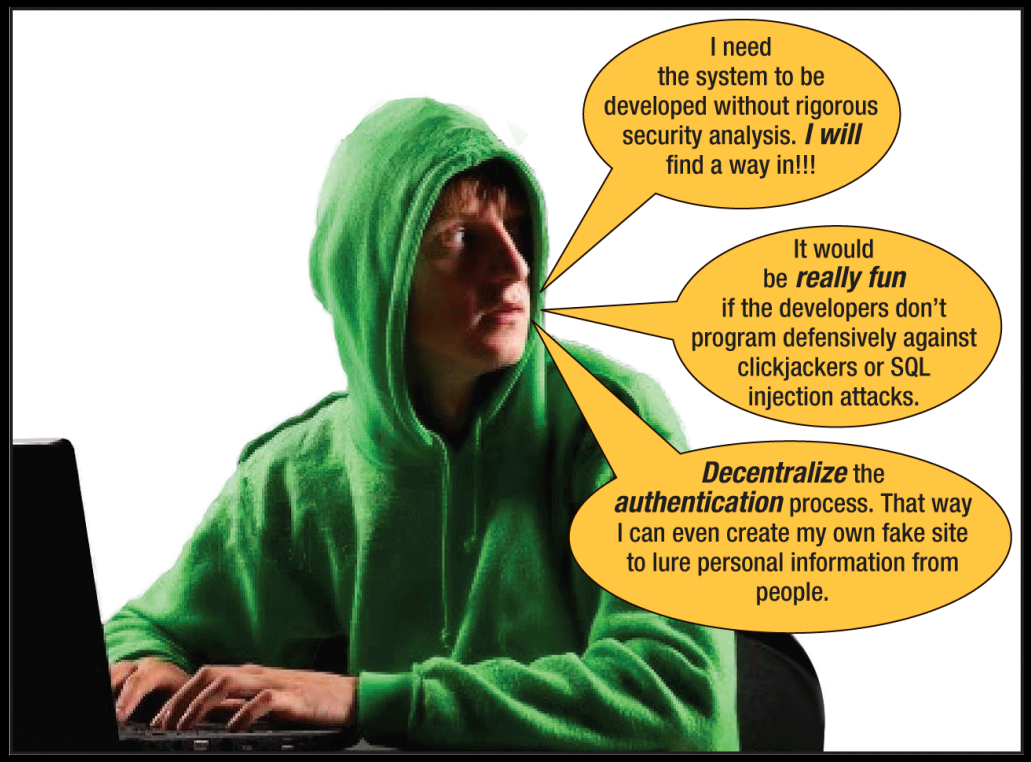
Security and Privacy

Jin Guo Nov 10th, 2020

Threat Modelling

- Used to inform defensive measures.
- Components
  - An abstraction of the system
  - Profiles of potential attackers (including their goals and methods)
  - A catalog of potential threats that may arise

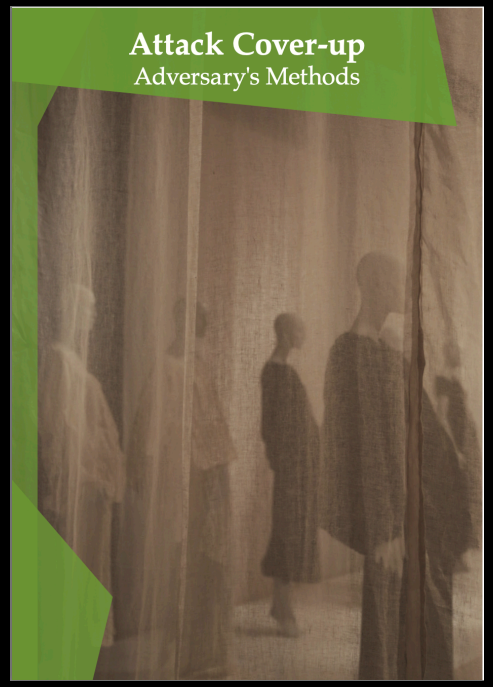
Persona Non Grata (PnG)

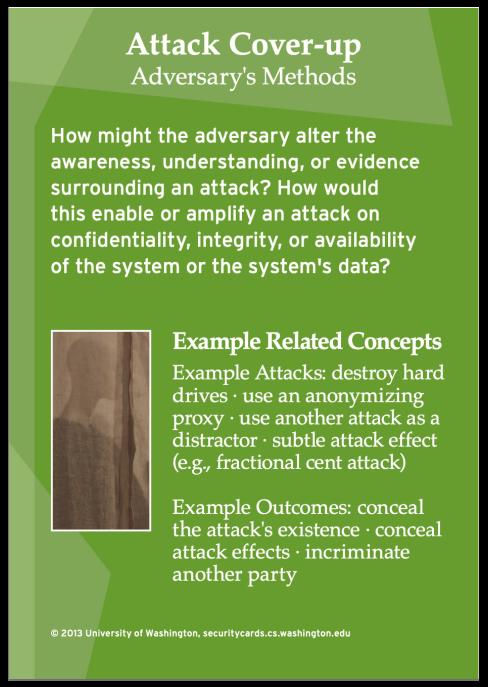


Cleland-Huang, J., 2014. How well do you know your personae non gratae?. *IEEE software*, *31*(4), pp.28-31.

Security Cards

HUMAN IMPACT
ADVERSARY'S MOTIVATIONS
ADVERSARY'S RESOURCES
ADVERSARY'S METHODS





STRIDE

Property	Threat	Definition
Authentication	**S**poofing	Impersonating something or someone else.
Integrity	**T**ampering	Modifying data or code
Non-repudiation	**R**epudiation	Claiming to have not performed an action.
Confidentiality	**I**nformation Disclosure	Exposing information to someone not authorized to see it
Availability	**D**enial of Service	Deny or degrade service to users
Authorization	**E**levation of Privilege	Gain capabilities without proper authorization

Privacy

Activity: Adversarial Goal and Method for ML

- Adversarial Goal:
  - Integrity
- Adversarial Goal:
  - Confidentiality and Privacy

1. Which part of the ML pipeline is vulnerable to the such adversaries?
2. What methods the attackers might use to achieve their goal?
3. How can we defend those attacks?

Adversarial Goal and Method for ML

- Adversarial Goal:
  - Integrity
- Adversarial Goal:
  - Confidentiality and Privacy

Input manipulation during training, inference (or online training)

Data pipeline manipulation

Improve the Robustness to Distribution Drifts

Membership Inference

Training data extraction

Learning and Inference with Privacy

Example: PATE

Private Aggregation of Teacher Ensemble