Datengetriebene Forschung

Ein Beitrag aus wissenschaftsreflexiver Perspektive

| GABRIELE GRAMELSBERGER | MATTHIAS MÜLLER |

Wachsende Datenfluten bewegen die Wissenschaft. In Europa soll Open Science diese Datenfluten erschließen und zugänglich machen. Was heißt das konkret für die Forschung, das Management von Forschungsdaten und die Wissenschaftsfreiheit?

Im März 2018 veröffentlichte die Europäische Kommission ihre Roadmap zur Implementierung der *European Open Science Cloud* (EOSC). Für 1,7 Millionen Europäische Forscher und Forscherinnen sowie für 70 Millionen Beschäftigte in Entwicklung und Technologie soll in den nächsten Jahren eine freie und offene Dateninfrastruktur aufgebaut werden. Dabei geht es um die Bildung eines gemeinsamen europäischen Datenraumes, der neben der EOSC auch den Datenaustausch mit der Wirtschaft regeln soll (SWD(2018) 125). Ziel der Europäischen Kommission ist es, eine *European Data Economy* zu schaffen, die in ihrer Wertschöpfungskette wissenschaftliche, soziale und ökonomische Akteure durch den freien Zugang zu Daten verbindet. Dazu soll die Wissenschaft ihre Daten nach dem FAIR-Prinzip zugänglich machen. FAIR steht für findable, accessible, interoperable, reusable data.

AUTOREN

Gabriele Gramelsberger ist Professorin für Wissenschaftstheorie und Technikphilosophie an der RWTH Aachen.

Matthias Müller ist Professor für Hochleistungsrechnen und Leiter des IT Centers an der RWTH Aachen.

»Die Idee einer induktiven Wissenschaft ist nicht neu, doch irgendetwas scheint fundamental anders an der digitalen Induktion zu sein.«

Sowohl die Allianz der deutschen Wissenschaftsorganisationen als auch der Rat für Informationsinfrastrukturen (RfII) begrüßen diese Initiative. Das Potenzial für die Wissenschaft wird in der Vereinfachung im Umgang mit Forschungsdaten, in der Sicherung, Aufbereitung und Zugänglichkeit von Daten, in der Entwicklung und Etablierung fachspezifischer und fachübergreifender Dienste, Standards und Schnittstellen sowie in der Verankerung des Managements von Forschungsdaten im Forschungszyklus gesehen. Was jedoch die kommerzielle Verwendung von Forschungsdaten in einer *European Data Economy* für die Wissenschaft bedeuten wird, ist noch unklar.

I. Forschungsdaten – ein wertvolles Gut

Daten, so wird an der aktuellen Diskussion um *Open Science* und Forschungsdatenmanagement deutlich, sind ein wertvolles Gut, das es professionell zu managen gilt. Wird die Entscheidungshoheit über die Forschungsdaten auf wissenschaftspolitischem Terrain zu diskutieren sein, so stellt sich aus wissenschaftsreflexiver Perspektive die Frage, was die aktuellen Diskussionen über die Transformation von Forschung epistemisch wie forschungspraktisch aussagen. Dass es sich dabei um die Folgen der Digitalisierung, des enormen Anstiegs an Sensor-, Rechen- und Speicherressourcen und dadurch der exponenziellen Zunahme an Daten handelt, ist offensichtlich. Dass diese Mengen an Daten – Petabytes in der täglichen Großforschung (Big Data), Megabytes in den unzähligen, einzelnen Laboren tagtäglich (Small Data) – die Professionalisierung des Datenmanagements wie auch der Automatisierung der Datenanalyse erfordern, zeigt sich an den aktuellen Bemühungen. Doch was bedeutet das für die wissenschaftliche Erkenntnisgenerierung?

Prominent wurde die 2008 angestoßene Debatte um „correlation is enough." Statt theoriebasierter Forschung und kausalitätsbasierter Erklärungen sollte eine datengetriebene Wissenschaft Korrelationen in großen Datenmengen finden und daraus rein induktiv Hypothesen und Vorhersagen generieren. Die Idee einer induktiven Wissenschaft ist nicht neu, bildete sie doch den Kern der wissenschaftlichen Revolution der Neuzeit. Schon Johannes Kepler gewann durch jahrelange Berechnungen per Hand aus den Forschungsdaten von Tycho Brahe die elliptische Form der Planetenbahnen als auch die mathematische Grundlage seiner Theorie. Doch irgendetwas scheint fundamental anders an der digitalen Induktion zu sein. Das Mehr an Daten durch die Sensorisierung der Umwelt

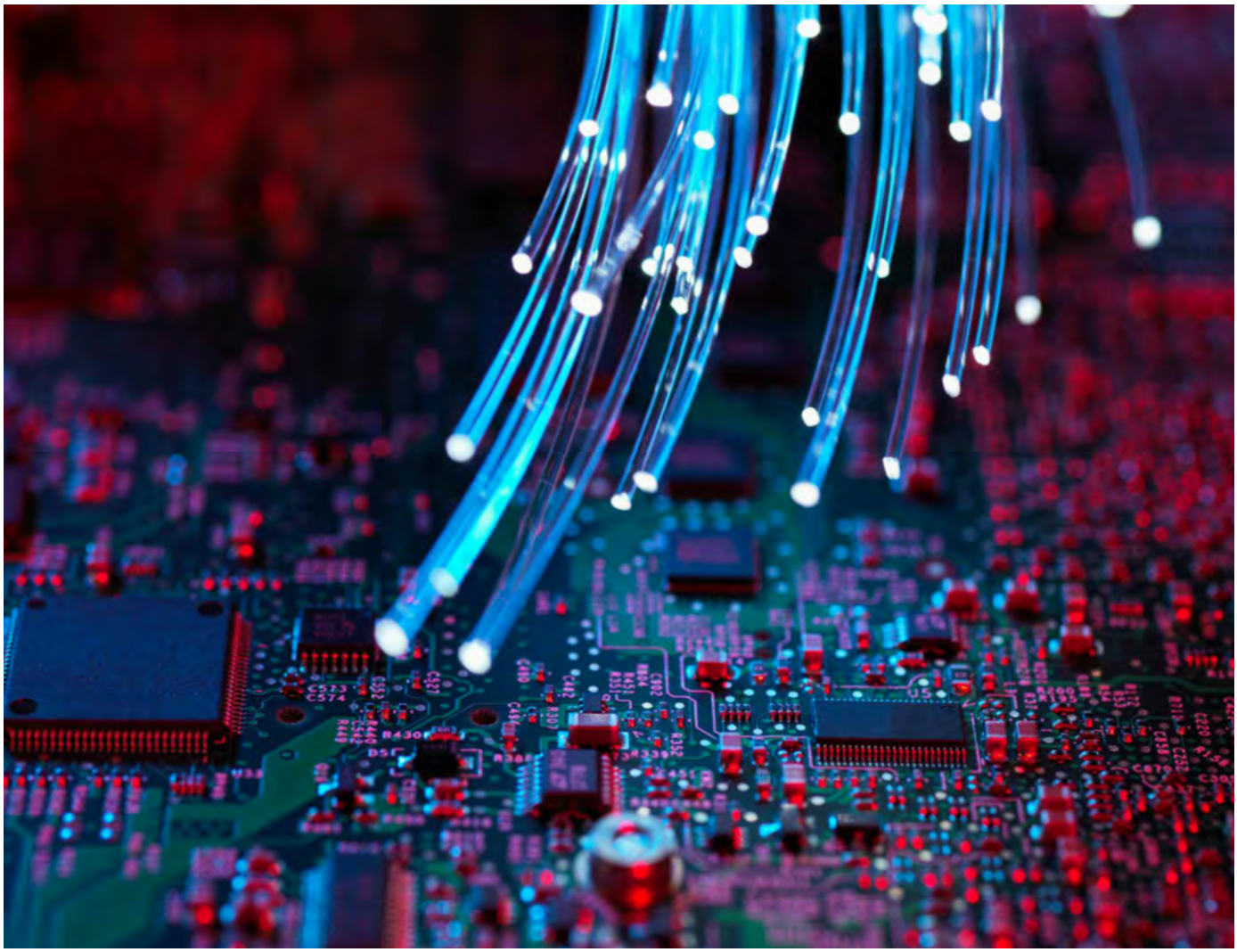

Foto: mauritius-images

ist jedoch nicht der einzig ausschlaggebende Punkt, sondern die Natur der Analyse verändert sich. Rechnete Kepler noch klassisch-symbolisch, so rechnen aktuelle Analyseverfahren anders: Sie erkennen Muster durch Vergleich, Klassifikation und Optimierung. Insbesondere Machine-Learning-Verfahren auf Basis von Künstlichen Neuronalen Netzen (KNNs) formalisieren menschliche Fähigkeiten der Informationsverarbeitung und wenden diese auf große Datenmengen an. Insofern ist es kein Zeichen von Güte, wenn ein zu großes KNN in den klassischen „brute-force"-Ansatz verfällt und jegliche Kombinationsmöglichkeit ausrechnet (Overfitting), statt zu lernen, Muster zu erkennen und einzuordnen.

Diese Analyseverfahren erschließen quantitative Lösungen neuer Problemklassen, auch wenn diese nicht mehr nachvollziehbar sind und aktuell viel Forschung in das Verständnis der Erkenntnisweise von KNNs fließt. Das bedeutet jedoch nicht, dass symbolisches Rechnen und „brute-force"-Ansätze obsolet wären. Im Gegenteil! Was neben der Datenanalyse die Anforderungen an Rechen- und Speicherressourcen in die Höhe treibt, sind Modell-basierte Prognosedaten in Form von Simulationsdaten. Ohne diese „in-silico Daten" oder „digital born data" würden Forschung und Entwicklung feststecken, denn Vorhersagewissen charakterisiert moderne Forschung und Entwicklung. Das Wetter von morgen, die Klimaerwärmung der kommenden Jahrzehnte, die Leistungsgrenzen von Motoren, die Strömungsdynamik neuer Designs, die Wirkung neuer Medikamente, die Vorhersage neuer Elementarteilchen, Moleküle und Materieeigenschaften gehören heute zur alltäglichen Routine wissenschaftlicher Erkenntnisgenerierung.

> »Ein unreflektierter Datenpositivismus birgt für Anwendungen ein gefährliches Fehlerpotenzial in sich.«

II. Big und Small Data

In der analytischen und prädiktiven Aussagekraft der Forschungsdaten liegt ihr epistemischer wie ökonomischer Wert. Dass diese analytische und prädiktive Aussagekraft mit zahlreichen Unsicherheiten behaftet ist, muss durch ein reflexives und transparentes Forschungsdatenmanagement deutlich gemacht werden. Denn in der Anwendung geht das Wissen um die Unsicherheiten schnell verloren. Ein unreflektierter Datenpositivismus birgt für Anwendungen jedoch ein gefährliches Fehlerpotenzial in sich und ergibt sich schon alleine aus der Generalisierung des Datenbegriffs. Nichts könnte jedoch heterogener sein als Daten. Sind beispielsweise die Myriaden an „sozialen Daten" tatsächlich Daten? Wie „The Parable of Google

Flu“ (*Science* 2014) zeigte, können Big-Data-Analysen nicht nur am Overfitting scheitern, sondern auch an der Adaptivität sozialer Daten: „When Google got flu wrong“ (*Nature* 2013). Die Frage nach der Reproduzierbarkeit von Daten betrifft jedoch auch wissenschaftliche Disziplinen wie die Psychologie, die Hirn- oder die Genomforschung. Ist alles – rein positivistisch gedeutet – ein Datum oder hält sich die Forschung an ihre altbewährten Wissenschaftskriterien? Die Frage ist nicht so leicht zu beantworten, denn die Natur der Datengenerierung verschiebt sich zunehmend in Richtung indirekter, heuristischer, probabilistischer und simulationsgestützter Datengenerierungsverfahren. Die Reduktion der Fehlerraten aktueller DNA-Sequenzierer beispielsweise wird nicht durch Verbesserung der Messungen, sondern durch Korrekturalgorithmen erzielt. Wendet man auf diese algorithmisch präzisierten Daten rein algorithmische Analyseansätze an, beispielsweise die indirekten Prognoseverfahren der funktionalen Geninterpretation, dann begibt sich die datengetriebene Forschung auf ein ähnlich putatives Terrain wie die approximative Simulation komplexer Systeme. Aus Fakten können schnell Datenartefakte werden. Doch im Unterschied zu den „digital-born“-Simulationsdaten, die empirisch evaluiert werden müssen, versteht sich die datengetriebene Forschung als empirisch.

Dies sind Probleme der automatisierten Generierung und Analyse von Hochdurchsatzdaten (Big Data). Was sich darüber hinaus beobachten lässt, ist, dass die neuen Analyseverfahren basierend auf KNNs zwar gut auf „virtuellen Daten“ funktionieren, also auf bereits digitalisierten Daten wie Bild- und Audiodaten, die an menschliche Wahrnehmungsfähigkeiten angepasst sind. Gut meint hier, dass es generalisierte Verfahren gibt, die für eine große Klasse an Problemen – Bilderkennung, Spracherkennung etc. – verwendet werden können und akzeptable Ergebnisse liefern. Der enorme Boom dieser Verfahren, gerade auch in der Automatisierung menschlicher Fähigkeiten, täuscht aber darüber hinweg, dass die Situation für „Real-Daten“, also klassische Messdaten im Labor, eine ganz andere ist. Hier gibt es bislang kaum generalisierbare Lösungen, denn die Heterogenität dieser Daten durch ihren Anwendungsbezug ist kaum zu überbieten. Unmengen an Small Data lagern in manuell oder proprietär gepflegten Speichersystemen in den Hochschulen und Forschungsinstituten vor Ort. Verfügt Big Data zunehmend über avancierte Infrastrukturen und standardisierte Analyseverfahren, so steht das Management dieser heterogenen Datenlandschaft am Anfang. Doch es ist zu vermuten, dass die Gesamtmenge der Small Data die der Big Data bei weitem übertrifft.

III. Forschungsdatenmanagement vor Ort

Wie also kann dieser Schatz an Small Data geborgen werden? Wird in Zukunft alles in der Cloud gespeichert oder in Forschungsdatenzentren (NFDI) ausgelagert? Oder ist der primäre Ort für das Management der Forschungsdaten die Hochschule oder Forschungsinstitution selbst? Sind es vor Ort die Bibliotheken oder Rechenzentren? Welcher Vorarbeiten bedarf es, Daten auszulagern und öffentlich zugänglich zu machen? Bereits eine sorgfältige Datenkuratierung stellen die Forscherinnen und Forscher vor große Herausforderungen.

»Aus Fakten können schnell Datenartefakte werden.«

Die Schnittstellen zwischen den verschiedenen „Datenakteuren“ – vor Ort, bundesweit, EU-weit – sind längst noch nicht definiert. Es bedarf also weit mehr als nur einfacher Datenverarbeitung. Dass hier Handlungsbedarf besteht, haben spätestens die zahlreichen Publikationen und Aufforderungen unter dem Stichwort „Forschungsdatenmanagement“ klar gemacht. Zwar gibt es an vielen Stellen Erfahrung im Umgang mit großen Datenmengen, doch selbst bei einer alleinigen Betrachtung der quantitativen Aspekte gibt es angesichts der absehbaren prinzipiellen Entwicklungsgrenzen heutiger Technologien das Kernproblem der Finanzierung der Datenspeicherung. Die Datenspeicherung für eine zukünftige Nachnutzung ist eine Zukunftsinvestition. Wie hoch diese sein muss, um den wissenschaftlichen Erkenntnisgewinn zu maximieren, kann im Moment niemand verlässlich beantworten. Wo immer sich Wissenschaftsgebiete einer datengetriebenen Herangehensweise zuwenden, muss die entsprechende Community diese Frage für sich beantworten. Bezogen auf das Forschungsdatenmanagement vor Ort, beispielsweise in Rechenzentren, könnte dies Folgendes heißen:

Das Rechenzentrum als Teilnehmer in Forschungsprojekten: Die Lösungen, die für Big Data in Forschungsprojekten erarbeitet wurden, lassen sich nicht einfach übertragen, um die Summe der Small Data Anforderungen zu bewältigen. Einerseits sind die erarbeiteten Lösungen oft zu spezifisch oder gar monolithisch, andererseits sind die Anforderungen und Ausgangslagen zu heterogen. Hier braucht es Forschungen zu einem geeigneten Datenmanagement, aber auch eine enge Zusammenarbeit zwischen Rechenzentren und Forschern vor Ort.

Das Rechenzentrum als Partner und Service Provider: Wo gut strukturierte Anforderungen und service-orientierte Rechenzentren zusammenkamen, wurden in der Vergangenheit praktikable und nachhaltige Lösungen entwickelt. Die aktuellen Herausforderungen bestehen nicht nur darin, diese Lösungen weiterzuentwickeln, sondern vor allem darin, dass die Anforderungen nicht klar definiert sind. Das Erarbeiten der Anforderungen gemeinsam mit dem Anwender, die Integration in eine dynamische Forschungs- und Infrastrukturlandschaft und die Anpassung an sich verändernde Ziele erfordern neue, ungewohnte Herangehensweisen. Das setzt auf Seiten der Rechenzentren nicht nur das entsprechend geschulte Personal und Personalmanagement voraus, sondern auch das Bewusstsein bei allen Akteuren, dass man als Partner gemeinsam die Ziele definiert und die Lösungen erarbeitet.

Neben der neuen, intensiveren Zusammenarbeit mit den Anwendern ist es erforderlich, dass die Rechenzentren die Zusammenarbeit untereinander intensivieren. Nur so lassen sich die Standards und Schnittstellen erarbeiten und definieren, die für eine Realisierung einer internationalen und fächerübergreifenden Forschungslandschaft notwendig sind. Ein solches reflexives und transparentes Forschungsdatenmanagement erfordert ein Qualitätsmanagement und eine Fehlerkultur, die bisher nur an wenigen Stellen gelebt wird.