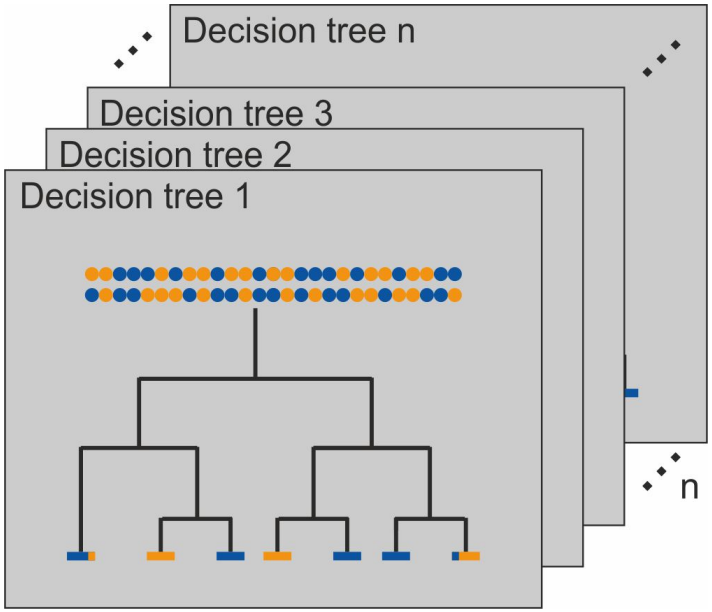
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Stephan Seifert

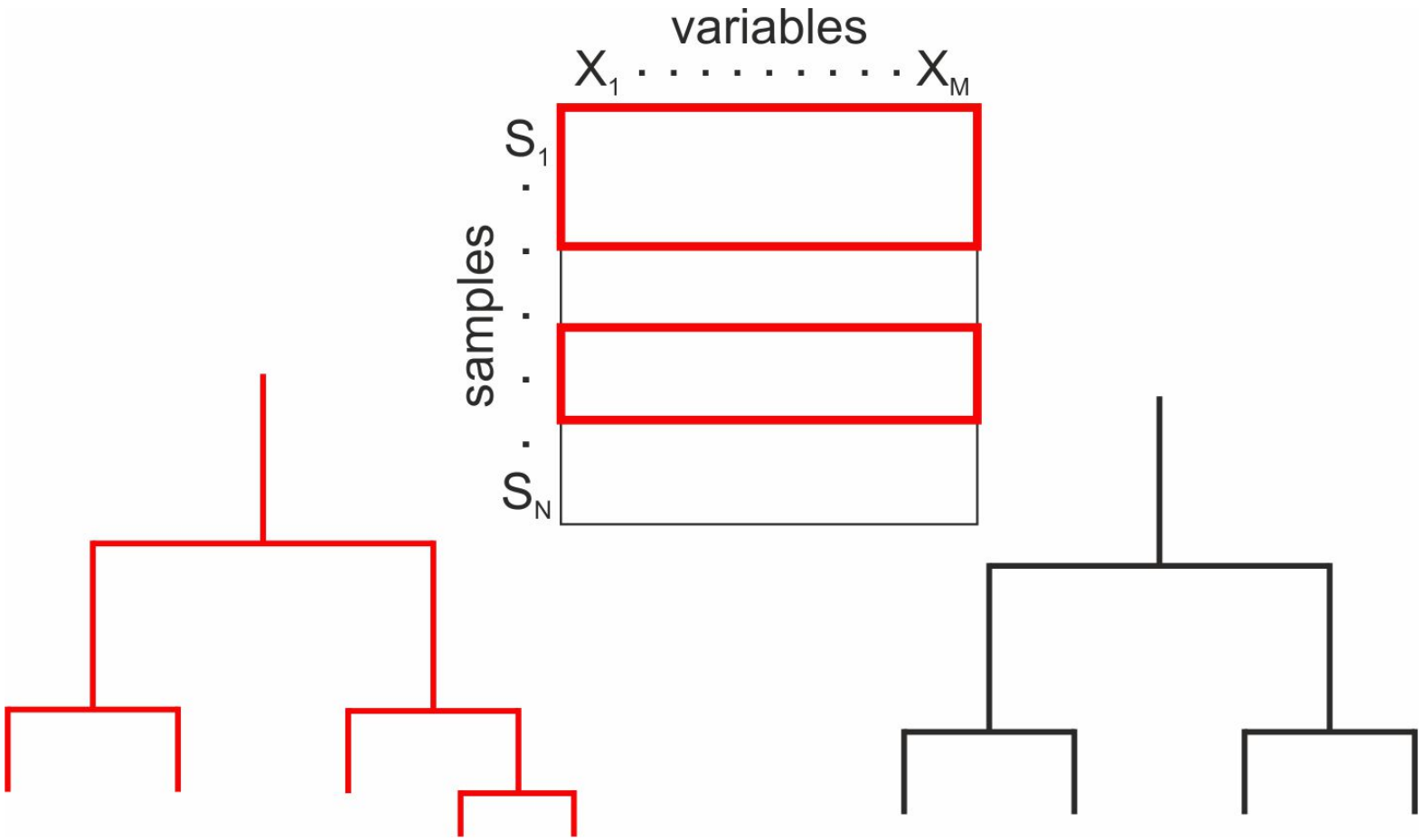
Selection of important and related variables using surrogate variables in random forests

Random Forest

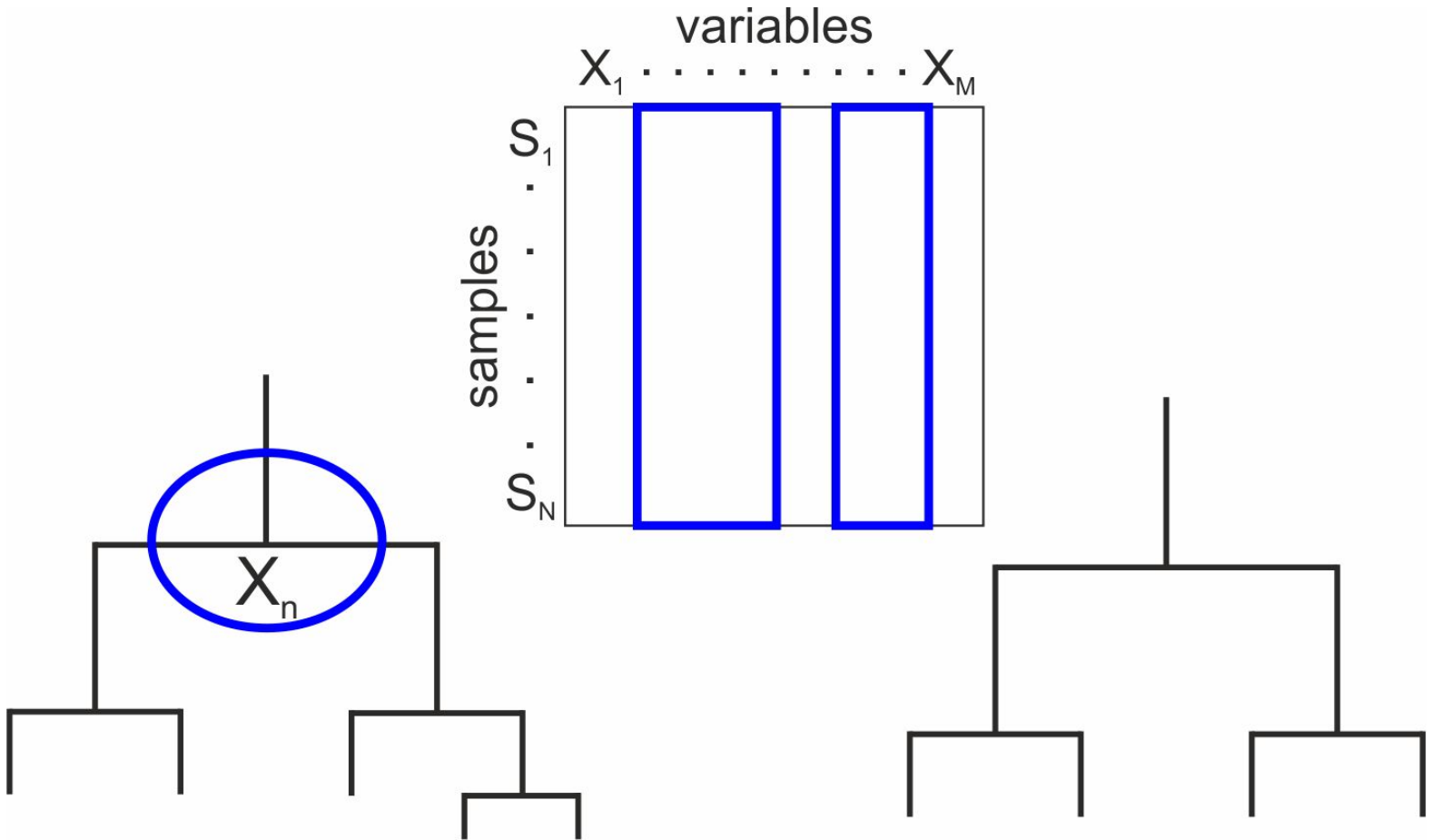


- Based on multiple decision trees
- Internal validation
- No distributional assumptions
- Different types of input variables
- Different outcomes
  (e.g. classification and regression)
- Can analyze high dimensional data
- Efficient implementations in R (ranger package)
- Multiple approaches for variable selection

Random Forest: Bootstrap samples to build each tree



Random Forest: Random subset of variables as candidates for each split



Variable importance

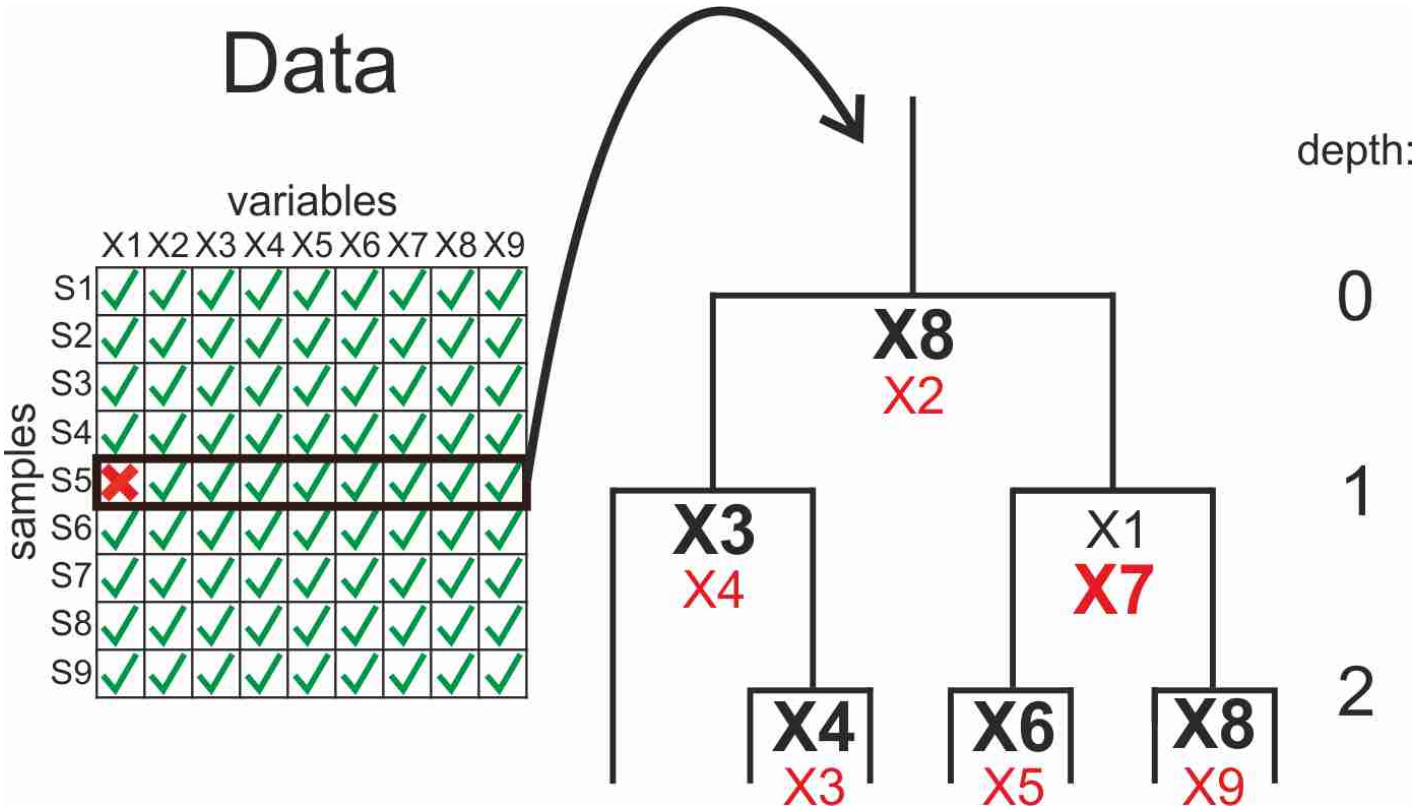
- Evaluation of each variable
- Relevance of variable for outcome
- Often applied: permutation importance
- Variable selection based on variable importance and a threshold (statistical test)
- Vita and Boruta top-performing methods in comparison study [1]

[1] F. Degenhardt, S. Seifert, S. Szymczak, Evaluation of variable selection methods for random forests and omics data sets. Brief. Bioinform. 2019, 20, 492-503.

Aims of variable selection

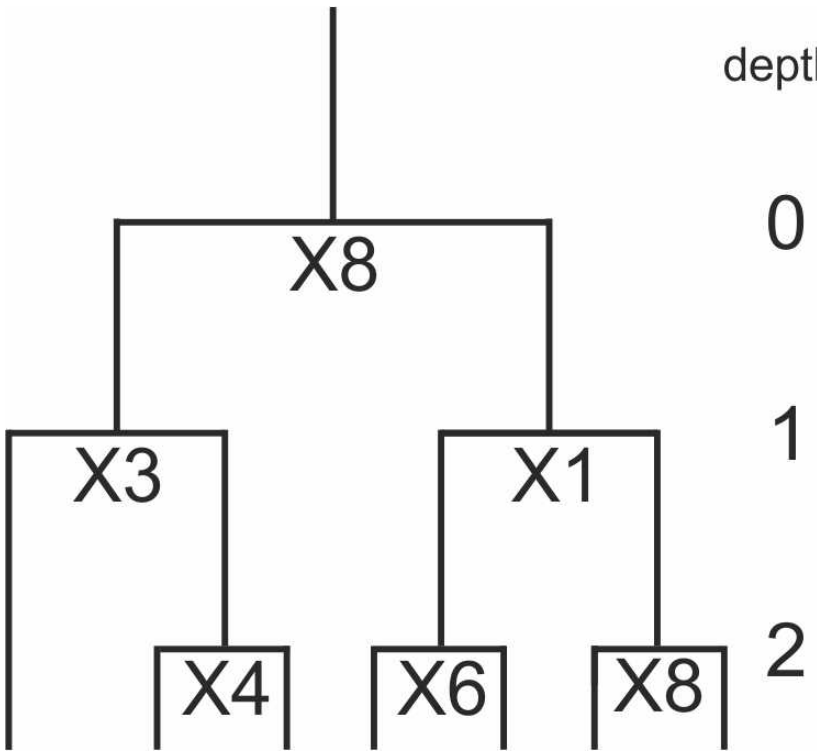
- Parsimonious model
  - Minimal set of variables
  - No redundant variables
    $\longrightarrow$ Selection based on permutation importance
- Information about underlying mechanisms
  - All relevant variables
  - Include redundant variables $\longrightarrow$ Variable selection based on tree structures

Surrogate Splits



L. Breiman, Classification and Regression Trees 1984, p. 140ff.

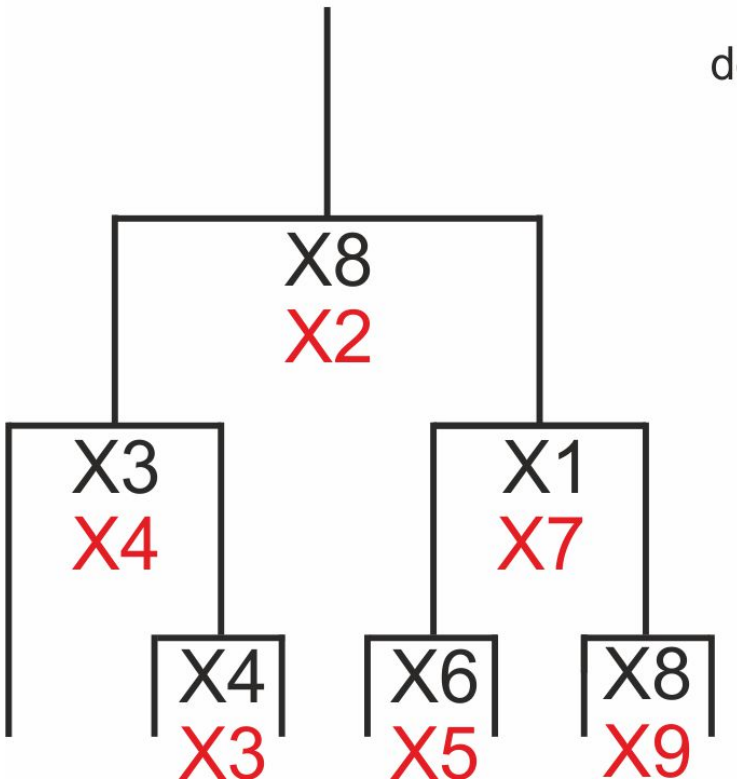
Minimal Depth



variable	minimal depth
X1	1
X2	3
X3	1
X4	2
X5	3
X6	2
X7	3
X8	0
X9	3

H. Ishwaran et al., High-Dimensional Variable Selection for Survival Data, J. Am. Stat. Assoc. 2010, 105, 205.

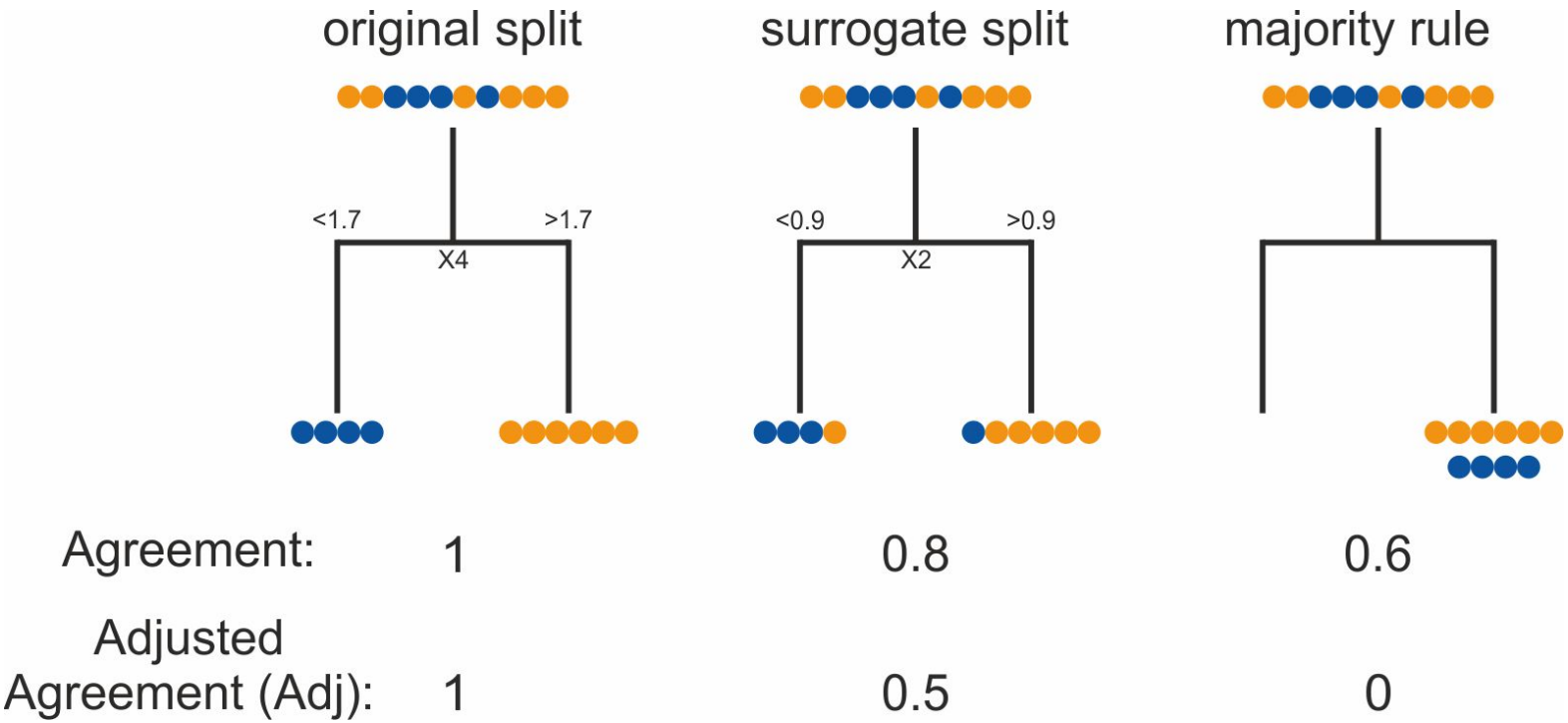
Surrogate Minimal Depth (crucial parameter: s)



variable	surrogate minimal depth
X_1	1
X_2	0
X_3	1
X_4	1
X_5	2
X_6	2
X_7	1
X_8	0
X_9	3

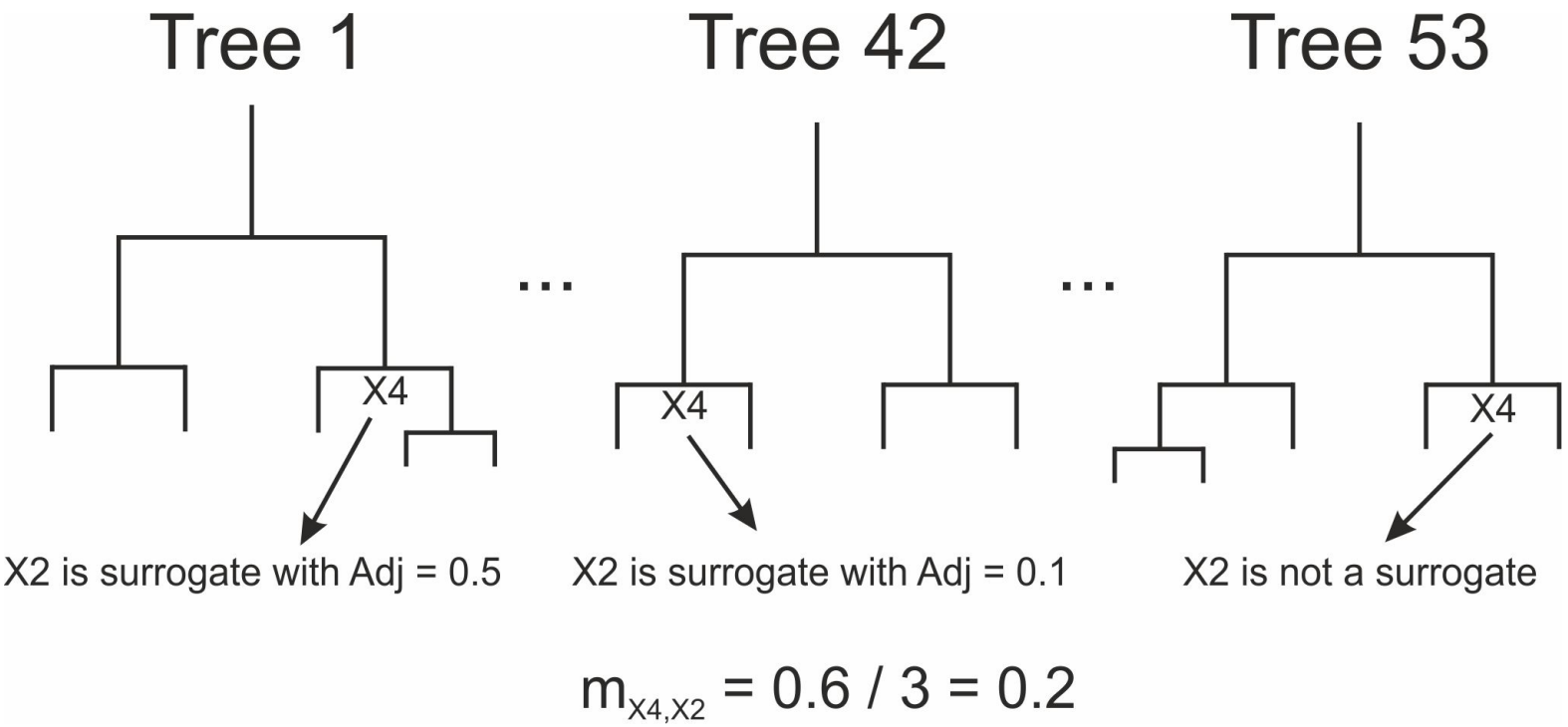
S. Seifert et al.,Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Identification of surrogate variables



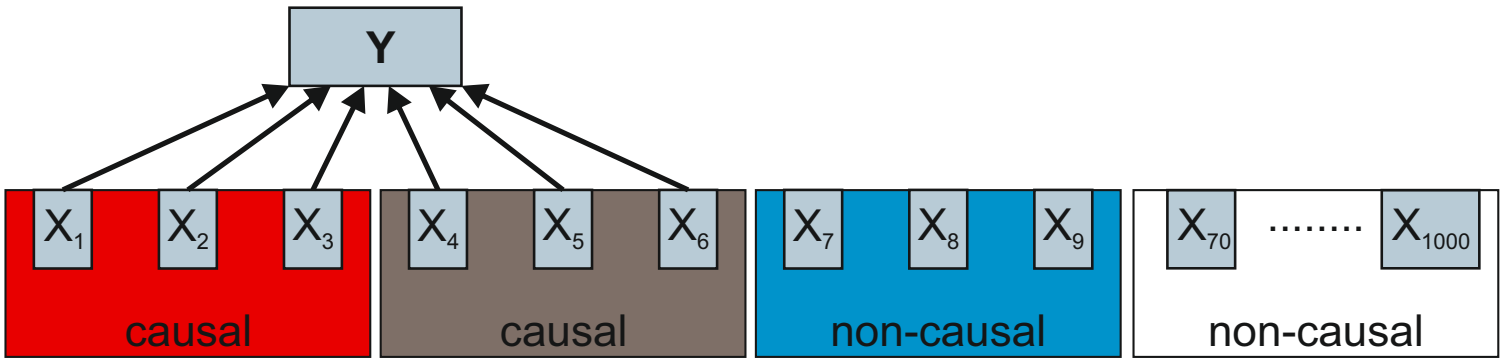
S. Seifert et al.,Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Mean adjusted agreement of X4 and X2



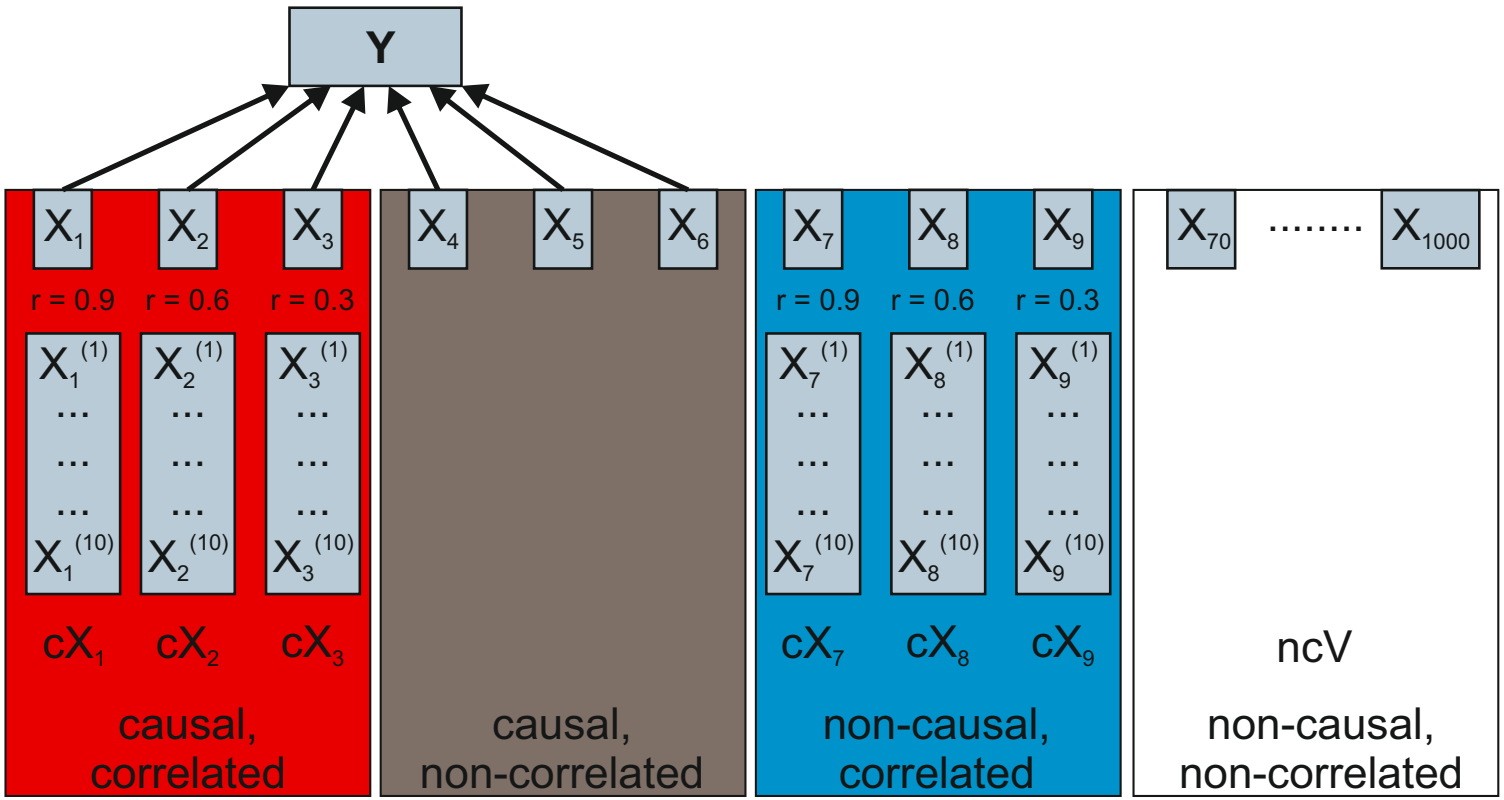
S. Seifert et al.,Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Simulation study: 50 replicates with 100 samples



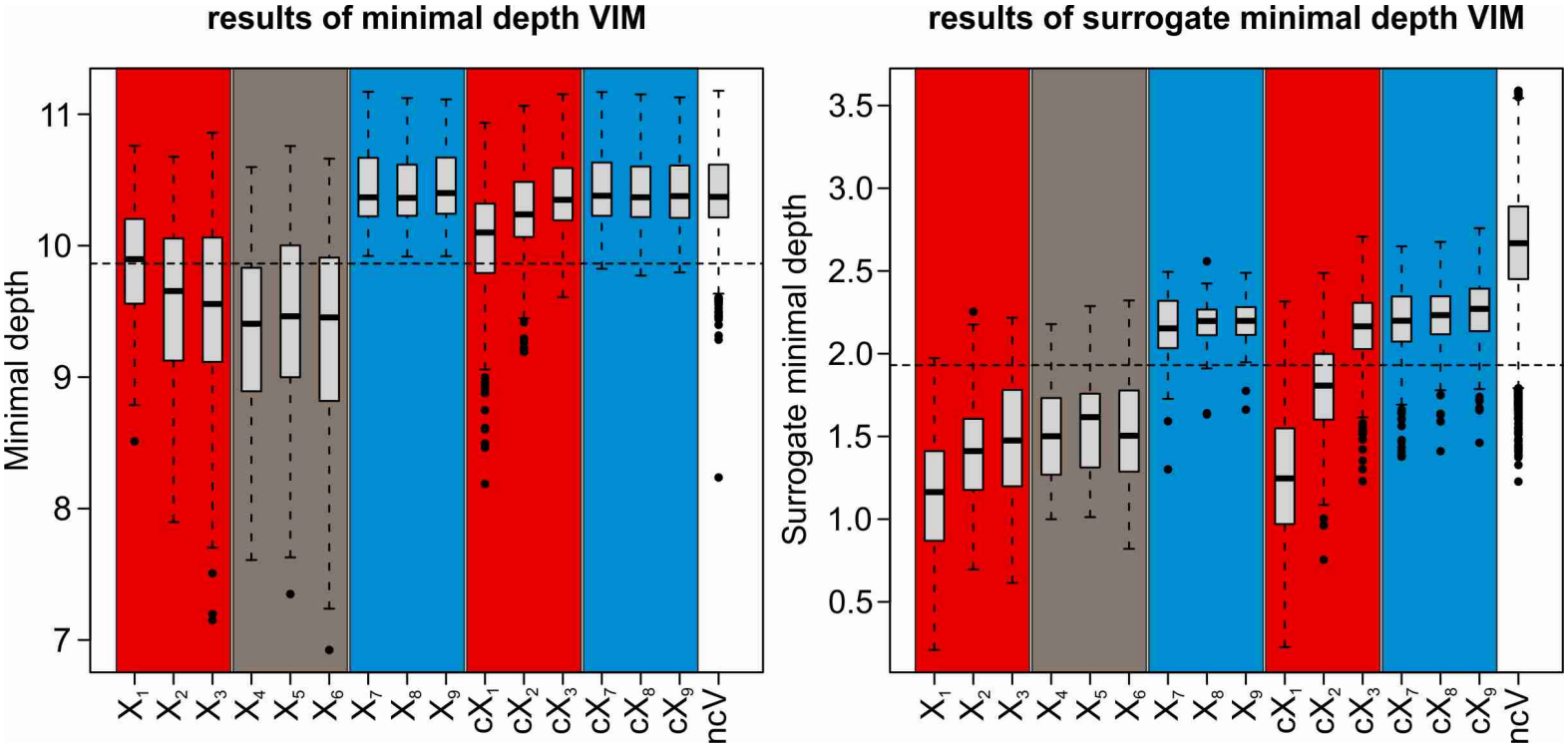
S. Seifert et al.,Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Simulation study: 50 replicates with 100 samples



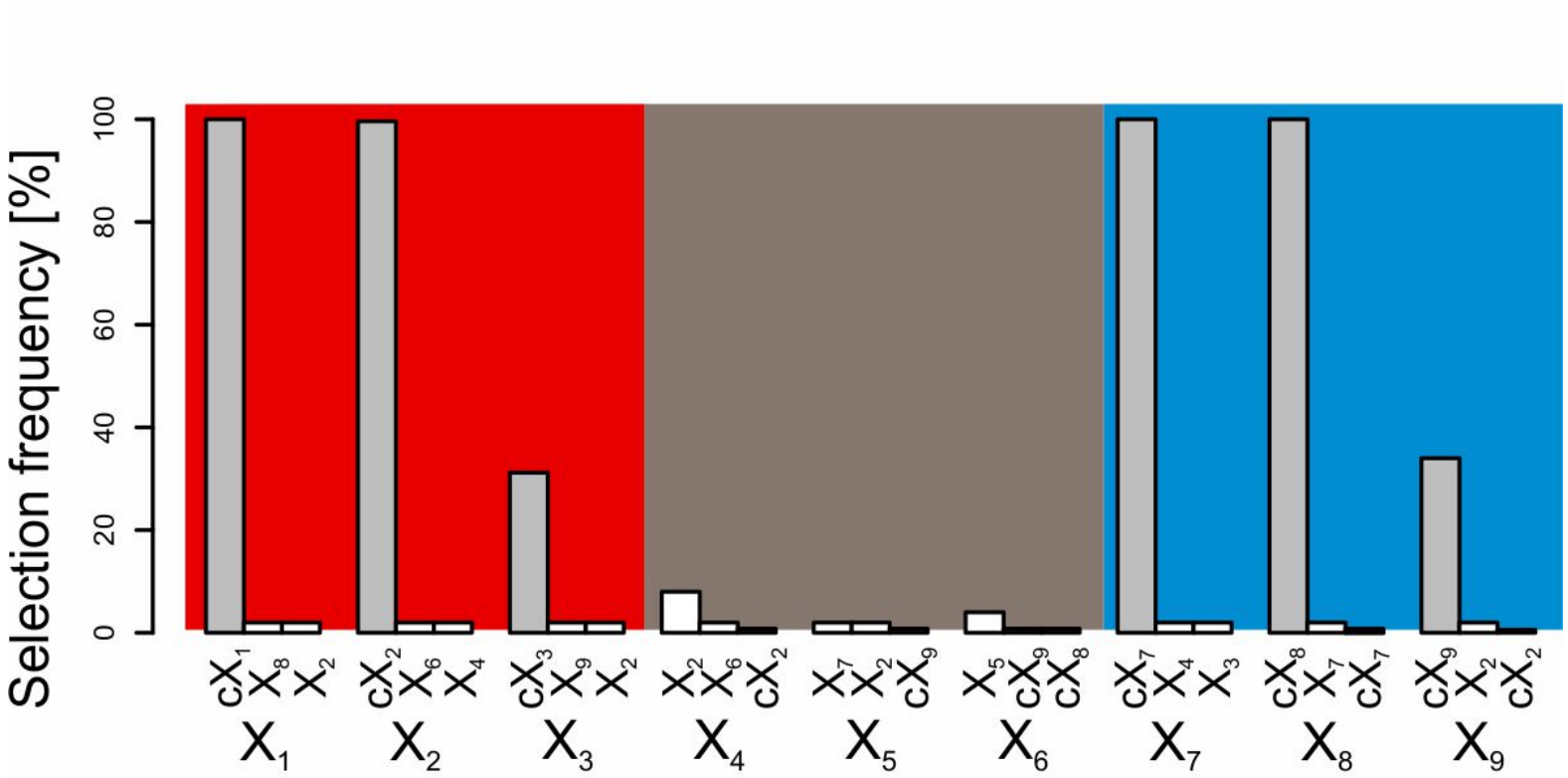
S. Seifert et al.,Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Simulation study: MD vs SMD with s=100



S. Seifert et al.„Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Simulation study: Variable relation analysis



S. Seifert et al.„Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 2019, 35, 3663-3671.

Applications for food profiling

metabolites

MDPI

Article

Opening the Random Forest Black Box of the Metabolome by the Application of Surrogate Minimal Depth

Soeren Wenck, Marina Creydt, Jule Hansen, Florian Gärber, Markus Fischer and Stephan Seifert *

Microchemical Journal 174 (2022) 107066

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Check for updates

Determination of the geographical origin of hazelnuts (*Corylus avellana* L.) by Near-Infrared spectroscopy (NIR) and a Low-Level Fusion with nuclear magnetic resonance (NMR)

Navid Shakiba [a,b], Annika Gerdes [a,b], Nathalie Holz [a], Soeren Wenck [b], René Bachmann [c], Tobias Schneider [a], Stephan Seifert [b], Markus Fischer [b], Thomas Hackl [a,b,*]

Applications on SERS data

SCIENTIFIC REPORTS
nature research

OPEN

Application of random forest based approaches to surface-enhanced Raman scattering data

Stephan Seifert[1,2]

ACS NANO
Cite This: *ACS Nano* 2019, 13, 9363−9375
www.acsnano.org

Optical Nanosensing of Lipid Accumulation due to Enzyme Inhibition in Live Cells

Vesna Živanović,[†,‡] Stephan Seifert,[§] Daniela Drescher,[†] Petra Schrade,[∥] Stephan Werner,[⊥] Peter Guttmann,[⊥] Gergo Peter Szekeres,[†,‡] Sebastian Bachmann,[∥] Gerd Schneider,[⊥] Christoph Arenz,[†,‡] and Janina Kneipp*[,†,‡]

Summary

- Surrogate Minimal Depth (SMD): random forest based variable selection including variable relations
- SMD can also be utilized to analyze variable relations
- Relation parameter shows the mutual impact of the variables on the model
- Broad applications, e.g. for food profiling and to analyze surface-enhanced Raman scattering data

R package: https://github.com/StephanSeifert/SurrogateMinimalDepth

Silke Szymczak

Sven Gundlach

Sören Wenck

e:Med SYSTEMS MEDICINE

C A U Kiel University Christian-Albrechts-Universität zu Kiel

SPONSORED BY THE Federal Ministry of Education and Research

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

H S F S