Was macht künstliche Intelligenz mit meinen Daten?

Benjamin Walczak

Digitale Woche Kiel 2019

12.09.2019

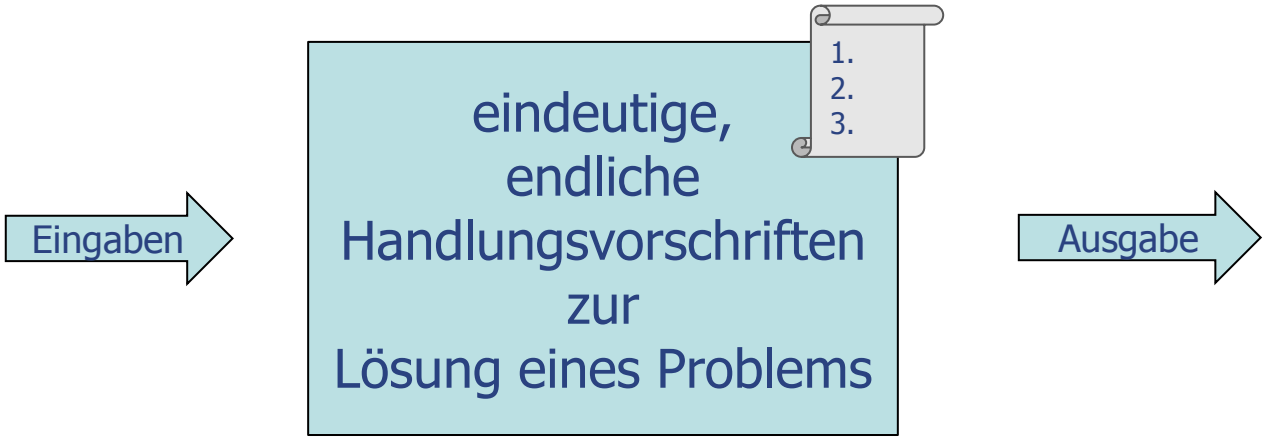
ULD

Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein

Übersicht

1. Kurzeinführung Künstliche Intelligenz
   - Algorithmen
   - Ausgewählte KI-Systeme
2. Künstliche Intelligenz und Datenschutz
   - Probleme
   - Anforderungen und Schutzmaßnahmen
   - Einsatz in öffentlicher Verwaltung

Algorithmen



- Von Menschen geschriebener Lösungsweg
- Gleiche Eingabe führt zu gleicher Ausgabe

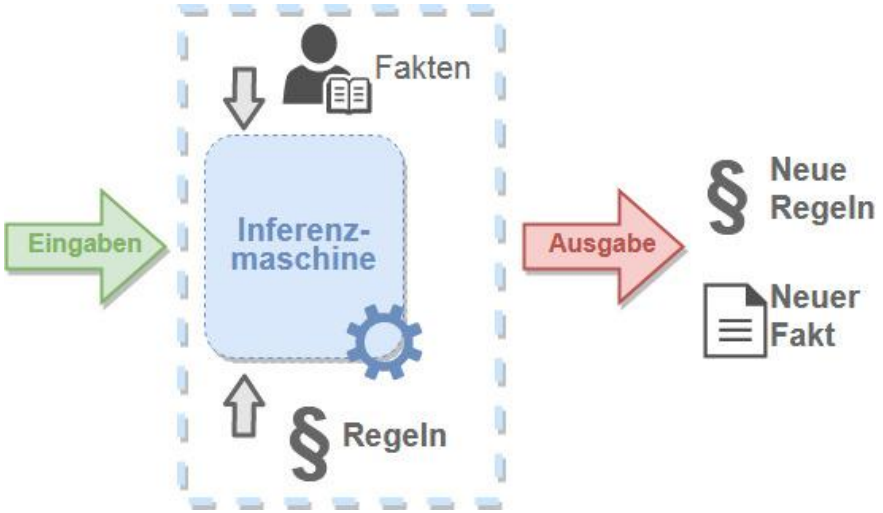
Beispiel: theoretische Führerscheinprüfung

Warum künstliche Intelligenz?

- Komplexe Probleme lassen sich nicht oder nur mit großem Aufwand durch (von Menschen programmierte) Algorithmen lösen.
- Mit starren Algorithmen kann man nur langsam auf sich ändernde (System-)Bedingungen reagieren.
- Wenn Systeme „lernen" können, können sie aus Fehlern und Erfolgen Schlüsse ziehen, um noch bessere Lösungen zu finden.

KI – erste Schritte

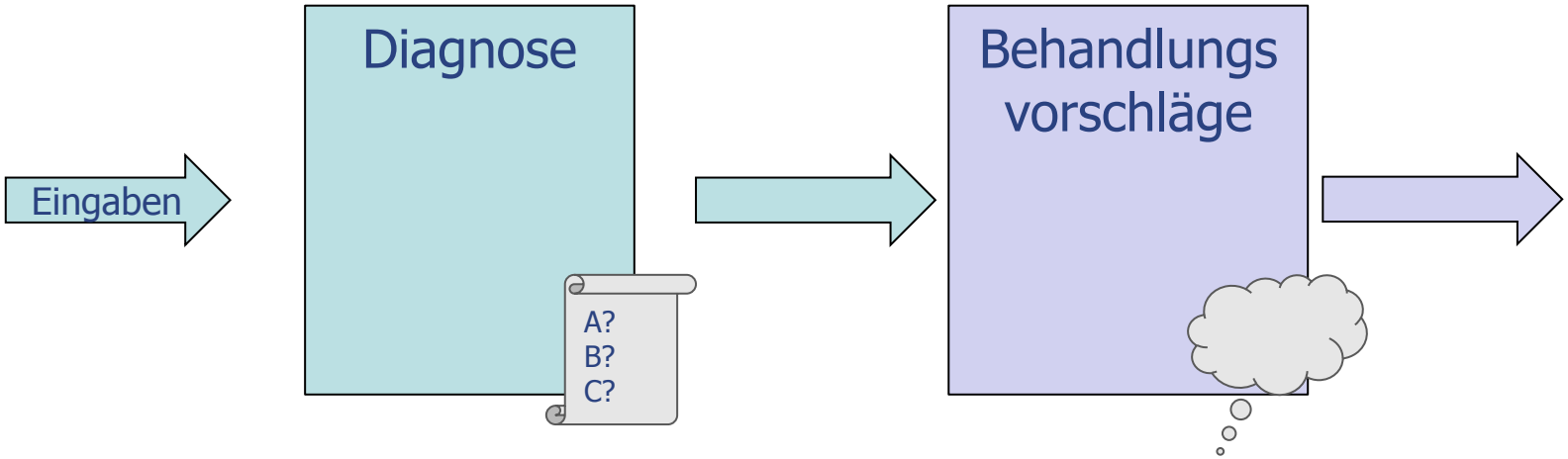
- **Expertensystem**
  Neue Regeln und Fakten werden aus bekannten Regeln und Fakten logisch erschlossen



- **Evolutionäre Algorithmen**
  Mithilfe evolutionärer Entwicklungen werden Merkmale optimiert

Beispiel Expertensystem: Blinddarmentzündung

- LEXMED



- Problem:
  Geht es um Diagnose und/oder Behandlungsvorschlag?
  Ist das einfach zu trennen?

Eingabewerte:

Personenangaben	nicht bekannt	Werte	
Geschlecht	○	● männlich ○ weiblich	?
Altersgruppe	○	○ 0-5 ○ 6-10 ○ 11-15 ○ 16-20 ● 21-25 ○ 26-35 ○ 36-45 ○ 46-55 ○ 56-65 ○ 65-	?
Untersuchungsergebnisse	**nicht untersucht**	**Werte**	
1. Schmerzquadrant	●	○ ja ○ nein	?
2. Schmerzquadrant	●	○ ja ○ nein	?
3. Schmerzquadrant	○	● ja ○ nein	?
4. Schmerzquadrant	●	○ ja ○ nein	?
Abwehrspannung	○	● lokal ○ global ○ keine	?
Loslassschmerz	○	● ja ○ nein	?
Erschütterungsschmerz	○	● ja ○ nein	?
Rektalschmerz	●	○ ja ○ nein	?
Darmgeräusche	○	○ schwach ● normal ○ vermehrt ○ keine	?
Sonographisch auffällig	●	○ ja ○ nein	?
Pathologisches Urinsediment	●	○ ja ○ nein	?
Rektaler Temperaturbereich	○	○ -37.3 ○ 37.4-37.6 ○ 37.7-38.0 ○ 38.1-38.4 ● 38.5-38.9 ○ 39.0-	?
Leukozytenbereich	○	○ 0-6k ○ 6k-8k ○ 8k-10k ○ 10k-12k ● 12k-15k ○ 15k-20k ○ 20k-	?

Abfragen		
Diagnose(4w) ?	Diagnose(3w) ?	Datenbankabfrage ?

Quelle: [2] Ertel, Künstliche Intelligenz

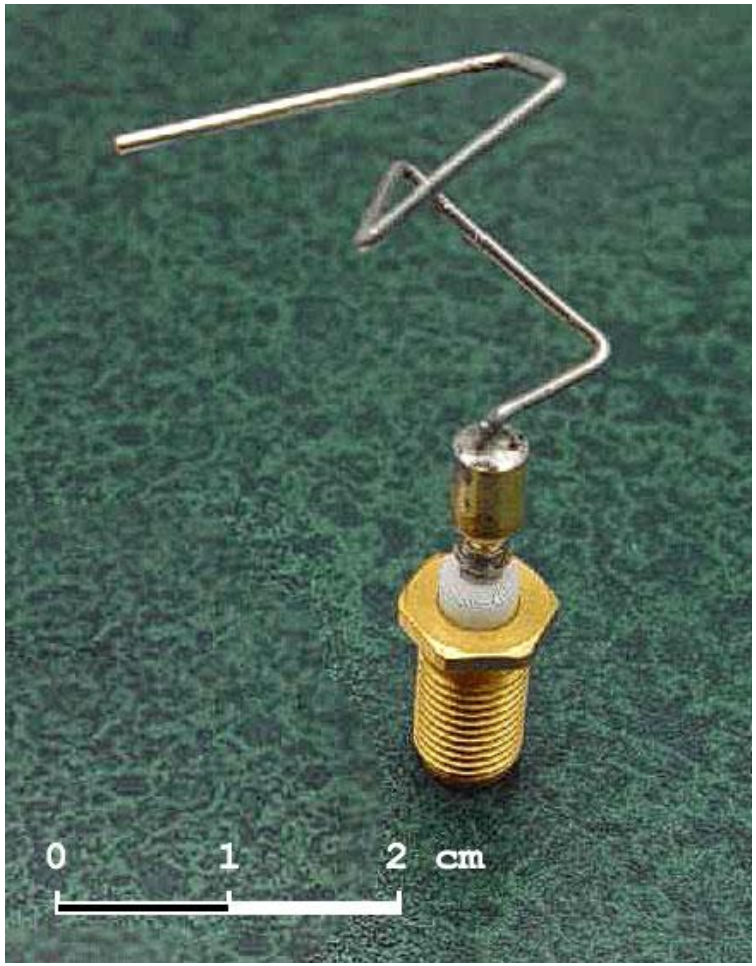
Entscheidungsmatrix

Therapie	Wahrscheinlichkeit für versch. Befunde				
	entzündet	*perforiert*	*negativ*	*andere*	
	0.25	0.15	0.55	0.05	
Operation	0	500	5800	6000	3565
Not-Operation	500	0	6300	6500	3915
Ambulant beob.	12000	150000	0	16500	26325
Sonstiges	3000	5000	1300	0	2215
Stationär beob.	3500	7000	400	600	2175

- Optimale Entscheidungen haben (Mehr-)Kosten 0.
- Ziel: Therapie mit den minimalen mittleren Kosten.

Quelle: [2] Ertel, Künstliche Intelligenz

Beispiel Evolutionärer Algorithmus: Antennen-Gestaltung



Quelle: [3] https://de.wikipedia.org/wiki/Datei:St_5-xband-antenna.jpg

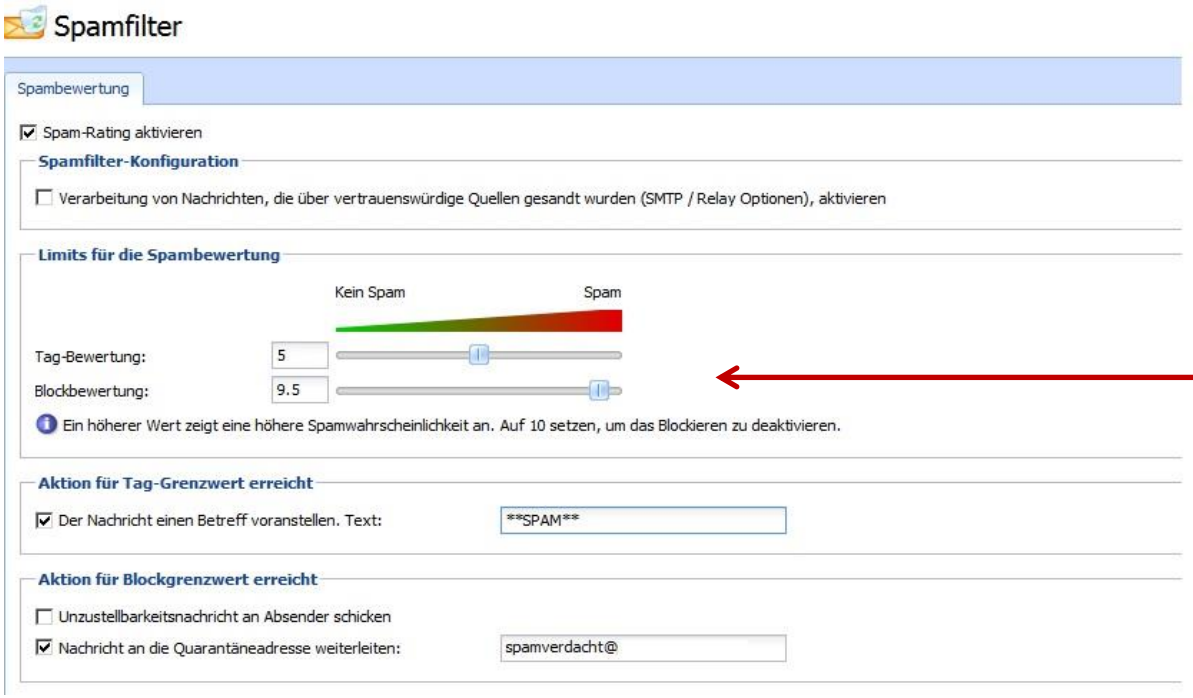
Die Antenne der Space-Technology-5-Satelliten wurde mit einem Evolutionären Algorithmus entwickelt.

Maschinelles Lernen durch Agenten

- Programmierter Agent bewertet Reaktionen der Umwelt und lernt daraus
- „Lernen" erfolgt durch Belohnen bzw. Bestrafen

Beispiel Maschinelles Lernen: Spam-Erkennung bei E-Mail

- Bewertung (Score)
- Tag/Markierung/Flag, in Abhängigkeit vom Scorewert
- Entscheidung, was passieren soll



Spamerkennung bei E-Mail

- Klassifikation als Spam mittels Wahrscheinlichkeiten – "Spam-Score":

X-Spam-Flag: YES
X-Spam-Score: 8.893

Bewertung einer E-Mail ("Scorewert")

X-Spam-Level: ********
X-Spam-Status: Yes, score=8.893 tagged_above=2 required=6.31
tests=[ADVANCE_FEE_4_NEW=2.9, BAYES_50=0.8, DKIM_SIGNED=0.1,
DKIM_VALID=-0.1, DKIM_VALID_AU=-0.1,
FREEMAIL_ENVFROM_END_DIGIT=0.25, FREEMAIL_FROM=0.001,
FREEMAIL_REPLY=1, HTML_MESSAGE=0.001, MISSING_HEADERS=1.021,
RCVD_IN_DNSWL_NONE=-0.0001, SPF_PASS=-0.001,
T_FILL_THIS_FORM_SHORT=0.01, T_KAM_HTML_FONT_INVALID=0.01,
UNCLAIMED_MONEY=2.427, URG_BIZ=0.573, URIBL_BLOCKED=0.001]

Zusammensetzung des Scorewertes aus verschiedenen Teilwerten=Teilbewertungen

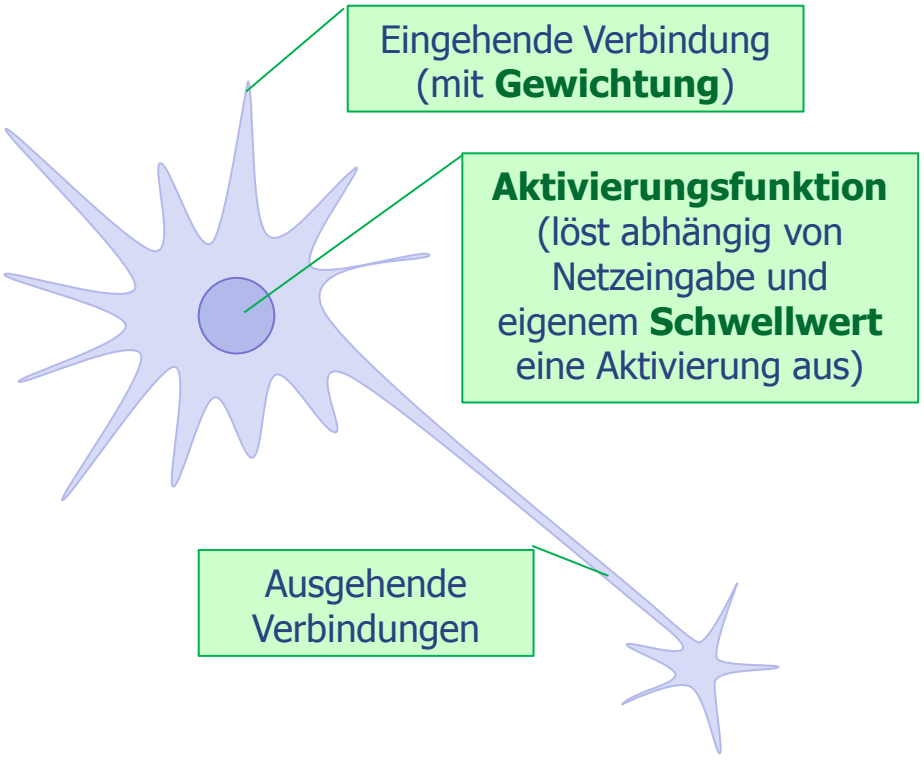
Fehler bei der Spamerkennung

- Agent-Umwelt-Beziehung:
  Spam-Mails werden vom Agenten erkannt, was zu neuartigen Spam-Mails führt, die möglichst noch nicht „erlernt" wurden.

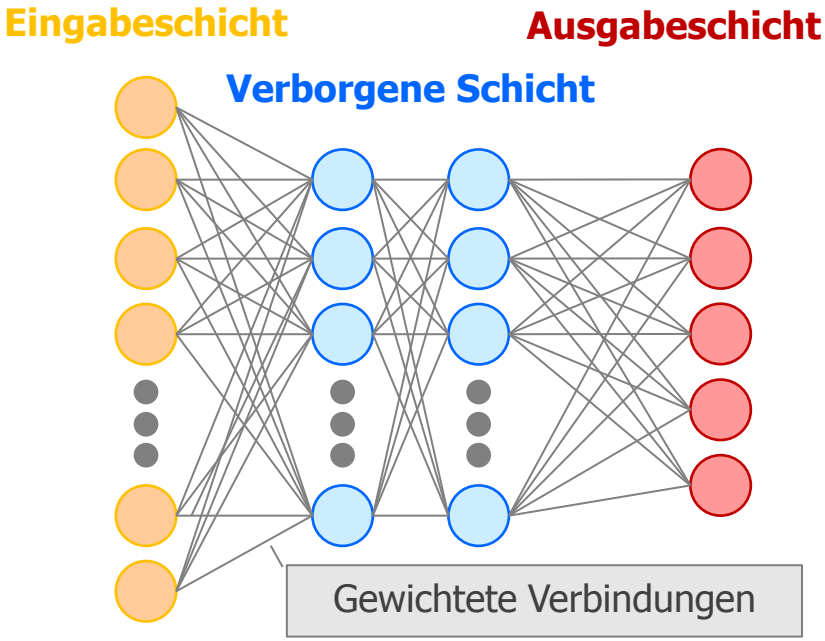
- Training des Agenten:
  - **Fehler 1. Art** als Spam erkannt, aber inhaltlich relevant
    ➔ schwere Bestrafung
  - **Fehler 2. Art** nicht als Spam erkannt, aber belanglos
    ➔ leichte Bestrafung

Künstliche Neuronale Netze

Künstliches Neuron



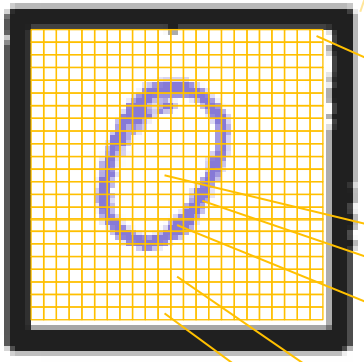
Neuronales Netz



Schon in diesem Beispiel:

6 Eingaben
\+ 8 Schwellwerte
\+ 60 Gewichtungen
= 74 variable Werte

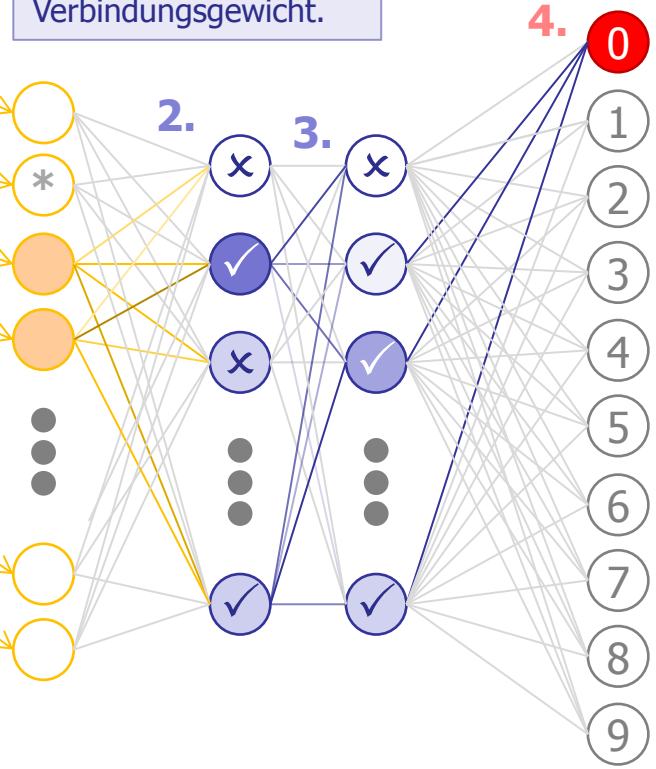
Künstliche Neuronale Netze
Erkennen von geschriebenen Ziffern



1. Jeder Pixel wird einem Eingabeneuron zugewiesen. Es ist aktiviert, wenn der Pixel farbig ist.

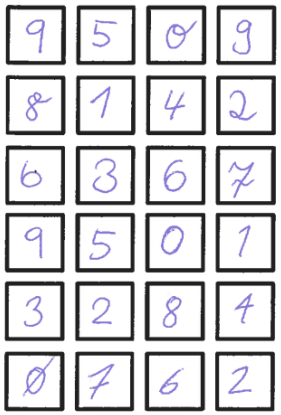
2. Die aktivierten Eingabeneuronen aktivieren weitere Neuronen – abhängig vom Verbindungsgewicht.

3. Sind aktivierte Neuronen (✓) über ihrem Schwellwert aktiviert, aktivieren sie andere Neuronen im Netz.



4. In einem optimalen Neuronalen Netz wird ein Ausgabeneuron aktiviert.

Trainingsdaten:



5. Training:
Um sich einem optimalen Neuronalen Netz anzunähern, muss es trainiert werden – z.B. mit solchen Ziffern, die unterschiedlich geschrieben wurden.

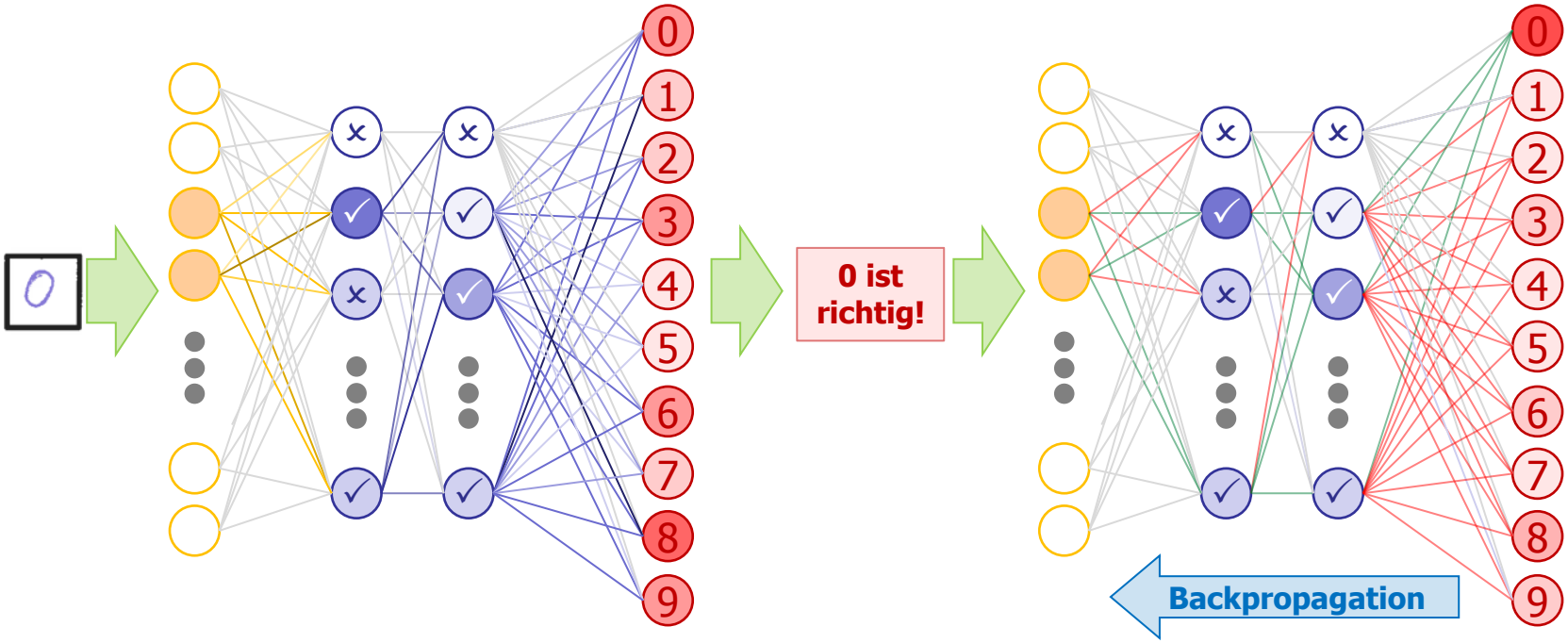
Netzanalyse am Rande:
Zum Beispiel der Pixel in der Mitte und dessen Eingabeneuron (*) sind beim Erkennen einer 0 kaum beteiligt – wohl aber bei anderen Ziffern wie 3, 4, 5, 6, 8 oder 9.

Künstliche Neuronale Netze
Training von geschriebenen Ziffern

Das Neuronale Netz wird mit einer Trainings-Ziffer aktiviert.

In diesem Beispiel hat das Neuronale Netz eine 8 „erkannt", d.h. das entsprechende Ausgabe-Neuron wurde am stärksten aktiviert.
Nun wird das Neuronale Netz mit der richtigen Lösung trainiert.

Die Gewichte der Verbindungen zwischen Neuronen werden angepasst: Von der 0 ausgehend rückwärts bis zu den Eingabe-Neuronen werden alle aktivierten Verbindungen verstärkt. Entsprechend werden für die anderen Ziffern die Verbindungen geschwächt.
Dieses Lernverfahren wird **Backpropagation** genannt.



Künstliche Neuronale Netze

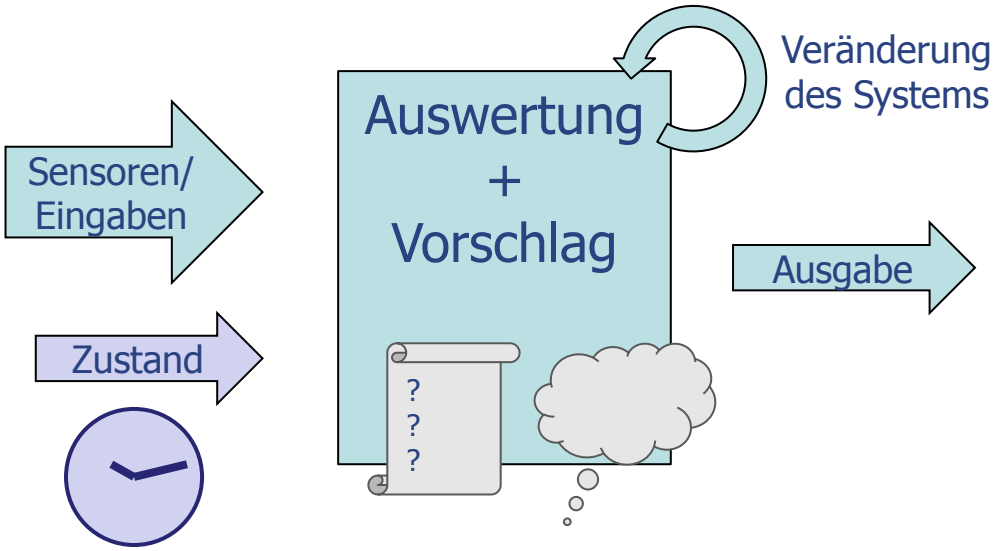
Einige Merkmale künstlicher Intelligenz in neuronalen Netzen:

- **Lernen** erfolgt zunächst auf Trainingsdaten und mit verschiedenen Verfahren (z.B. überwacht, unüberwacht)
- **Erfahrungen/Wissen** ändert sich mit jeder Nutzung. Der Systemzustand ist also ein Berechnungsparameter.
- **Bewertung** einer Lösung erfolgt durch Kostenfunktion, die vorab definiert werden muss.
- **Nachvollziehbarkeit** einer konkreten Berechnung nur schwer möglich, da ein Künstliches Neuronales Netz über eine sehr große Zahl neuronaler Verknüpfungen verfügt.

Schema für Künstliche Intelligenz

Typ	Eingaben	Wissen	System	Bewertung	Zustand	Ausgabe
Experten-system	Werte	Wissensbasis (Fakten+Regeln)	Interferenz-maschine	Logische Schlüsse/ Entscheidungsbaum	stabil	Erschlossene Fakten
Evolutionäre Algorithmen		Anfangspopulation (Zustände+Gene)	Evolut. Entwicklung	Fitnessfunktion	Entwicklung der Population	Entwickelte Population
Maschinelles Lernen	Umwelt (via Sensoren)	Lerngrundlagen	Programm. Agent	Belohnung/ Bestrafung	Änderung von Agent (+Umwelt)	Entscheidung + Verbesserter Agent
KNN	Eingabe-Neuronen	Trainingsdaten	Neuronales Netz	Kostenfunktion	Neuronales Netz	Ausgabe-Neuronen

Künstliche Intelligenz als Black box?



- Welche Sensoren/Eingaben haben überhaupt Einfluss auf das konkrete Ergebnis?
- Welche gelernten Informationen werden berücksichtigt?
- Nachvollziehbarkeit im Nachhinein z.B. durch Prüfen auf Plausibilität (nur geringe Abweichungen, …)

Probleme mit KI aus Datenschutzsicht

- **mangelnde Robustheit** bei kleinen Veränderungen in den Trainingsdaten, z.B. „katastrophisches Vergessen“
- **großer Aufwand** für überwachtes Lernen
- **garantierte Vertrauensniveaus** fehlen, um z.B. Angriffe gegen ein KI-System einzuschätzen (Manipulierbarkeit)
- **Mangel an Interpretierbarkeit und Erklärbarkeit**
- **Gefahr der Diskriminierung**

Anforderung der DSGVO

Die betroffene Person hat das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung - einschließlich Profiling - beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt.

Artikel 22 DSGVO

Konventionelle Schutzmaßnahmen (auf der nicht-kognitiven Ebene)

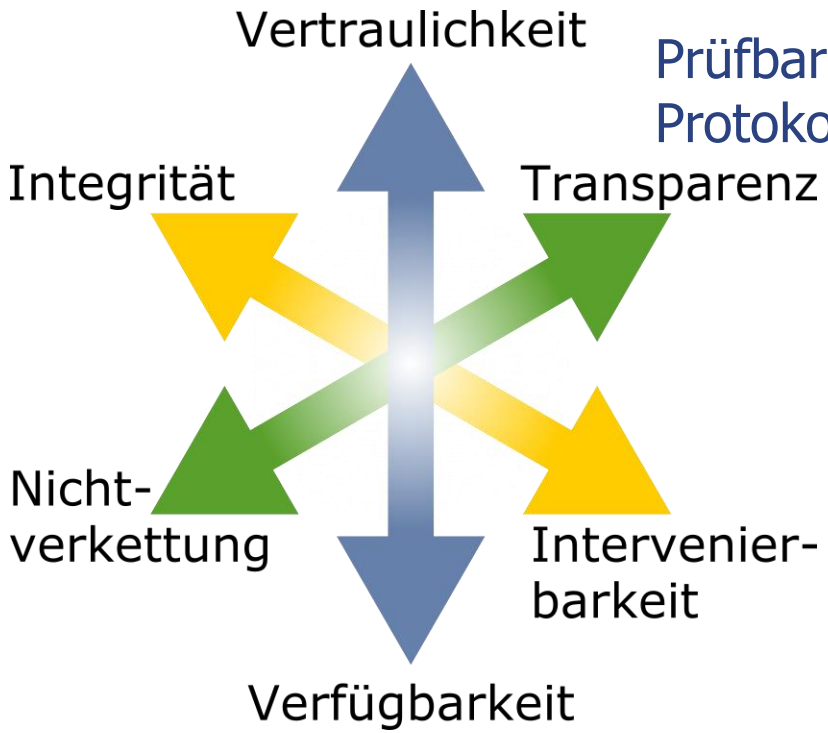
- **redundante Systeme**, Ausfallzeiten, Reparaturvereinbarungen
- **Härten der Systeme** (keine undefinierten Dienste, Nutzer, Zugriffsrechte)
- **Zertifikate bzgl. Verschlüsselung und Integritätssicherung**
- **Umsetzen der Nachweispflichten** durch Spezifikation-, Dokumentation- und Protokollierung der Aktivitäten auf sämtlichen Layern;
- **Interventionsmöglichkeiten** für Betroffene
- **Pseudonymisierung und Anonymisierung** von Daten

Weitere Schutzmaßnahmen

Zugriffe von außen unterbinden

Schutz gegen Manipulation und Fehler des Systems

Prüfbarkeit, Dokumentation, Protokollierung

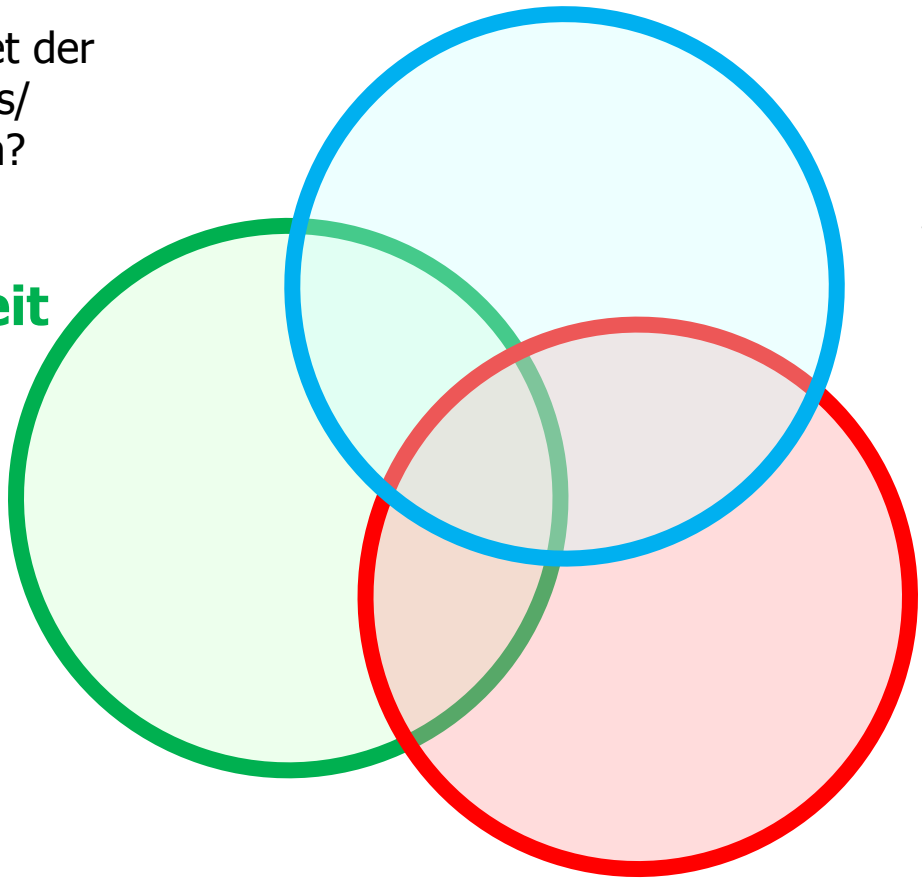


Datenkuration

Grad der Automatisierung

Ersatzsystem; vollständigen Verzicht ermögliche

Einsatz von KI-Systemen in der öffentlichen Verwaltung



Vielen Dank
für die Aufmerksamkeit

Gibt es eine Diskriminierung?

Quellen

[1] Graphik zum Wikipedia-Artikel "Problem des Handlungsreisenden",
https://de.wikipedia.org/wiki/Problem_des_Handlungsreisenden,
abgerufen 05.09.2018
URL der Graphik: https://de.wikipedia.org/wiki/Datei:TSP_Deutschland_3.png

[2] Wolfgang Ertel, Folien zum Buch
Grundkurs Künstliche Intelligenz, *Springer-Verlag, 2011*
www.hs-weingarten.de/~ertel/kibuch, Stand 14. Januar 2018

[3] Graphik zum Wikipedia-Artikel "Evolutionärer Algorithmus",
https://de.wikipedia.org/wiki/Evolution%C3%A4rer_Algorithmus, abgerufen 05.09.2018
URL der Graphik: https://de.wikipedia.org/wiki/Datei:St_5-xband-antenna.jpg