Universität Regensburg

To fit is to overfit

How the negligence of prediction performance blurs model quality

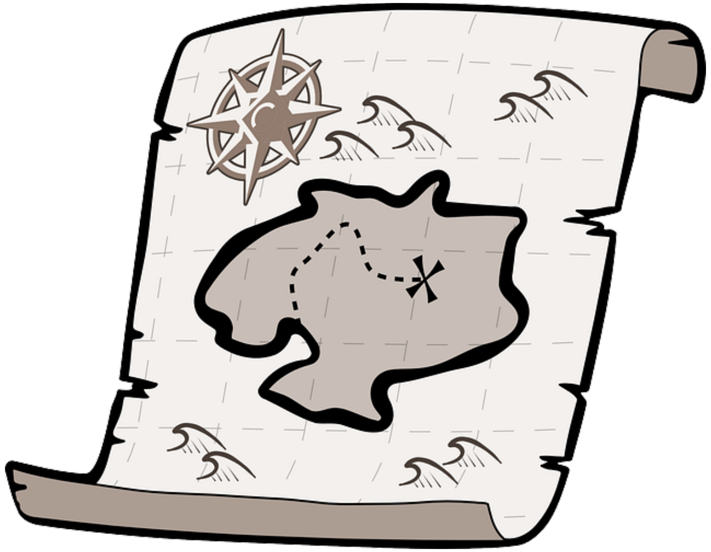
Sven Hilbert & Elisabeth Kraus

Universität Regensburg

Topics

Map of the topics covered in this talk

- Goals of an empirical science
- Comparison of two cultures of modeling (in empirical science)
- Short overview predictive modeling
- Prediction and explanation
- Over- and underfitting
- Resampling
- Short summary



Goals of empirical science

Universität Regensburg

1. Description

- **Descriptive statistics:** Summary statistics and plots, to make the data accessible

2. Explanation

- **Statistical inference:** Estimation of parameters to model the patterns within the data sample, assumptions about probability distributions

3. Prediction

- **Predictive modeling:** prediction of novel data, after training a model through resampling

➢The overarching goal is generalization

https://eclecticlight.co/2016/02/17/the-story-in-paintings-enlightened-by-science/

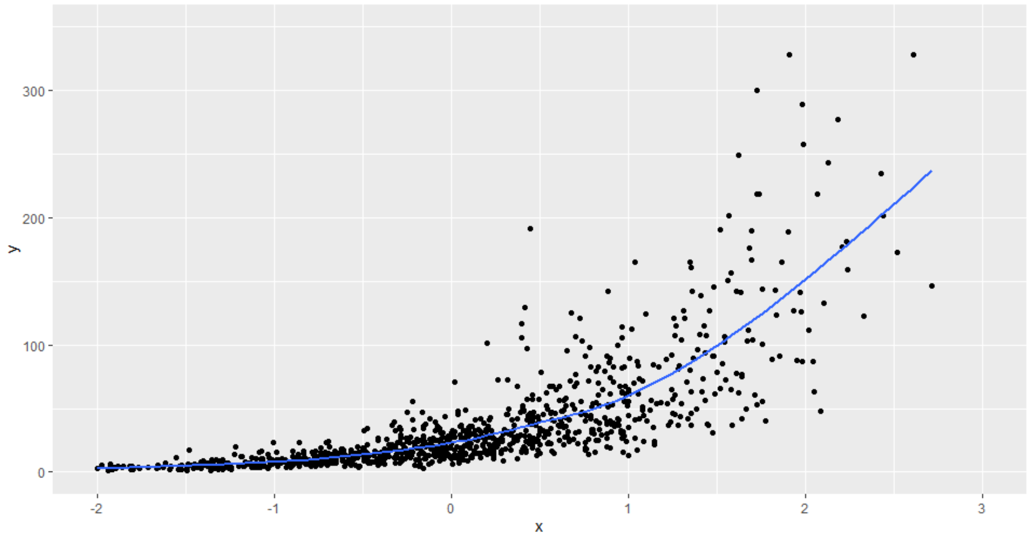
Explanation and prediction

Universität Regensburg

Leo Breiman (2001): ‚*Two cultures of statistical modeling*‘

1. Strong **theoretical assumption** of a given stochastic model, a data-generating process
   - ➢ e.g., linear or exponential relationship
   - (Classical) Inference statistics
     - Focus on **explanation** and model assumptions
     - *p*-values for inference

2. Treatment of **data-generating process** as **unknown,** use of flexible algorithmic models
   - Predictive modeling, machine learning
     - Focus **prediction performance**
     - Estimation of generalization error



Assumptions of classical models

- **General Linear Model**
  - Normal distribution of the residuals

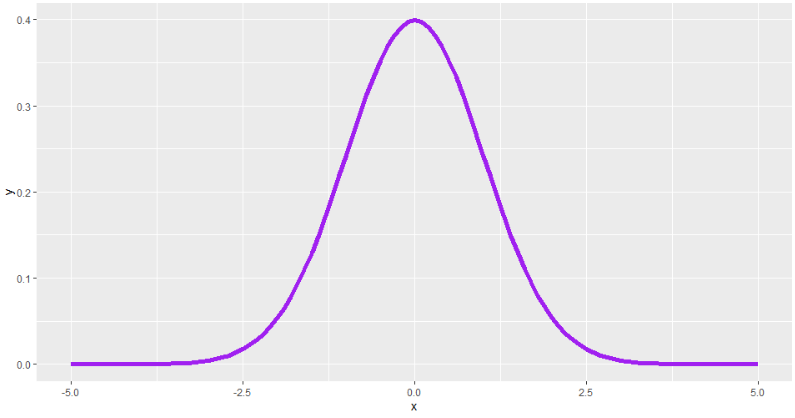
$$\varepsilon \sim N(0; \sigma^2)$$

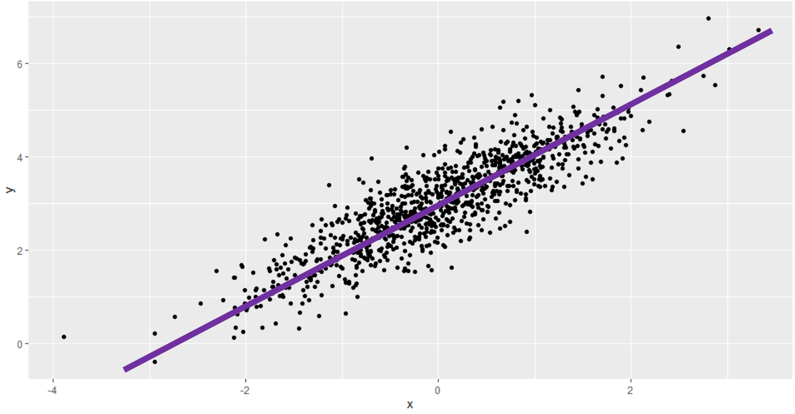
  - Linear relationships

$$y = \beta x + \varepsilon$$

- **Generalized Linear Model**

$$y = g(\beta x + \varepsilon)$$





Short overview predictive modeling

Universität Regensburg

Model types

- Tree-based methods (CART)
  - Random forest, boosting
- Kernel-based methods
  - Support vector machines
- Deep Learning
  - Neural network models

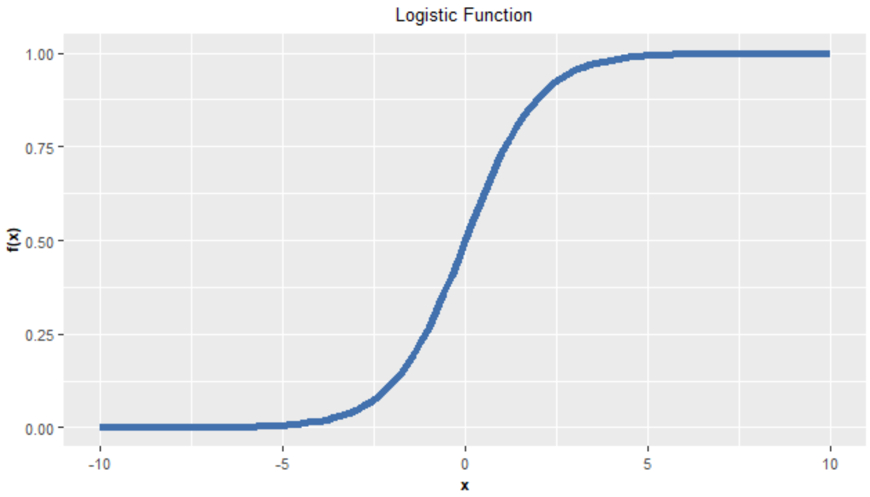
Characteristics

- Optimized for the **prediction** of **novel data**
- Often without directly interpretable parameters
- Highly functional with large amounts of variables
- Use of **resampling**

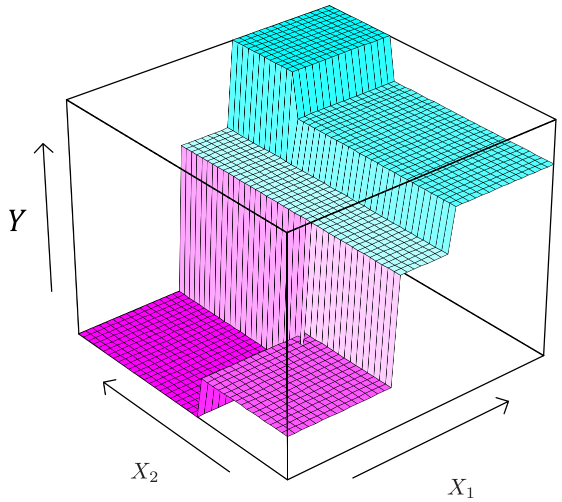
http://thecontextofthings.com/2016/04/18/transparency-in-business/

Comparison of classification models

- Classically, we use **logistic regression models** for (dichotomous) categorization
  - Interpretable parameters, but little flexibility when fitting to data
- **Tree-based models** are more flexible
  - However, interpretability often difficult and limited



$$\frac{e^x}{1+e^x}$$



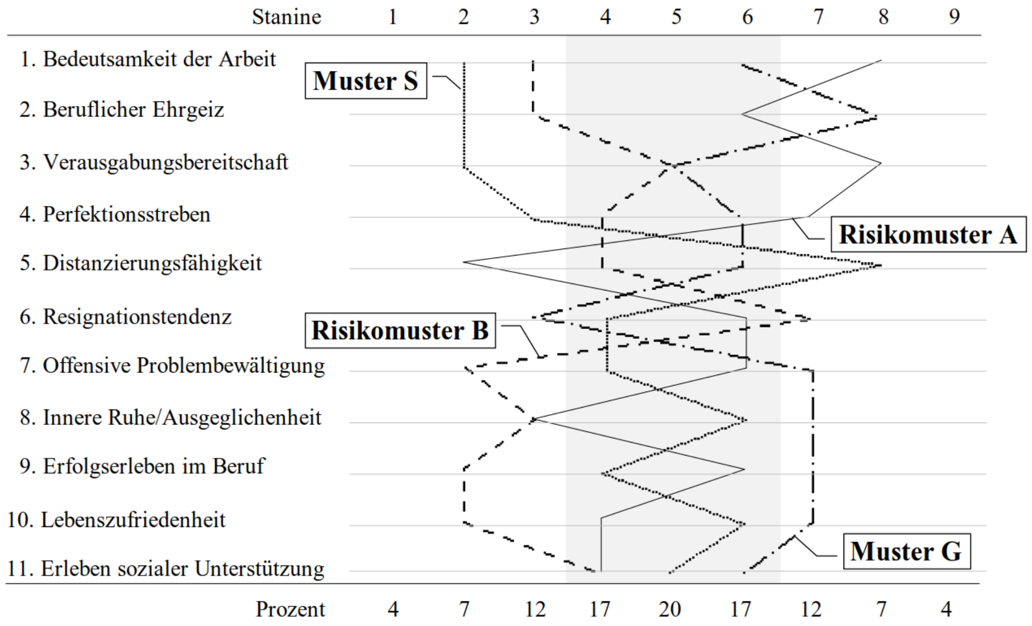
?

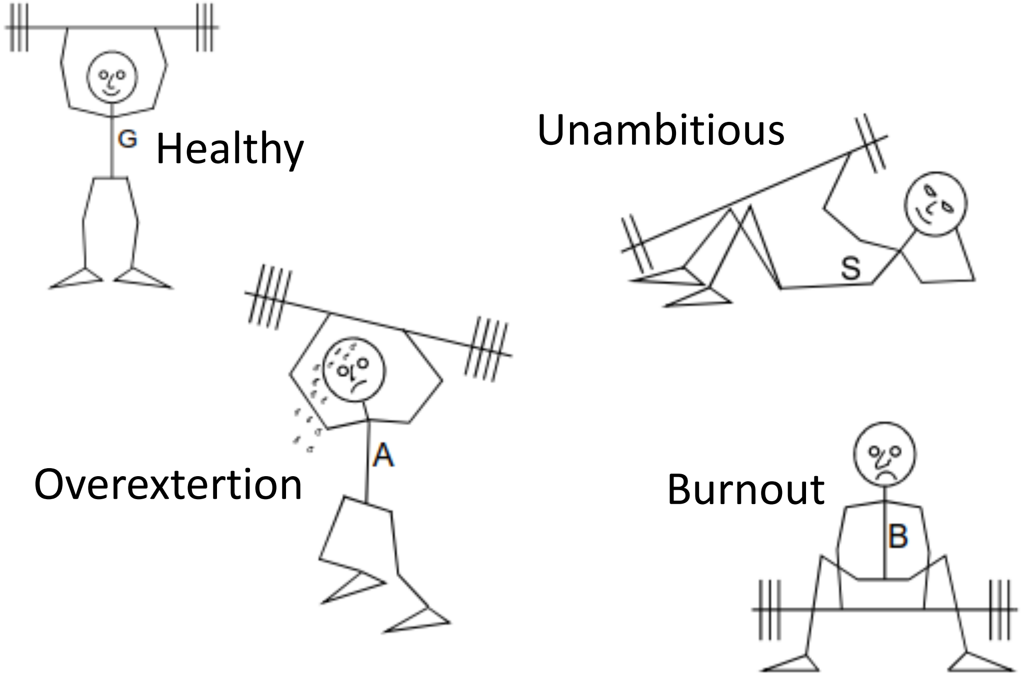
Grafik: Efron, B., & Hastie, T. (2016). *Computer age statistical inference*(Vol. 5). Cambridge University Press., S. 125

Exemplary Study Personality Types

Universität Regensburg

- AVEM: Pattern of Work-related Coping Behavior (Schaarschmidt & Fischer, 1996), modeled with a sample of *N* = 478 teachers





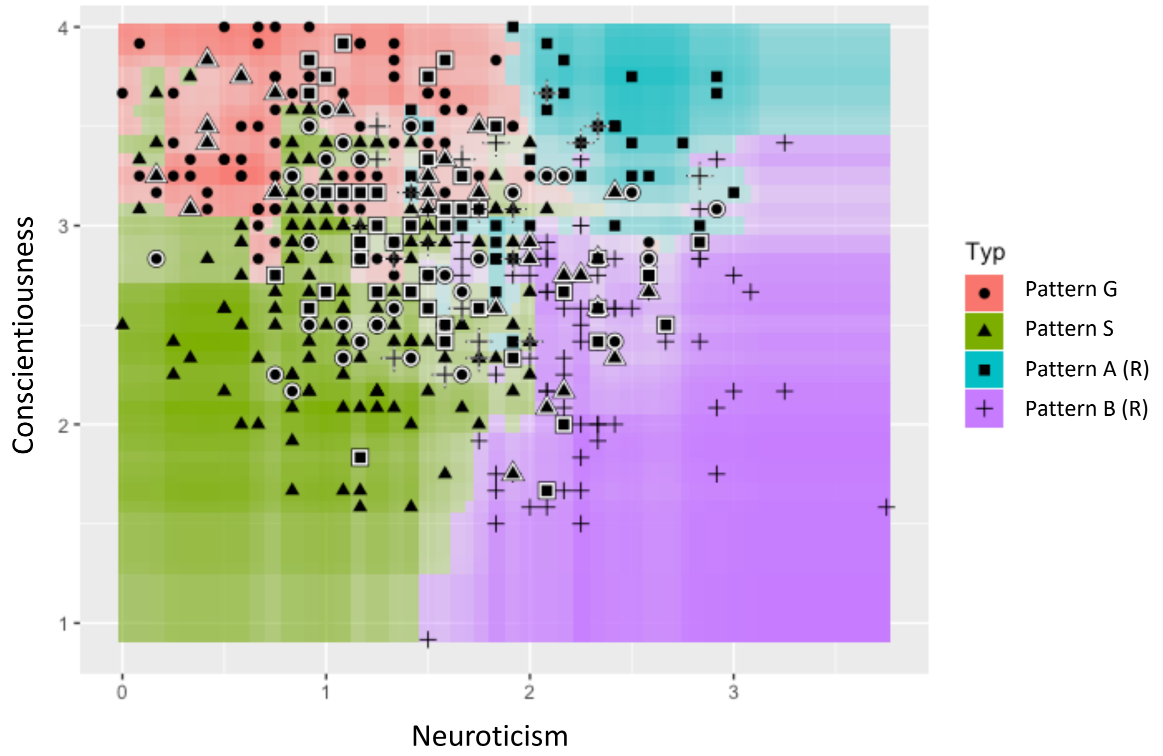
➢ Prediction using the Big Five personality traits, Motivation, and Competence

Prediction and explanation for AVEM patterns

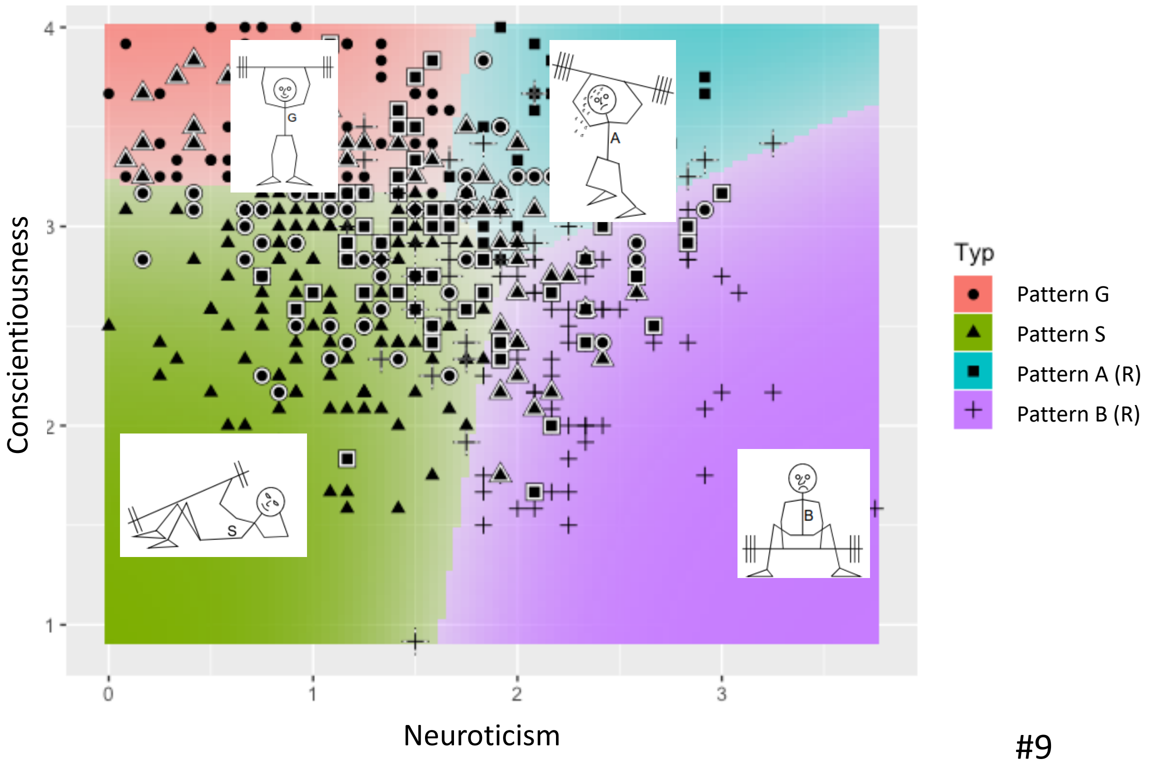
Universität Regensburg

Two models: **Random forest** and **multinomial regression**

Model with highest prediction performance



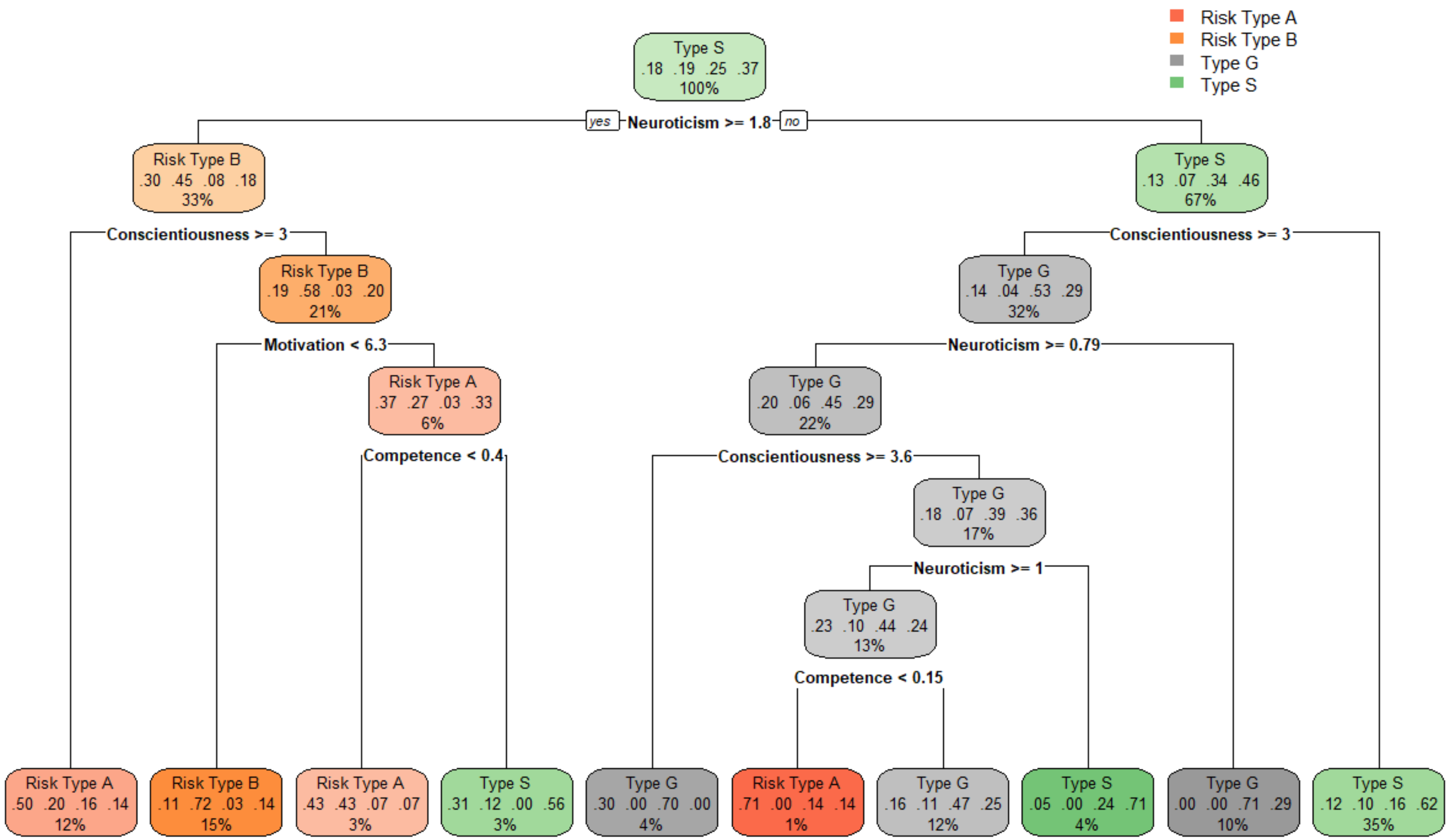
Model with most a priori assumptions



Universität Regensburg

Example decision tree

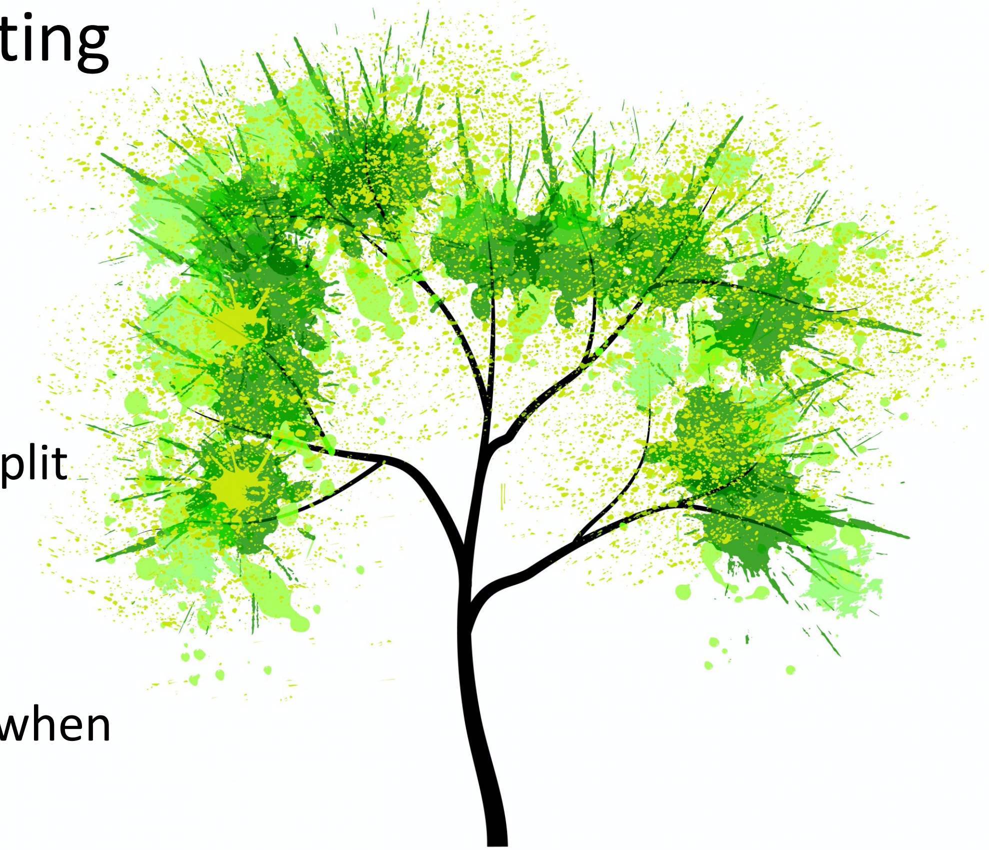
Classification of four AVEM coping patterns



CART overfitting

Many tree-based machine learning algorithms integrate measures to actively avoid overfitting

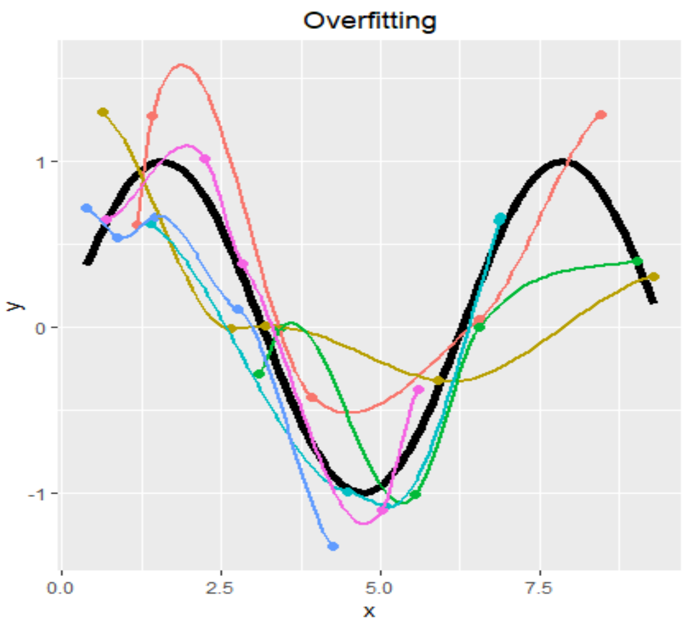
- **Random forest**
  - Bootstrapping
    - Bootstrap the cases for each tree
  - Split-variable randomization
    - Randomly select only *m* out *p* variables for each split
- **Boosting**
  - Early stopping
    - Stop improving the model fit to the training data when the test set performance stops improving



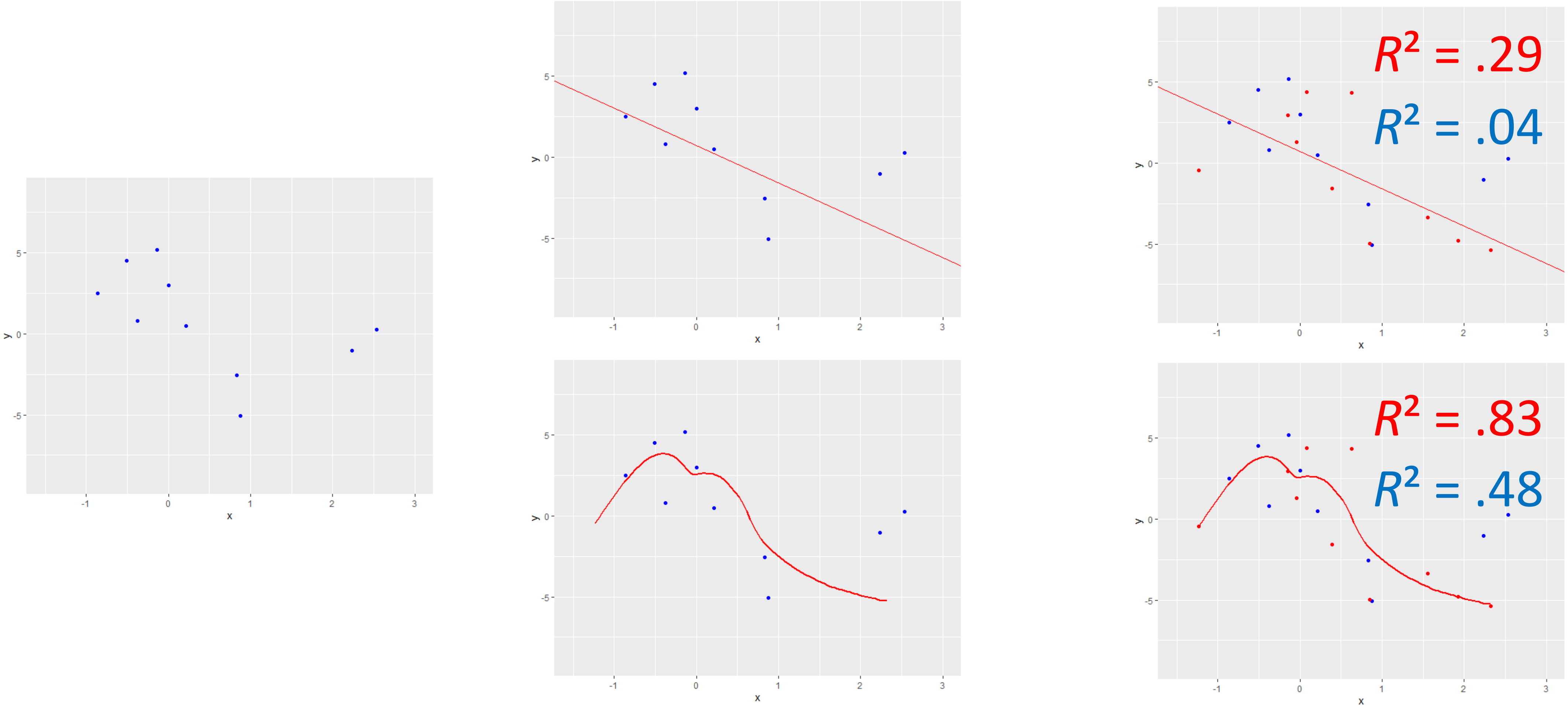
Universität Regensburg

Over- and underfit

- **Overfitting**
  - Model adapts too much to the sample data
- **Underfitting**
  - Model adapts too little to the sample data



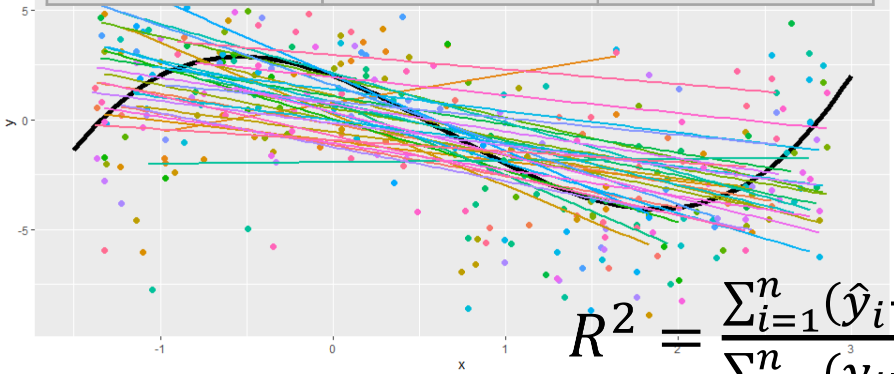
Consequences for estimates of model quality



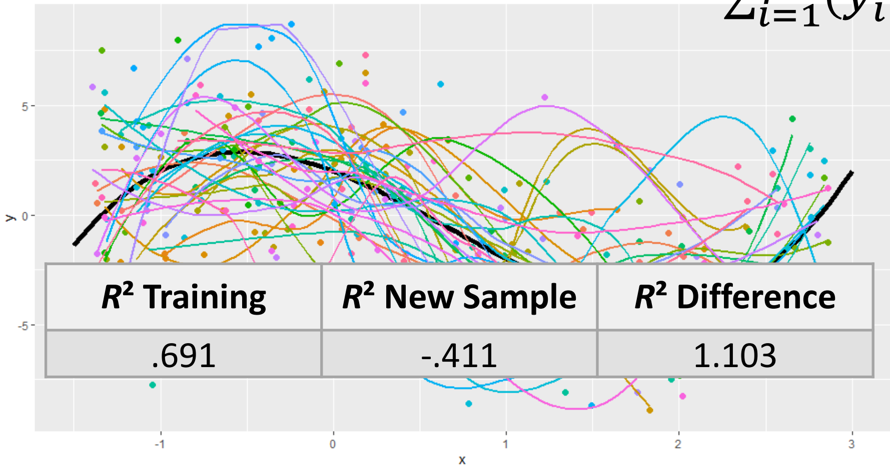
Bias, variance, and the amount of data

Universität Regensburg

R^2 Training	R^2 New Sample	R^2 Difference
.328	-.195	.523

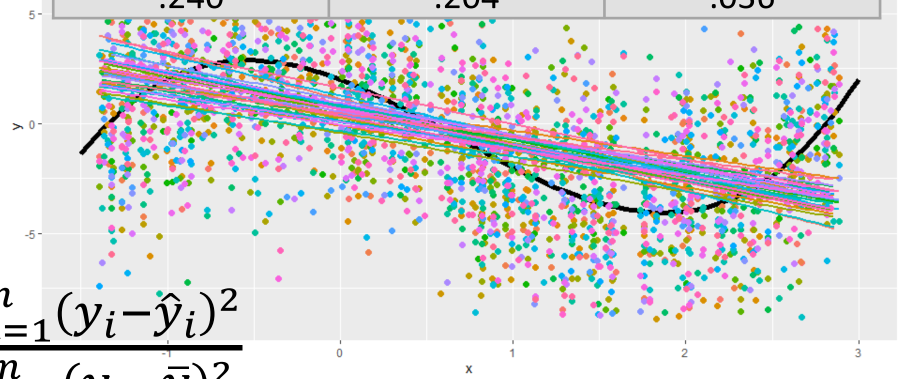


R^2 Training	R^2 New Sample	R^2 Difference
.691	-.411	1.103

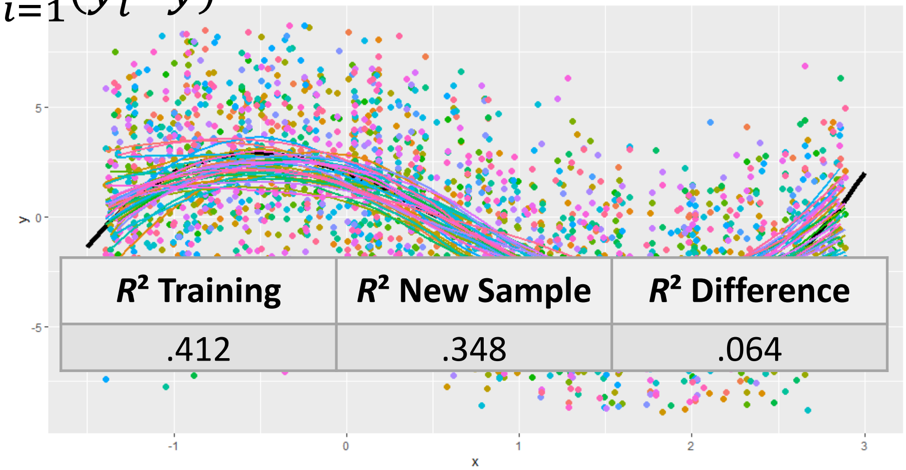


$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

R^2 Training	R^2 New Sample	R^2 Difference
.240	.204	.036



R^2 Training	R^2 New Sample	R^2 Difference
.412	.348	.064

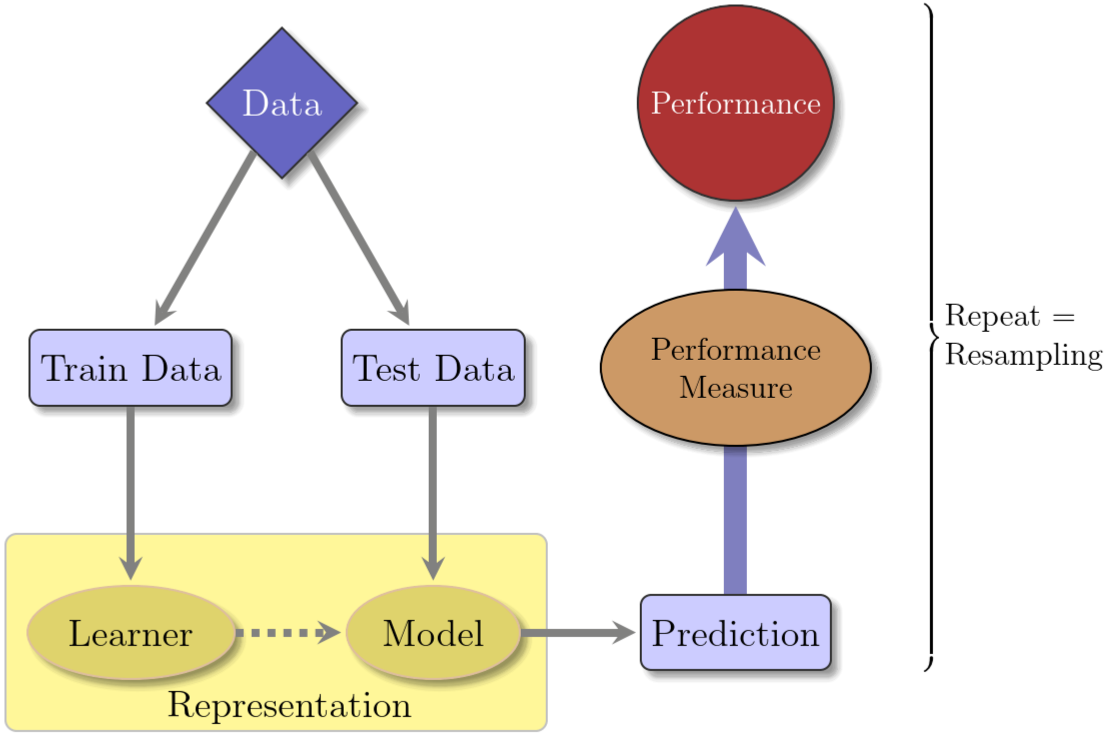


Universität Regensburg

Resampling

- **Use of training and test sets**
  - A learner is trained through resampling to become a model
- **Performance measure for the generalization error**
  - Comparison of different model types

➢ Model with **most accurate prediction** is used



Universität Regensburg

Generalization error

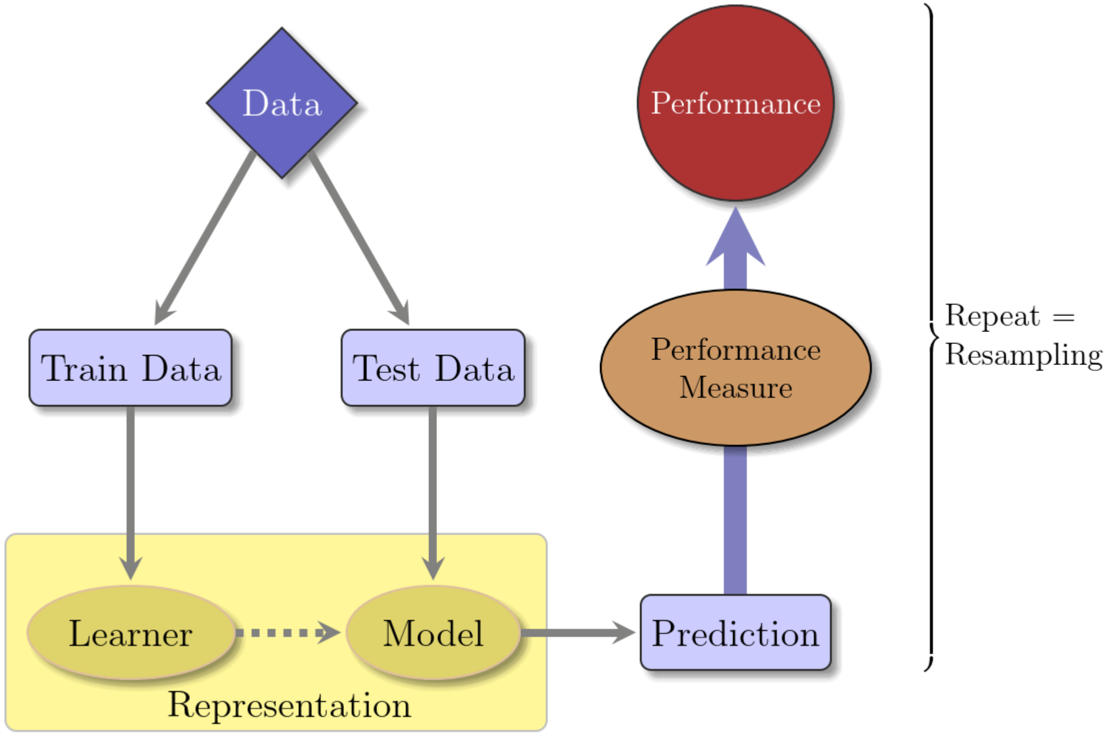
Estimation of the generalization error

- Categorization

$$MMCE = \frac{\#Misclassifications}{\#Total\ Classifications}$$

- Regression

$$MSE = \frac{\sum_{i=1}^{n}(x_{i.Predicted} - x_{i.True})^2}{n}$$



Summary: Assessing model quality

Ideas for **increasing model quality**

- Model assessment through **prediction performance**
  - Avoid **overfitting** and over-interpretation of *p*-values
  - Combine **prediction** with **description** and **explanation**
    - ➢ Use the **head**

- **Continuous evaluation** of models
  - Repeated estimation of the generalization error

- Another important aspect: **Open Science**

➢ Simulation code available at: https://osf.io/whqmx/

Universität Regensburg

Thank you

Universität Regensburg

References

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, *16*(3), 199-231.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference* (Vol. 5). Cambridge University Press.

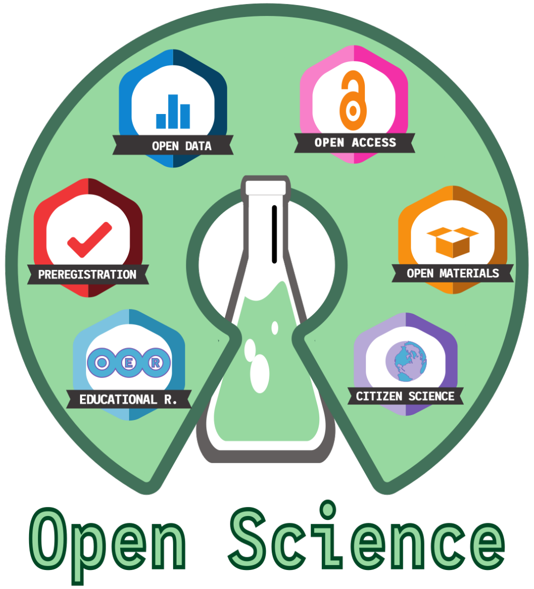
Schaarschmidt, U., & Fischer, A. (1996). *AVEM: arbeitsbezogene Verhaltens- und Erlebnismuster*. Swets Test Services.

Appendix

Open Science

- Open Science is a crucial aspect of trustworthy empirical research
  - ➢ Making the data publicly available is an important contribution to model evaluation
  - ➢ Public storage makes it possible to build new models from existing data

- A broad data base is the one of the most important foundations for the estimation of valid models

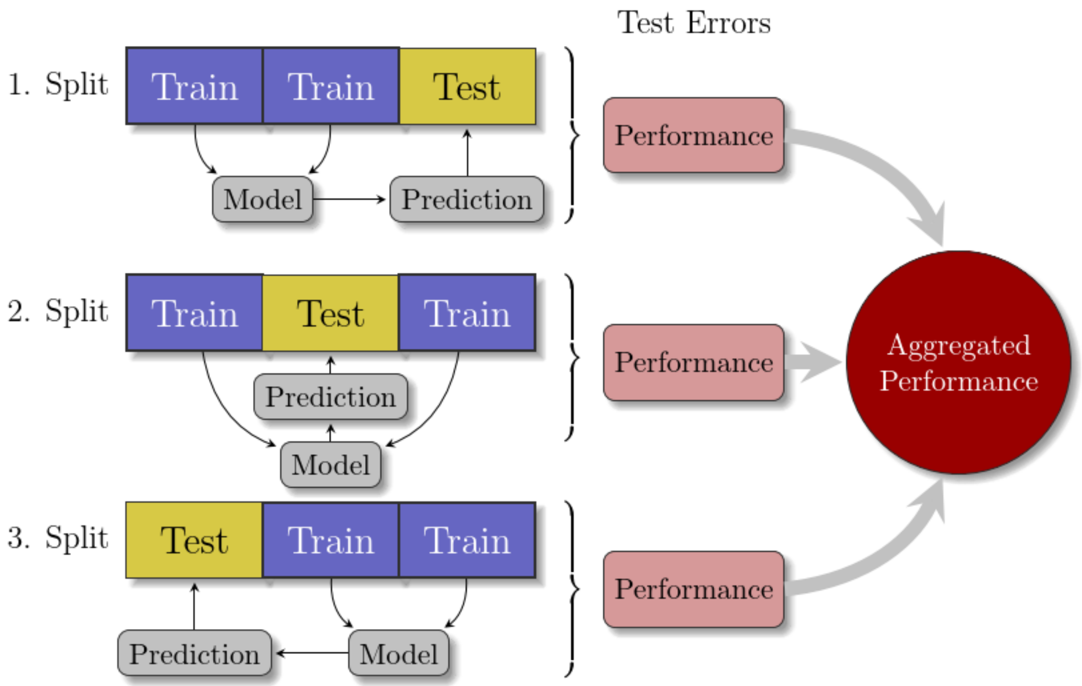


www.osf.io

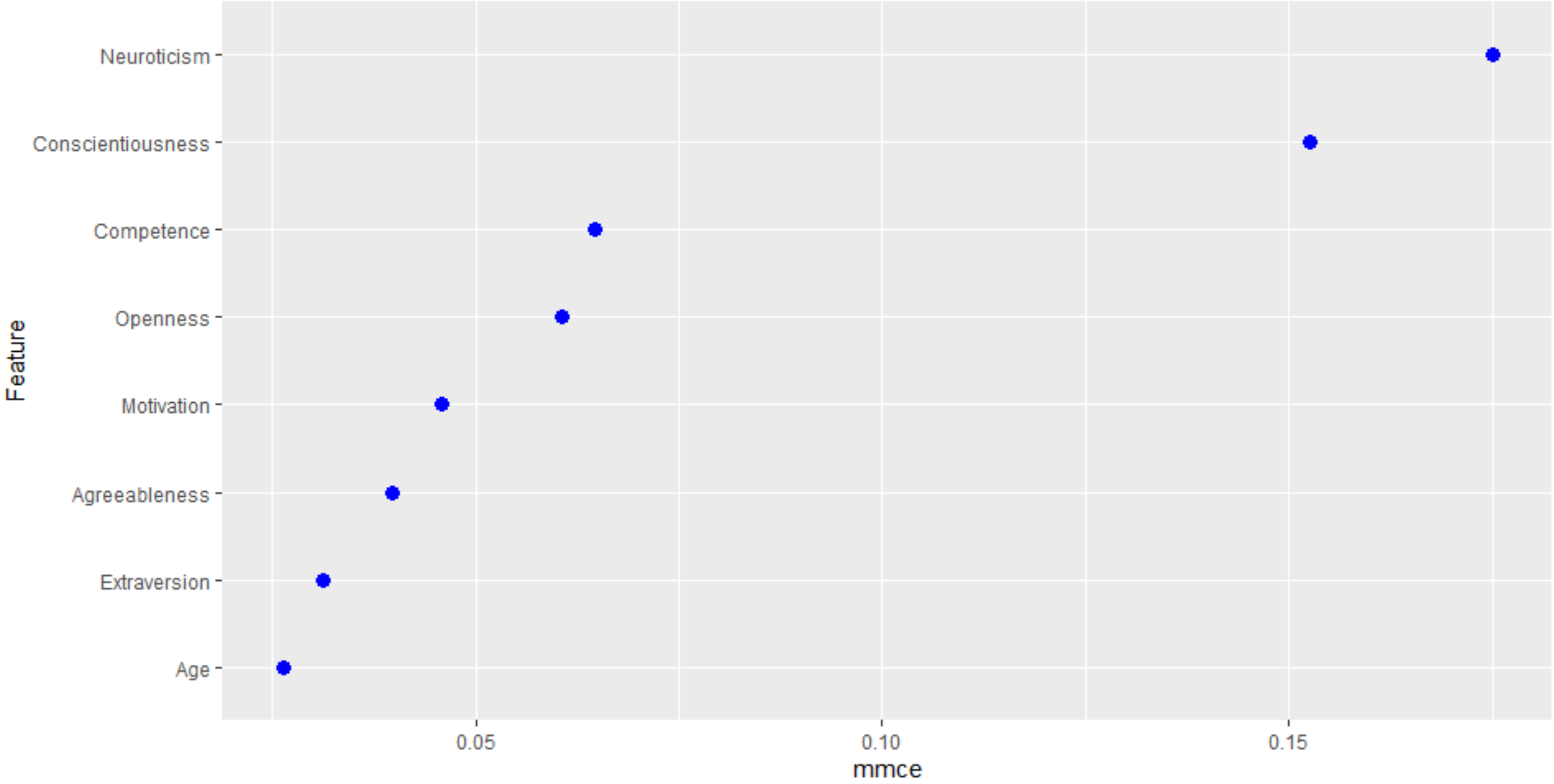
Example three-fold cross validation

Universität Regensburg

- Recycling of the sample data
  - ➢Division in multiple (sub-)sub-samples for training and testing



Variable (Permutation) importance



Overfit and test sample performance