ID de Contribution: **5** Type: **Non spécifié**

Interpreting Deep Neural Networks towards Trustworthiness

mercredi 19 avril 2023 12:30 (30 minutes)

In this talk, I will start by describing our contextual decomposition (CD) method to interpret neural networks, which attributes importance to features and feature interactions for individual predictions.
Using CD to interpret DL models in a cosmology problem led us to develop an adaptive wavelet distillation (AWD) interpretation method.
AWD is shown to be both outperforming deep neural networks and interpretable in the motivating cosmology problem and an external validating cell biology problem.
Finally, I will address the need to quality-control the entire data science life cycle to build any model for trustworthy interpretation.

Orateur: Prof. YU, Bin (Berkeley)