DISKRIMINIERENDE ALGORITHMEN UND DATENHAVARIEN

DIE GESELLSCHAFTLICHE VERANTWORTUNG ZUKÜNFTIGER DATENWISSENSCHAFTLER

Data Science Night FHNW

27. Mai 2021

David M. Sommer

Ich

- David M. Sommer
- Doktorand an der ETH Zürich
- Forschungsschwerpunkte:
  - *Privacy in Machine Learning*
  - *Anonymous Communication*
  - *Differential Privacy*

- Interessiert an den gesellschaftlichen Seite von Informationstechnologien

- Teile dieses Vortrags wurden bereits mit Linus Petermann am Winterkongress 2021 der Digitalen Gesellschaft Schweiz unter dem Titel „Diskriminierende Algorithmen: Technische oder soziale Frage?” gehalten.


Share of market activity
25%
75%
Projected sales of main production 2013

Public Transport Planning and the Technological Revolution

A paradigm shift in public transport is on the horizon, but the right ecosystem of policies, regulations, infrastructure, and skills must first be in place.

India is using facial-recognition to reunite missing children with their families

Police in New Delhi recently trialled facial recognition technology and identified almost 3,000 missing children in four days.

Canada Post reports data breach to 44 large businesses, 950K customers affected

By Twinkle Ghosh · Global News

Posted May 26, 2021 1:24 pm · Updated May 26, 2021 3:51 pm

Hackers leak personal data of 279M Indonesians

Government investigates suspected data leak, sale on hacker platform

Nicky Aulia Widadio | 21.05.2021

Dataset trained Microsoft chatbot to spew racist tweets

In March 2016, Microsoft learned that using Twitter interactions as training data for machine learning algorithms can have dismaying results.

ce Journal

Black Players Fight NFL Brain Injury Payout

m Based on Race

In March, Brod

discriminato

vor 1 Woche

Online-Shopping

Süchtig nach dem besten Preis

Um Menschen zum Kaufen zu bewegen, setzen Online-Plattformen auf Hirnforschung und künstliche Intelligenz. Das soll dynamischen Preisen zum Durchbruch verhelfen.

Daniel Stadelmann, Harry Stitzel

Freitag, 14.05.2021, 18:53 Uhr

Daten und AI: Gesellschaftliche Herausforderungen

- Werden mehr und mehr gesammelt, eingesetzt und verwertet
  - *Viel positives aber auch viel negatives Potential*
  - *Gefahr der übermassigen Vernetzung und Verselbstständigung*
    - US Credit Score System
    - SCHUFA-Score in Deutschland
    - weniger tragisch in der Schweiz
- Datenqualität/-bias/-ursprung
- Von der Einzelbetrachtungsweise zum Gemeinschaftsdenken
- Technologie beeinflusst Gesellschaft und vice versa
- Gesellschaftliche Normen können sich ändern
  - *Was passiert dann mit den Daten und Algorithmen?*
- Nicht komplett durchregulierbar (Geschäftsgeheimnis & Innovationshindernd)
  - *Verantwotrung bei Entwicklern und Umsetzer*

→ Ihr Datenwissenschaftler habt was zu tun ;)

Aufbau

- KI-Systeme: Begriff, Einsatz, Verbreitung
- Datenschutz
- Beispiel Diskriminierungsproblematik
- Technische Grundlagen: Funktionsweise, Problematiken und Widersprüche
- Rechtliche Grundlagen und mögliche Rahmenbedingungen
- Was tun? Perspektiven für Gesellschaft und Politik
- Fazit

ADM Systeme: Begriff, Einsatz

- KI, AI, Machine Learning, Neuronale Netzwerke…
  - *Unklare / verschwommene Definitionen*

→ **Automated Decision Making** (ADM) Systeme

- Heutiger Haupteinsatzzweck
  - *Effizienzsteigerung in Prozessen*
    - Ermöglicht Verarbeitung viel grösserer Datenmengen
  - *Sammlung und Verwertung von vorher unerreichbaren Informationen*
    - Z.B.: Bewegungsdaten, Gesundheitsdaten via Smartphones
- Grosse Bandbreite
  - Automatisierung von Entscheidungs- und Dienstleistungsprozessen
  - Polizei- und Nachrichtendienste

Datenschutz!

- Ideologisch: Schutz des Rechts auf informationelle Selbstbestimmung & Wahrung der Privatsphäre
- Praxis: Schutz vor missbräuchlicher Datenverarbeitung & Datenauskunftsrecht
- Warum: Wie viel Macht wollen wir den Datensammlern geben?
  - *Eigene Bewegungsfreiheit: Stasi (DDR) oder Fichen-Skandal (CH)*
  - *Wählermanipulation mit Cambridge Analytica*
  - *(Facebook Muslims List for Trump)*
- Datenschutzgesetze verlangen oft
  - *Zweckbindung, Korrektheit, Rechenschaftspflicht, zeitliche Speicherbegrenzung*
- Fokusierung auf den Einzelnen oder nur bestimmte Gruppen / Attribute
  - *Rechte sollte nicht für die die gelten, die sich wehren können*
  - *Benötigen Instrumente wie Sammelklagen (kollektiver Rechtsbehelf)*

- Regelmässig riesige Datenlecks
  - *Wöchentlich mehrere Lecks mit über 1Million Betroffenen*
  - *CAM4 data breach 10 Milliarden Datenreihen (März 2020)*
  - *Yahoo data breach 3 Milliarden Accounts (2017)*

→ **Minimalprinzip: Wenig Daten gespeichert – wenig Daten abgeflossen**

Hilft Datenschutz bei AMD Systemen?

- Informationen sind an vielen Verschiedenen Orten gespeichert
  - *Schwer, eine Übersicht zu erhalten*
  - *Daten werden dort auch genutzt / von Algorithmen verwertet*
- Datenschutz greift nur bei persönlichen Daten
  - *Bsp. nicht bei: Optimale Platzierung von Feuerwehrwachen (basierend auf öffentlichen Daten)*
- Algorithmen können trainiert werden, ohne Privatsphäre Einzelner zu verletzen (z.B. Differential Privacy)
  - *Keine Verhindert von Gruppennachteilen*

- **Schweiz:** Neues Datenschutzgesetz, Art. 21 (ab 2022)
  - Recht auf Informierung und Überprüfung bei / von vollautomatisierten Entscheiden
- **Europäische Union**, DS-GVO Art. 22 Abs. 1
  - *Recht auf nicht ausschließlich automatisierte Verarbeitung bei rechtlicher Wirkung*

→ Enge Definition, schränkt nur Extremfälle ein, nur Einzelfallperspektive

→ **Kein ausreichender Schutz durch Datenschutz!**
  - *Datenschutz verminderte extensives Verteilung von (potentiell fehlerhaften) Informationen*
  - *Datenschutz ist ein Schritt in die richtige Richtung, reicht aber nicht.*

Automatisierte Selektion in Bewerbungsverfahren

- Einfaches Beispiel:
  - *Algorithmus: Vorselektion Bewerbungsgespräche*
  - *Trainiert an alten Bewerbungsunterlagen, sucht Muster für «eingestellt»*
  - *Optimales Muster «eingestellt»: Schweizer Pass, Wohnort Zürich, Zweisprachig*
- Aktuelle Bewerber*innen:

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Direkte und indirekte Diskriminierung

- Selektion nach Nationalität → Direkte Diskriminierung
  - *Lösung: Attribut entfernen*
- Selektion nach Wohnort → Indirekte Diskriminierung
  - *Starker Zusammenhang mit nicht-schweizerischer Nationalität*
  - *Lösung?*

→ Reproduktion sozialer Normen

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Wandelnde Normen, unkritische Algorithmen

- Früher: Ungleichbehandlung von Frauen in jungem Alter
  - *Gerechtfertigt und mit «sachlichem Grund» (potentielle Schwangerschaft)*
- Heute: Rechtfertigung wird zunehmend prekär

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Geschlecht*	*Alter*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	w	30
Uljanow	Schweiz, Russland	MA ZHAW (Fachhochschule)	Deutsch, Englisch, Russisch	m	32
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	w	43

Soziale Prägung der Technik

- Diskriminierung nicht «nur» technisches Problem
  - *Algorithmus reproduziert einfach*
- Bestimmung von Diskriminierung ist offen und wird gesellschaftlich ausgehandelt
  - *Was ist «sachlich»?*
- Vorstellungen und Ordnung von Welt prägen Algorithmen

  → *Immer eine Übertragung von Werten und Normen in Technik*

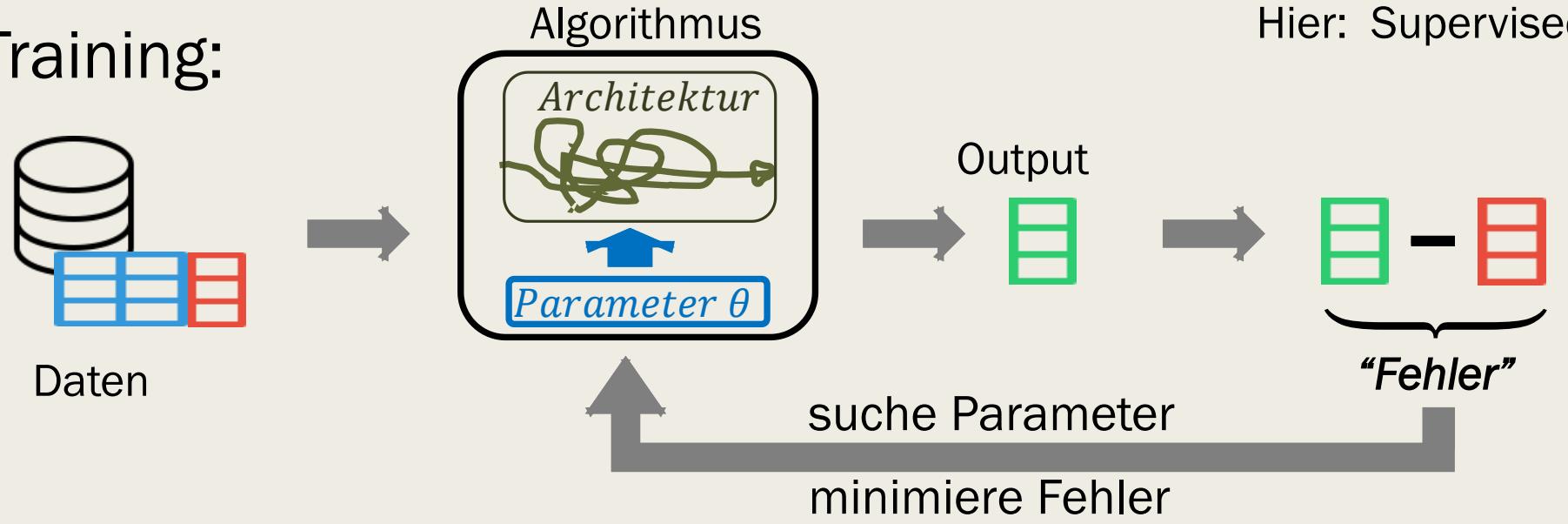
Algorithmen-Arten

- Supervised Algorithms
  - *Labels für Training vorhanden: Bild A (Sample) ist Hund (Label)*
  - *Bsp.: Logistische Regression, Multi-Layer-Perceptron*
  - *Häufigste Klasse*
- Unsupervised Algorithms
  - *Keine Labels*
  - *Bsp.: Clustering Algorithmus (z.B. k-means)*
- Reinforcement-Learning
  - *Algorithmus lernt mit Feedback-Funktion*
  - *Bsp.: smarter Staubsauger fährt, bis er Wand erreicht*

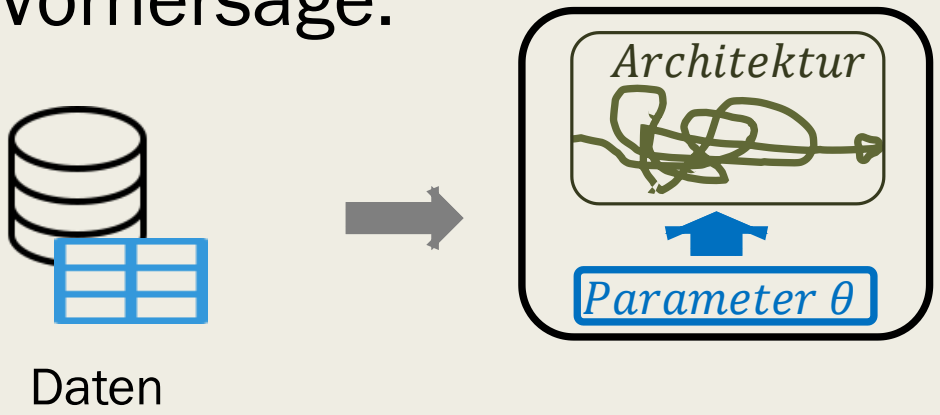
Wie funktionieren diese Algorithmen?

Hier: Supervised Classification

Training:



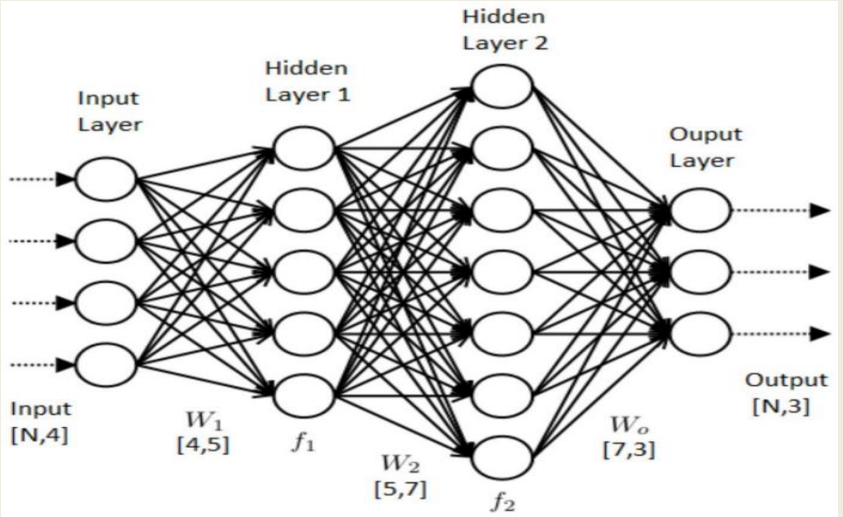
Vorhersage:



Qualitätsmessung, oft: Vorhersagegenauigkeit (im Schnitt)

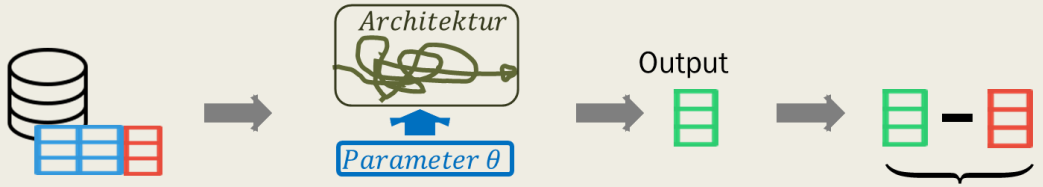
Kausalität

- Kausalität:
  → *Warum ein Algorithmus wie klassifiziert*
- Sinn ist nicht immer ersichtlich
  → Algorithmus findet unbekannte Korrelation
- Kann auch falsche Schlüsse ziehen
  → Wolf wird an Schnee im Hintergrund erkannt
- Zunahme Komplexität: Z.B. Neuronale Netze:
  → Milliarden an Parameter
  → Kausalität schwierig



Quelle Bilder:
https://s3.amazonaws.com/media.eremedia.com/wp-content/uploads/sites/4/2019/12/16105353/AI-wolf-or-husky.jpg
https://miro.medium.com/max/700/1*ZB6H4HuF58VcMOWbdpcRxQ.png

Bias (Verzerrung)



- Rohdaten beinhaltet bereits Diskriminierung
  - *Früher wurden mehr Schweizer eingestellt*
    ⟶ Algorithmus empfiehlt eher Schweizer
- *Population Bias*: Untergruppen waren nicht Teil des Trainings:
  - *Seifenspender funktioniert bei dunkler Hautfarbe nicht*
- *Representation Bias*: Untergruppen werden vernachlässigt
  - *Niedriger Anteil ⟶ grosser "Fehler" Einzelner im Schnitt vernachlässigbar*
- *Algorithmic Bias*: Architektur-/Design-Entscheidungen
  - *Algorithmus schafft es nicht, auf wichtige Details einzugehen*
  - *Definition von Fehler hat enorme Auswirkungen auf Resultat*

Fairness?

- Angesprochene Algorithmen unterscheiden zwingend.
- Wie bekommt man diese “fair”
  - *Anpassung der Rohdaten*
  - *Korrektur beim „Fehler“*
- Was ist fair?
- Individuelle vs. Gruppen-Fairness:
  - *Bsp: Priorisierte Einstellung von Frauen in Männerdomäne*
  - *Für Parität im Schnitt: Frau hat deutlich höhere individuelle Einstellungs-Chancen*
- Generell: Prioritisierung eines Zusammenhangs schwächt meistens einen anderen
  - Grund: andere Daten-Ausgangsstruktur --> andere Dynamik
- Trade-off zwischen Effizienz und Fairness
  - *Schwächt Nützlichkeit des Algorithmus*
- Algorithmen zwar neutral, reproduzieren aber Datenstruktur und "künstliche" Zwänge

→ Fairness ist technisch nicht eindeutig, benötigt gesellschaftliche Diskussion

Stand ADM Regulationen

- Transparenz beim Einsatz
  - *Kennzeichnung, komplette Veröffentlichung?*
- Recht der Zivilgesellschaft auf Einblick und nach Rechtfertigung durch Verantwortliche
- **Sanktionen!**
  - *Klare Haftung, von Menschen oder Unternehmen*
  - *Keine Auslagerung an andere Juristriktionen (Google vs. Spanien,* CJEU, 2014*)*

- EU: Gesetzesvorschlag veröffentlicht 21. April 2021 [0]
  - *Biometrische Massenüberwachung unzureichend verboten („Reclaim your face „)*
- Schweiz: diverse Strategiepapiere [1,2], kein konkreter Gesetzesvorschlag
  - *Europarat (Nicht EU): Committee on Artificial Intelligence (CAHAI) [3]*
  - *Schweiz ist Mitglied (Exekutive ist involviert)*
  - *29. April 2021: Multi-Stakeholder Consultation geschlossen*

[0] EU-Rat: Proposal for a Regulation laying down harmonised rules on artificial intelligence (2021)
[1] Bericht an Bundesrat: Herausforderungen der künstlichen Intelligenz (Ende 2019)
[2] Leitlinien für den Umgang mit künstlicher Intelligenz in der Bundesverwaltung (November 2020)
[3] https://www.coe.int/en/web/artificial-intelligence/cahai

Stand ADM Regulationen II

- Generell: 3-4 Kategorien
  - *Low risk: keine/wenig Regulation (z.B. Korrektheitskontrollen von Formularen)*
  - *High risk: weitgehende Regulationen (z.B. gezielte politische Werbung)*
  - *Verboten: z.B. Biometrische Identifikation für Massenüberwachung*

- Zwei Ansätze scheinen sich durchzusetzen:
  - *Kritikalitätsansatz: Regulierungsanforderungen je nach Schadensrisiko*
  - *Schadensansatz: Bei Verstoss kann geklagt werden (privatwirtschaftlich)*

- Unterscheidung zwischen Staat und Privatwirtschaft
- Idee: Keine Unterscheidung zwischen Entscheiden von Mensch oder Maschine

Gesellschaftlicher Umgang mit ADM Systemen

- Weiterentwicklung und schnelle Verbreitung von ADM Systemen absehbar
- **Schleier der Neutralität und Technikgläubigkeit aufbrechen:**
  - *Verschleiert Risiken*
  - *Werte- und Normübertragungen in Technik sichtbar machen und reflektieren*
- Gefahr: Verstärkung sozialer Normen, versteckte Diskriminierung
  - *Algorithmen ohne «Gewissen» → Gefahr der Verstärkung sozialer Normen*
  - *Konventionsdruck*
- ADM-Systeme zum Teil von Debatten machen:
  - *Datenbasis: Welche/Wessen Daten, welche Filter*
  - *Struktur/Architektur: Art des Algorithmus, Erneuerungszeitpunkte*
  - *Einsatz: Wann, Wofür, Wie? → Ob überhaupt?*
- Positiver Seiteneffekt: verbesserte statistische Erkennbarkeit von Diskriminierung

Fazit

- Reproduktion und Verstärkung sozialer Normen und bestehender Ungleichbehandlung
  - *Diskriminierende Gesellschaft → diskriminierende Daten & Algorithmen*
- Technische Lösungen beschränkt
  - *Diskriminierung technisch nicht komplett auflösbar*
  - *Widersprüche in Fairnessdefinitionen*
  - *→ Gesellschaftliche „Begleitung“ und Diskussion technischer Entwicklung notwendig*
- Technikgläubigkeit verführt zur Verantwortungsabgabe
- Problem ist nicht lösbar aus der Perspektive von Einzelpersonen
  - *Kollektive Ansicht benötigt*
- Regulierungsrahmen ungenügend / nicht vorhanden
  - *Intensive Diskussionen, Regulierungen absehbar*
  - *Zivilgesellschaft vs. Unternehmens- und Regierungsbestrebungen stark asymmetrisch*
  - *Rechte müssen erkämpft werden*

Fazit für Datenwissenschaftler

- Datenschutz reicht nicht
  - *Daten können trotz/je nach Datenschutzgesetz ausgewertet werden*
  - *Es gibt automatisierte oder datengetriebene Entscheide, die nicht von persönlichen Daten abhängen.*
- Regulationen werden vermutlich nicht tief genug gehen können
  - *Es liegt viel Entscheidungskompetenz bei euch Datenwissenschaftler*
  - *Wirtschaftlichen Druck motiviert Regulationen bis ans Maximum auszunutzen*
- Seid euch bewusst wie gesellschaftliche Normen die Algorithmen und Datenkolletionen beeinflussen
  - *Und umgekehrt.*

Weiterführende Links und Literatur

Hagendorff, Thilo (2019): *Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze*, in: *Österreichische Zeitschrift für Soziologie*, Jg. 44, Nr. 1, S. 53–66, doi: 10.1007/s11614-019-00347-2.

Orwat, Carsten (2019): *Diskriminierungsrisiken durch Verwendung von Algorithmen*, (Studie) Berlin: Antidiskriminierungsstelle des Bundes.

Automating Society Report Schweiz (2020) - AlgorithmWatch /CH

Sommer, David (2021): *Machine-Learning-Leitlinien des Bundes*, [online] https://www.digitale-gesellschaft.ch/2020/12/17/machine-learning-leitlinien-des-bundes-entscheidende-gefahren-nicht-benannt/

Heise: Kollektiver Datenschutz (https://www.heise.de/hintergrund/Kollektiver-Datenschutz-Was-dahinter-steckt-und-warum-er-noetig-ist-6054822.html)

Bilderquellen

- https://www.capgemini.com/de-de/news/gartner-magic-quadrant-data-analytics/
- https://www.computerweekly.com/de/meinung/Das-Potenzial-von-kuenstlicher-Intelligenz-in-Unternehmen
- https://www.computerweekly.com/visuals/German/article/ai-robot-macine-learning-blackboard-adobe.jpg
- https://www.dvz.de/fileadmin/_processed_/6/c/csm_14d02405a_Datenberg_Big-Data_Online_Illu-Luedemann_3860bd8469.png