TEXAS A&M
Institute of
Data Science

SEMINAR SERIES

Challenges and Opportunities of Self-Consuming Loops in AI

Presented by Dr. Richard Baraniuk, Rice University

There are many arguments for training deep learning models on synthetic data, including 1) with the advent of trillion parameter models, we are simply running out of training data; 2) synthetic training data is much cheaper to source than real data; 3) some important domains lack large training data sets. Moreover, even if you think you are training your model on clean, real data, it is likely that your dataset has been polluted inadvertently with synthetic data. Our preliminary research in this space has shown that synthetic data training creates a feedback loop that, over generations of models, can amplify artifacts, increase biases, and reduce the models' quality and diversity (aka, "model collapse" or "MADness"). The potential negative ramifications of model collapse reach across the entire spectrum of machine learning applications, from generative models to decision making system to signal and image processing systems. It is imperative that we design new machine learning models and training paradigms that are at least robust and at best immunized against synthetic data. In this talk, we show that this is indeed possible.

Richard G. Baraniuk is the C. Sidney Burrus Professor of Electrical and Computer Engineering at Rice University and the Founding Director of OpenStax and SafeInsights. His research interests lie in new theory, algorithms, and hardware for machine learning, signal processing, and sensing. He is a Member of the National Academy of Engineering and American Academy of Arts and Sciences and a Fellow of the National Academy of Inventors, AAAS, and IEEE.

November 10th
2:00 - 3:00 pm
WEB 232 and Zoom