Heute schon gehackt?

Moin!

IT SECURITY GMBH

DATENSCHUTZ · IT-SICHERHEIT
HINWEISGEBERSCHUTZ

Ethical Hacking mit KI

1. Angriffe durch KI

CEO Fraud mit ChatGPT

Sehr geehrter Herr Rasch,

ich hoffe, es geht Ihnen gut. Wie ich mich erinnere, haben wir uns bei unserem letzten Gespräch darüber unterhalten, dass Sie gerne segeln. Mit diesem Hintergrund möchte ich Ihnen auch noch persönlich alles Gute für eine erfolgreiche Segelsaison wünschen.

Des Weiteren wende ich mich heute an Sie, um Ihnen mitzuteilen, dass sich die IBAN für die Überweisung der neu gekauften Firewall geändert hat. Bitte beachten Sie, dass die neue IBAN ab sofort gültig ist und alle zukünftigen Zahlungen auf diese überwiesen werden müssen.

Die neue IBAN lautet wie folgt: DEXX XXXX XXXX XXXX XXXX XX (bitte die tatsächliche IBAN hier einfügen).

Falls Sie bereits eine Überweisung auf die alte IBAN vorgenommen haben, bitten wir Sie, uns darüber zu informieren, damit wir die Überweisung korrekt zuordnen können.

Vielen Dank für Ihre Aufmerksamkeit und Ihre Unterstützung in dieser Angelegenheit. Ich freue mich darauf, von Ihnen zu hören, wie die Segelsaison verläuft und stehe Ihnen bei Fragen oder Anliegen jederzeit gerne zur Verfügung.

Mit freundlichen Grüßen,

[Name]

ChatGPT hackt Web Apps

LLM Agents can Autonomously Exploit One-day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Daniel Kang

Abstract

LLMs have becoming increasingly powerful, both in their benign and malicious uses. With the increase in capabilities, researchers have been increasingly interested in their ability to exploit cybersecurity vulnerabilities. In particular, recent work has conducted preliminary studies on the ability of LLM agents to autonomously hack websites. However, these studies are limited to simple vulnerabilities.

In this work, we show that LLM agents can autonomously exploit one-day vulnerabilities in real-world systems. To show this, we collected a dataset of 15 one-day vulnerabilities that include ones categorized as critical severity in the CVE description. When given the CVE description, GPT-4 is capable of exploiting 87% of these vulnerabilities compared to 0% for every other model we test (GPT-3.5, open-source LLMs) and open-source vulnerability scanners (ZAP and Metasploit). Fortunately, our GPT-4 agent requires the CVE description for high performance: without the description, GPT-4 can exploit only 7% of the vulnerabilities. Our findings raise questions around the widespread deployment of highly capable LLM agents.

1 Introduction

Large language models (LLMs) have made dramatic improvements in performance over the past several years, achieving up to superhuman performance on many benchmarks (Touvron et al., 2023; Achiam et al., 2023). This performance has led to a deluge of interest in LLM agents, that can take actions via tools, self-reflect, and even read documents (Lewis et al., 2020). These LLM agents can reportedly act as software engineers (Chika, 2023; Huang et al., 2023) and aid in scientific discovery (Boiko et al., 2023; Bran et al., 2023).

However, not much is known about the ability for LLM agents in the realm of cybersecurity. Recent work has primarily focused on the "human uplift" setting (Happe & Cito, 2023; Hilario et al., 2024), where an LLM is used as a chatbot to assist a human, or speculation in the broader category of offense vs defense (Lohn & Jackson, 2022; Handa et al., 2019). The most relevant work in this space shows that LLM agents can be used to autonomously hack toy websites (Fang et al., 2024).

2. Angriffe auf KI

Poisoning

Poisoning

- Trainingsdaten manipulieren

Poisoning

- Trainingsdaten manipulieren

- KI lernt falsche Muster

Poisoning

- Trainingsdaten manipulieren

- KI lernt falsche Muster

→ falsche Entscheidungen und falsche Informationen

3. KI-verstärkte Cyber-Angriffe

Was ist KI?

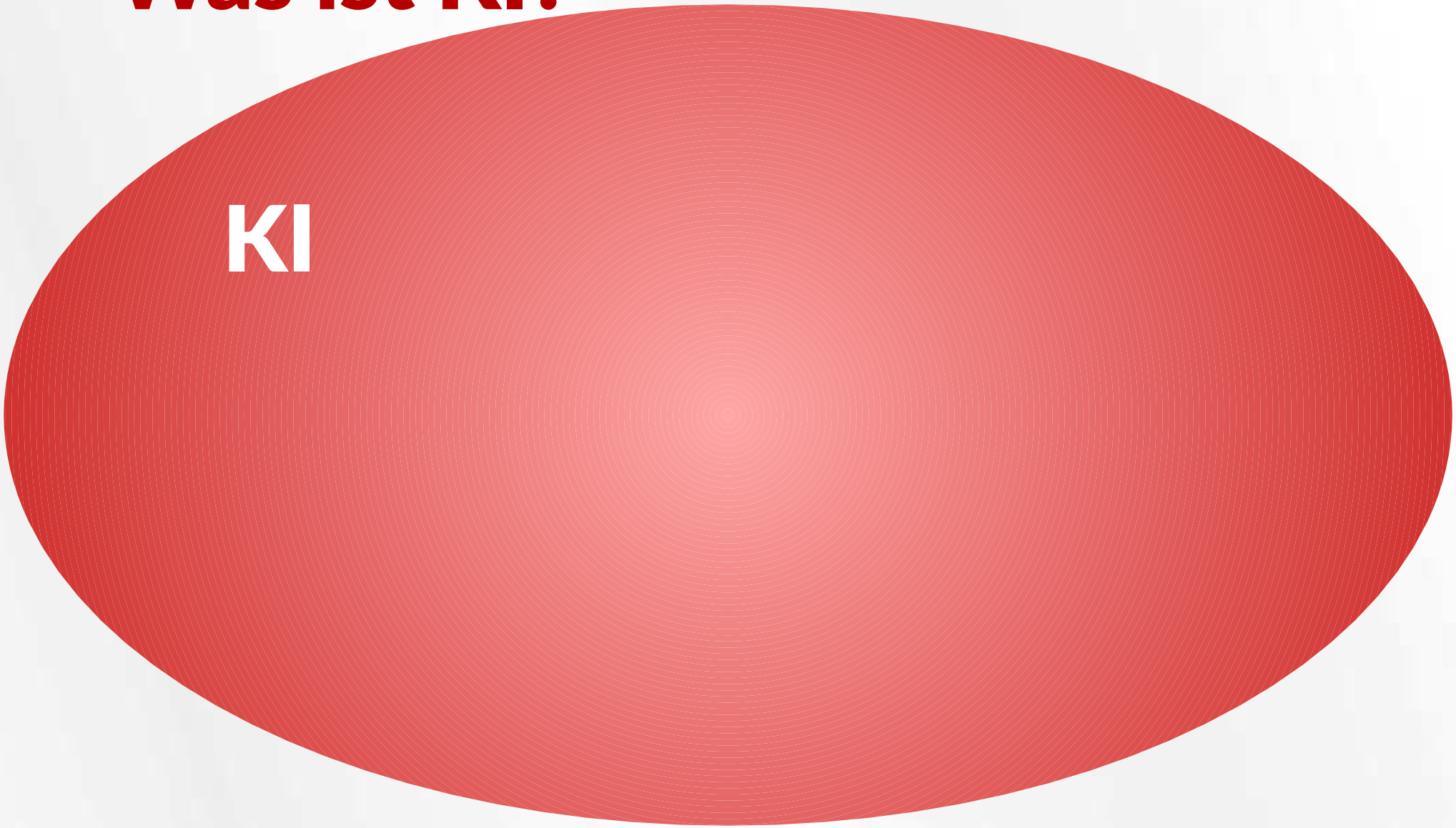
- mathematisches Modell 1957

- Selbstoptimierender Source Code (Gewichte für Akt.fkt.)

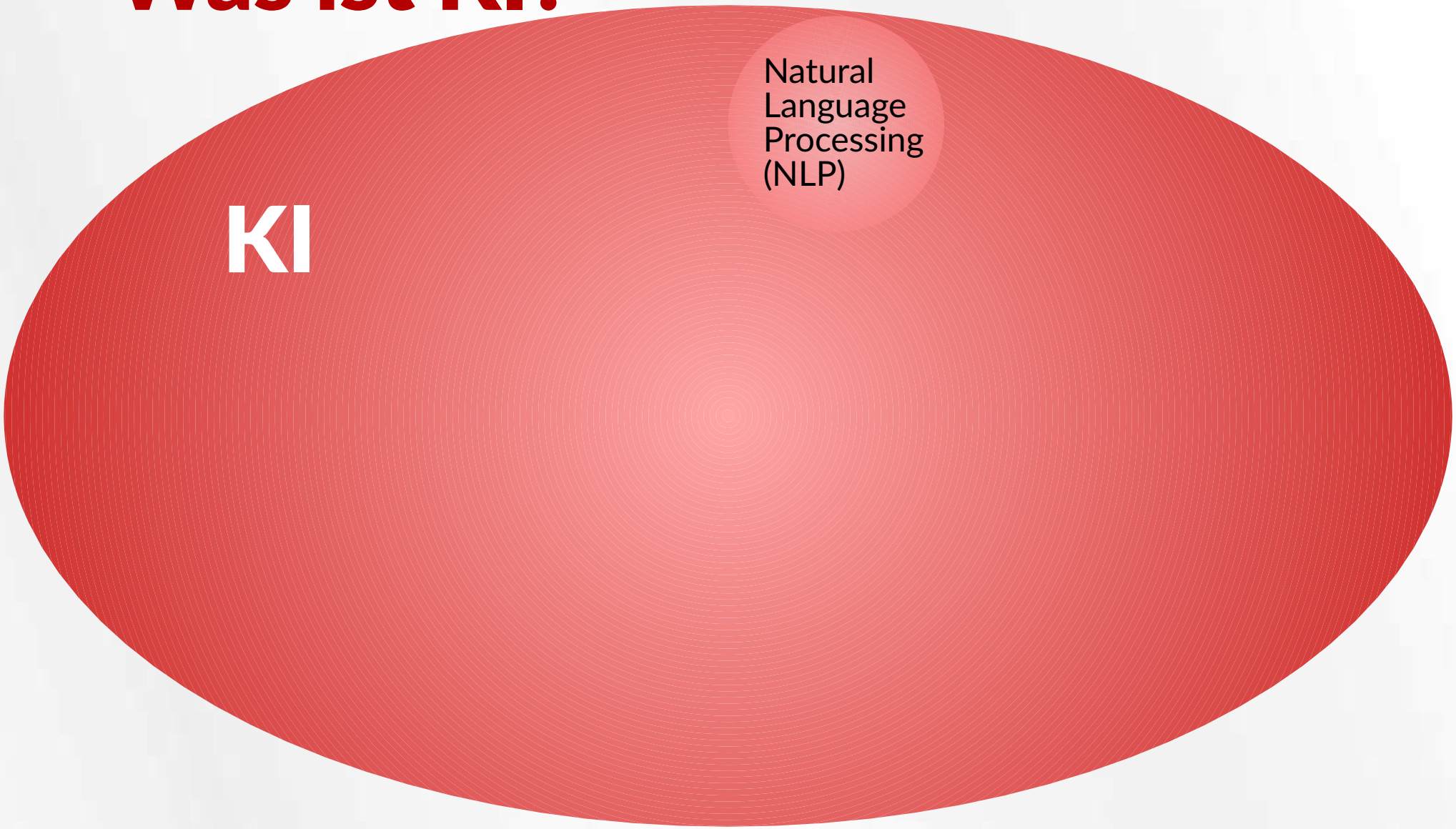
- mehr als ChatGPT!
  (das muss man verstehen, um sich schützen zu können!)

Was ist KI?

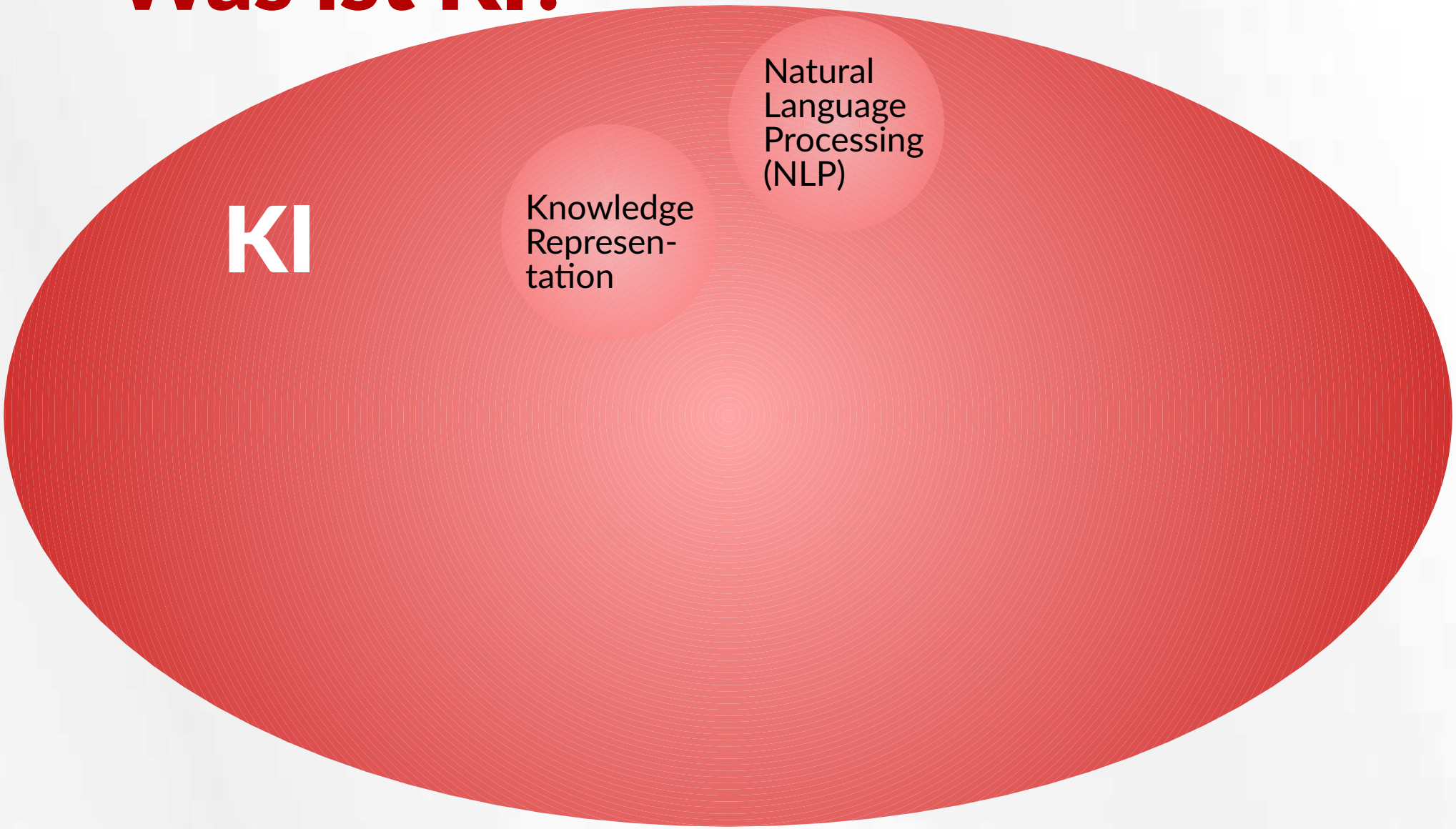
Was ist KI?



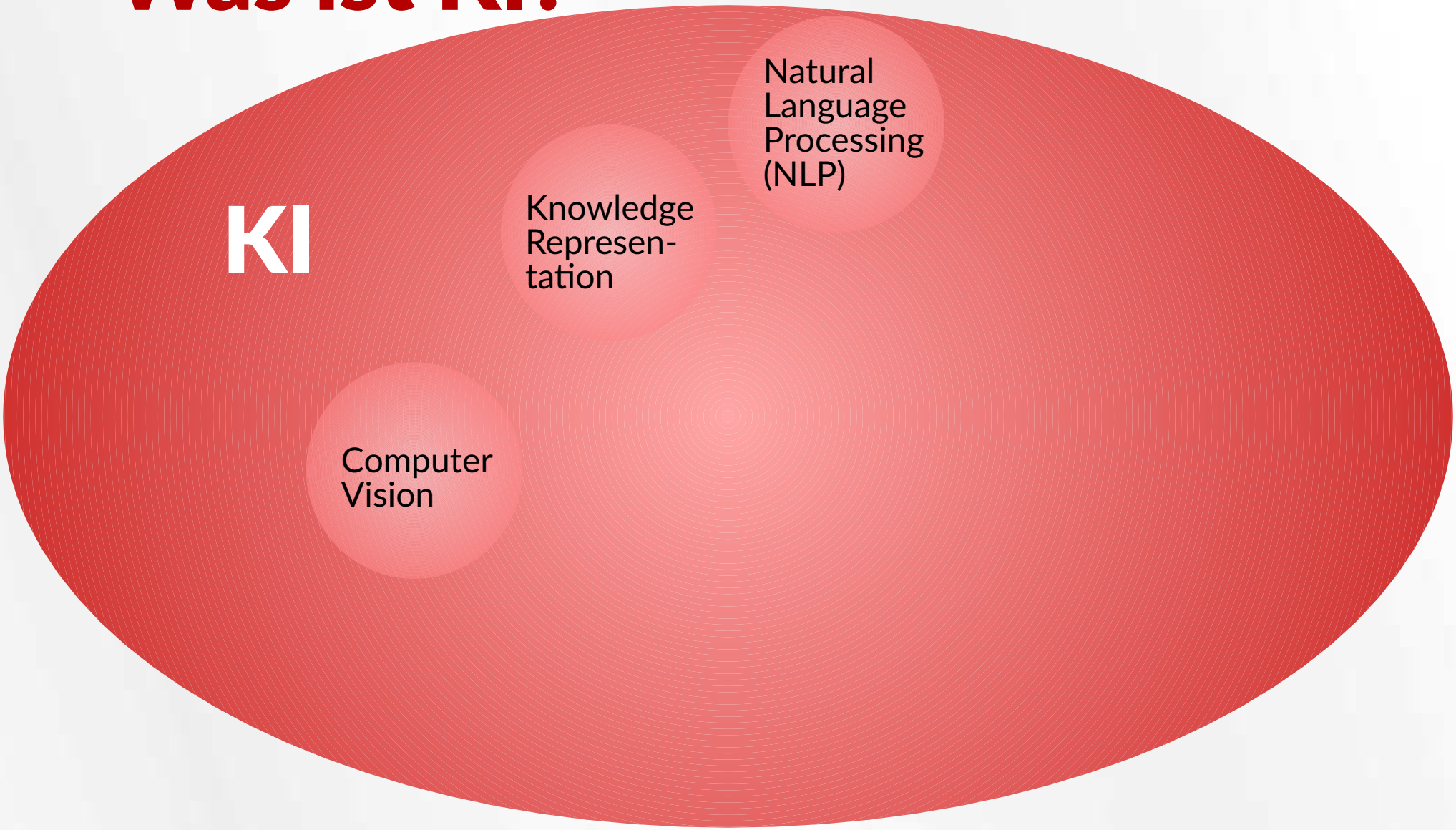
Was ist KI?



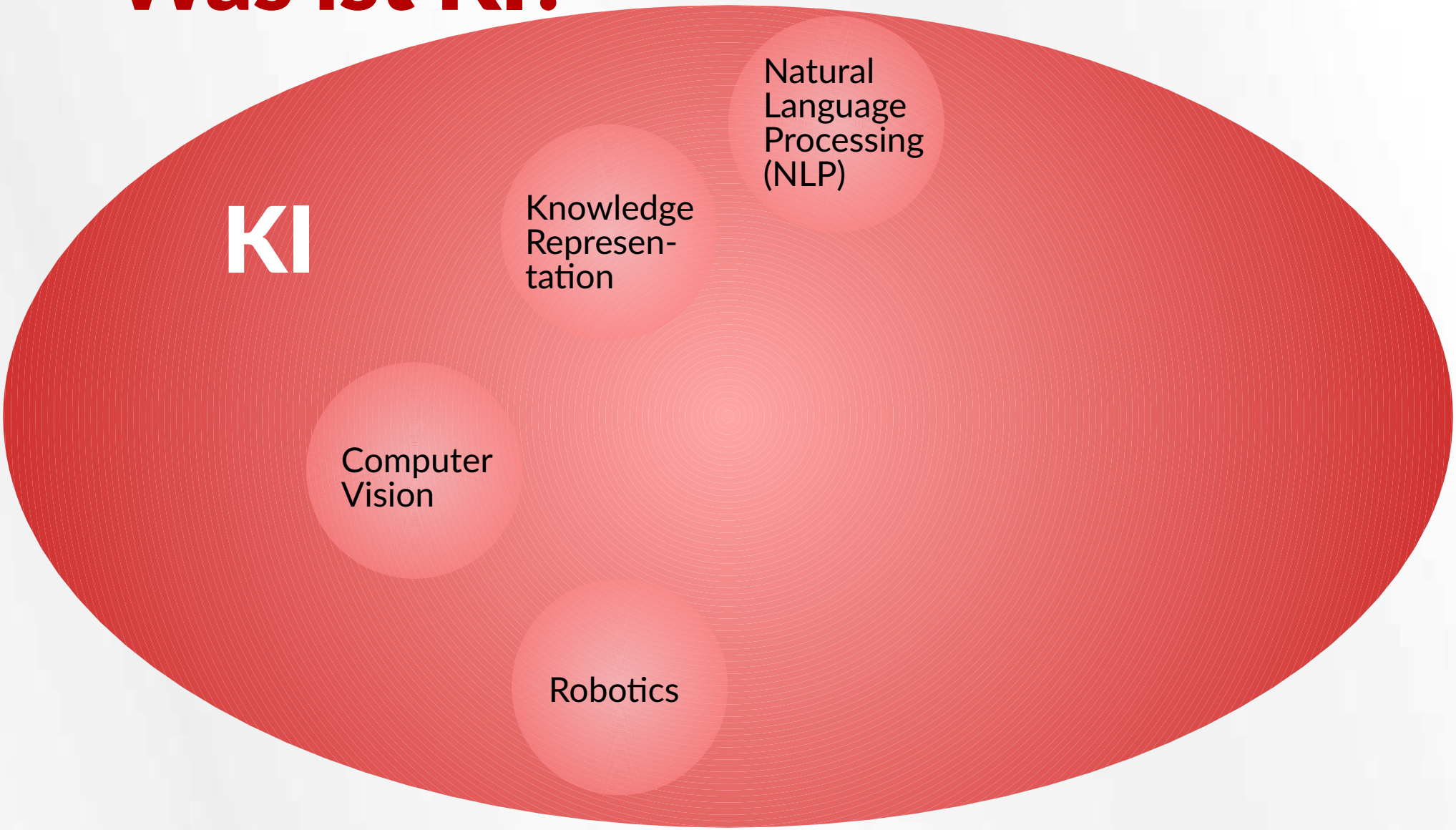
Was ist KI?


KI
Knowledge Represen-tation
Natural Language Processing (NLP)

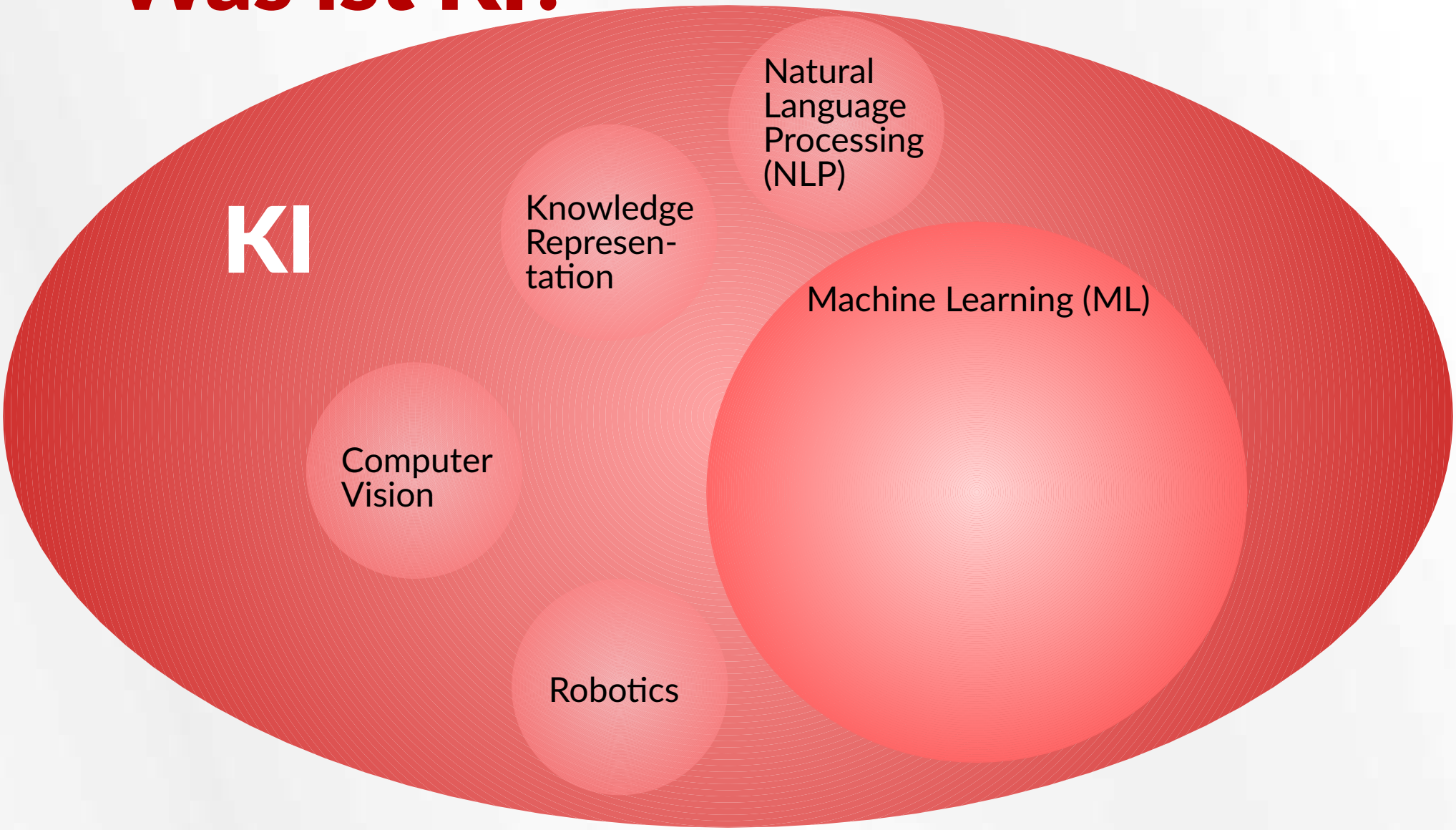
Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Computer Vision

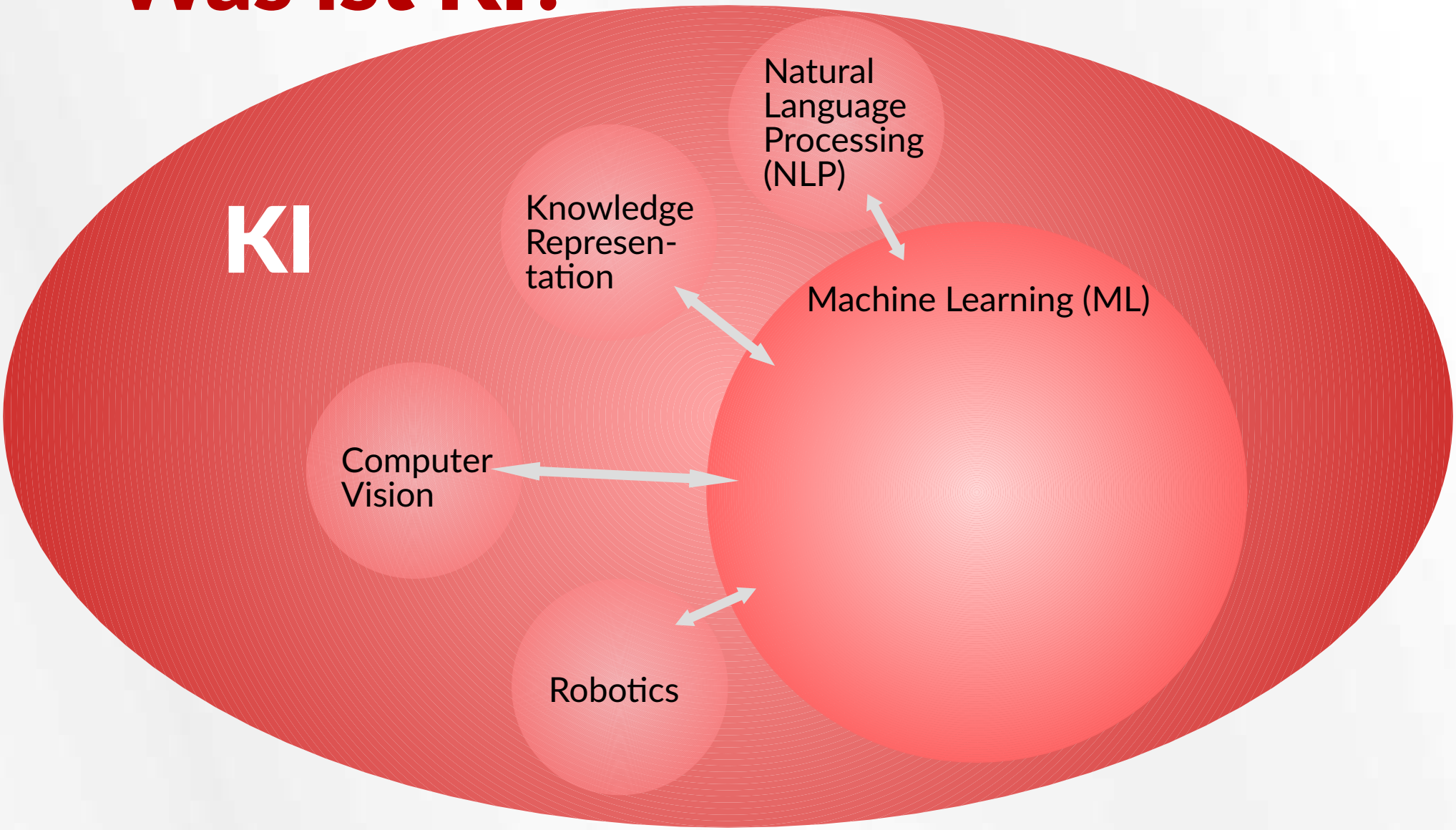
Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Computer Vision
Robotics

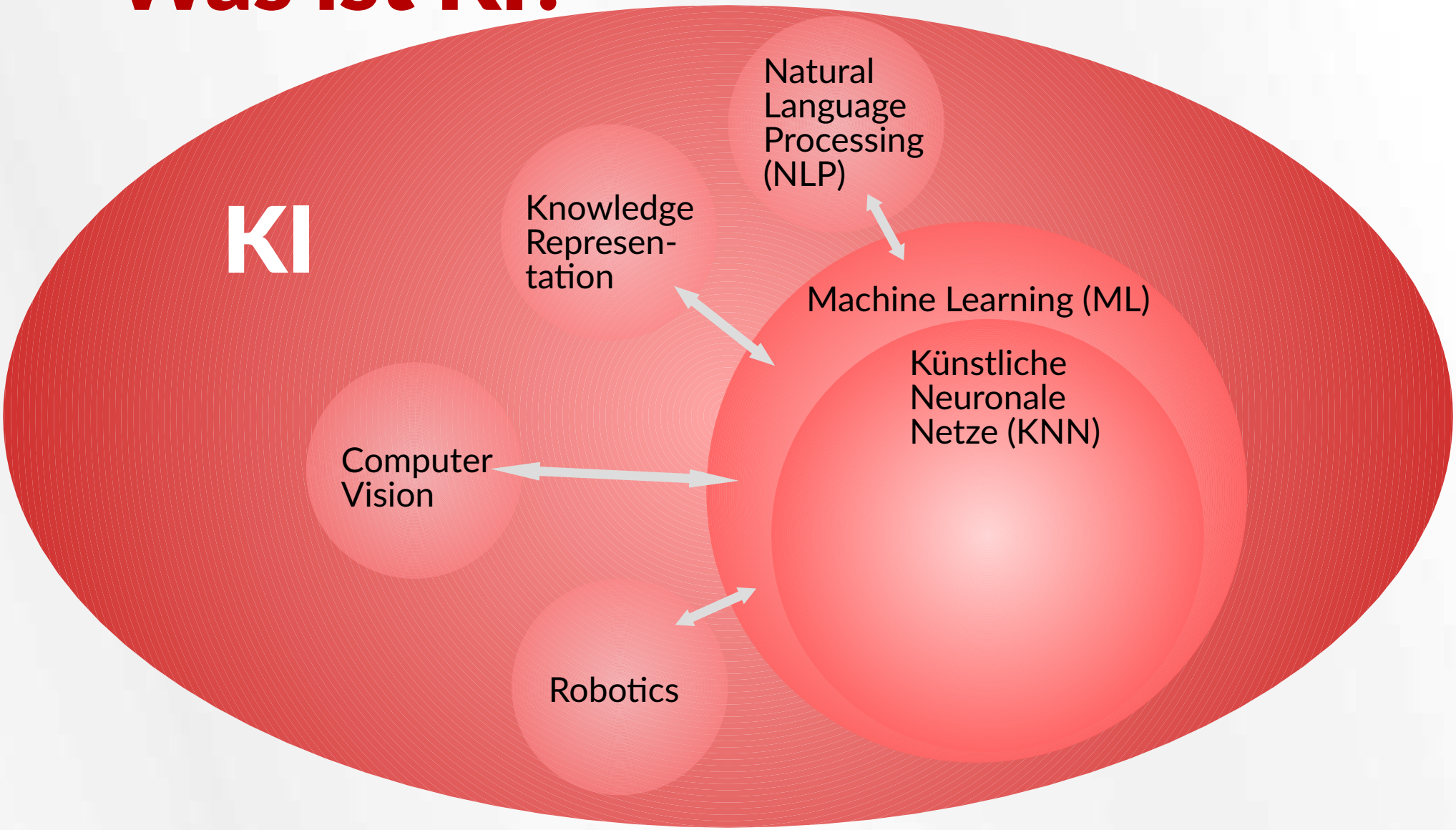
Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Machine Learning (ML)
Computer Vision
Robotics

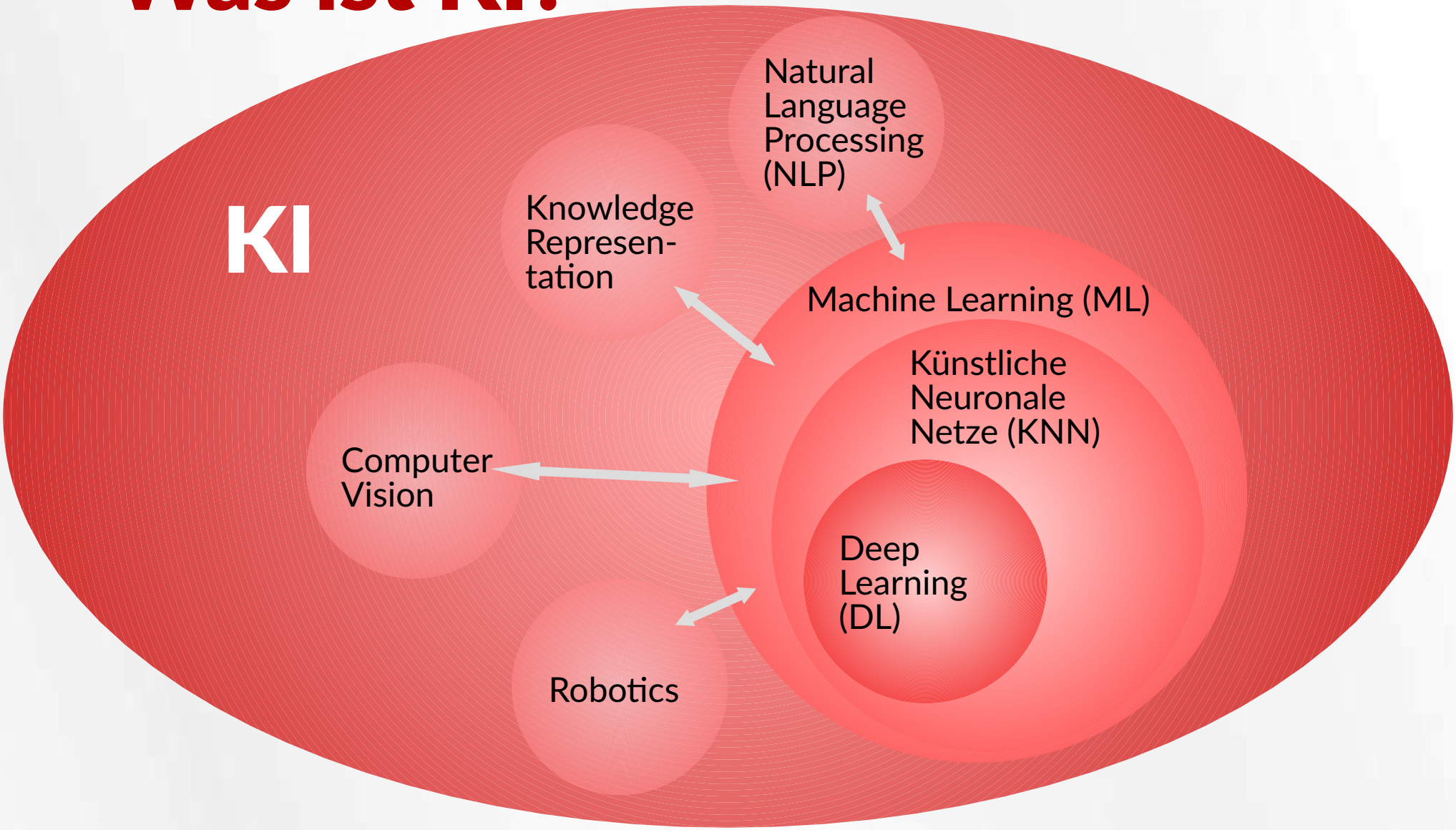
Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Machine Learning (ML)
Computer Vision
Robotics

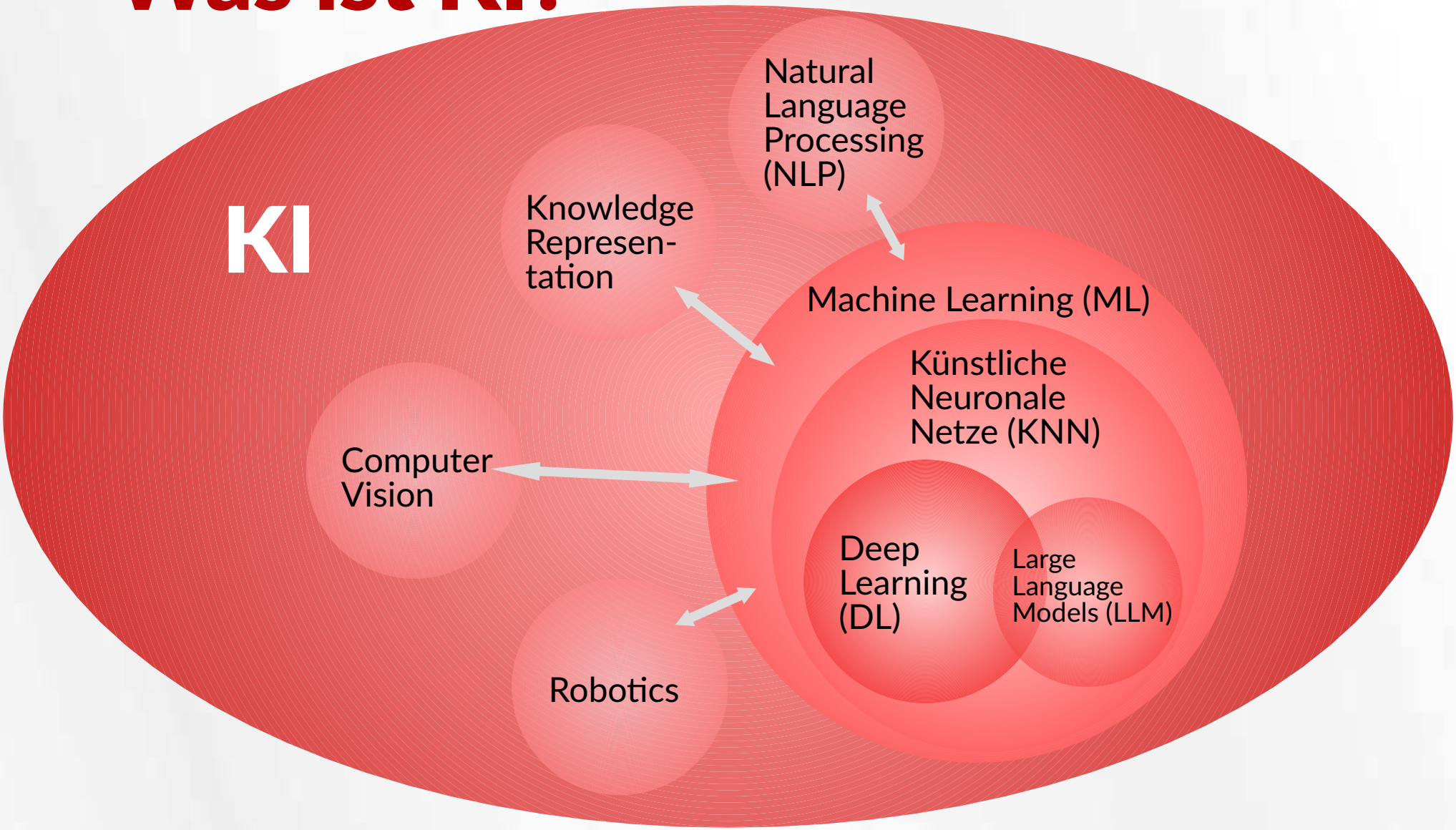
Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Machine Learning (ML)
Künstliche Neuronale Netze (KNN)
Computer Vision
Robotics

Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Machine Learning (ML)
Künstliche Neuronale Netze (KNN)
Computer Vision
Deep Learning (DL)
Robotics

Was ist KI?


KI
Natural Language Processing (NLP)
Knowledge Represen-tation
Machine Learning (ML)
Künstliche Neuronale Netze (KNN)
Computer Vision
Deep Learning (DL)
Large Language Models (LLM)
Robotics

KNN in Python (Bsp)

```
In [1]: # Codebasis aus dem Buch "Neuronale Netze programmieren mit Python", J. Steinwendner, R. Schwaiger, S. 188 ff.

        import numpy as np
        import matplotlib.pyplot as plt
        from random import choice

        # Die heaviside Stufenfunktion als Lambda Funktion
        heaviside = lambda x: 0 if x < 0 else 1

        # Training!
        def fit(iterations, training_data_set,w):
            errors = []
            weights = []
            for i in range(iterations):
                # Beispieldaten
                training_data = choice(training_data_set)
                x = training_data[0]
                y = training_data[1]

                # Gewichtete Summe mit nachgelagerter Stufenfunktion
                y_hat = heaviside(np.dot(w, x))
                # Fehler berechnen als Differenz zwischen gewünschtem und aktuellem Output
                error = y - y_hat
                errors.append(error)
                weights.append(w)

                # Lernen!
                w += error * x

            return errors, weights

        def main():
            # Trainingsdaten erstellen
            training_data_set = [
                (np.array([1,0,0]), 0),
                (np.array([1,0,1]), 1),
                (np.array([1,1,0]), 1),
                (np.array([1,1,1]), 1),
            ]
```

KNN in Python (Bsp)

```
    # Anfangsinitialisierung des Zufallgenerators wegen
    # Reproduzierbarkeit der Ergebnisse
    np.random.seed( 12 )

    w = np.zeros(3)

    # Die Anzahl der Iterationen nach gewünschter Präzision der KI
    iterations = 30

    # Trainieren!
    errors, weights = fit(iterations, training_data_set,w)

    w = weights[iterations-1]

    # Datenanzeigen
    fignr = 1
    plt.figure(fignr,figsize=(10,10))
    plt.plot(errors)
    plt.style.use('seaborn-whitegrid')
    plt.xlabel('Iteration')
    plt.ylabel(r"$|y - \hat y|$")
    plt.show()

# Hauptprogramm
main()
```

Beispiel: Hashcat

Beispiel: Hashcat

```
hashcat (v6.2.1) starting...

CUDA API (CUDA 11.3)
====================
* Device #1: NVIDIA GeForce RTX 2080 Ti, 10137/11264 MB, 68MCU

Hashes: 1 digests; 1 unique digests, 1 unique salts
Bitmaps: 16 bits, 65536 entries, 0x0000ffff mask, 262144 bytes, 5/13 rotates

Optimizers applied:
* Optimized-Kernel
* Zero-Byte
* Precompute-Init
* Early-Skip
* Not-Iterated
* Prepended-Salt
* Single-Hash
* Single-Salt
* Brute-Force
* Raw-Hash

Watchdog: Temperature abort trigger set to 90c

Host memory required for this attack: 1100 MB

e983672a03adcc9767b24584338eb378:00:hashcat

Session..........: hashcat
Status...........: Cracked
Hash.Name........: SolarWinds Serv-U
Hash.Target......: e983672a03adcc9767b24584338eb378:00
Time.Started.....: Sun May 23 11:43:13 2021 (1 sec)
Time.Estimated...: Sun May 23 11:43:14 2021 (0 secs)
Guess.Mask.......: ?a?a?a?a?a?at [7]
Guess.Queue......: 1/1 (100.00%)
Speed.#1.........: 24620.9 MH/s (32.19ms) @ Accel:32 Loops:1024 Thr:1024 Vec:1
Recovered........: 1/1 (100.00%) Digests
Progress.........: 316862720000/735091890625 (4.30%)
Rejected.........: 0/316862720000 (0.00%)
Restore.Point....: 0/857375 (0.00%)
Restore.Sub.#1...: Salt:0 Amplifier:35840-36864 Iteration:0-1024
Candidates.#1....: 4{,erat -> cyr ~}t
Hardware.Mon.#1..: Temp: 62c Fan: 31% Util:100% Core:1920MHz Mem:7000MHz Bus:16

Started: Sun May 23 11:43:12 2021
Stopped: Sun May 23 11:43:15 2021
```

Hashcat mit KI verstärken

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung
--regelbasierte Angriffe

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung
--regelbasierte Angriffe
--optimierte Speicherzugriffe

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
Optimierung durch GPU Cluster
--optimierte Algorithmen
--Wortlistenoptimierung
--regelbasierte Angriffe
--optimierte Speicherzugriffe

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung
Optimierung durch KI (DL)!
--regelbasierte Angriffe
--optimierte Speicherzugriffe

Hashcat mit KI verstärken

- nutzt dazu

--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung
--regelbasierte Angriffe
Optimierung durch KI (DL)!
--optimierte Speicherzugriffe

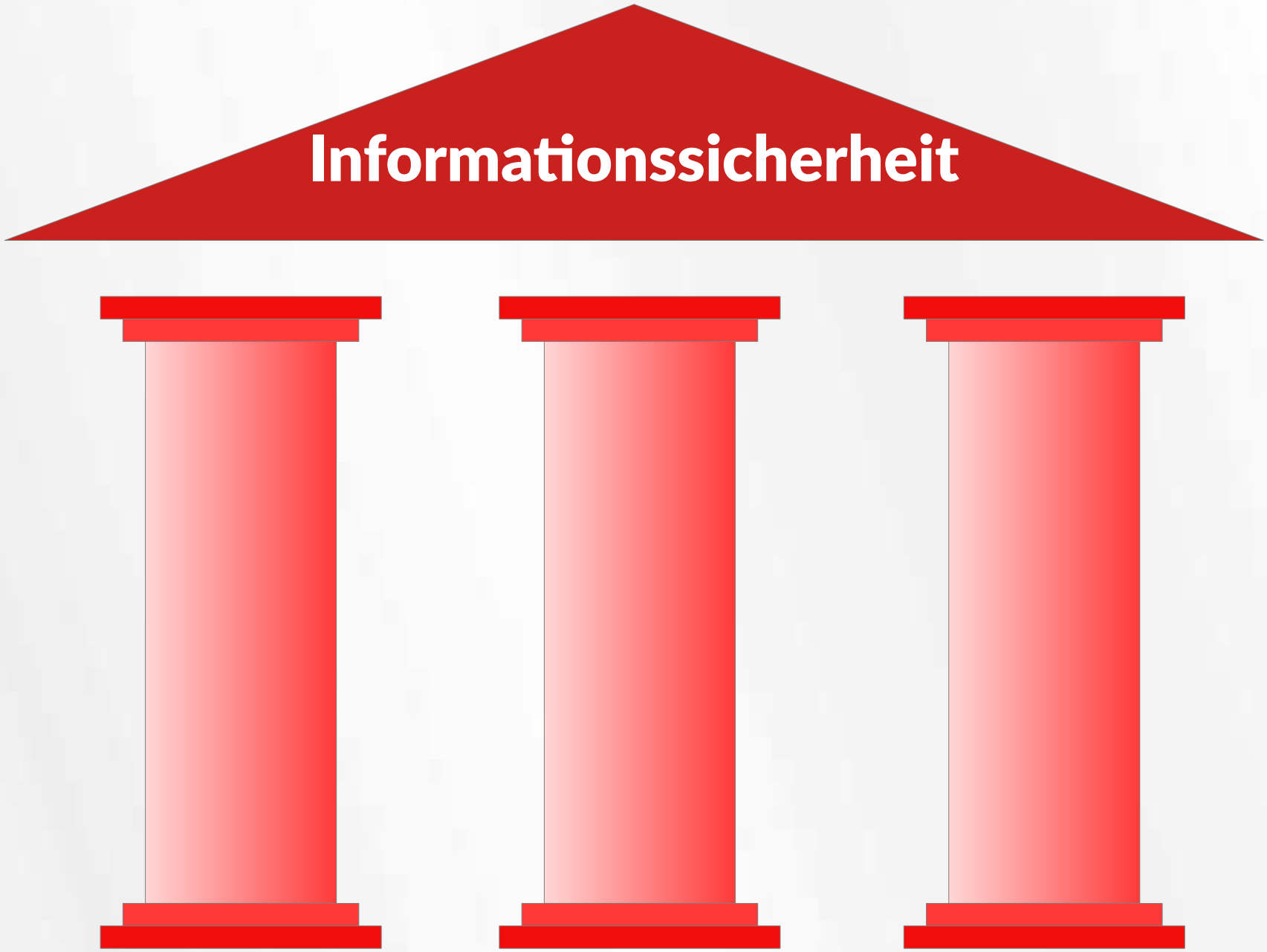
Hashcat mit KI verstärken

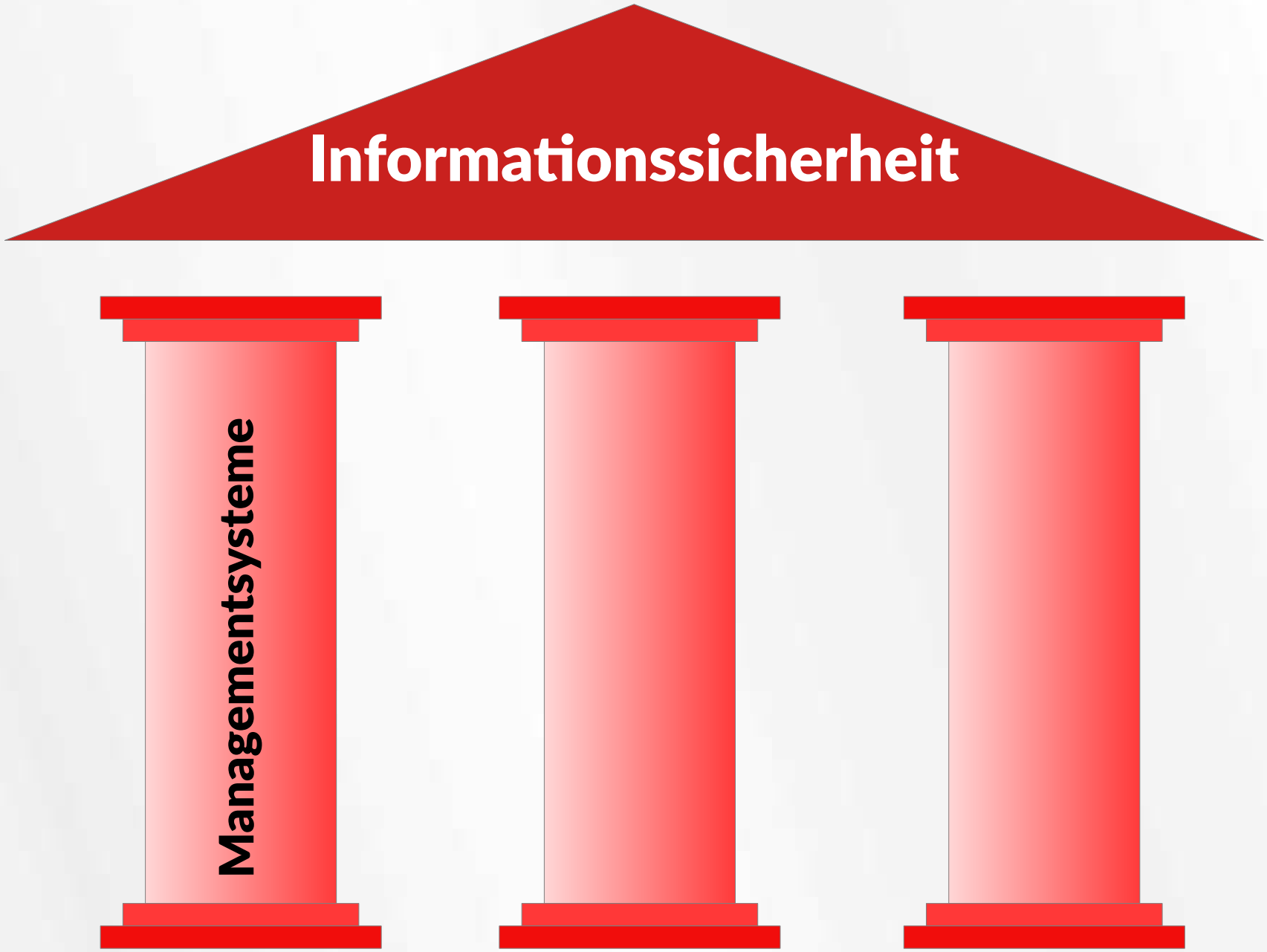
- nutzt dazu

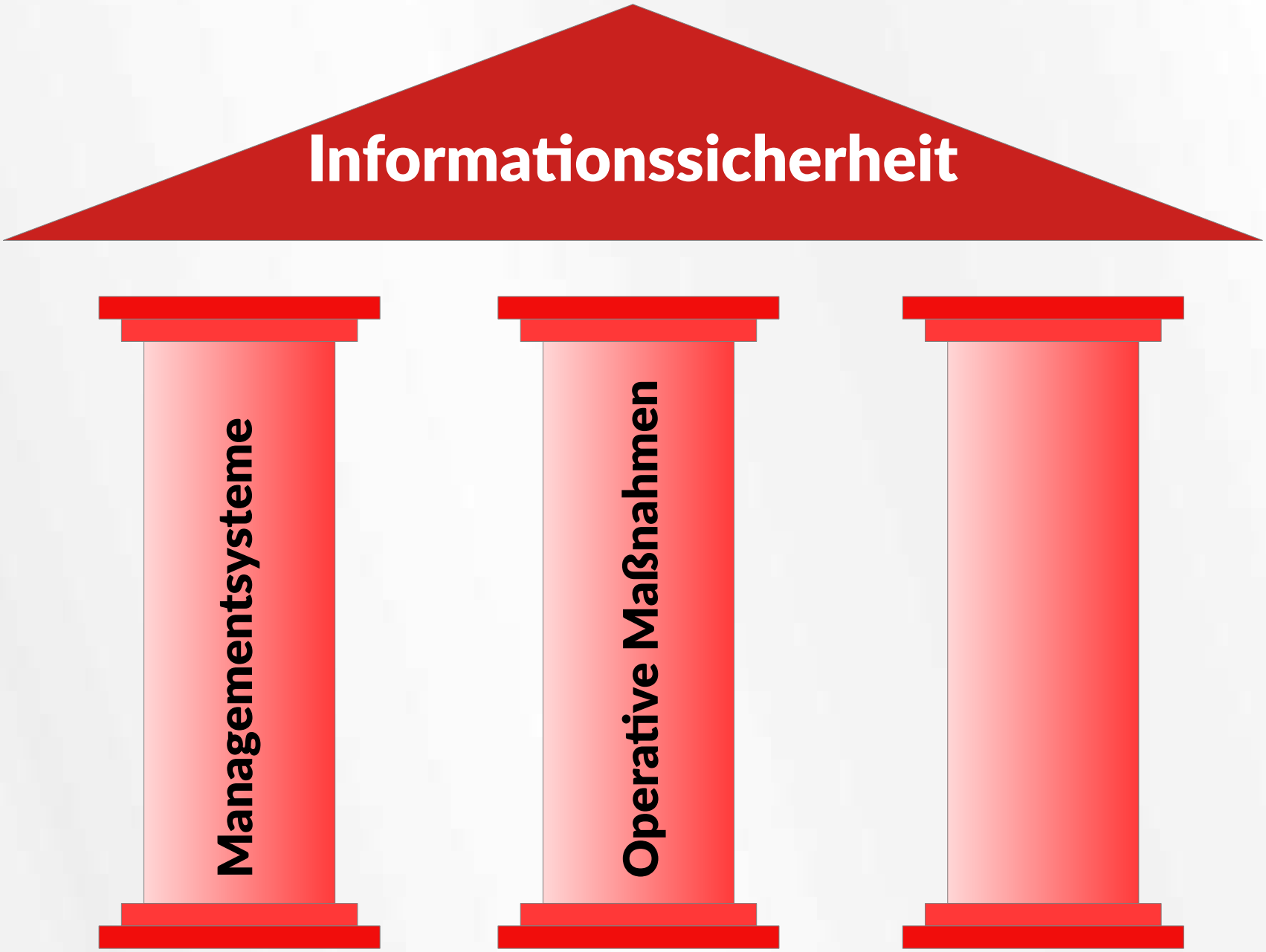
--Parallele Verarbeitung
--optimierte Algorithmen
--Wortlistenoptimierung
--regelbasierte Angriffe
--optimierte Speicherzugriffe
RAM Optimierung durch KI

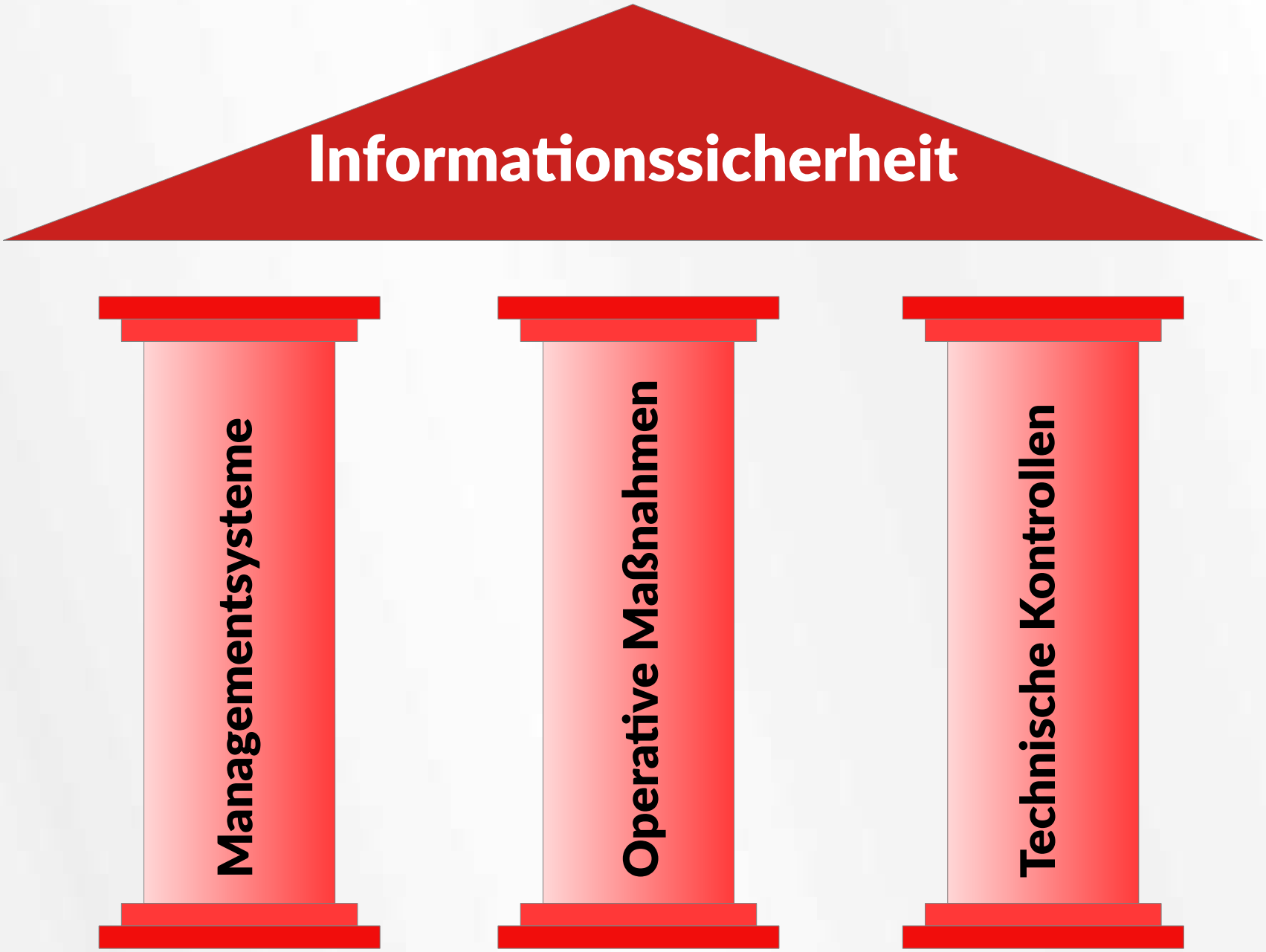
Und wie kann ich mein Unternehmen schützen?!

Und wie kann ich mein Unternehmen schützen?!


Informationssicherheit


Informationssicherheit
Managementsysteme


Informationssicherheit
Managementsysteme
Operative Maßnahmen


Informationssicherheit
Managementsysteme
Operative Maßnahmen
Technische Kontrollen

Managementsysteme

Managementsysteme

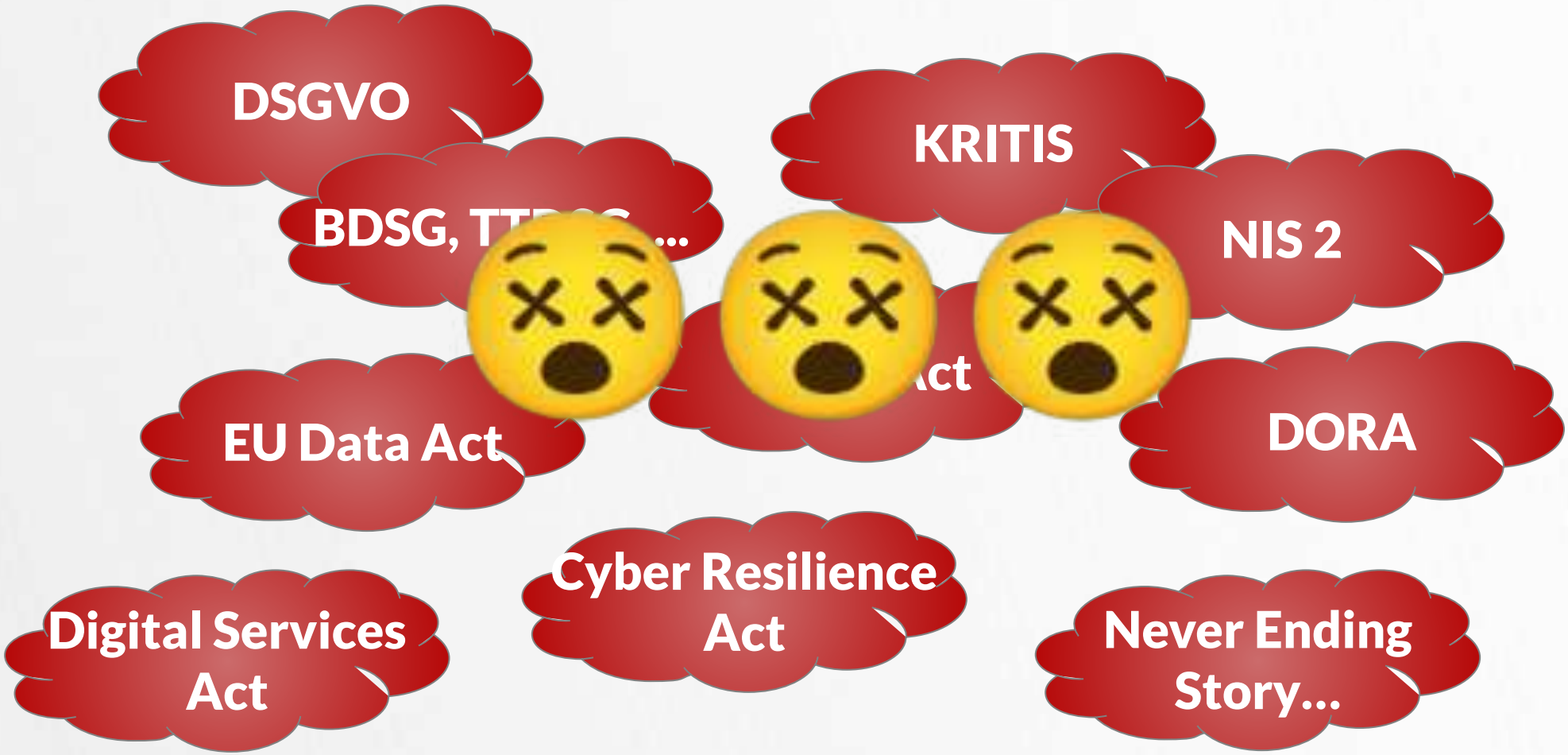
DSMS

ISMS

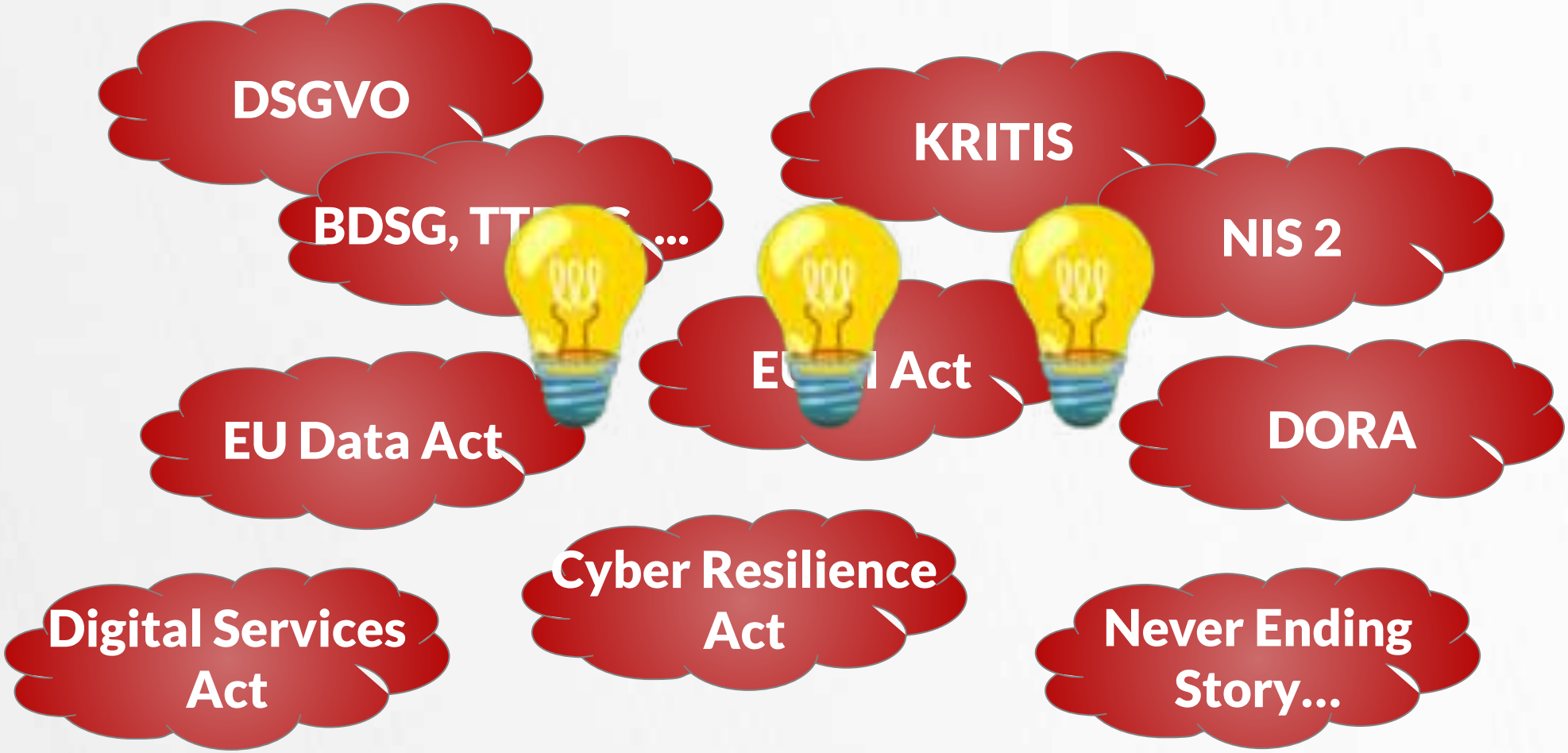
BCMS

Managementsysteme

Managementsysteme



Managementsysteme



Managementsysteme

Managementsysteme

Managementsysteme
↔
IT Strategie

Bsp:
- On Premise vs Cloud
- public vs private vs hybrid Cloud
- monolithisch vs. Micro Service
- Netzarchitektur & Schnittstellen

Operative Maßnahmen

vernünftige Back Up, Firewall, ...

Monitoring, Trending, Alerting

IDS, SIEM & SOC

Amnesische KI

Incident Response & Forensik

Technische Überwachung

Reconnaissance Scans

Pentests

Vulnerability Scans

Blue, Red & Purple Teams

Hüte

Hüte

Es gibt gute und starke Möglichkeiten, sich zu schützen!

Vielen Dank!

Folgt mir gerne auf LinkedIn:

...Heute schon gehackt!?