Christian Ekhart

Big Data und Metadaten

DEUTSCH

ABSTRACT

Big Data und Metadaten sind insbesondere im Gefolge der zahlreichen Skandale, die sich um Datenmissbrauch entwickelt haben, in aller Munde. Allgemein werden diese Begriffe oft emotional als Gefährdung verstanden. Dieser Beitrag versucht eine allgemein verständliche Begriffsklärung, eine Einschätzung der Situation im Hinblick auf die Potentiale von *Künstlicher Intelligenz* in der Datenverarbeitung und einen kurzen Ausblick auf die mittelfristige Entwicklung zu geben.

ENGLISH

Everyone is talking about big data and meta data in the wake of several scandals surrounding the misuse of data. The emotional conversation often paints these terms as a threat. This article aims to provide a universal and comprehensible definition, an assessment of the situation regarding the potential of artificial intelligence *in data processing, and an overview of medium-term developments.*

| BIOGRAPHY

Christian Ekhart, Dipl.-Ing., ist CEO von icomedias und Principal Consultant mit Schwerpunkt Healthcare und Public Safety. Er war 1995 Gründer einer der ersten Internet-Agenturen in Österreich, wurde mit mehreren österreichischen Staatspreisen ausgezeichnet und war zweimal Microsoft Partner of the Year für Mobile Business Formulare – HybridForms. Er arbeitet für internationale Einrichtungen sowie Verwaltungs-, Polizei- und Katastrophenschutz-Behörden im Bereich vernetzte und mobile Anwendungen mit sicherheitskritischen Daten.

| KEY WORDS

Anonymität; Big Data; Metadaten; Musterbezüge; Profiling

Einleitender Ausblick

Dieser Beitrag tendiert zum Essay, mit vielen Verweisen auf Wikipedia[1], aber er wurde vor dem Hintergrund von 30 Jahren Erfahrung mit Datennetzen, großen Datenmengen und Markttendenzen verfasst. Es geht vornehmlich um eine Betrachtung der letzten Jahre und den Versuch eines Ausblickes, wobei ein solcher – auch bei umfassender Datenlage und reicher Erfahrung – immer risikobehaftet ist.

Sehr oft haben uns in den letzten Jahrzehnten insbesondere technologische Entwicklungen überrascht, die in dieser Form nicht vorhersehbar waren. Daher ist jede Prognose ein Aufzeigen von konkreten Möglichkeiten, ja, Wahrscheinlichkeiten, keineswegs jedoch Sicherheiten – die Geschichte zeigt immer wieder unerwartet große Veränderungen, unerklärbare scheinende Beharrung oder auch enorme Resilienz. Wir sind dabei einer Gleichzeitigkeit unterschiedlichster Eindrücke ausgesetzt, die zum Teil in sich widersprüchlich scheinen, etwa in den gleichzeitigen Erfahrungen der Unverzichtbarkeit mobiler Datenendgeräte und deren zeitraubendem und datensammelndem Potential. Das ist allerdings nichts völlig Neues – seit der Mensch technologische Entwicklungen macht und nutzt, sieht er sich mit diesbezüglichen Dilemmata konfrontiert, die zugleich Chancen und Risiken bedeuten. Oder kurz gesagt: Kognitive Dissonanz ist normaler Bestandteil des Lebens; sie wird am Beispiel von Big Data und Metadaten vielleicht nur besonders deutlich.

Daten und Metadaten

Dass die Menge an Daten enorm und exponentiell wächst, das wissen – und nutzen wir. Die Steigerung der Datenmenge kommt nur zu einem geringen Maß zustande, weil wir mehr „Schriftsteller“ und Produzenten von Content haben oder weil mehr Content ins Digitale gehoben wird, wobei „heben“ durchaus eine Steigerung der Zugänglichkeit und der Verarbeitbarkeit meint.

Nein: Es fallen extrem viel mehr Daten an durch ubiquitäre Digitalisierung und eine ganz extrem gesteigerte Granularität der Datenquellen. Und mit IoT[2] – dem Internet der Dinge mit Sensoren an allen physischen Enden der Welt – stoßen wir seit wenigen Jahren eine weitere große Dimension auf.

Daten können auch mehr werden: Durch zeitliche und koinzidente Verbindungen zu anderen Daten entstehen Metadaten.

1 Für tieferes systematisches Eintauchen ist Wikipedia durchaus zu empfehlen; gerade die technischen Beiträge dort sind meist gut recherchiert bzw. redigiert. Außerdem sind dort zahlreiche weiterführende Referenzen zu finden.

2 Vgl. https://en.wikipedia.org/wiki/Internet_of_things (oft auch als „web of things“ bezeichnet).

Ich verwende mehrfach in diesem Text als einfache Beispiele elektronische Publikationen, den Kauf von Artikeln im Onlineversand sowie Social Media wie Facebook. Die Anwendbarkeit ist aber universell und real deutlich weiter gefasst.
Ich möchte nun zunächst einige Begriffe so definieren, wie ich sie anwende.

- *Daten:* Das sind „Nutzinhalte“ oder Content, also etwa dieser Artikel.
- *Metadaten 1:* Daten, die diese Nutzinhalte beschreiben. Also beschreibend (Titel, Autor, Datum…), strukturell zuordnend (Hierarchien, Taxonomien, Orte…) und administrativ (Einordnung, Rechte…).
- *Metadaten 2:* Einerseits Daten und Protokolle über die Nutzung (wann erstellt, wann/von wem abgerufen, wie lange wo verfügbar…), andererseits und zusätzlich dazu dynamische Verbindungen zu anderen Daten mit Musterbezügen.
- *Metadaten 3:* Daten über die kollektive (Massen-) Nutzung von Angeboten. Das sind Nutzung, Zustimmung und Ablehnung, Kommentierung, Verweildauern, Re-Publikation etc.
- *Musterbezüge:* Das sind Relationen zu anderen Daten (und Metadaten) aufgrund von ähnlichen oder gegenteiligen Mustern in Daten wie in Metadaten. Als Beispiel sei die bekannte Empfehlung weiterer Produkte bei Amazon genannt. Die Erkennung von Mustern findet oft nicht auf Basis systematischer und vordefinierter kausaler Planung statt. Vielmehr wird Big Data[3] dafür genutzt, automatisch Zusammenhänge bzw. Koinzidenzen zu erkennen, vielfach ohne für die „menschliche“ Betrachtung ersichtliche und direkte Kausalität. Musterbezüge können hoch dynamisch sein. Themenverwandte Publikationen zu einem Artikel können beispielsweise laufend erscheinen oder geändert werden Und auch diese Dynamik kann als Protokoll zu weiteren – oft ganz wesentlichen – Metadaten werden.
- *Profilierung:* Das fügt eine weitere Metaebene hinzu. Daten, Nutzung, Nutzer und derlei Bezüge sind sehr unterschiedlich, mitunter nur oberflächlich sehr verschieden, mitunter qualitativ. Profilierung ist eine Art der Klassifizierung und Sammlung für bestimmte Anwendungsfälle. So kommt es zu einer benutzerspezifischen Zuordnung einer Vorliebe für Küchenutensilien und französische Kochbücher beim Stöbern im Angebot von Amazon und in weiterer Folge zu entsprechenden Angeboten.

3 Vgl. https://en.wikipedia.org/wiki/Big_data.

Data Science, Big Data und Business Intelligence

Bis vor wenigen Jahren war die Datenverarbeitung auf überwiegend (für heutige Verhältnisse) geringere Datenmengen und vordefiniert algorithmische Prozesse beschränkt. Das war begründet im Fehlen der (erheblichen) Rechnerkapazitäten – und es war auch nicht am gedanklichen Horizont.
Das hat sich grundlegend geändert: Durch deutliche Weiterentwicklungen in der Hardware, zunehmende Vernetzung und dynamische Ressourcen durch Cloud-Computing stehen heute exponentiell erweiterte Möglichkeiten zur Verfügung.[4]

Die Dynamisierung der Ressourcen erweitert die Möglichkeiten im Datenraum exponentiell.

Ohne ersichtliche und direkte Kausalität: Big Data ist die Verarbeitbarkeit und Durchführung dessen mit (sehr!) großen Datenmengen. Sehr große Datenmengen in begrenzter Zeit sind etwa die Betriebs- und Sensordaten der Jet-Triebwerke[5] eines Verkehrsflugzeuges am Flug über den Atlantik von 10-600 TB in zehn Stunden.
Viele der dabei erfassten Daten weisen keinerlei oder zumindest keine direkte Kausalität mit Ereignissen (wie etwa der Ausfall eines Triebwerkes) auf.
Mit Data Science, Business Intelligence und Maschinenlernen bzw. Künstlicher Intelligenz werden diese Daten und Medatataen analysiert, Muster gefunden und verfeinert, Cluster gebildet und visualisiert. Die dazu verwendeten Systeme sind „lernfähig", perfektionieren ihre Vorgangsweise also auch laufend. Damit werden mehrere Daten bzw. Daten-Änderungen gesamt zu einer mit einer Eintritts-Wahrscheinlichkeit bewerteten Vorankündigung für ein Ereignis.
Oft tauchen diese Muster auf, bevor diese Information über (eigentlich) primäre Quellen verfügbar wird. Ebenso kann dies Ereignisse vorwegnehmend beeinflussen:

- Thyssen-Krupp[6] unterzieht Aufzugsanlagen einer zeitnahen vorsorgenden (und damit planbaren = billigeren) Wartung, wenn sich ein Ausfall ankündigt – und vermeidet diesen dadurch.
- Der Musikdienst Spotify spielt den Abonnenten neue Musik vor, die sie hören wollen: Profilierung und das dynamische Nutzungsverhalten anderer Nutzer dieses Profils.

4 Nur um die Größenordnung zu verdeutlichen: Ein typischer Heimcomputer der 1980er war der Commodore C64; weit verbreitet, aber bei weitem nicht in jedem Haushalt vorhanden (weltweit wurden ca. 30 Mio. Exemplare verkauft). Die Geräte waren nicht vernetzbar. Heute steht in praktisch jedem Haushalt mindestens ein Gerät, das ca. 3.400mal so schnell ist und fast 250.000mal so viel Speicher hat, nämlich eine Standard-Notebook der aktuellen Generation. Dass Netzwerkfähigkeit aus keinem Endgerät wegzudenken ist, braucht man nicht eigens zu erwähnen.

5 Vgl. http://aviationweek.com/connected-aerospace/internet-aircraft-things-industry-set-be-transformed.

6 Vgl. https://max.thyssenkrupp-elevator.com/en/.

- Google erstellt geografische „Heatmaps“ von Grippe-Erkrankungen, und zwar auf der Basis von Suchdaten – deutlich bevor die Patienten zum Arzt gehen, googeln sie ihre eigenen Symptome, was den Suchalgorithmen auffällt.
- Erdbeben tauchen unmittelbar in den Karten von Fitnesstracker-Apps auf, und zwar durch unüblich gehäufte Aktivität in bestimmten Regionen, weil Menschen aufschrecken und sich etwa nachts bewegen.
- Facebook filtert und priorisiert (wenige) Beiträge (vieler) anderer Nutzer auf Basis von Nutzerverhalten und Profilierung.[7] Amazon macht es ebenso.

Big Data und alles damit ist auch Big Business – und Big Politics.

Profile, Filterblasen und Echokammern

Der unmittelbar sichtbare Vorteil dieser „Datenveredelung“ sind zugeschnittener Content, bessere Vorhersagen (etwa bei Google Maps Navigation im Auto) und weniger Informationsüberfluss bzw. weniger intellektuelle Lese- und Einordnungsanforderung. Big Data weiß, was wir wollen – und hilft uns dabei, es zu finden bzw. präsentiert es uns schon unaufgefordert.

Muster und Profilierungen sind selbstverstärkend – sie fokussieren Inhalte mit jeder entsprechenden Nutzung in zunehmendem Maß.

Das ist nicht vorrangig negativ. Dabei übergehe ich das Datenschutzproblem vorerst bewusst – ob es gut ist, dass außer mir noch andere über Interessen, Vorlieben und Verhalten fast alles weiß, ist ein anderes Thema. Dazu später etwas mehr unter dem Stichwort Selbstbestimmung.
Die Anwendung dieser Muster und Profilierungen birgt ganz wesentlich eine andere Gefahr: Diese Filterung hebt vorrangig vermeintlich gewünschte Inhalte in die Aufmerksamkeit der Nutzer. Und mit jeder Nutzung verfestigt sich dieser Wunsch einerseits für das filternde System – andererseits vielfach für den so versorgten Nutzer.
Sehr vereinfacht formuliert: wer Wienerschnitzel liebt, wird nur noch mit Gebackenem versorgt. Doch wenn Sie Gemüse lieben, wollen Sie sich vielleicht nicht ausschließlich vegan ernähren…

7 Die im Frühjahr 2018 bekannt gewordene Affäre der von Facebook an Cambridge Analytica weitergegebenen Daten ist nur eine Metaform des Grundproblems. Die von Cambridge Analytica angewandten Strategien wurden bzw. werden weiterhin natürlich auch Facebook-intern angewandt.

Komplexer: Meinungen, Denkmuster und Gruppen ähnlich Denkender formieren, verstärken und verfestigen sich durch positiven Zuspruch und durch Abschwächung ausgewogener Information.
Dies wird durch Big Data unterstützt, trifft aber nicht ausschließlich darauf zu. Verglichen mit wenigen (Broadcast)-Medien vor dem Internet-Boom gibt es heute viel mehr oft sehr einschlägige Informations-Quellen und -Produzenten. Eine Selektion und damit letztlich die Filterung erfolgt auch eigenverantwortlich! im Extremfall endet man selbst gefangen in den Echokammern bevorzugter Angebote, wofür man aber letztlich niemandem – außer dem eigenen Selektionsverhalten – die unmittelbare Schuld geben kann.

Erkenntnisse durch Big Data und Machine Learning

Machine Learning[8] und Künstliche Intelligenz[9,10] ermöglichen nicht nur das Erkennen bisher unbekannter Zusammenhänge, sondern auch das Finden ganz neuer Lösungsansätze. Dies kann ganz wesentlich zu neuen Erkenntnissen führen. Beispielhaft seien angeführt:

- Aufnahmen bildgebender Verfahren in der Medizin können vorsortiert und auffällige Bereiche darin markiert werden;
- Erkennung von gesellschaftlich anstößigen oder verbotenen Inhalten und die darauf folgende Streichung der betreffenden Inhalte z. B. aus Suchergebnissen oder deren Löschung aus Datenbanken;[11]
- Überprüfung von Sozialleistungen durch Umfeldanalyse[12] der Bezieher;
- Predictive Policing – Berechnung von z. B. Einbruchswahrscheinlichkeiten oder Autodiebstählen in Parkgaragen und der vermehrten Präsenz von Polizeikräften;
- Predictive Maintenance – vorausschauende Wartung von z. B. Aufzugsanlagen, Flugzeug-Triebwerke.

Die Anwendungsmöglichkeiten sind natürlich wesentlich zahlreicher als diese kurze Auflistung.
Die Basis für alles maschinelle Erkennen und Lernen sind große Datenmengen. Die herausragende Eigenschaft von Machine Learning ist die enorme Geschwindigkeit und iterative Anpassung. Mit einem dynamischen Teil des Datenbestandes wird erkannt und gelernt – und dies wird gegen

8 Vgl. https://en.wikipedia.org/wiki/Machine_learning.
9 Vgl. https://medium.com/iot-forall/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-3aa67bff5991.
10 Vgl. https://en.wikipedia.org/wiki/Artificial_intelligence.
11 Hier machen wir private Konzerne und Maschinen zu vorsorglichen Wächtern oder sogar Richtern.
12 Beispiele dafür gibt es etwa in Großbritannien. Hier wird eine Abwägung „Datenschutz vs. berechtigtes Interesse der Gesellschaft“ vorgenommen.

den übrigen Datenbestand verifiziert. Das erinnert nicht nur an Evolution und Mutation - nein, diese Ansätze werden ganz methodisch verwendet. Die Ergebnisse liegen – ähnlich wie die Produkte evolutiver Prozesse – oft jenseits des menschlich erwartbaren. Es entsteht Neues; und das Neue erschafft möglicherweise bzw. sogar wahrscheinlich selbst noch Neueres. Ein interessantes Beispiel – weil aufgrund der unerwarteten Ergebnisse und der unkontrollierbaren Entwicklung abgeschaltet – ist die autonome Entwicklung einer künstlichen, neuen Sprache[13] zwischen zwei Botsystemen.

Datensammeln über Dinge, Umwelt und Nutzung

Das IoT – „Internet of Things“ oder das „Internet der Dinge“ ist die großflächige und (oft) sehr detaillierte und hochvolumige Sammlung von Daten. Das können Betriebsparameter und millionenfache Sensorik-Daten eines Jet-Triebwerks auf einem Atlantikflug sein, oder – natürlich in geringerem Umfang – die Messdaten aus einem Personenaufzug. Dazu kommen noch Wetterdaten, Auslastungen, Korrelationen mit Tages- und Wochenrhythmen, Feiertagen etc.,verschnitten mit Ereignissen wie Ausfällen und Reparaturen. All das fließt in Muster- und Lernmodelle, die dazu dienen, Ereignisse wie Ausfälle und Schäden von Anlagen vorab zu erkennen und Maßnahmen auszulösen. So kann etwa ein Monteur zu einer Aufzugsanlage geschickt werden, die mit einer hohen Wahrscheinlichkeit in Kürze einen Ausfall haben wird. Diese vorausschauende Wartung erfolgt aber nicht auf einer starren Zeitbasis („alle halben Jahre“), sondern individuell für jede Anlage auf Basis der dort erfassten Daten.
Das Modell kann sehr ähnlich für die Vorhersage von (stark) erhöhten Diebstahlhäufigkeiten von PKWs in Tiefgaragen oder von Haus- und Wohnungseinbrüchen eingesetzt werden.[14]
Wie weit sich solche Methoden eventuell selbst wieder bremsen (bedeutet ein verhinderter Einbruch nicht eine Verlagerung an andere und für diesen Zweck besser geeignete Orte und damit eine Änderung des Datenmodells?) ist derzeit noch nicht hinreichend fassbar.

Datensammeln? Das machen wir selbst!

Um bei den polizeilichen Beispielen von Diebstahl und Einbrüchen zu bleiben: auch andere Daten fließen stark in diese Rechenmodelle ein. Aller-

13 Vgl. https://www.huffingtonpost.com.au/2017/08/02/facebook-shuts-down-ai-robot-after-it-creates-its-own-language_a_23058978/.

14 Die filmische Fiktion wird hier von der Wirklichkeit mit technischen Mitteln eingeholt: Minority Report (Steven Spielberg, US 2002).

dings sammeln wir selbst diese ergänzenden Daten selbst und stellen sie quasi öffentlich zur Verfügung: Und zwar im Wege der Social Media.
So häufen sich Bilder und Texte über bestimmte Inhalte etwa vor der Bildung von Demonstrationen (etwa im Umfeld der Wall Street-Proteste[15] in New York 2017). Und mit diesen selbst veröffentlichen, weiterveröffentlichten (etwa „Retweets" auf Twitter oder „teilen" auf Facebook) sowie mit den Geo-Daten sowie mit der Unterstützung von Bild-/Texterkennung in Fotos entsteht ein zeitnahes Lagebild. Sicherheitsdienste setzen diese Daten, Analysen und Prognosen schon seit mehreren Jahren ein – die Systeme entwickeln sich mit immer mehr Daten teilautonom weiter und erreichen immer bessere Genauigkeiten.

Viele Daten für Profile entstehen als „Nebenprodukte" fast unbewusst durch die Nutzung von standardmäßig eingerichteten und durchaus nützlichen Services bzw. Apps.

Nicht nur in Social Media „gespiegeltes Leben" und Verbindungen zu anderen „ähnlichen" Publikationen und Produzenten („Likes") sagen sehr viel über die Nutzerinnen und Nutzer aus. Dazu kommt noch die „Selbstvermessung" durch Dinge wie Fitness-Tracker, regelmäßige Laufstrecken,[16] Gesundheitsdaten und viele mehr. Vielfach sind wir also selbst die Sensoren – und sorgen gleichzeitig für eine strukturierte Publikation dieser Daten.
Mit Funktionen wie Geo-Bezügen etwa auf Bild- und Videomaterial und der Erkennung und Klassifizierung dieser Medieninhalte, entstehen hier weitere Daten (aus dem Bereich Content und Meta).

Datensammeln? Das machen andere!

Die meisten Daten sammeln wir also letztlich selbst, oder stellen sie selbst zur Verfügung. Welche Bahn- oder Flugtickets wir wann mit welchem Zahlungsmittel und welchen Beginn- und Endpunkten wir (für wen) buchen, welche Telefonnummern oder Internet-Adressen wir (wiederholt) abrufen, von wo wir das tun – und welche Arten von Inhalten wir dabei konsumieren – all das geht noch wesentlich ohne detailliertes Wissen über die „echten" Nutzinhalte oder Motivationen.
Andererseits werden unsere Daten im Alltag ja intensiv „geerntet". Mautsysteme der Autobahn erkennen KFZ-Kennzeichen und damit Zeit und Ort

15 Vgl. https://en.wikipedia.org/wiki/Occupy_Wall_Street.
16 Über häufige und regelmäßige Lauf-Trackings in wenig bewohnten Gebieten konnten unlängst mit hoher Wahrscheinlichkeit die Standorte geheimer Militärstützpunkte „erraten" werden. Vgl. https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

unserer Bewegung, Geschwindigkeitskontrollsysteme ebenso. Wasser- und Stromversorger wissen durch intelligente Zähler „(Smart meter“) immer besser über unser Nutzungsverhalten Bescheid. Banken und Kreditkarten- bzw. Bezahlsystembetreiber kennen unsere Konsumvorlieben ebenso wie der Online-Versandhandel, Auskunftssysteme, Suchmaschinen etc.

Daten aus dem Bereich Gesundheit, Soziales und Risikofaktoren sind für Versicherungen wertvolle Entscheidungsgrundlage für Vertragsabschlüsse.

Hier sammeln andere sehr viele Daten. Und zumindest die Daten im „eigenen Bereich“ sind wertvolle Rohstoffe, die schon gehoben und genutzt oder zumindest systematisch protokolliert und für spätere Verwendung gelagert werden. So kennt Amazon unser Einkaufsverhalten und unsere Vorlieben und reagiertmit zielgerichteten Angeboten hoher Kaufwahrscheinlichkeit und dynamischen Preisen für die (eigene) Ertragsoptimierung. Bei Facebook werden Inhalte und deren Verstärkung mit Werbung verbunden und dynamisch optimiert (wodurch die oben erwähnten „Echokammern“ entstehen). Optimiert ist das alles auf zeitlich (beste „Nutzungsbereitschaft“) und inhaltlich (Bestärkung und offene Aufnahme) optimal angepasste individuelle Ansprache.

Dazu kommen noch Daten aus dem Bereich Soziales, Gesundheit und zu Versicherungsleistungen (die z. B. für Versicherungen hoch relevant für die Entscheidung für oder gegen einen neuen Vertragsabschluss werden können).[17]

Sofern (!) wir von allen diesen Daten überhaupt wissen, haben wir heute weitgehend noch selbst die Kontrolle (durch selektive Nicht-/Nutzung von Services), und diese Daten sind derzeit überwiegend auf die jeweiligen Verarbeitungsinseln der Anbieter beschränkt. Hinzu kommt die Möglichkeit, auf der Basis des Datenschutzgesetzes und der Datschschutz-Grundverordnung Auskunft über die und Löschung der eigenen Daten zu begehren.

Jedoch: Um solches zu begehren, muss man einerseits wissen, wer über meine Daten verfügt, und diese Firmen müssten im Geltungsbereich der entsprechenden Regulierungen ansässig sein. Beide Fälle sind nicht die Regel. Weiters:

Der sehr weit und einfach geöffnete Zugriff der Finanzämter auf Bankdaten oder von Sicherheitsbehörden auf Telekommunikations-Verbindungsdaten reißt diese Grenzen teilweise ein – derzeit noch auf eher individueller Basis. Begehren wie systematische und verdachtsunabhängige Screenings

17 Im April 2018 wurde über eine Regierungsvorlage in Österreich die Möglichkeit geschaffen, dass Dritte auf Gesundheitsdaten (die „ELGA“) von Bürgerinnen und Bürgern zugreifen können. Trotz Anonymisierung ist hier dem Missbrauch Tür und Tor geöffnet – das einfache Ersetzen eines Namens mit einer Kennzahl z. B. ist angesichts der heuristischen Möglichkeiten einer maschinellen Analyse nur wenig wirksam. Anonymisierung ist nicht mehr mit einfachen Mitteln zu erreichen. Vgl. https://www.sn.at/politik/innenpolitik/gesundheit-sakte-elga-regierung-will-daten-der-oesterreicher-fuer-forschung-oeffnen-26528515.

dürfen aber nicht überraschen und werden politisch und wirtschaftlich konsequent betrieben.

Datensammeln, Spurenlesen und Anonymität

Die Nutzung von Anonymisierungssystemen, oder alleine schon die Abschaltung der für das Nutzer-Tracking notwendigen Funktionen ist ebenso ein Metadatum, das bei Nutzung auffällt und erfasst wird. Dazu bedarf ein noch keiner NSA- und Geheimdienst-Kontroll-Phantasie, diese Daten fallen jetzt schon offen an. Auch wenn z. B. bei der Nutzung des TOR-Netzwerkes oder dem Übermitteln stark verschlüsselter Inhalte (wahrscheinlich) nicht bekannt ist, WAS übermittelt wurde, reicht die Informationen über das „sichere Übermitteln" möglicherweise schon für Änderungen in einem Verdachts-Scoring.
Oft ist die Nutzung nur insofern anonym, als einem Inhalt kein Klarnutzer bzw. Klarname zugeordnet werden kann. Vielfach kann diesen Nutzungen aber ein Identifier, Browser-Fingerprint oder ein ähnliches „wahrscheinlich" eindeutiges Wieder-/Erkennungsmerkmal zugeordnet werden.
Anonymität hat oft mit (Daten-)Sparsamkeit zu tun. Und Sparen ist unbequem. Zusätzlich kann die beste Sparsamkeit, Tarnung und Anonymisierung mit dem einen Mal umkippen, wenn an einem Punkt der Nutzung Klarname und anonymes Wiedererkennungsmerkmal zusammen auftreten. Bei guten Datensammlungen sind dann alle anonymen Nutzungen auf einmal historisch aufgerollt.

Mit Daten bezahlen – und eine Meinung

Alles kostet etwas – und allermeistens zahlen wir selbst (über Umwege). Zumindest in den meisten Fällen. Es gilt das ökonomische „There ain't no such thing as a free lunch".[18]
Meistens werden vordergründig kostenlose Angebote wie Suchmaschinen, Websites von Zeitungen usw. mit Werbung bezahlt. Diese Werbung macht in erster Linie Sinn, wenn sie zielgerichtet ist – also auf Basis von gesammelten Nutzungs-Metadaten und Klassifizierungen auf diesem oder übergeordneten Werbe-Systemen. „Störungsverminderer" wie Ad-blocker[19] sind daher nicht sehr beliebt und verhindern oft die kostenlose Nutzung bzw. sind diese Angebote und Websites erst nach einer Abschaltung des Ad-

18 Vgl. https://en.wikipedia.org/wiki/There_ain%27t_no_such_thing_as_a_free_lunch.
19 Vgl. https://techcrunch.com/2018/04/20/adblock-plus-v-axel-springer/.

Blockers bzw. White-listing[20] zugänglich. Das ist Teil einer gegenseitigen Nutzungsübereinkunft – und auch grundsätzlich in Ordnung (wenn es in den AGB klar kommuniziert wird). Schließlich bekommt man ja auch etwas, was man will.

Mitunter ist schon alleine die Sammlung und Anreicherung von Daten für die Nutzung für deren Weitergabe an Werbenetzwerke eine Art der Zahlung. Auch bei Bezahl-Angeboten (die einen Konsum „ohne störende Werbung" versprechen) ist nicht automatisch sichergestellt, dass nicht ebenso Daten gesammelt werden – eben unter einem anderen Gesichtspunkt und mit anderer Methodik.

Über andere Daten wie persönliche bei Amazon, Twitter oder Facebook– haben wir weitgehend noch die Nutzungskontrolle– auch wenn es zweifellos immer wieder Datenmissbrauch[21] gibt.

Die Datenschutz-Grundverordnung (DSGVO) räumt dem Nutzer wieder mehr Kontrolle über seine Daten ein, gefährdet aber dadurch die Existenz einiger Services, an die man sich inzwischen gewöhnt hat.

Der Einsatz von Profilierung und zielgerichteter individueller Kommunikation sind – da sei mir eine persönliche Einschätzung erlaubt – nicht geeignet, demokratische Entscheidungen wie die Entscheidung zum Brexit oder die US Präsidentenwahl maßgeblich und anders verändernd zu beeinflussen als es über klassische Massenmedien seit vielen Jahrzehnten möglich ist. Damit ist das nur genauer möglich. Der Rest sind festgelegte demokratische Regeln. Wenn man weitreichend ändernde Entscheidungen für Brexit mit einfacher Mehrheit macht, dann ist das so entschieden – auch wenn ein ohnedies knapper Ausgang dadurch ebenso knapp auch kippen kann.

In letzter Zeit gibt es immer mehr Druck auf die großen Daten-Verarbeiter wie Facebook, keine Daten mehr mit anderen Anbietern zu teilen. Doch dadurch verfestigen sich andererseits exklusive Datenmonopole. Werbung kann daher etwa nur mehr durch Facebook zielgerichtet angeboten werden. Steigende Anforderungen im Datenschutz wie etwa GDPR/EU-DS-GVO[22] und die Begrenzung der Weitergabe an Dritte führen zu mehr und besserer Privacy. Wir haben wieder (?) mehr Kontrolle über die Daten.

Das bedeutet natürlich auch, dass die Geschäftsmodelle vieler Anbieter so nicht mehr funktionieren – bis auf die großen Anbieter mit hinreichend großen eigenen Dateninseln.

20 Aufhebung des Ad-blockers für dieses Angebot im lokalen Browser.
21 Vgl. https://www.forbes.com/sites/jodywestby/2018/03/27/what-is-amazing-about-the-facebook-cambridge-analytica-story.
22 Vgl. https://en.wikipedia.org/wiki/General_Data_Protection_Regulation.

Wenn meine Daten mir gehören und oft mit Daten bezahlt wird, könnte ich ja entscheiden, dass ich meine Daten freiwillig teilen möchte. In der Tat passiert genau das etwa im Bereich „Öffentliche Sicherheit“ und „Grenzübertritt“. Bestimmte Daten werden dazu schon ganz offiziell abgeglichen. Für die Einreise in die USA kann von den Sicherheitsbehörden gezielt die Öffnung der eigenen Social-Media Zugänge aufgetragen werden. Diese laufen dann durch ein (KI und Big Data) Screening und Scoring. Ist dabei alles OK, geht es schneller bei der Einreise. Das legitime Sicherheitsinteresse des Staates, die persönliche Bequemlichkeit und der Anspruch auf Privacy haben ein veritables Vereinbarkeitsdilemma.
Zu dieser Freiwilligkeit gibt es seitens des World Economic Forum entsprechende Überlegungen[23] der Systematisierung so eines Screenings.

Ein persönliche Einschätzung und ein Ausblick

Big Data, Künstliche Intelligenz und Machine Learning werden – nach aller Erfahrung der letzten wenigen Jahre – häufiger, genauer und schneller. Damit lässt sich viel erreichen und viele Erkenntnisse deutlich außerhalb traditioneller Erwartungsrahmen gewinnen. Das ist vorerst eine sehr neutrale Tatsache.
Umfang und Geschwindigkeit dieser Entwicklung gehen über alle bisherigen Erfahrungswerte hinaus. Der Vorsprung des menschlichen Geistes wird – quantitativ betrachtet – kleiner und bisweilen schon mit höherer Taktzahl[24] überholt; nicht nur durch enorme Rechenkraft zur vielstufigen numerischen Lösung aller möglichen Resultate, sondern durch eigenständig weiter-lernende und optimierte Lösungsansätze: Machine Learning. Ähnlich unvollkommen wie der Mensch, anders, aber ebenso erfolgreich.
Das Grundübereinkommen der westlich demokratischen Gesellschaften sind Regeln, Teilhabe, Repräsentation und Gewaltentrennung im Interesse der– nie vollkommenen – Freiheit und Sicherheit. Der Einzelne verzichtet abgesehen von wenigen gesetzlichen Ausnahmefällen auf Gewaltausübung und übergibt diese als Monopol an den Staat. Dieser sorgt im Gegenzug für Sicherheit. Dies wird immer wieder angepasst an geänderte Rahmenbedingungen auf der Basis von neuer Erkenntnisse und und freier mehrheitlicher Entscheidung. Diese Prozesse benötigen Zeit.
Veränderungen werden kommen, und in Bereichen, die wir bisher als nicht betroffen angenommen haben. Roboter haben von Menschen schwere[25] und hoch präzise[26] sowie repetitive Tätigkeiten übernommen. Sehr erfolg-

23 Vgl. http://www3.weforum.org/docs/WEF_The_Known_Traveller_Digital_Identity_Concept.pdf.
24 Dies bezieht sich auf einen – letztlich hinkenden – Vergleich: Die parallele Signalverarbeitung im Gehirn kann letztlich nicht mit jener der derzeitigen Computer verglichen werden; diese verwenden daher für AI-Prozesse „brute force“ im Sinne von extrem hoher Taktung.
25 Etwa die Fertigung von KFZ-Teilen, die Tätigkeit von Schweißrobotern etc.
26 Etwa in chirurgischen Behandlungen, so beim Einfräsen von künstlichen Hüftgelenken.

reich – und mit deutlicher Auswirkung auf die Arbeitswelt. Derzeit stehen wir auf der Stufe zur „Denk-Erkennungs-Übernahme“ etwa im Bank- und Versicherungsbereich[27] mit künstlich intelligenter Sachbearbeitungsautomation.

Ich knüpfe an die Einleitung an: Wir leben in einer divergenten Welt und haben dennoch angesichts der Unzahl der unterschiedlichen Interaktionsebenen eigentlich erstaunlich wenige Divergenzerfahrungen. In meinem Stamm-Café bevorzuge ich die Bedienung durch das Personal, das mich bereits kennt und das mir ungefragt den Espresso bringt. Doch auch die automatische Musikauswahl meines Streaming-Dienstes funktioniert erstaunlich gut: Beide kennen mich auf je eigene Weise.
Die Geschichte zeigt oft genug unerwartete große Veränderungen, unerklärbare scheinende Beharrung wie auch enorme Resilienz der Gesellschaft. Das bezieht sich auch auf die aktuellen technologischen Entwicklungen, von denen ich nicht jeden Aspekt uneingeschränkt optimistisch sehe, aber dennoch das positive Potential für größer halte als die Gefahren. Oder kurz gesagt: Die kognitive Dissonanz, die wir im Augenblick im Umgang mit Big Data erleben, ist normal, wir nennen das „Leben leben“. Und es kann gut gehen – und besser werden.

27 Vgl. https://www.bankingtech.com/2018/03/will-ai-replace-humans-in-the-insurance-industry/.