LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Seminar Kick-Off

- Explaining Black-Box Predictions: A Game-Theoretic Approach to Reliable Model Interpretability
- Research Seminar in Explainable AI: Reproducibility and Critical Evaluation of Methods

14.10.2025

Chair for Statistical Learning and Data Science

Head of Chair: Prof. Dr. Bernd Bischl

Research:

- **Interpretable Machine Learning / Explainable AI**
- Automated Machine Learning and Optimization
- Causal and Fair Machine Learning
- Empirical Machine Learning
- Machine Learning for Survival Analysis
- Methods Beyond Supervised Learning
- Probabilistic Machine and Deep Learning
- Research Software Engineering

Find us here: https://www.slds.stat.uni-muenchen.de/

Goals of this Seminar

- Topic: know about *Advances in Interpretable Machine Learning*
- Literature Search: know how to navigate scientific publications
- Writing: how to write a scientific report using a journal template
- Presentation: give a great presentation about a scientific topic
- → Perfect preparation for a Master's thesis in this field
- **Seminars:**
  - Explaining Black-Box Predictions: A Game-Theoretic Approach to Reliable Model Interpretability
  - Research Seminar in Explainable AI: Reproducibility and Critical Evaluation of Methods
  - Results of a similar Seminar HERE
- **Check Guidelines for Seminars which contains info about**
  - Scope, deliverables, and roles (presentation, report, discussant/reviewer)
  - Suggested resources/tools and potential collaboration expectations (in case of teamwork)
  - Grading criteria (presentation and final report)

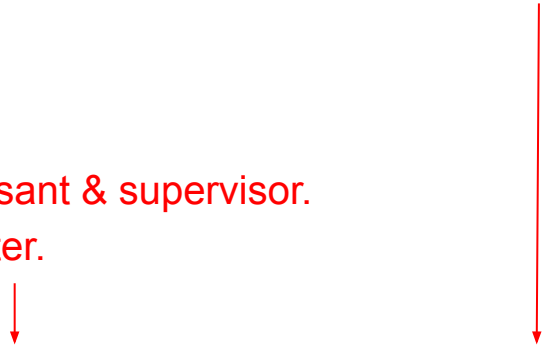
Structure of this Class

1. **Phase 1: Introduction**
   - Today (14.10): Introduction, Orga, Topics
   - 21.10.2026 from **14:30 - 16:00**: Topic Assignment and Q&A Session
2. **Phase 2: Presentation Slots**
   - Tue, 03.02.2026 09:00 - 18:30 (~10 slots) -> Research Seminar?
   - Tue, 10.02.2026 09:00 - 18:30 (~10 slots) -> Game-theoretic Seminar?
   - Backup: 11.02.2026 and 12.02.2026 possible
3. **Phase 3: Report** (Final report: 15.03.2026)

→ Participation in all meetings/presentations of **your seminar** is mandatory

Structure of this Class - Timeline

- **Presentation (relative to X):**
  - X − 7 days: Presenter sends slide draft to discussant & supervisor (ready for review).
  - X − 3 days: Discussant returns feedback.
  - X: Presentation.
- **Report (fixed dates):**
  - March 1: Presenter sends current draft to discussant & supervisor.
  - March 8: Discussant returns feedback to presenter.
  - March 15: Final report due.



Red dates only relevant for single speakers who work alone on a topic (not for teams), since we expect that within a team you review each other's work anyways.

Structure of this Class - Single vs. Team

Individual (single speaker) + role as discussant/reviewer for another student

- Talk: 30 min + 10 min discussion (±10%; discussant leads).
- Report: 20 - 40 pp (excl. refs/appx), lower bound ok.
- **Discussant/Reviewer** duty (for another student's seminar topic):
  a. Write ≤ 3-page review 7 days before report submission (include into your report)
  b. Provide actionable feedback 3 days before presentation to other student
  c. Lead Q&A and discussion phase

Team (no discussant/reviewer needed)

- Talk: 35 min to 40 min + 5 to 10 min discussion (±10%), same report limits.
- Contribution transparency for teamwork:
  a. Add lead authors to each section in your report (first author, second author)
  b. Add dedicated "Author Contributions" section
  c. Work on GitHub, each member commits under own account
     -> Use issues, branches, PRs, peer review, so contributions are traceable
     -> Ensure fair distribution of tasks

Presentation

- Overview, examples, deep dive into research questions of paper
- Synthesize the topic
  - Bring across the common problem
  - Discuss how this embeds into related work
- Critically reflect on the paper
  - Strengths: strong experiments, timely/impactful problem, computationally fast, …
  - Weaknesses: bad experimental setup, wrong claims, unrealistic assumptions, …
- Describe your mini contribution
  (e.g., new extension, comparison, simulation, illustration of limitation, etc.)
- Suggestions:
  - Bring examples to better illustrate the method
  - Go beyond summarizing the paper to demonstrate deep understanding of the topic
  - Look for additional material on the topic (often there are videos, blog posts, etc.)
    - ➢ Can give a broader view, often authors note things not written in the paper

Report

Context & Literature

- Place the paper in the broader literature context and conduct a thorough literature review
- Clearly motivate the problem addressed, mention any classical approaches to solve the given problem
- Summarize prior, parallel, and follow-up works that address the same/a similar problem
- Discuss the paper’s potential impact

Core analysis and mini contribution

- Explain contributions, methods, key results, assumptions, and limitations with minimal illustrative examples
- Go beyond summarizing: demonstrate deep understanding and critical thinking
  - State or illustrate strengths and weaknesses of the paper/method
  - Think of what would you have done differently
- Add a small original element (e.g., tiny methodological extension, ablation/comparison, new simulation, stress test) and report setup and findings

Academic Standards

- Write with clarity, structure, rigor and ensure reproducibility
- Use proper notation, consistent citations, and a curated bibliography

Communication

- Materials will be uploaded on the Moodle page
- Announcement will be made on Moodle
- Communication with me should be done via email
    - only use LMU university addresses

FAQ

- Does the content of the report has to be organized in the same way as the presentation slides?
  - No. In fact, the report should go beyond the presentation content.
- Should I discuss other techniques or papers in my report?
  - Definitely! You should not stick to the paper we provided. If there is a really similar technique, it might make sense to discuss it in detail and point out the differences.

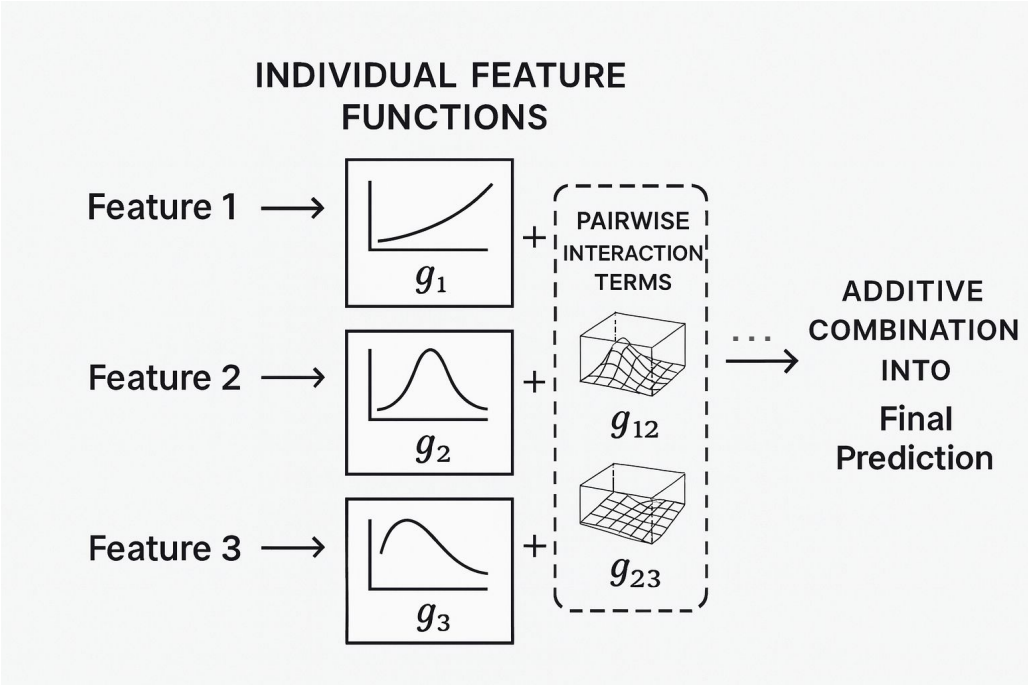
Questions?

Reminder:

- Use LaTeX
- Use the JMLR style file
- 20-40 pages

Functional Decomposition to obtain Interpretability

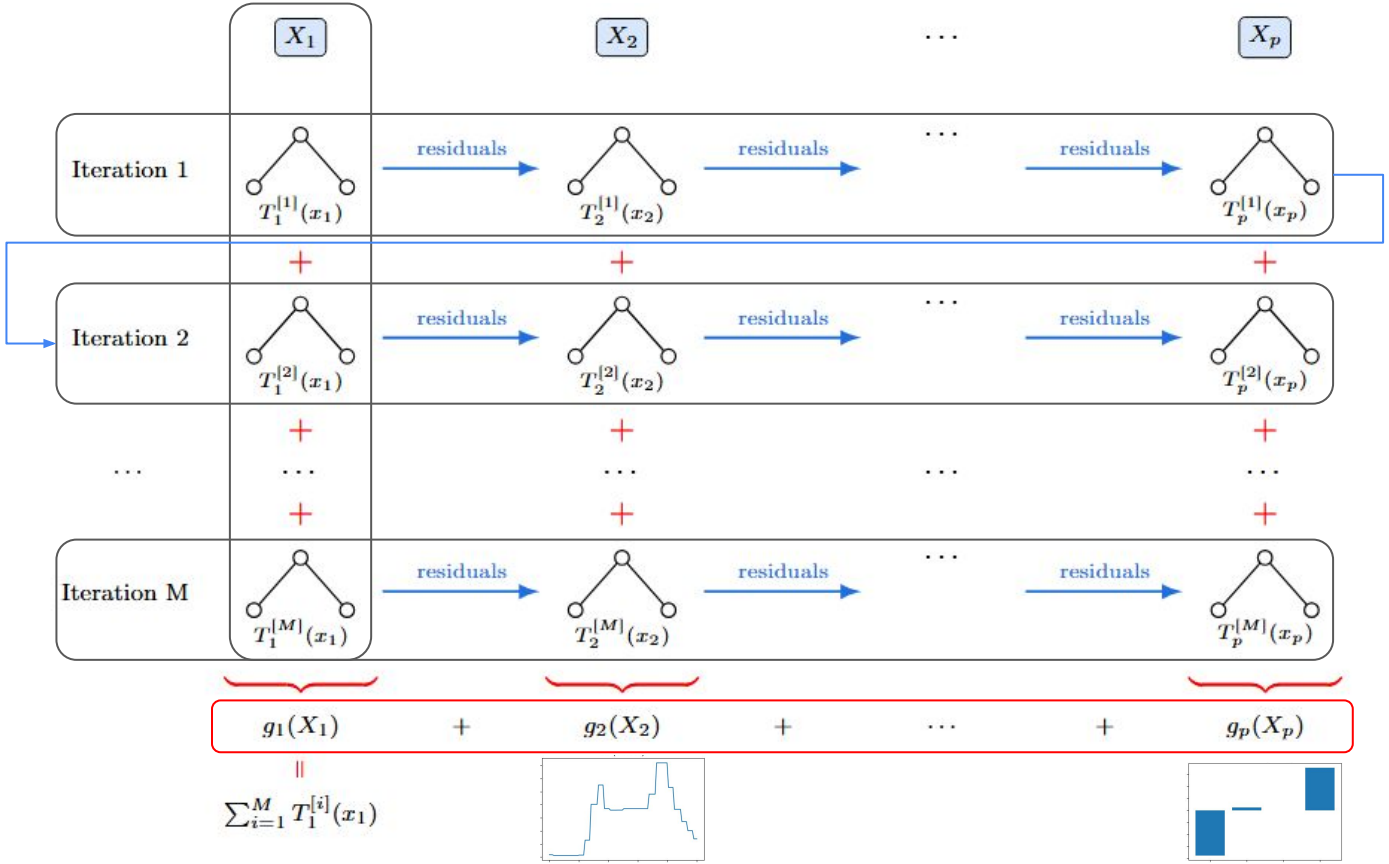
$$f(x_1, \ldots, x_p) = g_0 + \sum_{j=1}^{p} g_j(x_j) + \sum_{1 \le i < j \le p} g_{ij}(x_i, x_j) + \cdots + g_{1\ldots p}(x_1, \ldots, x_p).$$



Post-hoc:
Given a trained black-box model, try to recover additive structure (no retraining)

Inherently interpretable:
Fit a model that directly learns this additive structure (GAMs, NAMs, EBMs/GAMI-Tree, …)

Example: Using Trees to obtain Additive Structure



Explainable Boosting Machines (no interactions)

Explaining Black-Box Predictions: A Game-Theoretic Approach to Reliable Model Interpretability

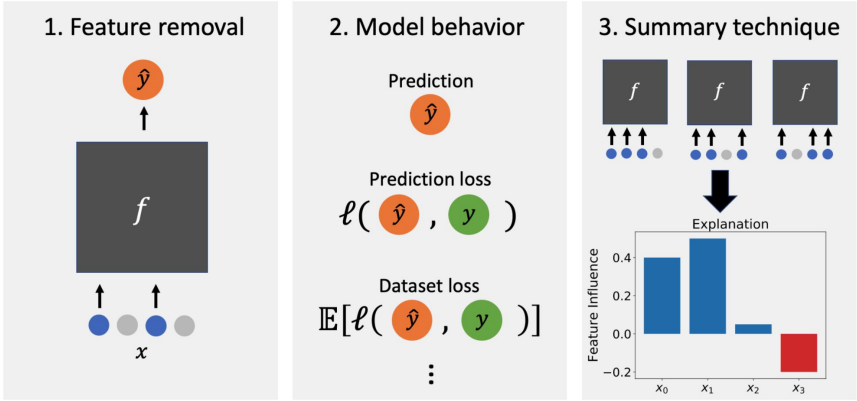


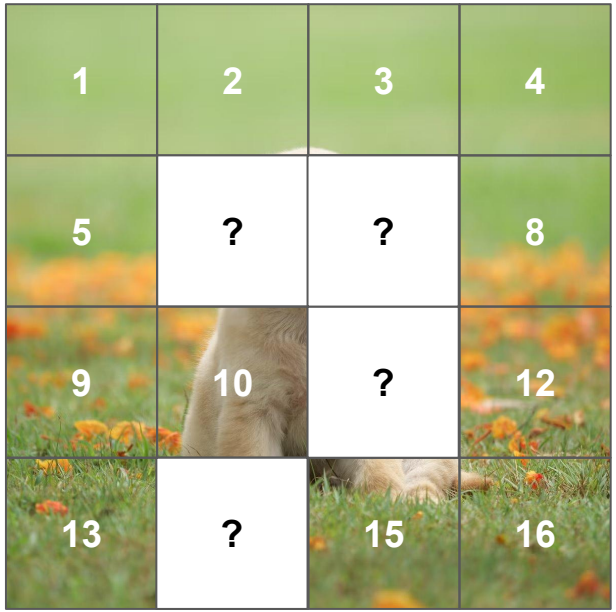
Figure 1: A unified framework for *removal-based explanations*. Each method is determined by three choices: how it removes features, what model behavior it analyzes, and how it summarizes feature influence.

Covert, Ian, Scott Lundberg, and Su-In Lee. "Explaining by removing: A unified framework for model explanation." *Journal of Machine Learning Research* 22.209 (2021): 1-90.

Explaining Black-Box Predictions locally

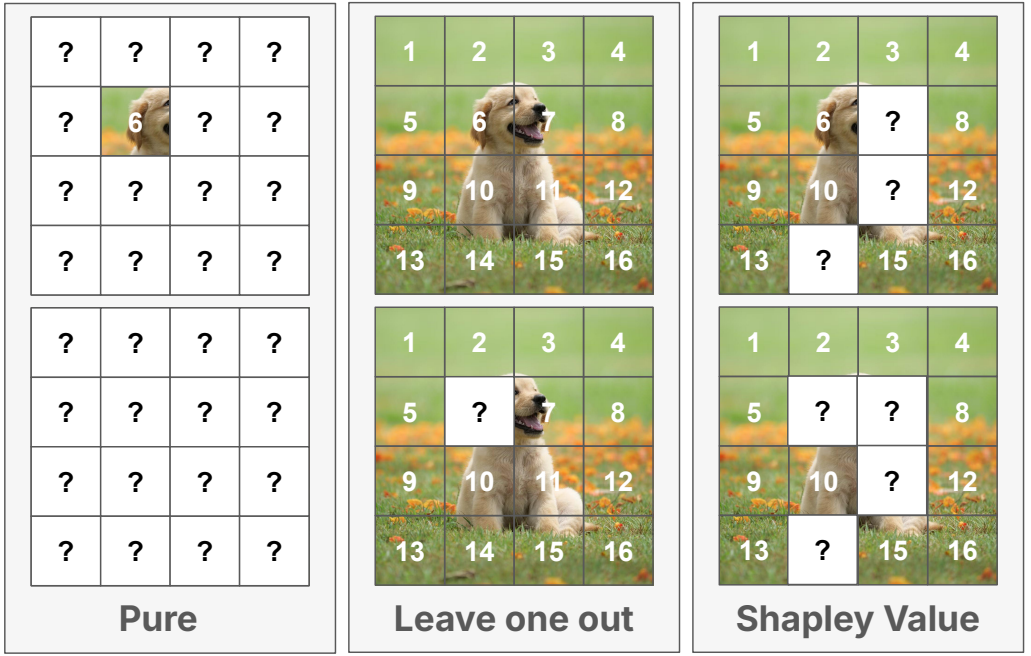
Feature Removal (masked predictions)

Output prediction for any subset



Summary Technique (choose subsets)

Compute difference in predictions with and without feature



Selection Process for each Seminar

1. Research Seminar in Explainable AI: Reproducibility and Critical Evaluation of Methods

Topics 

Topic Assignment Poll 

2. Explaining Black-Box Predictions: A Game-Theoretic Approach to Reliable Model Interpretability

Topics 

Topic Assignment Poll