SECURITY, ADVERSARIAL LEARNING, AND PRIVACY

Christian Kaestner

with slides from Eunsuk Kang

Required reading: Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering." (2018), Chapter 25 (Adversaries and Abuse) Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press. Chapter 19 (Managing AI Risk)

Recommended reading: Goodfellow, I., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, *61*(7), 56-66. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (pp. 43-58).

LEARNING GOALS

- Explain key concerns in security (in general and with regard to ML models)
- Analyze a system with regard to attacker goals, attack surface, attacker capabilities
- Describe common attacks against ML models, including poisoning attacks, evasion attacks, leaking IP and private information
- Measure robustness of a prediction and a model
- Understand design opportunities to address security threats at the system level
- Identify security requirements with threat modeling
- Apply key design principles for secure system design
- Discuss the role of AI in securing software systems

SECURITY AT THE MODEL LEVEL

- Various attack discussions, e.g. poisioning attacks
- Model robustness
- Attack detection
- ...

SECURITY AT THE SYSTEM LEVEL

- Requirements analysis
- System-level threat modeling
- Defense strategies beyond the model
- Security risks beyond the model
- ...

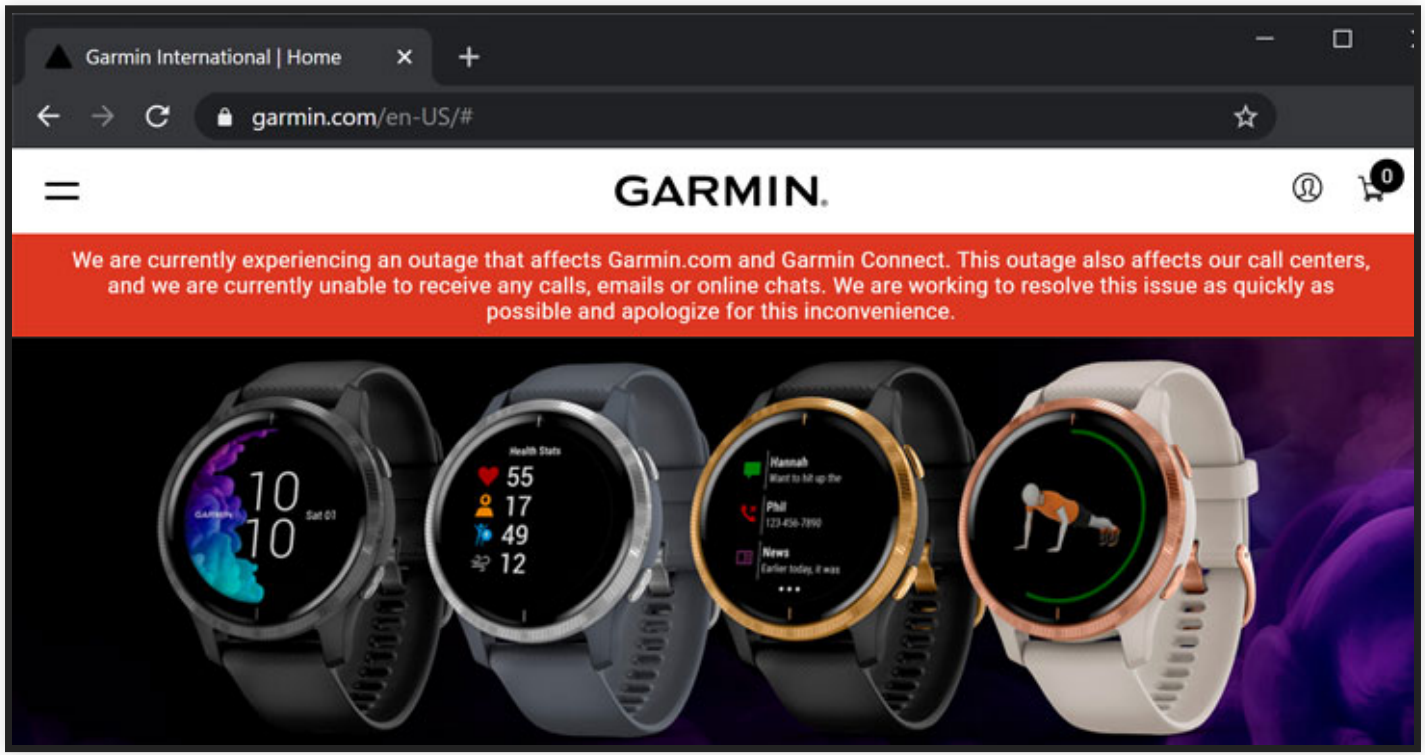
SECURITY

ELEMENTS OF SECURITY

- Security requirements (policies)
  - What does it mean for my system to be secure?
- Threat model
  - What are the attacker's goal, capability, and incentive?
- Attack surface
  - Which parts of the system are exposed to the attacker?
- Protection mechanisms
  - How do we prevent the attacker from compromising a security requirement?

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


Garmin International | Home
garmin.com/en-US/#
GARMIN.
0
We are currently experiencing an outage that affects Garmin.com and Garmin Connect. This outage also affects our call centers, and we are currently unable to receive any calls, emails or online chats. We are working to resolve this issue as quickly as possible and apologize for this inconvenience.
10
10
Sat 01
Health Stats
55
17
49
12
Hannah
Phil
123-456-7890
News

OTHER SECURITY PROPERTIES

- Authentication (no spoofing): Users are who they say they are
- Integrety (no tampering): Data is changed only through authorized processes
- Non-repudiation: Every change can be traced to who was responsible for it
- Confidentiality (no inform. disclosure): Information only accessible to authorized users
- Availability (no denial of service): Critical services must be available when needed by clients
- Authorization (no escalation of privilege): Only users with the right permissions can access a resource/perform an action

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 application site should backup all applications in case of a server failure.
- The acceptance notices can only be sent out by the program director.

CIA OF AN ML MODEL

*What are security concerns of a **ML model** for ranking applications?*

- **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

Speaker notes

Many examples: Confidentiality attacks: try to infer sensitive labels for data (e.g. training instances) Integrity: cause a model to misclassify a data point, e.g. spam as nonspam Availability attack: Misclassify many data points to make a model essentially useless

UNDERSTANDING ATTACKER GOALS

WHY THREAT MODEL?

WHAT IS THREAT MODELING?

- 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?

ATTACKER GOALS AND INCENTIVES

- What is the attacker trying to achieve? Undermine one or more security requirements
- Why does the attacker want to do this?

Example goals and incentives in Garmin/college admission scenario?

Speaker notes

- Access other applicants info without being authorized
    - Modify application status to “accepted”
        - Submit applications that get accepted
        - Cause expense by making the model useless and forcing manual evaluations or poor outcomes
    - Cause website shutdown to sabotage other applicants

ATTACKS ON ML MODELS

SCENARIO: RANKINGS AND REVIEWS ON WEB SHOP

Antique Box Ugears, 3D Mechanical Treasure Models, Self-Assembling Precut Wooden Gift, DIY Craft Set

4.5 out of 5 stars ⌄ 261

$41.90 ~~$44.90~~

FREE Delivery for Prime members

Only 1 left in stock - order soon.

More Buying Choices

$38.89 (44 new offers)

ROKR 3D Wooden Puzzle for Adults-Mechanical Train Model Kits-Brain Teaser Puzzles-Vehicle Building Kits-Unique Gi...

4.0 out of 5 stars ⌄ 44

$22.99

prime FREE One-Day

Get it **Tomorrow, Jul 26**

Ages: 14 years and up

Wooden Puzzles for Toddlers, Aitey Wooden Alphabet Number Puzzles Toddler Learning Puzzle Toys for Kids Ages 2 3 4 (Set of...

4.5 out of 5 stars ⌄ 283

$23.99

prime FREE One-Day

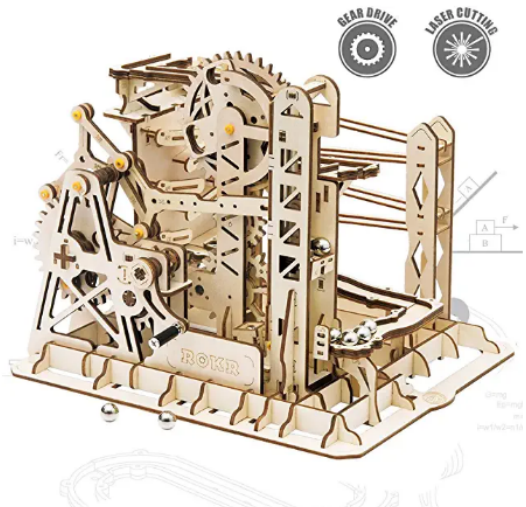
Get it **Tomorrow, Jul 26**

More Buying Choices

$22.79 (2 used & new offers)

Ages: 14 years and up

Ages: 12 months and up



ROKR 3D Assembly Wooden Puzzle Brain Teaser Game Mechanical Gears Set Model Kit Marble Run Set Unique Craft...

4.5 out of 5 stars ⌄ 172

$30.99

Unidragon Wooden Jigsaw Puzzles - Unique Shape Jigsaw Pieces Best Gift for Adults and Kids Alluring Fox 7 × 9.2 in (18 ...

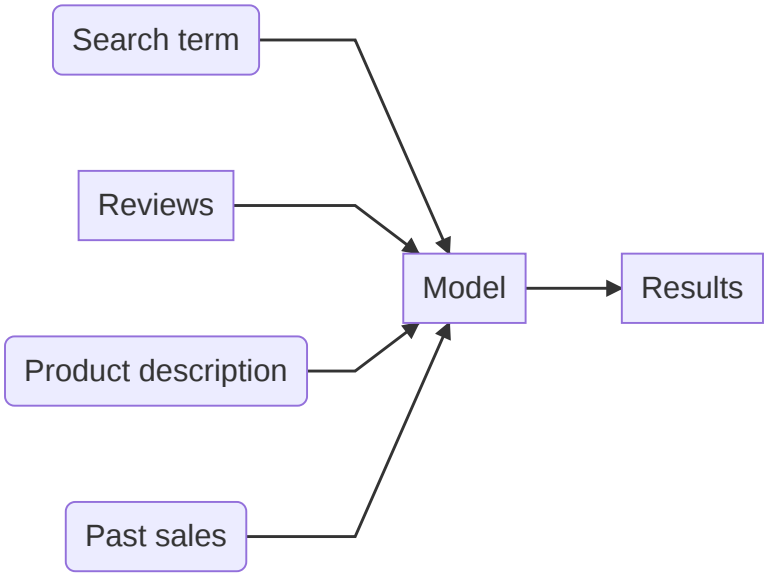
5.0 out of 5 stars ⌄ 13

$40.99 $59.99

KINGZHUO Hexagon Tangram Classic Handmade Wooden Puzzle for Children and Adults Challenging Puzzles Brain...

4.5 out of 5 stars ⌄ 263

$9.98


Search term
Reviews
Product description
Past sales
Model
Results

SCENARIO: SPAM FILTER

CAPABILITIES

How can an attacker interact with / influence the model?

ATTACK VECTORS

- Influence the training data ("causative attack", "poisioning attack")
- Influence the input data ("exploratory attack", "evasion attack")
- Influence the telemetry data

Examples in spam filter scenario?

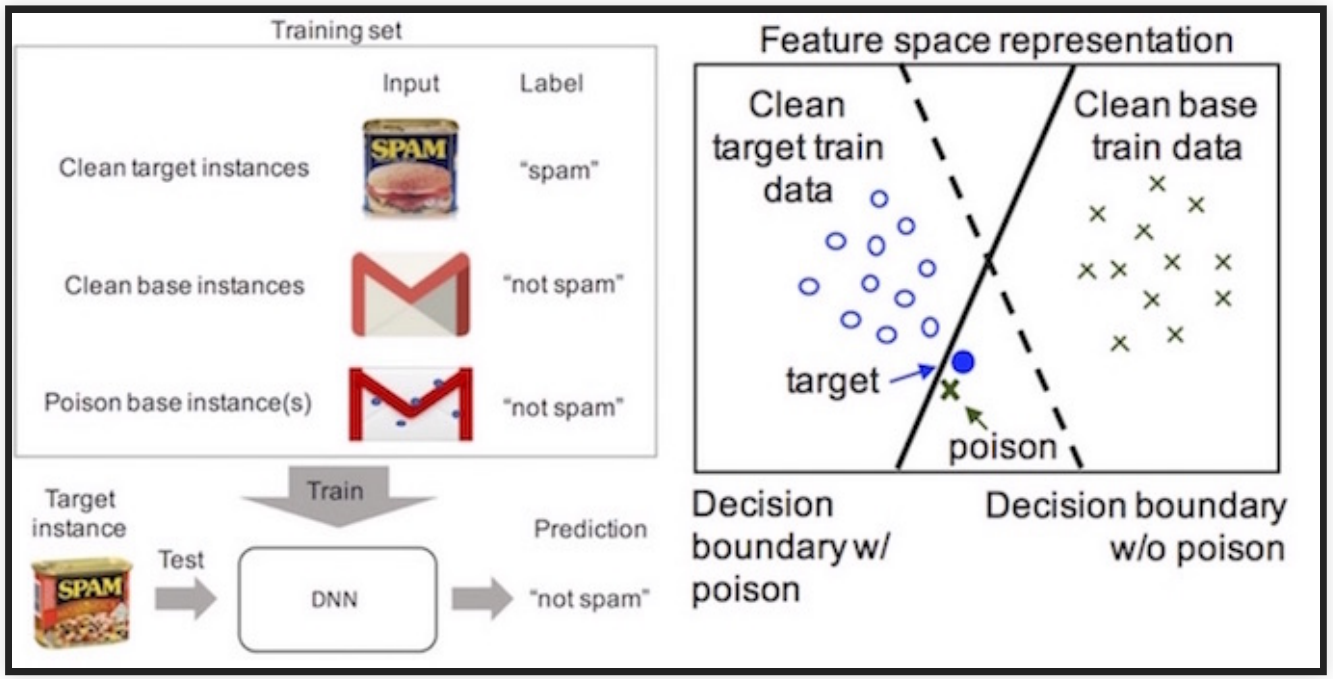
POISONING ATTACK: AVAILABILITY

- Availability: Inject mislabeled training data to damage model quality
  - 3% poisoning => 11% decrease in accuracy (Steinhardt, 2017)
- Attacker must have some access to the training set
  - models trained on public data set (e.g., ImageNet)
  - retrained automatically on telemetry

Speaker notes

- Example: Anti-virus (AV) scanner
    - Online platform for submission of potentially malicious code
        - Some AV company (allegedly) poisoned competitor's model

POISONING ATTACK: INTEGRITY



- Insert training data with seemingly correct labels
- More targeted than availability attacks
    - Cause misclassification from one specific class to another

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

MANY DIFFERENT KINDS OF ATTACKS ON TRAINING DATA

- Correlated outlier attack: add spurious features to malicious instances to misclassify benign instances
- Red herring attack: add spurious features to early malicious instances, then send malicious payload without those features

POISONING ATTACK IN WEB SHOP?

Antique Box Ugears, 3D Mechanical Treasure Models, Self-Assembling Precut Wooden Gift, DIY Craft Set

4.5 stars 261

$41.90 ~~$44.90~~

FREE Delivery for Prime members

Only 1 left in stock - order soon.

More Buying Choices

$38.89 (44 new offers)

Ages: 14 years and up

ROKR 3D Wooden Puzzle for Adults-Mechanical Train Model Kits-Brain Teaser Puzzles-Vehicle Building Kits-Unique Gi...

4.0 stars 44

$22.99

prime FREE One-Day

Get it **Tomorrow, Jul 26**

Ages: 14 years and up

Wooden Puzzles for Toddlers, Aitey Wooden Alphabet Number Puzzles Toddler Learning Puzzle Toys for Kids Ages 2 3 4 (Set of...

4.5 stars 283

$23.99

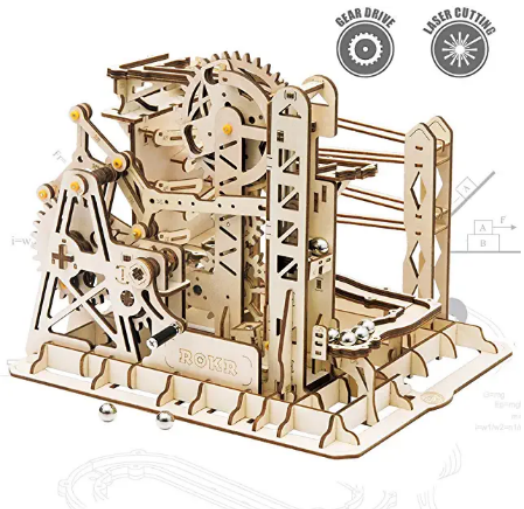
prime FREE One-Day

Get it **Tomorrow, Jul 26**

More Buying Choices

$22.79 (2 used & new offers)

Ages: 12 months and up



ROKR 3D Assembly Wooden Puzzle Brain Teaser Game Mechanical Gears Set Model Kit Marble Run Set Unique Craft...

4.5 out of 5 stars 172

$30.99

Unidragon Wooden Jigsaw Puzzles - Unique Shape Jigsaw Pieces Best Gift for Adults and Kids Alluring Fox 7 × 9.2 in (18 ...

5.0 out of 5 stars 13

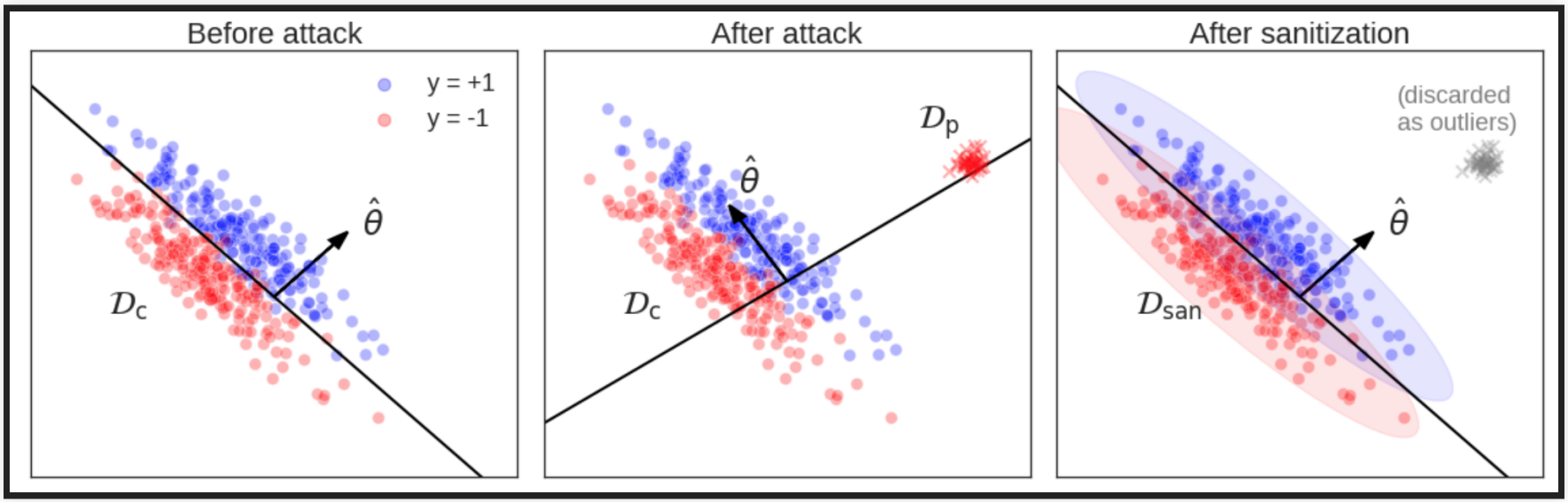
$40.99 $50.99

KINGZHUO Hexagon Tangram Classic Handmade Wooden Puzzle for Children and Adults Challenging Puzzles Brain...

4.5 out of 5 stars 263

$8.98

DEFENSE AGAINST POISONING ATTACKS



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

DEFENSE AGAINST POISONING ATTACKS

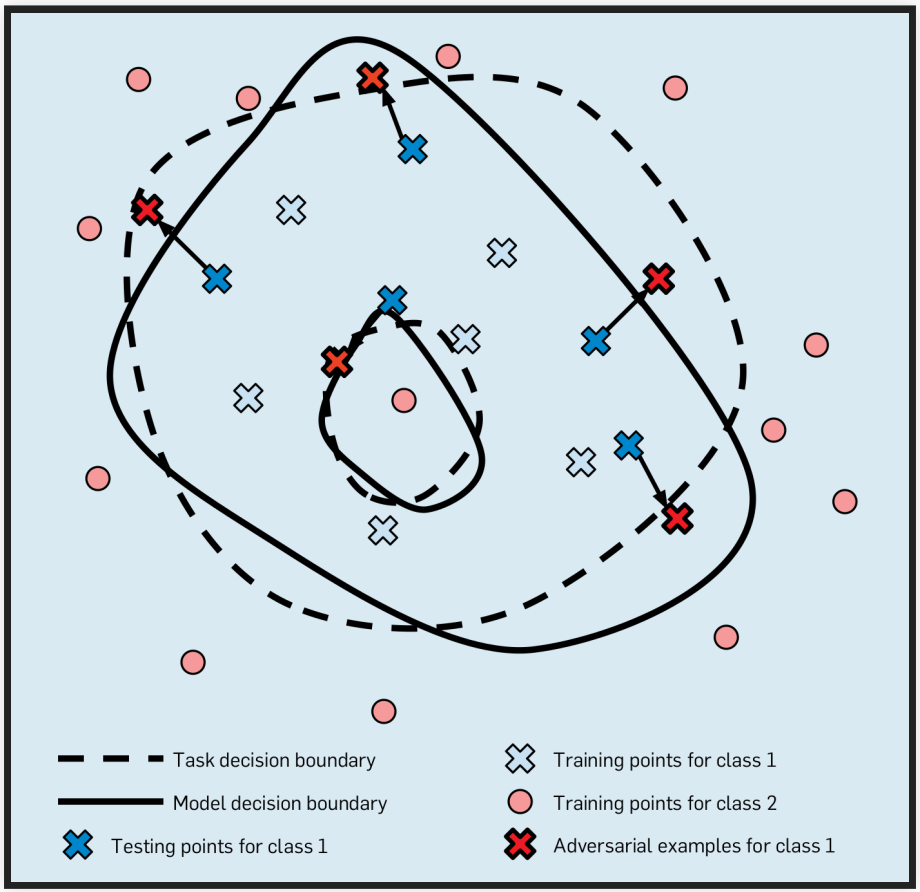
- Anomaly detection & data sanitization
  - Identify and remove outliers in training set
  - Identify and understand drift from telemetry
  - See data quality lecture
- Quality control over your training data
  - Who can modify or add to my training set? Do I trust the data source?
  - Use security mechanisms (e.g., authentication) and logging to track data provenance
- Slow down retraining, monitor model quality
- Debug models + explainability (e.g., influential instances)
- Use models that are robust against noisy training data

ATTACKS ON INPUT DATA (EVASION ATTACKS, ADVERSARIAL EXAMPLES)

- Add noise to an existing sample & cause misclassification
  - achieve specific outcome (evasion attack)
  - circumvent ML-based authentication like FaceID (impersonation attack)
- Attack at inference time

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

TASK DECISION BOUNDARY VS MODEL BOUNDARY



From Goodfellow et al (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, *61*(7), 56-66.

GENERATING ADVERSARIAL EXAMPLES

- see counterfactual explanations
- Find similar input with different prediction
  - targeted (specific prediction) vs untargeted (any wrong prediction)
- Many similarity measures (e.g., change one feature vs small changes to many features)
  - $x^* = x + argmin\{|z| : f(x+z) = t\}$
- Attacks more affective which access to model internals, but also black-box attacks (with many queries to the model) feasible
  - With model internals: follow the model's gradient
  - Without model internals: learn surrogate model
  - With access to confidence scores: heuristic search (eg. hill climbing)

EXAMPLE OF EVASION ATTACKS

Spam scenario? Web store scenario? Credit scoring scenario?

Antique Box Ugears, 3D Mechanical Treasure Models, Self-Assembling Precut Wooden Gift, DIY Craft Set

4.5 out of 5 stars 261

$41.90 ~~$44.90~~

FREE Delivery for Prime members

Only 1 left in stock - order soon.

More Buying Choices

$38.89 (44 new offers)

ROKR 3D Wooden Puzzle for Adults-Mechanical Train Model Kits-Brain Teaser Puzzles-Vehicle Building Kits-Unique Gi...

4.4 out of 5 stars 44

$22.99

prime FREE One-Day

Get it **Tomorrow, Jul 26**

Ages: 14 years and up

Wooden Puzzles for Toddlers, Aitey Wooden Alphabet Number Puzzles Toddler Learning Puzzle Toys for Kids Ages 2 3 4 (Set of...

4.5 out of 5 stars 283

$23.99

prime FREE One-Day

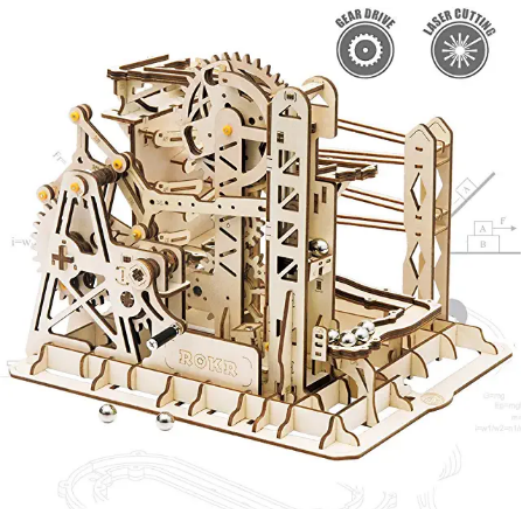
Get it **Tomorrow, Jul 26**

More Buying Choices

$22.79 (2 used & new offers)

Ages: 14 years and up

Ages: 12 months and up



ROKR 3D Assembly Wooden Puzzle Brain Teaser Game Mechanical Gears Set Model Kit Marble Run Set Unique Craft...

4.5 out of 5 stars 172

$30.99

Unidragon Wooden Jigsaw Puzzles - Unique Shape Jigsaw Pieces Best Gift for Adults and Kids Alluring Fox 7 × 9.2 in (18 ...

5.0 out of 5 stars 13

$40.99 $50.00

KINGZHUO Hexagon Tangram Classic Handmade Wooden Puzzle for Children and Adults Challenging Puzzles Brain...

4.5 out of 5 stars 263

$8.98

RECALL: GAMING MODELS WITH WEAK FEATURES

Does providing an explanation allow customers to 'hack' the system?

- Loan applications?
- Apple FaceID?
- Recidivism?
- Auto grading?
- Cancer diagnosis?
- Spam detection?

Gaming not possible if model boundary = task decision boundary

DISCUSSION: CAN WE SECURE A SYSTEM WITH A KNOWN MODEL?

- Can we protect the model?
- How to prevent surrogate models?
- Security by obscurity?

- Alternative model hardening or system design strategies?
- 

EXCURSION: ROBUSTNESS

property with massive amount of research, in context of security and safety

DEFINING ROBUSTNESS:

- A prediction for x is robust if the outcome is stable under minor perturbations of the input
  - $\forall x^{'} . d(x, x^{'}) < \epsilon \Rightarrow f(x) = f(x^{'})$
  - distance function d and permissible distance ϵ depends on problem
- A model is robust if most predictions are robust

ROBUSTNESS AND DISTANCE FOR IMAGES

- slight rotation, stretching, or other transformations
- change many pixels minimally (below human perception)
- change only few pixels
- change most pixels mostly uniformly, eg brightness

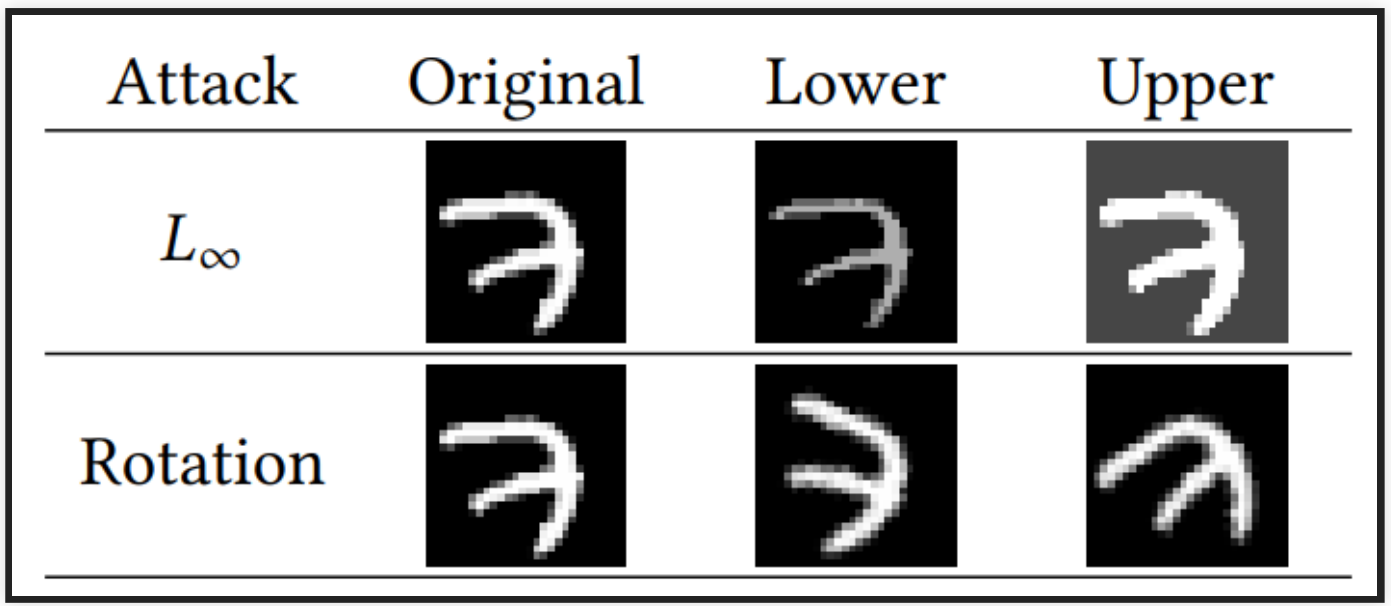


Image: Singh, Gagandeep, Timon Gehr, Markus Püschel, and Martin Vechev. "An abstract domain for certifying neural networks." Proceedings of the ACM on Programming Languages 3, no. POPL (2019): 1-30.

ROBUSTNESS AND DISTANCE

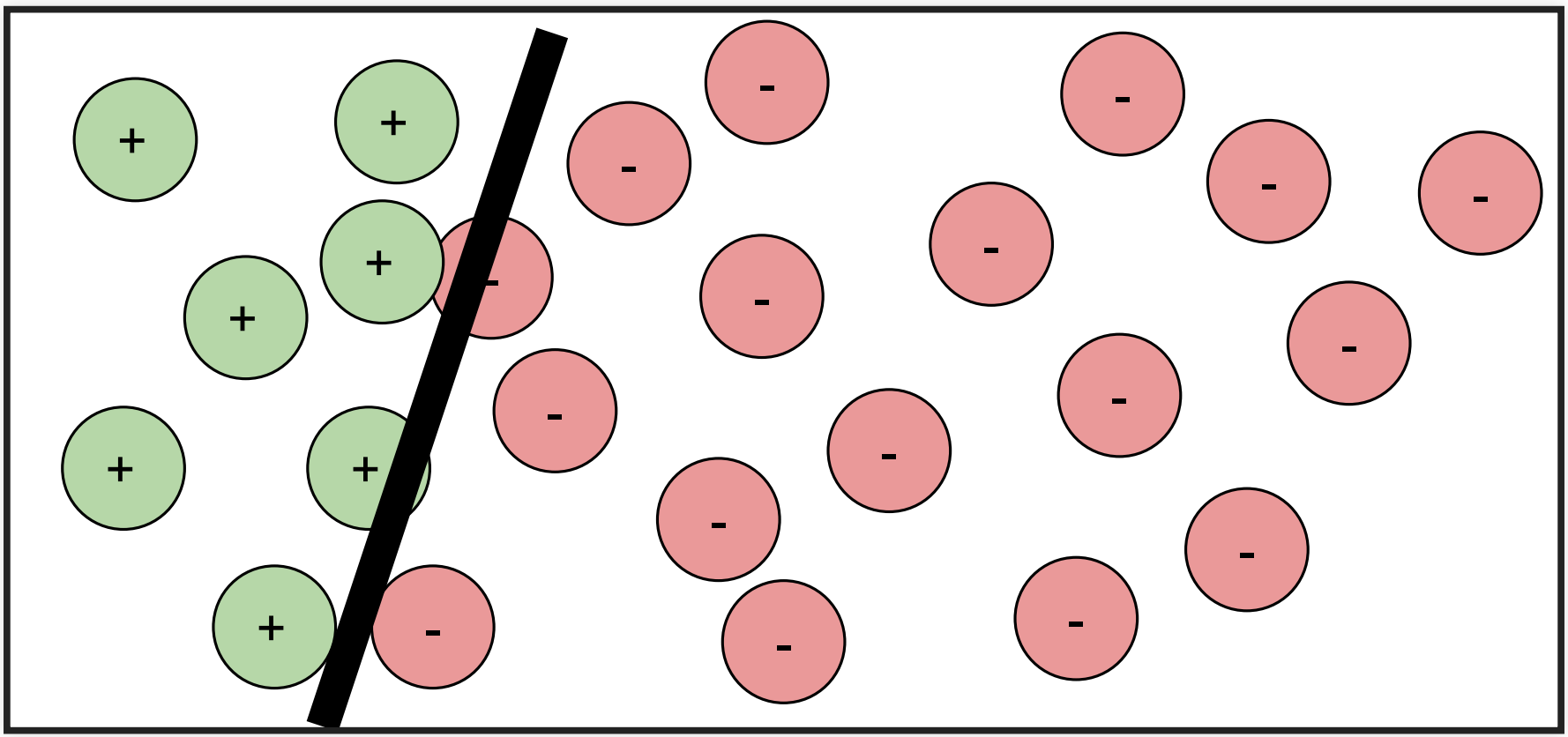
- For text:
  - insert words
  - replace words with synonyms
  - reorder text
- For tabular data:
  - change values
  - depending on feature extraction, small changes may have large effects
- ...

note, not all changes may be feasible or realistic; some changes are obvious to humans

realistically, a defender will not anticipate all attacks and corresponding distances

NO MODEL IS FULLY ROBUST

- Every useful model has at least one decision boundary (ideally at the real task decision boundary)
- Predictions near that boundary are not (and should not) be robust



ROBUSTNESS OF INTERPRETABLE MODELS

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE
IF age between 21–23 and 2–3 prior offenses THEN predict arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215.

DECISION BOUNDARIES IN PRACTICE

- With many models (especially deep neural networks), we do not understand the model's decision boundaries
- We are not confident that model decision boundaries align with task decision boundaries
    - The model's perception does not align well with human perception
- Models may pick up on parts of the input in surprising ways

ASSURING ROBUSTNESS

- Much research, many tools and approaches (especially for DNN)
- Formal verification
  - Constraint solving or abstract interpretation over computations in neuron activations
  - Conservative abstraction, may label robust inputs as not robust
  - Currently not very scalable
  - Example: Singh, Gagandeep, Timon Gehr, Markus Püschel, and Martin Vechev. "An abstract domain for certifying neural networks." Proceedings of the ACM on Programming Languages 3, no. POPL (2019): 1-30.
- Sampling
  - Sample within distance, compare prediction to majority prediction
  - Probabilistic guarantees possible (with many queries, e.g., 100k)
  - Example: Cohen, Jeremy M., Elan Rosenfeld, and J. Zico Kolter. "Certified adversarial robustness via randomized smoothing." In Proc. International Conference on Machine Learning, p. 1310--1320, 2019.

PRACTICAL USE OF ROBUSTNESS?

Current abilities: Detect for a given input whether neighboring inputs predict same result

PRACTICAL USE OF ROBUSTNESS

- Defense and safety mechanism at inference time
  - Check robustness of each prediction at runtime
  - Handle inputs with non-robust predictions differently (e.g. discard, low confidence)
  - Significantly raises cost of prediction (e.g. 100k model inferences or constraint solving at runtime)
- Testing and debugging
  - Identify training data near model's decision boundary (i.e., model robust around all training data?)
  - Check robustness on test data
  - Evaluate distance for adversarial attacks on test data

(most papers on the topic focus on techniques and evaluate on standard benchmarks like handwitten numbers, but do not discuss practical scenarios)

INCREASING MODEL ROBUSTNESS

- Augment training data with transformed versions of training data (same label) or with identified adversaries
- Defensive distillation: Second model trained on "soft" labels of first
- Input transformations: Learning and removing adversarial transformations
- Inserting noise into model to make adversarial search less effective, mask gradients
- Dimension reduction: Reduce opportunity to learn spurious decision boundaries
- Ensemble learning: Combine models with different biases

- Lots of research claiming effectiveness and vulnerabilities of various strategies

More details and papers: Rey Reza Wiyatno. Securing machine learning models against adversarial attacks. Element AI 2019

DETECTING ADVERSARIES

- Adversarial Classification: Train a model to distinguish benign and adversarial inputs
- Distribution Matching: Detect inputs that are out of distribution
- Uncertainty Thresholds: Measuring uncertainty estimates in the model for an input

More details and papers: Rey Reza Wiyatno. Securing machine learning models against adversarial attacks. Element AI 2019

ROBUSTNESS IN WEB STORE SCENARIO?

Antique Box Ugears, 3D Mechanical Treasure Models, Self-Assembling Precut Wooden Gift, DIY Craft Set

4.5 out of 5 stars 261

$41.90 ~~$44.90~~

FREE Delivery for Prime members

Only 1 left in stock - order soon.

More Buying Choices

$38.89 (44 new offers)

Ages: 14 years and up

ROKR 3D Wooden Puzzle for Adults-Mechanical Train Model Kits-Brain Teaser Puzzles-Vehicle Building Kits-Unique Gi...

4.0 out of 5 stars 44

$22.99

prime FREE One-Day

Get it **Tomorrow, Jul 26**

Ages: 14 years and up

Wooden Puzzles for Toddlers, Aitey Wooden Alphabet Number Puzzles Toddler Learning Puzzle Toys for Kids Ages 2 3 4 (Set of...

4.5 out of 5 stars 283

$23.99

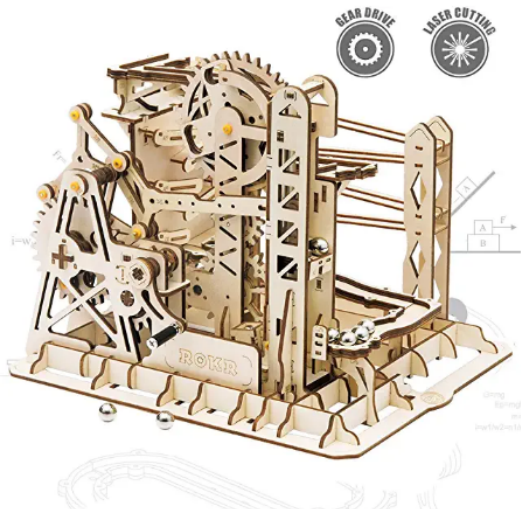
prime FREE One-Day

Get it **Tomorrow, Jul 26**

More Buying Choices

$22.79 (2 used & new offers)

Ages: 12 months and up



ROKR 3D Assembly Wooden Puzzle Brain Teaser Game Mechanical Gears Set Model Kit Marble Run Set Unique Craft...

4.5 out of 5 stars 172

$30.99

Unidragon Wooden Jigsaw Puzzles - Unique Shape Jigsaw Pieces Best Gift for Adults and Kids Alluring Fox 7 × 9.2 in (18 ...

5.0 out of 5 stars 13

$40.99 $50.99

KINGZHUO Hexagon Tangram Classic Handmade Wooden Puzzle for Children and Adults Challenging Puzzles Brain...

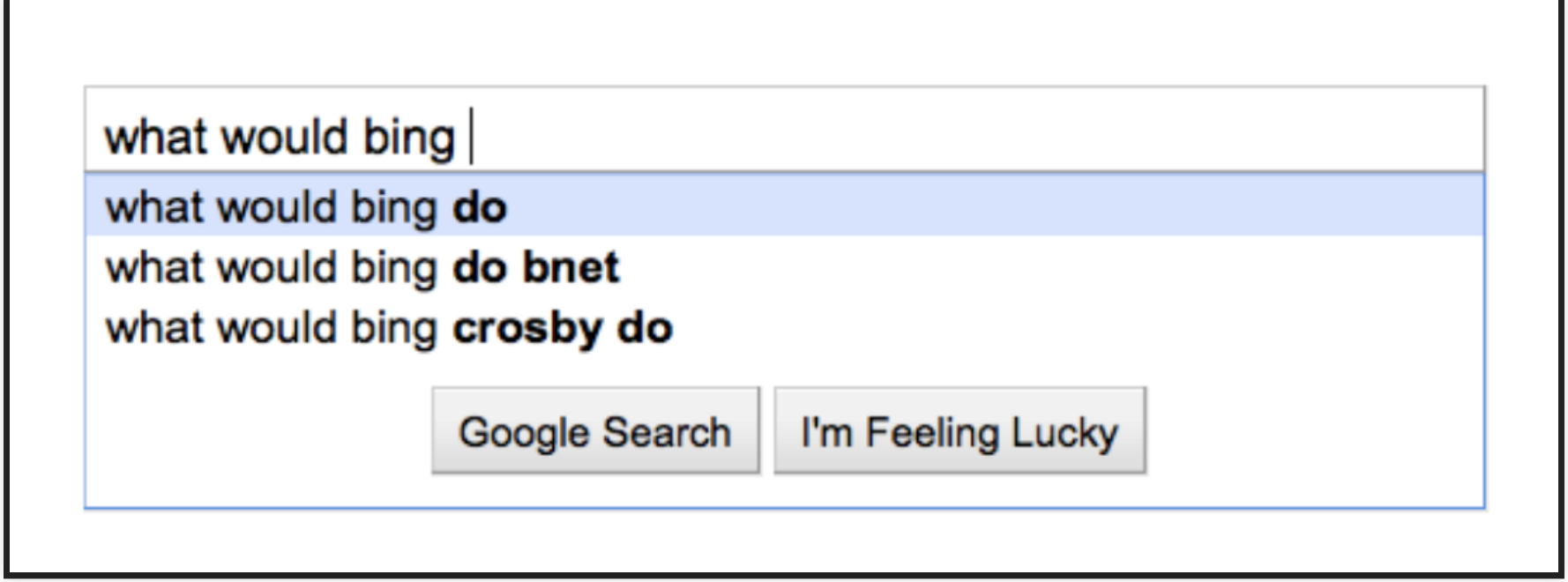
4.5 out of 5 stars 263

$8.98

IP AND PRIVACY

RYAN SINGEL 02.01.11 02:31 PM

Google Catches Bing Copying; Microsoft Says 'So What?'


what would bing
what would bing do
what would bing do bnet
what would bing crosby do
Google Search
I'm Feeling Lucky

INTELLECTUAL PROPERTY PROTECTION

- Depending on deployment scenario
- May have access to model internals (e.g. in app binary)
- May be able to repeatedly query model's API
    - build surrogate model (*inversion attack*)
    - cost per query? rate limit? abuse detection?
- Surrogate models ease other forms of attacks

RYAN SINGEL 03.12.10 02:48 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine.

Speaker notes

"an in-the-closet lesbian mother sued Netflix for privacy invasion, alleging the movie-rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest."

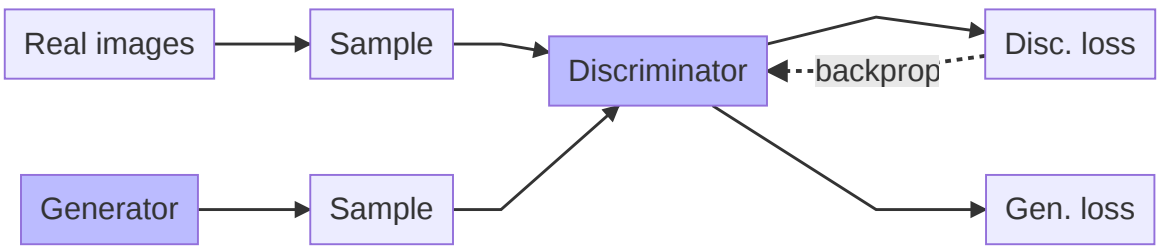
Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. 2015.

PRIVACY

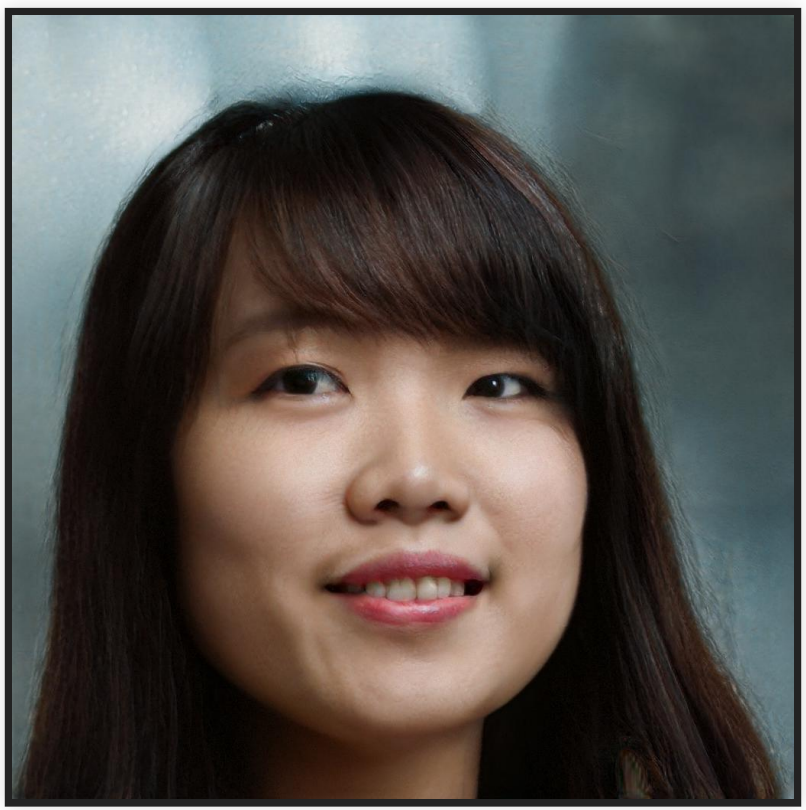
- Various privacy issues about acquiring and sharing training data, e.g.,
  - DeepMind receiving NHS data on 1.6 million patients without their consent
  - Chest X-rays not shared for training because they may identify people
  - Storage of voice recordings of voice assistants
- Model inversion attacks: Models contain information from training data, may recover information from training data
  - Extract DNA from medical model
  - Extract training images from face recognition model

Kyle Wiggers. AI has a privacy problem, but these techniques could fix it. Venturebeat, 2019

GENERATIVE ADVERSARIAL NETWORKS



PROTOTYPICAL INPUTS WITH GANS



Speaker notes

- Generative adversarial networks: 2 models, one producing samples and one discriminating real from generated samples
    - Learn data distribution of training data
    - Produce prototypical images, e.g. private jets
    - Deep fakes

PRIVACY PROTECTION STRATEGIES

- Federated learning (local models, no access to all data)
- Differential privacy (injecting noise to avoid detection of individuals)
- Homomorphic encryption (computing on encrypted data)

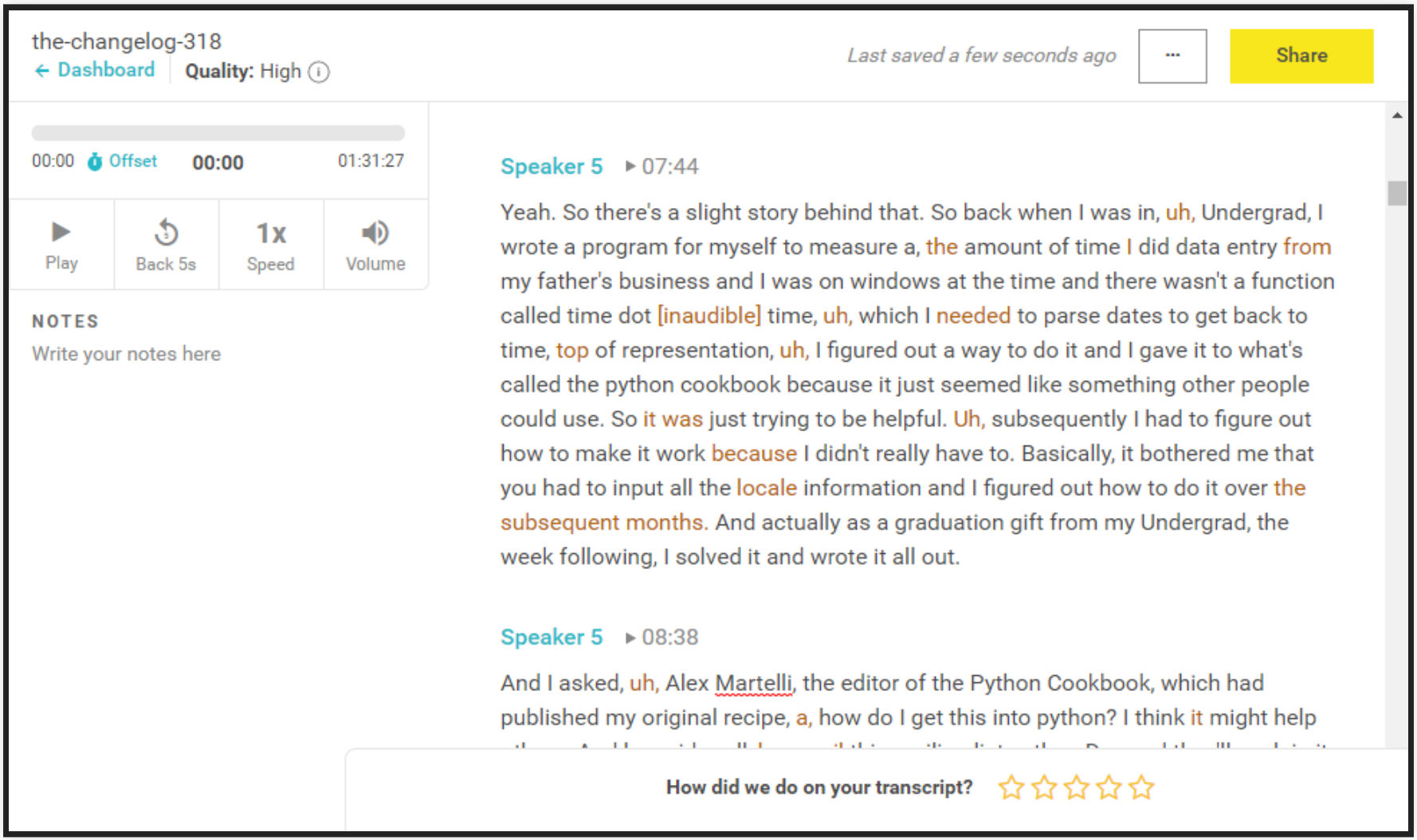
- Much research
- Some adoption in practice (Android keyboard, Apple emoji)
- Usually accuracy or performance tradeoffs

Kyle Wiggers. AI has a privacy problem, but these techniques could fix it. Venturebeat, 2019

SECURITY AT THE SYSTEM LEVEL

security is more than model robustness

defenses go beyond hardening models


the-changelog-318
← Dashboard
Quality: High
Last saved a few seconds ago
...
Share
00:00
Offset
00:00
01:31:27
Play
Back 5s
1x
Speed
Volume
NOTES
Write your notes here
Speaker 5 ▸ 07:44
Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.
Speaker 5 ▸ 08:38
And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help
How did we do on your transcript?

Speaker notes

At a price of $.25 per min it iss possibly not economical to train a surrogate model or inject bad telemetry

Jeffrey N. Fritz Top Contributor: Amazon Echo VINE VOICE

★★★★★ **Fun to Build Detailed Steam Engine Model**

Reviewed in the United States on September 17, 2019

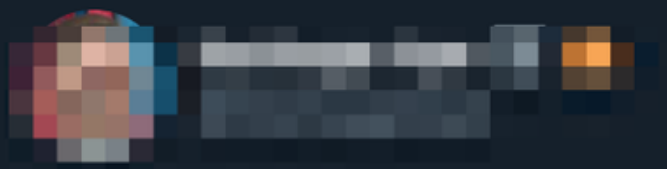
Verified Purchase

The wooden steam engine model made by ROKR is called a “3D Puzzle Kit.” I completed without great difficulty it over the span of two days. The model is made from laser cut wood parts that need to be punched out (carefully) from eight large flat wooden panels. The individual parts are labeled by board and number. There is no glue used, the pieces are all pressed together (again carefully.)

The model is fairly large at 14 inches long, 9 1/2 inches high and 2 inches wide. It weighs almost 3

Speaker notes

Raise the price of wrong inputs



5G Corona

5G is responsible for the Coronavirus.

Get the facts about COVID-19

1:52 AM · Jun 7, 2020 · Twitter for iPhone

35 Retweets **919** Likes

Speaker notes

source https://www.buzzfeednews.com/article/pranavdixit/twitter-5g-coronavirus-conspiracy-theory-warning-label

Shadow banning also fits here

Copyright strike basics

This content is about copyright strikes. If you're looking for information about Community Guidelines strikes, which are different than copyright strikes, go to our Community Guideline strikes basics.

If you get a copyright strike, that means your video has been taken down from YouTube because a copyright owner sent us a complete and valid legal request asking us to do so. When a copyright owner formally notifies us that you don't have their permission to post their content on the site, we take down your upload to comply with copyright law.

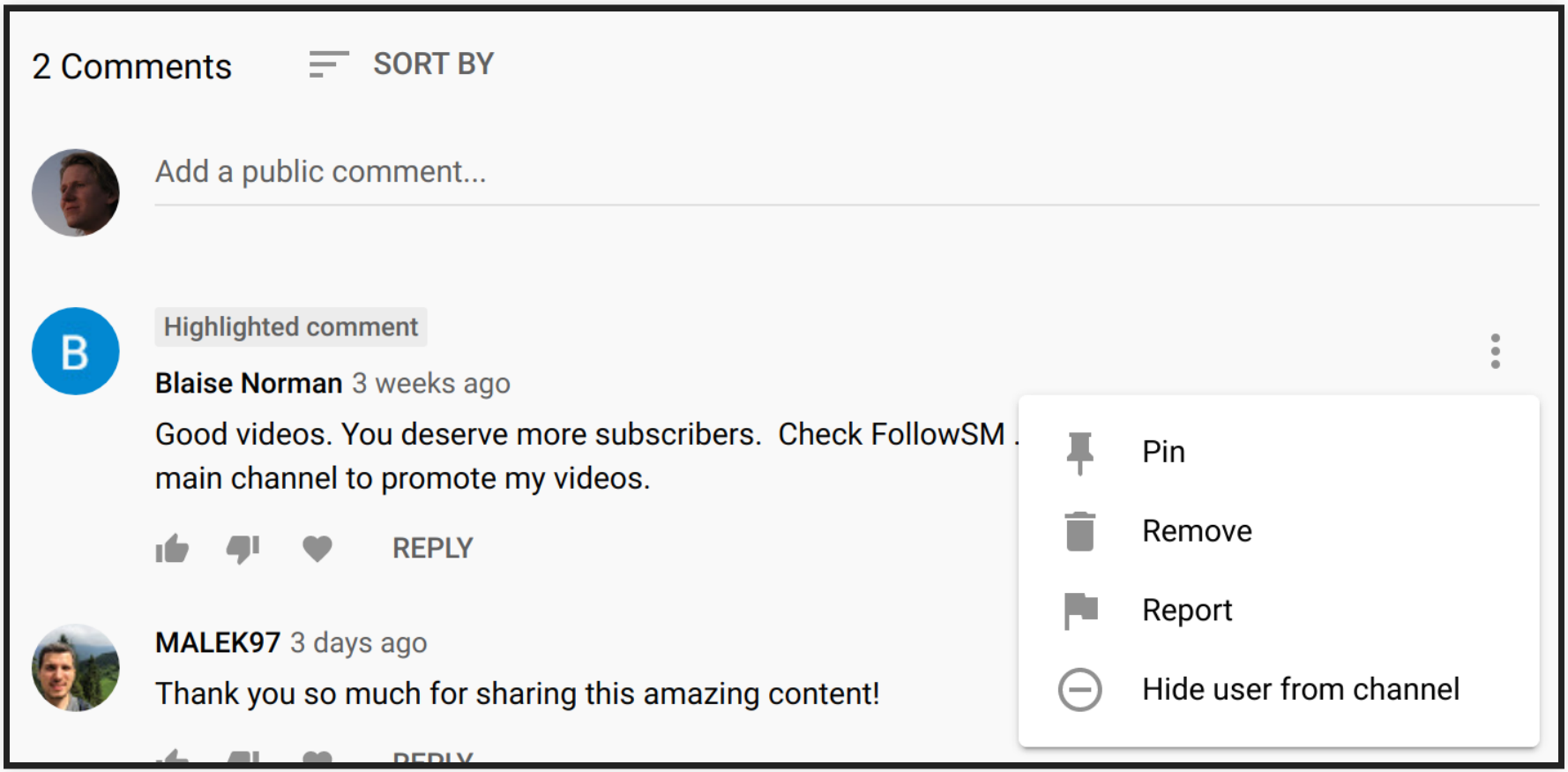
A video can only have one copyright strike at a time. Keep in mind that videos can be removed from the site for reasons other than copyright. Also, Content ID claims don't result in a strike.

Deleting a video with a strike won't resolve your strike. Learn how to resolve a copyright strike below.

What happens when you get a copyright strike

Speaker notes

Block user of suspected attack to raise their cost, burn their resources


2 Comments
SORT BY
Add a public comment...
Highlighted comment
Blaise Norman 3 weeks ago
Good videos. You deserve more subscribers. Check FollowSM .
main channel to promote my videos.
REPLY
MALEK97 3 days ago
Thank you so much for sharing this amazing content!
Pin
Remove
Report
Hide user from channel

Speaker notes

Reporting function helps to crowdsource detection of malicious content and potentially train a future classifier (which again can be attacked)

5

As per mroutis on the DVC Discord server:

1. `dvc unprotect` the file; this won't be necessary if you don't use `symlink` or `hardlink` caching, but it can't hurt.
2. Remove the .dvc file
3. If you need to delete the cache entry itself, run `dvc gc`, or look up the MD5 in `data.dvc` and manually remove it from `.dvc/cache`.

Edit -- there is now an issue on their Github page to add this to the manual: https://github.com/iterative/dvc.org/issues/625

share improve this answer follow

edited Sep 17 '19 at 13:12

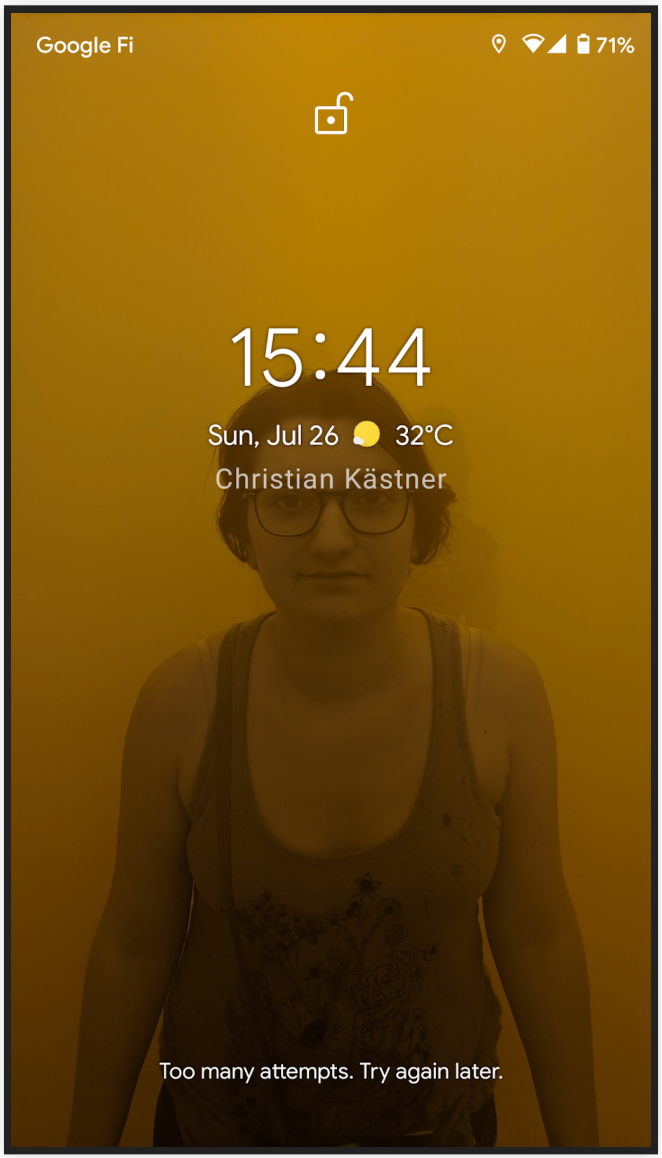
answered Sep 17 '19 at 4:18

shadowtalker

7,356 ● 2 ● 28 ● 65

Speaker notes

See reputation system


Google Fi
71%
15:44
Sun, Jul 26
32°C
Christian Kästner
Too many attempts. Try again later.

Speaker notes

Block system after login attempts with FaceID or fingerprint

SYSTEM DESIGN QUESTIONS

- What is one simple change to make the system less interesting to abusers?
- Increase the cost of abuse, limit scale?
- Decrease the value of abuse?
- Trust established users over new users?
- Reliance on ML to combat abuse?
- Incidence response plan?

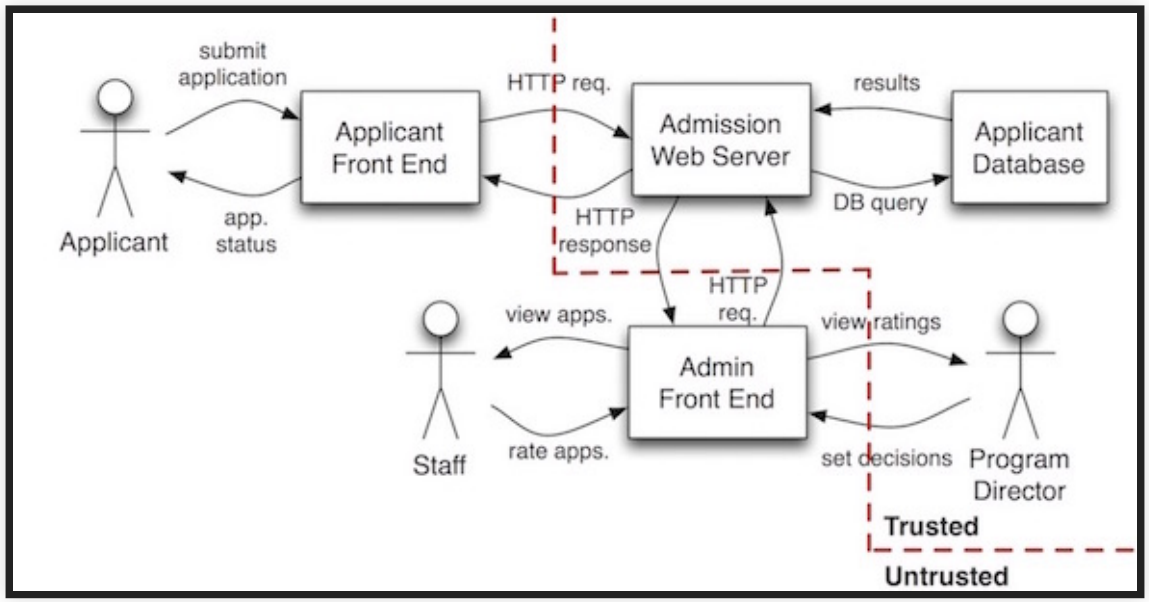
Examples for web shop/college admissions AI?

THREAT MODELING

THREAT MODELING

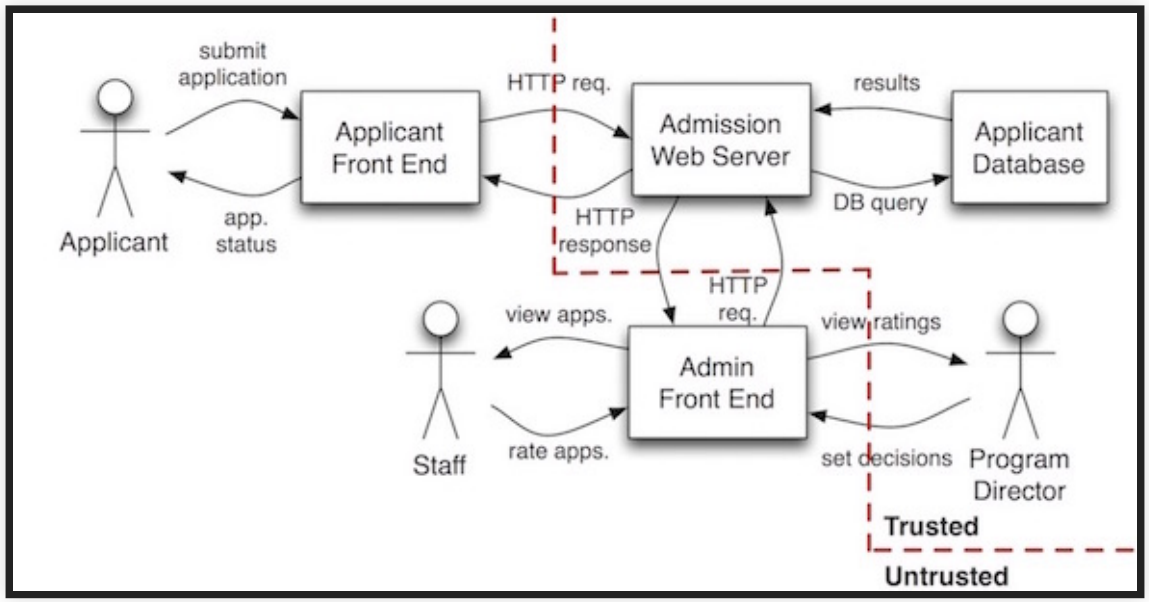
- Attacker Profile
  - **Goal**: What is the attacker trying to achieve?
  - **Capability**:
  - Knowledge: What does the attacker know?
  - Actions: What can the attacker do?
  - Resources: How much effort can it spend?
  - **Incentive**: Why does the attacker want to do this?
- Understand how the attacker can interact with the system
- Understand security strategies and their scope
- **Identify security requirements**

ATTACKER CAPABILITY



- Capabilities depends on system boundary & its exposed interfaces
- Use an architecture diagram to identify attack surface & actions
- Example: Garmin/College admission
    - Physical: Break into building & access server
    - Cyber: Send malicious HTTP requests for SQL injection, DoS attack
    - Social: Send phishing e-mail, bribe an insider for access

ARCHITECTURE DIAGRAM FOR THREAT MODELING



- Dynamic and physical architecture diagram
- Describes system components and users and their interactions
- Describe thrust boundaries

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

- Systematic inspection to identifying threats & attacker actions
  - For each component/connection, enumerate & identify potential threats using checklist
  - e.g., Admission Server & DoS: Applicant may flood it with requests
  - Derive security requirements
- Tool available (Microsoft Threat Modeling Tool)
- Popularized by Microsoft, broadly used in practice

OPEN WEB APPLICATION SECURITY PROJECT

OWASP Top 10 Application Security Risks - 2017

A1:2017-Injection

Injection flaws, such as SQL, NoSQL, OS, and LDAP injection, occur when untrusted data is sent to an interpreter as part of a command or query. The attacker's hostile data can trick the interpreter into executing unintended commands or accessing data without proper authorization.

A2:2017-Broken Authentication

Application functions related to authentication and session management are often implemented incorrectly, allowing attackers to compromise passwords, keys, or session tokens, or to exploit other implementation flaws to assume other users' identities temporarily or permanently.

A3:2017-Sensitive Data Exposure

Many web applications and APIs do not properly protect sensitive data, such as financial, healthcare, and PII. Attackers may steal or modify such weakly protected data to conduct credit card fraud, identity theft, or other crimes. Sensitive data may be compromised without extra protection, such as encryption at rest or in transit, and requires special precautions when exchanged with the browser.

A4:2017-XML External Entities (XXE)

Many older or poorly configured XML processors evaluate external entity references within XML documents. External entities can be used to disclose internal files using the file URI handler, internal file shares, internal port scanning, remote code execution, and denial of service attacks.

A5:2017-Broken Access Control

Restrictions on what authenticated users are allowed to do are often not properly enforced. Attackers can exploit these flaws to access unauthorized functionality and/or data, such as access other users' accounts, view sensitive files, modify other users' data, change access rights, etc.

A6:2017-Security Misconfiguration

Security misconfiguration is the most commonly seen issue. This is commonly a result of insecure default configurations, incomplete or ad hoc configurations, open cloud storage, misconfigured HTTP headers, and verbose error messages containing sensitive information. Not only must all operating systems, frameworks, libraries, and applications be securely configured, but they must be patched/upgraded in a timely fashion.

- OWASP: Community-driven source of knowledge & tools for web security

THREAT MODELING LIMITATIONS

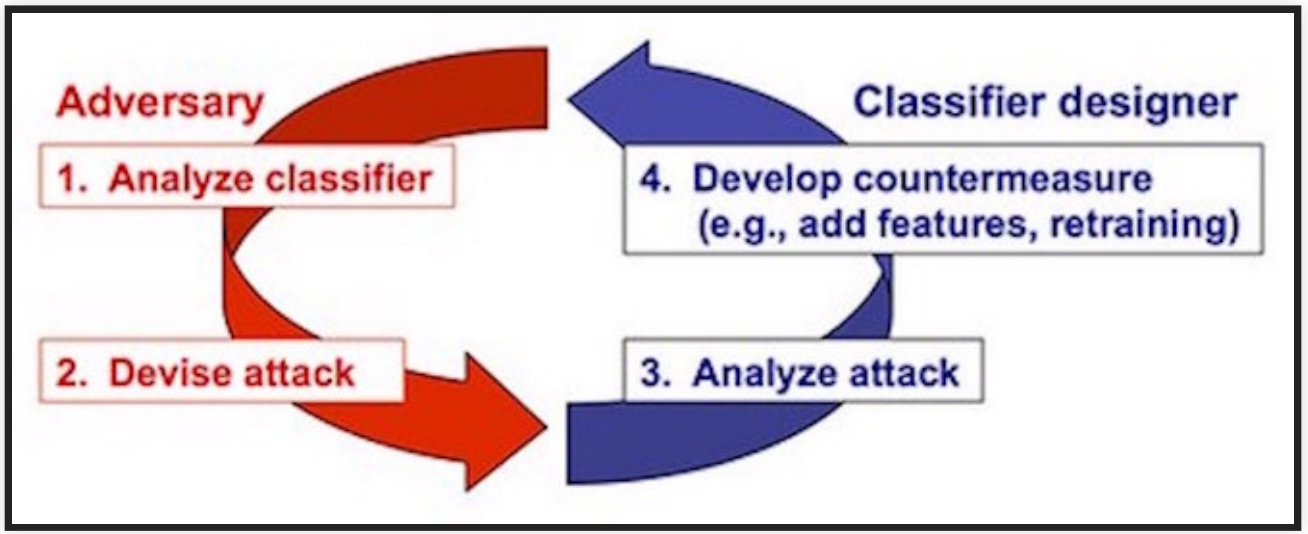
- Manual approach, false positives and false negatives
- May end up with a long list of threats, not all of them relevant
- Need to still correctly implement security requirements
- False sense of security: STRIDE does not imply completeness!

THREAT MODELING ADJUSTMENTS FOR AI?

THREAT MODELING ADJUSTMENTS FOR AI?

- Explicitly consider origins, access, and influence of all relevant data (training, prediction input, prediction result, model, telemetry)
- Consider AI-specific attacks
    - Poisoning attacks
    - Evasion attacks
    - Surrogate models
    - Privacy leaks
    - ...

STATE OF ML SECURITY



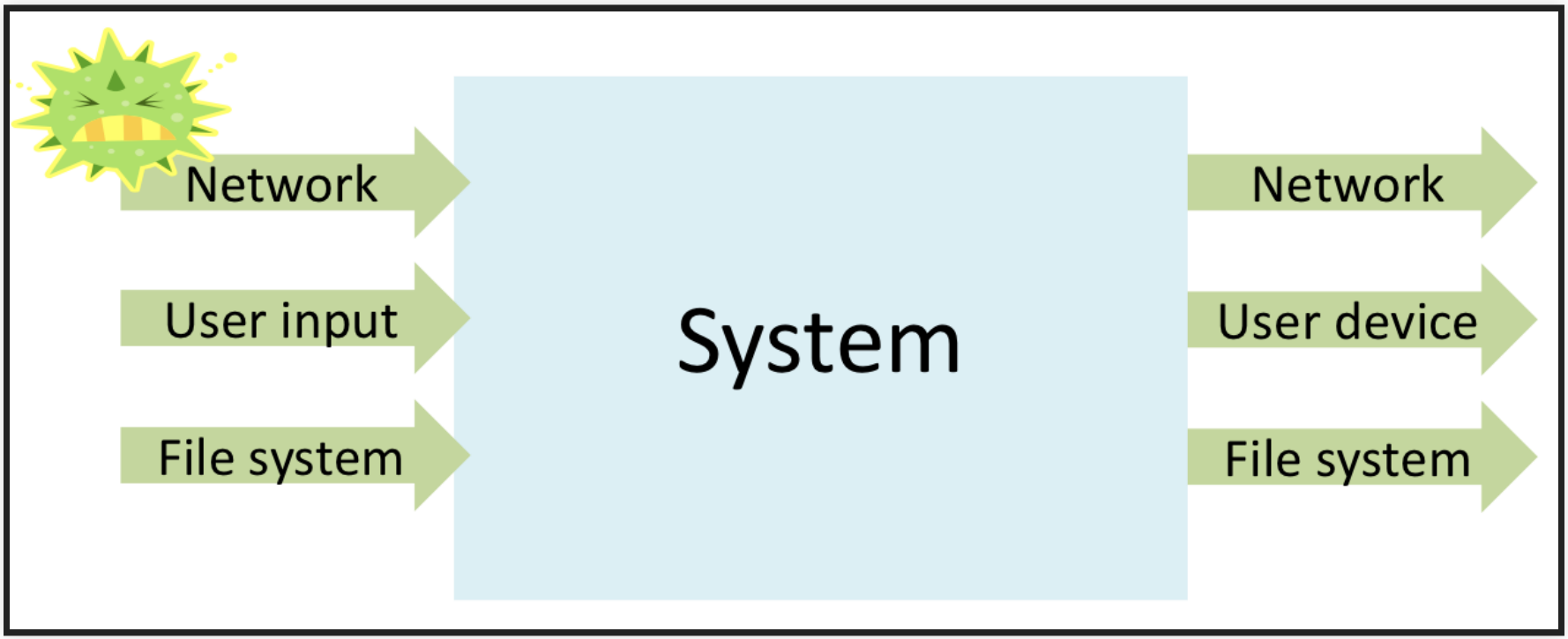
- On-going arms race (mostly among researchers)
  - Defenses proposed & quickly broken by noble attacks
- *Assume ML component is likely vulnerable*
  - Design your system to minimize impact of an attack
- Remember: There may be easier ways to compromise system
  - e.g., poor security misconfiguration (default password), lack of encryption, code vulnerabilities, etc.,

DESIGNING FOR SECURITY

SECURE DESIGN PRINCIPLES

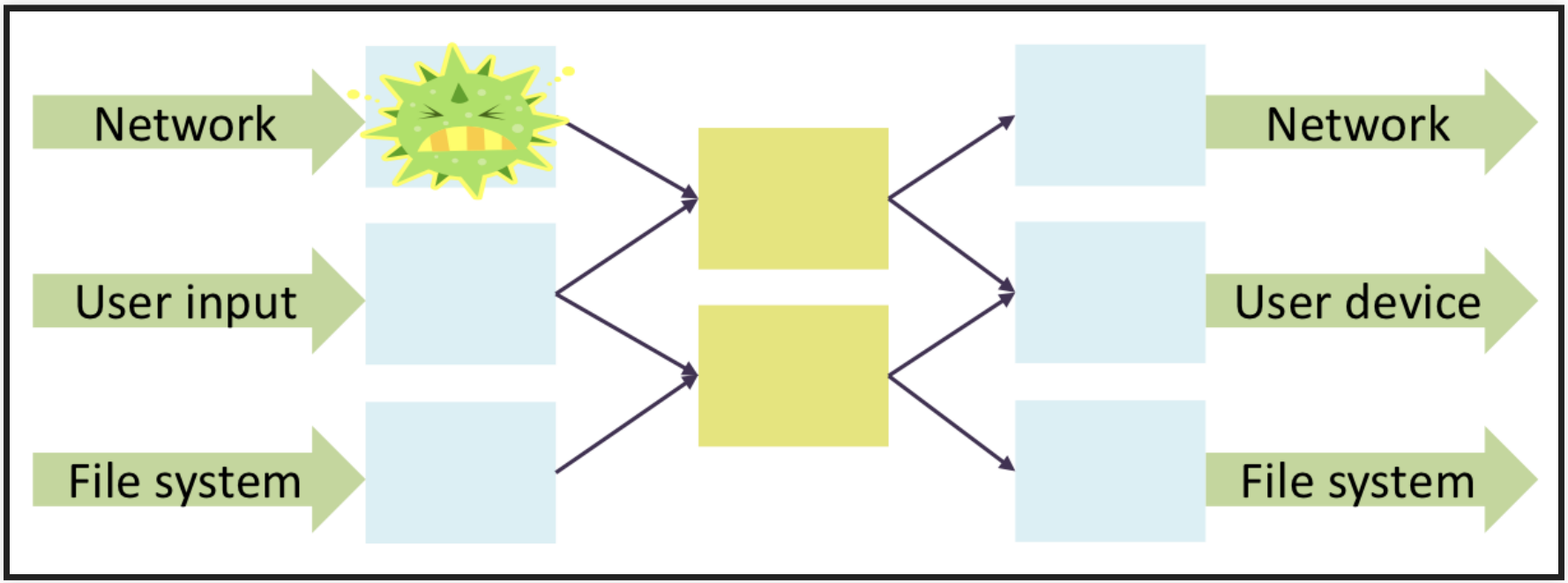
- Principle of Least Privilege
  - A component should be given the minimal privileges needed to fulfill its functionality
  - Goal: Minimize the impact of a compromised component
- Isolation
  - Components should be able to interact with each other no more than necessary
  - Goal: Reduce the size of trusted computing base (TCB)
  - TCB: Components responsible for establishing a security requirement(s)
  - If any of TCB compromised => security violation
  - Conversely, a flaw in non-TCB component => security still preserved!
  - In poor system designs, TCB = entire system

MONOLITHIC DESIGN



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

COMPARTMENTALIZED DESIGN

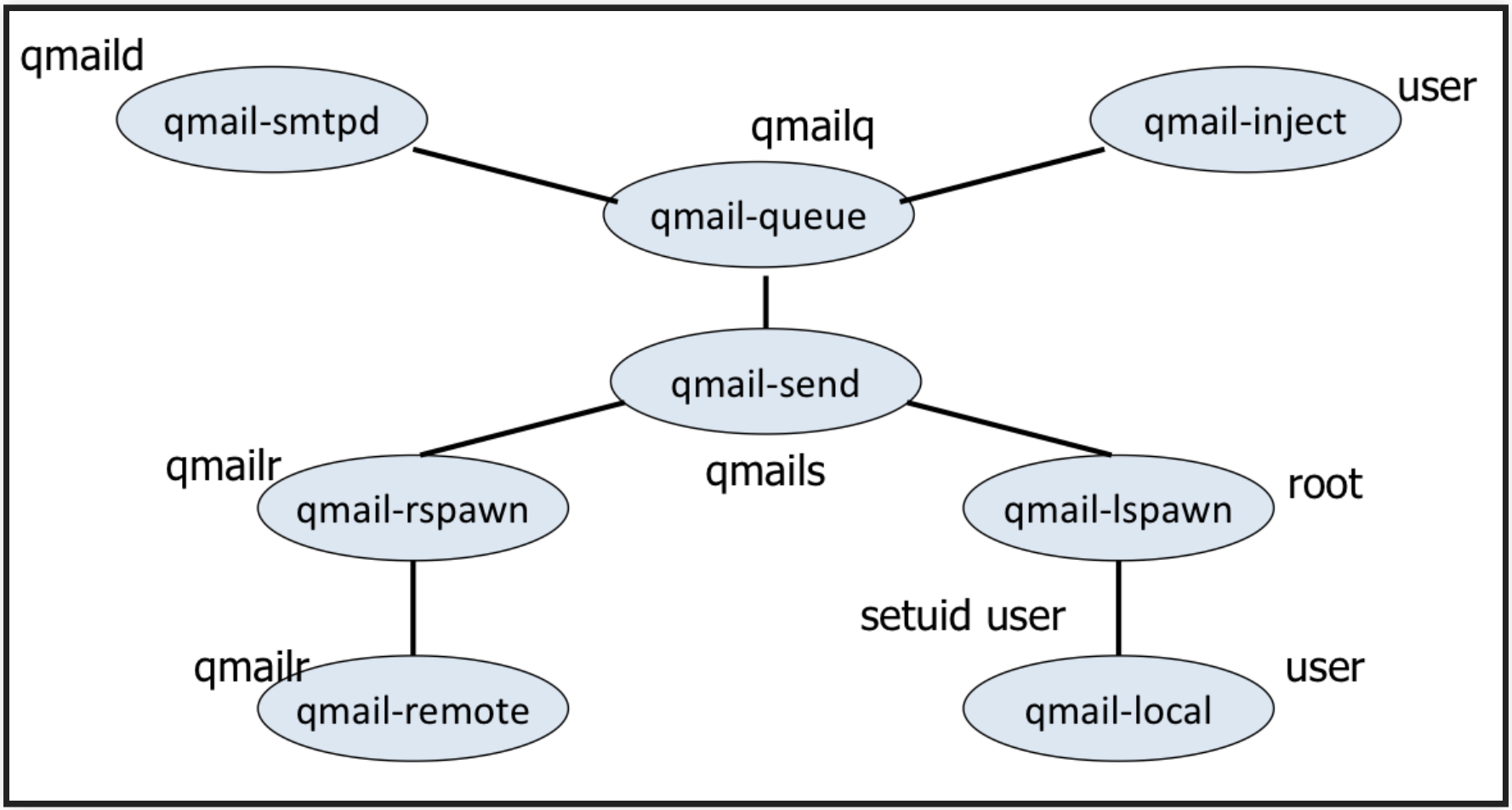


Flaw in one component => Limited impact on the rest of the system!

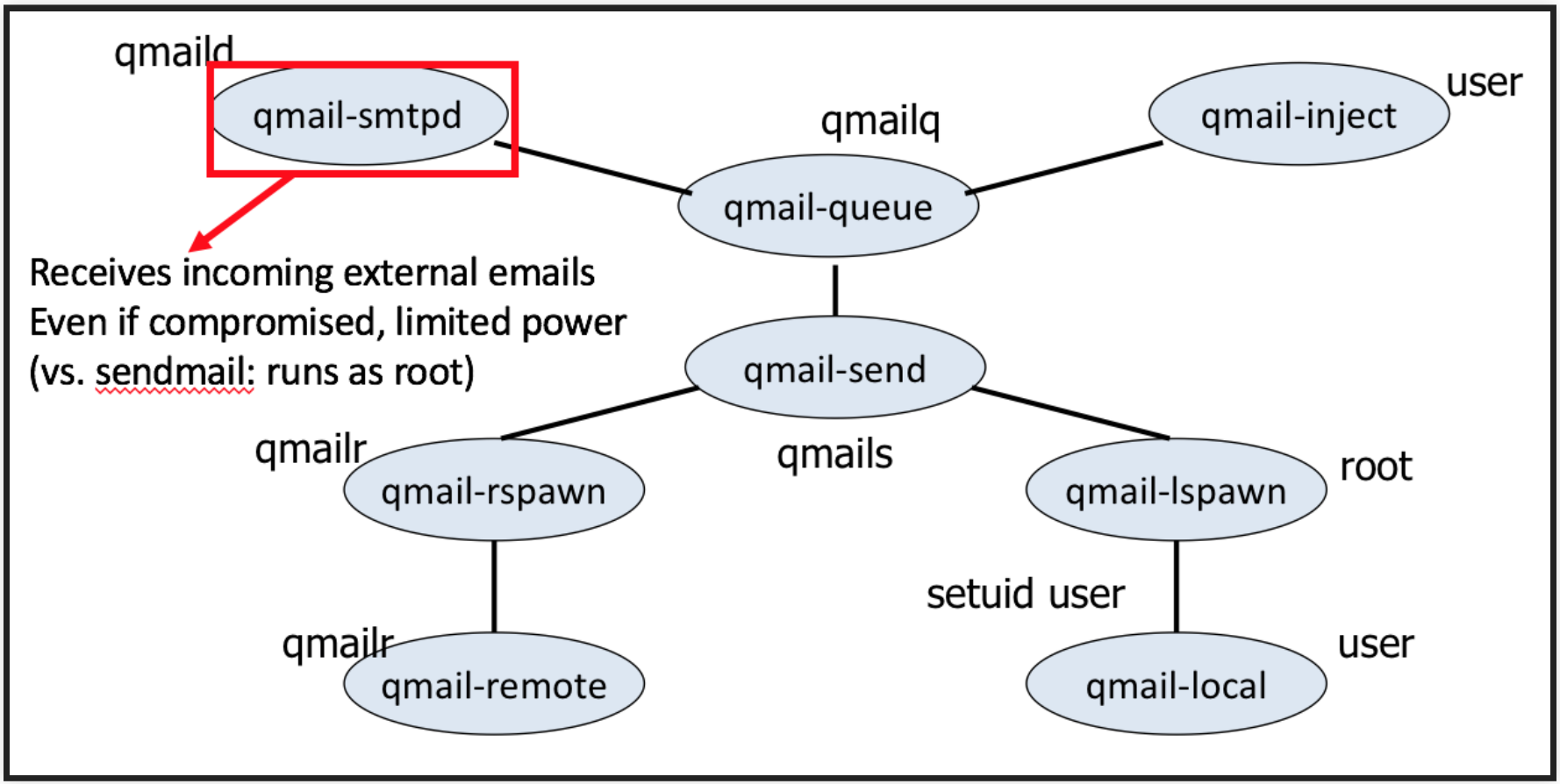
NON-ML EXAMPLE: MAIL CLIENT

- Requirements
  - Receive & send email over external network
  - Place incoming email into local user inbox files
- Sendmail
  - Monolithic design; entire program runs as UNIX root
  - Historical source of many vulnerabilities
- Qmail: “Security-aware” mail agent
  - Compartmentalized design
  - Isolation based on OS process isolation
  - Separate modules run as separate “users” (UID)
  - Mutually distrusting processes
  - Least privilege
  - Minimal privileges for each UID; access to specific resources (files, network sockets, …)
  - Only one “root” user (with all privileges)

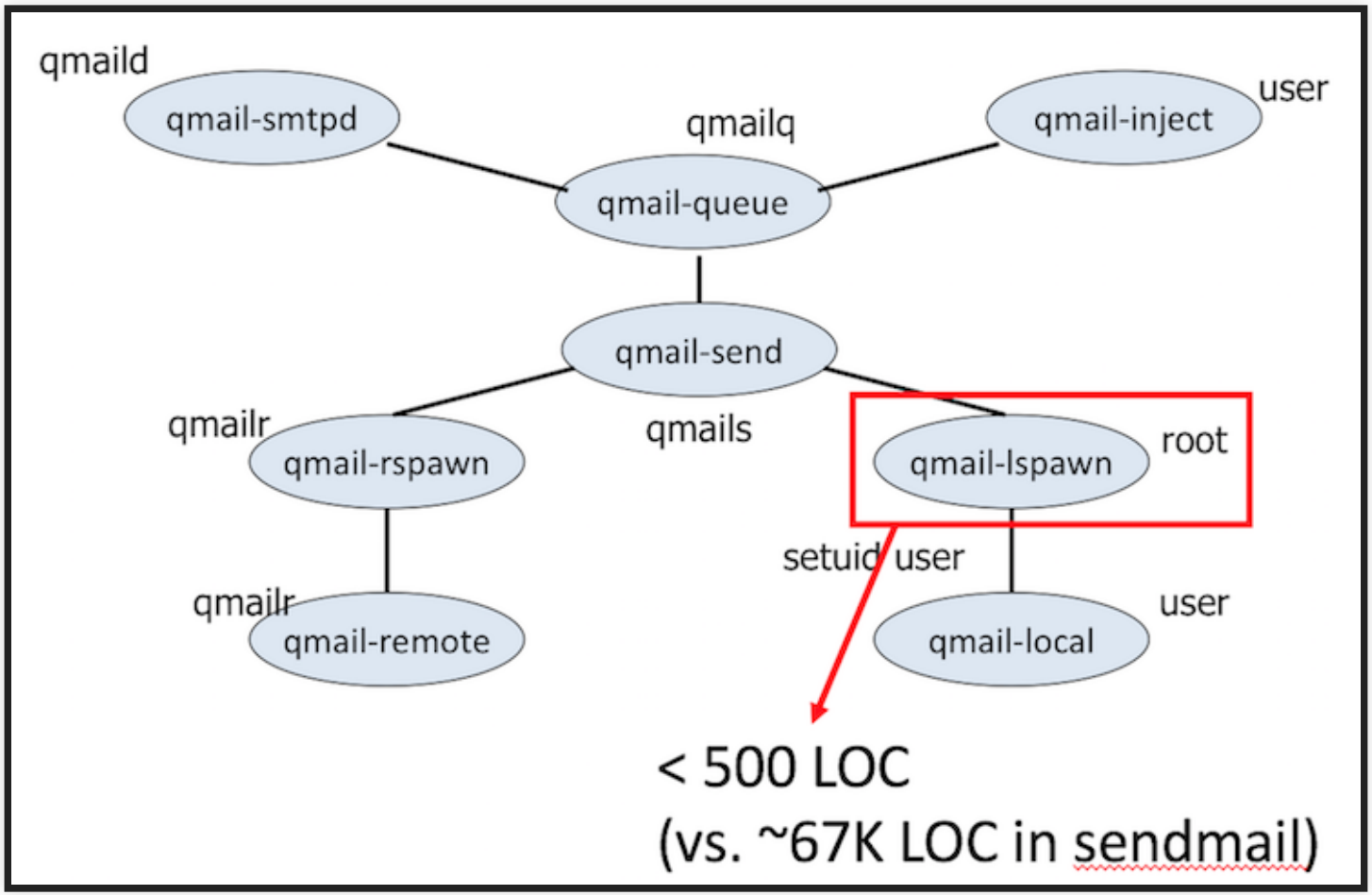
QMAIL ARCHITECTURE



QMAIL ARCHITECTURE



QMAIL ARCHITECTURE



AI FOR SECURITY

builtin TECH JOBS

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

Alyssa Schroer

July 12, 2019 Updated: July 15, 2020

By the year 2021, cybercrime losses will

MANY DEFENSE SYSTEMS USE MACHINE LEARNING

- Classifiers to learn malicious content
  - Spam filters, virus detection
- Anomaly detection
  - Identify unusual/suspicious activity, eg. credit card fraud, intrusion detection
  - Often unsupervised learning, e.g. clustering
- Game theory
  - Model attacker costs and reactions, design countermeasures
- Automate incidence response and mitigation activites
  - Integrated with 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 AI-ENABLED SYSTEMS TOO

- AI/ML component one part of a larger system
- Consider entire system, from training to telemetry, to user interface, to pipeline automation, to monitoring
- AI-based security solutions can be attacked themselves


EQUIFAX

Speaker notes

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

SUMMARY

- Security requirements: Confidentiality, integrity, availability
- ML-specific attacks on training data, telemetry, or the model
    - Poisoning attack on training data to influence predictions
    - Evasion attacks to shape input data to achieve intended predictions (adversarial learning)
    - Leaks of model IP (surrogates) and training data
- Robustness as a measure of prediction stability w.r.t to input perturbations; verification possible
- Security design at the system level
    - Influence costs and gains
    - Security mechanisms beyond the model
- Threat modeling to identify security requirements
- AI can be used for defense (e.g. anomaly detection)
- **Key takeaway**: Adopt a security mindset! Assume all components may be vulnerable in one way or another. Design your system to explicitly reduce the impact of potential attacks