Angriffe mit KI

Urs Wagner

cnlab Herbsttagung 2024: KI und Sicherheit
Gleisarena, Zürich, 4. September 2024

cnlab

Verteidiger

Angreifer

Microsoft und OpenAI

- Analysierten Aktivitäten von bekannten «Threat Actors»
- Identifizierten «Techniques, Tactics and Procedures» (TTPs) mit LLMs

Reconnaissance: Information zu

- Organisation
- Infrastruktur
- Technologien
- Personen

Entwicklung von Fähigkeiten / Verständnis

- Schwachstellen
- Software
- Systeme

Umgehung von Sicherheitsmassnahmen

- Captchas

Unterstützung bei Entwicklung von Tools

- Skripts
- Malware
- Payloads

Social Engineering

- Phishing-Mails
- Übersetzungen
- Impersonation

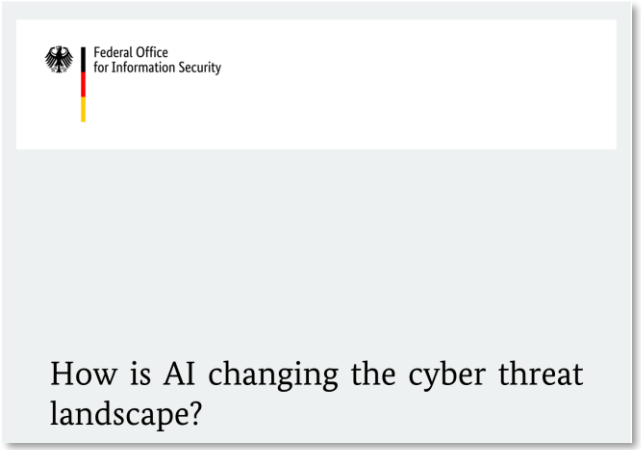
https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

Microsoft und OpenAI - Konklusion

- Konklusion der Studie:
  - ChatGPT als **Produktivitätstool** für Angriffe
  - **Keine** besonders neu- oder **einzigartigen Angriffstechniken** beobachtet
- Aber: Konklusion nur basierend auf Verhalten auf den Online-Diensten!

Was sagen Behörden?

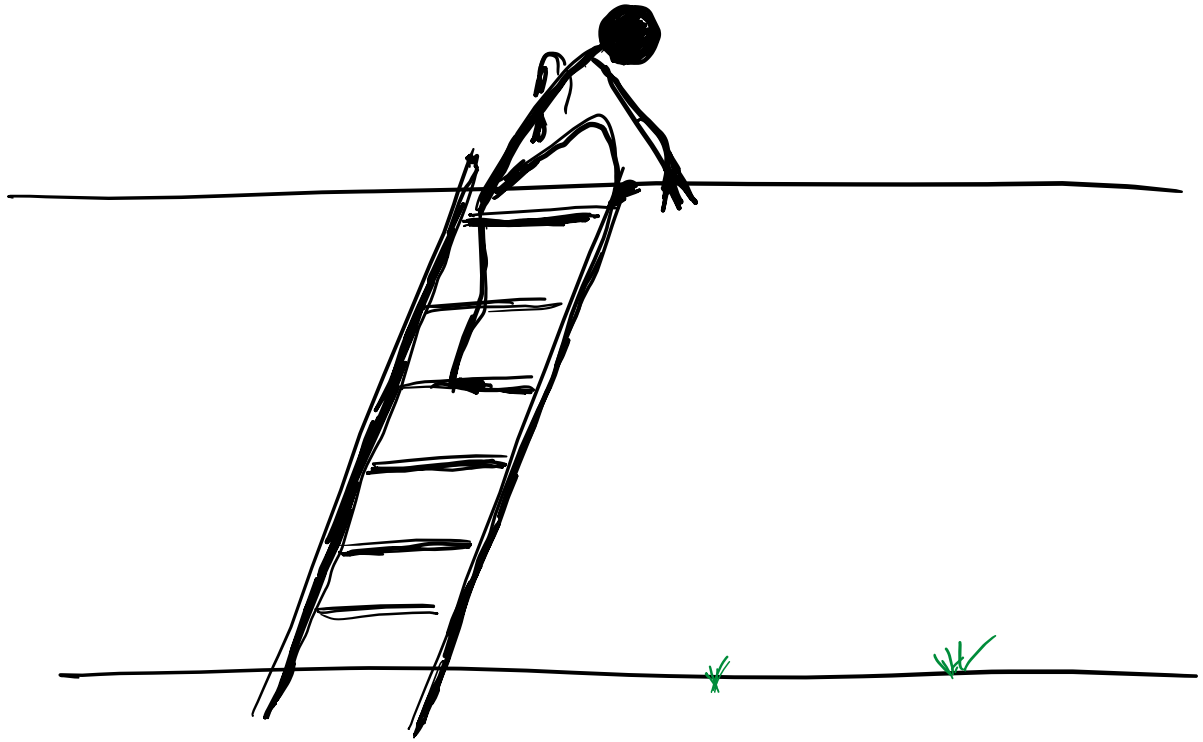
- NCSC (UK) und BSI (Deutschland)
- Hauptaussagen zu Angriffen mit KI:
  - reduzieren Einstiegshürden
  - erlauben erhöhte Qualität und Umfang von Angriffen
  - POC-Status, aber nicht «production ready»:
    - automatische Generierung und Mutation von Malware
    - automatische Exfiltration
  - mittelfristig nicht in Sicht:
    - Agenten, welche eigenständig beliebige Infrastruktur kompromittieren können



https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/How-is-AI-changing-cyber-threat-landscape.pdf

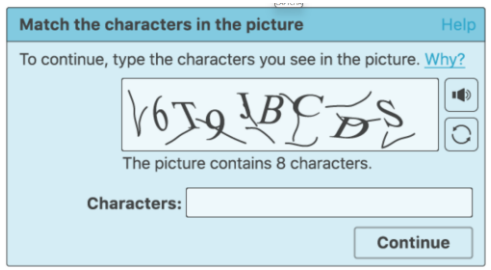
https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

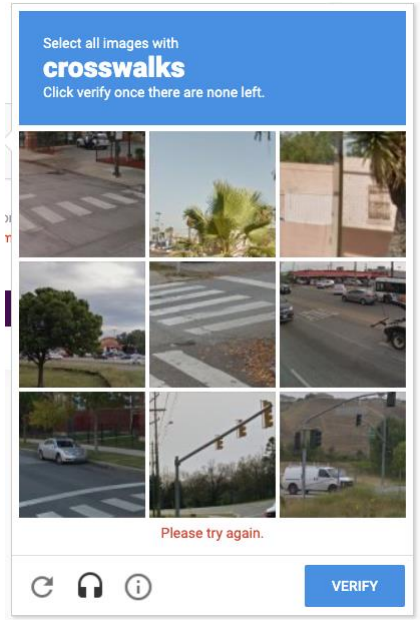
Umgehung von Sicherheitsmassnahmen

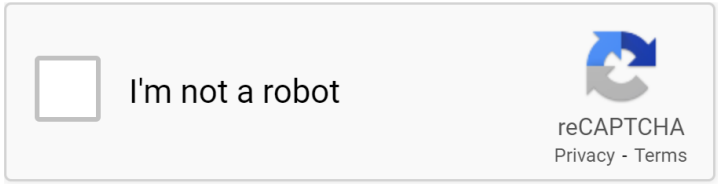


CAPTCHA

- «**C**ompletely **A**utomated **P**ublic **T**uring Test to tell **C**omputers and **H**umans **A**part»
- Einfach für Menschen, schwierig für Maschinen: Beschränkung von Zugriffen







Passwörter

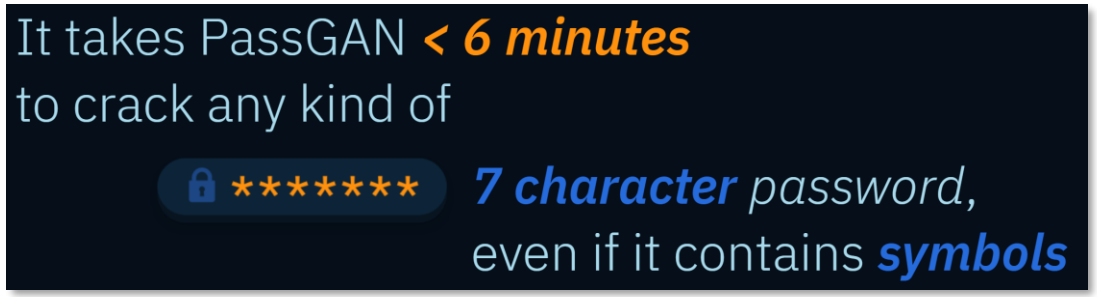

NEWS

AI can crack most passwords faster than you can read this article

Artificial intelligence is accelerating the ability to crack weak passwords quickly.

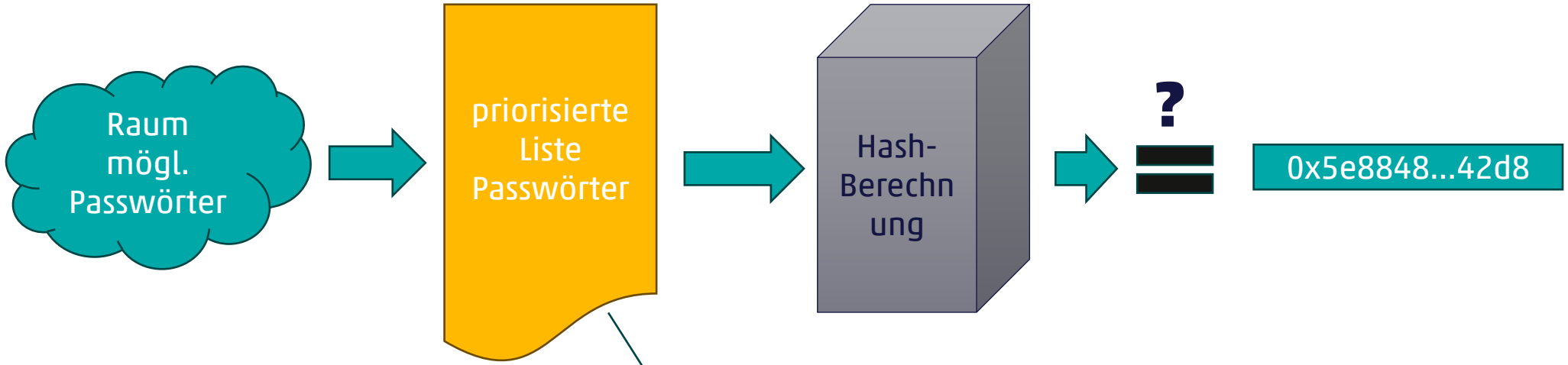
https://www.pcworld.com/article/1782671/ai-can-crack-most-passwords-faster-than-you-can-read-this-article.html

https://www.securityhero.io/ai-password-cracking/



- PassGAN, ein KI-Tool von «Home Security Heros»
- Generative Adversarial Network (GAN) trainiert mit einem Password-Breach von RockYou (2010, 15.7 Millionen Passwörter)

Wie funktioniert (offline) Password-Cracking?



Einfluss auf Speed	Grösse	Güte der Priorisierung Generierungsgeschw.	Rechenpower Hashfunktion

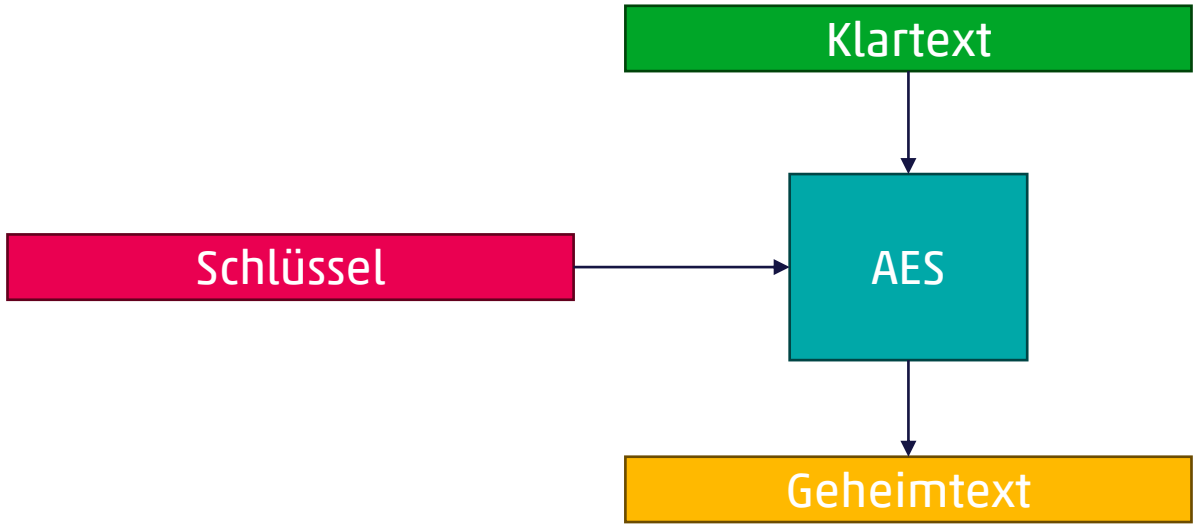
KI kann einen Beitrag leisten

KI ist nicht schnell und liefert Duplikate

nutzlos für zufällige Passwörter!

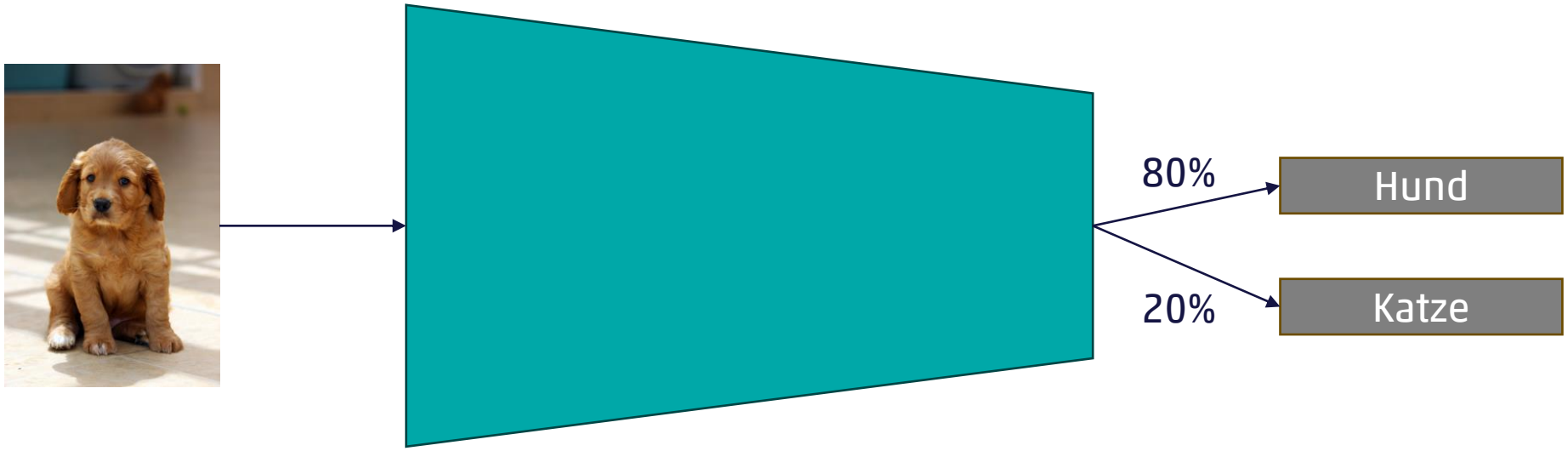
Gute Passwörter bleiben gut, schlechte Passwörter bleiben schlecht

Kryptografie am Beispiel von AES



- Ignoriert für den Moment:
  - Block-Mode
  - IV (Randomization)

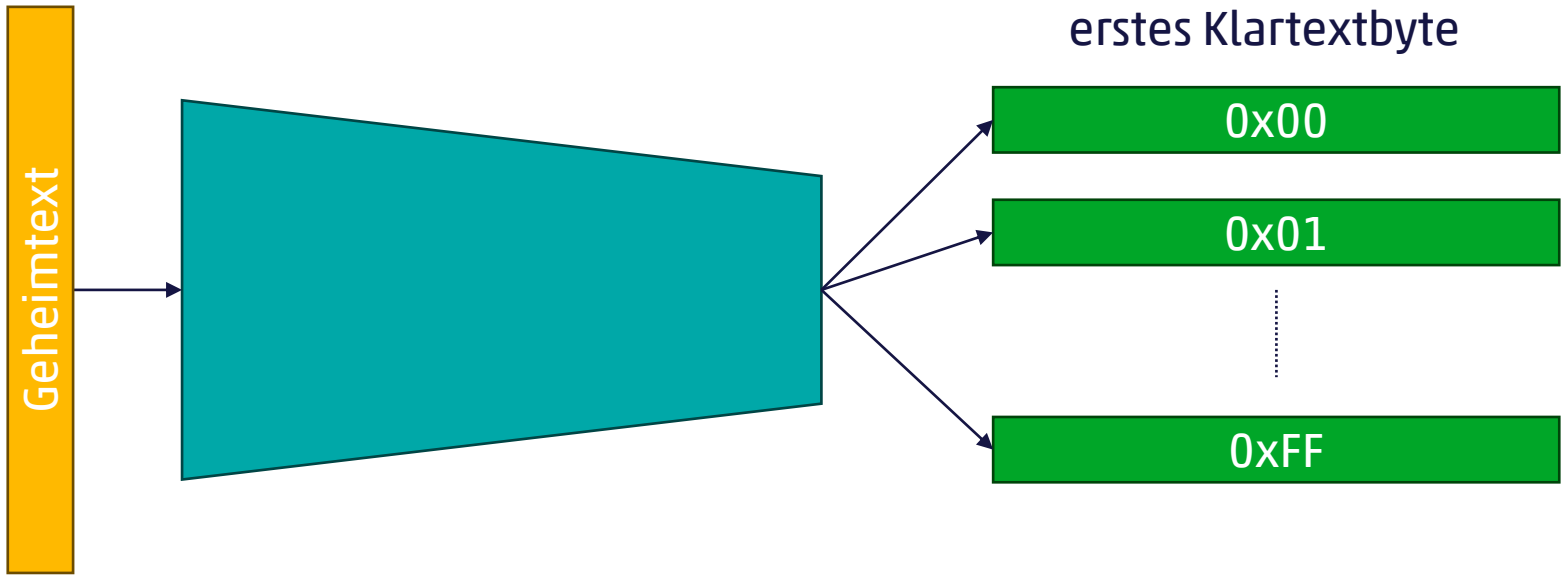
Deep-Learning / Klassifizierung



Idee 1: Brauche Deep-Learning Algorithmen um den Klartext zu finden

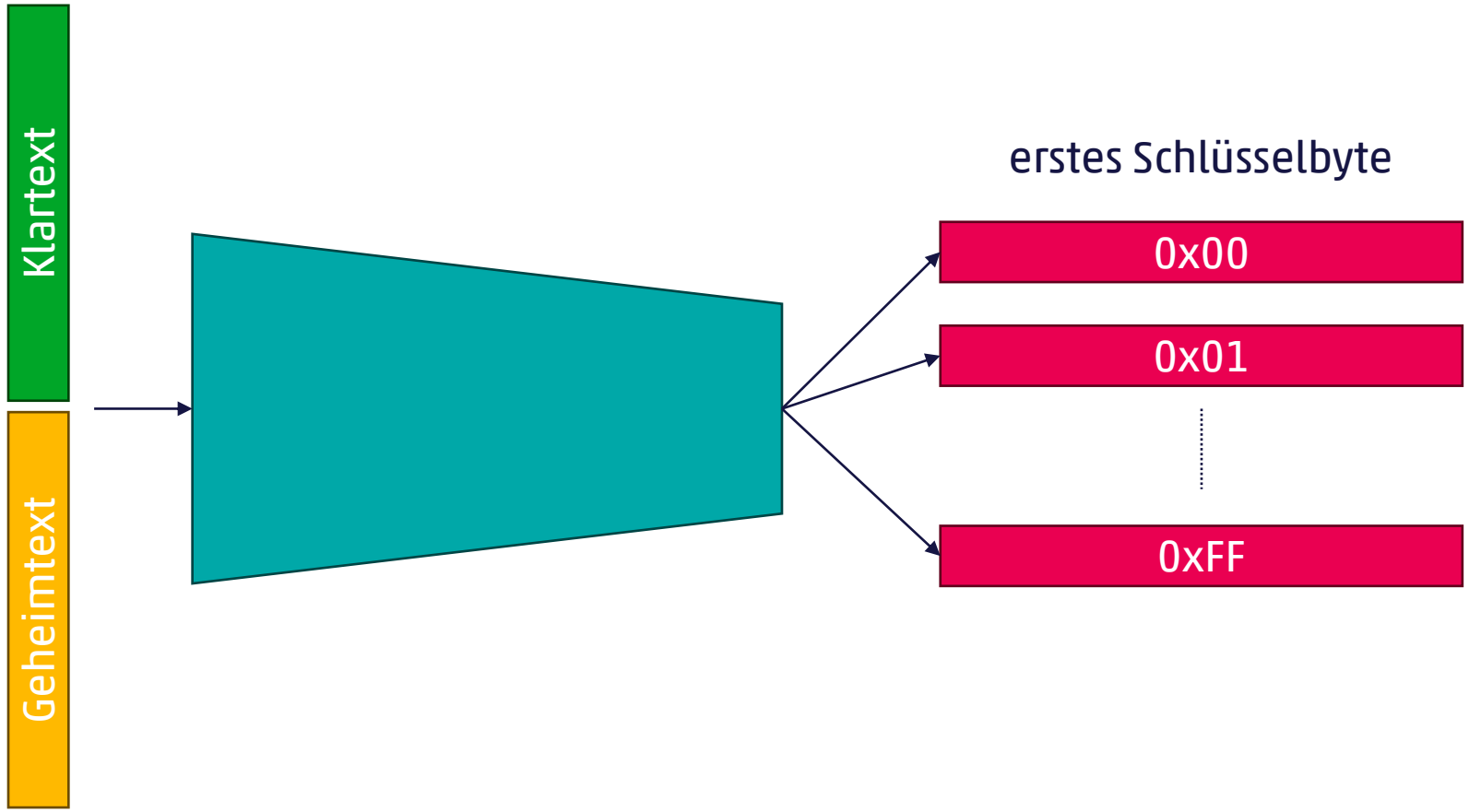
- Schlüssel fix

1. Trainiere das Modell mit vielen bekannten Klar- und Geheimtextpaaren
2. Benutze das trainierte Modell, um zu einem Geheimtext den Klartext (erstes Byte) zu extrahieren



Idee 2: Brauche Deep-Learning Algorithmen um den Schlüssel zu finden

1. Trainiere das Modell mit vielen bekannten Klar- und Geheimtextpaaren und Schlüsseln
2. Benutze das trainierte Modell, um zu einem Klar- und Geheimtextpaar den Schlüssel zu extrahieren



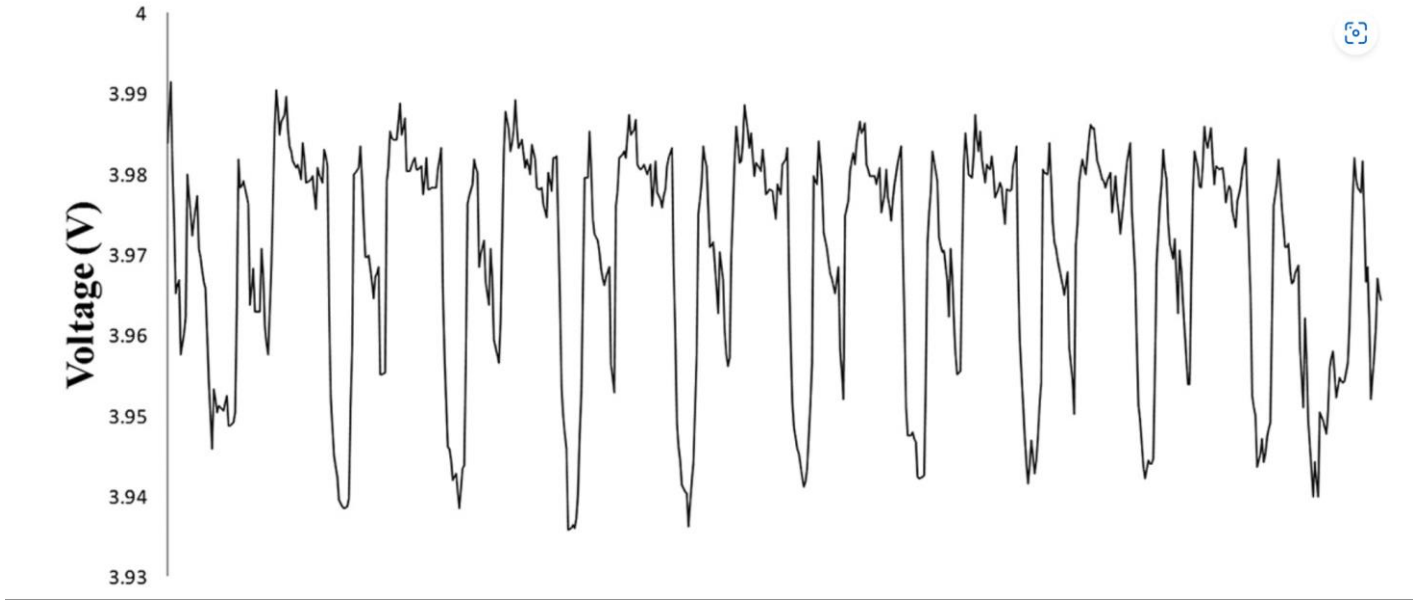
AES – Design Ziel

- Output von AES soll **zufällig** aussehen. Auch bei Kenntnis von vielen Klar- und Geheimtext-Paaren
  - soll nicht auf den Klartext eines neuen Geheimtexts geschlossen werden können
  - soll nicht auf den Schlüssel geschlossen werden können
- D.h. auch KI kann nicht lernen
  - es fehlt die Struktur/Muster
  - bzw. eine Struktur/Muster würde eine Schwäche bedeuten
- Geschichte spricht für AES:
  - Standardisiert in 2001
  - Sehr verbreitet
  - State-of-the-Art
  - Design Ziel oft gechallenged
- Design-Ziel gilt auch für andere etablierte Algorithmen

Erfolgreicher Angriff unwahrscheinlich.

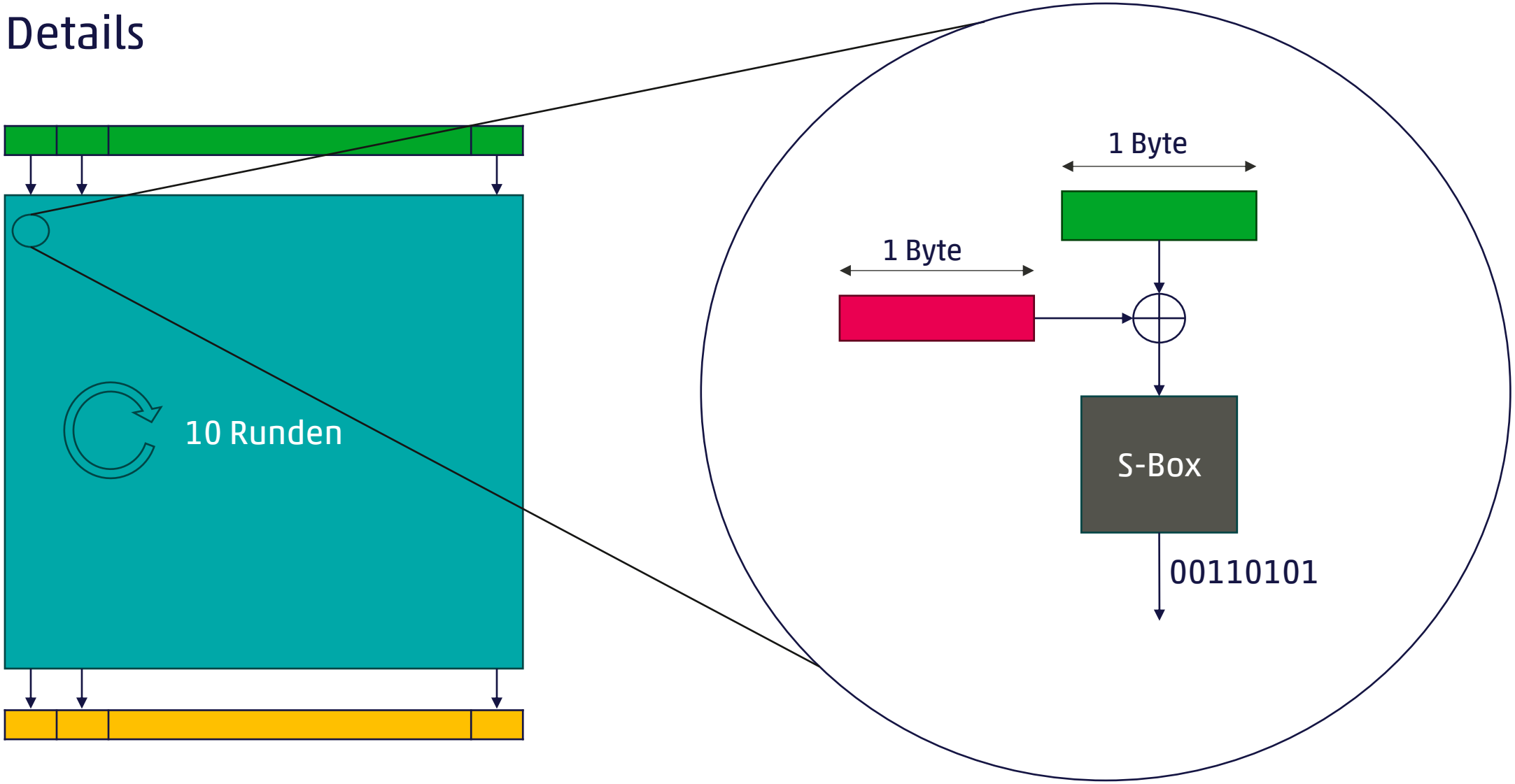
… aber was ist mit Side-Channel Angriffen?

- Angriffe auf Implementationen
- Schlüsselextraktion von
  - TPMs
  - Auto-Fernbedienung
  - …
- Side-Channels:
  - Timing
  - Strombedarf
  - Hitze
  - Elektromagnetische Strahlung



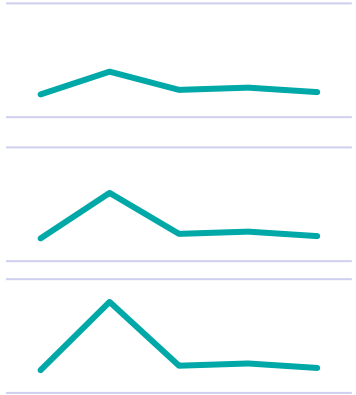
Lo, O., Buchanan, W. J., & Carson, D. (2016). Power analysis attacks on the AES-128 S-box using differential power analysis (DPA) and correlation power analysis (CPA). *Journal of Cyber Security Technology*, *1*(2), 88–107. https://doi.org/10.1080/23742917.2016.1231523

AES - Details

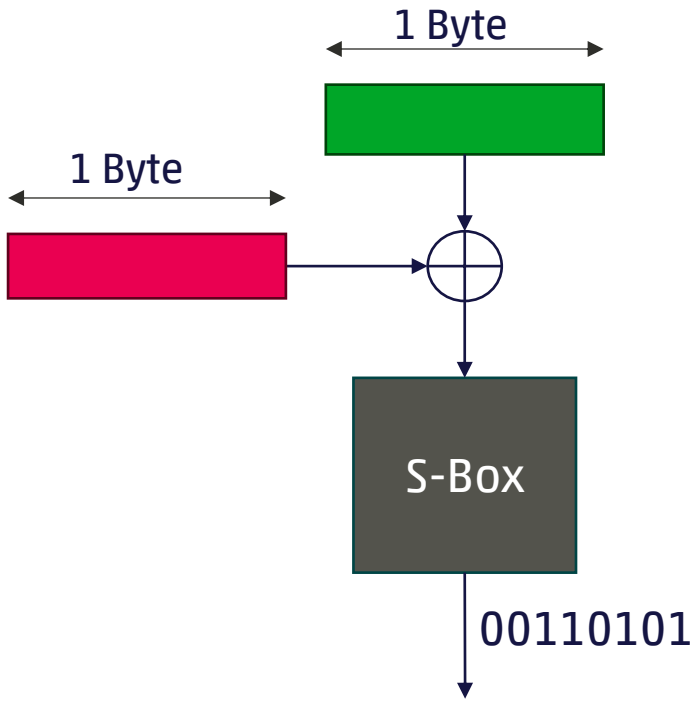


AES – Side Channel Angriffspunkte

- Output S-Box:
  - Strombedarf abhängig von Anzahl Einsen (Hamming-Weight)
  - Möglich, Templates zu erstellen:

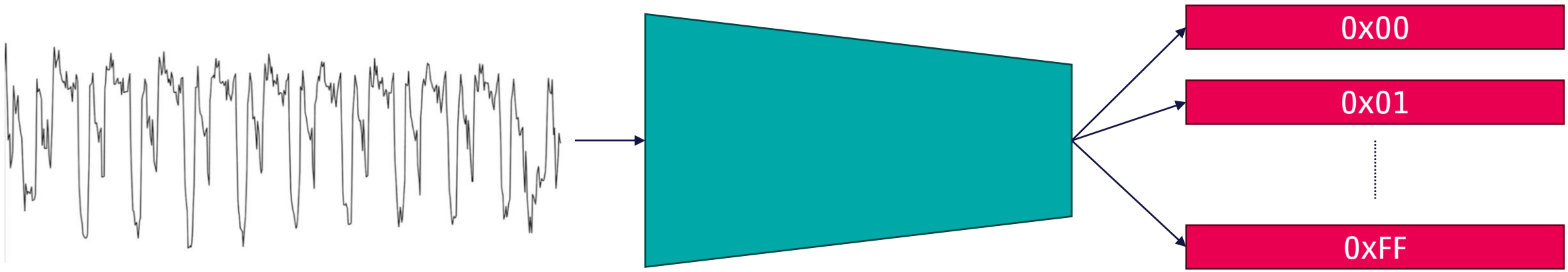


- Angriff:
  1. Rate das Schlüsselbyte
  2. Chiffriere verschiedene Plaintexts
  3. Schaue, ob Power-Traces mit erwartetem Korrelieren
- Viele «händische» Arbeit
  - Traces ausrichten
  - Interessante Stellen finden
  - Statistisches Modell erstellen



Idee: Trainiere ML-Modell mit Power-Traces um den Schlüssel zu finden

- Trainiere das Modell mit vielen bekannten Powertraces bei bekanntem Klartext und Schlüsseln
- Modell «lernt» die statistische Abhängigkeit der Traces vom Schlüssel



- Gewisse Modelle können gebraucht werden, um die relevanten Stellen der Implementation zu finden
- Kann auch defensiv verwendet werden

Erfolg?

- GPAM: Deep-Learning-Architektur
- Erfolgreich gegen **geschützte** Implementationen von AES und ECC

Generalized Power Attacks against Crypto Hardware using Long-Range Deep Learning

Elie Bursztein[1], Luca Invernizzi[2], Karel Král[2], Daniel Moghimi[1], Jean-Michel Picod[2] and Marina Zhang[1]

[1] Google, Sunnyvale, USA, scaaml@google.com
[2] Google, Zurich, Switzerland, scaaml@google.com

"there is a pressing need to devise new countermeasures that are resilient to ML attacks"

Fazit

- Sind mit KI Angriffe möglich, die vorher unmöglich waren? – Nein
- Kann KI Angreifern helfen? – Ja
- Lässt KI das ewige Pendel zwischen Angreifer und Verteidiger auf Seite der Angreifer schlagen? – Nein
- Müssen Verteidiger die Entwicklungen im Bereich KI verfolgen? – Ja, wie mit jeder technischen Entwicklung

Vielen Dank für Ihre Aufmerksamkeit_

Urs Wagner

info@cnlab-security.ch
+41 55 214 33 40

cnlab security AG
Obere Bahnhofstrasse 32b
CH-8640 Rapperswil-Jona
Switzerland

cnlab