Veranstaltung | Talk

Concept-Level Explainable AI

Speaker: Prof Wojciech Samek (Fraunhofer HHI und TU Berlin)

Fr
05.05.

Uhrzeit
14.00–15.00

Ort
Campus Schöneberg
Haus B Room B 2.06 and online
Badensche Straße 50–51
10825 Berlin

Google Maps

Kosten
kostenfrei

Anmeldung
Online on BigBlueButton – Access code: 841500

To the event

Veranstalter/in
Institute for Data-Driven Digital Transformation (d-cube) in Kooperation mit der Methodenwerkstatt Statistik

Zum Institut

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