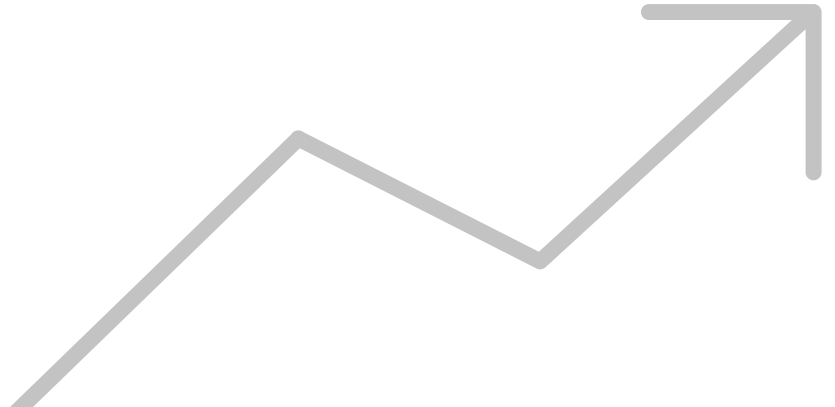
DESTATIS
Statistisches Bundesamt

An automated machine-learning pipeline for statistical matching

uRos 2023, 13.12.2023

Theresa Küntzler, Destatis



Statistical Matching? Pipeline?

Statistical Matching …

- is a technique to add variables to a data set that are initially only available in a second data source.
- allows to enrich data sets without the need for additional data collection.
- requires the estimation of a suitable model.

The statistical matching pipeline …

- enables automated and fast execution of statistical matching,
- while complying with methodological standards (e.g. nested resampling, standard hyperparameter search spaces).

Icons: freepik.com / edited

Methods and Technology

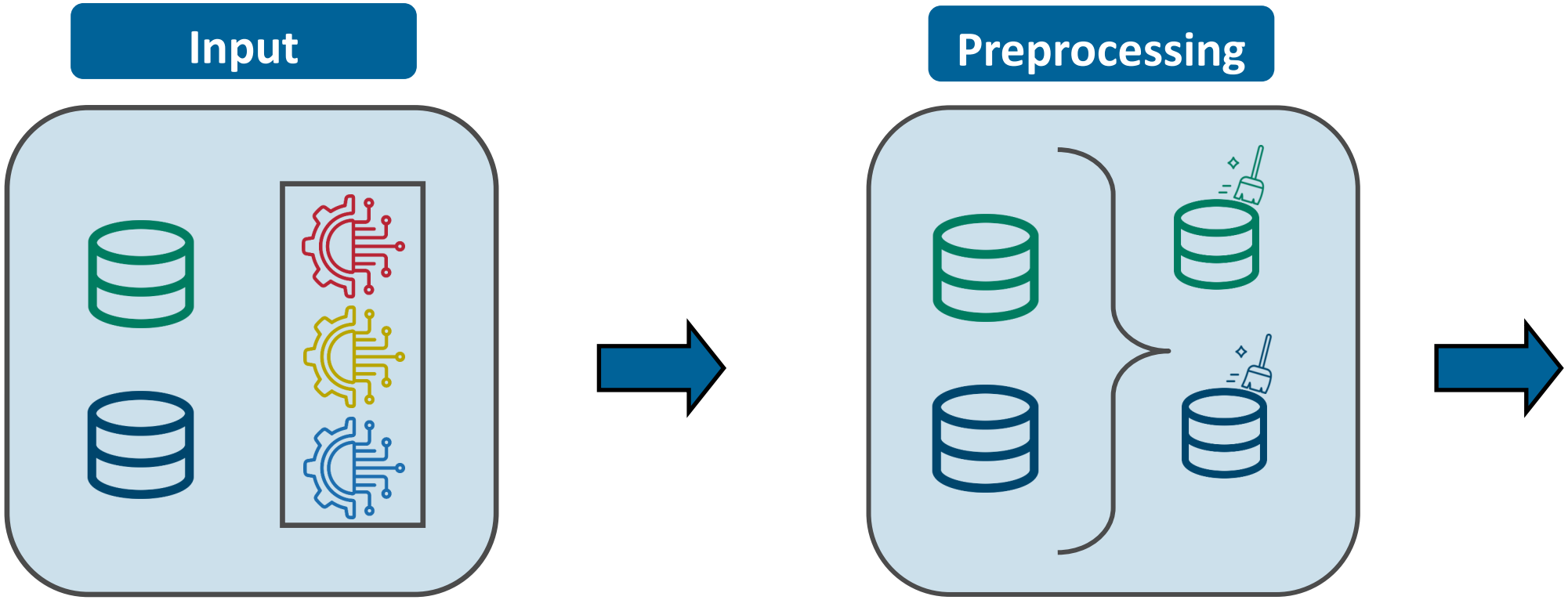
Implementation using mlr3 (Lang et al. 2019) to ensure a highly flexible pipeline

Implementation of standard hyperparameter search spaces (Becker 2023)

Selection of the best-fitting algorithm via nested resampling (Bischl et al. 2023; Simon 2007)

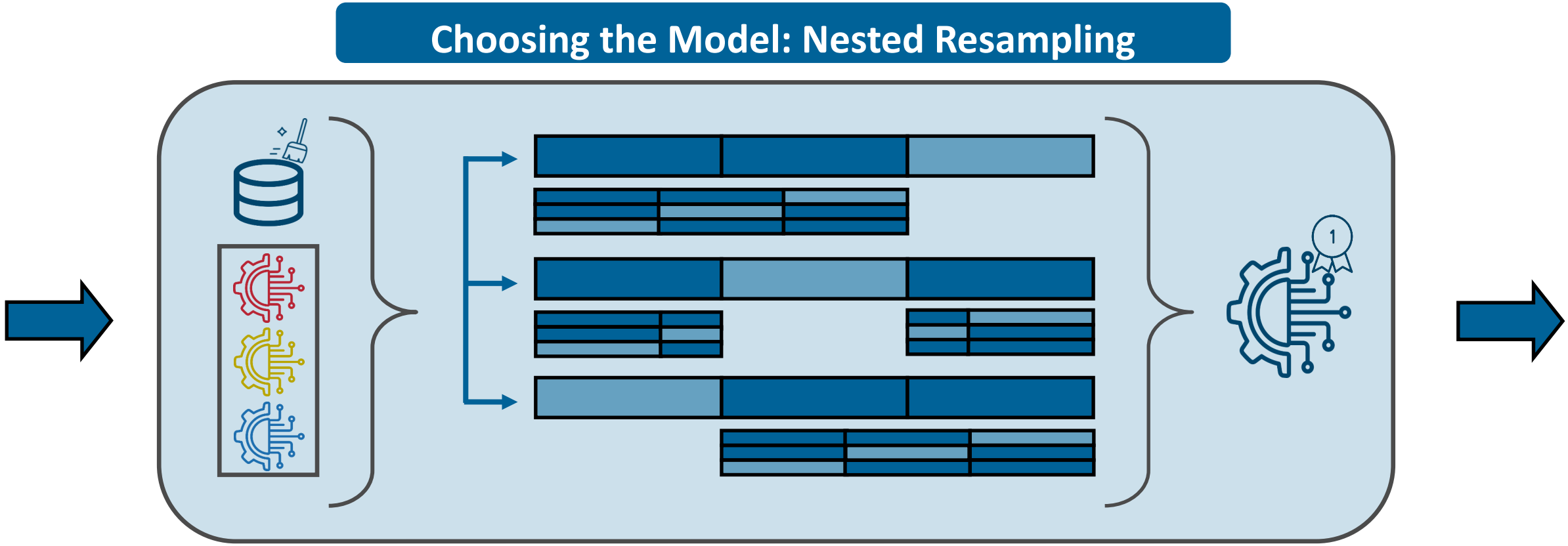
Icons: freepik.com / edited; Logos: r-project.org/logo, mlr3book.mlr-org.com, github.com/ropensci/targets, github.com/ropensci/tarchetypes, github.com/Rdatatable/data.table, github.com/tidyverse/ggplot2, github.com/tidyverse/tibble, github.com/tidyverse/dplyr, github.com/tidyverse/tidyr

The Statistical-Matching-Pipeline 1



Icons: freepik.com / edited

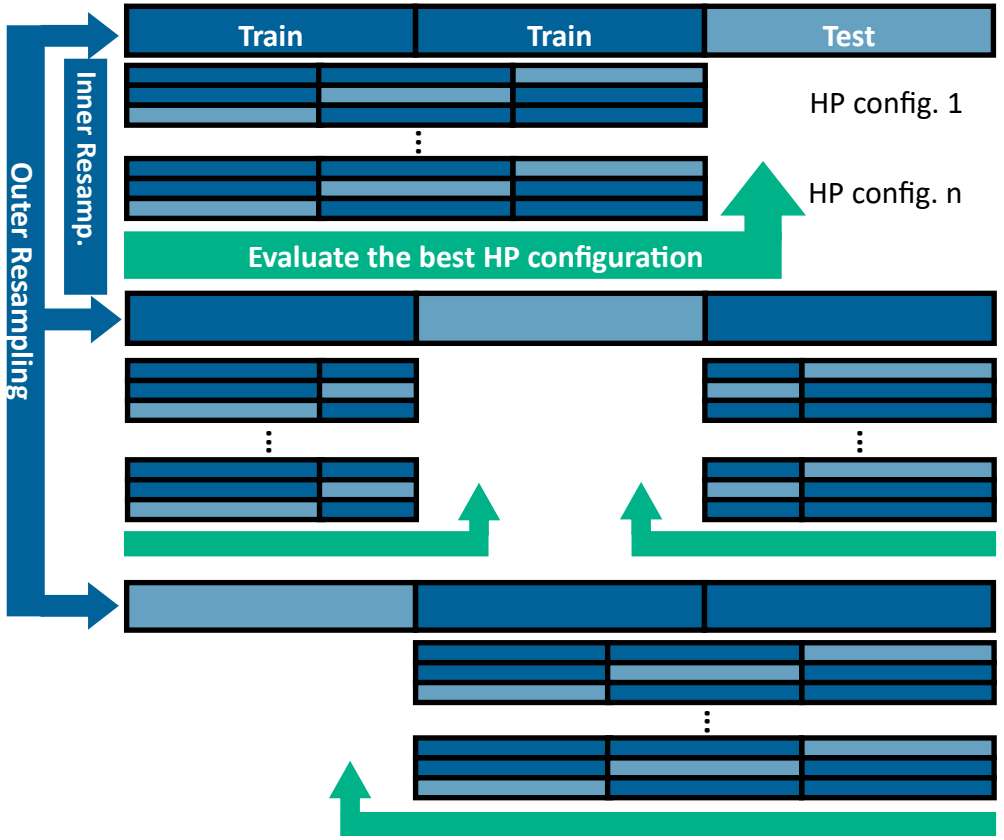
The Statistical-Matching-Pipeline 2



Icons: freepik.com / edited

Zoom: Nested Resampling

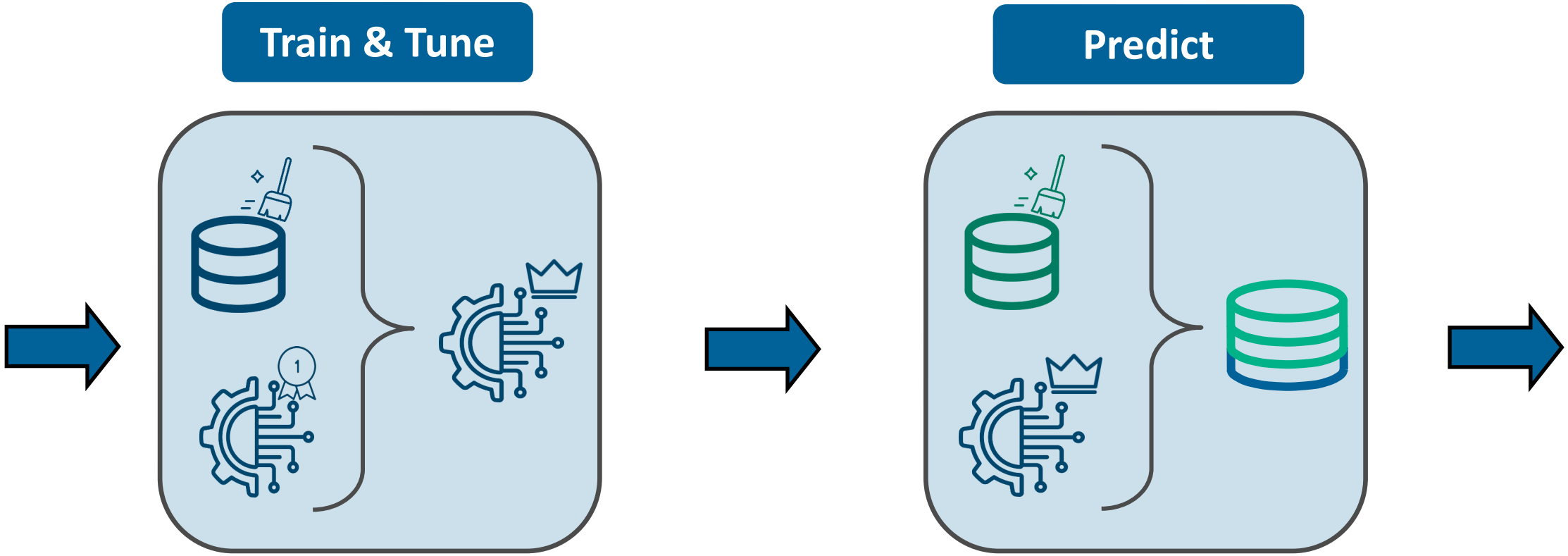
1. **Outer resampling: Split data in k folds**
2. **Inner resampling: Split outer training sets again in k folds**
3. **Optimize hyperparameters (HP) on each inner resampling (e.g. via grid search)**



4. **Evaluate each top HP-configuration with outer test set**
5. **Performance estimate: Average perfomance estimate on outer test sets**
6. **Repeat for each algorithm**

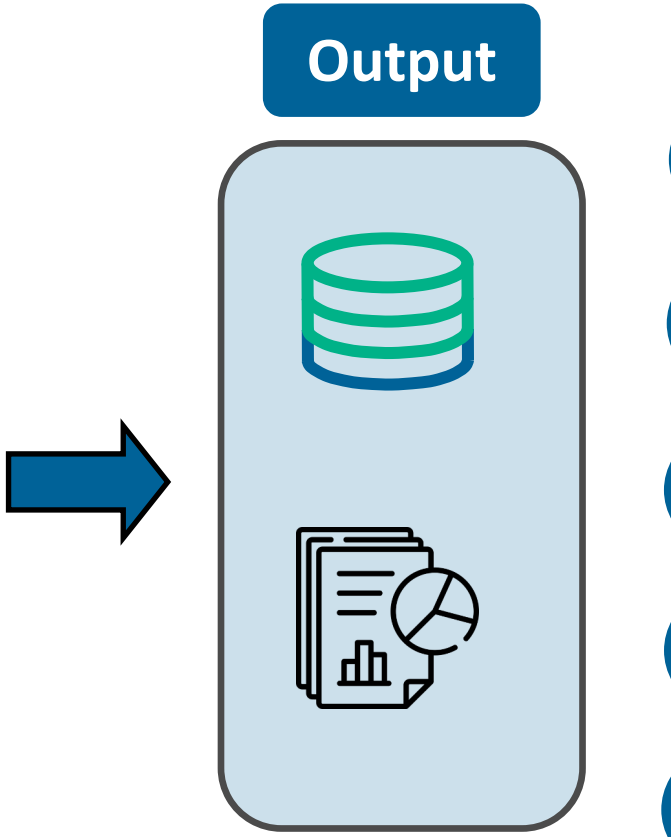
Figure after Bischl, B. et al (Eds.) (2024). "Applied Machine Learning Using mlr3 in R". CRC Press., Figure 4.5.
Bischl, B. et al. (2023). "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1484.

The Statistical-Matching-Pipeline 3



Icons: freepik.com / edited and own icons

The Statistical-Matching-Pipeline 4



- Numbers of observations and variables, variable types, missings, correlations
- Comparisons: Differences in distributions, correlations
- Inner resampling: Hyperparameters and performance measures
- Outer resampling: performance measures
- Plot of relationship of interest

Icons: freepik.com / edited and own icons

Next Up:

- Additions to the report
- Testing … developping … testing …
- Move to an R-package

Appendix

Foundations and Advances of Machine Learning in Official Statistics

International Conference

- 3rd to 5th April, 2024 in Wiesbaden (Germany)
- Topics:
  - Mathematical and statistical questions surrounding the use of ML in official statistics
  - Practical experiences and pilot projects
  - Overarching questions (ethical, legal, technical, organizational)

Call and Info: www.destatis.de/ml-conference

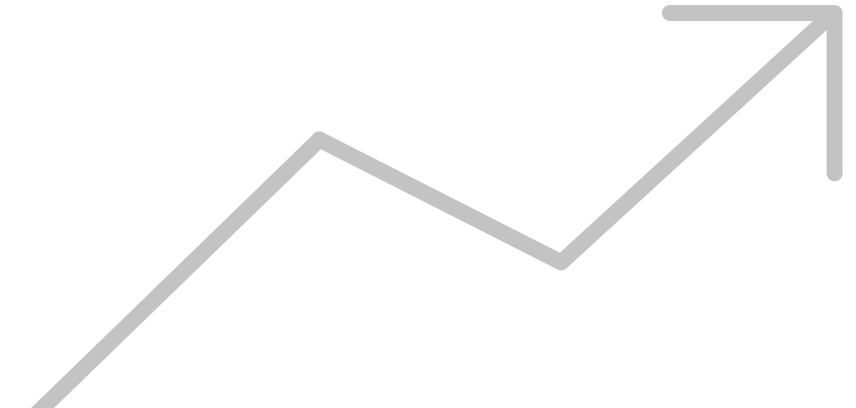
DESTATIS
Statistisches Bundesamt

Contact

Theresa Küntzler

Artificial Intelligence, Big Data

Theresa.Kuentzler@destatis.de



References

Becker, M. (2023). “mlr3tuningspaces: Search Spaces for 'mlr3'”. https://mlr3tuningspaces.mlr-org.com, https://github.com/mlr-org/mlr3tuningspaces.

Bischl, B. et al. (2023). “Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1484.

Bischl, B. et al (Eds.) (2024). “Applied Machine Learning Using mlr3 in R”. CRC Press.

Lang, M. et al. (2019). “mlr3: A modern object-oriented machine learning framework in R”. Journal of Open Source Software.

Simon, R. (2007). “Resampling Strategies for Model Assessment and Selection”. In Fundamentals of Data Mining in Genomics and Proteomics, edited by Werner Dubitzky, Martin Granzow, and Daniel Berrar, 173–86. Boston, MA: Springer US.