Wirtschafts- und
Sozialwissenschaftliche
Fakultät

MARKETING AND DATA SCIENCE GROUP
Dr. Aseem Behl

Telefon +49 7071 29-78202
aseem.behl@uni-tuebingen.de
www.uni-tuebingen.de/wiwi/marketing

DS500 Data Science Project – Winter Term 2026/27

As part of the M.Sc. „Data Science in Business and Economics”

Last updated: September 22, 2026

1 Introduction

All students in the master program *M.Sc. Data Science in Business and Economics* enrol in an implementation-focussed module *Data Science Project*. With 12 ECTS, the project has a substantial weight in the curriculum. This is also reflected in our expectation that students show a high degree of independence in their work in this module. Students work together in groups of 2-4 persons.

The basic idea of this module is that students **independently** design and program the entire process of a data science project from start to finish.

This requirement places a focus on **automation**, which means that the procurement of data, the structuring, reading, validation, modification, and analysis take place (as far as possible) without the intervention of the analyst or user of the analysis.

A data science project should include the following components:

1. **Data acquisition**: The project acquires and reads in the data, for example, by scraping data from web pages, by using APIs, or by reading text documents, images, or video data. It is also possible for students to create a new dataset semi-manually themselves or to use an existing dataset as the basis for the project. The complexity and the effort of the data acquisition are considered in the grading of the project.
2. **Data preparation and validation**: The data are read in and made ready for further analysis, computation, and/or estimation. Among other things, this means that unstructured data (e.g., texts) are converted into computable formats and that different data sources are combined meaningfully. The data should be validated, for example, by generating descriptive statistics and figures as well as by identifying implausible observations and outliers. After having defined the criteria for validation, this process should also take place mainly without the users' intervention.

3. **Data analysis**: The primary focus of the project is a relevant substantive question, which can, but does not necessarily have to be of an economic nature. The corresponding analysis should be implemented with adequate methods (traditional statistics or machine learning).
4. **Validation of the results**: It is crucial to validate the results of the analysis by suitable approaches. This is to make sure that the results from your analysis are helpful and practically relevant.
5. **Presentation of the results**: All project outputs (e.g., tables, graphs, results) should be embedded in an interactive environment, for example, in the form of a webapp. Results should be presented in an appealing and informative way, according to the state-of-the-art of data visualization.
6. **Automation**: The entire project with the above-mentioned steps should run automatically. When new data periodically arrives, a user who did not create the code can evaluate and analyse it without having to "touch" the code.
7. **Version control and collaboration:** As soon as the topic, data sources, structure, and type of analysis are defined, the code, including the version history, must be organized in a GitHub repository.
8. **Documentation**: The whole project should be well-documented so that external parties can easily understand and review what is happening in your code and how different components of your architecture work together. This includes a meaningful structure of your repo (directory structure and file names) and comprehensibly commented code, among others. In addition, there should be a document that guides the reader through your repo (e.g., a README) that includes, for example, which files must be executed and in what order.

2 Difference between a master thesis and the Data Science Project

A master thesis focuses on solving a relevant research problem, which must be embedded in the relevant literature. The analyses are described in the text, and the results are printed in the form of static tables or graphs. The thesis then interprets the results and draws conclusions. There are no requirements for automation, and the reader "consumes" the text without interacting with the data themselves.

	Master thesis	**Data Science Project**
Elaborate research problem	High weight	Low weight
Literature work	High weight	Low weight
Automation	Low weight	Very high weight
Reader/user interacts with data	Low weight	Very high weight
Conclusion	Mostly author	Rather reader/ user

3 Evaluation/Grading

Students present their results (as a group) twice: In an interim presentation, they present the plan and first steps. In the final presentation, they show the completed project, the central structure of the code, the architecture behind their “product”, and how users can interact with the result.

The project is evaluated according to the following criteria:

1. Complexity of the problem
   a. How complex are the different components of the project?
   b. How novel is the problem and the proposed solution?
2. Quality of the execution
   a. How is the user experience of the final product?
   b. How well does the automation work? How well is the error-handling?
   c. Is the implementation logically consistent?
   d. Are the product/results appropriately validated?
3. Quality of the documentation
   a. Are the project architecture and code structure easily understandable?
   b. Is the code easily readable by an external party?

For assessment, students will produce the following components:

1. A presentation that motivates the central question, derives its relevance, and refers to relevant literature as appropriate. Furthermore, students present their project’s architecture and technical implementation as well as the finished “product”, e.g., a website/dashboard or a comparable way to interact with the data as a user. The presentation also contains a discussion in which supervisors and other students ask questions.
2. A verbal summary (max. 3 sentences) of their Data Science Project.
3. A short video (approx. 3 minutes) that clearly presents the project and the result to a broader circle of users. The video can be a screencast where only the presenter's voice is heard while the final product is shown, or it can be a video presentation in which the presenters are also visible. The target group is a non-technical audience; e.g., in a business context, it would be a manager, not a fellow data scientist.
4. Students provide the code by granting the supervisors access to the GitHub repository or by providing a link to the repository if it is public.
5. Since the dataset might be very large, there is no need to upload it to GitHub or ILIAS. However, be prepared to grant supervisors access, e.g., by a dedicated database user.
6. The interface (e.g., webapp) must be available to other students and teachers, e.g., by hosting it on the bwCloud.

Evaluation is done at the individual level, not at the group level. Thus, the group members’ contributions to the project and presentations will be considered in the evaluation.

All components except for the presentation slides must be submitted one week prior to the final presentation (see schedule below).

4 Schedule

Until Sunday, October 11, 2026	Registration via alma	
Wednesday, October 14, 2026 14:00 – 15:00	Kick-off; assignment of topics, supervisors, and teams	SR E01 (Mohlstr. 36)
Wednesday, November 4, 2026 9:00 – 13:00	Workshop “Presentation Skills”	SR331 (Mohlstr. 36)
Wednesday, November 25, 2026 14:00 – 18:00	Interim presentations	SR209 (Mohlstr. 36)
Friday, January 22, 2027 23:59	Submission of project (except slides for final presentation)	---
Thursday, January 28, 2027 20:00	Submission of slides for the final presentation	---
Friday, January 29, 2027 9:00 – 14:00	Final presentations	SR209 (Mohlstr. 36)

Note: Participation on all sessions is mandatory. Times might change depending on the number of participants.

5 Topics

In addition to the topics listed below, students are encouraged to propose their own topic ideas. If you would like to do so, please send your topic proposal at least a week prior to the kick-off date via e-mail to Dr. Aseem Behl (aseem.behl@uni-tuebingen.de) and Jakob Zgonc (jakob.zgonc@uni-tuebingen.de).

Topic 1: Reported deforestation vs. actual situation based on satellite images

Deforestation is commonly monitored through national reports and international statistics, which are used to track environmental change and evaluate climate policies. However, reported figures may not always reflect the actual extent of forest loss on the ground due to reporting delays, methodological differences, or political incentives. Satellite imagery provides an independent and consistent way to observe changes in forest cover over time, allowing for a data-driven comparison between reported deforestation and observed land-cover change.
This project investigates whether officially reported deforestation levels correspond to the patterns visible in satellite data. The comparison could focus on identifying discrepancies in magnitude, timing, or spatial distribution of forest loss.

Topic 2: Gaps in the Bike Lane Network

Many cities aim to promote cycling as a safe and sustainable mode of transportation, but bike lane networks are often incomplete: dedicated cycling infrastructure may abruptly end, forcing cyclists to share busy roads or take long detours. These discontinuities in the bike lane network can reduce safety and discourage cycling. The goal of this project is to use spatial and mobility data to identify such gaps by analyzing where cyclists frequently travel on streets without dedicated bike infrastructure. By combining street network data, existing bike lanes, and indicators of bicycle usage, the project should develop a data-driven approach to detect and prioritize missing links in the network. The results should highlight locations where new bike lanes or bike streets would most improve network continuity and provide evidence to support policy makers and urban planners in planning future cycling infrastructure.

Topic 3: Course Planning and Recommendation for Master's Students

New master's students often face the challenge of planning their studies under several constraints: mandatory courses, prerequisites, and the fact that not all courses are offered every semester. At the same time, students want to take the courses that best match their interests, but poor planning early on can make this difficult later in the program. The goal of this project is to develop a module planning and recommendation system that helps students create a feasible multi-semester study plan before or at the beginning of their studies. Using program requirements, course offerings from alma, and scheduling constraints, the system should recommend course combinations that allow students to complete all requirements while maximizing their ability to take the courses they are most interested in, without compromising progress toward their degree.

Topic 4: Optimal Locations for Wind and Solar Energy Parks
The transition to renewable energy requires large areas for wind turbines and solar parks, but identifying suitable locations is a complex planning problem. Potential sites must satisfy multiple criteria, including natural resource availability, legal constraints, environmental protection, land use, and economic feasibility. In Germany, for example, renewable energy installations must comply with local land-use plans minimum distances to residential areas, and environmental regulations. At the same time, practical factors such as proximity to the power grid, terrain slope, distance to urban areas, and existing infrastructure strongly influence the economic viability of a project. The goal of this project is to develop a data-driven approach to identify and rank suitable areas for wind or solar parks to provide decision support for planners and policymakers involved in renewable energy expansion.

Topic 5: Tracing the Origin of AI-Generated Content
With the rapid spread of generative AI, it is becoming increasingly difficult to determine whether content was created by humans or generated by AI systems. At the same time, questions arise about which model may have produced the content, whether it has been post-processed or edited, and whether parts of it might originate from copyrighted material in training data. The goal of this project is to develop an app that analyzes uploaded content (e.g., text, images, or audio) and estimates the likelihood that it was generated by AI. Using features such as stylometric patterns, visual artifacts, audio characteristics, and machine learning models, the system should provide an interpretable assessment of the content's origin, possible model signatures, and the extent of human post-processing. Such a tool could help educators, journalists, and policymakers better assess the authenticity and provenance of digital content.

Topic 6: Explainable Portfolio Tracker with Multi-Factor Risk and Return Attribution
This project is about building an explainable portfolio tracker that combines investment analytics with modern data science. In addition to recording trades, holdings, and performance, the system should analyze the economic drivers behind a portfolio's behavior. Students will model portfolio returns and risk using a multi-factor framework, identifying exposures to systematic sources such as market, sector, region, and style factors including value, size, and momentum. The goal is to move from simple tracking to interpretable portfolio explanation. The project integrates data engineering, statistics, financial modeling, and visualization. Students will reconstruct holdings over time, compute standard performance measures, estimate factor exposures, and decompose both return and risk into meaningful components. More advanced versions may compare classical linear factor models with machine learning methods that capture nonlinear interactions. The final output should be an interactive analytical tool that helps users understand not only how a portfolio performed, but also why.