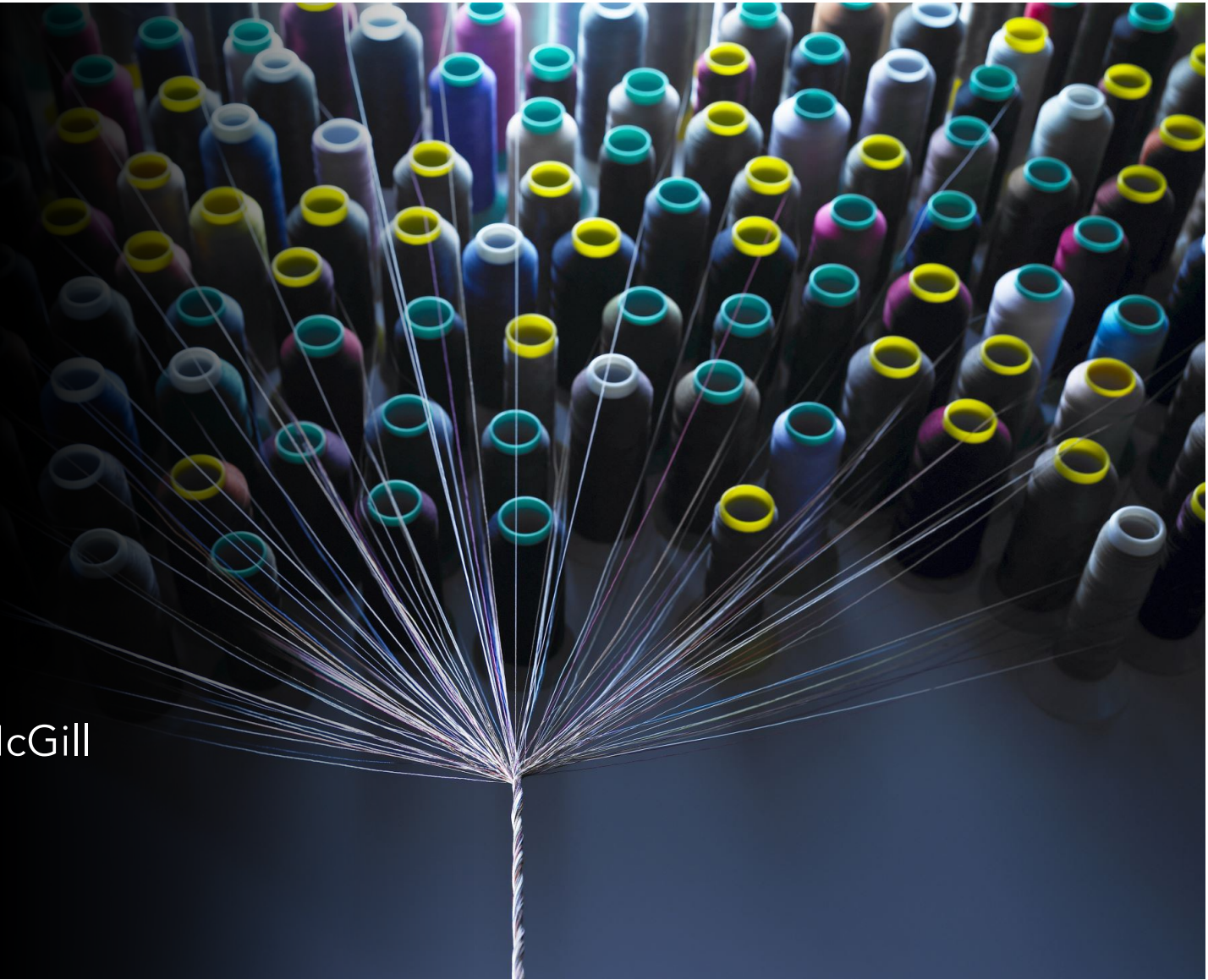


Security

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

Whether it guarantees access to the data entrusted to it exclusively to those who have the appropriate rights to have access to that data.

- Integrity

Whether the authenticity of data or operations can be verified and the data or functionality has not been modified or made unauthentic through unauthorized activity

- Availability

Whether authorized actors are able to access system functionality and/or data whenever they have the need or desire to do so.

Privacy: *the right of not having that information exposed to unauthorized third parties.*

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.

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.

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.

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

Input and output Data

ML Code

ML models

Data gathering
Techniques

ML Applications

Hardware

Reputation

Hardware Design

.....

Agenda

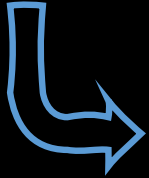
- Security Basics
- Threat Modeling
- Common Threats for ML Systems

Threat Modeling

Define security requirements

Threat Modeling

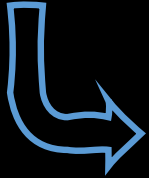
Define security requirements



Understand the application
(decompose and diagram)

Threat Modeling

Define security requirements



Understand the application
(decompose and diagram)



Identify the threats

STRIDE

	Threat	Property Violated	Threat Definition
S	Spoofing identify	Authentication	Pretending to be something or someone other than yourself
T	Tampering with data	Integrity	Modifying something on disk, network, memory, or elsewhere
R	Repudiation	Non-repudiation	Claiming that you didn't do something or were not responsible; can be honest or false
I	Information disclosure	Confidentiality	Providing information to someone not authorized to access it
D	Denial of service	Availability	Exhausting resources needed to provide service
E	Elevation of privilege	Authorization	Allowing someone to do something they are not authorized to do

<https://insights.sei.cmu.edu/blog/threat-modeling-12-available-methods/>

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)

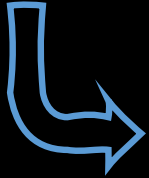
Example Outcomes: conceal the attack's existence · conceal attack effects · incriminate another party

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

<https://securitycards.cs.washington.edu>

Threat Modeling

Define security requirements



Understand the application
(decompose and diagram)

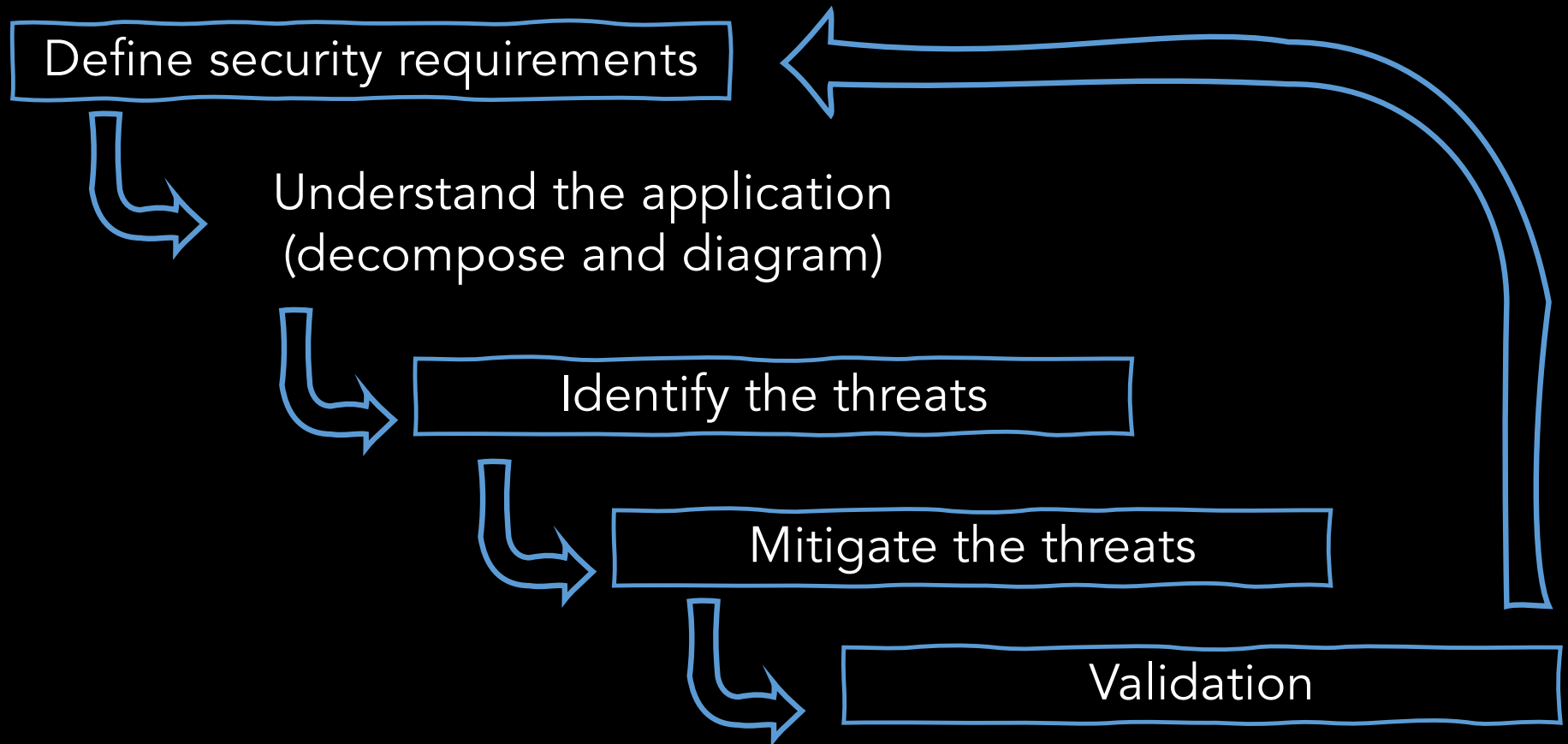


Identify the threats



Mitigate the threats

Threat Modeling



Agenda

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

Activity

- Consider the ML application of a hiring system.
- 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

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.

Model Inversion

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



Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Fredrikson, M., Jha, S. and Ristenpart, T., 2015, October. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (pp. 1322-1333).

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.

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)

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.

Review

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

- Next

Privacy and Wrap Up