Discipline: Information Systems / Operations Research

1. Language
English

2. Title
Machine Learning

3. Lecturer
Professor Dr. Stefan Lessmann, School of Business and Economics, Humboldt-University of Berlin
https://www.wiwi.hu-berlin.de/de/professuren/bwl/wi/personen/hl
stefan.lessmann@hu-berlin.de

4. Date and Location
The course will be offered over four days, comprising lectures, tutorials, and discussion sessions.
Course period: 6th – 9th April 2027

aQua-Institut GmbH für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH,

Maschmühlenweg 8 – 10, Raum Emmy Noether I + II

37073 Göttingen

5. Course Description
5.1 Abstract and Learning Objectives

Machine learning (ML) has become an important methodological toolkit for empirical research in business and economics. It enables researchers to analyze complex, high-dimensional data, develop accurate predictive models, learn representations from data, and address research questions that are difficult to study with conventional statistical methods alone. At the same time, the growing availability of powerful software libraries, foundation models, and generative AI shifts attention from merely implementing ML algorithms toward selecting appropriate methods, designing credible empirical studies, critically evaluating model outputs, and understanding the boundaries of predictive analysis.

The course provides an introduction to modern ML, emphasizing its use in empirical research. It addresses participants with heterogeneous prior experience. To that end, we start from the foundations of supervised learning and model evaluation. Covering these foundations, we progress rapidly toward contemporary developments such as neural representation learning, transformers, foundation models, modern methods for tabular data, and machine learning for time series and longitudinal data.

Throughout the course, methodological concepts are connected to questions of empirical research design. Participants discuss how ML can support prediction, measurement, pattern discovery, and other empirical research tasks; how ML studies should be evaluated; and under which circumstances predictive modeling alone is insufficient. The latter perspective motivates an introduction to related research directions including causal inference, algorithmic decision-making, and responsible machine learning.

After completing the course, participants should be able to:

- understand the main principles underlying contemporary supervised machine learning;
- select and critically evaluate suitable ML approaches for empirical research problems;
- design appropriate validation and model-comparison strategies;
- understand the main ideas behind neural networks, representation learning, transformers, and foundation models;
- recognize methodological challenges arising from tabular and longitudinal data;
- interpret and critically assess predictions from complex ML models;
- identify research questions for which ML can provide substantive or methodological value;
- distinguish predictive questions from causal and decision-oriented research questions and recognize important limitations of predictive ML.

The course targets Ph.D. students who are new to machine learning as well as participants who already have practical ML experience and want to deepen their understanding of recent methodological developments and their implications for empirical research.

5.2 Content

The course progresses from the foundations of predictive machine learning toward contemporary methods and, finally, toward the use and limitations of ML as an empirical research methodology.

The first part establishes the methodological foundations required for subsequent topics. It introduces ML from the perspective of empirical research and discusses the relationship between statistical modeling and predictive ML. Participants revisit core approaches to supervised learning, including regularized models, tree-based methods, ensembles, and boosting. Particular attention is paid to model evaluation, validation, hyperparameter tuning, data leakage, and the design of credible out-of-sample comparisons.
The second part introduces major developments in modern ML. Starting from neural networks and representation learning, the course develops the ideas underlying attention, transformers, pretraining, transfer learning, and foundation models. Natural language serves as an illustrative context where useful, whereas structured, tabular as well as longitudinal data represent the clear focus of the course. Specifically, we discuss contemporary ML for tabular data, including tabular data foundation models, and examine the challenges that arise when observations are temporally ordered or repeatedly observed. ML for time series, panel, and

longitudinal data is considered with emphasis on temporal dependence, appropriate validation strategies, changing data distributions, and selected recent developments including time-series foundation models.

The third part turns from predictive performance toward interpretation and empirical research. Participants study approaches to interpretable and explainable ML and discuss what model explanations can and cannot reveal about empirical relationships. Building on the methods covered earlier, the course then considers different roles that ML can play in research, including prediction, measurement, pattern discovery, and methodological investigation. The course concludes by examining the boundaries of predictive ML. In particular, it distinguishes prediction from causal inference and decision-making and discusses how issues such as interventions, economic consequences, constraints, and algorithmic fairness can give rise to additional research questions.

Throughout these parts, examples and discussions emphasize applications in business and economics and the design of scientifically credible ML-based research.

5.3 Course Schedule

The course consists of 15 sessions distributed over four days. Lecture sessions are complemented by interactive formats that encourage participants to connect the methodological material to their own research.
The following course outline distinguishes between lecture (L), discussion (D), and practice (P) sessions.

Pre-course stage		
		Study papers from reading list
		Prepare a short 5-min presentation of your research (see below)
		Familiarize with Python and Jupyter notebooks
Day 1		
		Arrival of participants
09:00	10:30	Welcome, Introduction, and ML Foundations (L)
10:30	11:00	Coffee break
11:00	12:30	Supervised Learning: Models, Regularization & Ensembles (L)
12:30	13:30	Lunch break
13:30	15:30	Science Slam (P)
15:30	16:00	Coffee break
16:00	17:30	Research Problems through an ML Lens (D)
Day 2		
09:00	10:30	ML Model Evaluation, Validation & Tuning (L)
10:30	11:00	Coffee break
11:00	12:30	Neural Networks & Representation Learning (L)
12:30	13:30	Lunch break
13:30	15:30	Audit and Critical Evaluation of an ML Analysis (P)
15:30	16:00	Coffee break

16:00	17:30	Transformers & Foundation Models (P)
Day 3		
09:00	10:30	Modern ML for Tabular Data & Tabular Foundation Models (L)
10:30	11:00	Coffee break
11:00	12:30	ML for Time Series & Panel/Longitudinal Data (L)
12:30	13:30	Lunch break
13:30	15:30	ML Research Design Clinic (P)
15:30	16:00	Coffee break
16:00	17:30	Interpretable & Explainable Machine Learning (L)
Day 4		
09:00	10:30	ML as an Empirical Research Method (D)
10:30	11:00	Coffee break
11:00	12:30	Beyond Prediction: Causality, Decisions & Responsible ML (L)
12:30	13:30	Lunch break
13:30	15:30	Closing session: Discussion of the course assignment & next steps (D)
Post-course stage		
6 to 8 weeks		Development of a computational essay demonstrating the use of ML in research. The assignment should ideally connect to the participant's own research. Details will be discussed during the course.

5.4 Course format

The course combines conceptual lectures with discussion, collaborative analysis, participant presentations, and practical work. The different formats are intended to accommodate participants with heterogeneous backgrounds while progressively connecting methodological concepts to empirical research.

Participants introduce their own research interests in a short science-slam format at the beginning of the course. These research problems provide examples that can be revisited throughout the subsequent sessions. Further interactive sessions include the critical assessment of an ML analysis and a research-design clinic in which participants consider how machine learning can contribute to empirical research questions.

Computer-based analysis remains an important component of the course, and examples use contemporary Python libraries and ML tools where appropriate. However, the practical sessions do not focus primarily on reproducing programming procedures. Instead, emphasis is placed on understanding modeling choices, critically analyzing them, designing appropriate empirical strategies, and making effective use of contemporary computational tools.

The course deliberately combines foundational material with selected recent developments. Participants without prior ML experience should acquire a solid conceptual basis for using ML in research, while more experienced participants are exposed to contemporary topics and broader questions concerning research design, model interpretation, foundation models, and the limits of predictive machine learning.

The final assignment allows participants to apply and further develop these skills in an ML-related empirical project that should ideally connect to their Ph.D. research.
The course language is English.

6. Preparation and Literature

6.1 Prerequisites

- Master-level education in Business, Economics, Information Systems, Computer Science, Engineering, or a related field.
- Prior experience with machine learning is **not required**. The course is designed to accommodate participants ranging from ML newcomers to Ph.D. students with ample experience in ML.
- Participants should be familiar with basic concepts from statistics or econometrics, including regression analysis, statistical inference, and the interpretation of empirical models.
- Some course activities and the final assignment involve working with data and ML software. Participants should therefore have basic familiarity with Python and Jupyter notebooks or acquire it before the course. Advanced programming skills are not required.
- Participants are not expected to implement ML algorithms from scratch. The course uses contemporary software libraries and computational tools, and participants may use AI-based coding assistance where appropriate.
- Participants should bring a research-oriented perspective. Having a current or prospective Ph.D. research question, dataset, or methodological problem in mind is encouraged, as some course activities connect ML concepts to participants' own research.

6.2 Essential Reading Material

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the opportunities and risks of foundation models*. https://arxiv.org/abs/2108.07258
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023
- Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31*(2), 87–106. https://dx.doi.org/0.1257/jep.31.2.87

6.3 Additional Reading Material

- Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11(1), 685-725. https://doi.org/10.1146/annurev-economics-080217-053433
- Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular

foundation model. *Nature*, *637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. http://dx.doi.org/10.1038/nature14539
- Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. https://dx.doi.org/10.1038/s42256-019-0048-x

6.4 To prepare
Participants are expected to study the essential reading material before the course. Familiarity with the additional readings is helpful but not required.

In addition, each participant should prepare a **five-minute presentation** for the research science slam on the first day of the course (see outline). The presentation should introduce the participant's research interests, current project, or a research problem they find interesting. The emphasis should be on the **research question, empirical setting, data, and methodological challenge**. A connection to ML is welcome, but **not required**. In particular, participants are not expected to know in advance whether or how ML is suitable for their research.

The presentation should be brief and accessible to Ph.D. students from different areas of business and economics („science slam“). Participants should avoid technical detail that is unnecessary for understanding the research problem and instead focus on what they study, why the question matters, and what analytical challenges they face.

Participants who are at an early stage of their Ph.D. and do not yet have a defined dissertation topic are equally welcome. In this case, the presentation may introduce a broader research area, an empirical phenomenon of interest, a dataset they would like to work with, or a methodological question they would like to explore during the course.

Because some course activities involve computational work, participants should also ensure before the course that they can work with Python and Jupyter notebooks at a basic level and that they have access to a suitable local or cloud-based Python environment. Advanced programming skills are not required.

7. Administration

7.1 Max. number of participants
The number of participants is limited to 20.

7.2 Assignments
none
7.3 Exam
After the course, participants can work on a machine learning assignment and write up results as a computational essay (i.e., Jupyter Notebook). The assignment is necessary to obtain 6 ECTS from the course.
Typically, each participant will work on a different modeling task. Ideally, the assignment task connects to participants' research. Alternative assignment topics include replicating a published paper or working on a Kaggle competition (http://www.kaggle.com). The course schedule leaves room to discuss possible topics for the assignment. Participants will submit their assignment solution roughly six weeks after the end of the course. The submitted notebooks will be graded according to the quality of the exposition, the complexity of the modeling tasks, and the degree to which ML concepts have been used successfully.

7.4 Credits
The course corresponds to a scope of 6 LP/ECTS

8. Working Hours

Working Hours	**Stunden**
Mandatory readings	30 h
Preparation for the course (programming, regression analysis, presentation)	40h
Active participation in class	30 h
Final exam (practical assignment to be completed and written-up after the course)	90 h
SUMME	**180 h**