Abstract:

Machine learning classifiers are becoming increasingly popular, and often achieve outstanding performance in testing. When deployed, however, classifiers can be thwarted by motivated adversaries who adaptively construct adversarial examples that exploit flaws in the classifier's model. Over the past few years, a vibrant research community has emerged focused on the problem of adversarial examples, instigated by (seemingly) surprising results showing how vulnerable deep learning classifiers are to adversaries finding small distortions to inputs that fool a classifier. In this talk, I will argue that adversarial examples are a much older problem, and that they stem from a fundamental problem of distinguishing correlation from causation. I will describe a new very simple strategy, feature squeezing, that can be used to harden classifiers by detecting adversarial examples. Feature squeezing reduces the search space available to an adversary by coalescing samples that correspond to many different inputs in the original space into a single sample. Adversarial examples can be detected by comparing the model's predictions on the original and squeezed sample. In practice, of course, adversaries are not limited to small distortions in a particular metric space. Indeed, in security applications like malware detection it may be possible to make large changes to an input without disrupting its intended malicious behavior. I'll report on an evolutionary framework we have developed to search for such adversarial examples that can automatically find evasive variants against state-of-the-art classifiers. This suggests that work on adversarial machine learning needs a better definition of adversarial examples, and to make progress towards understanding how classifiers and oracles perceive samples differently.
Bio:
David Evans (https://www.cs.virginia.edu/evans/) is a Professor of Computer Science at the University of Virginia where he has been a faculty member since 1999. He leads the Security Research Group (https://www.jeffersonswheel.org), with current research foci on multi-party computation, adversarial machine learning, and web security. He is the author of an open computer science textbook (http://www.computingbook.org) and a children's book on computability (http://www.dorimic.org). He won the Outstanding Faculty Award from the State Council of Higher Education for Virginia, an All-University Teaching Award. He was Program Co-Chair for the 24th ACM Conference on Computer and Communications Security (CCS 2017) and the 30th (2009) and 31st (2010) IEEE Symposia on Security and Privacy. He grew up in a suburb of Detroit, Michigan, and has SB, SM and PhD degrees in Computer Science from MIT.