ELECTRICAL & COMPUTER ENGINEERING **BIO-SEMINAR** **Spring 2020** TEXAS A&M UNIVERSITY Engineering

When: **Friday 12:40 – 13:30, March 27, 2020**

Where: https://tamu.zoom.us/j/514754727

Speaker: **Mengnan Du**

Ph.D. Student in Prof. Xia (Ben) Hu's DATA Lab
Department of Computer Science and Engineering
Texas A&M University

Title: **Towards Interpretable and Credible Deep Neural Networks**

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 by 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. On the other hand, interpretability alone is insufficient for DNNs to be credible, unless the provided explanations conform with the well-established domain knowledge. That is to say, correct evidence should be adopted by the networks to make predictions. In this talk, I will present our efforts to tackle the black-box problem and to make powerful DNN models more interpretable and credible. Firstly, 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). Secondly, I will focus on a novel training method to regularize the interpretations of a DNN with domain knowledge, aiming to develop more credible DNNs.

Bio: Mengnan Du is currently a 3rd year Ph.D. student in Computer Science at the CSE department of Texas A&M University, under the supervision of Dr. Xia Ben Hu. His research is on interpretable machine learning, with a particular interest in the areas of DNN interpretability. He has had more than 10 papers published by several prestigious data mining conferences including KDD, WWW, ICDM, CIKM and WSDM. Three of his papers were selected for the Best Paper Candidate at WWW'19, the Best Paper Candidate at ICDM'19, INFORMS 2019 Best Refereed Paper Finalists, respectively. One of his papers was highlighted on the cover page of Communications of the ACM, January 2020 issue.