School of Data Science
香港城市大學
City University of Hong Kong

ONLINE SEMINAR

# Deep Neural Networks Explainability: Algorithms and Applications

Date: 22 February 2022 (Tuesday)

Time: 10:00am - 11:00am

Seminar link: https://cityu.zoom.us/j/99422424079

## ABSTRACT

Deep neural networks (DNN) have achieved extremely high prediction accuracy in a wide range of fields such as computer vision, natural language processing, and recommender systems. Despite the superior performance, DNN models are often regarded as black-boxes and criticized for the lack of interpretability, since these models cannot provide meaningful explanations on how a certain prediction is made. Without the explanations to enhance the transparency of DNN models, it would become difficult to build up trust and credibility among end-users. In this talk, I will present our efforts to tackle the black-box problem and to make powerful DNN models more interpretable and trustworthy. First, I will introduce post-hoc interpretation approaches for predictions made by two standard DNN architectures, including Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). Second, I will introduce the usage of explainability as a debugging tool to improve the generalization ability and fairness of DNN models.

## Dr Mengnan DU

### GUEST SPEAKER'S PROFILE

Dr Mengnan Du is currently a Ph.D. student in Computer Science at Texas A&M University, under the supervision of Dr. Xia Ben Hu. His research is on the broad area of trustworthy machine learning, with a particular interest in the areas of explainable, fair, and robust DNNs. He has had around 40 papers published in prestigious venues such as NeurIPS, AAAI, KDD, WWW, NAACL, ICLR, CACM, and TPAMI. He received over 1,200 citations with an H-index of 11. Three of his papers were selected for the Best Paper (Candidate) at WWW 2019, ICDM 2019, and INFORMS 2019, respectively. His paper on Explainable AI was also highlighted on the cover page of Communications of the ACM, January 2020 issue. He served as the Registration Chair of WSDM'22, and is the program committee member of conferences including NeurIPS, ICML, ICLR, AAAI, ACL, EMNLP, NAACL, etc.

**Enquiries: sdscgo@cityu.edu.hk**

**All are welcome**