API

OPEN ACCESS

Publiziert im Januar 2021
DOI: 10.15460/apimagazin.2021.2.1.52

Zeitschriftenartikel

Begutachtet

Begutachtet:

Helena Häußler
HAW Hamburg
Deutschland

Erhalten: *29. November 2020*
Akzeptiert: *3. Dezember 2020*
Publiziert: *28. Januar 2021*

Empfohlene Zitierung:

ALBY, Tom, 2021: Data Science: Von der Sprache der Daten zur Sprache der Algorithmen. In: *API Magazin* **2**(1) [Online] Verfügbar unter: DOI 10.15460/apimagazin.2021.2.1.52

Data Science: Von der Sprache der Daten zur Sprache der Algorithmen

Tom Alby[1]*

[1] Humboldt-Universität zu Berlin, Deutschland / Lehrbeauftragter für Data Science an der Hochschule für Angewandte Wissenschaften, Hamburg, Deutschland

* Korrespondenz: redaktion-api@haw-hamburg.de

Zusammenfassung

Die Datafizierung unserer Welt erfordert nicht nur ein gehobenes Maß an Datenkompetenz in Bezug auf Erhebung und Analyse von Daten, sondern auch ein mindestens grundlegendes Verständnis der gegenwärtigen Machine Learning-Techniken sowie deren potenziellen Auswirkungen. Dieser Beitrag ordnet die verschiedenen Begriffe wie Machine Learning, Data Science sowie Künstliche Intelligenz ein und beschreibt die Anforderungen und möglichen Folgen der Anwendung von Data Science-Techniken.

Schlagwörter: Data Science, Data Literacy, Datenkompetenz, Künstliche Intelligenz

Abstract

The datafication of our world does not only require an enhanced level of data literacy with respect to data acquisition and analysis but also a minimum basic understanding of current machine learning techniques and their potential impact. The article defines the boundaries between machine learning, data science and artificial intelligence and describes requirements and possible consequences of the application of data science techniques.

Keywords: Data Science, Data Literacy, Data Competence, Artificial Intelligence

 DOI: 10.15460/apimagazin.2021.2.1.52

1 Data Science – Versuch einer Definition

Wie Christine Gläser in der vorherigen Ausgabe des API-Magazins feststellte, bestimmen Begriffe wie Künstliche Intelligenz, Machine Learning und weitere „Buzz Words“ den Diskurs der Datafizierung, so dass Datenkompetenz zu einer Schlüsselkompetenz wird, die es in fast allen Wissenschaften zu erwerben gilt (Gläser 2020). Die in dem Beitrag diskutierte Ridsdale-Matrix bildet verschiedene Grade der Datenkompetenz ab, von der Einführung in Daten zu der „Evaluation Decisions Based on Data“.

So genau wie die Risdsale-Matrix zwischen den Graden der Datenkompetenz unterscheidet, ist die Abgrenzung zwischen Data Science und den anderen „Buzz Words“ nicht. Denn obwohl der Beruf des Data Scientists von der Harvard Business Review als „the sexiest job of the 21st century“ bezeichnet wurde (Davenport und Patil 2012), besteht keine einheitliche Sicht darüber, was Data Science eigentlich ist. So bezeichnet Dhar Data Science als „the study of the generalizable extraction of knowledge from data“ (Dhar 2013). Eine andere Sicht bezieht sich vor allem auf die Größe und Komplexität von Datensätzen („Big Data“), aus denen sinnvolle Informationen generiert werden können (Banton 2019). Dass Data Scientists außerdem Kompetenzen in Mathematik, Machine Learning, Statistik, Künstliche Intelligenz, Datenbanken und Optimierung besitzen sollen (Dhar 2013), macht die Definition sicherlich nicht einfacher, zumal die einzelnen Disziplinen schon seit Jahrzehnten über Verfahren verfügen, die heute in Data Science-Projekten verwendet werden. Viele Algorithmen, die in Data Science-Kursen vermittelt werden, sind zum Teil seit langem bekannte Verfahren der Statistik wie zum Beispiel Support Vector Machines (1963) oder k Means (1957). Auch der Begriff „Data Science“ ist nicht neu; 1974 hatte ihn Peter Naur als Alternative zu Computer Science vorgeschlagen (Naur 1975). Konzepte der heutigen Data Science-Techniken wurden sogar schon 1962 von John Tukey erwähnt, der forderte, dass die Statistik sich um den Bereich des Lernens aus Daten erweitern sollte (Tukey 1962). Ist Data Science also nur alter Wein in neuen Schläuchen?

Tatsächlich besteht ein großer Unterschied zwischen den Anfängen des maschinellen Lernens und dem heutigen Data Science-Kontext darin, dass nun praktisch jede*r Computerbesitzer*in über so viel Rechenkapazitäten verfügt, dass die Verfahren nicht nur einer kleinen Gruppe von Wissenschaftler*innen mit Zugang zu einem Großrechner vorbehalten sind (Pulu 2020). Ebenso sind viele kostenlose Softwarebibliotheken entstanden, so dass Machine Learning-Modelle allen Interessierten offenstehen. Zu guter Letzt entstehen immer mehr Daten aus verschiedenen Quellen, die übers Netz miteinander verbunden und in Echtzeit ausgewertet werden können (Dhar 2013). Aus diesen drei Faktoren ergeben sich neue Möglichkeiten, die eine Interdisziplinarität erfordert, da die Herausforderungen von einer Disziplin allein nicht gelöst werden können.

 DOI: 10.15460/apimagazin.2021.2.1.52

Um Werte aus Daten schöpfen zu können, ist allerdings eine weitere Kompetenz notwendig: Das zu lösende Problem zu definieren. Ein funktionierendes Machine Learning-Modell zu entwickeln ist das eine, aber zu verstehen, wie dieses auch ein echtes Problem löst, ist häufig eine noch größere Herausforderung, auch durch die hohen Erwartungen der Anwender*innen (siehe Abschnitt 2). Der entstehende monetäre Wert durch die Anwendung eines Modells muss zudem den gleichzeitig entstandenen Kosten entgegengehalten werden, um den Return on Investment zu beurteilen. Denn neben der*dem „reinen" Data Scientist werden weitere Rollen für alle Tätigkeiten in einem Data Science-Projekt benötigt:

- Identifikation und Definition des Problems, das durch ein Machine Learning-Modell gelöst werden soll
- Design der Applikationsarchitektur
- Bereitstellung der notwendigen Infrastruktur
- Projektmanagement
- Daten-Akquise
- Daten-Reinigung
- Modellierung und Optimierung
- Testen von Modellen
- Industrialisierung von Modellen
- Sicherstellen des Einhaltens einer Daten-Ethik

Nicht alle diese Tätigkeiten können oder sollten von einer*m Data Scientist ausgeführt werden. So existiert in größeren Teams zum Beispiel ein*e Data Engineer, die*der Daten und Infrastruktur zur Verfügung stellt und die entstandenen Modelle in ein Produktionssystem einbaut. Auch Datenschutzexpert*innen, die die Datensammlung und die Auswirkungen von Modellen überwachen, gehören in ein Data Science-Projekt eingebunden. Denn selbst wenn der Begriff Data Science zunächst vor allem aus der Informatik- und Statistiksicht definiert wurde, so ist deutlich, dass Data Science-Projekte ein breiteres Kompetenzprofil erfordern. Gleichzeitig kann sich Data Science nicht nur auf die Performanz von Modellen konzentrieren, sondern muss dies im Kontext einer gesamtheitlichen betriebswirtschaftlichen, aber auch ethischen Betrachtung tun.

2 Künstliche Intelligenz: Dieses Mal mehr als ein Hype?

Data Science wird auch in Verbindung mit dem Begriff „Künstliche Intelligenz" genannt. Die Vorstellung einer Künstlichen Intelligenz ist von der Fantasie befeuert, dass Maschinen selbständig denken und Aktionen ausführen können. Noch aber hat keine Software den Touring-Test bestanden, in dem Menschen eine Maschine nicht von einem anderen Menschen unterscheiden können. Bereits zwei Mal in der Vergangenheit war der Bereich der Künstlichen Intelligenz jedoch derart „gehyped", dass man davon ausging, dass der Durchbruch ganz nah war. In den 1960er Jahren

DOI: 10.15460/apimagazin.2021.2.1.52

war der General Problem Solver von Simon und Newell ein Beispiel für den Glauben an die scheinbar unbegrenzten Möglichkeiten der KI, in den 1980er Jahren waren es die Expertensysteme, die zumindest innerhalb eines begrenzten Wissensgebiets erfolgsversprechend zu sein schienen. Beide Male wurden die Erwartungen nicht erfüllt. Wir befinden uns demnach heute in einem dritten KI-Frühling. Dass dieser mehr Chancen haben könnte als die beiden Frühlinge davor, ist durch die im letzten Abschnitt genannten Aspekte deutlich geworden.

Unterschieden wird zwischen starker und schwacher Künstlicher Intelligenz. So ist die schwache Intelligenz das, was wir heute in Alexa, Siri und dem Google Assistant sehen. Diese Systeme können nicht tatsächlich selbst denken, sondern nur in klar definierten Kontexten agieren. Eine starke KI hingegen kann sich außerhalb der trainierten Domänen bewegen, und hier besteht der Unterschied zu den reinen Machine Learning-Ansätzen. Man könnte Maschinelles Lernen also als Teilgebiet der Künstlichen Intelligenz verstehen (Müller 2019).

Erste Ansätze einer starken KI sind in solchen Systemen wie Googles Deepmind zu erkennen. Ob diese Systeme aber so weit kommen werden wie HAL 9000 in Kubricks Film „2001: A Space Odyssey“ ist heute noch nicht vorhersagbar. Zwar glauben einige Computerwissenschaftler*innen an die sogenannte technische Singularität, also einem Zeitpunkt, an dem Maschinen mindestens so intelligent sind wie wir. Aber die Chance eines weiteren KI-Winters ist auch nicht komplett unwahrscheinlich. So hat Watson, das KI-System von IBM, bereits eine Bauchlandung in der Gesundheitsindustrie erlitten; die Versprechen konnten nicht erfüllt werden (Strickland 2019). Geschürt durch Werbespots, in denen Watson so klug wie ein gutgelaunter HAL 9000 klingt, sind die Erwartungen über die tatsächlichen derzeitigen Möglichkeiten gestiegen. Dies führt zwangsläufig zu Enttäuschungen, die der Disziplin nicht zum Vorteil gereichen.

3 Brave New Data Science World

Die Enttäuschungen, die durch die genannten anfänglichen Misserfolge entstehen, sollten aber nicht darüber hinwegtäuschen, dass auch Erfolge erzielt werden. Wir bekommen hervorragende Musikempfehlungen basierend auf der Auswertung von Millionen von digitalen Musik-Bibliotheken, Kreditentscheidungen müssen nicht mehr manuell geprüft werden, sondern passieren automatisch, und immer mehr ziehen auch intelligente Assistent*innen in unseren Alltag ein. Wir als Nutzer*innen generieren dabei Daten, die dazu führen, dass die automatischen Entscheidungen für alle besser werden. Denn je mehr ein System genutzt wird, desto mehr Daten fließen wiederum in ein Modell ein, verbunden mit Zielvariablen, ob eine Empfehlung erfolgreich war oder ein Kredit komplett zurückgezahlt wurde.

DOI: 10.15460/apimagazin.2021.2.1.52

Je mehr diese Modellbildung aber in einer Art „Black Box“ stattfindet, desto weniger kann ein Mensch nachvollziehen, warum eine Maschine eine Entscheidung getroffen oder eine Empfehlung generiert hat. Zwar existieren Bestrebungen, dass Modelle möglichst transparent sein sollen, aber angesichts der vielen verschiedenen einbezogenen Informationen kann dies auch ein naiver Wunsch sein (Wissenschaftliche Dienste 2017). Umso wichtiger ist es, dass die Grundprinzipien der Algorithmen, mit denen gearbeitet wird, sowie auch ein allgemeines Verständnis der damit verbundenen Prozesse einer größeren Gemeinschaft vermittelt werden. Denn mit den entwickelten Modellen entstehen nicht gewollte „Nebenwirkungen“, die schmerzhafte Konsequenzen mit sich bringen können.

Ein offensichtliches Problem ist zum Beispiel die potenzielle Diskriminierung durch Machine Learning-Modelle. Eine Maschine weiß nicht, wann sie diskriminiert, und sie weiß auch nichts von den Gesetzen, die Diskriminierung verhindern sollen. Geschlecht, Religion und andere sensible Merkmale können zu automatisierten Entscheidungen führen, die die Betroffenen und ihre Interessen diskriminieren. So wurde kurz nach der Vorstellung der Apple-Kreditkarte ein Fall bekannt, bei dem ein Paar sich wunderte, dass die Frau weniger Kreditlimit bekam als der Mann, obwohl sie mehr Geld verdiente (Vigdor 2019).

Auch die Inhalte, die wir im Netz sehen, sind durch Algorithmen bestimmt. Was bedeutet es für unsere Meinungsbildung, wenn uns nur die Inhalte angezeigt werden, die ein Algorithmus für uns vorgesehen hat? Die Lektüre von Zeitungen beinhaltete zumindest das Potential, dass auch Artikel gelesen werden, die nicht primär unseren Interessen entsprechen und uns somit über den Tellerrand gucken lassen. Wenn aber alles darauf optimiert wird, dass der nächste Klick stattfindet, um weitere Werbung einblenden zu können, wie können dann die Inhalte angezeigt werden, die uns herausfordern und unseren Horizont erweitern? Und was, wenn ein Algorithmus nicht auf Werbeklicks aus ist, sondern auf eine bestimmte Meinungsbildung (Berghel 2018)?

4 Ist Data Science auch eine Schlüsselkompetenz?

Aus den im vorherigen Abschnitt genannten und bereits spürbaren Auswirkungen einer Data Science-zentrierten Welt, ergibt sich die Notwendigkeit, dass das Feld nicht nur den Data Scientists überlassen wird, die sich vor allem auf die Algorithmen und deren Performance fokussieren. Datenkompetenz für diejenigen außerhalb der MINT-Fächer muss über das Sammeln und Auswerten von Daten hinausgehen, denn ansonsten sind Daten-Kompetente nur noch die Wasserträger der Datenwissenschaftler*innen. Insbesondere in der Informationswissenschaft ist abzusehen, dass Technisierung und Datafizierung sich gegenseitig verstärken und Data Science-Verfahren auch in die Teilbereiche Einzug halten, die bisher wenig oder

DOI: 10.15460/apimagazin.2021.2.1.52

gar nicht Gebrauch von Machine Learning machten. Eine Suchmaschine basiert heutzutage nicht mehr vorrangig auf einfachen Verfahren wie TF/IDF; Google setzt mit BERT heute bereits beim Verstehen einer Suchanfrage an (Devlin et al 2018).
Damit ist nicht gemeint, dass unbedingt jede*r zum Data Scientist werden muss. Aber je mehr zumindest ein grundlegendes Verständnis von Data Science vorhanden ist, desto mehr werden die Vor- und Nachteile der jeweiligen Verfahren nachvollzogen und die Datenwissenschaftler*innen auch herausgefordert. Dass dies notwendig ist, belegen die oben genannten Beispiele. Die Güte eines Modells hängt vor allem von der Datenqualität, aber auch von der sauberen Definition des zu lösenden Problems sowie den Kompetenzen drum herum ab. Es reicht daher nicht, nur Daten zu verstehen, wenn Maschinen aus diesen Daten Informationen extrahieren.

5 Fazit

Dieser Beitrag soll die Bedeutung von Data Science und die möglichen Konsequenzen der Ausbreitung dieses Felds für verschiedene Sparten der Wissenschaft, insbesondere der Informationswissenschaft darlegen. Die Informationswissenschaft darf Data Science nicht als außenstehende Disziplin betrachten, sondern muss sich als Teil der Datenwissenschaft etablieren. Durch ihr Verständnis von der Repräsentation und Präsentation von Informationen bringt sie eine Kompetenz mit, die Data Science-Projekte nicht nur vereinfachen, sondern auch erfolgreicher gestalten können. Gleichzeitig bedeutet dies, dass nach der Sprache der Daten auch die Sprache der Algorithmen vermittelt werden muss.

DOI: 10.15460/apimagazin.2021.2.1.52

Literatur

BANTON, Caroline, 2019: Data Science. [Online] Verfügbar unter: https://www.investopedia.com/terms/d/data-science.asp

BERGHEL, Hal, 2018. Malice Domestic: The Cambridge Analytica Dystopia. In: Computer, 51(5). Verfügbar unter: https://doi.org/10.1109/MC.2018.2381135

DAVENPORT, Thomas H. und PATIL, D.J., 2012: Data Scientist: The sexiest job of the 21st century. In: Harvard Business Review October 2012 [Online] Verfügbar unter: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

DEVLIN, Jacob und CHANG, Ming-Wei und LEE, Kenton und TOUTANOVA, Kristina, 2018: „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. arXiv:1810.04805

DHAR, Vasant, 2013: Data science and prediction. In: Communications of the ACM 56(12) [Online] Verfügbar unter: DOI 10.1145/2500499

GLÄSER, Christine, 2020: Wer spricht die Sprache der Daten? Data Literacy in der Lehre am Department Information. In: API Magazin 1(2) [Online] Verfügbar unter: DOI 10.15460/apimagazin.2020.1.2.48

MÜLLER, Tobias, 2019: Spielarten der Künstlichen Intelligenz: Maschinelles Lernen und Künstliche Neuronale Netze. [Online] Verfügbar unter: https://blog.iao.fraunhofer.de/spielarten-der-kuenstlichen-intelligenz-maschinelles-lernen-und-kuenstliche-neuronale-netze/

NAUR, Peter, 1975: Concise Survey of Computer Methods. New York: Petrocelli Verlag. ISBN 0-88405-314-8

PULU, Tibi, 2020: Your smartphone is millions of times more powerful than the Apollo 11 guidance computers. [Online] Verfügbar unter: https://www.zmescience.com/science/news-science/smartphone-power-compared-to-apollo-432/

STRICKLAND, Eliza, 2019: IBM Watson, Heal Thyself. In: IEEE Spectrum, April 2019. [Online] Verfügbar unter: https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

TUKEY, John W., 1962: The Future of Data Analysis. In: Annals of Mathematical Studies 33(1)

DOI: 10.15460/apimagazin.2021.2.1.52

VIGDOR, Neil, 2019: Apple Card Faces Inquiry After Charge It is ‘Sexist’. In: New York Times Nov. 11, 2019, Section B [Online, Zugriff am: 2020-11-28] Verfügbar unter: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

WISSENSCHAFTLICHE DIENSTE DES DEUTSCHEN BUNDESTAGS, 2017: Algorithmen im Medienbereich – Gesetzlicher Regelungsbedarf. [Online] Verfügbar unter: https://www.bundestag.de/resource/blob/529616/bbe3de30880170a7b710e5c8732b7c06/WD-10-048-17-pdf-data.pdf