GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

CiBreed Center for Integrated Breeding Research

6th CEN Conference | Warsaw 2026
Power of Data - Shaping the Future of Life Sciences

From Data to Decisions

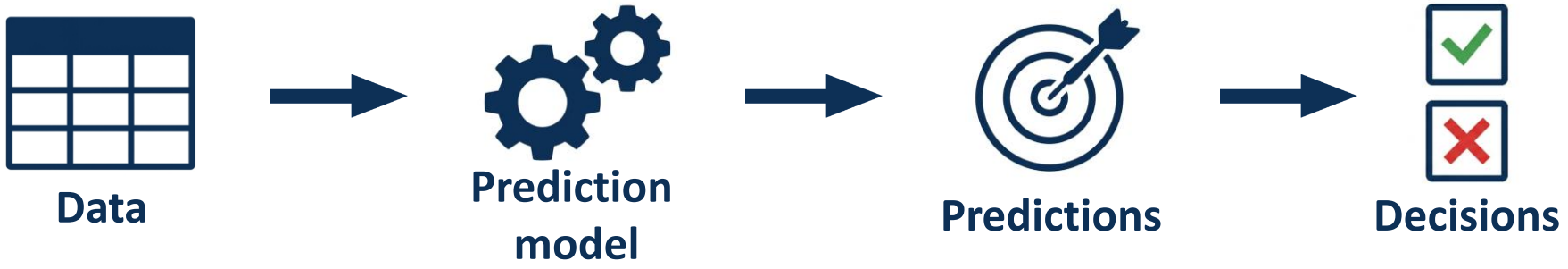
Enhancing the Reliability of Random Forest Predictions with optRF

Dr. Thomas Martin Lange

Postdoctoral researcher
Breeding Informatics
Georg-August University of Göttingen

Prediction-Based Decision-Making Processes

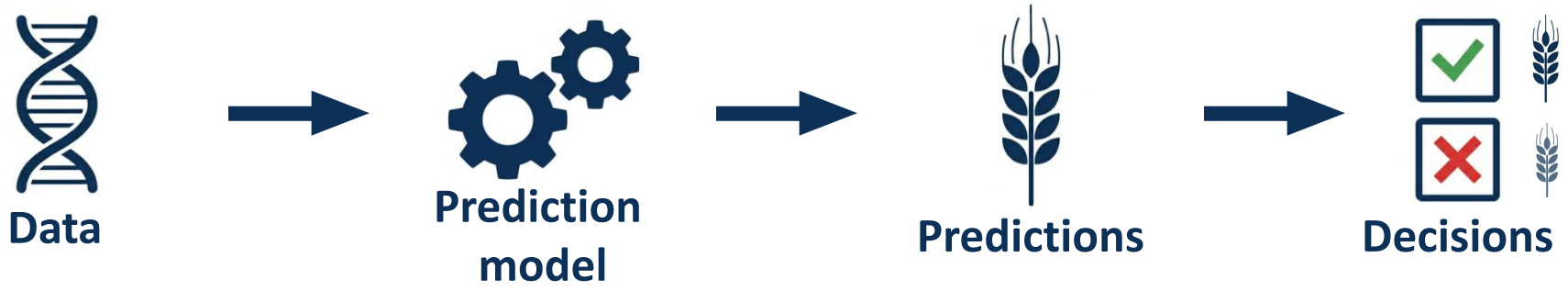
- Derive a prediction model that describes the mathematical relationship between predictors and the response
- With the prediction model, the response can be predicted for new data
- The predictions can be used to make decisions
- This workflow is applied in various fields from clinical diagnostics and drug development to environmental risk assessment



Genomic selection – A Special Case of Decision-Making

- Genomic selection is a special case of data-driven decision-making which follows the same workflow
- A prediction model is created that models the relationship between the genotype and a certain phenotype (e.g. the yield of wheat plants)
- Based on the predicted phenotype, the top-performing individuals can be early selected

→ This workflow necessitates a reliable prediction model to make accurate decisions



Random forest prediction models

- Random forest is a widely used prediction model from the field of machine learning
- Especially advantategous to predict complex responses
  - Non-parametric (Montesinos López et al., 2022)
  - Handles large, noisy data (Jannink et al., 2010)
  - Integrates non-additive interactions between predictors (Lange et al., 2023)
- Applications in various fields of biometrics

Precision medicine
Personalising treatment for gastrointestinal cancers (Mohammadi et al., 2024)

Risk assessment
Predicting the risk of diabetes (Wang et al., 2021)

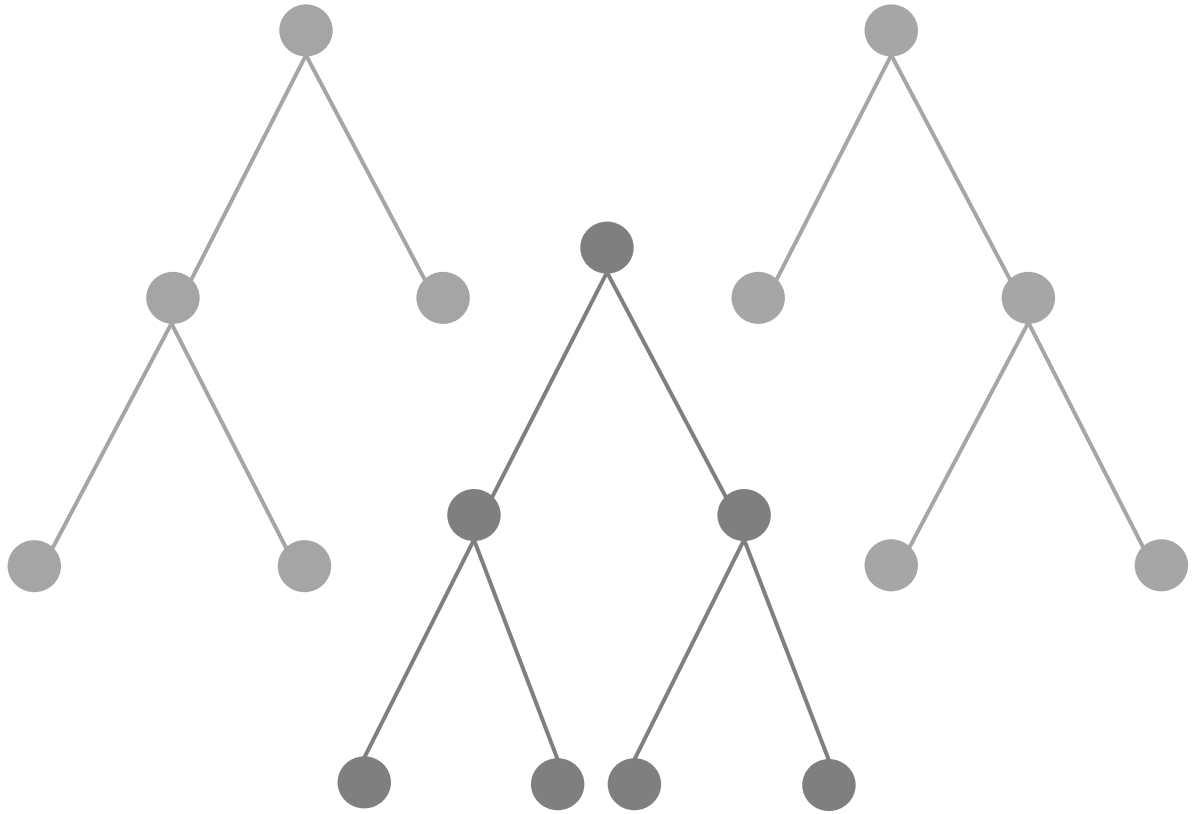
Ecology
Predicting nesting locations of birds (Cutler et al., 2007)

Drug development
Predicting adverse drug reactions (Roberts-Nuttall et al., 2026)

How does random forest work?

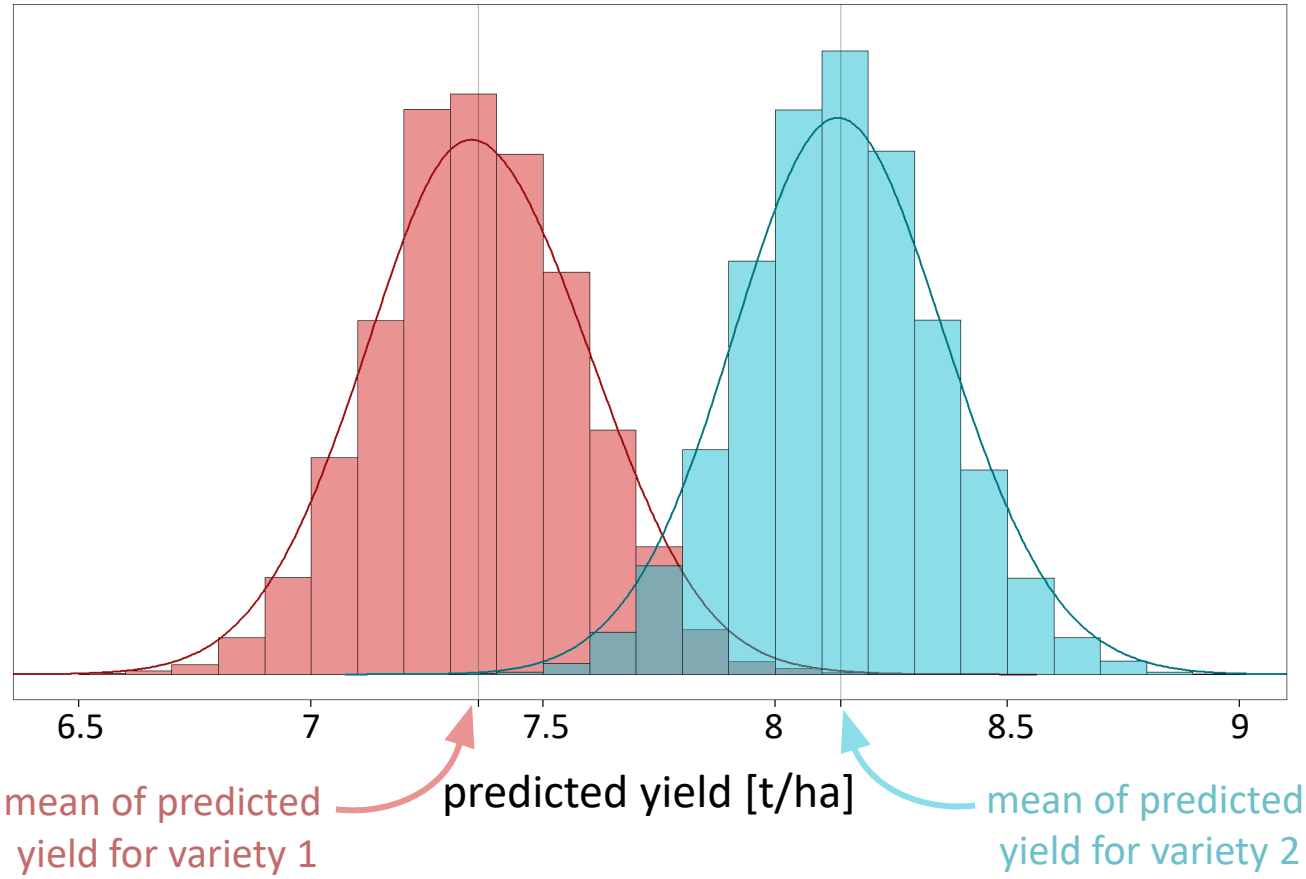
- Random forest works by growing multiple decision trees
- The final random forest prediction is the average of the decision tree predictions
- To increase accuracy, randomness is introduced in the decision trees by variable selection and bootstrapping
- This randomness makes random forests naturally **non-deterministic**

  → Predictions and decisions can change even with the **same input data** just by chance



How non-determinism can affect 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

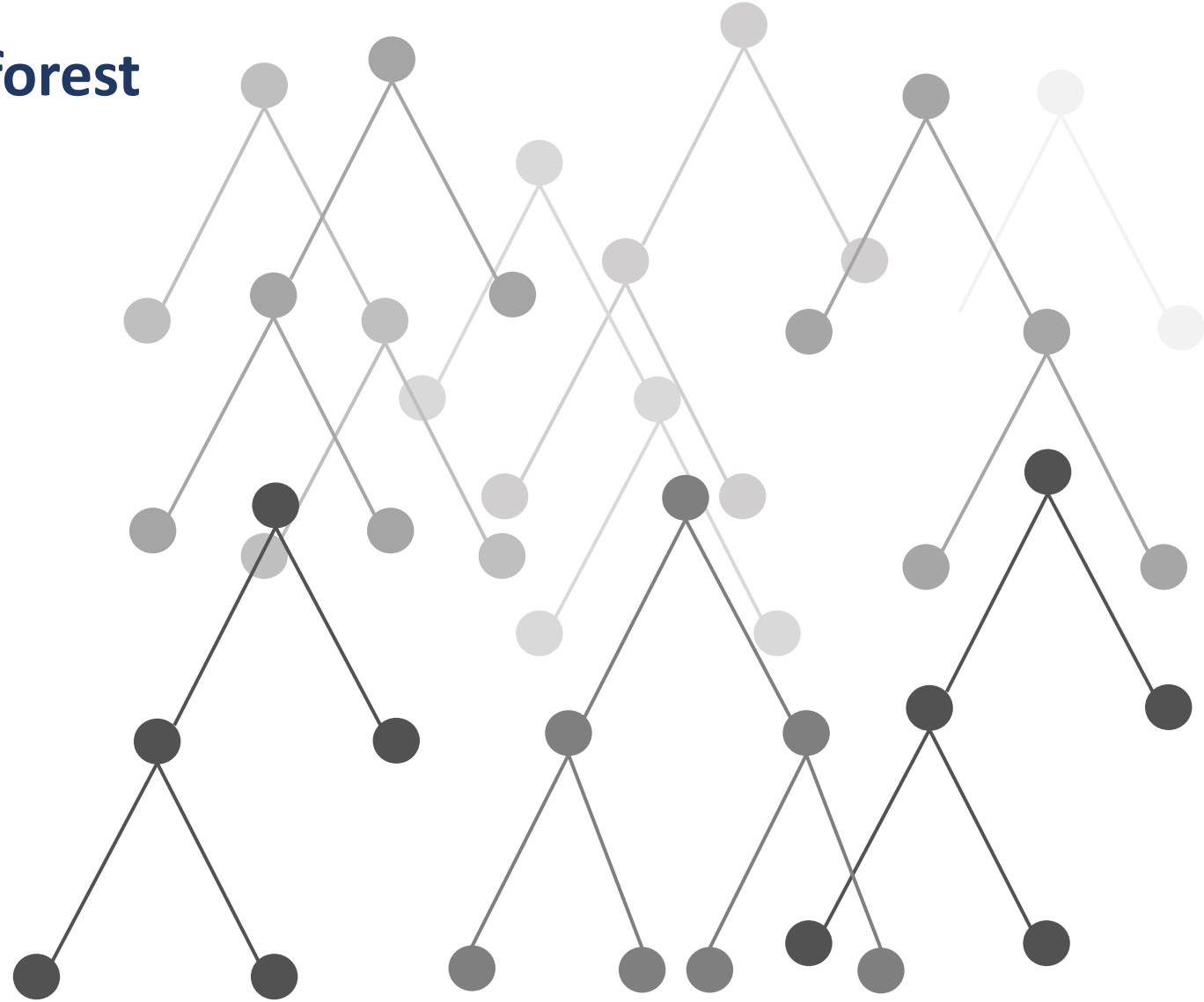


Increasing the stability of random forest

- More decision trees will lead to more stable predictions
- Disadvantage:

  Growing a decision tree requires computation time

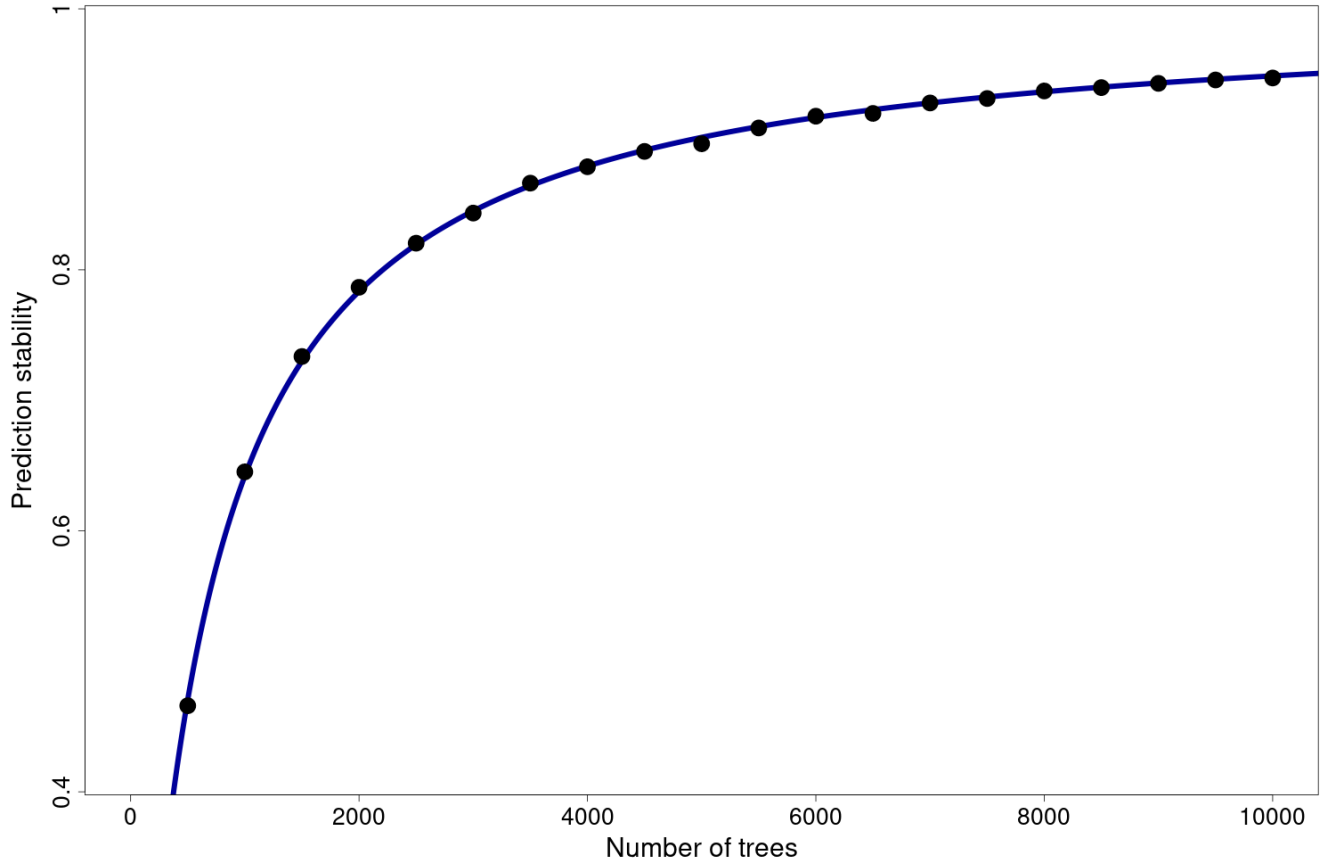
  → Growing 1,000 decision trees takes twice as long as growing 500 decision trees



Relationship between stability and the number of trees

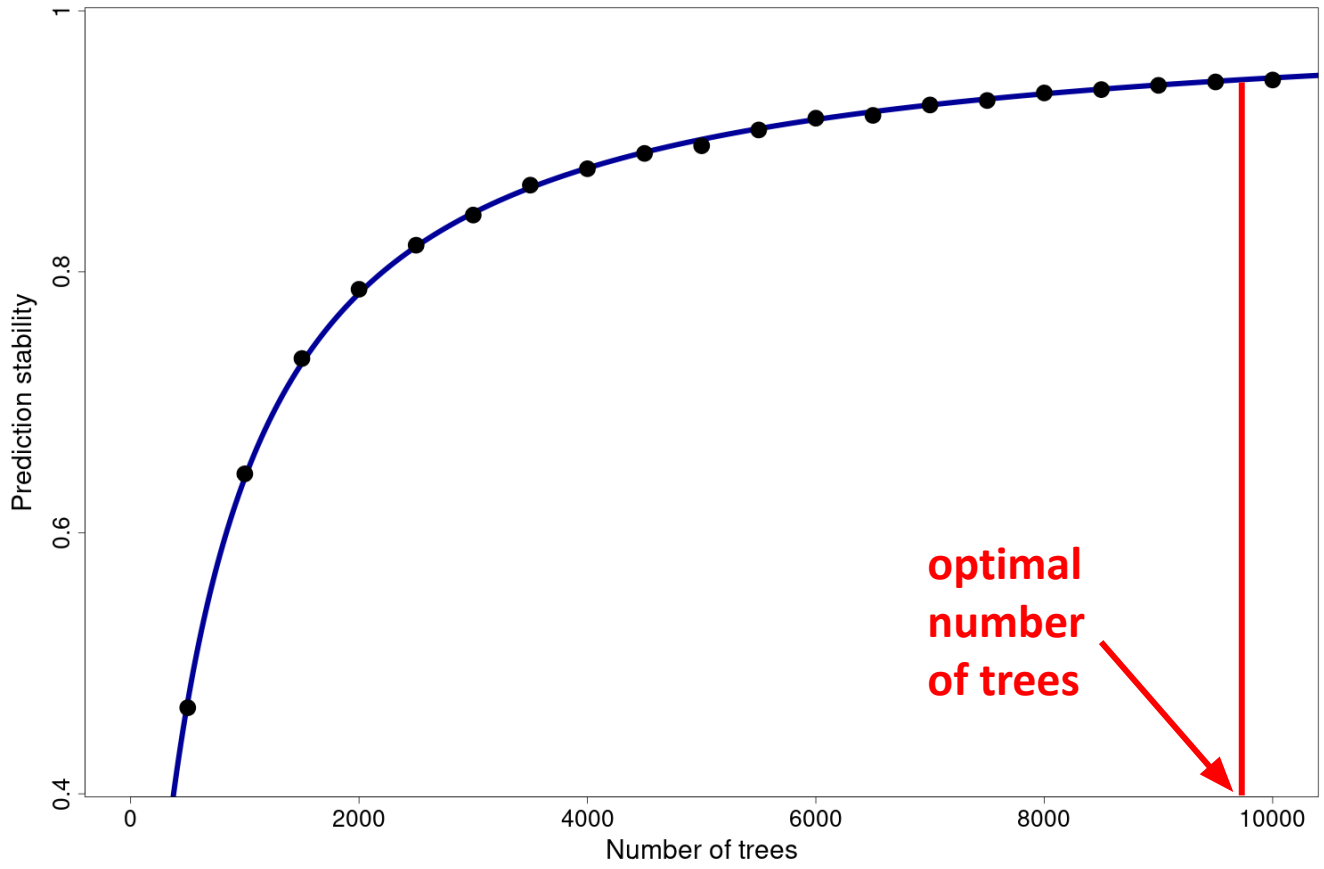
- The variability of the predictions of a non-deterministic model can be measured as the *prediction stability* by repeating the model fitting process
- The prediction stability (s) increases non-linearly with higher number of trees (t)
- This relationship can be modelled using a two parameter logistic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

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



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 predictors
  - Number of observations in the input data
  - Data set structures



The optRF package

- The R package optRF is available on CRAN
- Automatically determines the optimal number of trees tailored to your specific data set
- Calculates the expected prediction stability of the resulting model
- Estimates the computation time for the optimised random forest model

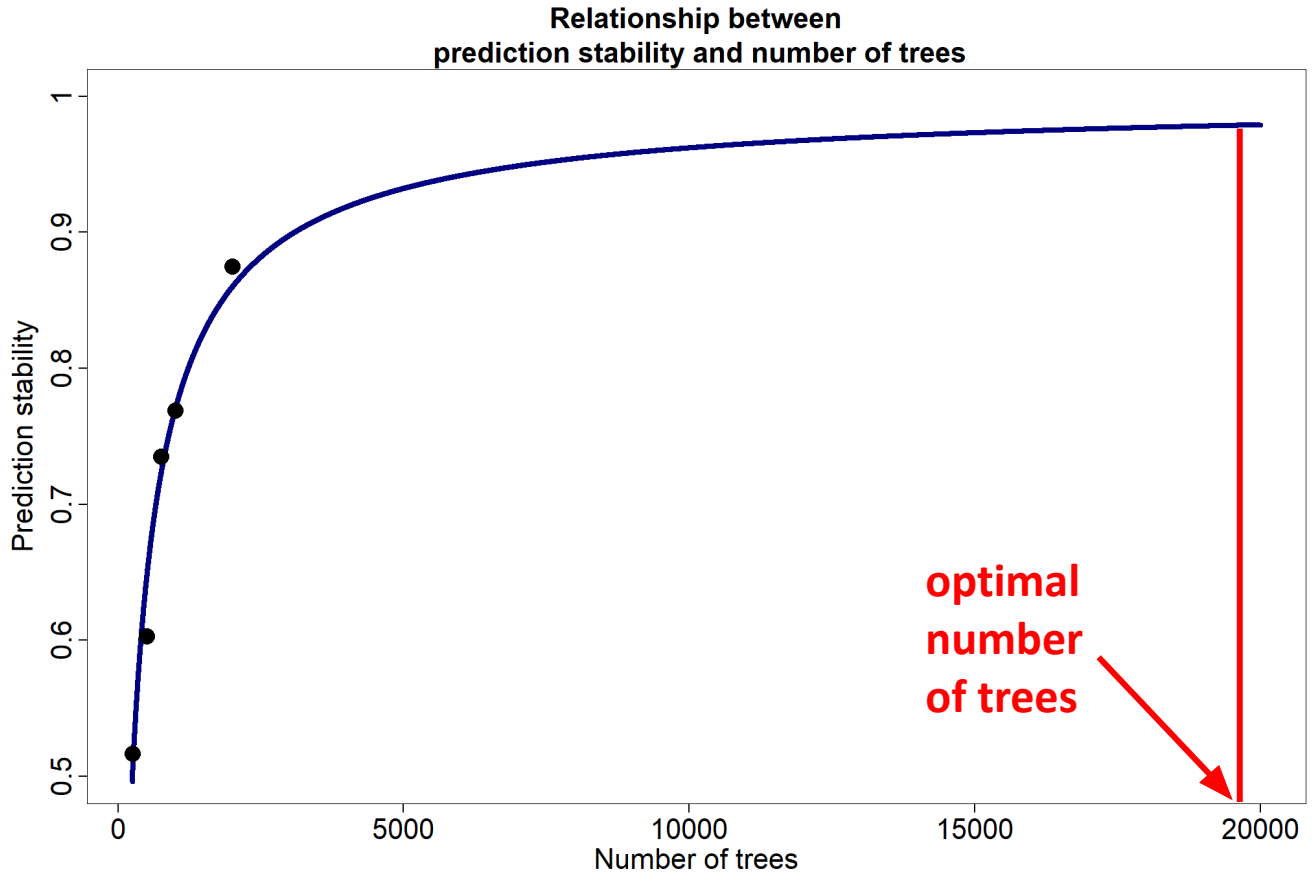
  → For the specific hardware and thread configuration when using the ranger function

```
> library(ranger)
> library(optRF)

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

> summary(optRFres)
  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)
```

Workflow of optRF

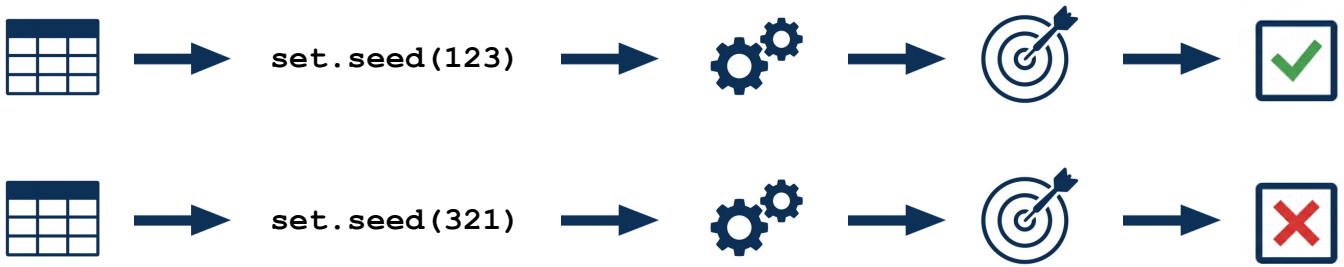
To determine the optimal number of trees, the optRF package

1) Quantifies the prediction stability for certain numbers of trees
2) Defines the relationship between the number of trees and the prediction stability using the 2PL model
3) Extrapolates the relationship to very large numbers of trees
4) Determines the optimal number of trees



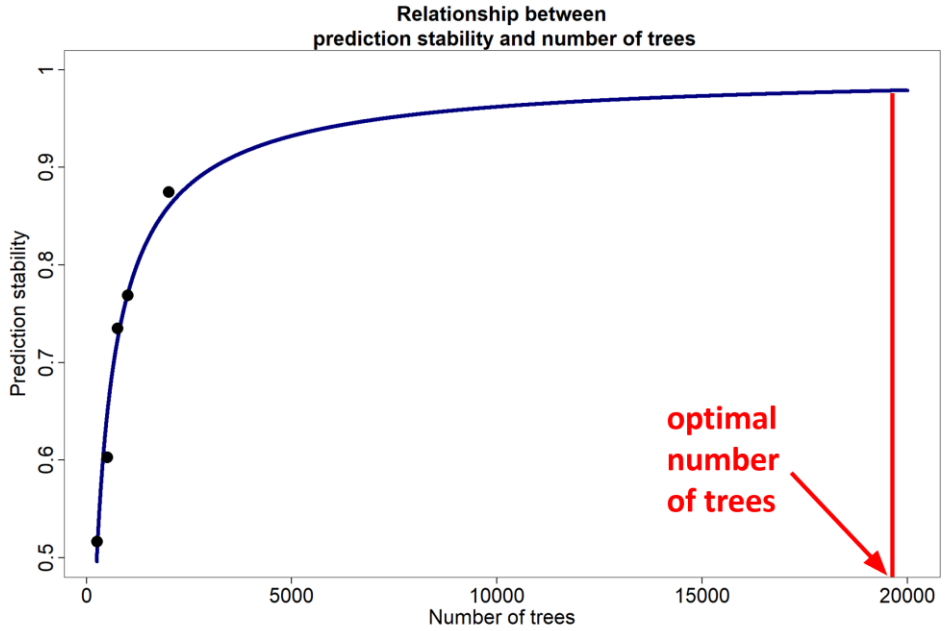
Reproducibility is Not Stability

- 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

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792.

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.

Lange, T. M., Heinrich, F., Kopisch-Obuch, F., Keunecke, H., Gültas, M., & Schmitt, A. O. (2023). Improving genomic prediction of rhizomania resistance in sugar beet (*Beta vulgaris* L.) by implementing epistatic effects and feature selection. *F1000Research*, *12*, 280.

Mohamadi, Z., Shafizadeh, A., Aliyan, Y., Shayesteh, S. F., Goudarzi, P., Khodabandeh, A., ... & Pouyan, K. (2025). The application of random forest-based models in prognostication of gastrointestinal tract malignancies: a systematic review. *Frontiers in Artificial Intelligence*, 8, 1517670.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature.

Roberts-Nuttall, J., Jones, A. M., Castellani, M., & Pham, D. (2026). An interpretable machine learning framework for adverse drug reaction prediction from drug-target interactions. *PLoS One*, 21(1), e0340900.

Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Qiu, L. (2021). Exploratory study on classification of diabetes mellitus through a combined random forest classifier. *BMC Medical Informatics and Decision Making,* 21, 1–14.