03-IMAP-LLML	Interpretable Machine Learning

Lehrform *(teaching format)* / **SWS** *(hours per week)*: 4K

Kreditpunkte *(credit points)*: 6

Turnus *(frequency)*: every second winter term

Inhaltliche Voraussetzungen *(content-related prior knowledge/skills)*: Machine learning basics

Sprache *(language)*: English

Lehrende *(teaching staff)*: Prof. Dr. Marvin Wright

Studiengang *(degree program)*	Module	Semester
Informatik (Master)	IMVP, IMVP-AI	ab 1.Sem.
AI and Intelligent Systems (Master)	AI-M-MLCS	from 2nd sem.
Management Information Systems (Master)	(MIS-INF3)	from 2nd sem.
Informatik (Bachelor)	(nur *Freie Wahl*)	

Lernergebnisse / *Learning Outcome*:

- Understand the theoretical and practical basics of interpretable machine learning (IML), including general concepts and specific methods
- Be aware of pitfalls and challenges with IML methods
- Be able to implement IML methods yourself in Python (or R or another language of choice)
- Be able to apply IML methods in practice (on real data)
- Be able to understand and explain the results of IML methods

---

Inhalte / *Contents*:

- Dimensions and scope of interpretable machine learning
- Interpretable models, e.g., (generalized) linear models, rule-based models
- Feature effects: individual conditional expectation, partial dependence, accumulated local effects
- Shapley values and Shapley additive explanation (SHAP)
- Feature importance: permutation feature importance, conditional feature importance
- Local interpretable model-agnostic explanations (LIME)
- Counterfactuals and adversarial examples

Hinweise *(remarks)*: The table lists only the primary / most specific modules to which this course is assigned.