Title: "Concept-Level Explainable AI".

Abstract: "The emerging field of Explainable AI (XAI) aims to bring transparency to today's powerful but opaque deep learning models. This talk will present Concept Relevance Propagation (CRP), a next-generation XAI technique which explains individual predictions in terms of localized and human-understandable concepts. Other than the related state-of-the-art, CRP not only identifies the relevant input dimensions (e.g., pixels in an image) but also provides deep insights into the model's representation and the reasoning process. This makes CRP a perfect tool for AI-supported knowledge discovery in the sciences. In the talk we will demonstrate on multiple datasets, model architectures and application domains, that CRP-based analyses allow one to (1) gain insights into the representation and composition of concepts in the model as well as quantitatively investigate their role in prediction, (2) identify and counteract Clever Hans filters focusing on spurious correlations in the data, and (3) analyze whole concept subspaces and their contributions to fine-grained decision making. By lifting XAI to the concept level, CRP opens up a new way to analyze, debug and interact with ML models, which is of particular interest in safety-critical applications and the sciences"

Bio: „Wojciech Samek is a professor in the EECS Department at the Technical University of Berlin and is jointly heading the AI Department at Fraunhofer Heinrich Hertz Institute. He studied Computer Science at Humboldt University of Berlin and received the PhD in Machine Learning from the Technical University of Berlin in 2014. He is Fellow at the BIFOLD - Berlin Institute for the Foundation of Learning and Data, the ELLIS Unit Berlin and the DFG Research Unit DeSBi. Furthermore, he is a senior editor of IEEE TNNLS, an editorial board member of Pattern Recognition, and an elected member of the IEEE MLSP Technical Committee and the Germany's Platform for Artificial Intelligence. He is co-author of more than 180 publications, co-editor of two Springer books on Explainable AI, and recipient of multiple best paper awards, including the 2020 Pattern Recognition Best Paper Award and the 2022 Digital Signal Processing Best Paper Prize."