Kolloquium „Statistische Methoden in der empirischen Forschung“

Wann: 10. November 2020, 17:00 – 18:30 Uhr

Wo: Online

Felix Bießmann (Beuth Hochschule für Technik Berlin)

Deep learning for missing value imputation in tables with non-numerical data

With the growing importance of machine learning (ML) algorithms for practical applications, reducing data quality problems in ML pipelines has become a major focus of research. In many cases missing values can break data pipelines which makes completeness one of the most impactful data quality challenges. Current missing value imputation methods are focusing on numerical or categorical data and can be difficult to scale to datasets with millions of rows. We release DataWig, a robust and scalable approach for missing value imputation that can be applied to tables with heterogeneous data types, including unstructured text. DataWig combines deep learning feature extractors with automatic hyperparameter tuning. This enables users without a machine learning background, such as data engineers, to impute missing values with minimal effort in tables with more heterogeneous data types than supported in existing libraries, while requiring less glue code for feature engineering and offering more flexible modelling options. We demonstrate that DataWig compares favourably to existing imputation packages in a comprehensive suite of real world experiments under missing-completely-at-random, missing-at-random and missing-not-at-random conditions.