



**Westfälische
Hochschule**

Gelsenkirchen Bocholt Recklinghausen
University of Applied Sciences

Künstliche Intelligenz *und* Cyber-Sicherheit

Prof. Dr. (TU NN)

Norbert Pohlmann

Institut für Internet-Sicherheit – if(is)
Westfälische Hochschule, Gelsenkirchen
<http://www.internet-sicherheit.de>

if(is)
internet-sicherheit.

- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

■ **Einordnung**

(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)

■ **Maschinelles Lernen**

(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)

■ **Künstliche Neuronale Netze**

(Idee, KNN, Deep Learning, ...)

■ **Anwendungen KI und Cyber-Sicherheit**

(Alert-System für Online-Banking, passive Authentifikation, ...)

■ **Angriffe auf maschinelles Lernen**

(Idee, Trainingsdaten, Verkehrszeichen, ...)

■ **Künstliche Intelligenz und Cyber-Sicherheit**

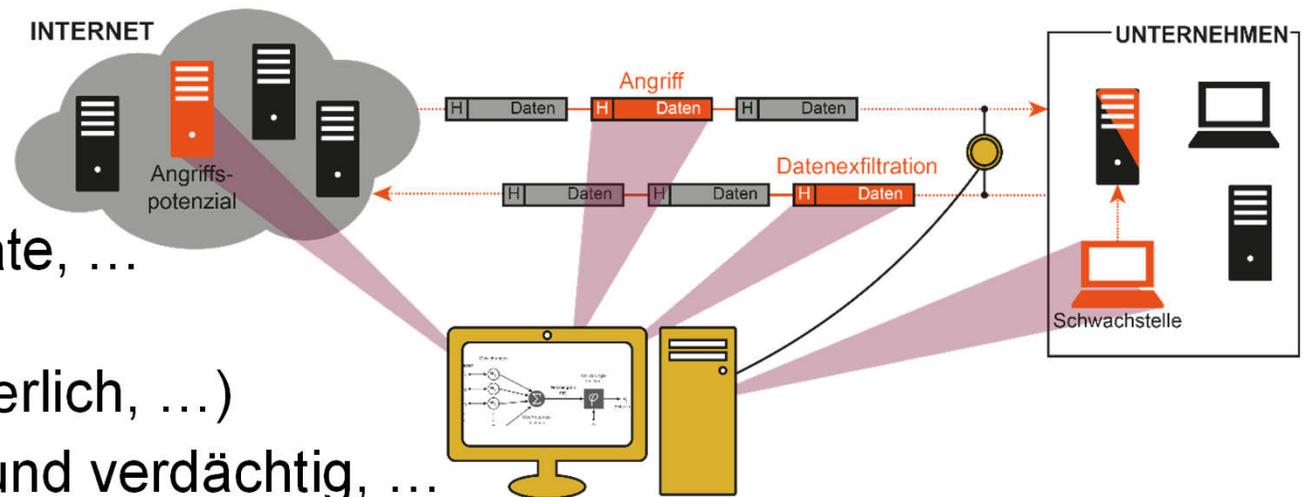
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)

■ **Ergebnis und Ausblick**

Künstliche Intelligenz → und Cyber-Sicherheit

- Erhöhung der **Erkennungsrate von Angriffen**

- Netzwerk, IT-Endgeräte, ...
- adaptive Modelle (selbständig, kontinuierlich, ...)
- Unterschied: normal und verdächtig, ...



- **Unterstützung / Entlastung von Cyber-Sicherheitsexperten**
(von denen wir nicht genug haben)

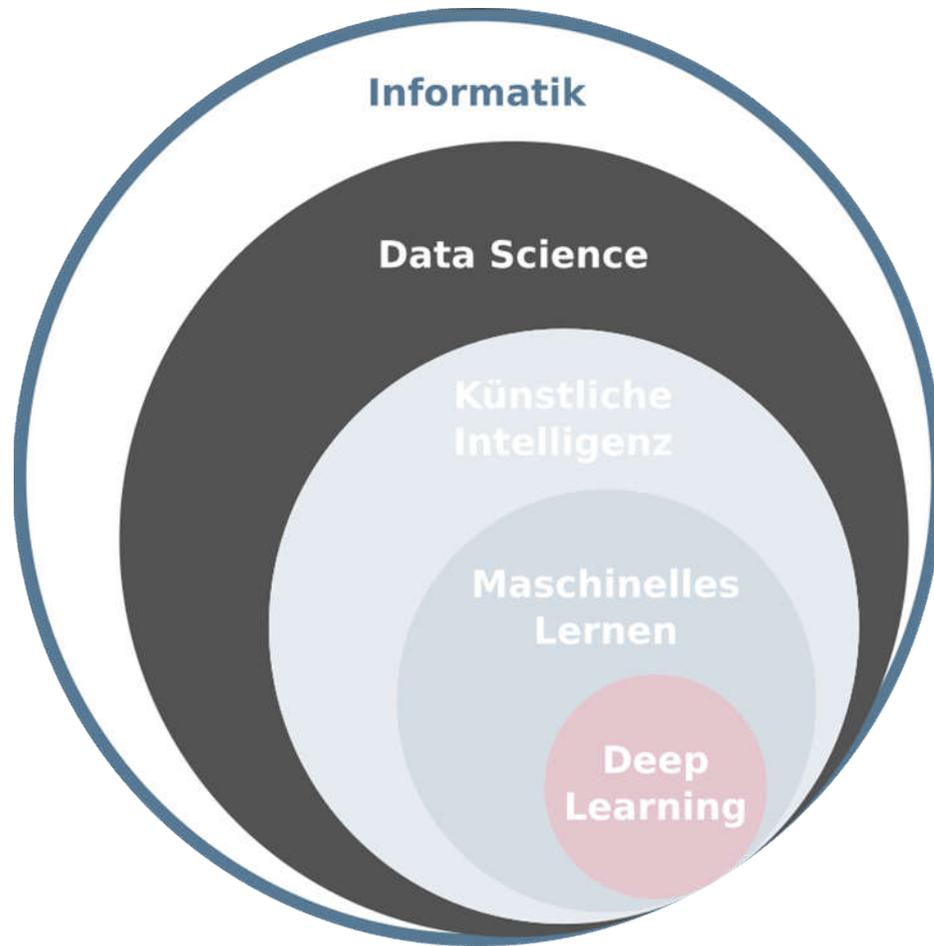
- Erkennen von **wichtigen** sicherheitsrelevanten Ereignissen (*Priorisierung*)
- **(Teil-)Autonomie** bei Reaktionen, ... Resilienz, ...

- **Verbesserungen** von bestehenden **Cyber-Sicherheitslösungen**

- KI leistet einen Beitrag zu einer erhöhten Wirkung und Robustheit
- Z.B.: Risikobasierte und adaptive Authentifizierung



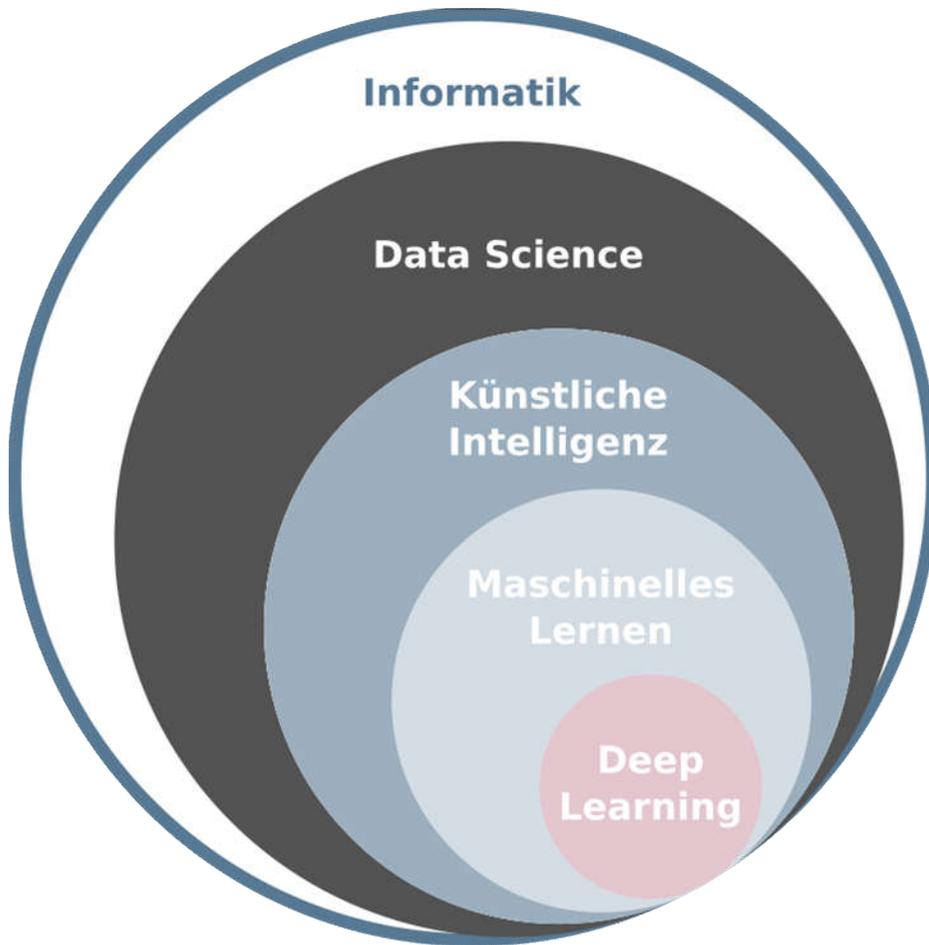
Einordnung → Data Science



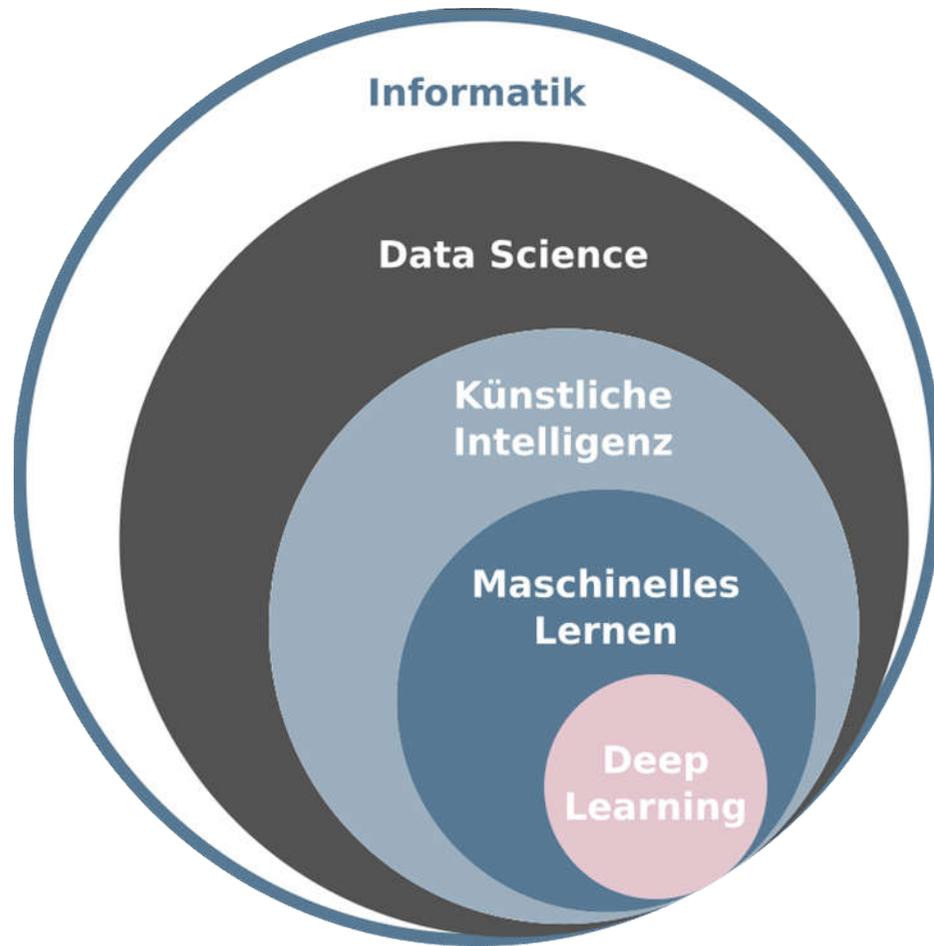
- **Data Science** bezeichnet generell die **Extraktion von Wissen** aus Daten.
- **Da es immer mehr Daten gibt, kann auch immer mehr Wissen daraus abgeleitet werden.**
(Wichtig: Daten müssen Informationen erhalten)
- **Abgrenzung zur künstlichen Intelligenz:**
 - Statistiken
 - Kennzahlen
 - Datenerhebung

Einordnung → Künstliche Intelligenz

- **Künstliche Intelligenz** ist ein Fachgebiet der Informatik
- setzt intelligentes Verhalten in Algorithmen um
- (Ziel)
 - **automatisiert** „**menschenähnliche Intelligenz**“ nachzubilden.
 - **Starke „Künstliche Intelligenz“** (Zukunft)
 - Superintelligenz
 - **Singularität** („**Maschine**“ **verbessert sich selbst**, sind **intelligenter als Menschen**)



Einordnung → Maschinelles Lernen

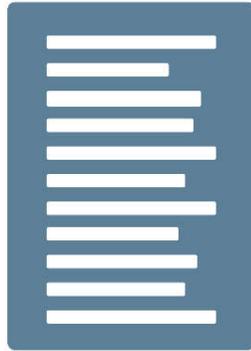


- **Maschinelles Lernen** ist ein Begriff für die „künstliche“ **Generierung von Wissen aus Erfahrung** (in Daten) durch Computer.
- In **Lernphasen** lernen entsprechende ML-Algorithmen aus Beispielen (*alte Daten*) **Muster und Gesetzmäßigkeiten**.
- Daraus erstehende Verallgemeinerungen können auf *neue Daten* angewendet werden.
- **Schwache „Künstliche Intelligenz“** (wird heute erfolgreich umgesetzt)

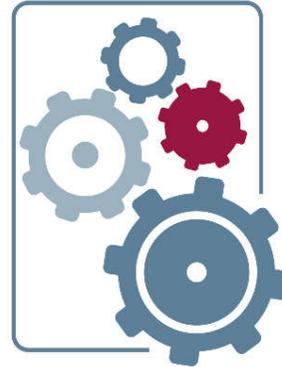
Maschinelles Lernen

→ Workflow

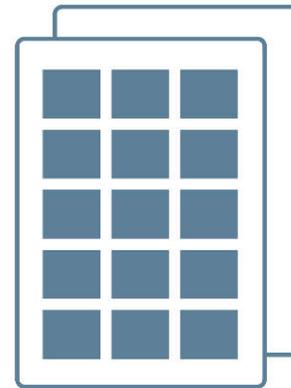
Eingabedaten



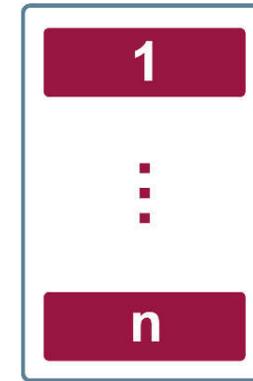
Algorithmus



Ergebnisse



Verwendung



Eingangsdaten

Qualität: Inhalt, Vollständigkeiten, Repräsentativität, ... Aufbereitung

Algorithmen (ML)

Support-Vector-Machine (SVM), k-Nearest-Neighbor (kNN), ... Deep Learning

Ergebnisse

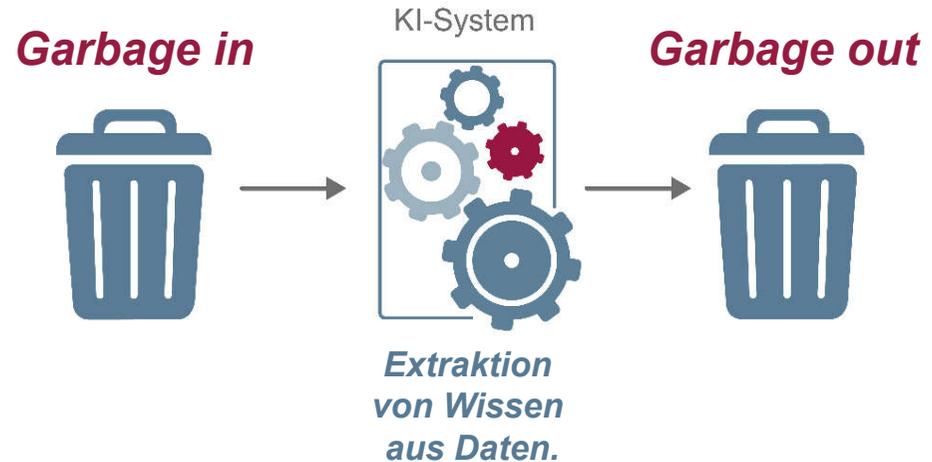
Ergebnisse aus der Verarbeitung (Algorithmus) der Eingangsdaten ...

Verwendung

Die Anwendung entscheidet, wie Ergebnisse verwendet werden (*Vertrauen*).

Vertrauenswürdigkeit → Qualität der Daten

Paradigma



Standards für die Datenqualität:

- Inalthöhe der Daten und Korrektheit
- Nachvollziehbarkeit (Datenquellen)
- Vollständigkeit und Repräsentativität
- Verfügbarkeit und Aktualität

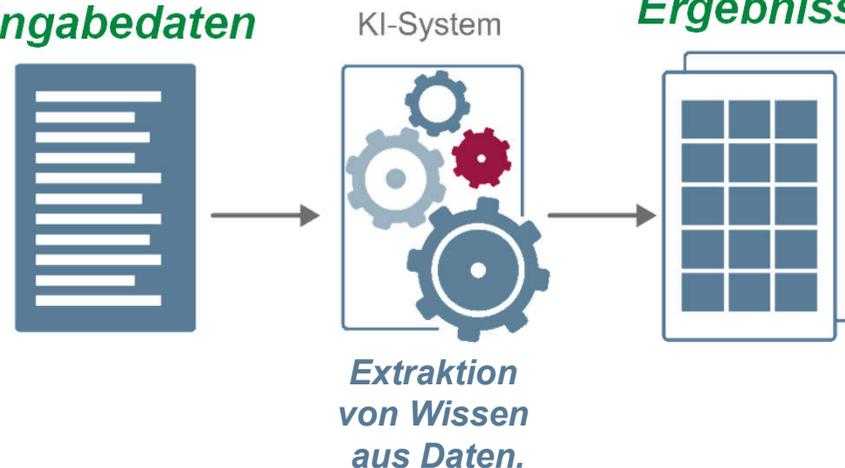
Qualitativ hochwertige und sichere Sensoren motivieren

hohe Datenqualität der Eingabedaten

qualitative, vertrauenswürdige Ergebnisse

Weitere Aspekte zur Erhöhung der Qualität:

- Datenpools etablieren
- Austausch von Daten fördern
- Interoperabilität schaffen
- Open Data Strategie puschen

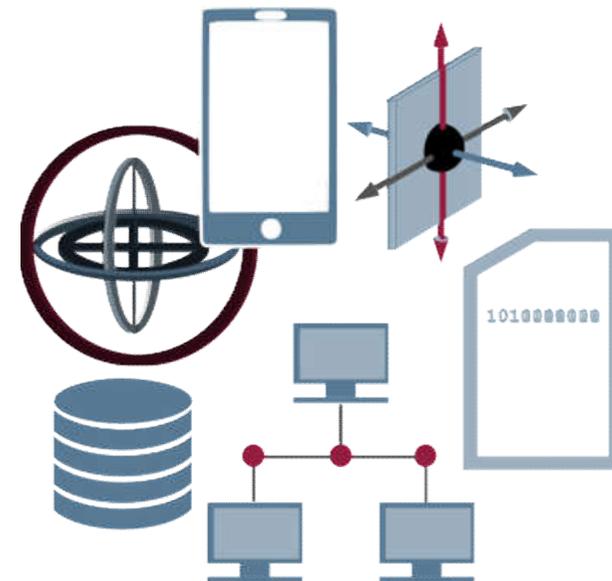


Erfolgsfaktoren – KI / ML

→ Eingabedaten

Erfolgsfaktor: Immer mehr vorhandene Daten

- **Smartphone, SmartWatch** (körpernah, personenorientiert)
 - Lage- und Beschleunigungssensoren, Nutzereingaben, Benutzerverhalten
- **Computer**
 - Nutzereingaben, Benutzerverhalten, Log Daten
- **Netzwerke, Netzwerkkomponenten (Router, Firewall, ...)**
 - Protokolldaten, Log Daten
- **Web-Dienste**
 - Benutzerverhalten, ...
- **IoT (Internet of Things)**
 - Sensorik und Aktorik
- **Auto, ...**



Erfolgsfaktoren – KI / ML

→ Leistungsfähige IT und Algorithmen

Erfolgsfaktor: **Leistungsfähigkeit** der IT-Systeme

- **enorme Steigerung** (CPU, RAM, ...) 20 CPU Kerne, 64 GB Arbeitsspeicher, 1 TB SSD, usw. Spezial-Hardware: GPUs, FPGA, TensorFlow PU (TPU),...
... Parallelisierung, Kommunikationsgeschwindigkeiten, spezielle Software-Frameworks, ...
- **leistungsfähige Cloud-Lösungen**, wie Amazon Web Services, Microsoft Azure, Google Cloud Platform und die IBM Cloud.

Erfolgsfaktor: **Algorithmen**

- Immer **bessere Algorithmen** (viel als OpenSource)
- Immer **mehr Erfahrungen** mit dem Umgang
- Immer **einfacherer Zugang** zu den Technologien und Diensten
- Beispiele: Support-Vector-Machine (SVM), k-Nearest-Neighbor (kNN), k-Means-Algorithmus, Hierarchische Clustering-Verfahren, Convolutional Neural Network

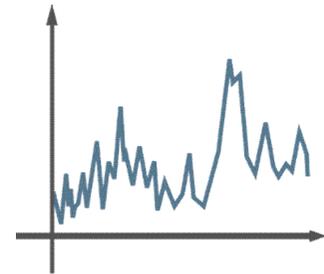
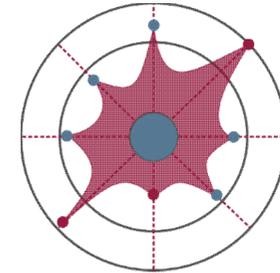


Künstliche Intelligenz

→ Ergebnisse und Verwendung

Ergebnisse sind **Modelle** zu den gelernten Eingabedaten

- **Nutzung** der Modelle führt zur konkreten **Anwendung**, z.B.:
 - **Klassifizierung** der Eingangsdaten, zur **Erkennung von Angriffen**
 - **Numerische Werte**, wie Hinweise zur **Verbesserung eines Produkts**
 - **Binäre Werte**, wie eine **erfolgreiche biometrischer Authentifizierung**

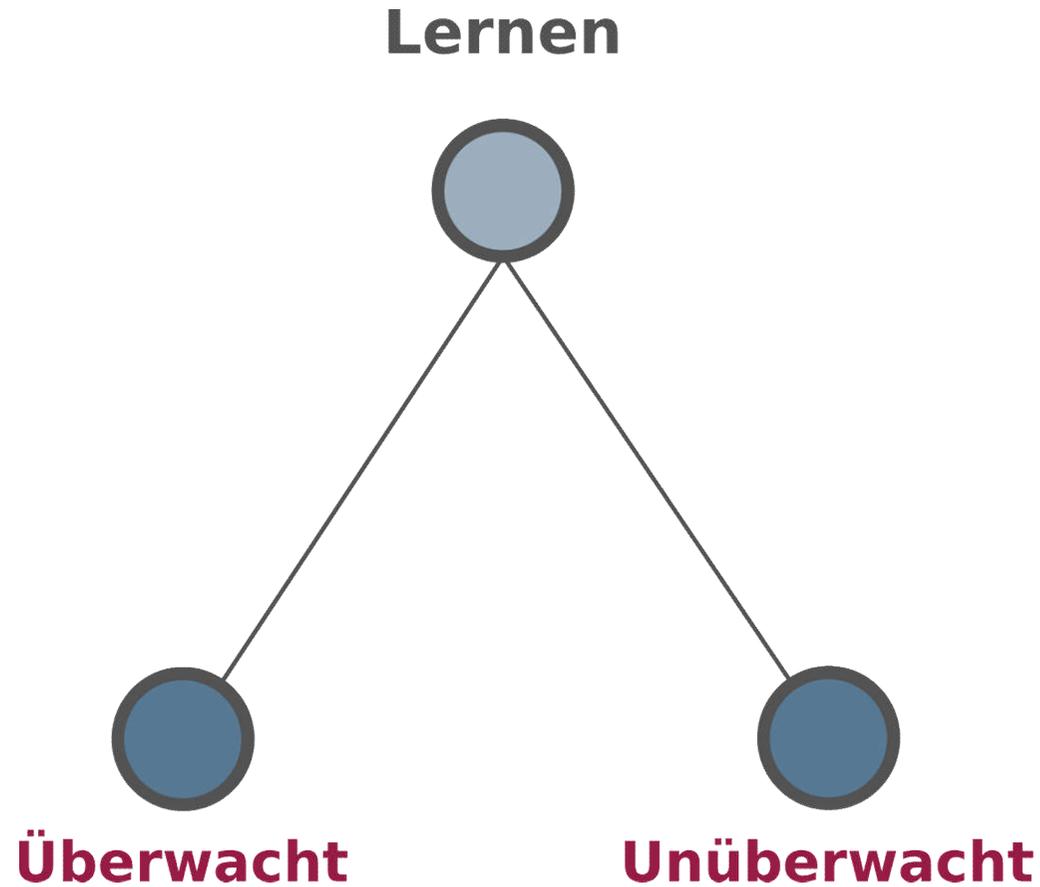


Verwendung: Policy, wie die Ergebnisse genutzt werden sollen.

- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Maschinelles Lernen

→ Kategorien des Lernens



ML-Algorithmus

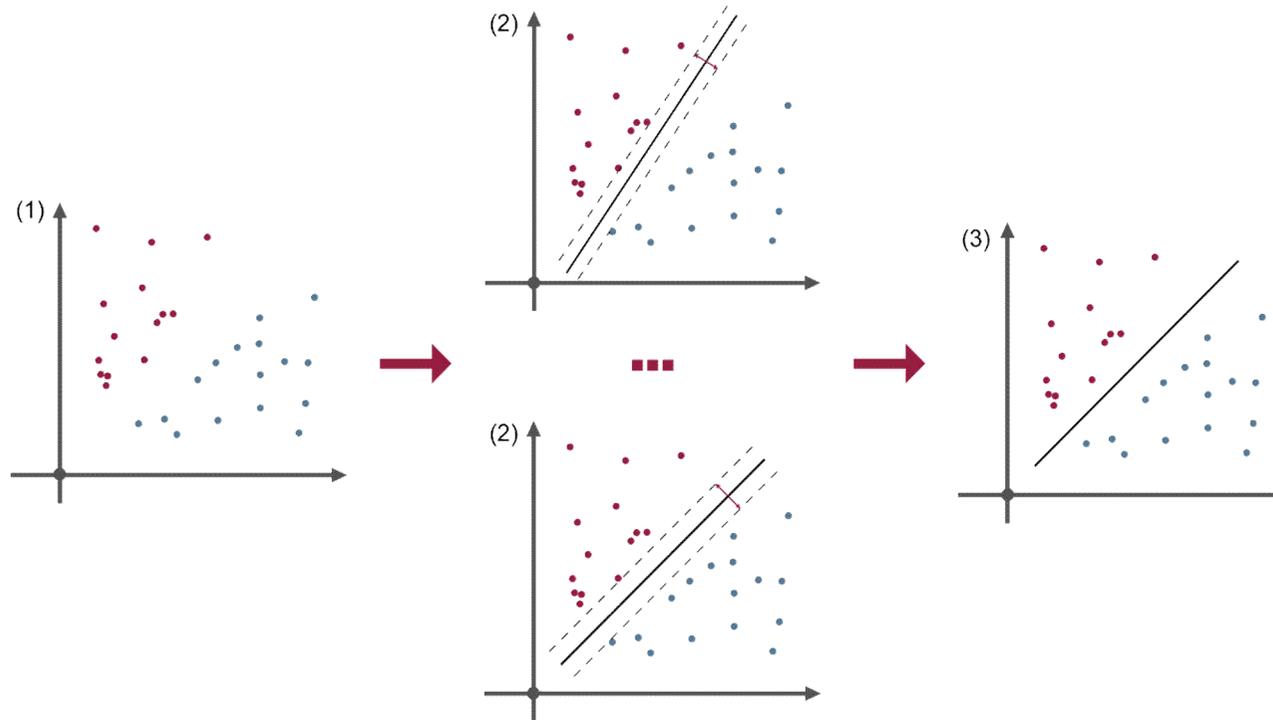
→ Überwachtes Lernen

- Ziele des überwachten Lernens
 - **Regression:** Vorhersagen von numerischen Werten
 - **Klassifizierung:** Einteilung von Daten in Klassen
- Beispiel: Erkennung von Spam-Mails
- Eingabedaten enthalten **erwartete Ergebnisse**
- **Einteilung der Daten in Trainings- und Testmengen**
(*kontinuierlich* lernen)
- Ziel: Selbständig Ergebnisse generieren
- **ML-Algorithmus, z.B.:**
 - Support-Vector-Machine (SVM)
 - k-Nearest-Neighbor (kNN)

ML-Algorithmus

→ Support-Vector-Machine (SVM)/Training

2-Dimensional



■ Input-Daten (1):

- bereits klassifizierte **Daten**
- **Abstandsmaß**

■ ML-Algorithmus (2):

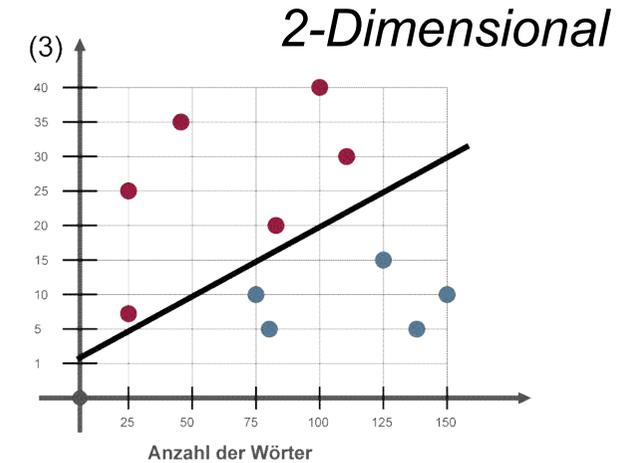
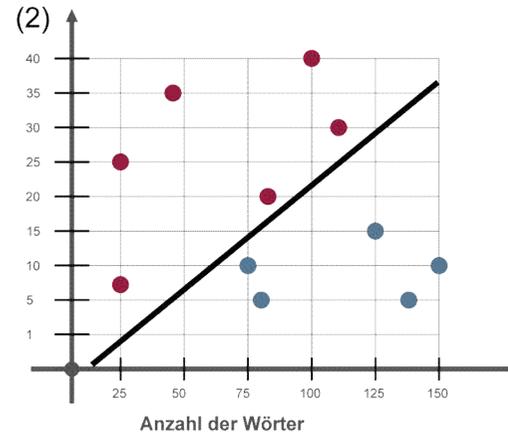
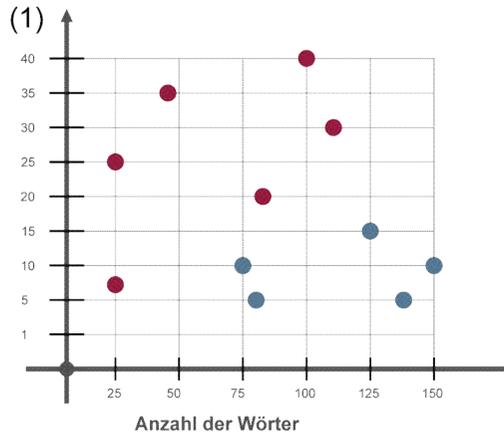
- **Ermitteln** von Geraden zur Trennung der Daten
- **Bewertung** durch Abstand zu den Punkten
- **Wahl** der Geraden mit maximalem Abstand zu beiden Klassen

■ Output (3):

- Gerade als **Modell** zur Klassifizierung

ML-Algorithmus

→ SVM - Beispiel Training (Spam)E-Mail



„Wissen aus Erfahrung“

Anzahl Wörter	25	25	47	75	79	82	100	110	125	140	150
Anzahl Wörter in Großbuchstaben	7	25	35	10	5	20	40	30	15	5	10
Spam-E-Mail	ja	ja	ja	nein	nein	ja	ja	ja	nein	nein	nein

■ Input-Daten (1):

- E-Mails mit entsprechender Klassifikation
Spam / kein Spam

■ ML-Algorithmus (2):

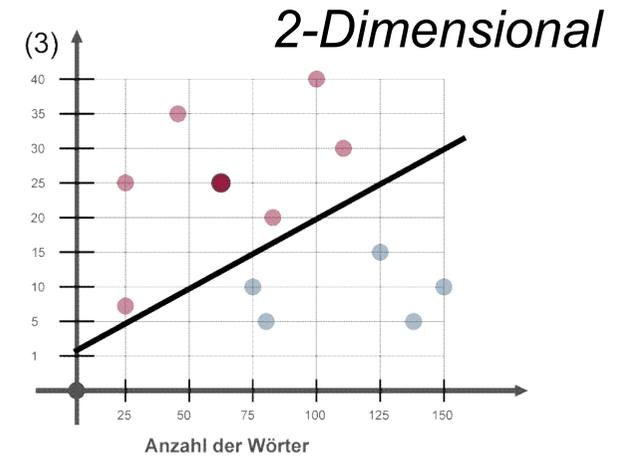
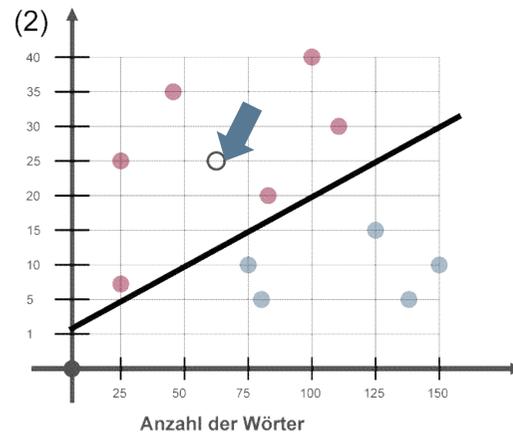
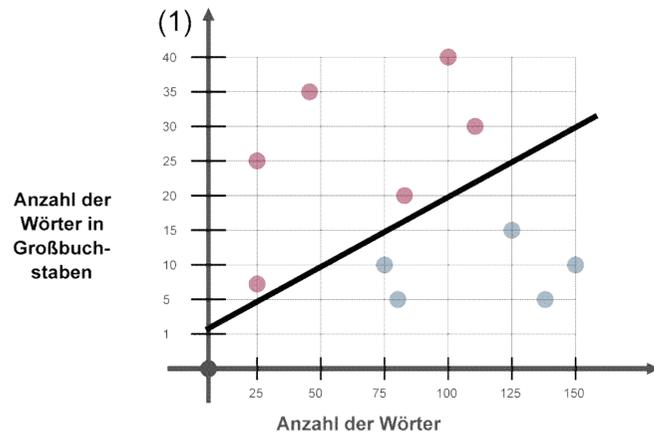
- Ermittlung der Geraden, welche die Daten trennen
- Bestimmung der besten Geraden

■ Output (3):

- Gerade als Modell zur Klassifizierung von E-Mails als **Spam** / kein Spam

ML-Algorithmus

→ SVM - Beispiel Spam - Erkennung



„auf neue Daten anwenden“

Anzahl Wörter	25	25	47	75	79	82	100	110	125	140	150	63
Anzahl Wörter in Großbuchstaben	7	25	35	10	5	20	40	30	15	5	10	25
Spam-E-Mail	ja	ja	ja	nein	nein	ja	ja	ja	nein	nein	nein	?

■ Input-Daten (1):

- **Modell** zur Erkennung von möglichen Spam-Mails
- **zu beurteilende E-Mail** (z.B.: 63/25)

■ ML-Algorithmus (2):

- Berechnung der Lage der zu untersuchenden **E-Mail (63/25)**

■ Output (3):

- Lage der Punkte zum Modell klassifiziert die E-Mail als **Spam-Mail**

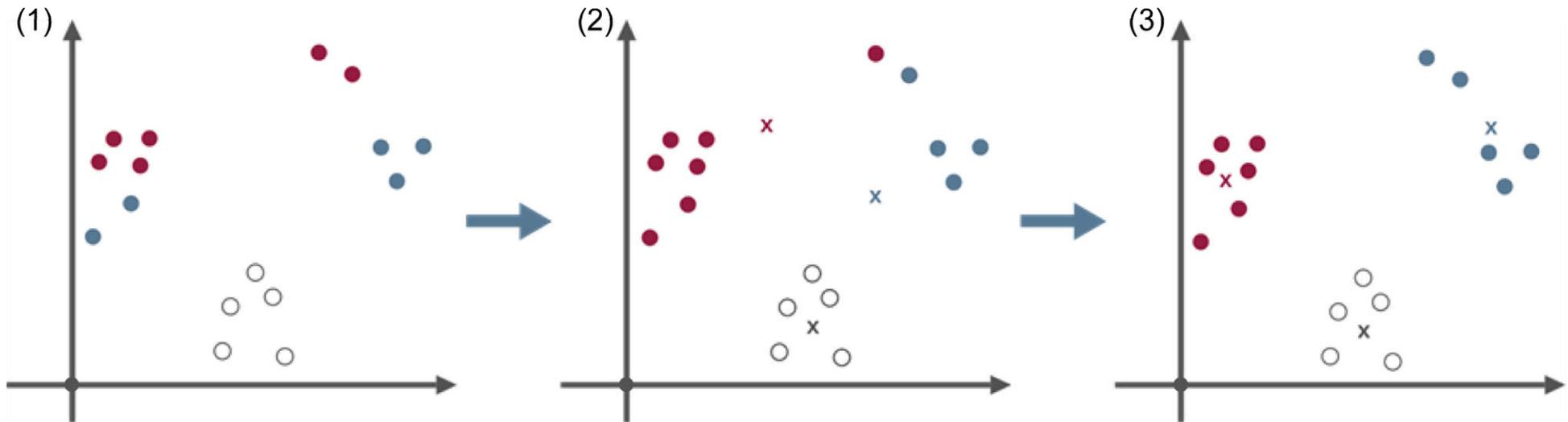
ML-Algorithmus

→ Unüberwachtes Lernen

- **Stärke im Suchen nach Mustern in unklassifizierten Daten**
- Erwartungshaltung an diesen Ansatz:
 - Muster erkennen, die vorher **anders nicht greifbar waren** (Komplexität)
- ML-Algorithmus lernt selbstständig
- Klassische Fehler werden in diesem Sinne nicht produziert
- **ML-Algorithmus**
 - Clustering setzt ähnliche Datengruppen miteinander in Verbindung, z.B.:
 - k-Means-Algorithmus
 - Hierarchische Clustering-Verfahren
- **Problem:** Lernt der ML-Algorithmus in die gewünschte Richtung?

ML-Algorithmus

→ k-Means-Algorithmus



■ Input-Daten:

- beliebige Daten
- Abstandsmaß
- Anzahl k Cluster
- Initiale Zuordnung der Elemente zu Clustern (z.B. zufällig)

■ ML-Algorithmus:

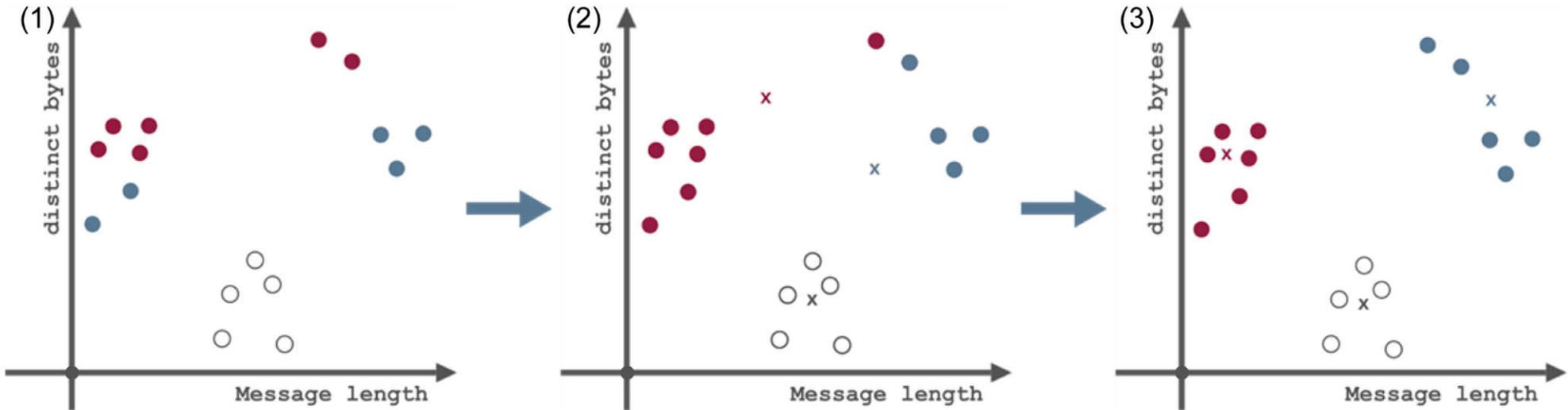
- Berechnung der **Schwerpunkte** (Zentroide)
- Zuordnung der Elemente zu Cluster mit dem nächsten Zentroid
- Neuberechnung der Zentroide und erneute Zuordnung

■ Output:

- **Einteilung** der Objekte in **k Cluster**

ML-Algorithmus

→ k-Means-Algorithmus - Beispiel



■ Input-Daten (1):

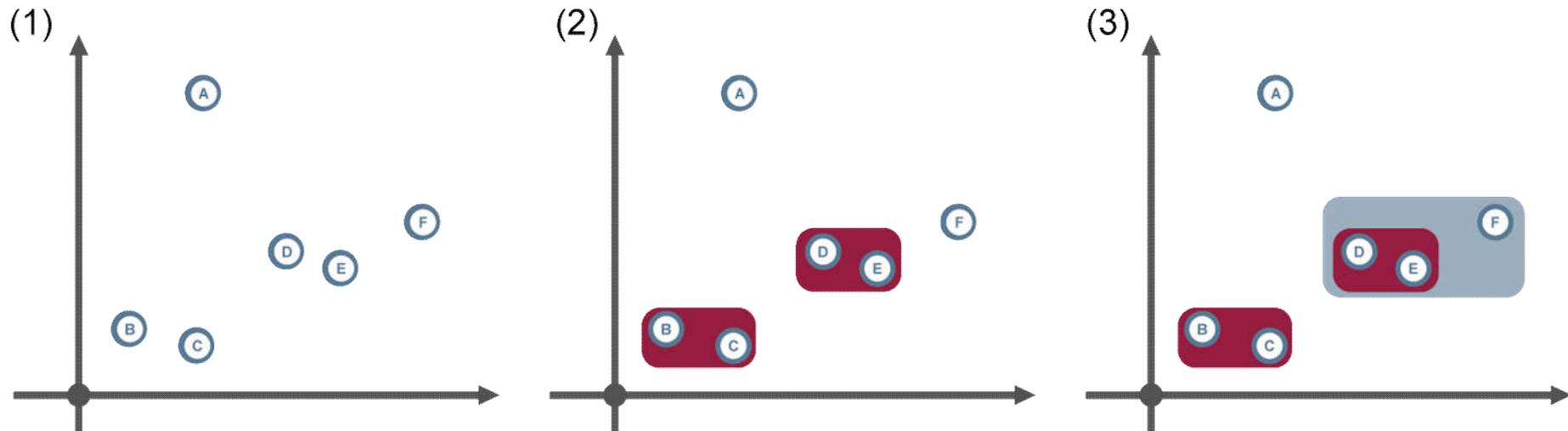
- Daten von Malware (*Palevo, Virut, Mariposa*)
- Abstandsmaß
- $k = 3$
- Initiale Zuordnung nach Message length, distinct bytes

■ ML-Algorithmus (2):

- Berechnung der Durchschnitte
- Zuordnung der Elemente zur Malwareart mit dem nächsten Zentroid
- Neuberechnung der Zentroide und erneute Zuordnung

■ Output (3):

- Einteilung der Malware in die drei Malwarearten
 - Rot = Virut
 - Weiß = Palevo
 - Blau = Mariposa

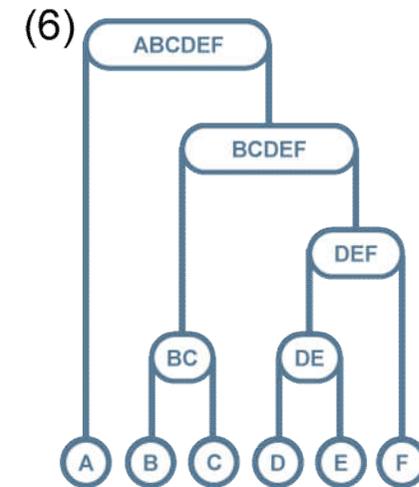
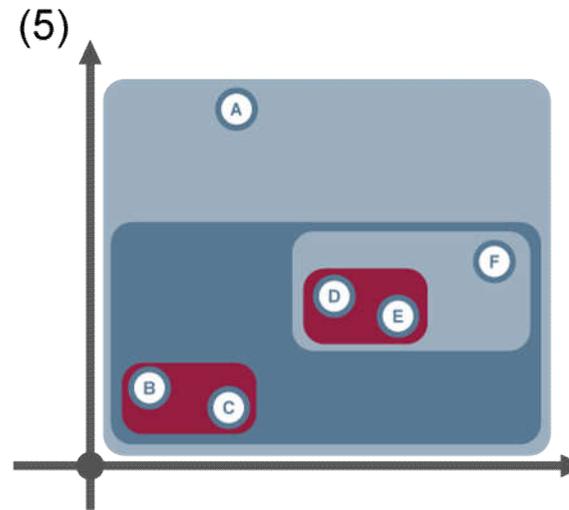
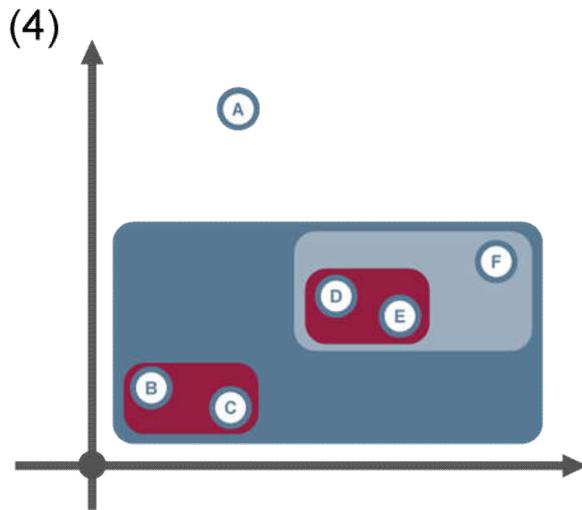


■ Input-Daten (1):

- beliebige Daten
- Ähnlichkeitsmaß

■ ML-Algorithmus (2 bis 5):

- jeder Datenpunkt ist ein eigenes Cluster
- ähnlichste Cluster werden zuerst zusammengeführt
- entstandene Cluster werden erneut als Eingabedaten verwendet
- iteratives Zusammenführen der Cluster induziert eine hierarchische Struktur



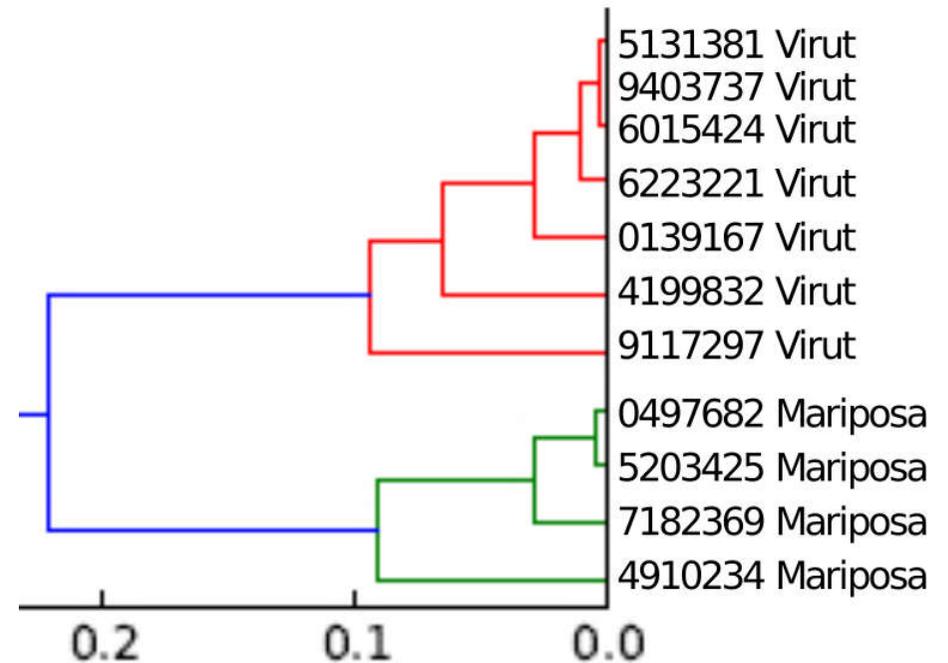
■ Output (6):

- Hierarchische Beziehungen zueinander in Form eines Binärbaums (Dendrogramm)

ML-Algorithmus

→ Hierarchische Clustering-Verfahren - Beispiel

- Clustering der Daten aus Botnet-Analyse
- Anwendung einer komplexen Distanzfunktion (Wertebereich [0, 1])
- Trennung der Familien-Cluster bei Distanz von ca. 0.1
- Einordnung der Daten in zwei Malware-Familien Virut und Mariposa

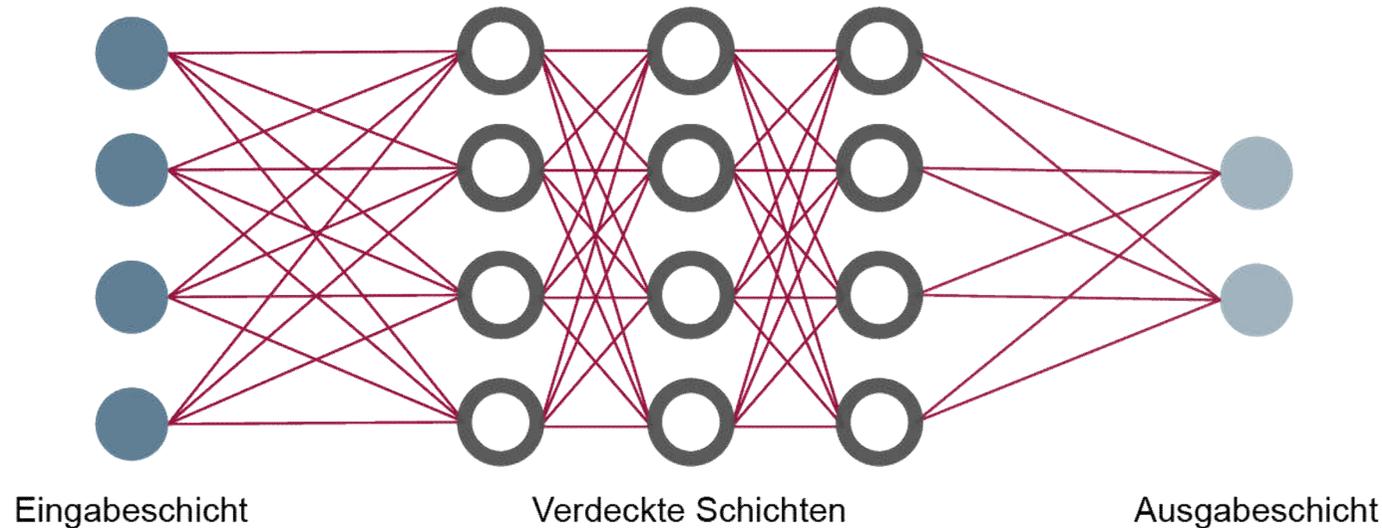


- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Künstlich Neuronale Netze (KNN)

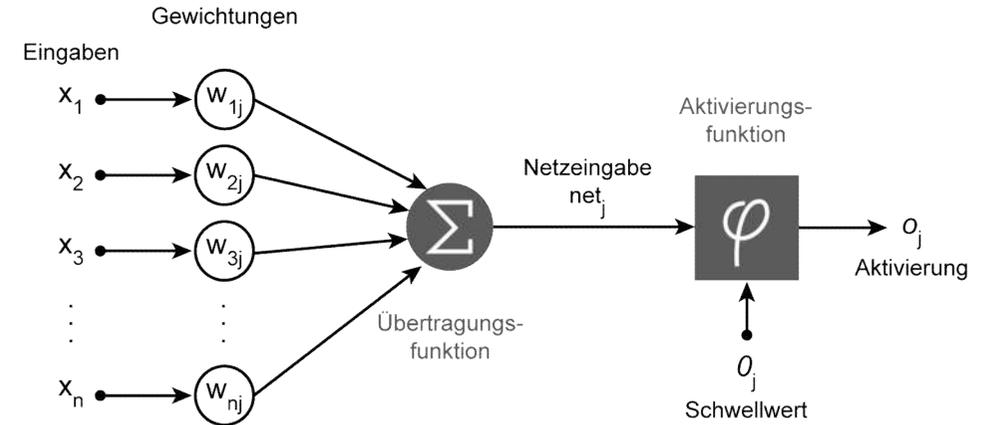
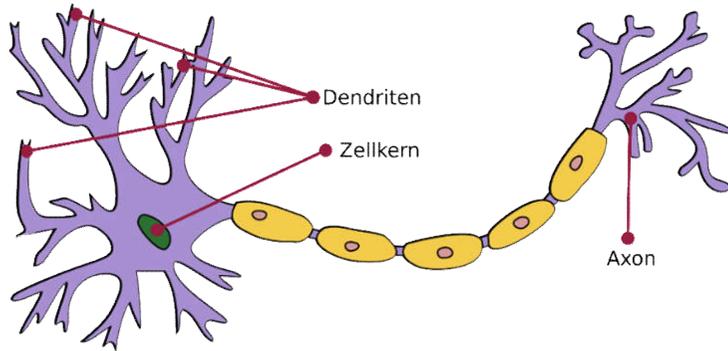
→ Netze aus künstlichen Neuronen (1/2)

- Vorlage ist die die biologische Struktur des Gehirns/Neurons
- Nutzen Gewichte und mathematische Funktionen (für die Informationsverarbeitung)
- Informationsverarbeitung über mehrere miteinander verbundene Schichten aus künstlichen Neuronen



Künstlich Neuronale Netze (KNN)

→ Netze aus künstlichen Neuronen (2/2)



■ Biologisches Neuron:

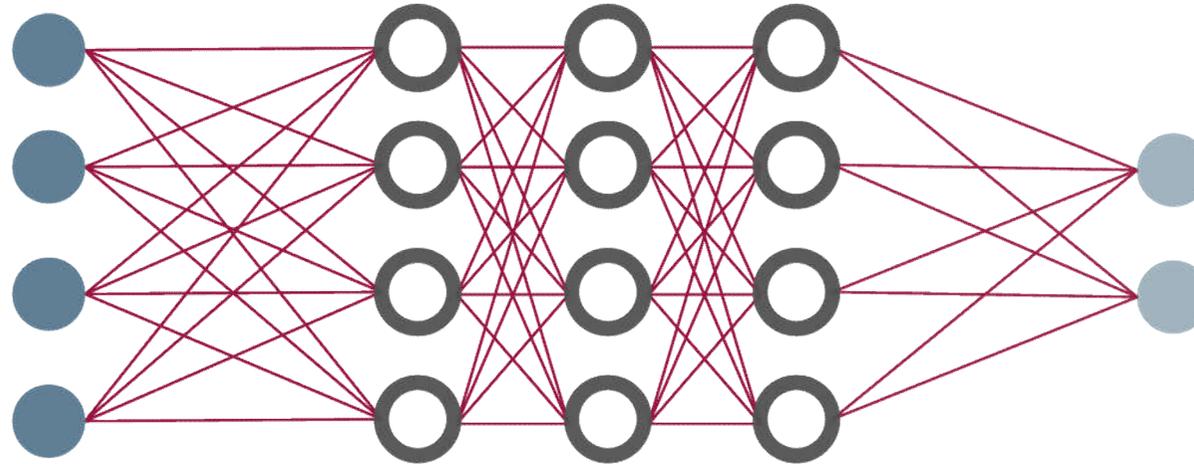
- Dendriten:
 - Reizaufnahme (Signaleingang)
- Axon:
 - Leitet die Informationen weiter (Signalausgang)
- Zellkern:
 - Reizverarbeitung (Signalverarbeitung)

■ Künstliches Neuron:

- Übertragungsfunktion:
 - Berechnet anhand der Summe der Wichtungen, der Eingaben, die Netzeingabe
- Aktivierungsfunktion/ Ausgabefunktion:
 - Ausgabe der Information
- Schwellenwert:
 - Wert eines Reizes, bei dem das Neuron aktiviert wird

Künstlich Neuronale Netze (KNN)

→ Schichten in einem KNN



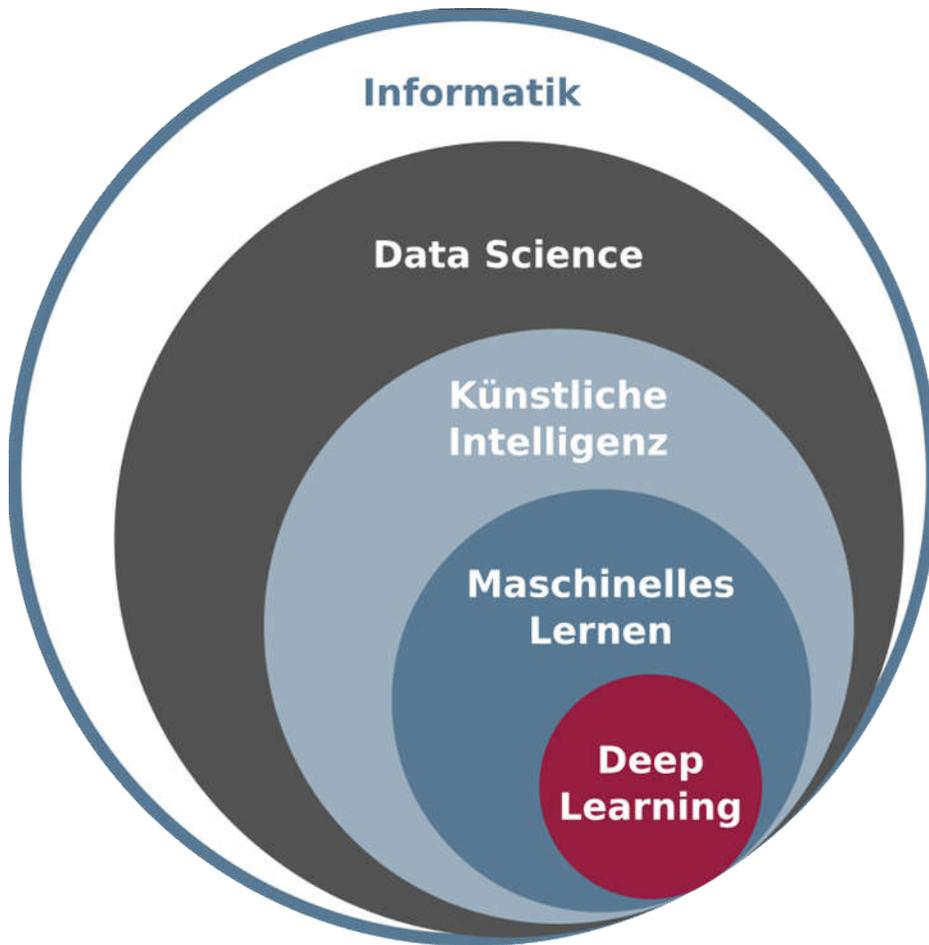
Eingabeschicht

Verdeckte Schichten

Ausgabeschicht

- **Eingabeschicht:**
 - Eingabeneuronen (z.B. Ohren, Retina oder Haut)
 - Eingabedaten werden in geeignete Repräsentation überführt
- **Verdeckte Schichten:**
 - Je nach Komplexität der Aufgabe 1-N verknüpfte Neuronen
 - Erkennung von simplen Mustern und Strukturen
 - Mit jeder Schicht werden immer komplexere Merkmale herausgefiltert
- **Ausgabeschicht:**
 - Ausgabe sämtlicher möglicher Repräsentationen der Ergebnisse

Einordnung → Deep Learning



- Maschinelles Lernen wird noch effektiver durch:
 - **Deep Learning**
- Deep Learning ist eine Spezialisierung des maschinellen Lernens
- *Nutzt vorwiegend neuronale Netze*
 - ***Erlaubt unvollständige Daten***
 - ***Erlaubt Rauschen und Störungen***
- Kommt dem „menschlichen Gehirn“ am nächsten

Deep Learning

→ Architekturen (1/2)

- Forschung durch **leistungsfähigere Hardware** und **steigende Datenverfügbarkeit** in letzten Jahren deutlich gestiegen
- Neben klassischen Feed-Forward-Netzen auch Recurrent Neural Networks handhabbar
 - Kanten können auch zu vorherigen Schichten zurückführen
- **Hohe Anzahl an Schichten**, welche nach Funktionsweise zusammengefasst werden können
- Verschiedene Architekturen haben sich für unterschiedliche Problemstellungen als besonders effektiv gezeigt
- **Bessere Skalierbarkeit**

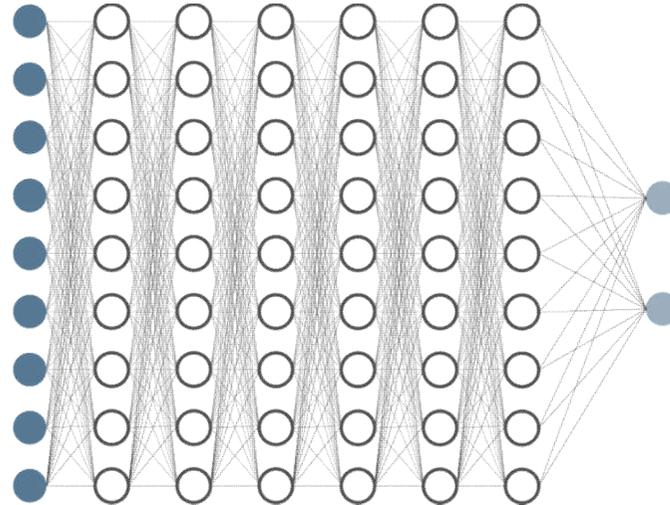
Deep Learning

→ Architekturen (2/2)

- **Convolutional Neural Networks (CNN):**
 - Zweidimensionales „Fenster“ wird über Daten „geschoben“
 - Einfluss durch benachbarte Felder wird berücksichtigt
 - Besonders erfolgreich bei Computer Vision (z.B. Handschrift-Erkennung)
- **Long Short-Term Memory Networks (LSTM):**
 - Spezialform eines Recurrent Neural Networks
 - Neuronen können Zustände über einen längeren Zeitraum speichern
 - Besonders erfolgreich bei gesprochener Sprache (Alexa, Siri, usw.)

Deep Learning

→ Handschrifterkennung - Beispiel



1010010010
1010110010
1010011111
1011001001
1010101101

Ziffer	0	1	2	3	4	5	6	7	8	9
Übereinstimmung	0 %	7 %	1%	0 %	4 %	0 %	0 %	85 %	0 %	3 %

■ Input-Daten (1):

- Bilddatei mit einer Zahl (7), die klassifiziert werden soll

■ ML-Algorithmus (2):

- Eingabedaten werden in den künstlichen Neuronen in den Schichten verarbeitet
- Z.B. mit Hilfe eines Convolutional Neural Network (CNN)

■ Output (3):

- Tabelle mit einer Verteilung der **Wahrscheinlichkeiten** für eine Übereinstimmung mit **einer Ziffer**

- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Anwendungen von KI und CS (1/2)

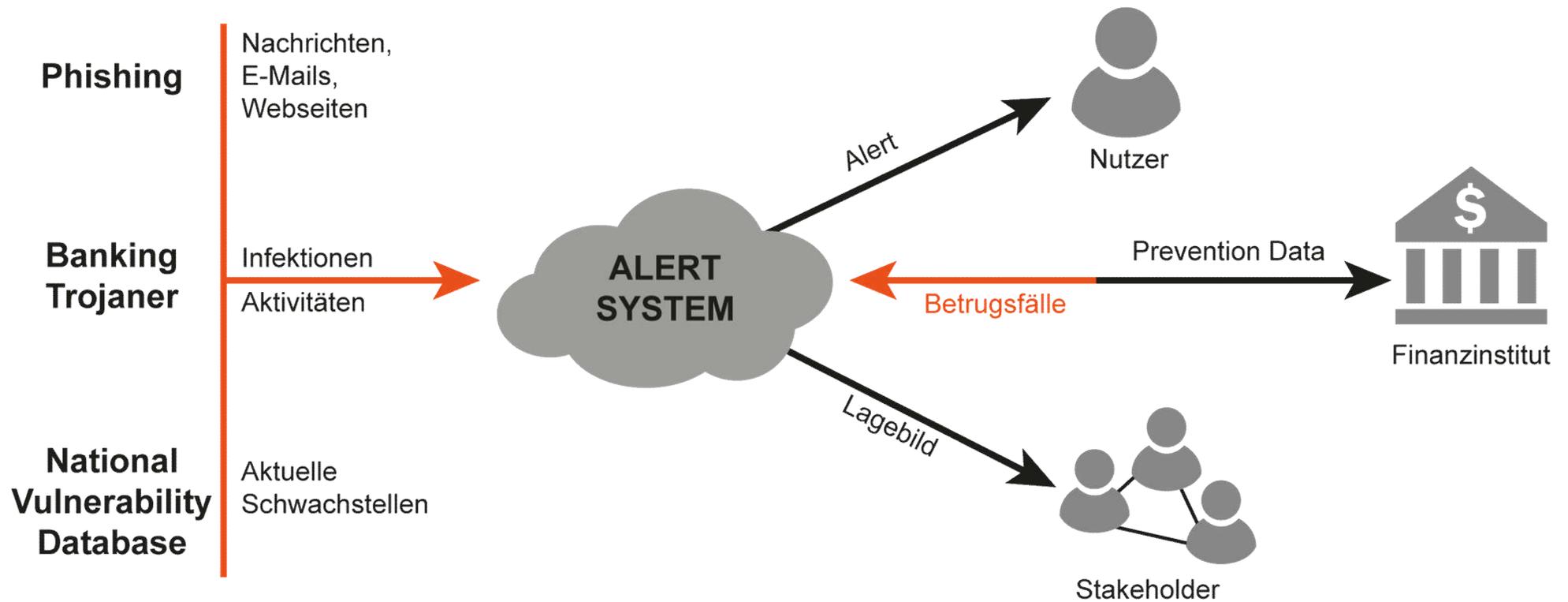
→ Alert-System für Online-Banking

- Wie könnte eine Lösung aussehen?
 - Tagesaktuelle Warnungen bei erhöhter Gefahrenlage (Online-Banking)
→ **damit der Bankkunde und die Bank reagieren können**
 - Aufklärung der Nutzer, wenn Gefahren vorliegen
→ **damit der Bankkunde sich „richtig“ verhalten kann**
- Ansatz des Alert-Systems
 - **Sicherheitskennzahlen** zum Betrug identifizieren
 - Mittels KI **Gefahrenlage bestimmen**
 - Nutzer und Bank **Warnen**



Alert-System für Online-Banking

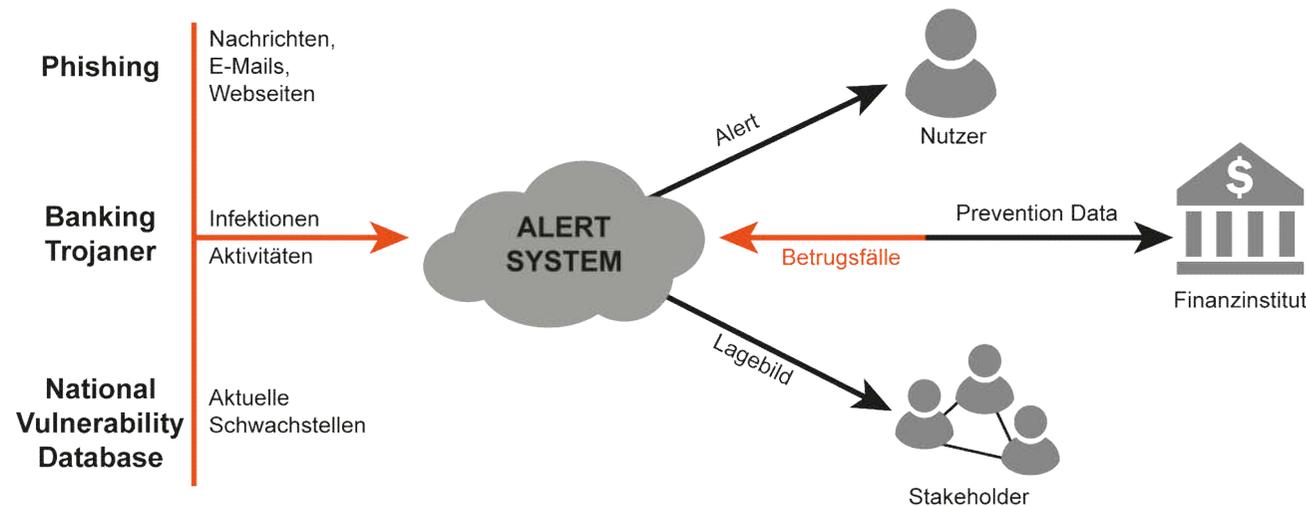
→ Konzept



Alert-System für Online-Banking

→ Zahlen für den Testzeitraum von 456 Tage

- 1.904 Nachrichten (Phishing-Angriff) – „Stackoverflow-Netzwerk“
- 5.589 **E-Mail** (Phishing-Angriff) – „Spam Archive“
- 2.776 Phishing-**Webseiten** – „PhishTank“
- 23.184 **Infektionen** von Banking-Trojaner (Malware) – Anti-Malwarehersteller
- 875 relevante **Schwachstellen** (NVD)
- 459 erfolgreiche **Betrugsfälle** im Online-Banking - Bankengruppe

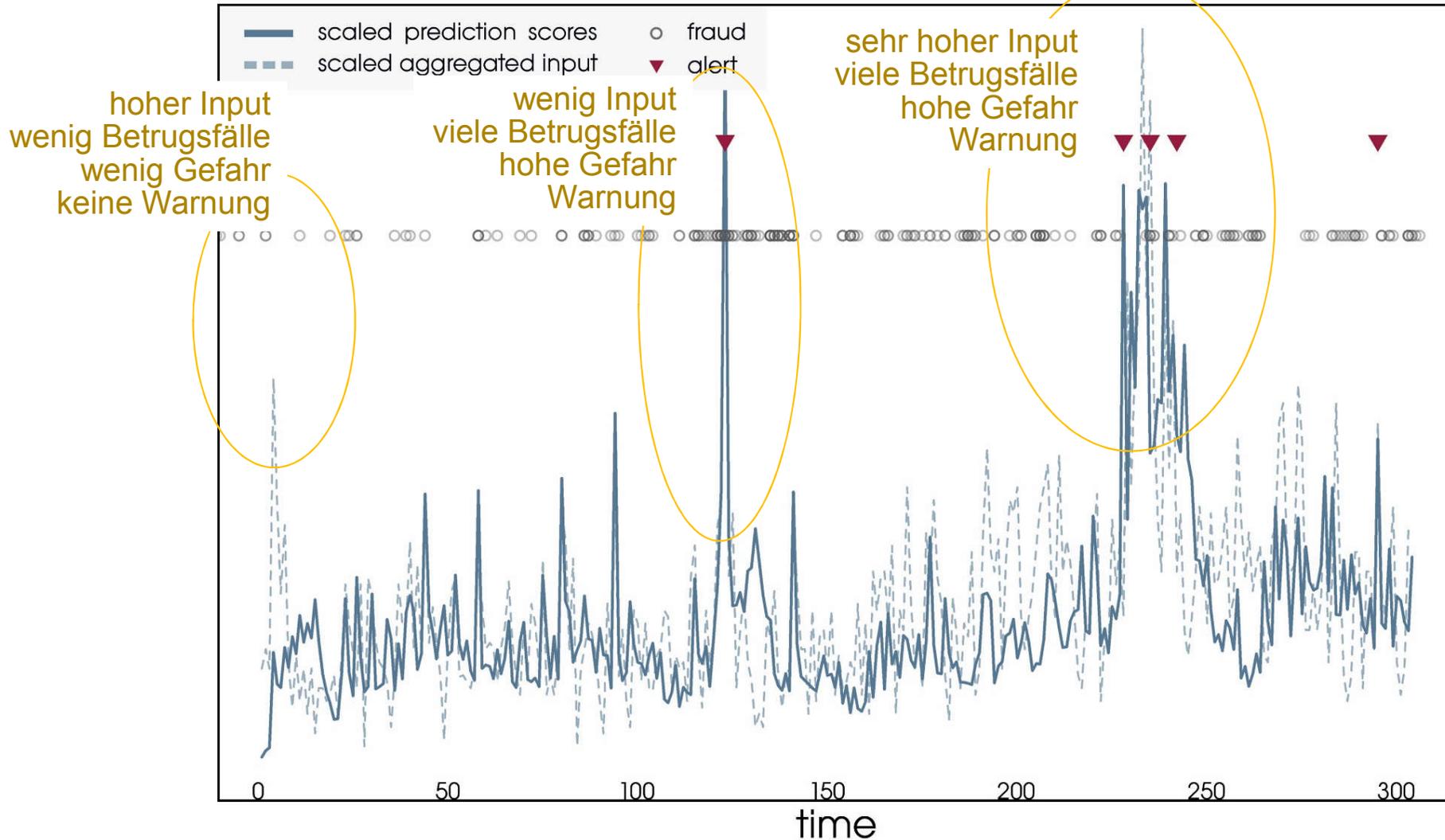


$\frac{1}{3}$ des Zeitraums zum Training (152 Tage) $\frac{2}{3}$ zur Evaluation (304 Tage)

Ergebnis einschätzen

→ k-Nearest Neighbor

k-Nearest Neighbor

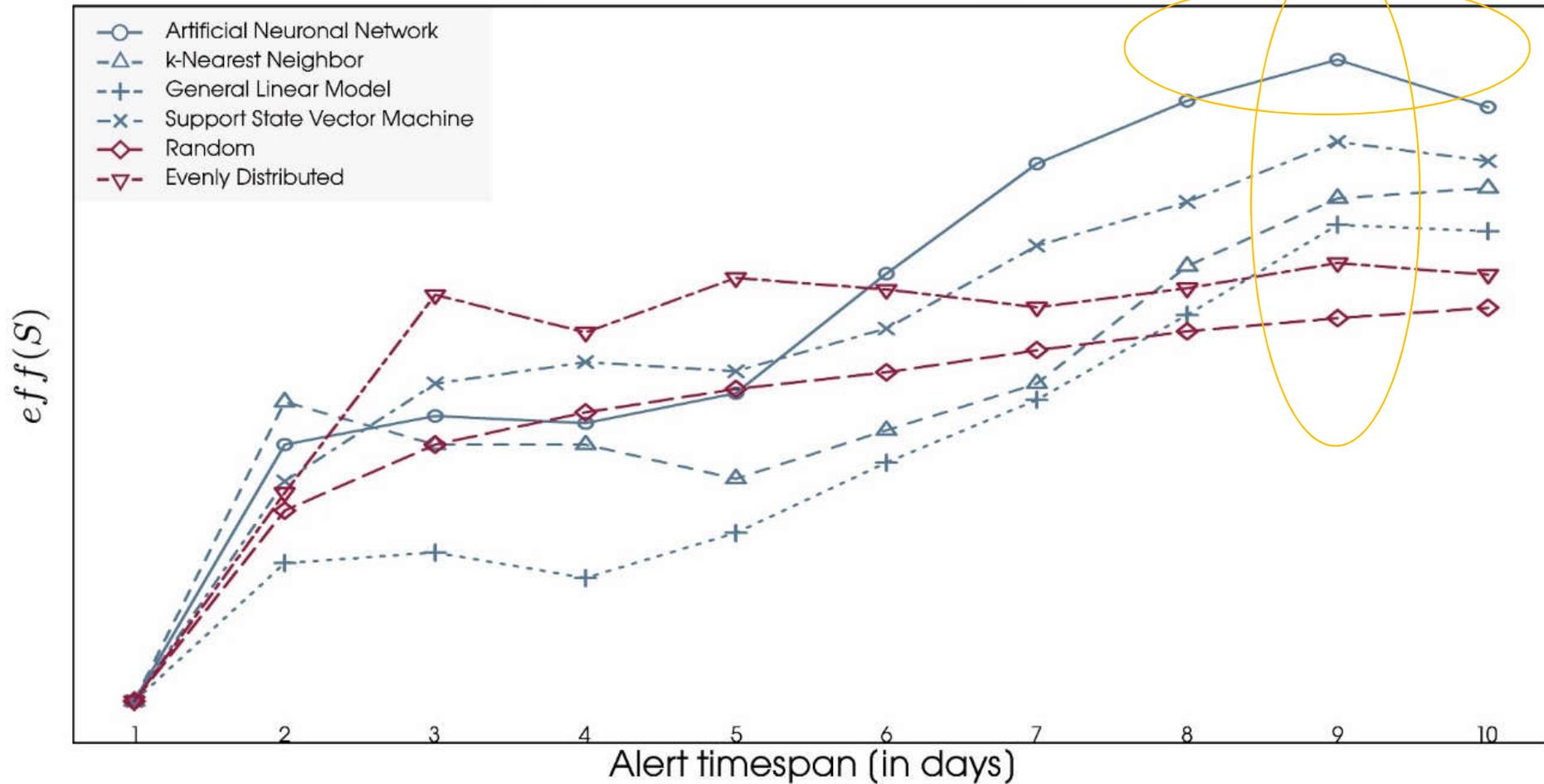


Ergebnisse

→ Vergleich der verschiedenen Verfahren

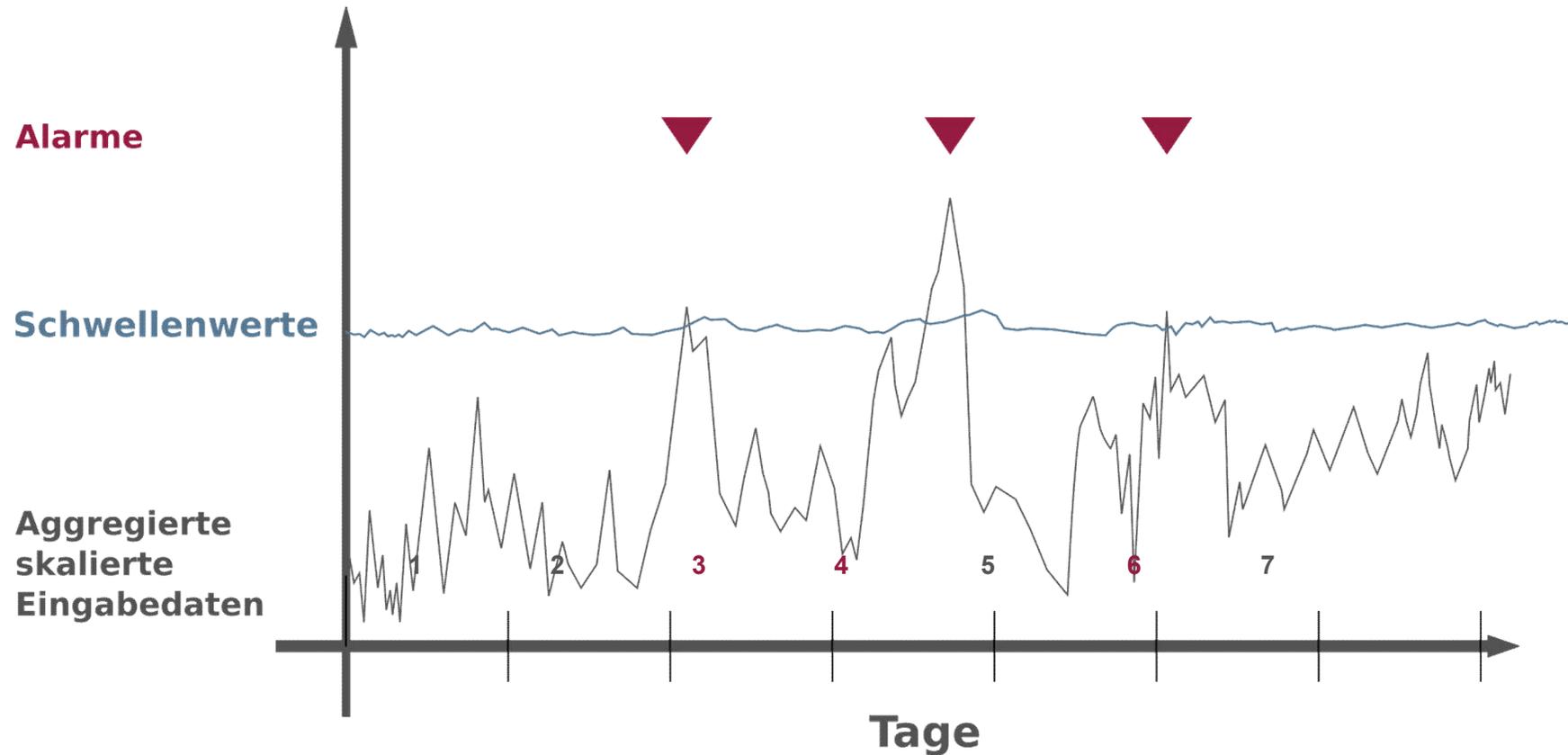
„Aber, drei Mal soviel Zeit für das Trainieren“

Comparison of the different approaches



Alert-System für Online-Banking

→ Ergebnis



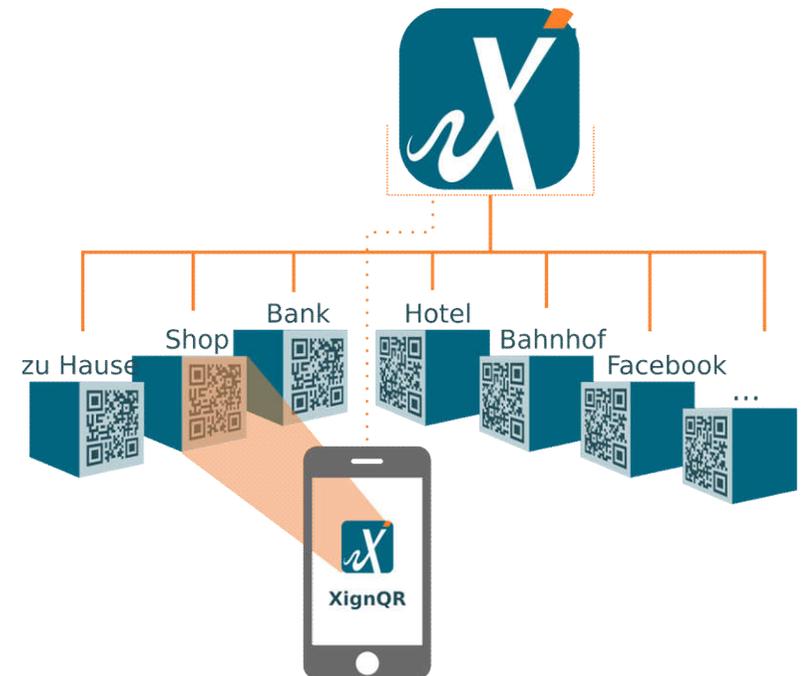
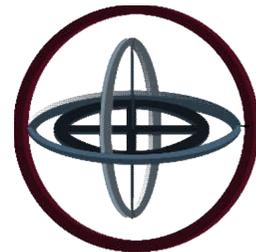
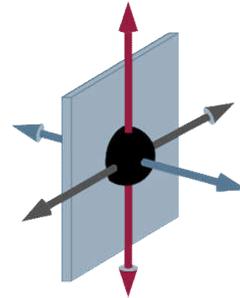
■ Output:

- Vorhergesagte Bedrohungswerte überschreiten an den Tagen 3, 4 und 6 den für dieses Alert-System eingestellten Schwellenwert
- da Schwellenwert überschritten wurde, wird ein Alarm ausgelöst

Anwendungen von KI und CS (2/2)

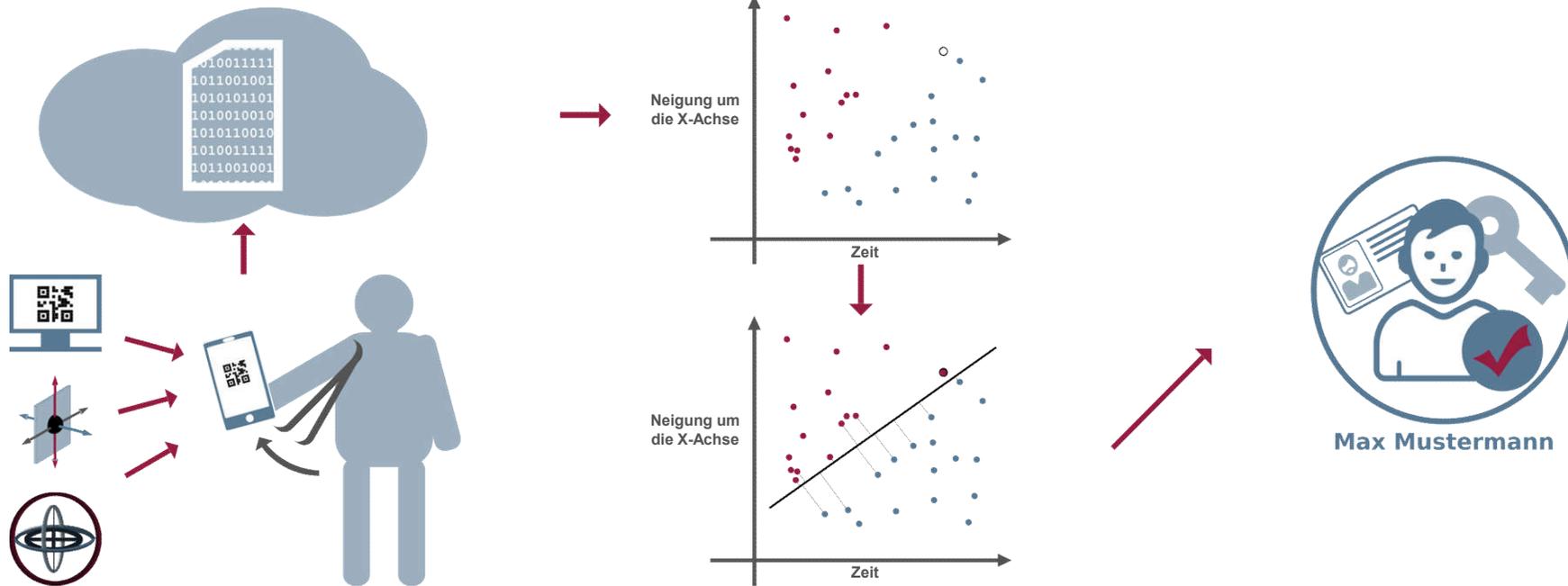
→ Passive Authentifikation - XignQR

- Ein Nutzer wird automatisch an der Art und Weise der Nutzung beim QR-Code Scannen erkannt.
- Während das gesamten Vorgangs werden passive biometrische Bewegungsdaten erfasst.
- Datenerfassung durch
 - **Beschleunigungssensor**
 - **Lagesensor**



Passive Authentifikation - XignQR

→ Support-Vector-Machine (SVM)



■ Input-Daten:

- Nutzer holt Gerät aus Hosentasche
- Erfassen von **Lage** und **Beschleunigung** des Smartphones

■ ML-Algorithmus:

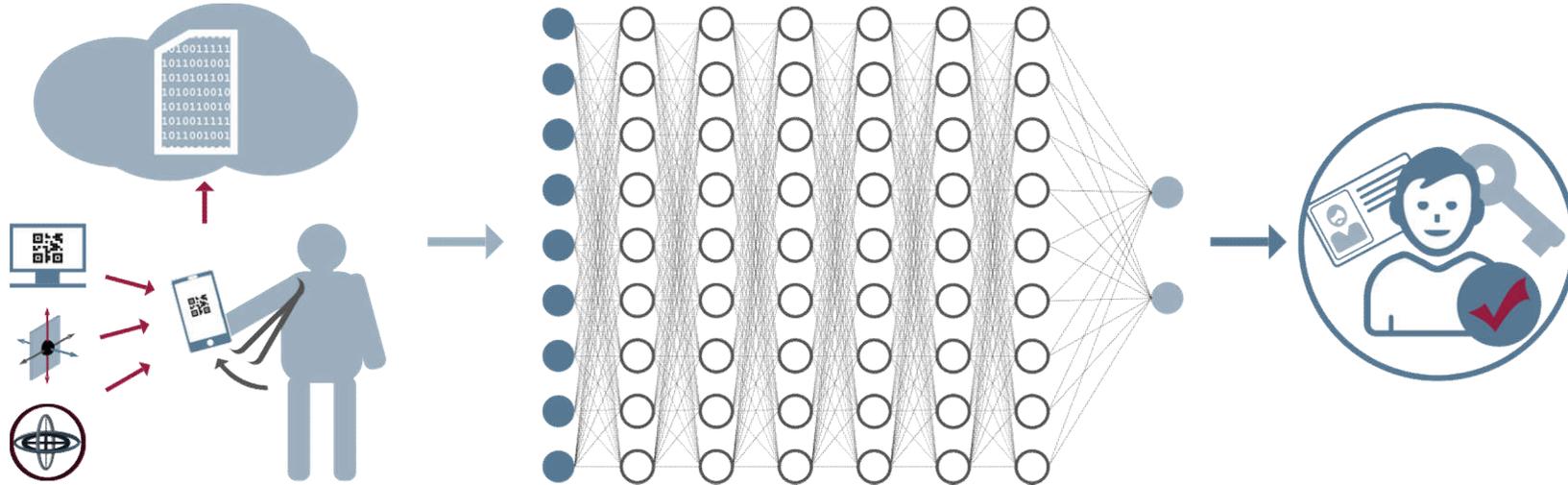
- Daten werden anhand der Hyperebene/des Modell klassifiziert
- rote Übereinstimmung ist **positive** Klassifizierung
- blau eine **negative** Klassifizierung (bspw. anderer Nutzer)

■ Output:

- Authentisierung ist entweder erfolgreich oder schlägt fehl (**95 %**)

Passive Authentifikation - XignQR

→ Neuronales Netz



Input-Daten:

- Lage und Beschleunigungsdaten des Nutzers werden erzeugt

ML-Algorithmus:

- Eingabedaten werden in den künstlichen Neuronen in den Schichten verarbeitet

Output:

Nutzer	Übereinstimmung
0	0,059 %
1	99,85 %
2	0,087 %

```
time, type, x, y, z
271, Accelerometer, -0.07606506, 9.173798, 3.6333618
277, Accelerometer, 1.0681152E-4, 9.146423, 3.5619507
279, Gyroscope, 0.027664185, 0.06774902, 0.02182006
...
```

```
[[5.9110398e-04 9.9853361e-01 8.7528664e-04]]
Predicted Class [1]
Predicted Person: Sandra Kreis
```

KI und Cyber-Sicherheit

→ Weitere Beispiele

- Logdatenanalyse
- Malware-Erkennung
- Security Information and Event Management (SIEM)
- Threat Intelligence
- Spracherkennung
- Bilderkennung (Ausweis, Video, ...)
- Authentifikationsverfahren
- Fake-News
- IT-Forensik
- Sichere Softwareentwicklung
- ...

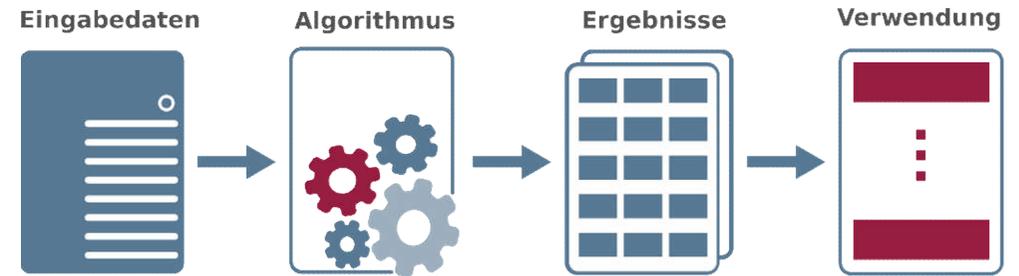
- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Künstliche Intelligenz / ML

→ Angriffe

- „Hacker“ greifen an und manipulieren den Workflow

- die Eingabedaten (Input)
 - gezielte Manipulation
- die Algorithmen
- die Ergebnisse (Output)
- die Verwendung



- **Angriffe auf die Privatsphäre**
(personenorientierte Daten, die verwendet werden)

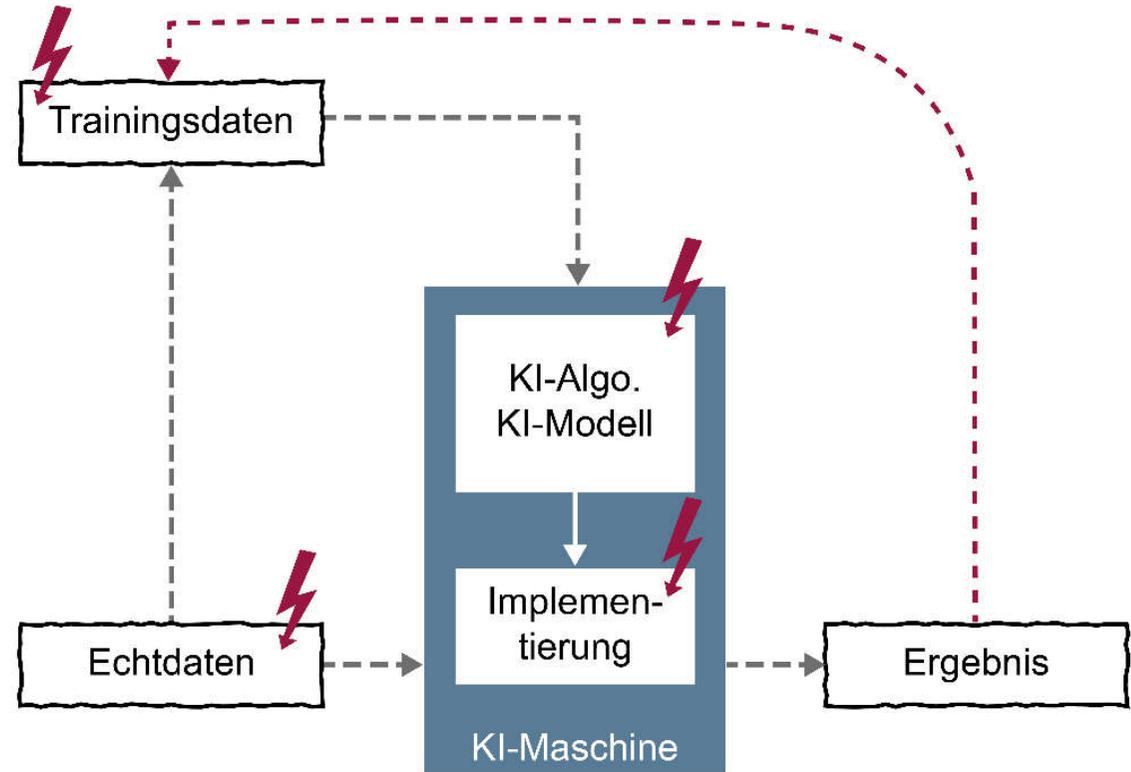
Vertrauenswürdigkeit → Qualität der Umsetzung

Stand der Technik an IT-Sicherheitsmaßnahmen zum Schutz

- der **Daten** (Training, Echt, Ergebnis),
- der **KI-Maschine** und
- der **Anwendung**

Schutzziele:

- **Integrität**
(Erkennen von Manipulation der Daten)
- **Vertraulichkeit**
(Wahrung von Geschäftsgeheimnissen)
- **Datenschutz**
(Schutz von personenbezogenen Daten)
- **Verfügbarkeit**
(der Anwendung und Ergebnisse)



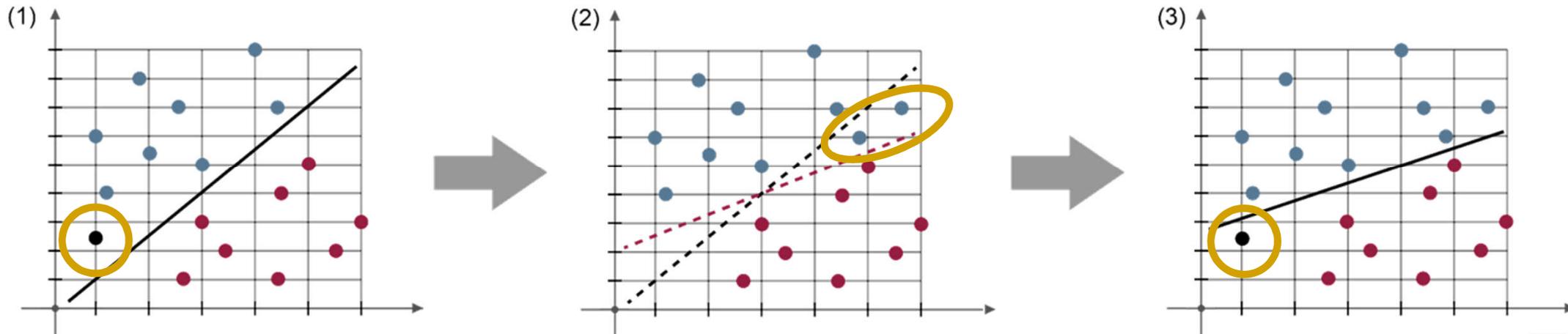
**Nutzung einer
qualitativ hochwertigen
KI-Technologie**

**Zusammenarbeit von erfahrenen
KI- und
Anwendungsexperten**

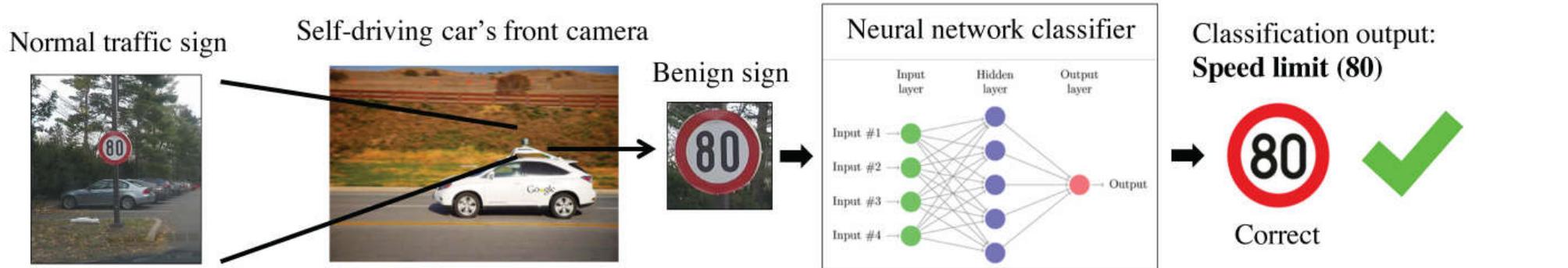
Angriffe auf maschinelles Lernen

→ Manipulation von Trainingsdaten

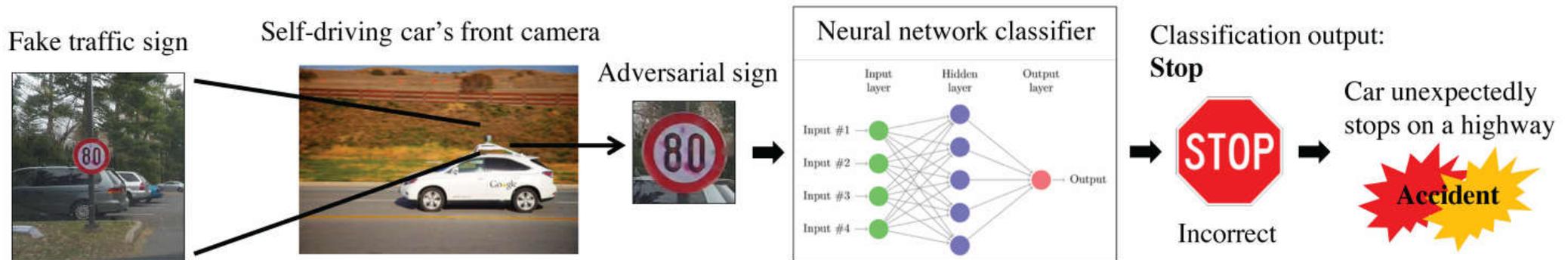
- (1) **Normale Klassifizierung** eines neuen Inputs.
(*neuer schwarzer Punkt gehört zur blauen Klasse*)
- (2) **Beispiel: Manipulation von Trainingsdaten**
 - Falsch klassifizierte Daten werden in den Trainingsprozess als Angriff einschleusen (*zwei weitere blaue Punkte*).
 - Dadurch wird die Gerade des Modells zur Klassifizierung manipuliert (*Gerade wird flacher*).
- (3) Damit kann ein Angreifer für **falsche Klassierungen** sorgen.
(*jetzt gehört der neuer schwarzer Punkt zur roten Klasse*)



Angriffe auf maschinelles Lernen → Manipulation von Verkehrszeichen



(a) Operation of the computer vision subsystem of an AV under *benign conditions*



(b) Operation of the computer vision subsystem of an AV under *adversarial conditions*

Fig. 1. **Difference in operation of autonomous cars under benign and adversarial conditions.** Figure 1b shows the classification result for a drive-by test for a physically robust adversarial example generated using our Adversarial Traffic Sign attack.

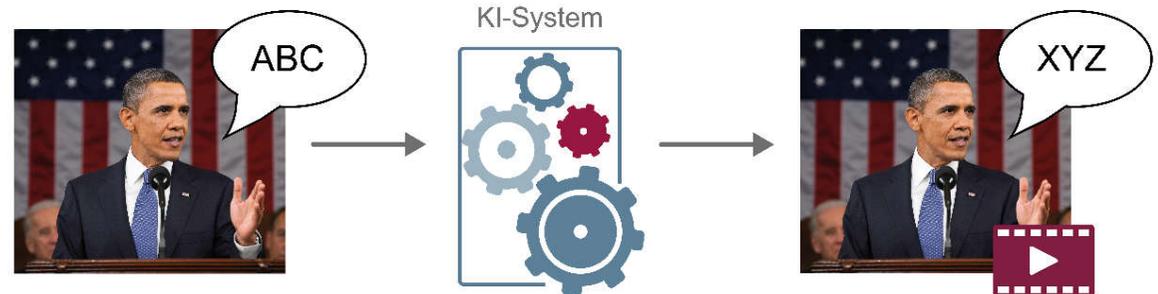
- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Künstliche Intelligenz

→ Angreifer verwenden KI

„Hacker“ verwenden KI ebenfalls für ihre Zwecke (Dual-Use)

- Schnelle Schwachstellensuche (bessere SW, schneller Angreifen)
- Social-Engineering (Chatbots, ...)
- Passwortknacker
- Neue Angriffsstrukturen und Vorgehensweisen
- Videomanipulation (Deep-Fake)
 - „Fake Obama Video“
 - „Make Putin Smile Video“



Künstliche Intelligenz

→ Allgemeine Herausforderungen

- **Datenschutz** (persönliche Daten ... Europäische Datenschutz-Grundverordnung)
- **Selbstbestimmung** („human in the loop“)
- **Diskriminierung** (ausgeglichene Daten ... Problem: gibt es nicht)
→ Frau/Mann, Herkunft, Ausbildung, ...
- **Vertrauenswürdigkeit** der Daten und Ergebnisse
→ KI-Siegel
- ...



Intelligente Algorithmen

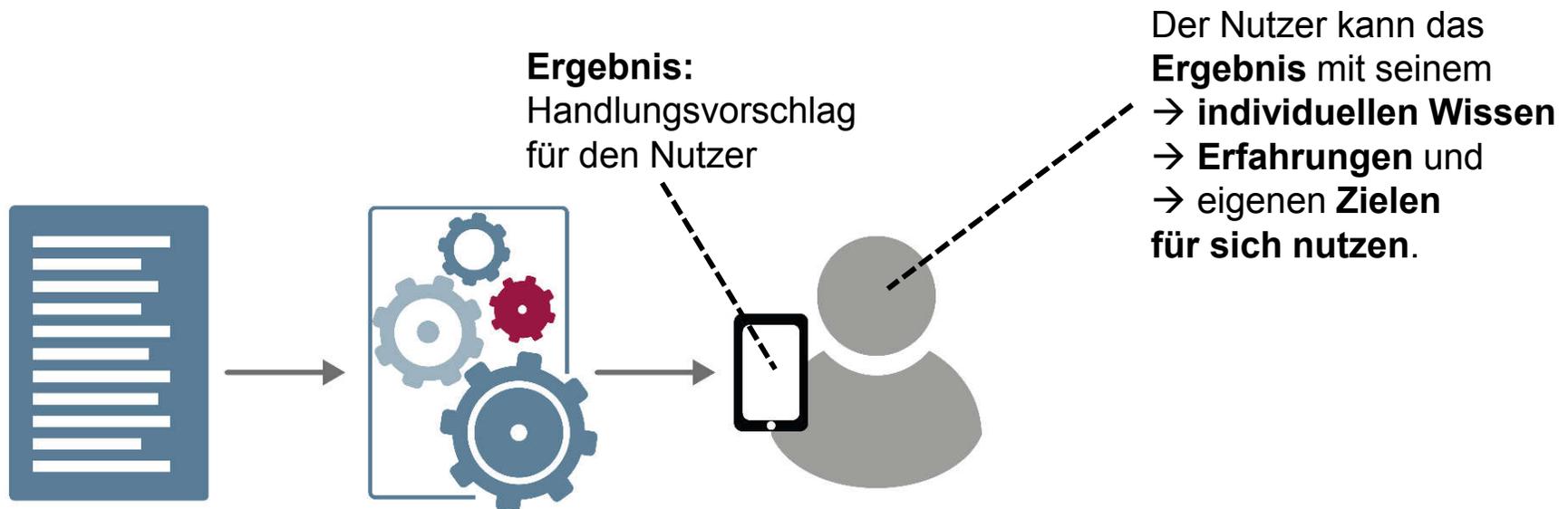
→ Chancen und Risiken

- **Individuelles Wissen** und **Komplexität des denkenden Menschen** sind Algorithmen überlegen! +
- **Algorithmen können schneller Wissen** aus vorhandenen **Daten auswerten!** +
- Individuelles Wissen + Algorithmen Wissen = ++++
- **Praktische Probleme:** Medizin / Watson
 - Diagnostik (*Maschine*)
 - Haftung (*Mensch*)

Vertrauenswürdigkeit

→ Nachvollziehbarkeit der Ergebnisse

- „Keep the human in the loop“
 - KI-Ergebnis muss als **Handlungsempfehlung** für den **Nutzer** verstanden werden.
 - Damit wird die **Selbstbestimmtheit** der Nutzer gefördert und die Vertrauenswürdigkeit erhöht.



- **Automatisierte Anwendungen** (z.B. autonomes Fahren)
 - Simulation, Test und **Validierung**
 - Verantwortung, **Haftung** und Versicherung

- **Einordnung**
(Idee, Data Science, KI, ML, Workflow, Erfolgsfaktoren, ...)
- **Maschinelles Lernen**
(überwacht/unüberwacht, SVM, k-Means, h-Clustering, ...)
- **Künstliche Neuronale Netze**
(Idee, KNN, Deep Learning, ...)
- **Anwendungen KI und Cyber-Sicherheit**
(Alert-System für Online-Banking, passive Authentifikation, ...)
- **Angriffe auf maschinelles Lernen**
(Idee, Trainingsdaten, Verkehrszeichen, ...)
- **Künstliche Intelligenz und Cyber-Sicherheit**
(Dual-Use, Herausforderungen, Chancen und Risiken, ...)
- **Ergebnis und Ausblick**

Künstliche Intelligenz und CS

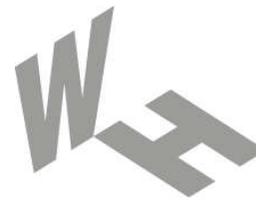
→ Ergebnis und Ausblick

- **KI/ML ist eine wichtige Technologie für die Zukunft, auch für Cyber-Sicherheit**
 - Erkennen von Bedrohungen, Schwachstellen, Angriffen, ...
 - Erkennen von Nutzern (Authentifikation)
 - Unterstützung von Cyber-Sicherheitsexperten
 - Vorschläge für Handlungsanweisungen
 - ...
- **Sehr gute Daten sind das Wichtigste**
 - Neue, bessere Sensoren (Daten mit sehr gutem Inhalt)
 - Zusammenarbeit und Austausch von Daten
 - ...
- **Technologische- und Daten-Souveränität wird immer wichtiger**

Künstliche Intelligenz

→ Einstiegspunkte

- **Amazon**
 - Amazon Machine Learning
 - Amazon Lex (Konversationsschnittstellen für Sprache und Text)
- **Microsoft**
 - Azure Machine Learning
 - Microsoft Cognitive Services (Bildanalyse und Gesichtserkennung)
- **Google**
 - Google Cloud Machine Learning Engine
 - Tensorflow
- **IBM**
 - IBM Machine Learning
 - **Watson**



**Westfälische
Hochschule**

Gelsenkirchen Bocholt Recklinghausen
University of Applied Sciences

Künstliche Intelligenz *und* Cyber-Sicherheit

Mit **Künstlicher Intelligenz** in die Zukunft!

Prof. Dr. (TU NN)

Norbert Pohlmann

Institut für Internet-Sicherheit – if(is)
Westfälische Hochschule, Gelsenkirchen
<http://www.internet-sicherheit.de>

if(is)
internet-sicherheit.

Wir empfehlen

- **Kostenlose App securityNews**



securityNews



- **7. Sinn im Internet (Cyberschutzraum)**

<https://www.youtube.com/cyberschutzraum>



- **Master Internet-Sicherheit**

<https://it-sicherheit.de/master-studieren/>



Besuchen und abonnieren Sie uns :-)

WWW

<https://www.internet-sicherheit.de>

Facebook

<https://www.facebook.com/Internet.Sicherheit.ifis>

Twitter

[https://twitter.com/ ifis](https://twitter.com/ifis)

YouTube

<https://www.youtube.com/user/InternetSicherheitDE/>

Prof. Norbert Pohlmann

<https://norbert-pohlmann.com/>

Quellen Bildmaterial

Eingebettete Piktogramme:

- Institut für Internet-Sicherheit – if(is)

Der Marktplatz IT-Sicherheit

(IT-Sicherheits-) Anbieter, Lösungen, Jobs, Veranstaltungen und Hilfestellungen (Ratgeber, IT-Sicherheitstipps, Glossar, u.v.m.) leicht & einfach finden.

<https://www.it-sicherheit.de/>

N. Pohlmann, S. Schmidt: „Der Virtuelle IT-Sicherheitsberater – Künstliche Intelligenz (KI) ergänzt statische Anomalien-Erkennung und signaturbasierte Intrusion Detection“, IT-Sicherheit – Management und Praxis, DATAKONTEXT-Fachverlag, 05/2009

D. Petersen, N. Pohlmann: "Ideales Internet-Frühwarnsystem", DuD Datenschutz und Datensicherheit – Recht und Sicherheit in Informationsverarbeitung und Kommunikation, Vieweg Verlag, 02/2011

M. Fourné, D. Petersen, N. Pohlmann: "Attack-Test and Verification Systems, Steps Towards Verifiable Anomaly Detection". In Proceedings der INFORMATIK 2013 - Informatik angepasst an Mensch, Organisation und Umwelt, Hrsg.: Matthias Horbach, GI, Bonn 2013

D. Petersen, N. Pohlmann: „Kommunikationslage im Blick - Gefahr erkannt, Gefahr gebannt“, IT-Sicherheit – Management und Praxis, DATAKONTEXT-Fachverlag, 4/2014

U. Coester, N. Pohlmann: „Verlieren wir schleichend die Kontrolle über unser Handeln? Autonomie hat oberste Priorität“, BI-SPEKTRUM Fachzeitschrift für Business Intelligence und Data Warehousing, 05-2015

U. Coester, N. Pohlmann: „Diskriminierung und weniger Selbstbestimmung? Die Schattenseiten der Algorithmen“, tec4u, 12/17

N. Pohlmann: „Künstliche Intelligenz und Cybersicherheit - Unausgegoren aber notwendig“, IT-Sicherheit – Fachmagazin für Informationssicherheit und Compliance, DATAKONTEXT-Fachverlag, 1/2019

N. Pohlmann: Lehrbuch „Cyber-Sicherheit“, Springer Vieweg Verlag, Wiesbaden 2019
ISBN 978-3-658-25397-4

Weitere Artikel siehe: <https://norbert-pohlmann.com/artikel/>