RUB
RESEARCH SCHOOL

Communicating Research

Doreen Riepel and Thorsten Eisenhofer

CASA
Cyber Security in the Age
of Large-Scale Adversaries

RUHR
UNIVERSITÄT
BOCHUM
RUB

CASA

DFG Cluster of Excellence

Cyber Security in the Age of Large-Scale Adversaries

Doreen Riepel

Last year doctoral student

Theoretical Cryptography

Thorsten Eisenhofer

Last year doctoral student

ML & Computer Security

Large-Scale Adversary

RUHR UNIVERSITÄT BOCHUM RUB

Communicating Research

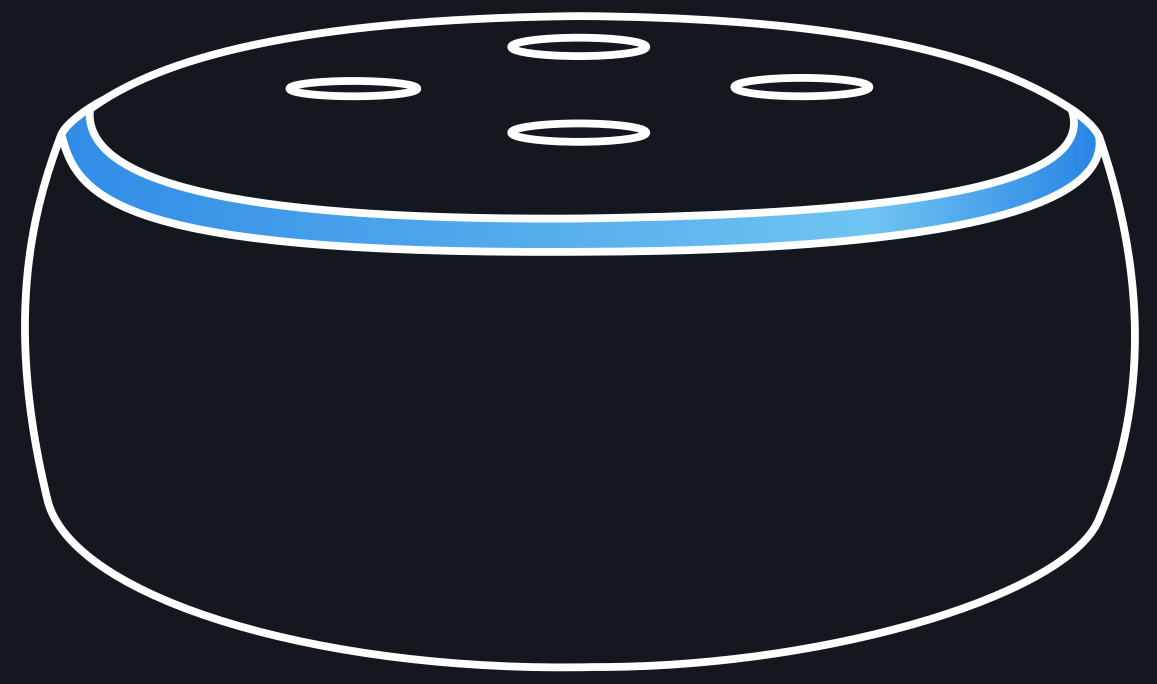
Beyond academia

Interdisciplinary communities

Beyond academia

Project about the security of voice assistants

Raw Audio Wave



Voice Assistant

I SOLEMNLY
SWEAR THAT I AM
UP TO NO GOOD

Transcription

Beyond academia

Different abstraction levels

- Reduced complexity
- Visualization vs. technical details

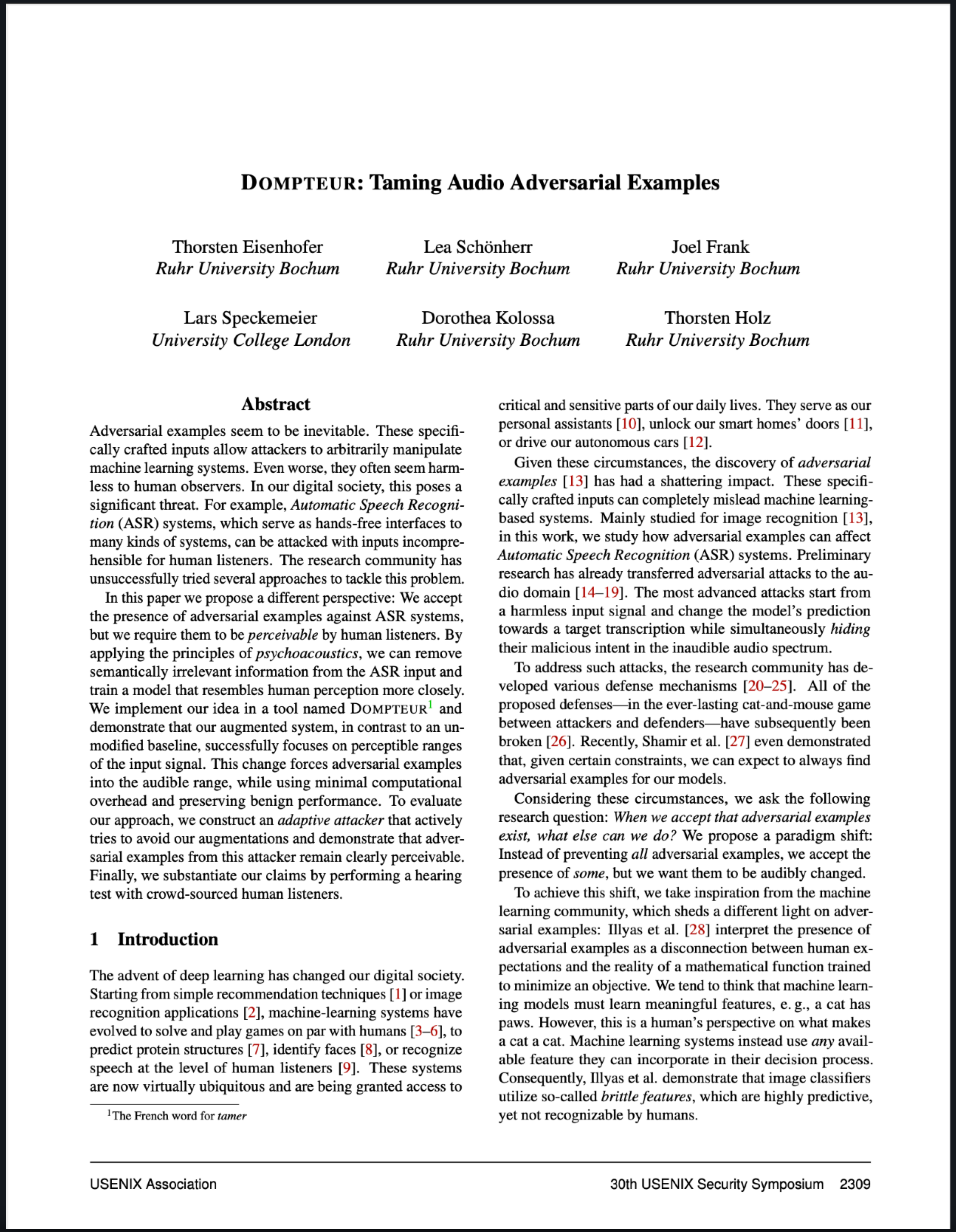
Beyond academia

Research Paper

News Article

Comic

Abstraction



DOMPTEUR: Taming Audio Adversarial Examples

Thorsten Eisenhofer
Ruhr University Bochum

Lea Schönherr
Ruhr University Bochum

Joel Frank
Ruhr University Bochum

Lars Speckemeier
University College London

Dorothea Kolossa
Ruhr University Bochum

Thorsten Holz
Ruhr University Bochum

Abstract

Adversarial examples seem to be inevitable. These specifically crafted inputs allow attackers to arbitrarily manipulate machine learning systems. Even worse, they often seem harmless to human observers. In our digital society, this poses a significant threat. For example, *Automatic Speech Recognition* (ASR) systems, which serve as hands-free interfaces to many kinds of systems, can be attacked with inputs incomprehensible for human listeners. The research community has unsuccessfully tried several approaches to tackle this problem.

In this paper we propose a different perspective: We accept the presence of adversarial examples against ASR systems, but we require them to be *perceivable* by human listeners. By applying the principles of *psychoacoustics*, we can remove semantically irrelevant information from the ASR input and train a model that resembles human perception more closely. We implement our idea in a tool named DOMPTEUR[1] and demonstrate that our augmented system, in contrast to an unmodified baseline, successfully focuses on perceptible ranges of the input signal. This change forces adversarial examples into the audible range, while using minimal computational overhead and preserving benign performance. To evaluate our approach, we construct an *adaptive attacker* that actively tries to avoid our augmentations and demonstrate that adversarial examples from this attacker remain clearly perceivable. Finally, we substantiate our claims by performing a hearing test with crowd-sourced human listeners.

1 Introduction

The advent of deep learning has changed our digital society. Starting from simple recommendation techniques [1] or image recognition applications [2], machine-learning systems have evolved to solve and play games on par with humans [3–6], to predict protein structures [7], identify faces [8], or recognize speech at the level of human listeners [9]. These systems are now virtually ubiquitous and are being granted access to critical and sensitive parts of our daily lives. They serve as our personal assistants [10], unlock our smart homes' doors [11], or drive our autonomous cars [12].

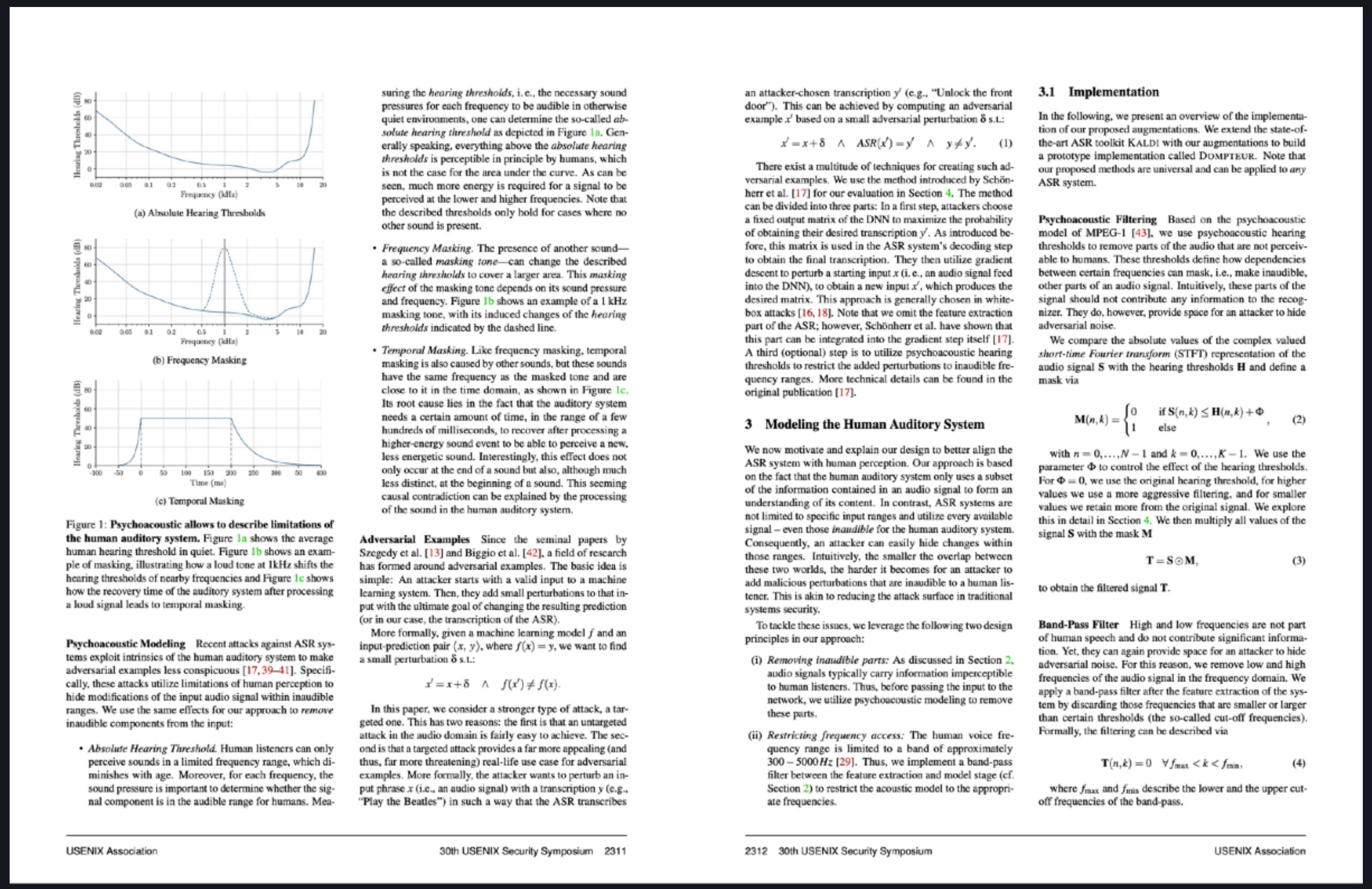
Given these circumstances, the discovery of *adversarial examples* [13] has had a shattering impact. These specifically crafted inputs can completely mislead machine learning-based systems. Mainly studied for image recognition [13], in this work, we study how adversarial examples can affect *Automatic Speech Recognition* (ASR) systems. Preliminary research has already transferred adversarial attacks to the audio domain [14–19]. The most advanced attacks start from a harmless input signal and change the model's prediction towards a target transcription while simultaneously *hiding* their malicious intent in the inaudible audio spectrum.

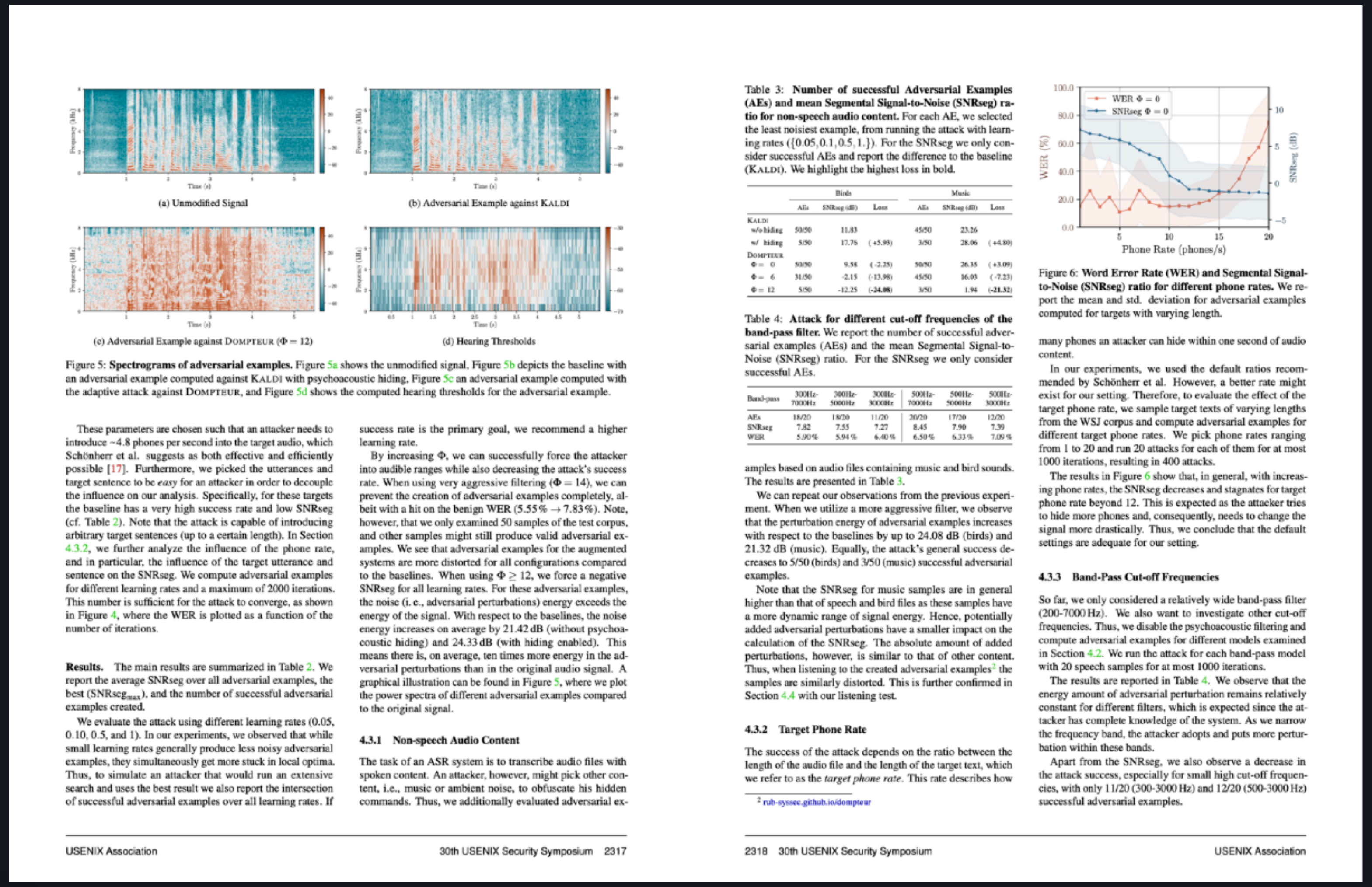
To address such attacks, the research community has developed various defense mechanisms [20–25]. All of the proposed defenses—in the ever-lasting cat-and-mouse game between attackers and defenders—have subsequently been broken [26]. Recently, Shamir et al. [27] even demonstrated that, given certain constraints, we can expect to always find adversarial examples for our models.

Considering these circumstances, we ask the following research question: *When we accept that adversarial examples exist, what else can we do?* We propose a paradigm shift: Instead of preventing *all* adversarial examples, we accept the presence of *some*, but we want them to be audibly changed.

To achieve this shift, we take inspiration from the machine learning community, which sheds a different light on adversarial examples: Illyas et al. [28] interpret the presence of adversarial examples as a disconnection between human expectations and the reality of a mathematical function trained to minimize an objective. We tend to think that machine learning models must learn meaningful features, e. g., a cat has paws. However, this is a human's perspective on what makes a cat a cat. Machine learning systems instead use *any* available feature they can incorporate in their decision process. Consequently, Illyas et al. demonstrate that image classifiers utilize so-called *brittle features*, which are highly predictive, yet not recognizable by humans.

[1] The French word for *tamer*





Eisenhofer et al.: *"Dompteur: Taming Audio Adversarial Examples"*, USENIX Security 2021

CASA

RUHR UNIVERSITÄT BOCHUM RUB

Beyond academia

Abstraction

RESEARCH PAPER

NEWS ARTICLE

COMIC

https://news.rub.de/sites/default/files/rubin_2019_2_web.pdf

CASA

RUHR UNIVERSITÄT BOCHUM

RUB

Beyond academia

Abstraction

RESEARCH PAPER

NEWS ARTICLE

COMIC

https://www.sueddeutsche.de/projekte/artikel/digital/smartspeaker-alexa-und-co-lauschen-versehentlich-e793169

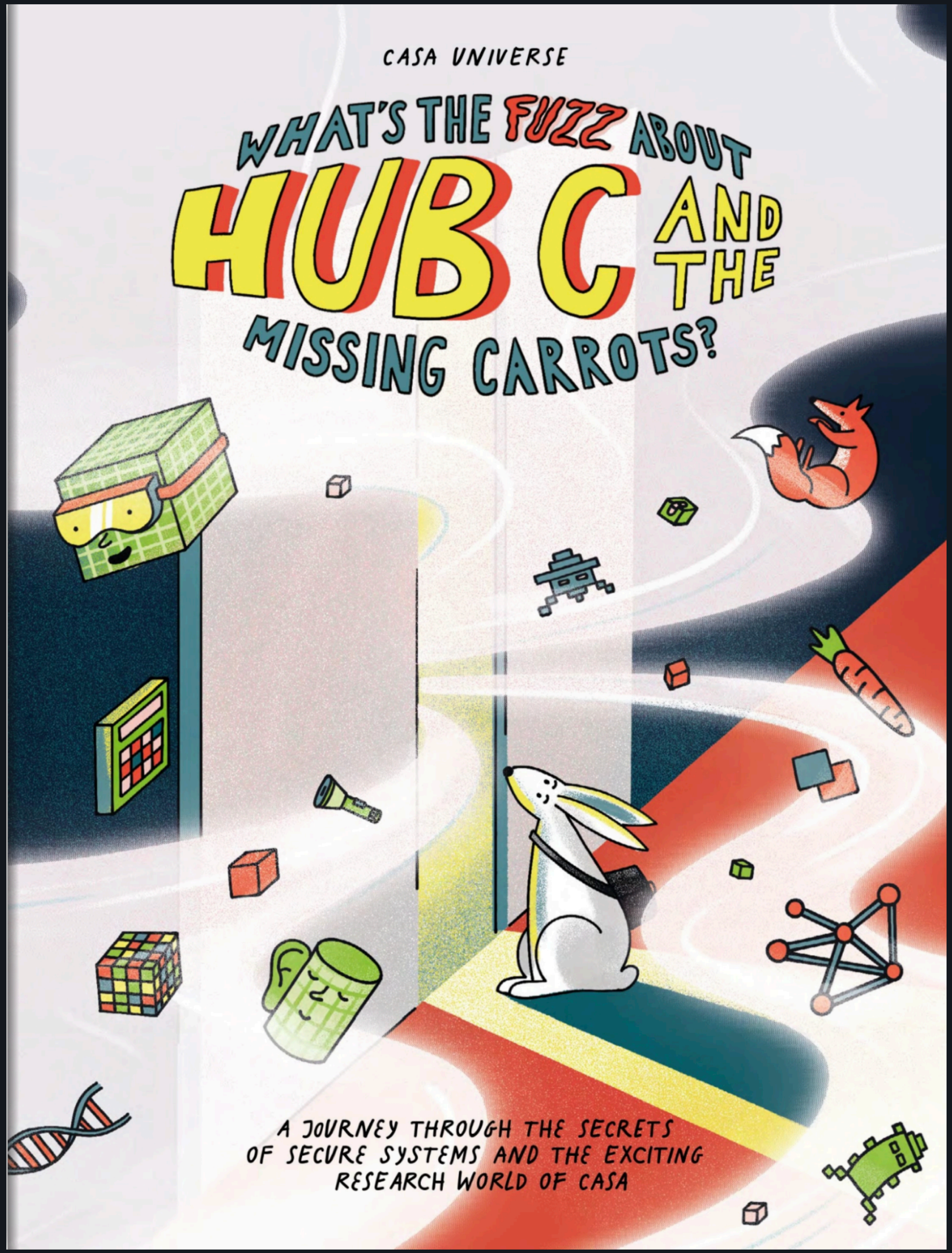
Beyond academia

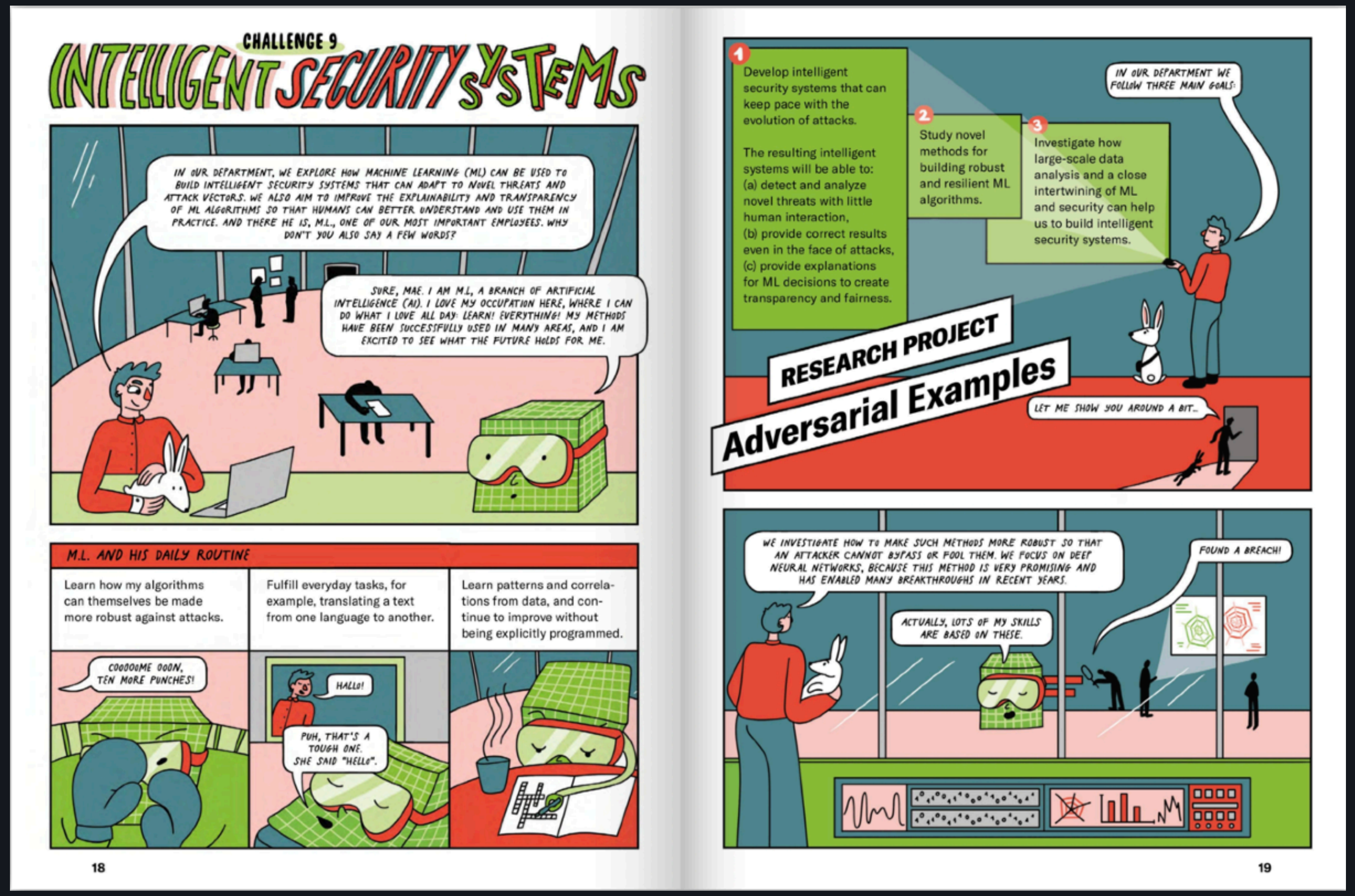
Abstraction

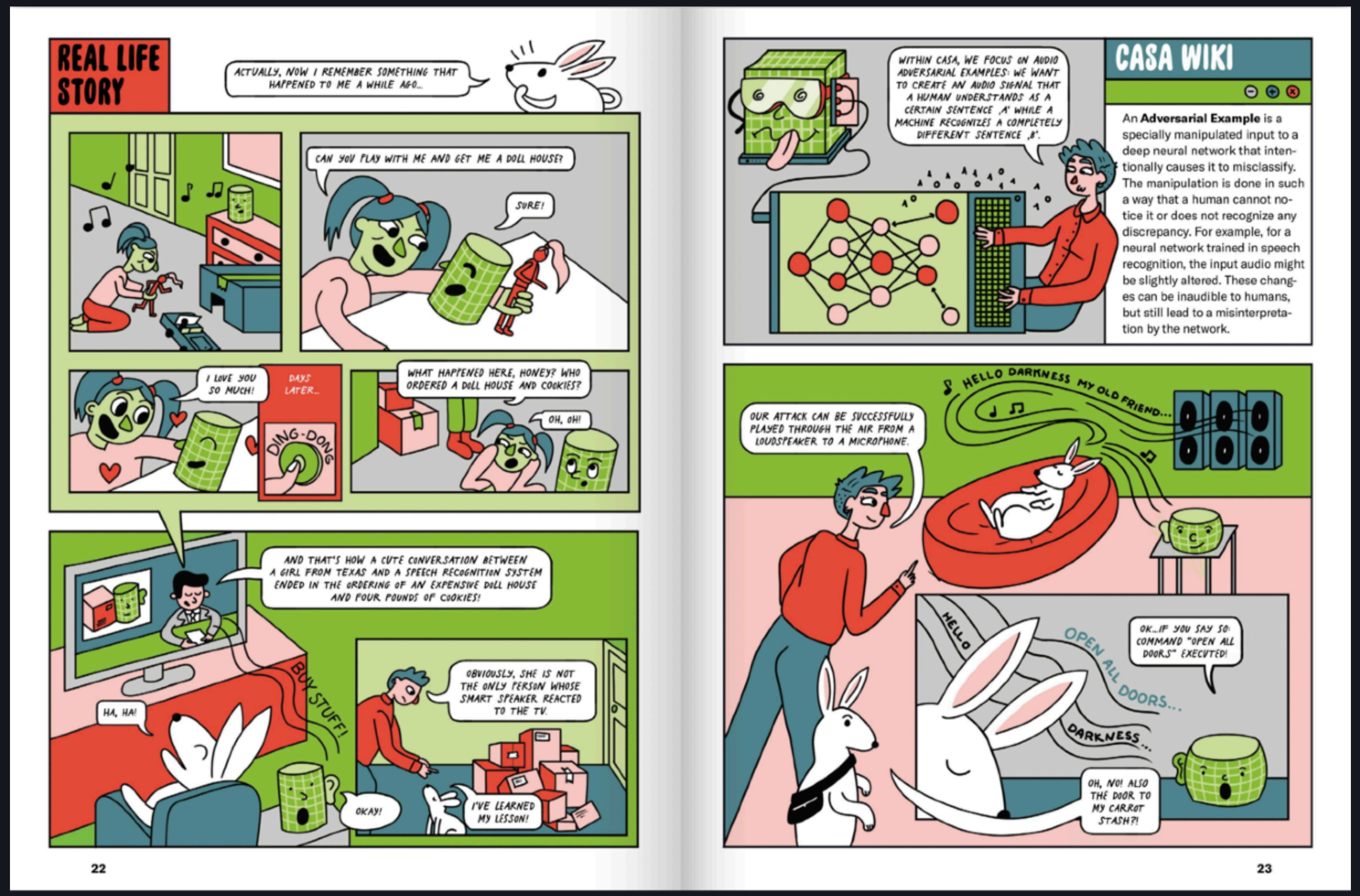
Research Paper

News Article

Comic







https://casa.rub.de/outreach/wissenschaftscomics

CASA

RUHR UNIVERSITÄT BOCHUM

RUB

Beyond academia

Different abstraction levels

- Reduced complexity
- Visualizations vs. technical details

Caveat: Overselling

- Work together and communicate simplifying assumptions
- Decline requests when you're not comfortable

Communicating Research

Beyond academia

Interdisciplinary communities

Interdisciplinary communities

Cryptography for trustworthy machine learning

Provably secure implementation of the right to be forgotten

Researchers from different fields

Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot

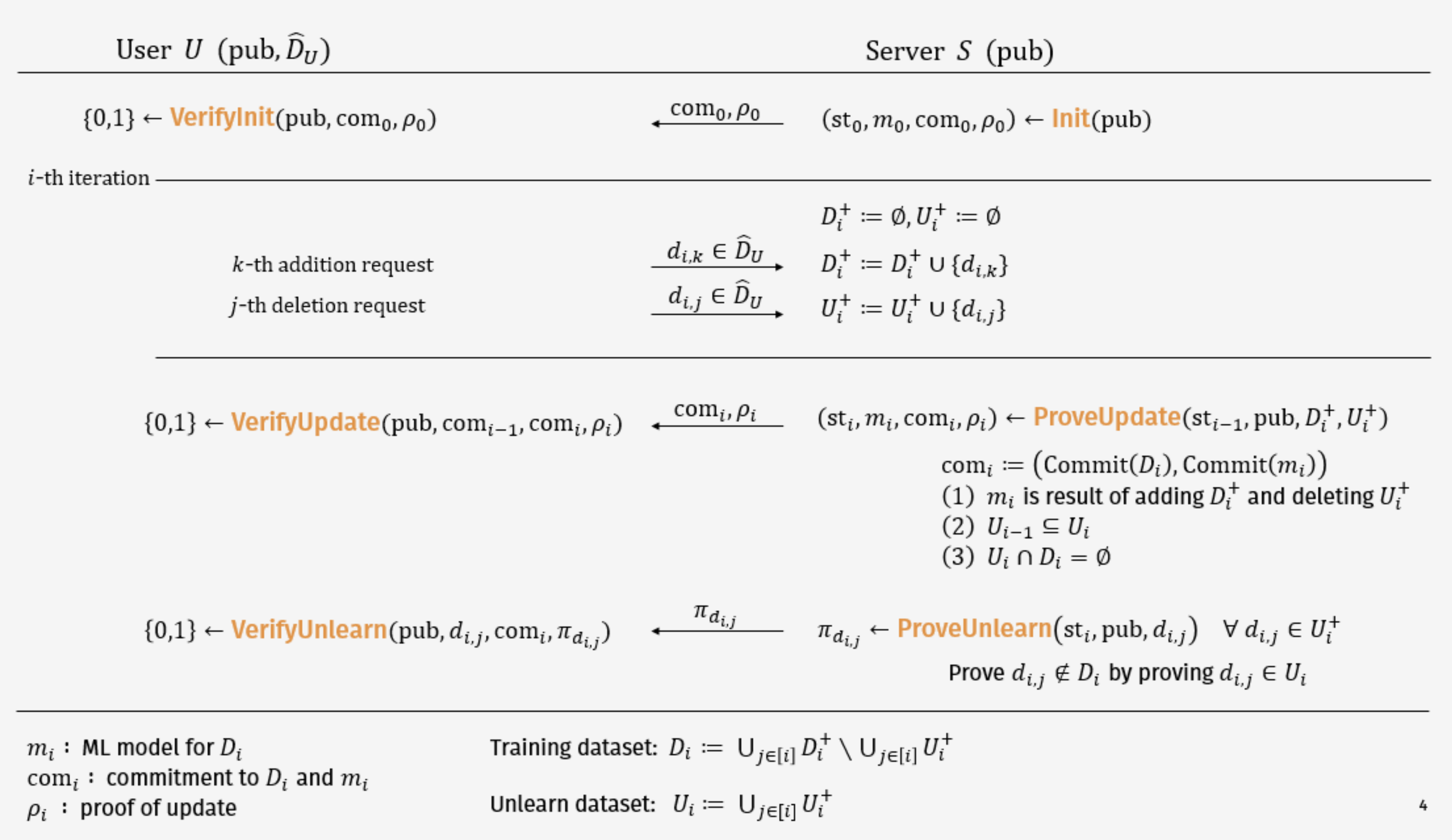
"Verifiable and Provably Secure Machine Unlearning", *In Submission*

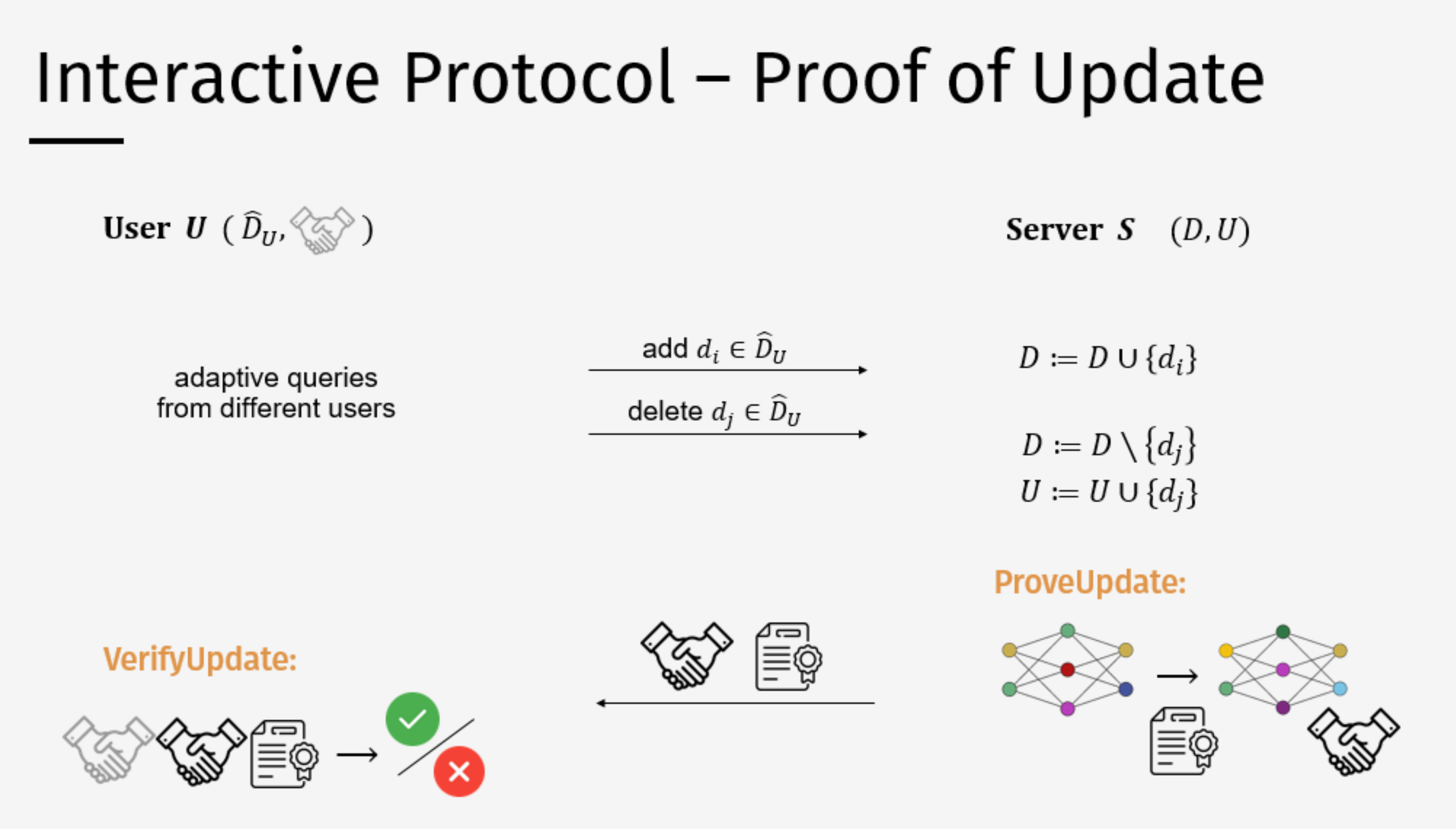
CASA

RUHR UNIVERSITÄT BOCHUM

RUB

Interdisciplinary communities





AUDIENCE: CRYPTOGRAPHERS

AUDIENCE: RESEARCHERS IN SECURITY / ML

CASA

RUHR UNIVERSITÄT BOCHUM RUB

Interdisciplinary communities

Work on interesting problems on the edge of communities

- Develop a common language across communities
- Opportunity to work with and learn from experts of different fields

Caveat: Underselling

- Solution might be trivial within each community