Einladung
zum Doktorandenkolloquium

Data-Efficient and Controllable Machine Learning
in Resource-constrained Domains
von Herrn Julka Sahib
am Mittwoch, 6.11.2024 ab 13:25 Uhr im SR 027, WiWi, Innstr. 27 und online per Zoom (die Zoomdaten wurden intern gemailt)

(**Betreuer:** Prof. Dr. Michael Granitzer)

Abstract:
Deep learning models are inherently data-intensive, requiring substantial volumes of labelled data to achieve optimal performance. This requirement poses a significant challenge in specialised fields where data is scarce and the cost of labelling is prohibitively high. Consequently, progress in applying machine learning in these areas is often slow and constrained by limited resources. This thesis addresses these challenges by focusing on two main objectives crucial for advancing machine learning in resource-constrained environments:

a) *building data-efficient pipelines to reduce dependency on large, annotated datasets,* and,

b) *developing controllable and interpretable representations to enhance the effectiveness and applicability of simulations.*

The first part of the thesis concentrates on creating methodologies that maximise the utility of limited data resources, thereby reducing reliance on extensive manual labelling. This segment includes innovative work in Deep Active Learning and Knowledge Distillation applied to planetary science applications, enhancing the efficiency of models under resource constraints. Additionally, this section explores the novel use of foundation models as active annotators, extending their application to the domain of rare human languages and testing their effectiveness in a new context.
The second part of the thesis focuses on the development of controllable and interpretable representations, vital for creating effective and practical simulations. Within the context of trajectory prediction application, this segment discusses methods to learn generative factors of variation both implicitly and explicitly, concluding with a chapter that introduces a new quantitative metric to measure the quality of disentangled representations.
Collectively, these initiatives address the challenges posed by the intensive data requirements of contemporary machine learning models. Providing scalable solutions that improve the practicality and effectiveness of machine learning in specialized domains, this work makes a relevant contribution to the field, facilitating broader application and greater operational efficacy of machine learning techniques in challenging environments.