Nordrhein-Westfälische Akademie
der Wissenschaften und der Künste

Nordrhein-Westfälische Akademie
der Wissenschaften und der Künste

Herausgegeben von der
Nordrhein-Westfälischen Akademie der Wissenschaften
und der Künste

Katharina Morik, Walter Krämer (Hg.)

DATEN – WEM GEHÖREN SIE, WER SPEICHERT SIE, WER DARF AUF SIE ZUGREIFEN?

Ferdinand Schöningh

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Internet: www.schoeningh.de

Herstellung: Brill Deutschland GmbH, Paderborn

ISBN 978-3-506-79248-8

INHALTSVERZEICHNIS

VORWORT

Big Data – Wem gehören sie, wer darf sie nutzen?

Wird das Duo Big Data/Digitalisierung eines Tages in den Geschichtsbüchern den gleichen Rang einnehmen wie das Schießpulver, der Buchdruck oder die Dampfmaschine? Gemeinsamkeiten gibt es genug. Diese Erfindungen haben die Welt verändert, keine Frage. Und sie konnten diese weltbewegende Kraft erst in Kombination mit anderen Innovationen schöpfen: Beim Schießpulver waren es die Fortschritte in der Metallherstellung und -verarbeitung, beim Buchdruck die kurz zuvor erst mögliche Massenherstellung von billigem Papier und bei der Dampfmaschine der gleichzeitige Aufschwung des Kohle-Bergbaus in Großbritannien. Genauso hätte auch die Erfindung des Elektronenrechners durch Konrad Zuse 1940 niemals ohne die nachfolgende Explosion der Möglichkeiten der Speichertechnik und Datengewinnung ihre epochale Wucht entfalten können (und wäre die Prognose des seinerzeitigen IBM-Chefs Watson – es gibt auf der Welt einen Bedarf von vielleicht fünf Computern – möglicherweise wirklich eingetroffen).

Erst die Kombination unterschiedlicher Innovationen verleiht einem Gesamtpaket oft die große Kraft. Informationen und Daten zum Beispiel gab es schon immer und nicht viel weniger als heute. Die Temperatur auf der Nordseite des Matterhorns am Weihnachtsabend existierte solange das Matterhorn existiert. Und dass Frau X hat am 10. Januar 1998 bei Aldi Süd in Mainz-Bretzenheim zwei Flaschen Rotwein der Marke Z gekauft hat, war schon seinerzeit ein

Fakt. Aber die Verbreitung dieser Information beschränkte sich auf Frau X und möglicherweise die junge Dame an der Kasse. Heute weiß es, sollte Frau X mit Kreditkarte bezahlt haben, im Prinzip die ganz Welt. Big Data ist kein Phänomen der Existenz, sondern der Verfügbarkeit von Daten. Und diese Verfügbarkeit nimmt heute durch das Zusammenwirken von immer effizienterer Rechner- und Speichertechnik auf der einen und immer ertragreicherer Datenfischerei auf der anderen Seite gigantische Ausmaße an. Über den größten Teil der Menschheitsgeschichte wussten wenige wenig, heute wissen viele viel und der Zug geht mit Volldampf in Richtung alle wissen alles.

Der vorliegende Sammelband gibt die Referate einer Fachtagung der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste wieder, die im Juni 2017 die daraus entstehenden Probleme des Datenschutzes, der Eigentumsrechte und des Zugriffs aus der Warte verschiedener Wissenschaften beleuchtet hat. Damit beteiligt sich die Akademie an der dringend nötigen Diskussion, wie diesen Herausforderungen durch die Informatik, die Statistik, durch Rechts- und Wirtschaftswisenschaften und durch Ingenieure zu begegnen ist. Schon im Mai 2015 hat die EU-Kommission bei der Erarbeitung ihrer Standards für den digitalen Binnenmarkt viele der im Weiteren angesprochenen Probleme identifiziert. Die Bitkom mit ihrem „Excellence in Big Data" Bericht (2016) und der deutsche Bundesverband IT-Mittelstand beschäftigt sich damit genauso wie die Arbeitsgemeinschaft Digitaler Neustart der deutschen Justizminister-Konferenz. Die Politik hat die Brisanz des Themas erkannt. Hier folgt eine Analyse aus der Sicht der Wissenschaft.

Der erste Vortrag von Katharina Morik führt die nötigen Grundlagen aus der Informatik ein. Kurz und knapp, aber allgemein verständlich, wird der zeitliche Ablauf der Digitalisierung dargestellt. Kaum eine Wissenschaft kommt mehr ohne Daten aus, oft auch so viele, dass sie nur mit maschinellen Lernverfahren zu bewältigen sind. Bei Geschäftsmodellen gibt es viel Spielraum für Entscheidungen: wollen wir unsere Daten lokal bei uns behalten? Fordern wir Privatheit erhaltende Dienstleistungen? Können wir jetzt einfach we-

niger arbeiten und Maschinen machen lassen? Der Abschnitt zu Potenzialen und Gefahren betont die gesellschaftliche Frage, wie das Potenzial der Daten zum allgemeinen Wohl ausgeschöpft werden kann.

Das Referat von Alexander Scheuch hat juristische Aspekte der modernen Datenexplosion zum Gegenstand. Wem gehört denn dieses Informationsuniversum? Sind Daten das gleiche wie Sachen und damit überhaupt erst eigentumsfähig? Wo endet das Recht auf Zugang zu, aber auch das Recht auf Verhinderung der Weitergabe von Daten aller Art? Seitdem das Bundesverfassungsgericht im Kielwasser der seinerzeitigen Volkszählungsdebatte das Recht auf informationelle Selbstbestimmung als Teil der Menschenwürde definiert, kommen auch noch ethische Aspekte hinzu. Diese juristische Perspektive wird ergänzt durch den Beitrag der europäischen Rechtsanwältin Luters-Thümmel, die uns durch den Dschungel der europäischen Initiativen, Konsultationen und Verordnungen führt.

Joachim von Braun beleuchtet die Bedeutung von Big Data für die Entwicklungspolitik und für das Schicksal der Armen dieser Erde. Man braucht keinen Hochleistungsrechner zu Hause auf dem Schreibtisch, ein billiges Smartphone reicht oft völlig aus, um etwa in den Genuss des modernen Mikro-Bankings zu kommen, das vor allem in Entwicklungsländern das Finanzwesen zum Guten revolutioniert. Und so wächst mit dem Spektrum der Anwendungen zugleich auch der Pool potentieller Nutzer mit vielfältigen gegenseitigen Verstärkungseffekten.

Kerstin Ludwig befasst sich mit den medizinischen und genetischen Aspekten der modernen Datenexplosion. Speziell bei der Genomanalyse haben sich inzwischen Möglichkeiten aufgetan, die nicht jeden Menschen glücklich machen. Oder wer würde sich über die Nachricht freuen, dass er oder sie im Dezember 2035 an Herzversagen stirbt? So genau sind die durch Gendiagnostik erzeugten Prognosen natürlich noch lange nicht, aber die Richtung ist aufgezeigt.

Joachim Schumann beleuchtet Big Data-indizierte neue Möglichkeiten in der Hydrologie. Wasser ist eine der kostbarsten natürlichen Ressourcen, aber leider hat man davon

zuweilen zu viel und ein anderes Mal zu wenig. So kamen etwa seit dem Jahr 2010 weltweit über 100.000 Menschen durch Flutkatastrophen zu Tode. Damit erschließt sich der Sinn besserer Prognosemethoden in der Hydrologie von selbst. Hier hat sich seit den berühmten Nilwasser-Abflussmessungen-Analysen von Harold Edwin Hurst in den 50er Jahren, die in der Statistik den Anstoß für die Theorie des langen Gedächtnisses gegeben haben, viel getan, mit aktuell geradezu explodierendem Datenvolumen, besonders durch Wettersatelliten und die Messung des Niederschlags und des Wasserabflussgeschehens aus dem Weltraum.

Der Ökonom Justus Haucap schließlich betrachtet das Big Data-Phänomen aus wettbewerbs- und ordnungspolitischer Perspektive. Dahinter steht die Befürchtung, dass nicht, wie zu Anfang formuliert, der Zug in Richtung alle wissen alles, sondern in die Richtung Wenige wissen alles laufen wird. Und diese Wenigen verfügen dann über beispiellose Möglichkeiten, dieses Wissen zu Geld zu machen. Siehe den gigantischen Marktwert von Unternehmen wie Google oder Facebook, der allein darauf gründet, dass Investoren diesen Unternehmen zutrauen, aus ihren Datenbeständen in Zukunft ebenso gigantische Gewinns abzuschöpfen. Damit hat das Sammeln und Verarbeiten von Massendaten auch bedeutende kartellrechtliche Aspekte. Wie lässt sich diese Spannung zwischen Datenschutz, Dateneigentum und Wettbewerb auflösen oder zumindest ausbalancieren? Traditionelle Maße für Marktmacht und Konzentration sind hier wenig hilfreich, hier steht die Wettbewerbs-Ökonomie vor großen Herausforderungen.

Und nicht nur die Wettbewerbsökonomie. Viele Wissenschaften sehen sich hier vor neuen, in den Lehrbüchern nicht vorkommenden Problemen, herkömmliche Denkmuster helfen nicht mehr weiter, ein neuer Ansatz ist gefragt. Auf dass das vorliegende Sammelwerk diese Suche nach neuen Ansätzen befruchten möge.

Dortmund, im Frühjahr 2018

Katharina Morik und Walter Krämer

DATEN – WEM GEHÖREN SIE, WER SPEICHERT SIE, WER DARF AUF SIE ZUGREIFEN?

Katharina Morik

1. Einführung

„Daten sind das neue Öl" sagte 2009 die liberale EU-Politikerin Meglena Kuneva[1]. Damit wollte sie die Bedeutung für die wirtschaftliche Entwicklung deutlich machen. Das World Economic Forum hat sich 2010 und 2011[2] mit dem Wert von Daten für die Betriebs- und Volkswirtschaft beschäftigt. Tatsächlich sind Daten von Öl sehr verschieden: ihr Vorkommen ist nicht begrenzt, sondern sie können erhoben werden; sie gehen nicht in den Produkten auf, die aus ihnen geschaffen werden; und sie können prinzipiell beliebig lange halten. Sie sind also kein Rohstoff, sondern eher ein Investitionsgut. Damit klingt bereits die Frage nach dem Besitz und der Verfügungsgewalt der Daten an. Gleichzeitig wird in der Öffentlichkeit der Aspekt des Besitzes bei Daten oft nicht wahrgenommen – alles ist kostenlos, alles wird geteilt! Es gibt immer mehr auf Daten beruhende Produkte und Dienstleistungen, die nicht mit Geld, sondern mit Daten oder gar nicht bezahlt werden (s. Brynjolfsson/McAfee/2014). Dies verschleiert, dass der Besitz an den und die Verfügungsgewalt über die Daten diskutiert und letztlich gesetzlich geregelt werden muss.

Mit diesem Aufsatz soll eine Einführung in die Diskussion aus Sicht der Informatik gegeben werden. Die Akademie der Wissenschaften und der Künste bietet die Möglichkeit zur interdisziplinären Diskussion: die Ingenieur- und

Wirtschaftswissenschaften betrifft das Thema ebenso wie die Jura. Die Klassensitzung am 13.6.2017 soll den Ausgangspunkt für die akademische Diskussion darstellen, an der sich alle Mitglieder beteiligen können.

Zunächst soll ein kurzer geschichtlicher Abriss deutlich machen, dass das Thema der Daten und der Datenanalyse keinesfalls neu ist. Damit ist die Grundlage gelegt für die verschiedenen Anwendungsgebiete. Einige werden vorgestellt, bevor am Ende .

Daten – geschichtlicher Abriss

Die Digitalisierung analog aufgenommener Signale ist früh für die Kommunikation untersucht worden im Hinblick auf das Abtasten und Rekonstruieren von Funktionen und die geeignete Kodierungen (Shannon/Weaver/1949) und wurde in der Elektrotechnik für den praktischen Einsatz vorangetrieben (Manoli/1988). Seit etwa 2015 wird der Begriff nun weiter gefasst, so dass er auch die Prozesse umfasst, die digitale Information verwenden. Insbesondere ist oft die Umgestaltung von Produktions- und Kommunikationsprozessen aufgrund des Vorhandenseins einer digitalen Repräsentation gemeint. Damit sind wir bei dem Thema der Daten und der durch sie ermöglichten Anwendungen angekommen.

Unter „Daten" verstehe ich hier die digitale Repräsentation von Information. Einerseits nehmen Daten Bezug auf ihre Gewinnung, andererseits auf ihre weitere Verarbeitung und Speicherung. Sinn und Bedeutung erhalten Daten erst durch eine Interpretation. Interpretierbar sind sie durch Programme aufgrund einer formalen Semantik. Dass auch Programme selbst Daten sind, führt manchmal zu Verwirrungen. Beschreibungen von Prozessen sind verschieden von den Prozessen selbst: Das Wort „Hund" bellt nicht. Wenn diese Beschreibungen jedoch maschinell interpretiert werden, kann es sein, dass sie den beschriebenen Prozess auch ausführen – der Roboterhund bellt. Wir müssen bei Daten also stets mit angeben, wie sie interpretiert werden. Dabei gibt es meist eine *Folge von Transformationen* mit je eigenen Interpretationen: für jede Anwendung bzw. jeden Dienst werden die Daten in eine geeignete Form überführt. Wir

haben also auf der Basis von Rohdaten eine Kette von Veredelungsprozessen mit ihren jeweiligen Daten bis hin zu verschiedenen Endprodukten. Gerade dies macht die Frage nach dem Besitz der Daten und ihren Nutzungsrechten kompliziert. Auf jeder Stufe sind andere Akteure an der Veredlung der Daten beteiligt und jede Nutzung auf einer Stufe beruht auf den Prozessen der vorangegangenen Stufen.

Datenbanken sind spätestens seit 1970 in relationaler Form mit der Anfragesprache SQL die typische Form, Daten zu speichern (Codd/1970). Insbesondere werden Kunden und die von ihnen gekauften Produkte oder mit ihnen geschlossenen Verträge gespeichert. Die Daten werden bei den Händlern, Energielieferanten, Versicherungs- und Telekommunikationsfirmen gespeichert. Früher wurden sie nur zur Abrechnung und zum Reporting genutzt. Durch den online-Handel sind die Datenmengen gestiegen und neue Dienste wie z.B. die Empfehlung von Produkten oder das gezielte Angebot neuer Vertragsvereinbarungen sind hinzugekommen. Diese neuen Dienste formen die für die Abrechnung erhobenen Daten für die Datenanalyse um und verknüpfen sie mit anderen Datenquellen. Die Akteure sind nun die Hersteller von Produkten oder Dienstleistungen (incl. Versicherungen), die Handelsplattformen wie z.B. amazon, die Informatikerinnen und Informatiker, die die Daten verknüpfen, analysieren und Dienste wie Empfehlungen oder personalisierte Werbung entwickeln und die Kunden.

Die statistischen Ämter der Bundesrepublik Deutschland bieten der Öffentlichkeit Informationen aus sorgfältig erhobenen Daten an. „Die Statistischen Ämter des Bundes und der Länder haben nach dem Bundesstatistikgesetz (BStatG) die Aufgabe, Daten zu erheben, aufzubereiten und zu veröffentlichen. ... Für die Auskunftgebenden besteht in der Regel Auskunftspflicht. Die amtliche Statistik garantiert die Geheimhaltung der erhobenen Einzelangaben.“[3] Die Daten bieten Transparenz für die Öffentlichkeit.

Open Data ist eine Zielsetzung, die Transparenz auch auf solche Daten bezieht, die nicht eigens für den öffentlichen Nutzen von staatlichen Institutionen erhoben wurden.

So gibt es beispielsweise in Europa offene Daten über soziale Fragen, Wissenschaften, Umwelt, Beschäftigung und Arbeit, Finanzen und Wirtschaft, Verkehr, Produktion, Bildung, Recht, und andere Themen[4]. Gewünscht ist die Öffentlichkeit staatlicher Daten, damit die Bürgerinnen und Bürger tatsächlich partizipieren und zu Akteuren werden können.

Das Internet stellt viele Daten in unterschiedliche Form zur Verfügung. Die einzige Bedingung für Inhalte, um im Internet angezeigt zu werden, ist die Befolgung des Protokolls, eine eindeutige Bezeichnung (URI: Uniform Resource Identifier) und die Darstellung in einem Browser-lesbaren Format, d.h. der Browser ist der Interpreter der Repräsentation und befolgt die eventuell angegebenen Deklarationen zur Semantik der Daten. Erster Schritt, um passende Informationen in den vielen Daten zu finden, ist das Indexieren der Webseiten. Es gibt eine Web-Seite, die die aktuelle Größe des indexierten Internets abschätzt[5], am 6.10.2017 waren von Google mindestens 46,2 Milliarden Seiten indexiert. Unterschiedliche Algorithmen für die Suche und das Sammeln (Crawling) bleiben Forschungsthemen. Die internationale Tagung WWW ruft auch für 2018 Beiträge zu Web crawling, Indexing und Suche auf.

Weiter geht die Idee des semantischen Internets, bei der die Autoren einer Seite ihre Inhalte und Präsentationsformate deklarieren sollen (Berners-Lee et al. 2001). So strebt das semantische Internet eine Taxonomie an, in die sich Inhalte kategorisieren lassen. Dies würde die Suche nach Information und Dienste anhand von Informationen erleichtern. Das Internet Consortium W3 stellt die Formalismen zur Deklaration und Formatierung zusammen. Browser lesen die Deklaration und interpretieren die Daten entsprechend. Die Annotationssprache XML ist für viele Daten und ihre Beschreibung inzwischen der Standard. Auch Programme werden oft mit ihren Parametern durch XML beschrieben. So können ganze Prozesse der Datenanalyse mit den Framework *streams* oder dem Werkzeug RapidMiner in XML beschrieben, die kurzen Beschreibungen versendet und in den Systemen dann zum Laufen gebracht werden.

Ohne verbindliche Taxonomien zu fordern, folgt die Idee der *Linked Open Data* dem Motto „everything is a link“. Der Name (URI) soll als Bezeichnung für Dinge und Personen und Sachverhalte genutzt werden. Die betreffende Web-Seite soll die üblichen Standards einhalten (RDF: Resource Description Format), so dass leicht von überall her darauf zugegriffen werden kann[6]. Hier wird aber keine vorgegebene Taxonomie angenommen oder vom Autor eine Annotation gefordert. Die Aufbereitung von Informationen aus unterschiedlichen Quellen erfordert dann mehr Arbeit von denen, die die Daten nutzen wollen. So hat beispielsweise das europäische Projekt vista-tv Information über Fernsehsendungen und die online Einschaltungen in Programmen des Internet-Fernsehens als Linked Open Data zugänglich gemacht[7]. Für viele Dienstleistungen oder Analysen ist eine solche Aufbereitung notwendige Voraussetzung. Die Wahrung der Privatheit der Internet-Fernsehbetrachter war dabei Aufgabe der Projektpartner BBC und Zattoo und wurde durch die Privatheit erhaltende Datenanalyse unterstützt.

Die Vernetzung von Geräten, insbesondere Smartphones, über das Internet hat eine weitere Zunahme des World Wide Webs zur Folge gehabt. Das Internet of Things (IoT) vernetzt Geräte, die selbst Daten sammeln. Einer Schätzung zufolge werden 2020 etwa 25 Milliarden Geräte mit dem Internet verbunden sein[8]. Die Frage, welche Daten bei dem lokalen Gerät bleiben und welche an eine Zentrale gesendet werden, ist entscheidend für die Wahrung von Privatheit.

Die App als gekapselte Dienstleistung mit Zugriff auf zugehörige Daten leitete zudem eine neue Sicht auf die Schnittstellen zwischen Benutzer, Datenbereitstellung, Telekommunikationsgesellschaft und Dienstleistungsanbieter ein: statt einer Schnittstelle zu Daten für viele unterschiedliche Dienste, ist nun eine Dienstleistung eine Entität, die separat, nur vom Betriebssystem abhängig, selbständig den Zugriff auf Daten gestaltet. Hier ist oft dem Benutzer nicht deutlich, welche Daten von seinem Smartphone an eine Zentrale übermittelt werden. Und welche lokal bleiben.

Gerade die Nutzung des Internets wurde unter dem Stichwort der „digital divide“ diskutiert: eine eher reiche,

mächtige, gebildete, überwiegend männliche Elite nutzt und gestaltet das Web (Muki Haklay 2012).

Datenanalyse – geschichtlicher Abriss
Die Analyse von Daten begann schon im 17. Jahrhundert für den Bevölkerungscensus mit Sterbetafeln und Wahrscheinlichkeiten des Überlebens. Die Statistik bleibt Grundlage der Datenanalyse. Allerdings sind Daten in der Statistik üblicherweise sorgfältig erhoben und haben eher wenige Beobachtungen mit wenigen Merkmalen.

Das maschinelle Lernen begann als Logik-basierte Analyse von Daten. Während überwiegend aussagenlogische Formalismen von Tripeln aus Entität, Merkmal und Wert genutzt wurden, gab es auch schon frühzeitig Formalismen der eingeschränkten Prädikatenlogik (Emde/etal/1983). Die induktive logische Programmierung ließ sich direkt auf Datenbanken aufsetzen (Morik/Brockhausen/1997) und konnte direkt ausführbare Regeln aus Daten lernen (Morik/etal/1993). Die Betrachtung von Relationen, insbesondere der Unabhängigkeit, führte später zu probabilistischen graphischen Modellen, die als Kombination von Logik und Statistik betrachtet werden könnten (Kersting 2006; Piatkowski/etal/2016). Gleichzeitig wurden frühzeitig Perzeptrons und dann neuronale Netze entwickelt, die zum Kanon der Lernverfahren gehören (Mitchell 1997).

Die Analyse sehr großer, nicht für die Analyse erstellter Datenbanken wurde als *Knowledge Discovery in Data* bzw. *Data Mining* bezeichnet. Die Analyse wurde oft auf das Finden korrelierter Merkmale reduziert, die häufigen Mengen. Algorithmisch wurde die Skalierbarkeit auf sehr große Datenmengen mit mehreren Verfahren erreicht[9].

Dass die neuronalen Netze heute unter dem Namen *Deep Learning* weltweit große Aufmerksamkeit erregen, mag an mehreren Gründen liegen: es gibt genügend Daten, genügend Rechenkapazität (auch Spezialprozessoren wie z.B. TPU) und genügend Werkzeuge (z.B. TensorFlow) für eine einfache Anwendung. Es handelt sich bei den *Convolution Neural Networks* und *Recurrent Neural Networks* um Klassen von Lernverfahren, deren Struktur und Parameter

sorgfältig gewählt werden müssen. Es ist keineswegs so, dass der Anwender nichts zu tun hat. Die meisten Erfolge wurden mit Bilderkennung erzielt, weil dort viele korrelierte Merkmale vorhanden sind. Es gibt aber auch Grenzen. Zum Beispiel wurden durch leichte Veränderungen Bilder nicht mehr richtig erkannt (Moosafi-Dezfooli/etal/2017).

Problematisch ist vor allem, dass es keine Theorie gibt, die die Lernbarkeit durch Deep Learning oder das Risiko der Überanpassung an die Daten, die zum Lernen verwendet wurden angibt. Auch so sind gelernte Modelle nicht verständlich, nur selten vom Menschen interpretierbar.

2. Anwendungsgebiete

Datengestützte Wissenschaft

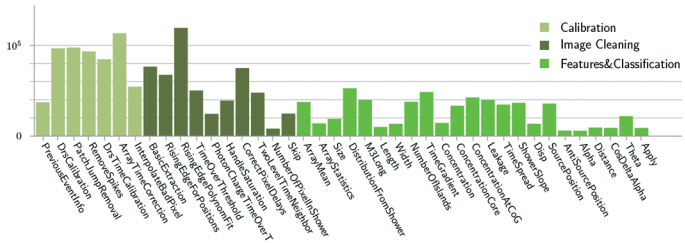
Prinzipiell können alle Wissenschaften von der Datenanalyse profitieren. Insbesondere wird aber gerade dort die Analyse genutzt, wo die Daten auf einer Ebene gemessen werden, die nicht direkt interpretierbar ist. Ein einfaches Beispiel soll diese Ebenen deutlich machen: die Handlung eines Menschen, z.B. essen, besteht auf der Ebene darunter in mehreren Teilhandlungen (den Löffel eintauchen, mit der Suppe darin zum Mund führen, den Mund öffnen, den Löffel im Mund leeren), die jeweils wieder auf der Ebene darunter aus kleineren Teilen bestehen. Werden nun die Nervenströme oder Muskelbewegungen gemessen oder das Essen gefilmt, so muss aus diesen Daten der Begriff „essen" gelernt werden. In einer Erzählung wäre der Begriff direkt vorhanden und ohne Bezug auf seine Realisierung gegeben. Die Datenanalyse ist also gefragt, wenn aus Daten auf einer Ebene Aussagen auf einer höheren Ebene erschlossen werden, kurz: ein Modell aus Daten erworben wird. Für die Überprüfbarkeit der Modelle ist dabei wichtig, dass die Analyseverfahren selbst eine theoretische Begründung haben. Wir wollen für die Empirie die „first principles" nicht aufgeben!

Lebenswissenschaften

Die Lebenswissenschaften verfügen über hochdimensionale Daten, die oft auch eine Graph- bzw. Netzstruktur aufweisen. Die Entschlüsselung des menschlichen Genoms hat viele Fragen eingeführt, die man nun mit Hilfe der Daten beantworten will, beispielsweise Verwandschaftsbeziehungen zwischen Gattungen und Wanderbewegungen der Urmenschen. Hier wirken die Lebenswissenschaften auf andere Disziplinen wie Anthropologie und Geschichte zurück. Die Analyse genetischer Daten ist natürlich insbesondere für die Medizin und die Pharmazie bedeutsam. Einerseits erhofft man sich ein besseres Verständnis der Prozesse in der Zelle (Masip/etal/2016), andererseits will man Heilungschancen vorhersagen (Lee/etal/2014) oder die Wirksamkeit von Substanzen untersuchen (Nicola/etal/2015). Auch hier ist die Verfügbarkeit in den Ländern sehr unterschiedlich geregelt. So gibt es die öffentlich verfügbaren US-amerikanischen Daten des *The Cancer Genome Atlas* mit Daten über Tumor- und normales Gewebe von mehr als 11 000 Patienten. Europäische Daten hingegen sind nur für einzelne Studien nach einer besonderen Prüfung durch den Ethikrat erhältlich.

Physik

Die Physik nimmt immer aufwändigere Messungen vor, um immer weitreichendere Fragen zu beantworten. Das CERN mit seinen Experimenten zur Teilchenphysik (LHCb Collaboration/2015) und die Experimente der Astroteilchenphysik IceCube (IceCube Collaboration(2014) liefern mehr Daten als sich Menschen auch nur ansehen können. Die Datenerhebung durch Cherenkov Teleskope wird von einem Ring um die Erde einzelner Teleskope (MAGIC, HESS, VERITAS) erweitert zu zwei Arealen von insgesamt mehr als 100 Teleskopen, um die Luftschauer noch genauer nach Gamma-Strahlen zu durchsuchen, die Aufschluss über Prozesse im Weltall geben können. Von den Rohdaten bis hin zu klassifizierbaren Daten ist es ein weiter Weg, wobei die Analyse nicht erst bei dem letzten Schritt ansetzt. Auf jeder Ebene sind Anwendungen von Lernverfahren möglich und liefern



die Daten für die nächste Ebene. Tatsächlich nehmen Ausreißererkennung, Zeitreihen-bereinigungen, Merkmalsauswahl und Aggregationen, die dem Lernen eines Klassifikators, der die Gammastrahlen von Hadronen trennt, vorausgehen, den größten Teil der Analyse ein. Eine Fülle von Programmen, die direkt auf den Datenströmen arbeiten, wurde in dem Framework *streams* entwickelt (Bockerman/etal/2015). Das Schaubild zeigt die Prozesse, die bei der Verarbeitung der einströmenden Daten vorgenommen werden müssen. Der hintere Teil enthält maschinelles Lernen im weitesten Sinne, nur der letzte Schritt ist die Anwendung einer gelernten Klassifikation, also maschinelles Lernen im engeren Sinne.

Diese Forschung ist ohne Daten, deren Speicherung und Analyse gar nicht zu bewerkstelligen. IceCube hütet die Originaldaten an der Madison University Wisconsin, USA, und wendet Algorithmen auf sie erst an, wenn der Analyseansatz an Simulationsdaten getestet und die Kollaboration (mehr als 200 WissenschaftlerInnen) den Ansatz geprüft haben. Die TU Dortmund betreibt gemeinsam mit anderen deutschen Forschungsinstitutionen ein eigenes Teleskop auf La Palma, FACT, um direkt auf den ersten Rohdaten arbeiten zu können. Es ist durchaus auch eine Sicherung von Einfluss auf die Wissenschaft, wer die Daten speichert und den Zugang dazu ermöglicht.

Sprachwissenschaft

Die Sprachwissenschaft hatte früh einen Bezug zur Datenanalyse. Die Sprachstatistik mit dem Gesetz von Zipf (Zipf/1949) brachte die Verteilung der Wörter in den wissenschaftlichen Diskurs ein, die Untersuchung der Lernbarkeit von Grammatik (Gold/2004; Angluin/Becerra-Bonache/ 2017) fügte Methoden der theoretischen Informatik hinzu. Die *Message Understanding Conference* (MUC) hatte das Ziel, automatisch aus Texten Inhalte zu extrahieren und brachte eine Reihe von Fragestellungen hervor: wie erkennt man Eigennamen von Dingen, Orten, Personen; wie erkennt man Relationen zwischen Eigennamen; welche Merkmale sind wichtig, welche Strukturen? Das Projekt *Never ending learning of language* extrahiert Fakten und Regeln aus Texten aus dem Internet, um so immer besser das Web „lesen" zu können (Yang/Mitchell/2016). Andere Fragestellungen betreffen die automatische inhaltliche Kategorisierung von Texten wovon handelt ein Text? (Joachims/2001). Die Untersuchung der in Texten zum Ausdruck kommenden Einstellungen (sentiment analysis) wurde meist verkürzt auf die positive oder negative Beurteilung von Produkten im Internet.

Für die deutsche Sprache gibt es eine hervorragende Datensammlung für die Verwendung einzelner Wörter in vielen Texten aus verschiedenen Zeiten: Digitales Wörterbuch der Deutschen Sprache. Hier sind auch syntaktische und etymologische Annotationen gegeben. Untersuchungen auf den Textbelegen zu einem Wort, z.B. Platte oder Leiter, können den Wandel der Sprache über der Zeit belegen. Durch die sorgfältig kuratierten Daten lassen sich linguistische Erklärungen mit Daten-gestützter Analyse leicht verbinden (Bartz/etal/2014).

Durch die sehr großen Mengen an Text, die Google zur Verfügung stehen, kann auf Regeln verzichtet und statt dessen mit einfacher Analyse Korrelationen festgestellt werden. Deshalb die Aussage von Peter Norvig „All models are wrong and increasingly you can succeed without them"[10]. Die Korrelation von Sätzen verschiedener Sprachen ergibt eine brauchbare maschinelle Übersetzung, wenn es genügend Daten in beiden Sprachen gibt. Man kann mit Topic