Applied Data Science (ADS) project acquisition form

Title of the project *

Synthetic data sets for the analysis of private data

Number of students for the project: (typically, projects have 2-3 students) *

3

Description (abstract size, approximately 200 words) *

Recent innovations allow us to generate synthetic data sets through the same techniques to solve missing data problems. For data that has restrictions with respect to privacy and confidentiality, we could use missing data methods to simply generate synthetic data versions of those data sets. Such synthetic data sets would hold the same statistical properties as the original real data but would completely circumvent the problem of privacy and data ownership. One requirement is that we cannot infer the original data values from the synthetic data; another requirement is that both data sets yield the same statistical inferences / conclusions. To test this, we answer the following two questions on simulation:

1. Can we discriminate between the original data and the synthetic data if we compose a joined data set with randomly selected rows from either set?
2. Are the inferences the same for the original data and the synthetic data. i.e. are the estimates unbiased and do they display proper variance (confidence validity).

Literature is available. No previous experience with incomplete data theory is required, students will receive a small private self-paced course in incomplete data theory to get them up to speed.

Organization name and names of internal supervisors involved. *

Utrecht University / Social Sciences / Department of Methodology and Statistics /

Names of supervisors from Utrecht University

Stef van Buuren / Gerko Vink / Hanne Oberman

Website address for additional information of organization or project *

https://github.com/amices

https://stefvanbuuren.name/fimd/

Short description of the available data. *

Data generated by simulation

Project domain *

Social and behavioural science

Optional: required courses in domain https://www.uu.nl/masters/en/applied-data-science/courses

Epidemiology and Big Data
Using data from routine care, registries, health devices and public repositories
Spatial data analysis and simulation modelling
Spatial Statistics and Machine Learning
Social Behavioural Dynamics
Network Analysis
Data Mining: Text, Images, Video
Personalisation for (Public) Media

Additional requirements (such as signing an NDA, clearance, etc.)

None

Optional: Add a pdf/word document with extra details