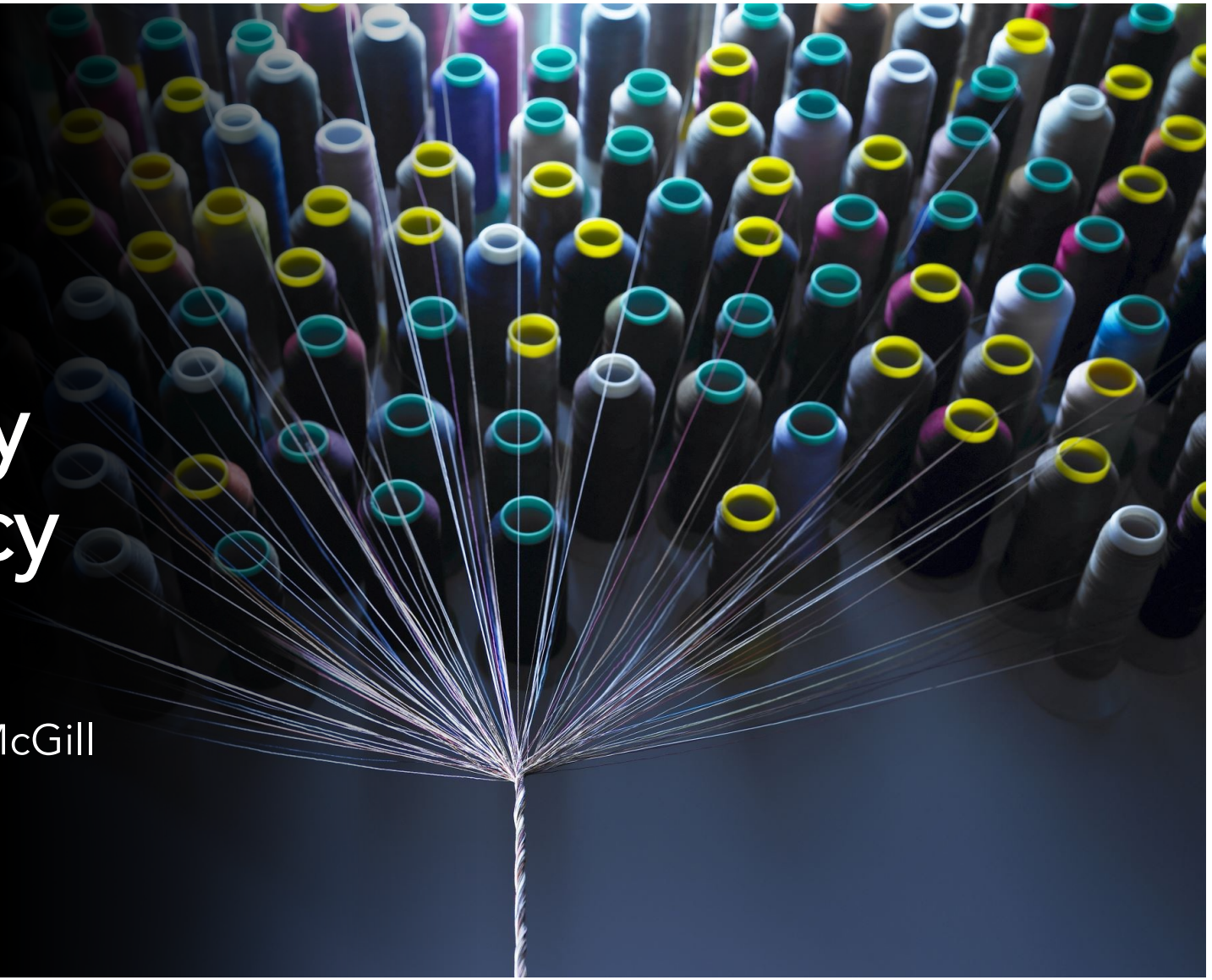
AI Security and Privacy



Jin L.C. Guo, SOCS McGill University

Agenda

- Security Basics
- Threat Modeling
- Common Threats for ML Systems

Security

Infrastructure Security

Application Security

Operational Security

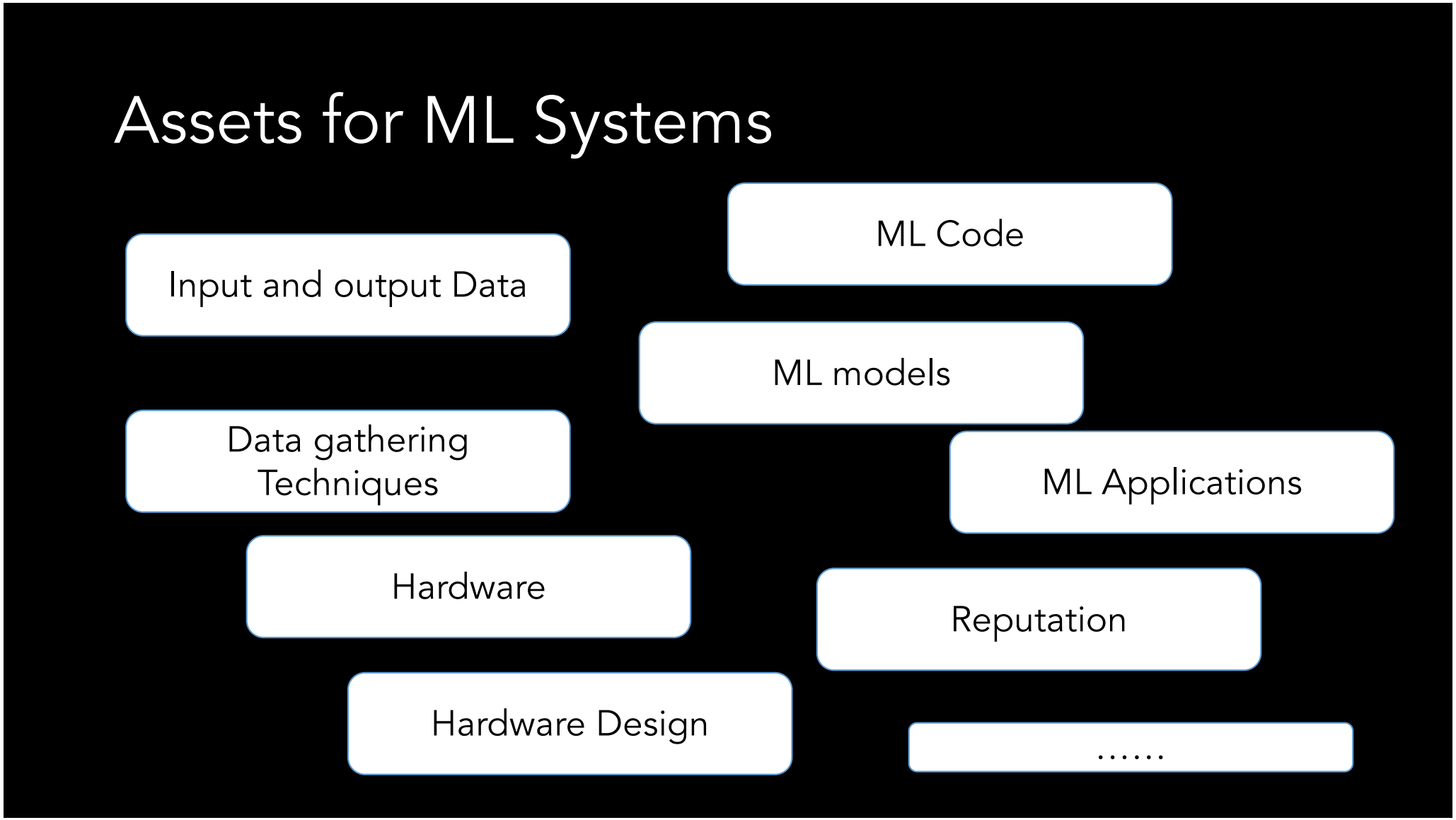
Primary Security Properties

- Confidentiality
- Integrity
- Availability

Security Terminologies

Term	Definition
Asset	Something of value that has to be protected. The asset may be the software system itself or the data used by that system.
Attack	An exploitation of a system's vulnerability where an attacker has the goal of causing some damage to a system asset or assets. Attacks may be from outside the system (external attacks) or from authorized insiders (insider attacks).
Control	A protective measure that reduces a system's vulnerability. Encryption is an example of a control that reduces a vulnerability of a weak access control system.
Exposure	Possible loss or harm to a computing system. This can be loss or damage to data or can be a loss of time and effort if recovery is necessary after a security breach.
Threat	Circumstances that have potential to cause loss or harm. You can think of a threat as a system vulnerability that is subjected to an attack.
Vulnerability	A weakness in a computer-based system that may be exploited to cause loss or harm.

Assets for ML Systems



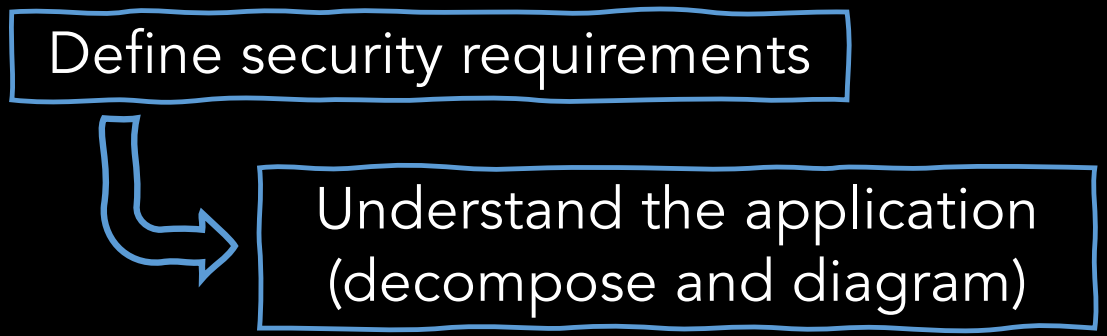
Agenda

- Security Basics
- Threat Modeling
- Common Threats for ML Systems

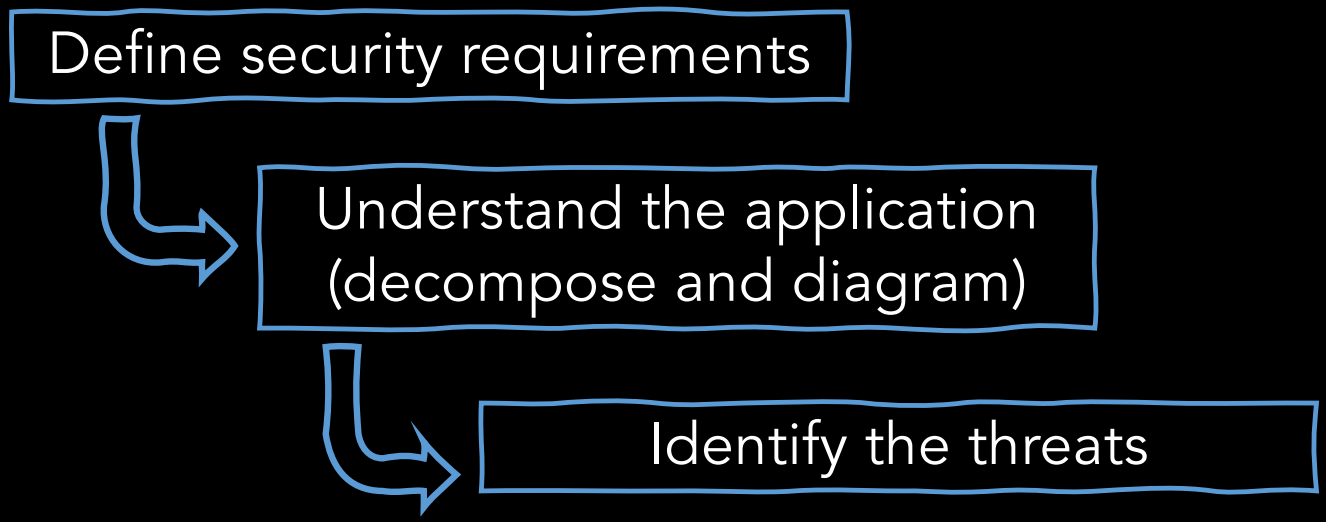
Threat Modeling

Define security requirements

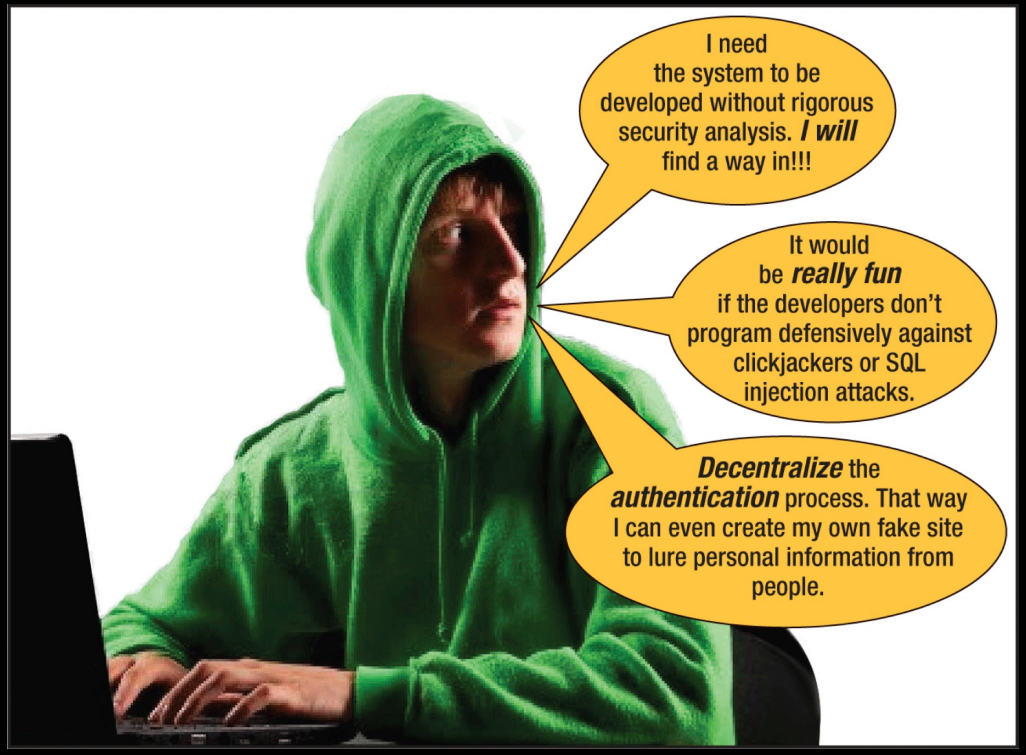
Threat Modeling



Threat Modeling



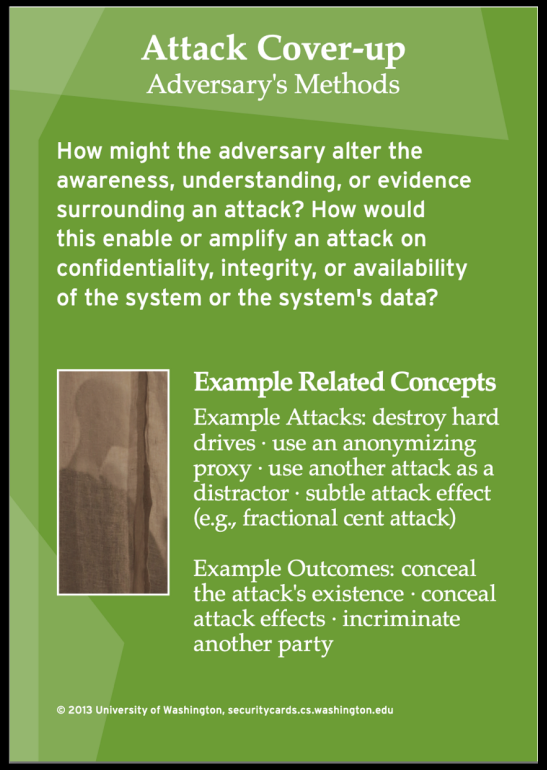
Persona Non Grata (PnG)



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

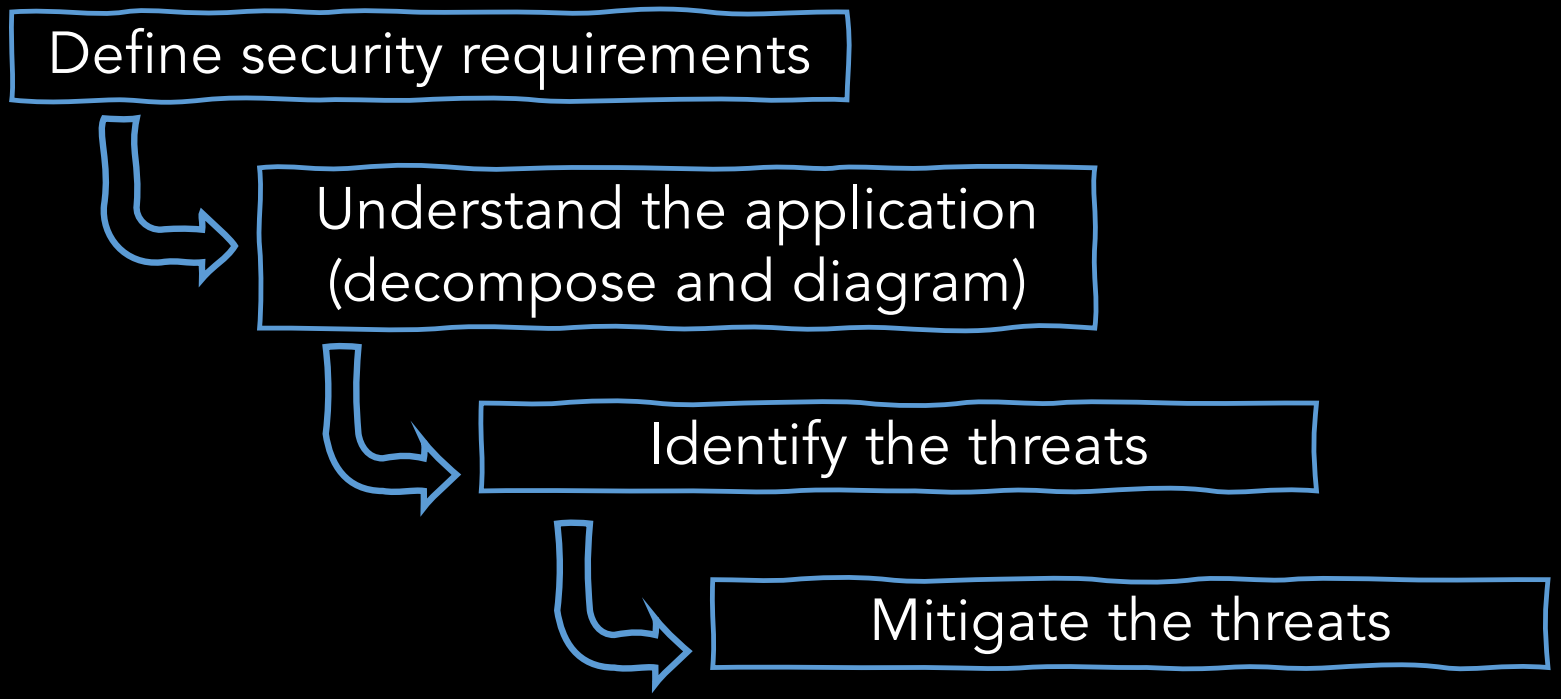
Security Cards

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

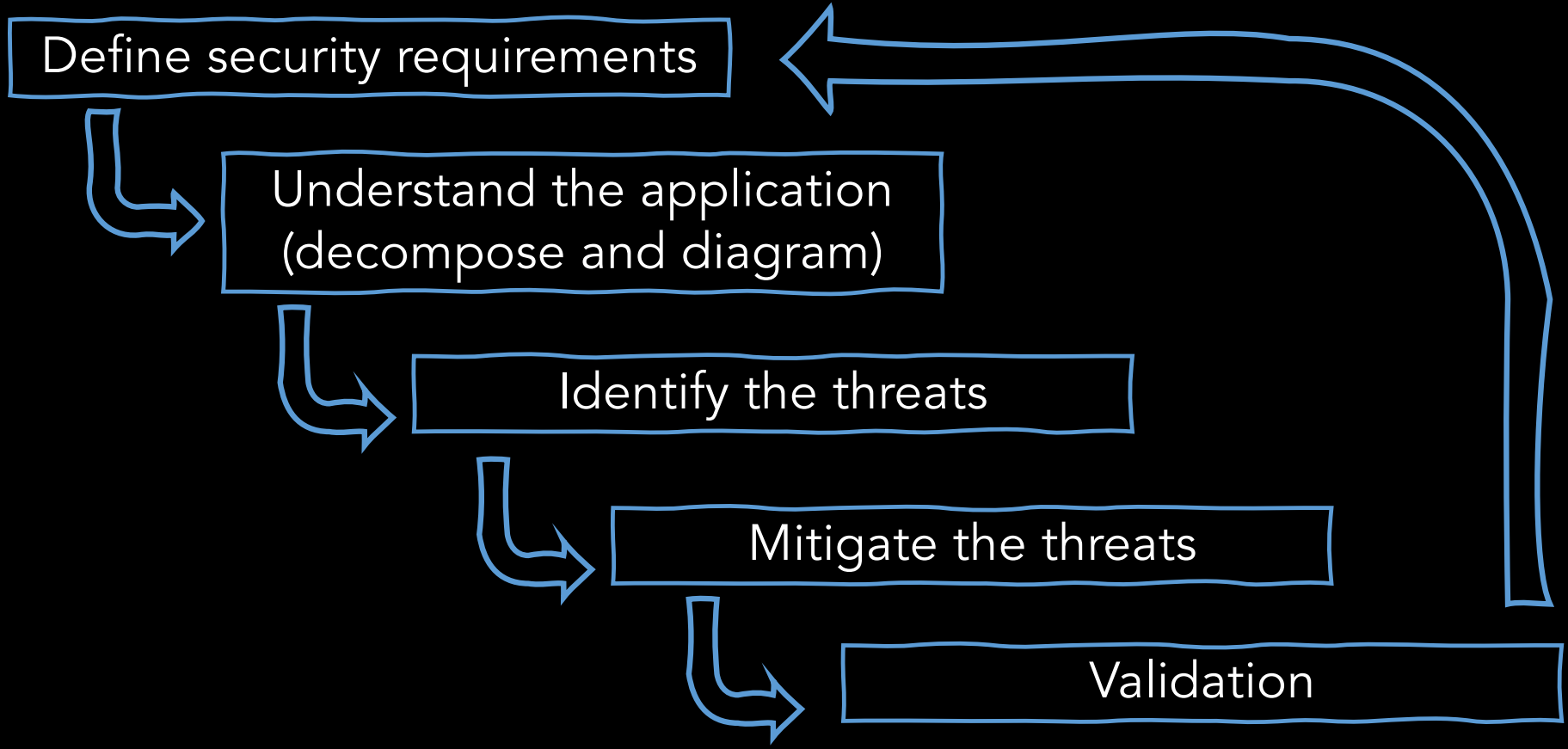


https://securitycards.cs.washington.edu

Threat Modeling



Threat Modeling



Agenda

- Security Basics
- Threat Modeling
- Common Threats for ML Systems

Activity

- Choose an application domain.
- Identify the assets that are most important for that application.
- Brainstorm the kind of attacks the ML systems might be subject to. What property (confidentiality, integrity, or availability) will be compromised? What knowledge is required to perform such attacks? What are the potential mitigation methods?

Perturbation attacks

- The attacker stealthily modifies the query to get a desired response

- Integrity

Image: Noise is added to an X-ray image, which makes the predictions go from normal scan to abnormal

Text translation: Specific characters are manipulated to result in incorrect translation. The attack can suppress specific word or can even remove the word completely

https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Data Poisoning attacks

- The attacker contaminated the machine model generated in the training phase, so that predictions on new data will be modified during the testing and deployment phase

- Integrity

In the Tay chatbot, future conversations were tainted because a fraction of the past conversations were used to train the system via feedback.

https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Model Inversion

- The attacker recovers the private features used in machine learning models.

- Confidentiality

> The researchers demonstrate that an adversary who knows only a little bit about an individual subscriber of Netflix can easily identify this subscriber's record in the dataset, uncovering their apparent political preferences and other potentially sensitive information.

Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111-125. IEEE, 2008.

https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Membership Inference attack

- The attacker can determine whether a given data record was part of the model's training dataset or not

- Confidentiality

Researchers were able to predict a patient's main procedure (e.g., Surgery the patient went through) based on the attributes (e.g., age, gender, hospital)

https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Model Stealing

- The attackers recreate the underlying model by legitimately querying the model. The functionality of the new model is same as that of the underlying model.

- Confidentiality

Researchers successfully emulated the underlying algorithm from Amazon, BigML. For instance, in the BigML case, researchers were able to recover the model used to predict if someone should have a good/bad credit risk (German Credit Card dataset) using 1,150 queries and within 10 minutes.

https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Review

- Security Basics
- Threat Modeling
- Common Threats for ML Systems

- Next

AI Security and Privacy (cont'd)