Website Fingerprinting in Tor: An Adversarial ML Approach

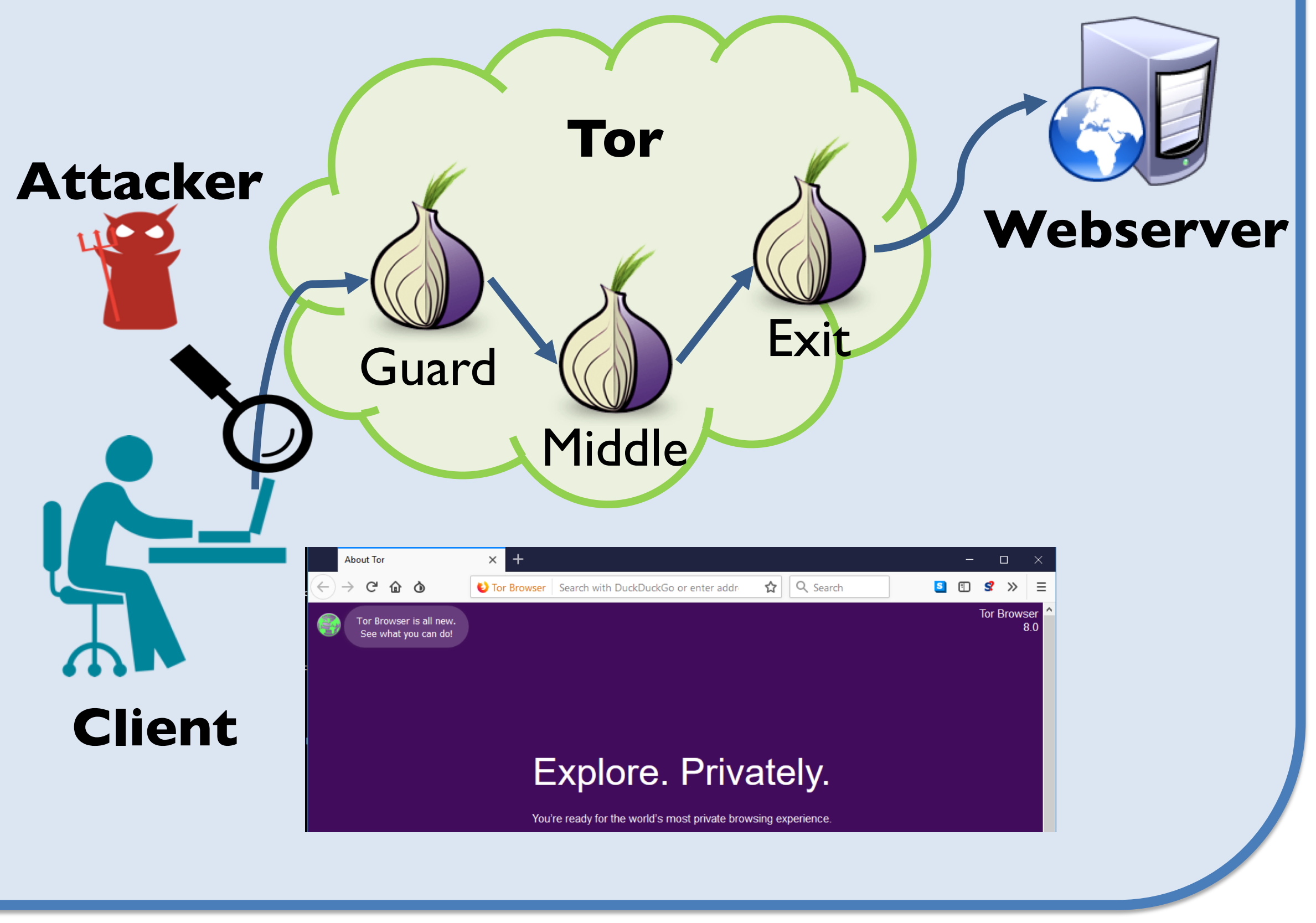
Matt Wright, Rochester Institute of Technology matthew.wright@rit.edu

http://www.rit.edu/cybersecurity

Collaborators: Claudia Diaz (KU Leuven), Kantha Gangadhara, Nick Hopper (U. Minnesota), Jack Hyland, Aneesh Joshi, Marc Juarez (KU Leuven), Nate Mathews, Se Eun Oh (U. Minnesota), Mike Perry (Tor Project), Mohammad Saidur Rahman, Payap Sirinam (Navaminda Kasatriyadhiraj Royal Air Force Academy)

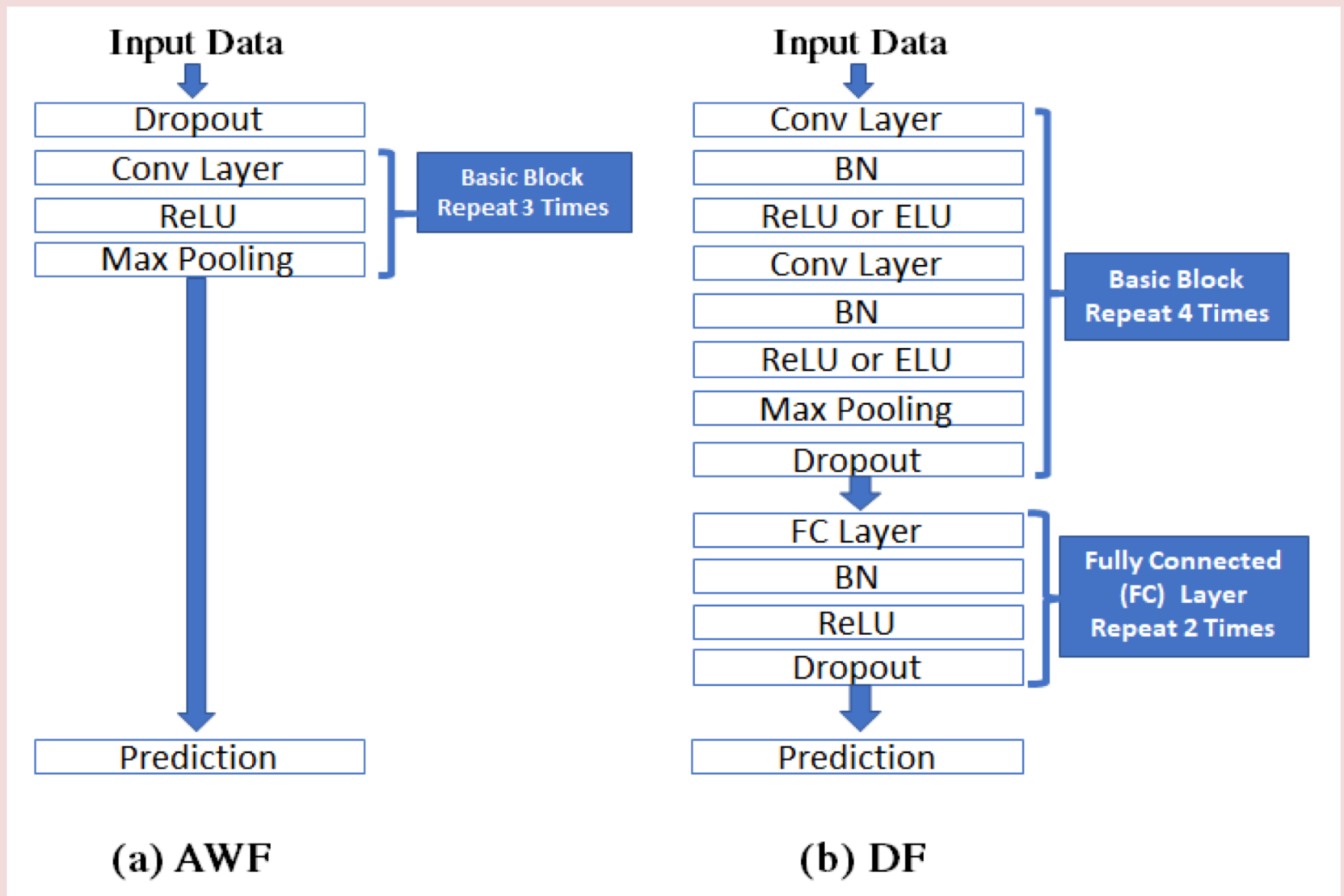
Challenge: Website Fingerprinting (WF)

- Tor protects user privacy online for 8M people/day
- But it can be attacked by a local eavesdropper who:
  1. Trains a ML classifier on traffic patterns
  2. Uses the classifier to uncover the client's activity

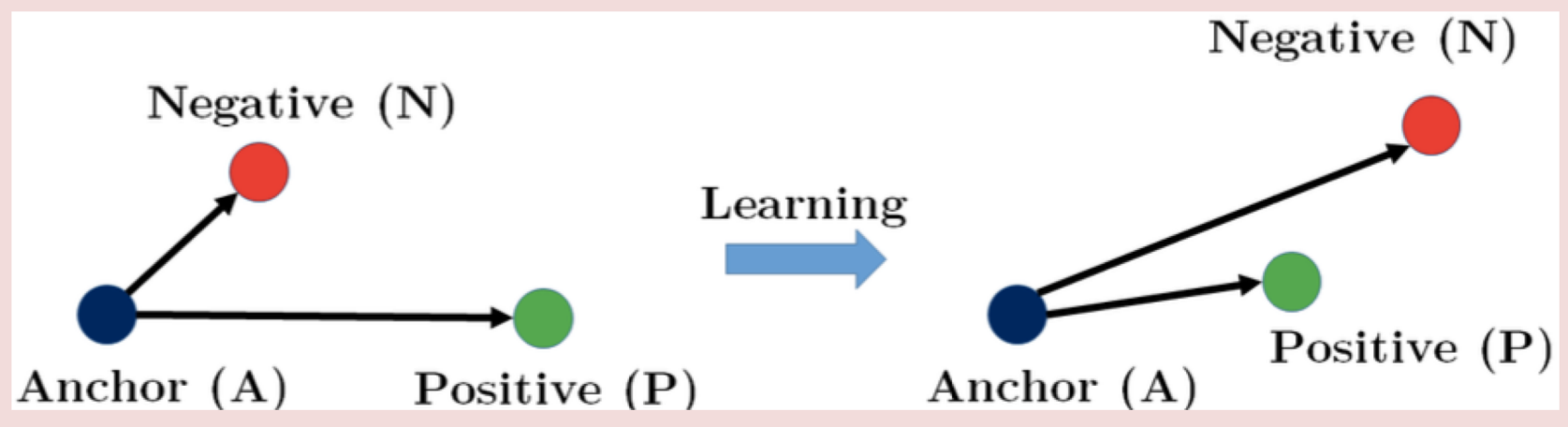


Contribution 1: New WF Attacks

- To better understand the threat of WF, we have explored two new attacks:



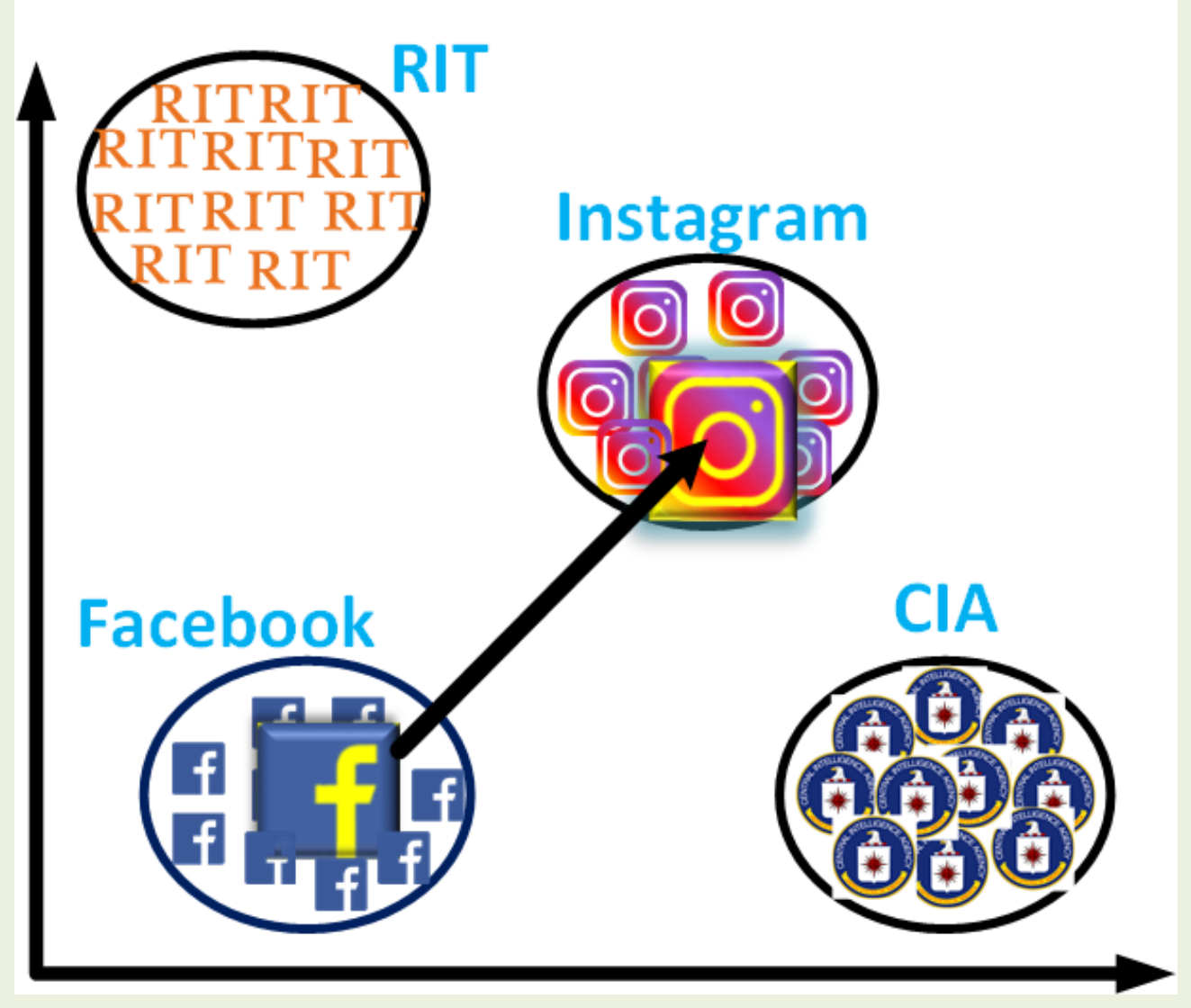
DF (CCS '18): Using a deep CNN, we achieve 98% closed world accuracy.



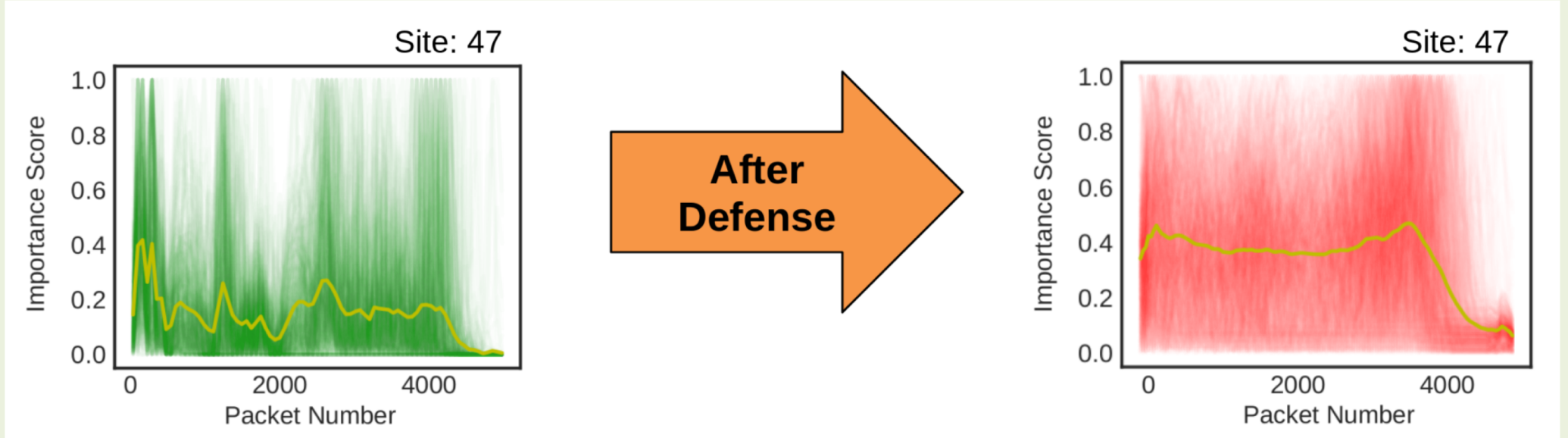
TF (CCS '19): Using "triplet networks," we get high accuracy with low data. This shows how an attacker could flexibly attack different users or sites of interest with modest resources.

Contribution 2: Defenses

- Against these deep-learning based attacks, we explore two techniques for improving WF defenses for more effectiveness with reasonable overhead:



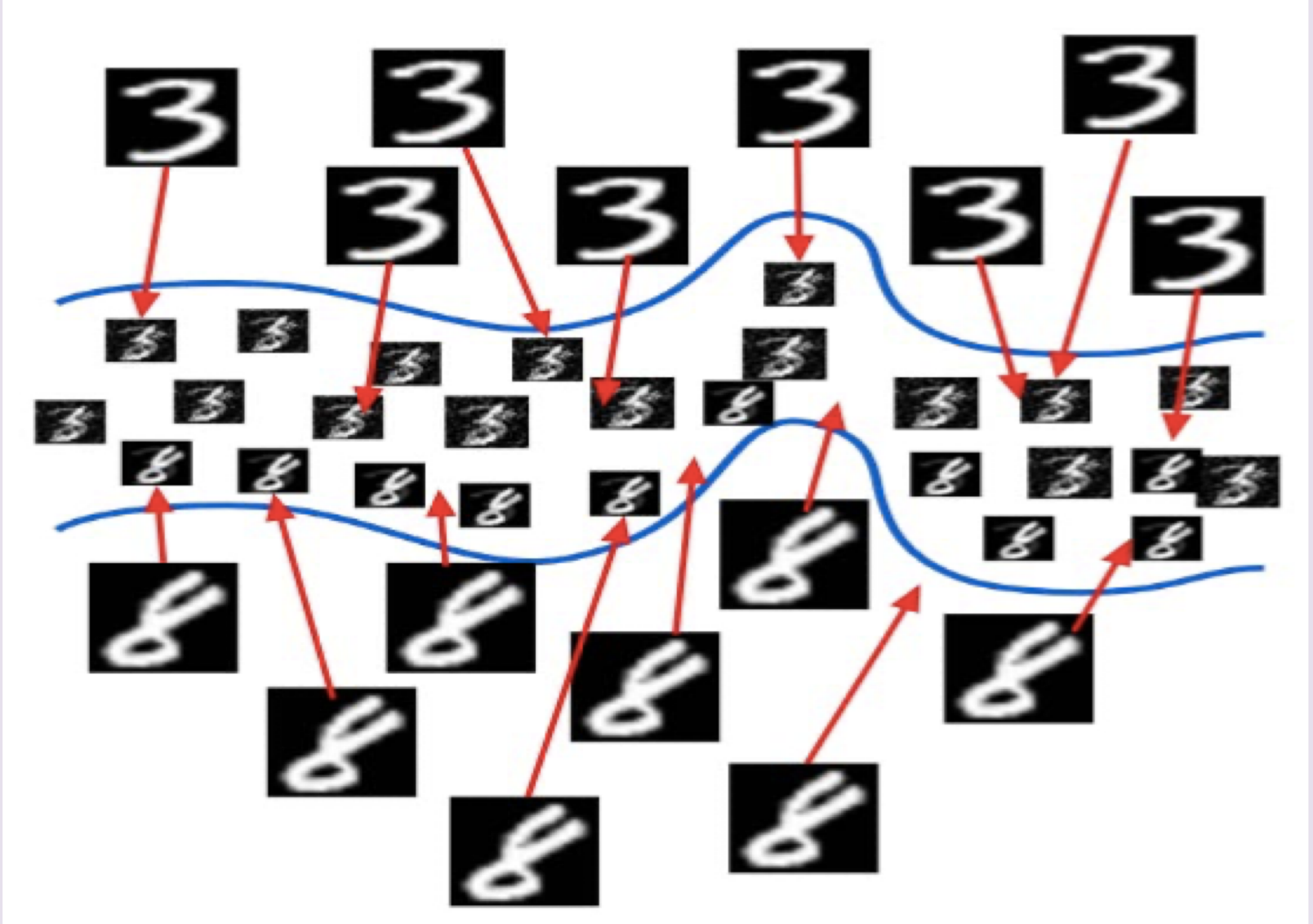
Mockingbird: Apply *adversarial examples* to evade classification.



RIS-PAD: Use deep learning visualization tools (*Grad-CAM*) to find important parts of the traffic trace for classification and pad more heavily there.

Contribution 3: Detecting Adversarial Examples

- Although it could help the WF attacker, we also explore defenses against adversarial examples, as they undermine security in other ML applications.



PadNet: Combines a padding class with *targeted gradient minimization*.

- PadNet starts with *adversarial training,* used to pad the boundary between classes
- *Targeted Gradient Minimization* (TGM) makes the neural network avoid moving towards the padding class.
- Overall effect is robust classification.

Scientific Impacts

- Applications of deep-learning techniques on network traffic traces
- Adversarial examples in traffic traces
- Exploring adversarial examples for defense
- Novel adversarial example technique that is robust to adversarial training
- Deep learning visualization on traffic traces

Broader Impact: *What do the websites you visit say about you?*

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS
SCIENTOLOGY
islam.com
BLACK LIVES MATTER
NRA
NARAL PRO-CHOICE AMERICA
THE CENTER THE LESBIAN, GAY, BISEXUAL & TRANSGENDER COMMUNITY CENTER
glaad
PRIDE AT WORK
betterhelp
CORE MENTAL HEALTH COUNSELING
Transitions COUNSELING SERVICES
Counseling & Life Coaching
WHITE SPRUCE COUNSELING

Tor protects this information, but only if it is secure against WF.

Broader Impact: Education

- New courses: *Anonymity & Tor* and *Deep Learning Security*
- Coming up: Podcast on Adversarial Machine Learning.

RIT

The 4th NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
October 28-29, 2019 I Alexandria, Virginia