Security of ML Systems

Thorsten Eisenhofer

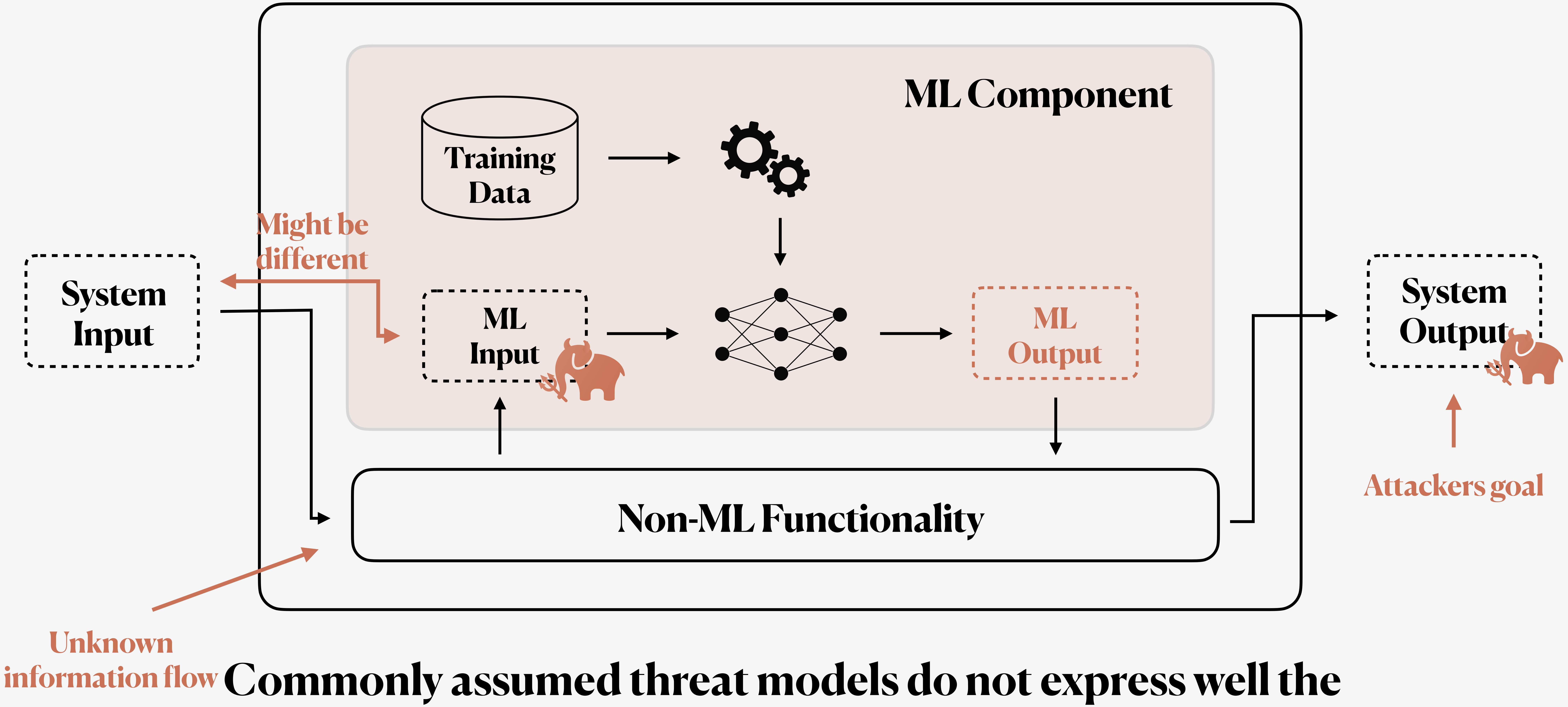
CASA
Cyber Security in the Age
of Large-Scale Adversaries

RUHR
UNIVERSITÄT
BOCHUM
RUB

The development of AI is as fundamental as the creation of the microprocessor, the personal computer, the Internet, and the mobile phone. [...] Entire industries will reorient around it. Businesses will distinguish themselves by how well they use it.

Bill Gates — March'23

ML Systems



Commonly assumed threat models do not express well the goals, capabilities and knowledge of real-world adversaries

Feature-problem-space attacks

Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, and Konrad Rieck
No more Reviewer #2: Subverting Automatic Paper Reviewer Assignment using Adversarial Learning
USENIX Security Symposium, 2023

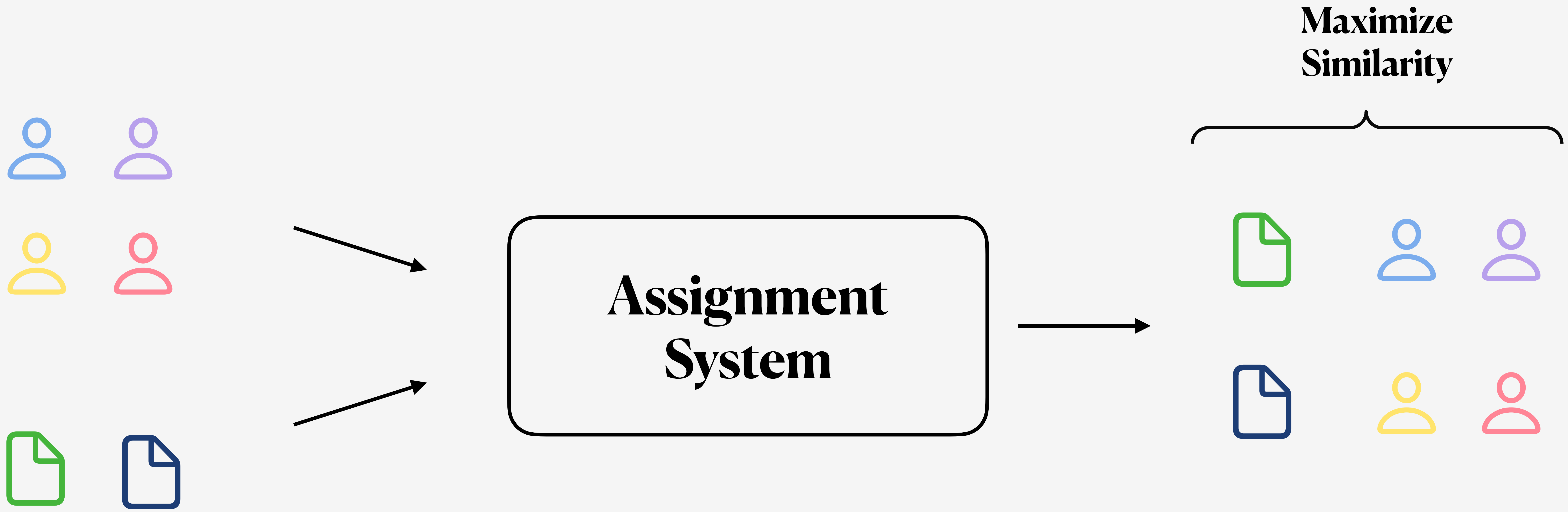
Domain-specific priors

Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, and Thorsten Holz
Dompteur: Taming Audio Adversarial Examples
USENIX Security Symposium, 2021

ML security beyond the model

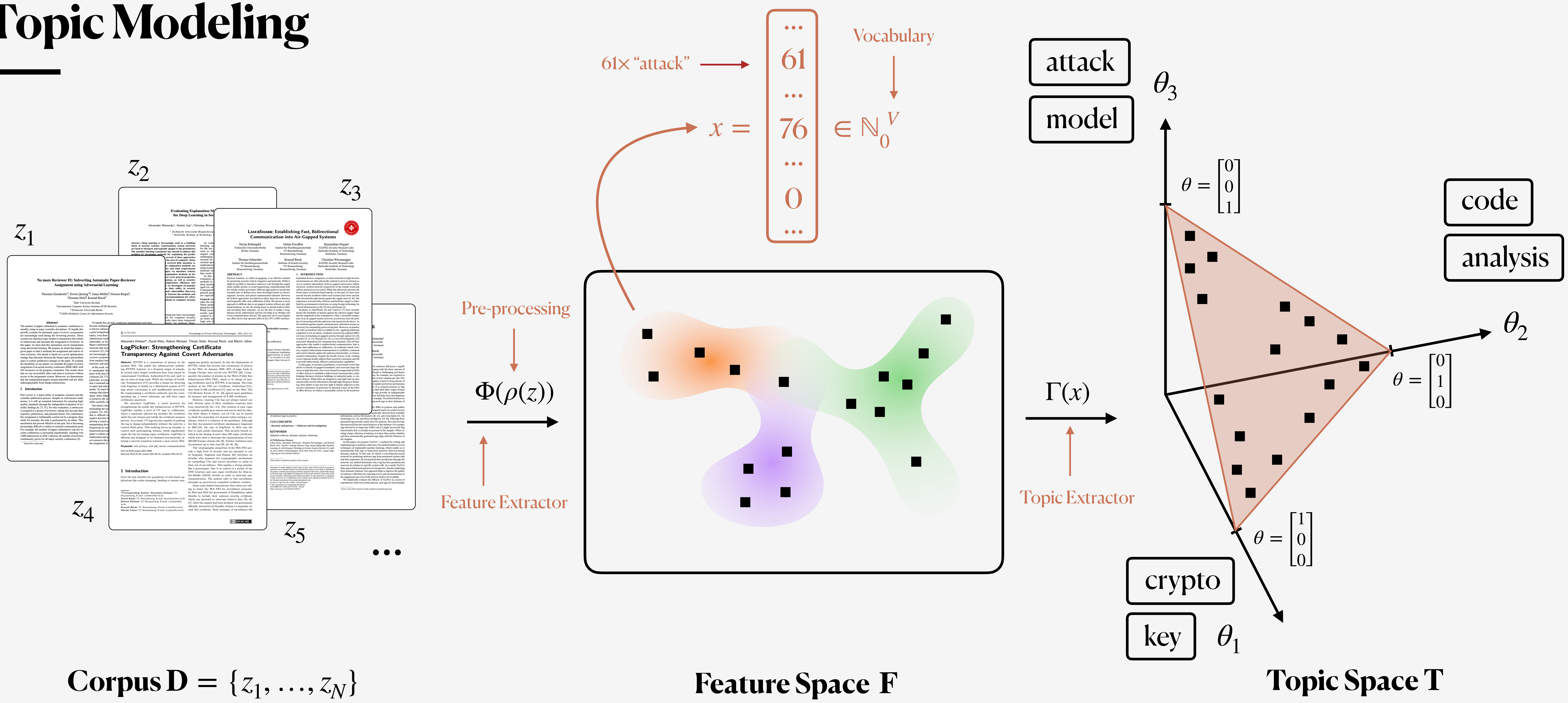
Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot
Verifiable and Provably Secure Machine Unlearning
In Submission

Assignment Systems



Use ML to distill submissions and reviewer expertise

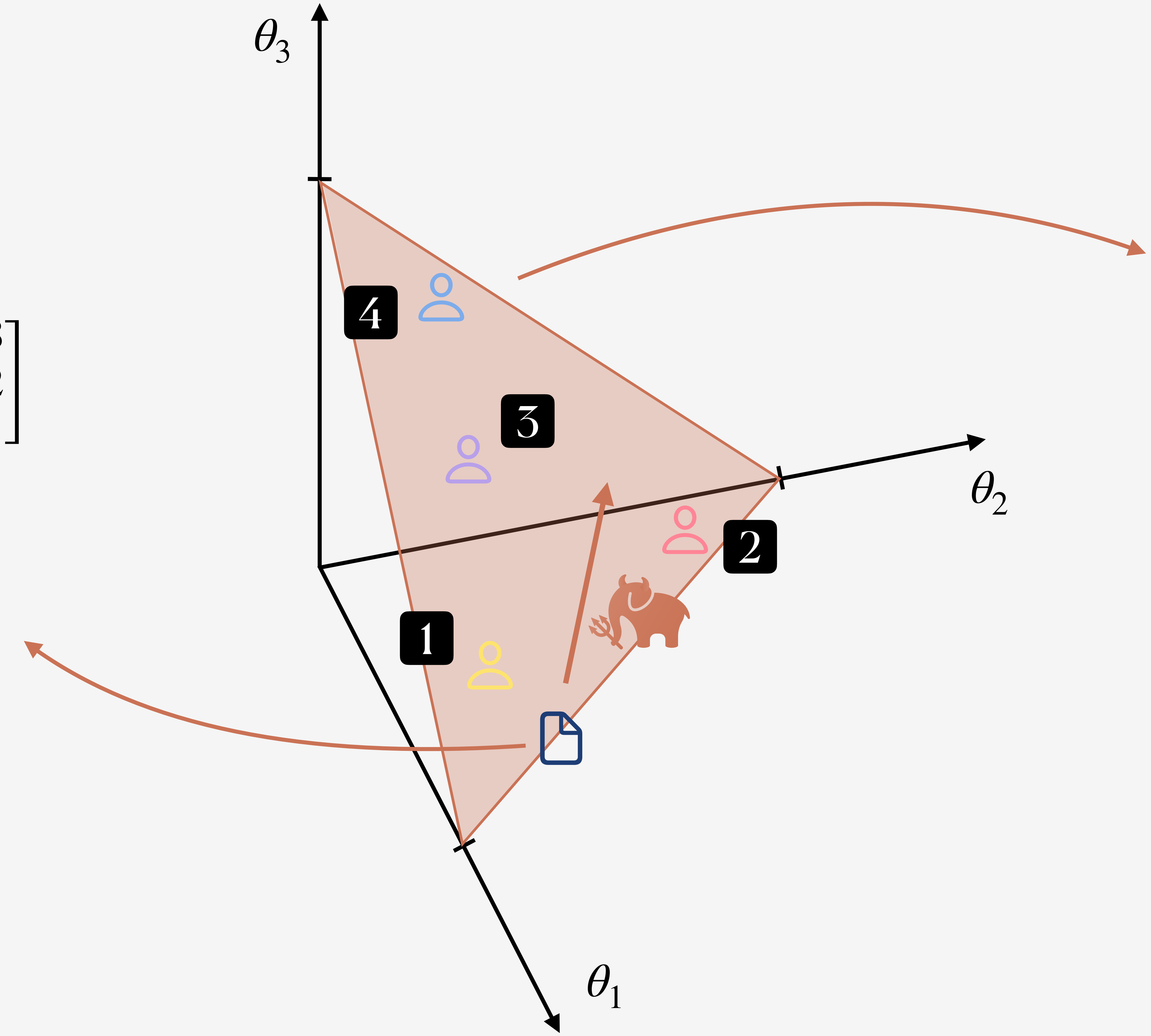
Topic Modeling



Topic Modeling

Submission Topics $\theta = \begin{bmatrix} 0.8 \\ 0.2 \\ 0 \end{bmatrix}$

80%	key	crypto
20%	code	analysis
0%	attack	model

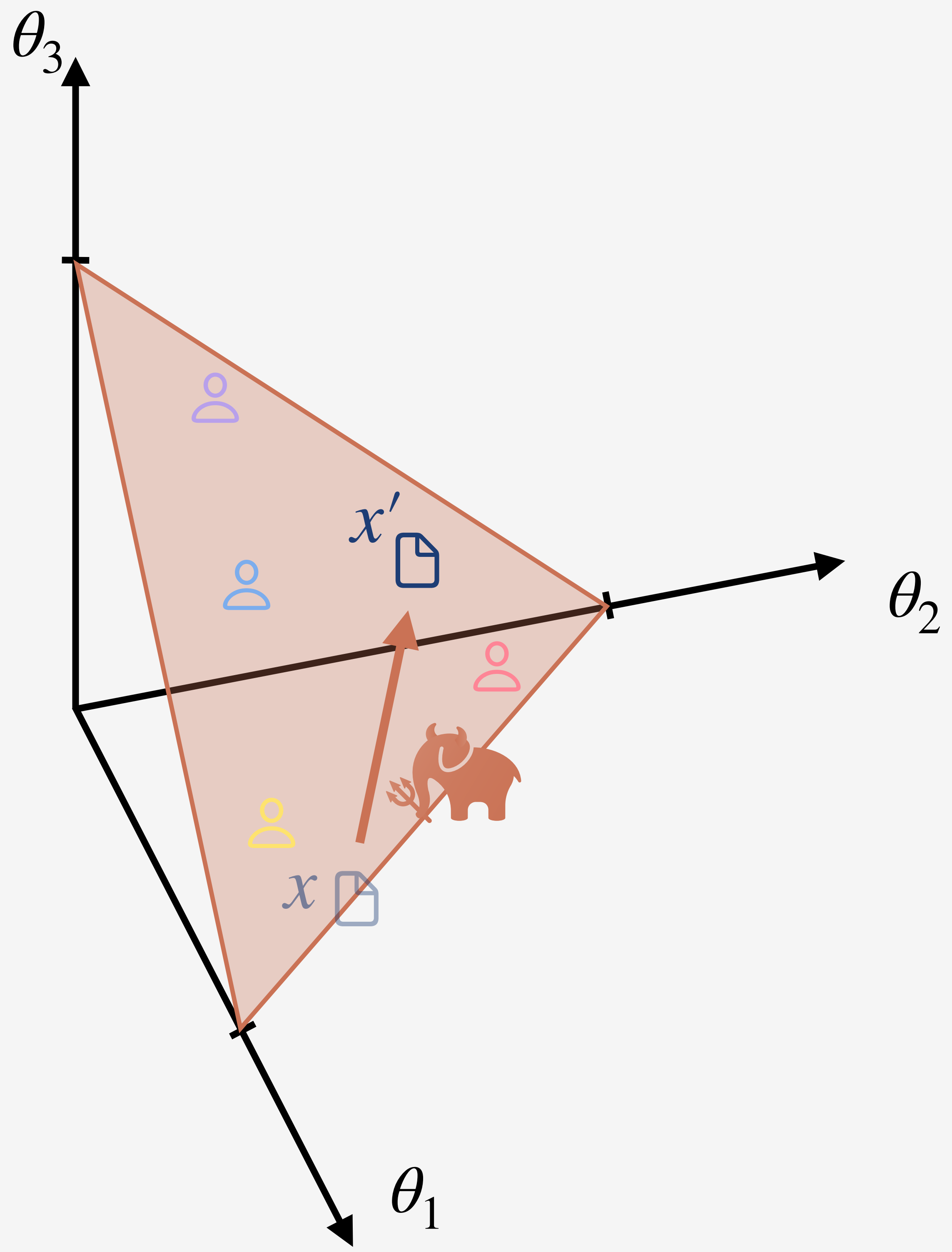


Reviewer Topics $\theta = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$

10%	key	crypto
20%	code	analysis
70%	attack	model

Goal: Manipulate submission to pick our own reviewers

Feature-space Attack



Let R_{sel} be the set of selected reviewer

Let R_{rej} be the set of rejected reviewer

Find $\delta \in \mathbf{F}$ s.t. $x' := x + \delta$ fullfils

1. $r \in R_{\text{sel}} \Rightarrow r \in R_{x'}$
2. $r \in R_{\text{rej}} \Rightarrow r \notin R_{\text{x}'}, \forall r \in \mathbf{R}$

Target assignment

subject to $\|\delta\|_1 \leq L_1^{\max}$ **and** $\|\delta\|_\infty \leq L_\infty^{\max}$

Total modifications per paper

Total modifications per word

Need to project changes back into the problem space!

Problem-space Attack

Transform input file to add/remove words: $\omega : \mathbf{Z} \to \mathbf{Z}, z \mapsto z'$

Format-/ and encoding-level

Hidden Box

Homoglyphs ← a ≠ а (u+0061, u+0430)

Text-level

Reference addition

Synonyms

Language models

Spelling mistakes

Chain several transformations

$$\Omega = \omega_k \circ \ldots \circ \omega_2 \circ \omega_1$$

Constraints

$$\Omega(z) \models \Upsilon \Leftrightarrow \Omega(z)$$

is plausible and semantic correct

Feature-problem-space Attack

Feature-problem-space attack

$$r \in R_{\text{sel}} \Rightarrow r \in R_{x'}$$

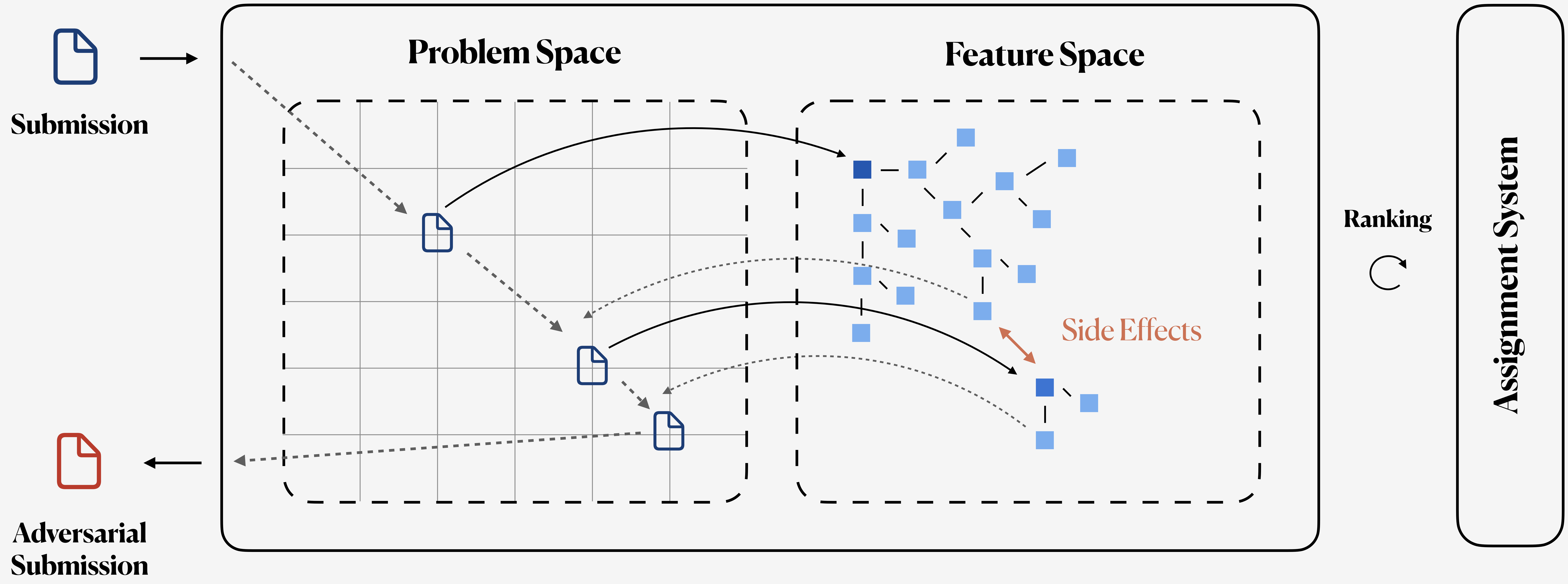
$$r \in R_{\text{rej}} \Rightarrow r \notin R_{\mathrm{X}'}, \forall r \in \mathbf{R}$$

subject to $\delta_1 \leq L_1^{\max}$ **and** $\delta_\infty \leq L_\infty^{\max}$

with $\mathbf{x} = \Phi(\rho(z))$, $\mathbf{x}' = \Phi(\rho(\Omega(z)))$ **and** $\delta = (\mathbf{x}' - \mathbf{x})$

We design a hybrid search strategy for this

Hybrid Search Strategy



Results

White-box setting

Remove *any* initially assigned reviewers

Scale to choose *all* of the assigned reviewer

Black-box setting

Use only *public knowledge* about a conference (e.g., the PC)

Success rate of up 90% to *select* and up to 81% to *reject* a reviewer

User study

Tested visible transformations

Detection precision of only 33% with a recall of only 8%

Feature-problem-space attacks

Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, and Konrad Rieck
No more Reviewer #2: Subverting Automatic Paper Reviewer Assignment using Adversarial Learning
USENIX Security Symposium, 2023

Domain-specific priors

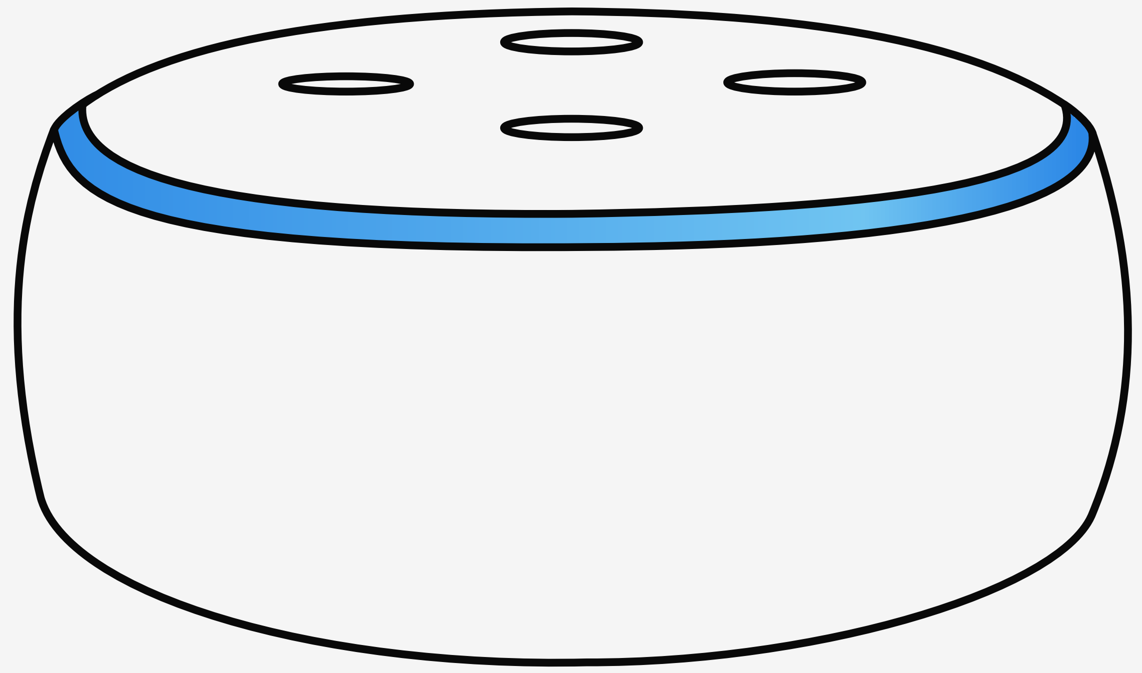
Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, and Thorsten Holz
Dompteur: Taming Audio Adversarial Examples
USENIX Security Symposium, 2021

ML security beyond the model

Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot
Verifiable and Provably Secure Machine Unlearning
In Submission

Voice Assistants

Raw Audio Wave

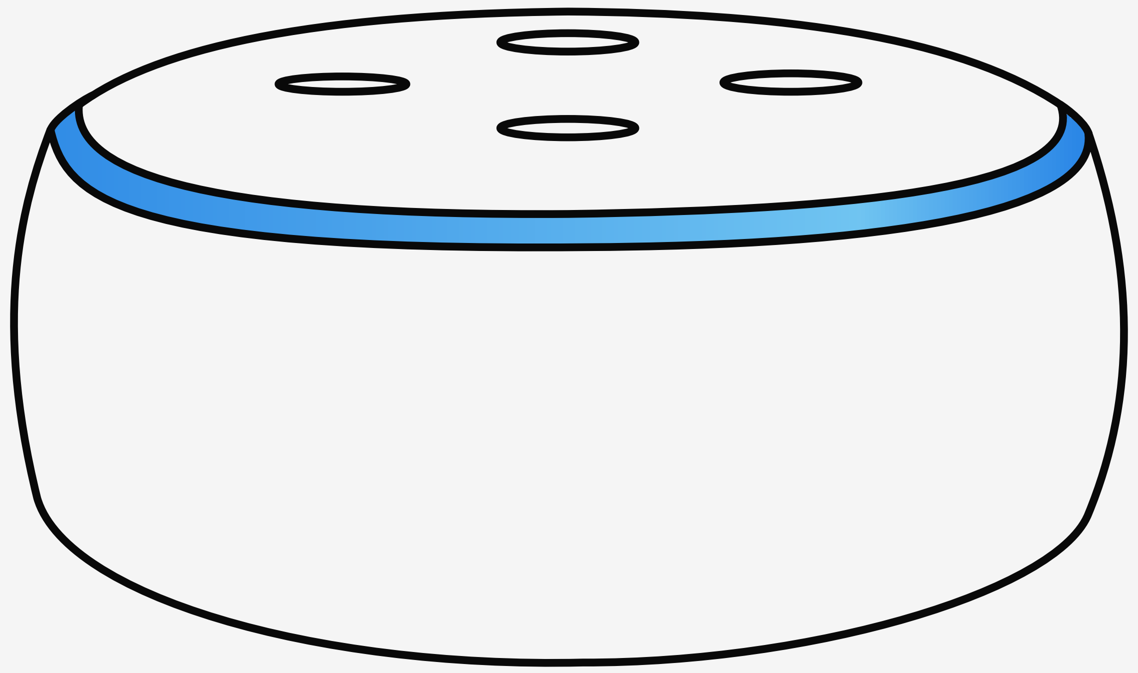


Voice Assistant

BIDS TOTALING SIX
HUNDRED FIFTY ONE
MILLION DOLLARS
WERE SUBMITTED

Transcription

Voice Assistants

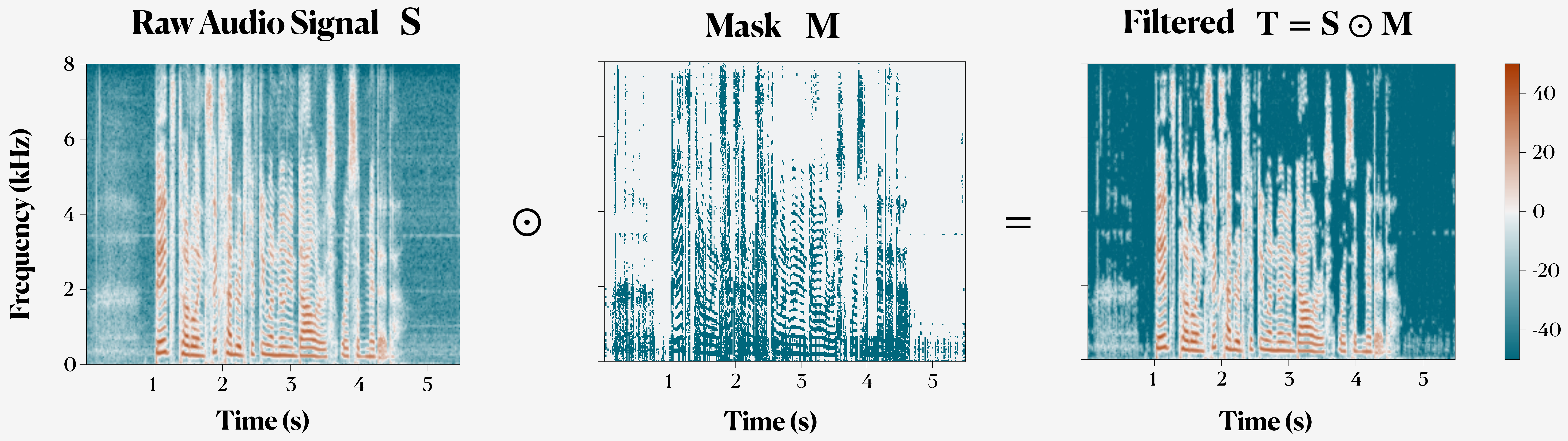


DEACTIVATE
SECURITY CAMERA
AND UNLOCK
FRONT DOOR

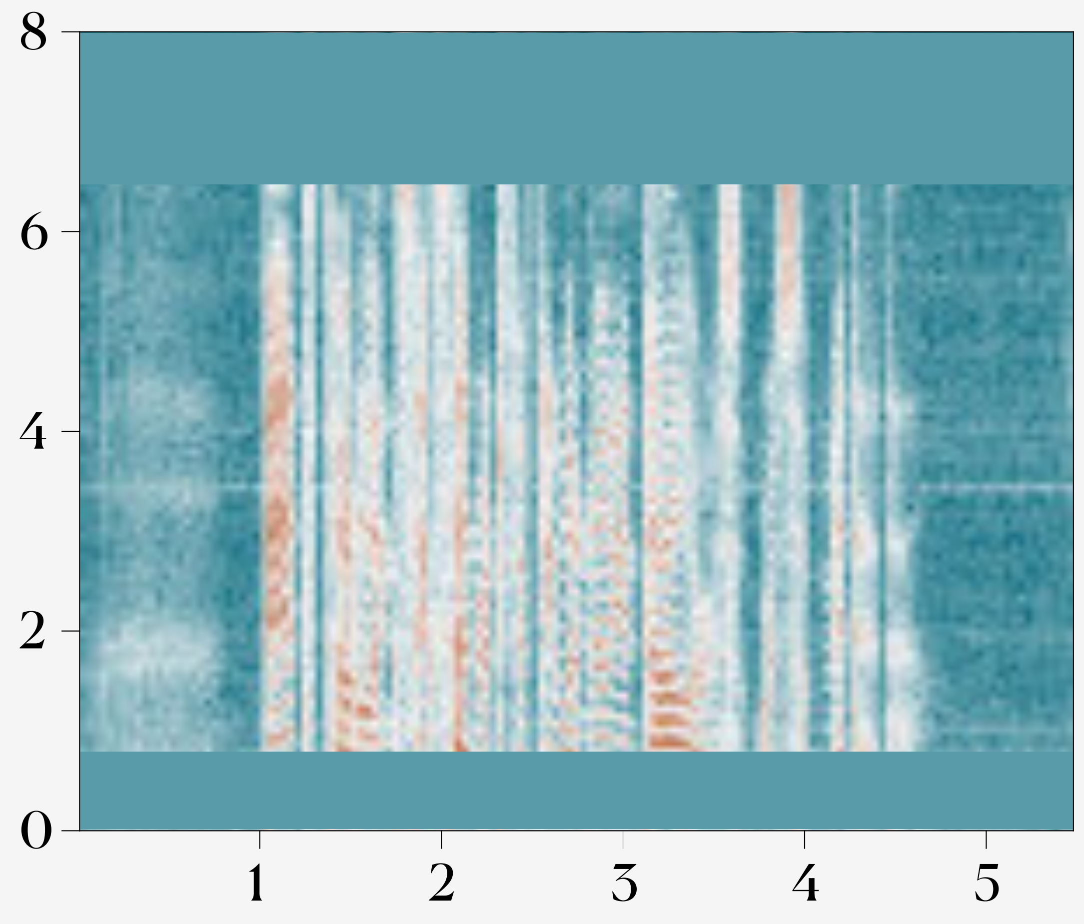
Raw Audio Wave → Voice Assistant → Transcription

Integrate knowledge on the human auditory system to improve robustness

Psychoacoustic Filtering

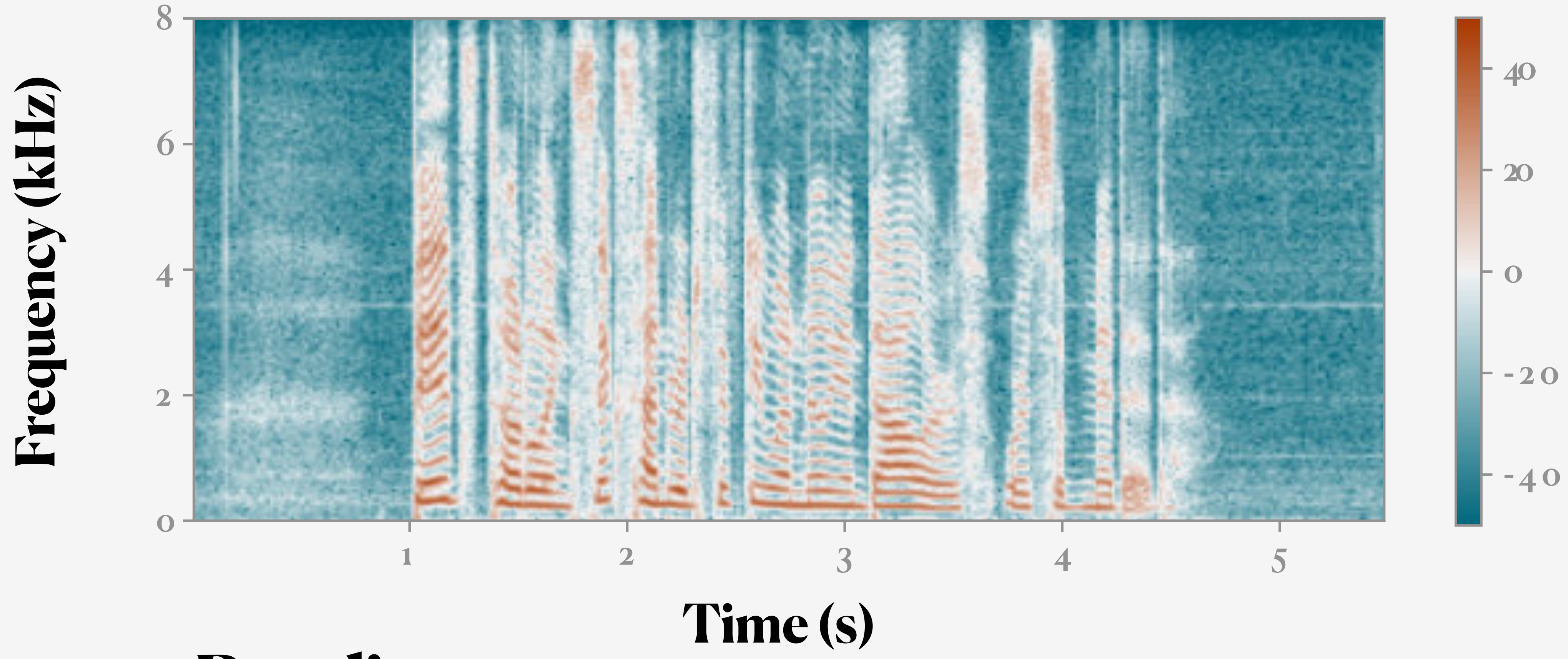


Band-Pass Filter



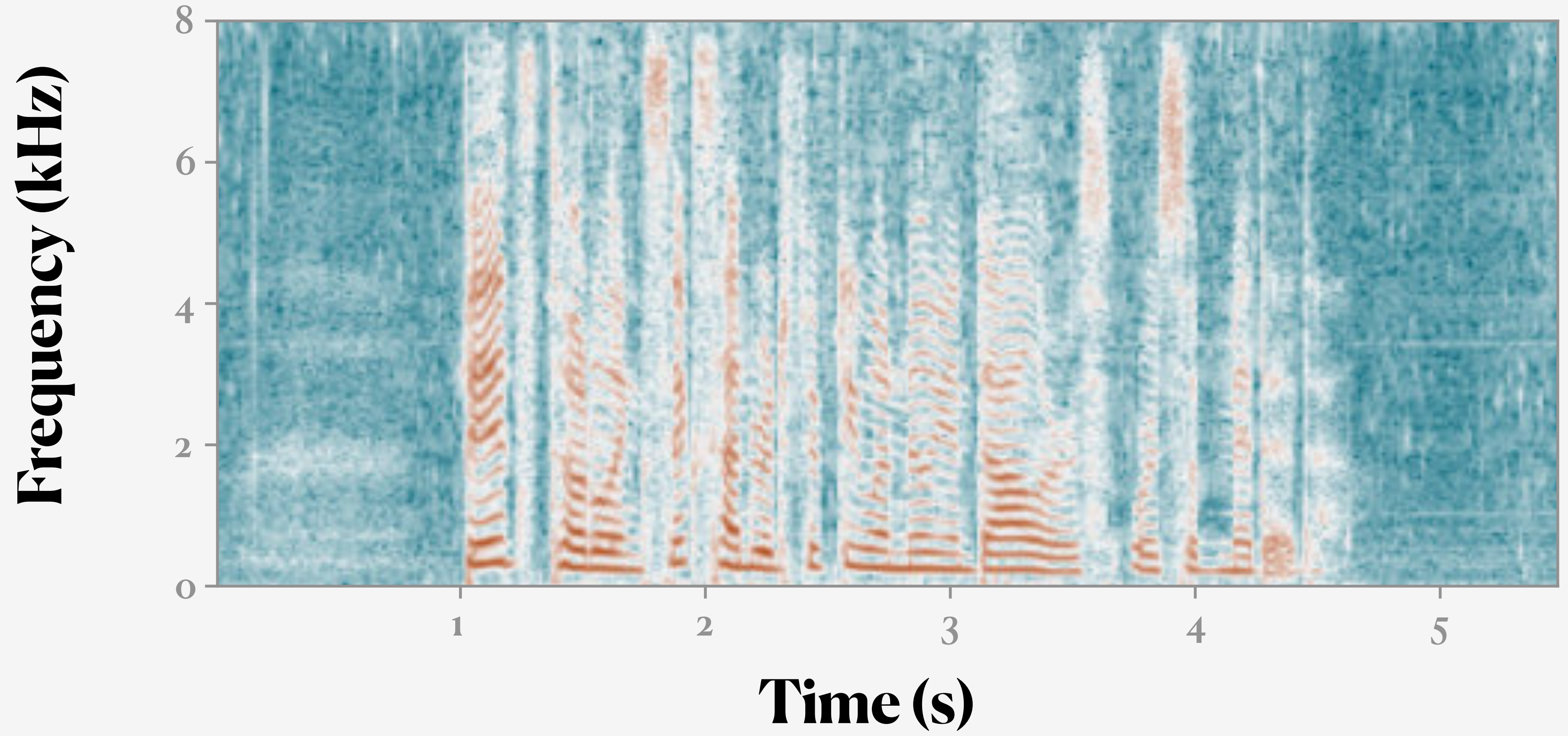
Results

Unmodified Signal



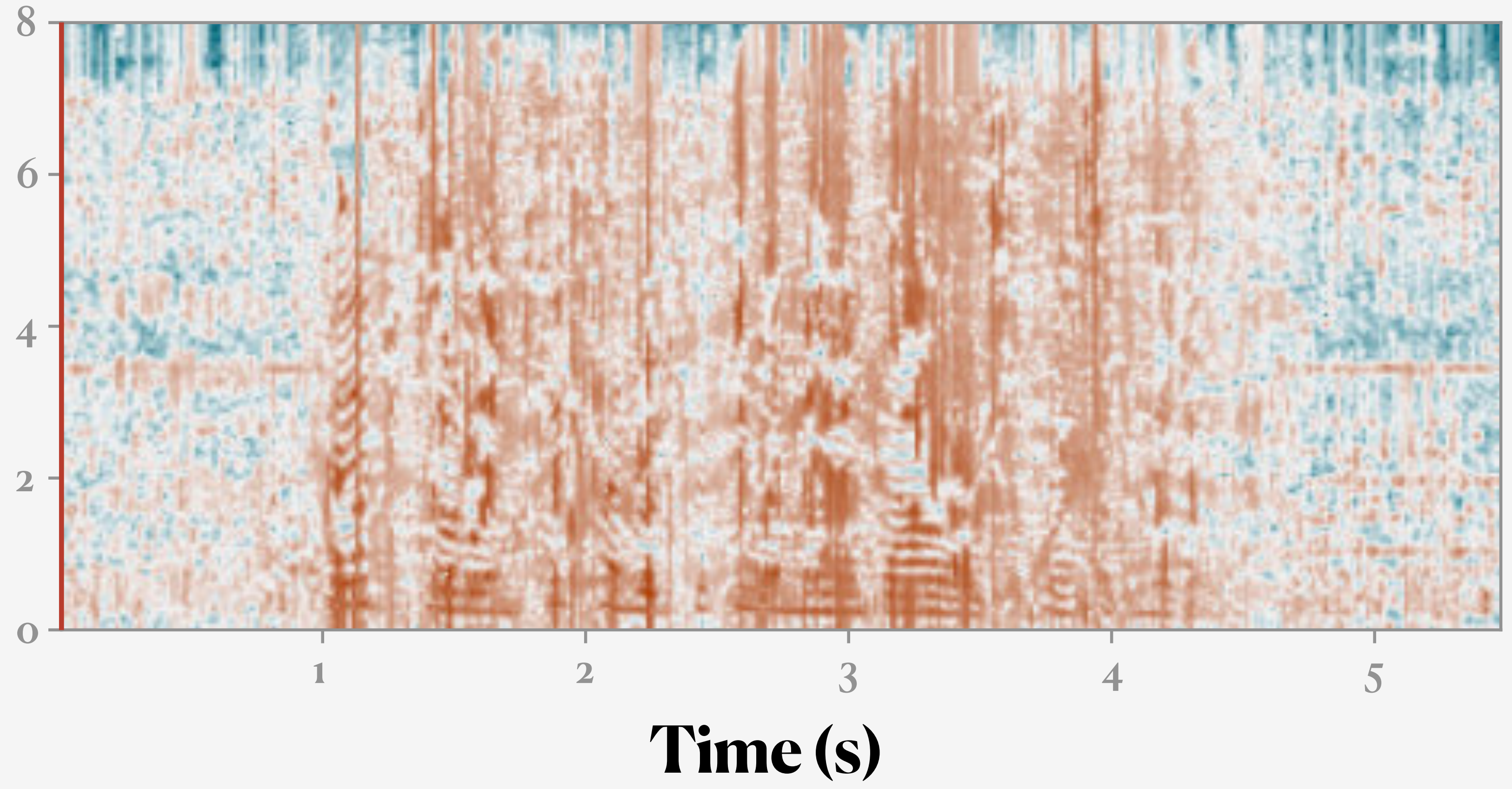
BIDS TOTALING SIX HUNDRED
FIFTY ONE MILLION DOLLARS
WERE SUBMITTED

Baseline



SEND SECRET FINANCIAL REPORT

Augmented System



SEND SECRET FINANCIAL REPORT

Feature-problem-space attacks

Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, and Konrad Rieck
No more Reviewer #2: Subverting Automatic Paper Reviewer Assignment using Adversarial Learning
USENIX Security Symposium, 2023

Domain-specific priors

Thorsten Eisenhofer, Lea Schönherr, Joel Frank, Lars Speckemeier, Dorothea Kolossa, and Thorsten Holz
Dompteur: Taming Audio Adversarial Examples
USENIX Security Symposium, 2021

ML security beyond the model

Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot
Verifiable and Provably Secure Machine Unlearning
In Submission

Machine Unlearning

Server

Delete my data d^* $\longrightarrow$

$$D = \{d_1, \cdots, d_N\} \longrightarrow \boxed{M_D}$$

$$D' = D\backslash\{d^*\} \longrightarrow \boxed{M'_{D'}}$$

M_D might leak about d^*

ok! $\longleftarrow$

Can we trust the server?

Machine Unlearning

Delete my data d^* $\longrightarrow$

Server

$$D = \{d_1, \cdots, d_N\} \longrightarrow \boxed{M_D}$$

$$D' = D \backslash \{d^*\} \dashrightarrow \boxed{M'_{D'}}$$

Cheaper not to
unlearn faithfully!

ok! $\longleftarrow$

Can we trust the server?

Goal: Prove that the unlearning actually happened

Verifiable Machine Unlearning

Verifiable Unlearning

1. Proof of Training
2. Proof of Unlearning

Proof of Training

Proof that M_D was obtained from D

Verified by all users

Proof of Unlearning

- Proof that d^* was removed from M_D
- Proof that $d^* \notin D$

Verified by owner of d^*

Capture consistency across model updates and evolving datasets

Results

Security definition for verifiable machine unlearning

Requires algorithmic definition

Iteration-based protocol

Verifiable computation allows for a generic instantiation

Interface that is applicable to any training and unlearning algorithm

Security proof based on cryptographic assumptions

High computational costs

Proof generation in the order of minutes even for small datasets (100 - 500 data points)

Application specific relaxations possible

Take Aways

Attacks against ML systems ≠ Attack against ML model

Domain-specific priors can help defend a system

Sometimes need to consider the history of a system

Thank you!

Fin