

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



Attack Cover-up
Adversary's Methods

How might the adversary alter the awareness, understanding, or evidence surrounding an attack? How would this enable or amplify an attack on confidentiality, integrity, or availability of the system or the system's data?

 **Example Related Concepts**
Example Attacks: destroy hard drives · use an anonymizing proxy · use another attack as a distractor · subtle attack effect (e.g., fractional cent attack)

Example Outcomes: conceal the attack's existence · conceal attack effects · incriminate another party

© 2013 University of Washington, securitycards.cs.washington.edu

STRIDE

Privacy

Property	Threat	Definition
Authentication	Spoofing	Impersonating something or someone else.
Integrity	Tampering	Modifying data or code
Non-repudiation	Repudiation	Claiming to have not performed an action.
Confidentiality	Information Disclosure	Exposing information to someone not authorized to see it
Availability	Denial of Service	Deny or degrade service to users
Authorization	Elevation of Privilege	Gain capabilities without proper authorization

Activity: Adversarial Goal and Method for ML

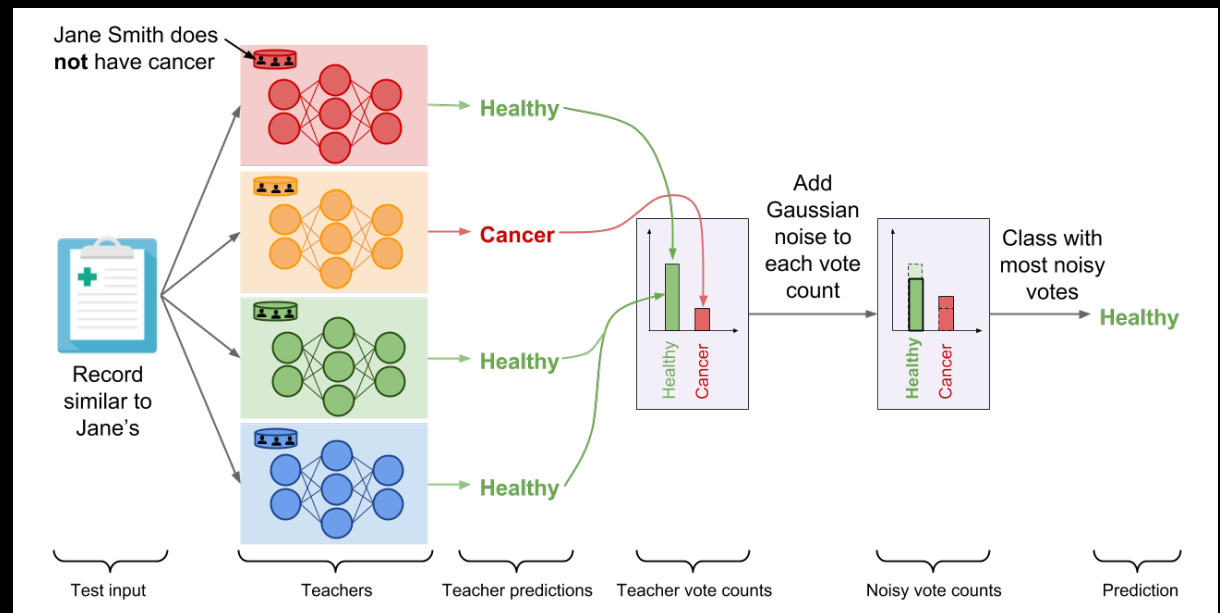
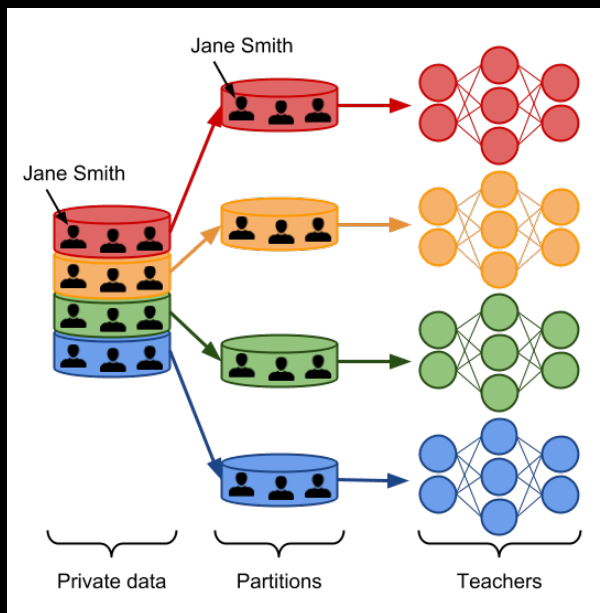
- Adversarial Goal:
 - Integrity
 - Adversarial Goal:
 - Confidentiality and Privacy
1. Which part of the ML pipeline is vulnerable to the such adversaries?
 2. What methods the attackers might use to achieve their goal?
 3. How can we defend those attacks?

Adversarial Goal and Method for ML

- Adversarial Goal:
 - Integrity
 - Input manipulation during training, inference (or online training)*
 - Data pipeline manipulation*
- Adversarial Goal:
 - Improve the Robustness to Distribution Drifts*
- Confidentiality and Privacy
 - Membership Inference*
 - Training data extraction*
 - Learning and Inference with Privacy*

Example: PATE

Private Aggregation of Teacher Ensemble



Privacy and machine learning: two unexpected allies?