Gefördert durch:

Bundesministerium
für Wirtschaft
und Technologie

aufgrund eines Beschlusses
des Deutschen Bundestages

DATENSOUVERÄNITÄT: FORTSCHRITT UND VERANTWORTUNG (Preprint)

Anne Schwerk[1], Jack Thoms[1], Tilmann Rabl[1,2], Volker Markl[1,2]

[1] DFKI GmbH
[2] Technische Universität Berlin

Dieser Artikel ist entstanden mit Unterstützung des Bundesministeriums für Wirtschaft und Energie (BMWi). Die Beiträge und Meinungen müssen nicht notwendigerweise mit der des BMWi und mit der Meinung der Bundesregierung übereinstimmen.

DATENSOUVERÄNITÄT: FORTSCHRITT UND VERANTWORTUNG

Daten bilden das Fundament der wirtschaftlichen Entwicklung; nicht nur, da sie uns eine sehr präzise quantitative Evaluierung eröffnen, sondern da sie uns auch die Möglichkeit bieten, Prozesse zu optimieren, zu automatisieren und durch datengetriebene Rückkopplung zu steuern. Dies macht Daten zu einem omnipräsenten Gegenstand aller unserer Lebensbereiche in wirtschaftlicher, wissenschaftlicher und gesellschaftlicher Hinsicht: Sensoren verbessern heute nicht nur unsere Autos und Häuser, sie ermöglichen auch informationsbasierte Medizin mit verbesserten Diagnosen und Therapien. Die Analyse und das maschinelle Lernen von Daten aus Medien und Sensoren in Echtzeit ermöglicht effizientere Steuerung von Maschinen, Optimierung von Verkehrsflüssen, oder die Erkennung und Prognose von gesellschaftlichen Trends. Cisco prognostiziert, dass die Datenerzeugung und -kommunikation im Internet im Jahr 2021 3,3 ZettaByte pro Jahr betragen wird[1]. Die Datenbestände sind proportional so enorm, dass 90% der heutigen Daten innerhalb der letzten zwei Jahre entstanden sind[2]. Diese Datendominanz wird sogar als die vierte wissenschaftliche Revolution angesehen, das sogenannte vierte Paradigma, in dem datenintensive Forschung als Motor aller Wissenschaften sowie gesellschaftlicher und wirtschaftlicher Entwicklung fungiert[3].

Die erfassten Daten werden nicht nur größer, sondern auch immer heterogener und vollständiger[4], wodurch Dienste und Produkte im Sinne einer Massenindividualisierung stetig personalisiert und verbessert werden. Dadurch entsteht jedoch auch das Potential für Diskriminierung, Nivellierung von Ungleichem, mangelnde Transparenz von Gründen für Entscheidungen und Handlungen, mangelnde Neutralität oder mangelnder Zugang zu Daten sowie Einschränkungen der informationellen Selbstbestimmung. Somit stellt sich die Frage, wie wir die Entwicklung von digitalen Diensten und Big Data fördern können und gleichzeitig den Herausforderungen begegnen können.

DAS ZEITALTER DER DATENÖKONOMIE UND DIE DIGITALE SOUVERÄNITÄT

Big Data sind in der wachsenden Datenökonomie zu einem eigenen Rohstoff geworden, den es gilt, zu erzeugen, zu veredeln und grenzüberschreitend zu tauschen. Oftmals wird Big Data mit einem Rohstoff wie Öl verglichen, was jedoch in der Praxis ein schlechter Vergleich ist, da Öl im Gegensatz zu Daten nach der Verwendung verbraucht ist. Besserer wäre der Vergleich mit einem Produktionsfaktor wie Boden, aus dem durch Anbau und Veredelung ständig Neues entsteht, ohne dass der Boden verbraucht wird. Das Tauschen von großen Datensätzen erfordert hierbei geeignete digitale Infrastrukturen und rechtliche Rahmenbedingungen im Hinblick auf Eigentum und Besitz am Produktionsfaktor Big Data. Die Kapitalisierung von Big Data in einer digitalisierten Volkswirtschaft und Gesellschaft erfordert nicht nur moderne Datenübertragungsnetze, sondern vielmehr eine komplette Informationsinfrastruktur, in der Daten gespeichert und rechtskonform analysiert werden können. Der derzeitige Fokus auf die Bereitstellung von Datenübertragungs- oder Berechnungs- und Speicherinfrastrukturen alleine ist nicht ausreichend. Nur durch eine Infrastruktur, die den Produktionsfaktor Big Data bereitstellt und Analysen und maschinelles Lernen darauf ermöglicht, wird die Digitalisierung eine der wirkungsvollsten Mittel zur Erhöhung der deutschen Wettbewerbsfähigkeit.

> »Digitale Souveränität« bezeichnet in diesem Sinne die Fähigkeit zu selbstbestimmtem Handeln und Entscheiden im digitalen Raum. Digital souveräne Systeme verfügen bei digitalen Schlüsseltechnologien und -kompetenzen, entsprechenden Diensten und Plattformen über eigene Fähigkeiten auf internationalem Spitzenniveau. Sie sind darüber hinaus in der Lage, selbstbestimmt und selbstbewusst zwischen Alternativen leistungsfähiger und vertrauenswürdiger Partner zu entscheiden, sie bewusst und verantwortungsvoll einzusetzen und sie im Bedarfsfall weiterzuentwickeln und zu veredeln (Bitkom)[5]

Bislang sind es vor allem US-amerikanische und zunehmend auch asiatische Unternehmen, die wichtige Dateninfrastrukturen für die digitale Wirtschaft bereitstellen. Die europäische Fragmentierung des Marktes und der Gesetzgebungen sind essentielle Barrieren für europaweite Entwicklungen von gemeinsamen Infrastrukturen und einschlägigen Kooperationen. Die Marktführer kommen hauptsächlich aus homogenen großen Märkten, in denen die Gesetze, Kultur und Sprache aufeinander abgestimmt sind. Darum ist es von großer Wichtigkeit, den europäischen Markt einheitlicher zu gestalten und einen souveränen und sicheren Datenaustausch durch gemeinsame digitale und rechtliche Rahmenbedingungen zu etablieren. Dies würde auch eine Schnittstelle für einen produktiven internationalen Austausch schaffen, Investitionen und Innovationen erleichtern und Europa zentral und wettbewerbsfähig positionieren.

GRENZEN DER GESELLSCHAFTLICHEN DATENSOUVERÄNITÄT

Der Anstieg der Digitalisierung erfordert den Zugang und die Verarbeitung von Daten, je größer die Datensätze und Möglichkeiten der Analyse, desto präziser

Preprint

werden die Aus- und Vorhersagen und somit der Wert für die Gesellschaft. Ein freier Datenzugang würde die digitale Wirtschaft zwar sehr fördern, allerdings kollidiert die freie Nutzung mit der souveränen Datenverwaltung der Einzelnen. Datensouveränität setzt voraus, dass sich Einzelne im Zeitalter der Big Data der Risiken und Potenziale ihrer Datenschätze bewusst sind und diese angemessen und unter Wahrung der informationellen Selbstbestimmung in unserer vernetzten Welt einsetzen können[6].

Diese Prämisse wird kontrovers diskutiert, da viele Datenanalysetechniken bisher keine genügende Transparenz bieten, deren genaue Verwendungen einzusehen und unter optimalen Privatrechts-Wahrungen zu handhaben und sie deshalb einen souveränen Datenumgang hemmen. Beispielsweise kann eine verpflichtende Daten-Pseudonymisierung nur bedingt Schutz bieten, da gerade bei unikalen und longitudinalen Datensets konkrete Rückschlüsse auf die untersuchten Personen möglich sind[7,8]. Ähnlich enthalten einige Massendatenanalysen, wie Genomsequenzierungen, tiefgründige Informationen über das Individuum und darüber hinaus auch über verwandte Individuen, die nicht direkt sequenziert worden sind. Dieses Privatrechts-Interdependenz Dilemma bildet ein komplexes Problem in der Einhaltung des Privatrechtes. Auch das Nutzen von Datenspuren durch Dritte (sog. Third-Party Services) ist unter dem Territorialprinzip nur schwer kontrollierbar[9], da unbemerkt Daten über verschiedene Geräte eines Individuums gesammelt werden können und der Großteil dieser App-verfolgten Daten in Drittländer wie die USA exportiert werden, wo diese von nur einigen wenigen Anbietern verwaltet werden und die Nutzung nicht mehr dem deutschen oder europäischen Rechtsrahmen unterliegt[10]. Diese Monopolisierung sowie der unfreie Datenumgang ist nur schwer zu reglementieren, da auch unter der europaweiten *General Data Protection Regulation* (GDPR) nicht deutlich ist, unter welche Gesetze solche transnationalen Drittleistungen fallen.

Ein weiteres Dilemma in der transparenten Datenhandhabung ist, dass das Recht auf Dateneinsicht (right to access) in der Praxis nur schwer zu implementieren ist; sowohl eine potentielle Einsicht in die elektronische Patientenakte[11], als auch eine simple Einsicht in die sozialen Mediendaten[12] ist oft technisch noch nicht umsetzbar. Die GDPR hat zwar die Kosten für solche Dateneinsichten entfernt, allerdings können Institutionen, wie Facebook, eine Datenbereitstellung verweigern, wenn der Aufwand hierfür unverhältnismäßig groß ist[12]. Solche Handhabungen sind zwar nicht-rechtskonform, erfordern aber einen enormen zeitlichen Aufwand und Durchsetzungsvermögen seitens der DatenerfragerInnen. Ein souveräner Datenumgang wird somit nicht direkt durch die GDPR erzwungen und beinhaltet bisweilen nur minimale gesellschaftliche Hilfeleistungen und Strafen für nichtkonforme Handhabungen.

Eine mögliche Stärkung Deutschlands kann folglich nur durch eine Öffnung der deutschen Rechtsbedingungen oder einer Erweiterung der Gesetze auf transeuropäische Unternehmen geschaffen werden. Gleichzeitig müsste der Transfer und die Speicherung und Verarbeitung von in Deutschland entstandenen Daten außerhalb des deutschen oder europäischen Rechtsrahmens eingeschränkt bzw. unter Androhung empfindlicher Strafen bis hin zur Abschaltung rechtsverletzender Dienste verboten werden. Ähnliche Ansätze werden bereits in China angewandt, wo fremde Unternehmen personenbezogene Daten, inkl. Datenspuren, nur innerhalb Chinas bewahren dürfen und ihre Dienste nur nach tiefgründigen Reviews anbieten können und ggf. auch blockiert werden[13]. Vietnam und Singapur haben diese Datensicherheitsbestimmungen bereits teilweise übernommen. Nur durch ähnliche Schritte lässt sich die deutsche oder europäische Datensouveränität zurückgewinnen. Ansonsten werden weiterhin mit deutschen Daten multinationale, insbesondere US-amerikanische Konzerne ihre Geschäftsprozesse optimieren und einen uneinholbaren Vorsprung gegenüber deutschen oder europäischen Wettbewerben haben. Ferner werden auch deutsche oder europäische Konzerne ihre datengetriebene Forschung vermehrt im Ausland durchführen, wo diese durch einfachere Regelungen zum Datenschutz umsetzbar sind.

Ohne einschlägige Gesetzesänderungen wird Deutschland weiterhin die derzeitige US-amerikanische Vormachtstellung in der Nutzung der deutschen Daten fördern; nationale und europäische Bestrebungen in der Algorithmen- und Systemforschung würden beeinträchtigt werden bzw. eher die Wertschöpfung in den (außereuropäischen) Ländern befördern, in denen der Produktionsfaktor Big Data vorhanden ist. Neben den legislativen Anforderungen müssten BürgerInnen durch DatentreuhänderInnen und zielgruppenspezifische Bildungsmaßnahmen für die zunehmende digitale Datenkomplexität gestärkt werden, um die Chancen der Datenbereitstellung mit den Risiken der Datennutzung einschätzen zu können.

GRENZEN DER IT-DATENSOUVERÄNITÄT

IT-WissenschaftlerInnen tragen oft die komplexe Verantwortung der Anwendungen und Konsequenzen von Massendatenanalysen. Dies ist eine große und interdisziplinäre Anforderung, die nicht nur WissenschaftlerInnen beantworten sollten, sondern auch

Preprint

AnwenderInnen und die Gesellschaft. Außerdem gilt es generell zwischen den verschiedenen IT-Wissenschaften zu unterscheiden, denn die System-Forschung fokussiert sich primär auf Datenverarbeitungsstrukturen, wohingegen die Algorithmen-Forschung und die Datenwissenschaften die Entwicklungen und Anwendung der Datenanalyse erstreben. Darum können auch nur Letztere konkrete Anwendungen und Manipulationsmöglichkeiten der Datenwissenschaft durch die Reinheit der Testdaten und Verifikation der Algorithmen Validität verantworten. Eine besondere Wichtigkeit kommt hier der Qualität der Rohdaten zu. Falls diese bereits Verzerrungen beinhalten (Selektions-, Beobachterverzerrung etc.), sind fehlerhafte, unfaire oder diskriminierende Entscheidungen durch Algorithmen möglich. Daher ist primär nicht der Algorithmus das Problem einer fehlerhaften Analyse, sondern meist sind hierfür fehlerhafte Trainingsdaten verantwortlich. Für valide Forschungsergebnisse ist es daher unabdingbar, viele und heterogene Primärdaten DatenwissenschaftlerInnen zur Verfügung zu stellen.

Weiter müssen Algorithmen validiert und verstanden werden, eventuell durch Visualisierung oder großflächige Anwendungstests, besonders in Bezug auf die externe Validität (Validität über verschiedene Datensets). Diese Validierungen und Tests sollte initial durch die EntwicklerInnen ausgeführt werden, liegen aber im Kontext der Anwendung unter Verantwortung der AnwenderInnen. Denn nur diese können später mit Realdaten die Anwendung validieren und auch Modelltests betreffend möglicher Diskriminierungen und anderer ethischer Konsequenzen ausführen. Ein weiteres inhärentes Problem der Massendatenanalysen sind Scheinkorrelationen, die durch verdeckte Variablen entstehen können und letztendlich zu irrtümlichen Aussagen führen. Hier liegt die Verantwortung bei den WissenschaftlernInnen, die die erzeugten Modelle nicht nur intern und extern validieren müssen, sondern auch theoretisch untermauern sollten, da durch rein Datengetriebene Modelle oft zu viele und irrelevante Variablen berücksichtigt werden und folglich falsche Schlussfolgerungen entstehen können.

Es ist deutlich, dass sich die Verantwortung der datengetriebenen Produkte über viele verschiedene Disziplinen erstreckt und nicht mit simplen Regelungen auskommen kann. Da Gesetze außerdem nur einen limitierten Schutz bei grenzüberschreitenden und komplexen Fragestellungen bieten können und darüber hinaus nur langsam implementiert werden, sollten alternativ ethische Richtlinien für eine *Good Big Data Practice* der Forschung und Wirtschaft erarbeitet und durchgesetzt werden[14], v.a. für die Arbeit mit personenbezogenen Daten. Diese müssten außerdem durch technische Richtlinien für eine *Good Big Data Practice,* wie beispielsweise physisch getrennte Datenrepositorien, *Privacy/Security-by-Design* inkl. Ende-zu-Ende Verschlüsselung unterstützt werden. Folglich ist es von nationaler Wichtigkeit, dass Europa seine eigenen Infrastrukturen und Dienste bereitstellt und den BürgernInnen somit alternative Datenmarktplätze bietet, die diesen Richtlinien folgen. Dies setzt allerdings voraus, dass eine datenintensive Forschung unterstützt wird und nicht durch einschlägige Gesetze und digitale Hemmnisse erschwert wird. Gleichzeitig muss jenseits der Projektförderung eine Dateninfrastruktur, d.h. ein Informationsmarktplatz bzw. eine Informationskommune für Deutschland und Europa aufgebaut werden, welche sich nicht nur durch öffentliche Förderung von Netzen und Rechenkapazität auszeichnet, sondern durch eine beständige Kuratierung, Datenspeicherung und Bereitstellung von öffentlichen Analysediensten.

BIG DATA UND DAS DIGITALE BEWUSSTSEIN IN DEUTSCHLAND

Die Industrienation Deutschland zeigt, trotz wirtschaftlicher Stärke, eine nur mittelmäßige Digitalisierung und trägt somit nicht aktiv zu Europas digitaler Vormachtstellung bei. In dem jährlich von der Europäischen Kommission (EC) erscheinenden Bericht über den Stand der Digitalisierung in Europa (*Digital Economy and Society Index*) belegt Deutschland nur den 11. von 28 Plätzen und überlässt Finnland und Schweden die Spitzenpositionen[15]. Erschreckend ist, dass Deutschland zusammen mit Portugal und Lettland die geringste digitale Verbesserung zum Vorjahr aufzeigt. An den digitalen Kompetenzen fehlt es nicht (Rang 7), allerdings bedroht der deutsche Fachkräftemangel im Informations- und Kommunikationstechnologie (IKT)-Bereich das Potenzial der Wirtschaft. Die Agentur für Arbeit meldet hierzu, dass es bereits seit Jahren an SoftwareentwicklernInnen und IT-AnwenderberaternInnen mangelt[16]. Bitkom bestätigt dies mit einem nur schwachen Wachstum der IT-Studien anfängerInnen und 51.000 offenen IT-Expertenstellen in 2016[17]. Dieses Loch ist verheerend, da das IKT-Feld als primärer Innovationsmotor der deutschen Forschung gesehen wird.

Ein zukunftsweisendes Konzept müsste bereits in der Schule die Kinder digitaler aufstellen und diese gezielt fördern. Das Gegenteil zeigt sich in Deutschland: Sogar Schulkinder zeigen bereits digitale Defizite auf und werden unzureichend motiviert, Medien wie z.B. Wikis oder Lern-Apps anzuwenden[18]. Nur jeder zehnte Lehrer nutzt digitale Anwendungen und jeder fünfte Lehrer gibt an, keinen Zugang zu WLAN in der Schule zu haben. Geschätzte

Preprint

2,8 Milliarden Euro pro Jahr müssten in Deutschlands Schulen investiert werden, um diese global kompetitiv aufzustellen[19]. Darüber hinaus ist Deutschlands Glasfasernetz eines der schlechtesten Europas[20] und es mangelt an vielen Schlüsseltechnologien, wie Datenanalyse, Mikroelektronik, Sensorik, Aktorik und *Embedded Software* – die durch den Ausbau von international ausgerichteten Forschungs- und Entwicklungsschwerpunkten stimuliert werden könnnten[21]. Gleichzeitig muss durch Abbau von Hemmnissen sichergestellt werden, dass kein *Brain-Drain* in andere Regionen stattfindet: derzeit wandern immer noch zu viele Experten ins Ausland ab, wo durch verbesserten Zugang zu Big Data und Analyseinfrastrukturen bessere Arbeitsbedingungen mit interessanteren Problemstellungen existieren.

DEUTSCHLANDS DATENSCHUTZ-KLIMA HEMMT DIE DIGITALE ENTWICKLUNG

Neben der unzureichenden Digitalisierung weist die deutsche Gesetzgebung eine besonders hohe Rechtssicherheit auf, auch im europäischen Vergleich. Der Datenschutz ist in Deutschland selbst ein Grundrecht, wohingegen in einigen europäischen Ländern das Recht auf informationelle Selbstbestimmung anderen Gesetzen, wie dem Öffentlichkeitsprinzip unterlegen ist und in den USA selbst gar nicht gesetzlich verankert ist. Besonders im Gesundheitsbereich, wo viele personenbezogene Daten erhoben werden, bildet der Datenschutz darum eine der größten Barrieren nationaler Forschungsumsetzungen, da die Forderungen des Datenschutzes, die Privatheit Einzelner zu bewahren, mit dem Recht auf Informationsfreiheit und öffentlicher Transparenz kollidieren. Die deutsche Gesetzgebung schützt NutzerInnen nicht nur durch das Grundrecht der informationellen Selbstbestimmung, sondern auch durch das Recht auf Gewährleistung der Integrität und Vertraulichkeit informationstechnischer Systeme (IT-Grundrecht) und durch das Telekommunikationsgeheimnis. Darüber hinaus verbietet der deutsche Urheber- und Datenbankschutz jegliches *Text und Data Mining* (TDM), inklusive der reinen Informationsextraktion, so lange keine direkte Einwilligung der Eigentümer vorliegt[22]. Dieses TDM Verbot wird auch unter der GDPR beibehalten; nur England hat eine Anpassung für nicht-kommerzielle Zwecke erreicht. Solch ein scharfer europäischer Datenschutz blockiert nicht nur die Forschung, sondern festigt auch hier wieder die US-amerikanischen Monopole, da die amerikanische *Copyright* Gesetzgebung keine TDM-Beschränkungen kennt. Hier sollte die deutsche Verfassung ein generelles Verbot von Urheberrecht oder Nutzungsbeschränkungen öffentlicher Datenbestände implementieren und das Urheberrecht Dritter durch spezifische Einräumungen der Nutzungsrechte regeln.

Der deutsche Datenschutz spiegelt allerdings ein generelles Verhalten der Nation wieder, nämlich ein großes Bedürfnis, Unsicherheiten und Veränderungen zu vermeiden, also das Gegenteil der US-amerikanischen *risk-taking* Mentalität. Diese Unsicherheitsvermeidung ist auch der Grund warum viele Startup-Ideen zwar in Deutschland entwickelt werden, aber letztendlich in Amerika mit ihren Produkten den Markt erstürmen, wo sie durch eine aktive Wagniskapitalbranche und für Risiken aufgeschlossene Unternehmern eine stimulierende Umgebung finden[23]. Was bleibt, sind etablierte und Franchise-basierte Unternehmen, das Gegenteil von Innovation und *Disruption*.

Diese gesellschaftlichen Barrieren können nur durch politische Rahmenbedingungen verändert werden, die die Deutsche Wagniskapitalbranche fördern und geeignete Strukturen für einen souveränen Datentausch und potentielle Innovationen schaffen. Dies gilt es auf rechtlicher sowie auf steuerlicher und gesellschaftlicher Ebene zu realisieren, damit auch Startups und v.a. kleine und mittlere Unternehmen wieder Zugang zum Markt finden. Denn gerade letztere sind unzureichend digitalisiert[15].

Box 1

Der Industrial Data Space formt einen dezentralen branchenübergreifenden virtuellen Raum für den souveränen Austausch von Daten und erarbeitet eine internationale Standardisierung und Zertifizierung des Datenaustausches[33]. Die Mitglieder verwalten und kontrollieren ihre Datenflüsse, ohne dass diese auf einer Cloud gespeichert werden müssen. Nur zertifizierte Mitglieder erhalten Zugang und wenn ein Datenaustausch angefragt wird, können die Bedingungen mit den Daten selber verbunden werden und somit individuell kontrolliert werden. Der sichere Datenaustausch lässt sich mittels einer Verbindungssoftware (IDS Connector) umsetzen, an der wiederum andere Service (Cloud-Plattformen, IoT etc.) angeschlossen werden können.

DIE PRIORITÄT DER PLATTFORM-ÖKONOMIE

Eine große Vormachtstellung in der Digitalisierung wird den Plattformen zuerkannt, welche bereits zu großen *disruptiven* Veränderungen auf dem internationalen Markt geführt haben – so wie Google, Apple, Facebook und Amazon. Plattformen haben nicht nur die Möglichkeit den gesamten Markt zu verändern, da sie kolossale Nutzergemeinden erreichen können und somit eine starke Marktdominanz schaffen, sondern auch da sie intrinsische Innovationswerte zeigen und somit selber

das Produkt einer neuen Servicewelt darstellen. Darum bilden Plattform-basierte Geschäftsmodelle die Grundlage des digitalen Fortschritts. Darüber hinaus stellen Plattformen eine Möglichkeit dar, Chancengleichheit und Fairness sicherzustellen und sollten somit primär durch unabhängige Förderinstitutionen eingerichtet werden.

In Deutschland gibt es mehrere Initiativen, die die Verbesserung und Homogenisierung der digitalen Infrastruktur fördern, wie z.B. der *Industrial Data Space*, (Box 1) der 2016 durch eine Kooperation der Fraunhofer-Gesellschaft, der Industrie und mehreren Bundesministerien entwickelt wurde. Diese Referenzarchitektur ist durch ihren modularen Charakter sehr flexibel und erlaubt eine hohe Datenprivatheit und Souveränität. Jedoch ist es kein offenes System und die Mitglieder müssen trotz des dezentralen Charakters eine umsatzorientierte Gebühr für die Nutzung bezahlen. Die Preisgestaltung ist ein zentrales Element der Attraktivität einer Plattform und könnte bei dem *Industrial Data Space* Modell zusätzlich durch das erzwungene Erwerben von Datendiensten verschiedener Anbieter durch den *App-Store* technologisch beeinflusst werden. Außerdem gibt es strategische Partnerschaften, die sich durch Abhängigkeiten innerhalb der Vereinsstrukturen abbilden.

Ein europäisches Projekt, DECODE (Box 2), versucht die Monopolisierung der Internetdienste und die Fragmentierung des IoTs durch einen neuen europäischen Standard von den USA abzuheben. Datensicherheit und -souveränität bilden zentrale Elemente, welche u.a. durch Normen wie das *Security und Privacy-by-Design*, sowie konstante Verschlüsselung der Daten umgesetzt werden. Die BürgerInnen halten nicht nur die Hoheit ihrer Daten, sondern werden auch durch *Data Commons*, sogenannten Datenkommunen, unterstützt. Transparenz und *Empowerment* sollen die NutzerInnen motivieren, Daten zu teilen und aktiv an der Umgebungsgestaltung teilzunehmen.

Box 2

DECODE wird mit 14 Partnern aus der Industrie, Forschung, Recht, und internationalen Gemeinden, wie der Stadt Amsterdam umgesetzt. Ziel ist es, eine Nutzer- und Daten-zentrierte Peer-to-Peer Infrastruktur zu entwerfen, die der Gemeinde alle Rechte und Verwaltungsfreiräume lässt. Daten sollen durch die BürgerInnen, das IoT, und Sensornetzwerke generiert werden und mit gesellschaftlicher Relevanz angewendet werden. Daten werden somit nicht mehr nur als Wirtschaftsgut gesehen, sondern als Gesellschaftsgut, die später durch Innovatoren, Startups, NGOs, Kooperationen und lokale Gemeinden verwertet werden und in Apps und Dienste, zugeschnitten auf die Bedürfnisse umgesetzt werden.

Beide dieser Konzepte leiden unter geringer Mitgliederbeteiligung, da sie keine großen Anreize schaffen, außerhalb der Datensicherheit und -souveränität, breitflächige Daten zu teilen. Außerdem können potentielle Daten-Transaktionskosten beim *Industral Data Space*, sowie ein Mangel an Analysemöglichkeiten und ein selektiver und limitierter Datenaustausch weitere Hindernisse bilden. Verständlicherweise geben Unternehmen nur Daten frei, wenn ihnen diese keinen Wettbewerbsnachteil bereiten. So zählt der *Industrial Data Space* bisher nur 81 Mitglieder, worunter sich nur sechs Universitäten und meist große Unternehmen befinden. Da so ein Konzept von den Mitgliedern lebt, bleibt es abzuwarten, ob es ein Markterfolg werden kann. Dagegen zählt die US-amerikanische Version dieses Konzeptes, das *Industrial Internet Consortium*, bereits 238 Mitglieder.

Eine Motivation zum Datenteilen

Ein zentraler Punkt des Erfolges von Datenplattformen liegt in dem Inzentiveren des Datenteilens und des Erreichens einer kritischen Nutzermasse, welche die Vitalität der Plattform bestimmt. Die meisten Webseiten erfordern Aufwand seitens der DatenspenderInnen. Sollte der Mehrwert der Plattform dann nur geringfügig sein, gibt es keinen Anreiz, den Service zu nutzen und Daten zu teilen. Außerdem sind *Usability* und *User Experience* weitere wichtige Punkte der Nutzerakzeptanz; eine Plattform muss vor allem intuitiv zu bedienen sein und schnell die wichtigsten Informationen zusammenfassen können. Darüber hinaus müsste es auch für EntwicklerInnen genügend Anreize geben, die Struktur zu nutzen und zu optimieren. Open-Source APIs (*application programming interfaces*) stellen zwar ein gewisses Risiko für die folglich geöffnete und somit auch geteilte Plattform dar, allerdings bilden gut dokumentierte Open-Source APIs einen großen Anreiz für EntwicklerInnen, da diese folglich Teil des Plattformumsatzes werden können und ihre eigenen Subsysteme und Innovationen in die Plattform integrieren können.

DatenspenderInnen könnten außerdem durch die Einführung eines Daten-Eigentumsrechts stimuliert werden, welches die Einzelnen befähigt, ihre personenbezogenen Datenmaterialien autonom und transparent einzusetzen. Solch eine Nutzer-zentrierte Verwaltung ist das Prinzip des konditionellen Datenspendens (Nutzer-zentrierte Datenkontrolle-by-Design), welches den SpendernInnen die Zustimmung der Datennutzung für bestimmte Zwecke ermöglicht und somit exklusive Datenrechte aberkennt[24],[25]. Eine ähnliche Idee des

Hasso-Plattner-Institutes, der sogenannte *Data Donation Pass*, erlaubt es, dass Personen ihre Daten für spezifische Zwecke spenden können und im Umkehrschluss medizinisches Personal oder Familienangehörige Zugang zu bestimmten Gesundheitsdaten erhalten. Die zentrale Verwaltung dieses Eigentums durch TreuhänderInnen oder die dezentrale durch *Block-Chain* Technologien könnte derartige Rechte sichern.

Open-Access und Open-Data sind elementare Voraussetzungen für das Datenteilen und stimulieren einer interaktiven Forschungsgemeinschaft, da der offene Zugang zu Daten (inklusive Metadaten und Software) und Informationen einen regen und kritischen Austausch ermöglicht. Open-Access und Open-Data sind außerdem Grundvoraussetzungen für die wissenschaftliche Entwicklung und ein zentrales Element der Informationsfreiheit. Darum ist bei öffentlichkeitsrelevanten Daten zu überlegen, ob Datensammler verpflichtet werden, die Informationen und Daten in frei zugänglichen Repositorien abzulegen, sobald wichtige strategische Ziele, wie die Veröffentlichung der Daten, erreicht sind. Dies ist unabdingbar für eine Entwicklung der Datenanalysetechniken und für sekundäre Datenwertschöpfungen, sowie die Verifizierung und Reproduzierbarkeit der Resultate. Außerdem würde dies weniger entwickelten Ländern die Möglichkeit bieten, ohne eigene Forschungsvorhaben am Datenfortschritt teilzunehmen[26]. Darüber hinaus ermöglichen offene Daten die breitflächige Validierung von Algorithmen bezüglich potentiell diskriminierender und unfairer Anwendungen.

In den USA gibt es bereits in der klinischen Forschung *The International Committee of Medical Journal Editors* (ICMJE), welche neue Standards setzen und ForscherInnen verpflichten die de-identifizierten Daten nach Veröffentlichung der Studienresultate offen zu sammeln[27]. Auch die Deutsche Forschungsgemeinschaft und das Bundesministerium für Bildung und Forschung verpflichten geförderte Institutionen zur Open-Access Veröffentlichung oder Ablage in öffentlich zugänglichen Repositorien. Ähnlich könnten ForscherInnen motiviert werden Daten zu teilen, indem der Wert ihrer Forschung an der Erlaubnis zur Datenteilung festgelegt wird, durch *Credits* oder einen Einfluss auf den *Impact Factor*. Zur Förderung von Sekundäranalysen gibt es bereits einige Initiativen, wie die *Special Interest Group Management of Data* (SIGMOD) und die *Very Large Data Base* (VLDB) Konferenz, welche Reproduktionen mit neuem Erkenntnisgewinn belohnen und als vollwertige Artikel veröffentlichen. Solche Förderungen sind unabdingbar wichtig für die Qualität und Entwicklung der Forschung, müssen aber parallel durch ethische Kodexe und einen interaktiven Austausch zwischen den WissenschaftlernInnen, Unternehmen und der Gesellschaft untermauert werden. Denn nicht jedeR DatenhändlerIn ist gleichermaßen verantwortlich, sollte sich aber der Konsequenzen und des Dilemmas bewusst sein und generell nach dem Prinzip des nicht-Schadens Handeln und forschen. Das letztendliche Ziel der Open-Access und Open-Data Bestrebungen sollte sein, die verteilten Datensilos in transnationale Open-Access Datenplattformen zu vereinen[28], die würde die Nutzergemeinden erhöhen, den (europäischen) Markt homogenisieren und zentral stellen und die Relevanz der Plattform durch den Informationsumfang steigern.

Generell ist hierbei anzumerken, dass zentrale Plattformen in der Praxis erfolgreicher sind. Dezentrale Plattformen sind in der Forschung und aufgrund des deutschen und europäischen Föderalismus zwar populär, unterliegen jedoch in Hinblick auf Einfachheit der Entwicklung und des Betriebs, *Economies of Scale*, sowie im Hinblick auf Kontrolle und Expansion regelmäßig den zentralen Lösungen (z.B. Google, Facebook, Amazon, Twitter etc.).

TECHNOLOGISCHE HERAUSFORDERUNGEN

Eine große Herausforderung im Hinblick auf digitale Souveränität und Datensouveränität für deutsche und europäische Unternehmen ist es, als *effective Follower* bei den Software-Technologien und Infrastrukturen zur Verwaltung und Analyse sowie des maschinellen Lernens von großen Datenmengen aufzuholen, die in den marktbeherrschenden amerikanischen und chinesischen Unternehmen eingesetzt werden. Die großen Datenplattformen wie Amazon oder Google haben eigene Infrastrukturen basierend auf eigenen Technologien mit großem Entwicklungsaufwand und *Know-How* erzeugt. Für einen deutschen oder europäischen Herausforderer wäre eine derartige Entwicklung prohibitiv. Dennoch ist es zum Erlangen einer digitalen Souveränität wichtig, mindestens einen Herausforderer einer Datenplattform zu schaffen. Vielversprechende Alternativen sind Open-Source Systeme, welche die Entwicklungen einer globalen Entwicklergemeinschaft einsetzen und somit konstant Nutzer-zentriert optimiert werden. Eine kontinuierliche Verbesserung und parallele Entwicklungsmöglichkeiten, die offene Systeme bieten, sind nötig, um Netzwerkeffekte zu erreichen und flexibel auf Nutzerbedürfnisse einzugehen. Ein erster Ansatz für eine deutsche Open-Source und Open-Data Infrastruktur wurde mit der *MCloud* realisiert, die zusammen mit dem Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) erarbeitet wurde und offene Verkehrsdaten bietet. Dennoch bleiben auch hier große Nutzergemeinden aus. Ein Nutzer-zentrierter Design Workshop des BMVI zeigte, dass die Gründe hierfür unvollständige Daten, langes Suchen und schlechte Datenvisualisierung sowie das Fehlen eines direkten Kontaktes zu den Datenlieferanten sind. Die DatenspenderInnen hingegen

Preprint

verlangen nach deutlicheren Normen und einer stärkeren Unterstützung in der Handhabung von Open-Data seitens der Behörden. Gleichzeitig fehlt in der *MCloud* bisher die Infrastruktur zur einfachen Datenanalyse und eine stetige Datenkuratierung mit Qualitätskontrolle sowie das Einführen von tagesaktuellen Daten, welche kritische Erfolgsfaktoren für eine Datenplattform darstellen.

Open-Source, Open-Data und Open-Access sind zwar Bedingungen für das Datenteilen und für technologische Entwicklungen, da diese den Fortschritt von TDM und anderen Datenanalyse- und Lerntechniken durch freizugängliche Daten ermöglichen[29], allerdings muss das System selber genügend Mehrwert bieten, NutzerInnen und EntwicklerInnen zu binden. Neben den bereits erwähnten wichtigen Effekten von *Usability* und *User Experience* bilden die Qualität und die Dokumentation der Programmierschnittstellen (APIs) die Grundlage für eine offene und einladende Infrastruktur. Diese bestimmen den Zugang und Datenfluss der Infrastruktur und ermöglichen so eine individuelle Gestaltung und den Fortschritt durch fremde EntwicklerInnen und NutzerInnen. Auf der anderen Seite profitieren die Plattform-BereitstellerInnen von externen Innovationen, neuen Partnernetzwerken sowie fortwährenden Aktualisierungen und professionellem Feedback. Es gilt darum, die Plattform nach außen so zu öffnen, dass sich innovative Modelle entwickeln können und die Bemühungen belohnt werden, während wichtige Kontrollmechanismen auf der Plattform bewahrt werden. Parallel sollte das Realisieren von geeigneten Infrastrukturen durch Technologie-Entwicklungen unterstützt werden. Diese betreffen nicht nur Werkzeuge, die das Verarbeiten von Daten erleichtern und gleichzeitig die Datenhoheit wahren, wie maschinelle De-identifizierungen und *Privacy-by-Design* gesteuerte Applikationen, sondern auch Techniken, die die Rückverfolgung, Reproduzierbarkeit und Identifizierung von Datenquellen ermöglichen[22].

Darüber hinaus ist eine grundlegende Bedingung von souveränen digitalen Infrastrukturen eine hohe Datensicherheit, diese ist auch eine Voraussetzung für die Gewährung des Privatrechts. Viele Unternehmen sind nicht nur unzureichend geschützt und benötigen ein antizipatives Beobachtungssystem und Risikobewertungen für Datensicherheitsverletzungen[30], sondern bilden ihr Personal in Bezug auf Cybersicherheit und –bewusstheit ungenügend aus. Da 82% der Datendiebstähle durch interne Personalfehler ermöglicht werden[31] und daher als größtes Risiko für Datenlecks gelten[32], wäre ein integriertes (öffentliches) Sensibilisierungs-Programm eine wichtige Maßnahme zusätzlich zu zielgruppenspezifischen Trainings. Darüber hinaus gibt es einen großen Bedarf für die Entwicklung von Methoden zur Identifikation von Datenlecks und auch für die Schaffung von angemessenen Verteidigungsmaßnamen, gerade im hochsensiblen Gesundheitsbereich.

EINE DATENGETRIEBENE WISSENSINFRASTRUKTUR: IT-AIRBUS

Oben wurde bereits beschrieben, dass der Zugang zu dem und die Verarbeitung des Produktionsfaktors Big Data eine große Herausforderung darstellt; in der Wirtschaft ist dieser mangelnde Zugang oftmals ein Hemmnis für die Gründung von innovativen Unternehmen, in der Wissenschaft ist es ein Hemmnis für die Durchführung von datengetriebenen Forschungsarbeiten. Dafür gibt es mehrere Gründe: (1) Zum einen sind relevante Daten aus Datenschutzgründen oder kommerziellen Interessen oftmals schwer für WissenschaftlerInnen und Data Scientists zugänglich. (2) Zum anderen steht vielen WissenschaftlernInnen mit Datenzugang selbst die für die Analysen erforderliche Rechnerinfrastruktur nicht zur Verfügung. Hierbei ist insbesondere anzumerken, dass bei großen Datenmengen Cloud-Computing an seine Grenzen stößt, falls die Daten erst in die Cloud übertragen werden müssen bzw. innerhalb der Cloud viele Datentransfers zur Analyse erforderlich sind und die Analysesoftware diese nicht automatisch minimiert. (3) Überdies ist die Erstellung von Datenanalyseprogrammen oftmals so komplex und aufwändig, dass es wenigen Experten aus dem Bereich der Informatik vorbehalten ist, Datenanalysen durchzuführen. Insbesondere besteht die Gefahr, dass WissenschaftlerInnen aus den Rechts-, Geistes- und Sozialwissenschaften (Stichwort: *Digital Humanities*) oder interessierte, aber weniger informatikaffine BürgerInnen, insgesamt von der Digitalisierung abgekoppelt werden.

Gleichzeitig existiert durch das Internet ein riesiger Datenschatz, der beispielsweise für Wirtschafts-, Rechts-, Sozial- und Geisteswissenschaften oder gesellschaftliche Studien eine riesige Chance bietet, Forschungen auf aktuellen Daten durchzuführen sowie eine statistisch signifikante Datenbasis zur Ableitung neuer Erkenntnisse durch statistische Methoden zu etablieren. Beispielsweise könnten WirtschaftswissenschaftlerInnen für eine Analyse die Frage stellen: „Wie ist das Durchschnittsalter von Vorstandsvorsitzenden in Unternehmen nach Land und Branche gruppiert?“, um Korrelationen zwischen Vorstandsalter und Innovationskraft von Unternehmen zu untersuchen und dies auch einem Ländervergleich zu unterziehen. Andere WissenschaftlerInnen könnten sich für die Anzahl von Firmenkäufen und Fusionen bei Unternehmen im Bereich der nachhaltigen Energiegewinnung interessieren. Ähnliche strukturierte datenbezogene Fragestellungen betreffen auch die Natur- und Ingenieurswissenschaften oder

Preprint

interessierte BürgerInnen, insbesondere im Hinblick auf Sensordaten oder das Internet of Everything. Jedoch können gegenwärtige Infrastrukturen derartige Fragen nicht beantworten, insofern diese eine Synthese einer Datenvielzahl aus dem Internet oder anderen Quellen erfordern und somit die reine Suche von Daten übersteigen.

Für die genannten Fragestellungen sind die Daten zwar prinzipiell im Internet verfügbar, allerdings müsste der Analyseprozess manuell bzw. durch programmierte Skripte mittels vieler Suchanfragen bei üblicherweise außereuropäischen Suchmaschinenanbietern realisiert werden, deren Ergebnisse dann extrahiert und integriert werden müssten, um das Analyseergebnis zu erhalten. Die zielgerichtete, automatisierte Beantwortung einer Fragestellung wird jedoch nicht unterstützt. Anstelle einer einfachen, sogenannten *deklarativen* Formulierung der Frage müssen die vielen Einzelschritte zur Berechnung des Ergebnisses von SoftwareentwicklernInnen in ein Programm zusammengefasst und auf einer skalierbaren Datenanalyseplattform adaptiert werden. Ferner können Suchmaschinen AnwenderInnen blockieren, falls diese eine Vielzahl von Suchanfragen automatisiert veranlassen und verhindern somit oftmals den Zugriff auf eine statistisch signifikante Datenmenge.

Forschungsarbeiten im Bereich des Datenmanagements, des maschinellen Lernens, der verteilten Systeme, des Supercomputing und der Rechnernetze haben im letzten Jahrzehnt Technologien geschaffen (z.B. Theseus) bzw. schaffen diese derzeit [z.B. auch in den deutschen Big Data Kompetenzzentren, dem Berlin Big Data Center (BBDC) und dem Competence Center for Scalable Data Services and Solutions (ScaDS)], um derartige komplexe Fragestellungen auf riesigen, heterogenen Datenquellen deklarativ zu spezifizieren und überdies die Informationsextraktion und -integration automatisiert unter Einsatz von Verfahren des maschinellen Lernens zu realisieren, sowie um die Antworten auf diese Fragestellungen massiv, parallel und effizient zu berechnen und als Analysedienste bereitzustellen. Aufgrund der Vorarbeiten ist ein erster Aufbau einer derartigen Infrastruktur inzwischen technologisch möglich.

Zur Erreichung einer adäquaten deutschen Datensouveränität im Hinblick auf Bildung und Forschung und zur Beantwortung der obengenannten Fragestellungen ohne ausgeprägte Programmierkenntnisse und Investitionen durch WissenschaftlerInnen oder BürgerInnen insgesamt zu ermöglichen, ist die Schaffung und der kontinuierliche Betrieb einer Wissensinfrastruktur erforderlich. Diese geht über die vom Rat für IT-Infrastrukturen diskutierte Struktur hinaus, indem sie – im Hinblick auf Datensouveränität – den Zugang und die Verarbeitung von Daten und die dafür erforderlichen Dienste in einem Open-Source und Open-Data-Umfeld in den Vordergrund stellt. Diese Infrastruktur sollte sowohl (1) kontinuierlich die Daten des Internets und aus wissenschaftlichen Projekten, als auch (2) die Berechnungskapazität, eine Softwareplattform und (3) eine auf deklarativen Konzepten zum einfachen Gebrauch basierende Benutzerschnittstelle zur Analyse und visuellen Interaktion mit den Daten zur Verfügung stellen. Ziel ist es dabei, das Internet und weitere Datenströme und -quellen (z.B. aus wissenschaftlichen Experimenten oder Medien) als dynamische, stetig wachsende Datenbank zu behandeln, wobei Informationsextraktions- und Integrationsverfahren zur Ableitung von semantischen Informationen aus den Rohdaten verwendet werden. Durch den Einsatz von Verfahren des maschinellen Lernens könnten Informationsextraktionsverfahren (z.B. das Erkennen von Dokumenten) durch Angabe von Beispielen realisiert werden (d.h. im obengenannten Beispiel könnte durch das Markieren von exemplarischen Dokumenten, die sich auf nachhaltige Technologien beziehen, ein Modell bzw. Sprachmuster gelernt werden, um diese automatisiert zu erkennen und zu analysieren). Daraus könnte ein kontinuierlich wachsender Wissensgraph aufgebaut werden, der semantische Schlussfolgerungen erlaubt und durch Bootstrapping die Informationsqualität weiterer Informationsextraktionsschritte erhöht. Neben den Daten des Internets könnten WissenschaftlerInnen auch weitere bereits annotierte, hochwertige Daten bereitstellen, um die Basis der Wissensinfrastruktur zu erweitern. Ferner könnten in dieser Infrastruktur die NutzerInnen im Sinne von Open-Source die Extraktions-, Integrations-, Analyse- und Informationsvisualisierungsmethoden teilen, um auf diese Weise von Vorarbeiten und Datenaufbereitungen anderer NutzerInnen direkt zu profitieren.

In Deutschland ist eine derartige Wissensinfrastruktur bisher nicht vorhanden. Rechenzentren fokussieren sich bislang auf die Bereitstellung von Rechen- und Datendiensten, häufig im Hinblick auf wissenschaftliches Rechnen und optimieren dafür die Datenverarbeitung. Normalerweise ist der Fokus jedoch nicht die Datenbereitstellung, sondern die leichtgewichtige semantische Integration heterogener Daten, sowie die Minimierung von Datentransfers, welche für ein effizientes Datenmanagement im Rahmen einer Wissensinfrastruktur nötig sind. Das Smart Data Innovation Lab (SDIL) fokussiert sich darauf, eine Menge von Datenquellen und Analysesystemen bereitzustellen, hat jedoch keine integrierte Sicht auf Daten und Analysen und ignoriert den für die Wissenschaften größten Datenschatz als Datenquelle: Das Internet. Weder Rechenzentren noch das SDIL stellen Datendienste

Preprint

bereit, die die gesamte Datenwertschöpfungskette behandeln und deklarativ und damit leicht zu bedienen sind, wie die beispielgetriebene Spezifikation von Datenanalysen, sowie eine Community, die Informationsextraktionen, Integrationsverfahren, Analysen, Ergebnisse und Visualisierung teilt und leicht als Basis für neue Algorithmen und Untersuchungen wiederverwenden kann.

Demzufolge gibt es eine Notwendigkeit für eine aus obengenannten Gründen zentralisierte Wissensinfrastruktur in Deutschland und Europa, welche (1) kontinuierlich semantisch angereicherte webbasierte Informationen bereitstellt und die Integration weiterer Datenquellen erlaubt. Darüber hinaus sollte diese Infrastruktur (2) eine einfache, beispielgetriebene, deklarative Spezifikation von Analysen ermöglichen, cloudbasierte Analysedienste anbieten und außerdem (3) eine Community zum Teilen und Wiederverwenden von Daten, Informationsextraktionsverfahren, -integrationsverfahren, -visualisierungen, Analyseverfahren und -ergebnissen realisieren. Eine derartige Wissensinfrastruktur bildet das Rückgrat der gesellschaftlichen Digitalisierung, welche eine ähnliche Bedeutung erhalten wird, wie sie die Vernetzung im Laufe des letzten Jahrzehnts insgesamt erlangt hat. Um die Führungsposition, Innovationskraft und internationale Wettbewerbsfähigkeit des Wissenschaftsstandorts Deutschland zu erhalten und die wissenschaftlichen Prozesse in die digitale Zukunft zu überführen, stellt eine entsprechende Infrastruktur einen kritischen Erfolgsfaktor dar. Letztendlich wird eine Wissensinfrastruktur auch ein wichtiger Baustein für die Zukunftsfähigkeit von Wirtschaft und Gesellschaft in Deutschland sein, da diese wichtiges Know-How für die Erstellung und den Betrieb solcher Infrastrukturen im kommerziellen Umfeld darstellt. Gleichzeitig sollte eine entsprechende Wissensinfrastruktur nach einer erfolgreichen Etablierung nicht nur WissenschaftlernInnen vorbehalten bleiben, sondern für alle BürgerInnen geöffnet werden, um Fragen zu beantworten und Recherchen durchzuführen.

Neben der Wissenschaft würde auch die Wirtschaft, insbesondere der Mittelstand und die Startups, welche selbst derartige Strukturen aus Kosten- und Kompetenzgründen nicht stemmen könnten, massiv von solch einem Informationsmarktplatz profitieren. Es wäre jedoch erst zu prüfen, ob die Öffnung solch einer Struktur für die Wirtschaft, den Mittelstand und die Startups den Vorzug erhalten sollte, oder ob parallel eine ähnliche Infrastruktur mit wirtschaftlichen Zielsetzungen aufgebaut werden müsste. Dies könnte beispielsweise im Rahmen einer neutralen Stiftung geschehen, die Daten, Analysen, Visualisierungen etc. als Dienste bereitstellt und es Unternehmen ermöglicht, darauf Mehrwertdienste zu entwickeln und anzubieten. Eine derartige Stiftung könnte sich nach einer staatlichen Anschubfinanzierung durch Erträge von der Nutzergemeinde finanzieren. Die Schaffung einer ähnlichen datengetriebenen Infrastruktur sollte jedoch als nationale oder europäische Aufgabe begriffen werden, welche politischen Willen und massive Investitionen erfordert. Hierbei bildet der Airbus eine vergleichbare Initiative, mit der ebenfalls die europäische Souveränität (im Flugzeugbau) gegenüber einem Wettbewerbsvorsprung von außereuropäischen Unternehmen behauptet werden kann. Solch eine Wissensinfrastruktur kann als ein IT-Airbus betrachtet werden, der mit ähnlicher Vehemenz sowie auch mit Agilität vorangetrieben werden müsste.

ZUSAMMENFASSUNG UND HANDLUNGSEMPFEHLUNGEN

Das Internet und die fortschreitende Digitalisierung eröffnen viele neue Freiheiten und Möglichkeiten, welche intrinsisch an komplexe Verantwortlichkeiten und ethische Fragestellungen gekoppelt sind. Die Frage nach einer dualen Lösung, die sowohl die Entwicklung von Big Data getriebenen Innovationen fördert, als auch die Autonomie der DatenspenderInnen, muss schlussfolgernd in einem datensouveränen Umgang der Einzelnen, der Unternehmen und der Öffentlichkeit zu finden sein. Um dieser Ambition gerecht zu werden, müssen sowohl Gesetzgebungen, ethische Kodexe, unabhängige Support-Gruppen und Bildungsförderungen ineinandergreifen.

Folgende Handlungsempfehlungen sollten zur Stärkung der deutschen Wirtschaft umgesetzt werden:

- Priorität der digitalen Infrastruktur-Entwicklung, inkl. geeigneter Analysemöglichkeiten, mit dem Ziel einer europäischen Markthomogenisierung
- Aufbau einer Wissensinfrastruktur bestehend aus kontinuierlich aktualisierten und kurierten Datenquellen sowie einem cloudbasierten Analysesystem nach den Prinzipien von Open-Source und Open-Data, das die gesamte Datenwertschöpfungskette vom Extraktion, Integration, Analyse, Visualisierung und Modell-bildung beinhaltet
- Europäischer Fokus der Datensouveränität auf technologischer Basis (IT-Airbus)
- Förderung von Open-Source, Open-Data und Open-Access als Leitfaden des digitalen Innovationsökosystems
- Öffentliche Anreize zur Datenteilung für Unternehmen und BürgerInnen
- Erzwingung von Open-Access und Open-Data bei öffentlichkeitsrelevanten Daten

Preprint

- Förderung der IKT-Bildung als nationale Priorität
- Förderung der Technologieforschung durch strategische Forschungs- und Entwicklungsschwerpunkte
- Entwicklung von bindenden ethischen Richtlinien für eine *Good Big Data Practice* der Forschung und Wirtschaft
- Entwicklung von bindenden technischen Richtlinien für eine *Good Big Data Practice* der Forschung und Wirtschaft
- Politische Rahmenbedingungen für die Förderung von einer aktiven Wagniskapitalbranche etablieren
- Datensouveränität stützen (*Commons*, DatentreuhänderInnen etc.) und ausbauen durch technische Souveränität, wie einen Datenspenderausweis
- Digitale Datenverwertungen durch Drittparteien gesetzlich angreifen oder national befähigen

REFERENZEN

1. Cisco. VNI Complete Forecast Highlights. (2016). Available at: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf.
2. Jacobson, R. 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? IBM (2013). Available at: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/.
3. Hey, T, Tansley, S & Tolle, K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009
4. Abadi, D. et al. The Beckman report on database research. Commun. {ACM} 59, 92–99 (2016).
5. Rohleder, B. Digitale Souveränität Positionsbestimmung und erste Handlungsempfehlungen für Deutschland und Europa. BITKOM (2015).
6. Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Deutscher Ethikrat. Stellungnahme (2017).
7. Hansson, M. G. et al. The risk of re-identification versus the need to identify individuals in rare disease research. Eur. J. Hum. Genet. 24, 1553 (2016).
8. Evans, B. Power to the People: Data Citizens in the Age of Precision Medicine. Vanderbilt J. Entertain. Technol. law 19, 243–265 (2017).
9. Markl, V., Krcmar, H. & Hoeren, T. Big Data Management Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen. Bundesministeriums für Wirtschaft und Technol. (2014).
10. Razaghpanah, A, Nithyanand, R, Vallina-Rodriguez, N, Sundaresan, S, Allman, M, Kreibich, C, Gill, P. Apps, Trackers, Privacy and Regulators: A Global Study of the Mobile Tracking Ecosystem. Proc. NDSS (2018).
11. Terry, K. Patient Access to Health Records Not Smooth or Easy: Report. Medscape Medical News (2018).
12. Solving Murder. What do they know? Econ. 23
13. The battle for digital supremacy. The Challenger. Econ. 19–22
14. Wissenschaftsfreiheit und Wissenschaftsverantwortung. Empfehlungen zum Umgang mit sicherheitsrelevanter Forschung. Deutsche Forschungsgemeinschaft, Deutsche Akademie der Naturforscher Leopoldina e.V. (2014).
15. European Commission, E. C. Bericht über den Stand der Digitalisierung in Europa 2017 – Länderprofil Deutschland. 1–11 (2017). Available at: ec.europa.eu/newsroom/document.cfm?doc_id=44307.
16. IT-Fachleute. Agentur für Arbeit. Berichte: Blickpunkt Arbeitsmarkt (2018).
17. Bitkom. Der Arbeitsmarkt für IT-Fachkräfte. (2016).
18. Stiftung, B. Digitalisierung an Schulen: Der Geist ist willig, das WLAN ist schwach. Themen (2017).
19. Breiter, Andreas; Zeising, Anja; Stolpmann, Björn, E. Impulse, die Schule machen. IT-Ausstattung an Schulen: Kommunen brauchen Unterstützung für milliardenschwere Daueraufgabe. Bertelsmann Stift. (2017).
20. Digitale Infrastruktur mit Schwächen. Innovationsindi

Preprint

kator. BDI Acatech (2017).
21. Frietsch, R; Lichtblau, K; Beckert, B. et al. Elektroin dustrie als Leitbranche der Digitalisierung – Innovati onschancen und Innovationshemmnisse für die Elekt roindustrie. Frankfurt (2016).
22. Smart-Data-Begleitforschung & Informatik. Daten als Wirtschaftsgut. WIRmachenDRUCK, Backnang 1–56 (2017). Available at: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/2017-11-22_smartdata_daten_wirtschaftsgut.pdf?__blob=publicationFile&v=3.
23. Maier, M. Start-ups in Deutschland und den USA. Aus druck unterschiedlicher Innovationskultur. Anal. und Argumente. Konrad Adenauer Stift. 99, (2011).
24. Strandburg, K; Frischmann, B; & Madison, M. Governing Medical Knowledge Commons. Cambridge Studies on Governing Knowledge Commons (Cambridge University Press, 2017).
25. Belli, L; Schwartz, M; Louzada, L. Selling your soul while negotiating the conditions: from notice and consent to data control by design. Health Technol. (Berl). 7, 453–467 (2017).
26. Boulton, G., Babini, D., Hodson, S. & Li, J. Open Data in a Big Data World. An international accord. Sci. Int. (2015).
27. Taichman, D. B. et al. Sharing Clinical Trial Data — A Proposal from the International Committee of Medical Journal Editors. N. Engl. J. Med. 374, 384–386 (2016).
28. Healthcare Subgroup, T. Big Data technologies in Healthcare. Needs, opportunities and challenges. Big Data Value Assoc. (2016).
29. JeTriaille, J-P, de Meeûs d'Argenteuil, J & de Franc quen, A. Study on the legal framework of text and data mining (TDM). Wolf Partners. Eur. Union. 1–118 (2014).
30. Hayes, D. R. & Cappa, F. Open-source intelligence for risk assessment. Bus. Horiz. (2018).
31. Fogarty, K. 82% of Data Breaches Due to Staff Errors; 4% of IT Trusts Users; IT is Still to Blame. ITWorld (2012).
32. Sanzgiri, A. & Dasgupta, D. Classification of Insider Threat Detection Techniques. in Proceedings of the 11th Annual Cyber and Informa tion Security Research Conference 25:1--25:4 (ACM, 2016).
33. Otto, B; ten Hompel, M; Wrobel, S. Industrial Data Space. in Digitalisierung (ed. Neugebauer, R.) 113–133 (Springer Vieweg, 2018).

Prof. Dr. rer. nat. Volker Markl
Volker Markl ist Direktor des vom Bundesministerium für Bildung und Forschung eingerichteten Berliner Big Data Centers und des Smart Data Forums. Volker ist überdies ordentlicher Professor und Inhaber des Lehrstuhls für Datenbanksysteme und Informationsmanagement an der Technischen Universität Berlin. Außerdem hält er eine Anstellung als außerordentlicher Professor an der Universität Toronto und ist der Leiter der Forschungsgruppe Intelligente Analyse von Massendaten am DFKI. Er hat mehr als 100 Forschungsarbeiten an erstklassigen wissenschaftlichen Einrichtungen veröffentlicht und hält 18 Patente und zahlreiche renommierte Auszeichnungen, wie den IBM Outstanding Technology Award. Er war Referent und Projektleiter der von der Deutschen Forschungsgemeinschaft geförderten Verbundforschung Stratosphere, aus der zahlreiche Top-Tier-Publikationen sowie das Big-Data-Analysesystem Apache Flink hervorgingen. Volker ist derzeit Präsident der VLDB Stiftung und wurde 2014 von der Deutschen Gesellschaft für Informatik zu einem der führenden digitalen Köpfe Deutschlands gewählt.

Prof. Dr. Tilmann Rabl
Tilmann Rabl ist Gastprofessor und Forschungsdirektor der Gruppe Datenbanksysteme und Informationsmanagement (DIMA) und Koordinator des Berlin Big Data Centers (BBDC). Er ist Professional Affiliate des Transaction Processing Performance Councils (TPC) und Mitglied des Steering Committees der Research Group der Standard Performance Evaluation Corporation (SPEC). Tilmann Rabl ist zudem CEO und Mitbegründer des mehrfach ausgezeichneten Startups bankmark.

Dr. Jack Thoms
Jack Thoms ist stellvertretender Leiter der Forschungsgruppe Intelligente Analyse von Massendaten am DFKI. Er leitet das Smart Data Forum, eine gemeinsam vom Bundesministerium für Wirtschaft und Energie (BMWi) und dem Bundesministerium für Bildung und Forschung (BMBF) getragene Wissens, Demonstrations- und Vernetzungsplattform für digitale Technologien und Innovationen in Deutschland.

Dr. Anne Schwerk
Anne Schwerk ist Projektmanagerin für den Bereich Medizin und künstliche Intelligenz am DFKI. Nach ihrer Promotion in der Klinik für Neurologie an der Charité arbeitete Anne für zwei Jahre bei TNO in den Niederlanden, wo sie u.a. die Pathologie automatisierte und digitalisierte.

Preprint