technische universität
dortmund

SFB
823

Data Science –
Is the Impact of Statistics Significant?

Katja Ickstadt and **Claus Weihs,**
TU Dortmund University, Dortmund, Germany

EDSC
Luxembourg, 8.11.2016

Premise

Statistics is the discipline to provide tools and methods to find structure in and to give deeper insight into data!

Data Science: Scientific discipline influenced by computer science, mathematics, operations research, and statistics as well as the applied sciences.

Statistical Methods: Crucial in all basic steps of data science.

Steps

Data achievement and enrichment: Design of experiments should be utilized for systematically reducing data bases and for tuning algorithms, as well as for imputation of missing data to fill gaps in data bases.

Data Exploration: Exploratory statistics is basic for preprocessing to learn about the content of a data base.

Data analysis: Statistical methods are fundamental to find structure in data and to make predictions: These methods comprise

- Multiple testing to correct for multiple usage of data,
- Classification methods to find and predict groups of data,
- Regression methods to find global and local relationships between features,
- Network analysis to learn network structure,
- Stochastic differential and difference equations to handle models from the applied sciences,
- Time series analysis to understand and predict temporal structure.

Integrative Data Analysis: Local models can be combined to global models by hierarchical and chain-like modelling.

Differential Data Analysis: Tests for comparing models, perturbation experiments, and analyzing structural breaks.

Model validation and model selection: Resampling methods and Meta-analysis as well as model averaging and the analysis of concept drifts are important tools.

Deployment: Visualization to interpret found structures, storing of models in an easy-to-update form.

Hypothesis

Yet, the role of statistics in data science is underrepresented as, e.g., compared to computer science, in particular in the areas of data achievement and enrichment as well as in advanced modelling.

Questions

What is on the data scientist`s wish list for the statistician?

What would data scientists suggest to strengthen the role of statistics in the field?

Wish List

-
-
-
-
-

Wish List

from the Statisticians point of View

- Advanced statistical models **>>** Black box tools
- Realistic, Sparse Models
- Model Comparison / Validation
- Uncertainty Reporting / Guarantees
- Automatization (when ever possible)