Das Ökosystem der Datenwissenschaften

Thomas Ludwig · Wolfgang E. Nagel
Ramin Yahyapour

Data Analytics umfasst einen wichtigen Teil im Lebenszyklus von Daten und ist damit strategischer Bestandteil des Aufgabenfeldes im Umfeld von Data Science. Gleichzeitig stützt sich Data Analytics auf Infrastrukturen zur Datenspeicherung, die aus Hardware und Software gebildet werden. Eine zusätzliche Wertschöpfung durch Daten erfolgt mittels einer Verteilung an Nutzer, die aus ihnen weitere Erkenntnisse gewinnen können. Der folgende Artikel beschreibt exemplarisch auf der Basis von ausgewählten Beispielen wichtige Aspekte dieses Themas.

Der Datenlebenszyklus

Die Daten unserer technisierten Welt unterliegen einem Lebenszyklus, den wir in vier Phasen einteilen können. Daten werden erzeugt und analysiert. Sie werden archiviert und verteilt. Die einzelnen Phasen unterstützen verschiedene Zielsetzungen, die in unterschiedlichen Wertschöpfungsszenarien von variierender Bedeutung sind (Abb. 1).

Beispielsweise erfolgt eine Archivierung zum Zwecke der Nachvollziehbarkeit bestimmter Auswertungen. In der Wissenschaft ist es eine Anforderung durch die Regeln der Guten Wissenschaftlichen Praxis, die in Deutschland die Aufbewahrung regelt [3]. In der pharmazeutischen Industrie sind die Vorgaben zur Archivierung gesetzlich geregelt. Archivierungskonzepte müssen auch darstellen, wann Daten zwingend zu löschen sind. Personenbezogene Daten unterliegen hier genauen Vorgaben. Darüber hinaus werden Daten aber auch langfristig gespeichert, z. B. weil ein sogenannter Dauerwert festgestellt wird. Die betrifft z. B. Daten von geschichtlicher Bedeutung oder von gesellschaftlichem Wert. Viele Datensätze werden aber auch schlicht „vergessen“ und bei den stark wachsenden Datenmengen stellen sie nach einiger Zeit auch keinen ökonomischen oder organisatorischen Faktor mehr dar, der besonderer Aufmerksamkeit bedürfte.

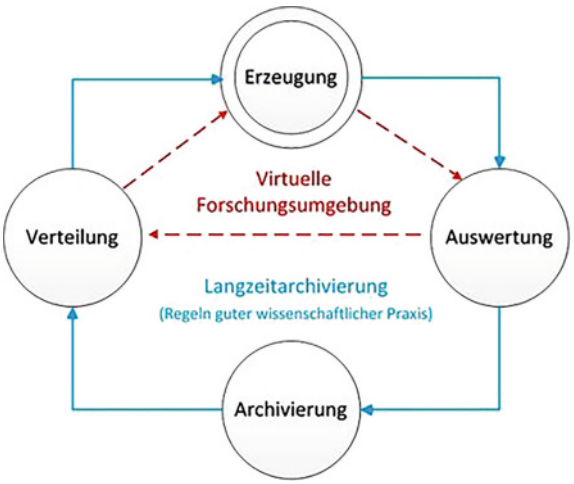

Abb. 1 Der Datenlebenszyklus umfasst vier Stufen, die in verschiedene Wertschöpfungskreisläufe eingebettet sind

https://doi.org/10.1007/s00287-019-01219-5

Thomas Ludwig
Deutsches Klimarechenzentrum und Universität Hamburg
E-Mail: ludwig@dkrz.de

Wolfgang E. Nagel
Zentrum für Informationsdienste und Hochleistungsrechnen und Technische Universität Dresden

Ramin Yahyapour
Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen und Universität Göttingen

Die Verteilung der Daten dient einer Nachnutzung durch Dritte. Die Daten müssen bestimmten Anforderungen genügen, die vom Nutzungszweck abhängen. Je längerfristiger und weitreichender eine Nachnutzung erwünscht ist, desto aufwendiger müssen die Daten durch Metainformationen annotiert werden, um diese Zwecke zu unterstützen. Unsere modernen Zeiten der Datenorientierung sehen Wertschöpfungen, die Zeiten und Anwendungsbereiche übergreifen. Als Beispiel seien historische Logbücher von Segelschiffen aus vergangenen Jahrhunderten genannt, deren Eintragungen als Informationen für modere Klimasimulationen verwendet werden können – nachdem man sie aufwendig in maschinenlesbare Form überführt hat [4].

Die Erzeugung von Daten erfolgt auf vielfältige Weise. Sie können als Ergebnisse von Computersimulationen als sogenannte „digital born data" entstehen. Beispiele sind hier Windkanalsimulationen in der Autoindustrie, Klimasimulationen in der Wissenschaft oder Finanzsimulationen in der Wirtschaft. Die zweite große Erzeugerquelle ist jegliche Art technischer Apparaturen. Gensequenzierer, Mikroskope und Teleskope sind Beispiele für typische Großerzeuger von Daten. Aber auch die moderne Gesellschaft mit ihren vernetzten Geräten produziert steigende Datenmengen durch z. B. Stromzähler, Überwachungskameras und vieles mehr. Die Gegebenheiten im Internet-der-Dinge (IOT) werden hier neue Randbedingungen schaffen. Eine weitere wichtige Datenquelle sind heute auch unsere Aktivitäten in sozialen Netzen, unsere Likes und Tweets, die für vielfältige Auswertungen genutzt werden.

Die Methoden und Anwendungsgebiete der Auswertung von Daten bilden den Kern des vorliegenden Themenheftes. Wir wollen uns hier auf die Frage der Forschungsinfrastrukturen konzentrieren und darstellen, wie Daten erzeugt, analysiert und ausgewertet, archiviert und verteilt werden.

Die stetigen Fortschritte in der halbleiterbasierten Technik der letzten Jahrzehnte haben dabei zu exponentiell ansteigenden Datenmengen geführt, mit denen die Datenhaltungsinfrastrukturen nur schwer mithalten konnten. Unsere Fähigkeiten und Methoden zur sinnvollen Verwaltung der Daten haben sich leider nur wenig weiterentwickelt. Beide Aspekte, Datenhaltung und Datenverteilung, sind Gegenstand aufwendiger Forschungs- und Entwicklungsarbeiten, die hier exemplarisch dargestellt werden. Nur wenn jede Phase des Datenlebenszyklus gleichwertig unterstützt wird, verwandeln sich Daten in Information, die zu gesellschaftlichem – und häufig auch wirtschaftlichem – Wert führt. Betrachten wir zunächst ein Beispiel für eine Forschungsinfrastruktur, die sich auf Datenhaltung und -verteilung konzentriert.

Klimamodellierung am Deutschen Klimarechenzentrum

Das Deutsche Klimarechenzentrum (DKRZ) ist eine nationale Forschungseinrichtung, die die Gemeinschaft der Klimaforscher und Forscher angrenzender Gebiete unterstützt. Seine Aufgabe besteht in der Bereitstellung hoher Rechenleistung zur Ausführung von Klimasimulationen und umfassender Systeme zur Speicherung und Archivierung von Daten. Das Datenmanagement, insbesondere die Archivierung und Verteilung, werden durch geeignete Dienstleistungen und spezielle Software-/Hardwaresysteme unterstützt.

Das DKRZ unterstützt den gesamten Datenlebenszyklus, wobei es sich von der Natur der Daten um „digital born data" handelt. Nahezu alle am DKRZ gelagerten Daten sind Ergebnisdaten von Klimasimulationen. Beobachtungsdaten von Satelliten und Messdaten sind nur zu einem kleinen Anteil auf den Systemen abgelegt. Klimasimulationen arbeiten mit sehr kleinen Eingabedatenmengen und erzeugen große, potenziell beliebig große Ausgabedatenmengen. Dies ist konträr zu Wetterberechnungen, die mit einer großen Datenmenge aus Geräten als Eingabedatenmengen starten und eine überschaubare Menge von prognostizierten Wetterdaten erzeugen.

Es sollen hier die Aspekte der Archivierung und Verteilung dargestellt werden. Die Fragen der visuellen Auswertung von Daten wurden im Artikel von Röber und Böttinger im vorliegenden Heft ausführlich diskutiert. Das DKRZ wurde 1987 als unabhängige Einrichtung gegründet und hat stets als besonderes Alleinstellungsmerkmal die Größe seines Festplattensystems und Bandarchivs gehabt. Insofern war das DKRZ von Anfang an eine Art Cloudservice-Provider. Die Forscher haben umfangreiche Datensätze mit den Hochleistungsrechnern erzeugt, diese dort auf den Platten und den Bändern gespeichert, sie vor Ort ausgewertet und für eine potenzielle Weiterverwendung bereitgestellt. Selten werden umfangreiche Datenmengen an einen anderen Ort außerhalb des DKRZ umkopiert. Jedoch

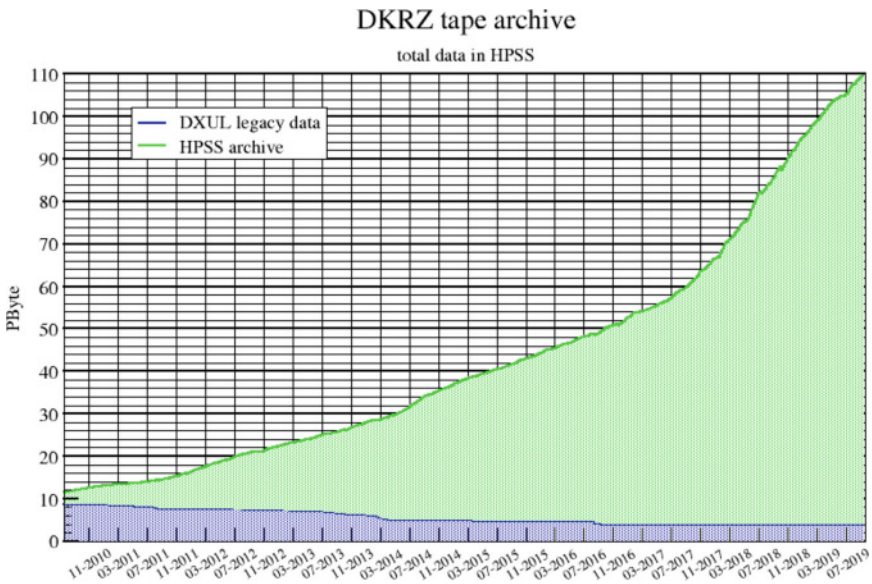

Abb. 2 Jährliches Anwachsen der Daten im HPSS-Bandarchiv. Man erkennt den Wechsel auf ein stärkeres Rechnersystem in den Jahren 2015/16

führt das DKRZ in zunehmendem Maße Simulationsdaten, die außerhalb des DKRZ gerechnet wurden, zur weiteren Analyse am eigenen Standort zusammen.

Die umfassende Bereitstellung von Systemen und Diensten stützt sich dabei auf eine Netzanbindung mit nur 2×10 Gbit/s, die noch für alle Nutzungsszenarien ausreichend ist. Das DKRZ startete 1988 mit einem CDC-205-System mit einer Rechenleistung von 0,2 GFlop/s – einer Leistung, die heute bereits von Smartphones des unteren Preisspektrums vielfach übertroffen wird. Seit 2015 betreibt das DKRZ einen Supercomputer von Bull/Atos mit 3,6 PFlop/s Spitzenleistung und einem Festplattensystem von 56 PByte. Der für 2020 vorgesehene Ausbau des Systems wird die Rechenleistung bei einer angenommenen Verdoppelung des Speichervolumens etwa vervierfachen. Damit gehört das DKRZ zu den absolut gesehen am höchsten mit Speicher versehenen Hochleistungsrechenzentren.

Die Notwendigkeit hierfür ergibt sich aus Sicht der Wissenschaft aus den Abläufen zur Erkenntnisgewinnung: Simulationen für globale Klimamodelle erzeugen hohe Datenvolumina durch den großen betrachteten Raum, und gleichzeitig wird über einen langen Zeitraum simuliert. 100 Jahre und mehr sind häufig für den Erkenntnisgewinn notwendig. Die Auswertung der Daten erfolgt zeitverzögert zur Erzeugung, aber eine Nachnutzung durch Dritte kann über Jahre und Jahrzehnte erfolgen. Dies erfordert insbesondere am DKRZ ein großes Datenarchiv. Man greift hier auf die bewährte Technik der Magnetbänder zurück. Die Bandbibliotheken am Zentrum bieten 65.000 Stellplätze für Bandkassetten. Versehen mit LTO-7-Kassetten, die 6 TByte-Daten speichern können, beträgt die Kapazität des Archivs knapp 400 PByte. In den letzten 10 Jahren konnte dabei eine Kapazitätserweiterung stets problemlos durch einen Übergang auf eine neue Kassettengeneration realisiert werden. Die Hardware der Bibliotheken wurde weitergenutzt, allerdings müssen jeweils die Laufwerke erneuert werden. Das aktuell gespeicherte Datenvolumen beläuft sich auf ca. 120 PByte, wobei allein im letzten Jahr ca. 40 PByte an Archivdaten aufgenommen wurden. Mit dem Vorgängermodell des Rechners wurden 8 PByte pro Jahr archiviert, das aktuelle System ist für knapp die 10-fache Rate ausgelegt, da auch das Festplattensystem etwa 10-mal so groß sein wird (siehe Abb. 2). Betriebswirtschaftlich belastet das Bandarchiv den Haushalt des DKRZ mit mehreren hunderttausend Euro jährlich – im Vergleich zu den mehr als zwei Millionen Stromkosten ein kleinerer Kostenfaktor. Daten vorheriger Rechnergenerationen sind ökonomisch schnell irrelevant: Die gesamte archivierte Datenmenge des Vorvorgängersystems ist aktuell nur einige wenige Prozent des aktuellen Systems. Trotzdem werden auch alte Daten immer wieder gelöscht, was man am langsamen Absinken der Datenmengen im ehemaligen DXUL-Archiv erkennen kann (Abb. 2).

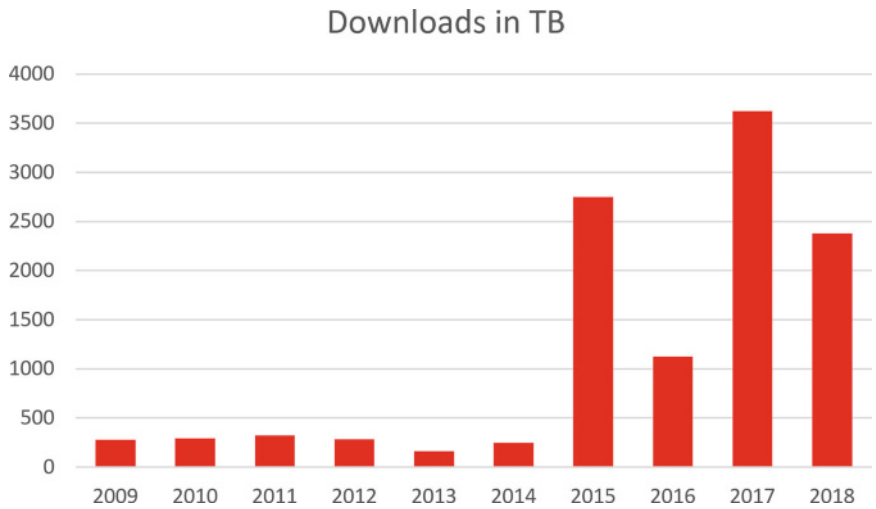

Abb. 3 Jährliches Volumen der aus dem WDCC am DKRZ weltweit heruntergeladenen Daten

Die Verteilung von Klimasimulationsdaten an nationale und internationale Nutzer erfolgt am DKRZ im Wesentlichen auf zwei Wegen: Das DKRZ betreibt ein Langzeitarchiv, das im Rahmen des *World Data Systems* die Rolle des *World Data Center Climate* (WDCC) einnimmt [5]. Weiterhin integriert es seine Daten in die Infrastruktur der *Earth System Grid Federation* (ESGF) [6]. In beiden Umgebungen werden Daten bereitgestellt, die aufwendig kuriert und qualitätsgesichert wurden. Allgemein gesprochen können Wissenschaftler auf räumliche und zeitliche Ausschnitte der Ergebnisdaten (meist) globaler Klimasimulationen zugreifen und diese Datensätze für die eigene wissenschaftliche Wertschöpfung verwenden. Die über Programmier- und Webschnittstellen ausgelieferten Daten sind überwiegend auf den Bändern der Bibliothek langzeitarchiviert.

Das WDCC stellt Daten auch über einen Zeitraum von weit über 10 Jahren bereit. Entsprechend sorgfältig erfolgt die Pflege der physikalischen Daten am Ort. Es handelt sich hier um den einzigen Datensatz, der am DKRZ mit einer örtlich entfernt gelagerten Sicherungskopie versehen ist. In Würdigung der am DKRZ etablierten Prozesse wurde das WDCC 2018 mit dem Zertifikat des Core Trust Seal für vertrauenswürdige Archive ausgezeichnet. Als abschließende Dienstleistung in diesem Zusammenhang ermöglicht das DKRZ die Zuordnung von DataCite DOIs („digital object identifiers“) zu Datensätzen, um diese in wissenschaftlichen Zusammenhänge zitierfähig zu machen. Dies sichert den Autoren von Datensätzen die Urheberschaft und ggf. die Reputation im Veröffentlichungsprozess. Die Gesamtheit der Prozesse inklusive der Zuordnung der DOIs orientiert sich an den FAIR-Prinzipien für Wissenschaftsdaten und zeichnet diese als „findable“, „accessible“, „interoperable“ und „reusable“ aus.

ESGF ist eine internationale Zusammenarbeit unter der Leitung eines Scientific Steering Board und eines Executive Committee. Es wird geleitet vom DKRZ und dem LLNL (USA). Die Finanzierung teilt sich auf in institutionelle Beiträge und zwei Großprojekte vom Department of Energy (DOE) der USA und der Europäischen Union. ESGF hat zum Ziel, eine Forschungsinfrastruktur zur Verwaltung, Analyse und Verteilung von Simulations- und Beobachtungsdaten dauerhaft bereitzustellen und weiterzuentwickeln. Der Erkenntnisgewinn in den Erdsystemwissenschaften soll vereinfacht und verbessert werden. ESGF lagert dafür die Daten in verschiedenen, global verteilten sogenannten Datenknoten. Das DKRZ gehört zu den Gründungsmitgliedern der ESGF und betreibt am Standort einen der zentralen europäischen ESGF-Knoten.

Allein über das Langzeitarchiv des DKRZ wurden 2018 mehr als 2 PByte an Daten an Nutzer weltweit ausgeliefert (siehe Abb. 3). Die Wirksamkeit koordinierter Datenspiegelungen an verschiedenen Standorten erkennt man z. B. am hohen Auslieferungsvolumen 2015, das in der Folge eines Ausfalls von Servern in den USA und weltweit entstand. Gleichzeitig kann man feststellen, dass die Archivsysteme auch eine gewisse Resilienz gegen politisch motivierte Beeinflussungen aufweisen.

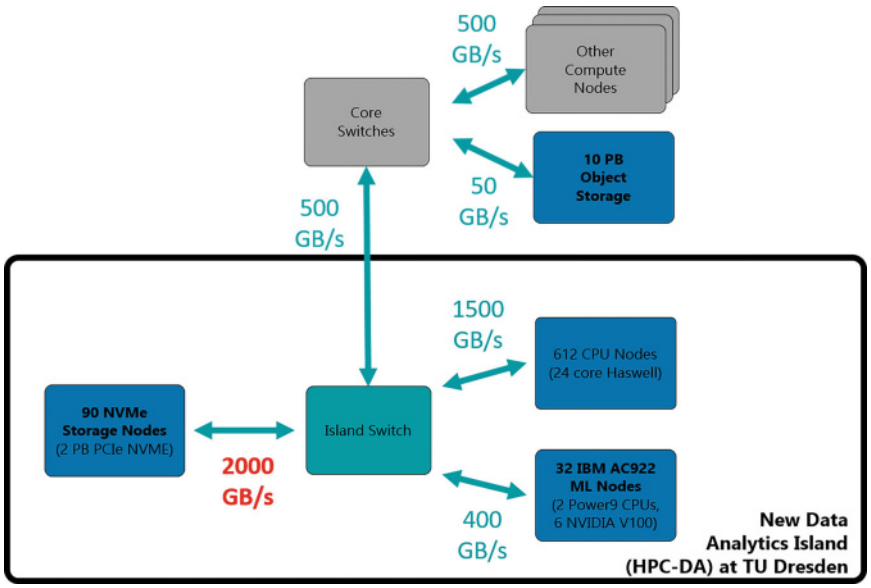

Abb. 4 An der TU Dresden implementierte HPC-Data Analytics Infrastruktur zur Auswertung von sehr großen Datenmengen

Das DKRZ betreibt somit das Datenökosystem für die deutsche Klimamodellierergemeinschaft und angrenzende Wissenschaftsbereiche. Die angebotenen Dienste sind auf den Erkenntnisgewinnungsprozess der Klimaforscher abgestimmt und decken den gesamten Lebenszyklus von Daten ab. Die Prozesse und Daten sind in internationale Verbünde integriert und unterstützen virtuelle Forschungsumgebungen auf der ganzen Welt. Ein zweites Beispiel einer Forschungsinfrastruktur erläutert spezifische Konzepte zur Datenhaltung und -analyse.

Infrastruktur und Methoden zu Data Analytics

Für die Datenwissenschaften haben momentan die Analyse- und Auswertungsmethoden der Künstlichen Intelligenz (KI) neuer Art – insbesondere Machine Learning und Tiefenlernen (Deep Learning) – einen sehr hohen Stellenwert. Der dramatische Fortschritt in der KI wurde in den letzten Jahren hauptsächlich durch zwei Faktoren bestimmt. Erstens nutzt das Maschinelle Lernen – und insbesondere das Deep Learning – Massendaten (Big Data), um Modelle für die Klassifizierung und Vorhersage zu entwickeln, die so genau sind wie die menschlichen Fähigkeiten oder sogar darüber hinaus gehen. Die zweite treibende Kraft ist die immer noch exponentielle Zunahme – getrieben durch Moore's Law – einer sehr hohen, effektiven Rechenleistung, entwickelt im Kontext des wissenschaftlichen Hochleistungsrechnens (High Performance Computing, HPC) und insbesondere getrieben durch die Hardwareentwicklung grafischer Ko-Prozessoren (GPUs). Derartige Karten waren lange Zeit ausschließlich im Umfeld der Spieleentwicklung im Consumer-Markt einsetzbar, haben aber in den letzten zehn Jahren breiten Eingang in das HPC gefunden. Durch die weiterhin rasante Entwicklung in der Mikroelektronik werden inzwischen GPU-Einzelkarten angeboten, deren Rechenleistung noch vor weniger als 15 Jahren durch mehr als 10 Racks konventioneller Rechentechnik erzeugt werden musste. Strategisch sind also viele der schnellen Erfolge in der KI durch diese Hardwareentwicklungen – einhergehend mit der Entwicklung von Softwaremethoden und Frameworks aus dem HPC-Umfeld und der Verfügbarkeit großer Datenmengen – erst möglich geworden.

Für die Auswertung großer Datenmengen benötigen diese schnellen Infrastrukturen eine balancierte Rechnerarchitektur, getragen von einer hohen Kommunikations- und I/O-Leistung (Ein-/Ausgabe). Sind in üblichen Anwendungsvarianten IO-Bandbreiten von einigen 10 GByte/s durchaus ausreichend, sind in diesem neuen Nutzungsumfeld für das HPC sehr viel höhere I/O-Leistungen wichtig. Exemplarische sei hier die an der TU Dresden neu aufgebaute Data-Analytics-Infrastruktur HPC-DA (Abb. 4) gezeigt, die nach einem bereits vielfach bewährten Inselkonzept realisiert ist und sich damit von der klassischen HPC-Welt abgrenzt. Eine moderne NVMe-Knoten-Infrastruktur (I/O-

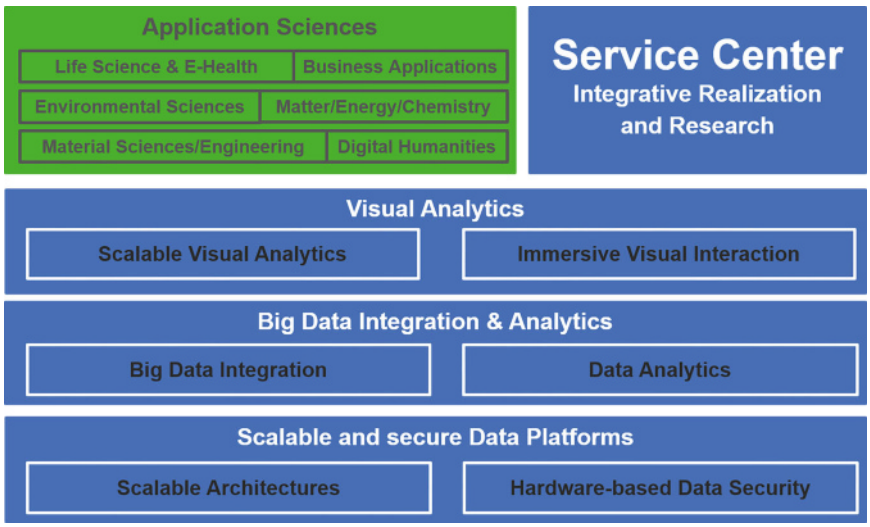

Abb. 5 Aufbau, Struktur und Fokusbereiche des Big-Data-Kompetenzzentrums ScaDS Dresden/Leipzig

Bandbreite etwa 2 TB/s) bilden den Kern, um die sich eine wassergekühlten ML-Infrastruktur auf der Basis IBM Power9 AC922 (192 Nvidia V100) und ein Staging-Bereich von 10 PB Object Storage gruppiert. Die TU Dresden stellt dieses Gesamtsystem – im Rahmen der Gauß-Allianz [2] – nach wissenschaftsgeleitetem Antrags- und Begutachtungsverfahren deutschlandweit als leistungsfähige Datenanalyseinfrastruktur für Forschung und Wissenschaft kostenfrei zur Verfügung.

Diese Architektur zeigt auch, in welcher Weise die spezialisierten Systeme für HPC und Big Data inhaltlich zusammenwachsen und so die für die Anwendungsszenarios jeweils notwendigen Komponenten kosteneffizient bereitstellen können. Gestützt wird dies durch einen nutzerzentrierten Software-Stack, der neben den etablierten Batch-Jobs, Scheduling-Mechanismen und einer HPC-Software-Umgebung mit klassischen numerischen Paketen auch Möglichkeiten zur Virtualisierung, eine Vielzahl von üblichen Big-Data- und ML-Frameworks (Hadoop, Flink, Sparc, Keras, Tensorflow, Caffee und vieles mehr) sowie einfache Zugangsmöglichkeiten – auch interaktiv – auf Teilsysteme bereitstellt. Am Standort Dresden wird dies gestützt durch die gemeinsam mit Partnern aus Leipzig erbrachte Beratungsleistung im Rahmen des Big-Data-Kompetenzzentrums ScaDS Dresden/Leipzig (Scalable Data Services and Solutions), das neben der strategischen Weiterentwicklung von Methoden – sowohl im Fachumfeld als auch im fokussierten Anwendungsspektrum – zusätzlich in signifikantem Umfang über das Servicezentrum Beratungsleistung für die deutsche Wissenschaftslandschaft bereitstellt (Abb. 5).

KI-Technologien dringen inzwischen bereits tief in die meisten Geschäftsfelder ein und beeinflussen unseren Alltag. Dennoch wurden in der Mehrzahl der Fälle bisher nur die niedrig hängenden Früchte geerntet, was auf eine Reihe von noch offenen Herausforderungen zurückzuführen ist, die es zu lösen gilt. Im Gegensatz zum klassischen maschinellen Lernen will die KI Probleme lösen, Muster identifizieren, mit den Nutzern interagieren und in der Lage sein, zu erkennen und zu verstehen. Um dies zu erreichen, benötigt die KI Zugang zu qualitativ hochwertigen Daten und formalisiertem Wissen. Mit der Kombination aus Forschung zur Wissensakquisition, -repräsentation und Grundlagenforschung zu KI-Methoden werden in den nächsten Jahren und sicher bedeutende Fortschritte bei wissensbasierten Methoden der künstlichen Intelligenz erzielt werden. Darüber hinaus müssen KI-Methoden systematisch in wissenschaftliche Analyse-Workflows eingebunden werden, die den Forschungsfortschritt in vielen anderen Forschungsbereichen beschleunigen können. Die Datenanalyse erfordert zunehmend hochinteraktive und iterative datengesteuerte Workflows von Trial-and-Error, die Zwischenergebnisse überprüfen und die Analyse in einem geschlossenen Regelkreis anpassen. Im Geschäftsleben muss die KI in die Entwicklung von Produktdesigns, Dienstleistungen und Geschäftsmodellen eingebunden werden. Es besteht auch ein großer Bedarf an Ver-

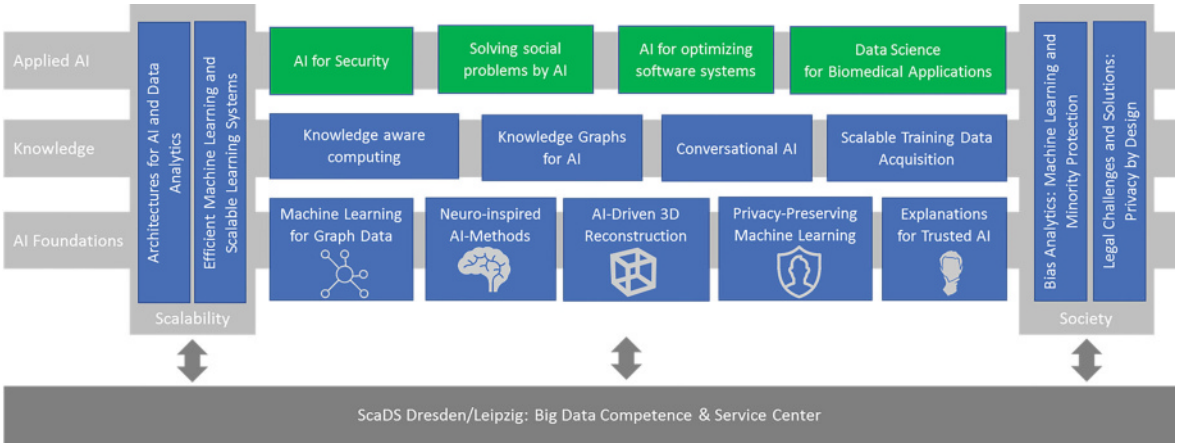

Abb. 6 Struktur und Forschungsbereiche des KI-Kompetenzzentrums ScaDS.AI Dresden/Leipzig

trauen, Transparenz und Rückverfolgbarkeit von KI-getriebenen Entscheidungen und Prozessen. Schließlich ist die Wahrung des Rechts auf Privatsphäre und die informationelle Selbstbestimmung der Bürger noch immer ein weitgehend ungelöstes Thema.

Zur Adressierung dieser Felder ist nach einem Begutachtungsprozess die Entscheidung gefallen, das Big-Data-Kompetenzzentrum ScaDS Dresden/Leipzig um Forschungs- und Entwicklungsbereiche zur KI-Forschung zu erweitern und mit ScaDS.AI Dresden/Leipzig (Abb. 6) mit erheblich erweitertem Personal und neuen Professuren diese Herausforderungen in den Datenwissenschaften forschungstechnisch und wissenschaftlich zu begleiten und damit Lösungen inhaltlich voranzutreiben. Der einfache Zugang zu leistungsfähigen Infrastrukturen, modernste Algorithmen, Methoden und Verfahren zur Datenauswertung und Wissensgenerierung sowie die Verfügbarkeit von Fachpersonal, das diese Entwicklungen sowohl in KMUs als auch in Leitindustrien einbringt und nutzbar macht, sind strategische Voraussetzungen für eine erfolgreiche wirtschaftliche Weiterentwicklung des Industriestandortes Deutschland.

Informationsinfrastrukturen

Da Wissenschaft zunehmend kollaborativ arbeitet, z. B. als Verbundforschung, wächst die Bedeutung von standortübergreifende Strukturen und deren Interoperabilität. Die Etablierung von geeigneten Standards in einem fachdisziplinären Ökosystem mit gemeinsamen Formaten und Diensten stellt dabei eine große Herausforderung dar. Die zuvor genannte Klimaforschung ist im Vergleich zu anderen Forschungsbereichen bereits weit fortgeschritten und nicht typisch. Die Klima-Community besitzt inzwischen zahlreiche gut akzeptierte Dateiformate, Tool-Bibliotheken und Infrastrukturen wie das DKRZ. Dies ist jedoch in den einzelnen Wissenschaften sehr unterschiedlich weit ausgeprägt. Hier sind insbesondere die Forschungsgebiete weiter fortgeschritten, die auf Datenaustausch und gemeinsame Strukturen aufgrund ihrer Natur zwingend angewiesen waren und sind, bspw. Teilchenphysik, Astrophysik oder eben die Klimaforschung. Andere Gebiete erkennen erst die Vorteile und Möglichkeiten bei der Nachnutzung von Daten und stehen am Beginn dieser Entwicklung. Auch durch Empfehlungen oder Vorgaben der Fördermittelgeber erlangt dies eine größere Sichtbarkeit. Die DFG hat gerade erst im Juli 2019 die „Leitlinien zur Sicherung guter wissenschaftlicher Praxis“ neu beschlossen. Auch hier findet sich bei den jeweiligen Leitlinien zu Methoden und Standards, der Dokumentation und der Herstellung von öffentlichem Zugang zu Forschungsergebnissen Elemente, die die Ausprägung eines entsprechenden Ökosystems mit geeigneten Standards nahelegen.

Da es zahlreiche Disziplinen gibt, die bisher noch wenig an gemeinsamen Strukturen ausgeprägt haben, stellt sich die Frage, wie dies unterstützt werden kann. Teilweise fehlen geeignete Governancestrukturen, um wissenschaftsgetrieben gemeinsame Standards zu definieren. Oder es fehlen die notwendigen Finanzierungsmechanismen, um Infrastrukturen langfristig zu etablieren. Oder es erfordert die notwendigen Anreize, um

Daten zu sichern und bereitzustellen. Die Gemeinsame Wissenschaftskonferenz (GWK) hat hierzu im November 2013 den Rat für Informationsinfrastrukturen (RfII) eingerichtet, um Politik und Wissenschaft bei der Weiterentwicklung der Informationsinfrastrukturen zu beraten. Die GWK folgte damit einer Empfehlung des Wissenschaftsrates aus dem Jahr 2012, in der ein koordiniertes Vorgehen bei der Weiterentwicklung dieser Strukturen gefordert wurde [10]. Der RfII hatte hierzu in einer ersten Phase bereits bestehende Strukturen analysiert und festgestellt, dass die häufig betriebenen Bottom-up-Entwicklungen wenig oder nicht nachhaltig in der Finanzierung angelegt waren und meist keine langfristig stabilen Strukturen ausprägten. Auf der anderen Seite neigten Top-down-Ansätze zu geringer Akzeptanz unter den Wissenschaftlerinnen und Wissenschaftlern. In seiner Empfehlung „Leistung aus Vielfalt" hat der RfII im Mai 2016 zahlreiche Maßnahmen vorgeschlagen, um bessere Bedingungen für Management für Forschungsdaten zu schaffen [8]. Hierzu gehört insbesondere die Umstellung der Fördermechanismen auf eine langfristige Finanzierung von Infrastrukturdiensten. Das heißt, es muss eine transparente Perspektive für eine mögliche Verstetigung von relevanten Diensten gegeben werden.

Als konkrete Ausprägung wurde eine die Einrichtung einer Nationalen Forschungsdateninfrastruktur (NFDI) vorgeschlagen, der die GWK im November 2018 mit dem Beschluss gefolgt ist, eine entsprechende Förderlinie ab 2020 zu etablieren. Gemäß GWKGBeschluss sollen in der NFDI „Datenbestände in einem aus der Wissenschaft getriebenen Prozess systematisch erschlossen, langfristig gesichert und entlang der FAIR-Prinzipien über Disziplinen- und Ländergrenzen hinaus zugänglich gemacht werden". Die Umsetzung erfolgt aktuell durch die DFG und hat in einem ersten Schritt über 50 Absichtserklärung von Konsortien geliefert, die in den Jahren 2019 bis 2021 eine Beantragung planen. Der Aufbau der NFDI wird sukzessive erfolgen. Der Dynamik der Wissenschaft folgend ist die NFDI als lernendes System angelegt, um sich in Förderformaten und Beteiligten den Bedarfen anzupassen.

Die so ausgeprägten Informationsinfrastrukturen können nicht losgelöst und völlig eigenständig betrachtet werden. So gibt es in vielen Bereichen bereits etablierte Strukturen, die zu berücksichtigen und sinnvoll einzubinden sind. So wird es nicht überraschend sein, dass langfristig die Erschließung und Zugänglichkeit von Datenbeständen in vielen Bereichen nicht von Diensten zur Erzeugung und Analyse von Daten zu trennen sein wird. Teilweise werden die Datenbestände auch zu umfänglich sein, um diese trotz immer leistungsfähigeren Netzwerkinfrastruktur und günstigerer Datenspeicher frei zu bewegen. Auch wird es ökonomisch nicht sinnvoll oder möglich sein, an allen Orten die benötigten Rechner- und Speichersysteme oder Forschungsgeräte vorzuhalten. Die Lizenzierung von kommerziellen Informationen wird hier ebenso berücksichtigt werden müssen. Diese Infrastrukturen werden daher in Bezug auf übergreifende Informations- und Infrastrukturknoten bzw. -hubs z. B. bei Rechenzentren, Bibliotheken, Forschungsdatenzentren in mehreren Dimensionen verbunden sein. Daher werden die verschiedenen fachdisziplinären Infrastrukturen ein vernetztes Ökosystem darstellen, in dem verschiedene Leistungserbringer, Datenerzeuger und -nutzer zusammenarbeiten.

Ebenso ist offensichtlich, dass diverse Dienste nicht disziplinärspezifisch ausgeprägt sein müssen bzw. sollten, um eine Vernetzung zwischen diesen Infrastrukturen zu vereinfachen. Offensichtliche Beispiele finden sich bei Fragen zur Authentifizierung und Autorisierung von Nutzern (AAI, Föderiertes Identitätsmanagement), der Referenzierung von Daten (Persistente Identifizierungsdienste) oder Standards zur Metadatenbeschreibung (Metadata Standard Registries). Auch wird es einen rechtlichen Rahmen für die Definition von Nutzungsbedingungen geben müssen (Use and Access Policies, Consent Modelle). In der NFDI wird dies aktuell unter dem Begriff der Research Data Commons bzw. als horizontale, konsortienübergreifenden Dienste diskutiert. Im internationalen Raum liefert RDA einen geeigneten Rahmen, um sich diesbezüglich abzustimmen.

Da viele hochinteressante Forschungsthemen gerade interdisziplinär an Grenzflächen der klassischen Disziplinen stattfinden, ist die Vernetzung und Durchlässigkeit zwischen diesen Informationsstrukturen sowohl national als auch international von großer Bedeutung. Neben den nationalen Bestrebungen, wie sich dies mit der NFDI in Deutschland oder ähnlichen nationalen Initiativen in anderen Ländern findet, muss dies international gedacht werden. Es finden sich hierzu Aktivitäten in

disziplinbezogenen Standardisierungsgremien oder in übergreifenden Ansätzen wie in der Research Data Alliance (RDA). Ebenso gibt es europäische Bemühungen, um mit der European Open Science Cloud (EOSC) eine föderierte Infrastruktur zu entwickeln [1]. Für Deutschland stellt sich daher die Frage nach einer Anschlussfähigkeit und Mitgestaltung dieser Entwicklungen. Die NFDI kann hier eine Antwort liefern, um innerhalb der Forschungsdisziplinen neue Strukturen zu schaffen, die eine Beteiligung und Unterstützung ermöglichen.

Der Weg zu einem stabilen, sich aber dennoch anpassenden, vernetzten Ökosystem von verschiedenen Infrastrukturen wird weiterhin eine Herausforderung bleiben. Durch verschiedene nicht abgestimmte Fördermaßnahmen ist auch künftig absehbar, dass weiterhin Parallelentwicklungen entstehen werden. Auch ist die hohe Dynamik der Wissenschaft nicht zu unterschätzen, während im Vergleich die Entwicklung und der Betrieb von Infrastrukturen in eher langsameren Zyklen erfolgt. Die Möglichkeit, neue Ansätze verfolgen zu können, muss letztlich kein Nachteil sein, sondern ist zwangsläufig notwendig, um den Innovationen der Wissenschaft Raum zu geben. Daher ist eine dauerhafte Beobachtung der Entwicklungen und regelmäßiges Nachjustieren notwendig.

Ebenso ist nicht zu übersehen, dass große Informationsbestände in der Wirtschaft bei einigen wenigen Firmen entstehen und dort als wertvolles Gut geschützt werden. Es ist absehbar, dass in einigen Disziplinen Wissenschaft nicht kompetitiv erfolgreich sein kann, wenn kein Zugang zu diesen Datenquellen besteht. Solche Monopolstrukturen sind nicht kompatibel mit den wissenschaftlichen Vorstellungen zu Open Data und Open Science. Dennoch wird man sich mit dieser Entwicklung auseinandersetzen und geeignete Kooperationsmodelle suchen müssen.

Fazit

Erfolgreiche Datenanalyse erfordert aufwendige Hardware-Software-Infrastrukturen zur Speicherung, Analyse, Archivierung und Verteilung von Daten. Nur in einem Ökosystem mit aufeinander abgestimmten Komponenten ist eine maximale Wertschöpfung aus Daten möglich. Die beschriebenen Beispiele erläutern Forschungsinfrastrukturen mit Alleinstellungsmerkmalen, die spezifische Nutzergruppen unterstützen. Sie repräsentieren jedoch nur eine kleine Auswahl der deutschen Zentren, die zu diesen Fragestellungen forschen, entwickeln und Systeme und Dienste bereitstellen. Darüber hinaus bemüht man sich auf nationaler Ebene um geeignete Strukturierungskonzepte, um aus Data Science einen Nutzen für die Gesellschaft zu schaffen. Die Bestrebungen sind in vielfältige europäische und internationale Ansätze eingebunden. Data Science kennt keine Grenzen – ethische Fragestellungen hierbei müssen anderen Themenheften vorbehalten bleiben.

Literatur

1. European Open Science Cloud (EOSC) (2019) Strategic Implementation Plan. Juli 2019. ISBN: 978-92-76-09175-2KI-03-19-507-EN-N, https://ec.europa.eu/info/publications/european-open-science-cloud-eosc-strategic-implementation-plan_en, letzter Zugriff: 3.10.2019
2. Gauß-Allianz, https://gauss-allianz.de/de/, letzter Zugriff: 3.10.2019
3. Gute wissenschaftliche Praxis – „Leitlinien zur Sicherung guter wissenschaftlicher Praxis", https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf, letzter Zugriff: 3.10.2019
4. https://www.oldweather.org/, letzter Zugriff: 3.10.2019
5. https://www.dkrz.de/up/systems/wdcc, letzter Zugriff: 3.10.2019
6. https://esgf-data.dkrz.de, letzter Zugriff: 3.10.2019
7. https://www.coretrustseal.org/, letzter Zugriff: 3.10.2019
8. RfII-Empfehlung (2016) „Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland", Juni 2016, URN: urn:nbn:de:101:1-201606229098
9. RfII-Fachbericht (2017) Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich. Juli 2017
10. Wissenschaftsrat (2012) Strategische Weiterentwicklung des Hoch- und Höchstleistungsrechnens in Deutschland. Positionspapier. Berlin, Drs. 1838-12
11. Wissenschaftsrat (2015) Empfehlungen zur Finanzierung des Nationalen Hoch- und Höchstleistungsrechnens in Deutschland. Stuttgart, Drs. 4488-15, (April 2015), https://www.wissenschaftsrat.de/download/archiv/4488-15.pdf, letzter Zugriff: 3.10.2019