GESELLSCHAFT
FÜR INFORMATIK

Data Literacy und Data Science Education:

Digitale Kompetenzen in der Hochschulausbildung

Policy Paper der Präsidiums-Task-Force „Data Science“ der Gesellschaft für Informatik e.V. in Zusammenarbeit mit Vertretern der Deutschen Mathematiker-Vereinigung e.V., der Deutschen Physikalischen Gesellschaft e.V. und der Gesellschaft Deutscher Chemiker e.V.

GI.DE

GESELLSCHAFT
FÜR INFORMATIK

IMPRESSUM

Herausgabe

Gesellschaft für Informatik e.V.
Spreepalais am Dom
Anna-Louisa-Karsch-Str. 2
10178 Berlin
gi.de

Stand

April 2018

Bildnachweis

Monsitj

GELEITWORT

Sehr geehrte Damen und Herren,

liebe Leserin, lieber Leser,

der Begriff „Data Science“ – die Datenwissenschaft – stammt aus den Anfängen der Datenhaltung und -analyse, die bis in die 1960er Jahre zurückgehen. Mit der zunehmenden Bedeutung von „Big Data“ rückte die Wissenschaft der Daten weiter in den Fokus. Der Schwerpunkt der „Data Science“ liegt dabei nicht bei den Daten selbst, sondern auf der Art und Weise, wie diese verarbeitet, aufbereitet und analysiert werden. Data Science beschäftigt sich mit einer zweckorientierten Datenanalyse und der systematischen Generierung von Entscheidungshilfen und -grundlagen, um Wettbewerbsvorteile erzielen zu können.

Im Unternehmensumfeld ist das Thema häufig im Bereich Business Intelligence angesiedelt. IT-Unternehmen, Banken und Beratungsfirmen suchen die auf große Datenmengen spezialisierten Analysten genauso wie Autohersteller, Versicherungen und Verwaltungsbehörden. Die Unternehmensberatung McKinsey geht für 2017 von 150 000 offenen Stellen allein in den Vereinigten Staaten aus. Auch in Deutschland ist der Bedarf groß. Ein Geschäftsüberblick der Universität Harvard kürte den „Data Scientist“ sogar zum „attraktivsten Beruf des 21. Jahrhunderts“.

In der Wissenschaft beschäftigt sich Data Science mit unterschiedlichen Bereichen und kann daher verschiedene akademische Hintergründe haben: Informatik, Statistik, Mathematik, Natur- oder Wirtschaftswissenschaften, einschließlich des maschinellen Lernens, des statistischen Lernens, der Programmierung, der Datentechnik, der Mustererkennung, der Prognostik, der Modellierung von Unsicherheiten und der Datenlagerung. Mittlerweile existiert eine Reihe von Data-Science-Bachelor- oder -Masterstudiengängen.

Aufgrund der wachsenden Bedeutung dieses neuen Wissenschaftsfeldes an der Schnittstelle zu verschiedenen Anwendungsbereichen – sowohl für Forschung als auch für die Lehre – hat die Gesellschaft für Informatik e.V. die Task-Force „Data Science“ ins Leben gerufen. Diese interdisziplinäre

Arbeitsgruppe geht der Frage nach, was einen Data Scientist in Abgrenzung zu bestehenden Wissenschaftsdisziplinen wie der Informatik ausmacht und welche Kompetenzen ein Datenwissenschaftler und eine Datenwissenschaftlerin mitbringen müssen. Wir wollen der Frage nachgehen, wie Universitäten und Hochschulen das Profil eines Data Scientist definieren und wie sich das vom Verständnis in Unternehmen abgrenzt.

In Abgrenzung dazu bedarf es auch grundlegender digitaler Kompetenzen in der Breite der Studierendenschaft: Eine Data Literacy. Data Literacy ist die Fähigkeit des planvollen Umgangs mit Daten. In Ergänzung zu spezialisierten Fachkräften – den Data Scientists – liegt der Fokus auf der bedarfsgerechten, Disziplinen übergreifendem Know-how, um datengestützt arbeiten und entscheiden können.

In einem interdisziplinären Workshop mit Beteiligung von Vertreterinnen und Vertretern der Deutschen Mathematiker-Vereinigung e.V., der Deutschen Physikalischen Gesellschaft e.V. und der Gesellschaft Deutscher Chemiker e.V., der Ende Januar 2018 unter dem Titel „Data Science: Vom Buzz-Word zu einer neuen Methodik des (wissenschaftlichen) Arbeitens im 21. Jahrhundert“, in Berlin stattfand, wurden verschiedene Perspektiven hinsichtlich „Data-Literacy“- und „Data-Science“-Kompetenzen in Deutschland aufgezeigt.

Im Koalitionsvertrag von CDU/CSU und SPD heißt es: „Die Hightech-Strategie wird weiterentwickelt und auf die großen gesellschaftlichen Herausforderungen fokussiert. […] Es gilt heute Data Science in allen Bereichen, insbesondere aber in den Hochschulen, auszubauen. Dazu muss der Umgang mit Daten zu einem zentralen eigenen Wissenschaftsfeld und einer eigenen Disziplin werden.“

Die Gesellschaft für Informatik e.V., die mit 20.000 Mitgliedern die größte Fachgesellschaft für Informatik im deutschsprachigen Raum ist, will diese Entwicklung maßgeblich mitgestalten.

Diese Publikation ist ein Beitrag dazu. Sie stellt eine Zusammenfassung der im Workshop vorgestellten Impulse dar und soll das Diskussionsfeld rund um den Themenbereich „Data Science“ aufspannen. Die folgenden Beiträge reflektieren dabei ausschließlich die Meinungen der Autorinnen und Autoren.

Wir wünschen viel Spaß bei der Lektüre und freuen uns über weitere Beiträge.

Ihr Peter Liggesmeyer

Past-President und Sprecher der Präsidiums-Task-Force Data Science der Gesellschaft für Informatik e.V.

INHALTSVERZEICHNIS

GELEITWORT 3

GRUNDLEGENDE DIGITALE KOMPETENZEN IN DER HOCHSCHULAUSBILDUNG 6

Data Literacy: Data Skills in der Breite der Hochschulausbildung 6
Data Literacy und das Modell der Schlüsselkonzepte des Datenmanagements 10

DATA SCIENCE EXPERTS: BEDARF UND MÖGLICHKEITEN DIESEN ZU DECKEN 14

Irgendwas mit Daten… Warum wir Data Scientists brauchen 14
Qualitätsanforderungen für Data Science in Deutschland 19
Data Science in der deutschen Hochschullandschaft 22
Data-Science-Kompetenzzentren als Schlüssel zum Erfolg 25
Data Science und die Qualität von Daten 28
Data Science im angloamerikanischen Kontext: Status quo eines modernen Studien- und Berufsbilds 31

DATA SCIENCE ALS INTERDISZIPLINÄRE HERAUSFORDERUNG FÜR WISSENSCHAFT 34

Data Science zur Gestaltung der Digitalen Transformation 34
Interdisziplinäre Daten-Laboratorien als Antwort auf die „Data Challenge“ 37
Data Science in den Sozialwissenschaften 40
Data Science aus Sicht der Mathematik 44
Die Bedeutung von Data Science für die Chemie 47

DATA SCIENCE EDUCATION IN DER PRAXIS 50

Data Science: Von der Wissenschaft zum Erfolgsfaktor für Unternehmen 50
Data Scientists in Deutschland: Über Talentmangel, den Lückenschluss zwischen Promotion und Industrie sowie firmeninterne Weiterbildung 53

AUSBLICK 56

Von Data Literacy bis Data Science: Der Handlungsbedarf in der deutschen Hochschullandschaft 56

GRUNDLEGENDE DIGITALE KOMPETENZEN IN DER HOCHSCHULAUSBILDUNG

Data Literacy: Data Skills in der Breite der Hochschulausbildung

Von Dr. Jens Heidrich, Fraunhofer-Institut für Experimentelles Software Engineering IESE, Kaiserslautern, und Daniel Krupka, Gesellschaft für Informatik e.V., Berlin

Data Literacy ist die Fähigkeit des planvollen Umgangs mit Daten. In Ergänzung zu spezialisierten Fachkräften – den Data Scientists – liegt der Fokus auf der bedarfsgerechten, Disziplinen übergreifenden Vermittlung von Wissen, um datengestützt arbeiten und entscheiden können. Data Literacy wird als eine zentrale Kompetenz für Digitalisierung und globale Wissensgesellschaft gesehen. Eine Herausforderung liegt darin, dass Kompetenzen aus verschiedenen Disziplinen integriert werden müssen und ein an die Bedürfnisse des Anwendungsfelds angepasstes Angebot kreiert werden muss.

Die Fähigkeit, planvoll mit Daten umzugehen und sie im jeweiligen Kontext bewusst einsetzen und hinterfragen zu können, wird über alle Studienrichtungen hinweg immer wichtiger. Eine Studie der Dalhousie University in Kanada (Ridsdale et al.[1]) fasst diese Fähigkeit unter dem Begriff Data Literacy zusammen. Dies beinhaltet die Kompetenzen, Daten erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anwenden zu können.

In der Lesart der kanadischen Forscher ist Data Literacy eine zentrale Kompetenz für die Digitalisierung und die globale Wissensgesellschaft in allen Sektoren und Disziplinen. Angesichts der zunehmenden Menge und der Verfügbarkeit von Daten stellt sich die Herausforderung, mit den Daten Wissen zu generieren und fundiert Entscheidungen treffen zu können.

Aus Sicht des Hochschulforums Digitalisierung braucht es hierfür „auf der einen Seite spezialisierte Fachkräfte – Data Scientists – und auf der anderen Seite in allen Sektoren und Fächern datengewandte Expertinnen und Experten sowie Führungskräfte, die in

[1] Ridsdale et al. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report, 2015.

ihren jeweiligen Gebieten Fragen zu den Daten formulieren und Datenanalysen fachlich interpretieren können, insgesamt also datengestützt arbeiten und entscheiden können."[2]

Die kanadischen Forscher haben dazu 22 Kompetenzen mit zugehörigen Fähigkeiten, Wissen und Aufgaben identifiziert und in 5 Kompetenzfelder gruppiert. Zusätzlich werden konzeptionelle Kompetenzen, Kernkompetenzen sowie fortgeschrittene Kompetenzen unterschieden. Darüber hinaus wurde eine Reihe von Empfehlungen erarbeitet, wie „Data Literacy" erfolgreich an Hochschulen vermittelt und verbreitet werden kann.

Conceptual Framework	Introduction to Data
Data Collection	Data Discovery and Collection
	Evaluating and Ensuring Quality of Data and Sources
Data Management	Data Organization
	Data Manipulation
	Data Conversion
	Metadata Creation and Use
	Data Curation, Security and Re-Use
	Data Preservation
Data Evaluation	Data Tools
	Basic Data Analytics
	Data Interpretation (Understanding Data)
	Identifying Problems Using
	Data Visualization
	Presenting Data (Verbally)
	Data Driven Decisions Making (DDDM)
Data Application	Critical Thinking
	Data Culture
	Data Ethics
	Data Citation
	Data Sharing
	Evaluating Decisions based on Data

Abbildung 1: Data-Literacy-Kompetenzen nach Ridsdale et al.

[2] Ausschreibung einer Studie „Übergreifende Kompetenzen und Studieninhalte in der digitalen Welt am Beispiel von Data Literacy".

Ausgehend von den Vorarbeiten der Dalhousie University arbeitet die Gesellschaft für Informatik zusammen mit dem Fraunhofer-Institut für Experimentelles Software Engineering IESE an einer Studie, um umsetzbares Wissen für Hochschulen und Fächer für die Curriculumentwicklung im Hinblick auf übergreifende Kompetenzen und Inhalte in der digitalen Welt anhand der Kompetenz der Data Literacy zusammenzustellen. Insbesondere soll die Studie beleuchten, wie die Kompetenz der Data Literacy von allen Studierenden immersiv erworben werden kann.

Dazu wurden aus 83 internationalen Studien- und Weiterbildungsangeboten 13 ausgewählt, die nun einer näheren Betrachtung unterzogen werden, um im Rahmen von Interviews und einem Workshop entsprechende Inhalte, Erfahrungen und Empfehlungen für die Vermittlung und Verbreitung von Data Literacy zusammenzustellen. Vorbilder für die erfolgreiche Etablierung entsprechender Angebote in Hochschulen finden sich insbesondere im angloamerikanischen Raum; aber auch innerhalb von Europa und Deutschland gibt es bereits vielversprechende Ansätze, die näher betrachtet werden.

Fazit und Ausblick

Erste „Lessons Learned“ und Empfehlungen, die auf der Literatur- und Desk-Research-Phase identifiziert werden konnten, sind:

1) Es ist eine große Herausforderung, möglichst früh das Bewusstsein für die Wichtigkeit von Data Literacy bei Studierenden und den Bildungsinstitutionen zu schaffen.
2) Es bedarf des Aufbaus einer disziplinunabhängigen Institution, die Forschende und Lehrende aus verschiedenen Kompetenzfeldern (wie informatische, mathematische und Domänen-Kompetenzen) zusammenbringt.
3) Es bedarf des Einsatzes kreativer Lehransätze mit technologischen Hilfsmitteln, u.a. „Hands-on“, „Modul-basiertes“ und „Projekt-basiertes“ Lernen in Workshops oder Labs mit realen Daten
4) Es bedarf einer genauen Betrachtung der Bildungsniveaus und Disziplinen, um die Angebote entsprechen anzupassen, denn nicht jede/r braucht das komplette Spektrum der Data-Literacy-Kompetenzen.
5) Data Literacy sollte in der gesamten Breite der Hochschulausbildung vermittelt werden.

Über die Autoren

Dr. Jens Heidrich ist Hauptabteilungsleiter Prozessmanagement am Fraunhofer-Institut für Experimentelles Software Engineering IESE. Der promovierte Informatiker leitet die Durchführung der Data-Literacy-Studie für das Hochschulforum Digitalisierung.

Daniel Krupka ist Geschäftsführer der Gesellschaft für Informatik e.V. und leitet die Berliner Geschäftsstelle der Fachgesellschaft. Dort ist der Verwaltungswissenschaftler u.a. verantwortlich für die Kontakte zu Politik, Wirtschaft und Gesellschaft.

Data Literacy und das Modell der Schlüsselkonzepte des Datenmanagements

Von Andreas Grillenberger und Prof. Dr. Ralf Romeike, Friedrich-Alexander-Universität Erlangen-Nürnberg

Data Literacy ist ein Bereich der informatischen Bildung, welcher sowohl mit den (Weiter-) Entwicklungen des Forschungsbereichs Datenbanken als auch den korrespondierenden gesellschaftlichen Anforderungen in Verbindung steht. In der berufsbezogenen Ausbildung und für die Allgemeinbildung gewinnt Data Literacy zunehmend an Bedeutung. Zur fachlichen Fundierung wird der Bezug auf das Modell der *Schlüsselkonzepte des Datenmanagements* vorgeschlagen und ein Data-Literacy-Kompetenzmodell skizziert.

Die Didaktik der Informatik erforscht und beschreibt informatische Bildungsgegenstände bezogen auf die berufliche Ausbildung und die Allgemeinbildung. Eine ihrer zentralen Aufgaben ist die Betrachtung tragfähiger Kernaspekte der Wissenschaft Informatik mit dem Ziel, nachhaltige Lerngegenstände und Kompetenzen zu ermitteln. Hierzu werden Ansätze wie die *Fundamentalen Ideen der Informatik*[3] oder die *Great Principles of Computing*[4] herangezogen, welche die Informatik oder eines ihrer Teilgebiete durch zentrale Begriffe, Konzepte, Ideen oder Prinzipien charakterisieren. Dadurch vermitteln diese einen Einblick in die betrachtete Wissenschaftsdisziplin, strukturieren sie verständlich und helfen Lehr- und Lerninhalte herauszudestillieren.

Ein zentraler Gegenstand der Informatik ist die Verwaltung und Verarbeitung von Daten. Deren Relevanz und Wahrnehmung hat in den letzten Jahren, insbesondere mit der Digitalisierung aller Lebensbereiche, deutlich zugenommen, wodurch sich ihre Bedeutung auch außerhalb der Informatik geändert hat. Sowohl für den alltäglichen als auch den beruflichen Umgang mit Daten sind heute grundlegende Kompetenzen essentiell, die oft unter dem Begriff Data Literacy subsumiert werden.

Zur Ermittlung der zentralen Inhalte und Kompetenzen, die hinter dem Begriff Data Literacy stehen, sowie zu deren fachlicher Fundierung kann das Modell der Schlüsselkonzepte des

[3] Schwill, Andreas (1993): Fundamentale Ideen der Informatik. In: Zentralblatt für Didaktik der Mathematik, 25(1).
[4] Denning, Peter J. (2003): Great Principles of Computing. In: Commun ACM, 46(11).

Datenmanagements[5] (vgl. Abbildung 2) verwendet werden. Dieses nimmt, in Anlehnung an o.g. Arbeiten zum Fundament der Informatik, eine Charakterisierung des Fachgebiets vor und stellt die zentralen Bereiche des Themenfelds Daten strukturiert dar.

Eine empirische Untersuchung der Inhalte und Kompetenzen, die in verschiedenen Data-Science-Studiengängen enthalten sind, liefert auf dieser Basis einen Entwurf eines Data-Literacy-Kompetenzmodells (vgl. Abbildung 3). Dieses Kompetenzmodell beinhaltet die Data-Literacy-Kompetenzen nach Ridsdale et al., stellt den Bereich aber mit einem stärkeren fachdidaktischen Fokus dar und legt einen Schwerpunkt auf die fachliche Fundierung und die dahinterstehenden Konzepte.[6]

Wie auch beim Kompetenzmodell der GI-Empfehlungen für Bildungsstandards Informatik für die Sekundarstufe I/II sind die Prozess- und Inhaltsbereiche des Data-Literacy-Kompetenzmodells eng miteinander verzahnt, wie folgende Beispiele verdeutlichen: Die Kompetenz „Daten mit Hilfe von Sensoren erfassen“ verbindet die Bereiche Daten und Datenquellen sowie Datenerfassung/-gewinnung miteinander, während „die kontinuierliche Erfassung von Daten durch und über uns beurteilen“ die Verknüpfung von Datenethik, legalen und gesellschaftlichen Aspekten mit der Datenerfassung/-gewinnung betont und „eine (einfache) korrelationsbasierte Datenanalyse auf geeigneten Daten durchführen“ die Grundsätze der Datenanalyse sowohl mit der Datenerfassung/-gewinnung als auch der Analyse, Visualisierung und Evaluation in Verbindung setzt.

Um das Kompetenzmodell weiter zu fundieren und auszudifferenzieren, aber auch um den Gegenstandsbereich Data Literacy und Data Science zu erschließen, ist weitere Forschung notwendig.

[5] Grillenberger, Andreas und Romeike, Ralf (2017): Key Concepts of Data Management: An Empirical Approach. In: Proceedings of the 17th Koli Calling International Conference on Computing Education Research, ACM, New York.

[6] Ridsdale et al. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report“, Report, 2015.

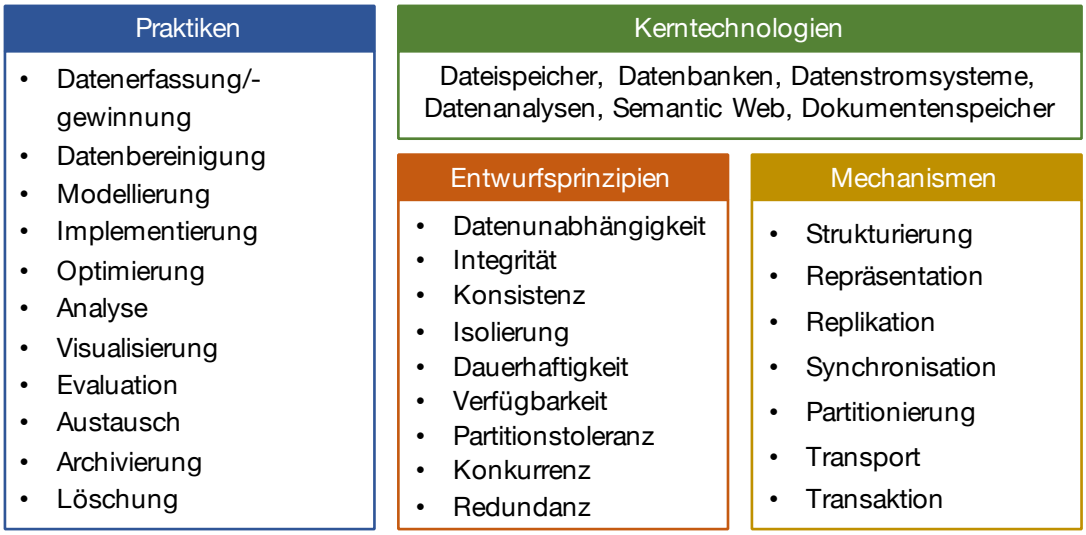

Abbildung 2: Modell der Schlüsselkonzepte des Datenmanagements

Abbildung 3: Entwurf eines Data-Literacy-Kompetenzmodells

Fazit und Ausblick

1) Data Science ist ein zunehmend wichtiger interdisziplinärer Forschungs- und Bildungsbereich, der eine starke Basis in der Informatik (insbesondere Datenmanagement) aufweist.
2) Data Literacy ist ein junger Gegenstandsbereich der informatischen Bildung, der aufgrund seiner steigenden Bedeutung sowohl für die berufsbezogene Ausbildung als auch im Hinblick auf die Allgemeinbildung auszudifferenzieren ist.
3) Für einen mündigen Umgang mit Daten und datenbasierten Systemen müssen alle Schüler und Studenten grundlegende Data-Literacy-Kompetenzen erwerben.
4) Kernideen der Data Science werden durch das Modell der Schlüsselkonzepte des Datenmanagements beschrieben.
5) Der Fachbezug von Data Literacy kann über das Modell der Schlüsselkonzepte des Datenmanagements hergestellt werden.
6) Weitere Forschung zur Fundierung und Ausdifferenzierung von Data Literacy und Data Science ist notwendig.

Über die Autoren

Andreas Grillenberger ist wissenschaftlicher Mitarbeiter an der Professur für Didaktik der Informatik der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seit mehr als vier Jahren beschäftigt er sich mit der Aufarbeitung des Themengebiets Datenmanagement aus informatikdidaktischer Sicht.

Ralf Romeike ist Professor für Didaktik der Informatik an der Friedrich-Alexander-Universität Erlangen-Nürnberg und Sprecher der GI-Fachgruppe Didaktik der Informatik. Ziel seiner Aktivitäten in Forschung und Lehre ist die Ausgestaltung informatischer Bildung, die Kinder und Jugendliche dazu befähigt, die digitale Gesellschaft zu verstehen und mitzugestalten.

DATA SCIENCE EXPERTS: BEDARF UND MÖGLICHKEITEN DIESEN ZU DECKEN

Irgendwas mit Daten... Warum wir Data Scientists brauchen

Von Gerald Swarat, Fraunhofer-Institut für Experimentelles Software Engineering IESE

Ein Data Scientist soll „Big Data“ beherrschbar und nutzbar machen, indem er oder sie von der Datensammlung über die Datenanalyse und Präsentation bis hin zur Ableitung von Handlungsempfehlungen den kompletten Big-Data-Lifecycle aus dem Effeff beherrscht. Die Nachfrage nach den Expertinnen und Experten ist enorm und steigt konsequent an, denn Unternehmen haben natürlich erkannt, dass sie Fachexperten brauchen. Der Bedarf muss mit unterschiedlichen Maßnahmen kurz-, mittel- und langfristig adressiert werden.

Data never sleeps: Für die riesigen, großteils unstrukturierten Datenbestände in Wirtschaft und öffentlicher Hand hat sich in den letzten Jahren mit großem Nachdruck ein Fachbegriff etabliert: Big Data. Allerdings sind diese Daten weder Informationen noch Wissen: „Sie sind erst wertvoll, wenn sie verfeinert und analysiert werden, sodass aus Rohdaten ‚Smart Data‘“[7]werden. Nur dann können die wissenschaftlichen, ökonomischen und sozialen Wirkungskräfte freigesetzt werden.

Im gleichen Maße, wie Schlagworte wie Big Data, Smart Data, Data Mining, Data Engineering, Predictive Analytics in die Diskussion geraten, steigt die Hoffnung auf eine neue Spezies – der Wunsch nach Data Scientists. Denn beinahe unmittelbar mit dem Aufkommen des Begriffs „Big Data“ folgte das Bekenntnis, mit dem Phänomen schier unübersichtlicher Datenmengen überfordert zu sein und diesen nicht mit herkömmlichen Strategien beizukommen oder gar Mehrwerte daraus ziehen zu können.

[7] Vgl. Markl, Volker: Gesprengte Ketten. In: Informatik Spektrum 01/2015.

Data Scientist – The Sexiest Job of the 21st Century?[8]

Was ist also ein Data Scientist? Ganz verkürzt dargestellt soll ein Data Scientist „Big Data" beherrschbar und nutzbar machen, indem er oder sie von der Datensammlung über die Datenanalyse und Präsentation bis hin zur Ableitung von Handlungsempfehlungen den kompletten Big-Data-Lifecycle aus dem Effeff beherrscht.

Einsatzfelder zeigen sich zur Genüge, so hat sich z.B. zunehmend der Bereich von algorithmischen Entscheidungssystemen als Paradebeispiel für Data Scientists angeboten, seitdem über deren Fehler in der Öffentlichkeit diskutiert wird. Bspw. bei der Bewertung von Angeklagten bezüglich ihres aktuellen oder künftig zu erwartenden kriminellen Verhaltens oder bei der Vergabe von Krediten.[9] In diesen Bereichen werden die Entscheidungen und Interpretationen von Data Scientists Auswirkungen sowohl auf die Gesellschaft im Allgemeinen als auch auf das Wohl und Wehe eines einzelnen Individuums besitzen, was zusätzlich zur technischen Souveränität eine klare Berufsethik erfordert. Es geht also darum, „die möglichen gesellschaftlichen Folgen von Softwaresystemen zu modellieren und zu antizipieren."[10]

Der Bedarf ist groß

Die Nachfrage ist trotz (oder gerade wegen) der sehr breiten und unscharfen Definition vorhanden und steigt konsequent an, denn Unternehmen haben natürlich erkannt, dass sie Fachexperten brauchen, die aus diversen Datenquellen Antworten auf die brennenden Fragen finden, um Prozessabläufe zu optimieren, den Kunden besser zu verstehen oder gar völlig neue Geschäftsmodelle zu eröffnen. Banken und Beratungsfirmen suchen die auf große Datenmengen spezialisierten Analysten genauso wie Autohersteller, Versicherungen und Verwaltungsbehörden – und in Zeiten der allgegenwärtigen Digitalisierung und Konzepten wie Industrie 4.0 und autonomer Systeme findet man fast in jeder Branche ein Einsatzfeld. Natürlich steigt mit diesem Markt auch die Nachfrage bei Studierenden.

Nur werden diese Bedürfnisse durch das heutige Ausbildungssystem noch nicht ausreichend gestillt; hinzu kommt, dass die Anforderungen an einen Data Scientist außerordentlich komplex sind. Es ist also höchste Zeit, Aufgaben und Kompetenzen des Berufsbildes zu diskutieren und ein Curriculum abzuleiten, das übergreifende Kompetenzen und Inhalte in der digitalen Welt, wie z.B. Wissenschaft, Arbeitswelt und Gesellschaft, zusam-

[8] Lt. Harvard Business Review (2012): http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

[9] Vgl. Zweig, Katharina (2018): Wo Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. Eine Publikation der Bertelsmann Stiftung, Gütersloh.

[10] Zweig, Katharina: S. 30.

menbringt. Dabei geht es einerseits tatsächlich um die Ausbildung von Fachexperten – den Data Scientists, die sich in der Tiefe mit den Verfahren und Technologien auskennen – und andererseits darum, in den einzelnen Disziplinen eine Art übergreifende Grundkompetenz in Data Literacy zu vermitteln.

In Deutschland wird Data Science bisher vor allem als Aufbaustudium nach einem Bachelor zum Beispiel in Mathematik oder Informatik angeboten. München war 2015 eine der ersten Universitäten, die einen Master in Data Science anboten, Mannheim und Darmstadt zogen nach. Außerdem gibt es kostenpflichtige Fortbildungsangebote für Berufstätige.

Volker Markl weist zu Recht darauf hin, dass in der Vergangenheit die Disziplinen der Datenanalyse und der skalierbaren Datenverarbeitung nicht eng verzahnt waren, was jedoch für einen souveränen Umgang mit großen Datenmengen mit geringer Latenz erforderlich ist. Zudem sind das Wissen der Anwendungsdomäne und die juristischen und gesellschaftlichen Implikationen zu beachten. Markl vergleicht deshalb die Wünsche, die auf den Data Scientist projiziert werden, mit der Suche nach der eierlegenden Wollmilchsau:

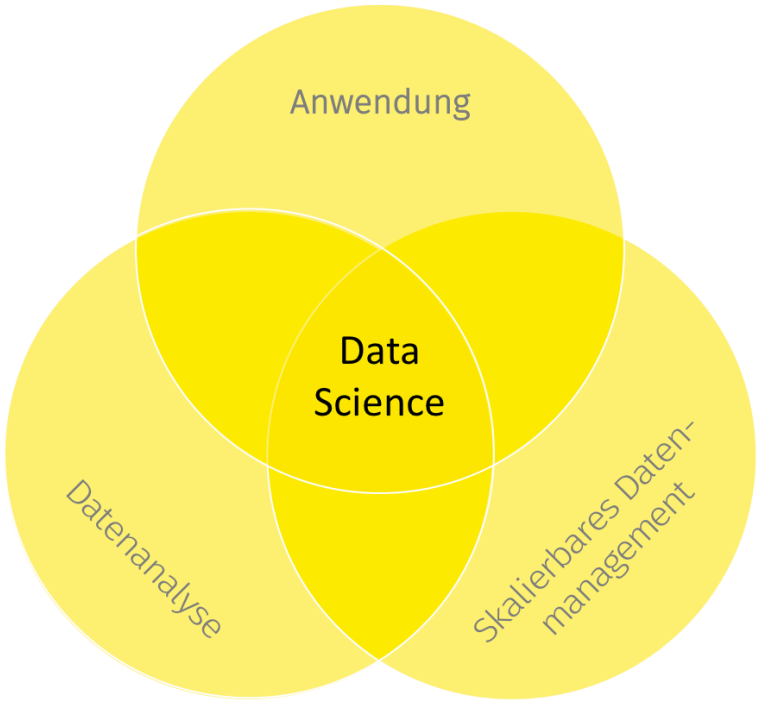

Abbildung 4: Die Anforderungen an den/die „Data Scientist“ ist enorm (Markl 2015)

Warum ist die eierlegende Wollmilchsau so wichtig?

Wenn wir Angst haben, dass wir alte Jobs verlieren, weil sie in der digitalen Welt überflüssig werden könnten, müssen neue geschaffen werden. Wenn wir Angst haben, dass uns die Datenflut übermannt und wir niemanden haben, der die Instrumente beherrscht, dann brauchen wir Berufsbilder, die das können. Wenn wir Angst haben, dass nur die bekannten Big Player der Digitalisierung, wie Google, Amazon, Facebook und Apple, Potenzial aus Big Data schöpfen, dann müssen wir Sorge tragen, dass unsere Unternehmen die notwendigen Kompetenzen über entsprechend ausgebildetes Personal aufbauen und nutzen können. Wenn wir Angst haben, dass die Technik zunehmend Macht über uns gewinnt (Stichwort KI/Algorithmen) und wir feststellen, dass sich immer mehr ethische Fragen aufdrängen, sollten wir Rahmenbedingungen wie einen digitalen Datencodex entwickeln und Data Scientists in Ethik schulen.

Wir stellen also fest, dass eine interdisziplinäre und ethische Perspektive notwendig ist, die bei den oft technischen Ausbildungen der heutigen Data Scientists nicht unbedingt gegeben ist. Zum anderen fehlen den Anwendern in den einzelnen Domänen oft Kompetenzen, wenn es darum geht, die Datengrundlage einzuschätzen und Ergebnisse richtig zu interpretieren. Diese Kompetenzen gilt es in der Breite aufzubauen. Es ist deshalb zu begrüßen, dass im Koalitionsvertrag Data Science Beachtung findet.

Handlungsempfehlungen

Um kurzfristig den Bedarf zu decken, sollten die betreffenden Mitarbeiter in den Unternehmen unterstützt und geschult werden (bspw. über Fortbildungen[11]). Dafür müssen mehr flexible Angebotsformen (Fernstudienoptionen und berufsbegleitende Optionen) bereitgestellt werden. Ein mittelfristiger Ansatz bringt die Unternehmen mit den anwendungsorientierten Wissenschaftseinrichtungen zusammen, um so die Bedarfsfelder mit den Kompetenzträgern zu vernetzen und ein Bewusstsein in der Öffentlichkeit zu erzeugen (z.B. in einem Kompetenzzentrum, das die Felder identifiziert, die den dringendsten Know-how-Aufbau benötigen und die Inhalte an verschiedene Bildungsniveaus und Disziplinen (nicht jeder braucht das komplette Spektrum) anpasst). Als langfristige Lösung ist ein im breiten Dialog entstandenes Curriculum in Deutschland zu entwickeln, das unter dem Schirm bestimmter Essentials den Universitäten genügend Freiraum überlässt, den Studiengang für sich auszugestalten.

11 https://www.academy.fraunhofer.de/de/weiterbildung/information-kommunikation/data-scientist-schulungen.html.

Über den Autor

Gerald Swarat ist studierter Historiker und Germanist. Er ist Leiter des Berliner Kontaktbüros des Fraunhofer-Instituts für Experimentelles Software Engineering IESE (Kaiserslautern). Swarat koordiniert die Aktivitäten des Instituts in der Bundeshauptstadt rund um die Themen Smart Ecosystems, Industrie 4.0, Datensouveränität und Smart Rural Areas, ist Ansprechpartner für Wissenschaft, Politik und Wirtschaft vor Ort in Berlin und stellvertretender Sprecher der Regionalgruppe Berlin-Brandenburg der GI.

Qualitätsanforderungen für Data Science in Deutschland

Von Dr. Hans-Josef Linkens, Bundesministerium für Bildung und Forschung

Um das Versprechen der Datenwissenschaften, aus Daten Informationen zu machen, zu erfüllen, sind offene Forschungsdaten wichtig. Denn ohne offene Daten ist ihre Nutzung, ihre Nachnutzung über die Grenzen von Regionen, Organisationen und Disziplinen hinweg nicht möglich. Um das deutsche Wissenschaftssystem im internationalen Wettbewerb weiter zu stärken, brauchen wir eine leistungsfähige Infrastruktur – und zwar nicht nur als Speicher, Rechner und Netze, sondern als Dienstleistungsangebote und Services. Und wir benötigen Personalentwicklung für Data Science auf allen Ebenen.

Der wissenschaftlichen Frage, wie aus Daten Informationen, Wissen und Innovationen gewonnen werden, kommt eine große Bedeutung zu. Die Politik hat sich bereits mehrfach mit den entsprechenden Rahmenbedingungen befasst. Zwei Empfehlungen haben das Thema breiter diskutiert und die wissenschaftspolitische Diskussion, die zu einer Nationalen Forschungsdateninfrastruktur und der EOSC geführt wird, geprägt und sollen hier kurz erwähnt werden:

Erstens: Die Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, die der Wissenschaftsrat 2012 veröffentlicht hat. Darin wird betont, dass die Gewährleistung des Zugangs zu den für die wissenschaftliche Arbeit und das Studium erforderlichen Daten und Wissensbeständen eine öffentliche Aufgabe ist. Zu den Empfehlungen gehörte, dass die Wissenschaft sich auf die „erforderlichen Kompetenzprofile verständigen“ solle.

Zweitens: Die Empfehlungen des Rats für Informationsinfrastrukturen (RfII) „Leistung aus Vielfalt“ von 2016. Hierin zieht der RfII u.a. den Schluss, eine nationale Forschungsdateninfrastruktur zu etablieren, um eine Grundversorgung mit Speichermöglichkeiten und Services anzubieten, Fragmentierung zu überwinden, Standards und Methoden zu vereinheitlichen und vernetzte Dienstleistungen anzubieten.

Um das Versprechen der Datenwissenschaften, aus Daten Informationen zu machen, zu erfüllen, sind offene Forschungsdaten wichtig. Denn ohne offene Daten ist ihre Nutzung über die Grenzen von Regionen, Organisationen und Disziplinen hinweg nicht möglich. Selbstverständlich soll „offen“ nicht „frei“ erhältlich bedeuten. Ebenso selbstverständlich gibt es Interessen, die zu berücksichtigen

sind, und Schutzbedarfe, die einzuhalten sind. Dennoch sollten Daten aus öffentlich geförderten Vorhaben zugänglich gemacht werden, um das Wissenschafts- und Innovationssystem zu stärken. Dieses Anliegen wird auch von der G20 und auf EU-Ebene verfolgt.

Die Grenzen zwischen dem, was die klassischen Wissenschaften und die Informationsinfrastrukturen leisten, werden unschärfer. Dies ist bereits heute zu bemerken – etwa bei Begutachtungen von wissenschaftlichen Vorhaben, in denen mehr und mehr Methodenkenntnis aus dem infrastrukturellen Teil des Antrags erforderlich ist.

Daraus folgt: Um die Qualität der Wissenschaft weiter zu befördern, benötigen wir eine leistungsfähige Infrastruktur – und zwar nicht nur als Speicher, Rechner und Netze, sondern als Dienstleistungsangebote und Services für das Forschungsdatenmanagement, für Mustererkennung, für Simulation, Visualisierung, Langzeitarchivierung, statistische Auswertung – für die gesamte Bandbreite der datenwissenschaftlichen Methoden, Techniken und Theorien, also dem, was man Data Science nennt. Hier brauchen wir generische Dienste wie auch fachspezifische Angebote. Es liegt dabei in der Verantwortung der wissenschaftlichen Communitys und einzelnen Fächer, ihre spezifischen Bedarfe zu beschreiben und Standards zu setzen. Auch die Wissenschaftspolitik ist in der Verantwortung, Rahmenbedingungen festzulegen.

Für die Bewältigung dieser Herausforderungen ist eine umfassende Personalentwicklung erforderlich. Angesichts des großen Bedarfs in Forschung und Lehre, in den Forschungseinrichtungen, aber auch in der Wirtschaft wird Datenexpertise auf allen Ebenen benötigt. Zum Bedarf gibt es verschiedene Aussagen, Studien und Abschätzungen – allen gemein ist: Der Bedarf ist erheblich. Es werden also sowohl spezialisierte Fachleute für die verschiedenen Methoden von Data Science als auch Kompetenzen in den jeweiligen Fachdisziplinen benötigt. Da die Informationsinfrastrukturen ein wesentlicher Bestandteil des Forschungsprozesses selber werden, werden auch Managementkompetenzen (Prozessmanagement, Kommunikation) erforderlich, um die Datenwissenschaften in den Forschungsprozess einzubinden.

Fazit und Ausblick

Vier Thesen für eine weitere Diskussion:

1) Eine Grundvoraussetzung für Datenwissenschaften sind offene Forschungsdaten – offen in einer intelligenten Form.
2) Die Grenzen zwischen dem, was die klassischen Wissenschaften und die Informationsinfrastrukturen leisten, werden unschärfer.

3) Um die Qualität der Wissenschaft weiter zu befördern, ist die Einrichtung einer leistungsfähigen Infrastruktur notwendig.

4) Eine Personalentwicklung auf allen Ebenen ist erforderlich.

Über den Autor:

Dr. Hans-Josef Linkens ist Referatsleiter im Referat Digitaler Wandel in Wissenschaft und Forschung des Bundesministeriums für Bildung und Forschung.

Data Science in der deutschen Hochschullandschaft

Von Dr. Klaus Wannemacher, HIS-Institut für Hochschulentwicklung e.V. (HIS-HE)

Eine sehr ausgeprägte Nachfrage nach Data Scientists am Arbeitsmarkt trifft auf ein sehr begrenztes Angebot an AbsolventInnen. Die Entwicklung des entsprechenden Studienangebots an deutschen Hochschulen befindet sich noch in einer Frühphase und ist bislang stark von staatlichen Hochschulen dominiert. Es ist ein Mangel an Bachelorstudiengängen, an weiterbildenden Angeboten und an alternativen Qualifizierungsformen erkennbar.

Im Rahmen des Forschungsvorhabens „Studienangebote im Bereich Data Science – Potenziale für Arbeitsmarkt und Hochschulentwicklung“ geht HIS-HE im Sinne einer Trendanalyse dem Entwicklungsstand im Bereich des Studienangebots für Data Science an den deutschen Hochschulen nach. Im Rahmen einer systematischen Literatur- und Dokumentenrecherche, einer bundesweiten Bestandsaufnahme sowie einer Expertenbefragung werden Tendenzen der Entwicklung des Studienangebots im Bereich Data Science erfasst und ausgewertet.

Eine im Rahmen des Projekts durchgeführte systematische Literatur- und Dokumentenrecherche förderte Studien zum Potenzial von Data Science für Unternehmen, arbeitsmarktanalytische und berufsfeldbezogene Untersuchungen u.Ä. zutage, doch kaum Studien zur Vermittlung einschlägiger Kompetenzen, zum Potenzial von Data Science für die Hochschulentwicklung und nahezu keine Untersuchung zum Studienangebot für Data Science.

Eine Deloitte-Studie zu „Data Analytics“, in deren Rahmen 291 deutsche Unternehmen ab 100 Mio. Euro Umsatz befragt worden waren, gelangte 2015 im Hinblick auf die Verfügbarkeit von Data-Science-Experten am Arbeitsmarkt zu einer kritischen Einschätzung: „Noch bietet der Hochschulsektor in Deutschland aber kaum dezidierte Studiengänge in diese Richtung; die Fähigkeiten bringen am ehesten Naturwissenschaftler mit.“ 47 Prozent der befragten Unternehmen sprachen sich für ein Eingreifen der Politik zur Erhöhung der Wettbewerbsfähigkeit im Auf- und Ausbau in spezialisierten Studiengängen aus.

Die grundsätzliche Situation des Studienangebots im Bereich Data Science an den deutschen Hochschulen ist noch immer wenig komfortabel, hat sich seit der Deloitte-Studie allerdings signifikant verändert, wie

Zwischenergebnisse der HIS-HE-Untersuchung belegen.

Die Bestandsaufnahme der Studienangebote im Bereich Data Science an den deutschen Hochschulen zeigt, dass mittlerweile rund 25 Bachelor- und Masterstudiengänge (sowie einige Kontaktstudienangebote/Zertifikatskurse) für Data Science existieren – bei steigender Tendenz. Die Studiengänge wurden mehrheitlich seit 2014 eingeführt und sind weit überwiegend generalistisch konzipiert; nur selten wird ein spezieller fachlicher Fokus verfolgt (z.B. Data Science in der Medizin). Die Studienangebote sind weit überwiegend an staatlichen Universitäten und staatlichen Fachhochschulen angesiedelt. Es handelt sich mehrheitlich um Masterstudiengänge, welche vielfach an zuvor absolvierte Informatik- oder Mathematikstudiengänge anknüpfen.

Gegenwärtig wird eine qualitative, explorative Expertenbefragung unter Vertreter(inne)n von Wirtschaftsverbänden und Unternehmen sowie in Wissenschaftseinrichtungen zur Validierung und Vertiefung der Projektergebnisse durchgeführt. Erste Ergebnisse der Expertenbefragung deuten darauf hin, dass

- die Nachfrage bei Studieninteressierten sehr ausgeprägt ist. Studiengangssprecher von Data-Science-Studiengängen berichten von jährlichen Ablehnungsquoten zwischen 50 und 95 Prozent.
- sich trotz hoher fachlicher Anforderungen zahlreiche Fachfremde (mit einem wissenschaftlichen Hintergrund in Betriebswirtschaftslehre, Naturwissenschaften, Psychologie etc.) auf Studienplätze bewerben. Sie holen fehlende Mathematik-/ Informatikkenntnisse zu Beginn des Studiums nach.
- dass widrige Rahmenbedingungen die Entwicklung weiterbildender Angebote durch Hochschulen hemmen (z.B. die MOOC-Konkurrenz aus den USA, heterogene Zielgruppen und die hochschulrechtlich verankerte Norm kostendeckenden Arbeitens bei weiterbildenden Angeboten).

Fazit und Ausblick

Es bedarf einer weiteren Ausdifferenzierung des Studienangebots für Data Science. Insbesondere sollten

- weiterbildende Teilzeitstudiengänge (Bachelor- und Masterniveau),
- weiterbildende Zertifikatskurse (mit ECTS) sowie
- weiterbildende Seminare und Workshops (ohne ECTS)

konsequent weiter ausgebaut werden. Während bei weiterbildenden Zertifikatskursen und Seminaren schon eine Anschubfinanzierung hilfreich sein kann, gilt es für Studiengänge dauerhafte Finanzierungswege sicherzustellen.

Über den Autor

Dr. Klaus Wannemacher ist wissenschaftlicher Mitarbeiter im Geschäftsbereich Hochschulmanagement des HIS-Instituts für Hochschulentwicklung (HIS-HE). Als Organisationsberater unterstützt er Hochschulen, außeruniversitäre Forschungseinrichtungen und Ministerien mit Grundlagenarbeiten, Beratungsleistungen und Forschungsprojekten sowie Angeboten zum Wissenstransfer mit einem Schwerpunkt auf dem Bereich Hochschule im digitalen Zeitalter.

Data-Science-Kompetenzzentren als Schlüssel zum Erfolg

Von Prof. Dr. Erhard Rahm, Universität Leipzig / ScaDS Dresden / Leipzig Competence Center for Scalable Data Services and Solutions

Erfolgreiche Data-Science-Lösungen erfordern die kombinierte Expertise zu Datenaufbereitung/Big Data, Data Mining und maschinellem Lernen sowie Anwendungshintergründen. Diese breite Expertise kann in der Regel nur durch ein Team von Spezialisten bereitgestellt werden, die idealerweise an Data-Science-Kompetenzzentren kooperieren. In der Hochschulausbildung sollte Data Science sowohl auf Bachelor- auch als auf Masterebene stark ausgeweitet werden.

Data Science beinhaltet Methoden zur umfassenden Analyse von Daten zur Gewinnung neuer Erkenntnisse in unterschiedlichsten Anwendungsgebieten. Das Spektrum der Analyseverfahren ist weit gefächert und reicht von einfachen Datenabfragen bis hin zu statistischen Data-Mining-Verfahren und Techniken des maschinellen Lernens. Wesentlich für die Aussagekraft und Qualität der Analyse ist eine umfassende Aufbereitung der Daten (Datenbereinigung und -konsolidierung), die oft aus unterschiedlichen Quellen stammen, teilweise unstrukturiert sind und zahlreiche Qualitätsmängel und Fehler beinhalten können. Die Daten sind zudem teilweise hoch-dynamisch (z.B. Sensordaten) und oft von sehr großem Umfang (Big Data). Diese Aspekte verlangen sehr leistungsfähige, hochparallele Lösungsansätze zur Datenverarbeitung, z.B. mit zahlreichen Prozessoren in lokalen Clustern.

Erfolgreiche Data-Science-Lösungen erfordern somit die kombinierte Expertise zu

- Big-Data-Technologien wie Datenaufbereitung/Datenintegration und skalierbare Ansätze zur parallelen Datenverarbeitung,
- zu Data Mining und maschinellem Lernen
- sowie die Anwendungsexpertise, um für die spezifischen Problemstellungen adäquate Verfahren zur Datenaufbereitung und Datenanalyse zu bestimmen.

Einzelne Personen werden in der Regel nur ein bis zwei dieser drei Bereiche ausreichend vertieft beherrschen, so dass anspruchsvolle Data-Science-Projekte in der Regel ein Team von Experten benötigen. Kompetenzzentren zu Data Science (bzw. Big Data oder Machine Learning) können die unterschiedlichen Kompetenzen bündeln und damit eine schnellere Nutz-

barkeit von Data Science in Wissenschaft und Praxis unterstützen. Dort können auch die entsprechenden Hardwareressourcen, in Kombination mit geeigneter Administrations- und Serviceunterstützung, bereitgehalten werden. Vorreiter solcher Zentren sind die beiden BMBF-geförderten Kompetenzzentren für Big Data, ScaDS Dresden/Leipzig sowie BBDC. Weitere Zentren sind geplant oder bereits an anderen Orten eingerichtet worden, um den hohen Bedarf an Data-Science-Lösungen in Deutschland zu adressieren.

Die Ausbildung zu Data Science sollte an den Hochschulen auf dem Bachelor- und Masterlevel stark ausgebaut werden, um der immensen Nachfrage nach akademisch ausgebildeten Data-Science-Spezialisten Rechnung zu tragen. Dies kann in eigenständigen Studiengängen erfolgen oder durch entsprechende neue Schwerpunkte in existierenden Studiengängen, z.B. der Informatik. Dabei sind in Informatikstudiengängen vor allem die Methoden in den beiden ersten der oben genannten Gebiete zu berücksichtigen. Die Anwendungsaspekte können im Rahmen eines Ergänzungsfachs sowie durch Praktika und in der Abschlussarbeit berücksichtigt werden.

Fazit und Ausblick

1) Die Einrichtung von Kompetenzzentren zu Data Science (Big Data, Machine Learning) hat sich bewährt, sollte durch Bund und Länder jedoch wesentlich verstärkt werden, jeweils mit standortspezifischen Forschungs- und Anwendungsschwerpunkten.
2) Umfassende Mehranstrengungen sind in der Data-Science-Ausbildung erforderlich, insbesondere zur Etablierung neuer bzw. angepasster Bachelor- und Masterstudiengänge an den Hochschulen.
3) Um einem Wildwuchs solcher Studiengänge entgegenzutreten, wäre es hilfreich, wenn die GI Empfehlungen zur inhaltlichen Ausgestaltung solcher Studienangebote erstellen könnte.

Über den Autor

Prof. Dr. Erhard Rahm ist Professor für Informatik (Datenbanken) an der Universität Leipzig und einer der beiden Direktoren des BMBF-geförderten Big-Data-Kompetenzzentrums ScaDS Dresden/Leipzig. Seine Forschungsgebiete sind Big Data und Datenintegration. Er ist Autor mehrerer Bücher und von über 200 wissenschaftlichen Publikationen. Seine Forschungsergebnisse wurden bereits mehrfach ausgezeichnet, u.a. mit den angesehenen VLDB Ten Year Best Paper Award sowie dem ICDE Influential Paper Award. Prof. Rahm ist gewähltes Mitglied im DFG-

Fachkollegium Informatik. In der GI ist er Sprecher des Fachbereichs Datenbanken und Informationssysteme (DBIS).

Data Science und die Qualität von Daten

Von Prof. Dr. Richard Lenz, Friedrich-Alexander-Universität Erlangen-Nürnberg

Im Zeitalter von „Big Data“ wird zunehmend der „Data Scientist“ als wichtiges Berufsfeld identifiziert. Er analysiert große Datenmengen mit teilautomatisierten Methoden und leitet aus seinen Ergebnissen Handlungsempfehlungen ab. Die Beurteilung der Datenqualität ist für den Data Scientist von zentraler Bedeutung. Dabei kann er aber nicht auf die traditionellen Methoden der Qualitätssicherung zurückgreifen, denn die Datenquellen entziehen sich in der Regel der Kontrolle der Datenkonsumenten. Data-Profiling-Methoden, neue Methoden der Datenqualitätsmessung und neue Methoden der Schema-Inferenz werden gebraucht.

Datenqualität ist ein schwer zu fassender Begriff, weil die Kriterien, nach denen die Qualität von Daten beurteilt wird, in der Regel vom Verwendungskontext abhängen. Sehr abstrakt wird Datenqualität oft als „fitness for use“ charakterisiert. Um Datenqualität zu messen und zu verbessern, gibt es heute zahlreiche Methoden, die aber nur selten domänenunabhängig sind und oft nicht auf andere Anwendungskontexte übertragbar sind. Proaktive Methoden zur Qualitätssicherung, wie TDQM (Total Data Quality Management), sind auf Big-Data-Szenarien kaum anzuwenden, denn die zahlreichen Datenquellen entziehen sich zielgerichteten qualitätssichernden Maßnahmen im Datenproduktionsprozess, wenn die Quellen nicht unter der Kontrolle der Datenkonsumenten stehen.

Ein Großteil der Aufwände des Data Scientist fließt somit in Aufgaben, geeignete Quellen zu finden, zu beurteilen und gegebenenfalls nachträglich bedarfsgerecht zu bereinigen und mit geeigneten Metadaten anzureichern. Von zentraler Bedeutung ist dabei für den Data Scientist insbesondere die verschiedenartigen Datenquellen gut zu verstehen, um beurteilen zu können, ob sie für den Verwendungszweck auch geeignet sind. Falsch interpretierten Daten sieht man nicht an, dass sie fehlerhaft, unscharf oder für den intendierten Verwendungszweck gänzlich ungeeignet sind, sie führen aber zu falschen Schlüssen und damit auch falschen oder unbegründeten Handlungsempfehlungen.

Data Profiling bezeichnet ein breites Spektrum an Methoden, die ein Data Scientist anwenden kann, um Datenquellen besser zu verstehen. Das reicht von der Analyse von Datentypen, Wertebereichen, Werteverteilungen, Schlüsseleigenschaften von Attributen, funktionalen Abhängigkeiten

bis zu bedingten funktionalen Abhängigkeiten und Inklusionsanalysen. Felix Naumann weist in einem vielbeachteten Artikel auf die besondere Bedeutung von Data Profiling im Zusammenhang mit Data Science hin und macht auf den dringenden Bedarf an neuen Methoden aufmerksam, die insbesondere für das Profiling nicht-relationaler Daten geeignet sind. Eine ähnliche Einschätzung finden Halevy et al., die im Zusammenhang mit den Erfahrungen aus Googles Goods-Projekt auf die hohe Bedeutung von Werkzeugen zur Verbesserung des Verständnisses von Datenquellen hinweisen.

Data-Profiling-Methoden helfen bei unbekannten Datenquellen, sind jedoch weit davon entfernt die Bedeutung von Daten vollständig erfassen und erklären zu können. Methoden der Schema-Inferenz würde man sich wünschen, damit automatisiert erkannt werden kann, welche Datenquellen vergleichbar oder sinnvoll verknüpfbar sind. Meist lässt sich das aber nicht automatisieren, und dann erfordert die semantische Einordung von Datenquellen einen hohen kognitiven Aufwand beim Data Scientist. Hat der Data Scientist die Quellen einmal verstanden, kann er sie in geeigneten Anfragen zielgerichtet weiterverwenden oder sinnvoll mit anderen Quellen verknüpfen. Das Wissen um die Bedeutung der Daten, das sich der Data Scientist erarbeitet hat, ist in diesen Anfragen implizit enthalten.

Im OCEAN-Projekt an der FAU Erlangen wird versucht das in Anfrageprotokollen versteckte Wissen um die Bedeutung von Datenquellen nutzbar zu machen. Auf diese Weise soll versucht werden, einmal unternommene Anstrengungen für die Datenintegration nicht verpuffen zu lassen, sondern wiederzuverwenden, in der Hoffnung, dass dadurch eine inkrementelle nutzungsorientierte Verbesserung des Datenverständnisses erreicht werden kann.

Fazit und Ausblick

Im Zusammenhang mit der Beurteilung von Datenqualitätsfragen im Bereich Data Science ergeben sich Fragestellungen in folgenden Bereichen:

- Präzisierung und Standardisierung elementarer Datenqualitätskriterien
- Standards zur systematischen Annotation von Quelldaten mit Qualitätsmerkmalen
- Neue Methoden zum Data Profiling für nicht-relationale Datenquellen
- Neue Methoden der Schema-Inferenz

Über den Autor

Prof. Dr. Richard Lenz ist Professor für Datenmanagement an der Universität Erlangen-Nürnberg. Er beschäftigt sich in seiner Forschung u.a. mit dem Thema Evolutionäre Informationssysteme. In diesem Zusammenhang ist auch das Thema Datenqualität von zentraler Bedeutung. Er ist Sprecher des Fachbereichs Informatik in den Lebenswissenschaften in der GI.

Data Science im angloamerikanischen Kontext: Status quo eines modernen Studien- und Berufsbilds

Von Dr. Tarek R. Besold, Lecturer in Data Science an der City, University of London

Im internationalen Kontext zählt die junge Disziplin Data Science zu den momentan sowohl bei Arbeitgebern als auch bei (gegenwärtigen oder zukünftigen) Berufseinsteigerinnen und -einsteigern gefragtesten Berufsfeldern. Die Data Scientist ist hierbei eine mit solider theoretischer Fundierung ausgebildete Allrounderin, die alle wichtigen Schritte des „Data Lifecycle" durchführen, ihre jeweilige Tätigkeit, wie auch erzielte Resultate, kritisch reflektieren und anschließend sowohl den Prozess als auch die erreichten Ergebnisse kommunizieren kann.

Data-Science-Studiengänge erfreuen sich im angloamerikanischen Raum großer Beliebtheit. Kleine und größere Unternehmen aus allen Wirtschaftszweigen, von digitalen Serviceprovidern zu klassischer fertigender Industrie, erhoffen sich Wettbewerbsvorteile durch datenunterstützte Entscheidungsfindung sowie durch (teilweise oder vollständig) datengetriebene Produkte. Studierende und Young Professionals, aber auch Arbeitnehmerinnen und Arbeitnehmer in vormals datenfernen Berufszweigen versprechen sich persönliche Karrierevorteile, aber auch nachhaltigere Karriereverläufe aus dem Erwerb der Fähigkeiten der Data Science.

Die genaue Interpretation von Data Science und die Schwerpunktlegung innerhalb eines Data-Science-Curriculums unterscheiden sich hierbei je nach Profillinie der jeweiligen anbietenden Institution. Generell ist eine Betonung anwendungsnaher und/oder anwendungsrelevanter Ausbildungsinhalte festzustellen. Ein zweiter genereller Trend kann in der gleichzeitigen, wenn auch etwas geringer gewichteten Fokussierung auf solide theoretische Grundlagen – vor allem in Masterstudiengängen, aber auch in grundständigen Studiengängen mit dezidierter Schwerpunktlegung auf Data Science als eigenständiger Disziplin – gefunden werden.

Konzeptuell dienen die theoretischen Lehreinheiten (wie z.B. Module zu statistischen Grundlagen, zu maschinellem Lernen oder zu Datenbanktechnologien im Kontext von Big Data) als Grundlage der angewandten Inhalte (z.B. Datenvisualisierung, Parallel Computing mit Schwerpunkt auf maschinellem Lernen, oder Verarbeitung großer Datenmengen). Zugleich stellt die theoretische Ausbildung die Nachhaltigkeit der erworbenen

Kenntnisse sicher und ermöglicht es der Data Scientist auch zukünftige Entwicklungen mitverfolgen und sich bei Bedarf innerhalb eines dynamischen Technologieumfelds weiterqualifizieren zu können.

Eine Sonderstellung innerhalb der Data-Science-Landschaft nimmt die Unterkategorie der Data Science für wissenschaftliche Zwecke ein. Während die benötigten Grundqualifikationen zu großen Teilen mit der Data Science für den Einsatz im wirtschaftlichen Zusammenhang überlappen, wird für die sinnvolle wissenschaftliche Anwendung in den meisten Fällen zusätzlich Expertise im jeweiligen Wissenschaftsgebiet benötigt. Diese dient u.a. der reibungslosen Kommunikation mit Fachexpertinnen und -experten, ist aber auch den in wissenschaftlichen Zusammenhängen auftretenden spezifischen Fragestellungen und Arbeits- und Datenkontexten geschuldet, welche in ihren speziellen Ausprägungen im Business- oder Industriekontext in vielen Fällen als eher exotisch betrachtet werden können.

Eine wichtige Rolle jenseits der theoretischen oder angewandten „technischen“ Ausbildung spielt im Data-Science-Zusammenhang die Sensibilisierung der Data Scientist für ein kritisches Herangehen an ihre Tätigkeiten – beginnend bei der Einschätzung bestehender und der Entwicklung neuer Anwendungsmöglichkeiten von Data Science, über das Hinterfragen der Validität von erzielten Resultaten, hin zur kritischen Reflexion des Umgangs mit Daten, der epistemischen Stellung von aus Daten gewonnenen Einsichten und der sowohl individuellen als auch sozialen Auswirkungen zunehmend datengetriebener Entscheidungsprozesse in allen Lebensbereichen. Gerade jetzt zu Beginn des „Datenzeitalters“ fällt es in die Verantwortung von Data Scientists als Domänenexpertinnen und -experten diese Entwicklung kritisch zu begleiten und, wo nötig, in entsprechende Bahnen zu leiten.

Als Voraussetzungen für die erfolgreiche Errichtung eines Data-Science-Ausbildungsprogramms spielt der Anschluss an vor Ort bereits vorhandenen Kompetenzen und Schwerpunkten eine wichtige Rolle. Aufgrund der anwendungsverankerten Natur der Inhalte, und des schnellen Fortschritts im Feld, ist es unerlässlich, dass Lehrende auf eigene Erfahrungen im Umgang mit den jeweiligen Lehrinhalten zurückblicken können, aktiv Empfehlungen zu Inhalten und Trends aussprechen, und die Weiterentwicklung der jeweiligen Themengebiete tagesaktuell mitverfolgen und – wenn irgend möglich – mitgestalten. Von einer Einrichtung eines Data-Science-Programms „ex nihilo“ ist daher abzuraten; sowohl im Sinne der Studierenden als auch im Sinne aller Programmbeteiligten auf Seite der Lehrenden.

Fazit und Ausblick

Angesichts des sich nach wie vor beschleunigenden technologischen Fortschritts hin zum Datenzeitalter und dem „Second Machine Age“ wird das Berufsbild der Data Scientist in den nächsten Jahren weiterhin an Popularität und Bedeutung gewinnen. Um diese in immer mehr Gebieten digitale Zukunft mitzugestalten, sind Unternehmen aller Branchen auf Data Scientists angewiesen, welche mit theoretischem Wissen und der Fähigkeit, dieses im Anwendungskontext vielseitig und produktiv zum Einsatz zu bringen, versehen sind. Dieser Bedarfskontext sollte sich auch im Ausbildungszusammenhang widerspiegeln, wo solide Grundlagenausbildung mit der notwendigen mathematischen und theoretischen Tiefe nahtlos in praktische Anwendungsfähigkeiten übergeht.

Über den Autor

Dr. Tarek R. Besold ist Lecturer in Data Science an der City, University of London. Dort forscht er an erklärbaren KI-Systemen, neurosymbolischer Integration sowie Themen an der Schnittstelle zwischen Künstlicher Intelligenz, künstlicher Kreativität und Kognitiven Systemen. Neben Aktivitäten als Organisator von wissenschaftlichen Konferenzen und Workshops sowie verschiedenen Herausgebertätigkeiten hat er die Rolle des Obmanns des DIN-NIA-Arbeitsausschusses zum Thema Künstliche Intelligenz (NA 043-01-42) inne.

DATA SCIENCE ALS INTERDISZIPLINÄRE HERAUSFORDERUNG FÜR DIE WISSENSCHAFT

Data Science zur Gestaltung der Digitalen Transformation

Von Prof. Dr. Key Pousttchi, Universität Potsdam

Der veränderte Umgang mit Daten ist ein wesentlicher Einflussfaktor der Digitalisierung. Aus Sicht der Wirtschaftsinformatik wird dabei einerseits Data Literacy zur conditio sine qua non für Organisationen aller Art, andererseits Data Science zum wichtigsten Gestaltungsmittel der Digitalen Transformation in allen drei Dimensionen Leistungserstellung, Leistungsangebot und Kundeninteraktion. Ihr Einsatz darf dabei nicht isoliert erfolgen, sondern muss sich in das strategische Gesamtkonzept einfügen – häufig wird er dieses maßgeblich beeinflussen oder sogar definieren.

Der Begriff Digitale Transformation bezeichnet erhebliche Veränderungen des Alltagslebens, der Wirtschaft und der Gesellschaft durch die Verwendung digitaler Technologien und Techniken sowie deren Auswirkungen. Typischerweise wird der Begriff im engeren Sinne für die Teilmenge entsprechender Veränderungen von Unternehmen und Branchen verwendet, wobei zwischen den Dimensionen Leistungserstellung, Leistungsangebot und Kundeninteraktion unterschieden wird. Treten die Veränderungen plötzlich und umbruchartig ein, wird hierfür der Begriff Disruption verwendet.[12]

Die Digitale Transformation beruht auf der mittelbaren und unmittelbaren Wirkung des Einsatzes digitaler Technologien und Techniken auf organisatorische und ökonomische Gegebenheiten einerseits und neuartige Produkte und Dienstleistungen andererseits.

Neben der stetig steigenden Rechenleistung und Miniaturisierung klassischer IT-Komponenten ist dabei deren allgegenwärtige Integration in Technik aller Art von Bedeutung, speziell in Verbindung mit:

- flächendeckendem Einsatz von Sensoren und Aktoren einschließlich Audio- und Videoaufzeichnung,
- Einsatz mobiler elektronischer Kommunikationstechniken zur Vernetzung und automatisierten Kommunikation mit sehr geringen Latenzzeiten,
- umfassender Erhebung, Archivierung und Verarbeitung sehr großer Da-

[12] Pousttchi, K.: Digitale Transformation. In: Enzyklopädie der Wirtschaftsinformatik. GITO, Berlin 2017. Online verfügbar: http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/technologien-methoden/Informatik--Grundlagen/digitalisierung/digitale-transformation

tenmengen mittels Big-Data-Techniken,
- verschiedenen Techniken maschinellen Lernens,
- fortgeschrittenen Formen der Mensch-Computer-Interaktion.

Insbesondere die Kombination dieser Faktoren führt zu neuen Potentialen für umfassende Automatisierung im kognitiven und gemischt mechanisch-kognitiven Bereich. Außerdem von Bedeutung für die Integration sind Techniken zur Simulation der Realität für den Menschen (Virtual Reality) und zur Ergänzung der Realität für den Menschen um elektronisch generierte Information (Augmented Reality).

Unter den Faktoren spielen Daten mit Abstand die wichtigste Rolle, ihre Generierung oder Verwendung bildet meist Zweck oder Grundlage des Einsatzes der anderen Faktoren.

Die Digitale Transformation findet in drei Dimensionen statt, die aufeinander aufbauen.

Leistungserstellungsmodell: Die erste Dimension der Digitalen Transformation umfasst den Einfluss auf die Erstellung von Produkten und Dienstleistungen einschließlich der dazu notwendigen Unterstützungsprozesse und der Organisation des Unternehmens. Seit 25 Jahren wissen wir, dass die Steigerung der Produktivität von Unternehmen durch den Einsatz von IT nicht in erster Linie ein technisches, sondern ein organisatorisches Problem darstellt und die Erzielung von Effizienz- und Effektivitätsvorteilen die prozessorientierte Umgestaltung des Unternehmens erfordert, damit der Technologieeinsatz an allen Ecken des "magischen Dreiecks" Kosten-Zeit-Qualität zu Verbesserungen führen kann.

Leistungsangebotsmodell: Die zweite Dimension der Digitalen Transformation umfasst den Einfluss auf die Produkte, Dienstleistungen und Erlösmodelle des Unternehmens. Im Mittelpunkt steht dabei die mittelbare und unmittelbare Wirkung des Einsatzes digitaler Technologien und Techniken auf die Verbesserung bestehender Produkte und Dienstleistungen, auf das Angebot neuer oder sogar neuartiger Produkte und Dienstleistungen sowie auf Veränderungen der zugehörigen Erlösmodelle.

Kundeninteraktionsmodell: Die dritte Dimension der Digitalen Transformation umfasst den Einfluss auf Art und Inhalt der Interaktion mit Kunden. Wesentliche Kennzeichen sind die kanalübergreifende und ganzheitliche Gestaltung der Kundenbeziehung und die Einbeziehung automatisierter Kommunikation und moderner Formen der Datenanalyse.

Die Bedeutung von Data Science für das Unternehmen ist dabei in den drei Dimensionen stetig steigend (vgl. Abb. 1): Während die Prozessgestaltung vor allem Data Literacy voraussetzt und Data Science eher in der Optimierung zum Einsatz kommt, bildet letztere im Leistungsangebotsmodell bereits ein wesentliches Element und wird im Kundeninteraktionsmodell zur strategisch entscheidenden Größe.

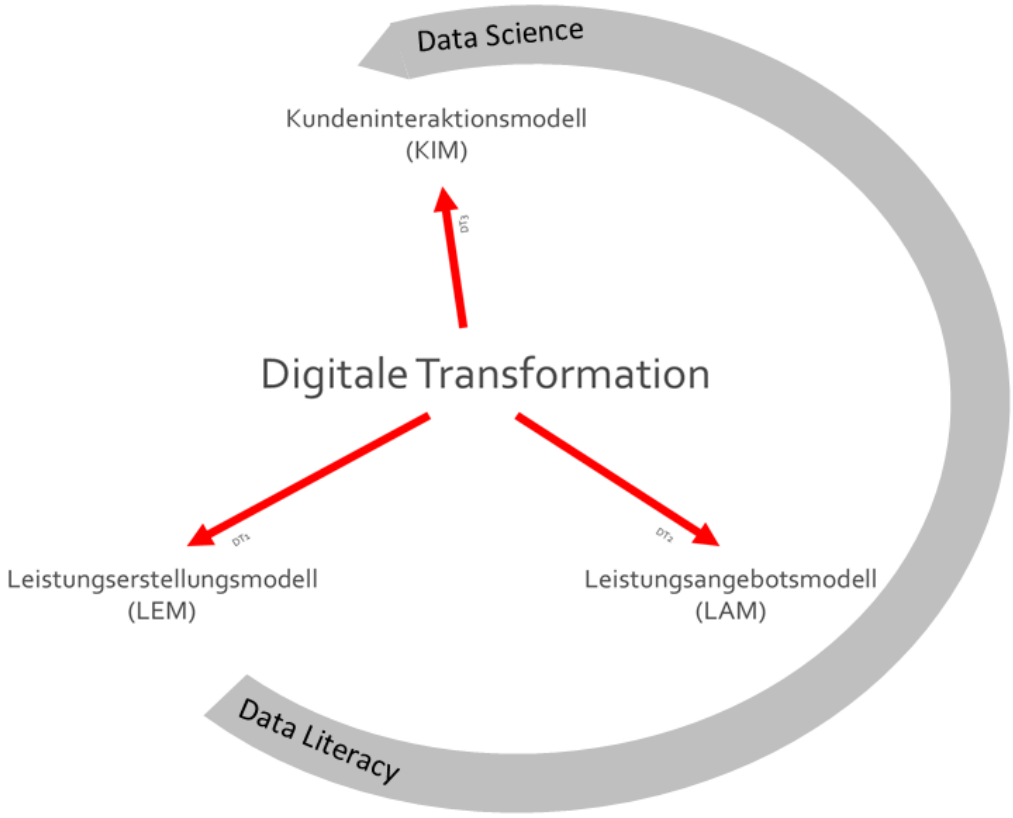

Abbildung 5: Einfluss von Data Literacy und Data Science auf die Dimensionen der Digitalen Transformation

Über den Autor

Prof. Dr. Key Pousttchi ist Inhaber des SAP-Stiftungslehrstuhls für Wirtschaftsinformatik und Digitalisierung an der Universität Potsdam. Ziel seiner Aktivitäten in Forschung und Lehre sind die systematische Erforschung der Digitalisierung und die Gestaltung der Digitalen Transformation von Wirtschaft und Gesellschaft mit ingenieurmäßigen Methoden. In der GI ist er Sprecher der Fachgruppe Mobilität und Mobile Informationssysteme (MMS).

Interdisziplinäre Daten-Laboratorien als Antwort auf die „Data Challenge“

Von Prof. Dr. Karl Mannheim, Julius-Maximilians-Universität Würzburg, Sprecher des Arbeitskreises Physik, moderne Informationstechnologie und Künstliche Intelligenz der Deutschen Physikalischen Gesellschaft e.V., und Dr. Kai Polsterer, Heidelberger Institut für Theoretische Studien

Data Scientists/Engineers verfügen über Grundlagen aus Informatik, Mathematik oder Physik sowie Spezialwissen über Statistik, Machine Learning, skalierbares Datenmanagement und Data Mining bzw. integrierte Systeme. Das Wissen sollte anwendungsorientiert vermittelt werden. Die Universitäten sind bestens aufgestellt, anwendungsnahe Fragestellungen in die Curricula einzubringen. Dazu eignen sich Forschungspraktika in „Data-Labs“ mit leistungsfähiger IT-Infrastruktur, Mentoren und interdisziplinärer Kommunikationskultur.

In Physik und Astronomie werden große und heterogene Datenströme durch die digitalisierte Auslese von Signalen aus Detektoren und Sensoren gewonnen und nach entsprechender Aufbereitung analysiert. Ziel ist es dabei, möglichst exakte Messungen auszuführen und eine konkrete Fragestellung probabilistisch im Sinne der Informationstheorie zu beantworten. Durch die Anwendung statistischer Methoden wird dabei ein grundlegendes Verständnis der Datenunsicherheiten gewonnen und die Datenqualität durch Validierungsverfahren sichergestellt. Artefakte sollen als Ergebnis einer Datenanalyse ausgeschlossen werden.

Aufgrund der Leistungsfähigkeit moderner breitbandiger Vielkanal-Digitalwandler sind die Anforderungen an die datenverarbeitende Hardware und Software teilweise enorm. Sie können Impulse für die gezielte Entwicklung neuartiger Komponenten oder ganzer Computerarchitekturen geben (z.B. Photonik, Quantum-Computing). Eine wichtige Rolle spielt auch das autonome Datenmanagement bzw. die Vorverarbeitung bei integrierten Systemen mithilfe von KI-Methoden.

Die Komplexität des Datenmanagements erreicht in vielen Forschungsprojekten bereits die Stufe einer „Data Challenge“ und muss als interdisziplinäre Aufgabe von Physik und Informatik gesehen werden. Eine Professionalisierung der Lösungsansätze mithilfe der wissenschaftlichen Methodik der Data Science wird als notwendig angesehen, und es besteht

dringend Handlungsbedarf auf der Ausbildungsseite an den Universitäten. Zwar stellen Programmierkenntnisse und die mathematisch fundierte Ausbildung von Physikern und Astronomen eine sehr gute Grundlage für die Behandlung datenwissenschaftlicher Probleme dar, sie reichen aber nicht aus, um strukturierte Beiträge zu großen „Data Challenges“ zu leisten. Eine Spezialisierung der Physiker und Astronomen auf die datenwissenschaftlichen Probleme würde auf Kosten der physikalisch-astronomischen Inhalte stattfinden und wird daher als nicht zielführend angesehen. Um den existierenden Bedarf von Physik und Astronomie an Informatik abzudecken, haben sich bereits interdisziplinäre Fachprofile wie die „Astro-Informatik“ ausgebildet.

Die Förderung der Interdisziplinarität von Physik/Astronomie, Mathematik und Informatik schafft eine optimale Win-win-Situation. Physik und Astronomie liefern Daten und konkrete Fragestellungen, während die Informatik die praktische Methodik für die Optimierung bzw. Realisierung der Datenanalyse liefern kann. Die Mathematik und Theoretische Physik ergänzen die Grundlagen für die korrekte wahrscheinlichkeitstheoretische Behandlung und Interpretation der Daten. Durch die Impulse aus Physik und Astronomie entstehen dann im Dreieck dieser Wechselbeziehungen neue Methoden der Datenanalyse (wie bisher z.B. die Monte-Carlo-Methoden oder die im Maschinenlernen benutzten Mean-Field-Methoden).

Im Ausland bereits vielfach erprobt, eignen sich Forschungspraktika in Kombination mit selbstständigem wissenschaftlichem Arbeiten bestens, um gemeinsame Lösungen zu erarbeiten. „Data Scientists“ sollten sich in verschiedenen Anwendungsfeldern auskennen und brauchen dafür ein geeignetes Ökosystem an den Universitäten. Dabei ist es wichtig, dass Studierende bereits im Bachelorstudium erste Erfahrungen mit der interdisziplinären Arbeit machen können, die sie dann im Masterstudium und im Promotionsstudium vertiefen. Dadurch etabliert sich die erforderliche Kommunikationskultur auch unter den betreuenden Professoren und Assistenten.

Als Ort für diese interdisziplinäre Kommunikationskultur empfehlen sich Daten-Laboratorien, die von den Vertretern verschiedenster Fachrichtungen für Forschungspraktika in der Data Science genutzt werden. Sie verfügen über eine erstklassige IT-Infrastruktur sowie betreuendes Personal. Data Science ist mehr als die bloße Anwendung von Methodik, sie beinhaltet die erkenntnisbringende kritische Auseinandersetzung mit den Ergebnissen und ihre wissenschaftliche Interpretation. Daten-Laboratorien sind der natürliche Ort für eine innovative, lebendige Gründerszene.

Fazit und Ausblick:

Komplexe Anwendungen sind Triebfedern für die Entwicklung einer zukunftsfähigen wissenschaftlichen Methodik der Data Science. Physik und Astronomie bieten eine große Diversität von Fragestellungen, die für interdisziplinäre Forschungspraktika geeignet sind. Zur Durchführung der Praktika müssen Universitäten mit Daten-Laboratorien ausgestattet werden, die über wissenschaftliche Datenzentren mit dedizierter Infrastruktur (z.B. GPU-Farmen, IO-optimierte Architekturen) und betreuendes Personal verfügen.

Über die Autoren:

Prof. Dr. Karl Mannheim ist Ordinarius für Astrophysik an der Julius-Maximilians-Universität Würzburg. Er beschäftigt sich mit der Analyse sehr großer Datenströme von astronomischen Observatorien. Er ist Sprecher der Arbeitsgruppe für das Square Kilometre Array (SKA) im deutschen GLOW-Konsortium und des Arbeitskreises für Physik, moderne Informationstechnologie und Künstliche Intelligenz in der Deutschen Physikalischen Gesellschaft (DPG).

Dr. Kai Polsterer ist Junior-Gruppenleiter für Astro-Informatik und entwickelt innovative Datenanalysemethoden am Heidelberger Institut für Theoretische Studien (HITS).

Data Science in den Sozialwissenschaften

Von Dr. Ingo Scholtes, Chair of Systems Design, ETH Zürich

Die Digitalisierung der Gesellschaft führt zu einer zunehmenden Verflechtung von Informatiksystemen mit gesellschaftlichen Akteuren und Prozessen. Eine Konsequenz ist die Verfügbarkeit großer Datensätze, welche ein digitales Abbild menschlichen Verhaltens, zwischenmenschlicher Interaktionen und sozialer Strukturen liefern. Die Analyse solcher Datensätze mittels Data-Science-Methoden wirft einerseits datenschutzrechtliche und ethische Fragen auf. Andererseits bietet sie das Potenzial, unser Verständnis gesellschaftlicher Prozesse grundlegend zu verbessern.

Mehr und mehr Aspekte unseres täglichen Lebens hinterlassen digitale Spuren. Dies gilt für die durch Digitalisierung geprägte Arbeitswelt ebenso wie für den privaten Internet- und Medienkonsum, das Einkaufsverhalten oder die zwischenmenschliche Kommunikation auf digitalen Plattformen. Hierbei anfallende Datensätze liefern einen immer genaueren „digitalen Schatten" menschlichen Verhaltens und sozialer Interaktionen.

In der wichtigen laufenden Debatte über Chancen und Risiken, die sich aus der Anwendung von Data-Science-Methoden auf solche Datensätze ergeben, steht meist die Abwägung zwischen ethischen sowie datenschutzrechtlichen Anforderungen im Vordergrund. Das gewaltige Potenzial für die Fortentwicklung der empirischen Sozialforschung im 21. Jahrhundert tritt demgegenüber häufig in den Hintergrund.

Dabei bieten Data-Science-Methoden neuartige Möglichkeiten für Exploration und Analyse großer Datensätze zu menschlichem Verhalten. Gleichzeitig versprechen moderne rechnergestützte Simulations- und Modellierungsverfahren, insbesondere durch den Abgleich mit empirischen Daten, Einsichten in jene Wirkmechanismen, die kollektiven sozialen Phänomenen zugrunde liegen. Die Kombination dieser Ansätze bietet das Potenzial, unser Verständnis gesellschaftlicher Phänomene grundlegend zu verbessern.

Die Untersuchung sozialwissenschaftlicher Fragestellungen mittels Data-Science Methoden liegt im Fokus eines interdisziplinären Forschungsgebiets, welches häufig unter dem Begriff „Computational Social Science" zusammengefasst wird. Einige Sozialwissenschaftler setzen die sich hieraus ergebenden Möglichkeiten mit einer methodischen Revolution gleich. Gleichzeitig entstehen aber

auch neue Herausforderungen, nicht nur für die Sozialwissenschaften, sondern auch für die Informatik. Gemessen an den Bekenntnissen von Hochschulen und Förderinstitutionen zur Bedeutung von Interdisziplinarität sind echte „inter"-disziplinäre Arbeiten an der Schnittstelle zwischen Informatik und Sozialwissenschaften, d.h. Arbeiten, in denen sozialwissenschaftliche Theorien mittels Informatikmethoden getestet und weiterentwickelt werden, leider nach wie vor die Ausnahme. Demgegenüber steht eine steigende Zahl „datengetriebener" Studien, welche Korrelationen und Muster in großen Datenmengen bspw. aus sozialen Medien aufzeigen, ohne jedoch Aufschluss über zugrundeliegende Wirkmechanismen geben zu können.

In der Tat erwecken viele dieser Studien den Eindruck, dass vielmehr die Analyse der Daten und nicht die Beantwortung einer wissenschaftlichen Frage im Fokus des Interesses steht. Sie können also lediglich ein erster Schritt in Richtung einer „theoriegetriebenen" und „datengestützten" Forschungsdisziplin sein, die den Namen Computational Social „Science" verdient. Denn die Erklärung sozialwissenschaftlicher Phänomene und das Aufzeigen kausaler Mechanismen erfordern mehr als nur Datenanalysekompetenzen.

Wissenschaftstheoretische Grundlagen sind hierzu von ebenso großer Bedeutung wie Domänenwissen zu soziologischen Theorien und Methoden. Darüber hinaus sind Kompetenzen in der Modellierung kollektiver Phänomene in komplexen Systemen wechselwirkender Agenten gefragt, welche in der statistischen und interdisziplinären Physik von großer Bedeutung sind.

Um sinnvoll von den Möglichkeiten Gebrauch zu machen, die sich dank neuer Datenquellen und Data-Science-Methoden ergeben, müssen Informatikcurricula daher ebenso weiterentwickelt werden wie sozialwissenschaftliche Studiengänge. Hierbei gilt es Studenten in die Lage zu versetzen u.a. folgende Fragen zu beantworten: Wie lassen sich sozialwissenschaftliche Theorien und Hypothesen mittels Data Science überprüfen? Wie aussagekräftig und repräsentativ sind Ergebnisse von Studien, welche bspw. auf öffentlich zugänglichen Daten aus sozialen Medien basieren? Wie können Stichprobenverzerrungen in solchen Daten erkannt, quantifiziert und gegebenenfalls korrigiert werden? Welche spezifischen Herausforderungen bspw. für Methoden des maschinellen Lernens oder der sozialen Netzwerkanalyse ergeben sich durch fehlerbehaftete, unvollständige und zeitlich aufgelöste Datensätze? Und welche Rolle spielt Data Science in der Theoriebildung?

Was also ist Computational Social Science? Eine „Hilfswissenschaft", die sich in die lange Liste erfolgreicher Computational Sciences und Bindestrich-Informatiken wie Bio-, Wirtschafts-, Medizin-, Geo-, Umwelt-,

Medien- und Agrarinformatik einreiht? Ich bin überzeugt, dass Computational Social Science in dieser Liste eine Sonderstellung hat. Denn neben dem unbestreitbaren Potenzial für die Sozialwissenschaften ergeben sich aus der Konvergenz sozialer und technischer Systeme auch neue Problemfelder für die Informatik. Denn die Rückkopplung technischer und sozialer Aspekte in Informatiksystemen führt zu einer Komplexität, die mit existierenden Ansätzen des Systementwurfs nur schwer beherrschbar ist.

Es ergeben sich wichtige Fragen, die sich zum Teil mit den von der Gesellschaft für Informatik formulierten Grand Challenges decken und deren Beantwortung eine quantitative Modellierung sozialer Aspekte erforderlich macht. Welche neuartigen systemischen Risiken ergeben sich in weltumspannenden soziotechnischen Systemen? Welche sozialwissenschaftlichen Erkenntnisse müssen im Entwurf resilienter technischer Systeme berücksichtigt werden? Inwiefern beeinflussen Mechanismen von Informatiksystemen (bspw. intelligente Empfehlungssysteme, Reputationsmechanismen etc.) soziale Phänomene wie Polarisierung oder Diskriminierung? Welche Methoden der Datenanalyse und -modellierung können wir nutzen, um solche Phänomene zu quantifizieren, vorherzusagen oder sogar zu beeinflussen? Und welche neuen Ansätze ergeben sich für Analyse und Management menschlicher Aspekte in der kollaborativen Softwareentwicklung?

Die Anwendung von Computational-Social-Science-Methoden auf große Datensätze soziotechnischer Systeme verspricht Antworten auf diese wichtigen Fragen. Aus diesem Grund ist Computational Social Science nicht nur für die Sozialwissenschaften, sondern auch für die Informatik von großer Bedeutung.

Fazit und Ausblick

Data-Science-Methoden sind für die Sozialwissenschaften von zunehmender Bedeutung. Sie liefern neue Ansätze zum Testen sozialwissenschaftlicher Theorien und erweitern den Methodenkanon der empirischen Sozialforschung substanziell.

1. Die Analyse unvollständiger, fehlerbehafteter und zeitlich aufgelöster Daten zu sozialen Systemen stellt Anforderungen, die von existierenden Methoden, bspw. der sozialen Netzwerkanalyse oder dem maschinellen Lernen, allenfalls zum Teil erfüllt werden. Für die Informatik ergeben sich damit neue Herausforderungen sowohl für die Forschung wie auch für die Ausbildung einer Generation kritischer Datenwissenschaftler.
2. Während die kritische Reflexion empirischer Forschungsmethoden in den Sozialwissenschaften wichtiger Aspekt der Ausbildung ist, fehlen in Informatikstudiengängen häufig sowohl statistische

wie auch wissenschaftstheoretische Grundlagen, die für den Einsatz von Data-Science-Methoden in den Wissenschaften von herausragender Bedeutung sind.

3. Aller Sonntagsreden zur Bedeutung interdisziplinärer Forschung zum Trotz bestehen erhebliche kulturelle und wissenschaftspolitische Hürden, die der effektiven Zusammenarbeit von Informatikern und Sozialwissenschaftlern im Wege stehen. Der Abbau dieser Hürden in Forschungsförderung, Ausbildung und wissenschaftlichen Anreizsystemen und Strukturen ist für die weitere Entwicklung der Sozialwissenschaften wie auch der Informatik von entscheidender Bedeutung.

Über den Autor

Dr. Ingo Scholtes ist Oberassistent am Lehrstuhl für Systemgestaltung der ETH Zürich. In seiner Forschung beschäftigt er sich mit Methoden zur Analyse unstrukturierter, fehlerbehafteter und zeitgestempelter Daten aus den Wissenschaften. Er ist Juniorfellow und gewähltes Mitglied des Präsidiums der Gesellschaft für Informatik. Gemeinsam mit Prof. Dr. Markus Strohmaier ist er Gründungsvorsitzender des Arbeitskreises Computational Social Science. Er ist zudem Mitglied im Fachverband Physik sozio-ökonomischer Systeme der Deutschen Physikalischen Gesellschaft. Am Institut für Informatik der Universität Zürich baut er aktuell eine neue, vom Schweizerischen Nationalfonds zur Förderung der wissenschaftlichen Forschung finanzierte Forschungsgruppe zu Data Analytics auf.

Data Science aus Sicht der Mathematik

Von Prof. Dr. Sebastian Stiller, Institut für Mathematische Optimierung an der Technischen Universität Carolo-Wilhelmina zu Braunschweig und Deutsche Mathematiker Vereinigung e.V. (DMV)

Data Science ist eine Methodenwissenschaft. Sie erforscht und entwickelt Methoden, um aus Daten Erkenntnisse abzuleiten, wenn die einfachen, allgemeinverständlichen Methoden der Empirie nicht mehr ausreichen. Kurz: Es geht darum in scheinbar „sinnlosen" Datenmengen Struktur zu erkennen. Nachhaltige Innovationen und ein verantwortlicher Umgang mit diesen Datenmengen erfordern das hochqualitative, anspruchsvolle Studium.

Fachlich gesehen sind „sinnlose Datenmengen" Punktwolken in einem sehr hochdimensionalen Raum. Dort erscheinen sie ohne Struktur. Data Science versucht darin Struktur, d.h. niederdimensionale, gut beschreibbare Objekte, zu finden, die die Punktwolke beschreiben. Diese Strukturen können unterschiedlich sein: algebraisch, topologisch, diskret (etwa Netzwerke) oder „modellfrei". Welche Struktur „gut" ist, hängt von der Anwendung, von der Berechenbarkeit und von den Daten ab.

Doch wozu braucht es Data Scientists? Die Methoden sind zu elaboriert, als dass Experten der Anwendungsgebiete in Wissenschaft, Technologie und Wirtschaft sie nebenher lernen könnten. Deshalb lohnt es sich für die Anwendungen ausgesprochene Methodenexperten mit Grundkenntnissen der Anwendung hinzuziehen.

Data Science ist eine junge Wissenschaft. Viele grundlegende Fragen sind noch offen. Viele zentrale Ergebnisse und Methoden werden erst gefunden werden. Deutschland hat sehr gute Voraussetzungen, um diese Innovation führend mitzugestalten. Das muss unser Ziel sein. Gegenwärtig stehen drei erfolgreiche Aspekte im Vordergrund:

(1) Lerntheorie (insbesondere neuronale Netze),
(2) Zentralitätsmaße in Netzwerken und
(3) statistische Verfahren wie etwa LASSO.

Reduziert man die Data-Science-Ausbildung auf die Vermittlung des bisher Bekannten, vergibt man die Chance auf Innovationsführerschaft. Deshalb brauchen wir grundlegend ausgebildete Data Scientists.

Data Science und klassische Empirie

Data Science entwickelt Methoden, die in mindestens drei Aspekten von der klassischen Empirie abweichen, denn Data Science versucht Erkenntnisse aus Daten zu generieren:

(1) ohne vorher aus anderen Gründen eine Hypothese gebildet zu haben,
(2) ohne dass die Experimente wiederholbar wären,
(3) selbst wenn nur schwache Korrelationen bestehen.

Die dabei verletzten, empirischen Prinzipien sind nicht falsch oder überholt. Aber man kann, wo sie nicht zu erfüllen sind, unter bestimmten Bedingungen und mit wesentlichen Abstrichen an dem Statut der Erkenntnis selbst (Stichwort Korrelation und Kausalität) Erkenntnisse gewinnen oder Hinweise dazu bekommen, welche klassischen, aufwändigen Experimente erfolgversprechend sind.

Data Science ersetzt nicht die klassische, wissenschaftliche Methodik, sondern ergänzt sie.

Risiken und verantwortlicher Umgang

Erkenntnistheoretisch sind Data-Science-Methoden oft prekär. Sie sollen es auch sein, denn es geht darum, auch die erkenntnistheoretischen Grenzbereiche auszunutzen. Es besteht die reale Gefahr, dass sich elaborierte Methoden und Systeme etablieren, deren Rolle in der Anwendung methodisch nicht zu rechtfertigen ist, deren Elaboriertheit aber verhindert, dass Anwender sich dieses Fehlers bewusst sind. Hier muss ein Data Scientist für einen verantwortlichen Umgang mit den prekären Methoden sorgen. Deshalb brauchen wir grundlegend ausgebildete Data Scientists.

Wie Data Science studieren?

Grundlegend ausgebildete Data Scientists müssen:

(1) die Vielfalt der Strukturen kennen und entwickeln können, mit denen Daten analysiert werden
(2) die Methoden kennen, mit denen diese Strukturen gesucht werden
(3) die Fragen und die Angemessenheit der Methoden für zumindest ein Anwendungsgebiet beurteilen können

Daher ist ein Data-Science-Studium ein Mathematikstudium mit Informatikanteil oder ein Informatikstudium mit einem stärkeren Mathematikanteil, jeweils mit Grundkenntnissen und Abschlussarbeiten in einer Anwendung. Für Ersteres benötigt man Strukturwissenschaften wie etwa Algebra, diskrete Mathematik sowie Stochastik und Statistik, für Zweiteres bedarf es tiefergehender Kenntnisse von Algorithmen, Datenstrukturen, Optimierung und Numerik. Zusätzlich sollte eine Erkenntnis- oder Wissenschaftstheorie gehört werden.

Fazit und Ausblick

Diese Ansprüche sind sehr hoch. Sollte man einen „Data Scientist light“ ermöglichen, für den nur die Verwen-

dung der existierenden Methoden gelehrt wird? Ist es ein deutscher Sonderweg, auf die hochqualitative Ausbildung zu setzen? Data Scientist light kann auch von Statistikern, Mathematikern und Informatikern mit Zusatzkursen geleistet werden. Nachhaltige Innovation und verantwortlicher Umgang erfordern das hochqualitative, anspruchsvolle Studium. Dies ist auch der Weg, den Top-Institutionen in den USA gehen. Wir sollten uns nicht am Mittelmaß orientieren.

Über den Autor

Prof. Dr. Sebastian Stiller ist angewandter Mathematiker. Er entwirft und analysiert Algorithmen. In Erlangen und Leuven studierte er Mathematik und Philosophie. Später forschte er an der TU Berlin und am Massachusetts Institute of Technology (MIT). Seit 2015 ist er Professor für Mathematik am Institut für Mathematische Optimierung der Technischen Universität Carolo-Wilhelmina Braunschweig und Vertreter der Deutschen Mathematiker Vereinigung e.V. (DMV).

Die Bedeutung von Data Science für die Chemie

Von Prof. Dr. Stefan M. Kast, Technische Universität Dortmund / Fachgruppe Computer in der Chemie (CIC) der Gesellschaft Deutscher Chemiker (GDCh)

Die in der Chemie erzeugten umfangreichen Datenmengen werden bislang nicht hinreichend genutzt, da die grundlegenden datenwissenschaftlichen Methoden in der Fachausbildung vernachlässigt werden. Während Data Science in der Biologie durch Studiengänge der Bioinformatik zumindest sichtbar ist, trifft dies für die Chemie praktisch nicht zu. Ziel ist, Chemie und benachbarte Naturwissenschaften sowohl als optionale Domain-Science-Komponenten in Vollstudiengängen Data Science zu verankern als auch koordinierte Spezialmodule mit datenwissenschaftlichem Hintergrund in den Fachstudien anzubieten.

Chemie und benachbarte Disziplinen wie chemische Biologie erzeugen umfangreiche Datenmengen unterschiedlicher Qualität aus den verschiedensten Quellen. Diese werden auch bislang schon in typischen Datenbanken (z.B. Reaktionen, Moleküleigenschaften, Spektren, Strukturen, Sequenzen, biologische Aktivität, Imaging) gesammelt, wobei die Unsicherheit mit zunehmendem „biologischem Charakter“ der Datenquelle im Regelfall zunimmt. Auch wenn Studierende regelmäßig mit diesen Daten konfrontiert werden bzw. mit ihnen arbeiten müssen, bleibt die fundierte statistische Analyse zum Großteil dem sehr kleinen Anteil theoretisch arbeitender Spezialisten vorbehalten, da die notwendigen Grundlagen in der Fachausbildung bislang weitgehend vernachlässigt werden.

Aus diesem Grund wird der prinzipielle Mehrwert von Wissen, das allein durch fundierte statistische Analyse, computergestützte Verarbeitung und maschinelles Lernen anhand von Daten erzeugt werden kann, weitgehend nicht realisiert. Dieses ungenutzte Potenzial hat massiven negativen Einfluss auch auf die Beschäftigungsaussichten der Studierenden, da Firmen wie z.B. BASF die Digitalisierung auf allen Ebenen massiv vorantreiben, ohne in ausreichendem Maß das hierfür geeignet qualifizierte Personal zu finden. Der Hebel zur Verbesserung der Situation liegt in der grundlegenden Ausbildung.

Insofern muss das Momentum der aufkommenden Data Science-Studiengänge, das sich auf viele Bereiche der MINT-Fächer bereits ausgewirkt hat, auf die Chemie und ihre Nachbarfächer übertragen werden. In der fachlichen Grundausbildung im Bereich Mathematik in der Chemie spielen statistische Aspekte praktisch

keine Rolle, stattdessen werden die Grundlagenmodule dominiert durch eine Untermenge der typischerweise für Ingenieurstudiengänge gelehrten Mathematikinhalte, von z.B. linearer Algebra bis Differentialgleichungen.

In bspw. der Biologie und Pharmazie wird grundlegende Statistik zwar regelmäßig angeboten, allerdings auf die Bereiche der höheren Mathematik verzichtet. Weder in der Chemie noch in den Nachbarschaftsdisziplinen spielen fundierte Programmier- oder Informatikkenntnisse eine Rolle in den Pflichtcurricula. Vielmehr hat die Biologie – im Gegensatz zur Chemie – durch die Etablierung von Bioinformatik-Studiengängen ein Angebot und eine Präsenz geschaffen, die dem Ideal der drei Säulen der Data-Science-Studiengänge (Mathematik/Informatik/„Domain Science“) am nächsten kommt und entsprechend ausgebildete Absolventinnen und Absolventen erzeugt, die vom Arbeitsmarkt auch unmittelbar absorbiert werden.

Da die Ausbildung der (Bio-) Chemie stark durch praktische Anteile und im Bereich der theoretischen Chemie hauptsächlich durch angewandte Quantenmechanik geprägt ist, wird eine Erweiterung der Curricula hin zu „Data Science“ sehr schwierig zu gestalten sein. Gleichzeitig wird eine „Chemieinformatik“ (obwohl in der Forschung ein etabliertes Feld) als eigenständiger Studiengang Akzeptanzprobleme haben und kaum zu realisieren sein.

Es ergeben sich somit zwei Perspektiven: Zum einen können ausgewählte, relevante Inhalte von „Data Science“-Modulen erarbeitet werden, die zunächst im Wahlpflichtbereich von Bachelor- und Masterstudiengängen in Chemie und benachbarten Fächern angeboten werden können. Nach einer Evaluationsphase der Akzeptanz und Relevanz für den Arbeitsmarkt könnte die Motivation für die Überführung in den Pflichtbereich gegeben sein. Zum anderen sollte (Bio-) Chemie als „Domain Science“-Komponente inhaltlich anhand relevanter Fragen aus der Praxis verankert und in Form von Modulinhalten vereinheitlicht werden. Studierende der Data Science erhalten auf diesem Weg eine weitere Wahlmöglichkeit eines bislang unterrepräsentierten Gebiets. Hierbei kann auf die Erfahrungen der Bioinformatik-Studiengänge zurückgegriffen werden.

Fazit und Ausblick:

- Die Inhalte von Data-Science-Grundlagenmodulen im Bachelor- und Masterbereich sollten bundesweit einheitlich, differenziert nach Chemie und benachbarten Disziplinen, formuliert werden.
- (Bio-) Chemie als „Domain Science“-Komponente in Data-Science-Studiengängen sollte in-

haltlich erarbeitet werden, analog zu den biologischen Inhalten in der Bioinformatik-Ausbildung.

- Diese beiden Aspekte erfordern die enge Abstimmung mit den Curricular-Kommissionen der Fachgesellschaften.

Über den Autor:

Prof. Dr. Stefan M. Kast vertritt das Gebiet Theoretische Physikalische Chemie an der Fakultät für Chemie und Chemische Biologie der Technischen Universität Dortmund und ist seit 2017 Vorstandsvorsitzender der Fachgruppe „Computer in der Chemie“ (CIC) der Gesellschaft Deutscher Chemiker (GDCh). Er forscht und lehrt im interdisziplinär geprägten Bereich der Modellierung und Simulation chemischer und biologischer Systeme, was auch die Verarbeitung und statistische Analyse großer Datenmengen erfordert.

DATA SCIENCE EDUCATION IN DER PRAXIS

Data Science: Von der Wissenschaft zum Erfolgsfaktor für Unternehmen

Von Thomas Bendig, Fraunhofer-Verbund IUK-Technologie, Berlin

Immer mehr Unternehmen erkennen das Potenzial der heute verfügbaren Datenmengen und müssen sich als „Data-driven Company" neu definieren, um im Wettbewerb bestehen zu können. Mit prädiktiven Modellen gewinnen sie aus den Daten Prognosen für Entscheidungen und Maßnahmen auf allen Geschäftsebenen. Dazu benötigen sie Teams mit einem besonderen Mix an Kompetenzen. Für diese Teams sind „Data Scientists", die Konzepte und Techniken aus Informatik, Statistik und Mathematik zu nutzen wissen, von enormer Bedeutung.

Die Fraunhofer-Gesellschaft hat diesen rasant wachsenden Bedarf, der sich nicht allein durch Absolventen neuer Studiengänge decken lässt, frühzeitig erkannt. Im Kontakt mit den Unternehmen zeigt sich, dass bei der strategischen Einführung von Big Data genau diese Kompetenzen noch an vielen Stellen fehlen und Unterstützung beim Aufbau von Big-Data-Know-how notwendig ist. An diesen Bedarf knüpft das Fraunhofer-Schulungs- und Weiterbildungsprogramm seit 2013 an: Mit flexibel gestalteten Fortbildungsmodulen richtet es sich an Führungskräfte, die fit für Big Data werden möchten, und an Fachkräfte, die sich kompakt zu Data Scientists weiterbilden möchten.

Business Developer profitieren von diesem Schulungsprogramm für die Unternehmensentwicklung – etwa für neue Geschäftsmodelle, individualisierte Angebote, smartere Produkte oder die Optimierung von Geschäftsprozessen. Analysten erfahren, wie sie mit maschinellen Lernverfahren prädiktive Modelle entwickeln, um neue Trends in Daten aufzuspüren, Prognosen zu erstellen und Handlungsoptionen abzuleiten. Software-Ingenieure lernen, mit modernen Datenbanken, verteilter Speicherung und hocheffizienten Technologien robuste, skalierbare Lösungen zu entwickeln, um Massendaten datenschutzkonform auszuwerten und die Ergebnisse sicher in die Unternehmens-IT einzuspeisen.

Die Fraunhofer-Gesellschaft steht für Spitzenforschung auf höchstem Niveau. Die Institute haben jahrzehntelange wissenschaftliche Expertise in den Bereichen Data Mining, maschi-

nelles Lernen und Mustererkennung und setzt diese für innovative Entwicklungen in nationalen und internationalen Forschungs- und Entwicklungsprojekten ein. Die Kursleiter sind Autoren wichtiger wissenschaftlicher Publikationen.

Das Angebot beinhaltet sowohl Schulungen zu Big-Data-Grundlagen als auch Module für spezifische Anwendungsfelder. Die drei- bis fünftägigen zertifizierten Schulungen mit jeweils zwei erfahrenen Dozenten und maximal zwölf Teilnehmenden bieten beste Möglichkeiten, auf alle Data-Science-Aspekte und individuelle Bedürfnisse einzugehen. Die Inhalte orientieren sich praxisnah am neuesten Stand der Wissenschaft.

Die Fraunhofer-Weiterbildungsangebote umfassen einerseits folgende branchenneutralen Kurse:

- Zertifizierter Data Scientist Basic Level
- Zertifizierter Data Analyst
- Zertifizierter Data Scientist Machine Learning
- Basic Data Analytics
- Data Analytics Potentials and Realization
- Big Data Architecture
- Big Data Analytics
- Visual Analytics
- Multimedia Analytics
- Linked Enterprise Information Integration
- Deep Learning

Darüber hinaus gibt es einige branchenorientierte Angebote:

- Data Scientist for Smart Energy Systems
- Energy Analyst
- Smart Data und Big Data für Industrie 4.0
- Data Scientist for Smart Buildings
- Multimedia Analytics
- Text Analytics in Life Sciences

Fazit und Ausblick

Ein Data Scientist sollte neben Kenntnissen in Mathematik, Statistik, IT und Programmierkenntnissen vor allem auch Fachwissen (Domain-Knowledge) aus der jeweiligen Branche mitbringen, da sonst die Interpretation der gewonnenen Aussagen schwierig bis unmöglich ist. Kompakte Data-Science-Weiterbildungsangebote an verschiedene Branchen haben sich als gute Möglichkeit erwiesen, Branchenexperten mit Data-Science-Kenntnissen auszustatten.

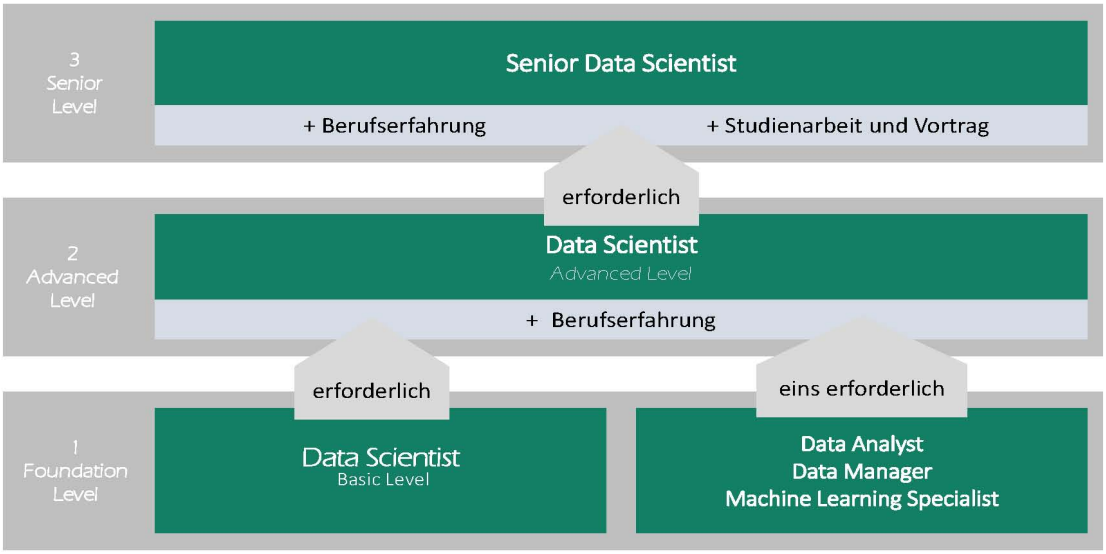

Abbildung 6: Die Kurse der Fraunhofer Academy haben unterschiedliches Niveau und bauen teilweise aufeinander auf

Über den Autor

Thomas Bendig ist Forschungskoordinator des Fraunhofer-Verbundes IUK-Technologie, der mit 21 Fraunhofer-Instituten größten europäischen Forschungsorganisation für angewandte IT-Forschung. Koordination, Strategieentwicklung und Technologietransfer bilden den Schwerpunkt seiner Tätigkeit an der Schnittstelle zwischen Forschung, Industrie und Politik.

Data Scientists in Deutschland: Über Talentmangel, den Lückenschluss zwischen Promotion und Industrie sowie firmeninterne Weiterbildung

Von Dr. Chris Armbruster, Director Data Science Retreat, Berlin

Wenn ich – oder auch Sie – auf LinkedIn Data Science in die Suche eingebe und auf „People“ klicke, erhalte ich mehr als 10 Millionen Ergebnisse. Aber für „Data Scientist“ sind es nur etwa 100.000 Personen, und für Deutschland weniger als 4000. Dieser greifbare Talentmangel muss und kann unter Zeitdruck in den nächsten fünf Jahren behoben werden, wenn Ausbilder, Industrie und auch die Politik zusammenwirken. Hier wird gezeigt, wie dies gelingt.

Wem die eigene Suche bei LinkedIn nicht solide genug erscheint, der möge die Recherche der New York Times nachlesen, wonach es weltweit nur etwa 10.000 KI-Experten gibt, welche neue Projekte initiieren und leiten können. Zudem beziffert das Tencent Research Institute die Zahl der KI-Talente auf insgesamt nur 300.000, wovon etwa 100.000 an Universitäten sind. Die Zahl der Absolventen wird weltweit auf jährlich nicht mehr als 20.000 geschätzt. Es kommt hinzu, dass die Industrie im großen Stil Professor/inn/en wie Doktorand/inn/en abwirbt, und die verbleibenden Experten oft nur teilweise zur Verfügung stehen wegen zusätzlicher Engagements in der Industrie. Der Grund ist Zeitdruck, wie am Beispiel autonomer Mobilität ersichtlich ist, wo in den nächsten zehn Jahren ein Wettlauf stattfinden wird, bei dem alle Teilnehmer um den Erfolg und zugleich ihre Eigenständigkeit kämpfen.

In diesem Szenario gibt es meines Erachtens zwei Wege den Fachkräftemangel binnen fünf Jahren zu beheben:

1. MINT-Talente zu finden, bei denen die Lücke zu Data Science und Künstlicher Intelligenz schnell und einfach geschlossen werden kann.
2. Die firmeninterne Weiterbildung und auch Umschulung von motivierten Mitarbeitern bei geeignetem fachlichem Hintergrund.

Beide Wege werden bereits beschritten; es fehlt aber noch an einem skalierbaren Modell.

Zu 1. Deutschland und Europa bilden überdurchschnittlich viele Doktoranden aus, wie folgender Vergleich verdeutlicht: In Deutschland, Großbritannien, Frankreich und Spanien schließen etwa so viele Doktoranden pro Jahr ab wie in den USA (um die 70.000), jedoch ist die Wirtschafsleis-

tung nur etwa halb so groß und auch die Bevölkerung deutlich kleiner (260 Mio. zu 330 Mio.). In Deutschland promovieren jedes Jahr etwa 30.000 junge Menschen, davon sind rund 60% MINT-Promotionen, mit einem steigenden Frauenanteil (nur in den Ingenieurwissenschaften ist die Verteilung mit 80/20 noch sehr ungleich).

Auf einer „Science to Data Science"-Tour durch deutsche Universitätsstädte verschaffe ich mir zurzeit einen Überblick. Erste Daten deuten darauf hin, dass Doktoranden in großer Zahl (d.h. einige Tausend) über die Karrierechancen in Data Science und Künstlicher Intelligenz nachdenken und auch prinzipiell die notwendigen Voraussetzungen mitbringen. Jedoch klappt der „Lückenschluss" zwischen Promotion und Industrie im Bereich KI und Data Science derzeit eher nur vereinzelt und zufällig.

Es gibt einige Gründe, die den „Lückenschluss" verhindern; so etwa, dass a) Kandidat/inn/en nicht wissen, wie sie ihre Fähigkeiten an die Industrie bringen; b) auch der Abschluss mehrerer Onlinekurse oft noch nicht zu einer Anstellung führt; c) die Industrie nicht weiß, wo die Talente zu finden sind und wie sie für einen Einstieg vorbereitet werden müssen; und d) die Hochschulen keine berufsorientierte Weiterbildung anbieten können oder wollen.

Zu 2. Deutsche Unternehmen investieren in die Onlineweiterbildung und dabei ist Sebastian Thruns Udacity oft der bevorzugte Partner, z.B. für Bertelsmann (15.000 Stipendien für Data-Science-Kurse), oder BMW, Bosch und Daimler (Kooperationspartner für den Kurs „self-driving car engineer").

Meine eigenen Recherchen und Nachfragen haben ergeben, dass insbesondere in der Autoindustrie (auch bei den Zulieferern) eine Vielzahl von Mitarbeitern Onlinekurse belegt und auch abschließt. Dabei kommt es immer häufiger vor, dass das Unternehmen die Kursgebühren übernimmt, Lerngruppenbildung ermutigt und auch Lernen während der Arbeitszeit ermöglicht. Soweit mir bekannt, schaffen einzelne Mitarbeiter den Umstieg zu Data Science oder KI.

Jedoch habe ich kein Beispiel vor Augen für eine systematische firmeninterne Umschulung („Upskilling"). Es gab Überlegungen zur Verzahnung von Onlinekursen mit Präsenztraining und langfristigem Mentoring für die Autoindustrie, und ich war daran beteiligt, aber ein tragfähiges Modell ist nicht entstanden. Die folgenden Gründe kann ich benennen: a) Mitarbeiter/innen systematisch über Monate für die Weiterbildung freizustellen scheint schwierig und die deutschen Gepflogenheiten zur Weiterbildung (z.B. mit einem Anspruch von 10 Tagen in 2 Jahren) helfen nicht; b) die Vorqualifizierung und Auswahl geeigneter Kandidaten scheint bisher zu aufwendig; c) es fehlt wohl auch des Öfteren ein Plan, was mit den neu qualifizierten Mitarbeitern kurzfristig sinnvoll zu erreichen wäre.

Zusammenfassend lässt sich beobachten: Doktoranden und Postdocs erreichen den Lückenschluss bisher am besten auf eigene Initiative, wobei Onlinekurse üblicherweise nicht ausreichen und Bootcamps sehr teuer sind, so dass die Zahl der qualifizierten Talente klein bleibt. In der Industrie ist ebenso Eigeninitiative gefragt, und mit Hilfe von Lerngruppen und Onlinekursen gelingt in der Autoindustrie einigen Hundert Mitarbeitern ein Einstieg in das Thema Data Science und KI. In diesem Format ist es aber unwahrscheinlich, dass dies nachhaltig die Wettbewerbsfähigkeit der Unternehmen befördert.

Lösungsvorschläge zur Skalierung

Mehrjährige Erfahrung in den USA wie Europa zeigt, dass eine Kombination aus Onlinekursen und Bootcamps einen qualifizierten Einstieg in Data Science und/oder Künstliche Intelligenz ermöglicht. Für Doktoranden wie Industriespezialisten ist dies üblicherweise ein Aufwand von 400 bis 600 Stunden. Die Erfahrung mit der Industrie, insbesondere den „hiring managers“ verdeutlicht dabei, dass Onlinekurse allein nicht ausreichen. Vielmehr kommt es darauf an, über eine gewisse Zeit, etwa 6 bis 10 Wochen, intensiv an einem ersten Praxisprojekt zu arbeiten, gegebenenfalls auch als Team.

Aus den obigen Beobachtungen und Überlegungen ergeben sich folgende Lösungsansätze:

1. Während der Promotion ließe sich der Lückenschluss recht kompakt erreichen durch eine Kombination aus begleitender Weiterbildung, kompaktem Training und Projektarbeit für Produktprototypen. Für Postdocs wäre das gleiche Format denkbar. Alternativ wäre auch ein Umstieg direkt im Anschluss an die Promotion denkbar. Wenn Industrie und Universitäten hier nicht zueinander finden, könnte auch ein privater Anbieter die Lücke füllen.
2. Unternehmen müssen neu planen. Der „Business Case“ für Data Science wie Künstliche Intelligenz ist hinreichend bekannt, und die Konsequenzen des Zu-Spät- oder Nicht-Handelns eigentlich auch. Wenn man erprobte Konzepte wie Sabbatical, Umstrukturierung und Ausgründung zusammenfügt, ergibt sich die Möglichkeit, neue und wendige Geschäftseinheiten in kürzester Zeit ans Laufen zu bringen.

Über den Autor

*Dr. Chris Armbruster ist Direktor des Data Science Retreat, Berlin. Seit Anfang 2018 läuft seine Kampagne 10.000 Data Scientists for Europe mit dem Ziel, binnen fünf Jahren mindestens 10.000 zusätzliche Talente zu finden und für einen qualifizierten Einstieg in die KI-getriebene Produktentwicklung **zu** trainieren.*

AUSBLICK

Von Data Literacy bis Data Science: Der Handlungsbedarf in der deutschen Hochschullandschaft

Von Prof. Dr. Michael Goedicke, Vizepräsident der Gesellschaft für Informatik e.V. von der Universität Duisburg-Essen, und Prof. Dr.-Ing. Peter Liggesmeyer, Sprecher der GI-Task-Force Data Science und Leiter des Fraunhofer IESE

Der Bedarf an Datenwissenschaftlerinnen und -wissenschaftlern in Deutschland und Europa ist enorm. Die Wirtschaft reagiert auf ihre Weise und entwickelt Weiterbildungsangebote, um dem Anspruch an eine fundierte und qualitativ hochwertige Ausbildung im Bereich Data Science zu erfüllen. Gleichzeitig darf das Feld nicht den Unternehmen überlassen werden. Indes fühlen sich viele Hochschulen mit der Entwicklung neuer Studiengänge überfordert. Um international nicht den Anschluss zu verlieren, bedarf es einer nationalen Anstrengung beim Aufbau von Bildungsinfrastrukturen im Bereich „Data Science". Die Hochschulpolitik muss diesen Bedarf adressieren.

Sowohl auf nationaler als auch auf der europäischen Ebene gibt es eine Reihe von Initiativen, die an der Fortentwicklung der Datenwissenschaften arbeiten. Die vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Big-Data-Kompetenzzentren Berlin Big Data Center (BBDC), ScaDS Dresden/Leipzig Competence Center for Scalable Data Services and Solutions, Smart Data Innovation Lab (SDIL) bspw. fokussieren stark auf die Big-Data-Forschung.

Das Technologieprogramm „Smart Data – Innovationen aus Daten" des Bundesministeriums für Wirtschaft und Energie (BMWi) ist auf die anwendungsnahe Forschung ausgerichtet und das Smart Data Forum, eine gemeinsame Initiative von BMWi und BMBF, unterstützt die Nutzung von Big-Data-Technologien und die Verbreitung von Data-Science-Know-how in mittelständischen Unternehmen. Auf der europäischen Ebene existiert mit dem EDISON Data Science Framework auch bereits eine Roadmap für die Data-Science-Bildung und -Schulung in Europa.

Die unterschiedlichen Beiträge der Experten aus der Gesellschaft für Informatik, der Deutschen Mathematiker-Vereinigung e.V., der Deutschen Physikalischen Gesellschaft e.V. und

der Gesellschaft Deutscher Chemiker e.V. zeigen, dass der Handlungsbedarf in der Hochschulausbildung im Bereich Data Literacy sowie Data Science in Deutschland sehr vielschichtig ist.

Data Literacy: Grundlegende Kompetenzen im Umgang mit Daten

Data Literacy ist ein relativ junger Gegenstandsbereich der Bildung in verschiedenen Anwendungsgebieten, der aufgrund seiner steigenden Bedeutung sowohl für die Forschung, die berufsbezogene Ausbildung als auch im Hinblick auf die Allgemeinbildung auszudifferenzieren ist.

Für einen mündigen Umgang mit Daten und datenbasierten Systemen sollten alle Studenten grundlegende Data-Literacy-Kompetenzen erwerben. Grundsätzlich muss möglichst früh das Bewusstsein für die Wichtigkeit von Kompetenzen im qualifizierten Umgang mit Daten sowohl bei den Studierenden als auch den Bildungsinstitutionen selbst geschaffen.

Data Literacy sollte daher in der gesamten Breite der Hochschulausbildung bereits auf Bachelorniveau vermittelt werden.

Dazu bedarf es disziplinenübergreifender Kollaborationsformen (wie z.B. des Aufbaus einer unabhängigen Institution), die Forschende und Lehrende aus den verschiedenen Kompetenzfeldern (wie informatische, mathematische und Domänen-Kompetenzen) zusammenbringt.

Der Einsatz kreativer Lehransätze mit technologischen Hilfsmitteln, u.a. „Hands-on“, „Modul-basiertes“ und „Projekt-basiertes“ Lernen in Workshops oder Labs mit realen Daten hat sich erfolgversprechend gezeigt. Dabei spielt auch die genaue Betrachtung der Bildungsniveaus und des fachlichen Hintergrunds des Anwendungsgebiets der Studierenden eine Rolle, um die Angebote entsprechend anzupassen, denn nicht jede/r braucht das komplette Spektrum der Data-Literacy-Kompetenzen in der gleichen Tiefe.

Data Science: Tiefgehende Kompetenzen im Umgang mit Big Data

Data Science ist ein zunehmend wichtiger interdisziplinärer Forschungs- und Bildungsbereich, der eine starke Basis in der Informatik (insbesondere Datenmanagement) und Mathematik (insbesondere Datenanalyse) aufweist.

Eine starke Nachfrage im Bereich Data Science am Arbeitsmarkt trifft aktuell ein noch begrenztes Angebot an Absolventen. Es besteht aktuell ein Mangel an insbesondere Bachelorstudiengängen und Angeboten für alternative Qualifizierungsformen. Letzteres stellt insbesondere im Rahmen der industriellen Weiterbildung einen wichtigen Bereich dar.

Es bedarf daher einer weiteren Ausdifferenzierung des Studienangebots

für Data Science. Insbesondere sollten weiterbildende Teilzeitstudiengänge (Bachelor- und Masterniveau), weiterbildende Zertifikatskurse (mit ECTS) sowie weiterbildende Seminare und Workshops (ohne ECTS) konsequent ausgebaut werden. Während bei weiterbildenden Zertifikatskursen und Seminaren schon eine Anschubfinanzierung hilfreich ist, sind für Studiengänge dauerhafte Finanzierungswege sicherzustellen.

Um einem „Wildwuchs“ solcher Studiengänge zu begegnen, wäre es hilfreich, wenn bspw. die Gesellschaft für Informatik e.V. Empfehlungen zur inhaltlichen Ausgestaltung solcher Studienangebote erstellen würde.

Um Data-Science- und Data-Literacy-Kompetenzen in die Breite zu tragen, ist eine klare Nomenklatur und Ausdifferenzierung der Kernkompetenzen von Data Literacy und Data Science notwendig.

Big-Data-Forschung, Datenqualität und -infrastrukturen

In der Big-Data-Forschung sollte weiterhin die Einrichtung von Kompetenzzentren zu Data Science im Allgemeinen und einzelnen Kompetenzbausteinen im Speziellen (wie Big Data, Machine Learning oder Deep Learning) helfen. Bisherige Bemühungen in diesem Bereich haben sich bewährt, sollten durch Bund und Länder jedoch wesentlich verstärkt werden. Dazu sollten standortspezifische Forschungs- und Anwendungsschwerpunkte geschaffen werden und insbesondere die Zusammenarbeit mit der ortsansässigen Industrie auf Basis realer Daten gestärkt werden, um die Potenziale von Big Data und AI zu eruieren. Dies muss insbesondere explizite Angebote für KMU umfassen, um Kompetenzen für datengetriebene Geschäftsmodelle und intelligente Systeme in die Breite der deutschen Unternehmen zu tragen.

Neben der Personalentwicklung auf allen Ebenen ist die Qualität der Wissenschaft weiter zu befördern und die Einrichtung leistungsfähiger Infrastrukturen notwendig.

Die Grenzen zwischen dem, was die klassischen Wissenschaften und die auf Informationsinfrastrukturen basierenden leisten, werden unschärfer. Eines ist klar: Eine Grundvoraussetzung für Datenwissenschaften sind offene Forschungsdaten – offen in einer intelligenten Form.

Auch das Thema der Datenqualität wird bei datengetriebenen Systemen immer wichtiger, denn es gilt: Garbage In, Garbage Out. Daher bedarf es im Zusammenhang mit der Beurteilung von Datenqualitätsfragen im Bereich Data Science einer Präzisierung und Standardisierung elementarer Datenqualitätskriterien bspw. in der systematischen Annotation von Quelldaten mit Qualitätsmerkmalen.

Die Einrichtung von interdisziplinären und hochschulübergreifenden Data-

Science-Zentren/-Laboratorien, die den Aufbau und die Vermittlung von Data-Literacy- und Data-Science-Kompetenzen auf Bachelor- und Masterniveau vorantreibt, kann dabei helfen die Lücke an Fachkräften und Experten im Umgang mit Big Data zu schließen.

Handlungsempfehlungen der Gesellschaft für Informatik

Um den identifizierten Handlungsbedarf schnell und konsequent zu adressieren, schlägt die Gesellschaft für Informatik kurzfristig, mittelfristige und langfristige Ansatzpunkte vor:

Kurzfristig geht es in erster Linie darum die Vernetzung der vielfältigen Aktivitäten in der Vermittlung von Data-Literacy- und Data-Science-Kompetenzen in Deutschland voranzutreiben und die Rahmenbedingungen für eine Weiterentwicklung zu schaffen. Zu dieser gehört auch, die Vernetzung der deutschen und europäischen Data-Science-Bildungsinitiativen zu unterstützen und eine systematische Evaluation von Best Practices zu fördern.

Darüber hinaus können einheitliche Standards – zumindest aber Leitlinien und Orientierungspunkte – für Kompetenzprofile und die Entwicklung (interdisziplinärer) Curricula der Data-Science-Ausbildung in der Hochschule hilfreich sein.

Zudem müssen Handlungsoptionen für eine flächendeckendere Verbreitung von Data-Literacy-Kompetenzen über alle einschlägigen Studienfächer hinweg erarbeitet werden.

Mittelfristig kann ein einzurichtendes nationales Forum „Data-Science-Education“ auf Bundesebene als Impulsgeber und Think-Tank dienen, um die Vernetzungsaktivitäten zu bündeln und die Aus- und Weiterbildung im Bereich „Data Science“ zu flankieren. Dieses Forum dient als Anlaufstelle für Hochschulen, die eigene Studiengänge aufsetzen wollen, und unterstützt bei der Entwicklung von hochschulinternen und -übergreifenden Data-Science-Laboren.

Langfristig muss eine nationale Hochschulstrategie „Datenwissenschaften“ entwickelt und implementiert werden, um dem immensen Bedarf an Datenwissenschaftlerinnen und Datenwissenschaftlern gerecht zu werden.

Über die Autoren

Prof. Dr. Michael Goedicke leitet an der Universität Duisburg-Essen die Arbeitsgruppe Spezifikation von Software-Systemen und ist Mitglied des Vorstands des Paluno – The Ruhr Institute for Software Technology. Er ist Fellow der Automated Software Engineering Conference (ASE) und Chair der Technical Assembly von IFIP und als Councilor Mitglied des Boards von IFIP. Er ist seit 2012 Mitglied des Präsidiums und seit 2018 Vizepräsident der Gesellschaft für Informatik e.V. (GI).

Prof. Dr.-Ing. Peter Liggesmeyer ist Leiter der Arbeitsgruppe Software Engineering: Dependability an der TU Kaiserslautern und wissenschaftlicher Leiter des Fraunhofer-Instituts für Experimentelles Software Engineering IESE. Von 2014 bis 2017 war er Präsident der Gesellschaft für Informatik e.V. und leitet als Past-President die Präsidiums-Task-Force Data Science.

www.gi.de