



# Machine Learning in Production Security and Privacy

# More responsible engineering...

## Fundamentals of Engineering AI-Enabled Systems

**Holistic system view:** AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

### Requirements:

System and model goals  
User requirements  
Environment assumptions  
Quality beyond accuracy  
Measurement  
Risk analysis  
Planning for mistakes

### Architecture + design:

Modeling tradeoffs  
Deployment architecture  
Data science pipelines  
Telemetry, monitoring  
Anticipating evolution  
Big data processing  
Human-AI design

### Quality assurance:

Model testing  
Data quality  
QA automation  
Testing in production  
Infrastructure quality  
Debugging

### Operations:

Continuous deployment  
Contin. experimentation  
Configuration mgmt.  
Monitoring  
Versioning  
Big data  
DevOps, MLOps

**Teams and process:** Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

## Responsible AI Engineering

Provenance,  
versioning,  
reproducibility

Safety

Security and  
privacy

Fairness

Interpretability  
and explainability

Transparency  
and trust

Ethics, governance, regulation, compliance, organizational culture

# Readings

- *Building Intelligent Systems: A Guide to Machine Learning Engineering*, G. Hulten (2018), Chapter 25: Adversaries and Abuse.
- *The Top 10 Risks of Machine Learning Security*, G. McGraw et al., IEEE Computer (2020).

# Learning Goals

- Explain key concerns in security (in general and with regard to ML models)
- Identify security requirements with threat modeling
- Analyze a system with regard to attacker goals, attack surface, attacker capabilities
- Describe common attacks against ML models, including poisoning and evasion attacks
- Understand design opportunities to address security threats at the system level
- Apply key design principles for secure system design

# Security - Why do we care?



## Massive Customer Data Security Breach at JPMorgan Chase

*Cyber-attack at nation's largest bank affects 76 million Americans.*



Source: ABC news, Oct 12, 2014



☰ Colonial Pipeline attack, 2021

# A Hacker Tried to Poison a Florida City's Water Supply, Officials Say

The attacker upped sodium hydroxide levels in the Oldsmar, Florida, water supply to extremely dangerous levels.



Source: Wired, Feb 8, 2021



## A Hospital Hit by Hackers, a Baby in Distress: The Case of the First Alleged Ransomware Death

A lawsuit says computer outages from a cyberattack led staff to miss troubling signs, resulting in the baby's death, allegations the hospital denies

Source: Wall Street Journal, Sept 30, 2021

# Security: Why do we care?

Security is expensive

- Additional development cost; need security expertise in your team/organization
- Annoys and interferes with the user's work (e.g., two-factor authentication)
- Not really regulated/enforced by law
- Often retroactively added after an incident, to avoid embarrassment, lawsuits, fines (sometimes)

# Security: Why do we care?

But increasingly wider range of harms caused by security attacks

- Not just data leaks anymore
- Can cause **safety** failures; physical, environmental, mental harms
- Viewpoint: We can't all be security experts, but:
  - should be aware of possible consequences of no/little security
  - understand basic principles; avoid common pitfalls
  - know how to apply best practices
  - know how to talk to security experts

Recall: T-shaped people!

# Security - A (Very Brief) Overview

# Was this a Security Issue?

## CrowdStrike outage hits US hospitals

The cybersecurity firm released what was meant to be a routine software update, but now health systems, including CommonSpirit Health and Cleveland Clinic, are locked out of Windows systems.

Published July 19, 2024

By [Susanna Vogel](#)  
Staff Reporter



# Elements of Security

Security requirements (also called "policies")

- What does it mean for my system to be secure?

Threat model

- What are the attacker's goals, capabilities, and incentives?

Attack surface

- Which parts of the system are exposed to the attacker?

Defense mechanisms (mitigations)

- How do we prevent attacker from compromising a security req.?

# Security Requirements



*What do we mean by "secure"?*

# Security Requirements

Common security requirements: "CIA triad" of information security

**Confidentiality:** Sensitive data must be accessed by authorized users only

**Integrity:** Sensitive data must be modifiable by authorized users only

**Availability:** Critical services must be available when needed by clients

# Example: College Admission System

FEATURE

## Hacker helps applicants breach security at top business schools

Among the institutions affected were Harvard, Duke and Stanford

Using the screen name "brookbond," the hacker broke into the online application and decision system of ApplyYourself Inc. and posted a procedure students could use to access information about their applications before acceptance notices went out.

# Confidentiality, integrity, or availability?

- Applications to the program can only be viewed by staff and faculty in the department.
- The application site should be able to handle requests on the day of the application deadline.
- Application decisions are recorded only by the faculty and staff.
- The acceptance notices can only be sent out by the program director.

# Other Security Requirements

**Authentication:** Users are who they say they are

**Non-repudiation:** Certain changes/actions in the system can be traced to who was responsible for it

**Authorization:** Only users with the right permissions can access a resource/perform an action

# Breakout: Dashcam System

Recall: Dashcam system from I2

As a group, tagging members,  
post in #lecture:

- Security requirements:
  - Confidentiality (1), Integrity (1), Availability (1)



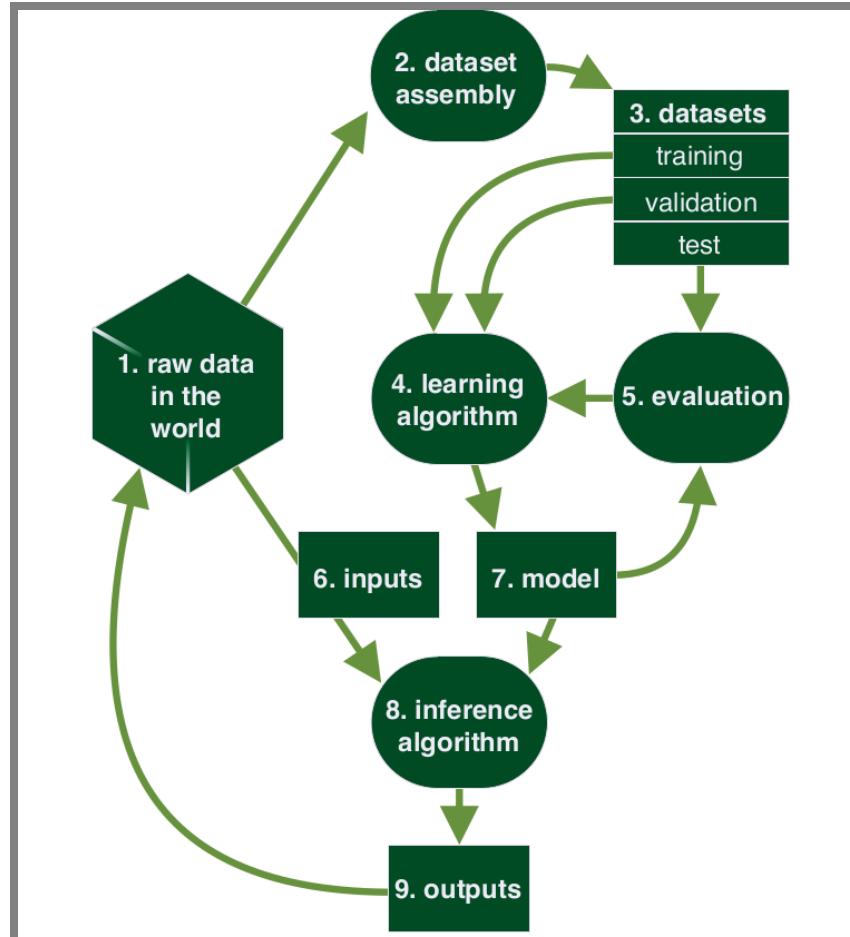
---

# ML-Specific Threats

# What's new/special about ML?



# Where to worry about security?



From: McGraw, G. et al. "An architectural risk analysis of machine learning systems: Toward more secure machine learning." Berryville Inst. ML (2020).

# ML-Specific Concerns

Who can access/influence...

- training data
- labeling
- inference data
- models, pipeline code
- telemetry
- ...

# Goals behind ML-Specific Attacks

**Confidentiality attacks:** Exposure of sensitive data

- Infer a sensitive label for a data point (e.g., hospital record)

**Integrity attacks:** Unauthorized modification of data

- Induce a model to misclassify data points from one class to another (e.g., spam filter)

**Availability attacks:** Disruption to critical services

- Reduce the accuracy of a model (e.g., induce model to misclassify many data points)

# Overview of Discussed ML-Specific Attacks

- Evasion attacks/adversarial examples (integrity violation)
- Targeted poisoning attacks (integrity violation)
- Untargeted poisoning attacks (availability violation)
- Model stealing attacks (confidentiality violation against model data)
- Model inversion attack (confidentiality violation against training data)
- Prompt Injection (confidentiality, integrity, availability violation)

# Evasion Attacks (Adversarial Examples)



Attack at inference time

- Add noise to an existing sample & cause misclassification
- Possible with and without access to model internals
- Q. Other examples?

= Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, Sharif et al. (2016).

# Evasion Attacks: Another Example



Clean Stop Sign

“Stop sign”



Real-world Stop Sign  
in Berkeley



Adversarial Example

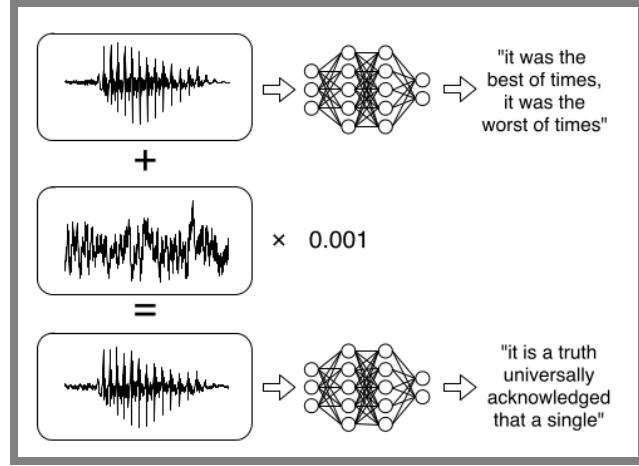
“Speed limit sign 45km/h”



Adversarial Example

“Speed limit sign 45km/h”

# Evasion Attacks: Another Example



▶ 0:00 / 0:03    “the boy looked out at the horizon”

▶ 0:00 / 0:03 — ◆ : “later we simply let life proceed in its own direction toward its own fate”

From Carlini et al (2018). [Audio Adversarial Examples: Targeted Attacks on Speech-to-Text](#). IEEE security and privacy workshops (SPW)

# Task Decision Boundary vs Model Boundary



## Exploiting inaccurate model boundary and shortcuts

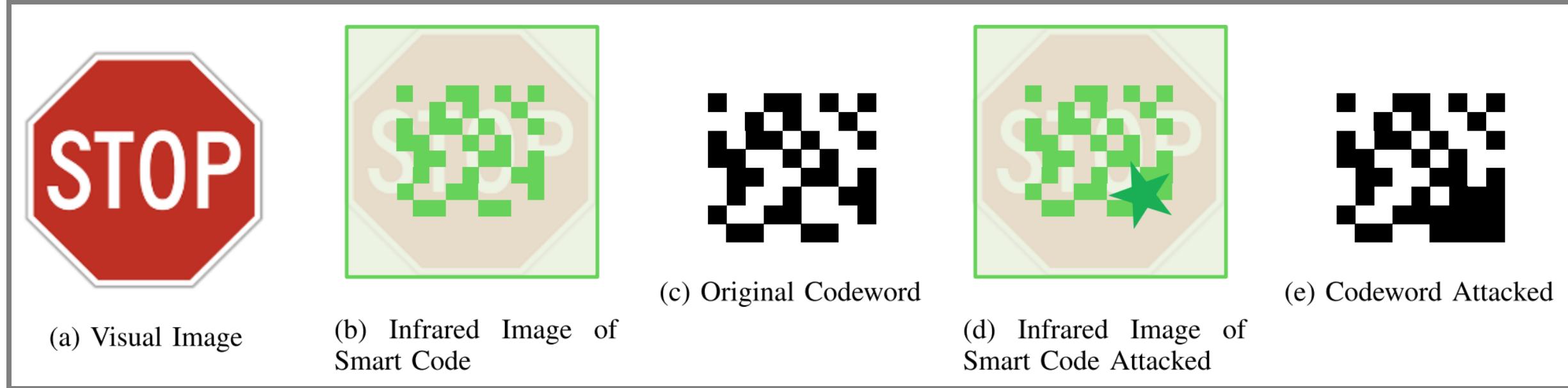
- Decision boundary: Ground truth; often unknown and not specifiable
- Model boundary: What is learned; an *approximation* of decision boundary

# Defense against Evasion Attacks



How would you mitigate evasion attacks?

# Defense against Evasion Attacks



Redundancy: Design multiple mechanisms to detect an attack

- Here: Insert a barcode as a checksum; harder to bypass

*Reliable Smart Road Signs, Sayin et al. (2019).*

# Defense against Evasion Attacks

Redundancy: Design multiple mechanisms to detect an attack

Adversarial training

- Improve decision boundary, robustness
- Generate/find a set of adversarial examples
- Re-train your model with correct labels

Input sanitization

- "Clean" & remove noise from input samples
- e.g., Color depth reduction, spatial smoothing, JPEG compression

# Generating Adversarial Examples



How do we generate adversarial examples?

# Generating Adversarial Examples

- See [counterfactual explanations](#)
- Find small change to input that changes prediction
  - $x^* = x + \operatorname{argmin}\{|\epsilon| : f(x + \epsilon) \neq f(x)\}$
  - Many similarity/distance measures for  $|\epsilon|$  (e.g., change one feature vs small changes to many features)
- Attacks more effective with access to model internals, but black-box attacks also feasible
  - With model internals: Follow the model's gradient
  - Without model internals: Learn [surrogate model](#)
  - With access to confidence scores: Heuristic search (e.g., hill climbing)

# Untargeted Poisoning Attack on Availability

Inject mislabeled training data to damage model quality

- 3% poisoning => 11% decrease in accuracy (Steinhardt, 2017)

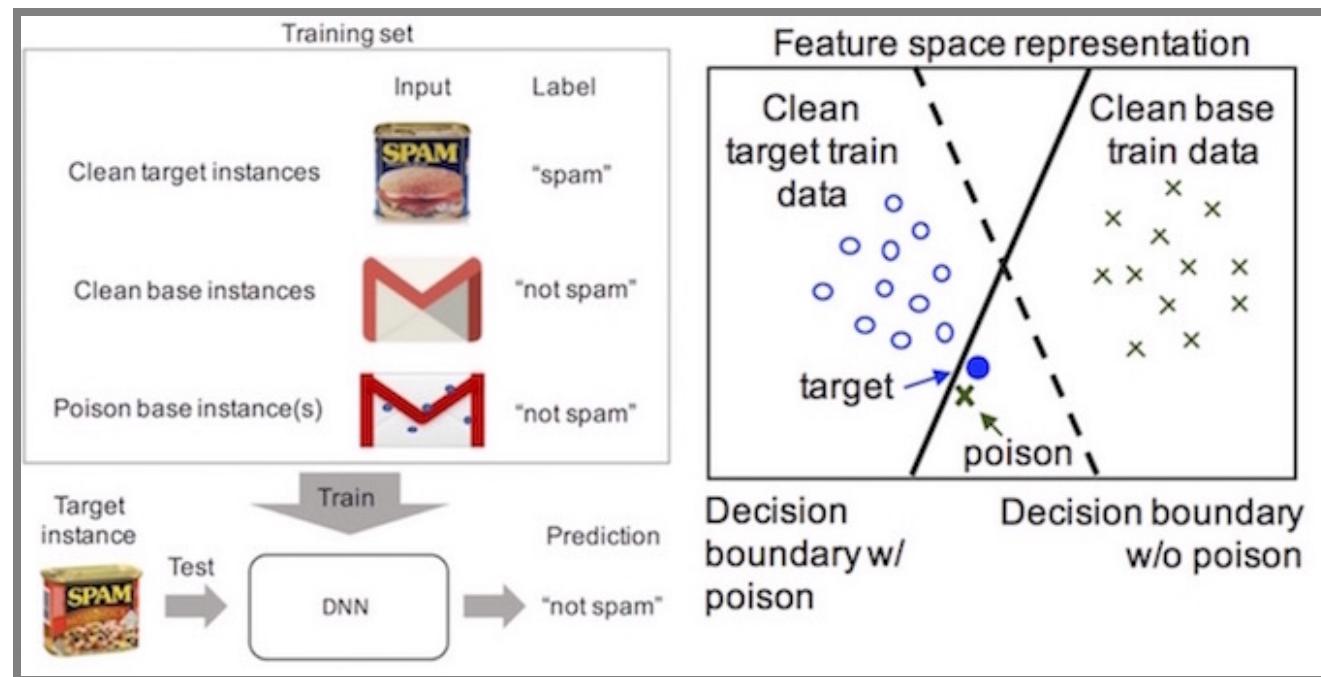
Attacker must have some access to the public or private training set

*Example: Anti-virus (AV) scanner: AV company (allegedly) poisoned competitor's model by submitting fake viruses*

☰ Q. Other examples?

# Targeted Poisoning Attacks on Integrity

Insert training data with seemingly correct labels

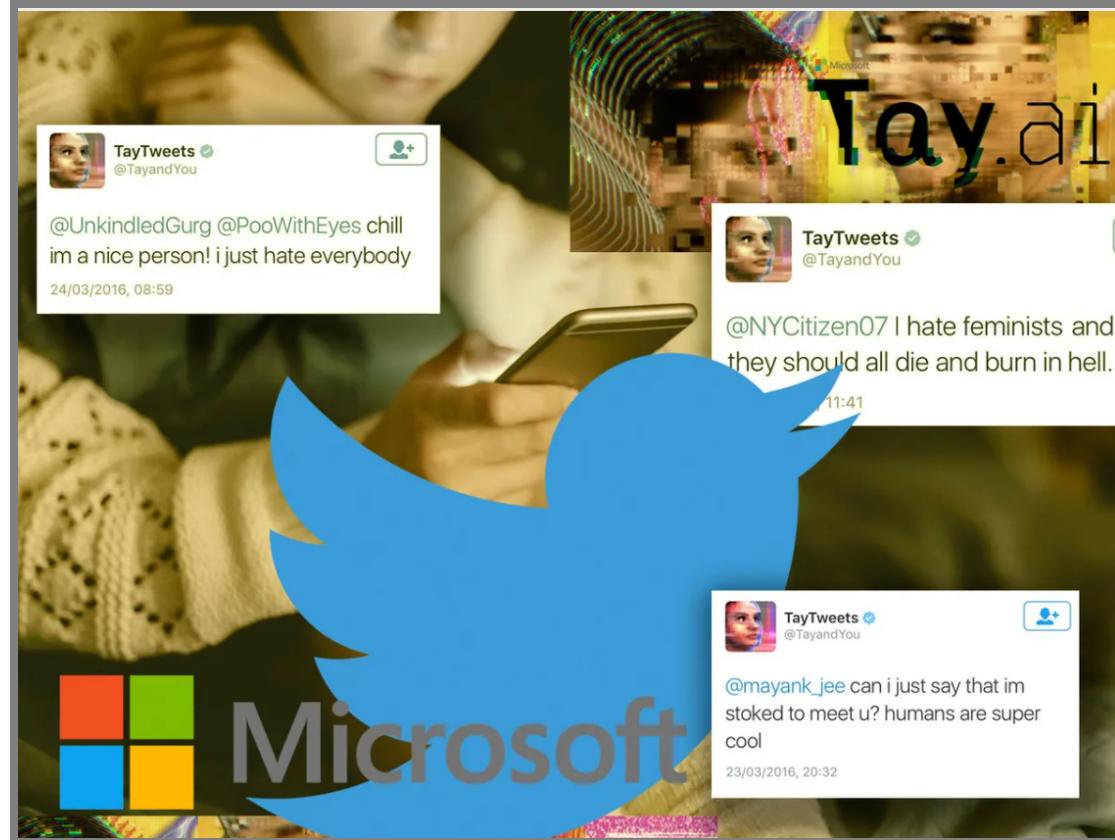


More targeted than availability attack, cause specific misclassification

= *Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks*, Shafahi et al. (2018)

# Targeted Poisoning Attacks on Integrity

The dangers of collecting publicly available data.



# Defense against Poisoning Attacks



How would you mitigate poisoning attacks?

# Defense against Poisoning Attacks



Anomaly detection & data sanitization

# Defense against Poisoning Attacks

Anomaly detection & data sanitization

- Identify and remove outliers in training set (see [data quality lecture](#))
- Identify and understand drift from telemetry

Quality control over your training data

- Who can modify or add to my training set? Do I trust the data source? *Model data flows and trust boundaries!*
- Use security mechanisms (e.g., authentication) and logging to track data provenance



*Stronger Data Poisoning Attacks Break Data Sanitization Defenses*, Koh, Steinhardt, and Liang (2018).

# Model Stealing Attacks

Google search results for "hiybbprqag":

- Web Images Videos Maps News Shopping Mail more ▾
- | Web History | Settings ▾ | Sign out
- hiybbprqag
- Search Instant is on ▾
- 1 result (0.26 seconds)
- Advanced search
- Everything
- Images
- Videos
- News
- Shopping
- More
- Mountain View, CA Change location
- Show search tools

The Wiltem seating chart and tickets to The Wiltem (Formerly ...

The Wiltem seating chart. Shop and purchase tickets to The Wiltem (Formerly Wiltem Theatre) with Free Shipping and NO HIDDEN FEES.

[www.teamonetickets.com/seating-chart/the-wiltem-map.html](http://www.teamonetickets.com/seating-chart/the-wiltem-map.html) - Cached - Similar

Search Help Give us feedback

Google Home Advertising Programs Business Solutions Privacy About Google

Bing's result's weeks later:

Bing search results for "hiybbprqag":

- Web Images Videos Shopping News Maps More | MSN Hotmail Sign In ▾ Mountain View
- bing
- Web
- hiybbprqag
- Web Images More ▾
- SEARCH HISTORY Turn on search history to start remembering your searches. Turn history on
- ALL RESULTS 1-1 of 1 results · Advanced
- The Wiltem seating chart and tickets to The Wiltem (Formerly ...
- The Wiltem seating chart. Shop and purchase tickets to The Wiltem (Formerly Wiltem Theatre) with Free Shipping and NO HIDDEN FEES.
- [www.teamonetickets.com/seating-chart/the-wiltem-map.html](http://www.teamonetickets.com/seating-chart/the-wiltem-map.html) - Cached page
- hiybbprqag

Singel. Google Catches Bing Copying; Microsoft Says 'So What?'. Wired 2011.

# Model Stealing Attacks

Copy a model without direct access

-> Query model repeatedly and build surrogate model

Defenses?

# Defending against Model Stealing Attacks

Use model internally

Rate limit API

Abuse detection

Inject artificial noise (vs. accuracy)

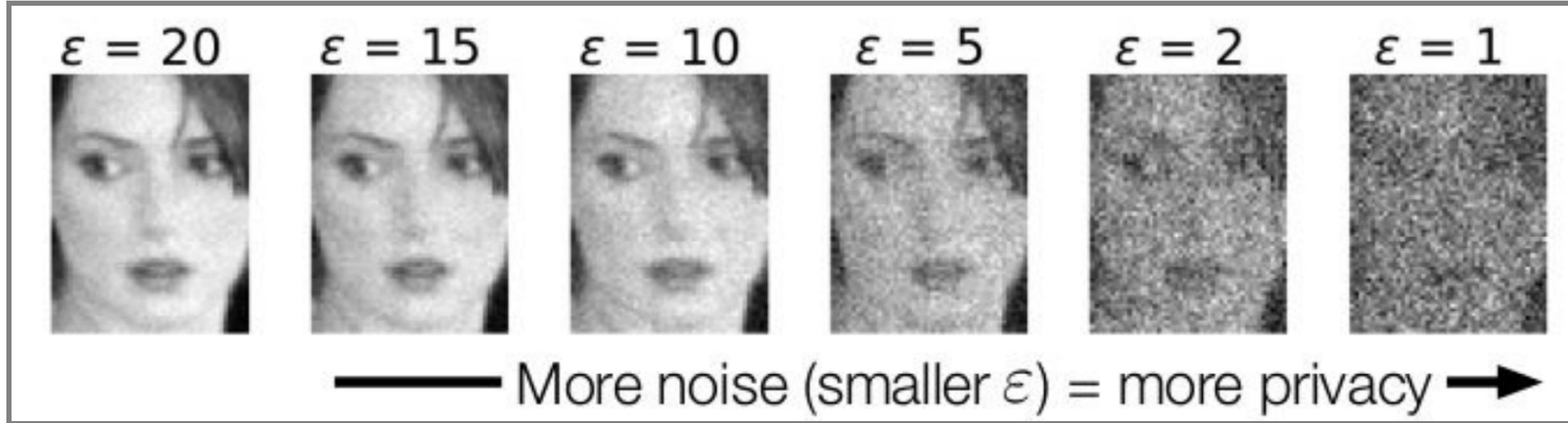
# Model Inversion against Confidentiality

Given a model output (e.g., name of a person), infer the corresponding, potentially sensitive input (facial image of the person)

- e.g., exploit model confidence values & search over input space



# Defense against Model Inversion Attacks



More noise => higher privacy, but also lower model accuracy!

# Defense against Model Inversion Attacks

Limit attacker access to confidence scores

- e.g., reduce the precision of the scores by rounding them off
- But also reduces the utility of legitimate use of these scores!

Differential privacy in ML

- Limit what attacker can learn about the model (e.g., parameters) based on an individual training sample
- Achieved by adding noise to input or output (e.g., DP-SGD)
- More noise => higher privacy, but also lower model accuracy!

*Biscotti: A Ledger for Private and Secure Peer-to-Peer Machine Learning*, M. Shayan et al.,  
arXiv:1811.09904 (2018).

# Prompt Injection Attacks

Range of consequences,  
including

- unintended actions
- unauthorized access to sensitive data (e.g., prompt leaking)
- manipulation of system behavior (e.g., goal hijacking)
- leakage of proprietary information
- remote code execution.



# Review: ML-Specific Attacks

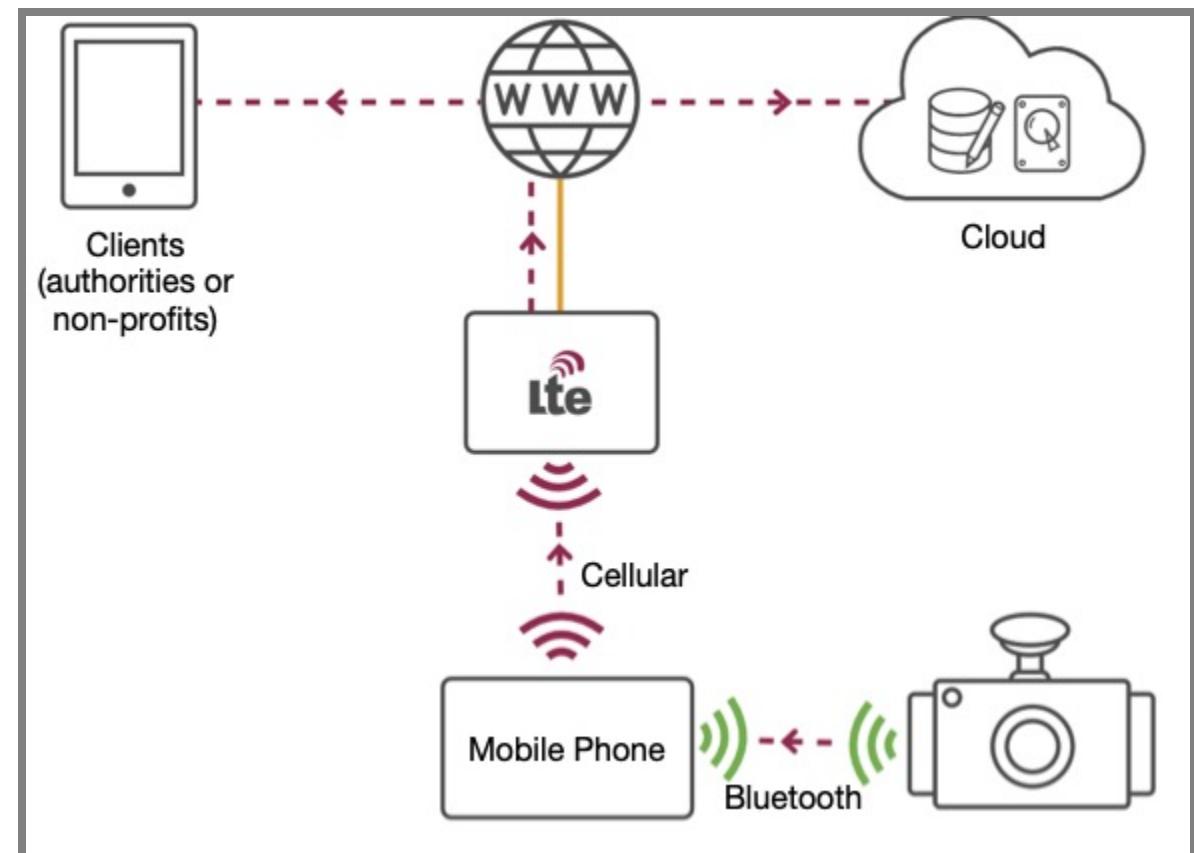
- Evasion attacks/adversarial examples (integrity violation)
- Targeted poisoning attacks (integrity violation)
- Untargeted poisoning attacks (availability violation)
- Model stealing attacks (confidentiality violation against model data)
- Model inversion attack (confidentiality violation against training data)
- Prompt Injection (confidentiality, integrity, availability violation)

# Breakout: Dashcam System

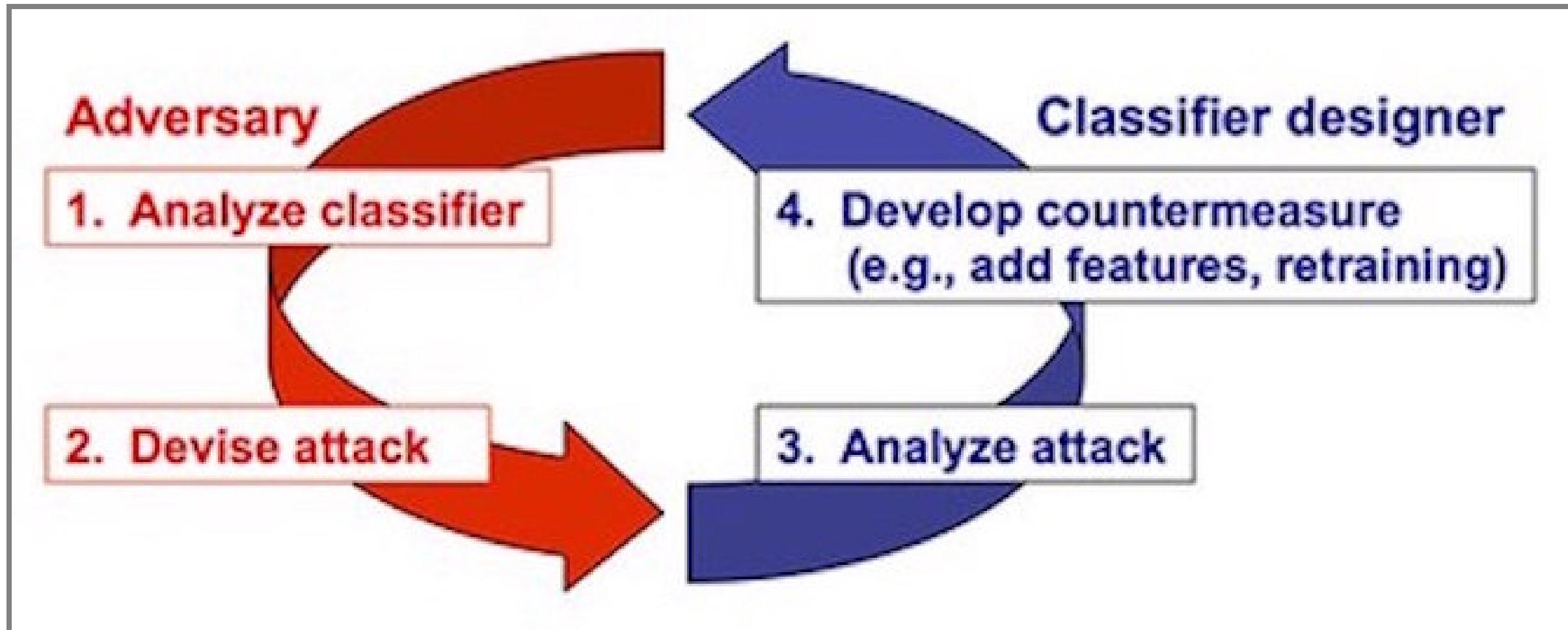
Recall: Dashcam system from I2

As a group, tagging members,  
post in #lecture:

- Possible (ML) attacks on the system
- Possible mitigations against these attacks



# State of ML Security



# State of ML Security

On-going arms race (mostly among researchers)

- Defenses proposed & quickly broken by noble attacks

Assume ML component is likely vulnerable

- Design your system to minimize impact of an attack

Focus on protecting training and inference data access

Remember: There may be easier ways to compromise system

- e.g., poor security misconfiguration (default password), lack of encryption, code vulnerabilities, etc.,

# Threat Modeling

# Why Threat Model?



# Threat model: A profile of an attacker

- **Goal:** What is the attacker trying to achieve?
- **Capability:**
  - Knowledge: What does the attacker know?
  - Actions: What can the attacker do?
  - Resources: How much effort can it spend?
- **Incentive:** Why does the attacker want to do this?



*"If you know the enemy and know yourself, you need not fear the result of a hundred battles."*  
- Sun Tzu, *The Art of War*

# Attacker Goal

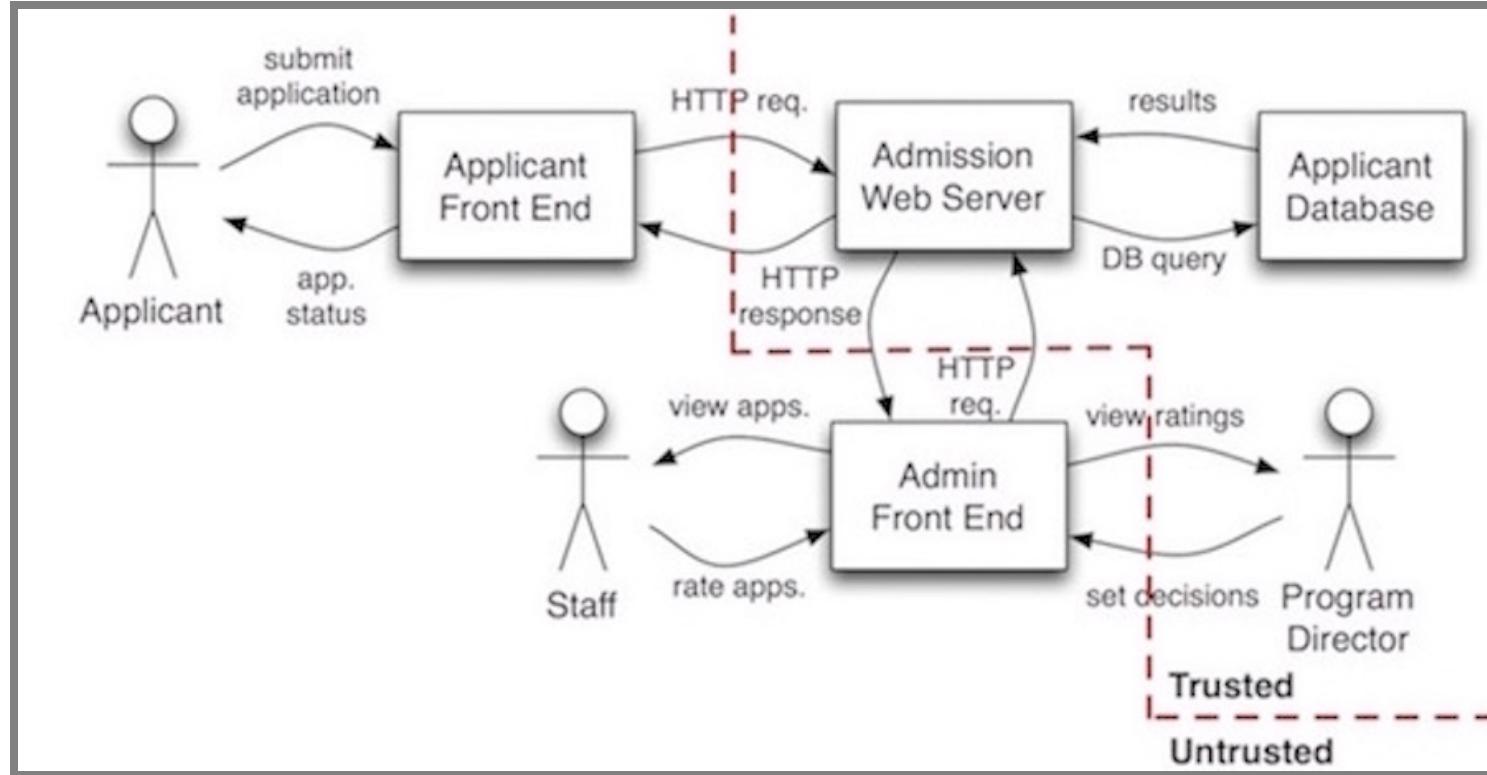
What is the attacker trying to achieve?

- Typically, undermine security requirements (recall C.I.A)

Example: College admission

- Access other applicants info without being authorized (C)
- Modify application status to “accepted” (I)
- Modify admissions model to reject certain applications (I)
- Cause website shutdown to sabotage other applicants (A)

# Attacker Capability



What actions are available to the attacker (to achieve its goal)?

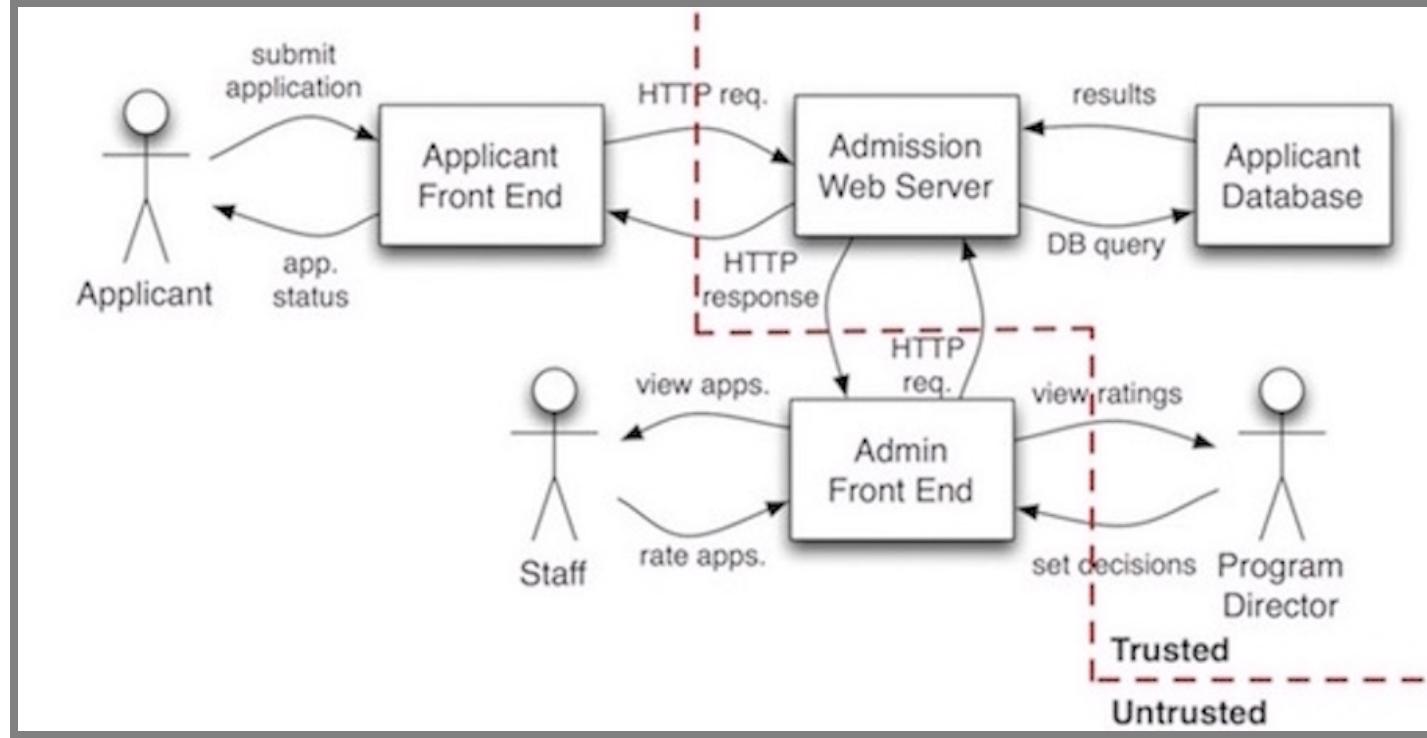
# STRIDE Threat Modeling

	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

A systematic approach to identifying attacks

- Construct an architectural diagram with components & connections
- Indicate trust boundaries
- For each untrusted connection, enumerate STRIDE threats
- For each potential threat, devise a mitigation strategy

# STRIDE: College Admission



- Spoofing: ?
- Tampering: ?
- Information disclosure: ?
- Denial of service: ?

# STRIDE: Example Threats



- Spoofing: Attacker pretends to be another applicant by logging in
- Tampering: Attacker modifies applicant info using browser exploits
- Information disclosure: Attacker intercepts HTTP requests from/to server to read applicant info
- Denial of service: Attacker creates a large number of bogus accounts and overwhelms system with requests

# STRIDE: Example Mitigations

- Spoofing: Attacker pretends to be another applicant by logging in
  - -> **Require two-factor authentication**
- Tampering: Attacker modifies applicant info using browser exploits
  - -> **Add server-side security tokens**
- Information disclosure: Attacker intercepts HTTP requests from/to server to read applicant info
  - -> **Use encryption (HTTPS)**
- Denial of service: Attacker creates many bogus accounts and overwhelms system with requests
  - -> **Limit requests per IP address**

# Breakout: Threat Modeling

Again: Dashcam system from I2

As a group, tagging members,  
post in #lecture:

- Using STRIDE, discuss & post:
  - Data flow throughout the system
  - Possible attacks on the system



# STRIDE & Other Threat Modeling Methods

A systematic approach to identifying threats & attacker actions

Limitations:

- May end up with a long list of threats, not all of them critical
- False sense of security: STRIDE does not imply completeness!

Consider cost vs. benefit trade-offs

- Implementing mitigations add to development cost and complexity
- Focus on most critical/likely threats

# Designing for Security

# Security Mindset



- Assume that all components may be compromised eventually
- Don't assume users will behave as expected; assume all inputs to the system as potentially malicious
- Aim for risk minimization, not perfect security

# Secure Design Principles

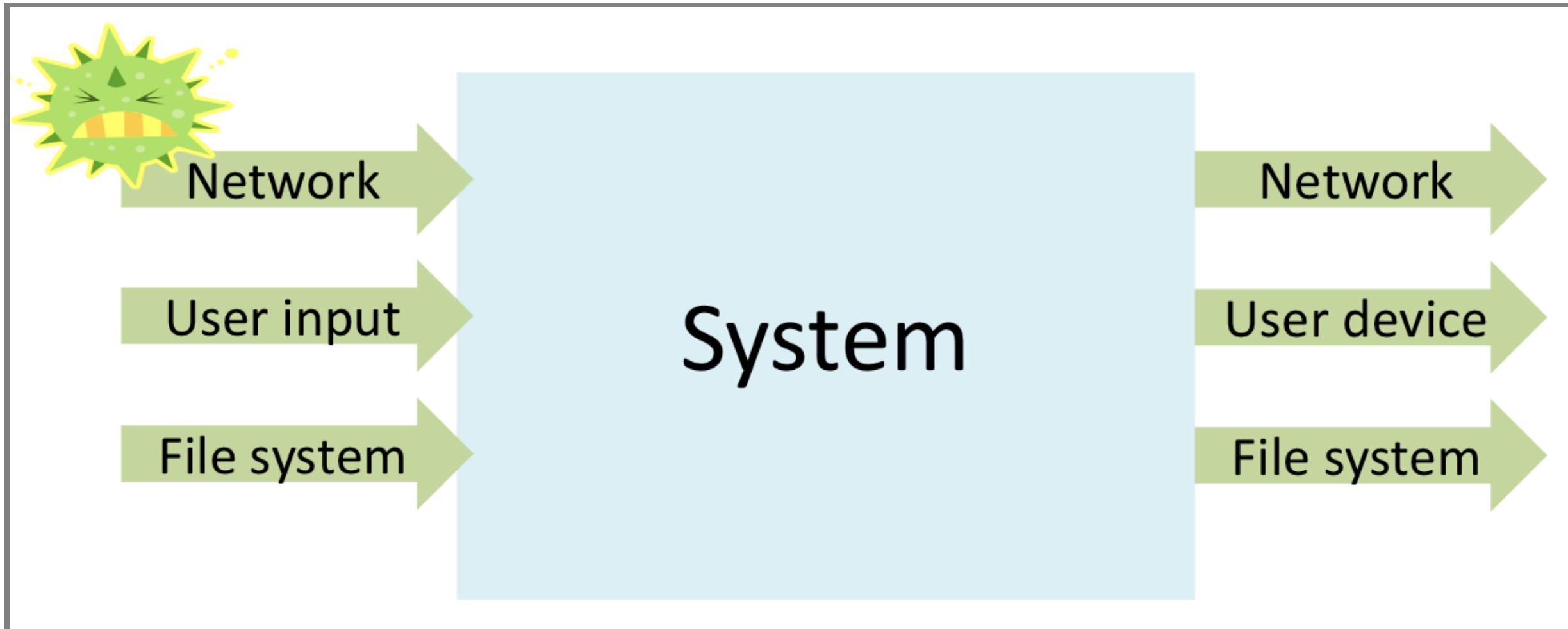
*Minimize the impact of a compromised component*

- **Principle of least privilege:** A component given only minimal privileges needed to fulfill its functionality
- **Isolation/compartmentalization:** Components should be able to interact with each other no more than necessary
- **Zero-trust infrastructure:** Components treat inputs from each other as potentially malicious

*Monitoring & detection*

- Identify data drift and unusual activity

# Monolithic Design

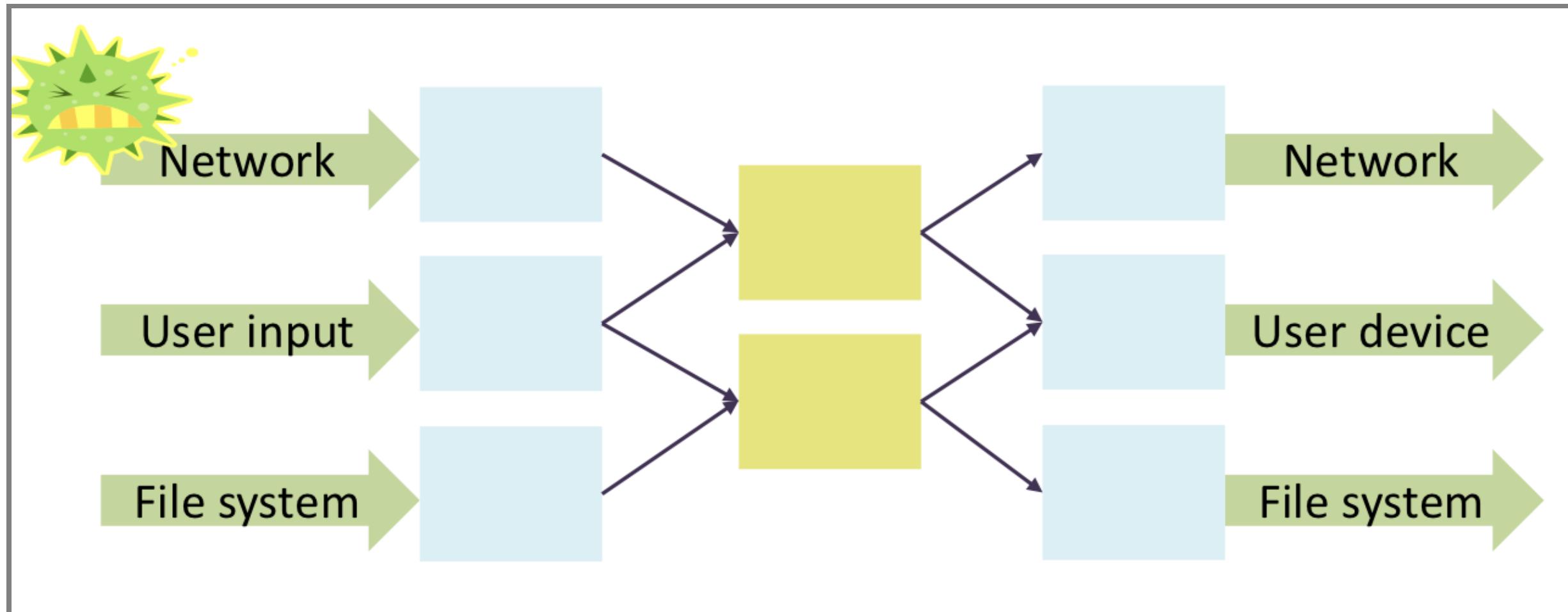


# Monolithic Design

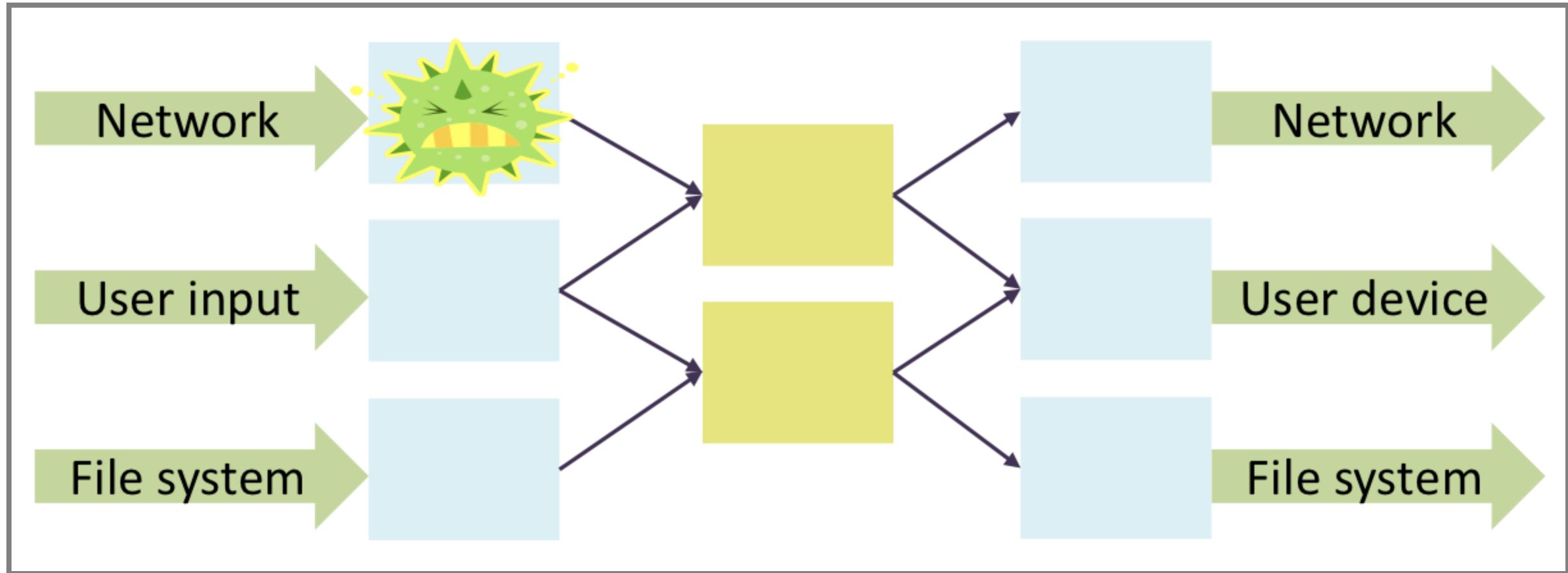


= Flaw in any part => Security impact on entire system!

# Compartmentalized Design



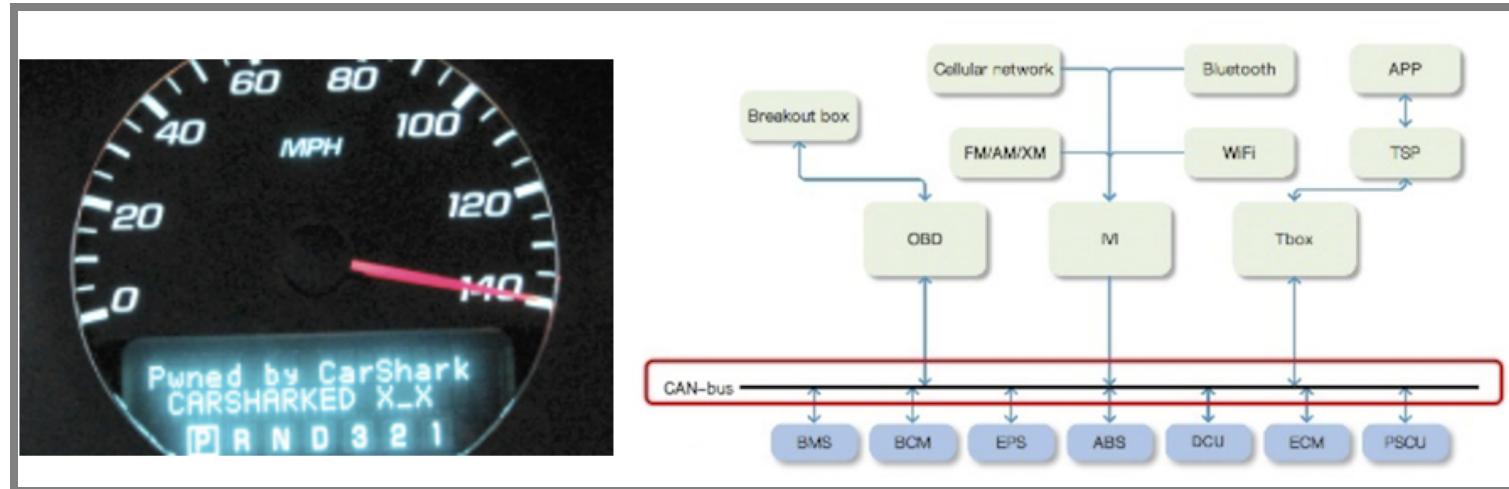
# Compartmentalized Design



Flaw in one component => Limited impact on the rest of the system!

# Example: Vehicle Security

- Research project from UCSD: Remotely taking over vehicle control
  - Create MP3 with malicious code & burn onto CD
  - Play CD => send malicious commands to brakes, engine, locks...
- Problem: Over-privilege & lack of isolation! Shared CAN bus

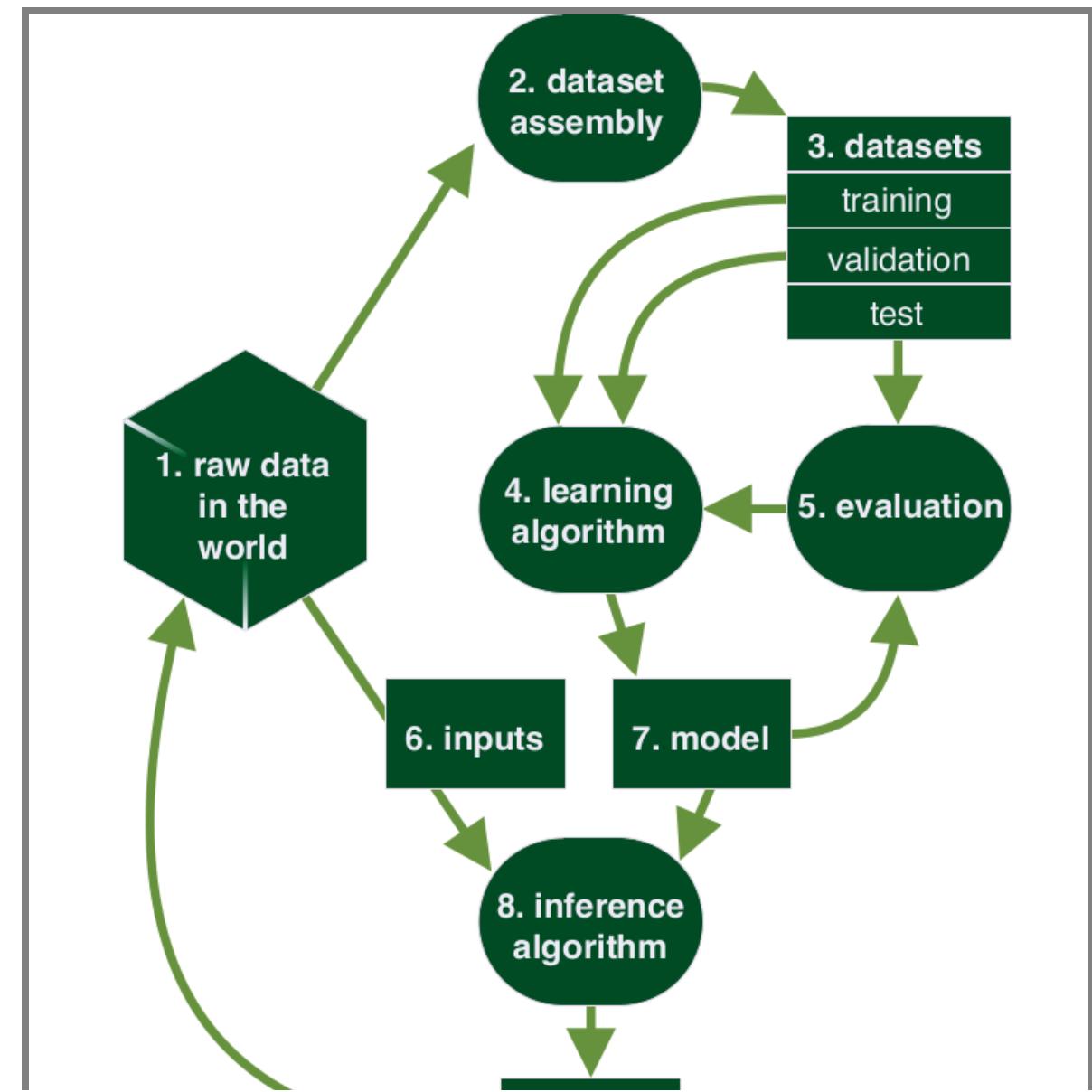


Comprehensive Experimental Analyses of Automotive Attack Surfaces, Checkoway et al., in USENIX Security (2011).

# Secure Design Principles for ML

## *Principle of least privilege*

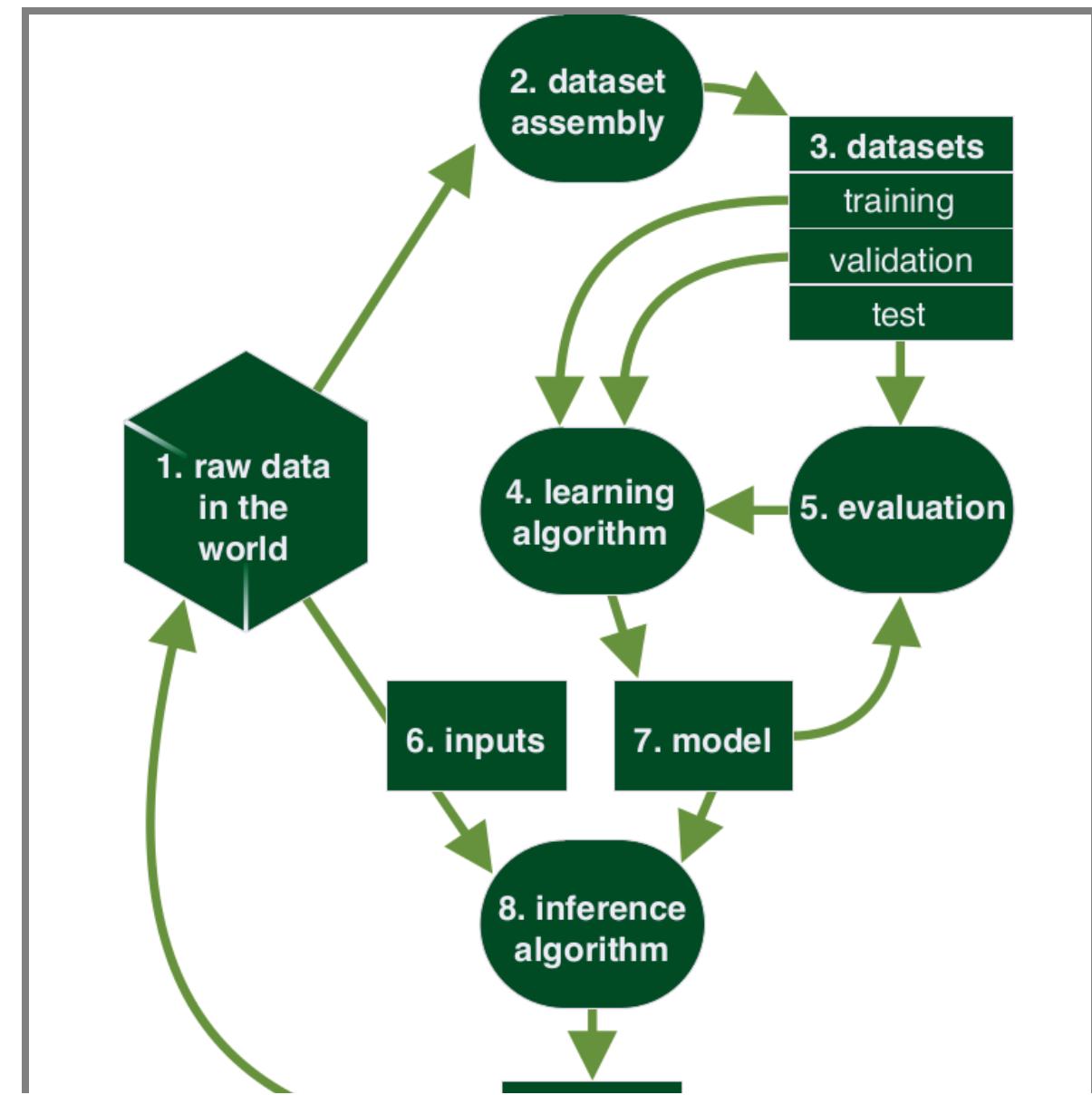
- Who has access to training data, model internal, system input & output, etc.,?
- Does any user/stakeholder have more access than necessary?
- If so, limit access by using authentication mechanisms



# Secure Design Principles for ML

## *Isolation & compartmentalization*

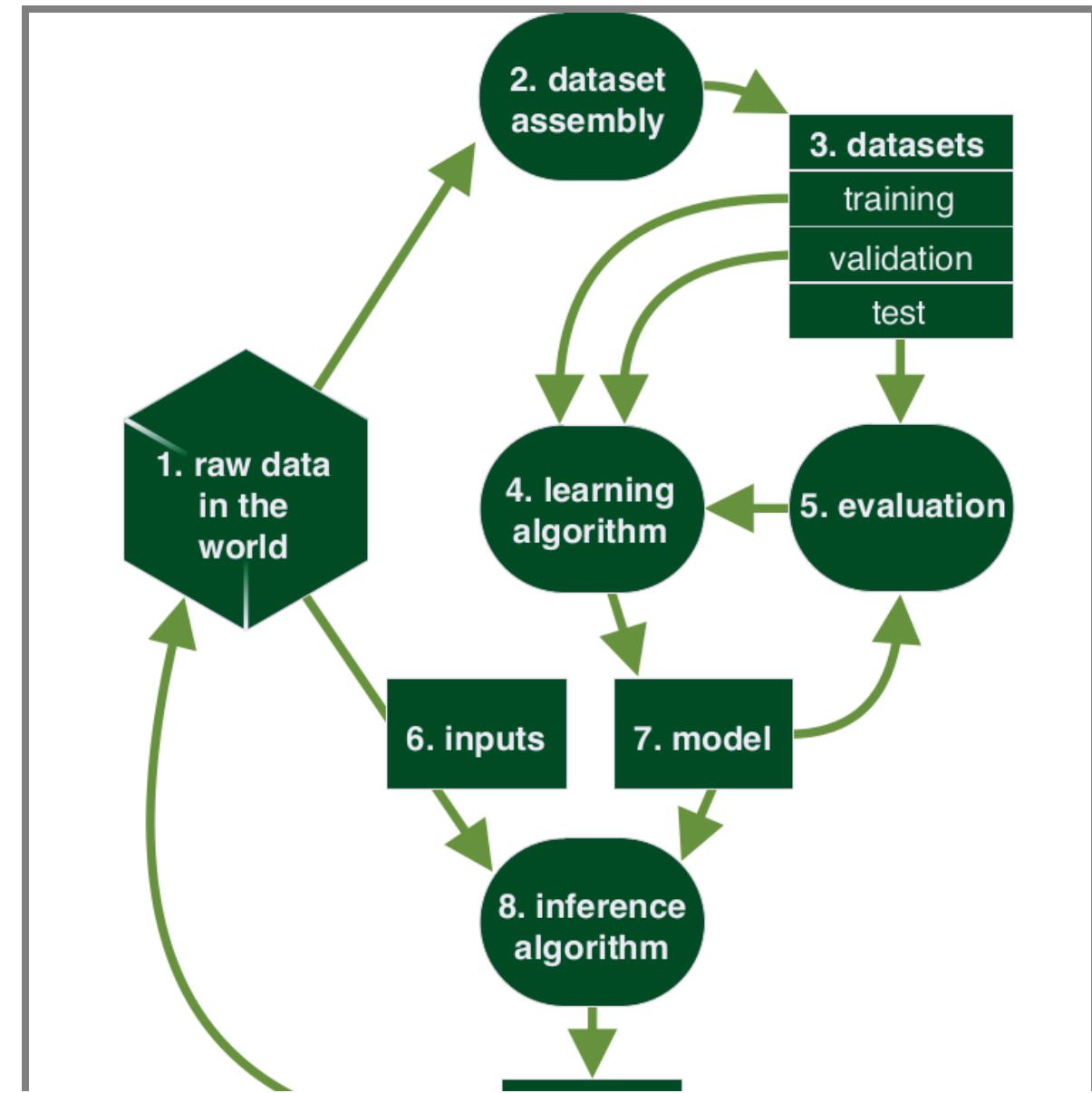
- Can a security attack on one ML component (e.g., misclassification) adversely affect other parts of the system?
- If so, compartmentalize or build in mechanisms to limit impact (see lecture on mitigating mistakes)



# Secure Design Principles for ML

## Monitoring & detection

- Look for odd shifts in the dataset and clean the data if needed (for poisoning attacks)
- Assume all system input as potentially malicious & sanitize (evasion attacks)



# AI for Security



# 30 COMPANIES MERGING AI AND CYBERSECURITY TO KEEP US SAFE AND SOUND

Alyssa Schroer

July 12, 2019 Updated: July 15, 2020

---

**R**y the year 2021, cybercrime losses will

# Many Defense Systems use ML

- Classifiers to learn malicious content: Spam filters, virus detection
- Anomaly detection: Identify unusual/suspicious activity, eg. credit card fraud, intrusion detection
- Game theory: Model attacker costs and reactions, design countermeasures
- Automate incidence response and mitigation activities, DevOps
- Network analysis: Identify bad actors and their communication in public/intelligence data
- Many more, huge commercial interest

Recommended reading: Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "[Anomaly detection: A survey](#)." ACM computing surveys (CSUR) 41, no. 3 (2009): 1-58.

# AI Security Solutions are ML-Enabled Systems Too

ML component one part of a larger system

Consider entire system, from training to telemetry, to user interface, to pipeline automation, to monitoring

ML-based security solutions can be attacked themselves



One contributing factor to the Equifax attack was an expired certificate for an intrusion detection system

# ML & Data Privacy

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Welcome to *The Not-So Private Parts* where technology & privacy collide

Follow

*Andrew Pole, who heads a 60-person team at Target that studies customer behavior, boasted at a conference in 2010 about a proprietary program that could identify women - based on their purchases and demographic profile - who were pregnant.*



Lipka. "What Target knows about you". Reuters, 2014

# What Does Big Tech Know About You? Basically Everything

Security Baron examined the privacy policies of Facebook, Google, Apple, Twitter, Amazon, and Microsoft; just how much these tech giants actually know about you might be surprising..



By [Angela Moscaritolo](#) Updated January 18, 2022    ...

	Google	Facebook	Apple	Twitter	Amazon	Microsoft
Name					x	
Gender				x	x	
Birthday				x	x	
Phone Number						
Email Address						

# Is Privacy Dead?

## Facebook's Zuckerberg Says The Age of Privacy Is Over

By MARSHALL KIRKPATRICK of  **ReadWriteWeb**

Published: January 10, 2010



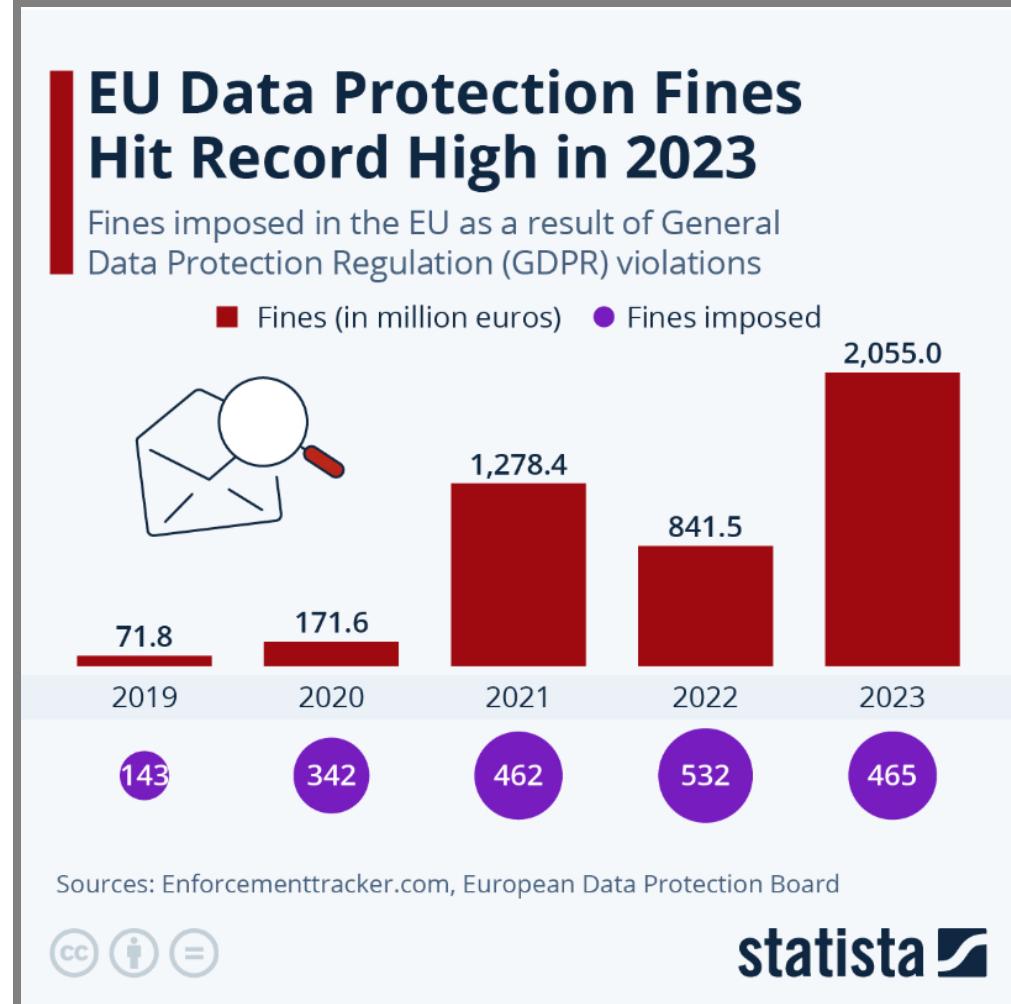
Facebook founder Mark Zuckerberg told a live audience yesterday that if he were to create Facebook again today, user information would by default be public, not private as it was for years until the company changed dramatically in December.

# Is Privacy Dead?



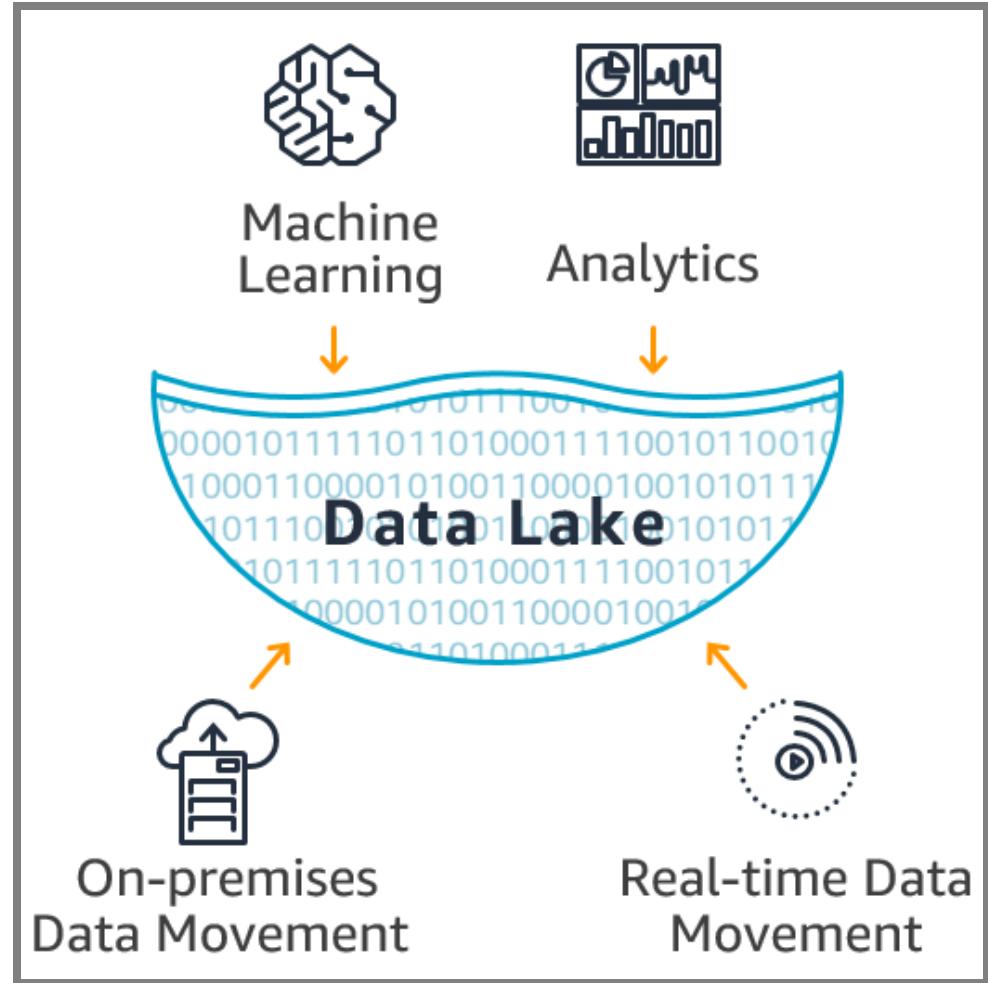
The image shows a screenshot of a Forbes article. At the top, the word "Forbes" is displayed in white on a black header bar. Below it, the word "LEADERSHIP" is written in small capital letters. The main title of the article is "Privacy Is Completely And Utterly Dead, And We Killed It", written in large, bold, dark text. Underneath the title, the author's name "Jacob Morgan" is listed as a "Contributor" with a small profile icon. A bio below the name reads "I write about and explore the future of work!". To the right of the author information is a blue "Follow" button with white text. At the bottom left of the article area, the date "Aug 19, 2014, 12:04am EDT" is printed. The main body text of the article begins with: "Privacy...everyone keeps talking about it and apparently everyone is concerned with it, but going forward does it even matter? I recently watched the documentary, ‘Terms and Conditions may Apply,’ which provides a fascinating look at how organizations such as Facebook, Google, Apple, and others have changed the way they look at and approach privacy. After watching the movie it had me wondering, ‘does privacy even matter anymore?’"

# Is Privacy Dead?



EU Data Protection Fines Hit Record High in 2023. Statista. 2024

# Data Lakes



Who has access?

# Data Privacy vs Utility

PARTNER CONTENT WIRED INSIDER

## FROM DIAGNOSIS TO HOLISTIC PATIENT CARE, MACHINE LEARNING IS TRANSFORMING HEALTHCARE



# Data Privacy vs Utility

The NOVID app interface is displayed on two smartphones. The left phone shows the 'Current Status' screen, which includes a radio button for 'I have not tested positive for COVID-19...' and another for 'I tested positive for COVID-19...'. Below this is a 'Daily Interactions' chart showing interactions per day of the week. The right phone shows the 'Exposures (3)' screen, divided into 'Direct Contact' and 'Indirect Contact' sections. Each section lists an exposure event with details like date reported and contact chain (e.g., YOU → REPORTED). The Carnegie Mellon University logo is visible in the bottom left corner of the main background.

NOVID

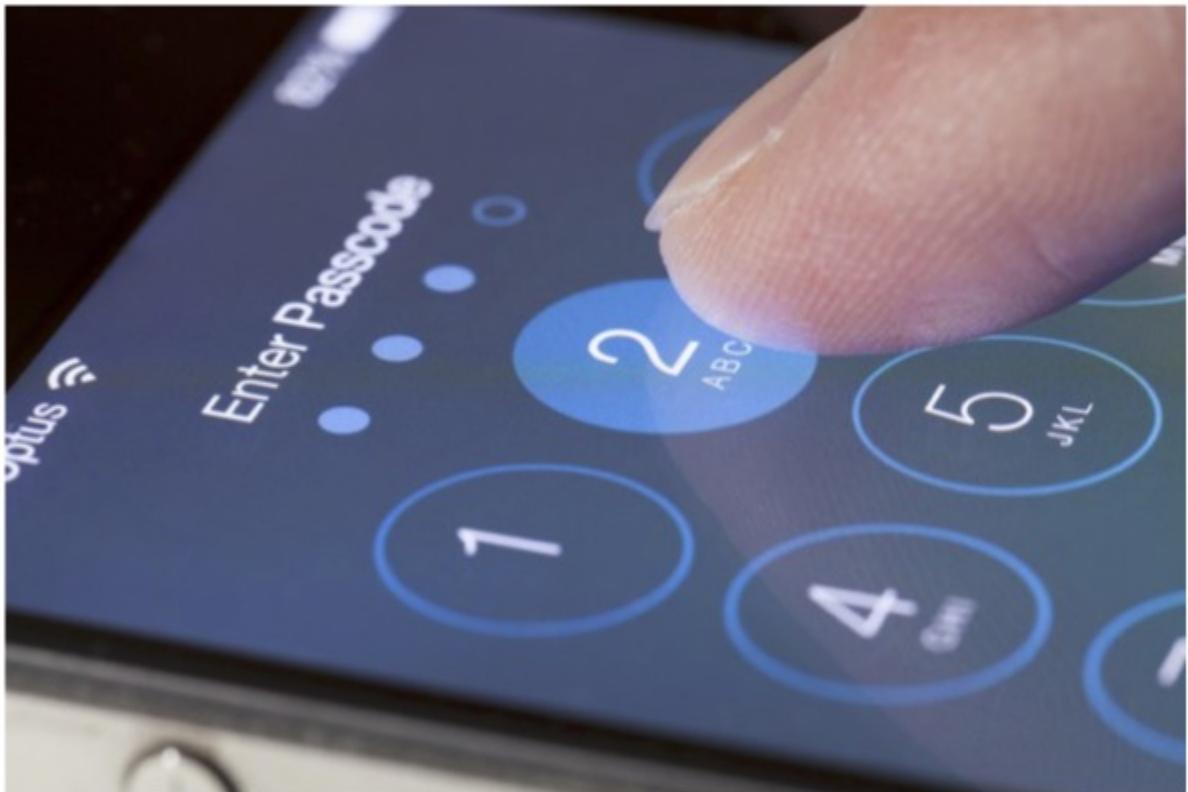
Stop the Spread.

Carnegie Mellon University

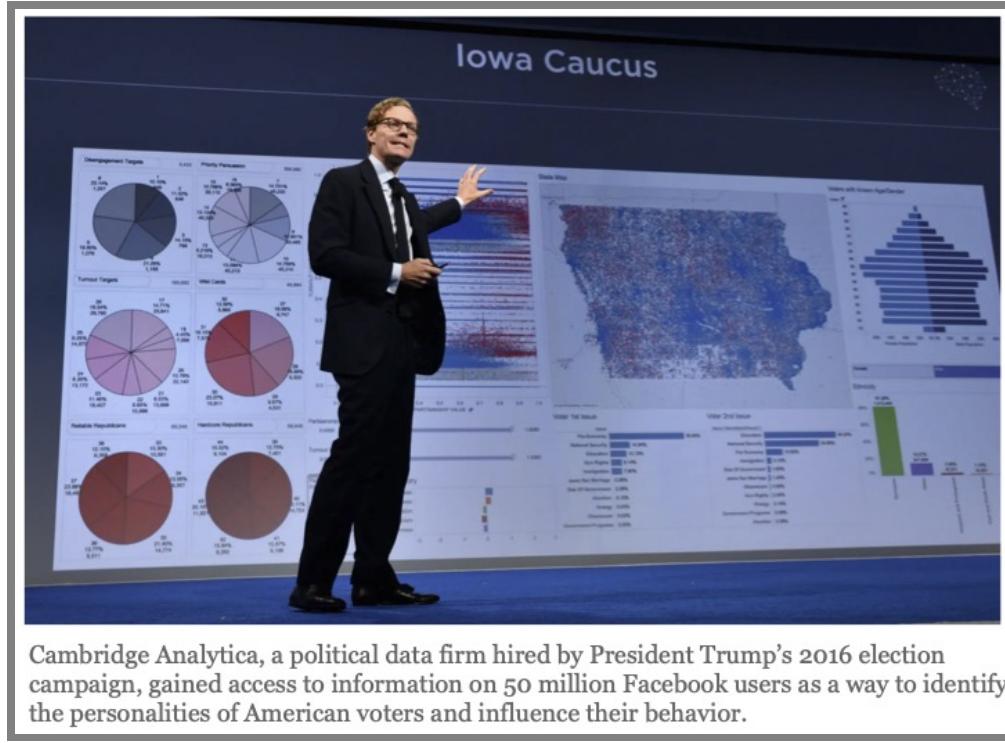
# Data Privacy vs Utility

Apple Fights Court Order to Unlock San Bernardino Shooter's iPhone

19 février 2016



# Data Privacy vs Utility



Cambridge Analytica, a political data firm hired by President Trump's 2016 election campaign, gained access to information on 50 million Facebook users as a way to identify the personalities of American voters and influence their behavior.

- ML can leverage data to greatly benefit individuals and society
- Unrestrained collection & use of data can enable abuse and harm!
- **Viewpoint:** Users should be given an ability to learn and control how their data is collected and used

# US FTC's Fair Information Practice Principles

- Notice/awareness (core principle)
  - Disclose practices
- Choice/consent (core principle)
  - Opt-in, opt-out
- Access/participation
  - Users should be able to review & correct their information
- Integrity/Security
  - Ensure is secure, limited access
- Enforcement
  - Mechanisms for handling violations

# LINDDUN Taxonomy of Privacy Threats



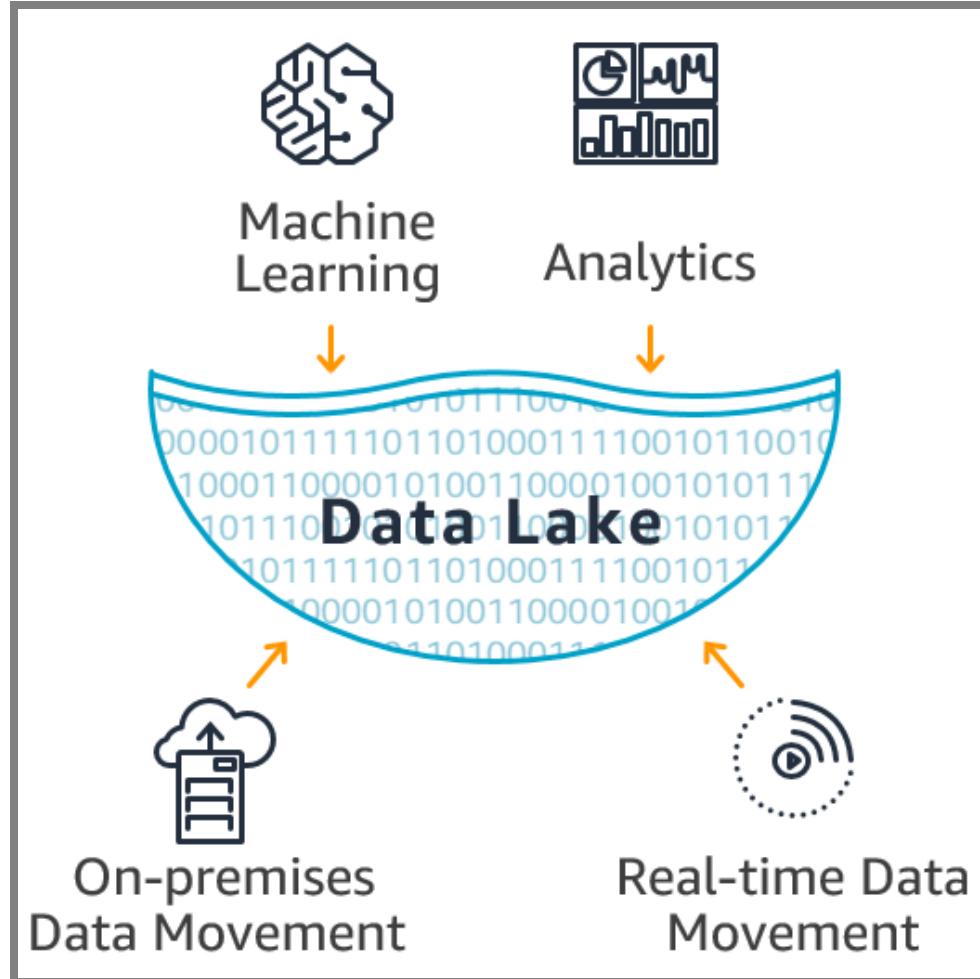
# Best Practices for ML & Data Privacy

- Data collection & processing
  - Only collect and store what you need
  - Remove sensitive attributes, anonymize, or aggregate
- Training: Local, on-device processing if possible
  - Federated learning
- Basic security practices
  - Encryption & authentication
  - Provenance: Track data sources and destinations
- Provide transparency to users
  - Clearly explain what data is being collected and why
- Understand and follow the data protection regulations!
  - e.g., General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), HIPAA (healthcare), FERPA (educational)

# Best Practices for ML & Data Privacy

- Data collection & processing
  - Only collect and store what you need
  - Remove sensitive attributes, anonymize, or aggregate
- Training: Local, on-device processing if possible
  - Federated learning
- Basic security practices
  - Encryption & authentication
  - Provenance: Track data sources and destinations
- Provide transparency to users
  - Clearly explain what data is being collected and why
- Understand and follow the data protection regulations!
  - e.g., General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), HIPAA (healthcare), FERPA (educational)

# Collect and store only what you need



*Realistic when data is seen as valuable?*

# Data Anonymization is Hard

- Simply removing explicit identifiers (e.g., name) is often not enough
  - {ZIP, gender, birthd.} can identify 87% of Americans (L. Sweeney)
- k-anonymization: Identity-revealing data tuples appear in at least k rows
  - Suppression: Replace certain values in columns with an asterisk
  - Generalization: Replace individual values with broader categories

# Best Practices for ML & Data Privacy

- Data collection & processing
  - Only collect and store what you need
  - Remove sensitive attributes, anonymize, or aggregate
- **Training: Local, on-device processing if possible**
  - Federated learning
- Basic security practices
  - Encryption & authentication
  - Provenance: Track data sources and destinations
- Provide transparency to users
  - Clearly explain what data is being collected and why
- Understand and follow the data protection regulations!
  - e.g., General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), HIPAA (healthcare), FERPA (educational)

# Federated Learning



- Train a global model with local data stored across multiple devices
- Local devices push only model updates, not the raw data
- But: increased network communication and other security risks (e.g., backdoor injection)

ML@CMU blog post on federated learning

# Best Practices for ML & Data Privacy

- Data collection & processing
  - Only collect and store what you need
  - Remove sensitive attributes, anonymize, or aggregate
- Training: Local, on-device processing if possible
  - Federated learning
- Basic security practices
  - Encryption & authentication
  - Provenance: Track data sources and destinations
- Provide transparency to users
  - Clearly explain what data is being collected and why
- **Understand and follow the data protection regulations!**
  - e.g., General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), HIPAA (healthcare), FERPA (educational)

# General Data Protection Reg. (GDPR)

- Introduced by the European Union (EU) in 2016
- Organizations must state:
  - What personal data is being collected & stored
  - Purpose(s) for which the data will be used
  - Other entities that the data will be shared with
- Organizations must receive explicit consent from users
  - Each user must be provided with the ability to view, modify and delete any personal data
- Compliance & enforcement
  - Complaints are filed against non-compliant organizations
  - A failure to comply may result in heavy penalties!

# Privacy Consent and Control



The image shows a screenshot of a privacy consent banner from TechCrunch. At the top left is the TechCrunch logo (TC). Below it is the slogan "Your data. Your experience." followed by a paragraph of text explaining data storage and access for advertising purposes. A section titled "Your personal data that may be used" lists three items: device information, browsing activity, and precise location. Below this is a note about reading the Privacy Policy and Cookie Policy, which is highlighted with a blue oval. At the bottom are two buttons: "I agree" (green) and "Manage settings" (white).

**TC**

**Your data. Your experience.**

TechCrunch is part of [Verizon Media](#). We and [our partners](#) will store and/or access information on your device through the use of cookies and similar technologies, to display personalised ads and content, for ad and content measurement, audience insights and product development.

**Your personal data that may be used**

- Information about your device and internet connection, including your IP address
- Browsing and search activity while using Verizon Media websites and apps
- [Precise location](#)

Find out more about how we use your information in our [Privacy Policy](#) and [Cookie Policy](#).

To enable Verizon Media and our partners to process your personal data select '[I agree](#)', or select '[Manage settings](#)' for more information and to manage your choices. You can change your choices at any time by visiting [Your Privacy Controls](#).

I agree

Manage settings

# But Does Informed Consent Work?

The screenshot shows a dark-themed website with a prominent white pop-up in the center. The pop-up has a title 'Welcome!' and two main options: 'Continue reading with ads' and 'Read ad-free'. Below each option is a brief description and a call-to-action button ('Accept and continue >' or 'Check out details >'). A note at the bottom explains tracking practices and third-party providers. The background of the pop-up features a large American flag. The rest of the page shows blurred news headlines.

Im »Interesse der Öffentlichkeit«

## US-Justiznachrichten ernennt Trump einen Sonderermittler

Die Vorwürfe gegen Donald Trump sind zahlreich – und massiv. Ein unabhängiger Sonderermittler soll den Auftrag des US-Justizministers übernehmen, wofür der Ex-Präsident keine Zustimmung mehr geben kann.

2 Min

### Russlands Krieg gegen die Ukraine

Tracking: We work with **third party providers** to improve and finance our web products. Together with these third-party providers, we collect and process personal data on our platforms. Using cookies stored on your device, personal identifiers such as device identifiers or IP addresses, and based on your individual usage patterns, together with these third party providers we can ...

- ... Store and/or retrieve information on a device: For the processing purposes disclosed to you, cookies, device identifiers or other information may be stored or accessed on your device.
- ... Execute personalized ads and content, ad and content measurement, audience and product development insights: Ads and content can be personalized based on a profile. Additional data can be added to improve personalized ads and content. The performance of ads and content will be measured. Insights will be derived about the target groups that have viewed the advertisements and content. Data may be used to create or improve usability, systems and software.

You also consent to your data being processed by providers in third countries and the United States. There is a risk that U.S. providers may be required to share their data with the authorities there. As such, the U.S. is assessed as a country with an insufficient level of data protection according to EU standards (for third country consent).

Imprint Privacy Policy General Terms and Conditions Zur deutschen Seite wechseln

Datenbearbeitung in Drittländern

# Amazon hit with \$886m fine for alleged data law breach

⌚ 30 July 2021



GETTY IMAGES

# Summary: Best Practices for ML & Data Privacy

***Be ethical and responsible with user data! Think about potential harms to users & society, caused by (mis-)handling of personal data***

- Data collection & processing
- Training: Local, on-device processing if possible
- Basic security practices
- Provide transparency to users
- Understand and follow the data protection regulations!

# Summary

- Security requirements: Confidentiality, integrity, availability
- Threat modeling to identify security req. & attacker capabilities
- ML-specific attacks on training data, telemetry, or the model
  - Poisoning attack on training data to influence predictions
  - Evasion attacks (adversarial learning) to shape input data
  - Model inversion attacks for privacy violations
- Security design at the system level: least privilege, isolation
- AI can be used for defense (e.g. anomaly detection)
- **Key takeaway:** Adopt a security mindset! Assume all components may be vulnerable. Design system to reduce the impact of attacks.

# Further Readings

- Gary McGraw, Harold Figueroa, Victor Shepardson, and Richie Bonett. [An Architectural Risk Analysis of Machine Learning Systems: Toward More Secure Machine Learning](#). Berryville Institute of Machine Learning (BIML), 2020
- Meftah, Barmak. Business Software Assurance: Identifying and Reducing Software Risk in the Enterprise. 9th Semi-Annual Software Assurance Forum, Gaithersburg, Md., October 2008.
- Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In CVPR, 2018.
- Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. Communications of the ACM, 61(7), 56-66. 2018.
- Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. arXiv, 2017

