of Trees

- The variability of the predictions of a non-deterministic model can be measured as the *prediction stability* by repeating the model fitting process
- The prediction stability ( $s$ ) increases non-linearly with higher number of trees ( $t$ )
- This relationship can be modelled using a two parameter logistic (2PL) regression model:

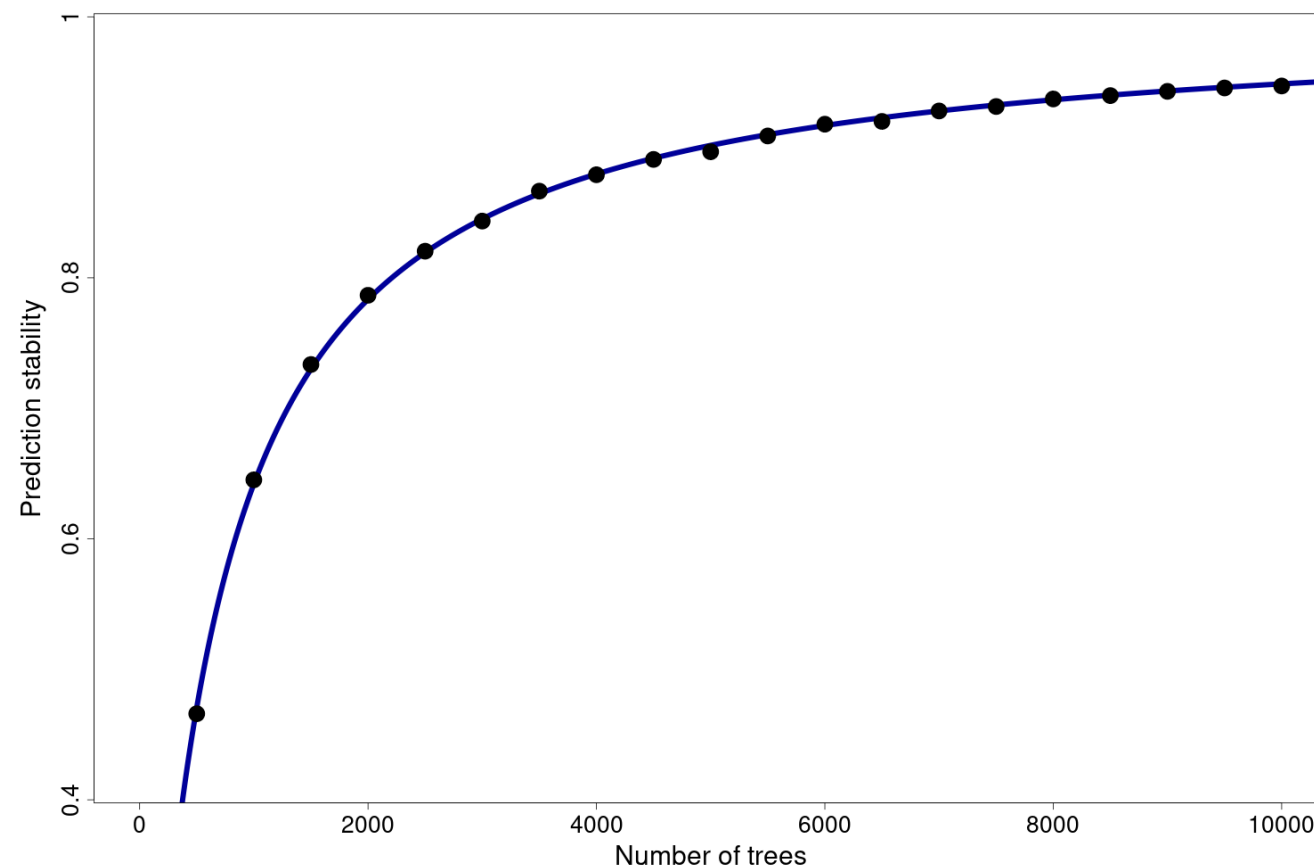
$$\hat{s}(t) = \frac{1}{1 + \left(\frac{\theta_1}{t}\right)^{\theta_2}}$$



## Relationship Between Stability and the Number of Trees

- The variability of the predictions of a non-deterministic model can be measured as the *prediction stability* by repeating the model fitting process
- The prediction stability ( $s$ ) increases non-linearly with higher number of trees ( $t$ )
- This relationship can be modelled using a two parameter logistic (2PL) regression model:

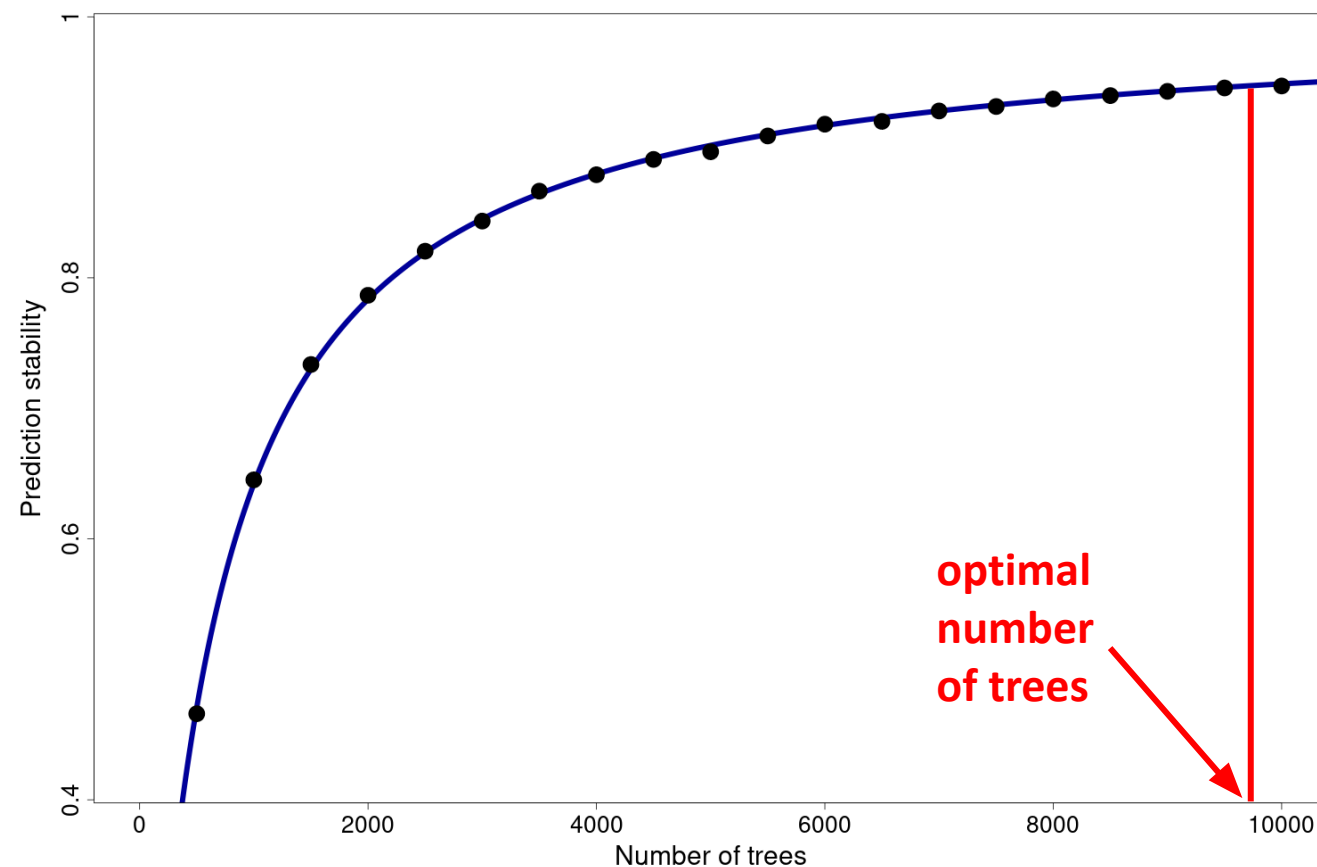
$$\hat{s}(t) = \frac{1}{1 + \left(\frac{\theta_1}{t}\right)^{\theta_2}}$$





## The Optimal Number of Trees

- With the 2PL model, the increase of prediction stability for each number of trees can be determined
- The optimal number of trees is where further trees lead to minimal gains in prediction stability ( $\leq 10^{-6}$ )
- But the optimal number of trees is data set dependent and is affected by
  - Number of genomic markers
  - Number of observations in the training data
  - Genetic architecture



## The optRF Package

- **optRF** is available on CRAN
- Workflow:
  - Calculates the prediction stability for small numbers of trees
  - Derives the 2PL model
  - Determines the optimal number of trees for your specific data set
- Output:
  - The optimal number of trees
  - Estimated prediction stability
  - Estimated computation time for the **ranger** function on the same machine

```
> install.packages(ranger)
> install.packages(optRF)

> library(ranger)
> library(optRF)

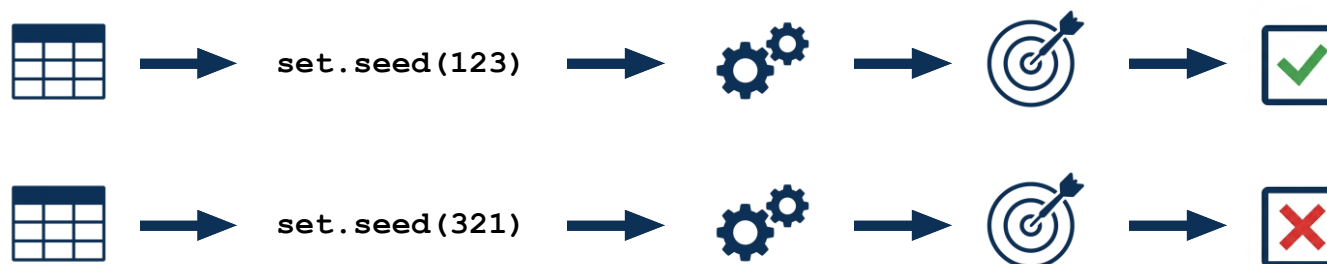
> optresult <- opt_prediction(y = Response,
+                             X = Predictors)
Recommended number of trees: 19000

> summary(optresult)
Recommended number of trees:      19000
Expected prediction stability: 0.978
Expected selection stability:  0.836
Expected computation time (s): 16.5

> RFmodel = ranger(y = Response,
+                  x = Predictors,
+                  num.trees = 19000)
```

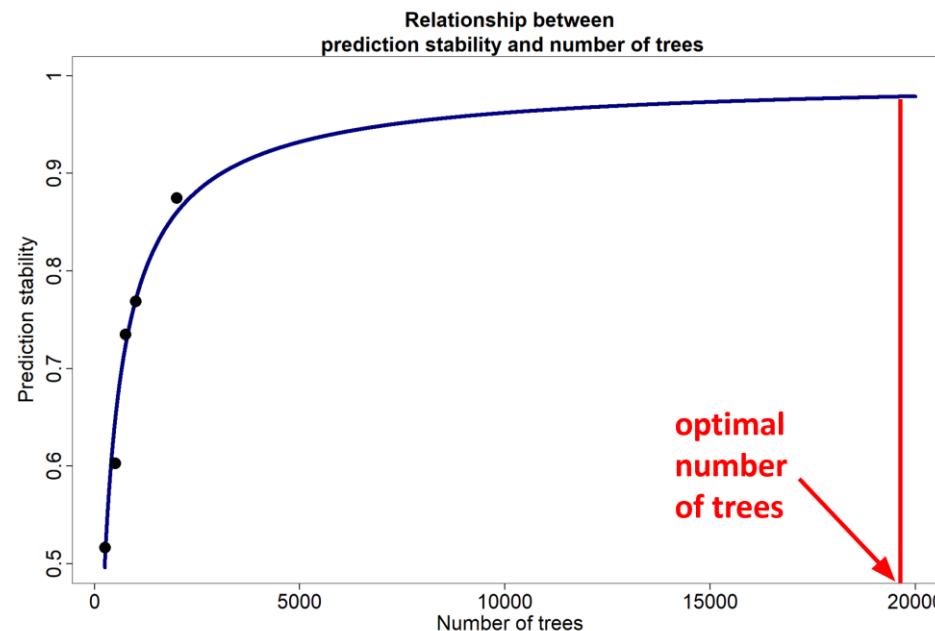
## Reproducibility is Not Stability

- One might be tempted to setting a seed to ensure reproducibility
- Setting a seed ensures that the same (quasi) random processes are repeatable but this does not eliminate the instability of the model
  - The risk of making a different decision when changing the seed still remains
- The goal of optRF is to provide a systematic way to reach stable predictions, regardless of the initial random state



## Take Home Messages

- Random forest is a non-deterministic prediction model
  - Predictions and decisions can change when repeating the analysis
- Increasing the number of trees increases stability and computation time
- The R package `optRF` determines the optimal number of trees for your research question and your data set
- While setting a seed ensures reproducibility of your predictions, it does not ensure stable predictions



## References

- Jannink, J. L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*, 9(2), 166-177.
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature.
- Lunetta, K. L., Hayward, L. B., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1), 32.
- Lange, T. M., Wutke, M., Bertram, L., Keunecke, H., Kopisch-Obuch, F., & Schmitt, A. O. (2021). Decision Strategies for Absorbance Readings from an Enzyme-Linked Immunosorbent Assay—A Case Study about Testing Genotypes of Sugar Beet (*Beta vulgaris* L.) for Resistance against *Beet Necrotic Yellow Vein Virus* (BNYVV). *Agriculture*, 11(10), 956.
- Merrick, L. F., Burke, A. B., Chen, X., & Carter, A. H. (2021). Breeding with major and minor genes: genomic selection for quantitative disease resistance. *Frontiers in plant science*, 12, 713667.
- González, A. M., Marcel, T. C., & Niks, R. E. (2012). Evidence for a minor gene–for–minor gene interaction explaining nonhypersensitive polygenic partial disease resistance. *Phytopathology*, 102(11), 1086-1093.

## Interested in More?

Find more information and resources at the project website

[www.uni-goettingen.de/en/optrf/707915.html](http://www.uni-goettingen.de/en/optrf/707915.html)



The original publication describing the methodology in detail



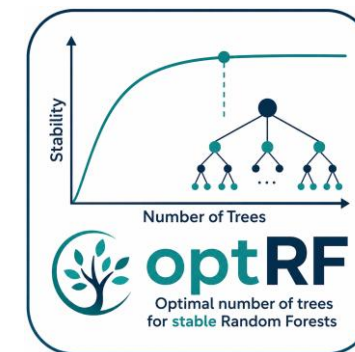
Links to the R package and the source code



A practical introduction to the R package



The slides of this presentation



Questions, feedback, collaboration ideas?



Dr. Thomas Martin Lange



Breeding Informatics Group  
Georg-August University of Göttingen



thomas.lange@uni-goettingen.de



**Thank you for the attention**