ADS Master Thesis Projects (15 February 2025)

G1 A data-driven approach to understanding small scale earthquakes in the laboratory

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Earthquakes are a common phenomenon in tectonically active areas, and form a danger to society. We have physical models to describe big normal earthquakes, but we have very little physical understanding of how eathquakes operate deeper in the earth where rocks are hotter. To make matters more complicated, some of those deeper earthquakes have a low magnitude, and we are unable to measure them directly. In an effort to understand such earthquakes better, we try to reproduce them in the lab. By performing mechanical tests on minerals using a machine called a nanoindenter, we can make direct measurements, and we observe that these minerals show little bursts of small earthquakes. These events show up as “jumps”, or “pop-ins”, in an otherwise smooth data curve. This project concerns the numerical analysis of the data from these lab experiments to try to come up with a predictive numerical model of when the small earthquakes will occur based on subtle changes in material properties during the experiments. An ideal outcome is to develop a model that predicts when such earthquakes happen before they do, as function of stress applied to the minerals.

G1.1 Subtopics

- Quantification of pop-ins in nanoindentation data
- Detecting mechanical property changes ahead of pop-ins in nanoindentation data

G1.2 Supervision

- UU supervisors: Alissa Kotowski, John Aiken, Hugo van Schrojenstein Lantman (h.w.vanschrojensteinlantman@uu.nl)
- External supervisors: Not indicated (Not indicated)

G1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: None given

G1.4 Additional information

- Data description: The currently available data consists of numerous lab experiments, each of which has results given as load versus depth datapoints taken during the duration of the experiment. Therefore, the entire data set will be a substantial amount of arrays with two columns, one array per experiment. More experiments are planned in the next few months.
- NDA: No NDA indicated
- Website: None

G2 Graphic model for air pollution mapping

Number of students: 2

Subject area: Health science,Geo science

External organisation: Not indicated

Current air pollution models predominantly rely on tabular data, such as road segments derived from road networks, which are typically treated as independent entities. However, the inherent connectivity of road networks offers valuable information that remains underutilized. This project proposes leveraging Graph Neural Networks (GNNs) to develop a graph-based air pollution model that captures the relational structure of road networks. The project aims at exploring popular GNN algorithms, such as Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT), to assess their effectiveness in modeling air pollution.

G2.1 Subtopics

- GAT for air pollution modelling
- GCN for air pollution modelling

G2.2 Supervision

- UU supervisors: Zhendong Yuan (z.yuan@uu.nl)
- External supervisors: Not indicated (Not indicated)

G2.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Human Network Analysis,Spatial statistics and machine learning
- Additional requirements: familiar with DL framework such as tensorflow or pytorch

G2.4 Additional information

- Data description: Hyperlocal air pollution data was collected from 2019-2020 from air quality sensors equipted on Google street view cars. Other ancillary data has already been used mulitple times in our previous pubilications.
- NDA: No NDA indicated
- Website: https://exposome.nl/news-events/news/air-quality-mapped-in-detail-by-project-air-view.html

G3 Reconstructing the research history of geography using transformer language models and knowledge graphs

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

To document, conserve and make available the research history of the human geography and spatial planning department at UU (SGPL) to a wider audience, we have collected a digital archive of student theses over the period from 1955 to 2008. The goal of this project is to use NLP extraction and knowledge graph construction methods to reconstruct part of the knowledge published in this archive, in particular, bibliometric details, geographic places studied, the studied research questions, the geographic concepts and data sources. The goal is to make these details openly accessible and queryable in terms of a knowledge graph. Students should work in a group on different knowledge aspects, and develop and test their own extraction and knowledge graph construction method for this purpose. The work can be done by 2 - 3 students.

G3.1 Subtopics

- Reconstructing the research history of SGPL: coverage of geographic places
- Reconstructing the research history of SGPL: concepts, methods and data sources

G3.2 Supervision

- UU supervisors: Simon Scheider (s.scheider@uu.nl)
- External supervisors: Not indicated (Not indicated)

G3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: Basic skills and interest in (e.g. Python-based) NLP, transformer-based language models, and knowledge graph construction (e.g. RDFLib). Some of these skills can be acquired during thesis process. Since the theses are in Dutch, some command of the language would be helpful, but native command is not a requirement.

G3.4 Additional information

- Data description: Thesis archive with approximately 375 theses of 90 pages on average, digitized using OCR technology.
- NDA: No NDA indicated
- Website: https://www.uu.nl/en/news/new-batch-of-open-science-projects-kick-off

G4 Spatial dimensions and prediction of loneliness in the Netherlands

Number of students: 2

Subject area: Health science,Geo science

External organisation: Not indicated

Loneliness has increasingly become a major health issue in modern times. The current urban spatial development, usages of technology, inadequate social cohesion, and disconnection with the natural environment have resulted in severe loneliness among different age and gender groups. However, loneliness is not only a personal phenomenon; rather, it can be influenced and determined by the environment in which people live, work, and play. The environment can be lonelygenic, which can vary over space. In existing literate, there is limited investigation on what spatial and environmental factors determine or explain the loneliness of a large population and how that varies between places. This project aims to investigate such a spatial dimension of loneliness in Dutch neighborhoods over multiple years, using diverse spatial and environmental predictors. It will critically examine how a new index of lonelygenic environment could be developed that is reproducible and help predict the future of loneliness in different areas in the Netherlands.

G4.1 Subtopics

- Spatial predictors of lonelygenic environment in the Netherlands
- Spatial dimensions of loneliness and how it varies over time and space

G4.2 Supervision

- UU supervisors: Dr SM Labib (s.m.labib@uu.nl)
- External supervisors: Not indicated (Not indicated)

G4.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data,Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: This project would require a moderate to high level of coding skills in Python and R; skills in QGIS will be useful.

G4.4 Additional information

- Data description: Most data must be extracted from Open data platforms (e.g., Open Street Map, Google Earth Engine, Environmental Atlas), Web servers, and APIs for national (e.g., PDOK, CBS) and international sources.
- NDA: No NDA indicated
- Website: None

G5 Analyzing Sydney's Car Routes and Travel Costs

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

This study examines drivers' route choices in Sydney, analyzing how uncertainty in travel times influences decisions between toll and toll-free routes across different sociodemographic groups. Using route data encoded as Google Maps polylines, we investigate how travellers respond to unreliable travel conditions through their routing decisions. By leveraging the Google Maps API and combining it with sociodemographic characteristics, we analyze how factors such as income, age, and residential location affect route choice behaviour. The analysis reveals patterns in how different population segments evaluate trade-offs between toll costs, travel time reliability, and route uncertainty.

G5.1 Subtopics

- How does travel time uncertainty influence the willingness to pay for toll roads across different sociodemographic groups in Sydney?
- How do drivers' route choice strategies and willingness to use toll roads vary across different times of day in Sydney?

G5.2 Supervision

- UU supervisors: Jaime Soza Parra (j.a.sozaparra@uu.nl)
- External supervisors: Not indicated (Not indicated)

G5.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial statistics and machine learning
- Additional requirements: Experience or willingness to learn to work with API connections

G5.4 Additional information

- Data description: 1 .- recorded route choices from Sydney drivers 2.- route characteristics obtained through the Google Maps API such as distance, typical travel times, and real-time traffic conditions for multiple days 3.- toll pricing data for Sydney's toll road network (location and API) 4.- sociodemographic information about the drivers including age, income, education level, and residential location
- NDA: No NDA indicated
- Website: None

G6 Travel guides as resource for geographic research

Number of students: 2

Subject area: Social and behavioural science

External organisation: Not indicated

Travel guides typically cover a variety of tourist information on a place or region, including tourist sights, accommodation, transportation and activities. The relative coverage of places and regions is often understood to reflect popularity among tourists (Antonescu, 2023). As such, guidebooks could be understood to reflect the stage of tourism destination development of the places included. It is not just the quantity of information that matters for understanding stages of tourism destination development but also how places are described. In fact, guidebooks are often also normative in the sense that they evaluate the attractiveness of a place or region. Resulting qualitative judgements make that guidebooks have the potential of directing tourist behaviour (Peel & Sørensen, 2016), thereby influencing tourism destination development. This project is a pilot to retrieve relevant information from a unique collection of recently digitalised lonely planet travel guides, where different editions of the Indonesia, Thailand, Italy and Greece travel guides that have appeared over time will be compared to reconstruct the evolution of tourism destinations. The challenge is to turn the descriptions of places in those OCR-readible pdfs of travel guides into a new dataset that captures for the places/regions/islands included their stage of tourism development, and syntheses the information provided in new quantitative indicators (e.g. some quantitative indicators of to what extent and how much they are covered). Next to challenges related to information retrieval and named entity recognition, there are challenges regarding the application of natural language processing to process place descriptions, such as topic modelling and a sentiment analysis. If this pilot is successful in the sense that it leads to useful and usable data, we envisage a larger follow-up scientific research project.

G6.1 Subtopics

- Obtaining geographical data from travel guides: topic modelling of place descriptions
- Ibtaining geographical data from travel guides: sentiment analysis of place descriptions

G6.2 Supervision

- UU supervisors: Evert Meijers (e.j.meijers@uu.nl)
- External supervisors: Not indicated (Not indicated)

G6.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Text and Media Analytics

- Additional requirements: For this project it would be nice if you have a passion for travel and don't think that travel guides are just for fossils - they are sold more than ever in this digital era!

G6.4 Additional information

- Data description: A series of recently digitalised lonely planet travel guides (Indonesia, Thailand, Greece, Italy), of which we have collected different editions through time. The earliest, first editions stem from the 1980s, so we can cover 30-40 years for each country.
- NDA: No NDA indicated
- Website: None

G7 Nature-based solutions on buildings and surroundings: identifying needs and solutions

Number of students: 2

Subject area: Geo science,Information and Computing science (including AI)

External organisation: Not indicated

Cities are increasingly facing adverse effects of climate change, such as increased heating, higher vulnerability to flooding, and more. Nature-based solutions (NBS) in cities, such as urban greening, are recognized as emerging low-cost yet effective solutions to fight back against the adverse impact of climate change. While urban NBS has obtained popularity as planning solutions, building level NBS implementation such as adding green roof, green façade, back garden did not receive enough attention. This is due to the fact the current state of NBS on and around buildings are less studied, also there is limited information on which buildings might have better potential to implement NBS. Furthermore, there is lack of investigation on how directly existing NBS conditions could relate to building energy performance, which could indicate where more NBS could be better implemented. Considering such knowledge gaps, this thesis project aims to develop and validate a new score for existing building's nature availability, accessibility, and integration levels. Then it will assess how such building greening score are correlated with building energy ratings and building valuation. This project will use openly available GIS and remote sensing data, along with energy rating information extracted from online. It will utilized spatial and machine learning models to evaluate these relations and identify where NBS would be better suited to implement more on buildings and its surrounding areas.

G7.1 Subtopics

- Nature-based solutions on buildings and surroundings: Creating a green rating
- Nature-based solutions on buildings and surroundings: relating NBS rating to energy and valuation

G7.2 Supervision

- UU supervisors: Dr. SM Labib (s.m.labib@uu.nl)
- External supervisors: Not indicated (Not indicated)

G7.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: None given

G7.4 Additional information

- Data description: Most data must be extracted from Open data platforms (e.g., Open Street Map, Google Earth Engine), Web servers, and APIs for national (e.g., PDOK, Funda) and international sources.
- NDA: No NDA indicated
- Website: None

G8 Developing accessibility profiles to understand travel and activity behaviour

Number of students: 2

Subject area: transportation science

External organisation: Not indicated

Accessibility is a key concept in transportation and land use planning. Accessibility as a concept dates back to the 1960s and has been defined as the ease with which individuals can reach specific destinations (e.g. shops, schools, jobs) given a certain location. Operationalisations of accessibility include cumulative accessibility measures and gravitation based accessibility measures, but also more straightforward measures such as the distance to the nearest destination of a specific type (e.g. a grocery shop) (Geurs and Van Wee, 2004). What these accessibility measures have in common is that they express accessibility to one activity/destination type (e.g. shops, schools, hospitals). However, people's ability to lead their desired life depends on the combination of accessibility levels to a variety of activities, which can be termed the accessibility profile of a location. According to this idea, some locations offer good accessibility to e.g. shops and cultural facilities and less accessibility to green facilities, whereas for other locations this may be the other way around. Whether a location 'suits' a person living there, then depends on his/her needs to participate in specific activities, according to their lifestyle. In particular, the accessibility profile may indicate limitations in accessibility from certain people. In addition, accessibility can be defined by different travel modes. Obviously, accessibility by car will usually be better than accessibility by public transport or bicycle. This may also affect the fit of a person and the accessibility profile of their residential location, based on their access to/ownership of travel modes. To formalize things a bit, each destination has an accessibility profile which is a combination of accessibilities Aam (to activity a by mode m), where the accessibility Aam can take the form of e.g. a cumulative opportunity measure or a gravitation based measure. Developing accessibility profiles is highly relevant from a policy perspective to support accessibility based transportation planning, which enables insight into how accessibility enables activity participation and influences travel mode use. The questions addressed in this thesis project are: 1. What accessibility profiles of locations can be defined based on accessibility measures Aam; 2. How can accessibility profiles be used to: a. Understand individuals' travel behaviour b. Understand individuals' activity participation 3. Can we identify individuals that 'underperform' (i.e. participate less in out of home activities than the average) according to their accessibility profile (based on their residence) and what factors explain this underperformance. The basic idea underlying this project is to carry out a latent class cluster analysis (LACC) (see Muchlisin et al., 2024) in which accessibility measures Aam are variables in the measurement model, and travel and activity characteristics are variables in the structural model. By estimating a LACC model, clusters of accessibility profiles are obtained, that best predict travel behaviour and/or activity participation. Reference Geurs, K. T., & Van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport

geography, 12(2), 127-140. Muchlisin, M., Soza-Parra, J., Susilo, Y. O., & Ettema, D. (2024). Unraveling the travel patterns of ride-hailing users: A latent class cluster analysis across income groups in Yogyakarta, Indonesia. Travel Behaviour and Society, 37, 100836.

G8.1 Subtopics

- 1. What accessibility profiles of locations can be defined based on accessibility measures
- 2. How can accessibility profiles be used to: a. Understand individuals' travel behaviour b. Understand individuals' activity participation

G8.2 Supervision

- UU supervisors: Dick Ettema (d.f.ettema@uu.nl)
- External supervisors: Not indicated (Not indicated)

G8.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences,Spatial data analysis and simulation modelling
- Additional requirements: None given

G8.4 Additional information

- Data description: Data about accessibility of activity locations is derived from existing data sets in UU. Data about activities and travel will be obtained from the Dutch national travel survey ODIN.
- NDA: No NDA indicated
- Website: None

G9 Mapping the suitability for residential development in the Netherlands using Formal Concept Analysis

Number of students: 3

Subject area: Geo science,Information and Computing science (including AI)

External organisation: Not indicated

The Netherlands is confronted with a ubiquitous shortage of housing which causes a large pressure on the housing market. To address this, the current government has voiced the ambition to construct 100,000 new houses per year. Part of the challenge is to find the space for new residences and some areas are more suitable than others for many different reasons. Geographic Information Systems (GISs) offer a variety of landuse suitability analysis (LSA) techniques, but a limitation is that these techniques tend to treat suitability criteria as independent from one another. The reasoning behind suitability of land for residential construction is usually highly complex, and disentangling this complexity into independent and unambiguous suitability factors is difficult. Formal Concept Analysis (FCA), a novel mathematical framework, may provide a solution by automating the factor clustering process, offering a principled approach for aggregating factors into suitability measures. The approach revolves around a concept lattice, a particular kind of hierarchically-structured graph. However, there is a lack of integration of FCA techniques into GISs, which means new methods need to be developed and their modeling accuracy needs to be validated. The research aim of this thesis project is to implement FCA methods into residential LSA to enhance the identification, classification, and prioritization of residential development sites. This topic is free for anyone interested in pursuing this research aim, but in particular the following objectives may be considered: - Identify and aggregate factors for residential LSA using a combination of GIS and FCA techniques and validate the results - Develop methods for introducing weighting into FCA-based LSA (e.g., using fuzzy FCA) - Develop map visualization techniques to display the information in a concept lattice on a 2-dimensional cartographic map. These topics are suitable for those with affinity with (the automation of) reasoning about concepts and spatial analysis. For the implementation of LSA and FCA, adequate proficiency in python is required. Experience with spatial analysis, e.g., in the INFOMSDASM and/or INFOMSSML courses is an advantage, but not prerequisite. For more information, feel free to send an e-mail to e.j.top@uu.nl For further reading, please see: https://en.wikipedia.org/wiki/Suitability_analysis https://en.wikipedia.org/wiki/Formal_concept_analysis https://en.wikipedia.org/wiki/Visual_variable (In particular for the third objective) Malczewski, J. (2004). GIS-based land-use suitability analysis: a critical overview. Progress in planning, 62(1), 3-65. (overview of LSA, but needs supplementation with recent publications) Karna, B. K., Shrestha, S., & Koirala, H. L. (2023). GIS based approach for suitability analysis of residential land use. Geographical Journal of Nepal, 35-50. (Recent example of using GIS for residential LSA) Ganter, B., & Wille, R. (2012). Formal concept analysis: mathematical foundations. Springer Science & Business Media. (Foundational book on FCA, other resources available)

G9.1 Subtopics

- Identification, aggregation, and validation of LSA factors using FCA
- Calibration and weighting techniques in FCA-based FCA
- Cartographic map visualization of concept lattices based on suitability factors

G9.2 Supervision

- UU supervisors: Eric Top (e.j.top@uu.nl)
- External supervisors: Not indicated (Not indicated)

G9.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: Understanding the technique of Formal Concept Analysis may require aptitude for abstract reasoning, logic, and philosophy

G9.4 Additional information

- Data description: Data is available from various sources ranging from CBS land use and neighborhood data to satellite imagery.
- NDA: No NDA indicated
- Website: None

G10 Mapping jobs and skills for green and digital transitions in Europe

Number of students: 3

Subject area: Geo science,Social and behavioural science,Information and Computing science (including AI),Law, Economics and Governance

External organisation: Not indicated

The green and digital transitions require an acceleration of the development and deployment of new green and digital technologies worldwide. At the same time, the impact of green and digital transitions on differs across jobs, industries and regions. In order to achieve a timely and just transition towards net zero, we need better evidences to inform policy making. The availability of large volume text data such as patents, scientific publications, trademarks, policy documents, job posts and web text open up the opportunity. With the recent enhanced computational capacity and advancements in language models, we can better map the jobs and skills related to green and digital transitions, policy instruments that support the transitions, and the potential impact of on labour markets and economies.

G10.1 Subtopics

- Which skills and jobs are relevant for the digital transitions?
- Which skills and jobs are relevant for the green transitions?
- The distribution of green and digital skills and jobs in European regions and sectors.

G10.2 Supervision

- UU supervisors: Deyu Li (d.li1@uu.nl)
- External supervisors: Not indicated (Not indicated)

G10.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Casual Inference Methods for Policy Evaluation,Spatial statistics and machine learning ,Text and Media Analytics
- Additional requirements: None given

G10.4 Additional information

- Data description: ready to use full text patent data from United States Patent and Trademark Office (USPTO) and European Patent Office (EPO) ready to use dataset of job vacancies in different European countries. descriptions of skills and occupations in European Skills/Competences, Qualifications and Occupations (ESCO). projects funded by the European Structural Investment Funds that aimed at reskilling and upskilling
- NDA: No NDA indicated
- Website: None

G11 High-resolution land cover data processing for use in large-scale land surface modelling

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

The representation of land use/cover (LULC) in land surface modelling is vitally important for the accuracy of hydrological process representation and results, considering the huge effects that LULC changes such as deforestation and agriculture have on hydrological cycles through alteration of evaporation and water runoff. Sources from which to retrieve the information required for land surface modelling have proliferated in recent years due to an increase in global land cover products derived from satellite data, along with the resolution of these products. We developed a methodology which uses recent high-resolution land cover data as the basis for creating data layers required for land surface modelling, considering different LULC classifications and temporal requirements.

G11.1 Subtopics

- Validation of global high resolution agricultural LULC data using observed national data
- Construction global high resolution LULC scenarios using existing data

G11.2 Supervision

- UU supervisors: Frances Dunn (f.e.dunn@uu.nl)
- External supervisors: Not indicated (Not indicated)

G11.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: Processing of large data and usage of high performance computing infrastructure is a plus.

G11.4 Additional information

- Data description: We produced global layers of 15 LULC classes as fraction data at 30sec as PCRaster maps, with a focus on accuracy of agriculture, particularly rice paddy areas. This needs validating using e.g. FAO national data, with additional questions on what is the consistency of the classification over time, how can we detect crop cultivation from this, and how to use the data to construct datasets of historical trends or future projections.
- NDA: No NDA indicated
- Website: None

G12 Global River Morphology: Correlating Environmental Influences

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Naturally occurring rivers have different forms or 'morphologies' defined by: the sinuosity of the channel, presence of depositional bars and the number of channels. With the increasing resolution and coverage of satellite imagery it is now possible to collect all relevant data on these morphological parameters but further work is needed to classify rivers into distinct types such as the commonly recognized varieties of meandering (sinuous single channel), single channel without regular sinuous pattern, braided (river bed contains mid-channel bars) and anastomosing (river consisting of multiple separate interweaving channels). A previous study by Nyberg et. Al. (2023) showed that image-recognition probably enables this type of automated classification but further work is needed to develop an image-recognition method for clustering river morphologies from un-classified binary images of rivers. This project proceeds from a former Applied Data Science project where self-supervised learning was used to cluster binary satellite images of river morphologies (thesis accessible at https://studenttheses.uu.nl/handle/20.500.12932/47120). Nyberg, B., Henstra, G., Gawthorpe, R.L. et al. Global scale analysis on the extent of river channel belts. Nat Commun 14, 2163 (2023). https://doi.org/10.1038/s41467-023-37852-8

G12.1 Subtopics

- self-supervised clustering of river morphologies from binary images
- self-supervised clustering of river morphologies from satellite photos

G12.2 Supervision

- UU supervisors: Daan Beelen (d.beelen@uu.nl)
- External supervisors: Not indicated (Not indicated)

G12.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial statistics and machine learning
- Additional requirements: no

G12.4 Additional information

- Data description: Global data on river morphologies from SWORD: https://www.swordexplorer.com/ as well as various publically available geographical datasets like DEMs https://www.earthdata.nasa.gov/data/instruments/srtm
- NDA: No NDA indicated
- Website: None

G13 Correlation of River Morphology with Environmental Factors

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Rivers naturally come in multiple morphological types like straight, meandering, braided and anastomosing. An ongoing research effort has now for the first time mapped all river morphologies worldwide using satellite imagery. In this project the student will increase our understanding of what determines the distribution of river morphologies by correlating river morphology data to geographical data on relevant environmental factors like slope, width, substrate type and climate. The aim is to use correlation matrices and support vector machines to spatially and numerically correlate morphological parameters like sinuosity, nr of river channels and braiding index to potentially relevant geographical data. The correlation will be done on the basis of all of the world's 150k mapped river reaches (reach: section of river with roughly constant morphology) and mapped nodes 1.3M (a short section of river with one bend). To achieve this, large geographical datasets have to be made concordant to enable correlation. Large computations can be performed on the faculty's central facilities.

G13.1 Subtopics

- Correlation of River reach morphology with Environmental Factors
- Correlation of River node morphology with Environmental Factors

G13.2 Supervision

- UU supervisors: Daan Beelen (d.beelen@uu.nl)
- External supervisors: Not indicated (Not indicated)

G13.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial statistics and machine learning
- Additional requirements: no

G13.4 Additional information

- Data description: River morphological information from: https://www.swordexplorer.com/ Various geographical data like: https://www.earthdata.nasa.gov/data/instruments/srtm
- NDA: No NDA indicated
- Website: None

G14 Predicting landscape change of fast-developing natural systems created in a laboratory environment.

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Estuaries, or river mouths dominated by ebb and flood, such as the Western Scheldt show dynamic patterns of channels and sand bars. These are quite relevant as shipping lanes to large ports and as habitats for fish and birds. However, the dynamics of these patterns are elusive, in part because of lack of data. Here we use overhead image series from laboratory experiments in the Metronome (www.uu.nl/metronome), which is a flume of 20 by 3 m in which estuaries form in the sand. The experimental estuaries are similar in all relevant aspects to the channels and sand bars in real systems. We are looking for creative students to identify and try methods to analyze spatial timelapse data of channel motion and sand bar change, to study whether these changes are predictable with machine learning. These techniques are urgently needed in research projects with societal stakeholders in the Wadden Sea and the Western Scheldt. During the project, the Metronome facility will be in action so you can visit it and dream away on the shores of the mini-beach.

G14.1 Subtopics

- Characterizing water depth from timelapse imagery and bathymetry data for an entire timeseries.
- Applying machine learning on timelapse imagery to predict future channel or sand bar migration.

G14.2 Supervision

- UU supervisors: Eise Nota (e.w.nota@uu.nl)
- External supervisors: Not indicated (Not indicated)

G14.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: None given

G14.4 Additional information

- Data description: We have for several experiments a lot of new gridded bathymetric and (calibrated, gridded) photographic data. The bathymetric data, was collected repeatedly on a dry bed. The imagery was collected much more frequently during the experiment with blue dye in the water so that the channel (depths) are visible and changes incremental. This culminated into a large dataset of over 220.000 simulated tidal cycles and 2.000.000 unique images.

- NDA: No NDA indicated
- Website: https://vimeo.com/383753627?turnstile=0.sOJSjTRN57FxXv_x9gxq11hL4_a3QIVd_4HAoi6qrb-J8Vz_Nas0RCe6vM4a_EGAwnjKyNJUJdRAJtXv5KzqBXf0e5uppX1AbYd1sIXXFfQbLep-ikxlhOt417K-qgDJOtykoCsvFO54ZS3RsUamfc5pYG4PL-q2OxXQw217Ej91KC4Wp21t6h1cB8WA2HwIL5lPyIRydKSCz7489YjTk_JgOmooH9RAORv_X3ynkJ40k3eAGr7TVrgQZySlCUbDMm1kdOmJvou9e_hbgbiQHN3Bei-TjRFK3qsUd9dW52urHr7NSSuSTg3IwIT0__OIB1GY_qdD8_XAeea_ZneIm3T648NMY-nqiNwl6rzz1QO40imRUcsGc5Gk0hwTJujUpMVE1ujh8Baa-z2PANEG-7LNtgx7tOqvEU9VuxmRTbEjTPbJNAT-XbjohwtEbCjv7YHyrpJLsOaafnuorWqg0ExSV_kCaCd8q_gIaoFdcEcFAorWkhLv1ZW0JpijqTrnt2tlpdyypsM6igqPqq7ucripvx9zGFtL5h46caoRz73hkVCo_TydIUFFtcIQ2_p5sERM9apHToqlm8FpLfyImNEox3IkMvfYuNxSBPxwn0_RPAZcJU0AMZZIUnghG9MLxr7hpIM4sSo_-SMf-HpupzPZZ4hUr8J0VL9r5Py5iYzHaWEd1_PgKvoyiN7E0hom8Q0_HpOI1EqpJMdbBvWRVhuHTYyLYJBr9ycR-nGr6BeRdpZ-ZsNAXP96Y645HVzDFPSEE6_zyYieYvbFZWYBKRxSFqMQ2NPs5B7g2ZcFMpwzlK9-7_MYpOds_STmEhAi4TKbNw2-ax5YDEXXHvAgTWzCpNHrSVWgthizRnXHUsQ.DRLmGCZdew6OdaofeG5MpQ.3c39bf9856e09b41fb6f5649b28e18bb26a938b370c0cc588311356a20d4422e

G15 A Physics Informed Neural Network (PINN) method with adaptive training data selection

Number of students: 2

Subject area: Geo science,Information and Computing science (including AI),Other:

External organisation: Not indicated

Partial differential equations (PDEs) are used in physics and many fields of science to describe the behavior of a various of phenomena, such as heat transfer, fluid mechanics, solid mechanics, electromagnetic and etc. Therefore, solving PDEs is a fundamental step to in exploring a wide range of physical phenomena and engineering problems. However, solving these equations is challenging that mostly relies on computationally expensive numerical methods. Recently, a new method called Physics-Informed Neural Network (PINN) has been introduced that uses artificial neural networks to solve PDEs. In this method, the governing equations and boundary conditions (that is, any known values in the study area) are directly embedded into the loss function and trained on allocated points in the area of interest. The accuracy of the results, though, is very sensitive to the point allocation strategy; for example, a random or gridded approach might miss crucial areas where the functions change abruptly. For our study, an adaptive training data selection algorithm is proposed. This approach primarily aims to improve the PINN method by adaptively selecting training data. Initially, the PINN method is trained with randomly selected training points. Then, an error estimate is then implemented based on the residuals of the PDEs (or loss function). Subsequently, the training data is adjusted to include more training data from regions with higher errors. Finally, the PINN model is retrained using the updated training set. This sequence can be applied recursively to arbitrarily improve model performance. With this method, the training data is intelligently selected to better capture the problem's underlying physics, in contrast to a random data split. The adaptive data selection approach for PINN is expected to yield more accurate solutions to PDEs, particularly for problems involving steep gradients and shocks, as the training data can be further refined around these regions. For this project, the study will focus on solving 1D and 2D advection-diffusion equations in both steady-state and transient cases (time-independent and time-dependent PDEs).

G15.1 Subtopics

- Training data re-selection (re-meshing)
- Training data refinement

G15.2 Supervision

- UU supervisors: Prof. dr. Derek Karssenberg (d.karssenberg@uu.nl)
- External supervisors: Not indicated (Not indicated)

G15.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Familiarity with machine learning and deep learning frameworks such as TensorFlow and/or PyTorch is required

G15.4 Additional information

- Data description: A basic PINN model for solving PDEs will be provided, and the student will mainly focus on developing an adaptive training data selection approach and integrating it with the PINN model.
- NDA: No NDA indicated
- Website: https://www.computationalgeography.org/

G16 Machine learning to identify water pollution hotspots in surface and groundwater globally

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

This thesis focusses on identifying water pollution hotspots in water systems globally (groundwater and surface water). Different machine learning techniques (e.g. Random Forest, LGBM regression) will be tested and the performances will be evaluated. A large global water quality monitoring dataset of water quality measurements for different pollutants (Jones et al, 2024; https://iopscience.iop.org/article/10.1088/1748-9326/ad6919) and various hydrological (e.g. streamflow), landuse and socio-economic datasets (e.g. population, GDP) will be used.

G16.1 Subtopics

- Machine learning to identify water pollution hotspots in groundwater systems globally
- Machine learning to identify water pollution hotspots in surface water systems globally

G16.2 Supervision

- UU supervisors: Prof. Dr. Michelle van Vliet (m.t.h.vanvliet@uu.nl)
- External supervisors: Not indicated (Not indicated)

G16.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: None given

G16.4 Additional information

- Data description: Global water quality monitoring data of for surface and groundwater (see Jones et al, 2024 for publication; https://iopscience.iop.org/article/10.1088/1748-9326/ad6919 In addition, global datasets of streamflow, groundwater levels, landuse, wastewater treatment, population counts, GDP and other water quality drivers will be provided by the supervisors or downloaded from online databases.
- NDA: No NDA indicated
- Website: None

G17 Hybrid modelling of mountain water resources

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Machine Learning based models that predict water resources such as river streamflow, flooding, or drought, mostly outperform process-based (physically-based) models regarding predictive performance, but have the disadvantage that they are mostly black-box models. This means they do not provide understanding of how the predictions are made by the model, i.e. the processes used and the internal system states and fluxes remain unknown. Dynamical system neural networks are a form of hybrid modelling that contribute to solving this issue by combining AI and process-based (physically-based) modelling. In this topic you will build upon an existing dynamical system neural network model of mountain hydrology. The model is structured like typical process-based models containing linked compartment models for components such as snow, subsurface water, evapotranspiration. However, the processes represented by these compartments are identified by using neural networks. This results in a grey-box model where the general structure of the system is known, as well as all internal storages and fluxes (e.g. snow cover, melting rate, subsurface water). The model thus far has been tested only on a single catchment in the Austrian Alps. In this project, you will apply the model to a number of other catchments, to evaluate its performance. If you are interested, it will also be possible to extend the model, to test it against artificially generated data, or to compare its performance to purely process-based models or purely AI-based models. For this topic it is required you have a background in machine learning, in particular machine learning with neural networks. The current model has been written in PyTorch and the project thus involves coding in Python.

G17.1 Subtopics

- Evaluation of dynamical system neural networks across a range of Alpine catchments
- Dynamical system neural network models trained on large data from numerical system modelling

G17.2 Supervision

- UU supervisors: Prof Dr Derek Karssenberg (d.karssenberg@uu.nl)
- External supervisors: Not indicated (Not indicated)

G17.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: None given
- Additional requirements: Python; PyTorch; Geosciences, Environmental Science, Physics or a related domain

G17.4 Additional information

- Data description: Existing climate reanalysis data available in the group or from https://www.ecmwf.int/en/era5-land. Streamflow data from https://grdc.bafg.de
- NDA: No NDA indicated
- Website: https://www.computationalgeography.org

G18 Creating a hybrid AI–land subsidence model for parameter estimation

Number of students: 2

Subject area: Geo science,Information and Computing science (including AI)

External organisation: Not indicated

Land subsidence is a critical problem in coastal and deltaic regions where sediment deficits and altered water tables, often driven by human activities, threaten both communities and ecosystems. In the Netherlands, the situation is particularly severe as approximately 50% of it coastal-deltaic plain now lies below mean sea level mainly due to soil consolidation, peat oxidation accumulated over centuries. The ongoing subsidence continues to increase economic costs. Numerical models such as Atlantis (Bootsma et al., 2020) provide valuable insights into the interactions between soft soil consolidation, peat oxidation, human interventions (e.g. agricultural drainage) in a changing climate. However, these models require spatially distributed parameters for their equations that are difficult and costly to obtain in traditional ways. Recently advances in remote sensing could assist in that task by providing observational data, but incorporating these large datasets into existing models is computationally demanding and remains a significant barrier to proper parameter estimation. Recently, much interest has been raised in Earth Sciences to develop differentiable modelling tools with the aim to integrate machine learning and physical models (Shen et al., 2023), which could streamline the data integration process. One potential application of this type of hybrid models is to replace physical parameters with an Artificial Neural Network (ANN) module that can be trained directly from the physical model output. Nevertheless, most physical models (including Atlantis) are not written in a language that supports automatic differentiation, which is necessary for backpropagating the loss to update the ANN weights during training. This project involves the development and testing of a differentiable version of the Atlantis land subsidence model in PyTorch, and the development of a hybrid modelling setup that implements an ANN module in place of the oxidation parameter in Atlantis, trained using remotely sensed observations of land subsidence in the municipality of Krimpenerwaard. Literature Bootsma, H. a. (2020). Atlantis, a tool for producing national predictive land subsidence maps of the Netherlands. Proceedings of IAHS, 415–420. doi:https://doi.org/10.5194/piahs-382-415-2020 Shen, C. a.-J. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. Nature Reviews Earth & Environment. doi:10.1038/s43017-023-00450-9

G18.1 Subtopics

- Development and benchmarking of a differentiable version of the Atlantis land subsidence model in PyTorch
- Implementation of an ANN module to replace the oxidation parameter in the (PyTorch-based) Atlantis land subsidence model

G18.2 Supervision

- UU supervisors: Oriol Pomarol Moya (o.pomarolmoya@uu.nl)
- External supervisors: Not indicated (Not indicated)

G18.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: None given
- Additional requirements: Having at least some experience in PyTorch and/or extensive Python knowledge is required. Having followed one or more courses in the spatial track is a plus.

G18.4 Additional information

- Data description: The data contains mainly remote sensing products such as InSAR vertical displacement measurements at high spatial and temporal resolution (brief overview: https://nwa-loss.nl/en/work-packages/wp1-measuring-and-monitoring/wp1-1-insar-time-series/). Also local extensometer and CO2 emissions measurements from various locations in the Netherlands. InSAR data is available at one observation per six days whereas other variables are collected at sub daily resolution collected since 2018. For further information please refer to: https://www.nobveenweiden.nl/wp-content/uploads/2021/11/NOBV-Data-analyse-2020-2021.pdf.
- NDA: No NDA indicated
- Website: None

G19 Artificial Intelligence to create maps - Can we make machine learning algorithms explicitly spatial?

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

In numerous domains, the creation of maps relies on spatial prediction. Observations, such as rainfall levels, air pollutant concentrations, or soil pH, are typically available only at discrete point locations. To generate continuous maps for e. g. arable fields, entire regions, countries, or even the entire globe, one uses other spatial predictor data like satellite images or elevation models and also tries to exploit "similarities" in the data. According to a fundamental principle of geography observations closer to each other are often more similar than observations further away. The extent to which close observations are similar, and the distance over which this relationship holds, is explicitly quantified through the concept of spatial auto-correlation using a kriging variogram. Machine learning (ML) techniques were not initially designed for spatial mapping and lack inherent knowledge of the relative locations of observations. Nevertheless, ML has become the predominant choice for spatial mapping due to its user-friendliness and good performance, particularly with large datasets featuring numerous predictor variables. Consequently, many studies resort to unsatisfactory ad-hoc methods to incorporate location information. However, ML algorithms could potentially directly account for spatial autocorrelation by integrating the variogram function into the learning process of the ML algorithm. The thesis focus is on the implementation of "ML kriging" using an existing R package, to apply it to environmental mapping and on the investigation of the prediction performance and behavior. Core part of this project can only be done in R.

G19.1 Subtopics

- Behavior and performance of ML kriging for detailed mapping of small areas
- Behavior and performance of ML kriging for mapping of large areas with many predictors

G19.2 Supervision

- UU supervisors: Madlene Nussbaum (m.nussbaum@uu.nl)
- External supervisors: Not indicated (Not indicated)

G19.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: None given

G19.4 Additional information

- Data description: To make conclusions comparable to other publications you will first work on soil test datasets often used for kriging methods (e.g. Meuse, Ebergoetzen datasets). Moreover, point observations from soil surveys in Switzerland will be provided (depending on case study 1000-3500 observations) along with a large number explanatory predictor maps (multi-scale terrain analysis, geology, climate). Depending on interest of the students, the analysis of other datasets of similar type can be discussed.
- NDA: No NDA indicated
- Website: None

G20 Where to take samples? Optimal choice of sampling location for mapping with machine learning

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Optimal selection of locations to make new observations of a property or to take samples for laboratory analysis is a common challenge in various mapping domains. Such endeavors are prominent in earth sciences and physical geography, encompassing activities such as soil or geological surveys, geomorphology studies addressing natural hazards, coastal formation, and extend beyond to domains like air pollution mapping. Traditionally, the positioning of new samples has been extensively explored within classical geostatistics and optimization strategies area available for kriging methods. However, contemporary mapping in most domains now relies on machine learning (ML) model prediction. ML requires a different spatial configuration of sampling. Although general optimization approaches to sample for ML predictions are available, several relevant open questions persist, especially with regard to practical application of such designs: 1) How sensitive are user choices regarding the environmental criteria used for a sampling design (e.g. resolution of slope map, type of remote sensing index) and is it possible to formulate general recommendations? 2) How can an sampling design for ML be computed in adequate time? How much less accurate is the final map depending on what we simplify? The thesis will focus on implementation of different variants of ML optimized sampling designs. The final prediction accuracy will be used as benchmark to compare performance, hence spatial mapping is also part of the thesis.

G20.1 Subtopics

- Evaluate sensitivity of user choices for ML optimized sampling design
- How much do we loose if we create simpler sampling designs?

G20.2 Supervision

- UU supervisors: Madlene Nussbaum (m.nussbaum@uu.nl)
- External supervisors: Not indicated (Not indicated)

G20.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: None given

G20.4 Additional information

- Data description: Sampling is too expensive to repeat the needed large number of times to form conclusions. Therefore, in the thesis you will simulate a map that is then taken

to be “ground truth”. To make this simulation as realistic as possible it is based on actually sampled data from German Ebergötzen case study area or from Swiss soil surveys (1000-3500 observed locations) and a large number of predictor variables. Depending on interest of the students, the use of other datasets of similar type can be discussed.

- NDA: No NDA indicated
- Website: None

G21 Physical or chemical constraints for machine learning based soil mapping

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

In numerous domains, the creation of maps relies on spatial prediction. Observations, such as rainfall levels, air pollutant concentrations, or soil pH, are typically available only at discrete point locations. To generate continuous maps for e. g. arable fields, entire regions, countries, or even the entire globe, one uses other spatial predictor data like satellite images or elevation models and also tries to exploit relationships in the data. Those relationships are commonly built up in a fully data driven way. In many domains, such as soil mapping, there are rarely enough data points to reliably train the full “behavior” of the mapped property. As a result we will have predictions that may not be possible from a chemical or physical viewpoint. For example, it is chemically not possible to have calcareous material (>0 g/mg) in soil at a pH of below 6 (due to buffer reactions). A machine learning based prediction will still result in soils being predicted with this impossible combination. The goal is hence to “teach” the prediction model about the known relationships, even for a small number of training data points. One way to achieve that, is to include constraints in the loss function. At the same time we evaluate models based on mean error performance by for example using R2. Such error metrics do not account for physically or chemically non-feasible model output. Expanding validation metrics by properly accounting for not just statistically correct predictions is therefore another objective to explore.

G21.1 Subtopics

- Evaluate physical constraints in neural network loss functions
- Explore model evaluation metrics that consider physical sound prediction

G21.2 Supervision

- UU supervisors: Madlene Nussbaum (m.nussbaum@uu.nl)
- External supervisors: Not indicated (Not indicated)

G21.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: None given

G21.4 Additional information

- Data description: Soil data sets will be made available from recent campaigns or from soil archives, e. g. from Swiss study sites in the Canton of Zurich (3500 surveyed locations). Alternatively, test data sets are available from German Ebergötzen case study

area. Depending on interest of the students, the use of other datasets of similar type can be discussed.

- NDA: No NDA indicated
- Website: None

G22 Subsidence-related damage risk assessment across the Netherlands using different machine learning approaches

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Land subsidence in delta regions can lead to various problems, including damage to infrastructure, increased flood risk, greenhouse gas emissions, and salinization of freshwater resources, particularly when considering the impacts of accelerated climate change and sea level rise (SLR). In the Living on Soft Soil (NWA-LOSS) research programme, different pathways for addressing land subsidence in the Netherlands have been developed by first identifying the targeted future states that align with addressing land subsidence problem in the Netherlands. Then, we developed a series of water and land management strategies and scenarios to reach these identified future states. However, this process required the creation of a novel subsidence risk map in order to better linking each specific area in the Netherlands with a specific management strategy and intervention measures according to the risk level of this area. This MSc thesis focuses on employing statistical assessment tools and machine learning approaches to get the most accurate results of the subsidence-related damage risk assessment across both urban and rural areas in the Netherlands.

G22.1 Subtopics

- Subsidence-related damage risk assessment across the rural areas in the Netherlands using different machine learning approaches
- Subsidence-related damage risk assessment across the urban areas in the Netherlands using different machine learning approaches

G22.2 Supervision

- UU supervisors: Dr. Muhannad Hammad (m.hammad@uu.nl)
- External supervisors: Not indicated (Not indicated)

G22.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: None given

G22.4 Additional information

- Data description: The project will employ a range of data sources, including remote sensing, hydrological, geological, topographical, environmental, and socio-economic datasets. Specific data types will include InSAR subsidence data, groundwater level

information, soil and geological maps, CO2 emission data, land use maps, elevation data, population density maps, and the damage probability data of buildings with shallow foundations, among others. These comprehensive datasets will serve as the foundation for mapping subsidence-related damage risks across both rural and urban areas, as well as for the application of machine learning techniques to assess the relative importance and interactions of various relevant variables.

- NDA: No NDA indicated
- Website: https://nwa-loss.nl/

G23 Drought forecasting using machine learning and time-series data

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

Drought is one of the major meteorological disasters that has impacts on ecosystem functioning and agricultural production. Developing effective tools to forecast drought events could be helpful for mitigation strategies. Satellite based observations such as EVI or vegetation greenness is widely used as a proxy for vegetation health and drought stress. The objective of this project is to predict vegetation time series during and after drought conditions in different climatic regions using machine learning techniques. With the use of different climatic and geographical variables we intend to reliably forecast the EVI at different time periods in advance. Sub-topics to be decided in consultation

G23.1 Subtopics

- Vegetation forecasting with a Random Forest model
- Vegetation forecasting with a LSTM model

G23.2 Supervision

- UU supervisors: S.L. Verhoeve (s.l.verhoeve@uu.nl)
- External supervisors: Not indicated (Not indicated)

G23.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: None given

G23.4 Additional information

- Data description: The data for this project consists of remote sensing time-series data on vegetation of a geographical region which has yet to be chosen. Monthly vegetation maps for at least 20 years are available. Additionally other spatio-temporal data will be used such as meteorological data (e.g. monthly temperature and precipitation; also timeseries) and geographical characteristics (e.g. elevation and land use; spatial data). The data is generally stored in netCDF files.
- NDA: No NDA indicated
- Website: None

G24 Comparison of (regional) climate change projections & spatial correlation to Water-Energy-Food (WEF) security in South Africa.

Number of students: 2

Subject area: Geo science,Other:

External organisation: Not indicated

The impacts of climate change on water, energy, and food (WEF) security represent a critical challenge for South Africa, where vulnerabilities are exacerbated by socio-economic and environmental inequalities. To better understand the impacts of climate change, it is important to gain an insight in the vast range of possible future climate outcomes, especially considering the high spatial variation in current climate in South Africa (covering 13 climate zones). For this project, I am looking for a student that is able to perform a spatial comparison of various climate change projections for South Africa, such as from global models (e.g., CMIP6 projections) and high-resolution regional climate models (e.g., using CORDEX downscaling principles, as done in [1, 2] for the continent of Africa). After integrating these datasets, the projected climate parameters (e.g., temperature, precipitation, and extreme weather patterns) under different emission pathways can be linked to a spatial database of WEF insecurity factors across South Africa's municipalities. These findings will offer insights in whether there is spatial correlation between projected impacts of climate change and current WEF insecurity. Thus it will enable the identification of regions most at risk of climate change. [1] Dosio, A., Jones, R.G., Jack, C. et al. What can we know about future precipitation in Africa? Robustness, significance and added value of projections from a large ensemble of regional climate models. Clim Dyn 53, 5833–5858 (2019). https://doi.org/10.1007/s00382-019-04900-3 [2] Dosio, A., Jury, M.W., Almazroui, M. et al. Projected future daily characteristics of African precipitation based on global (CMIP5, CMIP6) and regional (CORDEX, CORDEX-CORE) climate models. Clim Dyn 57, 3135–3158 (2021). https://doi.org/10.1007/s00382-021-05859-w

G24.1 Subtopics

- How do the climate impacts in South Africa compare across global and regional climate model ensembles?
- Preferably only one student. Potentially: In which municipalities do detrimental climate impacts overlap with low WEF security status in South Africa?

G24.2 Supervision

- UU supervisors: Menno Straatsma, Inge Ossentjuk (m.w.straatsma@uu.nl; i.m.ossentjuk@uu.nl)
- External supervisors: Not indicated (Not indicated)

G24.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough

- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: None given

G24.4 Additional information

- Data description: For the climate models, there are many open-source datasets available, such as the global CMIP6 projections. But there is also already data of downscaled projections through CORDEX. For the WEF security, I have a dataset consisting of different variables of the water, energy and food systems for different security aspects (such as availability, affordability, etc.). The data has both the raw data as well as a processed dataset (processed in Python).
- NDA: No NDA indicated
- Website: None

G25 Data-driven geological learning: Transformer-based subsurface applications

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

This thesis investigates the use of transformer models to automate the interpretation of subsurface information (borehole data) into geological units, addressing the sequential aspect of the prediction. Borehole data presents a sequential challenge, as geological units (classes) have strict spatial and sequential relationships. For instance, geological units have superposition relationships based on stratigraphic principles and spatial patterns that result from spatial features such as geological faults. While many classifiers can be applied to borehole data, the problem's sequential nature is fundamental to accurate geological interpretations. For this reason, transformer models, a type of deep learning architecture that captures long-range dependencies in sequential data, may be well-suited for this task. This research aims to explore the extent to which transformers can learn, or be informed, from data the spatial and geological relationships of geological units of the subsurface, facilitating more accurate and automated interpretation approaches.

G25.1 Subtopics

- Data-driven geological learning: Transformer-based subsurface applications
- Improving Borehole Geological interpretations with Fault-Aware Transformers

G25.2 Supervision

- UU supervisors: Sebastian Garzón (j.s.garzonalvarado@uu.nl)
- External supervisors: Not indicated (Not indicated)

G25.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial statistics and machine learning ,Transformers: applications in language and communication
- Additional requirements: A Geology background would be useful, but it is not a limitation.

G25.4 Additional information

- Data description: The dataset consists of geological descriptions from approximately 26,000 boreholes in The Netherlands, maintained by the Geological Survey (GDN, as denoted in Dutch). This collection forms the input for the Dutch subsurface's Digital Geological Model (V.02.2). Detailed geological and lithological descriptions accompany each borehole. The dataset includes over 500,000 subsurface intervals, classified into 36

distinct geological units. For the geological faults, we will use the HIKE- European Fault Database, that contains major geological faults for The Netherlands.

- NDA: No NDA indicated
- Website: None

G26 Using graph theory to study tidal channel networks

Number of students: 2

Subject area: Geo science

External organisation: Not indicated

The tide forms a network of channels in coastal basins such as the Westerschelde and the Waddenzee. The characteristics of these channel networks affect unique wildlife habitats, shipping, and the coastal defence of the Netherlands. Studying them can reveal critical information about the state of the Dutch coast. In this project, you will work with a schematic representation of the channel network of the Westerschelde (and possibly also the Waddenzee) to examine the topology of the network and the relationship between the channels and their associated sand bars (spaces between the channels). You will apply both standard and novel analysis methods, and contribute to the development of a new method for sediment budgeting.

G26.1 Subtopics

- Channel network topology
- Channel-bar relationships

G26.2 Supervision

- UU supervisors: Abigail Hillen-Schiller (a.j.hillen-schiller@uu.nl)
- External supervisors: Not indicated (Not indicated)

G26.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: Familiarity with (some) Python geospatial packages is a must. A background in geography or Earth sciences is a plus.

G26.4 Additional information

- Data description: The data for this project are derived from bathymetry (depth) measurements collected by Rijkswaterstaat every 1-2 years in the period 1955-2023. The dataset comprises two types of GIS data: 1. Raster data - bathymetries (DEMs) for each year in which measurements were taken. 2. Vector data - lines representing the channels of the network for each of the bathymetries.
- NDA: No NDA indicated
- Website: None

H1 The Smoking Gun Part 2

Number of students: 3

Subject area: Media studies,Information and Computing science (including AI)

External organisation: Not indicated

NDP Nieuwsmedia, the Dutch assocation of news organizations (newspapers, news websites, news tv) has commissioned the Data School (Faculty of Humanities) to study the memorization of copyrighted data from their partners. We know that GPT4 is trained on mC4 (and other data), and that mC4 contains harvested archives of many news titles. So far we have clearly established memorization of copyrighted data, in data from the Netherlands, Flanders, and several other European countries, but many more research questions remain.

H1.1 Subtopics

- Comparing recitation of in-training news content versus out-of-training content
- Repeated prompt-based checking of content recitation: variations and limitations
- Cross-comparing LLM recitation capacities: model and training set size

H1.2 Supervision

- UU supervisors: Antal van den Bosch (a.p.j.vandenbosch@uu.nl)
- External supervisors: Not indicated (Not indicated)

H1.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Transformers: applications in language and communication
- Additional requirements: Proficiency in a first language other than Dutch or English means that the student could do the experiments on content in his/her own language,

H1.4 Additional information

- Data description: The core data is the mC4 dataset, available from Hugging Face at https://huggingface.co/datasets/allenai/c4 .
- NDA: No NDA indicated
- Website: https://www.ndpnieuwsmedia.nl/

H2 Eco-friendly LLMs: Memory-based Language Modeling

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

The current state of the art in LLMs relies on the Transformer architecture. One disadvantage of these systems is their large ecological footprint, rooted in their reliance on GPU/TPU training. A classic LM architecture, memory-based language modeling (MBLM), relies on CPU computing and RAM. A new version of MBLM is developed aimed at portability, scalability and offering a low ecological footprint. Part of the development is the evalation and benchmarking of MBLM, and finetuning MBLM to instruct/chat modi.

H2.1 Subtopics

- Instruct and chat finetuning of MBLM
- Evaluating and benchmarking MBLM
- Optimizing MBLM: Training set sizes, emulating attention and forgetting

H2.2 Supervision

- UU supervisors: Antal van den Bosch (a.p.j.vandenbosch@uu.nl)
- External supervisors: Not indicated (Not indicated)

H2.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Transformers: applications in language and communication
- Additional requirements: None given

H2.4 Additional information

- Data description: Training data for MBLM consists of publicly available textual databases such as MADLAD (https://huggingface.co/datasets/allenai/MADLAD-400) and instruct/chat finetuning databases such as listed here for Dutch: https://huggingface.co/collections/BramVanroy/geitje-7b-ultra-65c1ee010ad80fd1f6a8f208
- NDA: No NDA indicated
- Website: https://github.com/antalvdb/mblm

H3 Sasta self assessment

Number of students: 2

Subject area: Information and Computing science (including AI),Other:

External organisation: Not indicated

SASTA is a web application that enables the automatic analysis of spontaneous language transcripts following established methods available in Dutch, enabling a more efficient assessment of the language profile of young children (1-8 years, following TARSP [Schlichting 2017] and STAP [Van Ierland et al. 2008]), for example those with suspected Developmental Language Disorder, and of patients with aphasia (following ASTA [Boxum et al.2013]). It is already used by researchers in the field of language development and by clinical linguists working in various clinical settings (e.g., Pento, NSDSK, Vogellanden). SASTA currently achieves good results, with an overall F1-score for datasets ranging from 74.6 % to 96.2%. However, it is certainly not perfect yet. A manual check of the results by human experts is often still necessary. If the human expert has to check a full transcript, the gain in efficiency by using SASTA is limited. This gain can be increased if SASTA could point out specific utterances that it considers or suspects to be problematic, i.e. if SASTA could do some form of self-assessment. The goal of this project is to investigate whether such a SASTA self-assessment procedure is feasible. For example, if the parser used in SASTA (Alpino, [Van Noord, 2006] for some reason cannot combine all words of an utterance into one sentence, it analyses the parts that it is able to identify as a sequence of utterances. Certain measures in TARSP, an established and commonly used method for the analysis of spontaneous language of young children, describe how all words and phrases in an utterance are grammatically related. Our goal is that SASTA automatically and correctly assigns the codes which signal these grammatical relations. However, it is highly likely that the application is currently not able to do this for the utterances that it has wrongly not parsed as a single sentence. It can identify such utterances as requiring a manual check, but this is not a completely trivial process. That is, certain utterances actually consist of multiple discourse parts and have been analysed correctly by the parser, e.g., kijk uit hoor, hij bijt in je tenen

H3.1 Subtopics

- Predicting analysis quality on the basis of parsing quality
- identifying language measures that might affect the analysis result

H3.2 Supervision

- UU supervisors: Jan Odijk (j.odijk@uu.nl)
- External supervisors: Not indicated (Not indicated)

H3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough

- Course requirements: Text and Media Analytics
- Additional requirements: familiarity with parsing structures for Dutch sentences is an advantage

H3.4 Additional information

- Data description: data provided by VKL and by Auris, CHILDES data
- NDA: No NDA indicated
- Website: https://sasta.hum.uu.nl/

H4 Multi-faceted automatic text quality assessment

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

This project will run in collaboration with Bookarang (part of NBD Biblion), which specialises in content-based book recommendations and automated book analysis. Often, this process involves short condensed descriptions of a book. The overarching question of this project is, How can we assess whether such short text is good or bad, given its particular use (inferring book genre; type of plot; style)? When it comes to short texts describing a book (be it 'blurbs' that appear on book covers, written by human annotators; or be it AI-generated book summaries), how can we automatically detect whether they are fluent, written well, contain information that is useful to characterise the book along different dimensions? which existing tools and models can help us with these tasks, and how can we improve them?

H4.1 Subtopics

- Automatic assessment of the general quality of AI-generated book summaries (with focus on factual aspects of the text)
- Automatic assessment of the general quality of AI-generated book summaries (with focus on stylistic aspects of the text)
- What makes a good book blurb for automatic book characterisation?

H4.2 Supervision

- UU supervisors: Lisa Bylinina (on the Bookarang/NBD Biblion side, Niels Bogaards, niels@bookarang.com) (e.g.bylinina@uu.nl)
- External supervisors: Not indicated (Not indicated)

H4.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: Proficiency in Dutch might be needed

H4.4 Additional information

- Data description: There are 300k human-written book blurbs available (in Dutch), with smaller subsets that come with different types of annotation. Additionally, there are several thousand AI-generated book summaries (in Dutch as well) that have been corrected by human annotators, with the history of edits available. The exact size and information available as part of these datasets will be determined closer to the start of the project.

- NDA: No NDA indicated
- Website: https://www.bookarang.com/ (part of https://www.nbdbiblion.nl/)

H5 Mapping professional networks at the intersection of culture and technology

Number of students: 2

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

The project hosted by https://culttech.at/ aims to create an interactive system that reveals the complex web of relationships within the field of contemporary field of 'culltech' (tech-oriented cultural projects). Using LinkedIn connection data from the CultTech Association's database, students would develop tools to analyze and visualize how different professionals and organizations are connected across locations and sectors. The research would identify key influencers and bridge-builders in the network — people who connect different communities or sectors — while also revealing geographic patterns and potential hubs. The project would involve coming up with, testing and comparing different approaches to analysing human network data in an insightful way.

H5.1 Subtopics

- Identifying clusters / professional groups / influencers / hubs
- Suggesting missing information: missing values in the database, potential collaborations

H5.2 Supervision

- UU supervisors: Lisa Bylinina (on the Culltech side, Pavel Yushin, pi@culttech.at) (e.g.bylinina@uu.nl)
- External supervisors: Not indicated (Not indicated)

H5.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Human Network Analysis
- Additional requirements: None given

H5.4 Additional information

- Data description: The students will work with the CultTech Association's database, which contains information about 2.5k people active in the field, with (sparse) data along 14 dimensions, such as location, professional group membership etc.
- NDA: No NDA indicated
- Website: https://culttech.at/

H6 News Portrayals of AI as "Agentic" - A Computational Analysis of How Journalists write about AI as a "Human-Like" Entity

Number of students: 2

Subject area: Media studies

External organisation: Not indicated

Here is a revised version of your text with typos and grammar errors corrected: AI is a hot topic in the news media, which have increased their critical coverage over time. As data-driven algorithms that automate tasks increasingly affect diverse societal sectors, they have become an important issue on the public agenda. However, one point of criticism is how we talk about AI as if it were somehow (super)human—using agentic-mentalistic vocabulary when describing it—which may misrepresent it as something it actually isn't: human-like. Examples include sentences that state that AI "reasons," "thinks," "considers," "judges," etc. The news media are an important site where lay audiences are normalized to the ways in which the public discusses different issues and topics, including AI. The way journalists portray AI through their choice of metaphors can influence the "official" way of speaking that we adopt around AI. In this project, we aim to trace the stylistic choices in global English-speaking news media in AI reporting over time using NLP methods such as word embeddings and (possibly LLM-supported) text classification. The goal is to chart the evolution of the language used in news coverage to portray AI and provide an empirical basis for reflecting on the implications for how societies perceive and make sense of algorithms.

H6.1 Subtopics

- Analysing news framing with word embeddings
- Semi-supervised text classification using LLMs for news framing analysis

H6.2 Supervision

- UU supervisors: d.gnuyen@uu.nl (d.nguyen@uu.nl)
- External supervisors: Not indicated (Not indicated)

H6.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics
- Additional requirements: None given

H6.4 Additional information

- Data description: Over 20k news texts from different international brands covering AI (plus some social media data).
- NDA: No NDA indicated
- Website: None

H7 Representation of Migrants in Dutch News 2000-2024

Number of students: 2

Subject area: Media studies

External organisation: Not indicated

Media discourses play an important role in shaping the representation of different social groups in the public sphere. Analysing how the news report about migrants over time can reveal patterns in framing practices that may allow empirically substantiated critique of misrepresentation, xenophobia, and stereotyping. Furthermore, the news supposedly capture different viewpoints on critical issues such as migration and thus serve as a (biased)reflection of political trends. In this project, we will dive into the potential of NLP methods to unpack media framing practices. This includes topics analyses, sentiment analyses, and, importantly, stance detection and argument mining. One particular challenging task is to accurately retrieve direct and indirect citations and match them with the correct sources as a means to quantify what societal actors get the room to speak abut migration in the media.

H7.1 Subtopics

- Analysing Media Representation with NLP Methods
- Detecting verbatim and indirect citations in news text for stance detection

H7.2 Supervision

- UU supervisors: Dennis Nguyen (d.nguyen1@uu.nl)
- External supervisors: Not indicated (Not indicated)

H7.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics
- Additional requirements: None given

H7.4 Additional information

- Data description: Over 68.000 news articles from major Dutch news outlets. Command of Dutch is helpful (but not a super strict requirement).
- NDA: No NDA indicated
- Website: None

H8 What you hear is what you get: An exploration of audio level feature extraction for music recommendations

Number of students: 3

Subject area: Media studies

External organisation: Not indicated

It is common today to get music recommendations based our listening habits and the habits of others. Such recommentations typically consider metadata based on pre-defined music genres, billboard like rankings, and trends as well as what other members of a certain music platform (who share similar musical taste) have been listening over time. But while these types of music classifications and recommendations are now becoming the norm, what about music's audio level features themselves? Why are these features typically not presented as part of the classification and recommendation process? Could a music classification and/or recommendation system be build on these audio level features? If so, what meaningfull musical descriminants can be extracted from audio files to drive such system? These are all questions this project aims at addressing while positioning audio and sound as a productive domain of research in data science. Students who are interested in this project should have a basic knowledge of (digital) signal processing and music/sound theory, as well as being at ease in developing software in python. A knowledge of machine learning is a plus yet not mandatory. Data collection (i.e. gathering music files) will be part of the project and depend on individual project orientation coupled with students' interest in music and sound.

H8.1 Subtopics

- What you hear is what you get: Post-genre, feature-based music recommendation system
- Implementing Mood-Based Music Recommendations through Spectral Feature Analysis and Instrument Separation
- broadening audio taste: creating a music recommendation system for diversity with audio features.

H8.2 Supervision

- UU supervisors: Dr. David Gauthier (d.gauthier@uu.nl)
- External supervisors: Not indicated (Not indicated)

H8.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Personalisation for (public) media
- Additional requirements: basic knowledge of (digital) signal processing and music/sound theory

H8.4 Additional information

- Data description: Students are responsible to come up with a custom data collection.
- NDA: No NDA indicated
- Website: None

H9 Computational policy analysis of EU budgets

Number of students: 3

Subject area: Media studies

External organisation: Not indicated

In this project students use open data to answer questions about the political economy of media industries. Industrial policy has moved to the centre stage of political debates in the European Union during 2024 with the new European Commission and the Draghi report on economic competitiveness. In response, we research who gets money from current EU budgets and why, focusing on digital infrastructures such as AI, 5G, quantum and satellite. Such research is essential to inform the new seven year budget cycle that will start in 2028. Students are expected to apply statistical techniques for finding patterns in the data, pose critical questions, formulate evidence-based policy recommendations, and visualise results in a persuasive way for decision makers. The civil society organisation Open Futures will support the project by providing expertise on how EU funding works, how policy priorities are set, and which questions are interesting to investigate from a public values perspective.

H9.1 Subtopics

- AI, Quantum, or Satellite? Sectorial distribution of EU subsidies
- Following the shifts and trends from sustainability to militarisation
- Deep dive into the telecommunications sector

H9.2 Supervision

- UU supervisors: Zuzanna Warso, Director of Research (zuzanna@openfuture.eu)
- External supervisors: Not indicated (Not indicated)

H9.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Data Ethics: Responsible data practices and value-sensitive design
- Additional requirements: Background in European politics is not required but it is an asset

H9.4 Additional information

- Data description: Our starting point will be the Financial Transparency System of the European Commission, containing tabular data about more than a million EC-funded projects.:https://ec.europa.eu/budget/financial-transparency-system/about.html In order to dig deeper in particular research questions, we will make use of other EU open data sources attached to specific funding instruments. Of particular note is the Digital Europe Programme on digital infrastructures.
- NDA: No NDA indicated

- Website: https://openfuture.eu

H10 Stakeholder engagement patterns in developing cutting-edge mobile networks

Number of students: 3

Subject area: Media studies

External organisation: Not indicated

Media scholars and social commentators have long claimed that our societies are defined and governed by digital media and communication infrastructures, hailing the Network Society or the Information Age. In this project, we mobilise computational methods to study how such new media is produced through debates and discussions in standards bodies. The focus is on geopolitical power struggles around the new generation of mobile telecommunications protocols called 5G, one of the defining media infrastructure of this decade. Students will map stakeholder engagement patterns that translate to power struggles using analytical techniques such as named entity recognition and longitudial network analysis.

H10.1 Subtopics

- Does China and India, the largest mobile markets, dominate 5G innovation?
- Identifying controversies during the standardisation stage of emerging technologies
- How to support civil society participation in standards bodies?

H10.2 Supervision

- UU supervisors: Maxigas (p.dunajcsik@uu.nl)
- External supervisors: Not indicated (Not indicated)

H10.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Human Network Analysis
- Additional requirements: None given

H10.4 Additional information

- Data description: The corpus for the project is the mailing list archive of relevant working groups in the 3GPP standards organisations. We will specifically focus on the development of Release 18, also known as 5G-Advanced, which has been finalised in 2024. 3GPP uses an open standards development model, so that all discussions are publicly archived, which is a great opportunity for academic research into how technology is made.
- NDA: No NDA indicated
- Website: http://datactive.github.io/bigbang/

H11 Open source intelligence in unstructured data sets

Number of students: 3

Subject area: Media studies

External organisation: Not indicated

Open source intelligence is about collecting publicly available information that is nonetheless often hidden from the spotlights of public debates, and analysing it to produce actionable insights into current affairs. In this project, we focus on China's bid for global hegemony through dominance in high-tech such as AI, chips and 5G. The main challenge we tackle will be to navigate troves of technical documents, making them meaningful and relevant using text mining and topic modelling techniques.

H11.1 Subtopics

- Bottom-up approach using machine learning for topic modelling
- Top-down approach using knowledge graphs for topic modelling
- Locating hot spots and sociotechnical controversies using metadata

H11.2 Supervision

- UU supervisors: Maxigas (p.dunajcsik@uu.nl)
- External supervisors: Not indicated (Not indicated)

H11.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics
- Additional requirements: Background in engineering or tolerance for reading technical documentation may be an advantage in this course

H11.4 Additional information

- Data description: We will work with a corpus of roughly hundred millon tokens, namely the standards documents that define what 5G is, how it works, and what it does. More precisely, we will work on the standards of 5G-Advanced, finalised last year by an international collaboration of engineers under the umbrella of the 3GPP as Release-18. 5G mobile service providers and 5G equipment vendors are currently using these documents to implement and deploy the new generation of mobile telecommunications infrastructures.
- NDA: No NDA indicated
- Website: https://www.3gpp.org/

F1 FInding the best parameters for Rapid Invisible Frequency Tagging

Number of students: 3

Subject area: Social and behavioural science,Information and Computing science (including AI)

External organisation: Not indicated

Invisible Frequency tagging (RIFT) is a novel EEG technique that allows researchers to measure spatial attention at very fine temporal resolutions. Individual objects present on the screen can be tagged by periodically changing their luminance at high frequencies (>60 Hz). This luminance change is invisible to the participant but shifting attention to an object leads to an increase in the EEG power at the tagged frequency. The technique is highly promising for attention research as well as Brain Computer Interface development because it is completely invisible to the observer and provides state-of-the-art temporal resolution. One current challenge are measurements from several tagged objects with varying locations, as different stimulus positions project differently onto the EEG electrodes. In this project you will work on developing an algorithm to improve the signal-to-noise ratio in an EEG dataset. The main goal is to compute spatial filters that flexibly adjust to the source of the signal in the EEG and combines theses sources for further analysis. Furthermore you will explore several other techniques to help denoise the signal and incooperate oscillatory phase in source separation.

F1.1 Subtopics

- Rhythmic Source Separation for RIFT
- Phase-based spatial filters for RIFT
- Denoising techniques using RIFT coherence

F1.2 Supervision

- UU supervisors: Samson Chota (s.chota@uu.nl)
- External supervisors: Not indicated (Not indicated)

F1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: None given

F1.4 Additional information

- Data description: 64 Channel EEG dataset form a visual search experiment.
- NDA: No NDA indicated
- Website: https://www.cap-lab.net/

F2 Saving time and sanity with AI

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

"Given the picture as we see it now, it's conceivable that within the next ten years, AI systems will exceed expert skill level in most domains, and carry out as much productive activity as one of today's largest corporations." ~ OpenAI ~ In today's academic world, the number of scientific papers on any topic is growing. While this wealth of textual data is a treasure trove for data scientists, it simultaneously presents significant challenges for anyone who wants to screen this literature systematically. Whether for medical or food safety guidelines, evidence based therapy recommendations, systematic reviews for scientific papers, or evidence-based policy recommendations, one needs to sift through thousands of studies to identify the rare relevant ones. This tedious, time-consuming, and error-prone process, often due to screening fatigue, usually demands months of work. Despite this effort, systematic overviews are indispensable for scholars, clinicians, policymakers, journalists, and the general public. The rapidly evolving field of AI offers promising solutions to the literature screening challenge using machine learning models and, very recently, large language models (LLMs). However, many of these AI-driven solutions emerge from tech companies that publish new, and hopefully better, models at an unprecedented rate. The rapid advancements in the field of AI outpace meticulous scientific evaluations, leaving many methods unrefined and unproven. At Utrecht University, we have developed an open-source tool that implements many different learning models and provides extensive infrastructure to run large-scale simulation studies. For more information, see the GitHub organization (www.github.com/asreview), the website (www.asreview.ai), the AI-lab (www.uu.nl/en/research/ai-labs/disc-ai-lab), or the publication in Nature Machine Intelligence: www.nature.com/articles/s42256-020-00287-7. For your thesis project, you will join the development team under the supervision of AI specialists.

F2.1 Subtopics

- Beat the default models! Conduct simulation studies mimicking the labelling process to test the performance of newly developed models available on HuggingFace.
- Hyperparameter Tuning: Improve the training of model hyperparameters and develop tools to adapt these depending on the data type.
- AI-authority Bias Detection: Identify and quantify biases induced by using AI in the literature screening process.

F2.2 Supervision

- UU supervisors: Rens van de Schoot (a.g.j.vandeschoot@uu.nl)
- External supervisors: Not indicated (Not indicated)

F2.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: None given
- Additional requirements: Proven Github skills are essential

F2.4 Additional information

- Data description: All thesis projects can make use of the Synergy data; a collection of labelled datasets available on Dataverse: https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/HE6NAQ
- NDA: No NDA indicated
- Website: https://github.com/asreview/

F3 Clustering brain morphometry in Autism Spectrum Disorder (ASD)

Number of students: 3

Subject area: Health science,Social and behavioural science

External organisation: Not indicated

Structural brain alterations in ASD have been reported for a long time now, but there is considerable heterogeneity in the literature on all aspects of these structural alterations. We propose that this is due to the inherent subgroups existing in the ASD (and general) population which have different general organization of their brain structure, which is overshadowing the effects of ASD itself and leading to heterogeneous findings in different studies. Using a large cohort of many existing datasets, we aim to use different stratification algorithms to stratify these structural brain data into more homogenous subgroups. Within each subgroup, we will subsequently explore the specific effects of ASD.

F3.1 Subtopics

- e.g. Stratification of brain structure in ASD using spectral clustering and DB-SCAN.
- e.g. Comparing clustering outcomes of brain structure in ASD between males and females.
- e.g. Comparing stratification of brain structure in ASD across childhood and adolescence..

F3.2 Supervision

- UU supervisors: Daan van Rooij (d.vanrooij@uu.nl)
- External supervisors: Not indicated (Not indicated)

F3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data,Spatial statistics and machine learning
- Additional requirements: The students will need to implement existing clustering algorithms in python independently, and run basic statistics on the outcome of any clustering results.

F3.4 Additional information

- Data description: - segmented structural MRI scans for around 1500 cases with ASD and a similar number of healthy controls. Each structural scan is subdivided into 52 cortical and 6 subcortical areas. Each cortical area has a thickness and a surface area measure. Each subcortical area has a volume measure. - demographic data, IQ, ASD symptom severity.
- NDA: No NDA indicated
- Website: None

F4 Machine-Learning-Based Dimensionality Assessment for Cognitive Diagnosis Models

Number of students: 3

Subject area: Health science,Social and behavioural science

External organisation: Not indicated

Cognitive diagnosis models (CDM) are statistical models to classify participants in diagnostic settings – especially in psychology and education. CDM are latent class models that can be used to categorize individuals based on categorical attributes describing relevant skills or describing specific impairments (e.g., de la Torre & Minchen, 2014). When developing these models, assessing the dimensionality of the construct that should be assessed (e.g., mathematical abilities), i.e., determining the number of attributes underlying the construct can be challenging. Nájera et al. (2021), for example, show that depending on different data conditions, the performance of available methods varies drastically. Therefore, in a current research project in cooperation with researchers from the Comillas Pontifical University and Autonomous University of Madrid, an alternative approach for dimensionality assessment in CDM is developed. It combines extensive data simulation with machine learning (ML) modeling/ supervised learning. This general idea has already been successfully applied to the dimensionality assessment in exploratory factor analysis (Goretzko & Bühner, 2020). In your thesis project, you will be focusing on the last step of the development process. You will benchmark different supervised learning algorithms, explore the possibilities of hyperparameter tuning and investigating the most promising ML models with tools from interpretable machine learning. De La Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicología Educativa, 20(2), 89-97. Goretzko, D., & Bühner, M. (2020). One Model to Rule Them All? Using Machine Learning Algorithms to Determine the Number of Factors in Exploratory Factor Analysis. Psychological Methods, 25(6), 776–786. https://doi.org/10.1037/met0000262 Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the number of attributes in cognitive diagnosis modeling. Frontiers in Psychology, 12, 614470.

F4.1 Subtopics

- Machine-Learning-Based Dimensionality Assessment for Cognitive Diagnosis Models: A comparison of ML Algorithms
- Machine-Learning-Based Dimensionality Assessment for Cognitive Diagnosis Models: Using IML to Make Sense of the Prediction Model
- Machine-Learning-Based Dimensionality Assessment for Cognitive Diagnosis Models: Applicability Analyses

F4.2 Supervision

- UU supervisors: David Goretzko (d.goretzko@uu.nl)
- External supervisors: Not indicated (Not indicated)

F4.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Excellent understanding of supervised machine learning:

  - model training
  - hyperparameter tuning
  - model evaluation (resampling)
  - ...

F4.4 Additional information

- Data description: extensive data simulation and feature engineering by cooperation partners: preprocessed dataset for training and testing of ML models available
- NDA: No NDA indicated
- Website: None

F5 Understanding vaccination preparedness and other COVID-related attitudes and behaviors using contextual data.

Number of students: 3

Subject area: Social and behavioural science

External organisation: Not indicated

How do intentions to vaccinate and other attitudes about COVID-related measures and behavior depend on where someone lives? Religion, political preferences, and someone's health situation may affect their intentions and attitudes, but also many other personal and contextual features can be important. Results show that vaccination intentions are difficult to understand (Chambon et al. 2022). Not only the beliefs about the efficacy of the vaccine, but also the perception of the social norm about vaccination is an important factor to explain the vaccination intentions (Kroese et al. 2024). The last aspect suggests that aspects related to the social context might be important and a more data-driven approach with more attention for contextual data might be relevant to understand vaccination intentions. Chambon, M., Kammeraad, W. G., van Harreveld, F., Dalege, J., Elberse, J. E., & van der Maas, H. L. (2022). Understanding change in COVID-19 vaccination intention with network analysis of longitudinal data from Dutch adults. npj Vaccines, 7(1), 114. Kroese, F., van den Boom, W., Buskens, V., van Empelen, P., Hulscher, M., Ruiter, R. A., ... & Lambooij, M. (2024). When and why do people change their minds in favor of vaccination? Longitudinal analyses of switching COVID-19 vaccination preferences. BMC Public Health, 24(1), 1-12.

F5.1 Subtopics

- COVID-19 attitudes and political values
- Glassbox models to understand vaccination preparedness
- Predicting vaccination preparedness including context related measures

F5.2 Supervision

- UU supervisors: Vincent Buskens (v.buskens@uu.nl)
- External supervisors: Not indicated (Not indicated)

F5.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences
- Additional requirements: None given

F5.4 Additional information

- Data description: LISS is a representative Dutch panel that includes several measures on COVID-related attitudes and norms. It also has information on religion, political values and many other individual characteristics as well as context measures. Potential for connecting the data to other datasets on contextual properties can be investigated.

- NDA: No NDA indicated
- Website: None

F6 Who Influences Your Future? Using Explainable AI to Understand the Effect of Networks on Fertility Intentions

Number of students: 3

Subject area: Social and behavioural science,Information and Computing science (including AI)

External organisation: Not indicated

Understanding why people want or decide to have children is crucial for studying societal trends, but it's not just about individuals—social networks play a big role. For example, a person's friends, family, and other connections can influence their decisions. While traditional research has studied these social influences, it's often based on small or selective networks. A recent study by Stulp et al. (2023) used a unique dataset with complete networks to show that individual factors like age were better predictors of fertility intentions than network factors like how connected the network is. However, this study only looked at a limited set of pre-defined network characteristics (e.g., number of friends wanting children or network density). What if networks hold hidden patterns that weren't captured by these predefined metrics? Graph Neural Networks (GNNs)---cutting-edge models in machine learning---can uncover complex patterns in networks without needing to define these characteristics upfront. But there's a catch: GNNs are often "black boxes," meaning it's hard to understand how they make predictions. In this project, you will use explainability tools to unlock the "black box" of GNNs and uncover new insights into how social networks shape fertility intentions. This project offers an exciting opportunity to gain hands-on experience with Graph Neural Networks—widely used by industry leaders like Google and Netflix—and cutting-edge explainable AI tools, while working on a real-world problem with a unique dataset.

F6.1 Subtopics

- What are the most important network features according to explainability tools such as GNNExplainer or Zorro?
- Do networks play a different role for different types of women?
- Pre-training GNNs to learn better information from networks

F6.2 Supervision

- UU supervisors: Javier Garcia-Bernardo (j.garciabernardo@uu.nl)
- External supervisors: Not indicated (Not indicated)

F6.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Experience training neural networks.

F6.4 Additional information

- Data description: The data comes from the LISS panel, a large, representative survey from the Netherlands. It includes: * Individual characteristics: Information about age, education, income, partnership status, number of children, and fertility intentions (whether respondents want more children in the future). * Network information: Each respondent listed up to 25 people they had contact with in the past year, described the type of relationship (e.g., partner, parent, or friend), and indicated whether these contacts knew one another. This produced a rich dataset of over 18,000 relationships from 738 women.
- NDA: No NDA indicated
- Website: None

F7 Embedding Quality in Language Models for Social Sciences

Number of students: 3

Subject area: Social and behavioural science

External organisation: Not indicated

Language models (LMs) generate text embeddings that map words, sentences and documents into meaningful numerical representations. These embeddings, however, can be unstable—small changes in data often lead to significant variations. This instability raises questions about how well these embeddings capture the information they are intended to encode, potentially leading to inconsistent results in real-world applications, such as sexism detection. Explainability can help validate how embeddings are generated by LMs, yet its potential as a solution to address embedding instability has not been explored. This research is aimed at the development of more-explainable models to reduce instability in LM- embeddings, thereby improving the quality of text generation.

F7.1 Subtopics

- How to evaluate the quality of embedding?
- LMs vs LLMs: Which result better quality?
- XAI for better embedding quality

F7.2 Supervision

- UU supervisors: Ayoub Bagheri (a.bagheri@uu.nl)
- External supervisors: Not indicated (Not indicated)

F7.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Transformers: applications in language and communication
- Additional requirements: None given

F7.4 Additional information

- Data description: We use data from the EXIST 2024 challenge (https://nlp.uned.es/exist2024/), which comprises datasets sourced from X (Twitter). The labeled dataset contains tweets in both English and Spanish, with the training set comprising 6920 tweets in both languages (3260 in English, and 3660 in Spanish). For simplicity, we focus exclusively on Task 1 which involves binary classification of tweets to determine whether they express content related to sexism or not.
- NDA: No NDA indicated
- Website: https://nlp.sites.uu.nl

F8 Mapping Economic Crime: Analyzing Money Laundering Patterns using Network Science

Number of students: 3

Subject area: Social and behavioural science

External organisation: Not indicated

Economic crime has become an integrated part of global financial flows, constituting 2% to 5% of global GDP—approximately $800 billion to $2 trillion. Every day, criminals disguise the illicit origin of vast amounts of money to gain access to legitimate financial systems. Information about such transactions is difficult to detect, as they are often facilitated by a network of individuals and corporations integrating illicit funds into the lawful economy. In 2013, the International Consortium of Investigative Journalists (ICIJ) published the Offshore Leaks Database [1], followed in subsequent years by other high-profile leaks such as the Pandora Papers, Paradise Papers, Bahamas Leaks, and Panama Papers. Together, these datasets contain information on more than 810,000 offshore entities implicated in money laundering investigations. Complementing this, a comprehensive dataset [2] of all company owners, directors, and addresses in the Netherlands provides additional insights into corporate structures and relationships within legitimate and potentially illegitimate networks. These datasets offer an unprecedented opportunity to analyze patterns in economic crime by combining global and country-specific perspectives. Network science is a powerful tool for analyzing this phenomenon, as it allows researchers to study the relationships between individuals and companies involved in illicit financial activities. By modeling these relationships as a network, we can uncover hidden structures and better understand how economic crime networks operate. Furthermore, network science can reveal broader insights by comparing these economic crime networks to other network types, such as social or ecological systems. This approach enables the identification of shared universal patterns—such as hierarchy or clustering—or the discovery of unique interaction structures specific to economic crime. This thesis project aims to understand whether the network structure and dynamics of economic crime reveal universal patterns common to other networks or present distinctive features unique to financial crime. To achieve this, Network Science techniques—such as motif analysis, centrality measures, and community detection—will be applied to both the ICIJ Offshore Leaks Database and the Dutch corporate ownership and director dataset. By combining these data sources, the research will investigate both global and localized manifestations of economic crime. Additionally, Graph Machine Learning tools, such as link prediction and node classification, will be employed to uncover hidden connections and predict the behavior of network components. Two possible RQs of this thesis project are: 1. What global/local structural patterns or distinctive features can be identified in the networks using network techniques, and how do these compare to other network types? 2. How can Graph Machine Learning techniques, such as link prediction and node classification, uncover hidden connections and predict relationships within global and localized networks? [1] ICIJ: How to Download This Database | ICIJ Offshore Leaks Database. https://offshoreleaks.icij.org/pages/database [2]

Garcia-Bernardo, J., Witteman, J., & Vlaanderen, M. (2022). Uncovering the size of the illegal corporate service provider industry in the Netherlands: a network approach. EPJ Data Science, 11(1), 23.

F8.1 Subtopics

- Finding recurrent structures in company ownership networks: motif analysis approach
- Universal properties in real-world networks: are company ownership networks similar to other social networks?
- Finding hidden connections: Graph Machine Learning on company ownership networks

F8.2 Supervision

- UU supervisors: Elena Candellone (e.candellone@uu.nl)
- External supervisors: Not indicated (Not indicated)

F8.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Human Network Analysis
- Additional requirements: The thesis is centered on knowledge of basic and slightly advanced concepts in network science. Python and/or R proficiency is necessary too.

  Packages in Python we might use:
  - pandas
  - numpy
  - NetworkX
  - igraph
  - graphtool

  Suggested readings:
  - http://networksciencebook.com/
  - https://github.com/CambridgeUniversityPress/FirstCourseNetworkScience
  - https://cambridgeuniversitypress.github.io/WorkingWithNetworkData/

F8.4 Additional information

- Data description: The available data for this thesis includes the ICIJ Offshore Leaks Database, which contains information on more than 810,000 offshore entities implicated in money laundering investigations, and an anonymized dataset of all company owners, directors, and addresses in the Netherlands, which provides detailed insights into corporate structures and networks. The datasets are in the form of a relationships list (in network science terms, an edge list).
- NDA: No NDA indicated
- Website: None

F9 Chain of Thought Approaches for LLMs

Number of students: 3

Subject area: Social and behavioural science,Information and Computing science (including AI)

External organisation: Not indicated

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks, yet unlocking their full potential requires innovative strategies to enhance their reasoning abilities. One promising approach is the use of Chain of Thought (CoT) prompting, which guides LLMs through structured sequences of intermediate reasoning steps before reaching a conclusion. By making each step in the reasoning chain explicit, CoT prompting aims to improve the comprehensibility and accuracy of LLM outputs. Given the growing importance of LLMs in applications ranging from automated customer support to complex decision-making systems, understanding and optimizing CoT prompting has become crucial. In this master project, students will apply and compare different CoT approaches on a range of tasks such as programming, question answering and sentiment analysis. Through systematic evaluation, students will uncover how different domains benefit from CoT reasoning, offering insights into its broader applicability. Students will explore task-specific prompt engineering and will contribute to knowledge on optimizing LLM prompts to improve accuracy and reliability on a task-by-task basis. [1] Yin, M. J., Jiang, D., Chen, Y., Wang, B., & Ling, C. (2025). Enhancing Generalization in Chain of Thought Reasoning for Smaller Models. arXiv preprint arXiv:2501.09804. [2] Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B. and Liu, T. (2023). A survey of chain of thought reasoning: Advances, frontiers and future. arXiv preprint arXiv:2309.15402. [3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, pp.24824-24837.

F9.1 Subtopics

- Evaluating the chain-of-thought approaches on text data.
- Evaluating the chain-of-thought approaches on math/programming tasks
- Improving chain-of-thought with explainable AI techniques

F9.2 Supervision

- UU supervisors: Anastasia Giachanou (a.giachanou@uu.nl)
- External supervisors: Not indicated (Not indicated)

F9.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication

- Additional requirements: None given

F9.4 Additional information

- Data description: The data are annotated datasets for sentiment analysis, question answering and programming tasks
- NDA: No NDA indicated
- Website: Department of Methodology and Statistics

F10 On targeting: Predictors of Consumer Welfare from Catastrophic Insurance

Number of students: 2

Subject area: Law, Economics and Governance,Other:

External organisation: Not indicated

Financial decisions are complex, and individuals often make suboptimal choices. According to behavioral economic theory, optimal financial decisions occur when people maximize their individual welfare. Conversely, suboptimal decisions imply people forgo some individual welfare and could be better off if they made a different decision. This study focuses on catastrophic insurance, a financial product that protects rural households from severe climate shocks like droughts (Barrett et al., 2024; Chantarat et al., 2023). However, this insurance uses satellite imagery to determine indemnity payments rather than actual data on asset losses, which results in heterogeneous welfare outcomes—for some households, this insurance enhances welfare while it reduces welfare for others (Harrison et al., 2020). Experimental data from households are often collected to assess this product's welfare impacts, but large-scale experimental data collection is costly. Instead, we aim to study whether standard consumer characteristics data typically collected by insurers, like age, education, and asset ownership, can predict individual welfare outcomes. Thus, the research question is: Which consumer characteristics predict whether financial decisions are welfare-increasing or welfare-reducing? In this project, your task is to identify the characteristics of households that predict welfare outcomes from insurance decisions. You will start by using feature selection techniques, such as Lasso and Elastic Net, to identify key predictors of welfare measures(The welfare measures are calculated and provided in the dataset) (Einav et al., 2018). Beyond that, you explore more advanced models like Change Point Detection, Random Forest, and Bayesian Regressions to refine the analysis. A critical part of the project will be evaluating your models' predictive accuracy and generalizability and suggesting ways to improve them. The findings of this research can guide the targeting of insurance products to households most likely to benefit.

F10.1 Subtopics

- From Data to Welfare Impacts: Consumer Characteristics in Catastrophic Insurance
- Advancing Predictive Modeling: Enhancing Welfare Outcome Estimation from Catastrophic Insurance Decisions

F10.2 Supervision

- UU supervisors: Mahdi Shafiee Kamalabad (m.shafieekamalabad@uu.nl)
- External supervisors: Not indicated (Not indicated)

F10.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough

- Course requirements: None given
- Additional requirements: None given

F10.4 Additional information

- Data description: In this study, you will work with a dataset comprising survey and lab-in-field experiment data from 2,416 low-income pastoral households who made purchase and non-purchase decisions for this insurance in the Borena Zone, Southern Ethiopia. The dataset also includes welfare estimates derived from insurance decisions (purchase and non-purchase), calculated using a structural welfare evaluation model. You can find the data in the following link: https://drive.google.com/file/d/1NEnJIrrXf_YktALJcPpi5mLOefMZeDAh/view
- NDA: No NDA indicated
- Website: None

F11 Psychometric network analysis: new approach to modeling psychopathology. Simulation study.

Number of students: 2

Subject area: Social and behavioural science

External organisation: Not indicated

Psychometric network analysis is an emerging framework in psychopathology that conceptualizes mental disorders as networks of symptoms (nodes) and their partial correlations (edges). Unlike the traditional latent variable approach, which assumes symptoms correlate due to an underlying latent cause, network approach posits that symptoms are correlated because they cause one another. This methodology has been applied to model networks of disorders such as depression, anxiety, obsessive-compulsive disorder, and post-traumatic stress disorder. However, several methodological challenges remain: What sample sizes are sufficient for estimating specific networks? Which models, regularization techniques, and hyperparameters best recover the true network structure? How do different network parameters influence these considerations? In this project, you will use bootnet, igraph and parSim packages to: 1. simulate synthetic networks with varying parameters, 2. simulate data based on these networks, 3. use different models (GGM with GLASSO regularization, Bayesian GLASSO, etc.) to estimate networks, 4. compare performances of the models. The goal of this project is to guide future researchers in successfully applying psychometric network methodology.

F11.1 Subtopics

- How do network parameters influence selection of models, regularization strategies, and hyperparameter tuning for datasets with continuous variables?
- How do network parameters influence selection of models, regularization strategies, and hyperparameter tuning for datasets with mixed variables?

F11.2 Supervision

- UU supervisors: Dr. Mahdi Shafiee Kamalabad (v.dvoriak@uu.nl)
- External supervisors: Not indicated (Not indicated)

F11.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences,Human Network Analysis
- Additional requirements: None given

F11.4 Additional information

- Data description: Data will be simulated based on network models with varying parameters.

- NDA: No NDA indicated
- Website: None

F12 Word boundaries, word order, and word structure: statistics meets linguistics

Number of students: 3

Subject area: Information and Computing science (including AI),Other:

External organisation: Not indicated

The concept of word is crucial in linguistics while its theoretical status is contested. We explore the fundamental unit of word through the trade-off between word structure and word order. We build upon Mosteiro & Blasi (2025): Word boundaries and the morphology-syntax trade-off (Coling-Rel 25), where word-pasting and word-merging methods were used to reproduce an order-structure trade-off reported in Koplenig et al (2017): The statistical trade-off between word order and word structure (PLOS ONE 12). We will investigate whether conventional word boundaries (i.e., spaces in English) are informationally optimal (do they minimize bits per character). We will compare the trade-offs in Mosteiro & Blasi (2025) and Koplenig et al (2017) to check that their functional forms are statistically equivalent. Finally, we will evaluate the same methodology on a new and previously annotated dataset.

F12.1 Subtopics

- Are conventional word boundaries informationally optimal?
- Do previously results hold when using a new dataset?
- is the observed trade-off a mathematical or linguistic property?

F12.2 Supervision

- UU supervisors: Pablo Mosteiro (p.j.mosteiroromero@uu.nl)
- External supervisors: Not indicated (Not indicated)

F12.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Python and scikit-learn. Sub-project 2 involves training a simple machine learning model such as logistic regression.

F12.4 Additional information

- Data description: The Parallel Bible Corpus contains 2000 translations of the Bible in 1460 languages in a verse-aligned parallel structure, covering over 40 language families from 5+ continents. Each translation is tokenized and Unicode-normalized. For sub-project 3, we use either Europarl or Teddi, both available online. Europarl is a set of proceedings from the EU Parliament and Teddi is a sample of non-translated texts, both in multiple languages.
- NDA: No NDA indicated

- Website: None

F13 Hidden uncertainty in data analysis: Understanding sources of variability in many-analyst projects

Number of students: 3

Subject area: Social and behavioural science,Other:

External organisation: Not indicated

What do we know already? Data analysis is supposed result in informative and reliable conclusions regarding a research question. Yet recent many-analysts and multiverse projects have convincingly exposed analytic variability: based on the same research question and the same data, researchers often apply different statistical procedures, which can result in substantially different conclusions. For instance, during the pandemic different analysis teams estimated the reproduction number R based on the same data, and reached vastly different conclusions ranging from the possibility that the pandemic was receding, to an R-value of 1.8. While we have long been aware of population heterogeneity is social science (e.g., cultural differences), this type of analytic heterogeneity is often ignored in the typical research pipeline. Yet the high variability in outcomes of many-analysts projects signals that the conclusions are relatively fragile and dependent on the specifics of the analysis plan. What do we not know yet? While many-analyst projects are gaining popularity and analytic variability has been convincingly demonstrated, little is known, however, about which exact analytic decisions are most consequential. For instance, in answering the question whether religious people are happier, does it matter which covariates are added, if one uses a multilevel regression model or a t-test, or operationalizes happiness as positive mental health or being cheerful? Moreover, do any characteristics of the analysts themselves make a difference; are analysts who a priori believe in the effect more likely to find it? Does seniority play any role? The effects of some of these features have been suggested in previous many-analysts projects, yet a systematic investigation is lacking. What will you be doing? In the current project, we will use data from 3 existing many-analysts projects to systematically evaluate what types of analytic decisions drive observed heterogeneity across analysis teams. We will catalogue what types of analytical choices analysts make in practice and identify potential clusters of these choices (e.g., modeling choices, variable selection, data cleaning) using machine learning techniques. You can choose to focus on comparing 2 or 3 different datasets or investigating different machine learning algorithms.

F13.1 Subtopics

- Using machine learning to identify clusters of sources of variability in many-analysts projects
- Identifying clusters of sources of variability across different many-analysts projects
- Using machine learning to identify clusters of sources of variability in many-analysts projects

F13.2 Supervision

- UU supervisors: dr. Suzanne Hoogeveen (s.hoogeveen@uu.nl)
- External supervisors: Not indicated (Not indicated)

F13.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: None given

F13.4 Additional information

- Data description: Data from 2 or 3 existing many-analysts projects will be used: - The many-analysts religion well-being project: ~100 analysis teams (Hoogeveen et al., 2023, https://www.tandfonline.com/doi/full/10.1080/2153599X.2022.2070255) - The skin-color red card soccer project: 29 analysis teams (Silberzahn et al, 2018, https://journals.sagepub.com/doi/full/10.1177/2515245917747646) - The immigration and support for social policies project: 73 analysis teams (Breznau et al., 2022, https://www.pnas.org/doi/abs/10.1073/pnas.2203150119)
- NDA: No NDA indicated
- Website: None

F14 What works? Detecting and evaluating education policies at scale

Number of students: 3

Subject area: Social and behavioural science,Information and Computing science (including AI)

External organisation: Not indicated

Good education is extremely important for combating inequality, for good governance, and for our economy. But how do we decide which education policies work? Education systems are often reformed, with very little evaluation of the results. So it is likely that there are many more small policy changes than even education experts know about – let alone can give sensible advice on. Can AI be used to detect these reforms and support their evaluation? In this project, we will take a first step towards answering this question. To do so, we will take as our gold standard data set an existing overview of historical educational policy changes constructed by hand, some of which will be held out for validation. Your task will be to leverage LLMs to see whether these policy changes could have been detected from official government records. Time permitting, we will also perform a few simple policy evaluation analyses using survey microdata and common analytic strategies such as difference-in-difference or instrumental variables analysis. You will: • gain experience working with both open- and closed-source LLMs and prompt engineering, with a focus on scientifically rigorous evaluation of the results; • learn more about quantitative analysis of government policy, especially education; • work on the cutting edge of policy research with potential applications far beyond the specific dataset used here. Any enthusiastic candidates are welcome to apply. No specific coursework within the ADS master is required, although some experience with LLMs and statistical analysis is always a plus. See also: Braga et al. (2013). https://doi.org/10.1111/1468-0327.12002

F14.1 Subtopics

- Detecting Dutch education policy changes in the Staatscourant
- Evaluating Dutch education policy changes
- Detecting education policy changes in UNESCO and other online records

F14.2 Supervision

- UU supervisors: Daniel Oberski (d.l.oberski@uu.nl)
- External supervisors: Not indicated (Not indicated)

F14.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Casual Inference Methods for Policy Evaluation,Text and Media Analytics
- Additional requirements: No specific requirements, though at least one student who reads Dutch would be best

F14.4 Additional information

- Data description: - List of educational policy changes in European countries compile by Braga (2013) - Staatscourant, OECD records
- NDA: No NDA indicated
- Website: None

F15 Coupling of brain structure and function associated with cognition and self-regulation in young adults.

Number of students: 2

Subject area: Health science,Social and behavioural science,Information and Computing science (including AI)

External organisation: Not indicated

Our Brain Plasticity research group focuses on the neurobiological aspects of brain development and aging. We are particularly interested in how changes in the structure and function of the brain are related to cognition and mental health, as well as social skills such as self-regulation and antisocial behavior. Part of this effort is to understand how the structure and function of the brain co-develop with age, and how variation in this relationship impacts behavior. Within this research project you will work together to employ a data analysis method of your own choosing to integrate the information from the structural and functional MRI scans of the brain into a model that best explains the target behaviors: one of you will focus on cognition, the other will focus on self-regulation.

F15.1 Subtopics

- Coupling of brain structure and function associated with cognition.
- Coupling of brain structure and function associated with self-regulation.

F15.2 Supervision

- UU supervisors: Prof. dr. Hilleke Hulshoff Pol (h.e.hulshoffpol@uu.nl)
- External supervisors: Not indicated (Not indicated)

F15.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences,Epidemiology and big data,Human Network Analysis,Spatial statistics and machine learning
- Additional requirements: We offer this project at various degrees of difficulty and computational complexity, from association analyses requiring only a basic knowledge of statistical analysis with linear models in R, to running deep-learning models in Python requiring skills in handling large quantities of data and affinity with TensorFlow or related tools.

F15.4 Additional information

- Data description: This project will use data from an existing cohort of N=1,200 young adults between the ages of 22 and 35 years with magnetic resonance imaging (MRI) scans of their brain and behavioral measures (https://www.humanconnectome.org/study/hcp-young-adult). Preprocessed MRI data

and scores of several behavioral measures are readily available for analysis. In particular, the behavioral measures we are most interested in are cognitive performances on the NIH ToolBox and related cognitive tasks, as well as measures related to self-regulation and impulsivity, such as performance on the Delay Discounting task and the Flankers Inhibitory Control task.

- NDA: No NDA indicated
- Website: https://brainplasticity.group/

UM1 Dive into data

Number of students: 3

Subject area: Health science,Information and Computing science (including AI)

External organisation: Not indicated

At Princess Maxima Center, our mission is to cure every child with cancer and ensure they have the best quality of life possible. As part of translational research team, you will contribute to bringing innovative research one step closer to the clinical practice. To make this possible, we will gather and assess new relevant medical insights from large datasets. We will dive into data and: 1) uncover hidden patterns in categorical (sub-topic 1) OR numerical data (sub-topic 2) OR 2) establish and train a predictive model (sub-topic 3).

UM1.1 Subtopics

- Dive into data: uncovering hidden patterns in large clinical dataset
- Dive into data: uncovering hidden patterns in large research and diagnostic dataset
- Dive into data: predicting patients' outcome

UM1.2 Supervision

- UU supervisors: Hannah Kunstek/PhD. Stefan Nierkens as first examiner (h.kunstek@prinsesmaximacentrum.nl)
- External supervisors: Not indicated (Not indicated)

UM1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Good to have:
  You show initiative and think outside the box.
  You are dedicated to finding solutions and willing to learn.
  You are open to work in a diverse international team.
  You are comfortable working with large datasets.

UM1.4 Additional information

- Data description: In this project we will work with large pre-processed dataset. The dataset contains categorical data from the clinic, and numerical data from research and diagnostics. No NDA is needed, however, discretion is advised. Data is universal. No affinity for science or medicine is needed. At the beginning of the project, you will get a brief introduction to the dataset, just to understand different variables that you will be working with. No scientific or medical knowledge is needed. We will work together to best respond to the questions of interest.
- NDA: No NDA indicated

- Website: https://research.prinsesmaximacentrum.nl/en/research-groups/nierkens-group

UM2 t2d and ad

Number of students: 2

Subject area: Health science

External organisation: Not indicated

This research investigates how visit-to-visit blood pressure variability (BPV) impacts the risk of dementia and cerebral small vessel disease (CSVD) across glycemic spectrums (normoglycemia, prediabetes, type 2 diabetes). CSVD markers, such as white matter hyperintensities and brain volume, may mediate the link between BPV and dementia, with genetic and cardiovascular risk factors potentially modifying these relationships. Understanding these pathways could inform strategies to prevent dementia in high-risk populations.

UM2.1 Subtopics

- The Relationship Between Visit-to-Visit Blood Pressure Variability and the Risk of Dementia Across Glycemic Spectrums. This sub-project focuses on analyzing how BPV impacts dementia risk, including its subtypes (Alzheimer's Disease, Vascular Dementia), while accounting for glycemic status and key confounders.
- The Mediating Role of Cerebral Small Vessel Disease in the Association Between Blood Pressure Variability and Dementia. This sub-project investigates whether CSVD markers (e.g., white matter hyperintensities, brain volume) mediate the relationship between BPV and dementia, exploring potential modifiers like genetic and cardiovascular risk factors.

UM2.2 Supervision

- UU supervisors: Dr. Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM2.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM2.4 Additional information

- Data description: The UK Biobank (UKBB) and SMART cohorts offer repeated blood pressure measurements for calculating BPV, brain MRI data for assessing CSVD markers (e.g., WMH, brain volume), and detailed dementia outcomes. Both include glycemic status (HbA1c, diabetes diagnoses), genetic data (e.g., ApoE), cardiovascular risk factors, and lifestyle information, enabling comprehensive analysis of the relationships between BPV, CSVD, and dementia.
- NDA: No NDA indicated

- Website: None

UM3 Predictive Value of Vascular Markers for Dementia and Stroke Across Glycemic States

Number of students: 2

Subject area: Health science

External organisation: Not indicated

The research proposal focuses on investigating the predictive value of vascular markers, specifically Intima-Media Thickness (IMT) and Pulse Wave Velocity (PWV), for dementia and stroke across different glycemic states (normoglycemia, prediabetes, and type 2 diabetes). Additionally, the study aims to explore sex-related differences in how these vascular markers influence the risk of cognitive decline and cerebrovascular events.

UM3.1 Subtopics

- The Role of Intima-Media Thickness (IMT) in Predicting Dementia and Stroke Across Glycemic States: A Cohort Study
- Pulse Wave Velocity (PWV) and Its Association with Cognitive Decline and Cerebrovascular Events in Varying Glycemic States

UM3.2 Supervision

- UU supervisors: Dr Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM3.4 Additional information

- Data description: The UK Biobank (UKBB) provides comprehensive data on vascular markers (IMT, PWV), glycemic states (normoglycemia, prediabetes, T2D), and health outcomes like dementia and stroke. It includes detailed information on demographic, lifestyle, and clinical factors, allowing for the exploration of sex differences and the predictive value of vascular markers across varying glycemic states.
- NDA: No NDA indicated
- Website: None

UM4 Characterizing Patients with Type 2 Diabetes and Dementia: A Clustering and Mortality Risk Analysis

Number of students: 2

Subject area: Health science

External organisation: Not indicated

Type 2 Diabetes (T2D) and dementia are complex, interrelated conditions that impact millions globally. These conditions share underlying mechanisms such as inflammation, vascular dysfunction, and metabolic disturbances, which contribute to their progression and impact on mortality. Understanding how the temporal relationship between T2D and dementia influences clinical outcomes can guide more effective risk stratification and personalized management strategies. While significant research has been conducted on individual conditions, the combined effects of T2D and dementia remain poorly understood, necessitating further investigation into their interplay and associated outcomes.

UM4.1 Subtopics

- Temporal Clustering of Type 2 Diabetes and Dementia: Exploring Demographic and Clinical Factors
- Genetic Pathways and Biomarkers in Type 2 Diabetes and Dementia Clusters: A Comparative Analysis

UM4.2 Supervision

- UU supervisors: Dr Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM4.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM4.4 Additional information

- Data description: The available data for this study comes from the UK Biobank (UKBB), a large-scale biomedical database that includes comprehensive health, genetic, and lifestyle information for over 500,000 participants aged 40-85. This dataset provides detailed clinical, demographic, genetic, and imaging data, along with follow-up information on mortality outcomes. Specifically, it includes variables such as HbA1c levels, dementia subtypes, inflammatory markers, genetic risk scores, and cause-specific mortality data, which will be used to analyze the relationship between Type 2 Diabetes and dementia across different temporal clusters.
- NDA: No NDA indicated
- Website: None

UM5 The Role of Glycemic Status, Healthy Lifestyle, and Genetic Factors in the Development of Neurodegenerative Disorders: A Mediation Analysis

Number of students: 2

Subject area: Health science

External organisation: Not indicated

This study explores how variations in glycemic status (normoglycemia, prediabetes, and Type 2 Diabetes), alongside healthy lifestyle behaviors and genetic predispositions, influence the risk of neurodegenerative disorders such as dementia and all-cause mortality. The study aims to investigate the direct and indirect pathways through which glycemic control and genetic factors mediate the development of neurodegeneration, with a focus on understanding how lifestyle interventions and genetic risk scores can modify these associations.

UM5.1 Subtopics

- The Mediating Role of Healthy Lifestyle in the Association Between Glycemic Status and Neurodegenerative Disorders
- The Interaction Between Genetic Predisposition and Glycemic Control in Predicting Neurodegenerative Risk

UM5.2 Supervision

- UU supervisors: Dr Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM5.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM5.4 Additional information

- Data description: The available data for this study is derived from the UK Biobank (UKBB), a large population-based cohort containing genetic, lifestyle, and clinical information from individuals aged 40-85. This dataset provides extensive data on health outcomes, including neurodegenerative disorders, as well as a variety of risk factors such as glycemic status, lifestyle behaviors (diet, physical activity, smoking, etc.), and genetic predispositions through polygenic risk scores (PRS). The UKBB allows for robust analyses of associations between these variables while enabling exploration of mediation and moderation effects.
- NDA: No NDA indicated

- Website: None

UM6 Cognitive Decline Trajectories Across the Glycemic Spectrum: A Longitudinal Study of Normoglycemia, Prediabetes, and Type 2 Diabetes

Number of students: 2

Subject area: Health science

External organisation: Not indicated

The research focuses on examining cognitive decline across the glycemic spectrum—normoglycemia, prediabetes, and Type 2 Diabetes (T2D)—and how it relates to age and sex differences. By using a longitudinal cohort design with neuroimaging and cognitive assessments, the study aims to understand how changes in brain health occur over time in individuals with varying glycemic statuses. The goal is to identify early indicators of cognitive decline, explore underlying mechanisms, and inform personalized interventions to mitigate risks associated with dementia and other cognitive impairments.

UM6.1 Subtopics

- Assessing the Impact of Glycemic Variability on Cognitive Function in Normoglycemia, Prediabetes, and Type 2 Diabetes
- Exploring Structural and Functional Brain Changes Associated with Cognitive Decline Across the Glycemic Spectrum: Insights from Neuroimaging and Cognitive Assessments

UM6.2 Supervision

- UU supervisors: Dr Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM6.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM6.4 Additional information

- Data description: SMART and the UK Biobank offer comprehensive datasets for studying cognitive decline trajectories. SMART includes longitudinal cognitive assessments, neuroimaging (e.g., MRI, PET), and biomarkers related to aging and neurodegenerative diseases. UK Biobank provides a large-scale collection of genetic, clinical, and health data, including detailed glycemic information (HbA1c) and cognitive assessments. These datasets allow for the exploration of cognitive decline in relation to glycemic status, age, and sex, integrating diverse clinical, imaging, and genetic data for a deeper understanding of brain health over time.
- NDA: No NDA indicated
- Website: None

UM7 The Impact of Glucose Dysregulation on Stroke, Dementia, and Mortality Risk in Individuals with Mild Cognitive Impairment: A Focus on Prediabetes and Type 2 Diabetes

Number of students: 2

Subject area: Health science

External organisation: Not indicated

The research focuses on understanding how glucose dysregulation, including prediabetes and Type 2 Diabetes (T2D), impacts the risk of stroke, dementia, and mortality in individuals with Mild Cognitive Impairment (MCI). It explores the interplay between glycemic control, genetic predisposition to diabetes, and adverse cognitive and cardiovascular outcomes. The study aims to provide insights into how these factors contribute to the progression of cognitive decline and associated complications, with a focus on personalized risk assessment and intervention strategies.

UM7.1 Subtopics

- The Role of Glycemic Control in Predicting Stroke and Mortality Risks in Individuals with Mild Cognitive Impairment and Type 2 Diabetes
- Genetic Predisposition to Diabetes and Its Impact on Dementia Progression and Cognitive Decline in patients with Mild Cognitive Impairment

UM7.2 Supervision

- UU supervisors: Dr Fariba Ahmadizar (f.ahmadizar@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM7.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Epidemiology and big data
- Additional requirements: None given

UM7.4 Additional information

- Data description: For this research topic, available data sources include the UK Biobank (UKBB) and Study of Memory, Ageing, and Risk (SMART) cohorts. These datasets provide comprehensive information on individuals with Mild Cognitive Impairment (MCI), including baseline demographic details, clinical measurements (e.g., HbA1c, blood pressure, BMI), genetic data for polygenic risk score (PRS) calculations, and long-term follow-up outcomes such as stroke, dementia, and mortality. Additionally, participants' cognitive function scores and treatment data (e.g., glucose-lowering medications, antihypertensive therapies) are available, allowing for detailed analysis of

the relationship between glucose dysregulation, glycemic control, and adverse outcomes.

- NDA: No NDA indicated
- Website: None

UM8 Building bricks of data and knowledge: automating outpatient clinic scheduling

Number of students: 3

Subject area: Health science,Information and Computing science (including AI)

External organisation: Not indicated

This project focuses on developing data-driven optimalization algorithms to enhance and automate scheduling processes at multidisciplinary outpatient clinics of the UMC Utrecht. The aim is to design intelligent systems that generate efficient and patient-friendly scheduling suggestions for complex appointments across multiple specialties, such as pre-operative screening, oral and maxillofacial surgery, and vascular medicine. By leveraging "bricks of knowledge," these algorithms will incorporate constraints like resource availability (e.g., doctors, rooms) and patient preferences to optimize scheduling efficiency. The outcome is a step toward patient-centric, self-service scheduling, transforming the way appointments are managed in modern healthcare settings.

UM8.1 Subtopics

- Algorithm Development for Constraint-Based Scheduling (such as doctor availability, room capacity, and clinic hours)
- Incorporating Multidisciplinary Dependencies in Scheduling Optimization
- Efficiency Metrics and Performance Evaluation of Scheduling Algorithms

UM8.2 Supervision

- UU supervisors: Joppe Nijman (j.nijman@umcutrecht.nl)
- External supervisors: Not indicated (Not indicated)

UM8.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Casual Inference Methods for Policy Evaluation,Epidemiology and big data,Spatial data analysis and simulation modelling ,Spatial statistics and machine learning ,Using data from routine care
- Additional requirements: An understanding of linear optimization methods is recommended to effectively contribute to the development of efficient scheduling algorithms (sub-topic 1).

UM8.4 Additional information

- Data description: Historical scheduling data from the electronic patient file of all UMC Utrecht outpatient clinics is available for analysis and model development. Additionally, we are currently mapping the workflows of outpatient clinic assistants to reconstruct patient care pathways based on the data, to be used for the modelling.
- NDA: No NDA indicated

- Website: https://www.picudatalab.com and https://umcutrecht.nl

R1 Subjective well-being: comparing different continents

Number of students: 3

Subject area: Social and behavioural science,Law, Economics and Governance

External organisation: Not indicated

The 21st century has witnessed an increasing number of initiatives to measure to explain people's well-being beyond consumption possibilities. Some examples are the annual World Happiness Report published by the United Nations, the Better Life Index of the OECD and Kate Raworth's Doughnut Economics (2017). Also, the number of scientific articles on happiness or subjective well-being has increased tremendously, both in long-standing (psychology and economics) journals and in more recently established specialized journals with 'quality of life' or 'happiness' in their title. A majority of these happiness studies pertain to OECD countries, while the other regions in the world have been studied less often. In the present project the underexposed link between economic development and SWB (subjective well-being) will be analysed in a.o. Latin America and Asia and whether institutions like trust in politics, social networks, religion, mitigate the negative impact of economic hardship on well-being. While doing so, we test whether the findings for developed (OECD) countries in Europe also hold for other continents, of which just a few countries are OECD member.

R1.1 Subtopics

- • Which countries in Latin America are comparable with respect to SWB and the influences of institutions?
- • Which countries in Asian are comparable with respect to SWB and the influences of institutions?
- Which institutions affect SWB and what are the differences between countries?

R1.2 Supervision

- UU supervisors: Dr. Yolanda Grift (Y.Grift@uu.nl)
- External supervisors: Not indicated (Not indicated)

R1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences
- Additional requirements: None given

R1.4 Additional information

- Data description: There are several interesting data sets, like: • European Social Survey (ESS): www.europeansocialsurvey.org • AmericasBarometer Survey (Lapop): https://www.vanderbilt.edu/lapop/ • Asian barometer: https://www.asianbarometer.org/ • OECD, IMF • MICS: https://mics.unicef.org/surveys

As an example. The Lapop data consist of almost all LA countries for several years. "LAPOP is the premier academic institution carrying out surveys of public opinion in the Americas, with over thirty years of experience. As a center for excellence in survey research, LAPOP uses "gold standard" approaches and innovative methods to carry out targeted national surveys; conduct impact evaluation studies; and produce reports on individual attitudes, evaluations, and experiences. The AmericasBarometer survey is the only scientifically rigorous comparative survey that covers 34 nations including all of North, Central, and South America, as well as a significant number of countries in the Caribbean."

- NDA: No NDA indicated
- Website: https://www.uu.nl/organisatie/utrecht-university-school-of-economics-use

R2 Poverty, migration and the labour market: the case of Suriname

Number of students: 2

Subject area: Social and behavioural science,Law, Economics and Governance

External organisation: Not indicated

Poverty in Suriname is mainly explained by education followed by labour market opportunities, area of residence and finally household and family characteristics. Poverty in the rural area's is related to underdeveloped infrastructure and lack of most basic needs as sanitation, clean water and electricity. Less than 5% of the population have access to all three of these basic needs. As for the urban area the poverty profile is quite different. More than 90% of all households have access to basic needs and education. Poverty in urban areas is mainly associated with the living arrangements of the households and the living conditions. Lack of adequate housing, overcrowding, having a decent job and medical insurance, lone parent households with less or no financial resources are most of the problems these households face. Suriname, a South American country with a Dutch colonial background, has a small population of approx. 575.000 inhabitants. Despite the country's huge amount of natural resources such as crude oil and gold, it suffers regularly from economic crises and (hyper)inflation. The country also has an estimated poverty incidence varying from about 20 percent in the urban region to 89 percent in the interior. The huge disparities between these regions in the standard of living and labour market opportunities leads to an ongoing internal and external migration. Currently, almost 70 percent of the total population lives in the urban region which consist of a land area of less than 1 percent of the total land surface. Although the standard of living in the urban area is relatively much better than the rural and interior area, it is far from satisfying. The inequalities between for example major ethnic groups or between married and unmarried couples are huge and needs further research.

R2.1 Subtopics

- Which assets are important in describing poverty?
- Which social economic indicators can be identified to explain poverty?

R2.2 Supervision

- UU supervisors: Dr. Yolanda Grift (Y.Grift@uu.nl)
- External supervisors: Not indicated (Not indicated)

R2.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Dynamics and causality in the social and behavioural sciences
- Additional requirements: n.a.

R2.4 Additional information

- Data description: The data sets are diverse: pooled cross-sectional, panel data: • Suriname Survey of Living conditions: https://mydata.iadb.org/Social-Protection/2022-Suriname-Survey-of-Living-Conditions • AmericasBarometer Survey (Lapop): https://www.vanderbilt.edu/lapop/ • Multiple Indicator Cluster Survey Unicef: https://mics.unicef.org/surveys • Access to 10% Surinamese census data
- NDA: No NDA indicated
- Website: https://openknowledge.worldbank.org/server/api/core/bitstreams/d0b85f03-01fb-4a8c-b22b-ceec56637d4c/content

O1 The 2025 VAST Challenge: Advancing Visual Analytics for Data-Driven Decision Making

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

The IEEE Visual Analytics Science and Technology (VAST) Challenge offers students a unique opportunity to tackle real-world problems using advanced visual analytics and machine learning techniques. Each year, the VAST challenge presents participants with complex scenarios and diverse datasets, providing valuable hands-on experience in data analysis. The competition includes three mini-challenges and a grand challenge, allowing participants to solve specific tasks while exploring ways to integrate insights from multiple data sources. This research project focuses on addressing the 2025 VAST Challenge by developing interactive visualization dashboards and employing analytical methods to solve complex, data-driven problems. Students will combine advanced visualizations with analytical workflows to uncover patterns, detect anomalies, and support decision making. This project not only enhances students' technical and problem-solving skills but also provides an opportunity to contribute innovative solutions to a globally recognized competition.

O1.1 Subtopics

- Exploratory Data Analysis with Interactive Visualizations — VAST Mini-Challenge 1
- Modeling and Analyzing Data Spaces with Machine Learning — VAST Mini-Challenge 2
- Dynamic Workflows for Collaborative Decision Making — VAST Mini-Challenge 3

O1.2 Supervision

- UU supervisors: Angelos Chatzimparmpas (a.chatzimparmpas@uu.nl)
- External supervisors: Not indicated (Not indicated)

O1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Casual Inference Methods for Policy Evaluation,Human Network Analysis,Spatial data analysis and simulation modelling ,Spatial statistics and machine learning ,Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: Basic knowledge of data visualization libraries (e.g., Vega, Plotly, D3.js) and familiarity with machine learning techniques (e.g., clustering, classification) are considered a plus for this project.

O1.4 Additional information

- Data description: The VAST Challenge datasets are diverse in nature, including structured, unstructured, spatiotemporal, or textual data, and are designed to simulate complex, real-world problems. These datasets are released annually in mid-April, giving students ample time to analyze and develop novel dashboards. Additionally, the submission deadline in mid-July aligns well with the schedule of the ADS Master's program, allowing students to integrate this work into their thesis projects and potentially publish their findings at IEEE VIS, the world's leading conference in visualization. Students can also benefit from analyzing solutions to previous challenges to learn from state-of-the-art techniques and refine their own approaches. For instance, the 2024 edition, available at https://vast-challenge.github.io/2024/, showcases award-winning solutions that were presented at the last IEEE VIS conference. These resources provide valuable insights and serve as inspiration for students tackling the new VAST challenge.
- NDA: No NDA indicated
- Website: https://vast-challenge.github.io/2024/

O2 Tonal patterns in classical music

Number of students: 1

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

Before the adoption of major and minor keys in music, compositions were primarily written in in one of the eight modes - or twelve, depending on whom you ask. The transition the modes to the keys took place between 1500 and 1700, a non-linear process. To gain a better understanding of this transition, one of the tools needed is an estimator of the mode of a composition. In music information retrieval, there are multiple key estimators, but these do not work on music composed in the modes. The features that we expect to be strong indicators of the mode are pitch (class) profiles and finals. Additionally, it would be interesting to see which other features are strongly related to modes. The aim of this project is to create a mode estimator by classifing recordings of early music into the mode they are composed in using a selection of (audio) features.

O2.1 Subtopics

- The mode estimator: predict modes of renaissance and early baroque music using pitch (class) profiles and final
- The mode estimator: exploring features related to modal compositions

O2.2 Supervision

- UU supervisors: Mirjam Visscher (m.e.visscher@uu.nl)
- External supervisors: Not indicated (Not indicated)

O2.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Knowledge of music theory is needed:
  keys, reading sheet music.

O2.4 Additional information

- Data description: The data available consists of: - 1000+ recordings of modal compositions - labels indicating the mode of the composition - pitch extractions - features such as pitch profile, pitch class profile, final - code for the extraction of additional features In timely consultation with your supervisor, it is possible to expand the dataset.
- NDA: No NDA indicated
- Website: https://www.projects.science.uu.nl/ics-cantostream/

O3 Using educational assessment questions to predict curriculum alignment

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Not indicated

This study is aimed at automatic evaluation of curriculum alignment. Curriculum alignment refers to the extent to which the curriculum, educational activities and assessment are aligned. Measuring this alignment is a time-consuming process, since all educational materials need to be annotated with information that can be linked to a description of the curriculum. In the current project, we aim to explore if AI can be used to automate the process of annotating assessment questions based on the content of the question itself, possibly enriched with structured information from statistics textbooks. Annotations can be for example about the measured topic, learning objectives or level of the assessment questions. This research makes a societal contribution by improving curriculum alignment, which benefits the quality of student learning.

O3.1 Subtopics

- Comparing machine learning and LLM approaches to predict question annotations
- Enriching question data with textbook information to improve machine learning predictions of question annotations
- How do question properties (length, type, complexity) influence accuracy of predicted annotations?

O3.2 Supervision

- UU supervisors: Matthieu Brinkhuis (UU) / Lientje Maas (Cito) (m.j.s.brinkhuis@uu.nl_lientje.maasATcito.nl)
- External supervisors: Not indicated (Not indicated)

O3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Part of the questions in the dataset are in Dutch, but it is not necessary to be proficient in Dutch.

O3.4 Additional information

- Data description: For this project, we use the publicly available item bank ShareStats (https://www.sharestats.nl/). This item bank contains thousands of questions (and answers) for statistics in higher education. These questions are annotated according to a taxonomy that is also provided and can be found on the provided website.
- NDA: No NDA indicated

- Website: This project is part of a collaboration between UU and Cito (www.cito.nl).

O4 Gravitational wave data analysis with machine learning-enhanced Markov chain Monte Carlo sampling

Number of students: 3

Subject area: Information and Computing science (including AI),Other:

External organisation: Not indicated

Gravitational waves, ripples in the fabric of space-time caused by the mergers of black holes and/or neutron stars, were first detected in 2015. The Einstein Telescope, a next-generation detector proposed for construction in the Netherlands, promises to significantly enhance the detector sensitivity and vastly increase the number of gravitational wave detections compared to current-generation detectors. While this advancement offers exciting opportunities, it also presents significant challenges for data analysis—particularly in estimating the parameters of the sources. Current methods are too slow to keep up with the expected increase in data volume. Recent research from Utrecht has successfully accelerated data analysis for current-generation detectors using machine learning techniques. However, these methods lack the robustness required to handle the simulated signals expected from the Einstein Telescope. This thesis aims to explore existing literature on machine learning-enhanced Markov chain Monte Carlo sampling to identify and address these outstanding challenges.

O4.1 Subtopics

- Comparison of the performance of normalizing flow architectures in adaptive Markov chain Monte Carlo
- Sequential tempering for multimodal recovery in adaptive Markov chain Monte Carlo
- Automated tuning of step-sizes in gradient-based Markov chain Monte Carlo samplers

O4.2 Supervision

- UU supervisors: Thibeau Wouters (t.r.i.wouters@uu.nl)
- External supervisors: Not indicated (Not indicated)

O4.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: All our code is written in Python. However, ideally, we seek students who are confident in programming so that they can easily work with GitHub to collaborate on possible solutions. Moreover, students will have to be able to distill existing codebases quickly and integrate them into their own code. Affinity with machine learning frameworks (PyTorch, TensorFlow, and especially JAX) is a bonus but not necessarily required.

O4.4 Additional information

- Data description: Only simulated data will be used.
- NDA: No NDA indicated
- Website: https://www.uu.nl/en/research/institute-for-gravitational-and-subatomic-physics

E1 Transfer learning with human motion data for explainable fall risk assessment

Number of students: 2

Subject area: Health science

External organisation: Kinetic Analysis, Claudius Prinsenlaan 12, 4811 DK, Breda

Human motion data is valuable for a wide range of medical applications. This study focuses on developing and benchmarking fall risk assessment models for older people using data captured by low-cost, markerless Microsoft Kinect devices. Specifically, the research will leverage pre-trained video models such as MoViNet and SlowFast, and fine-tune them with domain-specific data on human motion. This transfer learning approach allows for the efficient adaptation of general activity recognition models to the specific task of fall risk assessment. As a baseline, these fine-tuned models will be compared with deep learning architectures trained from scratch. A critical aspect of this project is also ensuring the explainability of model predictions. Understanding why a model classifies certain movements or postures as high-risk can provide insights that support clinical decision-making.

E1.1 Subtopics

- Pre-trained models benchmarking for transfer learning with human motion data
- Fair and explainable transfer learning with human motion data

E1.2 Supervision

- UU supervisors: Not indicated (h.l.bodlaender@uu.nl)
- External supervisors: Sofia Yfantidou, Ph.D. (sofia@kinetic-analysis.com)

E1.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Proficiency in English

E1.4 Additional information

- Data description: There are data from N=331 individuals (more than 1.5M of rows) collected with a Microsoft Kinect camera during rehabilitation exercise sessions. The data correspond to individuals with diverse fall risk levels, and are accompanied by demographics allowing for fairness assessments.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: https://www.kinetic-analysis.com/

E2 Data-driven algorithms for patient safety in the nursing ward

Number of students: 2

Subject area: Health science

External organisation: Radboud University Medical Center

Radboudumc is a leader in the field of continuous monitoring of vital signs (heart rate, blood pressure etc.) in patients within nursing wards. This implementation already resulted in a 35% reduction in ICU admissions, showcasing significant improvements in patient safety and care quality. To further reduce workload and improve quality of care, the next step is to develop and prospectively clinically assess AI-algorithms that recognize trends and patterns in this longitudinal data, generating alerts for changes in clinical health status, and serve as decision support tool. Within this project, we aim to develop and adopt data-driven AI-algorithms to enhance complex clinical reasoning and early intervention.

E2.1 Subtopics

- Vital sign variability and patterns in relation to well-defined clinical events
- Vital sign based scores to predict well-defined clinical events

E2.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Bas Bredie, MD, PhD; internist-project lead continuous monitoring (bas.bredie@radboudumc.nl)

E2.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Using data from routine care
- Additional requirements:

E2.4 Additional information

- Data description: Sofar about 15.000 unique patients have been monitored during their admission. Available data of these admissions are: - high resolution vital sign monitor data (500 Hz) - 1-minute vital sign monitor data - large set of clinical data during admission in hospital - dataset with clinical endpoints
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: www.radboudumc.nl

E3 Enhancing Benchmarking Insights Through Machine Learning Applications

Number of students: 2

Subject area: Health science

External organisation: Menzis - Lawickse Allee 130, 6709 DZ Wageningen

Health insurers assess every year where healthcare can be delivered more cost-effectively without compromising quality. To achieve this, benchmark analyses are conducted to find out where spending exceeds proxies for demand. Using propensity score matching (a technique from microeconometrics), each patient is linked to comparable patients and health care costs are compared accordingly. We repeat this analysis for every core region in our patient population. In this topic, participating students will take the lead in developing new tools: Tool 1: One further refinement would be to use machine learning (such as decision trees, random forests or XGBoost) to help us learn what kind of patients score worse than their estimated benchmark and why. And how different regions differ from each other. Possible findings could be that - for Groningen - diabetes patients are generally worse off than expected while - for Utrecht - the elderly care seems suboptimal. While 'diabetics' and 'the elderly' are very large groups, machine learning helps us to pinpoint more exactly where problems arise (e.g. 'diabetics of a particular age, who had billing codes x,y,z indicating foot problems). Tool 2: The findings of our benchmarking analysis reveal significant variation in healthcare costs between mental health institutions, which cannot be attributed to differences in patient populations. To formulate more concrete action points for mental health institutions, Menzis aims to further dissect this variation in healthcare costs. For example, could it be that certain mental health institutions keep patients in care longer than other providers? To answer this question, we first need to benchmark the duration of patient treatment. We aim to achieve this by predicting treatment duration based on a wide range of patient characteristics. This is a complex task, as many factors can influence treatment duration. Fortunately, we have access to a wealth of data, including even the psychologist's estimated treatment duration (which is adjusted over time). In this project, you will use this rich dataset to develop the most accurate possible prediction of treatment duration. This prediction can then be used in discussions with mental health providers. By doing so, you will contribute to more efficient care and potentially help reduce the long waiting lists in the mental health sector.

E3.1 Subtopics

- Machine Learning Applications in Benchmarking: Discovering Localized Healthcare Inefficiencies
- Predicting Mental Health Treatment Duration: A Machine Learning Approach

E3.2 Supervision

- UU supervisors: Not indicated (G.T.Barkema@uu.nl)
- External supervisors: Arthur Hayen, PhD (hayen.a@menzis.nl)

E3.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Casual Inference Methods for Policy Evaluation,Epidemiology and big data,Spatial statistics and machine learning
- Additional requirements: Proficiency in Dutch comes in handy, but it is not a hard requirement. The ideal undergraduate background is also in data science or related fields (econometrics or mathematics).

E3.4 Additional information

- Data description: The data used in this project are rich claims data, complemented with person-level data on demographics and socio-economic status. From claims-level data, information on someone's (mental) health can be derived. You will also have access to these proxies. A final source of data is based on the psychologists' assessment of the treatment duration of their patient. Interestingly, these data are supplied to Menzis at the start of the treatment and can be adapted later on with future claims (e.g. a second session).
- NDA: The external partner requires a specific NDA
- Website: https://www.menzis.nl/

E4 Leveraging Large Language Models to Analyze and Evaluate Hospital Action Plans for Reducing Waiting Lists

Number of students: 2

Subject area: Health science,Information and Computing science (including AI)

External organisation: Menzis

This year, the Dutch Healthcare Authority instructed health insurers to better manage waiting lists in hospital care. Health insurers must hold hospitals more accountable for waiting lists and better monitor their progress on this issue. Menzis now closely monitors the waiting lists and has recently started requesting action plans from hospitals for treatments with waiting times longer than the national guidelines. Hospitals submit these action plans in various formats. Of course, some plans are more detailed than others and also the general quality of each plan varies. It takes a lot of time to manually evaluate all these plans. Therefore, Menzis wants to use a Large Language Model. This model must be able to extract the correct information from the submitted documents and put them in a general Menzis format. Additionally, the model must be able to assess the action plans on several dimensions. For example, whether the problem is well described, whether the actions are sufficiently concrete, and whether a timeline is included. In this thesis, you will develop the necessary prompts based on GPT-4. You will do this using fictitious action plans. Blog post: https://www.skipr.nl/nieuws/zorgverzekeraars-moeten-zorgaanbieders-strenger-aanspreken-op-wachttijden/ Important: This project will only host 1 student, but you will work closely alongside 2 other data science students Menzis is hosting.

E4.1 Subtopics

- Leveraging Large Language Models to Analyze and Evaluate Hospital Action Plans for Reducing Waiting Lists
- N/A (see project description)

E4.2 Supervision

- UU supervisors: Not indicated (i.velegrakis@uu.nl)
- External supervisors: Arthur Hayen, PhD (hayen.a@menzis.nl)

E4.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: Given the topic, proficiency in Dutch is needed.

E4.4 Additional information

- Data description: The main dataset will be a set of fictituous actions plans, based on the hospital action plans that we already received. In part, your data collection will be more

qualitative in nature. An interesting direction would be to let our managers score the fictituous action plans and see whether their scores are close to the scores given bij GPT-4 (and why they deviate).

- NDA: The external partner requires a specific NDA
- Website: https://www.menzis.nl/

E5 Mining for novel insights from the general public to enhance healthcare policies

Number of students: 2

Subject area: Social and behavioural science,Other:

External organisation: Rijksinstituut voor Volksgezondheid en Milieu, Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven

A major challenge for national health institutes is to align health care policies with questions, concerns and doubts that exist in the general public, so as to adapt and better motivate their policies. The vast stream of information that is generated through questionnaires, contact channels and social media makes it difficult to filter out the useful information for these ends. The aim of this project is to identify the common questions, concerns and doubts over time that are voiced in the open text fields of questionnaires and in (transcribed) telephone calls and text messages directed to the Dutch National Institute for Healthcare and the Environment. You will work with approaches from the field of Natural Language Processing.

E5.1 Subtopics

- Identifying concerns, experiences and questions from natural language in a generalizable way
- Recommendation and retrieval of insights from language data in interaction with an end user

E5.2 Supervision

- UU supervisors: Not indicated (f.a.kunneman@uu.nl)
- External supervisors: Mart Stein (mart.stein@rivm.nl)

E5.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: Proficiency in Dutch and is preferable.

E5.4 Additional information

- Data description: Dataset 1 - Infopunt: Incoming emails and transcriptions of telephone calls with questions and remarks from the general public. A sample will be drawn with a focus on one of the case studies (e.g.: vaccinations, cancer screening, nitrogen, heat waves or pandemic preparedness). Dataset 2 - Periodic questionnaire regarding health care policies during the Covid 19 pandemic.
- NDA: The external partner requires a specific NDA

- Website: https://www.rivm.nl/over-het-rivm/strategisch-programma-rivm (projecttab: Onderzoek naar Textmining methoden (EFFICIENT))

E6 Time-Series Forecasting at Lufthanca Cargo to predict Revenue and tonnage

Number of students: 2

Subject area: Geo science,Information and Computing science (including AI),Other:

External organisation: Lufthansa Cargo AG Frankfurt Airport, Gate 21 Building 322 D-60546, Frankfurt am Main

The project idea involves exploring time series forecasting to predict our monthly tonnage and revenue. We are currently developing a model based on a single variable and SARIMA, which already outperforms our Excel-based approach. However, we strongly believe the forecast can be even further improved by incorporating multiple variables into the model. This could be the scope of the proposed thesis. The project would require extensive data processing from various sources, such as capacity, macroeconomic, and holiday datasets. It would also involve comparing different algorithms, hyperparameter tuning, evaluating different prediction time horizons, and understanding the dynamics of our diverse sales regions

E6.1 Subtopics

- Time-Series Forecasting: Predicting monthly tonnage of Lufthansa Cargo
- Time-Series Forecasting: Predicting monthly revenue of Lufthansa Cargo

E6.2 Supervision

- UU supervisors: Not indicated (m.j.vankreveld@uu.nl)
- External supervisors: Luitwin Mallmann (luitwin.mallmann@dlh.de)

E6.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Proficiency in Python and machine learning, ideally with knowledge on time series forecasting, is required.

  Experience with SQL and Databricks and familiarity with Forecasting/Budgeting processes is a plus.

E6.4 Additional information

- Data description: Revenue Data: Historical revenue data from Lufthansa Cargo, segmented by the five sales regions (DACH, Americas, Europe, Africa & India, Asia), covering the period since 2010. Tonnage and Pricing Data: Corresponding data on chargeable weight (tonnage) and yield (pricing). Additional Variables for Prediction: Additional datasets will be incorporated based on their predictive utility. These include:

Capacity Data Holiday Data Macroeconomic Variables: Key indicators such as quarterly GDP, GDP export/import figures, and inflation rates, sourced via the OECD API.

- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: https://www.lufthansa-cargo.com/de/home

E7 Ball speed in Football

Number of students: 3

Subject area: Information and Computing science (including AI),Other:

External organisation: Forward Football B.V., Van Marwijk Kooystraat 10-A, Amsterdam

How can we adapt our existing football events (f.e. passes, dribbles, interceptions, etc.) and position data to calculate and visualise team and individual player's ball speed in passing? What additional metrics or data points can be incorporated to more accurately measure and assess an individual player's impact on ball speed?

E7.1 Subtopics

- Defining ball speed areas: How can we define boundaries of areas on a football pitch based on the player's ball speed in passing, movement patterns and clustering algorithms? Can we define areas with higher ball speeds in comparison to other areas?
- How can we detect or count the impact of a player's ball speed in relation to game play definitions like 'changing sides', creating goal chances', 'space dominance', 'passing forward lines', etc.
- How can the insights gained from individual ball speed analysis be effectively communicated to coaches, players and other stakeholders in a user-friendly and actionable manner?

E7.2 Supervision

- UU supervisors: Not indicated (G.T.Barkema@uu.nl)
- External supervisors: Ryan Stepfner (ryan@forward.football)

E7.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements: - Knowledge of football
  - Knowledge of visualisation (3D) libraries or frameworks might be and advantage for the 3rd sub-topic

E7.4 Additional information

- Data description: We have availability of match tracking data in terms of 15 timestamps per second with all X,Y coordinates of all players and the ball. Also we have accordingly events and timestamps includiing the X,Y positions of all kind of events like passes, dribbels, etc. Next we have data in the form of parameters that were calculated like goal chances and the involved players, the passes and shots related, etc.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: www.forward.football

E8 Sport & Memory: ReFrame

Number of students: 3

Subject area: Social and behavioural science,Media studies,Information and Computing science (including AI)

External organisation: Sport & Memory vof, Boudewijn van Roonstraat 26, 6824AG

Sport & Memory provides stories stories and images about the sports heroes from their youth.

E8.1 Subtopics

- Recognition of emotional moments in historic sportvideo's
- Personalization of historic sports memories
- Sentiment analysis of sports narratives

E8.2 Supervision

- UU supervisors: Not indicated (G.T.Barkema@uu.nl)
- External supervisors: Klaas Jan Bolt (klaasjan@sportandmemory.com)

E8.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Personalisation for (public) media ,Text and Media Analytics
- Additional requirements:

E8.4 Additional information

- Data description: Our Sports historian Jurryt van de Vooren has access to every public domain dataset with regards to sport.
- NDA: No NDA is required
- Website: www.sportandmemory.com

E9 Generative AI for AI driven data management

Number of students: 2

Subject area: Information and Computing science (including AI)

External organisation: Knights Analytics - Domstraat 6 Utrecht

Generative AI has seen a surge in popularity in recent years with the development of the transformer model and the commercialization of transformer-powered chatbots (chatgpt) and retrieval augmented generation (RAG) systems. However, the usefulness of transformers is not only limited to chatbots or search. In recent years, transformers have proven useful for automating enterprise data management tasks that have traditionally been costly and time-consuming, such as record deduplication and linking. Knights Analytics' data management platform leverages transformers to automate the most crucial data management tasks. In this project we will investigate how transformer models can be effectively used for two specific but important data management tasks. The first task is the calculation of name similarity scores. Traditionally, name similarity has been tackled using either string or phonetic matching. However, vector embeddings (e.g. name2vec) and transformer-based extraction of semantic content from names promise to deliver better matching results for use in downstream data quality workflows. In this project, we will fine-tune transformer models specifically for organization name matching. We will then compare embedding-based name matching with string similarity and phonetic matching. The second task is schema fusion. This is the problem of how to fuse the schema of one or more data sources that contain attributes referring to the same concept in multiple different ways and is a critical task for data integration and knowledge graph fusion. This problem is particularly thorny when the attributes come with a latent taxonomy structure: in an ESG context, "carbon emissions" and "Greenhouse emissions" refer to the same property, while "scope 1 emissions" refers to a subtype of carbon emissions. We will build on the LEAPME framework for schema fusion and investigate how to build a property similarity graph that can be used for hierarchical schema fusion.

E9.1 Subtopics

- Name similarity with transformers for entity matching
- schema fusion with latent hierarchical structure

E9.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Riccardo Pinosio (riccardo@knightsanalytics.com)

E9.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Transformers: applications in language and communication

- Additional requirements: familiarity with version control (github etc), solid knowledge of machine learning methods and techniques, ability to chart an independent course into a terra incognita.

E9.4 Additional information

- Data description: To fine tune models for string similarity we will use public datasets like GLEIF or the UK company house data. We will use information extracted from public ESG reports, or research papers to perform schema fusion experiments.
- NDA: The external partner requires a specific NDA
- Website: https://www.knightsanalytics.com/

E10 Machine learning to guide medical diagnostics: development of algorithms to predict outcome of bloodcultures in the Emergency Department

Number of students: 2

Subject area: Health science

External organisation: St. Antonius Hospital Nieuwegein/Utrecht

A bloodstream infection, to the public known as blood poisoning, is an infection where bacteria from gut, skin, lungs or urinary tract enter the bloodstream. This may lead to sepsis, a life-threatening condition where an inappropriate immune response causes organ dysfunction. In Europe, mortality rates of sepsis are estimated to be around 25-30 %. A recurrent dilemma for a clinical physician seeing a patient with signs of a possible bloodstream infection, is whether to order a specialized medical microbiology test ('blood culture') or not. This test measures the presence of bacteria in the (otherwise sterile) bloodstream. The returned result of such cultures is, in roughly 9 out of 10 cases, that no true infection is apparent, indicating the redundancy of a significant proportion of these tests. Not only is this an unnecessary burden on the hospital budget, but also on the patient, as false-positive results may lead to unnecessary medication and prolonged hospital stays. Machine learning applications are thought to be able to guide decision making in two aspects of this problem. First, they may help to correctly identify patients at risk, thus reducing the number of patients for which cultures have to be ordered. Secondly, they may assist in prediction of the species of bacteria causing the infection, thus assisting in guiding choice of antibiotics. Both applications can be of great importance in reducing use of antibiotics and thus, antibiotic resistance.

E10.1 Subtopics

- Machine learning to guide medical diagnostics: predicting the outcome of blood cultures on a species level
- Machine learning to guide medical diagnostics: reduction of positive class misclassification through subgroup discovery methods

E10.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Hanneke Boon (h.boon@antoniusziekenhuis.nl)

E10.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Data Ethics: Responsible data practices and value-sensitive design ,Epidemiology and big data,Using data from routine care
- Additional requirements:

E10.4 Additional information

- Data description: A cleaned dataset containing patient-level clinical and laboratory data for patients with a reported blood culture is available (> 27.000 admissions to the Emergency Department).
- NDA: The external partner requires a specific NDA
- Website: https://www.antoniusziekenhuis.nl/medische-microbiologie-immunologie

E11 How to optimise seat probability in NS trains?

Number of students: 2

Subject area: Information and Computing science (including AI),Other:

External organisation: NS

NB Voertaal binnen NS (en van alle documentatie) is Nederlands, dus opdracht ook in het NL: NS en IenW willen graag dat zoveel mogelijk mensen kunnen zitten tijdens hun reis. Om hierop te sturen is er vanaf 2025 een nieuwe KPI waarover gerapporteerd wordt: zitplaatskans 2e klas. Om hierop te kunnen sturen is het belangrijk dat we een aantal weken of zelfs maanden van te voren de zitplaatskans al kunnen voorspellen. Er zijn verschillende factoren die invloed hebben op deze KPI, bijvoorbeeld: het plan dat is ontworpen, de kwaliteit van de reizigersprognoses, treinen die korter of langer rijden dan gepland en verstoringen op de dag van de uitvoering. Het huidige model kijkt alleen naar het aantal geplande aantal staminuten en gebruikt een vaste factor om deze om te rekenen naar de verwachte zitplaatskans. Echter, bevat het plan nog veel meer variabelen die mogelijk nuttig zijn om mee te nemen.

E11.1 Subtopics

- Voorspelmodel variant 1: focus op kenmerken in de uitvoering die voorspellend zijn
- Voorspelmodel variant 2: focus op kenmerken in het plan die voorspellend zijn

E11.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Loes Knoben (Loes.Knoben@ns.nl)

E11.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Beheersing van de Nederlandse taal is essentieel (zie boven)

E11.4 Additional information

- Data description: De beschikbare data bevat alle NS treinactiviteiten, zowel van de planning als van de uitvoering, met daarbij het aantal reizigers, het aantal zitplaatsen en allerlei bijbehorende kenmerken. Er is ook verstoringsinformatie over treinen aanwezig. Eén maand aan data is ongeveer 1 miljoen regels. Data is representatief vanaf begin 2023.
- NDA: The external partner requires a specific NDA
- Website: www.ns.nl/en/about-NS

E12 Prepayment model monitoring

Number of students: 2

Subject area: Information and Computing science (including AI),Other:

External organisation: Triodos Bank NV, Hoofdstraat 10, 3972 LA Driebergen-Rijsenburg

Triodos Bank uses a Prepayment Model to predict the early repayment (prepayment) of loans and mortgages. This model needs to be monitored to ensure it continues to work properly. For each loan or mortgage, for each month in the upcoming 30 years the model predicts a probability of prepayment. The project can exist of several elements (depending on the students' preferences and capabilities): * Aligning model development data with production data. This problem has the following challenges: - The data used for model development was cleaned up, mismatches with production data need to be identified. - Some data is missing, incomplete or incorrect. - The model development set uses forward looking information, which is not available in the production dataset. - For all the above, assumptions need to be set, preferably based on statistical methods and the impact of these assumptions on performance monitoring needs to be estimated. * Develop a method to monitor the model long-term performance. This problem has several challenges: - to have realisations of a 1-year ahead prediction we need to use data of 1 year ago, and for 2-years ahead we need to look 2 years back etc. This means we have more data on the 1-year ahead prediction than on the 2-year ahead prediction. This can cause a bias in the performance monitoring. - besides prepayments the customer also makes regular repayments which need to be identified and correctly excluded. * Developing a data dashboard to summarise and visualise the results. This has the following challenges: - Model performance has several dimensions, the performance is measured using several methods, over a longer period, over several portfolios and other cross sections. This is a lot of information that needs to be displayed in a meaningful way. - Discuss with stakeholders which information they need for decision making. * Add statistical tests to the model monitoring Besides the visual representation, a statistical test gives an indication of performance. This has the following challenges: - Select an appropriate statistical test for a performance metric, such as R^2, AUC, tests for normality etc. - Take into account the existing correlation between loans and time periods. - Create a method to draw a final conclusion (model is fit / not fit for use).

E12.1 Subtopics

- Long-term performance monitoring of a prepayment model
- Performance metrics for a prepayment model

E12.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Annet Holst (annet.holst@triodos.com)

E12.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Depending on the chosen methods, programming in SAS might be necessary. General programming experience is a big plus. We have experts in-house to help students with this specific programming language. Only students who have a BSN, live in the Netherlands and have a proof of enrolment at preferably a Dutch university (European university is also OK) qualify for an internship at Triodos Investment Management / Triodos Bank. The students can work from home, but are expected in the office (Driebergen-Zeist) at least once a week for a progress meeting. Of course, they are welcome in the office every day if they want to.

  Students will receive a renumeration for 32 hours a week of EUR 400 per month per student.
  Students will receive a Triodos laptop on which all the work needs to be done, as they will be working with customer data. No data can be copied outside of the Triodos environment.

E12.4 Additional information

- Data description: There is a model development dataset available in Databricks (SQL / Python / PySpark development environment). This dataset consists of historical data per loan per month with about 10 years history. This dataset is cleaned up and can be used for brainstorming/developing methods. There is a production dataset with monthly data on loan level on which the model is applied. This dataset is available in SAS Analytics Software. The model monitoring is set up based on this data, which is loaded into Excel. Depending on the chosen methods, the students will either work in SAS or in Excel.
- NDA: The external partner requires a specific NDA
- Website: triodos.nl

E13 AI based road user classification in radar measurements

Number of students: 3

Subject area: Information and Computing science (including AI),Other:

External organisation: Radarxense B.V., Kwekerijweg 2A, 3709 JA Zeist

The radar tracks objects. From all the information on the track, it needs to determine the class of object (truck, bike, car, pedestrian). Properties like signal strength, speed, shape of the object are be used in this classification. The current detection chain works is a manually tuned decision three and there is room for improvement. Your task would be to replace the existing tree and see if it is possible to improve the classification performance using AI. Should the project be succesful, we intend to take the model into production.

E13.1 Subtopics

- Improve the classification performance using AI w.r.t. the existing implementation
- Develop techniques to mitigate overfitting problems in limited data sets with radar data
- TBD

E13.2 Supervision

- UU supervisors: Not indicated (m.j.vankreveld@uu.nl)
- External supervisors: Sjors Hettinga (sjors.hettinga@radarxense.com)

E13.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial statistics and machine learning
- Additional requirements: In this project you will work with radar point clouds (range, angle speed, signal strength). This requires to learn some domain specific aspects. Secondly you will sometimes need to transform a polar coordinate system to a cartesian coordinate system. Knowledge of interpolation techniques or willingness to learn this is mandatory. We would expect the student (max 2) to be on-site 2 to 3 days/week (Mo/Tuesday/Thursday) in our office in Zeist. Radarxense offers a renumuration.

E13.4 Additional information

- Data description: We have more than 45 devices running in the field producing data 24/7 (200 GB/month). For all these devices there is more than 3 hours of annotated video data available. You will be provided with a tool that aligns the annotations and the radar data and provides labelled training data. This will be the starting point for your assignment.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: https://www.radarxense.com/

E14 Domain transfer in satellite and aerial imagery

Number of students: 3

Subject area: Geo science,Information and Computing science (including AI)

External organisation: Readar, Princetonlaan 6, 3584CB Utrecht

Satellite and aerial imagery can have very different radiometric properties due to difference in light conditions, type of camera and in-camera post processing. Can we transfer the style of a source image such that a target image looks the same? Readar extracts information from aerial imagery into valuable information. We detect solar panels to help the energy transition and identify asbestos roofs to prevent health and environment hazards. We currently train our models to cope with differences between images. But we could use simpler and faster training procedures if we would be apply to apply a domain transfer to the images, to make sure they have comparable radiometric properties before we start our detection models. What we offer: we prepare all data, a processing framework and compute to perform your experiments. You will do the desk research into interesting models and inject them into our data framework to perform the experiments. Our office is at UtrechtINC where we welcome you to work in close collaboration with our team.

E14.1 Subtopics

- Domain transfer between two series of aerial imagery
- Domain transfer from calibrated satellite imagery to uncalibrated aerial imagery
- Selection of seasonal invariant areas for optimal domain transfer

E14.2 Supervision

- UU supervisors: Not indicated (Lynda.Hardman@cwi.nl)
- External supervisors: Sven Briels (svenbriels@readar.com)

E14.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements:

E14.4 Additional information

- Data description: Six datasets of aerial imagery each covering the entire Netherlands (from 2021 until 2024). Structured metadata giving time of acquisition, camera type, intrinsic and extrinsic parameters. Multiple datasets of satellite imagery with varying coverage (due to clouds). All data is available via our data framework allowing you to focus on the research.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: www.readar.com

E15 Safety signal detection in pooled clinical data using AI/ML.

Number of students: 1

Subject area: Health science

External organisation: JANSSEN PHARMACEUTICA N.V., TURNHOUTSEWEG 30, 2340 BEERSE Belgium

In the data-driven landscape of clinical development of new therapeutics, the volume of data from human clinical trials that must be screened for potential safety signals is enormous. When a data point falls outside the expected reference ranges, healthcare professionals are tasked with determining whether the value is clinically significant. At present, data analysis primarily occurs in a vertical manner, concentrating on a single compound within a specific indication. However, a horizontal examination of data from clinical trial populations across multiple studies and compounds could provide valuable insights into both how diseases present and how safety signals are detected. This approach could help us understand variations in safety assessment data that are independent of the study drug, identify predisposing factors that contribute to the likelihood of certain safety signals or clinical presentations of diseases, and detect differences in these presentations and assessments based on factors such as gender, age, and ethnicity. Having access to this comprehensive information can mitigate the risk of incorrectly attributing a safety signal to an investigational medicinal product when it may simply be background noise.
Furthermore, it could aid in identifying specific subpopulations that may benefit more from treatment and those that are more susceptible to adverse events.

E15.1 Subtopics

- • Unsupervised ML: detect clusters of clinical trial subjects based on both physiological measurements and covariates such as age, gender, ethnicity, etc in combined data from multiple clinical trials
- • Supervised ML (predictive toxicology): Develop algorithms to predict safety signals based on baseline physiological measurements across therapeutic compounds/diseases.

E15.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Eva Vets (evets1@its.jnj.com)

E15.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: affinity/experience with clinical data preferred but not mandatory

E15.4 Additional information

- Data description: Curated and annotated data from multiple clinical trials, available in a structured (database) format
- NDA: The external partner requires a specific NDA
- Website: na

E16 Improving accuracy of text comparisons in databases

Number of students: 2

Subject area: Information and Computing science (including AI)

External organisation: UWV, La Guardiaweg 116-162 1043 DL Amsterdam

This thesis will focus on building an effective algorithm for comparing text fields in databases. Currently, when information is compared, such as names of people or street numbers, this is done with sql functions or by hand. Using tooling from R/Python, scripts can be written that dicern more effectively between these fields.

E16.1 Subtopics

- Comparing text fields in databases: Street names
- Comparing text fields in databases: People's names

E16.2 Supervision

- UU supervisors: Not indicated (i.velegrakis@uu.nl)
- External supervisors: Hans-Jelle Wolse (hans-jelle.wolse@uwv.nl)

E16.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Casual Inference Methods for Policy Evaluation,Transformers: applications in language and communication
- Additional requirements:

E16.4 Additional information

- Data description: Person one will use a set of street names. The other will use a set of names of people. For the second person additional details will need to be agreed upon regarding the sharebility of the dataset itself.
- NDA: The external partner requires a specific NDA
- Website: None

E17 Enhancing Explainable AI in Generative Models: Designing a Truthfulness Metric for CynthAI©

Number of students: 2

Subject area: Information and Computing science (including AI),Law, Economics and Governance

External organisation: Crowe Foederer, Beukenlaan 60, 5651 CD Eindhoven

CynthAI© is Crowe Foederer's cutting-edge self-service data analytics platform combining the power of generative AI, business intelligence, and advanced analytics to empower users with real-time insights. It allows end users to interact with their data throughs natural language to get the analytical insights that matter to them without requiring a deep knowledge of programming or statistics. The technology behind CynthAI© relies on custom LLM agents developed in-house. A crucial feature of the platform is the Smart Narrative, which generates explanations for data insights and the underlying algorithms, i.e. how did the platform come up with the insights being presented. However, the current feature lacks a way to quantify the truthfulness of the generated narratives, especially in scenarios where the integrity of the data is critical (e.g., for auditors and accountants). The goal of this thesis project is to develop a methodology to assess the truthfulness and completeness of the transactional data, for example by comparing a user's prompt to a pre-defined reference data set. The truthfulness metric would be integrated into the platform's (CynthAI©) Smart Narrative system, ensuring that AI responses are more reliable and verifiable, particularly in contexts requiring rigorous data validation.

E17.1 Subtopics

- • Quantify the truthfulness of AI-generated narratives, for example by comparing the generated text against a pre-defined reference data set.
- • Create a completeness ratio: Ensure that the system can flag discrepancies when the completeness ratio is low.

E17.2 Supervision

- UU supervisors: Not indicated (f.a.kunneman@uu.nl)
- External supervisors: Leon Gerritsen & David Adewunmi (l.gerritsen@crowefoederer.nl)

E17.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements: It would be most beneficial for them to come to our office in Eindhoven for the first few days. At Crowe Foederer, we offer a remuneration of 824 euros per month for interns and students.

E17.4 Additional information

- Data description: We currently apply CynthAI to multiple business lines within Crowe Foederer, from accounting to finance. Students will have access to a selection of high quality financial datasets that they can use for development and reference.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: https://www.foederer.nl/

E18 Metric Selection For Root Cause Analysis of Cloud Infrastructure

Number of students: 2

Subject area: Information and Computing science (including AI)

External organisation: SUE B.V.

Root Cause Analysis in cloud/edge infrastructure is the process of finding the original cause of a failure in infrastructure. This is still a largely manual work and require extensive investigation before finding the origin of the issue. We want to research the possibilities to automate and improve on existing methods in order to minimise the down-time. In this research we are looking into discovering the relevant metrics for finding the root cause of an issue and the related root cause based on it. This project is important as while the root cause has not been found, the IT stack of a institution/company is not able to function properly which is extremely costly.

E18.1 Subtopics

- Metric selection for Cloud Root Cause Analysis
- Root Cause Analysis of Cloud Microservice-based infrastructure

E18.2 Supervision

- UU supervisors: Not indicated (i.velegrakis@uu.nl)
- External supervisors: Nathan Keyaerts (nathan.keyaerts@sue.nl)

E18.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: None given
- Additional requirements: At least a starter knowledge of IT infrastructure [Cloud, Containerisation,...] is highly recommended. We would like the students to be [on average] at least once a week at the office. We offer a 350 euro compensation and transport from the Geldermalsen station to the office.

E18.4 Additional information

- Data description: The following dataset is available for the project: https://lemma-rca.github.io/ LEMMA-RCA is a collection of multi-modal datasets with various real system faults to facilitate future research in RCA. It is multi-domain, consisting of real-world applications such as microservice systems. The datasets are released under the CC BY-NC 4.0 license and hosted on Huggingface, the codes are available on Github.
- NDA: The external partner requires a specific NDA
- Website: https://sue.nl

E19 Trip data and driver classification

Number of students: 1

Subject area: Social and behavioural science,Information and Computing science (including AI)

External organisation: Posthuma Partners, James Wattstraat 77, 1097 DL Amsterdam

An attempt is made to calculate a driver score based on the registered driving behaviour of the driver. The score tells something about the likelihood to be involved in an accident. Alternatively, an attempt is made to classify drivers in different risk groups, again based on the registered driving behaviour.

E19.1 Subtopics

- Driver scoring
- Driver classification / driver outliers

E19.2 Supervision

- UU supervisors: Not indicated (A.P.J.M.Siebes@UU.nl)
- External supervisors: Marco Nijmeijer (nijmeijer@posthuma-partners.nl)

E19.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements:

E19.4 Additional information

- Data description: We have a dataset of about 1 million trips (mostly taxi's) gathered over a period of 7 months. During a trip, data is sampled every 5 seconds. This includes location, direction, speed and distance covered. In addition, so called "acceleration events" are logged: moments of high acceleration (forward, backward, left or right).
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: www.posthuma-partners.nl

E21 Host discovery and workload assesment in a data center network

Number of students: 3

Subject area: Information and Computing science (including AI)

External organisation: Ditio GmbH i.G., Zürcherstrasse 30, 8142 Uitikon

Our startup focusses on the digital transfiromation of large financial institutions like banks. Specifically by helping them transition from an on-premise to a cloud or hybrid inrastructure. Financial institutions often manage portfolios of 500-1000 applications, making manual classification and mapping of applications a time-consuming and tedious process. By leveraging PCAP (Packet Capture) files, which capture network traffic, we can explore automated approaches to identify communication patterns and provide structured input for knowledge graphs. This will enable more efficient cloud migration processes by reducing the manual workload.

E21.1 Subtopics

- Unsupervised learning: How can unsupervised learning methods cluster communication patterns in PCAP files to identify and group application-level network traffic?
- Supervised learning: How can supervised learning models be trained on labeled PCAP data to automatically identify and classify application-level network traffic? (The feasibility of this avenue depends on the availability of labeled PCAP data.)
- Generative AI: How can generative AI models be used to annotate and interpret PCAP files to identify applications and their characteristics?

E21.2 Supervision

- UU supervisors: Not indicated (i.velegrakis@uu.nl)
- External supervisors: Torsten Boettjer (torsten.boettjer@gmail.com)

E21.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning ,Text and Media Analytics
- Additional requirements: A solid understanding of computer system- and network architectures is required

E21.4 Additional information

- Data description: PCAP files capture packets of data transmitted in a data center network. Each packet includes source and destination information, when the packet was captured. and the type of communication that is used (e.g., HTTP, TCP, UDP). This information can be used to map the topology in a data center and allows to classify communication requirements of applications on node level.
- NDA: The confidentiality clause contained in the UNL agreement is sufficient

- Website: www.ditio.cloud (not life yet)

E23 Exploring NLP Models for Generative Storytelling

Number of students: 3

Subject area: Health science

External organisation: Monkey Moves, Australiëlaan 5 Utrecht

At Monkey Moves, we aim to engage parents and young children in life long physical activity by creating dynamic and interactive solutions. We believe that by incorporating image recognition technology into our products, we can enhance the user experience and provide more enticing content. We are looking for a team of talented students that is interested in helping us create engaging stories based on images taken in living rooms or playgrounds. This project revolves around the development of generative language models to create engaging, age-appropriate, and well-aligned (with existing story-arcs and available playground equipment) stories. Students will explore GPT4ALL-compatable models (or equivalent).

E23.1 Subtopics

- How well does WizardLM perform in generating engaging and developmentally appropriate stories for young children?
- How well does Nous-Hermes perform in generating engaging and developmentally appropriate stories for young children?
- How well does Orca-Mini perform in generating engaging and developmentally appropriate stories for young children?

E23.2 Supervision

- UU supervisors: Not indicated (a.gatt@uu.nl)
- External supervisors: Leendert van Gaalen (leendert.vangaalen@monkeymoves.com)

E23.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Text and Media Analytics ,Transformers: applications in language and communication
- Additional requirements:

E23.4 Additional information

- Data description: A meta story is available, character descriptions, existing (audio) stories, access to a story writer, a pilot prompt
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: monkeymovesplay.com

E24 Evaluating optimal methods for comparing LLM outputs

Number of students: 2

Subject area: Information and Computing science (including AI),Other:

External organisation: a.s.r. (Archimedeslaan 10, 3584 BA Utrecht)

Evaluating the quality of output from Large Language Models (LLM's) is inherently subjective. This lack of a definitive ground truth makes it challenging to assess whether suggested changes genuinely enhance the quality of the output. However, as data professionals, it is crucial to adopt a data-driven approach to monitor and validate our improvements. At a.s.r., we leverage a variety of LLM functionalities to streamline processes and unlock new sources of textual information. Particularly for Retrieval-Augmented Generation (RAG) applications, we aim to compare different setups—such as models, prompts, and settings—to identify the most effective configuration. This comparison can be achieved through methods like using an LLM-as-a-Judge or employing metrics. The advantage of RAG applications lies in the ability to facilitate a ground truth, making a final validation possible. Additionally, similar evaluations can be conducted for pure LLM applications, where the answers are inherently subjective and lack ground truths. Your insights and findings will help improve our tools and methodologies for various applications.

E24.1 Subtopics

- Using LLM's to find the optimal method for evaluating general performance of RAG applications
- Invetigating non-LLM methods for determining the optimal evaluation of RAG applications

E24.2 Supervision

- UU supervisors: Not indicated (i.velegrakis@uu.nl)
- External supervisors: Tomaz de Jonge (tomaz.de.jonge@asr.nl)

E24.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: None given
- Additional requirements: Asr offers a renumeration.

E24.4 Additional information

- Data description: As the project entails multiple applications, multiple datasets will be used. In general they consist of a collection of questions and answers (LLM-based and/or human). This can range from medical information to, for example, policy documents.
- NDA: The external partner requires a specific NDA

- Website: https://www.asr.nl/

E25 Classification of logistics real estate

Number of students: 1

Subject area: Geo science,Information and Computing science (including AI),Law, Economics and Governance

External organisation: Statistics Netherlands, Henri Faasdreef 312

The commercial real estate market can provide important insights into the economic health of a region, which is why, since 2019, Statistics Netherlands publishes a quarterly price index series of commercial real estate. Currently this index series is divided into the categories of rental housing, industry, offices, and shops, based on the classification by the dutch governmental register of addresses and buildings, the 'BAG'. There is a need however for structural classification into other real estate segments that cannot be directly derived from the BAG. One interesting segment is logistics real estate, because of the expected growth of the sector and it's popularity with investors. Within Statistics Netherlands we are aiming to differentiate logistics real estate from the currently used commercial real estate segments. If we are successful we can improve our price index series by introducing more homogeneous groups and providing better differentiated analysis possibilities.

E25.1 Subtopics

- Classification of logistic real estate based on 3d BAG and basic registers
- Classification of logistic real estate based on satellite images

E25.2 Supervision

- UU supervisors: Not indicated (Lynda.Hardman@cwi.nl)
- External supervisors: Geert Raaijmakers (g.raaijmakers@cbs.nl)

E25.3 Requirements

- Programming knowledge: Knowledge of Python and R, plus the skills covered in the ADS Master are enough
- Course requirements: Spatial data analysis and simulation modelling ,Spatial statistics and machine learning
- Additional requirements:

E25.4 Additional information

- Data description: There are multiple datascources that could accomodate the proces of classifying logistic real estate. Think of the ABR (Algemeen Bedrijven Register), real estate transaction data, BAG (Basisregistratie adressen en gebouwen) also available in 3D.
- NDA: The external partner requires a specific NDA
- Website: None

E26 Translating Legal Texts to B1 Dutch Language Level

Number of students: 2

Subject area: Information and Computing science (including AI)

External organisation: Voorrecht-rechtspraak

Legal texts, such as contracts and general terms and conditions, often contain complex language that many citizens find difficult to understand. VoorRecht aims to make these texts more accessible by developing a tool that automatically rewrites legal documents to a B1 language level, enabling more people to understand their rights and obligations. A key challenge is maintaining legal nuances while simplifying the text. This research focuses on how machine learning and NLP can help with: â€¢ Detecting complex legal terms and sentence structures and translating them into B1 language level â€¢ Evaluating texts for legal completeness â€¢ Assessing readability and comprehensibility according to B1 guidelines

E26.1 Subtopics

- Detection of complex legal terms and sentence structures and translating them into B1 language level
- Evaluation of legal completeness of B1 translated texts

E26.2 Supervision

- UU supervisors: Floris Bex (f.j.bex@uu.nl)
- External supervisors: Thomas Helling (helling@dataconsultancy.nl)

E26.3 Requirements

- Programming knowledge: This project needs advanced programming skills (e.g. students will need to implement and train complex models, develop algorithms, etc)
- Course requirements: Text and Media Analytics,Transformers: applications in language and communication
- Additional requirements: '- Prefer Dutch speaking
  - We're not completely sure about the extent of the programming skills, but it is required to be able to build complex code pipelines and for machine learning, possibly train models explicitly trained on B1 text?

E26.4 Additional information

- Data description: Case law, model agreements, and potentially advanced (reasoning) LLMs that can assist with evaluation. https://www.ishetb1.nl
- NDA: The confidentiality clause contained in the UNL agreement is sufficient
- Website: www.voorrecht-rechtspraak.nl

ProjectID	How likely is it that you will be hiring once the projects are completed?
E1	Neither likely nor unlikely
E2	Somewhat likely
E3	Somewhat likely
E4	Somewhat likely
E5	Somewhat unlikely
E6	Somewhat likely
E7	Somewhat likely
E8	Extremely likely
E9	Somewhat likely
E10	Somewhat unlikely
E11	Neither likely nor unlikely
E12	Somewhat likely
E13	Somewhat likely
E14	Extremely likely
E15	Not indicated
E16	Somewhat likely
E17	Somewhat likely
E18	Somewhat likely
E19	Extremely likely
E21	Somewhat unlikely
E23	Somewhat unlikely
E24	Somewhat likely
E25	Somewhat likely
E26	Somewhat likely