**A theoretical look at adversarial examples and dataset poisoning**

Tom Goldstein, University of Maryland

Neural networks can solve challenging computer vision and natural language problems with human-like accuracy. However, it has been shown that most neural networks are easily fooled by "adversarial examples" - inputs that contain small perturbations and patterns that result in large changes in the behavior of the network. This property of neural networks can be exploited by bad actors to manipulate neural networks and take control of autonomous systems. This talk investigates adversarial examples from a theoretical perspective. After introducing the idea of adversarial attacks, and presenting a new class of "poisoning" attacks, I will discuss how classical theorems from high-dimensional probability can be used to derive fundamental bounds on the robustness of neural networks.