Prerequisites: A background of undergraduate linear algebra and probability is expected. CS37300 Data Mining & Machine Learning will be useful. Registration is by instructor's permission.

Learning objectives: By the end of the course, the students will be familiar with the importance of explainable AI and state-of-the-art techniques for ML interpretability.

Course Syllabus: This course is designed to provide the students an understanding of the importance of interpretability in machine learning through model design and data attribution.

As machine learning is becoming more and more ubiquitous, practitioners are designing models of varying complexity to suit their predictive needs. In this course, our focus would be to study machine learning models beyond their predictive capabilities to develop a deeper understanding of workings of the model to answer questions like — Which data points are most important for a prediction being made? Which features or model components are most relevant and can we attribute quantifiable sub-tasks to certain model components? How can we make complicated models like neural networks more understandable? Can we take our understanding of human learning and apply them to machine learning ? What are the possible biases that can arise in a machine learning model and how can we address them ? Answering such questions can be crucial in sensitive fields to disentangle how a machine learning model works and to allow for possibly debugging it if need be.

The course will be a mix of lectures and paper presentations from relevant recent publications in various ML fields including explainable AI and machine vision. We will also often employ tools from optimization and ML theory. Each student should expect to present one or two papers. Class attendance and participation is highly encouraged. There will be a class project due towards the end of the course and students will be also be asked to peer-review project presentations.

Grading:

Class participation: 5%
Project proposal: 10%
Project presentation: 20%
Project final report: 15%
Peer-reviews: 20%
Paper presentation: 30%