



Detlev Frick · Andreas Gadatsch
Jens Kaufmann · Birgit Lankes
Christoph Quix · Andreas Schmidt
Uwe Schmitz *Hrsg.*

Data Science

Konzepte, Erfahrungen, Fallstudien und Praxis

EBOOK INSIDE



Springer Vieweg

Data Science

Detlev Frick · Andreas Gadatsch ·
Jens Kaufmann · Birgit Lankes · Christoph Quix ·
Andreas Schmidt · Uwe Schmitz
(Hrsg.)

Data Science

Konzepte, Erfahrungen, Fallstudien
und Praxis

Hrsg.

Detlev Frick
FB Wirtschaftswissenschaften
Hochschule Niederrhein
Mönchengladbach, Deutschland

Jens Kaufmann
FB Wirtschaftswissenschaften
Hochschule Niederrhein
Mönchengladbach, Deutschland

Christoph Quix
FB Elektrotechnik/Informatik
Hochschule Niederrhein
Krefeld, Deutschland

Uwe Schmitz
FB Wirtschaft, FH Dortmund
Dortmund, Deutschland

Andreas Gadatsch
FB Wirtschaftswissenschaften
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Deutschland

Birgit Lankes
FB Wirtschaftswissenschaften, FH Niederrhein
Mönchengladbach, Deutschland

Andreas Schmidt
FB Wirtschaftswissenschaften
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Deutschland

ISBN 978-3-658-33402-4 ISBN 978-3-658-33403-1 (eBook)
<https://doi.org/10.1007/978-3-658-33403-1>

Die Deutsche Nationalbibliothek verzeichnetet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert durch Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2021

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung der Verlage. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag, noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung/Lektorat: Sybille Thelen

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.

Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Geleitwort: Den Menschen im Fokus – Datenschutz als Erfolgsfaktor für Big Data Technologien

Big Data ist nun wirklich kein neuer Trend mehr. Big Data befasst sich mit der Frage, wie enorme Mengen unterschiedlichster Daten aus unterschiedlichsten Quellen möglichst in Echtzeit so ausgewertet werden können, dass sich hierauf tragfähige Entscheidungen stützen lassen. Es geht hier also immer darum, Bestehendes zu nutzen, um daraus Mehrwerte zu generieren. Wir reden über enorme Chancen. Chancen, das Unsichtbare sichtbar zu machen. Und dies idealerweise zum Wohle aller.

Diese Chancen gehen mit Risiken einher. Ich denke, wir sind gut beraten, die Chancen und Risiken technologischer Entwicklungen differenziert zu betrachten, insbesondere, wenn hierbei personenbezogene Daten genutzt werden.

Personenbezogene Daten sind im höchsten Maße individuelle Informationen. Gerade in der digitalisierten Welt haben personenbezogene Daten eine herausragende Bedeutung. Denn sie fallen hier besonders viel und vielfältig an, sie bilden unser Leben ab. Digitalisierung wird integraler Bestandteil unseres Lebens – und damit auch die digitale Datenvielfalt.

Es ist auch gerade erst diese Individualität, aus der sich der – oft auch wirtschaftliche – Nutzen personenbezogener Daten speist. Diese Individualität ist es aber auch, die diese Daten besonders schützenswert macht. Deshalb ist es bei allen datengetriebenen Geschäftsmodellen mit Personenbezug wichtig, die Interessen des Individuums stets und zuallererst im Blick zu behalten.

Mir ist wichtig, dass der Datenschutz nicht als destruktives Element verstanden wird. Es geht ihm nicht darum, Innovationen einzuschränken oder zu erschweren. Datenschutz sucht vielmehr den Ausgleich. Den Ausgleich zwischen den Interessen einer Datennutzung durch Dritte und des Grundrechts der Souveränität eines jeden Einzelnen. Dies wird schnell vergessen, wenn vom „neuen Öl“ die Rede ist und Datenschutz fälschlicherweise als Bremsschuh für Innovationen gebrandmarkt wird.

Es ist vielmehr so: Chancen nutzen heißt auch, den Datenschutz als Erfolgsfaktor zu verstehen. Europäischer Datenschutz ist kein Show-Stopper, sondern kann globaler Game-Changer werden. Denn der europäische Datenschutz bietet zahlreiche

gute Gestaltungsmöglichkeiten für den skizzierten Interessenausgleich, etwa die Anonymisierung oder Pseudonymisierung personenbezogener Daten.

Wir alle tun gut daran, die Privatsphäre zu schützen, auch und gerade um einen Freiraum zur unbeobachteten persönlichen Entfaltung zu belassen. Der Datenschutz hat also stets das Individuum im Fokus und schafft damit gleichzeitig die Voraussetzungen für eine freiheitliche digitale Gesellschaft. Wer Innovationen mit Big Data schaffen will, die dem Menschen dienen, berücksichtigt deshalb naturgemäß die Regeln des Datenschutzes. Proaktiver Datenschutz, also den Schutz der individuellen Daten von Anfang an mitdenken, ist deshalb keine Innovationsbremse, sondern das Mittel um Vertrauen in neue Technologien und ihre Möglichkeiten zu schaffen, das notwendig ist, um sie erfolgreich in den Markt zu bringen. Dieses Vertrauen ist dann tatsächlich das „neue Öl“ für digitale Technologien.

Bonn

Prof. Ulrich Kelber

Bundesbeauftragter für den Datenschutz und die Informationsfreiheit

Vorwort

„Daten sind das neue Öl, aber Informationen sind das echte Gold!“ Das Schürfen dieses Goldes bedarf im digitalen Zeitalter keiner Westernmentalität, sondern neben technischer Lösungen ausgereiftem Fachwissen und digitaler Kompetenzen. Daten müssen effizient verwaltet, im Unternehmen systematisch analysiert und zur Digitalisierung von Geschäftsmodellen erfolgreich implementiert werden. Diese neuen Herausforderungen für Fachkräfte und Experten in Wirtschaft und Wissenschaft sowie fortgeschrittene Studierende mit Interesse an Big Data und Data Science werden in diesem Buch erstmals in den drei Rollen des „Data Strategist“, „Data Architect“ und „Data Analyst“ auf wissenschaftlichem Niveau mit dem erforderlichen Praxisbezug beschrieben.

Die Autoren sind dazu nicht zufällig ausgewählt worden, sondern kommen aus einem gemeinsamen Projekt der wissenschaftlichen Weiterbildung für die digitale Wirtschaft an der Hochschule Niederrhein in Kooperation mit der Hochschule Bonn-Rhein-Sieg und der FH Dortmund. In diesem auf den Bereich Data Science fokussierten Projekt wurde zu jeder der drei genannten Rollen ein sog. Certificate of Advanced Studies (CAS) als wissenschaftliches Weiterbildungsangebot für die Wirtschaft entwickelt, das sich aus einzelnen Zertifikatskursen zusammensetzt.

In moderierten Workshops wurde ausgehend von den für die genannten Tätigkeiten erforderlichen Kompetenzen die Inhalte, Fallbeispiele und Methoden für die Kurse erarbeitet und die einzelnen Curricula passgenau aufeinander abgestimmt. Über die Qualitätssicherung in Form zu genehmigender Prüfungsordnungen und Modulbeschreibungen für jeden Kurs wird die Kompetenz der Autoren durch die Fachgruppen der beteiligten Fachbereiche der Hochschulen weiter ergänzt. Die Kursdurchführung wurde vom Projektteam wissenschaftlich begleitet, intensiv evaluiert und abschließend ausgewertet. Diese Ergebnisse und Erfahrungen sind hier in dem vorliegenden Buch aufbereitet und um anwendungsorientierte Praxisbeispiele zielführend ergänzt. Der Leser erhält somit nicht nur eine wissenschaftliche Beschreibung der Fachkräfteprofile im Markt von Big Data und Data Science aus der Sicht von Lehrenden von drei Hochschulen für angewandte Wissenschaften mit ihrem expliziten Praxisbezug. In diesem Buch sind auch die Erfahrung der ersten Kursdurchführung und das Feedback der

Teilnehmer, die sämtlich aus der Wirtschaft mit einer entsprechenden Berufspraxis ausgewählt wurden, enthalten. Somit ist der Praxisbezug von zwei Seiten garantiert, aus der anwendungsorientierten Wissenschaft und aus der Berufspraxis.

Ein bunter Strauß spannender Themen wartet auf den Leser. Die Implementierung von Big Data-Technologien, die Gestaltung von Informationsarchitekturen und systematische Analyse von Unternehmensdaten bis hin zur Konzeption datenbasierter Geschäftsmodelle, alle Bereiche werden so praxisnah dargestellt, das Fach- und Führungskräfte aller Branchen die erworbenen Kenntnisse und Fähigkeiten direkt in ihrem Aufgabengebiet einsetzen können. Diese Lektüre wird sich lohnen!

Prof. Dr. Thomas Meuser
Leitungsteam des Cyber Management Campus Mönchengladbach
der Hochschule Niederrhein

Grußwort

Im Zuge der digitalen Transformation investieren viele Unternehmen in den Aufbau von Dateninfrastrukturen und Data Science-Teams, die den Weg zum „datengetriebenen“ Unternehmen ebnen sollen. Dies ist die konsequente Weiterentwicklung der klassischen Business Intelligence und scheint ein kleiner Schritt zu sein, schließlich ist man die strukturierte Arbeit mit Daten bereits gewohnt.

In der Praxis tauchen jedoch ungewohnte Hindernisse auf: Die angehäuften Daten müssen erst einmal zugänglich gemacht werden oder in auswertungsfähige Formate verwandelt werden. Dies erfordert Personen, die Bild- oder Textdaten in strukturierte Tabellen transformieren, um entscheidungsrelevante Informationen daraus generieren zu können.

Statistische Applikationen werden zudem in relativ neuen Software- und Programmierumgebungen entwickelt. Wenn eine solche Anwendung in den täglichen Betrieb überführt werden soll, muss diese in die Konzern-IT eingebettet werden und plötzlich müssen viele Anforderungen aus Governance- oder Compliance-Sicht erfüllt werden: Daran scheitern leider viele gute Data Science-Ideen.

Der Schlüssel zum datengetriebenen Unternehmen ist daher Multidisziplinarität. Neues Wissen und neue Rollen sind notwendig, um Data Science erfolgreich zu machen. Gleichzeitig müssen Prozesse, Organisation und IT-Strukturen überdacht werden, so dass sich der erwähnte „kleine Schritt“ sehr schnell auf die gesamte Enterprise Architektur auswirken kann.

Junge Unternehmen demonstrieren, wie eine analytische Organisation von Anfang an aufgebaut werden kann. Viele etablierte Unternehmen müssen dies oft noch lernen, um den Anschluss nicht zu verlieren. Da nicht genug auf diese Herausforderungen der Digitalisierung hingewiesen werden kann, freue ich mich über dieses Werk, das informiert, sensibilisiert und viele praktische Tipps beinhaltet!

Ulrich Dommer
Partner Consulting, KPMG AG Wirtschaftsprüfungsgesellschaft, Düsseldorf

Grußwort: Data Science – Weiterbildung für die Zukunft

Data Science ist nicht länger nur ein anhaltender Trend, sondern inzwischen auch in der Praxis angekommen. Viele Unternehmen setzen mathematisch-statistische Methoden sowie verschiedene Formen Künstlicher Intelligenz ein, um neue Geschäftsmodelle zu entwickeln, Prozesse zu optimieren und neue Formen der Kundeninteraktion einzuführen.

Das verfügbare Potenzial an Wissen in den Unternehmen reicht Stand heute häufig nicht aus. Es mangelt vielfach an Personal in der IT und in den Fachabteilungen. Daraus resultierend lassen sich zahlreiche Möglichkeiten von Data Science noch nicht ausschöpfen.

Der vorliegende Sammelband „Data Science“ spannt einen weiten Bogen, in dem er sich mit der gleichnamigen Thematik aus verschiedenen Perspektiven und auf mehreren Ebenen auseinandersetzt. Die historische Einführung aus Sicht der Wissenschaft ist für die Einordnung vieler Spezialthemen von Bedeutung. Die Fachbeiträge aus Wissenschaft und Praxis decken vielfältige Themenfelder ab und bieten einen spannenden Einblick in die vielfältigen Möglichkeiten der zukunftsweisenden Data Science.

Das Buch baut auf den wissenschaftlichen Zertifikatskursen zu „Big Data und Data Science“ der Hochschule Niederrhein auf. Die Autorenschaft kommt sowohl aus dem wissenschaftlichen als auch aus dem praktischen Umfeld, was das Werk für eine breite Zielgruppe besonders interessant macht. So kann es gleichermaßen als Einführung wie auch als Vertiefung oder als Nachschlagewerk genutzt werden.

Als Anbieter von IT-Lösungen und Services sind wir auf sehr gut ausgebildete Mitarbeiterinnen und Mitarbeiter angewiesen, ohne die wir unsere Leistungen nicht erbringen könnten. Wir wünschen dem Werk daher eine hohe Verbreitung in der wissenschaftlichen Ausbildung und in der Praxis.

Waldemar Zgrzebski
Geschäftsführer, Bechtle GmbH & Co. KG, IT-Systemhaus Bonn/Köln

Data Science – Entwicklungslinien und Trends

Zusammenfassung

Data Science ist als Begriff genauso viel oder wenig Trend, Hype oder Mode wie viele andere Begriffe der (Wirtschafts-)Informatik zuvor. Sie steht in bester Tradition aller Forschungs- und Anwendungsfelder der letzten Jahrzehnte, die sich mit der Generierung und Bereitstellung entscheidungsrelevanter Informationen befasst haben. Im ständigen Wechselspiel zwischen Technologieorientierung und Geschäftsorientierung folgt die Data Science den aktuellen technischen Möglichkeiten und umfassenden Datenbeständen und nutzt als übergreifendes Konstrukt die Methoden maschinellen Lernens genauso wie die geschäftlich motivierten Analysemethoden von Business Analytics. Schafft sie es, die vielen unterschiedlichen Disziplinen erfolgreich zu einem Zusammenwirken zu bewegen, zeichnet sich eine vielversprechende Zukunft für die Data Science ab.

Will man die Entwicklungslinien von innovativen Verfahren in der Informatik nachvollziehen, muss man in der Regel deutlich weiter in die Vergangenheit gehen als man gemeinhin vermutet. So hat sich bei fast allen Hype-Themen gezeigt, dass sie selten bahnbrechend oder überraschend neu sind. Wie in allen anderen Wissenschaften bauen neue Erkenntnisse auf den Errungenschaften früherer Generationen auf – seien es geglückte oder weniger geglückte Ansätze ehemaligen Ideenreichtums. Für die Informatik ist es symptomatisch, dass neue Ideen sich wegen noch fehlender technologischer Unterstützung nicht durchsetzen können, später unter besseren Rahmenbedingungen aber dann als Innovationen umsetzbar werden. Hinzu kommt ein Effekt, dass geschickte Kombinationen von verfügbaren Algorithmen, Verfahren oder Technologien diese Innovationen erst entstehen lassen. Typisch für die IT-Branche mit ihren stark umworbenen Märkten ist die Proklamation solcher Hype-Themen in kurzen Abständen. Der Wirtschaftsinformatik bleibt dann meist nur die nachträgliche Klärung, ob es sich um Trends oder Moden handelte.

Folgt man den Hype-Zyklen von Gartner zur vermeintlichen Vorhersage von aufkommenden Technologien, so vermittelt sich ein durchaus dynamisches Bild dieser Wellen. Über viele Jahre konnte man einen festen Trend und damit auch Wachstums treiber feststellen. Dabei war und ist die Bereitstellung entscheidungsrelevanter

Informationen für das Management von Unternehmen und Organisationen offensichtlich ein andauerndes Problemfeld, das buchstäblich Generationen von (Wirtschafts-) Informatikern beschäftigt hat. Gerade hier ist die Abfolge der „neuen“ Ansätze zur Problemlösung auffällig. Die Pendelbewegung zwischen Technologieorientierung und Businessorientierung findet man zum Beispiel u. a. bei den Begriffen „Data Warehousing“ und „Business Intelligence“ um die Jahrtausendwende. Der seit einigen Jahren andauernde Hype um „Data Science“ schließt sich hier lückenlos an – diesmal mit einem Ausschlag des Pendels in Richtung Technologie, insbesondere Daten und Algorithmen.

Eigentlich ist der Begriff schon in den Anfängen der Informatik durch Peter Naur in den 1960er Jahren geprägt worden. Der algorithmische Kern entspringt den bekannten Verfahren der Statistik und der künstlichen Intelligenz (KI), welche zum maschinellen Lernen geführt haben. Wiederum haben die Entwicklung von KI und Expertensystemen in den 1980er Jahren die prinzipielle Machbarkeit gezeigt, sind aber nicht zur Marktreife gekommen. Lediglich in Marktnischen hielten sich derartige wissensbasierte Systeme, ansonsten sorgte das Scheitern für ein signifikantes Abschwellen der KI-Welle. Erst mit dem Aufkommen der ersten Erfolge von „Big Data“ war das Interesse wieder vorhanden. Datafication oder die allumfassende Sammlung und Auswertung von polystrukturierten Daten beliebigen Formats in Echtzeit haben gezeigt, dass sich entscheidungsrelevante Informationen aus diversen Quellen generieren lassen. Die vielfältigen Anwendungsbereiche, welche anfänglich noch als „use cases“ krampfhaft gesucht wurden, überzeugten nicht zuletzt im Bereich „social media analytics“. Die hierzu eingesetzten „Data Scientists“ bei den Internetgiganten Google, Amazon etc. prägten ein neues Berufsbild, welches vertiefte Statistik- und KI-Kenntnisse forderte. Verbunden mit den hybriden Datenhaltungskonzepten aus klassischem Data Warehousing und Hadoop-Clustern sowie parallelen Hochleistungsrechnern formierten sich die neuen digitalen Ökosysteme. Eine frühere Ausrede der Systementwickler, dass die Algorithmen verfügbar wären, aber die Daten nicht, entfiel damit. Somit brachte die zweite KI-Welle unter dem Big-Data-Dach eine Renaissance der Künstlich Neuronalen Netze (KNN). Die Mustererkennung mit Deep Learning, wobei zahlreiche Zwischenschichten in die KNN eingesetzt werden, ist vielversprechend und überzeugend. Wiederum sind die Verfahren nicht neu, aber aufgrund der Datenverfügbarkeit und Rechengeschwindigkeit nun einsetzbar. Der digitale Fußabdruck jeglicher Objekte und Subjekte ist per Datenanalyse aufspürbar, was für zunehmenden Sprengstoff in der öffentlichen Diskussion um ethische Grundfragen sorgt. Somit hat die technologische Lösung unseres Informationsproblems zu einem neuen Problem bei der Informationsnutzung geführt. Hinzu kommen vielschichtige Fragestellungen um den Einsatz von KI in autonomen Systemen, die bisher nur ansatzweise beantwortet sind.

Die Entwicklungslinien von Data Science hängen direkt mit der Entstehung der Statistik und der künstlichen Intelligenz in Form von maschinellem Lernen zusammen. Gepaart mit Datenhaltungskonzepten und hochperformanten Computern, welche

wiederum ihre eigene Entwicklungsgeschichte haben, ist damit eine komplexe Werkzeugbank entstanden, die den handelnden Akteuren viel abverlangt.

Nicht zuletzt seit den Publikationen von Fayyad ab 1996, der den Begriff „Knowledge Discovery in Databases“ (KDD) prägte, ist allgemein bekannt, dass die Generierung von Wissen aus Daten einen Prozess darstellt. Dieser führt stufenweise von der inhaltlichen Fragestellung über die Datenvorverarbeitung, die eigentliche algorithmische Mustererkennung (Data Mining) und Interpretation (Erklärungsmodell) bis zur Modellimplementierung. Im Rahmen des Modelleinsatzes werden Anpassungsnotwendigkeiten entstehen, die zu einem erneuten Durchlauf des Prozesses führt. Dieser Betrachtung folgen viele Prozessmodelle wie der etablierte Industriestandard CRISP-DM („cross-industry standard process for data mining“) und auch das neu entwickelte DASC-PM („Data Science Process Model“).

Aufgrund der komplexen Aufgabenstellung entlang der Prozessphasen stellen sich vielfältige Anforderungen an die Datenanalysten. Die Bemühung um curriculare Erweiterungen einschlägiger Studiengänge zum Abschluss „Data Scientist“ ist daher allerorten an Hochschulen erkennbar. Man darf sich aber nicht täuschen, denn die algorithmische Befähigung alleine reicht nicht. Die lange Geschichte der Ausbildung zu Business-Intelligence-Experten hat gezeigt, dass ein Gleichklang von Technologie, Business und Organisation gefordert ist, um im Alltagseinsatz der Projekte gewappnet zu sein. Die Führung und die Integration von agilen Teams mit spezifischen Fachkenntnissen sind unverzichtbar, denn ein Einzelkämpfer steht den komplexen Aufgaben machtlos gegenüber.

Nicht nur die Kernaktivitäten der Datenanalyse sind herausfordernd, auch die Ausgestaltung der Digitalisierungsstrategie und die Klärung architektonischen Fragen zur Datenhaltung müssen behandelt werden. Nachfolgende Beiträge im Sammelwerk werden hierzu Antworten liefern.

Die starke Akzentuierung auf KI und Machine Learning (ML) deuten auf eine enge Bindung des Forschungsfeldes an die Kerninformatik hin. Dennoch hat das Thema „Business Analytics (BA)“ aus der Wirtschaftsinformatik dieses Teilgebiet immer eingeschlossen. Dabei kann BA als Sammlung unterschiedlicher Methoden und Technologien verstanden werden, welche dazu dienen, Erkenntnisse aus verfügbaren Daten für unternehmerische Entscheidungen zur Steuerung der Geschäftsprozesse zu gewinnen. BA grenzt sich von Business Intelligence (BI) dadurch ab, dass verstärkt auf die datengetriebene Analyse zur Planung und Prognoserechnung gesetzt wird. Damit steht die Zukunftsorientierung im Vordergrund. Unter dem Begriff Advanced Analytics werden in diesem Zusammenhang gerade die Methoden des maschinellen Lernens und der Statistik erfasst, welche die Ableitung von Vorhersagemodellen mit Kausalzusammenhängen ermöglichen, die deutlich über die Fähigkeiten von explorativen und vergangenheitsorientierten Datenanalysen der Business Intelligence hinausgehen. Also sind Data Science und Advanced Analytics prinzipiell wesensgleich. Der Unterschied entsteht dort, wo die Anwendungsdomäne den ökonomischen Bereich verlässt.

Im Fokus von Data Science steht die Entwicklung von einsetzbaren Entscheidungsmodellen, die wahlweise als interaktive Entscheidungsunterstützung oder als autonome „Entscheidungsmaschine“ genutzt werden können. Im Rahmen der Datenanalyse erhofft man sich als Resultat plausible und interpretierbare Muster, welche als Regelwerke die Entscheidungsmodelle bilden. Zur Aufdeckung der Ursache-Wirkungs-Beziehungen kommen vielfältige Verfahren des Data Mining zum Einsatz. Diese können grob in überwachte und unüberwachte Lernverfahren gegliedert werden. In die erste Gruppe fallen Vorhersagen auf der Basis von klassischen Regressionsverfahren, Klassifikationen mit Entscheidungsbäumen oder Künstlich Neuronalen Netzwerken sowie Zeitreihenanalysen. Den Verfahren ist gemeinsam, dass sie aus Datenbeständen die bekannte Abhängigkeit der zugrunde liegenden Variablen erlernen und als Prognosemodell zur Verfügung stellen. In die zweite Gruppe der unüberwachten Lernverfahren gehören Assoziationsanalysen sowie Clustering. Diesen Algorithmen stehen keine Lernmuster zur Verfügung, stattdessen ermitteln sie eigenständig Hypothesen aus dem Datenmaterial. Entscheidungsmodelle entstehen erst aus der Kombination von derartigen Datenanalysen und der mathematischen Optimierung. Die Verfahren des Operations Research (Simulation, lineare und nichtlineare Optimierung, stochastische Optimierung etc.) können auf den Kausalanalysen aufsetzen und den Erklärungsmodellen eine Zielfunktion hinzufügen. Hierdurch können optimale Handlungsalternativen ermittelt werden, so dass ein Übergang von Deskription über Prädiktion zur Präskription stattfindet.

Insgesamt finden sich vielfältige und verzweigte Wurzeln der Entwicklungsgeschichte von Data Science, die zumindest schlaglichtartig angeklungen sind. Zurzeit zeichnet sich das Bild einer prospektiven Zukunft des Teilbereichs der Informatik ab. Voraussetzung bleibt aber ein positives Zusammenwirken der unterschiedlichen Disziplinen und Akteure, um dem gemeinsamen Ziel der Automatisierung und Digitalisierung näher zu kommen.

Univ.-Prof. Dr. Peter Chamoni

Inhaltsverzeichnis

Teil I Data Strategist Digitalisierung von Geschäftsmodellen – Big Data Technologien erfolgreich implementieren

1	Big Data	3
	Uwe Schmitz	
1.1	Grundlagen	3
1.2	Architektur und Bausteine	6
1.3	Datengetriebene Geschäftsmodelle	15
1.4	Exemplarische Einsatzmöglichkeiten	17
	Literatur	24
2	Data Literacy als ein essenzieller Skill für das 21. Jahrhundert	27
	Andreas Schmidt, Thomas Neifer und Benedikt Haag	
2.1	Notwendigkeit von Data Literacy	28
2.2	Data Literacy als Begriff	30
2.3	Data Literacy Skills im Detail	32
2.4	Konzepte zur Implementation von Data Literacy in Lehre und Praxis	34
2.5	Fazit	38
	Literatur	39
3	Management von Big Data Projekten	41
	Andreas Gadatsch und Dirk Schreiber	
3.1	Konzeptioneller Rahmen des Informationsmanagements	41
3.1.1	Überblick	42
3.1.2	Aufgabenorientiertes Ebenenmodell	42
3.1.3	Integriertes Informationsmanagement	44
3.1.4	Einordnung von Big Data	45
3.2	Digitalisierung von Geschäftsmodellen mit Big Data	46
3.2.1	IT-Governance und Digitalisierung	46
3.2.2	Von der IT-Strategie zur Business Digitalstrategie	49

3.2.3	Management von Big Data.....	53
3.2.4	Vorgehensmodelle zur Einführung von Big Data	54
3.2.5	Messung des Reifegrades von Organisationen.....	57
3.2.6	Auswirkungen von Big Data auf die Organisation	59
Literatur.....		60
4	Digital Leadership.....	63
Wilhelm Mülder		
4.1	Führung im Digitalzeitalter	63
4.2	New Work.....	64
4.2.1	Mobile Arbeitsplätze	65
4.2.2	Flexible Arbeitszeiten.....	65
4.2.3	Veränderte Arbeitsinhalte.....	66
4.2.4	Neue Arbeitsorganisation.....	67
4.3	New Workforce	68
4.3.1	Beschäftigungseffekte der Digitalisierung.....	68
4.3.2	Rekrutierung von Generation Z	68
4.4	Digital Leader.....	71
4.4.1	Persönlichkeitsmerkmale	71
4.4.2	Führungscompetenzen.....	72
4.4.3	Virtuelle Führung.....	72
4.5	Konzepte und Methoden für Digital Leadership	73
4.5.1	SCRUM	73
4.5.2	Design Thinking.....	76
4.5.3	Servant Leadership.....	77
4.5.4	VOPA + Modell	77
4.6	Fazit	79
Literatur.....		80

Teil II Data Architect: Informationsarchitekturen gestalten – Daten effizient verwalten

5	Data Engineering	85
Christoph Quix		
5.1	Aufgaben des Data Engineering.....	86
5.2	Architekturen zum Daten-Management.....	87
5.3	Datenmodellierung und Metadaten-Management	91
5.4	Datenaufbereitung und Datenintegration.....	93
5.5	Datenbank-Management-Systeme: SQL, NoSQL und Big Data.....	99
5.6	Fazit	102
Literatur.....		103

6 Data Governance	105
Detlev Frick	
6.1 Einführung	105
6.1.1 Begriffliche Einordnung	105
6.1.2 Datenstrategie	107
6.2 Data Governance Framework	109
6.2.1 Strategie	109
6.2.2 Aufbauorganisation	111
6.2.3 Richtlinien, Prozesse und Standards	112
6.2.4 Messen und Beobachten	113
6.2.5 Technologie	114
6.2.6 Kommunikation	116
6.3 Data Quality Management (DQM)	117
6.4 Fazit	118
Literatur	118
7 Einsatz von In-Memory Technologien	121
Uwe Schmitz	
7.1 Einleitung	121
7.2 Definition und Abgrenzung In-Memory Technologien	123
7.3 Anforderungen an den Einsatz einer In-Memory-Technologie	127
7.4 Bewertung	129
7.5 Fazit	131
Literatur	131
8 Big-Data-Technologien	133
Christoph Quix	
8.1 Einleitung	133
8.2 Skalierbarkeit und Fehlertoleranz	134
8.3 Volume – Management von großen Datenmengen	137
8.4 Velocity – Kontinuierliche Verarbeitung von Datenströmen	142
8.5 Variety – Unterstützung für die Zusammenführung von heterogenen Daten	145
8.6 Fazit	148
Literatur	148
9 Information Data Models: Das Fundament einer guten Information Strategy	149
Christian Rupert Maierhofer	
9.1 Drei Thesen aus Sicht eines Praktikers	150
9.2 It's all about the information	152

9.3	Das Heute und seine Hürden	152
9.4	Wie es dazu gekommen ist.	153
9.5	Die Enterprise Architektur	154
9.6	Drei Formen der Informations-Architektur und deren Auswirkungen	155
9.6.1	Das Gestern und leider noch das Heute. Der anwendungs-zentrierte Ansatz (The Application Centric Approach)	155
9.6.2	Das Heute und die Morgendämmerung, der datengesteuerte Ansatz (The Data Driven Approach)	156
9.6.3	Das überfällige Übermorgen, die datenzentrische Architektur (The Data Centric Architecture)	159
	Literatur.	162
Teil III Data Analyst: Auswerten, Präsentieren, Entscheiden – Systematische Datenanalyse im Unternehmen		
10	Reporting multidimensionaler Daten und Kennzahlen.	167
	Detlev Frick und Birgit Lankes	
10.1	Betriebswirtschaftliche Motivation	167
10.1.1	Kennzahlen und ihre Anwendung	168
10.1.2	Auswahl von Kennzahlen.	169
10.2	Daten und Business Intelligence	170
10.2.1	Datenmodellierung.	171
10.2.2	Datensicherung.	172
10.2.3	Harmonisierung	173
10.2.4	Daten-/Informationsqualität.	173
10.2.5	Datenbereitstellung	174
10.3	Reporting/Berichtswesen	174
10.3.1	Berichtsgrundformen	176
10.3.2	Anforderungen an Berichte	177
	Literatur.	177
11	Fundamentale Analyse- und Visualisierungstechniken.	179
	Jens Kaufmann	
11.1	Einleitung und Begriffswelt.	179
11.2	Lineare Regression.	182
11.2.1	Basisidee und Begrifflichkeiten	182
11.2.2	Beispiel und Ergebnisinterpretation.	183
11.2.3	Prüfen der Voraussetzungen und Variablentransformation	185
11.3	Einfache Klassifikationsverfahren	186
11.3.1	k-Nearest-Neighbors	186
11.3.2	Naive Bayes	187

11.3.3	Entscheidungsbäume	188
11.4	Clustering-Verfahren	189
11.4.1	Hierarchische Verfahren	189
11.4.2	Partitionierende Verfahren	191
11.5	Assoziationsanalyse	191
11.6	Ergänzende Überlegungen, Software und Tools	192
	Literatur	193
12	Fortgeschrittene Verfahren zur Analyse und Datenexploration, Advanced Analytics und Text Mining	195
	Jens Kaufmann	
12.1	Einleitung	195
12.2	Datenexploration und -darstellung	196
12.3	Principal Component Analysis	197
12.4	Random Forests	200
12.5	Logistische Regression	200
12.6	Entscheidungsbewertung	201
12.7	Zeitreihenanalyse	202
12.8	Text Mining	205
12.9	Weitere Analysemöglichkeiten	207
	Literatur	208
13	Datenbasierte Algorithmen zur Unterstützung von Entscheidungen mittels künstlicher neuronaler Netze	209
	Daniel Retkowitz	
13.1	Datenbasierte Algorithmen und maschinelles Lernen	209
13.1.1	Maschinelles Lernen	210
13.1.2	Lernverfahren	211
13.2	Künstliche neuronale Netze	212
13.2.1	Netzarchitekturen	212
13.2.2	Grenzen künstlicher neuronaler Netze	213
13.3	Beispielhafte Anwendungsfelder	214
13.4	Entwicklungsprozess	216
13.5	Entwicklungsplattformen und Werkzeuge	217
13.5.1	TensorFlow und PyTorch	218
13.5.2	Ausführungsmodi	219
13.5.3	Deployment und Betrieb	220
13.6	Fazit und Ausblick	222
	Literatur	223
14	Künstliche Neuronale Netze – Aufbau, Funktion und Nutzen	225
	Anja Tetzner, Tom Kühne, Peter Gluchowski und Melanie Pföh	
14.1	Einleitung	226

14.2	Aufbau	227
14.2.1	Künstliches Neuron	227
14.2.2	Künstliche neuronale Netze	229
14.3	Lernen künstlicher neuronaler Netze	233
14.3.1	Überwachtes Lernen – Lernen mittels Backpropagation	234
14.3.2	Unüberwachtes Lernen – Lernen mittels Wettbewerbslernen	235
14.4	Nutzenpotenziale und Herausforderungen	236
14.5	Fazit	238
	Literatur	238
15	Bayesian Thinking in Machine Learning	241
	Thomas Neifer, Andreas Schmidt, Dennis Lawo, Lukas Böhm und Özge Tetik	
15.1	Bayesian Thinking	242
15.2	Bayes in Machine Learning	245
15.2.1	Bayes in Regressionsverfahren	245
15.2.2	Bayes in Klassifikationsverfahren	249
15.3	Naive Bayes Classifier	251
15.3.1	Grundlagen	251
15.3.2	Methodik	252
15.4	Fazit	254
	Literatur	254
Teil IV	Anwendungsorientierte Data Science	
16	Text Mining: Durchführung einer Sentiment Analysis mit SAP HANA	259
	Patrick Bachmaier	
16.1	Einleitung	259
16.2	Grundlagen	260
16.3	Umsetzung	261
16.3.1	Vorgehensmodell	261
16.3.2	Implementierung	263
16.4	Fazit	273
	Literatur	274
17	Weiterbildung in Data Science	277
	Christoph Quix	
17.1	Kompetenz-Rahmenwerke für Data Science	278
17.2	Studiengänge zu Data Science	280

17.3	Berufliche Weiterbildung zu Data Science	283
17.3.1	Zertifikatsprogramm der Fraunhofer Gesellschaft zu Data Science	284
17.3.2	Zertifikatsstudien der Hochschule Niederrhein	285
17.3.3	Zertifikatslehrgang zum Data Scientist der Bitkom Akademie	287
17.4	Fazit	288
	Literatur	289
18	Plattformökonomie für Data Plattformen	291
	Valeria Knoll und Alexa Scheffler	
18.1	Motivation	291
18.2	Begriffshaushalt	292
18.2.1	Plattformen und Plattformökonomie	292
18.2.2	Data Plattform	294
18.3	Design-Prinzipien für Data Plattformen	296
18.3.1	Netzwerkeffekte durch gemeinsam genutzte Datenobjekte	296
18.3.2	Strategien für die Aktivierung von Plattformteilnehmern	297
18.3.3	Einfacher Zugang durch Self-Service	298
18.3.4	Effektives Matching durch Metadaten	299
18.4	Monetarisierung	299
18.5	Zusammenfassung und Fazit	300
	Literatur	302
19	Akzeptanz und Nutzung von maschinellem Lernen und Analytics im Rechnungswesen und Controlling	305
	Markus Eßwein, Domenica Martorana, Martina Reinersmann und Peter Chamoni	
19.1	Eine Herausforderung für die Finanzfunktion	306
19.2	Nutzerakzeptanzforschung zu maschinellem Lernen	307
19.3	Befragung von Führungskräften	308
19.3.1	Strukturgleichungsmodell	308
19.3.2	Umfrage	308
19.4	Aktuelle Nutzung und Treiber	310
19.4.1	Ergebnisse der Befragung	310
19.4.2	Treibermodell zur Nutzung und Akzeptanz	315
19.5	Handlungsempfehlungen und Ausblick	317
	Literatur	318
20	Durch Daten zu neuen Geschäftsmodellen und Prozessoptimierungen – im Kontext von Car-Sharing	321
	Eva Schoetzau	
20.1	Kurze Einführung	321

20.2	Durch Daten zu neuen Ideen und Optimierungen	322
20.3	Umdenken im Unternehmen	325
20.4	Durch ständige Überwachung zur stetigen Anpassung	328
20.5	Mit ‚Lessons Learned‘ zur Optimierung von Geschäftsmodellen und -prozessen	331
20.6	Fazit	335
	Literatur	335
21	Einsatz von Logit- und Probit-Modellen in der Finanzindustrie	337
	Uwe Rudolf Fingerlos und Alexander Pastwa	
21.1	Einleitung	337
21.2	Logit- und Probit-Modelle	338
21.3	Datengrundlage	340
21.4	Modellierung	343
21.5	Überprüfung der Modellannahmen	347
21.6	Vorstellung der Ergebnisse	348
21.7	Vergleichende Beurteilung	351
	Literatur	354
	Stichwortverzeichnis	357

Herausgeber- und Autorenverzeichnis

Über die Herausgeber



Prof. Dr. Detlev Frick (Jahrgang 1956). Studium der Wirtschaftswissenschaften mit Schwerpunkt Wirtschaftsinformatik bei Prof. Dr. Jörg Biethahn an der Universität Gesamthochschule Duisburg mit Abschluss als Diplom-Ökonom absolviert. Anschließend wissenschaftlicher Mitarbeiter an der Ruhr-Universität Bochum am Lehrstuhl von Prof. Dr. Roland Gabriel und Promotion zum Dr. rer. oec. an der Gerhard-Mercator-Universität Duisburg (Gutachter: Prof. Dr. Roland Gabriel und Prof. Dr. Bernd Rolfes).

Tätigkeit als festangestellter und freiberuflicher SAP-Berater. Ab 1995 Projektleiter in der Softwareentwicklung (Individualsoftware). Beteiligung an Softwareprojekten in der Größenordnung von 10 bis 140 Mitarbeitern. Von 1999 bis 2001 verantwortlich für den Bereich Methoden und Standards der SAP-Systeme im zentralen Informationsmanagement des Konzerns Deutsche Telekom AG. Von 2001 bis 2004 Kompetenzmanager und Projektleiter der T-Systems Nova in der BU Essen und dort verantwortlich für den Themenbereich SAP. Durchführung von zahlreichen SAP-Projekten. Engagement beim Aufbau des Qualitätsmanagementsystems.

Zum SS 2004 Berufung als Professor für Betriebswirtschaftslehre, insb. Wirtschaftsinformatik an die HS Niederrhein.

Die anwendungsbezogene Lehre und Forschung umfasst die Fachgebiete Standardanwendungssoftware (insb. SAP), Projektmanagement, Business Intelligence und Data Science.

Zahlreiche Beratungsprojekte, Vorträge, Seminare, Workshops und Publikationen zu den vorgenannten Fachgebieten.



Prof. Dr. Andreas Gadatsch ist Inhaber der Professur für Betriebswirtschaftslehre, insbesondere Wirtschaftsinformatik, Leiter des Masterstudiengangs Innovations- und Informationsmanagement sowie Wissenschaftlicher Leiter des Data Innovation Labs im Institut für Management der Hochschule Bonn-Rhein-Sieg.

Er ist Gründungsmitglied des Big Data Innovation Centers der Hochschulen Bonn-Rhein-Sieg, Niederrhein und der FH Dortmund. Die aktuellen Projekte beschäftigen sich mit den Auswirkungen von Big Data auf das Informations- und Geschäftsprozessmanagement.

Er ist Autor von über 340 Fachpublikationen zur Wirtschaftsinformatik, davon 28 Bücher und Herausgeberbände, z. T. in mehreren Auflagen.



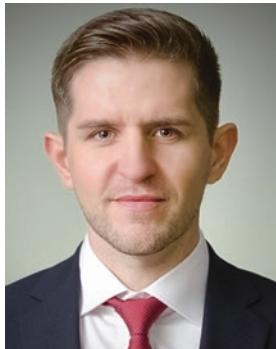
Prof. Dr. Jens Kaufmann ist Inhaber der Professur für Wirtschaftsinformatik, insb. Data Science an der Hochschule Niederrhein. Zuvor war er mehrere Jahre in der Beratung bei Horváth & Partners sowie im Bereich des Global CIO bei der ERGO Group AG in Düsseldorf tätig. Er dozierte als Gastprofessor an der University of North Carolina in Charlotte, NC, USA, und beschäftigt sich in Lehre und Forschung schwerpunktmäßig mit der Anwendung von Data Science und ihrem Transfer in die betriebliche Praxis.



Birgit Lankes ist seit 2013 Lehrkraft für besondere Aufgaben am Fachbereich Wirtschaftswissenschaften der Hochschule Niederrhein. Nach einem Fachhochschulstudium der BWL hat sie zunächst in der IT der Hochschule gearbeitet und hier erste Erfahrungen mit SAP gesammelt. Später hat sie die Lehrenden in der Forschung unterstützt und sich stetig im Bereich SAP weitergebildet. Mit einem internationalen Team hat sie gemeinsam mit Prof. Dr. Frick ein Curriculum zum SAP Solution Manager entwickelt, dass weltweit von Lehrenden eingesetzt wird. Seit 2013 hält verschiedene Veranstaltungen im SAP- und BI-Kontext.



Prof. Dr. Christoph Quix ist seit 2019 Professor für Wirtschaftsinformatik und Data Science im Fachbereich Elektrotechnik und Informatik an der Hochschule Niederrhein. Am Fraunhofer-Institut für Angewandte Informationstechnik FIT in St. Augustin leitet er im Forschungsbereich Life Science Informatics die Abteilung High-Content Analysis. Vorher hatte er eine Vertretungsprofessur für Data Science an der RWTH Aachen inne, an der er auch seine Habilitation (2013) und Promotion (2003) abgeschlossen hat. Seine Forschungsschwerpunkte sind Datenintegration, Management von großen, heterogenen Datenmengen und Metadaten-Management. Er hat mehr als 100 Publikationen in internationalen, wissenschaftlichen Zeitschriften und Konferenzen.



Andreas Schmidt ist wissenschaftlicher Mitarbeiter an der Hochschule Bonn-Rhein-Sieg und dort neben Lehr- und Forschungstätigkeiten im Bereich Data Literacy und dem Themenfeld Future Skills insbesondere mit dem Aufbau des Data Innovation Labs im Institut für Management betraut. Er hat an der Hochschule Bonn-Rhein-Sieg Innovations- und Informationsmanagement studiert und seither verschiedene berufliche Stationen im Bereich der Trend- und Zukunftsforschung durchlaufen – zuletzt dabei bei KPMG als Berater im Bereich Trend Analytics im Zuge der Entwicklung und Anwendung innovativer, datenbasierter Trendanalyse-Methoden zur Beantwortung aktueller gesellschaftlicher, wirtschaftlicher und technologischer Fragestellungen.



Prof. Dr. Uwe Schmitz studierte bis 1998 Betriebswirtschaftslehre an der Hochschule Niederrhein. Direkt im Anschluss hatte er bis zum Jahr 2009 verschiedene Positionen und Funktionen bei der SAP AG inne und war verantwortlich für diverse internationale Großprojekte bei namhaften DAX-Unternehmen. Berufsbegleitend erfolgte seine externe Promotion an der TU Chemnitz im Jahr 2005. Seit 2009 ist Dr. Uwe Schmitz Professor für Wirtschaftsinformatik an der Fachhochschule Dortmund, wo er zurzeit Vorsitzender der Fachgruppe Wirtschaftsinformatik und Leiter der Wirtschaftsinformatikstudiengänge (Bachelor und Master) ist. Er ist auch Vorsitzender de Big Data Innovation Centers der Hochschulen Bonn-Rhein-Sieg, Niedersachsen und der FH Dortmund, sowie Autor diverser wissenschaftlicher Artikel und Buchbeiträge.

Autorenverzeichnis



Patrick Bachmaier Jahrgang 1993, absolvierte das Bachelor- sowie Masterstudium in Wirtschaftsinformatik an der Hochschule Niederrhein in Mönchengladbach. Parallel zum Studium konnte Herr Bachmaier Praxiserfahrung im Bereich Data Warehouse, Business Intelligence und Data Science sammeln. Er arbeitet seit mehr als zwei Jahren als IT-Architekt im Bereich Data Analytics. Parallel zu dieser Tätigkeit besuchte er, neben weiteren Fortbildungen im Bereich Data Analytics, den Zertifikatsstudiengang CAS Data Analyst an der Hochschule Niederrhein, welchen er 2020 erfolgreich abschloss. Herr Bachmaier lebt mit seiner Frau in Moers am Niederrhein.



Lukas Böhm ist Wissenschaftlicher Mitarbeiter und Doktorand an der Universität Siegen. Dort forscht er zu nachhaltiger Mobilität. Weiterhin ist er Wissenschaftlicher Mitarbeiter an der Hochschule Bonn-Rhein-Sieg, wo er im Kompetenzzentrum Usability als KI-Trainer Workshops zu Data Science und Text Mining durchführt. Nach einem Bachelorstudium in Wirtschaftsinformatik an Europäischen Fachhochschule Brühl hat er Information Systems an der Universität zu Köln studiert.



Prof. Dr. Peter Chamoni war seit 1995 Inhaber des Lehrstuhls für Wirtschaftsinformatik, insbesondere Business Intelligence an der Mercator School of Management der Universität Duisburg-Essen. Nach dem Studium der Mathematik und Betriebswirtschaft promovierte er an der Ruhr-Universität Bochum in Operations Research und habilitierte sich dort zum Thema „Entscheidungsunterstützungssysteme und Datenbanken“. Seitdem erschienen von ihm zahlreiche Publikationen zum Thema „Data Warehouse und Business Intelligence“. Auf einschlägigen nationalen und internationalen Tagungen ist er Organisator, Autor und Fachgutachter. Neben der Wissenschaft und der Lehre im Masterstudiengang „Business Analytics“ nimmt die Arbeit in Praxisprojekten für ihn einen hohen Stellenwert ein. Er war

Mitgründer und Vorsitzender des Aufsichtsrats der cundus AG sowie Präsident des TDWI Germany e. V.

Seit dem Wintersemester 2019/2020 ist er im Ruhestand. Als Honorarprofessor an der TU Bergakademie Freiberg in Sachsen nimmt er weiterhin Lehraufgaben in der Wirtschaftsinformatik wahr.



Ulrich Dommer, Jahrgang 1974, Dipl.-Kfm. und MBA, beschäftigt sich seit über 20 Jahren mit SAP-Technologie, Datenarchitekturen, Business Intelligence und Predictive Analytics.

Als Unternehmensberater erarbeitet er gemeinsam mit seinen Kunden effiziente Lösungen für das Management und die Aufbereitung von Daten zu Steuerungszwecken. Herr Dommer begleitet insbesondere Transformationsprogramme im SAP-Umfeld aus Sicht der Unternehmensarchitektur und dem Aufbau von nachhaltigen Informations- und Steuerungssystemen.

Seine beruflichen Stationen umfassten seit seiner Ausbildung bei den Duisburger Grillo-Werken verschiedene Unternehmensberatungen. Ab 2007 führte er die Geschäfte der auf SAP BI-Lösungen spezialisierten CONOGY GmbH, die 2018 in die KPMG AG Wirtschaftsprüfungsgesellschaft integriert wurde. Seitdem verantwortet er Aktivitäten zur digitalen Transformation als Partner im Beratungsbereich der KPMG.

Herr Dommer ist Autor zum Thema Predictive Analytics mit SAP und Beirat im Big Data Innovation Center, über das er mit den Herausgebern dieses Werks verbunden ist.



Dr. Markus Eßwein ist interner Auditor für Finance & Accounting bei Henkel AG & Co. KGaA. Nach seinem Studium des Wirtschaftsingenieurwesens mit technischer Fachrichtung Elektrotechnik an der Technischen Universität Darmstadt arbeitete er als Strategieberater für den Bereich CFO & Enterprise Value bei Accenture Strategy. 2019 promovierte er zur Digitalisierung des Rechnungswesens und Controlling am Lehrstuhl für Wirtschaftsinformatik, insbesondere Business Intelligence an der Mercator School of Management der Universität Duisburg-Essen. Seit 2013 ist er Mitglied und Mitorganisator des Kompetenzzentrums Unternehmenssteuerungssysteme/Arbeitskreises Digital Finance der Schmalenbach-Gesellschaft für Betriebswirtschaft e. V.



Aktuell arbeitet **Dr. Uwe Rudolf Fingerlos** als Risikomanager mit Fokus auf der Performance-Messung und Governance interner Kreditrisikomodelle bei einer spanischen Großbank. Zuvor war er ebenfalls bei einer spanischen Großbank als Teamleiter im Bereich Forschung und Entwicklung für die Kreditrisikomodellierung mit Schwerpunkt auf internen Modellen, IFRS9-Modellen sowie Modellen zur Quantifizierung operationeller Risiken tätig. Überdies verfügt Dr. Fingerlos über Berufserfahrung als Manager und Data Scientist im Geschäftsfeld Risk Advisory (Service Line Regulatory Risk) bei der Deloitte GmbH Wirtschaftsprüfungsgesellschaft (Frankfurt am Main, Deutschland) und der Niederösterreichischen Gebietskrankenkasse (St. Pölten, Österreich) mit Schwerpunkten auf der statistisch-ökonomischen Datenanalyse sowie dem Datenmanagement. Nach seinem Studium promovierte Dr. Fingerlos im Jahr 2014 im Fachbereich Volkswirtschaftslehre an der Wirtschaftsuniversität Wien (Österreich).



Prof. Dr. Peter Gluchowski leitet den Lehrstuhl für Wirtschaftsinformatik, insb. Systementwicklung und Anwendungssysteme, an der Technischen Universität in Chemnitz und konzentriert sich dort mit seinen Forschungsaktivitäten auf das Themengebiet Business Intelligence & Analytics. Er beschäftigt sich seit mehr als 25 Jahren mit Fragestellungen, die den praktischen Aufbau dispositiver bzw. analytischer Systeme zur Entscheidungsunterstützung betreffen. Seine Erfahrungen aus unterschiedlichsten Praxisprojekten sind in zahlreichen Veröffentlichungen zu diesem Themenkreis dokumentiert.



Benedikt Haag studiert Betriebswirtschaftslehre an der Hochschule Bonn-Rhein-Sieg. Er arbeitet als Studentische Hilfskraft im Bereich Wirtschaftsinformatik. Dort unterstützt er beim Aufbau eines Data Innovation Labs sowie in der Erstellung eines Data Literacy Curriculums.



Prof. Ulrich Kelber ist seit dem 7. Januar 2019 der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit. Er ist verheiratet und hat fünf Kinder. Der Dipl.-Informatiker arbeitete nach dem Studium zunächst am Forschungszentrum Informationstechnik, danach als Wissensmanagement-Berater in einem Software-Unternehmen.

Im September 2000 rückte er in den Bundestag nach und vertrat als direkt gewählter Abgeordneter seine Heimatstadt Bonn bis zum Januar 2019.

Von 2005 bis 2013 war er stellvertretender Vorsitzender der SPD-Bundestagsfraktion und koordinierte die Politikbereiche Verbraucherschutz, Ernährung, Landwirtschaft, Umwelt, Naturschutz, Reaktorsicherheit sowie Nachhaltigkeit. Vom Dezember 2013 bis April 2018 war er Parlamentarischer Staatssekretär im Bundesministerium der Justiz und für Verbraucherschutz mit dem Schwerpunkt Verbraucher- und Datenschutz.



Valeria Knoll ist seit 2019 als Data Consultant bei der AXA Konzern AG in Köln tätig. Sie unterstützt die Bereiche Vertrieb und Finance bei strategischen und operativen Fragestellungen rund um die Daten. Davor war sie bei der Allianz Technology AG als Projektmanagerin tätig, nach fast 4 Jahren als Beraterin für Finance bei BearingPoint. Valeria erwarb ihr Bachelor an der Kiev-Mohyla Akademie in der Ukraine und Master an der Otto-von-Guericke Universität Magdeburg, beides im Bereich Finance. Zusammen mit ihrem Mann erzieht sie eine Tochter.



Tom Kühne ist als wissenschaftlicher Mitarbeiter und Nachwuchsforscher an der Professur für Wirtschaftsinformatik, insb. Systementwicklung und Anwendungssysteme, der Technischen Universität Chemnitz tätig. Neben den Themenbereichen Informationssicherheit und Datenbanken liegen seine Forschungsschwerpunkte in der Anwendung und Nutzung von Verfahren des maschinellen Lernens und der Künstlichen Intelligenz. Insbesondere Künstliche Neuronale Netzwerke stehen dabei im Fokus seiner Forschung.



Dennis Lawo ist Wissenschaftlicher Mitarbeiter und Doktorand an der Universität Siegen. Dort forscht er zu nachhaltigem Lebensmittelkonsum. Weiterhin ist er Wissenschaftlicher Mitarbeiter an der Hochschule Bonn-Rhein-Sieg, wo er im Kompetenzzentrum Usability als KI-Trainer Workshops zu Data Science und Text Mining durchführt. Er hat Information Systems an der Universität zu Köln studiert.



Christian Rupert Maierhofer ist 46 Jahre alt und hat einen Abschluss als Industriekaufmann (IHK) sowie als Betriebswirt (VWA). Er ist seit 2017 General Director A/V Software Solutions 360° bei der Bechtle GmbH & Co. KG, IT-Systemhaus Bonn/Köln, nachdem er zuvor dort andere Managementpositionen innehatte. Davor leitete er mehrere Jahre sein eigenes Unternehmen (CRM Design).

Die Gründung der Abteilung A/V Software Solutions 360° hatte für ihn eine „Matrix“-ähnliche Erfahrung. Die blaue oder die rote Pille? Er hat sich damals für die rote Pille entschieden und sagt heute, dass er es zwar nicht bereut hat, aber die IT hinter den Kulissen schon als „etwas sinnfremd“ bewertet.

Seitdem hat er sich mit Enterprise Architekturen sowie Sicherheits- und Skalierungsmodellen beschäftigt und musste feststellen, dass der „wahre Wert“ erstens in der Softwareentwicklung liegt und zweitens, dass das schlagfertigste Konstrukt innerhalb der Informationsverarbeitung die „Community“ ist. Er hat gelernt, dass die IT ein Spiegel der Gesellschaft ist und die Hürden der Digitalisierung wir Menschen selbst sind. „It's all about the information“ sagte Sir Ben Kingsley im Film „Senakers“ – Die lautlosen im Jahre 1992 – besser und treffender könnte man es 2020 auch nicht formulieren.



Dr. Domenica Martorana ist Scientist in der Produktentwicklung bei QIAGEN GmbH. Nach ihrem Studium der molekularen und zellulären Biologie an der Philipps-Universität Marburg promovierte sie zur Stressantwort in Pilz- und Humanzellen im Fachbereich Mikrobiologie und Genetik an der Georg-August Universität Göttingen sowie am Cancer Research Institute des Oslo University Hospitals in Norwegen. Während ihres Studiums erlangte sie in internationalen Wettbewerben in zwei aufeinanderfolgenden Jahren eine Goldmedaille. Im Rahmen ihrer Dissertation beschäftigte sie sich intensiv mit der automatisierten Auswertung großer Datensätze sowie der statistischen Analyse von Daten aus kleinen Stichproben.



Prof. Dr. Thomas Meuser hat seit 1998 eine Stiftungsprofessur für Informatik an der Hochschule Niederrhein in Krefeld inne. Zurzeit ist er dort Dekan des Fachbereichs Elektrotechnik und Informatik.

Prof. Dr. Meuser studierte und promovierte an der RWTH Aachen anschließend folgte der Wechsel zu den Philips Forschungslaboren in Aachen. Seine Lehr- und Forschungsinteressen liegen im Bereich der Kommunikationssysteme und Cyber-Sicherheit. Weitere Aktivitäten beziehen sich auf die digitale Lehre. So leitet er seit 2005 das Cisco Networking Academy Program in Deutschland und ist darüber hinaus Mitglied verschiedener internationaler Design- und Forschergremien in diesem weltweiten Lehrprogramm. Von 2017 bis 2019 führte er an der Hochschule Niederrhein als akademischer Leiter das Projekt ‚Weiterbildung für die digitale Wirtschaft‘.

Seine praktischen Erfahrungen in der Wirtschaft basieren auf einer mehrjährigen Beratertätigkeit, der Tätigkeit als Fachleiter für Informations-Sicherheitsmanagementsysteme und der langjährigen Beschäftigung als CIO eines mittelständischen Elektronikunternehmens. Seit 2019 ist Prof. Meuser Mitglied des Fraunhofer-Instituts FKIE in Bonn und aktuell arbeitet er im Gründungsteam des Cyber Management Campus MG der Hochschule Niederrhein mit.



Prof. Dr. Wilhelm Mülder studierte Wirtschaftswissenschaften an der Universität Essen und promovierte im Bereich Wirtschaftsinformatik zum Thema „Implementierung von computergestützten Personalinformationssystemen“.

Nach Tätigkeit als Software-Entwickler und Berater bei zwei Software-Unternehmen ist er als Professor für Wirtschaftsinformatik an der Hochschule Niederrhein, Fachbereich Wirtschaftswissenschaften in Mönchengladbach tätig. Die wichtigsten Lehr- und Forschungsschwerpunkte sind E-Business, M-Business, Digitalisierung der Wirtschaft und Internet der Dinge. Zu diesen Themen verfasste er mehrere Fachbücher und zahlreiche Fachartikel.

Er leitet das Forschungsinstitut GEMIT (Geschäftsprozessmanagement und IT). Ferner ist er Sprecher der Fachgruppe „Informationssysteme in der Personalwirtschaft“ innerhalb der Gesellschaft für Informatik e. V., Bonn, sowie Mitherausgeber der Fachzeitschrift „HR-Performance“ im Datakontext-Fachverlag, Frechen.



Thomas Neifer ist Wissenschaftlicher Mitarbeiter und Doktorand an der Universität Siegen. Er hat Innovations- und Informationsmanagement an der Hochschule Bonn-Rhein-Sieg studiert und promoviert im Bereich der Verbraucherinformatik. Er forscht zu Empfehlungs- und Reputationsmechanismen im Kontext von nachhaltigem Lebensmittelkonsum und Mobilität. Darüber hinaus wirkt er als Wissenschaftlicher Mitarbeiter an der Hochschule Bonn-Rhein-Sieg bei der Entwicklung einer offenen Datenintegrationsplattform mit. Als Dozent lehrt er dort Data Analytics, Statistik und Volkswirtschaftslehre.



Dr. Alexander Pastwa ist seit 2015 Senior Manager im Bereich Financial Industry Risk & Regulatory bei der Deloitte GmbH Wirtschaftsprüfungsgesellschaft. Er unterstützt nationale und internationale Kunden bei der fachlichen Konzipierung, technischen Umsetzung und Einführung von Standard- und Ad-hoc-Reporting-Anwendungen zur Erfüllung regulatorischer Anforderungen (z. B. im Kontext der MaRisk, BCBS #239); vorwiegend im Finanzsektor. In seinem Bereich und in den Projekten verantwortet Dr. Pastwa die Themenfelder Business Intelligence, Datenqualitätsmanagement und Data Governance. Zuvor war Dr. Pastwa als Senior Manager und stellvertretender Bereichsleiter bei der

SKS Unternehmensberatung in den Themenbereichen Risiko-management und BI-basiertes Reporting tätig. Nach seinem Studium der Wirtschaftswissenschaft promovierte er im Jahr 2009 am Lehrstuhl für Wirtschaftsinformatik der Ruhr-Universität Bochum.



Dr. Melanie Pföh arbeitet als wissenschaftliche Mitarbeiterin an der Professur Wirtschaftsinformatik, insb. Systementwicklung und Anwendungssysteme, und hat bereits mehrjährige Erfahrung im Bereich Data Mining sowie in der Konzeption und der Implementierung analytischer Informationssysteme. Ihren forschungsbezogenen Schwerpunkt setzte sie im Verlauf verschiedener Forschungsprojekte und ihrer Promotion im Bereich der Entscheidungsunterstützung im Kontext privater und öffentlicher Energiesysteme, welche in verschiedenen Veröffentlichungen dokumentiert sind.



Dr. Martina Reinersmann ist Lehrbeauftragte im Bereich Data Science an der FOM Hochschule für Oekonomie & Management und an der Hochschule Niederrhein in Mönchengladbach. Nach ihrer Promotion am Lehrstuhl für Wirtschaftsinformatik der Ruhr-Universität Bochum und der Universität Duisburg-Gesamthochschule arbeitete sie als freiberufliche Dozentin in berufsbegleitenden Weiterbildungsmaßnahmen der Industrie- und Handelskammer NRW (IHK) und der VWA – Verwaltungs- und Wirtschaftsakademie. Von 1997 bis 2007 hatte sie die Leitung der Geschäftseinheit „Analytical Applications“ bei der Alldata Systems GmbH und ab 2002 bei der ScaleOn GmbH & Co. KG (später: Bayer Business Solutions GmbH) inne. In den darauffolgenden Jahren war sie sowohl als selbstständige SAP BI Senior Consultant als auch als wissenschaftliche Mitarbeiterin an der Fakultät für Betriebswirtschaftslehre am Lehrstuhl für Wirtschaftsinformatik, insbesondere Business Intelligence, der Universität Duisburg-Essen tätig.



Prof. Dr. rer. nat. Daniel Retkowitz ist seit 2017 Professor für Wirtschaftsinformatik, insbesondere Software Engineering am Fachbereich Wirtschaftswissenschaften der Hochschule Niederrhein. Er studierte Informatik an der RWTH Aachen und der Chalmers University of Technology in Göteborg, Schweden. In seiner Dissertation beschäftigte er sich mit der Softwareentwicklung für Smart Homes. Anschließend war er mehrere Jahre bei einem IT-Dienstleister in der Versicherungsbranche als Projektleiter in der Systementwicklung sowie als IT-Architekt und Senior Software Developer tätig und dozierte als Lehrbeauftragter an der FH Aachen. In Forschung und Lehre beschäftigt er sich schwerpunktmäßig mit den Bereichen Software Engineering und Machine Learning.



Dr. Alexa Scheffler verantwortet seit 2020 den spartenübergreifenden Bereich Customer Analytics & Insights bei der AXA Konzern AG in Köln. Davor war sie bei AXA als Leiterin des Data Management Offices tätig, wo sie u. a. die Datenstrategie für den AXA Konzern entwickelte und am Aufbau einer Data Plattform mitwirkte. Vor ihrer Tätigkeit bei AXA war sie bei Capgemini als Beraterin für Enterprise Architektur. Ihr Ausbildungshintergrund ist ein Studium der Informatik an der Universität Augsburg und eine akademische Promotion in Wirtschaftsinformatik, bei der sie sich mit der ökonomischen Bewertung von In-Memory Datenbanken beschäftigte.



Eva Schoetzau hat einen Abschluss als B.A. in Betriebswirtschaft von der HS Niederrhein. Zunächst Ausbildung zur Fremdsprachenkorrespondentin in Englisch. Nach dreijähriger Berufstätigkeit u. a. bei einer internationaltätigen Wirtschaftsprüfungsgesellschaft folgte das Studium der Betriebswirtschaft. Praktische Erfahrungen während des Studiums erfolgten unter anderem in den Bereichen des Produktmanagements, Public Relations und des Online Marketings in der FMCG- und Medienbranche. Nach dem Studium mit dem Schwerpunkt Marketing mehrjährige, fundierte Erfahrungen in den Bereichen Kommunikation, (digitalem) Projektmanagement und der Beratung gesammelt. Weiterbildungen im Bereich des agilen Projektmanagements u. a. als Product Owner und Scrum Master konnten bereits

erfolgreich in der Praxis angewendet werden. Derzeit berufsbegleitendes Masterstudium des Business Development Managements an der Europäischen Fernhochschule Hamburg sowie eine Weiterbildung zur psychologischen Beraterin und Coach.



Prof. Dr. Dirk Schreiber hat seit 2000 eine Professur für Betriebswirtschaftslehre, insb. Informationsmanagement am Fachbereichs Wirtschaftswissenschaften der Hochschule Bonn-Rhein-Sieg inne. Im Rahmen der akademischen Selbstverwaltung ist er seit mehr als 10 Jahren in der Fachbereichsleitung engagiert. Prof. Dr. Schreiber studierte und promovierte an der Universität Siegen. Danach wechselte er zum Sparkasseninformatik-Zentrum in Bonn. Seine praktischen Erfahrungen in der Wirtschaft basieren darüber hinaus auf einem mehrjährigen Engagement als CIO eines mittelständischen Unternehmens der metallverarbeitenden Industrie.

Sein zentrales Lehr- und Forschungsfeld ist die Internet-Ökonomie, zu der er bereits 2010 ein Lehrbuch veröffentlicht hat, das mittlerweile in mehreren Auflagen erschienen ist.

Seit 2016 ist Prof. Schreiber Gründungsdirektor des Instituts für Management der Hochschule Bonn-Rhein Sieg. Aktuell arbeitet er an der Gründung des Instituts für Verbraucherinformatik an der Hochschule Bonn-Rhein-Sieg mit.

Özge Tetik studiert Betriebswirtschaftslehre an der Hochschule Bonn-Rhein-Sieg und ist als Studentische Hilfskraft und Tutorin in der Wirtschaftsinformatik tätig. Darüber hinaus unterstützt Sie als Studentische Hilfskraft das Mittelstand 4.0-Kompetenzzentrum Usability.





Dr. Anja Tetzner ist als wissenschaftliche Mitarbeiterin an der Professur Wirtschaftsinformatik, insb. Systementwicklung und Anwendungssysteme, an der Technischen Universität Chemnitz tätig. Der Fokus ihrer Tätigkeit liegt auf den Themenbereichen künstliche neuronale Netze und Business Intelligence, insbesondere dem Teilgebiet des analytischen Kundenbeziehungsmanagements. Im Rahmen ihrer Promotion legte sie den Schwerpunkt ihrer Forschungsarbeit auf den Einsatz künstlicher neuronaler Netze zur Identifikation von Ironie, welche die strukturierte Untersuchung der Eignung verschiedener künstlicher neuronaler Netze zur automatisierten Erkennung von Ironie in vordergründig informeller textueller Kommunikation forcierte.



Waldemar Zgrzebski ist seit 2005 Geschäftsführer des Bechtle IT-Systemhauses Bonn.

Bechtle ist ein zukunftsstarker IT-Infrastrukturdienstleister und Digitalisierungspartner für Industrie, Mittelstand und Verwaltung. Der Standort Bonn gehört zur börsennotierten Bechtle AG mit Hauptsitz in Neckarsulm. Mit 75 IT-Systemhäusern in der DACH-Region und IT-Handelsgesellschaften in 14 Ländern Europas ist Bechtle das größte IT-Systemhaus Deutschlands, starker Partner für zukunftsfähige IT-Architekturen und europaweit führend im IT-E-Commerce.

Waldemar Zgrzebski ist verheiratet, hat vier Kinder und lebt in Bonn.

Abbildungsverzeichnis

Abb. 1.1	Klassifizierung von Big Data-Technologien	7
Abb. 1.2	Allgemeine Kosten für das erste Jahr pro TB im Vergleich Entnommen aus: Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 36	8
Abb. 1.3	Data Mining – Instrumente.	10
Abb. 1.4	Eine Donut-Cloud für „Big Data“	11
Abb. 1.5	Ein Flare-Chart für Professor-Student-Beziehungen	11
Abb. 1.6	Ein Beispiel für Dashboards auf Smartphones und mobilen Endgeräten	12
Abb. 1.7	Social Media Prisma.	19
Abb. 1.8	Anwendungsziele bei Nutzung von Big Data für Marketing und Vertrieb.	20
Abb. 1.9	Anwendungsziele bei Nutzung von Big Data für Distribution und Logistik	21
Abb. 1.10	Anwendungsziele bei Nutzung von Big Data für Finanz- und Risikocontrolling	23
Abb. 1.11	Anwendungsziele bei Nutzung von Big Data für Produktion, Service und Support	24
Abb. 2.1	Häufigkeit der Erwähnung von „Data Literacy“ in der Forschungsdatenbank ScienceDirect im Zeitverlauf	31
Abb. 2.2	Data Literacy als Schnittstellenbegriff	32
Abb. 2.3	Data Literacy Kompetenzrahmen.	35
Abb. 2.4	Framework zur Integration von Data Literacy Initiativen in Unternehmen.	38
Abb. 3.1	Aufgabenorientiertes Ebenenmodell nach Krcmar (2015)	43
Abb. 3.2	Modell des Integrierten Informationsmanagements nach Zarnekov 2005	44

Abb. 3.3	Ebenen des Integrierten Informationsmanagements nach Zarnekov 2005	46
Abb. 3.4	Auswirkungen der Digitalisierung auf die menschliche Arbeit nach Kornwachs (2018)	47
Abb. 3.5	Reifegradmodell zur Digitalisierung nach Krafft (2018)	48
Abb. 3.6	Rollen im Informationsmanagement im Wandel	50
Abb. 3.7	Klassisches versus Agiles Informationsmanagement	51
Abb. 3.8	Ableitung der IT-Strategie aus der Unternehmensstrategie	51
Abb. 3.9	Entwicklung einer Digitalstrategie	52
Abb. 3.10	Bitkom-Portfolio für Big Data (2013)	54
Abb. 3.11	Data Science Pipeline	55
Abb. 3.12	CSC Vorgehensmodell für Big Data	56
Abb. 3.13	BITKOM Vorgehensmodell	57
Abb. 3.14	DASC-PM v1.0 – Ein Vorgehensmodell für Data- Science-Projekte	58
Abb. 3.15	Vorgehensmodell für Big Data – Modell Austria	59
Abb. 4.1	Scrum-Methode	74
Abb. 4.2	Design Thinking	76
Abb. 4.3	Servant Leadership	78
Abb. 4.4	Digital Leadership Tools im VOPA + Modell	78
Abb. 5.1	Schematische Darstellungen einer Data-Warehouse- (links) und einer Data-Lake-Architektur (rechts)	89
Abb. 5.2	Schritte während der Datenaufbereitung und -integration	94
Abb. 6.1	Data Governance als Teil der Data Economy	108
Abb. 6.2	Data Governance-Spannungsfeld	108
Abb. 6.3	Treiber für Data Governance	109
Abb. 6.4	Data Governance Framework	110
Abb. 6.5	Einordnung DGO und Data Governance	111
Abb. 6.6	Prozess zur Nutzung eines Data Catalogs	115
Abb. 6.7	Fachliche Taxonomie	116
Abb. 6.8	Data Asset Catalog	116
Abb. 6.9	Data-Profiling-Analyse als iterativer Prozess	118
Abb. 7.1	Technische Architektur der RDBMS	124
Abb. 7.2	Beispiele für Logs und Savepoints	125
Abb. 7.3	Spalten- und Zeilenorientierte Datenbanken	126
Abb. 8.1	Beispiel für Sharding und Replikation mit Replikationsfaktor 3	136
Abb. 8.2	Namenode und Datanodes in HDFS und Ablauf einer Leseoperation	138
Abb. 8.3	Schematischer Ablauf eines Map-Reduce-Jobs	140
Abb. 8.4	Architektur eines Kafka-Systems	144
Abb. 9.1	Mindset	151
Abb. 9.2	Application Centric	156

Abb. 9.3	Data Driven Business	157
Abb. 9.4	Data Centric Architecture.	160
Abb. 11.1	Boxplot von 11 Beobachtungen (Körpergröße von 11 Personen in cm)	181
Abb. 11.2	Darstellung einer linearen Regressionsanalyse (Körpergröße in Abhängigkeit vom Alter)	183
Abb. 11.3	Ausgabe eines multiplen linearen Regressionsmodells	184
Abb. 11.4	Wertentwicklung eines Kontos (links) und logarithmierte Darstellung (rechts)	186
Abb. 11.5	K-Nearest-Neighbors-Klassifikation mit $k=3$ (links) und $k=5$ (rechts)	187
Abb. 11.6	Klassifikation als 2D-Darstellung (links) und als Entscheidungsbaum (rechts).	188
Abb. 11.7	Datenbasis (links) mit identifizierten Cluster-Zugehörigkeiten der Datenpunkte und Dendrogramm (rechts) eines complete-linkage-Clusterings auf diesen Daten.	190
Abb. 12.1	Matrixplot mit Histogrammen und Korrelationskoeffizienten (229 Länder).	197
Abb. 12.2	Biplot (links) und Varianzerklärungsanteile (rechts) einer Hauptkomponentenanalyse von Fußballdaten.	199
Abb. 12.3	Konfusionsmatrix (links, threshold = 0,5) und ROC-Kurve (rechts) eines fiktiven medizinischen Tests	202
Abb. 12.4	Komponentendarstellung einer zerlegten Zeitreihe (Eheschließungen in Deutschland von 1990 bis 2019)	204
Abb. 12.5	Bag of Words und Darstellung der Berechnung der Kosinus-Ähnlichkeit für zwei Dokumente.	206
Abb. 13.1	Maschinelles Lernen in der Softwareentwicklung	211
Abb. 13.2	Entwicklungsprozess für künstliche neuronale Netze	216
Abb. 14.1	Künstliches Neuron	227
Abb. 14.2	Übersicht Aktivierungsfunktionen	229
Abb. 14.3	Exemplarischer Aufbau eines künstlichen neuronalen Netzes.	230
Abb. 14.4	Schematischer Aufbau eines CNN	231
Abb. 14.5	Schematischer Aufbau einer Selbstorganisierenden Karte.	232
Abb. 14.6	Schematischer Aufbau eines RNN	233
Abb. 15.1	Satz Quelle: Eigene Darstellung in Anlehnung an Lantz (2015)	243
Abb. 15.2	a priori und a posteriori Verteilungen der Parameter θ	247
Abb. 15.3	Trace-Plot des MCMC-Samplers	248
Abb. 15.4	Vergleich der Konfidenzintervalle für tatsächliche und vorhergesagte Werte	250
Abb. 15.5	Klassifikationsmodell	251

Abb. 16.1	Neue Quelle: Screenshot der SAP HANA Web IDE	264
Abb. 16.2	Neue Quelle: Screenshot der SAP HANA Web IDE	264
Abb. 16.3	Neue Quelle: Screenshot der SAP HANA Web IDE	265
Abb. 16.4	Neue Quelle: Screenshot der SAP HANA Web IDE	265
Abb. 16.5	Tabelleninhalt Quelle: Screenshot der SAP HANA Web IDE	266
Abb. 16.6	Tabelleninhalt Quelle: Screenshot der SAP HANA Web IDE	268
Abb. 16.7	Tag Quelle: Screenshot von SAP HANA Studio	272
Abb. 16.8	Tag Quelle: Screenshot von SAP HANA Studio	273
Abb. 18.1	Exemplarische Architektur einer Data Platform.	295
Abb. 18.2	Netzwerkeffekte einer Data Platform.	297
Abb. 18.3	Entwicklung des ROI entlang der Menge der verfügbaren Daten (Szenario 1)	301
Abb. 18.4	Entwicklung des ROI entlang der Menge der verfügbaren Daten (Szenario 2)	301
Abb. 19.1	Konstrukte und Indikatoren des Modells	309
Abb. 19.2	Tatsächliche Nutzung, Anwendungsbereiche, Anwendungsbereiche und Aufgaben	311
Abb. 19.3	Daten, Algorithmen, Entscheidungsschritte und Vorteile.	313
Abb. 19.4	Einstellung, Vertrautheit und Treiber	314
Abb. 19.5	Geschätzte Parameter des Strukturgleichungsmodells.	315
Abb. 21.1	Logit, Probit und Odds	340

Tabellenverzeichnis

Tab. 4.1	Charakteristische Merkmale von Generation Y und Generation Z	70
Tab. 15.1	Datensatz des Spam-Beispiels.	244
Tab. 15.2	Ergebnisse von OLS- und Bayesianischer Regressionsanalyse.	249
Tab. 15.3	Häufigkeitsverteilung des Frucht-Beispiels.	252
Tab. 16.1	Abfrageergebnisse schwach positiver Sentiments.	269
Tab. 16.2	Anzahl Tokens je Sentimentklasse.	272
Tab. 17.1	Kompetenzgruppen des EDISON Data Science Competence Framework.	280
Tab. 17.2	Master- und Bachelor-Studiengänge zu Data Science	282
Tab. 18.1	Funktionen von Data Platform, Data Warehouse und Data Lake.	295
Tab. 18.2	Elemente eines Metadaten Repositories	299
Tab. 19.1	Informationen zu den Teilnehmern an der Umfrage	310
Tab. 19.2	Gesamteffekte der Konstrukte.	316
Tab. 21.1	Deskriptivstatistiken im Trainings- und Testdatensatz	342
Tab. 21.2	WOE-Kodierung und Information Value.	345
Tab. 21.3	Ergebnisse	349
Tab. 21.4	Modellgüte.	353

Teil I

Data Strategist Digitalisierung von Geschäftsmodellen – Big Data Technologien erfolgreich implementieren

Uwe Schmitz: Einführung in Big Data (Begriffe, Architekturen, Geschäftsmodelle)

Andreas Schmidt: Data Literacy als ein essenzieller Skill für das 21. Jahrhundert

Andreas Gadatsch und Dirk Schreiber: Management von Big Data Projekten

Wilhelm Mülder: Digital Leadership



Big Data

1

Uwe Schmitz

Zusammenfassung

Die gespeicherten Daten-Volumina wachsen weltweit rasant an. Sowohl aus den sozialen Netzwerken als auch aus dem Umfeld des „Internet of Things“ werden Daten gesammelt, die meist unstrukturiert sind. Damit wachsen die Herausforderungen. Die Datenmengen müssen gespeichert, verwaltet und analysiert werden. Um sie nutzen zu können und damit Werte zu generieren, werden attraktive Anwendungsbereiche benötigt. Erst das passende Geschäftsmodell macht aus dem Datenvorrat einen potenziellen Schatz. Diverse Praxisbeispiele zeigen den Einsatz von Big Data Lösungen auf.

1.1 Grundlagen

Eine präzise Definition für Big Data gibt es nicht. Das englische Wort Big Data beschreibt eine große Menge an strukturierten und unstrukturierten Daten. Daher wird häufig versucht mit verschiedenen Kriterien sich einer Definition zu nähern. Zu nennen sind hier 5 Kriterien, die in der englischen Übersetzung mit dem Anfangsbuchstaben „V“ beginnen.

Als erstes Kriterium ist Volume (Datenmenge) zu nennen. Es ist ein „grundlegendes V“ und definiert die enormen Mengen an Daten, die z. B. Unternehmen und auch Privatpersonen täglich „produzieren“. Die Bezugsgröße ist der Datenumfang.

U. Schmitz (✉)

FB Wirtschaft, Fachhochschule Dortmund, Dortmund, Deutschland

E-Mail: uwe.schmitz@fh-dortmund.de

Das zweite Kriterium ist Variety (Vielfalt). Es ist ebenfalls ein „grundlegendes V“ und betrifft die strukturierte Einordnung von Daten. Hier sind zwei Aspekte bedeutsam:

- die Vielfalt der Datenformate & Datenquellen,
- unterschiedliche Strukturierungsgrade von Daten.

Die Vielfalt der Datenformate und Datenquellen stellt eine große Herausforderung beim Einsatz von Big Data Technologien dar, da eine Transformation der Daten in eine bestimmte Form u. a. für automatische Auswertungen notwendig ist.

Der Strukturierungsgrad der Daten kann drei unterschiedliche Ausprägungen haben: strukturiert, halbstrukturiert und unstrukturiert.

Zu den strukturierten Daten gehören Daten, die eine bestimmte Anzahl an Feldern aufweisen und daher in Form einer Tabelle gespeichert werden können. Jedes Feld besitzt dabei eine vordefinierte Struktur. Diese Daten werden hauptsächlich in den relationalen Datenbanken gespeichert und verarbeitet. Als Beispiel für strukturierte Daten können Kunden-, Artikel- oder Personalstammdaten angeführt werden. Letzteres werden bspw. wiederum durch Attribute wie Adresse oder das Geburtsdatum genauer beschrieben.

Zu den halbstrukturierten Daten gehören beispielsweise E-Mails. Einerseits besteht eine E-Mail aus einem Absender, den Empfängern, dem Betreff und dem textuellen Inhalt. Aus dieser Sicht ist die E-Mail strukturiert. Doch der Text der E-Mail ist meist unstrukturiert. Deshalb wird eine E-Mail zu den halbstrukturierten Daten gezählt. Zu den reinen unstrukturierten Daten gehören Video- und Audiodateien sowie Bilder.

Die Vielfalt der Datenquellen wie Transaktions-, Protokoll-, Sensor-, Geo-, Audio- sowie Ereignisdaten ist ebenfalls mit unterschiedlichen Datenformaten verbunden.

Als drittes Kriterium ist Velocity (Geschwindigkeit) zu nennen. Es ist ebenfalls ein „grundlegendes V“ und betrifft die Geschwindigkeit, mit der Daten generiert, ausgewertet und weiterverarbeitet werden. Hier sind wiederum zwei Aspekte bedeutsam:

- die Verarbeitungsgeschwindigkeit der Daten,
- die Änderungsdynamik der Daten.

Die Verarbeitungsgeschwindigkeit betrifft die Geschwindigkeit mit der Daten verarbeitet werden. Meist erfolgt dies heutzutage im Bruchteil von Sekunden bzw. in Echtzeit, z. B. aufgrund des Einsatzes einer In-Memory-Technologie. Die Verarbeitungsdynamik wird insbesondere dann zu einer Herausforderung für Unternehmen, wenn eine Echtzeit-Verarbeitung der Daten notwendig ist.

Die Veränderungsdynamik ist mit der Geschwindigkeit verbunden, mit der sich Daten und die Beziehungen zwischen den Daten ändern. Eine hohe Veränderungsdynamik ist beispielsweise bei Streaming-Daten oder Aktienkursen an Börsenplätzen kennzeichnend.

Die Änderungsdynamik kennzeichnet die Geschwindigkeit, mit der sich Daten und Beziehungen zwischen Daten sowie deren Bedeutung ändern. Dies betrifft bspw. Sensordaten oder Finanzmarktdaten oder die Daten in sozialen Netzwerken.

Das vierte Kriterium ist Veracity (Richtigkeit). Es ist ein „zusätzliches V“ und betrifft die Wahrhaftigkeit und Glaubwürdigkeit von Daten. Diese ist wiederum von der Datenqualität bestimmter Datentypen abhängig, z. B. zukünftige Wetterdaten, schwankende Wirtschaftsdaten oder mögliche Kaufentscheidungen der Kunden. Ein Problem liegt in deren Unvorhersehbarkeit und den damit verbundenen Unsicherheiten. Teilweise muss diese Unsicherheit akzeptiert werden, weil keine Bereinigungsmethoden existieren, die diese Unsicherheit beheben können.

Hierzu ein Beispiel: Stromerzeuger müssen einen bestimmten Prozentsatz des Stroms aus erneuerbaren Energiequellen erzeugen. Wind und Sonne sind aber nicht genau vorhersagbar. Der Einsatz verschiedener analytischer Methoden, die Kombination und der Kontextbezug der Daten aus mehreren weniger zuverlässigen Quellen schafft zwar eine präzisere Wetteranalyse, kann jedoch keine „100 %“ zuverlässige Aussage schaffen.

Das fünfte Kriterium ist Value (Business Mehrwert). Es ist ein „zusätzliches V“ und betrifft die Wertigkeit der Daten, die den Kosten gegenübergestellt werden müssen. Dies betrifft insbesondere Investitionen, die Unternehmen zum Aufbau eigener Datenplattformen und der erforderlichen zusätzlichen Infrastruktur tätigen müssen.

Anwendungsbeispiel Netflix: Netflix versucht so genau wie möglich vorherzusagen, was ihre eigenen Abonnenten sehen möchten, indem sie eindeutige Benutzerprofile erstellen. Dazu verwenden Netflix neben Stamm- und Clickstream-Daten auch unstrukturierte Daten aus Videos und Audiodaten. Dies sind bspw. auch Daten aus den Bereichen Gesichtserkennung und Farbanalyse. Anhand dieser Daten kann der Betrachter mit Subgenres und deren Konsumwahrscheinlichkeit verknüpft werden. Unter Verwendung dieser detaillierten Benutzerprofile begann Netflix auch Originalinhalte zu erstellen. Ein Beispiel hierfür ist die selbst produzierte Serie „House of Cards“, die u. a. auf der (damaligen) Beliebtheit des Schauspielers Kevin Spacy erstellt wurde.

Jede Netflix-Webseite sieht für die über 150 Mio. Benutzer anders aus, indem ein ausgeklügeltes Empfehlungssystem bereitgestellt wird, das Daten aus dem Anzeigeverlauf der Benutzer, Daten aus der Videometadatenplattform (Schauspieler, Regisseure und Genre sowie Benutzerbewertungen und Anzeigeverlauf) verwendet.

Übertragen auf die o.g. Kriterien ergibt sich fortfolgende Einordnung:

Volumen

Das Data Warehouse von Netflix ist über 60 Petabyte groß und nimmt ständig zu, da täglich mehr als 500 Mrd. Ereignisse erfasst werden. Täglich werden 3 PB neuer Daten erfasst, diese werden jedoch sofort gelesen und verarbeitet, und es wird nur eine Minderheit gespeichert.

Variety – Vielfalt

Netflix verwendet strukturierte (Stamm- und Metadaten), semi-strukturierte (Clickstreams) und unstrukturierte Daten (Videos), die mit einer Farbanalyse- und Gesichtserkennungssoftware analysiert werden.

Velocity- Geschwindigkeit

Netflix überträgt täglich über 125 Mio. Stunden an Inhalten und sammelt kontinuierlich Nutzerverhaltensdaten (500 Mrd. tägliche Ereignisse).

Veracity – Richtigkeit

Netflix automatisiert den Prozess der Analyse unstrukturierter Videodaten, indem Schnappschüsse von Szenen erstellt werden und so erkennt die Software mit hoher Wahrscheinlichkeit die aktuelle Handlung. Jedoch gibt es keine offiziellen Daten für die Genauigkeit dieser Daten.

1.2 Architektur und Bausteine

Große Herausforderungen bestehen für Unternehmen beim Einsatz von Big Data Technologien darin, Prozesse und Technologien für die Ermittlung und Auswertung der Datenn Mengen in bereits bestehende Geschäftsprozesse zu integrieren. Die Schwerpunkte liegen hierbei in der Integration unterschiedlich strukturierter Daten, der Verwaltung dieser Daten sowie der Möglichkeit zur schnellen Anpassung und Flexibilität. Insofern unterscheiden sie sich diese Systeme von transaktionalen Systemen, wie z. B. ERP-Systemen zur Durchführung und Überwachung von Geschäftsprozessen. Diese Systeme beinhalten meist nur strukturierte Daten. Zudem unterscheiden diese Systeme sich auch von klassischen analytischen Systemen wie Data Warehouse-Anwendungen mit einer zentralen Datenhaltung und qualitativen, strukturierten Daten.

Als Basis für die Klassifizierung von Big Data-Technologien soll in diesem Beitrag eine Taxonomie verwendet werden, die von der Bitkom erstellt wurde. In dieser Taxonomie werden sechs Bausteine betrachtet, wobei jeder Bausteinspezifische Aufgaben zu erfüllen hat (vgl. Abb 1.1).

Der erste Baustein beinhaltet Technologien, die zur Datenhaltung dienen. Zu diesen gehören NoSQL-Datenbanken und Hadoop sowie die In-Memory-Technologie, die aufgrund ihrer steigenden Bedeutung in einem eigenen Kapitel behandelt wird.

Der zweite Baustein bezieht sich auf die Technologien, die sich mit dem Datenzugriff befassen. Als Beispiel kann hier das Complex Event Processing (CEP) angeführt werden.

Der dritte Baustein beinhaltet Methoden zur analytischen Datenverarbeitung wie Data Mining, Text Mining, Machine Learning und Geodatenanalyse. Diese Schicht kann als analytische Schicht bezeichnet werden.

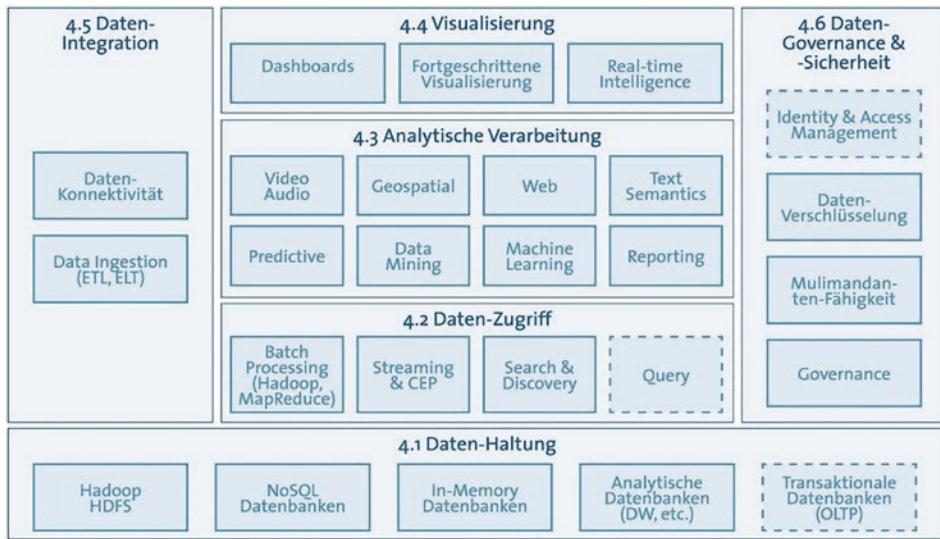


Abb 1.1 Klassifizierung von Big Data-Technologien (Quelle: in Anlehnung an n Anlehnung an: Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 23)

Der vierte Baustein umfasst Visualisierungstechniken, die zur Datenpräsentation dienen. Die anderen Schichten der Taxonomie beziehen sich auf Sicherheitsaspekte (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 23).

Bezüglich des fünften Baustein Data Governance soll stellvertretend das Thema Datenschutz nachfolgend erörtert werden.

Der sechste Baustein betrifft das Thema Datenintegration. Die Technologien, auf welche im Folgenden detailliert eingegangen wird, sind in der Abb 1.1 dargestellt.

Nachfolgend werden beispielhaft einzelne Technologien für jeden Baustein kurz vorgestellt.

Datenhaltung

Hadoop ist ein Open-Source-Framework, das zur parallelen Verarbeitung von großen Datenmengen in einem Computercluster dient. Dieses Werkzeug wird in der Literatur oft im Kontext „Big Data“ erwähnt.

Hadoop setzt keine festgelegte Struktur und Semantik der Daten bei ihrer Speicherung im Gegensatz zu relationalen Datenbankmanagementsystemen (RDBMS) voraus. Diesem Werkzeug liegt ein sog. Hadoop Distributed File System (HDFS) zugrunde. Das HDFS stellt die Hochverfügbarkeit der Daten sicher, indem eine redundante Datenspeicherung in einem Cluster mit mehreren Rechnern erzielt wird. Die Datenverarbeitung erfolgt parallel auf jedem Rechner aus dem Cluster mittels der MapReduce-Methode. Es existieren zurzeit unterschiedliche Hadoop-Distributionen,

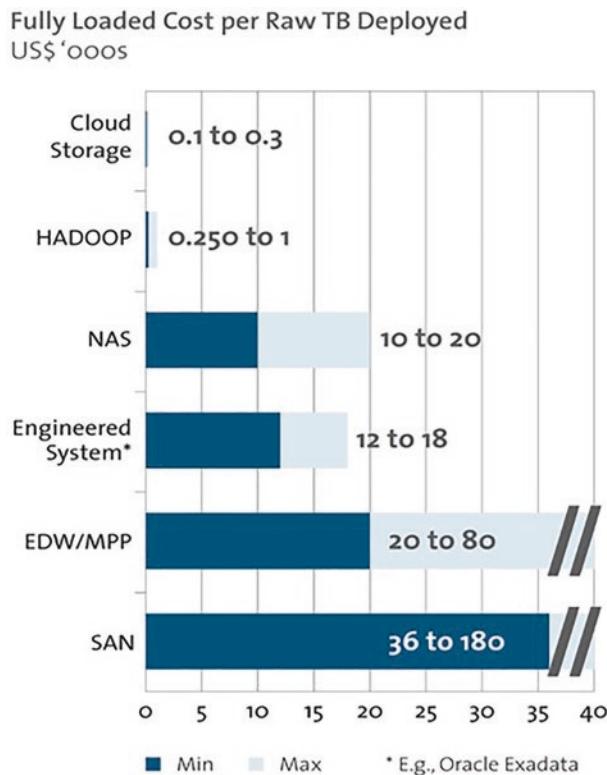
wie Cloudera oder Hortonworks, welche die Nutzung von Hadoop für Programmierer wesentlich erleichtern (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 35, 38, 41).

Ein Vorteil von Hadoop besteht darin, dass die Verarbeitungsgeschwindigkeit der Daten beim steigenden Datenvolumen fast unverändert bleibt. Dies wird dadurch erreicht, indem neue Rechner dem Cluster hinzugefügt werden. Hadoop weist dabei ein hohes Skalierbarkeitspotenzial auf: von wenigen TB bis hin zu mehr als hundert PB (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 24, 36). Da Hadoop ein Open-Source-Framework ist und keine Bindung an einen bestimmten Hardware-Hersteller hat, sind allgemeine Kosten für den Einsatz und die Wartung von einem zusätzlichen TB viel niedriger als bei anderen Software-Herstellern. Der Vergleich von allgemeinen Kosten für ein TB Daten ist in der Abb. 1.2 dargestellt.

Es zeigt sich, dass die Cloud Storage die kostengünstigste Lösung darstellt. NoSQL-Datenbanken gehören zu den neuen Datenbanktechnologien. Die Eigenschaften, welche die Mehrheit der NoSQL-Lösungen aufweisen, sind (vgl. Edlich, NoSQL, S. 2–3):

- meist nicht relationale Datenmodelle,
- i. d. R. flexible Schemata mit wenigen Restriktionen,

Abb. 1.2 Allgemeine Kosten für das erste Jahr pro TB im Vergleich Entnommen aus: Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 36



- hohe horizontale Skalierbarkeit,
- meist Open Source,
- Sicherstellung einer hohen Verfügbarkeit der Daten,
- primär zur Speicherung von großen Datenmengen konzipiert, sogar im PB-Bereich,
- Ergänzung zu RDBMS.

Datenverarbeitung

Die In-Memory-Technologie ermöglicht es, Daten nicht mehr auf der Festplatte, sondern direkt im Hauptspeicher kontinuierlich zu halten. Diese Technologie verbreitet sich u. a. aufgrund gesunkenener Preise für Hauptspeicher aktuell (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 24). Mithilfe der In-Memory-Technologie werden die Daten im Hauptspeicher gespeichert. Bei ausreichender Kapazität des Hauptspeichers kann eine Datenbank vollständig im Hauptspeicher gehalten werden. Trotz sinkender Hauptspeicherpreise ist die Datenhaltung im Hauptspeicher im Vergleich zur Speicherung auf einer der Festplatte i. d. R. aufwendiger. Infolgedessen wurde ein sog. „Temperatur-Modell“ entwickelt. Bei diesem Modell werden im Hauptspeicher nur die Daten gehalten, die zur Analyse von großer Bedeutung sind. Diese Daten werden als „Hot“-Daten gekennzeichnet. Die für die Analyse seltener benötigten Daten werden dann auf der Festplatte vorgehalten. Diese Daten gehören zu den „Cold“-Daten (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 127–128).

Ein weiterer Aspekt, der zusammen mit der In-Memory-Technologie häufig erwähnt wird, ist die spaltenorientierte Datenorganisation. Dabei werden relationale Daten im Hauptspeicher spaltenweise gehalten. Aufgrund einer erhöhten Leistungsfähigkeit von Netzwerkkomponenten sowie Fortschritten im Bereich verteilter Berechnungen ist es möglich, die In-Memory-Technologie in einem Rechner-Cluster umzusetzen (Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 127 f.).

Datenverarbeitungsmethoden

Text Mining ist eine Methode zur Analyse von Fließtext, um darin Muster aufzudecken. Fließtext stellt eine Menge von unstrukturierten Daten dar. Jedoch besitzt jede natürliche Sprache ihre eigene Semantik und Grammatik. Damit lassen sich Erkenntnisse aus einem Text durch Text Mining ermitteln. Eine große Herausforderung für Text Mining besteht darin, dass die grammatischen Regeln für jede Sprache unterschiedlich sind. Eine Disziplin, die sich mit der Verarbeitung von natürlichen Sprachen befasst, ist das Natural Language Processing (NLP). NLP ist darauf ausgerichtet, die semantische Analyse eines Textes durchzuführen, um den Aufbau des Textes sowie die Sprache zu ermitteln. Daraus können wiederum die Bedeutung des Textes und die Stimmung des Verfassers abgeleitet werden (vgl. Freiknecht, Big Data in der Praxis, S. 376–379).

Text Mining spielt im Kontext Big Data eine bedeutende Rolle im Social Media Marketing. Beispielsweise ermöglicht es diese Methode, den sozialen Netzwerken relevante Erkenntnisse bezüglich neuer Produkttrends zu entnehmen, indem die Nach-

richten von Nutzern analysiert werden (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 58).

Zur analytischen Verarbeitung gehört auch das Data Mining. Einige Data Mining – Instrumente sind in der Abb. 1.3 dargestellt.

Visualisierungstechniken

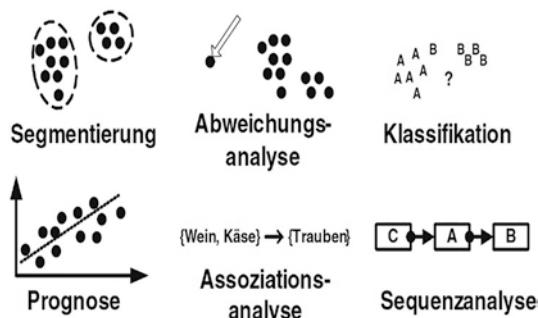
Mit dem Beginn der „Big Data-Ära“ wurden neue Darstellungstypen entwickelt. Diese Darstellungstypen können große Datenmengen in der Form präsentieren, damit die Anwender die Beziehungen zwischen Datensätzen ermitteln sowie eventuelle Messfehler erkennen können (vgl. Freiknecht, Big Data in der Praxis, S. 325). Die neuen Darstellungstypen dienen als Ergänzung zu bereits existierenden Darstellungstypen, wie Balken- oder Tortendiagrammen. Im Folgenden werden kurz zwei Darstellungstypen erörtert.

Der erste Darstellungstyp ist die Word-Cloud. Diese Darstellungsform eignet sich besonders für die Darstellung von Begriffen aus einer großen Menge an textuellen Daten (vgl. Freiknecht, Big Data in der Praxis, S. 336). Eine andere mögliche Darstellung ist die „Donut“-Cloud, welches hier für das Wort „Big Data“ und die damit verbundenen Begriffe in der Abb. 1.4 dargestellt wird.

Der dritte Darstellungstyp ist das Flare-Chart (vgl. Abb. 1.5). Mithilfe eines solchen Charts können Zusammenhänge zwischen unterschiedlichen Entitäten schnell erkannt werden. In der folgenden Abbildung ist ein Flare-Chart dargestellt, das Relationen zwischen Professoren und Studenten repräsentiert. Bei der Auswahl eines Professors werden in roter Farbe die Beziehungen zu den Studenten hervorgehoben, die von ihm betreut werden.

Außer den neuen Darstellungstypen kommen auch traditionelle Werkzeuge bei der Big Data-Analyse zum Einsatz, die bereits vor der Entwicklung der Big Data Technologien entwickelt wurden. Als Beispiel können hier Dashboards angeführt werden. Ein Dashboard setzt sich aus unterschiedlichen grafischen Bausteinen zusammen, die gleichzeitig auf einem Bildschirm abgebildet werden (vgl. Abb. 1.6).

Abb. 1.3 Data Mining – Instrumente (Quelle: Müller, R. M., Lenz, H.-J.: Business Intelligence, 1. Aufl., Berlin/Heidelberg, Springer Vieweg, 2013)



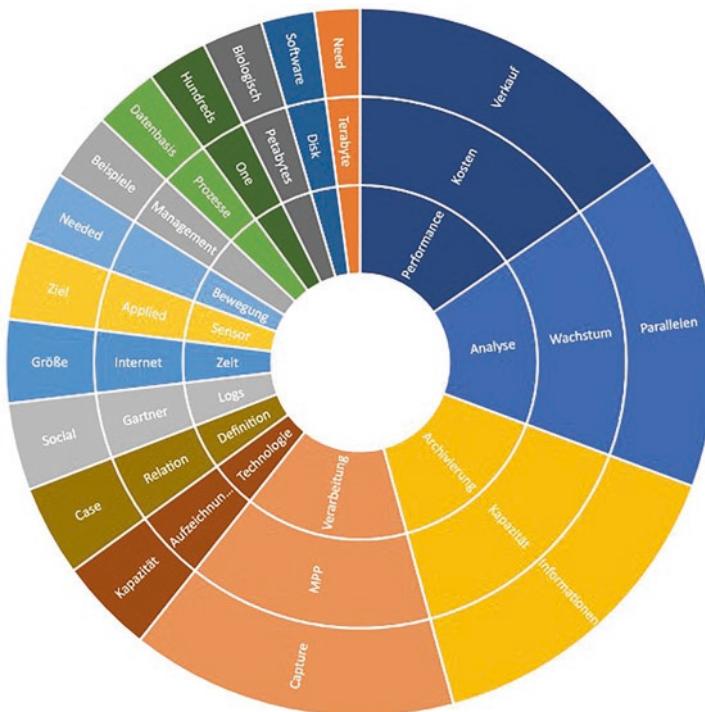


Abb. 1.4 Eine Donut-Cloud für „Big Data“ (Eigene Darstellung)

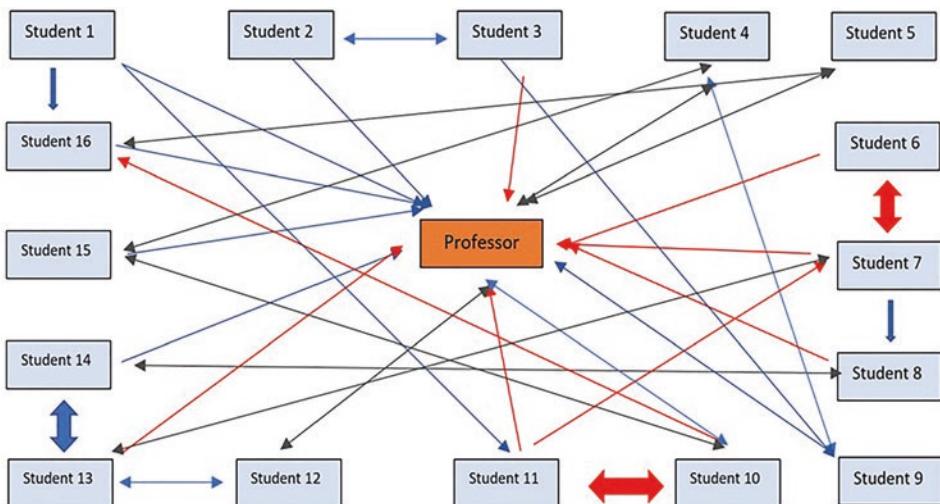


Abb. 1.5 Ein Flare-Chart für Professor-Student-Beziehungen (Eigene Darstellung)

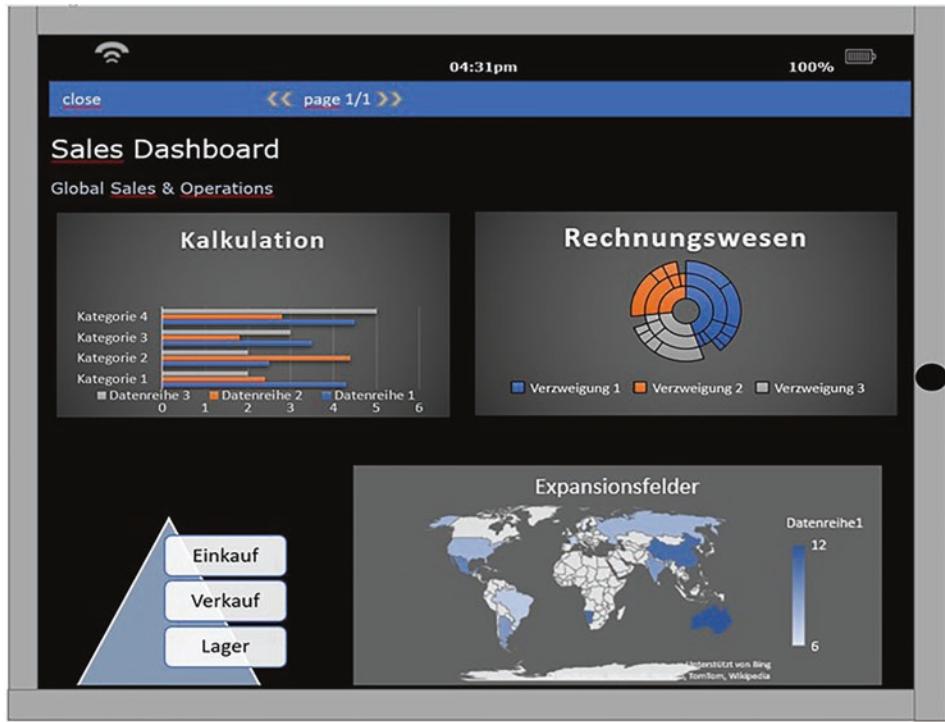


Abb. 1.6 Ein Beispiel für Dashboards auf Smartphones und mobilen Endgeräten (Eigene Darstellung)

Der Nutzer kann interaktiv auswählen, ob er die Daten in verdichteter oder detaillierter Form analysieren möchte. Des Weiteren stehen Filtermöglichkeiten zur Verfügung, die es beispielsweise erlauben, bestimmte Zeiträume einzuschränken. Auf diese Weise kann der Nutzer komplexe Zusammenhänge zwischen den Datensätzen verstehen (vgl. Bitkom, Big-Data-Technologien – Wissen für Entscheider, S. 76).

Bezüglich des fünften Baustein Data Governance soll stellvertretend das Thema Datenschutz nachfolgend erörtert werden.

Laut einer BARC-Studie halten 53 % der Unternehmen den Datenschutz für ein schwerwiegendes Problem bei der Einführung des Big Data-Projektes (vgl. Bange und Janoschek: Big Data Analytics 2014, S. 9). Dieses Problem bezieht sich auf die rechtmäßige Verwendung personenbezogener Daten. Zu diesen Daten gehören „solche Angaben, die persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbarer Person betreffen“. Dies sind beispielsweise Name, Adresse, Konfession, Gesundheits- oder Bankdaten (vgl. Bitkom: Management von Big-Data-Projekten., S. 24). Da die Kosten für den Hauptspeicher gesunken und Datenanalysewerkzeuge leistungsfähiger geworden sind, ist das Sammeln personenbezogener Daten durch das Unternehmen sowie durch Dritte stark angestiegen (vgl. Brücher: Rethink Big Data.,

S. 190). Damit Unternehmen bei der Nutzung personenbezogener Daten gesetzeskonform agieren, sollte eine Analyse der folgenden Aspekte berücksichtigt werden:

- Internationale rechtliche Rahmenbedingungen,
- Berücksichtigung der Interessen einer Person hinsichtlich der Verarbeitung ihrer Daten,
- Anonymisierung und Pseudonymisierung von personenbezogenen Daten.

Internationale rechtliche Rahmenbedingungen

Hier ist u. a. die EU-Datenschutzgrundverordnung mit ff. Gesetzen relevant:

- Art 5 Grundsätze für die Verarbeitung personenbezogener Daten
 - Transparenz
 - Zweckbindung
 - Datenminimierung
 - Richtigkeit
 - Speicher(dauer)begrenzung
 - Integrität und Vertraulichkeit
- Art 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling
- Art 35 Datenschutz-Folgeabschätzung

Zudem gelten in verschiedenen Ländern unterschiedliche Gesetze hinsichtlich der Verarbeitung und Speicherung personenbezogener Daten. In der Europäischen Union sind diese Gesetze etwas restriktiver als in den USA. Damit verbunden sind Fragen zum Umgang mit personenbezogenen Daten, wenn US-Unternehmen ihre Standorte in Europa haben (vgl. Heuer: Kleine Daten, große Wirkung. Big Data einfach auf den Punkt gebracht., S. 34). Dies führt beispielsweise zu folgendem Szenario. Die Daten werden in Deutschland gespeichert. Die Verarbeitung dieser Daten erfolgt aber in den USA. Nach der Verarbeitung werden die Daten in Deutschland per Internet übertragen und wieder gespeichert.

Die Auseinandersetzung mit dem obigen Szenario bedarf einer Ausarbeitung internationaler rechtlicher Standards. Diese Standards sollten den entsprechenden Schutz personenbezogener Daten sicherstellen. Andererseits dürfen sie nicht zu restriktiv sein, damit das Unternehmen überhaupt Nutzen aus diesen Daten gewinnen kann.

Des Weiteren sollte eine Transparenz der rechtlichen Rahmenbedingungen sowie eine kompetente Beratungshilfe bei der Einführung von Big Data-Technologien durch den zuständigen Gesetzgeber und die Datenschutzbehörden gewährleistet sein.

Berücksichtigung der Interessen einer Person hinsichtlich der Verarbeitung ihrer Daten

An dieser Stelle kann das Verbotsprinzip aus dem deutschen Bundesdatenschutzgesetz (BDSG) angeführt werden: Personenbezogene Daten dürfen verarbeitet und

genutzt werden, wenn die Person dies eingewilligt hat oder eine Rechtvorschrift diese Operationen explizit erlaubt (vgl. Das Bundesministerium der Justiz und für Verbraucherschutz: Bundesdatenschutzgesetz: § 4 Absatz 1).

Dieses Prinzip ist im Marketingbereich von großer Bedeutung. Es wird beispielsweise versucht Kundenbedürfnisse zu ermitteln und mittels Big Data-Verfahren ein Angebot individuell an diese Bedürfnisse anzupassen. Dieses Verfahren benötigt personenbezogene Daten, um korrekte Ergebnisse zu liefern. Dazu muss eine explizite Einwilligung der jeweiligen Kunden erfolgen. Der Kunde wird diese Einwilligung erteilen, wenn er einen Nutzen und keinen Schaden in der Freigabe seiner Daten sieht. Deshalb ist es für das Unternehmen wichtig, den ganzen Datenverarbeitungsprozess transparent zu machen und damit ein gewisses Maß an Vertrauen des Kunden zu gewinnen (vgl. Bitkom: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte., S. 44).

Personenbezogene Daten sind auch im Risikomanagement bedeutsam, z. B. Daten zur Betrugserkennung. Hier werden Verfahren eingesetzt, um Kunden vor dem Missbrauch seiner Kreditkarte durch Diebe und Betrüger zu schützen. Der Zahlungsdienstleister sollte die personenbezogenen Daten nur für diesen Zweck verwenden und die Daten nicht in die „Hände von Dritten“ geraten (vgl. Bitkom: Management von Big-Data-Projekten., S. 43).

Anonymisierung und Pseudonymisierung von personenbezogenen Daten

Laut des BDSG ist Anonymisieren „das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können“ (Das Bundesministerium der Justiz und für Verbraucherschutz: Bundesdatenschutzgesetz., S. 5). Um die Daten zu anonymisieren, muss das Unternehmen alle Identifikationsmerkmale, wie Nachname und Vorname löschen. Anschließend erfolgt das Aggregieren von anderen Merkmalen, was die Identifikation der Person unmöglich macht. Beispielsweise kann das Geburtsdatum durch das Geburtsjahr ersetzt werden. Die Anonymisierung erleichtert das Erfassen personenbezogener Daten, da so ein hoher Grad von Vertrauen der Kunden erreicht werden kann. Außerdem trägt es zu einem einfacheren Einhalten von datenschutzrechtlichen Rahmenbedingungen bei.

Die Aggregation der Daten ist hierbei besonders bedeutsam. Falls nicht aggregierte anonymisierte Daten an einen Dritten weitergegeben werden, kann dieser mittels der Verwendung von bestimmten Datenquellen die Person genau identifizieren (vgl. Bitkom: Management von Big-Data-Projekten., S. 45). Das hat das folgende Beispiel bewiesen.

Der amerikanische Online-Dienst America Online (AOL) hat in 2006 erledigte Suchanfragen veröffentlicht: insgesamt 20 Mio. Anfragen von 657.000 Anwendern. Das Ziel dieser Aktion war, den Wissenschaftlern die Möglichkeit zu geben, interessante Erkenntnisse aus diesen Daten zu ziehen. Die Identifikationsmerkmale wurden durch eine Nummer ersetzt, das Aggregieren wurde nicht ausgeführt. Nach ein paar Tagen haben die Journalisten der New York Times einen Anwender mit der Nummer 4.417.749 identi-

fizierte – eine Frau aus dem Staat Georgia. Damit wurde ein großer Skandal ausgelöst. Dem Leiter der Technologieabteilung und zwei anderen Mitarbeitern wurde gekündigt (vgl. Mayer-Schönberger und Cukier: Big Data., S. 195).

AOL wendete im o.g. Fall das Pseudonymisieren an, die ist „das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren“ (vgl. Das Bundesministerium der Justiz und für Verbraucherschutz: Bundesdatenschutzgesetz., S. 5). Dieses Verfahren kommt in solchen Bereichen zum Einsatz, wo die Verarbeitung jedes Profils eine wichtige Rolle spielt. Zu diesen Bereichen gehört beispielsweise die medizinische Forschung, bei der die Krankenhistorie analysiert wird (vgl. Bitkom: Management von Big-Data-Projekten., S. 7). Wie das Beispiel von AOL gezeigt hat, ist es nicht empfehlenswert, pseudonymisierte Daten an Dritte weiterzugeben.

Der sechste Baustein betrifft die Datenkonnektivität. Hierzu dient u. a. ein ETL-Prozess (Extraktions-, Transformations- und Ladeprozess), der traditionell in einem Business Intelligence System eingesetzt wird. Beim ETL-Prozess werden Daten aus mehreren gegebenenfalls unterschiedlich strukturierten Datenquellen in einer Zieldatenbank zusammengefasst und verbunden.

1.3 Datengetriebene Geschäftsmodelle

Bei der Überlegung datengetriebene Geschäftsmodelle zu entwickeln stehen ff. Fragen im Vordergrund:

1. Werden primär bzw. ausschließlich eigene und bestehende Datenbestände im Rahmen des Geschäftsmodells genutzt?
2. Zielt das Geschäftsmodell primär auf das aktuelle bestehende Geschäft und laufende Prozesse ab oder geht es um die Etablierung eines neuen Geschäftszweigs oder Geschäftsmodells?

Im Folgenden werden zunächst vier strategische Grundtypen untersucht.

Beim Geschäftsmodell Optimierung kann ein Mehrwert durch Auswertung bereits existierender Datenbestände entstehen. Dieser Ansatz bezieht sich auf die zugrunde liegende notwendige IT-Infrastruktur um das Ablegen, Verarbeiten, Analysieren und Darstellen immer größerer Datenmengen zu bewerkstelligen. Diese Vorgehensweise bietet einen Einstieg ins „Big Data-Business“ für viele Unternehmen. Vorreiter sind die Anbieter von Low-Cost-Flügen, die ihre IT-Systeme mit einer Vielzahl weiterer Parameter, z. B. aus dem Online-Verhalten, kombiniert und optimiert haben. Große Unternehmen überführen meist große Bestandteile alter Datenpools in neue Formate und Speichersysteme, um sie schnell und vor allem flexibel auswerten und visualisieren zu können.

Beim Geschäftsmodell Monetarisierung wird versucht mit bereits existierenden Daten neue Produkte zu schaffen. Ein Beispiel hierfür ist die anonymisierte Auswertung der Nutzer- und Standortdaten von Telefonnutzern zur Optimierung lokalisierter Dienste und

ortsbezogener Werbung oder dem Weiterverkauf von Transaktionsdaten. Ebenso entwickeln Internetunternehmen auf Basis der Nutzungsdaten und des Suchverhaltens neue Analysedienste. Zunehmend vermarkten auch Einzelhandelsunternehmen anonymisierte Transaktionsdaten an ihre Lieferanten aus dem Umfeld der Konsumgüter- und Lifestyle-Industrie.

Beim Geschäftsmodell Leverage werden bestehende Geschäftsmodelle und Dienstleistungen durch neue Daten „gehebelt“. In diesem Zusammenhang sind Reiseunternehmen zu nennen, die durch die Integration detaillierter Wetterprognosen sowohl ihre Marketingaktivitäten als auch die Auslastung ihrer Zieldestinationen optimieren können.

Zu diesem Geschäftsmodell gehört auch ein optimiertes Verkehrsmanagement in Metropolregionen über Mautsysteme, die den Verkehrsfluss über Preisadjustierungen steuern.

Das Geschäftsmodell Disrupt wird auch als „Kalifornischer Ansatz“ bezeichnet und stellt die „Königsklasse“ bei der Entwicklung neuer Geschäftsmodelle für die Data Economy dar. Hier werden auf Basis der Sammlung und Digitalisierung neuer Datenbestände neue Produkte/ Services geschaffen. Ein Beispiel hierfür sind die Nutzung von Energiedaten, um ortsbezogene Leistungsprognosen für die Betreiber von Solar- und Windparks anzubieten. Ebenso können durch digitale Kartographie von Städten neue Services, z. B. für die Hotellerie und Immobilienwirtschaft, angeboten werden. Weiterhin lassen sich die Geschäftsmodelle der Social Media-Monitoring Provider zu den „Disrupt“ -Modellen zählen, die das Feedback und Einstellungen der Nutzer zu verschiedenen Themen, Produkten und Markten zu aussagekräftigen Online-Analysen zusammentragen.

Neben der strategischen Einordnung lassen sich Big Data-Geschäftsmodelle auch nach Kategorien einordnen, die sich mit wachsender Reife des Marktes langsam als Produkt- bzw. Service-Gattungen etablieren.

Dazu gehört das Geschäftsmodell Analytics-as-a-Service.

Derzeit ist dies eine der wichtigsten und wachstumsstärksten Produkt- bzw. Dienstleistungskategorien. Hierbei werden Analysen und Prognosen über Cloud-basierte Plattformen bzw. gehostete IT-Infrastrukturen bereitgestellt. Diese beziehen sich häufig auf bestimmte Datentypen (Wetterdaten, Kundendaten, Social Media-Daten, Internetnutzungsdaten etc.) oder Unternehmensfunktionen (CRM, FuE, Controlling). Meist werden bestimmte Branchen, die sehr spezifische Analyse-Bedarfe sowie sehr spezifische Datentypen und Datenmengen verarbeiten müssen, adressiert.

Generische „Analytics-as-a-Service“ sind nicht so stark verbreitet, da hier meist eine hohe Anpassung und Integration von Daten und Features erfolgen muss.

Das Geschäftsmodell Data-as-a-Service ist bereits in einigen Branchen (z. B. der Online-Werbebranche) langjährig etabliert. Hier werden Nutzungsdaten u. a. über Cookies und Browser-Add-ons gesammelt, aggregiert und weiterverkauft, um die Zielgenauigkeit der Werbung zu erhöhen. Dazu gehört auch Datenstreams aus verschiedenen Social Media-Diensten zu extrahieren, zu normieren und mit relevanten Metadaten anzureichern. So werden die Ergebnisse des Social Media-Monitorings präziser und

entsprechende Social Media-Kampagnen lassen sich effizienter planen und umsetzen. Die Zusammenführung und Aufbereitung von Daten zum Zweck des Weiterverkaufs stellt hierbei ein attraktives Geschäftsmodell dar. Jedoch bedarf die Umsetzung entsprechender Vorhaben eine gründliche juristische Evaluierung sowie eine Vorab-Überprüfung der Kundenmeinung.

Das Geschäftsmodell „Data-infused Products“ beschäftigt sich mit der Entwicklung und dem Verkauf von „Data-infused Products“. Hierbei steht meist die Aufwertung bestehender Produkte durch mehr Daten-Intelligenz im Vordergrund. Hierzu gehören bspw. intelligente Stromzähler, die Gebäudeautomation sowie Werkzeuge oder Haushaltsgeräte, die entweder ressourcenärmer betrieben werden können oder dem Nutzer via Display und Steuerungskomponenten eine bessere und individuellere Handhabung bieten. Die IT-gestützte Sensorik im Auto und ein Großteil der Sicherheitsfunktionen gehören ebenso wie neue Devices (z. B. Wearables) dazu. Dies bieten neue Funktionen für Nutzer, wie beispielsweise Armbänder mit Sensoren zur Überwachung von Herzfrequenz und eingebautem Schrittzähler oder Skibrillen mit integriertem Head-Up-Display und GPS – zur Messung von Geschwindigkeit und Navigationsassistenz auf der Piste. Im Freizeitbereich verbreiten sich tragbare bzw. integrierte Kameras, deren hochauflösende Video-Signale übertragen und in Echtzeit verarbeitet werden müssen. Da sich die Bandbreite in den kommenden Jahren weiter erhöhen wird (z. B. durch 5G Netzbau) kann man davon ausgehen, dass mit sinkenden Kosten für die eingesetzten Technologien auch viele weitere neue Anwendungsgebiete gefunden werden, z. B. könnten Kinderwagen mit einer Sicherheits-Sensorik ausgestattet werden, die aktuell noch in PKW's verbaut wird.

Beim Geschäftsmodell Datenmarktplätze und Daten-Aggregatoren schaffen Marktplatzbetreiber Plattformen und einheitliche Standards für den Verkauf und die Nutzung verschiedener Datensätze oder Datenstreams. So bieten diese Modelle zum Vertrieb von Datenpaketen und Services an. Marktforschungs- und Beratungsunternehmen können so ihre Daten und Expertise auf der Plattform zu einheitlichen Konditionen vertreiben. Kunden erhalten Zugang über standardisierte APIs, sodass Abfragen, Visualisierungen oder auch die Programmierung eigener Analytik-Anwendungen möglich werden. Ein Beispiel hierfür ist auch aggregierte Location-Data für Software-Entwickler, eCommerce- und Internet-Unternehmen. Analysten gehen davon aus, dass sich die Datenmarktplätze branchen- und anwendungsspezifisch entwickeln werden.

1.4 Exemplarische Einsatzmöglichkeiten

Die Einsatzmöglichkeiten von Big Data Technologien sind vielfältig und betreffen sämtliche Unternehmensfunktionsbereiche. Der Einsatz von Big Data Technologien wird beispielhaft für das Themengebiet Social Media Marketing untersucht und anschließend anhand weiterer konkreter Unternehmensbeispiele in der Forschung und Entwicklung, in der Logistik, im Finanz-Risikocontrolling, dargestellt.

Social Media Marketing

Der Begriff Social Media beschreibt soziale Netzwerke bzw. soziale Plattformen, die zur gegenseitigen Kommunikation von Internetnutzern dienen. Unter Kommunikation wird in diesem Sinne nicht nur der Dialog zwischen den Nutzern, sondern darüber hinaus das Teilen von Erfahrungen oder Eindrücken sowie positive und negative Meinungsäußerungen hinsichtlich Marken, Produkten, Unternehmen o.ä. verstanden (vgl. Hilker 2010, S. 11). Die Internetnutzer haben durch die sozialen Medien die Möglichkeit Inhalte aktiv mitzugestalten. Das Spektrum der sozialen Plattformen ist umfangreich und umfasst soziale und berufliche Netzwerke, Foren und Blogs sowie Portale, auf denen bspw. Videos oder Fotos hochgeladen und betrachtet werden können. Dieses, über den Konsum von Inhalten hinausgehende, interaktive Mitgestalten von Inhalten der Nutzer, wird als Web 2.0 bezeichnet (vgl. Kleemann et al. 2012, S. 9).

Die Abb. 1.7 zeigt eine große Auswahl an Social Media Plattformen. Es stellt die „Social Media Welt“ mit all ihren relevanten Kommunikationskanälen dar.

Soziale Medien eignen sich sowohl für den internen als auch externen Unternehmenseinsatz. Bei einem externen Einsatz von Social Media sind vor allem die Marketing- und PR-Abteilung involviert, um das Unternehmen online zu platzieren. Als Plattformen dienen hierzu bspw. Twitter, Facebook, Blogs, YouTube oder Foren. Im Rahmen des internen Einsatzes dient Social Media zur Kommunikation innerhalb des Unternehmens. Hier werden vor allem Unternehmens-wikis oder Social-Bookmark-Dienste eingesetzt (vgl. Pfeiffer und Koch, 2011, S. 18).

Social Media Marketing ist eine Form des Online Marketing. Hierbei steht die Formulierung von Strategien und Maßnahmen zur Positionierung eines Unternehmens, deren Marke oder Produkte auf den verschiedenen Social Media Plattformen im Fokus (vgl. Friedrich 2012, S. 17). Dabei spielt die Bekanntmachung der eigenen Inhalte eine genauso wichtige Rolle wie der Kontaktaufbau zu den Nutzern der sozialen Plattformen (Kunden und potenzielle Kunden, Geschäftspartner etc.), um über die geteilten Inhalte, wie beispielsweise aktuelle Produktangebote, Produktneuerungen o.ä., kommunizieren zu können. Ein wichtiger Aspekt des Social Media Marketing ist somit die Nutzer und deren Meinung zu beachten sowie zu akzeptieren, ihnen „zuzuhören“ und auf sie mittels einer angemessenen Antwort einzugehen (vgl. Weinberg, Ladwig und Pahrmann 2014, S. 9). Das Social Media Marketing kann hinsichtlich des Marketing-Mix mit seinen vier Bereichen Produkt-, Preis-, Kommunikations- und Distributionspolitik, unterstützend eingesetzt werden. Die vielseitigen Einsatzmöglichkeiten gehen dabei von markenunterstützender Werbung über Gewinnspielaktionen bis hin zum Social Commerce (vgl. Lehning et al. 2015, S. 80). Ebenso zählen das Social Media Monitoring und das Social Customer Relationship Management zum Aufgabenbereich des Social Media Marketing. Generell kann im Social Media Marketing zwischen zwei Ansätzen unterschieden werden:

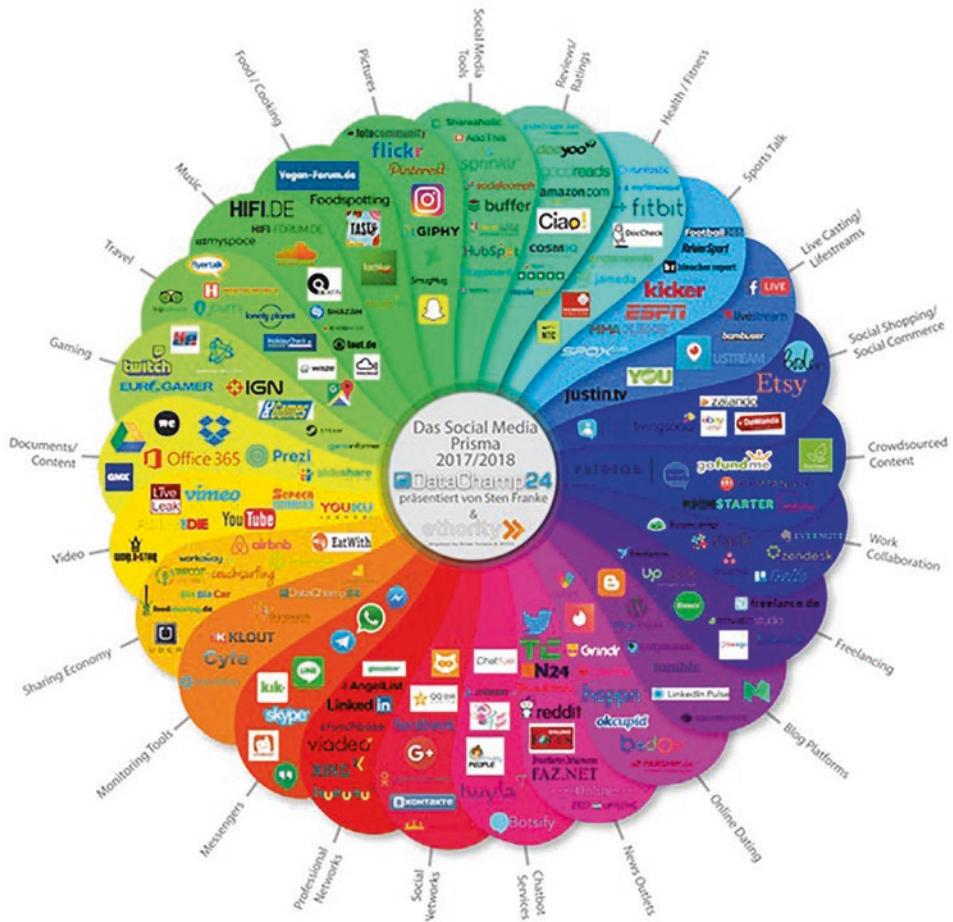


Abb. 1.7 Social Media Prisma (Quelle: <https://ethority.de/social-media-prisma/>)

Proaktiver Ansatz im Social Media Marketing

Der proaktive Ansatz ist vor allem durch die Kommunikation des Unternehmens zum Kunden gekennzeichnet. Die Kunden werden auf den Social Media Plattformen direkt angesprochen. Beispielsweise wird ihnen eine Feedback-Möglichkeit eröffnet, sei es in Form einer Facebook-Seite, eines Unternehmensblogs oder anderer sozialer Netzwerke. Darüber hinaus können die (potenziellen) Kunden bei diesem Ansatz aktiv in bestimmte Marketingaktivitäten eingebunden werden, bspw. bei der Produktgestaltung (Crowdsourcing). Der proaktive Ansatz verspricht, aufgrund seines Beziehungsaufbaus zu den Kunden, langfristig einen größeren Erfolg als der reaktive Ansatz.

Reaktiver Ansatz im Social Media Marketing

Der reaktive Ansatz zeichnet sich insbesondere durch die abwartende Haltung des Unternehmens aus. Die Social Media Plattformen werden im Rahmen des Social Media Monitoring nach Kommentaren und Meinungen über das Unternehmen oder deren Produkte durchsucht. Anschließend wird dann vom Unternehmen auf die Aussagen reagiert und geantwortet. Die Verhinderung einer negativen Unternehmensdarstellung durch negative Kommentare ist dabei das primäre Ziel, weshalb die Unternehmen entsprechend darauf eingehen und die Kunden „aufklären“. Dieser Ansatz eignet sich vor allem dafür, dass Nutzerverhalten zu studieren und einen Überblick über die Kommunikation im Social Web zu erhalten (vgl. Grabs und Bannour 2011, S. 66 ff.).

Im Social Media Marketing ist es sinnvoll, dass Social Media Guidelines erstellt werden, die festlegen, wie sich u. a. Mitarbeiter im Social Web verhalten sollten. Die Richtlinien geben den Mitarbeitern eine Orientierung was bspw. hinsichtlich ihres Verhaltens, der Kommunikation gegenüber Dritten oder dem Umgang mit vertraulichen Informationen beachtet werden sollte. Darüber hinaus können in den Guidelines Informationen zum Datenschutz, zur Sicherheit und zum Urheberrecht festgehalten werden (vgl. Aßmann und Röbbeln 2013, S. 71 ff.).

Eine Übersicht über die Anwendungsziele bei Nutzung von Big Data für Marketing und Vertrieb liefert die Abb. 1.8.



Abb. 1.8 Anwendungsziele bei Nutzung von Big Data für Marketing und Vertrieb (Bange und Janoschek 2014, S. 34)

Unternehmensbeispiel Telefónica

Telefónica ist ein spanischer Telekommunikationskonzern, der ein Projekt mit dem Namen „Smart Steps“ etabliert hat. Das Projekt bestand darin, ortsbzogene Daten von Telefónica-Nutzern mithilfe des Mobilfunknetzes zu sammeln, zu anonymisieren und an Dritte zu verkaufen. Die Anonymisierung der Daten verhinderte die Identifikation von Personen. Mit diesen Daten kann bspw. der Besitzer eines Einkaufszentrums erfahren, wie viele Menschen zu einer bestimmten Uhrzeit einen bestimmten Ort besucht haben. Somit kann die Personalplanung verbessert und Personalkosten reduziert werden.

Außerdem kann ein Mobiltelefonhersteller mit Mobilfunkdaten die Signalstärke seiner Geräte sowie die auf sie einwirkenden Faktoren analysieren. Mittels dieser Analyse kann der Hersteller die Empfangsleistung der Mobiltelefone verbessern (vgl. Mayer-Schönberger und Cukier: Big Data, S. 136).

Forschung und Entwicklung

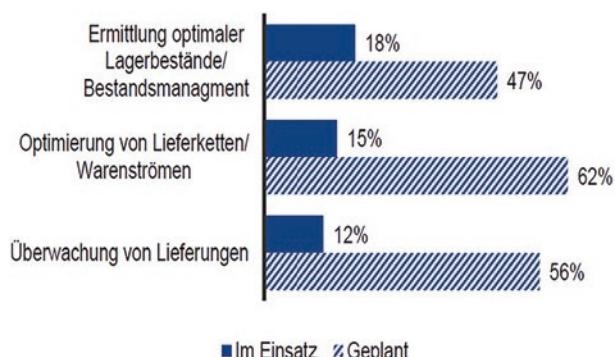
Eine Übersicht über die Anwendungsziele bei Nutzung von Big Data für Forschung und Entwicklung liefert die Abb. 1.9.

Unternehmensbeispiel UPS

UPS gehört zu den größten internationalen Logistikunternehmen mit über 495.000 Mitarbeitern. Es besitzt mehr als 125.000 Transportfahrzeuge (vgl. UPS, 2020).

Das Unternehmen beabsichtigt eine kontinuierliche Überwachung der Lieferungen, eine Optimierung der Routen sowie eine Reduzierung der Lieferkosten. Zur Zielerreicherung wurde jedes Transportfahrzeug mit einem Sensorsystem ausgestattet. Diese Sensoren erfassen beispielsweise die Daten über die Geschwindigkeit, die Richtung, den Benzinverbrauch sowie andere technische Parameter und übermitteln diese in eine Datenbank. Außerdem verfügen die Transportfahrzeuge über einen Global Positioning System (GPS)-Empfänger. Die dadurch erfassten GPS-Daten werden zur Analyse der Verhaltensweise und der Gewohnheit des Fahrers verwendet. Anhand der gesammelten Daten wird beispielsweise analysiert, wie oft der Fahrer gestoppt oder die falsche Route

Abb. 1.9 Anwendungsziele bei Nutzung von Big Data für Distribution und Logistik (Bange und Janoschek 2014, S. 34)



ausgewählt hat. Des Weiteren werden technische Parameter bewertet, um die Pannen vorhersagen zu können (vgl. Datafloq, o. J.).

Mittels des Sensorsystems kann UPS die Daten über ein bestimmtes Fahrzeug zu jedem Zeitpunkt erfassen. Der Umfang der gesammelten Daten beträgt mehr als 16 PB. Des Weiteren sollen die erfassten Daten sofort analysiert werden, damit die beste Route ermittelt werden kann. Seitdem das Sensorsystem eingeführt und eigene Algorithmen zur Analyse der erfassten Daten entwickelt wurden, konnte das Unternehmen bei den täglichen Routen insgesamt 85 Mio. Meilen Wegstrecke einsparen (ca. 30 Million Dollar täglich). Die Algorithmen analysieren auch die Daten über den technischen Zustand des Fahrzeuges und können so vorhersagen, wann der jeweilige Transporter einer Wartung unterzogen werden muss. Dies hat wiederum zur Verminderung der Pannen während der Abwicklung einer Lieferung geführt (vgl. Davenport und Dyché: Big Data in Big Companies., S. 4).

Darüber hinaus konnte UPS den Kunden einen Service namens „My Choice“ zur Verfügung stellen. Dieser Service ermöglicht, die Lieferzeit per Internet zu kontrollieren. Außerdem ist es möglich, den aktuellen Standort einer bestimmten Lieferung zu überwachen (vgl. Datafloq, o. J.). Dieser Service hat zur Steigerung der Kundenzufriedenheit beigetragen.

Mittels der GPS-Daten kann UPS das Verhalten eines jeden Fahrers analysieren. Anhand der Analyse kann festgestellt werden, ob der Fahrer zum Zweck der Verbesserung seiner Qualifikation an einer zusätzlichen Schulung teilnehmen soll, z. B. wenn der Fahrer häufig die richtige Abfahrt während der Fahrt verpasst (vgl. Datafloq, o. J.). Dadurch entsteht ein Zeitverlust und infolgedessen ein Profitverlust des Unternehmens.

Des Weiteren konnte durch Analyse der Sensor- und GPS-Daten festgestellt werden, dass das Linksabbiegen an Kreuzungen vermieden werden sollte. Dadurch wurde die Anzahl der Unfälle vermindert. Darüber hinaus wurde der Benzinverbrauch reduziert, weil keine Standzeiten wegen des Gegenverkehrs entstanden. Diese beiden Tatsachen werden nun bei der Berechnung der optimalen Routen in Betracht gezogen (vgl. Mayer-Schönberger und Cukier: Big Data., S. 114–115).

Eine Übersicht über die Anwendungsziele bei Nutzung von Big Data für Finanz- und Risikocontrolling liefert die Abb. 1.10.

Unternehmensbeispiel United Overseas Bank

Die United Overseas Bank ist eine Bank aus Singapur. Der gesamte Prozess zur Bewertung des Gesamtrisikos der Bank hat in der Vergangenheit ca. 18 h gedauert. Dabei werden ungefähr 8,8 Mrd. komplexe Berechnungen ausgeführt. Das Unternehmen zielt auf eine wesentliche Beschleunigung des Berechnungsprozesses bei der Berücksichtigung aller marktrelevanten Parameter (weit über 100.000) ab. Für die Problemlösung hat die Bank eine analytische Software-Lösung sowie eine In-Memory-Technologie eingeführt. Dabei wird die hohe Verarbeitungsgeschwindigkeit der Lösung genutzt, mit der alle relevanten Parameter ausgewertet werden. Der gesamte

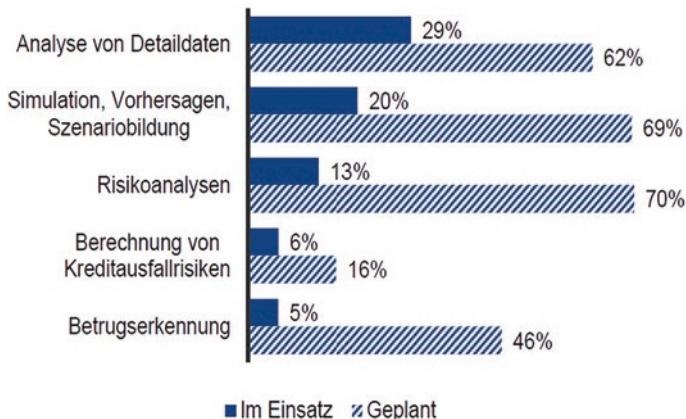


Abb. 1.10 Anwendungsziele bei Nutzung von Big Data für Finanz- und Risikocontrolling (Bange und Janoschek 2014, S. 35)

Berechnungsprozess vom Risiko dauert nach der Einführung der Big Data-Technologien nun wenige Minuten. Dabei können neue marktrelevante Parameter während der Ausführung der Berechnungen einbezogen werden (vgl. Bitkom: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte, S. 80).

Produktion, Service und Support

Eine Übersicht über die Anwendungsziele bei Nutzung von Big Data für Produktion, Service und Support liefert die Abb. 1.11.

Unternehmensbeispiel Vestas

Vestas ist einer der größten Produzenten und Betreiber von Windkraftanlagen (WKA) in der Welt. Dieses Unternehmen hat bereits mehr als 50.000 Wind-Turbinen in mehr als 70 Ländern aufgebaut. Die Auswertung der möglichen Positionen für WKA soll erheblich beschleunigt werden. Bei dieser Operation wird berechnet, „wie viel Energie in den nächsten Jahrzehnten dort erzeugt werden wird, wie hoch der Ertrag ist und wie schnell die Anlage sich rechnet“ (vgl. Digitalbusiness 2012). Ursprünglich hat die Auswertung mehrere Wochen gedauert. Zur Zielerreichung hat Vestas eine Big Data-Analysesoftware eingeführt und auch die Leistungsfähigkeit des Rechnersystems stark verbessert. Um die beste Position für den Aufbau einer WKA festzustellen, sollen verschiedene Faktoren, wie Geländehöhe, Satellitenbildern, Bewaldung, Stromnetzanbindung und historische Wetterdaten für die letzten elf Jahre so schnell wie möglich analysiert werden. Dabei sind die Qualität der Daten und die Unvorhersehbarkeit der Wetterdaten sowie die Ungenauigkeit der Satellitenbilder herausfordernd (vgl. Brücher: Rethink Big Data., S. 64). Zur Auswertung der Position einer WKA benötigt das Rechnersystem zurzeit wenige Stunden (vgl. Digitalbusiness 2012: Herren über den Datensturm). Des-



Abb. 1.11 Anwendungsziele bei Nutzung von Big Data für Produktion, Service und Support (Bange und Janoschek 2014, S. 32–33)

halb können Kundenanfragen viel schneller verarbeitet werden. Des Weiteren ermöglichen neue Rechnerkapazitäten eine bessere Position für WKA auszuwählen, was zur Reduzierung von Kosten pro Kilowattstunde des erzeugten Stroms führt. Außerdem verringern sich die Ausfallzeiten einer WKA, weil materialbelastende Turbulenzen mitberücksichtigt werden (vgl. Brücher: Rethink Big Data, S. 64).

Literatur

- Aßmann, S., Röbbeln, S.: Social Media für Unternehmen. Galileo Press, Das Praxisbuch für KMU. Bonn (2013)
- Bange, C., Janoschek, N.: :Big Data Analytics 2014, BARC - Business Application Research Center 2014. <https://barc-research.com/wp-content/uploads/2014/06/BARC-big-data-analytics-2014-EN.pdf> (2014). Zugegriffen: 02. Juli 2019
- BITKOM.: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. – Leitfaden des BITKOM. verfügbar unter <https://www.bitkom.org/Bitkom/Publikationen/Leitfaden-Big-Data-im-Praxiseinsatz-Szenarien-Beispiele-Effekte.html>(2012). Zugegriffen: 14. Dez. 2019
- Bitkom: Big-Data-Technologien – Wissen für Entscheider. Online verfügbar unter: <https://www.Bitkom.org/Bitkom/Publikationen/Big-Data-Technologien-Wissen-fuer-Entscheider.html>(2014). Zugegriffen: 07. Mai 2020
- Bitkom: Management von Big-Data-Projekten (2013)

- Brücher, C.: Rethink Big Data. MITP, Heidelberg (2013)
- Bundesministerium der Justiz und für Verbraucherschutz: Bundesdatenschutzgesetz. https://www.gesetze-im-internet.de/bdsg_2018/ (2018). Zugegriffen: 12. Juli 2019
- Datafloc: <https://datafloc.com/read/ups-spends-1-billion-big-data-annually/273> (o. J.). Zugegriffen: 26. Juni 2020
- Davenport, T., Dyché, J.: Big Data in Big Companies. https://docs.media.bitpipe.com/io_10x/io_102267/item_725049/Big-Data-in-Big-Companies.pdf (2013). Zugegriffen: 02. Juli 2019
- Digitalbusiness: Herren über den Datensturm, <https://www.digitalbusiness-cloud.de/herren-ueber-den-datensturm/> (2012). Zugegriffen: 02. Juli 2019
- Edlich, S.: NoSQL. Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Hanser (2011)
- Freiknecht, J.: Big Data in der Praxis. Carl Hanser, München (2014)
- Friedrich, M.: Social Media Marketingerfolg messen und analysieren. Wiley-VCH, Weinheim (2012)
- Grabs, A., Bannour, K.-P.: Follow me Erfolgreiches Social Media Marketing mit Facebook Twitter und Co. Galileo Press, Bonn (2011)
- Heuer, S.: Kleine Daten, große Wirkung. Big Data einfach auf den Punkt gebracht. https://www.lfm-nrw.de/fileadmin/lfm-nrw/nrw_digital/Publikationen/DK_Big_Data.pdf (2013). Zugegriffen: 28. März 2019
- Hilker, C.: Social Media für Unternehmer. Wie man Xing, Twitter YouTube und Co. erfolgreich im Business einsetzt. Linde, Wien (2010)
- Kleemann, F., Eismann, C., Beyreuther, T., Hornung, S., Duske, K., Voß, G.G.: Unternehmen im Web 2.0. Zur strategischen Integration von Konsumentenleistungen durch Social Media. Campus, Frankfurt a. M (2012)
- Lehning, T., Steiner, R., Holzer, M., Dürr, A.: Marketing-IT / ITMarketing. Eine Verständigungshilfe. Vogel Business Media GmbH & Co. KG, Würzburg (2015)
- Mayer-Schönberger, V., Cukier, K.: Big Data. Die Revolution, die unser Leben verändern wird, 1. Aufl. Redline Wirtschaft, München (2013)
- Pfeiffer, T., Koch, B.: Social Media Wie Sie mit Twitter Facebook und Co Ihren Kunden näher kommen. Addison-Wesley, München (2011)
- UPS.: FACT Sheet, <https://pressroom.ups.com/pressroom/ContentDetailsViewer.page?ConceptType=FactSheets&id=1426321563187-193> (2020). Zugegriffen: 23. Juni 2020
- Weinberg, T., Ladwig, W., Pahrmann, C.: Social Media Marketing Strategien für Twitter, Facebook & Co, 4. Aufl. O'Reilly GmbH & Co KG, Köln (2014)



Data Literacy als ein essenzieller Skill für das 21. Jahrhundert

2

Andreas Schmidt, Thomas Neifer und Benedikt Haag

„We are drowning in information, while starving for wisdom“
(Naisbitt 1982).

Zusammenfassung

Die heutige Geschäftswelt wird bestimmt durch Daten. Während das Problem normalerweise nicht ein Mangel an Daten ist, ist es vielmehr die Unfähigkeit, mit Daten adäquat umzugehen und daraus zielführende Schlussfolgerungen abzuleiten. Dabei ist Datenkompetenz heute genauso wichtig wie Lesen und Schreiben und eine Schlüsselqualifikation der Vierten Industriellen Revolution. Obgleich den meisten Unternehmen die Bedeutung von Daten in einer immer digitaleren Geschäftswelt durchaus bewusst ist, gibt es dennoch immer noch große Wissenslücken. Dieser Beitrag soll den Begriff Data Literacy und die damit verbundenen Fähigkeiten komprimiert beleuchten, während darüber hinaus mögliche Ansätze zur Integration des Themenfelds in Lehre und Praxis vorgestellt werden.

A. Schmidt (✉)

Wirtschaftswissenschaften, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Deutschland

E-Mail: mail.schmidt@email.de

T. Neifer

Wirtschaftswissenschaften, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Deutschland

E-Mail: thomas.neifer@h-brs.de

B. Haag

Wirtschaftswissenschaften, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Deutschland

E-Mail: benni-haag@web.de

2.1 Notwendigkeit von Data Literacy

Im Zuge eines Interviews mit der internationalen Unternehmens- und Strategieberatung McKinsey sagte Google's Chief Economist Dr. Hal R. Varian im Jahr 2009:

„The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.“ (Varian [2009](#)).

Was Varian bereits 2009 postulierte, ist auch heute weiterhin Realität. So identifiziert eine vom Karrierenetzwerk LinkedIn durchföhrte Analyse zum Thema „The Skills Companies Need Most in 2020“ den Bereich **Daten- Analyse und Interpretationsfähigkeit** als eine der – im Vergleich zur Rangliste vom Vorjahr – am stärksten wachsenden beruflichen Fähigkeit (vgl. Pate [2020](#)).

Analysen von McKinsey und dem Stifterverband im Zuge des Hochschulbildung-report 2020 (vgl. Stifterverband für die Deutsche Wissenschaft e. V. [2020](#)) sowie Untersuchungen des World Economic Forums (WEF) unterstreichen diesen Fähigkeitsbedarf. So weist das WEF in der Studie „Jobs of Tomorrow – Mapping Opportunity in the New Economy“ das Cluster „Data and AI“ als eines der sieben wichtigsten Berufscluster mit aufstrebenden Perspektiven hinsichtlich der Gegenwart und insbesondere der Zukunft aus (vgl. Racheva et al. [2020](#), S. 18). Weitere Forschungsarbeiten des WEF kommen darüber hinaus zu dem Ergebnis, dass für die nächsten Jahre unter anderem länder- als auch branchenübergreifend, die verbreitete Adoption von Big Data Analytics als ein maßgeblicher Treiber des Wandels für das Geschäftswachstum von Unternehmen sein wird (vgl. Leopold et al. [2018](#), S. VII und 39 ff.).

Gestützt werden die Ergebnisse der Analysen des WEFs auch von den bisherigen Erkenntnissen des fortlaufenden Studienprojekts „The Data Literacy Project“ des Global Data Analytics Leaders Qlik. In einer weltweit durchgeföhrten Studie wurde untersucht, inwieweit es einen Zusammenhang zwischen der Leistung eines Unternehmens und der Datenkompetenz seiner Mitarbeiter gibt. Konkret wurde dabei anhand eines Data Literacy Index aufgezeigt, dass Unternehmen mit ausgeprägten Datenkompetenzen in der Belegschaft einen bis zu fünf Prozent höheren Unternehmenswert aufweisen als andere (vgl. Qlik [2018a](#)).

Trotz den positiven Berufsaussichten und dem hohen Wertschöpfungsversprechen sieht die Realität hinsichtlich der Fähigkeiten zu Datenkompetenzen in Unternehmen und bei jungen Erwachsenen in der Breite jedoch ernüchternd aus.

Zentrale Ergebnisse belegen den Handlungsbedarf zum Thema Datenkompetenz:

- **Die unternehmensweite Datenkompetenz ist gering:**

Lediglich **24 %** der Mitarbeiter des unteren und mittleren Managements fühlen sich sicher im Umgang mit Daten und können diese adäquat lesen, analysieren und ziel-führend für datenbasierte Entscheidungsfindungen einsetzen.

- **Selbst Führungskräfte gehen hinsichtlich Datenkompetenzen nicht voran:**

Nur **32 %** der sogenannten „C-Suite“, d. h. die oberste Hierarchie-Ebene eines Unternehmens, werden als datenkompetent wahrgenommen, was eine negative Strahlkraft hinsichtlich der Notwendigkeit der Thematik auf das mittlere und damit gleichzeitig auch das untere Management zur Folge hat.

- **Die junge Generation und damit zukünftige Arbeitnehmer sind unvorbereitet im Umgang mit Daten:**

Gerade einmal **21 %** der 16- bis 24-jährigen verfügen über Datenkompetenzen, was darauf hindeutet, dass Schulen sowie Hochschulen und Universitäten ihre Schüler und Studenten nicht entsprechend gut genug für die neue Arbeitswelt der Gegenwart und Zukunft vorbereiten (vgl. Qlik 2018b, S. 4).

Unabhängig davon wächst die heute generierte Datenmenge unaufhaltsam weiter. So prognostizierte das US-amerikanische Marktforschungs- und Beratungsunternehmen IDC im Jahr 2016 eine Verzehnfachung des Datenwachstums bis zum Jahr 2025, wo weltweit ein Datenvolumen von 163 Zettabyte erreicht werden soll. Ferner zeichnet sich eine zunehmende Verlagerung bei den Datenquellen ab, sodass bis 2025 ein Großteil der Daten nicht wie bisher von Privatnutzern, sondern vielmehr mit einem Anteil von 60 % von Unternehmen generiert werden, was Entscheidern in der Wirtschaft die Möglichkeit eröffnet, neue und innovative Geschäftsmodelle auf Grundlage dieser Datenbasis und den damit verbundenen Erkenntnissen umzusetzen (vgl. Seagate 2017).

Angesichts der oben aufgeführten Ergebnisse des „Data Literacy Project“ bzgl. einer Standort-bestimmung zur Verbreitung grundlegender Datenkompetenzen ist es wenig verwunderlich, dass in der unternehmensweiten Praxis von einem sogenannten „**Data Science Skill Gap**“ die Rede ist, d. h. die Nachfrage nach Datenwissenschaftlern übersteigt das Angebot – obgleich die Fach-zeitschrift „Harvard Business Review“ einst im Oktober 2012 die Rolle des Data Scientist als „The Sexiest Job of the 21st Century“ bezeichnete (vgl. Davenport und Patil 2012, S. 70–76). So zeigt ein im Jahr 2019 von der etablierten Online-Jobsuchmaschine Indeed generierter Report für den US-amerikanischen Raum auf, dass die Nachfrage nach Datenwissenschaftlern um 31 % gegenüber dem Vorjahr und um 256 % seit dem Jahr 2013 anstieg. Demgegenüber nahmen die Suchanfragen von datenwissenschaftlich qualifizierten Arbeitssuchenden zum Vorjahr mit lediglich 14 % deutlich langsamer zu, was auf eine Lücke zwischen Angebot und Nachfrage hindeutet und damit den Data Science Skill Gap tendenziell bestätigt (vgl. Flowers 2019).

Gemäß dem IT-Beratungs- und Marktforschungsinstitut Gartner soll der Data Science Skill Gap dabei teilweise über Automatisierung ausgeglichen werden können. So sagte Gartner im Jahr 2017, dass:

„More than 40 percent of data science tasks will be automated by 2020, resulting in increased productivity and broader usage of data and analytics by citizen data scientists [...].“ (Gartner 2017)

Gartner definiert dabei einen sogenannten „**Citizen Data Scientist**“ als eine datenkompetente Person, die im Rahmen einer Datenanalyse Modelle generiert, welche u. a. fortschrittliche diagnostische Analysen oder prädiktive und präskriptive Funktionen verwenden, ihre eigentliche Hauptaufgabe im Unternehmen jedoch in der Regel außerhalb des Bereichs der Statistik und Analyse liegt. Unterstützt werden sie dabei über automatisierte Lösungen entlang des Data Science Prozesses und sind so in der Lage, anspruchsvolle Analysen durchzuführen, für die zuvor mehr Fachwissen erforderlich gewesen wäre, sodass sie fortschrittliche Analysen liefern können, ohne über die Fähigkeiten eines konkreten Data Scientist zu verfügen (vgl. Gartner 2017).

Automatisierungsansätze und Potenziale werden dabei sowohl wissenschaftlich sowie anwenderorientiert diskutiert (vgl. Saad et al. 2019; Dietz 2019), als auch bereits in der Praxis durch verschiedene Lösungsanbieter wie u. a. vgl. Alteryx, DataRobot oder dotData (2020) gelebt und so Data Science zunehmend demokratisiert. Nichtsdestotrotz Bedarf es auch im Rahmen der Arbeit mit automatisierten Tools und dabei insbesondere für das Verständnis im Zuge der Ergebnisauswertung ein gewisses Datenverständnis.

Subsummieren lassen sich die eingangs im Kapitel beschriebenen Fähigkeiten von Hal Varian unter dem zunehmend an Bedeutung gewinnenden Begriff Data Literacy. Auskunft darüber was konkret unter dem Begriff zu verstehen ist, zeigt der nachfolgende Kapitelpunkt auf.

2.2 Data Literacy als Begriff

Der Begriff **Data Literacy** kann insgesamt als ein aufstrebender Themenbereich betrachtet werden. Belege hierfür liefern eine Analyse zum weltweit relativen Suchinteresse nach dem Begriff „Data Literacy“ via Google Trends, mit dem Ergebnis einer bis heute stetig aufsteigenden Tendenz des Suchinteresses ab 2015 (vgl. Google LLC 2020), während eine Analyse über die Metasuchmaschine der vom Verlag Elsevier betriebenen wissenschaftlichen Forschungsdatenbank ScienceDirect ein ähnliches Bild zum Begriff zeichnet, siehe Abb. 2.1.

Definitionsversuche zum Begriff Data Literacy werden in der Literatur divers und aus unterschiedlichen Richtungen diskutiert. Eine in der Literatur weitläufig etablierte und oft zitierte Definition des Begriffs fußt auf dem Knowledge Synthesis Report von Ridsdale et al. (2015), in dem Data Literacy auf Basis eines ausführlichen Literature Reviews auf sehr prägnante Weise als die Fähigkeit definiert wird, „Daten auf kritische Art und Weise zu sammeln, zu managen, zu bewerten und anzuwenden“ (Risdale 2015).

Eine eher praxisnahe und anschaulichere Definition stammt vom IT-Beratungs- und Marktforschungsinstitut Gartner, die Data Literacy definieren als:

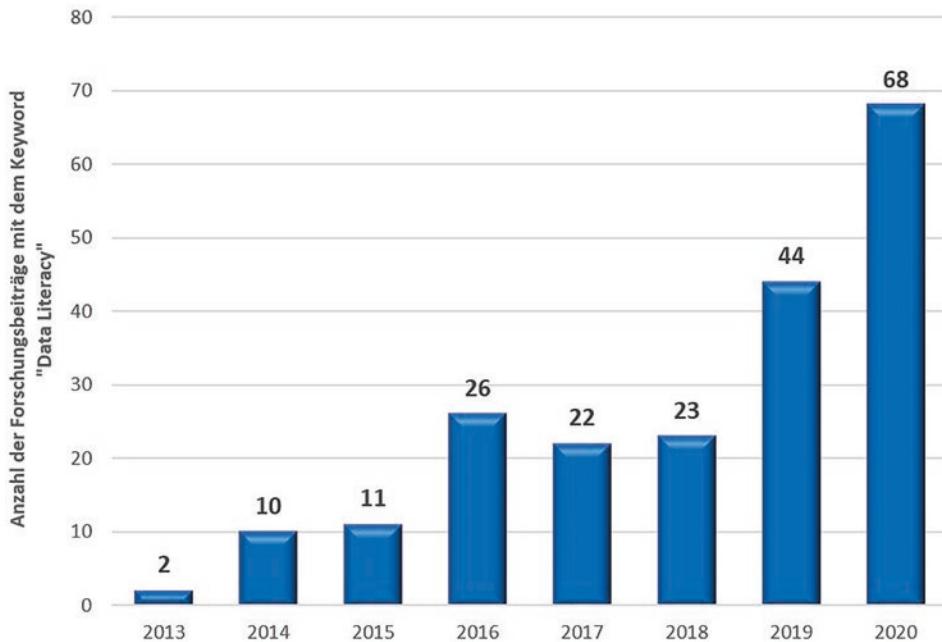


Abb. 2.1 Häufigkeit der Erwähnung von „Data Literacy“ in der Forschungsdatenbank ScienceDirect im Zeitverlauf (Eigene Darstellung)

„the ability to read, write and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied – and the ability to describe the use case, application and resulting value.“ (Gartner 2019)

So legt die Definition von Gartner nicht nur den Fokus auf die Fähigkeiten im Umgang mit Daten samt technischem und methodischem Hintergrundwissen, sondern bindet auch unternehmerische Aspekte mit ein, die sich auf die Implementierung entsprechender Datenprodukte und des damit verbundenen Wertzuwachs im Rahmen eines Datenanalyse-Projektes beziehen.

Data Literacy wird ferner in der Literatur als ein Schnittstellenbegriff definiert, der fließende Übergänge mit einer Reihe von weiteren Literacies, siehe Abb. 2.2 aufweist. Data Literacy interagiert dabei mit allen sechs dieser Ansätze und baut in Teilen auf ihnen auf. Die unterschiedlichen Abgrenzungen zueinander sind jedoch bis heute nicht eindeutig geklärt (vgl. Risdale 2015; Schüller et al. 2019, S. 16 und 23).

Inwiefern sich dabei der konkrete Kompetenzrahmen im Rahmen von Data Literacy ausgestaltet, soll nachfolgender Kapitelpunkt im Detail beleuchten.

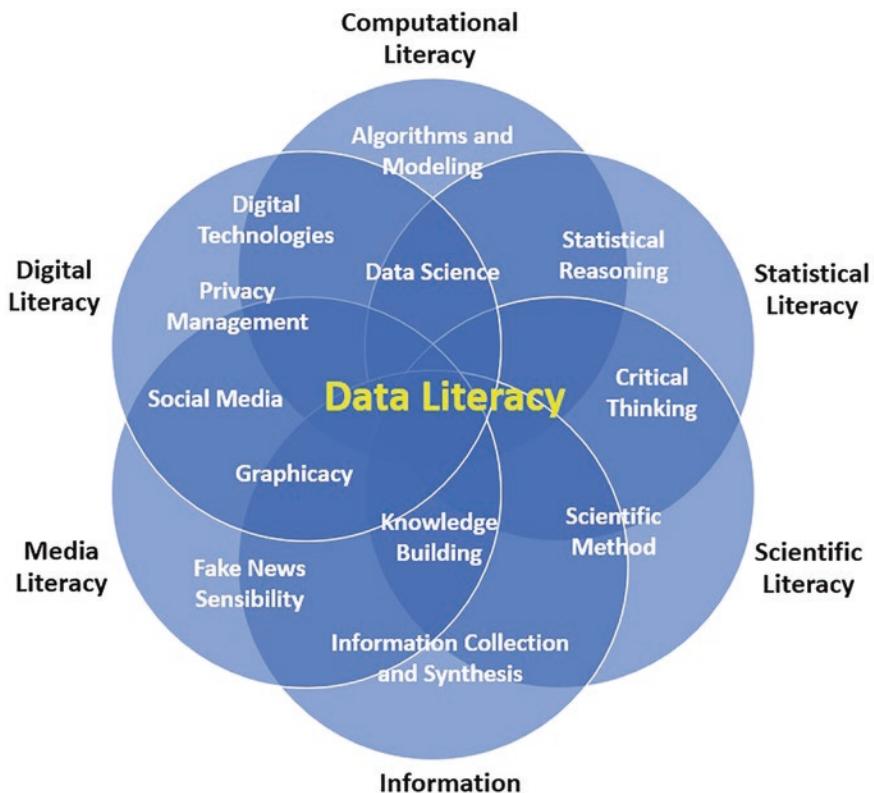


Abb. 2.2 Data Literacy als Schnittstellenbegriff (Eigene Darstellung)

2.3 Data Literacy Skills im Detail

Obgleich es viele wissenschaftlich diskutierte Frameworks zu den notwendigen Skills von Data Literacy gibt (vgl. Bhargava et al. 2015; Matthews 2016; Pedersen & Caviglia 2019; Schüller et al. 2019), stellt sich der bereits zuvor in Abschn. 2.2 erwähnte Knowledge Synthesis Report von Ridsdale et al. (2015) als die derzeit wissenschaftlich etablierteste und am häufigsten zitierte Quelle zu Data Literacy dar. Ridsdale et al. (2015) teilt dabei relevante Data Literacy Skills über eine Synthese aus 32 Publikationen in fünf Hauptbereiche auf (vgl. Ridsdale et al. 2015, S. 38). In Ergänzung dazu wird das Framework von Schüller et al. (2019, S. 90–108) als eine der neusten Studien für ein Data Literacy Framework herangezogen, um das Framework von Ridsdale et. al (2015) in fehlenden Punkten zu ergänzen. Die identifizierten Hauptbereiche mit den dazugehörigen Fähigkeiten lassen sich im Wesentlichen wie folgt beschreiben:

Conceptual Framework – Introduction to Data:

Dies bedeutet ein generelles Grundverständnis für Daten entwickeln, d. h. was sind Daten (Daten-Informationen-Wissen Pyramide), welche Daten-Typen gibt es und in welcher Form können sie vorliegen (Strukturierungsgrad der Daten) sowie ein Verständnis für verschiedene Datenkonzepte wie u. a. Big Data (siehe hierzu auch Big Data Literacy vgl. Sander 2020) und dabei insgesamt den Mehrwert von Daten verstehen und zu Problemlösungen nutzen können.

Data Collection:

Auf Basis der für ein Datenprojekt jeweiligen Fragestellung, adäquate, qualitative sowie nutzbare Datenquellen identifizieren können und über geeignete Kenntnisse sowie rechtliche Rahmenbedingungen zur Sammlung jeweiliger Daten für weitergehende explorative Analysen verfügen (wie u. a. Bulk Downloads, APIs, Web Scraping).

Data Management:

Bezieht sich auf ein Grundverständnis für die Organisation von Daten über entsprechende Daten-Management-Methoden sowie -Werkzeuge und die Einbettung der Daten in eine für die Analyse entsprechend notwendige Infrastruktur. Zudem sind damit Aspekte im Rahmen eines Data Cleaning sowie der Datenkonvertierung gemeint, wo es darum geht, die für einen Datenanalyse-Prozess vorliegenden Daten auf Anomalien und Unregelmäßigkeiten zu prüfen und diese mit geeigneten Methoden zu beheben. Ferner werden unter dem Oberbegriff die Bereiche Datenkonvertierung, d. h. das Wissen über verschiedene Datentypen und Konvertierungsmethoden, als auch das Metadatenmanagement mit Zuweisung von passenden Metadatenbeschreibungen zum Originaldatensatz verstanden, während ein allgemeines Wissen zu Data Security Thematiken den Themenkomplex abschließen.

Data Evaluation and Communication:

Umfasst im Wesentlichen ein Verständnis für alle Prozessschritte zur Wertschöpfung von Daten. Dazu gehören in erster Linie Kenntnisse über die relevanten Datenanalyse-Tools, -Techniken und -Methoden und deren passende Auswahl und Anwendung. Ferner sind grundlegende Kenntnisse zur Datenanalyse, insbesondere dabei im Bereich der Statistik erforderlich. Hierzu zählt z. B. die Durchführung einer explorativen Datenanalyse. Dabei werden entweder neue Fragestellungen aus Daten generiert oder aber eine bereits formulierte Fragestellung kritisch in Bezug auf vorliegende Daten u. a. auf Plausibilität oder relevante Auffälligkeiten geprüft. Zu beiden Szenarien bedarf es die Fähigkeiten, Daten und z. B. daraus generierte Tabellen, Graphen oder Diagramme lesen, interpretieren und daraus bedeutsame Erkenntnisse identifizieren zu können. Im weiteren Verlauf benötigt es Erfahrung darin, die entsprechend passende Modellierungsmethode (z. B. aus dem Bereich Machine Learning) auszuwählen, auf die Daten anzuwenden

und fortwährend das Modell auf Schwächen und Artefakte (z. B. Overfitting) zu untersuchen sowie diesen entgegenzuwirken. Die abschließend generierten Ergebnisse lassen sich anschließend mittels eines Verständnisses über geeignete Daten-Visualisierungsmethoden als auch Data-Storytelling organisieren sowie Handlungsempfehlungen visuell und zielgruppengerecht kommunizieren, um auf der Basis datenbasierte Entscheidungen zu treffen.

Data Application:

Unter dem Aspekt Data Application werden eher übergeordnete Kompetenzen im Rahmen von Datenanalysen verstanden. Darunter fallen u. a. Projektmanagement-Kompetenzen zur Planung, Koordinierung und Dokumentation sowie Durchführung eines Datenprojekts. Dabei ist u. a. die Rede von der Fähigkeit zu fortwährendem kritischem Denken während eines Datenanalyse-Projekts sowie einem Bewusstsein für datenethische Fragestellungen. Auch bezieht sich der Bereich im Sinne einer Datenkultur auf die Förderung und Unterstützung einer Umgebung in der durch Kommunikation und Netzwerken innerhalb eines Unternehmens auf die Wichtigkeit der Nutzung von Daten hingewiesen und in einem offenen Lernklima zu einer aktiven Exploration mit Daten ermutigt wird. Für den Bereich der Forschung sind darüber hinaus Kenntnisse zu Datenzitiermethoden relevant sowie ein Verständnis dafür, wie und welche Daten sich rechtlich sowie ethisch korrekt teilen und verbreiten lassen.

Abgeschlossen wird der Themenkomplex hinsichtlich einer regelmäßigen Evaluation von datenbasierten Entscheidungen, indem über einen Feedback-Loop vorangegangene Entscheidungen oder Lösungen mit neuen Daten validiert oder falsifiziert werden, sodass entweder Entscheidungen beibehalten oder auf neuer Datenbasis neue Entscheidungen getroffen werden. Hierfür ist es notwendig, die Hintergründe jeweiliger datenbasierter Entscheidungen zu verstehen und argumentativ bewerten zu können.

Nachfolgende Abb. 2.3 fasst noch einmal die verschiedenen zuvor genannten Aspekte anschaulich zusammen.

2.4 Konzepte zur Implementation von Data Literacy in Lehre und Praxis

Dieser Kapitelpunkt soll sich der Frage widmen, inwiefern bereits Best-Practice Ansätze zur Integration von Data Literacy in Lehre und Praxis existieren. Bevor dabei die Praxis beleuchtet wird, soll nachfolgend zunächst ein Blick auf die Bildungslandschaft und dort diskutierte Ansätze geworfen werden.

Mit einer dezidierten Auseinandersetzung hinsichtlich der Integration von Data Literacy in die Lehre hat sich das Hochschulforum Digitalisierung im Zuge eines Literature Reviews auseinandergesetzt (vgl. Heidrich et al. 2018), während ferner der Stifterverband über den Förderwettbewerb "Data Literacy Education – Future Skills"

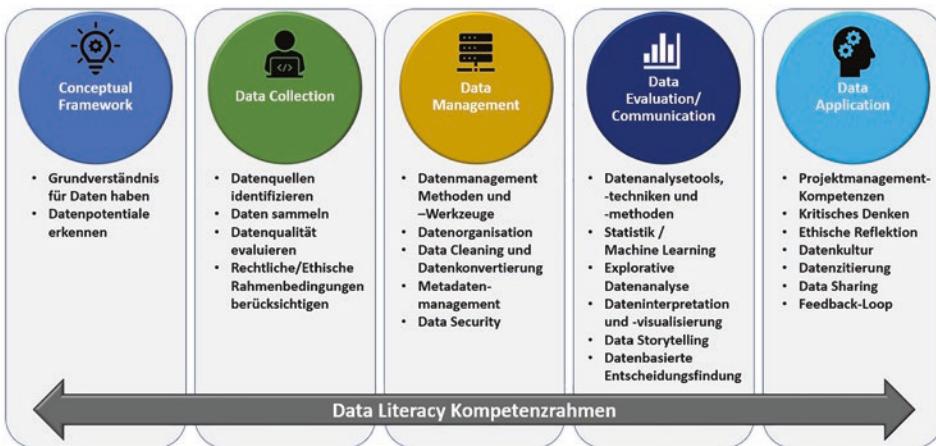


Abb. 2.3 Data Literacy Kompetenzrahmen (Eigene Darstellung)

insgesamt drei Hochschulen für herausragenden Ideen zur Vermittlung von Datenkompetenzen an Studierende aller Fächer ausgezeichnet hat – darunter die Leuphana Universität in Lüneburg, die Hochschule Mannheim sowie die Georg-August-Universität Göttingen (vgl. Stifterverband für die Deutsche Wissenschaft e. V. 2019).

Im Folgenden sollen diese drei als Leuchtturmprojekte fungierenden Gewinner des Förderwettbewerbs kurz skizziert werden.

Leuphana Universität Lüneburg

Das **Konzept DATA DRIVEN x (DATAx)** zielt darauf ab, Data Literacy in das Studienmodell der Leuphana zu integrieren. Das „x“ im Programmnamen steht dabei für das gemeinsame Präfix der vier englischen Begriffe „Exploration“, „Expertise“, „Experience“ und „Excitemen“ und soll die Leitideen von DATAx widerspiegeln. So sollen die Studierenden im Zuge eines Ansatzes des forschenden Lernens eigenständig Daten analysieren und Zusammenhänge entdecken (x-ploration). Dabei lernen die Studierenden, indem sie selbst praktische Erfahrungen machen und parallel dazu von den Erfahrungen der Lehrenden oder aber der involvierten Praxispartner profitieren (x-perience). Im Ergebnis sollen Studierende ihr Data-Science-Wissen als sogenannte Change Agents in Wirtschaft, Politik und Gesellschaft einbringen (x-pertise). Flankierende Best-Practice-Vorträge und Hands-On-Trainings sollen ferner Begeisterung für Daten schaffen und die intrinsische Motivation für ein lebenslanges Lernen in diesem Bereich fördern (x-citement). Um all dies zu erreichen, vermittelt DATAx zunächst im Online-Selbststudium Grundlagenwissen in den Bereichen Mathematik, Statistik und Programmierung. Die Methodenausbildung wird an Inhalte der Data Literacy Education angepasst und für alle Bachelor-Studierenden geöffnet. Zur vertiefenden Anwendung

bearbeiten im weiteren Verlauf die Studierenden in einem "Open Data Hacking Space" mit Echtdaten von Praxis- und Kooperationspartnern Praxisprojekte in der sie Datenanalysen und -visualisierungen umsetzen und ihre Ergebnisse veröffentlichen (für ausführliche Informationen siehe vgl. Leuphana Universität Lüneburg [2020](#)).

Hochschule Mannheim

Das Mannheimer **Modell Data Literacy Education (MoDaL)** sieht drei Maßnahmen vor: Auf der Stufe unimodal1 entwickeln Studienanfänger aller Fächer Kernkompetenzen im planvollen, verantwortlichen und kritischen Umgang mit Daten, während das Format als Ringvorlesung mit alternierenden Praxisübungen angelegt ist. Erfahrungen haben dabei gezeigt, dass der Fokus der Vorträge auf Praxisbezug und Anwendungsorientierung liegen muss und die Referent*innen ihre Impulsvorträge anschaulich aufbereiten sowie ein hohes Maß an Interaktion ermöglichen sollten. Auf der Stufe bimodal2 bearbeiten Studierende in interdisziplinären Kleingruppen weitgehend selbst organisiert im Learning-by-Doing Prinzip ein praxisorientiertes Datenprojekt. Auf der Stufe trimodal3 führen Studierende im Hauptstudium aller Bachelor-Programme in Kooperation mit Praxispartnern datengestützte Forschungsprojekte und Abschlussarbeiten durch und haben darüber die Möglichkeit alle in den vorherigen Stufen gelernten Themenkomplexe realitätsnah zu vertiefen (für ausführliche Informationen siehe vgl. Hochschule Mannheim [2020](#)).

Georg-August-Universität Göttingen

In dem **Projekt „Daten Lesen Lernen“** am Göttingen Campus entwickelt das Projektteam erstens die für Bachelorstudierende aller Fächer zugängliche Lehrveranstaltung Data Literacy Basics, die grundlegende Datenkompetenzen praxisorientiert und forschungsnah vermittelt. Dabei wird von den Studierenden ein 14-wöchiger Kurs durchlaufen, der beim Erlernen einer Skriptsprache anfängt und ferner Thematiken rund um die Bereiche Daten sammeln, lesen, schreiben und säubern sowie erkunden aufgreift. Flankiert wird der Basics-Kurs dabei durch Aspekte der statistischen Analyse sowie Ethik- und Datenschutzthemen, während eine Projektbearbeitung sowie Präsentation den Kurs abschließen. Zweitens wurde ein DataLab als Schnittstelle zwischen den verschiedenen Fächern, der regionalen Wirtschaft und gesellschaftlichen Akteuren aufgebaut, das eng mit der Lehrveranstaltung verknüpft ist und über Tutorials oder Data Consulting die praktische Anwendung von Datenanalysen ermöglicht. Um die Lehrveranstaltung nahtlos mit konkreten praktischen Anwendungsprojekten zu verknüpfen, kuratiert das Projektteam drittens eine qualitätsgeprüfte Sammlung von sogenannten Open Educational Resources in denen verschiedene Datenkompetenzbereiche vertieft werden können (für ausführliche Informationen siehe vgl. Georg-August-Universität Göttingen [2020](#)).

Wirft man den Blick auf Vorgehensweisen zur Einbettung von Data Literacy in der Praxis, so stellt man eine Forschungslücke in der wissenschaftlichen Literatur fest, die es künftig zu schließen gilt. Ein erster Ansatz für ein Framework zur Integration von Data Literacy in Unternehmen stammt aus dem Studienprojekt „The Data Literacy Project“ des Global Data Analytics Leaders Qlik. Konkret besteht das vorgeschlagene Framework dabei aus sechs Stufen, die sich wie folgt ausgestalten:

Step 1 – Planning and Vision:

Zu Beginn eines Data Literacy Programms ist es dabei notwendig eine Vision sowie ein anschließend definierten Plan mit Zielen und zeitlich festgelegten Meilensteinen zu entwickeln. Ausgangspunkt und Treiber sollte ein Datenexperte aus dem oberen Management, wie z. B. der Chief Data Officer sein, der durch Enthusiasmus für das Thema als Vorbild vorangeht, wodurch potenzielle Widerstände (wie z. B. fehlendes Budget oder fehlende Akzeptanz durch Mitarbeiter) gegen ein Data Literacy Programm reduziert werden.

Step 2 – Communication:

Ein definierter Kommunikationsplan ist im Implementierungsprozess von besonderer Bedeutung und sollte stets das „Warum“ hinter einer Initiative für ein Data Literacy Programm erklären – d. h. was bedeutet Datenkompetenz und warum ist es wichtig und welche langfristigen Vorteile sind damit für das Unternehmen und die Teilnehmer selbst verbunden (z. B. Förderung der Karriere durch berufliche Entwicklung). Ferner sollten die einzelnen Prozessschritte eines Data Literacy Programms jederzeit klar kommuniziert werden, um so ein hohes Maß an Transparenz und Orientierung zu gewährleisten.

Step 3 – Workforce assessment:

An dieser Stelle ist es notwendig über ein geeignetes Assessment (vgl. Schüller und Busch 2019, S. 36 ff.; Jones 2020; CAVORIT Consulting GmbH 2020) den IST-Zustand zu den vorhandenen Datenkompetenzen im jeweiligen Unternehmen zu erfassen. Die anschließend identifizierten Kompetenzniveaus der Organisation sollten dann über verschiedene Personas charakterisiert und für den jeweiligen Typ individuelle Lernpläne erstellt werden.

Step 4 – Cultural learning:

Die Implementation einer Data Literacy Initiative sollte wie ein herkömmliches Changemanagement Projekt angesehen werden. Ziel ist es dabei nicht, die Unternehmenskultur grundlegend zu verändern, sondern vielmehr weiterzuentwickeln und Data Literacy darin sukzessive als ein grundlegendes Element einzubetten sowie im Arbeitsalltag zu etablieren.

Step 5 – Prescriptive learning:

In diesem Punkt erhalten die Teilnehmer eines Data Literacy Programms ihrem in Step 3 identifizierten Lernstand entsprechende Lernpläne mit verschiedenen Lernangeboten. Diese Lernpläne in Form einer Roadmap sollen Mitarbeitern anschließend Orientierung geben, wo Defizite bestehen und welche Schritte notwendig sind, um ihre Datenkompetenzen auszubauen. Unternehmen sollten in dem Zusammenhang weiter die Wichtigkeit von Data Literacy für die Organisation hervorheben und Mitarbeitern zeitlichen und explorativen Raum geben und dazu ermutigen, sich regelmäßig mit Lerneinheiten auseinanderzusetzen.

Abb. 2.4 Framework zur Integration von Data Literacy-Initiativen in Unternehmen



Step 6 – Measurement:

In diesem letzten Punkt geht es darum, kontinuierlich das Data Literacy Programm über vorab definierte Metriken auf seinen Wirkungsgrad hin zu evaluieren. Dies kann entweder über Umfragen oder der Beobachtung hinsichtlich der Unternehmenskultur erfolgen (z. B. vermehrter Einsatz von Daten zu Entscheidungsfindungen in Meetings oder über die Anzahl abgeschlossener Kurse als auch Anzahl verliehener Zertifikate im Rahmen eines Lernprogramms).

Neben dem Wirkungsgrad des Data Literacy Programms, sollten darüber hinaus auch die Ausgestaltung der einzelnen sechs Schritte kontinuierlich in ausgewählten Intervallen evaluiert werden, um mögliche Schwachstellen zu identifizieren und daraufhin Verbesserungen vorzunehmen.

Nachfolgende Abb. 2.4 stellt das Framework nochmal anschaulich dar.

2.5 Fazit

Nahezu jedes Unternehmen muss inzwischen mit riesigen Datenmengen umgehen. Aber Daten zu sammeln und zu besitzen heißt noch lange nicht, sie auch zu verstehen. Obgleich Daten und Informationen zunehmend als die neue Sprache der Geschäftswelt fungieren, zeigen Untersuchungen, dass neben Mitarbeitern des unteren und mittleren Managements selbst Führungskräfte oft Defizite haben, Daten zu verstehen und mit ihnen zu arbeiten. Mangelnde Datenkompetenz bremst dabei viele Teams aus und blockiert die digitale Transformation der gesamten Organisation. Je besser Mitarbeiter

in der Lage sind, Daten zu lesen, mit ihnen zu arbeiten, sie zu analysieren und mit ihnen zu argumentieren, desto besser können sie auch ihre Aufgaben erfüllen und wirkungsvoll zur Zukunft ihres Unternehmens beitragen. So werden Mitarbeiter benötigt, die in der Lage sind, Daten zielgerichtet zu sammeln, zu managen, zu bewerten und anzuwenden. Dabei geht es im Wesentlichen nicht zwingend um hochspezialisierte Fachkräfte wie Data Scientists sondern vor Allem um die Mitarbeiter der Fachabteilungen. Auch sie müssen in der Lage sein, datengestützt zu arbeiten und zu entscheiden. Dazu gehört, in ihren jeweiligen Arbeitsbereichen datenbezogene Fragen zu formulieren, die Daten kritisch zu hinterfragen und Datenanalysen fachlich zu interpretieren. Die dafür erforderlichen als „Data Literacy“ bezeichneten Fähigkeiten sind Kernkompetenzen, über die Mitarbeiter heute verfügen müssen, damit die digitale Transformation gelingt.

Literatur

- Alteryx (Hrsg.): <https://www.alteryx.com/de> (2020). Zugegriffen: 28. Aug. 2020
- Bhargava, R. et al.: Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data. Data-Pop Alliance White Paper Series. Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute) and Internews (2015)
- CAVORIT Consulting GmbH (Hrsg.): Data Literacy Test - Exact Measurement of Data Literacy (2020). <https://www.dataliteracy.de/#test>, Zugegriffen: 28. Aug. 2020
- DataRobot (Hrsg.): <https://www.datarobot.com/> (2020). Zugegriffen: 28. Aug. 2020
- Davenport, T. H., Patil, D. J.: Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review 90(10), 70–76. (2012)
- Dietz, F.: Automation in Data Science <https://towardsdatascience.com/automation-in-data-science-f11fe389d49b> (2019). Zugegriffen: 28. Aug. 2020
- DotData (Hrsg.): <https://dotdata.com/> (2020). Zugegriffen: 28. Aug. 2020
- Flowers, A.: Data Scientist: A Hot Job That Pays Well <https://recruiting-indeed.de/wp-content/uploads/2020/06/Data-Scientist-report-final-PDF.pdf> (2019). Zugegriffen: 28. Aug. 2020
- Gartner (Hrsg.): Gartner Says More Than 40 Percent of Data Science Tasks Will Be Automated by 2020 <https://www.gartner.com/en/newsroom/press-releases/2017-01-16-gartner-says-more-than-40-percent-of-data-science-tasks-will-be-automated-by-2020> (2017). Zugegriffen: 28. Aug. 2020
- Gartner (Hrsg.): A Data and Analytics Leader's Guide to Data Literacy <https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy/#:~:text=Gartner%20defines%20data%20literacy%20as,case%2C%20application%20and%20resulting%20value.> (2019). Zugegriffen: 28. Aug. 2020
- Georg-August-Universität Göttingen (Hrsg.): Daten Lesen Lernen <https://www.uni-goettingen.de/de/daten+lesen+lernen/592287.html> (2020). Zugegriffen: 28. Aug. 2020
- Google LLC (Hrsg.): Analyse zum Begriff „Data Literacy“ über die Applikation Google Trends <https://trends.google.de/trends> (2020). Zugegriffen: 28. Aug. 2020
- Heidrich, J. et al.: Future Skills: Ansätze zur Vermittlung von Data Literacy in der Hochschulbildung. Hochschulforum Digitalisierung 114 (2018)
- Hochschule Mannheim (Hrsg.): Mannheimer Modell Data Literacy Education (modal) <https://www.modal.hs-mannheim.de/> (2020). Zugegriffen: 28. Aug. 2020

- Jones, B.: The Data Literacy Score - A Team-Based Assessment <https://dataliteracy.com/data-literacy-score/> (2020). Zugegriffen: 28. Aug. 2020
- Leopold, T. A. et al.: The Future of Jobs Report 2018 https://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf (2018). Zugegriffen: 28. Aug. 2020
- Leuphana Universität Lüneburg.: DataX <https://www.leuphana.de/universitaet/entwicklung/lehre/projekte/datix.html> (2020). Zugegriffen: 28. Aug. 2020
- Matthews, P.: Data literacy conceptions, community capabilities. *J. Community Inform.* **12**(3), 47–56 (2016)
- Naisbitt, J.: Megatrends: Ten New Directions Transforming Our Lives. (1982)
- Pate, D.: The Top Skills Companies Need Most in 2020—And How to Learn Them <https://learning.linkedin.com/blog/top-skills/the-skills-companies-need-most-in-2020and-how-to-learn-them> (2020). Zugegriffen: 28. Aug. 2020
- Pedersen, A. Y., Caviglia, F.: Data Literacy as a Compound Competence. In Antipova, T., Rocha, A. (Hrsg.) Digital Science: DSIC18: The 2018 International Conference on Digital Science, B 850, S. 166–173. Springer, Cham (2019). DOI: https://doi.org/10.1007/978-3-030-02351-5_21.
- Qlik (Hrsg.): The Data Literacy Index The \$500m Enterprise Value Opportunity Results Summary https://thedataliteracyproject.org/files/documents/Qlik%20-%20The_Data_Literacy_Index_October_2018.pdf (2018a). Zugegriffen: 28. Aug. 2020
- Qlik (Hrsg.): Lead with Data™ How to Drive Data Literacy in the Enterprise https://thedataliteracyproject.org/files/downloads/Qlik%20-%20How%20to%20drive%20Data%20Literacy%20within%20the%20Enterprise_October%202018.pdf (2018b). Zugegriffen: 28. Aug. 2020
- Racheva, V. et al.: Jobs of Tomorrow Mapping Opportunity in the New Economy https://www3.weforum.org/docs/WEF_Jobs_of_Tomorrow_2020.pdf (2020). Zugegriffen: 28. Aug. 2020
- Risdale, C. et al.: Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report. (2015). <https://doi.org/10.13140/RG.2.1.1922.5044>
- Saad, F. A. et al.: Bayesian synthesis of probabilistic programs for automatic data modeling. In: Proceedings of the ACM on Programming Languages, 3 (POPL): 1 (2019). DOI: <https://doi.org/10.1145/3290350>
- Sander, I.: What is critical big data literacy and how can it be implemented? *Internet Policy Rev.* **9**(2), 1–22 (2020). <https://doi.org/10.14763/2020.2.1479>
- Schüller, K. et al.: Future Skills: Ein Framework für Data Literacy https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_AP_Nr_47_DALI_Kompetenzrahmen_WEB.pdf (2019). Zugegriffen: 28. Aug. 2020
- ScienceDirect.: In: Elsevier (Hrsg.) <https://www.sciencedirect.com/> (2020). Zugegriffen: 28. Aug. 2020
- Schüller, K., Busch, P.: Data Literacy: Ein Systematic Review. (2019)
- Seagate (Hrsg.): Studie von IDC und Seagate: Weltweite Datenmenge verzehnfacht sich bis 2025 auf 163 ZB <https://www.seagate.com/de/de/news/news-archive/seagate-advises-global-business-leaders-and-entrepreneurs-pr-master/> (2017). Zugegriffen: 28. Aug. 2020
- Stifterverband für die Deutsche Wissenschaft e.V. (Hrsg.): Land und Stifterverband fördern Hochschulen mit drei Millionen Euro für die Vermittlung von Datenkompetenzen https://www.stifterverband.org/pressemittelungen/2019_11_18_data_literacy_education_nrw (2019). Zugegriffen: 28. Aug. 2020
- Stifterverband für die Deutsche Wissenschaft e.V. (Hrsg.): welche Fähigkeiten werden in Zukunft benötigt? Hochschul-Bildungs-Report 2020 <https://www.hochschulbildungsreport2020.de/2019/welche-faehigkeiten-werden-in-zukunft-benötigt> (2020). Zugegriffen: 28. Aug. 2020
- Varian, H.: Hal Varian on How the Web Challenges Managers. *McKinsey Quarterly* 1(2.2), <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers> (2009). Zugegriffen: 28. Aug. 2020



Management von Big Data Projekten

3

Andreas Gadatsch und Dirk Schreiber

Zusammenfassung

Big Data Projekte unterscheiden sich von klassischen IT-Projekten wegen der Vernetzung mit Geschäftsmodellen hinsichtlich der Erarbeitung von Anforderungen, der wirtschaftlichen Tragweite und der Vorgehensweise in der organisatorischen Umsetzung. Der Beitrag führt zunächst in die Rahmenkonzepte des klassischen Informationsmanagements ein, um die aktuellen Trends einzuordnen. Er geht anschließend auf die aktuellen Fragestellungen und Auswirkungen der Digitalisierung ein und thematisiert hieran die Herausforderungen von Big Data Projekten. Schließlich stellt er bekannte Ansätze von Vorgehensmodellen für die Durchführung von Big Data-Projekten vor und geht auf die Reifegradmessung von Big Data in Organisationen ein. Abschließend werden Auswirkungen auf das Informationsmanagement diskutiert.

3.1 Konzeptioneller Rahmen des Informationsmanagements

Der Abschnitt skizziert zunächst die an dem logistischen Prinzip orientierte Grundidee des Informationsmanagements. Er stellt zwei im deutschsprachigen Raum bedeutsame Rahmenkonzepte zum Informationsmanagement vor. Abschließend wird aufgezeigt, inwieweit sich Big Data Vorhaben in diese Rahmenkonzepte einordnen lassen.

A. Gadatsch (✉) · D. Schreiber

FB Wirtschaftswissenschaften, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Deutschland
E-Mail: andreas.gadatsch@h-brs.de

D. Schreiber

E-Mail: dirk.schreiber@h-brs.de

3.1.1 Überblick

Informationsmanagement umfasst die betrieblichen Aktivitäten, die zum adäquaten Einsatz der Ressource Information in einem Unternehmen beitragen. Dabei lässt sich die geforderte Adäquatheit anhand des aus der Materialwirtschaft bekannten logistischen Prinzips spezifizieren. Das logistische Prinzip, angewendet auf die Ressource „Information“, fordert die Vorhaltung

- der richtigen Information (vom Empfänger verstanden und benötigt),
- zum richtigen Zeitpunkt (für Entscheidungen ausreichend),
- in der richtigen Menge (so viel wie nötig, so wenig wie möglich),
- am richtigen Ort (beim Empfänger verfügbar),
- in der erforderlichen Qualität (ausreichend detailliert und wahr, unmittelbar verwendbar) (Augustin 1990).

Das betriebliche Informationsmanagement behandelt somit die Frage, wie sich Informations- und Kommunikationstechnologien bestmöglich zur Erreichung der Unternehmensziele nutzen lassen.

Zur Darstellung der daraus resultierenden vielfältigen Aufgaben im Informationsmanagement haben sich Rahmenkonzepte zum Informationsmanagement bewährt. Es existiert mittlerweile eine Vielzahl von Rahmenkonzepten zum Informationsmanagement. Nachfolgend sollen zwei Rahmenkonzepte zum Informationsmanagement, die insbesondere im deutschsprachigen Raum große Verbreitung erfahren haben, für diesen Beitrag vorgestellt werden.

3.1.2 Aufgabenorientiertes Ebenenmodell

Das aufgabenorientierte Ebenenmodell ist ein von Krcmar entwickeltes Rahmenkonzept zum Informationsmanagement (vgl. Krcmar 2015). Es verbindet die Ideen der aufgabenorientierten Konzepte (vgl. beispielsweise Heinrich und Lehner 2005), die eher aufzählend aber ohne sachlogische Struktur das Informationsmanagement beschreiben, mit den Ebenenmodellen zum Informationsmanagement (vgl. beispielsweise Wollnik 1988). Diese betonen basierend auf dem Gliederungsmerkmal „Techniknähe“ die daraus resultierende Sachlogik, gehen allerdings nur implizit auf die Sachaufgaben des Informationsmanagements ein. Der Grundaufbau des aufgabenorientierten Ebenenmodells ist in Abb. 3.1 illustriert und wird nachfolgend, basierend auf der ausführlichen Darstellung in (vgl. Krcmar 2015), kurz beschrieben.

Es besteht aus den drei Ebenen

- Management der Informationswirtschaft,
- Management der Informationssysteme und
- Management der Informations- und Kommunikationstechnik.

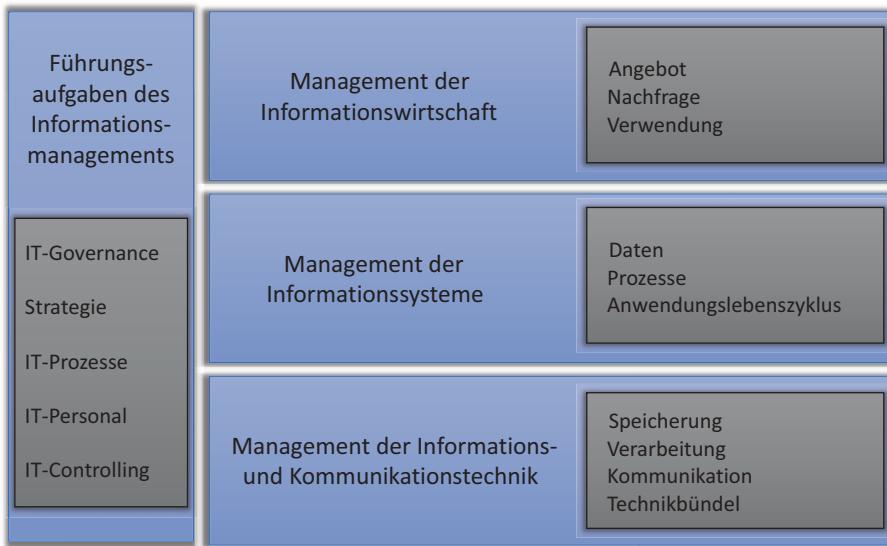


Abb. 3.1 Aufgabenorientiertes Ebenenmodell nach Krcmar (2015)

Das Management der Informationswirtschaft befasst sich mit Informationsnachfrage und -angebot. Es berührt alle Bereiche eines Unternehmens, die Informationen nachfragen und/oder bereitstellen. Es stellt Anforderungen an die betrieblichen Informationssysteme und übernimmt die von diesen zur Verfügung gestellten Unterstützungsleistungen.

Auf der Ebene der Informationssysteme sind betriebliche Anwendungssysteme, beispielsweise ERP- oder Data Warehouse Systeme, von zentraler Bedeutung, die durch das Informationsmanagement zu entwickeln und bereitzustellen sind. Entsprechend sind die dazu erforderlichen Daten und Prozesse zu gestalten und zu managen. Auf dieser Ebene werden Anforderungen an die Ebene der Informations- und Kommunikationstechnik spezifiziert und Unterstützungsleistungen von dieser empfangen.

Auf der Ebene der Informations- und Kommunikationstechnik wird die Technikinfrastruktur bereitgestellt und verwaltet. Sie liefert die physische Basis für die betrieblichen Anwendungssysteme.

Im Rahmen des Informationsmanagements existiert eine Reihe von Führungsaufgaben des Informationsmanagements, die sich nicht eindeutig einer Ebene zuordnen lassen. Dazu zählen

- die Gestaltung der IT-Governance,
- die Erarbeitung und Durchführung der IT-Strategie,
- das Management der IT-Prozesse,
- das Management des IT-Personals und
- das IT-Controlling.

Damit lässt sich Informationsmanagement als das Management der Informationswirtschaft, der Informationssysteme, der Informations- und Kommunikationstechniken sowie der alle diese Elemente betreffenden Führungsaufgaben unter Berücksichtigung der in allen 4 Bereichen anfallenden Gestaltungsaufgaben definieren. Somit erfordert ein umfassendes Informationsmanagement sowohl Managementfähigkeiten als auch informationstechnisches Wissen.

3.1.3 Integriertes Informationsmanagement

Auch das in Zarnekow et al. (2005) ausführlich vorgestellte Rahmenkonzept des integrierten Informationsmanagements versteht Informationsmanagement nicht nur als eine informationstechnologische Aufgabe, sondern auch als eine Führungsaufgabe. In Ergänzung zu den zuvor genannten Aspekten betont es explizit eine Marktorientierung, die sich aus der Beziehung zwischen IT-Leistungsanbieter und IT-Leistungsnachfrager ergibt. Diese bildet das zentrale Element einer Wertschöpfungs- und Lieferkette (Supply Chain), wobei beide Marktseiten unternehmensintern oder -extern positioniert sein können. Die Grundidee des integrierten Informationsmanagements ist in Abb. 3.2 illustriert und wird nachfolgend, basierend auf der ausführlichen Darstellung in (Zarnekow et al. 2005) kurz beschrieben.

Der Source-Prozess des IT-Leistungsnachfragers beinhaltet die zum Einkauf der IT-Leistungen notwendigen Aufgaben. Die eingekauften Leistungen fließen in den Make-Prozess ein, der alle Aufgaben zum Management der IT-Leistungserstellung umfasst. Dazu zählen – sich orientierend an dem klassischen „Plan, Build, Run“-Ansatz – das Management des Produktprogramms, das Management der Produktentwicklung und

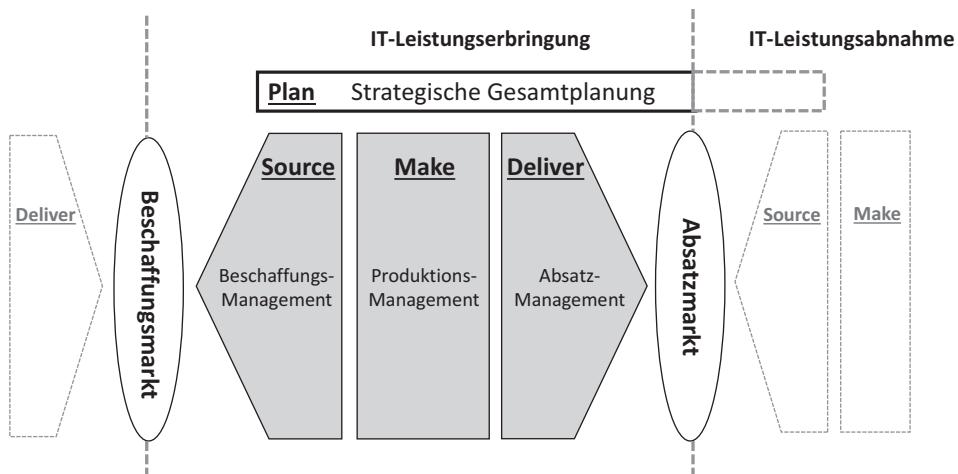


Abb. 3.2 Modell des Integrierten Informationsmanagements nach Zarnekow 2005

das Management der Produktion. Im Deliver-Prozess des IT-Leistungsanbieters sind die Aufgaben zum Management der Kundenbeziehung zusammengefasst. Sofern der Kunde eine Fachabteilung der eigenen Unternehmung ist, kann man von einem internen Markt sprechen, ansonsten liegt ein externer Markt vor.

Mit der beschriebenen Grundbeziehung zwischen IT-Leistungsanbieter und IT-Leistungsnachfrager können in beide Richtungen Lieferketten gebildet werden. So kann ein IT-Leistungsabnehmer seine erworbenen Produkte seinerseits seinen Kunden anbieten. Umgekehrt kann ein IT-Leistungsanbieter über seine Source-Funktion seinerseits IT-Leistungen nachfragen.

Plan- und Enable-Prozesse begleiten die zuvor vorgestellten Kernprozesse. Der Plan-Prozess subsumiert die erforderlichen Führungs- und Governance-Aufgaben. Der Enable-Prozess fasst die Querschnittsaufgaben zusammen, die die Kernprozesse unterstützen. Beispielhaft seien hier das Finanzmanagement, das Personalmanagement oder das Qualitätsmanagement genannt.

Die bereits im aufgabenorientierten Ebenenmodell vorgestellte Idee der Ebenenbildung zur Strukturierung und Konkretisierung der Aufgaben des Informationsmanagements findet man in ähnlicher Form auch im Modell des integrierten Informationsmanagements. So werden die Kernprozesse Source, Make und Deliver jeweils in die Ebenen

- Rahmenbedingungen (strategisch),
- Zielsetzungen (taktisch) und
- Umsetzung (operativ)

untergliedert, wie in Abb. 3.3 illustriert.

3.1.4 Einordnung von Big Data

Legt man das aufgabenorientierte Ebenenmodell zugrunde, so ist erkennbar, dass die Big Data-Thematik alle Ebenen des Informationsmanagements berührt. Auf der Ebene des Managements der Informationswirtschaft gilt es zunächst zu eruieren, welche Daten für Big Data Projekte von den Fachabteilungen benötigt werden und wie diese grundsätzlich angeboten werden können. Auf der Ebene des Managements der Informationssysteme ist zu erarbeiten, wie die betrieblichen Anwendungssysteme diese Daten bereitstellen. Neben traditionellen relationalen Datenstrukturen sind dabei insbesondere auch Non First Normal Form-Datenstrukturen zu diskutieren, die beispielsweise im Kontext von SAP-HANA eine zentrale Rolle spielen. Daran wird auch erkennbar, dass diese Frage auch die techniknahe Ebene des Managements der Informations- und Kommunikationstechnik berührt. Darüber hinaus sind auch ebenenübergreifende Führungsaufgaben wie beispielsweise das IT-Personalmanagement oder das IT-Controlling für das Management von Big Data Projekten relevant.

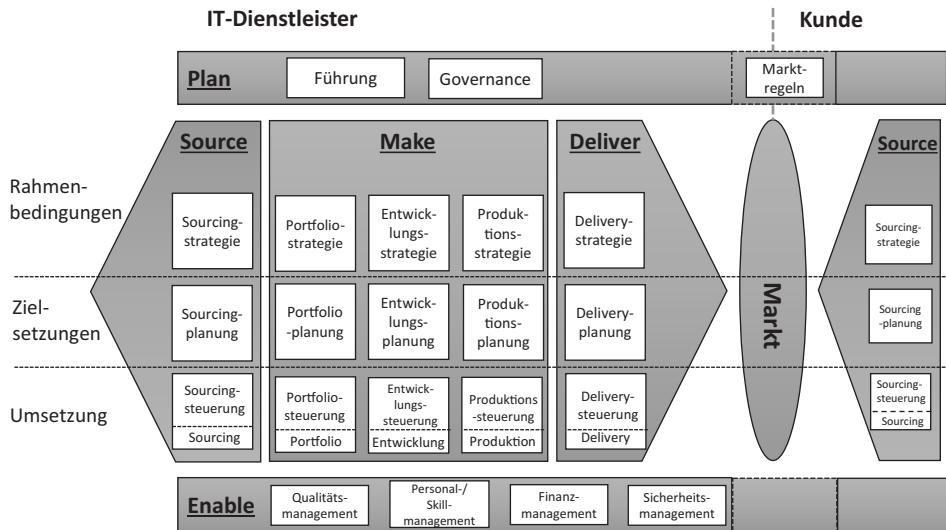


Abb. 3.3 Ebenen des Integrierten Informationsmanagements nach Zarnekov 2005

Auch bei Anwendung des integrierten Informationsmanagements zeigt sich, dass Big Data Projekte alle Prozessarten des Modells tangieren können. Im Deliver-Prozess sind die Anforderungen der Kunden an das Big Data Projekt zu eruieren. Um diese zu bedienen, sind die entsprechenden Big Data-Anwendungssysteme im Make-Prozess zu erstellen bzw. zu betreiben. Ggf. sind die für das Big Data Vorhaben erforderlichen IT-Ressourcen über den Source-Prozess zu beschaffen. Die zu beachtenden unternehmensinternen -und externen Regularien sind im Plan-Prozess abgebildet. Unterstützungsauflagen beispielsweise zum Management der Mitarbeiter oder der Finanzausstattung des Big Data Projekts lassen sich dem Enable-Prozess zuordnen.

3.2 Digitalisierung von Geschäftsmodellen mit Big Data

Der Abschnitt geht auf die Möglichkeiten der Digitalisierung von Geschäftsmodellen mit Hilfe von Big Data- Methoden ein. Er skizziert zunächst den Rahmen im Kontext der IT-Governance und der Digitalisierungsdiskussion, um dann die Brücke von der Digitalstrategie zu Einführungskonzepten zu bilden.

3.2.1 IT-Governance und Digitalisierung

Digitalisierung und Arbeit Die Auswirkungen der Digitalisierung auf breite Teile der Gesellschaft waren unter anderem Gegenstand einer interdisziplinären Fachtagung

der Akademie der Wissenschaften (acatec) (vgl. Spath 2018). Demnach wandeln sich Unternehmen weg von der reinen Profitorientierung hin zu sinnstiftenden Aktivitäten, klassische Hierarchien mit Anweisungen und Autoritätsdenken weichen Netzwerkorganisationen bei denen der Teamgedanke im Vordergrund steht. Langfristige Planung wird ersetzt durch agiles Arbeiten mit der Möglichkeit Experimente zu wagen.

Nicht alle Personen werden im Kontext der Digitalisierung positive Erfahrungen machen. Die von Kornwachs (2018) präsentierte typisierte Verteilung der Verschiebungen der Nachfrage nach Arbeitskräften macht dies sehr deutlich (vgl. Abb. 3.4). Die aufkommenden neuen Berufsbilder (z. B. Data Scientist) erfordern mehr Personen mit höherer Qualifikation. Insgesamt entsteht ein gewaltiger Nachqualifizierungsbedarf, teils auch bei Personen mit akademischer Ausbildung, da viele regelbasierte Tätigkeiten ersetzt werden können.

Insgesamt jedoch scheint es sich für Unternehmen jedoch wirtschaftlich langfristig zu lohnen, in Digitalisierung zu investieren. Eine Studie der Rheinischen Fachhochschule Köln zeigt, dass ein hoher Digitalisierungsgrad den Geschäftserfolg (gemessen am Umsatz) bei den betrachteten kleineren- und mittleren Unternehmen steigert (vgl. Buehler und Steimel 2018).

Allerdings ist bei der Umsetzung der Digitalisierung Eile geboten, denn Auto-Branche hat 62 Jahre benötigt, um die Zahl von 50 Mio. Automobilisten zu erreichen. Die Computerbranche benötigte dafür 14 Jahre, die Internetbranche benötigte nur noch 7 Jahre, dem chinesischen Chat-Dienst WeChat gelang es in 1 Jahr, das Videospiel «Pokémon» benötigte nur noch 19 Tage bis zur nahezu vollständigen Marktdurchdringung (vgl. Rüttig 2018). Satirische Artikel befürchten schon, dass auch der „Mensch zur Ware“ wird (vgl. Wetzel 2020).

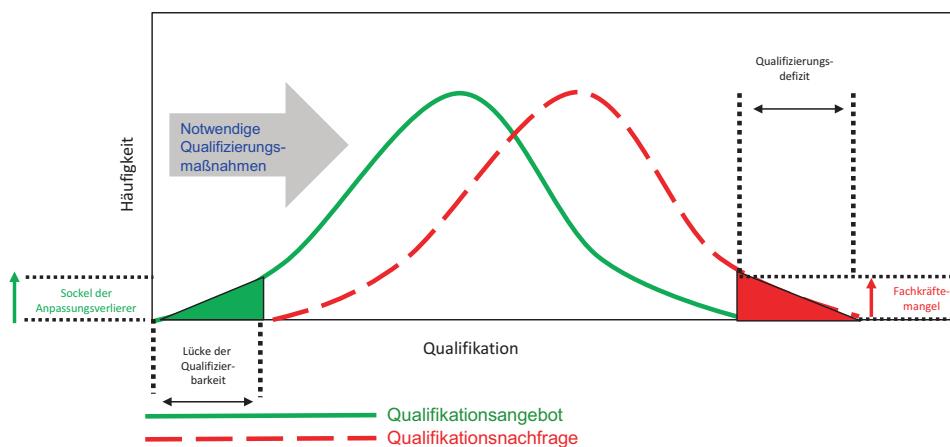


Abb. 3.4 Auswirkungen der Digitalisierung auf die menschliche Arbeit nach Kornwachs (2018)

Digitalisierung und Reifegrad

Die Reifegrade der Unternehmen im Hinblick auf die Umsetzung der Digitalisierung sind noch nicht als sehr hoch einzustufen. Wenn man das Reifegradmodell von Krafft (2018) zugrunde legt, welches aus der Sichtweise eines Praktikers entwickelt wurde, dürften viele Unternehmen aber schon den Reifegrad 3 oder 4 auf einer 6-stufigen Skala erreicht haben (vgl. Abb. 3.5.).

Die zunehmende Digitalisierung verdrängt klassische Berufsbilder in vielen Lebensbereichen. So sind klassische Tätigkeiten von Steuerberatern, Wirtschaftsprüfern, Controllern oder Bankmitarbeitern stark regelbasiert. Diese Tätigkeiten können in Zukunft zumindest teilweise durch Software-Roboter (Bots) ersetzt werden und fallen damit der Digitalisierung zum Teil zum Opfer. Insbesondere für Traditionssunternehmen stellt die Digitalisierung eine große Herausforderung dar, weil sie nicht, wie Startup-Unternehmen auf der „grünen Wiese“ beginnen können, sondern gewachsene Strukturen transformieren müssen.

So werden z. B. Banken, durch ihre „IT-Altlasten“ massiv daran gehindert, innovative digitale Konzepte umzusetzen, die von jungen Startups (z. B. „N26“) bereitgestellt werden. So können die seit langem genutzten, batchorientierten Informationssysteme der großen Bankunternehmen Sofortüberweisungen per App ins Ausland nur mit viel Mühe und Aufwand realisieren, weil ihre Architektur dafür nicht konzipiert worden ist (vgl. Freund 2019). Sogenannte FinTechs oder US-amerikanische Tech-Riesen mit eigenen Bezahlungen, betreffen diese Legacy-Schulden nicht, denn sie setzen auf Ökosystemen auf, die aus eigens entwickelter Software entstanden sind und sich besonders leicht an neue Aufgaben anpassen lassen (vgl. Freund 2019).

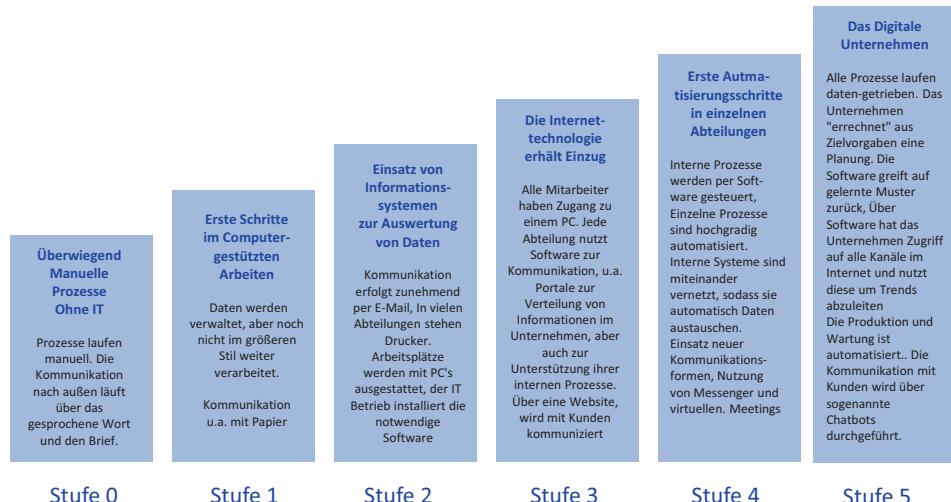


Abb. 3.5 Reifegradmodell zur Digitalisierung nach Krafft (2018)

Rollen im Informationsmanagement

Die Rollen im Informationsmanagement sind seit Jahrzehnten im Wandel (vgl. Abb. 3.6). Zunehmend repräsentieren die Rollen eine weg von Technik hin zum Geschäftszweck wandelnde Bewegung. Die aktuellste Entwicklungsstufe für Führungskräfte im Informationsmanagement (Chief Digital Officer, kurz CDO) verbindet die technische Ebene des klassischen IT-Leiters mit der Ebene der Geschäftsmodelle und damit der Rolle des Business Developments. Nur wenn die Vernetzung dieser Aufgaben gelingt, sind die Unternehmen für zukünftige Aufgaben gerüstet.

Klassisches versus Agiles Informationsmanagement

In der Praxis scheitern viele IT-Projekte, obwohl klassische IT-Projektmanagementmethoden genutzt werden. Die Kernprobleme klassischer Projektmanagementmethoden sind die hohe Unklarheit über das Ziel des Projektes, die zu wählende Vorgehensweise und die zukünftigen Anforderungen zu Beginn des Projektes. Die anfangs erstellten Projektpläne sind daher zwangsläufig ungenau und werden im Projektverlauf schnell von der Realität überholt. Als Lösungsansatz werden daher agile Methoden vorgeschlagen, welche auf detaillierte Projektplanungen weitgehend verzichten. Die Begründung für deren Ansatz ist einfach: Wenn die klassischen Methoden nicht funktionieren, dann sollten sie auch nicht angewendet werden. Stattdessen wird darauf vertraut, dass beim agilen Ansatz ein gemischtes Team mit erfahrenen Mitgliedern Freiheiten erhält, das Problem zu lösen. Viel Transparenz, Freiheit und eine zeitnahe Abstimmung des Teams und eine dezentrale Verantwortung ersetzen eine detaillierte zentrale Planung des Projektes. In Abb. 3.7 sind die zentralen Unterschiede des klassischen und des agilen Informationsmanagements gegenübergestellt.

3.2.2 Von der IT-Strategie zur Business Digitalstrategie

Klassischer Begriff der IT-Strategie

Viele Jahre lang galt die Aussage, dass aus der Geschäftsstrategie eine IT-Strategie abgeleitet wird (vgl. z. B. den Ansatz von Krcmar 2005). Eine Darstellung der Zusammenhänge ist in Abb. 3.8 zu sehen. In der Vergangenheit war dieser Ansatz auch sinnvoll, da andere Projektaufgaben zu bewältigen waren. Aber schon bei den ERP-Einführungsprojekten der 1980er Jahre war vielen Projektmanagern klar, dass die statischen Vorgehensmodelle nicht umsetzbar waren. Zudem wurde das Informationsmanagement in die reine „Umsetzer-Rolle“ gedrängt. Die „IT“ wurde nur als „Enabler“ für das maßgebliche „Business“ bezeichnet. Diese Denkschule ist noch heute vielen Führungskräften im Informationsmanagement verinnerlicht. Mangelnde Rückkopplungen durch das Informationsmanagement verhindern in einem solchen Konzept jedoch innovative IT-getriebene Lösungen. Dementsprechend waren auch die Inhalte der klassischen IT-Strategien sehr technisch geprägt (vgl. die Ergebnisse einer empirischen Studie in Gadatsch et al. 2017 oder die Strategiebeispiele in Hanschke 2009, S. 44–45).

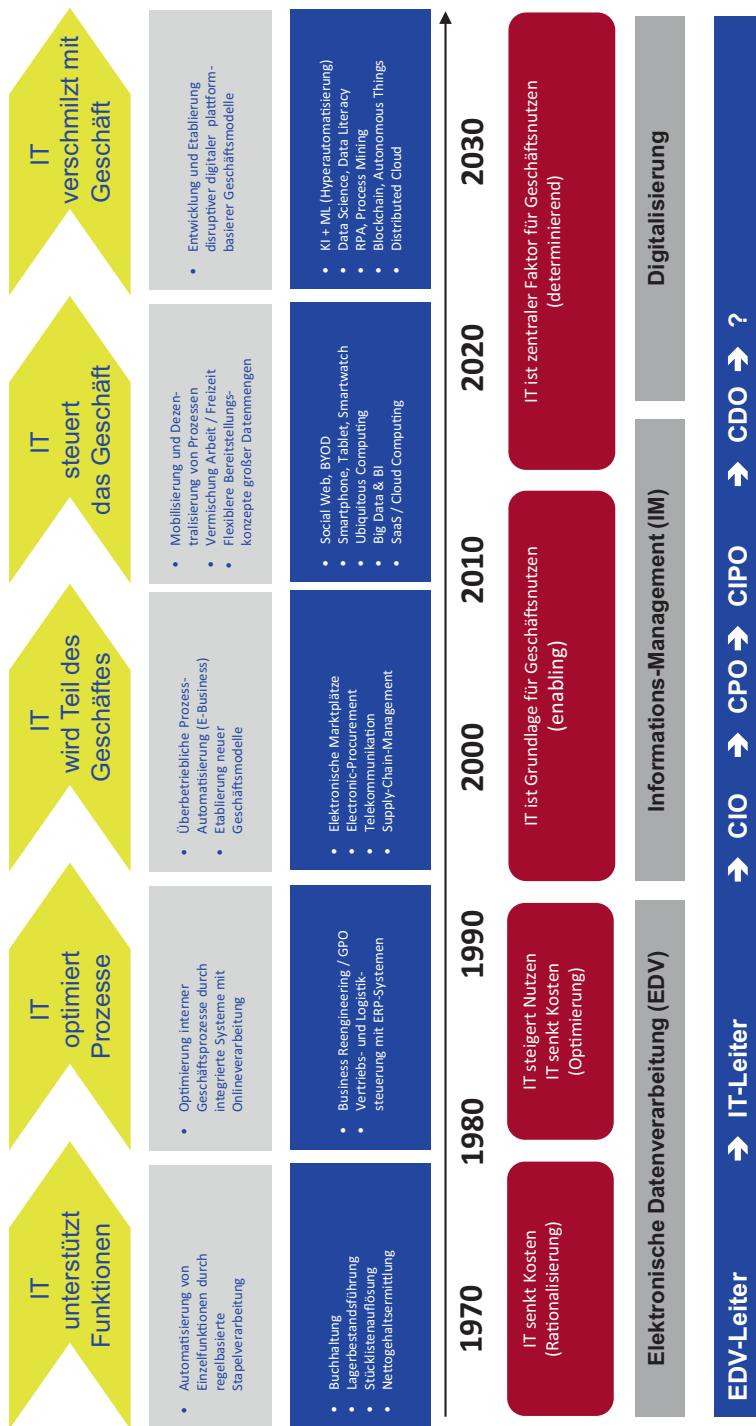


Abb. 3.6 Rollen im Informationsmanagement im Wandel

Traditionelles Informationsmanagement	Agiles Informationsmanagement
Fokussiert auf langfristige, stabile Anforderungen	Fokussiert auf kurzfristige, sich ändernde Anforderungen
Strebt nach Perfektion	Strebt nach Geschwindigkeit
Kümmert sich um Bedarf des aktuellen Geschäfts	Kümmert sich um den Bedarf des zukünftigen Geschäfts
Detaillierte Planung und Umsetzung (Wasserfallmodell)	Prototyping und iterative Entwicklung
Genehmigungsbasierte Governance, wenig Freiraum	Prozessbasierte Governance, mehr Freiraum
Service-Level-Agreements und Kennzahlen zur Steuerung im Regelkreis	Kundenfeedback, schnelle Reaktion für Verbesserung

Abb. 3.7 Klassisches versus Agiles Informationsmanagement

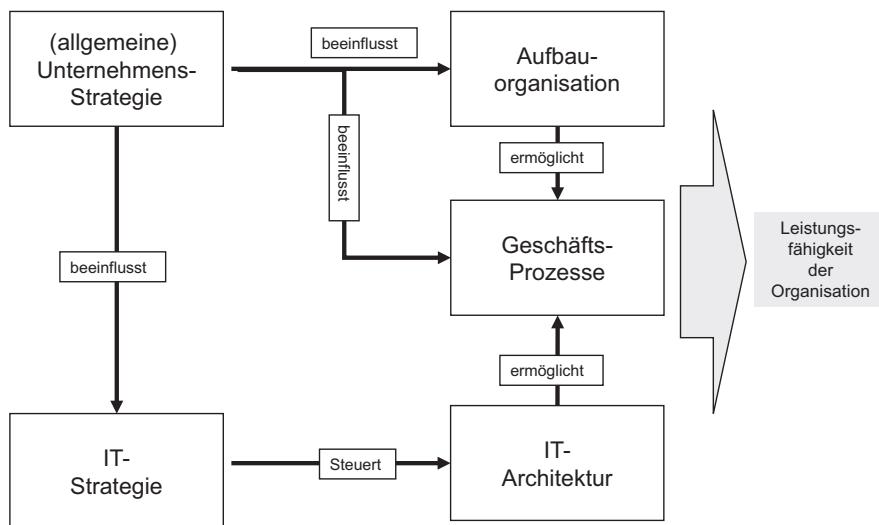


Abb. 3.8 Ableitung der IT-Strategie aus der Unternehmensstrategie

Die Forschungsergebnisse von Bharadwaj et al. (2013) beschreiben ein übergreifend festzustellendes Phänomen: Unternehmen verschmelzen ihre Unternehmens- und IT-Strategie zu einer „Digitalstrategie“ bzw. „Digitalen Geschäftsstrategie“ auf Basis meist neuer oder erweiterter Geschäftsmodelle. Folgt man den Ergebnissen, muss das in Abb. 3.8 dargestellte Schema wie in Abb. 3.9 dargestellt verändert werden.

Ein aktuelles Beispiel für eine Digitalstrategie finden wir in einem Wohnungsbauunternehmen aus Dortmund, der DOGEWO21 (vgl. Freitag 2019). Die Inhalte der

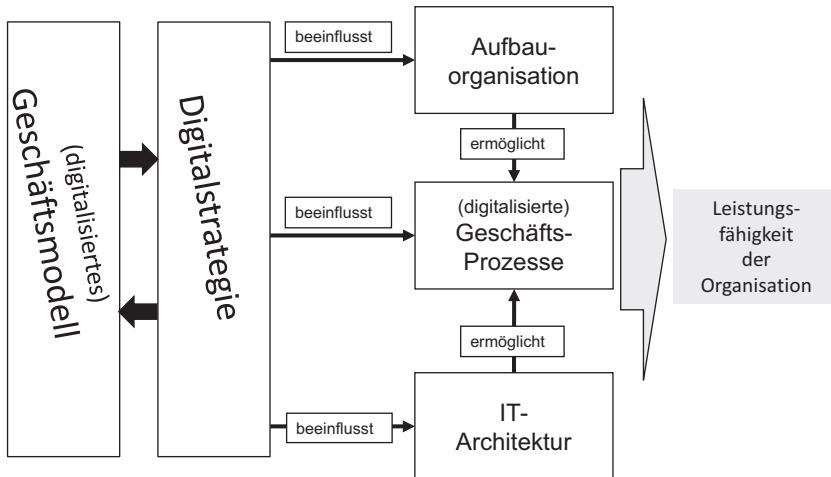


Abb. 3.9 Entwicklung einer Digitalstrategie

Strategie sind z. B.: „Hohe Kundenzufriedenheit erhalten“, „Auf bewährten Kontaktkanälen erreichbar bleiben (Telefon, Mail, Außenbüros in der Stadt u. a.)“, „KEIN Outsourcing von Kernkompetenzen, diese sind als wichtigste Grundlage digital abzubilden“, „Bei der Digitalisierung Kunden und Mitarbeiter mitnehmen“ sowie „Durchdachte und sorgfältige Implementierung (Mitarbeiter)“. Ein weiteres Beispiel hat das Traditionssunternehmen „Villeroy und Boch“ vorgestellt (vgl. Nau 2019).

Digitalstrategie der Villeroy & Boch

- **Data-Science:** Wie können wir Daten nutzen, um unser Geschäft besser zu machen? Welche Daten haben wir vorliegen darüber, was unsere Kunden wollen? Wie können wir z. B. Bilderkennung zur Qualitätssicherung nutzen?
- **Marketing und Sales:** Die Distribution und Kommunikationsstruktur hat sich extrem verändert, heute optimieren wir beispielsweise Produkte für Google oder bieten Services wie den Onlinebadplaner.
- **Prozessdigitalisierung:** Wie können wir mit digitalen Tools besser und schneller werden?
- **Neue Geschäftsmodelle und Innovation:** Wie muss sich unser Geschäftsmodell verändern, und welche neuen digitalen Geschäftsmodelle müssen wir an den Markt bringen?
- **Skalierung:** Wie skaliere ich mein Vorhaben in die Fläche, wie muss also der Rahmen aussehen, um die Menschen mitzunehmen? ◀

3.2.3 Management von Big Data

Einordnung in die Entwicklungsphasen der Wissenschaft

Eine sehr bekannte Veröffentlichung zu den Entwicklungsstufen der Wissenschaft betont die Bedeutung von Big Data (vgl. Hey 2009). Demnach war die erste experimentelle Phase der Wissenschaft mit der Sammlung und Auswertung von Daten beschäftigt. Die nächste Phase der „Theoretischen Wissenschaft“ beschäftigte sich mit der Entwicklung von Gesetzen und Regeln zur Beschreibung der Umwelt, Beweisen und dem Widerlegen von Formeln oder Theoremen. Später folgte die „Berechnende Wissenschaft“, welche computerbasiert die Simulation von Abläufen verfolgte, um daraus Erkenntnisse zu gewinnen über das Verhalten der Welt, z. B. im Bereich der Klimaforschung. Aktuell erleben wir den Übergang in die Datenintensive Wissenschaft (Data Science), d. h. der Nutzung empirischer Daten um daraus neue Erkenntnisse zu gewinnen, dem Aufbau von sehr komplexen Modellen (z. B. der Vorhersage von Klimaveränderungen oder der Vorhersage von Krankheitsverläufen).

Beispiel Textilfabrik

Ein typisches Beispiel aus der Industrie ist die Überwachung von Maschinen in einer Textilfabrik durch das kontinuierliche Sammeln und Analysieren von Daten in Echtzeit. Die Anbindung unterschiedlicher Datenquellen wie Sensoren, Maschinen oder Systeme ermöglicht es, die Anomalien in gesammelten Daten zu identifizieren und dadurch eine Vorhersage von zukünftigen Störungen oder Problemen zu ermöglichen. Entscheider können somit agieren, bevor ein Problem eintritt und müssen nicht kurzfristig reagieren, wenn das Problem schon existiert und ggf. schnelle, wenig optimale Lösungen finden. ◀

Beispiel Polizei

Auch Polizeibehörden nutzen datengetriebene Technologien. So hat das Unternehmen Clearview ein Geschäftsmodell entwickelt, bei dem öffentlich zugängliche Namen und Gesichter verknüpft und in einer Datenbank gespeichert werden. Die Informationen werden allerdings nur an Ermittler verkauft, welche das Bild einer gesuchten Person hochladen und es mit einer Datenbank abgleichen können (vgl. Beuth und Horchert 2020). ◀

Situation in der Praxis

Eine Studie des Institutes der Deutschen Wirtschaft (2019) hat ergeben, dass etwa 83 % der Unternehmen, die Big Data einsetzen, dies mit dem Ziel tun, ihre Prozesse zu optimieren. Die Durchdringung in der Praxis wird noch uneinheitlich gesehen. Im Mittelstand gibt es ein eher pessimistisches Bild, wie die folgende Aussage belegt: „... Technisch ist das Thema Big Data längst im Mittelstand angekommen. Allein, es findet

dort bislang nur wenig Widerhall...“ (vgl. Mittelstandswiki 2018). Dies verwundert, da die Potenziale für Big Data bereits seit vielen Jahren bekannt sind und diskutiert werden (vgl. Seufert 2014). Die Aussage „Große Unternehmen setzen Big Data öfter ein als KMU“ wurde in einem führenden IT-Management-Magazin veröffentlicht und dürfte kaum verwundern (vgl. Vaske 2015).

Big-Data als Katalysator für Geschäftsmodelle und -prozesse

Konzepte aus dem Konzept Big Data lassen sich in allen Unternehmens- und Gesellschaftsbereichen nutzen. Der Branchenverband Bitkom hat schon 2013 einen Portfolioansatz veröffentlicht, der zeigt, welche Dimensionen grundsätzlich denkbar sind (vgl. Abb. 3.10). Demnach können die Maßnahmen auf zwei Ebenen eingeordnet werden: Datenverwendung (vorhandene oder neue Daten) sowie das Geschäftsmodell (vorhandenes oder neues Geschäftsmodell).

3.2.4 Vorgehensmodelle zur Einführung von Big Data

In den ersten Jahren des „Big Data Hypes“ hatten viele Unternehmen keine konkreten Vorstellungen, wie Big Data konkret eingesetzt werden kann und wie die Einführung methodisch fundiert zu realisieren ist (vgl. Lixenfeld 2015). Als mögliche Gründe werden genannt: fehlende (Big Data)-Strategie, Unkenntnis über fachliche und technische Möglichkeiten von Big Data Technologien, fehlende Erfahrungen sowie fehlende Big Data spezifische Vorgehensmodelle.

Neues Business	Monetarisierung	Durchbruch
	<ul style="list-style-type: none"> Häufig lassen sich mit existierenden Daten neue Geschäftsmodelle kreieren, sofern die Nutzung der Daten rechtlich zulässig ist Beispiel: Anonymisierte Auswertung der Nutzer- und Standortdaten von Telefonnutzern zur Optimierung von lokализierten Diensten und von ortsbasiiger Werbung 	<ul style="list-style-type: none"> Königsklasse bei der Entwicklung neuer Geschäftsmodelle. Schaffung neuer Produkte oder Geschäftsmodelle <u>mit neuen Daten</u> Beispiel: Ortsbezogene Leistungsprognosen für Betreiber von Solar- und Windparks der Startup-Firma Enercast, digitale Kartographie von Städten durch Google Streetview
Vorhandenes Business	Optimierung	Aufwertung
	<ul style="list-style-type: none"> Einstieg in das Big-Data-Business. Bessere Nutzung von unternehmenseigenen Datenbeständen Beispiel: Rückschlüsse aus Kauf- und Online-Verhaltens der Kunden ziehen Vorreiter sind Anbieter von Billig-Flügen, die ihre Gewinn-Management-Systeme mit einer Vielzahl Parameter, z. B. aus dem Online-Verhalten, kombiniert und optimiert haben. 	<ul style="list-style-type: none"> Bestehende Geschäftsmodelle und Dienstleistungen lassen sich auch durch neue Daten aufwerten. Beispiel: Integration von Wetterprognosen in Marketingaktivitäten von Reiseunternehmen, Verkehrsmanagement in Metropolen über Mautsysteme, die den Verkehrsfluss über Preisumpassungen steuern

Vorhandene Daten

Neue Daten

Abb. 3.10 Bitkom-Portfolio für Big Data (2013)

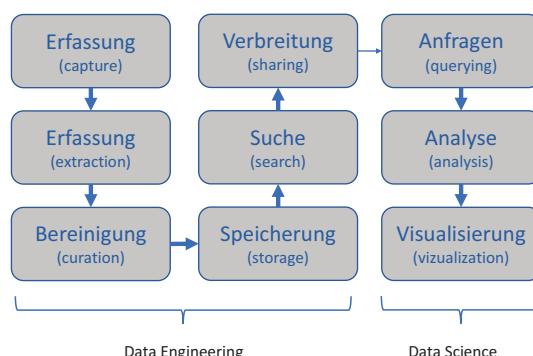
Durch die Vernetzung von Geschäftsmodellen, Geschäftsprozessen und Informations-technologie besteht für Big Data-Projekte der Bedarf, die bislang verwendeten Vorgehensmodelle anzupassen und die spezifischen Anforderungen dort einzuarbeiten. Bislang wurden nur vergleichsweise wenige Vorgehensmodelle für Big Data Projekte veröffentlicht. Sie werden nachfolgend kurz skizziert.

Felix Naumann „Data Science Pipeline“.

Das von Naumann (2020) vorgestellte und allgemein anerkannte Modell der „Data Science Pipeline“ (vgl. Abb. 3.11) ist sehr eng an den Prozess des Data Engineering (Infrastruktur für die Bereitstellung der Daten) und Data Science (Analyse der Daten) angeknüpft und setzt einen vorgelagerten Prozess der Strategie- und Geschäftsmodellentwicklung voraus. Es umfasst insgesamt zahlreiche Einzelschritte, die von der Erfassung bis zur Visualisierung der Daten reichen:

- **Erfassung (capture):** Auslesen von Sensoren und Erfassung von Nutzerdaten.
- **Extraktion (extraction):** Selektion relevanter Daten mit Fachexpertise.
- **Bereinigung (curation):** Korrektur fehlender, unvollständiger oder inkonsistenter Daten, z. B. können Nullwerte ausgelassen, ersetzt oder ergänzt werden.
- **Verbreitung (sharing):** Klärung zahlreicher sozialer und organisatorische Aspekte, wie die Daten zu verteilen sind, so ist z. B. u. U. eine Erlaubnis einzuholen (Gesetze, Betriebsrat) und Widerstände bei Mitarbeitern zu berücksichtigen.
- **Suche (search):** Sicherstellung der Nachvollziehbarkeit der verwendeten Daten.
- **Speichern (storage):** Klärung des Speicherungsverfahrens, z. B. Festplatte, Cloud.
- **Anfragen (querying):** Klärung der Fragen der Datenrelevanz, z. B. muss festgelegt werden, welche Daten Spalten und Zeilen zu verwenden sind, einzelne Daten müssen ggf. angereichert werden, z. B. die Postleitzahl mit dem Ortsnamen.
- **Analyse (analysis):** Hier geht es um den Kernaspekt von Data Science, der Anwendung verschiedener Methoden wie z. B. Machine Learning, Clustering, Classification um konkrete Ergebnisse zu erzielen.
- **Visualisierung (visualization):** Interpretation der Ergebnisse mit geeigneten Verfahren.

Abb. 3.11 Data Science Pipeline (Naumann 2020)



CSC: How To Do a Big Data Project – A Template for Success

Der Fokus des vom IT-Dienstleister CSC herausgegebenen Vorgehensmodells liegt auf der technischen Umsetzung bei einem vorgegebenen Use Case (vgl. CSC 2015). Die wesentlichen Elemente sind in Abb. 3.12 dargestellt. Es ist ein klassisches vierstufiges Phasenmodell, welches als erste Phase die Definition des „Use Cases“ enthält, welcher in den nachgelagerten Stufen mit Methoden des Big Data Umfeldes umgesetzt wird.

BITKOM Vorgehensmodell (Buschbacher et al. 2014).

Ein stark auf die Belange von Big Data Projekten zugeschnittenes Vorgehensmodell wurde schon sehr früh von den Mitgliedern des BITKOM-Branchenverbandes veröffentlicht, welche sich in einem „Big Data Arbeitskreis“ mit den seinerzeit noch neuen Herausforderungen beschäftigten (vgl. z. B. Buschbacher et al. 2014). Es ist in sieben Projektphasen zur Einführung von Big Data und eine achte Optimierungsphase zur ständigen Verbesserung der neuen Daten- und Systemlandschaft gegliedert (vgl. Abb. 3.13).

DASC-PM v1.0: Ein Vorgehensmodell für Data-Science-Projekte

Kürzlich wurde ein frei verfügbares Vorgehensmodell (Creative Commons Licence) veröffentlicht, das praktisch und wissenschaftlich durch Mitautoren aus dem Data Science Umfeld validiert wurde. Es beinhaltet eine Strukturierung der Aufgaben eines Data Science Projektes nach zahlreichen Schlüsselbereichen und geht damit einen anderen Weg als die klassischen Vorgehensmodelle. Die Schlüsselbereiche sind Daten, Analyseverfahren, Nutzbarmachung, Nutzung, Domäne, Wissenschaftliches Vorgehen sowie IT-Infrastruktur (vgl. Schulz und Neuhaus 2020). Der kompakte Life-Cycle des Modells

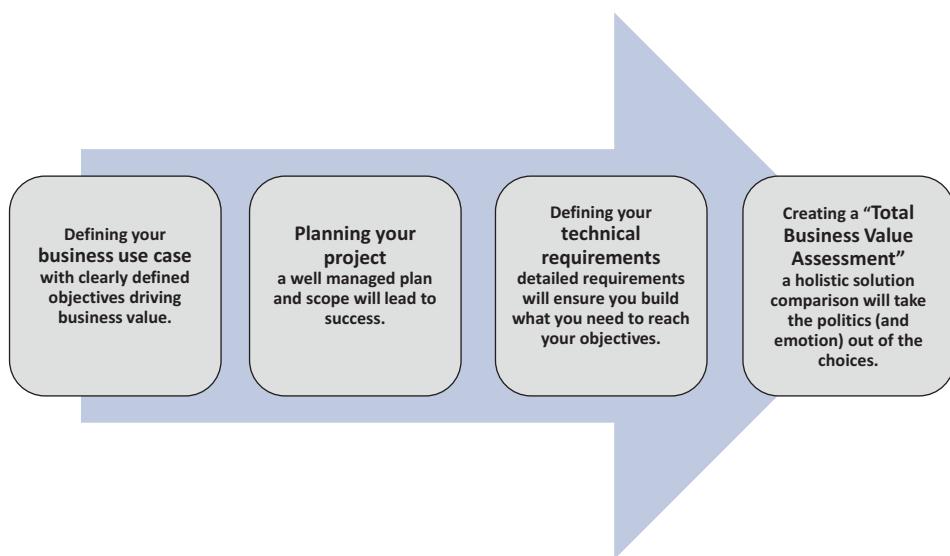


Abb. 3.12 CSC Vorgehensmodell für Big Data (CSC 2015)



Abb. 3.13 BITKOM Vorgehensmodell (Buschbacher et al. 2014)

ist in Abb. 3.14 dargestellt. Die dort dargestellten durchgezogenen Pfeile zeigen den primären Pfad bei der Verwendung des Vorgehensmodells, die gestrichelten Pfeile bilden Rückkopplungen zu vorherigen Phasen ab, die im Projektverlauf auftreten können (vgl. Schulz und Neuhaus 2020, S. 24).

Vorgehensmodell zur Big Data Einführung – Modell Austria

In Österreich wurde ein Vorgehensmodell für die Big Data-Einführung konzipiert, welches auch ein hierauf aufbauendes Reifegradmodell enthält (vgl. Huber-Meir und Köhler 2019). Das in Abb. 3.15 abgebildete Vorgehensmodell besteht aus acht Phasen, deren Ergebnisse ineinander einfließen, und deren Anwendung von dem gewählten Projektmanagementmodell abhängig ist und angepasst werden kann. Die Inhalte des Modells basieren auf umfangreichen Projektanalysen verschiedener untersuchter Leitprojekte aus dem Umfeld Mobility, Handel, Katastrophenmanagement und Weltraum.

3.2.5 Messung des Reifegrades von Organisationen

Die Reifegradmessung speziell im Hinblick auf Big Data wurde bislang noch wenig in der Literatur behandelt. Ein kürzlich veröffentlichtes Reifegradmodell stammt aus dem bereits erwähnten Projekt aus Österreich (vgl. Huber-Meir und Köhler 2019). Es stellt den Reifegrad in sechs Ebenen (von null bis fünf) dar:

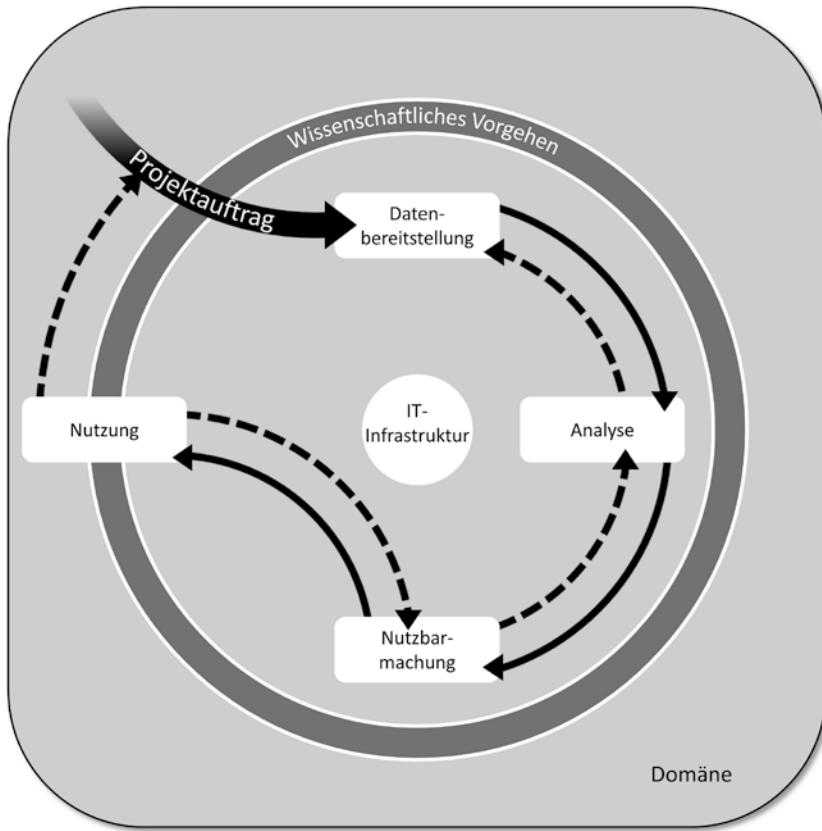


Abb. 3.14 DASC-PM v1.0 – Ein Vorgehensmodell für Data-Science-Projekte (Schulz und Neuhaus 2020)

- Level 5: Nachhaltiges Big Data Business,
- Level 4: Gesteuerte Big Data Projekte,
- Level 3: Prozess für Big Data Projekte,
- Level 2: Evaluierung von Big Data Technologien,
- Level 1: Big Data Kompetenzerwerb,
- Level 0: Keine Big Data Projekte.

Beim Level 0 hat die Organisation noch keine Aktivitäten im Big Data Bereich gestartet, es wurden noch keine Erfahrungen gesammelt und nur klassische Technologien für die Datenverwaltung und Analyse eingesetzt.

Beim Level 1 hat die Organisation erste Initiativen zum Erwerb von Big Data spezifischen Kompetenzen gestartet. Diese Phase ist oft durch hektisches und unkoordiniertes Vorgehen gekennzeichnet. Oft scheitern erste Versuchsprojekte.



Abb. 3.15 Vorgehensmodell für Big Data – Modell Austria

Beim Level 2 wurden erste spezifische Kompetenzen erworben und mögliche Einsatzbereiche der Technologien definiert. Es werden erste isolierte Versuchsbälle gestartet, wo der technologische Kompetenzerwerb erleichtert werden sollte.

Beim Level 3 sind einige initiale Projekte vor dem Abschluss und erste Evaluierungsergebnisse liegen vor. Die Big Data Verantwortlichkeiten sind definiert und strategische Maßnahmen für die Umsetzung von weiteren Projekten und deren Integration in Geschäftsprozesse sind gesetzt. Big Data hat sich in der Geschäftsstrategie etabliert und findet auch beim Top-Management die notwendige Unterstützung.

Beim Level 4 ist Big Data ein fester Bestandteil der Organisationstrategie, es existiert ein Prozessmanagement für Big Data relevante Projekte. Die Strategie wird ausgebaut.

Beim Level 5 ist Big Data ein zentraler Bestandteil der Strategie und wird zur Optimierung von Geschäftsprozessen eingesetzt. Die IT-Infrastruktur setzt Big Data Technologien ein und bietet diese der Organisation für die Umsetzung neuer Projekte an.

3.2.6 Auswirkungen von Big Data auf die Organisation

Big Data wird in erster Linie mit Technik in Verbindung gebracht, anstatt mit neuen Geschäftsideen und Produkten (vgl. z. B. Computerwoche 2019). Eine schon etwas zurückliegende Literaturanalyse wissenschaftlicher Beiträge mit Bezug zu Big-Data in Journals, die einem Peer Review Prozess unterliegen (MIS Quarterly, WIRTSCHAFTSINFORMATIK u. a.), hat ergeben, dass es eine hohe Aufmerksamkeit für Big-Data in Praxis und Theorie gibt. Deren Anwendung ist in vielen Disziplinen denkbar (Klima,

Finanzen, Medizin, Verhaltensforschung, Astronomie, Business Intelligence). Allerdings wird das Thema technisch getrieben, untersucht wird vor allem die Bereitstellung von Daten. Die fachliche Nutzung der Daten ist noch unklar (vgl. Pospiech und Felden 2013, S. 7).

Aktuell wird in Stellenanzeigen sehr häufig nach Data Scientisten gesucht, welche sich mit den Aufgaben im Kontext von Big Data beschäftigen sollen. Deren Berufsbild ist von folgenden Fähigkeiten geprägt (vgl. Wrobel et al. 2015, S. 370–377):

- Verständnis von Unternehmens-Zielen und ihrer Verbindung zu Analytics,
- Solides Grundverständnis datengetriebener Modellbildung mit analytischen Methoden,
- Fähigkeit zur Identifikation und Verknüpfung von Datenquellen,
- Beherrschung der notwendigen Algorithmen und Werkzeuge für Analyse und Verknüpfung,
- Engineering-Wissen über Realisierbarkeit, Skalierbarkeit und Kosten,
- Fähigkeiten zur Übernahme von Verantwortung, Leadership, Networking,
- Urteilsfähigkeit bezüglich Werten und Normen und
Kommunikatives Talent zur Übersetzung von Ergebnissen in die Business-Welt.

Die hohen Anforderungen stehen den realen Gegebenheiten oft entgegen. So sind nach einer US-Untersuchungen viele Data Scientisten zu 60 % ihrer Arbeitszeit mit dem Bereinigen von Daten beschäftigt und können nur 4 % für die Entwicklung von Algorithmen aufwenden (vgl. Forbes 2016).

Literatur

- Augustin, S.: Information als Wettbewerbsfaktor: Informationslogistik – Herausforderung an das Management. Verlag TÜV Rheinland, Köln (1990)
- Beuth P., Horchert, J.: Clearview kennt dich, Der Spiegel, 24.01.2020, 18.00 Uhr, <https://www.spiegel.de/netzwelt/gesichtserkennungs-app-clearview-kennt-dich-a-00000000-0002-0001-0000-000169122938> (2020). Zugegriffen: 27. Febr. 2020
- BITKOM (Hrsg.): Leitfaden Management von Big-Data Projekten, Berlin (2013)
- Bharadwaj, A., Sawy, E., Omar, A., Pavlou, P., Venkatraman, N.: AI business strategy: Toward a next generation of insights. MIS Quarterly **37**(2), 471–482 (2013)
- Buehler, K., Steimel, B.: Digitale Dividende im Mittelstand. Management Summary, Köln (2018)
- Buschbacher, F., Konrad, R., Mußmann, B., Weber, J.: Big Data-Projekte: Vorgehen, Erfolgsfaktoren und Risiken. In: Gleich, R., Klein, A. (Hrsg.). Der Controlling-Berater Bd. 35, (2014)
- Computerwoche (Hrsg.): Technologie, Big-Data, <https://www.computerwoche.de/k/big-data,3457> (2019). Zugegriffen: 21. Aug. 2019
- CSC (Hrsg.): How To Do a Big Data Project: A Template for Success, Whitepaper, <http://www.infochimps.com/resources/how-to-do-a-big-data-project-a-template-for-success/> (2015). Zugegriffen: 02. Sept. 2015

- Forbes (Hrsg.): Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task”, Gil Press, Forbes, March 23rd, 2016, <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-surveysays/> (2016). Zugegriffen: 26. Febr. 2020
- Freund, J.: Museums-IT“ bremst Digitalisierung aus <https://www.geldinstitute.de/it-itk/2019/09/kernprozesse---warum-sich-die-heutige--bank-it-nicht-retten-laes.html> (2019). Zugegriffen: 03. Sept. 2019
- Freitag, A.: Auf Knopfdruck im Bilde, Systemübergreifender Zugriff auf alle Informationen zum Mietobjekt, Essen, 02.07.2019, Vortragsunterlagen (2019)
- Gadatsch, A.: Grundkurs Geschäftsprozessmanagement, 9. Aufl. Springer, Wiesbaden (2020)
- Gadatsch, A., Kütz, J., Freitag, S. : Ergebnisse der 5. Umfrage zum Stand des IT-Controlling im deutschsprachigen Raum. In: Schriftenreihe des Fachbereiches Wirtschaft Sankt Augustin, Hochschule Bonn-Rhein-Sieg, Bd. 34, Sankt Augustin (2017)
- Hanschke, I.: Strategisches Management der IT-Landschaft. Carl Hanser, München (2009)
- Heinrich, L.J., Lehner, F.: Informationsmanagement: Planung, Überwachung und Steuerung der Informationsinfrastruktur, 8. Aufl. Oldenbourg, München (2005)
- Hey, T.: The Fourth Paradigm: Data-intensive Scientific Discovery, Microsoft Research (2009)
- Huber-Meir, M., Köhler, M.: Best Practice für Big Data Projekte, Leitfaden für #Big Data aus #Austria, https://www.ffg.at/sites/default/files/allgemeine_downloads/thematische%20programme/IKT/big_data_in_austria-leitfaden.pdf (2019). Zugegriffen: 30. Oct. 2019
- Institut der Deutschen Wirtschaft (Hrsg.): Big Data: Neuland für viele Unternehmen, <https://www.iwd.de/artikel/big-data-neuland-fuer-viele-unternehmen-438143/> (2019). Zugegriffen: 06. Aug. 2019
- Kornwachs, K.: Digitalisierung – Revolution oder Gestaltungsauftrag?, Dialogreihe „Innovation und Verantwortung“, Tutzing, 12. bis 13. November 2018, Digitalisierung und Arbeitswelt, Vortragsunterlagen. (2018)
- Krafft, K.: 5 Stufen der Digitalisierung, wie das digitale Unternehmen arbeitet, CIO Magazin, 16.01.2018 <https://www.cio.de/a/wie-das-vollstaendig-digitale-unternehmen-arbeitet,3574107> (2018). Zugegriffen: 04. Nov. 2019
- Krcmar, H.: Informationsmanagement, 6. Aufl. Springer Gabler, Berlin (2015)
- Lixenfeld, C. (CIO-Magazin): Ideen fürs Business fehlen Sinnloser Big-Data-Aktionismus, CIO-Magazin, 31.08.2015 https://www.cio.de/a/sinnloser-big-data-aktionismus,3245804?tap=1d3ab1058233ec6b59f18b263745438f&r=564614537252300&lid=445720&pm_ln=14 (2015). Zugegriffen: 02. Sept. 2015
- Mittelstandswiki: Big Data im Mittelstand, https://www.mittelstandswiki.de/wissen/Big_Data_im_Mittelstand (2018). Zugegriffen: 19. Oct. 2018 – 23. Oct. 2018
- Nau, E.: Villeroy & Boch, Keramik Wetterfest machen, 13.11.2019, https://www.cdo-insight.com/organisation/villeroy-boch-keramik-wetterfest-machen-2780/?utm_source=+CleverReach+GmbH+%26+Co.+KG&utm_medium=email&utm_campaign=13-11-2019+CDO+Insight+13.11.2019&utm_content=Mailing_13496095 (2019)
- Naumann, F.: Kurs „Data Engineering und Data Science – Klarheit in den Schlagwort-Dschungel“ (2020). Zugegriffen: 16. Jan. 2020
- Pospiech, M., Felden, C.: Stand der wissenschaftlichen Betrachtung: Zu viele Daten, zu wenig Wissen. BI-Spektrum 1, 7–13 (2013)
- Rütti, N.: Der Agile Mitarbeiter im Digitalen Strudel, NZZ, 06.12.18 <https://www.nzz.ch/wirtschaft/der-agile-mitarbeiter-im-digitalen-strudel-ld.1442307> (2018). Zugegriffen: 29. Juli 2019
- Schulz, M., Neuhaus, U. (Hrsg.): DASC-PM v1.0, Ein Vorgehensmodell für Data-Science-Projekte, Hamburg und Elmshorn, <https://www.nordakademie.de/forschung/data-science-process-model> (2020). Zugegriffen: 25.02.2020

- Seufert, A.: Entwicklungsstand, Potentiale und zukünftige Herausforderungen von Big Data – Ergebnisse einer empirischen Studie. *HMD* **51**(1), 412–423 (2014). <https://doi.org/10.1365/s40702-014-0039-7>
- Sieber, R.: Die IT der Zukunft hat zwei Geschwindigkeiten: agil und stabil. Von: <https://different-thinking.de/die-der-zukunft-hat-zwei-geschwindigkeiten-agil-und-stabil/> (2016). Zugriffen: 10. Apr. 2018 (modifiziert)
- Spath, D.: Arbeit in der digitalen Transformation:, Dialogreihe „Innovation und Verantwortung“, Tutzing, 12. bis 13. November 2018, Digitalisierung und Arbeitswelt, Vortragsunterlagen (2018)
- Vaske, H.: Digitale Transformation lässt keinen Stein auf dem anderen, CIO-Magazin, 06.11.2015. <https://www.cio.de/a/digitale-transformation-laesst-keinen-stein-auf-dem-anderen,3231493> (2015). Zugriffen: 20. Febr. 2019
- Wetzel, M.: Haben Sie schon Amazon Prime für Ihr Liebesleben? <https://blog.zeit.de/teilen/2020/02/05/amazon-dating-onlinedating-plattform-morgan-gruer/> (2020). Zugriffen: 05. Febr. 2020
- Wollnik, M.: Ein Referenzmodell des Informationsmanagements. *Information Management* **3**(3), 34–43 (1988)
- Wrobel, S., Voss, H., Köhler, J., Beyer, U., Auer, S.: Big Data. Big Opportunities, in: Informatik Spektrum **38**(5), 370–378 (2015)
- Zarnekow, R., Brenner, W., Pilgram, U.: Integriertes Informationsmanagement Strategien und Lösungen für das Management von IT-Dienstleistungen. Springer, Berlin (2005)



Digital Leadership

4

Wilhelm Mülder

Zusammenfassung

Die digitale Transformation verändert die Art und Weise, wie wir arbeiten, erfordert andere Qualifikationen seitens der Arbeitnehmer und verlangt neue Formen der Führung. Die Führungskräfte von morgen müssen bereit sein, ihren Führungsstil zu verändern und digitale Tools im Rahmen virtueller Führung anzuwenden. Für die erfolgreiche Verankerung von Digital Leadership im Unternehmen existieren verschiedene Konzepte und Methoden. Hierbei wird die Teamarbeit gestärkt. Ein Digital Leader muss Macht und Kontrollbefugnisse abgeben und agiert überwiegend als Coach und Motivator für sein Team.

4.1 Führung im Digitalzeitalter

Die Digitalisierung sorgt für enorme Veränderungen in Wirtschaft und Gesellschaft. Unternehmen entwickeln und vermarkten neue digitale Geschäftsmodelle, die Kunden erwarten im Zuge der Digitalisierung besseren Service und unmittelbare Bedürfnisbefriedigung, im Privatbereich ist ein Leben ohne Internet und Smartphone für die meisten Menschen überhaupt nicht mehr vorstellbar. Die digitale Transformation verändert komplett unsere Arbeitswelt: die Art und Weise, wie wir zusammenarbeiten, wird sich massiv wandeln, es werden andere Qualifikationen seitens der Beschäftigten benötigt und letztlich müssen die Mitarbeiter anders als bisher geführt werden. Wenn digitaler Wandel gelingen soll, dann muss der Wille zur Veränderung bei allen Akteuren,

W. Mülder (✉)

Institut GEMIT, Hochschule Niederrhein, Mönchengladbach, Deutschland

E-Mail: muelder@hs-niederrhein.de

insbesondere aber bei den verantwortlichen Managern existieren. Führung in Zeiten der digitalen Transformation wird als „**Digital Leadership**“ bezeichnet (vgl. Krug et al. 2018, S. 49). Im Vordergrund steht die Lenkung und Entwicklung eines bislang hauptsächlich analog agierenden Unternehmens in Richtung digitales Unternehmen. Die Mitarbeiter müssen bei diesem Transformationsprozess „mitgenommen“, d. h. motiviert und für neue Aufgaben qualifiziert werden. Teilweise werden junge, gut ausgebildete Fachkräfte neu eingestellt, die einen anderen Führungsstil erwarten als die angestammte Belegschaft. Letztlich bedeutet Digital Leadership auch, dass Mitarbeiter mithilfe digitaler Tools geführt werden, dies setzt ein „digitales Mindset“ beim Management voraus. Führungskräfte, die ihre Mitarbeiter im Digitalzeitalter und mittels digitaler Medien führen, werden als „**Digital Leader**“ bezeichnet. Als **Digitale Unternehmen** bezeichnen wir diejenigen, die einen digitalen Transformationsprozess zumindest teilweise schon umgesetzt oder sich von vornherein auf digitale Geschäftsmodelle konzentriert haben.

Neben Digital Leadership existieren weitere Begriffe, die sich mit neuen Führungsmethoden beschäftigen. Bei „**Agiler Führung**“ steht die Förderung der Selbstorganisation von Mitarbeitern und agilen Teams im Vordergrund. „Agil“ bedeutet hierbei die Fähigkeit eines Individuums, flexibel auf Anforderungen zu reagieren und sich der Umwelt anzupassen. „**Generationenorientierte Führung**“ berücksichtigt die unterschiedliche digitale Reife einzelner Arbeitnehmergenerationen und geht insbesondere auf veränderte Erwartungen und Motive jüngerer Fachkräfte (Generation Y und Z) ein. Bei „**Virtueller Führung**“ werden Mitarbeiter an unterschiedlichen Orten und teilweise zu unterschiedlichen Zeiten über digitale Kanäle geführt (vgl. Lindner und Greff 2019, S. 632 ff.).

4.2 New Work

Die digitale Transformation verändert die Arbeitswelt in starkem Maße. Einerseits wird betont, dass Mitarbeiter mit ihren (digitalen) Kompetenzen, aber auch mit ihrer Kreativität und Flexibilität den wichtigsten Anteil an einer erfolgreichen Digitalisierungsstrategie haben (vgl. Kofler 2018, S. 7). Andererseits gibt es Befürchtungen, dass ein Großteil der menschlichen Arbeit zukünftig durch Maschinen oder Algorithmen ersetzt wird. Im Rahmen von Digital Leadership muss sich jede Führungskraft rechtzeitig und intensiv mit diesen – für die betroffenen Mitarbeiter existenziellen – Fragen auseinandersetzen. Für die zukünftige Neugestaltung der Arbeitswelt werden die Begriffe „Arbeit 4.0“ sowie „New Work“ verwendet. **Arbeit 4.0** umschreibt neuartige Tätigkeiten und Beschäftigungsverhältnisse, die im Zuge der vierten industriellen Revolution (Industrie 4.0) entstehen. **New Work** basiert auf dem Konzept des Sozialphilosophen Frithjof Bergmann mit den zentralen Merkmalen Selbstständigkeit, Handlungsfreiheit sowie Teilhabe an der Gemeinschaft und steht heute für grundlegende Änderungen der Erwerbsarbeit als Folge der Digitalisierung (vgl. Bergmann 2005; Fischer et al. 2018, S. 90).

4.2.1 Mobile Arbeitsplätze

In der Arbeitswelt 4.0 wird die physische Anwesenheit einer Person an ihrem angestammten Arbeitsplatz im Büro oder in der Fabrik immer seltener erforderlich sein. Möglich wird dies durch die Digitalisierung sämtlicher Arbeitsdokumente und die vom jeweiligen Nutzerstandort unabhängigen Zugriffs- und Interaktionsmöglichkeiten über Internet bzw. Mobilfunk (5G). Zukünftig muss ein Kranführer seinen Kran nicht mehr zwangsläufig vor Ort bedienen, sondern könnte dies auch von Zuhause aus tun. Der Chirurg kann bei Verwendung von OP-Robotern die Operation aus der Distanz mit einem Joystick durchführen.

Mobile Arbeitsplätze kennen wir heute bereits in Form von Homeoffice, als Arbeit beim Kunden oder als Arbeit von unterwegs (z. B. Bahn, Hotelzimmer). Zukünftig wird die Zahl der mobilen Beschäftigten steigen. Die Arbeitnehmer gewinnen dank der neuen technischen Möglichkeiten an Freizeit (sie verbringen als Berufspendler weniger Zeit im Verkehrsstau), sie erreichen mehr Flexibilität (ein Jobwechsel ist nicht mehr zwangsläufig mit dem Wohnortwechsel verbunden) und es lassen sich andere Lebensentwürfe verwirklichen, etwa die bessere Vereinbarkeit von Beruf und Familie (Mülder 2016a, S. 37).

Selbst wenn die Mitarbeiter regelmäßig an einem festen Arbeitsort zusammentreffen, muss dies nicht bedeuten, dass weiterhin jeder über seinen „eigenen“ Schreibtisch oder sein „eigenes“ Büro verfügt. Stattdessen können in einem Gebäude unterschiedlich gestaltete Arbeitsräume zur Verfügung stehen, die der Einzelne je nach individuellem Bedarf stunden- oder tageweise „buchen“ kann. Eine Projektgruppe nutzt beispielsweise einen Raum, der besonders die Kreativität und Kommunikation fördert. Falls jemand längere Zeit per Telefon bzw. Videokonferenz kommunizieren muss oder in Ruhe ein neues Konzept erarbeiten möchte, stehen hierfür abgeschottete Kommunikations- und Denkräume zur Verfügung. Für den Ideenaustausch oder kurze kollegiale Gespräche eignen sich Räume, die wie Lounges eingerichtet sind mit bequemen Sitzgelegenheiten für wenige Personen oder das Treffen an Stehtischen in der Kaffeeküche (vgl. Creusen et al. 2017, S. 86). Zur Verbesserung der Zusammenarbeit und Kommunikation werden Einzel- und Zweierbüros zunehmend ersetzt durch größere, flexibel gestaltbare Arbeitsräume, die auch als Co-Working-Spaces bezeichnet werden. Der Digital Leader hat in dieser flexiblen Bürolandschaft auch nicht mehr das Sonderrecht auf ein Einzelbüro, sondern sucht sich wie seine Mitarbeiter täglich seine passende Arbeitsplatzumgebung aus.

4.2.2 Flexible Arbeitszeiten

Charakteristisch für New Work ist eine stärkere Orientierung am Output, also dem Arbeitsergebnis und weniger am Input, den geleisteten Arbeitszeiten (vgl. Walwei 2018, S. 357). Die Arbeitnehmer wünschen eine stärkere Selbstbestimmung und

Individualisierung ihrer Arbeitszeiten. Schon seit mehreren Jahren ist eine Abkehr vom früher üblichen 8-Stunden-Tag und der 5-Tage-Woche zu beobachten. Es existieren z. B. Gleitzeit-, Teilzeit- und Jahresarbeitszeitmodelle. Statt Mehrarbeitsvergütung wird Freizeitausgleich gewährt. Geleistete Arbeitszeiten werden nicht unmittelbar vergütet, sondern längerfristig „angespart“ und ermöglichen später Sabbaticals (Langzeiturlaub) oder den früheren Eintritt ins Rentenalter (Altersteilzeit). Das Unternehmen kann durch flexible Arbeitszeitmodelle seinen wechselnden Kapazitätsbedarf mit den vorhandenen Kräften decken und die Beschäftigten können ihre präferierten Arbeitszeiten in Abhängigkeit von der jeweiligen Lebenssituation bestimmen.

Andererseits verschwimmen in der neuen Arbeitswelt die Grenzen von Privatleben und Berufsalltag zunehmend. Viele Arbeitgeber akzeptieren, dass die Mitarbeiter private Angelegenheiten über ihren Laptop während der Arbeitszeit erledigen. Oftmals sind diese Personen aber auch nach Feierabend noch für ihr Unternehmen erreichbar und tätig, indem beispielsweise E-Mail auch noch am späten Abend beantwortet werden. Es sind Regelungen, wie z. B. Betriebsvereinbarungen erforderlich, die sowohl die erhöhten Anforderungen an Flexibilität – bedingt durch die neuen Techniken – und den Schutzbedarf der Arbeitnehmer miteinander in Einklang bringen. Andernfalls drohen gesundheitliche Risiken aufgrund der „Always-On-Mentalität“ (Mülder 2016b, S. 384).

Digital Leader benötigen mehr Zeit für die Einsatzplanung ihrer Mitarbeiter, weil individuelle Arbeitszeitwünsche mit immer differenzierteren betrieblichen Bedarfen in Einklang gebracht werden müssen. Hinzu kommt, dass sich v. a. qualitative Arbeitsergebnisse wesentlich schlechter messen bzw. kontrollieren lassen als die geleisteten Arbeitszeiten.

4.2.3 Veränderte Arbeitsinhalte

Digitalisierung bietet die Chance, dass körperlich anstrengende und gesundheitlich gefährliche Arbeiten reduziert und eines Tages ganz verschwinden werden. In den ersten drei industriellen Revolutionen wurden Produktivitätssteigerungen durch stärkere Arbeitsteilung erreicht. Die Komplexität der Tätigkeiten wurde immer stärker reduziert, sie beschränkte sich oftmals auf wenige Handgriffe und leicht erlernbare Routineabläufe. Diese einfachen Arbeiten können zukünftig vermehrt von Cyber-Physischen Systemen übernommen werden. Durch das Zusammenwirken von Mensch und Maschine entstehen neue Synergien, die zu weiteren Produktivitätsfortschritten führen. Der Arbeitnehmer muss komplexere Tätigkeiten übernehmen und wird hierzu von der Maschine unmittelbar angeleitet. Beispielsweise kann die Wartung einer Maschine fehlerfrei erfolgen, wenn die Maschine erkennt, dass der Monteur das falsche Werkzeug nutzt oder zu wenige Schmierstoffe verwendet. Zukünftig können derartige Hilfen per Sprache oder über eine Datenbrille gegeben werden (Mülder 2016a, S. 38).

Auf der anderen Seite werden wissensintensive, kreative und soziale Aufgaben zunehmen, sie sind in absehbarer Zeit auch nicht durch Automatisierung oder Künst-

liche Intelligenz ersetzbar. Von einem „Arbeitnehmer 4.0“ werden digitale und nicht-digitale Kompetenzen erwartet. Digitale Kompetenzen werden nicht nur gebraucht, um informationstechnische Systeme zu bedienen – dies würde der heutigen Anwenderschulung im Umgang mit einer bestimmten Software entsprechen – sondern auch um Potenziale neuer Technologien zu erkennen und diese frühzeitig einzusetzen. Zur Nutzung neuartiger Analyseverfahren im Rahmen von Big Data benötigt ein Controller zum Beispiel Kenntnisse über geeignete Analysemethoden, die erforderliche Datenbasis und die geeignete Analysesoftware. Zu den nicht-digitalisierbaren Kompetenzen zählen z. B. Flexibilität, Selbständigkeit, Kommunikationsfähigkeit und Problemorientierung (vgl. Walwei 2018, S. 353).

Der Digital Leader muss sein Team auf die neuen Arbeitsinhalte vorbereiten. Dies lässt sich vor allem durch Schulungs- und Personalentwicklungsmaßnahmen erreichen.

4.2.4 Neue Arbeitsorganisation

Arbeit lässt sich im Digitalzeitalter völlig anders organisieren. Während Arbeitsaufträge heute noch primär an die festangestellten Mitarbeiter vergeben werden, erfolgt die Vergabe zukünftig über Online-Plattformen an selbstständige **Crowdworker**. Der Arbeitgeber bzw. die Führungskraft agiert hierbei als Auftraggeber und vergibt klar definierte Arbeitsaufträge an eine Masse von potenziellen Auftragnehmern („Crowd“). Die gesamte Kommunikation, Abwicklung und Bezahlung erfolgt über die digitale Arbeitsplattform.

Bestimmte Tätigkeiten wie z. B. Erstellung von Texten, Übersetzung, Testen oder empirische Umfragen können **ortsungebunden** erledigt werden. In der Regel sind bestimmte Qualifikationen, z. B. Sprachkenntnisse, auf Seiten der Crowdworker erforderlich. Die Bezahlung erfolgt nach geleisteter Menge. Auch **ortsgebundene** Tätigkeiten werden über Internet-Plattformen vermittelt. Bekannt geworden sind in den letzten Jahren vor allem private Anbieter von Zimmern, die in direkter Konkurrenz zum Hotelgewerbe stehen, Auslieferung von Lebensmitteln, Pizza etc. in Großstädten, sowie private Personenbeförderungsdienste, die als Bedrohung des Taxigewerbes angesehen werden.

Der Auftraggeber muss für die Aufgabenerledigung keine Mitarbeiter dauerhaft oder als Leiharbeitnehmer einstellen und demnach auch keine Sozialbeiträge abführen. Die Crowdworker arbeiten auf eigenes Risiko, sie müssen sich selbstständig versichern und haben keine Arbeitsplatzsicherheit. Sie erhalten jedoch ein hohes Maß an zeitlicher und örtlicher Flexibilität. Crowdworker werden auch nicht wie „normale“ Mitarbeiter geführt, es dominiert vielmehr eine Kunden-Lieferanten-Beziehung (vgl. Mülder et al. 2018, S. 111 ff.; Walwei 2018, S. 354 f.). Prinzipiell lässt sich die Idee von Crowdworking auch unternehmensintern verwirklichen. Die Beschäftigten behalten ihren Arbeitnehmerstatus, sie arbeiten aber nicht mehr in festgefügten Strukturen und Hierarchien.

4.3 New Workforce

Der digitale Wandel kann zum Verlust von Arbeitsplätzen führen, weil neuartige Fähigkeiten und Fertigkeiten von den Mitarbeitern verlangt werden. Die meisten Beschäftigten müssen diese neuen digitalen Qualifikationen durch Weiterbildungsmaßnahmen erwerben. Zusätzlich müssen jüngere hochqualifizierte Arbeitnehmer von außen rekrutiert werden. Diese sogenannte Generation Z oder Digital Natives starten ihre Karriere jedoch mit anderen Wertvorstellungen und Zielen als die bisherige Stammbelegschaft.

4.3.1 Beschäftigungseffekte der Digitalisierung

Die Frage, ob Digitalisierung zu einem Arbeitsplatzabbau oder nicht führt, wurde in den letzten Jahren kontrovers diskutiert. Auslöser war eine US-amerikanische Studie von Frey und Osborne (2013). Insgesamt wurden 702 Berufe im Hinblick auf ihre Automatisierbarkeit und mögliche Arbeitsplatzverluste überprüft. Die Verfasser kamen zu dem Ergebnis, dass ungefähr 47 % der amerikanischen Beschäftigten in Berufen mit einem hohen Automatisierungsrisiko arbeiten. Besonders stark betroffene Branchen sind Transport, Logistik, Produktion sowie allgemeine Verwaltungstätigkeiten. Nach einer Analyse von Brzeski und Burk (2015) sind in Deutschland langfristig ca. 18 Mio. Arbeitsplätze bedroht, das sind 59 % aller sozialversicherungspflichtigen Beschäftigten. Durch Roboter (hierzu zählen auch beispielsweise Drohnen, die zukünftig Pakete ausliefern, aber auch Koch-, Garten- und Serviceroboter) und Automatisierung sind insbesondere Bürokräfte, Post- und Zustellkräfte, Verkäufer, Reinigungskräfte und Gastronomieservicekräfte bedroht. An anderer Stelle wird für 15 % der sozialversicherungspflichtigen Arbeitnehmer (ca. 4,7 Mio. Menschen) in Deutschland ein Jobverlust als Folge der Digitalisierung erwartet (vgl. Dengler und Matthes 2015).

Andere Forscher errechnen positive Beschäftigungseffekte als Folge der Digitalisierung. Rüßmann (2015) prognostizierte die Entstehung von bis zu 390.000 zusätzlichen Jobs in Deutschland innerhalb von 10 Jahren. Das Institut der deutschen Wirtschaft sieht keine Anzeichen dafür, dass Unternehmen als Folge der digitalen Transformation vermehrt Arbeitsplätze abbauen (vgl. Stettes 2016; Schmelzer und Losse 2018, S. 64). Die meisten Forscher sind sich jedoch einig, dass einfache Tätigkeiten zunehmend von Maschinen und Programmen übernommen werden, während kreative und technische Qualifikationen auf absehbare Zeit nicht ersetzbar sind.

4.3.2 Rekrutierung von Generation Z

Die informationstechnischen Kompetenzen, die im Zuge des digitalen Wandels benötigt werden, lassen sich nicht komplett durch Schulung der bisherigen Mitarbeiter vermitteln,

zusätzlich müssen qualifizierte Mitarbeiter von außen rekrutiert werden. Aufgrund des demografischen Wandels und des damit verbundenen „War for Talents“ stehen hierfür selten berufserfahrene Kandidaten zur Verfügung, überwiegend müssen junge Berufseinsteiger angesprochen werden. Damit später die Integration der „Neuen“ in vorhandene Teams und Strukturen gelingt, müssen deren Erwartungen und Motive vom Digital Leader erkannt und berücksichtigt werden.

Die gesuchte Zielgruppe von jungen, gut ausgebildeten und technikaffinen Personen wird als „Generation Z“ bezeichnet. Sie besitzt typische Werte- und Verhaltensmuster, die sich deutlich von den Einstellungen anderer Altersgruppen (Babyboomer, Generation X und Y) unterscheiden (vgl. Tab. 4.1).

Als „*Digital Natives*“ werden um die Jahrtausendwende geborenen Personen bezeichnet, die mit Internet, Social Media und Smartphone aufgewachsen sind. Sie unterscheiden sich von den „*Digital Immigrants*“, also vor der Jahrtausendwende Geborenen durch ihre intensive, vorurteilsfreie aber teilweise auch unkritische Techniknutzung und andere Wertvorstellungen (vgl. Scholz 2015, S. 68 ff.). „Wissensarbeiter“ ist eine weitere Bezeichnung für den momentan besonders nachgefragten Mitarbeiter-typus. Sie sind kreativ, flexibel, innovativ, teamorientiert, gut vernetzt und international einsetzbar (vgl. Urbach und Ahlemann 2018, S. 80).

Die umworbene Zielgruppe reagiert selten auf traditionelle Stellenanzeigen, sondern muss unkonventionell und direkt angesprochen werden. Eine Möglichkeit ist hierbei das Event-Recruiting, ein Mix aus fachlichen Herausforderungen, spielerischen Elementen, gemeinsamen Kennenlernen und offenen Karrieregesprächen. In einem Hackathon (Wortschöpfung aus Hack und Marathon) bearbeitet ein Team innerhalb einer Zeitspanne von 1 oder 2 Tagen ein gemeinsames Problem, z. B. wird eine neue Software entwickelt. Mehrere Teams können um die beste Lösung konkurrieren, am Ende werden die Ergebnisse vorgestellt und von einer Jury bewertet. DevCamps (Developer Camps) funktionieren ähnlich, im Vordergrund steht das gemeinsame Lernen und Entwickeln. Im Mittelpunkt des Barcamps steht die Wissensvermittlung und der gemeinsame Austausch. Es gibt kein vorher festgelegtes Programm und keine im Voraus geplanten Vorträge. Jeder Teilnehmer kann sich mit seinen Themen einbringen. Das Ergebnis hängt stark vom Engagement des Einzelnen und den Vereinbarungen im Team ab. Die hier vorgestellten Events lassen sich auch als interne Workshops durchführen, etwa bei agilen Projekten. Ein Digital Leader betritt mit der Organisation dieser Netzwerk-Veranstaltungen meistens Neuland. Die Teilnehmer kommen primär aus fachlichem Interesse, wollen sich untereinander austauschen und Spaß haben. Zu viel Werbung oder gar der Versuch, einen Kandidaten von einem anderen Unternehmen direkt abzuwerben, kommt bei den Teilnehmern nicht gut an und kann – da oftmals Influencer und Netzwerker teilnehmen – sogar zu einem erheblichen Imageschaden für das Unternehmen führen (vgl. Warkentin 2017).

Auch der eigentliche Rekrutierungsprozess ändert sich in digitalen Unternehmen. Bei Peer Recruiting nimmt nicht nur der Vorgesetzte am Bewerbungsinterview teil, sondern auch zukünftige Kollegen oder sogar das komplette Team (vgl. Frank 2020).

Tab. 4.1 Charakteristische Merkmale von Generation Y und Generation Z (vgl. Scholz 2014a; Mörsdorf 2020)

	Generation Y (geboren ab ca. 1980)	Generation Z (ca. ab 1995)
Lebensstil	Leben im Hier und Jetzt	Anspruchsvoll, finanzielle Anreize wichtig, international orientiert
Arbeitsleben	Arbeit muss Spaß machen	Arbeit und Privatleben müssen Spaß machen
	Work Life Blending	Klare Trennung Beruf und Privatebenen
		Modernste Technik am Arbeitsplatz
		Einzelkämpfer
	Flexibel und anpassungsbereit	Keine langfristige Bindung an Unternehmen
	Selbstständige Arbeitsweise	
	Keine Führungsposition um jeden Preis	
Kommunikations-medien	Soziale Medien wie z. B. Facebook, E-Mail	Mobil, neuere soziale Medien wie z. B. WhatsApp, Instagram, Always On

Beispiel

Bei dem Internet-Telefonie-Anbieter Sipgate werden Peers an den Bewerbungsinterviews beteiligt. Ein Bewerber arbeitet einen Tag „zur Probe“ in der zukünftigen Abteilung. Die Peers haben bei der Einstellung ein Vetorecht und können sich auch gegen einen Kandidaten entscheiden. Auch kurz vor Ablauf der Probezeit geben die Peers ein Feedback.

Google beteiligt ebenfalls zukünftige Kollegen an den Einstellungsgesprächen. Nur wenn alle am Interview Beteiligten sich einig sind, erfolgt eine Kandidatenzusage. ◀

4.4 Digital Leader

4.4.1 Persönlichkeitsmerkmale

Im Mittelpunkt steht hierbei die Frage, über welche persönlichen Eigenschaften ein Digital Leader im Idealfall verfügen sollte. Falls es gelingt, die relevanten Persönlichkeitsmerkmale zu bestimmen, die mit dem Führungserfolg korrelieren, ließen sich mittels psychodiagnostischer Verfahren geeignete Kandidaten rekrutieren und auf zukünftige Führungsaufgaben vorbereiten. Die Führungsforschung konzentrierte sich bis Mitte des 20. Jahrhunderts auf die „Great-Man-Theorie“, die auch als Eigenschaftstheorie bezeichnet wird. Aus mehreren Studien, in denen unterschiedliche Charaktereigenschaften untersucht wurden, lassen sich 5 Merkmalsgruppen identifizieren, die einen korrelativen Bezug zum Führungserfolg haben (Lippold 2017; Rosenstiel 2003, S. 7 f.; Jung et al. 2016):

- Befähigung (Intelligenz, Originalität, Urteilskraft),
- Leistung (Schule, Wissen, sportliche Erfolge),
- Verantwortlichkeit (Zuverlässigkeit, Initiative, Ausdauer),
- Partizipation (Kooperationsbereitschaft, Anpassungsfähigkeit, Humor),
- Status (sozioökonomische Position, Popularität).

Der Soziologe Max Weber prägte im Jahr 1922 als Erster den Begriff „**charismatische Führung**“ (vgl. Weber 2002). Die Ausstrahlung einer Führungskraft beeinflusst demnach in starkem Maße das Verhalten der Mitarbeiter. Sie setzen ihm absolutes Vertrauen gegenüber, sind loyal und akzeptieren ihn. Der charismatische Vorgesetzte seinerseits besitzt starken Machtwille, hohes Selbstbewusstsein, visionäre Kraft und Dominanzstreben (vgl. Lippold 2017, S. 347). Charismatische Führungseigenschaften werden oftmals auch Gründern und Top-Managern von digitalen Unternehmen zugesprochen, beispielsweise Bill Gates (Microsoft), Steve Jobs (Apple) oder Mark Zuckerberg (Facebook). Die eigenschaftstheoretischen Führungsansätze sind leicht nach-

vollziehbar und eignen sich zur Beschreibung des „idealen“ Führungsverhaltens. Die bisherigen empirischen Studien konzentrieren sich allerdings nicht auf Digital Leader. Unklar ist ferner, ob Führungseigenschaften angeboren oder erlernbar sind, ob einzelne Merkmale oder ihre Kombination entscheidend sind und ob erfolgreiche Eigenschaften situativ unterschiedlich zu bewerten sind (vgl. Scholz 2014b, S. 1162).

4.4.2 Führungskompetenzen

Ein Digital Leader muss über bestimmte Fähigkeiten verfügen, um sein Team erfolgreich durch einen digitalen Transformationsprozess zu leiten. Neben allgemeinen Kompetenzen wie z. B. Kommunikationsfähigkeit, Entscheidungsfähigkeit, Teamfähigkeit, Fach- und Methodenwissen gehört hierzu insbesondere die Digitalkompetenz. Während Digital Natives diese in der Regel bereits besitzen, müssen Führungskräfte den professionellen Umgang mit digitalen Tools noch lernen. Im Einzelnen sollte ein Digital Leader folgende Kompetenzen besitzen (vgl. Creusen et al. 2017, S. 55):

- Sicherer Umgang mit Kommunikations- und Kollaborationstools,
- Befähigung, selbstgesteuerte Recherchen und Informationssuche im Web durchzuführen,
- Fähigkeit zur Entwicklung neuer Geschäftsmodelle, digitaler Produkte und Services,
- Erkennen von Kundenproblemen und Kundenbedürfnissen,
- Kenntnisse über Agile Projektarbeit, agile Methoden,
- Erkennung zukünftig benötigter Qualifikationen, Formulierung neuer Jobprofile, z. B. Data Scientist.

4.4.3 Virtuelle Führung

Hierunter versteht man die Führung von Teams, die räumlich verteilt oder gar international verstreut sitzen. Die Kommunikation erfolgt hierbei primär über Medien wie z. B. E-Mail, Chat oder Videokonferenz. Falls die textbasierte schriftliche Kommunikation überwiegt, fehlen alle im Face-to-Face-Gespräch ansonsten erkennbaren nonverbalen Äußerungen. Bei Videokonferenzen ist z. B. nicht erkennbar, was der einzelne Teilnehmer gerade tut und wie groß sein Interesse am diskutierten Thema ist, wenn die eigene Kamera deaktiviert wird. Virtuelle Führung gelingt, wenn einige Voraussetzungen beachtet werden (vgl. Franken und Franken 2018, S. 114):

- Persönliche Treffen
Selbst wenn virtuelle Zusammenkünfte dominieren und für alle bequemer sind, sollten regelmäßig persönliche Begegnungen stattfinden. Hierdurch entstehen engere Bindungen und gegenseitiges Vertrauen wird aufgebaut.

- Sensibel kommunizieren

Bei digitaler Kommunikation sollte auf zweideutige Formulierungen und Reizwörter verzichtet werden. Ironie, Humor, Zustimmung und Ablehnung werden Online schlechter erkannt als im persönlichen Austausch.

- Achtsamkeit verbessern

Digitale Kommunikation ist meistens sachlicher und kürzer als persönliche Treffen. Trotzdem sollte die Führungskraft ihre Mitarbeiter nach dem aktuellen Befinden befragen und persönliche Beziehungssignale aussenden, wie z. B. Lob.

- Informationsaustausch strukturieren

Hierzu zählen die Festlegung verbindlicher Kommunikationszeiten, die Vorbereitung einer Agenda, die grafische Veranschaulichung, die Protokollierung von Ergebnissen sowie die Einbeziehung sämtlicher Teilnehmer.

4.5 Konzepte und Methoden für Digital Leadership

4.5.1 SCRUM

Die agile Vorgehensweise bildet eine Alternative zu exakt geplanten, sequenziellen Projektverläufen, die nach dem „Wasserfall-Modell“ organisiert sind (vgl. Abts und Mülde [2017](#), S. 431 ff.). Agile Methoden werden inzwischen auch außerhalb der IT im gesamten Unternehmen angewandt (vgl. Nowotny [2017](#)). Die wichtigsten Prinzipien sind hierbei:

- Inkrementell: flexible Reaktion auf Änderungen;
- Schnell: in kurzen Zeitabständen werden neue Teilergebnisse geliefert;
- Iterativ. Es wird in mehreren kleinen Schritten vorgegangen;
- Änderungsfreundlich: der Kunde bzw. zukünftige Software-Nutzer kann jederzeit Änderungsvorschläge einbringen;
- Kundenzentriert: der Kunde steht immer im Mittelpunkt;
- Eigenverantwortlich: das Team organisiert die Arbeit selbstständig und verantwortet das Ergebnis gemeinschaftlich; Kommunikation und Zusammenarbeit in der Gruppe sind besonders wichtig.

Scrum ist die bekannteste agile Methode. Der Begriff stammt vom Rugby-Sport, bedeutet Gedränge und beschreibt das Zusammenspiel der Mannschaft. Acht Spieler pro Mannschaft drängen sich zusammen und versuchen den Ball aus dem Haufen zu bekommen. Als Methode steht Scrum für kooperative und selbst organisierte Teamarbeit. Die Aufgaben werden schrittweise in mehreren kurzen Zyklen (2 bis 4 Wochen), die als „Sprints“ bezeichnet werden, bearbeitet. Die Scrum-Methode umfasst Rollen, Meetings und Artefakte Abb. [4.1](#).

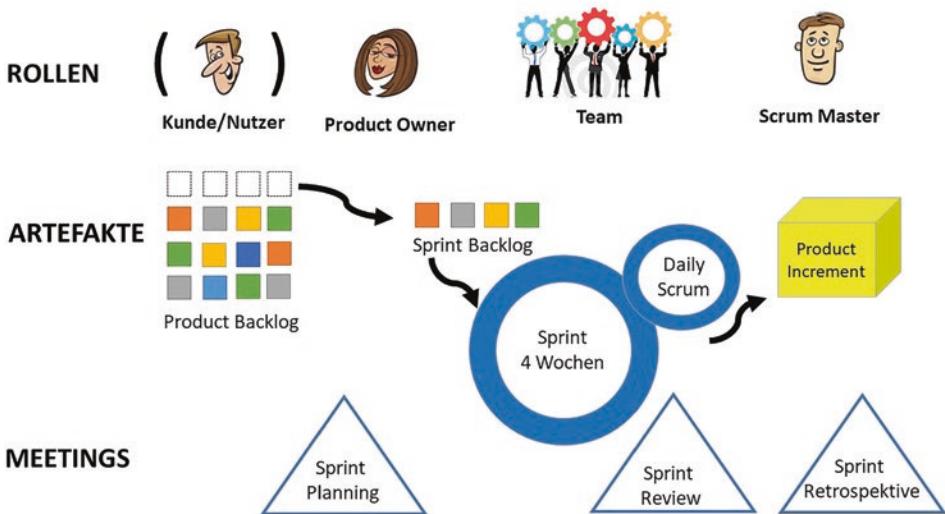


Abb. 4.1 Scrum-Methode

Rollen

Der Product Owner definiert die fachlichen Anforderungen im Projekt. Er priorisiert die Produkteigenschaften und bewertet nach jedem Sprint die Teilergebnisse. Das Entwicklungsteam erledigt die Aufgaben in einem Sprint selbstständig und eigenverantwortlich. Es ist interdisziplinär zusammengesetzt. Während eines Sprints darf das Team nicht "gestört" werden. Der Scrum-Master ist für die Einhaltung der agilen Regeln zuständig. Als Moderator und Coach fördert er die Zusammenarbeit und Kommunikation im Team. Der zukünftige Kunde bzw. Nutzer zählt zwar nicht zum Scrum-Team, spielt aber dennoch eine wichtige Rolle. Product Owner und Kunde stimmen sich regelmäßig untereinander ab und der Kunde gibt laufendes Feedback zu den erarbeiteten Zwischenergebnissen bzw. Prototypen.

Meetings

Die Kommunikation untereinander ist wichtig bei Scrum. Die maximale Zeitspanne für die einzelnen Meetings wird im Voraus festgelegt, um unnötige und unproduktive Diskussionen zu vermeiden. Im „Sprint Planning“ kommt das gesamte Scrum-Team zusammen und plant die Aktivitäten für den nächsten Sprint. Bei einem 4-wöchigen Sprint sollten hierfür maximal 8 h angesetzt werden. Beim Daily Scrum trifft sich das Entwicklungsteam täglich zur selben Zeit und am selben Ort. In maximal 15 min werden die Ergebnisse des Vortags diskutiert und es werden die aktuell anstehenden Aufgaben festgelegt. Lassen sich Hindernisse nicht sofort lösen, werden sie an den Scrum Master weitergeleitet, der sich um die Beseitigung kümmert. Sprint Reviews werden jeweils am

Ende eines Sprints durchgeführt. Der Product Owner begutachtet hierbei die erzielten Ergebnisse und aktualisiert das Product Backlog. Das gesamte Scrum-Team, aber auch z. B. der zukünftige Nutzer können ein direktes Feedback geben. Die Sprint Retrospektive findet unmittelbar nach dem Sprint Review statt und schließt den Sprint ab. Die Zusammenarbeit des Scrum-Teams steht im Mittelpunkt. Durch konstruktives Feedback soll der Arbeitsprozess kontinuierlich verbessert werden.

Artefakte

Hierzu zählen schriftliche Dokumente und (Teil-)Ergebnisse, die im Projektverlauf entstehen. Ein *Product Backlog* umfasst alle bekannten Anforderungen und Produktbestandteile, die im Projektverlauf erfüllt bzw. erreicht werden sollen. Das Product Backlog kann sich ändern aufgrund von zusätzlichen Kundenwünschen oder neuen Technologien. Die Einträge werden vom Product Owner priorisiert und im Projektverlauf immer weiter verfeinert.

Das *Sprint Backlog* enthält die für den jeweiligen Sprint zu erledigenden Aufgaben. Die Ergebnisse des Sprints werden als *Product Increment* bezeichnet. Bei Software-Entwicklung besteht das Increment aus lauffähiger Software, die vom ersten Prototypen oder Click-Dummy bis zur fertigen Produkt ständig verbessert wird.

Durch die Vermeidung langer Planungszeiten und der schnelleren Erreichung vorläufiger Resultate verkürzt sich der Entwicklungsprozess. Es kann direkt auf Kundenanforderungen reagiert werden. Die Präsentation von Zwischenergebnissen und Prototypen ermöglicht ein schnelles Feedback vom Kunden. Fehler bzw. falsch verstandene Anforderungen werden nicht erst nach monatelanger Entwicklungszeit erkannt. Die Arbeit in sich selbst organisierenden Teams steigert die Motivation der Entwickler. Die agile Vorgehensweise hat sich in den letzten Jahren als Erfolg versprechend herausgestellt: 39 % agiler Projekte konnten erfolgreich abgeschlossen werden, bei Projekten nach der Wasserfall-Methode lag diese Quote bei lediglich 11 %. Bei agiler Vorgehensweise scheiterten nur 9 % aller Projekte, während bei der Wasserfall-Methode 29 % erfolglos endeten (vgl. Kofler 2018, S. 228; Hastje und Wojewoda 2015).

Nachteilig bei Scrum ist, dass es sich lediglich um ein grobes Rahmenwerk handelt. Vor Einführung muss jedes Unternehmen die agilen Scrum-Ideen an die unternehmensindividuellen Gegebenheiten anpassen. Mitarbeiter müssen bereit sein, sich auf die neue Arbeitsweise umzustellen. Scrum kommt weitgehend ohne detaillierte Planungsphasen aus. Bei komplexen Projekten fehlt somit die genaue Vorstellung von dem endgültigen Produkt. Die regelmäßige Kommunikation zwischen allen Projektbeteiligten erfordert einen hohen Abstimmungsaufwand. Bei sehr großen Projekten entsteht das Problem der Koordination zahlreicher kleinerer Scrum-Teams untereinander (vgl. Abts und Mülder 2017, S. 440).

4.5.2 Design Thinking

Designer sind für unorthodoxe Ideen und kreative Lösungen bekannt. Mit Design Thinking wird die für Designer typische Arbeitsweise im Unternehmenskontext angewandt. Im Mittelpunkt steht der persönliche Austausch, die Ideengenerierung in einem möglichst interdisziplinären Team und die Verwendung analoger – nicht digitaler – Hilfsmittel wie beispielsweise Post-it-Zettel, Flipcharts. Es darf experimentiert werden, es dürfen mehrere Vorschläge entstehen, die anschließend auch wieder verworfen werden. Originelle Ideen sind erwünscht, unkonventionelle Perspektiven sollen eingenommen werden. Im Mittelpunkt sämtlicher Vorschläge steht der Kunde, der ein bestimmtes Bedürfnis oder Problem hat. In einem iterativen Prozess werden daher 6 Schritte durchlaufen Abb. 4.2

Zunächst einmal muss das Team den Kunden und seine Bedürfnisse verstehen. Zu diesem Zweck können Kunden befragt werden oder in ihrer realen Umgebung beobachtet werden. Mit Hilfe von *Personas* kann man sich besser in den Kunden oder zukünftigen Nutzer hineinversetzen. Es handelt sich dabei um fiktive Personen, die stellvertretend für reale Benutzergruppen mit ähnlichen Bedürfnissen und Wertvorstellungen stehen. Sie werden anschaulich beschrieben mit ihrem Beruf, Wohnort, sozialen Status, Familiensituation, private Präferenzen und ihren speziellen Bedürfnissen in Bezug auf das zu entwickelnde Produkt. Eine weitere Form der Veranschaulichung ist die „Customer Journey“. Hierbei wird eine typische Kundenreise mit allen Berührungs punkten zum Unternehmen bzw. Produkt beschrieben. Wenn beispielsweise ein neues Tool zur Datenanalyse im Handel entwickelt werden soll, könnte die Customer Journey sämtliche Berührungs punkte zum Unternehmen beschreiben, an denen Daten entstehen. Die Customer Journey beginnt z. B. mit Produktrecherchen im Internet. Weitere Stationen wären der Bezug eines Newsletters, Feedback und „Likens“ in sozialen Medien, die Anmeldung eines Online-Accounts, der Kaufvorgang (im Online-Shop oder im Ladengeschäft), Rücksendung, Reklamation, Treuebonus für Stammkunden.

Im nächsten Schritt, der Synthese, werden alle bisher gesammelten Informationen zu einem Gesamtbild zusammen gefügt. Danach werden neue Ideen entwickelt. Im

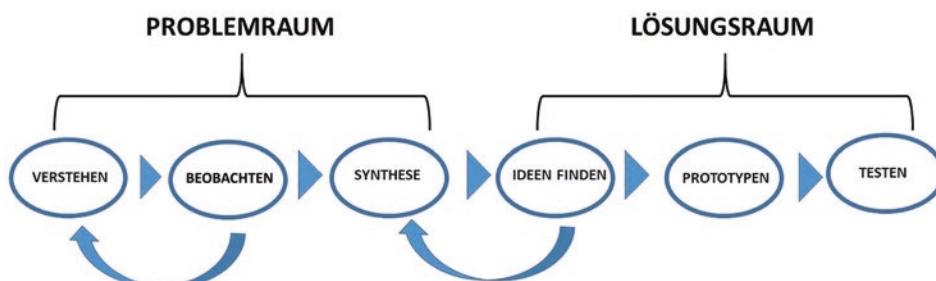


Abb. 4.2 Design Thinking (vgl. Graef und O'Mahony 2019, S. 1082)

gemeinsamen Brainstorming sollen mehrere, auch unkonventionelle Vorschläge formuliert und möglichst auch visualisiert werden. Danach entwickelt das Team bereits die ersten Prototypen. Hier wird noch nicht mit endgültigen Materialien gearbeitet, sondern z. B. mit Papier, Pappe oder Legosteinen. Bei einer Software-Entwicklung entsteht ein „Click Dummy“ oder „Mock up“, der aber bereits erste Benutzer-Interaktionen ermöglichen sollte. Es ist möglich, dass Prototypen auch komplett verworfen werden, nach dem Prinzip „Fail fast, fail cheap“. Letzter Schritt ist das Testen des Prototyps durch den Kunden und die kontinuierliche Erweiterung zum fertigen Produkt (vgl. Graef und O’Mahony 2019, S. 1083 f.; Nowotny 2017, S. 167 ff.).

Bei Design Thinking sind alle Teammitglieder gleichberechtigt. Ein Digital Leader kann zwar mitwirken, aber nicht in seiner Führungsrolle. Die Methode fördert ohne Zweifel unorthodoxes und kreatives Vorgehen. Allerdings müssen auch bei Design Thinking letztlich verwertbare und verkaufbare Produkte entstehen. Wenn im Unternehmen neue Produkte bislang ingenieurmäßig geplant wurden mit hohen Ansprüchen an Qualität und Zuverlässigkeit, dann wird die neue Methode zunächst einer kritischen Überprüfung standhalten müssen.

4.5.3 Servant Leadership

Die Anwendung agiler Prinzipien im gesamten Unternehmen kann zur Auflösung starrer hierarchischer Organisationsstrukturen führen. Die Mitarbeiter bestimmen selbst darüber, wie sie ihre Aufgaben bewältigen. Die Führungsverantwortung wird aufgeteilt, Führender und Geführter werden als gleichberechtigte Akteure betrachtet. Der Mitarbeiter wird als fachlicher Experte anerkannt, der Vorgesetzte agiert als Coach und Beschleuniger bei der Problemlösung. (vgl. Lippold 2017, S. 365 f.)

Servant Leadership Abb. 4.3 verlagert die fachliche Macht und Entscheidungskompetenz auf die Fachexperten. Beim Top-Management verbleiben die rechtliche Macht und Vertretungskompetenz. Das mittlere Management wird nicht mehr benötigt und kann neue Rollen übernehmen, wie z. B. als Scrum-Master, der ohne disziplinarische Führungsverantwortung dafür sorgt, dass agile Prozesse eingehalten und Hindernisse für das Team beiseite geräumt werden (vgl. Kofler 2018, S. 184 f.).

4.5.4 VOPA + Modell

Das VOPA+ Modell umfasst verschiedene Prinzipien, Methoden und Tools zur Führung im Digitalzeitalter Abb. 4.4.

Vernetzung bildet eine wichtige Voraussetzung für einen abteilungs- und regional-übergreifenden Erfahrungsaustausch. Persönliche Netzwerke entstehen z. B. in gemeinsamen Kaffeeküchen, beim regelmäßig organisierten gemeinsamen Frühstück,

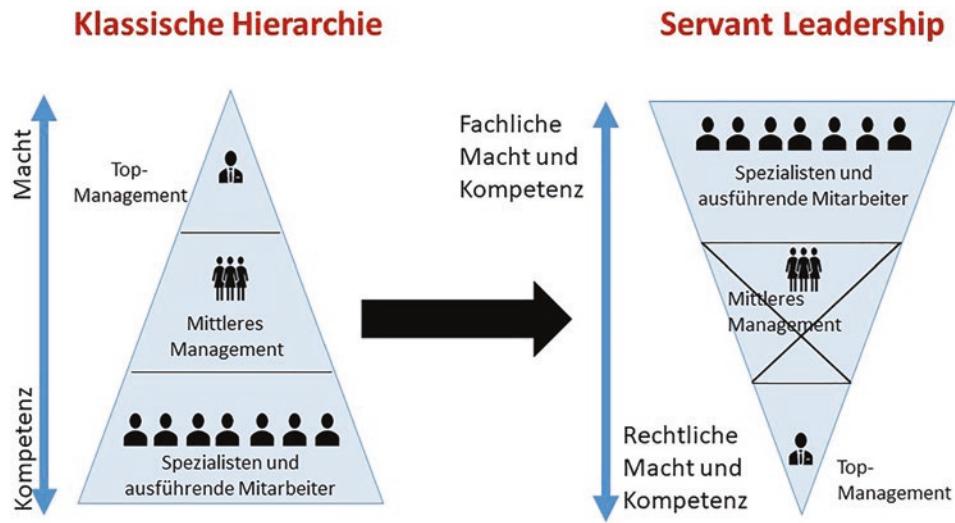


Abb. 4.3 Servant Leadership (in Anlehnung an Kofler 2018, S. 185)

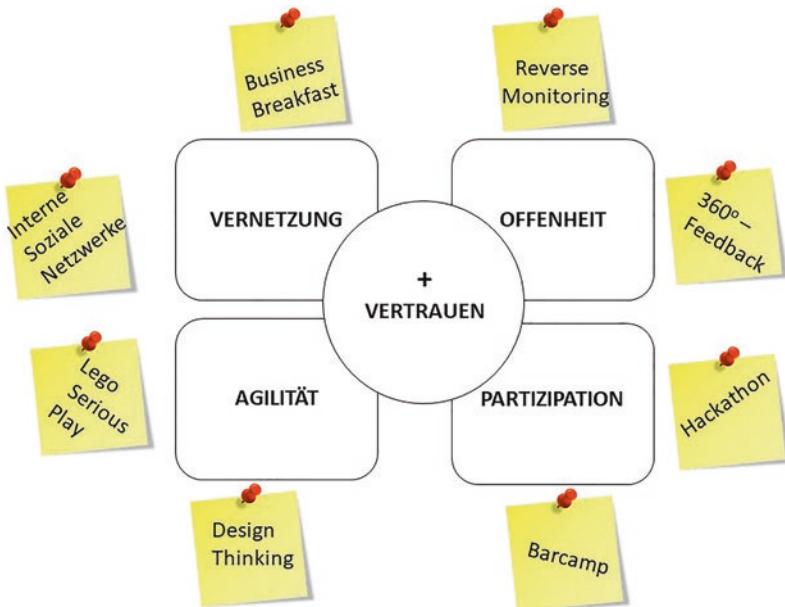


Abb. 4.4 Digital Leadership Tools im VOPA + Modell (in Anlehnung an Petry 2018, S. 314)

beim Feierabendbier. Zur Unterstützung können Softwaretools wie beispielsweise interne soziale Netzwerke genutzt werden.

Der Digital Leader sollte **offen** sein für technische Neuerungen; er sollte offen Feedback geben und ebenso offen für Kritik anderer sein. Als Methoden eignet sich hierzu beispielsweise Reverse Monitoring. Junge, digital versierte Mitarbeiter geben ihre Erfahrungen an ältere Kollegen oder Vorgesetzte weiter. Beim 360° – Feedback werden nicht nur wie bisher üblich die eigenen Mitarbeiter durch den Chef beurteilt, sondern es erhalten auch Vorgesetzte ein Feedback von ihren Mitarbeitern, gleichrangige Kollegen beurteilen sich gegenseitig und ggf. wird auch von Externen, z. B. Kunden ein Feedback eingeholt (vgl. Scholz 2014b, S. 506).

Digital Leader sind nicht allwissend, mehr denn je sind sie auf Fachwissen und Expertise anderer angewiesen. Die Bereitschaft, das Wissen bereitwillig zu teilen, dürfte in einem streng hierarchisch strukturierten Unternehmen kaum vorhanden sein. Bei **partizipativer** Führung können alle ihre Ideen und Meinungen einbringen und gemeinsam die bestmögliche Entscheidung treffen. Barcamp, Hackathon und DevCamp eignen sich als partizipative Workshop-Formate. Die *Lego Serious Play-Methode* fördert den gemeinsamen kreativen Problemlösungsprozess. Genutzt werden hierzu die uns allen bekannten Legosteine. Durch das Bauen und Gestalten werden alle, auch ansonsten eher introvertierte Teilnehmer am Arbeitsprozess beteiligt. Alle erhalten die gleiche Aufgabe und „bauen“ ihre eigene Lösung. Bei der anschließenden Diskussion erhalten alle ähnliche Redeanteile, jeder kann und soll sich zu den Beiträgen der anderen äußern. Das Wissen der unterschiedlichen Teilnehmer wird durch diese Methoden allen anderen zugänglich gemacht. Durch diesen spielerischen und haptischen Ansatz werden neue Ideen gefördert, die Kommunikation verbessert und Problemlösungen beschleunigt (vgl. Lego 2020; Petry 2018, S. 317).

Die digitale Transformation erfordert schnelles und flexibles Reagieren. **Agile** Methoden wie SCRUM und Design Thinking eignen sich hierfür am besten.

4.6 Fazit

Die Digitale Transformation verändert massiv die Art und Weise, wie wir arbeiten. Die meisten Arbeitnehmer müssen neue Qualifikationen erwerben. Für viele neuartigen Tätigkeiten müssen junge, gut ausgebildete Menschen eingestellt werden. Sie sind anders zu motivieren als die langjährig Tätigen und stellen hohe Anforderungen an Job, Karriere und ihren Vorgesetzten. Auch Führung verändert sich grundlegend im Zeitalter der Digitalisierung. Digital Leader müssen sich rechtzeitig auf ihre neue Führungsrolle vorbereiten: Sie müssen Erfahrungen in der Anwendung der neuen Tools erwerben, sie müssen Kontrolle und Macht an die Experten und Projektteams abgeben und müssen agile und unkonventionelle Arbeitsmethoden zulassen.

Literatur

- Abts, D., Mülder, W.: Grundkurs Wirtschaftsinformatik, 9. Aufl. Springer Fachmedien, Wiesbaden (2017)
- Bergmann, F.: Die Freiheit leben. Arbor, Freiamt (2005)
- Brzeski, C., Burk, I.: Die Roboter kommen – Folgen der Automatisierung für den deutschen Arbeitsmarkt (2015). www.ing-diba.de/pdf/ueber-uns/presse/publikationen/ing-diba-economic-research-die-roboter-kommen.pdf (2015). Zugegriffen: 30. Apr. 2015
- Creusen, U., Gall, B., Hackl, O.: Digital Leadership – Führung in Zeiten des digitalen Wandels. Springer, Wiesbaden (2017)
- Dengler, K., Matthes, B.: Folgen der Digitalisierung für die Arbeitswelt: Substituierbarkeitspotenziale von Berufen in Deutschland, Nürnberg 2015. IAB – Forschungsbericht 11. <https://doku.iab.de/forschungsbericht/2015/fb1115.pdf> (2015). Zugegriffen: 07. Juli 2020
- Fischer, S., et al.: Implikationen von Arbeit 4.0 auf die Personalarbeit. In: Werther, S., Bruckner, L. (Hrsg.) Arbeit 4.0 aktiv gestalten, S. 87–161. Springer, Berlin (2018)
- Frank, S.: Peer Recruiting: Wenn Mitarbeiter Mitarbeiter einstellen. <https://espiridon.com/peer-recruiting-wenn-mitarbeiter-mitarbeiter-einstellen-1178> (2020). Zugegriffen: 08. Juli 2020
- Franken, R., Franken, S.: Wandel von Managementfunktionen im Kontext der Digitalisierung. In: Hirsch-Kreinsen, H., Ittermann, P., Niehaus, J. (Hrsg.) Digitalisierung industrieller Arbeit, 2. Aufl., S. 98–120. Nomos, Baden-Baden (2018)
- Frey, C. B., Osborne, M.A.: The future of employment: how susceptible are jobs to computerisation, Working Paper Oxford Martin School (17.09.2013) https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (2013). Zugegriffen: 17. Juli 2020
- Graef, M., O'Mahony, N.: Design Thinking. WISU **10**(2019), 1082–1084 (2019)
- Hastje, S., Wojewoda, S.: Standish group 2015 chaos report – q&a with jennifer lynch (04.10.2015) <https://www.infoq.com/articles/standish-chaos-2015/> (2015). Zugegriffen: 15. Juli 2020
- Jung, R.H., Heinzen, M., Quarg, S.: Allgemeine Managementlehre, 6. Aufl., Berlin (2016)
- Kofler, T.: Das digitale Unternehmen. Springer, Berlin (2018)
- Krug, P., Weiß, M., Lang, J.: Digital Leadership: Führung im Zuge der digitalen Transformation. Wirtschaftsinformatik & Management **10**(6), 48–59 (2018)
- Lego: Die Lego®Serious Play® Methode, <https://legoinhe.de/die-lego-serious-play-methode/> (2020). Zugegriffen: 12. Juli 2020
- Lindner, D., Greff, T.: Führung im Zeitalter der Digitalisierung – was sagen Führungskräfte? HMD **53**(3), 628–646 (2019)
- Lippold, D.: Marktorientierte Unternehmensführung und Digitalisierung. Oldenbourg, Berlin (2017)
- Mörstedt, A. B.: Erwartungen der Generation Z an die Unternehmen (Vortragsunterlagen) o.O., <https://www.pfh.de/fileadmin/Content/PDF/forschungspapiere/vortrag-generation-z-moerstedt-ihk-goettingen.pdf> (2020). Zugegriffen: 07. Juli 2020
- Mülder, W.: Arbeit 4.0: Smarte Maschinen für Smart Workers? Wissensmanagement **4**, 37–39 (2016)
- Mülder, W.: Arbeitswelt 4.0 – Rolle der Arbeitnehmer in einer hochtechnisierten Industrie. Zeitschrift für wirtschaftlichen Fabrikbetrieb **111**(6), 383–385 (2016)
- Mülder, W., Endregat, R., Witten, E.: Grundlagen der Unternehmensführung, Herne (2018)
- Nowotny, V.: Agile Unternehmen, 3. Aufl., Göttingen (2017)
- Petry, T.: Ansätze zur adäquaten Führung im Digitalen Zeitalter – Darstellung einer Digital Leadership-Toolbox. In: Petry, T., Jäger, W. (Hrsg.). Digital HR, Freiburg u.a. S. 311–325 (2018)

- Rosenstiel, L.V.: Führung zwischen Stabilität und Wandel. Vahlen, München (2003)
- Rüßmann, M. et al.: Industry 4.0, April 2015. www.zvw.de/media.media.72e472fb-1698-4a15-8858-344351c8902.f.original.pdf (2015). Zugegriffen: 16. Juli 2020
- Schmelzer, T., Losse, B.: Keine Angst vor Robotern. Wirtschaftswoche 15, S. 64 (2018)
- Scholz, H.C.: Generation Z, 1 Aufl., Wiley-VCH, Weinheim (2014a)
- Scholz, H.C.: Personalmanagement, 6. Aufl., Springer Gabler, München (2014b)
- Scholz, H.C.: Generation Z: Digital Native oder digital naiv? HR Performance 1(6871), 1 (2015)
- Stettes, O.: Arbeitswelt der Zukunft – Wie die Digitalisierung den Arbeitsmarkt verändert, Forschungsberichte aus dem Institut der deutschen Wirtschaft, Köln 2016. <https://www.econstor.eu/bitstream/10419/157155/1/IW-Analyse-108.pdf> (2016). Zugegriffen: 07. Juli 2020)
- Urbach, N., Ahlemann, F.: Der Wissensarbeitsplatz der Zukunft: Trends, Herausforderungen und Handlungsempfehlungen. In: Hofmann, J. (Hrsg.) Arbeit 4.0 – Digitalisierung, IT und Arbeit, S. 79–93. Springer, Wiesbaden (2018)
- Walwei, U.: Die digitale Wirtschaft: Was ändert sich am Arbeitsmarkt? In: Bär, V.C., Grädler, T., Mayr, R. (Hrsg.) Digitalisierung im Spannungsfeld von Politik, Wirtschaft, Wissenschaft und Recht, Bd. 2, S. 345–361. Springer Gabler, Berlin (2018)
- Warkentin, N.: Barcamp: Definition, Methode, Regeln, o.O. <https://karrierebibel.de/barcamp/#Barcamp-Definition-Was-ist-ein-Barcamp> (2017) (08.07.2020)
- Weber, M.: Wirtschaft und Gesellschaft, 5. Aufl. Mohr Siebeck, Tübingen (2002)

Teil II

Data Architect: Informationsarchitekturen gestalten – Daten effizient verwalten

Christoph Quix: Data Engineering

Detlev Frick: Data Governance

Uwe Schmitz: Einsatz von In Memory Technologien

Christoph Quix: Big Data Technologien

Christian Rupert Maierhofer: Information Data Models: Das Fundament einer guten
Information Strategy



Data Engineering

5

Christoph Quix

Zusammenfassung

Die Digitalisierung schafft eine stetig steigende Flut von Daten unterschiedlichster Art in Unternehmen, die in den Geschäfts- oder Produktionsprozessen an verschiedenen Stellen benötigt werden. Die Verfügbarkeit und Nutzbarkeit von Daten ist für die Steuerung eines Unternehmens auf verschiedenen Ebenen unentbehrlich. Zwar haben viele Unternehmen diese Notwendigkeit erkannt, jedoch stellen die vielfältigen Daten-Management-Lösungen, die in den letzten Jahren unter dem Schlagwort „Big Data“ entstanden sind, die Unternehmen vor die Herausforderung ein geeignetes „Ökosystem“ für das Daten-Management im Unternehmen aufzubauen.

Data Engineering beschäftigt sich mit verschiedenen Aspekten, die für ein effizientes und effektives Daten-Management notwendig sind. Anhand von verschiedenen Vorgehensmodellen für Data Science werden zunächst die Aufgaben charakterisiert, die zum Data Engineering zählen. Die wesentlichen Aufgaben, wie zum Beispiel die Konzeption einer Architektur für das Daten-Management, Datenmodellierung und Datenintegration, werden dann in den folgenden Abschnitten im Detail diskutiert.

C. Quix (✉)

FB Elektrotechnik/Informatik, Hochschule Niederrhein, Krefeld, Deutschland
E-Mail: christoph.quix@hs-niederrhein.de

5.1 Aufgaben des Data Engineering

Der Begriff Data Engineering ist in der Fachliteratur nicht genau definiert und wird oft im Kontext der Begriffe „Data Management“ und „Information Engineering“ verwendet. Wir haben uns im Rahmen dieses Buches für den Begriff „Data Engineering“ entschieden, da wir einerseits die Repräsentation von Informationen als digitalisierte Daten als Teil des Data Engineering sehen, also weniger die Entwicklung von konzeptuellen Informationsmodellen. Andererseits zählen wir zum Data Engineering technische Aufgaben wie die Definition von Datenstrukturen oder die Konfiguration von Daten-Management-Systemen, um eine möglichst gute Performance bei der Verarbeitung von Daten zu erreichen. Zwar benutzen wir in diesem Kapitel auch den Begriff Daten-Management, meinen aber damit nicht das Daten-Management im weiteren Sinne, das zum Beispiel auch organisatorische Fragen umfasst, die im Themenfeld Data Governance oder auch Datenqualitätsmanagement betrachtet werden. Den Begriff Daten-Management benutzen wir in diesem Kapitel quasi synonym für Data Engineering.

Betrachtet man verschiedene Vorgehensmodelle für Datenanalyse, wie zum Beispiel den Prozess zu Knowledge Discovery in Databases (KDD, Fayyad et al. 1996) oder CRISP-DM (Cross-Industry Standard Process for Data Mining, Wirth und Hipp 2000), so beschäftigt sich Data Engineering vor allem mit den Aufgaben, die vor der eigentlichen Datenanalyse stattfinden. Dazu zählen unter anderem das Verstehen der vorhandenen Daten, die Auswahl der Daten und die Aufbereitung der Daten, sodass sie mit geeigneten Methoden analysiert werden können. Im „Vorgehensmodell für Data-Science-Projekte“ (Schulz und Neuhaus 2020) werden die Aktivitäten des Data Engineering im Schlüsselbereich „Daten“ erfasst, wozu Datenbeschaffung, -integration, -bereinigung, -transformation und -speicherung gezählt werden.

Während beim KDD-Prozess und beim CRISP-DM Daten am Anfang als gegeben angenommen werden (die Daten sind einfach „da“), wird im Vorgehensmodell nach (Schulz und Neuhaus 2020) der Punkt der Datenbeschaffung ebenfalls miteinbezogen. In der Praxis stellt sich oft heraus, dass die Datenbeschaffung ein äußerst zäher Prozess ist, der viel Kommunikation mit verschiedenen Unternehmensbereichen (IT und Fachabteilungen) erfordert, um die Datensilos in einem Unternehmen erschließen zu können. Eine wichtige Aufgabe des Data Engineering ist daher die Definition einer entsprechenden Daten-Management-Architektur für das Unternehmen, in der Daten zusammengeführt und miteinander verknüpft werden können. Diesen Punkt werden wir in Abschn. 5.2 genauer betrachten.

In CRISP-DM stehen das Verständnis der Geschäftsprozesse („Business Understanding“) und das Datenverständnis („Data Understanding“) am Anfang des Analyseprozesses. Das Business Understanding bildet die Schnittstelle zwischen dem Geschäftsprozessmanagement und der datenorientierten Aufgaben, da ohne ein Verständnis der Prozesse im Unternehmen, die Daten produzieren und konsumieren, ein detailliertes Verständnis der Daten nicht entwickelt werden kann. Data Science – und damit auch Data

Engineering – wird aber heute nicht nur im betriebswirtschaftlichen Kontext angewandt, sondern auch zur Analyse von Produktionsdaten (Industrie 4.0), medizinischen Daten oder auch Mobilitätsdaten (zum Beispiel zur Verkehrsanalyse oder beim autonomen Fahren) eingesetzt, um nur einige weitere Anwendungsbeispiele zu nennen. Von daher ist für das Data Engineering ein „Domain Understanding“ erforderlich, um die vorliegenden Daten korrekt interpretieren zu können.

Während sich die domänenspezifischen Kenntnisse nicht allgemein formalisieren lassen, sollte das Verständnis über die Daten in einem Datenmodell formalisiert werden. Klassischerweise werden bei der Datenmodellierung vor allem die Strukturen von Daten definiert, zum Beispiel durch ein relationales Schema. Bei der Datenbank-Entwicklung oder Anwendungsentwicklung geschieht dieser Schritt meist zu Beginn des Projekts, um eine Struktur für die neu zu erfassenden Daten zu beschreiben. Im Kontext von Data-Science- und Big-Data-Projekten sind die Daten jedoch oft schon in diversen Datenquellen vorhanden. Hierbei geht es dann mehr darum, die Daten zu verstehen, Verknüpfungen und Regeln in den Daten zu erkennen, und dann in einem formalen Datenmodell zu dokumentieren. Dieses „Reverse Engineering“ der Datenmodelle sollte durch die Erfassung von weiteren Metadaten ergänzt werden, die zum Beispiel eine inhaltliche Beschreibung der Daten geben oder eine Verknüpfung zu Domänenstandards herstellen. Die Datenmodellierung und das Metadaten-Management werden wir genauer in Abschn. 5.3 betrachten.

Die Hauptaufgabe des Data Engineering ist die Aufbereitung und Integration von Daten. Es wird vielfach berichtet, dass das „Data Wrangling“ (Heer et al. 2019) den meisten Aufwand innerhalb eines Data-Science-Projekts erfordert (Lohr 2014; Press 2014; Council 2019), weil die Daten sehr heterogen sind und oft nicht da^{5.4} gewünschte Format und die erforderliche Qualität für eine Datenanalyse haben. Ein Data Engineer muss also eine Vielfalt von Methoden und Werkzeugen beherrschen, um Daten in das gewünschte Format zu bringen. In Abschn. 5.4 geben wir einen kurzen Überblick über die erforderlichen Schritte bei der Datenaufbereitung und -integration.

Das Kapitel schließt ab mit einem Überblick über verschiedene Systeme zum Daten-Management. Dazu gehören natürlich auch die klassischen Datenbank-Management-Systeme, NoSQL-Systeme und neuere Big-Data-Systeme. Letztere werden wir aber noch in einem separaten Kapitel detailliert betrachten.

5.2 Architekturen zum Daten-Management

Daten werden heute in Unternehmen in komplexen, heterogenen Systemlandschaften verwaltet, die aus einer Vielzahl von verschiedenen Daten-Management-Systemen bestehen (Brodie 2010). Zwar wird für die Mehrzahl der Informationssysteme nach wie vor ein relationales Datenbank-Management-System zur Datenspeicherung eingesetzt, mit steigender Popularität von NoSQL-Datenbank-Management-Systemen oder anderen Speziallösungen für das Daten-Management, ist die Heterogenität in den letzten Jahren

noch weiter gewachsen und führt zunehmend zu „Polyglot Persistence“ (Sadalage und Fowler 2012). In einer Daten-Management-Architektur sollte das Zusammenspiel der verschiedenen Systeme geregelt werden. Insbesondere muss ein Konzept dafür entwickelt werden, wie die Daten aus den unterschiedlichen Systemen verfügbar gemacht werden können, so dass sie in Data-Science-Projekten genutzt werden können. Hierbei geht es nicht darum, ein Architekturkonzept für das Unternehmen zu wählen, sondern vielmehr die Ko-Existenz von verschiedenen Architekturen und Systemen zu gewährleisten.

Durch den steigenden Spezialisierungsgrad von Software-Systemen, Reorganisationen im Unternehmen, Abteilungsdenken oder kaum abgestimmte IT-Projekte in verschiedenen Unternehmensbereichen entstehen oft informationstechnische Insellösungen, auf deren Daten oft nur mit großem Aufwand zugegriffen werden kann. Für eine umfassende Datenanalyse sind jedoch gerade diese Datenquellen wichtig. Daher sollten die Daten dieser Quellen in einem einheitlichen System zusammengeführt werden.

Im Laufe der 1990er Jahre wurden Data-Warehouse-Systeme vorgeschlagen, um innerhalb eines Unternehmens die Daten aus verschiedenen operativen Systemen zusammenzuführen (Jarke et al. 2003). Einer der Hauptgründe für die Einführung von Data-Warehouse-Systemen war die Trennung von OLTP (On-Line Transaction Processing) und OLAP (On-Line Analytical Processing). Bei OLTP werden die operativen Geschäftstransaktionen unterstützt (z. B. Erfassung einer Bestellung oder Buchung einer Rechnung), was auf Datenbankebene durch kurzlebige und eher einfache Transaktionen umgesetzt werden kann. OLAP hingegen beruht auf länger laufenden analytischen Anfragen, bei denen ein größerer Datenbestand ausgewertet werden soll (z. B. alle Verkaufsaktivitäten in einem Quartal). Diese beiden Transaktionsmodelle konnte damals nicht in einem System umgesetzt werden, auch weil für die beiden Konzepte die Daten in unterschiedlichen Datenstrukturen und Schemata verwaltet werden sollten. Daher wurden Data-Warehouse-Systeme vorgeschlagen, die die Daten aus den operativen OLTP-Systemen extrahieren, aufbereiten und in bereits voraggregierter Form (z. B. Umsatzzahlen nach Zeitraum, Produktkategorie oder Region) in der Data-Warehouse-Datenbank zur Verfügung stellen.

Die Aufbereitung der Daten aus den heterogenen OLTP-Systemen stellte sich als eine der Hauptherausforderungen in diesem Kontext heraus, weshalb sich um diese ETL-Prozesse (Extraktion-Transformation-Laden) ein eigenes Forschungsgebiet und ein separater Markt mit ETL-Werkzeugen herausbildete (Simitsis und Vassiliadis 2018). Da die ETL-Prozesse durchaus komplex und zeitaufwendig sein können, wurden anfangs die Daten im Data Warehouse nur täglich aktualisiert. Mittlerweile können die ETL-Prozesse die Daten im Data Warehouse in nahezu Echtzeit aktualisieren, was den Begriff „Real-Time Data Warehousing“ geprägt hat. ETL-Werkzeuge kommen heute nicht nur bei Data-Warehouse-Systemen zum Einsatz, sondern werden für jegliche Art von Datentransformationen genutzt.

Eine schematische Darstellung einer Data-Warehouse-Architektur ist in der linken Hälfte von Abb. 5.1 dargestellt. Zusätzlich zum Data Warehouse können die Daten noch

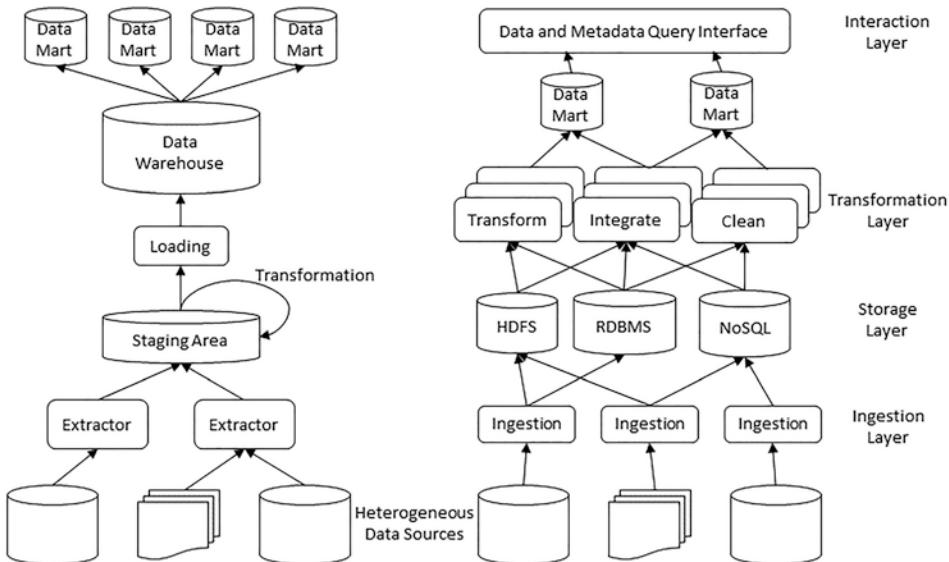


Abb. 5.1 Schematische Darstellungen einer Data-Warehouse- (links) und einer Data-Lake-Architektur (rechts)

in nachgelagerten Data Marts abgelegt werden, die nur einen Teil der Daten des Data Warehouse beinhalten, der für eine bestimmte Anwendung relevant ist.

Data-Warehouse-Systeme sind mittlerweile eine etablierte Technologie und diverse Anbieter bieten Standard-Software-Pakete für Data-Warehouse-Systeme oder ETL-Prozesse an. Die meisten mittleren und größeren Unternehmen setzen Data-Warehouse-Systeme als Teil ihrer Business-Intelligence-Systeme bzw. Führungsinformationssysteme ein. Der Data-Warehouse-Ansatz ist vor allem geeignet für Anwendungsfälle, in denen die Datenstrukturen der Datenquellen und benötigten ETL-Prozesse sehr stabil sind und sich nur selten ändern. Einerseits ist die Erstellung der ETL-Prozesse sehr aufwendig, daher wäre eine Änderung an den Datenquellen bzw. der gewünschten Data-Warehouse-Struktur ebenfalls sehr aufwendig in der Umsetzung. Andererseits ist diese Konstanz von der Anwendung her gewünscht, weil man häufig langfristige Zahlen miteinander vergleichen will. Würde sich die Berechnungsgrundlage häufig ändern, wäre eine sinnvolle Analyse von langfristigen Daten nicht möglich.

Bei Big-Data-Projekten ist diese Beständigkeit aber nicht immer gegeben. Zwar gibt es auch Big-Data-Probleme, bei denen man den Data-Warehouse-Ansatz einsetzen kann (z. B. Thusoo et al. 2010a, 2010b), jedoch steht bei den meisten Big-Data-Systemen das Schema-on-Read-Konzept im Vordergrund. „Schema-on-Read“ heißt, dass die Daten zunächst ohne ein vorgegebenes Schema abgelegt werden. Erst beim Lesen wird ein Schema definiert – und dabei auch eine Abbildung von der Struktur in der Datenquelle in das gewünschte Schema. Im Gegensatz dazu wird beim Data-Warehouse-System

der Schema-on-Write-Ansatz verfolgt. Die Schemata der Datenquellen und der Data-Warehouse-Datenbank sind bekannt bevor Daten mittels ETL-Prozessen in die Data-Warehouse-Datenbank geschrieben werden. Business-Intelligence- und OLAP-Anwendungen, die auf das Data-Warehouse-System zugreifen, erwarten, dass die Daten in einem relationalen Schema vorliegen.

Das Schema-on-Write-Modell ist aber für Big Data nicht passend, da es eine größere Zahl, mehr Heterogenität und eine größere Agilität bei den Datenquellen gibt. Auch sind beim Aufbau eines Big-Data-Systems nicht alle möglichen Anwendungsfälle vorhersehbar, sodass ein integriertes Schema nicht a-priori definiert werden kann. Insgesamt ist das Vorgehen im Big-Data-Kontext mehr auf Ad-Hoc-Analyse und Datenexploration ausgelegt, anstelle von standardisiertem Reporting wie bei Data-Warehouse-Systemen. Daher sollte für die Datenbereitstellung bei Big-Data-Projekten eine andere Architektur gewählt werden, die eine größere Flexibilität erlaubt.

In Abb. 5.1 ist auf der rechten Seite die Architektur eines Data-Lake-Systems dargestellt. Die Idee zu Data Lakes wurde erstmals 2010 vorgestellt und wurde seit etwa 2015 intensiver in Wissenschaft und Wirtschaft diskutiert(Quix und Hai 2019). Im Gegensatz zu Data-Warehouse-Systemen werden bei Data-Lake-Systemen die Daten in ihrer ursprünglichen Form in die Storage-Ebene des Systems übernommen. Ein solcher Ansatz passt zu den Big-Data- und NoSQL-Systemen, die üblicherweise nicht die Definition eines Schemas erfordern, bevor Daten abgelegt werden können. Insbesondere Hadoop erlaubt über sein verteiltes Dateisystem (Hadoop Distributed File System, HDFS) die Speicherung von unstrukturierten oder semi-strukturierten Daten und eignet sich daher sehr gut zur Realisierung der Storage-Ebene in einem Data-Lake-System.

Zur Datenspeicherung können aber auch gleichzeitig noch weitere Daten-Management-Systeme eingesetzt werden, zum Beispiel klassische relationale Datenbank-Management-Systeme für tabellarische Daten (oder auch wie Hive als Add-On-System für Hadoop, Thusoo et al. 2010b), NoSQL-Datenbank-Management-Systeme für spezielle Datenstrukturen (zum Beispiel XML- oder JSON-Dokumente, Graphen). Wie gesagt sollen die Daten im Data Lake in ihrer ursprünglichen Form abgelegt werden, d. h. Harmonisierung und eine Überführung in ein einheitliches Schema wie bei Data-Warehouse-Systemen ist hier nicht vorgesehen. Die Ingestion-Ebene realisiert im Wesentlichen eine Kopieroperation, wobei allerdings auch neben den eigentlichen Daten auch Metadaten aus den Datenquellen extrahiert bzw. separat erfasst werden sollen. Der Mehrwert des Data-Lake-Systems gegenüber einem Zugriff auf die Datenquellen sollte vor allem durch eine einheitliche Datenabfrageschnittstelle realisiert werden. Apache Spark (Zaharia et al. 2016) bietet zum Beispiel mit DataFrames, Konnektoren und SparkSQL eine Möglichkeit, Daten aus verschiedenen heterogenen Systemen abzurufen und in einem einheitlichen Rahmenwerk zusammenzuführen.

Das Metadaten-Management hat bei Data-Lake-Systemen eine noch größere Bedeutung als bei Data-Warehouse-Systemen. Während bei Data-Warehouse-Systemen die meist relationalen Datenbank-Management-Systemen eine ausreichende

Selbstauskunft zu den Schemata ihrer Datenbanken geben können, ist dies im Kontext von Data-Lake-Systemen aufgrund der unstrukturierten Daten nicht immer der Fall. Zudem ist durch die Komplexität der ETL-Prozesse sichergestellt, dass nur kuratierte Daten ins Data Warehouse gelangen. Da ein Data Lake in der Lage sein soll, alle Daten aufzunehmen, ist hier bei der Datenaufnahme (Ingestion) schon darauf zu achten, dass Metadaten extrahiert werden und ein Mindestmaß an Datenqualitätsanforderungen bei der Aufnahme überprüft werden. Ansonsten besteht die Gefahr, dass der Data Lake versumpft, da Daten nicht mehr ausreichend beschrieben und somit nicht auffindbar und nutzbar sind.

Metadaten sind auch wichtig für die Anfrageverarbeitung im Data-Lake-System. Eine integrierte Abfrageschnittstelle hilft nicht, wenn man nicht weiß, in welchen Daten-Management-Systemen die gewünschten Daten befinden. Die Metadaten sind also in erster Linie ein Verzeichnis der vorhandenen Datensätze im Data Lake. Auf weitere Aufgaben des Metadaten-Managements gehen wir im folgenden Abschnitt ein.

5.3 Datenmodellierung und Metadaten-Management

Datenmodellierung wird vor allem bei der klassischen Datenbank- und Anwendungsentwicklung eingesetzt und scheint im Kontext von Big Data weniger populär zu sein. Bei der klassischen Datenmodellierung folgt man in der Regel einem Top-Down-Ansatz, d. h. man fängt mit einem konzeptuellen Datenmodell an, verfeinert dies in ein logisches Datenmodell (zum Beispiel ein relationales Schema) und setzt es dann in einem Datenbank-System als physisches Modell um, indem man Indexstrukturen und andere Zugriffsstrukturen definiert.

Im Allgemeinen ist das Ziel der Modellierung die Erstellung eines Modells, das die Grundregeln oder wichtigsten Elemente eines komplexen Sachverhalts beschreibt. Modelle stellen eine vereinfachte Abbildung der Wirklichkeit dar, wobei vor allem die Elemente abgebildet werden, die für die geplante Anwendung des Modells relevant sind. Bei der Datenmodellierung erreicht man die Vereinfachung in der Regel durch Abstraktion, d. h. Details werden ausgeblendet oder verallgemeinert. Zum Beispiel werden Personen mit Name und Geburtsdatum modelliert, Details wie Haarfarbe oder Körpergröße werden weggelassen, wenn sie nicht für die Anwendung notwendig sind. Durch das Datenmodell werden die Strukturen vorgegeben, mit denen die Objekte der realen Welt in einer Anwendung beschrieben werden sollen.

Da es bei Data Science meist um die Analyse von existierenden Datensätzen geht und bei Big Data bereits viele Datensätze mit gegebenen Strukturen verfügbar sind, hat sich die Rolle der Datenmodellierung geändert. Es geht weniger darum, ein Schema für Daten zu definieren, die in Zukunft in einer Anwendung erfasst werden sollen, sondern vielmehr um die Beschreibung der Strukturen, Zusammenhänge und Regeln für existierende Datensätze. Dies ist ein Bottom-Up-Vorgehen, bei dem man von den

existierenden Datensätzen ausgeht, logische Datenmodelle ableitet und schließlich die semantischen Zusammenhänge in einem konzeptuellen Datenmodell beschreibt. Die Erstellung des logischen Datenmodells wird sehr gut durch verschiedene Werkzeuge unterstützt. Zum Beispiel können aus relationalen Datenbank-Systemen die Schemata direkt ausgelesen werden, aber auch für semi-strukturierte Daten wie JSON oder XML kann man anhand von gegebenen Datensätzen ein Schema ableiten. Solche Methoden zur Schemaextraktion werden von den meisten Datenintegrationswerkzeugen unterstützt.

Darüber hinaus kann mit Data Profiling (Abedjan et al. 2018) weitere Detailinformationen über die Datensätze gewinnen. Bei der Schemaextraktion reichen die Angaben aus, die für die zum Beispiel für die Erstellung eines relationalen Schemas erforderlich sind. Beispielsweise werden für Attribute Datentypen wie Integer und String erkannt. Mit Data Profiling kann man genauere Wertebereiche oder Muster in den Datensätzen erkennen, zum Beispiel dass die Spalte „Alter“ nur Integer-Werte von 0 bis 120 enthält oder dass in der Spalte „Datum“ die Zeichenketten dem Muster „DD.MM.YYYY“ entsprechen. Da man mit Data Profiling auch sehr leicht Ausreißer und fehlerhafte Daten erkennen kann, ist es auch relevant für die Datenaufbereitung (siehe Abschn. 5.4).

Für das Data Engineering ist es wichtig, eine Beschreibung der Datensätze (in Dateien, Datenbanken oder sonstigen Quellsystemen) in Form von Datenmodellen zu haben. Ein Datenmodell sollte zumindest die Strukturen, Verknüpfungen und Regeln bzw. Einschränkungen (Constraints) der Daten beschreiben. Modellierungssprachen wie die „Data Definition Language“ (DDL) von SQL, XML Schema oder auch die Unified Modeling Language (UML) erfüllen diese Voraussetzungen und eignen sich daher für die Darstellung von logischen Datenmodellen. Aufgrund der objektorientierten Eigenschaften kann UML auch für die konzeptuelle Modellierung benutzt werden, wobei man dann weniger auf die exakte Definition von Datentypen achtet, sondern mehr auf die semantischen Beziehungen zwischen Datenobjekten.

Neben den Datenmodellen sollten aber noch weitere Metadaten zu den Datensätzen erfasst werden. Unabhängig davon welche Architektur für das Daten-Management gewählt und in welchen Systemen die Datensätze verwaltet werden, sollte ein Metadaten-Management-System genutzt werden, um eine Übersicht der vorhandenen Datensätze zu bekommen. Google hat ein solches Verzeichnis zunächst für die firmeninternen Datensätze entwickelt (Halevy et al. 2016), mittlerweile gibt es die „Google Dataset Search“ für im Internet offen verfügbare Datensätze (<https://datasetsearch.research.google.com/>). Neben dem Datenformat kann man dort auch nach Datensätze mit Nutzungsrechten oder Themengebieten suchen.

Der Aufbau und die Verwaltung eines solchen Verzeichnisses innerhalb eines Unternehmens ist Teil der Data Governance. In einem Data-Lake-System sollte ebenfalls ein solches Verzeichnis vorhanden sein. Dabei muss auch festgelegt werden, welche Metadaten für einen Datensatz erfasst werden sollten. Dazu gehören zum Beispiel folgende Punkte:

- *Inhalt:* Schlagworte, Themen, Beschreibung
- *Herkunft:* Quellsystem, Kontext der Datenerfassung (z. B. Ort, Zeit)
- *Datenqualität:* Messwerte für Datenqualitätseigenschaften (z. B. Vollständigkeit)
- *Kontakt:* Ansprechpartner für die Datenquelle, Verantwortliche
- *Verfügbarkeit:* Zugriffsmöglichkeiten, Lizenzinformationen, Nutzungseinschränkungen

Ein entsprechendes Metadaten-Modell kann individuell für die Bedürfnisse des Unternehmens erstellt werden, jedoch bietet es sich an, auf bereits entwickelte, ausgereifte Modelle zurückzugreifen. Das World-Wide-Web-Consortium (W3C) hat zum Beispiel das Data Catalog Vocabulary (DCAT) definiert, ein Metadaten-Modell in Form eines RDF-Vokabulars zur Beschreibung von Datensätzen. Auch bei Datenintegrationswerkzeugen kann man üblicherweise bei der Erfassung von Datenquellen weitere Metadaten angeben.

5.4 Datenaufbereitung und Datenintegration

Die Aufbereitung und Integration von Daten, sodass sie für bestimmte Anwendungen oder auch eine Datenanalyse nutzbar sind, ist ein Schwerpunkt des Data Engineering. Da die Daten und die darauf aufbauenden Anwendungen sehr heterogen sind, ist es schwierig für diesen Prozess ein allgemeingültiges Vorgehen oder generelle Methoden zu entwickeln. Trotz der Fortschritte in den letzten Jahren beim Daten-Management in verschiedenen Bereichen der Datenaufbereitung und Datenintegration, bleiben diese Schritte ein arbeitsintensiver Prozess. Auch in absehbarer Zukunft braucht man menschliche Intelligenz, um die Heterogenität zwischen den Datensätzen aufzulösen, um Verknüpfungen zwischen den Datensätzen zu erkennen und um Regeln für die Transformation zwischen verschiedenen Datensätzen zu definieren.

Das konkrete Vorgehen in einem Datenaufbereitungs- und -integrationsprojekt ist also immer vom vorliegenden Anwendungsfall abhängig. Nichtsdestotrotz kann man einige Schritte identifizieren, die in den meisten Projekten erforderlich sind. Abb. 5.2 gibt einen Überblick über diese Schritte, die sich in eine Daten-Ebene und Schema-Ebene unterteilen. Im Folgenden werden wir die einzelnen Schritte genauer diskutieren.

Exploration & Profiling Bevor mit der eigentlichen Aufbereitung der Daten begonnen werden kann, ist eine Untersuchung der vorhandenen Daten erforderlich. Das heißt, es müssen die Strukturen, Zusammenhänge und Regeln erkannt werden, die in den Datensätzen vorhanden sind und für die Zusammenführung der Daten relevant sind. Hierbei geht es nicht um eine tiefergehende Analyse der Daten mit Machine Learning (das ist eine spätere Aufgabe für die eigentliche Datenanalyse), sondern es geht um die Zusammenhänge, die man auch in einem Datenmodell abbilden würde (zum Beispiel Schlüssel- und Fremdschlüsseleinschränkungen). Die Methoden, die man in diesem

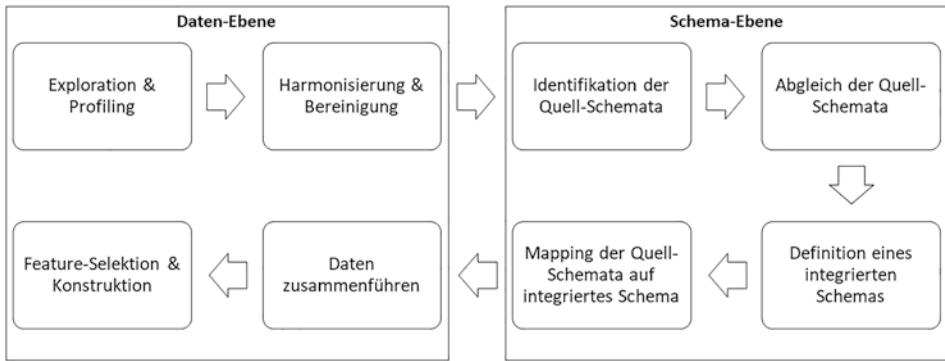


Abb. 5.2 Schritte während der Datenaufbereitung und -integration

Kontext einsetzt, sind also identisch mit den in Abschn. 5.3 besprochenen Methoden zum Reverse Engineering von Datenmodellen. Wenn ein gutes Metadaten-Management vorhanden ist, sind die meisten der benötigten Informationen schon vorhanden und müssen an dieser Stelle nicht separat ermittelt werden. Das zeigt auch einen Vorteil eines gut organisierten Metadaten-Managements für Datensätze: die Untersuchung der Datensätze wird nicht wiederholt für jedes Datenintegrationsproblem durchgeführt, sondern das Wissen kann im Metadaten-Management-System abgelegt werden und bei Bedarf abgerufen werden. Dieser Schritt sollte als Ergebnis ein umfassendes Verständnis der Datensätze haben, d. h. die Strukturen, die Wertebereiche und -verteilungen und eventuell vorhandene Datenqualitätsprobleme sollten bekannt sein. Im Gegensatz zur Datenmodellierung oder dem nachfolgenden Schritt der Identifikation der Quellschemata steht hier die Arbeit mit den Daten im Vordergrund.

Harmonisierung & Bereinigung In diesem Schritt sollen die ersten groben Probleme in den Datensätzen behoben werden. Hierzu gehört zum Beispiel die Überführung aller Datensätze in ein einheitliches syntaktisches Format, zum Beispiel Verwendung der Dateiformate CSV und JSON anstelle von Excel- oder anderen proprietären Formaten. Dabei sollte auch die Repräsentation von Daten angepasst werden, zum Beispiel Datumsformate einheitlich entsprechend dem Muster „DD.MM.YYYY“ oder einheitliche Zahlenformate. Dabei fallen in der Regel auch Fehler in den Daten auf, die eine Sonderbehandlung erfordern. Dafür sollte aber zunächst geklärt werden, ob es sich wirklich um Fehler handelt, oder ob nicht eine besondere Information in der uneinheitlichen Darstellung verborgen ist. Ein Unternehmen nutzte zum Beispiel zwei Zahlenformate (4,31 und 4.31) um darzustellen, dass es sich in dem einen Fall um die ursprünglichen Messwerte handelt und im anderen Fall um einen gerundeten, bereinigten Wert. Ähnliches gilt auch bei der Behandlung von Ausreißerwerten: ist es wirklich ein Messfehler, der ignoriert werden sollte, oder steckt gerade in dem Ausreißer die wertvolle Information, die auf ein Problem im Produktionsprozess hindeutet?

Im Kontext der Harmonisierung spielt auch ein gutes Stammdaten-Management eine wichtige Rolle. Wenn es für Materialien, Produkte, Kunden, Lieferanten usw. unterschiedliche Repräsentationsformen gibt, müssen diese vereinheitlicht werden. Dies sollte mithilfe eines Referenzdatensatzes erfolgen, der entweder im unternehmensinternen Stammdatensystem gepflegt wird oder durch eine externe Quelle bereitgestellt wird. Ein Werkzeug zur Harmonisierung und Bereinigung von Datensätzen ist zum Beispiel OpenRefine. Dort können Referenzdaten von Wikidata eingebunden werden, die sich sehr gut eignen, um Namen von Organisationen zu vereinheitlichen.

Die Datenbereinigung (Data Cleaning) ist ein Teil des Datenqualitätsmanagements und hat die Verbesserung der Datenqualität als primäres Ziel. Datenbereinigung ist aber ein reaktives Datenqualitätsmanagement, d. h. die Fehler sind in den Daten schon vorhanden, und man kann mit verschiedenen Datenbereinigungsmethoden nur versuchen, diese Fehler zu beseitigen. Im Gegensatz dazu steht das proaktive Datenqualitätsmanagement, dass durch die Definition von Qualitätszielen und -metriken eine kontinuierliche Überwachung der Datenqualität im Unternehmen erlaubt. Um eine ausreichende Datenqualität sicherzustellen, wird durch Analyse der gesamten Datenverarbeitungsprozesse versucht, möglichst früh in der Prozesskette eine gute Datenqualität zu erreichen. Dies könnte zum Beispiel dadurch erreicht werden, dass bereits bei der Datenerfassung (in einem Dateneingabeformular) bereits entsprechende Überprüfungen durchgeführt werden und fehlerhafte Daten erst gar nicht aufgenommen werden können.

Identifikation der Quell-Schemata Auch in diesem Schritt können sich die Investitionen in Metadaten-Management und einer Erfassung der Datenmodelle der Datenquellen bezahlt machen. Wie schon beim Schritt Exploration & Profiling dargestellt, fokussiert der vorherige Schritt auf das Erkennen von Mustern und Abhängigkeiten in den vorliegenden Datensätzen. In diesem Schritt liegt der Schwerpunkt auf der Arbeit mit den Datenmodellen, d. h. der Verfeinerung der erkannten Regeln und Strukturen, so dass ein valides Datenmodell entsteht. Das Datenmodell (zum Beispiel als relationales Schema oder XML Schema) ist erforderlich, da die folgenden Schritte auf Schema-Ebene erfolgen. Das Ergebnis dieses Schritts sind detaillierte Datenmodelle der Datenquellen, die im nächsten Schritt zum Abgleich der Datenquellen genutzt werden können.

Abgleich der Quell-Schemata Um Daten aus verschiedenen Datenquellen zusammenführen zu können, muss man zunächst die Ähnlichkeiten der Schemata untersuchen. In den Schemata bzw. Datenmodellen sollte daher das zuvor gesammelte Wissen über die Strukturen, Wertebereiche und Abhängigkeiten der Daten vorhanden sein. An dieser Stelle sollen noch keine genauen Transformationsregeln definiert werden, es geht zunächst um das Erfassen von Korrespondenzen zwischen den verschiedenen Datenquellen, zum Beispiel, dass das Attribut „ZIP“ in einer Datenquelle dem Attribut „PLZ“ in der anderen Datenquelle entspricht.

Eine Unterstützung durch intelligente Werkzeuge, sogenannte Schema Matching Tools (Bellahsene et al. 2011), scheint für diese Aufgabe vielversprechend zu sein, jedoch kann der Schritt nicht vollständig automatisiert werden. Vor allem in den 2000er Jahren war Schema Matching ein sehr aktiv bearbeiteter Forschungsbereich. Ähnliche Technologien werden beim Ontology Matching (Ochieng und Kyanda 2018) angewendet, da man Ontologien auch als Datenmodelle sehen kann. Zwar hat es im Laufe der Zeit einige Fortschritte bei den Matching-Werkzeugen gegeben, bei realen Integrationsaufgaben erreichen diese aber oft nur eine Genauigkeit von 60–70 %. Das heißt, dass eine Reihe der gefundenen Korrespondenzen nicht korrekt sind und daher nachträglich manuell entfernt werden müssten. Andererseits werden nicht alle vorhandenen Korrespondenzen durch die Algorithmen erkannt, so dass man noch individuell weitere Korrespondenzen. Da man sich nicht auf die Vollständigkeit des Ergebnisses eines Schema-Matching-Werkzeugs verlassen kann, ist also immer manuelle Nacharbeit erforderlich. Die Komplexität und Heterogenität von Datenmodellen sind groß, um sie vollständig durch automatisierte Werkzeuge aufzulösen.

Durch die Schema-Matching-Werkzeuge werden verschiedene Eigenschaften der Schemata bzw. der Datensätze analysiert, die auch ein Mensch beim Abgleich der Schemata nutzen würde. Zu den Eigenschaften gehören beispielsweise:

- *Bezeichnungen der Schema-Elemente:* Neben dem einfachen Vergleich der Zeichenketten (zum Beispiel „ProduktNr“ und „ProductNo“) können auch Methoden eingesetzt werden, die mit Thesauri oder Wörterbüchern Ähnlichkeiten zwischen den Begriffen erkennen können (zum Beispiel „Mitarbeiter“ und „Angestellter“). Allerdings werden in vielen Schemata nur Kurzbezeichnungen oder Abkürzungen benutzt, so dass diese Methode schnell an ihre Grenzen stößt. Nichtsdestotrotz ist der Namensvergleich eine grundlegende Methode für das Schema Matching.
- *Datentypen:* Ein Vergleich zwischen Datentypen kann nicht wirklich zum Erkennen von Ähnlichkeiten benutzt werden, sondern dient mehr zum Ausschluss von bereits erkannten Korrespondenzen. Wenn ein Attribut den Datentyp „Float“ hat, und ein anderes Attribut den Datentyp „Date“, dann ist eine Übereinstimmung unwahrscheinlich.
- *Wertebereiche:* Beim Profiling kann man zum Beispiel auch Histogramme für die Wertverteilung eines Attributs analysieren. Wenn die Histogramme von zwei Attributen die gleiche Wertverteilung aufweisen (zum Beispiel, weil beide das Alter von Personen darstellen), dann ist eine Ähnlichkeit wahrscheinlich.
- *Struktur:* Eine wichtige Information ist die Strukturähnlichkeit. Man kann die Schemata als Graph- oder Baumstrukturen betrachten und dann aus der Ähnlichkeit von zwei Elementen die Ähnlichkeit von benachbarten Elementen ableiten. Zum Beispiel kann man durch Vergleich der Namen feststellen, dass die Relationen „Address“ und „Adresse“ ähnlich sind, dann kann man daraus schlussfolgern, dass die Attribute diese Relationen auch Übereinstimmungen haben sollten.

- *Referenzmodelle:* Über den Abgleich mit Referenzmodellen oder bereits vorhandenen, validierten Korrespondenzen mit anderen Schemata könnte man durch logisches Schlussfolgern oder den Einsatz von Machine Learning auch weitere Korrespondenzen lernen. Deep Learning wurde erfolgreich für komplexe Aufgaben eingesetzt, die man zuvor nicht mit anderen Machine-Learning-Algorithmen lösen konnte (zum Beispiel Bilderkennung). Beim Schema Matching fehlt allerdings die große Datenmenge mit Trainingsdaten, die zum Trainieren von Deep-Learning-Algorithmen eingesetzt werden könnten. Daher haben entsprechende Ansätze bisher noch keine guten Ergebnisse erzielt.

Es gibt also verschiedene Ansätze, den Abgleich der Schemata zu unterstützen, aber letztlich muss das Ergebnis immer durch einen Menschen überprüft werden, zumindest in den Fällen, wo eine hohe Genauigkeit und Ergebnisqualität erforderlich ist. Sind Ungenauigkeiten und Fehler erlaubt, weil sie eventuell sowieso in der Masse der Daten untergehen, kann man auch mit den Schema-Matching-Methoden ausreichende Ergebnisse erzielen.

Definition eines integrierten Schemas Das integrierte Schema kann man quellenorientiert oder anwendungsorientiert definieren. Quellenorientiert heißt, dass man quasi die Vereinigungsmenge der Quellschemata als Basis nimmt. Dabei berücksichtigt man die im vorherigen Schritt erkannten Korrespondenzen, um ähnliche Elemente zusammenzuführen und Redundanzen im integrierten Schema zu entfernen. Auch dieser Prozess kann durch Werkzeuge unterstützt werden, muss aber auch wieder letztlich durch manuelle Arbeit ergänzt oder überprüft werden.

Man kann auch umgekehrt vorgehen und zunächst unabhängig von den Datenquellen ein (integriertes) Datenmodell für die geplante Anwendung bzw. Datenanalyse definieren. Das würde der im Abschn. 5.3 besprochenen Top-Down-Datenmodellierung entsprechen, da man sich nicht an vorhandenen Datensätzen orientiert, sondern von den Anforderungen der Anwendung ausgeht. Der Vorteil dieses Vorgehens ist, dass man ein für die Anwendung besser passendes Datenmodell erhält und ein besseres Verständnis darüber hat, was eigentlich für die Anwendung erforderlich ist. Daraus kann man auch ableiten, welche Informationen in den Datenquellen nicht vorhanden sind und eventuell durch neue Datenerfassungsmethoden oder externe Datenquellen beschafft werden sollten.

Auch bei einem quellenorientierten Vorgehen sollte der integrierte Datenbestand immer mit dem Bezug zu einer konkreten Anwendung aufgebaut werden. Data-Warehouse-Systeme haben einen größeren Anwendungsbereich, daher ist in diesem Kontext ein generelles Schema das Ziel der Schemaintegration. In der Data-Warehouse-Architektur gibt es aber auch Data Marts, die die Daten für eine bestimmte Anwendung bereitstellen. In einem Data-Lake-System sollte daher die Erstellung von anwendungspezifischen Data Marts angestrebt werden, da die Erstellung eines integrierten Schemas für alle Daten im Data-Lake-System nicht erfolgsversprechend ist.

Mapping der Quell-Schemata auf integriertes Schema In diesem Schritt müssen die Transformations- und Integrationsregeln definiert werden, die die Daten aus den Datenquellen extrahieren und in einen einheitlichen Datenbestand überführen. Diese Mappings kann man in der Regel als Anfragen an die Datenquellen definieren (sofern es sich um Datenbank-Management-Systeme mit einer Anfrageschnittstelle handelt) oder über ein Datenintegrationswerkzeug wie von Informatica oder Talend realisieren. Die notwendigen Vorbereitungen und Vorerlegungen zur Definition der Mappings wurden in den vorhergehenden Schritten schon getroffen und können daher diesen Schritt erleichtern.

Datenintegrationswerkzeuge haben ihre Stärke darin, Konnektoren für viele Daten-Management-Systeme und Dateiformate bereitzustellen, mit denen die Daten aus diesen Datenquellen extrahiert werden können. Gleichzeitig bieten diese Werkzeuge auch Funktionen zur Datentransformation und zur Zusammenführung der Daten an, mit denen die heterogenen Daten aus den Quellen in ein einheitliches System überführt werden können. Da sich die Datenaufbereitung und -integration wie hier dargestellt aus mehreren Schritten zusammensetzt, unterstützen diese Werkzeuge auch die Umsetzung der Schritte in einem zusammenhängenden Prozess.

Die Definition der Mappings findet wie das Formulieren einer SQL-Anfrage zunächst auf der Schemaebene statt. Der nächste Schritt beschäftigt sich mit dem Zusammenführen von Daten.

Daten zusammenführen Während die vier vorherigen Schritte auf der Schemaebene durchgeführt wurden und sich nicht mit der konkreten Datenlage beschäftigten, geht es nun um die konkrete Zusammenführung von Datensätzen. In der Fachliteratur wird dieser Schritt als Record Linkage bezeichnetnet (Koudas et al. 2006), d. h. es geht darum zu erkennen, welche Datensätze aus den verschiedenen Datenquellen dem gleichen Objekt entsprechen. Zum Beispiel müssen bei der Zusammenführung von Kundendaten unterschiedliche Schreibweisen bei Namen und Adressdaten berücksichtigt werden. Für dieses Problem gibt es auch eine Fülle von Methoden. Es kommt auf die besonderen Anforderungen im Einzelfall an, um eine passende Methode auswählen zu können.

Feature-Selektion und -Konstruktion Die Auswahl von geeigneten Features (bzw. Merkmalen oder Attributen) für eine Datenanalyse ist der letzte Schritt der Datenaufbereitung und -integration. Dieser Schritt kann auch schon als Teil der Datenanalyse angesehen werden, da die Auswahl der Features stark von der gewählten Datenanalysemethode abhängig ist. Verfahren wie Support Vector Machines sind für hoch-dimensionale, numerische Daten sehr gut geeignet, während Entscheidungsbäume eher für diskrete Daten mit wenigen Dimensionen geeignet sind. Beim Einsatz von Deep Learning kann man eventuell sogar die Rohdaten direkt verwenden. Generell sollte man aber bei der Wahl und Zusammenstellung der Features für eine Datenanalyse die folgenden Grundregeln beachten:

- *Redundante und irrelevante Daten vermeiden:* Will man eine Klassifikation durchführen und die Datensätze in unterschiedliche Klassen unterteilen, so sollten die Daten auch Merkmale enthalten, mit denen man die Klassen unterscheiden bzw. Klassen bilden kann. Ein Schlüssel (eindeutiger Wert für alle Datensätze) ist zum Beispiel nicht hilfreich, da alle Datensätze unterschiedliche Werte haben und man daraus keine Gemeinsamkeiten für eine Klasse ableiten kann. Umgekehrt sind Merkmalen mit den gleichen Werten für alle Datensätze auch nicht hilfreich. Solche Merkmale sollten entfernt werden, um die Komplexität bei der Datenanalyse zu verringern.
- *Aussagekräftige Features wählen:* Die Zusammenhänge zwischen den Daten sollten direkt erkennbar und keine Berechnungen erfordern. Wenn man zum Beispiel Grundstücke in verschiedene Klassen unterteilen will, sollte man nicht Werte wie Länge, Breite oder GPS-Koordinaten wählen, sondern besser Grundstücksgröße oder Lage.
- *Aussagekräftige Repräsentation wählen:* Wie oben bereits erwähnt, kommt es auf den Algorithmus an, ob eine numerische oder kategoriale Darstellung von Merkmalen günstiger ist. Das Merkmal Körpergröße könnte man als numerischen Wert angeben oder nach einem bestimmten Schema in Kategorien (z. B. klein, mittel, groß) aufteilen.
- *Vergleichbare Skalen bei numerischen Daten:* Insbesondere bei neuronalen Netzen oder Deep Learning ist es wichtig, dass die numerischen Werte vorher auf eine einheitliche Skala normalisiert werden (zum Beispiel 0 bis 1 oder -1 bis 1), um den Einfluss von Merkmalen mit sehr großen numerischen Werten zu reduzieren. Auch bei anderen Verfahren, die mit numerischen Vektordaten arbeiten (zum Beispiel Support Vector Machines) sollte diese Normalisierung durchgeführt werden.
- *Datenmenge durch Aggregation bzw. Generalisierung reduzieren:* Wenn sehr viele einzelne Datensätze zur Verfügung stehen, ist die Datenanalyse komplexer und aufwendiger. Je nach Ziel der Datenanalyse ist eine Reduktion der Datenmenge durch Aggregation oder Generalisierung möglich. Bei Verkaufsdaten betrachtet man zum Beispiel nicht jede einzelne Verkaufstransaktion, sondern alle Transaktionen in Zeitraum (Tag, Woche, Monat, ...) oder in einer geografischen Region.

Die dargestellten acht Schritte stellen eine Zusammenfassung der wichtigsten Aspekte bei der Datenaufbereitung und -integration dar. Im Einzelfall müssen nicht alle Schritte durchlaufen werden oder auch eine andere Reihenfolge ist möglich. In Forschung und Praxis werden aber die oben aufgeführten Aktivitäten immer wieder bei Datenintegrationsprojekten durchgeführt.

5.5 Datenbank-Management-Systeme: SQL, NoSQL und Big Data

Da das Data Engineering sich mit den technischen Aspekten des Daten-Managements beschäftigt sind tiefergehende Kenntnisse über Daten-Management-Systeme unerlässlich. Zunächst wollen wir die Begriffe Datenbank-Management-System und

Daten-Management-System klären. Unter einem Datenbank-Management-System verstehen wir ein Software-System, dass über eine Schnittstelle mit einer definierten Sprache, das Anlegen, Löschen, Ändern und Abfragen von Daten erlaubt (Geisler und Quix 2020). Dabei sollten auch Transaktionen unterstützt werden und eine Persistenz der Daten gewährleistet werden, d. h. wenn das System die Rückmeldung gegeben hat, dass Daten erfolgreich geschrieben wurden, dann sollten die Daten auch noch einem Absturz verfügbar sein. Die klassischen relationalen Datenbank-Management-Systeme und die meisten NoSQL-Systeme erfüllen diese Eigenschaften. Bei einigen Big-Data-Systemen sehen wir nicht alle dieser gewünschten Eigenschaften und erfassen sie deshalb unter dem Oberbegriff Daten-Management-Systeme. Hadoop zum Beispiel unterstützt keine definierte Sprache zum Anlegen oder Abfragen von Daten, nur mit zusätzlichen Software-Paketen wie Hive kann man Daten über SQL abfragen. Andererseits bieten Systeme wie Apache Spark zwar eine komplexe Funktionalität zum Anlegen und Abfragen von Datensätzen in verschiedenen Datenformaten, realisieren aber selbst keine Datenspeicherung.

Im Rahmen dieses Kapitels nutzen wir diese Unterscheidung zur Abgrenzung zwischen den beiden Kapiteln Data Engineering und Big-Data-Technologien. Da Data Engineering sich auch mit der längerfristigen Bereitstellung und Verwaltung von Daten beschäftigt, besprechen wir in diesem Abschnitt Datenbank-Management-Systeme. Für das Data Engineering sind natürlich auch Big-Data-Systeme relevant, aufgrund der generellen Bedeutung von Big-Data-Technologien für das Thema Data Science widmen wir diesen Systemen aber ein separates Kapitel.

Zurück zu den Datenbank-Management-Systemen: zunächst kann man diese Systeme in die klassischen relationalen und neueren NoSQL-Systeme unterscheiden. Die relationalen Datenbank-Management-Systeme haben eine lange Entwicklungshistorie seit den 1970er Jahren und können auf den weitverbreiteten und akzeptierten Standard SQL als Datenbanksprache aufbauen. Datenintegrität, Konsistenz und Transaktions-sicherheit im Mehrbenutzerbetrieb sind nur einige der bevorzugten Eigenschaften, aufgrund derer unternehmenskritische Anwendungen oft relationale Systeme nutzen. Im Laufe der Jahre wurden aber auch andere Datenmodelle für Datenbank-Systeme vorgeschlagen, zum Beispiel objekt-orientierte Datenmodelle oder XML in den 1990er Jahren. Auch wenn diese Modelle einige Vorteile gegenüber dem relationalen Modell hatten, konnten sich diese Modelle nicht durchsetzen. Letztlich haben sie aber dazu geführt, dass die Funktionalität der relationalen Systeme erweitert wurde, zum Beispiel um objekt-relationale Funktionen oder einen XML-Datentyp mit entsprechenden Abfragemöglichkeiten.

Ein Problem konnte allerdings nicht von den relationalen Datenbank-Management-Systemen gelöst werden: kostengünstige Skalierbarkeit und Fehlertoleranz bei stark verteilten Anwendungen. Diese Anforderungen gewannen mit der steigenden Popularität von großen Internet-Plattformen wie Amazon, Ebay oder Google anfangs der 2000er Jahre an Bedeutung. Für verteilte Anwendungen wurde auch das CAP-Theorem definiert (Brewer 2000). CAP steht für drei Anforderungen an ein verteiltes

Daten-Management-System: Konsistenz (Consistency), Verfügbarkeit (Availability) und Toleranz bei Netzwerk-Partitionierungen. Das Theorem besagt, dass nur zwei von drei Anforderungen erfüllt werden konnten. Die klassischen relationalen Systeme mit keiner oder nur eingeschränkter Verteilung fokussierten auf Konsistenz und Verfügbarkeit und waren aufgrund dessen nicht für stark verteilte Anwendungen geeignet. Bei einem verteilten System muss man aber fehlertolerant gegenüber Netzwerkunterbrechungen sein, da man in einem verteilten System nicht vermeiden kann, dass einzelne Server oder Teile des Netzwerks nicht erreichbar sind.

Ein weiteres Problem war die Inkompatibilität der Datenmodelle: in relationalen Datenbank-Systemen arbeitet man mit normalisierten Relationen, deren Inhalte man zwar über Join-Abfragen einfach zusammenführen kann, die Änderung der Daten über mehrere Relationen erfordert aber eine komplexe Anwendungslogik und Nutzung von Transaktionen. Andererseits nutzt man in Web-Anwendungen objekt-orientierte oder andere verschachtelte Datenstrukturen (z. B. JSON) und arbeitet mit komplexen Objekten, die zum Beispiel Daten einer Web-Session oder eines Nutzerprofils repräsentieren.

Software-Entwickler begannen daher für ihre Anwendungen besser geeignete Daten(bank)-Management-Systeme zu entwickeln, die einerseits ein für die Anwendungen besser passendes Datenmodell als SQL hatten, andererseits direkt ein verteiltes Daten-Management über mehrere Server-Knoten unterstützten. Aus diesem Grund waren die Systeme auch fehlertolerant gegenüber Netzwerkpartitionierungen, und hatten nach dem CAP-Theorem die Wahl, die Anforderung nach jederzeit konsistenten Datenbeständen oder nach Verfügbarkeit zu erfüllen. Da eine Internet-Anwendung jederzeit verfügbar sein sollte, entschieden sich die meisten Systeme für Verfügbarkeit und erlauben daher zwischenzeitlich inkonsistente Datenbestände. Diese Art der relaxierten Konsistenzgarantie wird auch Eventual Consistency genannt. Da gleichzeitig auch Open-Source-Software populär wurde und sich ein großes Interesse für solche Plattformen bildete, wurden diese Systeme auch kostenlos zur Verfügung gestellt. Mittlerweile gibt es die meisten Systeme weiterhin in einer kostenlosen Open-Source-Variante, für weitere Funktionalität für unternehmenskritische Anwendungen stellen die Unternehmen, die sich um die NoSQL-Systeme gebildet haben, kostenpflichtige „Enterprise Editions“ zur Verfügung. Erst einige Jahre später wurde für diese Systeme der Begriff „NoSQL“ eingeführt, der allerdings nur eine Eigenschaft betonte: die neuen Datenbank-Management-Systeme unterstützten nicht mehr SQL als primäre Abfragesprache.

Grundlegend kann man bei den NoSQL-Datenbank-Management-Systemen vier Datenmodelle unterscheiden:

- *Key-Value*: Datenobjekte werden unter einem Schlüssel gespeichert. Der Zugriff auf die Datenobjekte ist über den Schlüssel oder auch einfache Abfragemechanismen möglich. Die Datenobjekte haben häufig eine baumartige Struktur, wie zum Beispiel ein JSON-Dokument.

- *Dokument-orientiert*: Auch hier werden die Daten als JSON-Dokumente abgelegt, das System unterstützt aber weitergehende Abfragemöglichkeit über die Struktur der Dokumente. MongoDB ist ein dokument-orientiertes Datenbank-Management-System, das derzeit das populärste NoSQL-System ist.
- *Wide Column*: Dieses Datenmodell ist dem relationalen Modell sehr ähnlich und bietet vergleichbare Abfragemöglichkeiten wie SQL, jedoch ist mehr Flexibilität bei den Spalten der Datensätze einer Tabelle möglich, d. h. nicht alle Datensätze müssen die gleiche Struktur haben.
- *Graph-orientiert*: Hierbei können Graphen mit komplexen Knoten und Kanten abgespeichert werden. Sowohl Knoten als auch Kanten können verschiedene Typen und Attribute haben. Abfragen können bestimmte mathematische Eigenschaften von Teilgraphen testen (z. B. Konnektivität, kürzeste Wege), sondern auch nach bestimmten Mustern im Graphen suchen.

Neben den bereits diskutierten Eigenschaften ist es bei den meisten Systemen nicht erforderlich, ein Schema für die Daten zu definieren. Das stellt auch einen wesentlichen Vorteil der NoSQL-Systeme dar, dass sie direkt genutzt werden können, ohne die mühselige Definition eines Schemas, die in relationalen Systemen erforderlich ist. Es sollte allerdings offensichtlich sein, dass man nicht ganz ohne eine Modellierung oder Strukturierung der Daten auskommt, da man die Daten in einer bestimmten Struktur ablegen und abfragen will. Im Unterschied zu relationalen Systemen ist nun aber einiges der Logik, die zur Überprüfung der Datenstruktur oder Integrität von neuen Daten notwendig ist, vom Datenbank-Management-System in die Applikation gewandert. Das erhöht die Komplexität der Anwendungen und den Implementierungsaufwand. Auf der anderen Seite hat man mehr Flexibilität und Skalierbarkeit durch den Einsatz der NoSQL-Systeme gewonnen. Welcher Teil schwerer wiegt, muss man für den jeweiligen Anwendungsfall entscheiden, aber man sollte sich bewusst sein, dass NoSQL-Datenbank-Management-Systeme nicht nur Vorteile haben.

5.6 Fazit

Hatte man in den 1990er Jahren den Eindruck, dass die relationalen Datenbank-Management-Systeme, insbesondere die kommerziell erfolgreichen Produkte, alle rivalisierenden Konzepte durch Marktmacht und schnelle Adaption von neuen Funktionalitäten verdrängen könnten, hatte sich dieses Bild in den 2000er Jahren geändert. Die Anforderungen für Internet-Anwendungen und Big Data konnten mit den existierenden Systemen nicht zufriedenstellend adressiert werden. Somit ist in den letzten zehn bis zwanzig Jahren eine große Vielfalt an verschiedenen Daten-Management-Systemen entstanden, die für spezielle Anforderungen eine Nischen-Lösung bereithält. Das führt zwar einerseits zu einer besseren Performance, aber andererseits zu einer größeren Heterogenität was den Begriff Polyglot Persistence geprägt hat.

Aktuell ist schwer abzuschätzen, welche Datenbank-Technologien in naher Zukunft sich neu oder weiter entwickeln werden. Graph-orientierte Datenbanksysteme existieren zwar schon, führen aber noch ein Nischendasein. Das Thema Wissensmanagement, die Verknüpfung von Informationseinheiten und damit die Verwaltung von Knowledge Graphs spielt für viele Unternehmen eine zunehmende Rolle. Daher ist eine Weiterentwicklung und steigende Popularität insbesondere bei den graph-orientierten Datenbank-Management-Systemen anzunehmen. Ein weiteres Thema ist die Verarbeitung von Datenströmen. Dies fällt zwar auch in die Kategorie von Big-Data-Systemen und wird daher im nächsten Kapitel detaillierter besprochen, aber Datenbank-Management-Systeme zur Verwaltung von Zeitserien (z. B. aus Sensordaten) bzw. zur Verarbeitung von Datenströmen in nahezu Echtzeit werden in naher Zukunft an Popularität gewinnen, da in immer mehr Anwendungen zeitnahe Reaktionen auf neue Daten gefordert sein werden (z. B. Industrie 4.0).

Für das Data Engineering bedeutet dies, dass die Systemlandschaften weiterhin an Komplexität gewinnen werden, und dass die Interoperabilität und Integration der Systeme und Daten auch künftig die wesentliche Herausforderung bleiben.

Literatur

- Abedjan, Z., Golab, L., Naumann, F., Papenbrock, T.: Data profiling. Synthesis Lectures on Data Management **10**(4), 1–154. Morgan & Claypool Publishers, Williston, VT, USA (2018)
- Bellahsene Z., Bonifati A., Rahm E.: Schema Matching and Mapping. Springer, Berlin (2011). DOI: <https://doi.org/10.1007/978-3-642-16518-4>
- Brewer E.A.: Towards robust distributed systems (abstract). In Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing. Portland (2000). DOI: <https://doi.org/10.1145/343477.343502>
- Brodie M.L.: Data Integration at Scale: From Relational Data Integration to Information Ecosystems In: Proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), S. 2–3. Perth, Australia, (2010) DOI: <https://doi.org/10.1109/AINA.2010.184>
- Council J.: Data Challenges Are Halting AI Projects, IBM Executive Says. Wall Street Journal. <https://www.wsj.com/articles/data-challenges-are-halting-ai-projects-ibm-executive-says-11559035800> (2019)
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Commun. ACM **39**(11), 27–34 (1996). <https://doi.org/10.1145/240455.240464>
- Geisler S., Quix C.: Database Management Systems (DBMS). In Schintler, L. A., McNeely, C. L. (Hrsg.). Encyclopedia of Big Data. Springer, Cham. (2020). DOI: https://doi.org/10.1007/978-3-319-32001-4_538-1
- Halevy A.Y., Korn F., Noy N.F., Olston C., Polyzotis N., Roy S., Whang S.E.: Goods: Organizing Google’s Datasets. Proceedings of the ACM SIGMOD International Conference on Management of Data S. 795–806. San Francisco (2016). DOI: <https://doi.org/10.1145/2882903.2903730>
- Heer, J., Hellerstein, J.M., Kandel, S.: Data Wrangling, In Encyclopedia of Big Data Technologies, Springer, Cham (2019). https://doi.org/10.1007/978-3-319-63962-8_9-1
- Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses, 2. Aufl. Springer, Berlin (2003)

- Koudas N., Sarawagi S., Srivastava D.: Record linkage: similarity measures and algorithms. In Proceedings of the ACM SIGMOD International Conference on Management of Data, S. 802–803. Chicago (2006). DOI: <https://doi.org/10.1145/1142473.1142599>
- Lohr S.: For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights. New York Times. <https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html> (2014)
- Ochieng P., Kyanda S.: Large-Scale Ontology Matching: State-of-the-Art Analysis. ACM Computing Surveys, 51(4):75:1–75:35. (2018) DOI: <https://doi.org/10.1145/3211871>
- Press G.: Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/> (2014)
- Quix C., Hai R.: Data Lake. In Sakr S., Zomaya A.Y. (Hrsg.). Encyclopedia of Big Data Technologies. Springer, Cham (2019). DOI: https://doi.org/10.1007/978-3-319-63962-8_7-1
- Sadalage P.J., Fowler M.: NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Pearson Education. Upper Saddle River, NJ, USA (2012)
- Schulz M., Neuhaus U.: DASC-PM v1.0 – Ein Vorgehensmodell für Data-Science-Projekte. Nordakademie. <https://www.nordakademie.de/forschung/data-science-process-model> (2020)
- Simitsis A., Vassiliadis P.: Extraction, Transformation, and Loading. In: Liu, L., Öszu, M.T. (Hrsg.). Encyclopedia of Database Systems, Bd. 2. Springer, New York, NY, USA, (2018). DOI: https://doi.org/10.1007/978-1-4614-8265-9_158
- Thusoo A., Shao Z., Anthony S., Borthakur D., Jain N., Sarma J.S., Murthy R., Liu H.: Data warehousing and analytics infrastructure at Facebook. In: Proceedings of the ACM SIGMOD International Conference on Management of Data S. 1013–1020. Indianapolis, USA. (2010). DOI: <https://doi.org/10.1145/1807167.1807278>
- Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Zhang N., Anthony S., Liu H., Murthy R.: Hive - a petabyte scale data warehouse using Hadoop. In Proceedings of the 26th International Conference on Data Engineering (ICDE 2010) S. 996–1005. Long Beach, California (2010). DOI: <https://doi.org/10.1109/ICDE.2010.5447738>
- Wirth R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 1 Aufl., S. 29–40. Springer, London. (2000)
- Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I.: Apache Spark: a unified engine for big data processing. Commun. ACM **59**(11), 56–65 (2016). <https://doi.org/10.1145/2934664>



Data Governance

6

Detlev Frick

Zusammenfassung

Die sogenannten „Megatrends“ Big Data und Industrie 4.0 bestimmen die aktuelle Forschung im Bereich der Data Governance. Die Anzahl der Datenquellen nimmt zu, wie auch die Menge der Daten und insbesondere deren unterschiedlicher Strukturierungsgrad. Die klassischen Werkzeuge für Datenhaltung und -verarbeitung und ein anforderungsgerechter Umgang mit den verfügbaren Datenmengen stößt an seine Grenzen. Die Manager in den Unternehmen verstehen Daten inzwischen als wertvolles Unternehmensgut und es wird dringend ein Ordnungsrahmen für das Datenmanagement in den Unternehmen benötigt.

Data Governance steht für ganzheitliches Management von Daten, die in einem Unternehmen oder einer Organisation verwendet werden. Es beinhaltet Richtlinien und Vorgehensweisen, um die Qualität, den Schutz und die Sicherheit der Daten zu gewährleisten und sorgt für die Einhaltung rechtlicher Vorgaben. Damit ist Data Governance für alle Mitarbeiter, die mit Daten zu tun haben, essentiell!

6.1 Einführung

6.1.1 Begriffliche Einordnung

Governance hat in den Unternehmen immer mehr an Bedeutung gewonnen. Governance wird oft übersetzt als Regierungs-, Amts- bzw. Unternehmensführung einer

D. Frick (✉)

Duisburg, Deutschland

E-Mail: detlev.frick@hs-niederrhein.de

politisch-gesellschaftlichen Einheit wie Staat, Verwaltung, Gemeinde, privater oder öffentlicher Organisation. Häufig wird es auch im Sinne von Steuerung oder Regelung einer Organisation (etwa einer Gesellschaft oder eines Betriebes) verwendet (vgl. (Governance, kein Datum)).

In einem Unternehmen spricht man also von Corporate Governance man den rechtlichen und faktischen Ordnungsrahmen für die Leitung und Überwachung von Unternehmen zum Wohlwollen aller relevanten Anspruchsgruppen meint. Der Ordnungsrahmen wird maßgeblich durch Gesetzgeber und Eigentümer bestimmt. Die konkrete Ausgestaltung obliegt dem Aufsichts- bzw. Verwaltungsrat und der Unternehmensführung (vgl. (Corporate Governance, kein Datum)).

Information und die Informationsverarbeitung ist für die Unternehmen von zentraler Bedeutung geworden. „*Data is the new oil*“ ist eine Aussage, die auf einen Artikel in The Economist im Jahre 2017 (vgl. (o.A. 2017)) zurück geht. Unter dem Titel „*The world's most valuable resource is no longer oil, but data*“ wird eine Regulierung der Internet Giganten wie Google, Facebook, usw. gefordert. Fakt ist, dass Informationen immer mehr zu einem eigenen Produktionsfaktor geworden sind und damit die Governance auch auf Informationen und IT auszuweiten ist. IT-Governance ist ein Ordnungsrahmen für ein effektives Management der IT (vgl. (Weill und Ross 2004).).

Die weitere Konsequenz dieser Entwicklung führt zur Data Governance. Hier geht in Abgrenzung zur IT-Governance um den Ordnungsrahmen für die digitalen Daten und Informationen in den Unternehmen und nicht um die IT(-Systemlandschaft).

Maria Villar, Theresa Kushner und Dave Wells haben eine recht aussagenkräftige Definition von Data Governance vorgenommen (vgl. (Villar et al. 2018)):

„Data governance is an emerging, cross-functional management program that treats data as an enterprise asset. It includes the collection of policies, standards, processes, people, and technology essential to managing critical data to a set of goals. Data governance also includes the oversight necessary to ensure compliance and to manage risk. A data governance program can be tailored to match an organization's culture, information maturity, priorities, and sponsorship.“

Im Projekt „*Data Economics and Management of Data driven business*“ (DEMAND), das aus einem Konsortium aus mehreren Instituten und Unternehmen besteht und die Entwicklung eines Ansatzes zur effizienten Datenbewirtschaftung in Unternehmen zum Ziel hat, wurde eine Definition von Data Governance entwickelt, die versucht, alle praxisrelevanten Aspekte der Datenbewirtschaftung zu berücksichtigen und zu einem Gesamtkonzept zu vereinheitlichen, das nicht nur für ein Unternehmen anwendbar, sondern auch in einem Datenökosystem im Rahmen einer Data Economy implementierbar ist (vgl. (DEMAND, kein Datum)):

Data Governance stellt das Rahmenwerk dar, welches die Grundlage für den Umgang mit und die Bewirtschaftung von Daten in einem Unternehmen für alle Stakeholder bildet.

Das Rahmenwerk beinhaltet sechs Dimensionen:

- Assets: Definition und Identifikation von Daten und deren ökonomischen Wert, Definition von einheitlichen, unternehmensübergreifenden Standards für die Datenbewertung und Einhaltung von Datenqualitätsstandards.
- Roles, Tasks & Responsibilities: Festlegung von Rollen für die Data Execution und von unternehmensübergreifenden Rollen für das Datenökosystem, Zuweisung von Zuständigkeiten für Daten und datengetriebene Prozesse.
- Processes: Überwachung der internen und unternehmensübergreifenden Datenprozesse, Überwachung des Teilens und der Nutzung von Daten, Entscheidungen über das Management und die Nutzung von Daten.
- Architecture & Tools: Unterstützung von Data Governance durch Technologie, Definition von Standards für die technische Umsetzung und die Auswahl der genutzten Tools für die Datenbewirtschaftung.
- Security: Definition von internen und unternehmensübergreifenden Standards zur Datensicherheit, Festlegung der Zugriffsrechte, Vorgehensweise bei Sicherheitsvorstößen
- Compliance: Sicherstellung der Einhaltung von internen/externen Anforderungen/Richtlinien an das Datenmanagement und den Datenschutz.

6.1.2 Datenstrategie

Der Einstiegspunkt in das „*Enterprise Data Governance*“ ist eine „*Enterprise Data Strategy*“. Für das Unternehmen muss eine eindeutige Vision erarbeitet werden, in der der Wert der Daten Aufgangspunkt der Betrachtung ist. Die Frage muss geklärt werden, ob es eine datengetriebene Wertschöpfung im Unternehmen gibt oder zumindest angestrebt wird. Gibt es also ein datengetriebenes Geschäftsmodell und wie muss die Datenwirtschaft im Unternehmen gestaltet werden?

Krotova und Eppelsheimer unterscheiden drei Perspektiven auf Data Governance: Die System-Perspektive, die Prozess-Perspektive und die Strategie-Perspektive (siehe Abb. 6.1). Aus der System-Perspektive definiert Data Governance die Regeln für die Datenarchitektur, also die technische Komponente der Datenbewirtschaftung im Unternehmen, womit sie an die IT-Governance anknüpft. Aus der Prozess-Perspektive beschreibt Data Governance die Rahmenbedingungen für operative datengetriebene Prozesse im Unternehmen (Data Execution). Diese beinhalten den kompletten iterativen Prozess des Data Managements, angefangen mit der Datenerhebung, und endend bei der Datenveräußerung oder Löschung. Aus der Strategie-Perspektive ist Data Governance ein Enabler der Datenstrategie. Im Allgemeinen stehen der Unternehmensführung zwei Handlungsoptionen bei der Strategiefindung zur Auswahl: Entweder kann der Fokus auf die Optimierung laufender Prozesse im Unternehmen oder auf die Entwicklung neuer

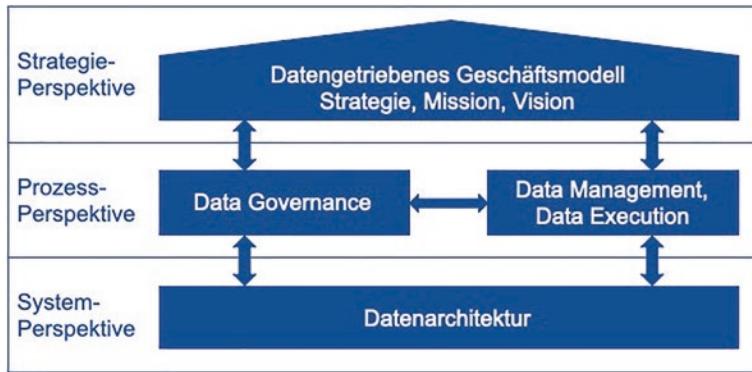


Abb. 6.1 Data Governance als Teil der Data Economy, in Anlehnung an Krotova und Eppelsheimer 2019, S. 9

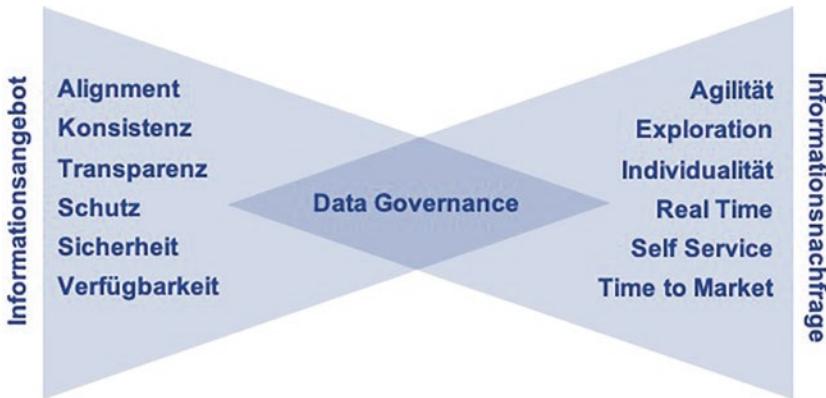


Abb. 6.2 Data Governance-Spannungsfeld, in Anlehnung an Gluchowski 2020a, S. 7

datengetriebener Geschäftsmodelle gelegt werden (vgl. (Krotova und Eppelsheimer 2019), S. 8).

Die Data Governance steht bei der Formulierung der Datenstrategie im Spannungsfeld zwischen den Datenanbietern (Informationsanbietern) im Unternehmen und den Datennachfragern (Informationsnachfragern), deren Ziele und Werte häufig stark voneinander abweichen (vgl. Abb. 6.2) von der Unternehmens-IT angeboten und die Fachabteilungen werden in der Regel die Daten nachfragen. Während bei den Anbietern Aspekte der Konsistenz, Transparenz und Verfügbarkeit unter gleichzeitiger Beachtung von den Anforderungen an Alignment, Datenschutz und Datensicherheit im Vordergrund stehen, wollen Datennutzer vor allem eine zeitnahe Informationsversorgung mit großer Flexibilität und Agilität sowie freien, individuellen Optionen zur selbstständigen



Abb. 6.3 Treiber für Data Governance, in Anlehnung an Iffert et al. 2013, S. 7

Exploration des Datenbestandes. Alle diese Wünsche können nicht gleichzeitig erfüllt. Es ist eine Aufgabe der Data Governance hier einen angemessenen Ausgleich herzustellen und den Ordnungsrahmen für die Beteiligten zu bilden (vgl. (Gluchowski 2020a), S. 7).

Iffert, Görlich und Grosser nennen einige wichtige Treiber für Data Governance (vgl. Abb. 6.3).

6.2 Data Governance Framework

Im Allgemeinen stellt ein Framework einen Rahmen, der die verschiedenen Themenbereiche behandelt und zusammenfasst, die im Data Governance zu regeln sind. Data Governance als eine zentrale Managementaufgabe in den Unternehmen muss Regelungen und Richtlinien in einer Reihe von Themenbereichen bereitstellen, um den Anwendern in den verschiedenen Unternehmensbereichen eine Orientierungsfunktion zu liefern. Welche Bereiche und Themen in einzelnen zu regeln gilt, ist sicherlich nicht für jedes Unternehmen in der gleichen Art und Weise zu beantworten.

Nachfolgend sollen zwei Varianten von Data Governance Frameworks kurz vorgestellt und diskutiert werden. Es handelt sich dabei um das Framework, dass sich an der Sichtweise von O’Neal orientiert ((O’Neal 2012), zitiert nach (Gluchowski 2020a), siehe Abb. 6.4). Weiterhin wird das Framework des „The Data Governance Institute (DGI)“ kurz vorgestellt (The Data Governance Institute (DGI) 2014).

6.2.1 Strategie

In der Strategie hat das Unternehmen seine Vision und seine Mission festzulegen, die mit dem Data Governance verfolgt werden sollen. Eine Vision ist ein Zukunftsentwurf und damit eine langfristige Zielvorstellung einer Unternehmung. Es handelt sich dabei nicht



Abb. 6.4 Data Governance Framework, in Anlehnung an O’Neal 2012, zitiert nach Gluchowski 2020a, S. 6

unbedingt um ein konkretes Ziel, sondern um eine unscharfe Vorstellung von dem, was man erreichen möchte. Die Vision ist aber notwendig, um daraus eine Mission und später auch Ziele ableiten zu können (Brecht 2012), S. 35).

Im Unterschied zur Vision bringt die Mission zum Ausdruck, welche Rolle das Data Governance generell im Unternehmen verfolgen soll. Mit der Mission soll das Selbstverständnis der Data Governance im Unternehmen deutlich werden. Entwickelt die Data Governance Regeln zum Beispiel gemeinsam mit den Beteiligten (im Sinne einer Daten-Demokratie) oder legt sie einfach diese nur fest (im Sinne einer Daten-Diktatur) (vgl. dazu (Dittmar und Fürber 2020), S. 20 ff.).

Ziele der Data Governance beschreiben konkrete, messbare Visionen. Die Ziele dienen dazu, die Vision erreichbar zu machen. Seiner beschreibt einige mögliche Ziele für das Data Governance (vgl. (Seiner 2012)):

Übergeordnetes Ziel:

- Maximierung des geschäftlichen Nutzens durch Datenverwendung bei gleichzeitiger Übereinstimmung mit den Unternehmenszielen.
Einzelne Zielstellungen:
 - Klar definierte und dokumentierte Kompetenzen, Rollen und Verantwortlichkeiten.
 - Abgestimmter Orientierungsrahmen für alle Stakeholder.
 - Eindeutige Vorgaben für Programmplanung, Priorisierung und Finanzierungsprozesse.
 - Permanenter Abgleich von Business / IT-Alignment.
 - Transparente Entscheidungsprozesse für Entwicklungs- und Betriebsaktivitäten.
 - Nachhaltige Wertmessung und Ergebnisberichtswesen.

6.2.2 Aufbauorganisation

Im Zentrum der Aufbauorganisation des Data Governance Frameworks stehen die Rollen mit den damit verbundenen Befugnissen, Verantwortlichkeiten und Zuständigkeiten. Weiterhin wird häufig zur Unterstützung ein Data Governance Office (DGO) genannt, welches die Zuständigkeit für die Koordination der anfallenden Data Governance-Aufgaben hat (vgl. (The Data Governance Institute (DGI) 2014), S. 14). Ob damit eine Person, ein Personenkreis oder eine Organisationseinheit im Unternehmen gemeint ist, hängt sehr stark davon ab, was mit der Data Governance erreicht werden soll. Wichtig sind die Kommunikation, das Vereinbaren von Regeln und die Begleitung von Entscheidungsprozessen zur Umsetzung der Daten-Strategie im Unternehmen. Dabei ist insbesondere die Festlegung von Entscheidungsrechten bzw. -befugnissen in Bezug auf die Daten wesentlich. Daraus ergeben sich gleichzeitig Rollendefinitionen für das Data Governance. Die Abb. 6.5 zeigt das Zusammenspiel der wichtigen Elemente in der Aufbauorganisation zu der Daten-Strategie und der IT-Governance (vgl. (Gansor und Totok 2015), S. 24 f.).

Einige wichtige Rollen im Data Governance haben Gansor und Totok und das Data Governance Institute (DGI) beschrieben (vgl. (Gansor und Totok 2015), S. 24 f.) und (The Data Governance Institute (DGI) 2014), S. 13 ff.). Eine ausführliche Beschreibung der Rollen findet sich auch bei Dittmar und Fürber (vgl. (Dittmar und Fürber 2020), S. 17 ff.)

- Data Owner: Senior Manager aus dem Fachbereich mit ausgeprägten Kenntnissen zur Datensemantik und zu fachlichen Datenanforderungen, der weitreichende datenbezogene Entscheidungskompetenzen besitzt und zugehörige Richtlinien definiert. Das Data Governance bezeichnet den Data Owner auch als Data Stakeholder.
- Data Steward: Mitarbeiter aus dem Fachbereich mit Daten- und IT-Verständnis und unternehmensweiter Sichtweise auf die Datenverwendung, der Empfehlungen

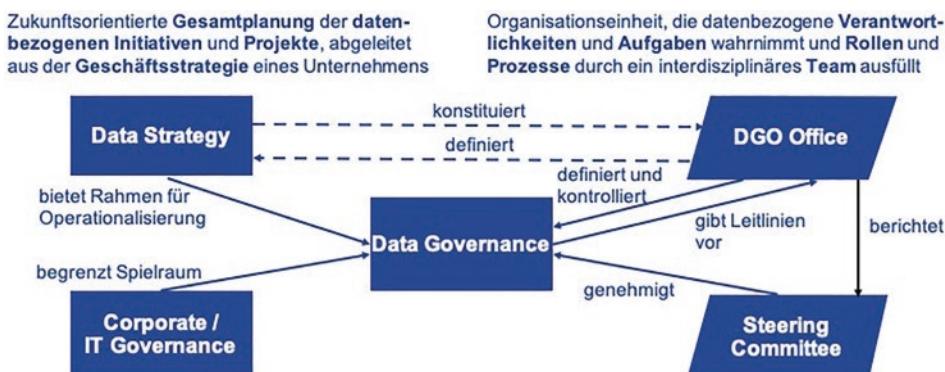


Abb. 6.5 Einordnung DGO und Data Governance

bezüglich der Datenzugriffe, Datenverteilung, Datensicherheit und Datenaufbewahrung ausspricht.

- Data Custodian (technischer Data Steward): Datenspezialist (Architekt, Modellierer, Administrator) aus der IT mit sowohl technischen als auch datenbezogenen Kenntnissen und Kompetenzen in Bezug auf die Datenverwaltung, -archivierung, -sicherung und -wiederherstellung sowie auf die Vermeidung von Datenverlusten oder -verfälschungen und auf unberechtigte Zugriffe.

6.2.3 Richtlinien, Prozesse und Standards

Eine Hauptaufgabe der Data Governance ist die Entwicklung, Vereinbarung und Etablierung von datenbezogenen Richtlinien, Standards und Prozessen. Richtlinien (Policies) stellen verbindliche Vorgaben dar, während Standards der Implementierung der Richtlinien dienen. Zu den Richtlinien sind auch geeignete Kontrollinstrumente und -maßnahmen zu entwickeln, um die Einhaltung der überwachen und sicherstellen zu können, dass die Vorgaben auch tatsächlich umgesetzt wurden.

Das Data Warehouse Institute hat folgende „*Common Data Governance Processes*“ benannt ((The Data Governance Institute (DGI) 2014), S. 16):

1. Aligning Policies, Requirements, and Controls
2. Establishing Decision Rights
3. Establishing Accountability
4. Performing Stewardship
5. Managing Change
6. Defining Data
7. Resolving Issues
8. Specifying Data Quality Requirements
9. Building Governance into Technology
10. Providing Stakeholder Care
11. Communications and Program Reporting
12. Measuring and Reporting Value

An den nachfolgenden ausgewählten Anwendungsfällen kann man sehr schön die Maßnahmen zur Umsetzung der Data Governance darstellen und die Wirkung auf die verschiedenen beteiligten Elemente (Systeme, Menschen, Prozesse, Daten) erkennen. Die Beispiele wurden aus Dittmar und Fürber ((Dittmar und Fürber 2020), S. 27 f.) entnommen:

- Schaffung von Datentransparenz:
 - System: Einführung eines zentralen Datenkatalogs
 - Prozesse: Definition und Umsetzung von Prozessen zur Veröffentlichung und Anreicherung von Metadaten

- Menschen: Etablierung von Datenverantwortlichkeiten
- Daten: Befüllung und Anreicherung eines Datenkatalogs mit Metadaten
- Herstellung und Aufrechterhaltung einer hohen Datenqualität:
 - System: Aufbau von Data-Monitoring- und Data-Cleansing-Systemen und Services, Etablierung von Quality Gates in der Datenproduktion
 - Prozesse: Etablierung eines standardisierten Datenqualitätsmanagements
 - Menschen: Etablierung von Verantwortlichkeiten für Datenqualität, Bereitstellung eines zentralen Data-Services-Teams zur Wahrnehmung von Datenqualitätsaufgaben
 - Daten: Data Profiling, Identifikation von Datenqualitätsproblemen über Datenqualitätsanalysen, Aufbau von Standardisierten Data Monitoring Reports und Entwicklung von Data-Cleansing-Strategien zur Bereinigung der Daten und Behebung der Ursachen
- Stärkere Datennutzung durch Datendemokratisierung:
 - System: Aufbau einer zentralen Rohdatenschicht, Aufbau von Self-Services-Analytics-Plattformen, Aufbau einer internen Datenaustauschplattform, APIs für den Datenzugriff
 - Prozesse: Berechtigungsprozesse für den Datenzugriff, Datenbereitstellungsprozesse
 - Menschen: Etablierung von Data Owner zur Freigabe von Daten
 - Daten: Bildung von Data Sets als teilbare Datenpakete

Die Schritte zur erfolgreichen Einführung einer Data Governance lassen sich je nach Unternehmen in verschiedene Schritte einteilen. Das Data Governance Institute beschreibt sieben Schritte (vgl. (The Data Governance Institute (DGI) 2014), S. 7) Bei Dittmar/Fürber finden sich folgende acht Schritte zur Einführung (vgl. (Dittmar und Fürber 2020), S. 26 ff.):

- Schritt 1: Ermittlung des Status Quo im Datenmanagement.
- Schritt 2: Ziele und Scope definieren.
- Schritt 3: Initiales Data Governance-Konzept inklusive Roadmap erstellen.
- Schritt 3: Initiales Data Governance-Konzept inklusive Roadmap erstellen.
- Schritt 4: Zustimmung von Stakeholder einholen und Sponsor finden.
- Schritt 5: Data Governance-Konzept ausarbeiten und Transformationsprogramm aufsetzen.
- Schritt 6: Roadmap fokussiert umsetzen.
- Schritt 7: Data Governance auf weitere Bereiche ausrollen.
- Schritt 8: Data Governance stabilisieren und kontinuierlich verbessern.

6.2.4 Messen und Beobachten

Data Governance ist keine einmalige Aktivität, sondern ein kontinuierliches Programm, das Weiterentwicklung und Verbesserungen als Zielsetzung hat. Es ist also auch

kontinuierlich zu Beobachten und zu Messen, welchen aktuellen Stand da Data Governance hat.

Durch regelmäßige Messungen mit dem Abgleich zu vorher definierten Zielgrößen lassen sich Auffälligkeiten und Abweichungen frühzeitig identifizieren und analysieren. Aus der Analyse heraus ergeben sich weitere evtl. bisher unbekannte Probleme und Konflikte, die in einem strukturierten Issue-Management erfasst und gelöst werden können.

Ein weiterer wichtiger Aspekt ist die Verfolgung der Strategie. Durch ein regelmäßiges Monitoring lassen sich feststellen, ob die Daten-Strategie auch tatsächlich zielgerichtet erfüllt werden kann.

6.2.5 Technologie

Im Bereich der Technologie sind die Themen zu betrachten, die durch die IT im Unternehmen realisiert werden müssen. Dazu gehört neben den Aspekten Datenschutz und Datensicherheit auch das Datenqualitätsmanagement (Data Quality Management), welches im Abschn. 6.3 noch näher betrachtet werden soll. Für alle Themen sind geeignete Werkzeuge bereitzustellen und Mitarbeiter entsprechend zu schulen.

Insbesondere das Thema der Metadaten ist von zentraler Bedeutung für die Unternehmen und rückt daher für das Data Governance in den Mittelpunkt. Die Unternehmen benötigen eine Übersicht über die vielfältigen Daten, die irgendwo in irgendwelchen Datentöpfen vorhanden sind. Das Management der Metadaten rückt damit immer stärker in den Fokus. Unter Metadaten versteht Informationen zu den abgelegten Produktdaten, die Aussagen liefern zu (vgl. (Gluchowski 2020a), S. 11):

- Der Bedeutung der Informationsobjekte und dem betriebswirtschaftlichen Kontext.
- Prozessinformationen hinsichtlich der Veränderung, Verknüpfung und logischer Zuordnung.
- Strukturangaben hinsichtlich des Datentyps, Wertebereich, Qualität.
- Administrative Informationen über Erstellungszeitpunkt, Zugriffshäufigkeit und Berechtigung.

Es sind also deutlich mehr Metadaten zu erheben und zu katalogisieren, als man das aus den bisherigen Data Repositories der Datenbanken kennt. Die Data Repositories können aber schon einen Teil der benötigten Metadaten liefern (z. B. administrative Angaben). Sie sind aber durch die Konzepte Data Catalog und Data Lineage zu erweitern.

Data-Lineage bzw. Datenherkunft (auch Data Provenance oder Data Pedigree, deutsch auch Datenabstammung und -stammbaum) bezeichnet in einem Data Warehouse-System (Datenlager) die Fragestellung, zu gegebenen aggregierten Datensätzen die ursprünglichen Datensätze zu bestimmen, aus denen sie entstanden sind. Üblicherweise werden in einem Data-Warehouse-System Daten aus verschiedenen

Quellen extrahiert, nach bestimmten Regeln transformiert und zur Analyse bereitgestellt (ETL-Prozess). Beim Data-Lineage muss der umgekehrte Weg beschrieben werden, um von Analyseergebnissen zu den Quellen zu gelangen. Dazu werden die Transformationen mathematisch modelliert, um für gegebene Ausgabewerte einer Transformation die dazugehörigen Eingabewerte zu bestimmen (vgl. (Data Lineage, kein Datum)).

Ein Data Catalog dient der Beschreibung der gespeicherten Daten aus technischer und fachlicher Sicht. Die technische Sicht könnte durch ein oder mehrere Datenbank Repository bewerkstelligt werden. Wichtig ist hier die fachliche (Business) Sicht auf die Daten. Damit lassen sich Daten einfacher finden und interpretieren. Dies gilt nicht nur für analytische Vorhaben, sondern auch für den „normalen“ operativen Betrieb. Die Fachseiten können direkter auf die für interessanten Daten zugreifen und sie ggf. auswerten. Es lassen sich so aber auch fachliche Unterschiede in den Daten besser erkennen und eine Analyse von inhaltlich nicht passenden Daten verhindern. Die Abb. 6.6 zeigt in groben Schritten den Prozess der notwendig ist, wenn ein Data Catalog im Unternehmen aufgebaut bzw. genutzt wird.

Wichtig ist dabei die fachliche Taxonomie, die die Kategorien für die Daten aus fachlicher Hinsicht liefert. Die Definition fachlicher Metadaten im Data Catalog basierend auf der Unternehmenstaxonomie ermöglicht es den Data Scientists und den Data Engineers die für ihre Fragestellung relevanten Daten zu suchen und diese Data Assets innerhalb des unternehmensweiten Data Lakes zu finden. Kategorien sind dabei die Geschäftsprozesse, die Geschäftsbereiche (Sparten), Produkte oder Services, die Organisation inklusive regionale Strukturen und Länder sowie die Zeit (siehe Abb. 6.7).

Im Data Catalog laufen die verschiedenen Metadaten zusammen (vgl. Abb. 6.8) und die verschiedenen Rollen können hier die Daten klassifizieren und mit weiteren Informationen anreichern. Die Nutzer der Daten, also die Data Scientists und die Data Engineers, können über den Data Catalog die Datenobjekte recherchieren, die für die



Abb. 6.6 Prozess zur Nutzung eines Data Catalogs, in Anlehnung an Eberlein 2017

Abb. 6.7 Fachliche Taxonomie, in Anlehnung an (Eberlein 2017)

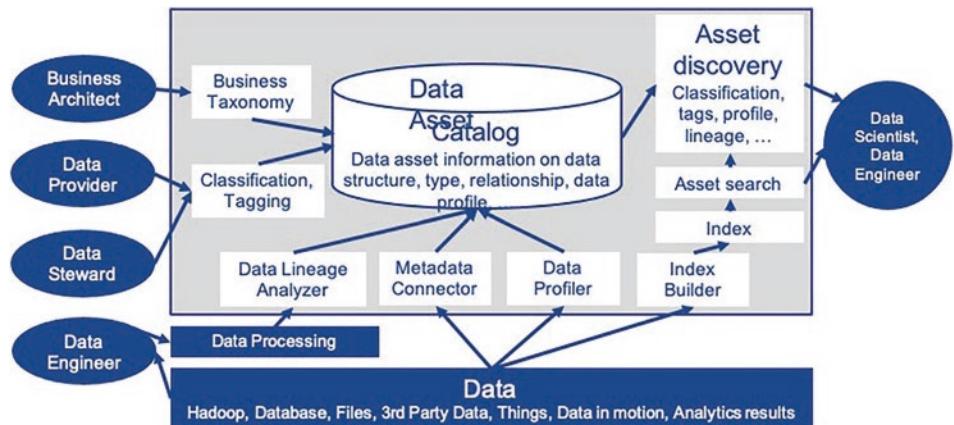
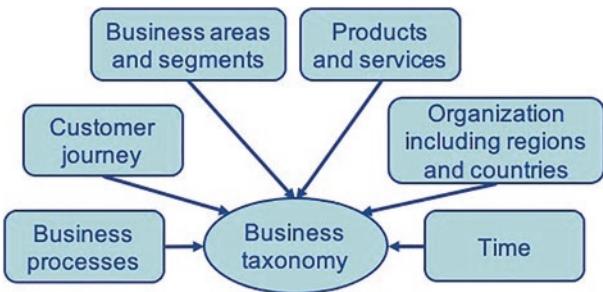


Abb. 6.8 Data Asset Catalog, in Anlehnung an Eberlein 2017

Aufgabenerledigung notwendig sind. Hier werden nicht nur die strukturierten Daten aus den operativen Systemen recherchierbar, sondern die übrigen unstrukturierten Daten in den Data Lakes.

6.2.6 Kommunikation

Nur mit einer zielgruppenorientierten Kommunikation kann die Data Governance erfolgreich sein. Hier geht es zum einen darum, dass die Richtlinien und Regelungen an die ausführenden Mitarbeiter im Unternehmen kommuniziert werden. Zum anderen sollten die zuständigen Mitarbeiter auch frühzeitig über eine Data Governance Initiative informiert und in die laufenden Aktivitäten eingebunden werden, um Akzeptanz zu schaffen.

Es geht aber auch um die Etablierung eines strukturierten Informationsaustauschs. Dies wird erreicht, indem ein Kommunikationsplan entwickelt wird, der regelt, wer Informationen benötigt, welcher Informationsbedarf besteht und wer für deren Bereitstellung verantwortlich ist.

Zur erfolgreichen Implementation eines Data Governance sollte auch ein Trainingsplan entwickelt werden, der die Schulungsmaßnahmen und den Schulungsumfang für die verschiedenen involvierten Gruppen im Unternehmen (z. B. Data Stewards) regelt.

6.3 Data Quality Management (DQM)

Eng mit der Data Governance verbunden ist das Data Quality Management (DQM) und das Master Data Management (MDM). Beide Themenbereiche sind in einer Data Governance unverzichtbar. Andersherum findet am aber in den Unternehmen durch ein DQM und MDM ohne das bereits eine Data Governance implementiert wurde. Es ist also auch schon in der „alten Welt“ von großem Vorteil, wenn mit qualitativ hochwertigen Daten gearbeitet wird. Unternehmen haben schon immer erkannt, dass in ihren Daten sowohl großes Potenzial für ein besseres Wirtschaften als auch attraktive Produkte schlummern. Doch es fällt den Unternehmen schwer, die Daten so zu nutzen, dass die Dinge tatsächlich besser werden. Ein Grund liegt häufig in Daten, die Fehler enthalten, sich widersprechen, nicht vollständig genug oder veraltet sind (vgl. (Iffert et al. 2020), S. 85).

Nachfolgende Prozessbereiche sind im Rahmen der DQM und MDM nach Apel, Behme, Eberlein und Merighi zu berücksichtigen (vgl. (Apel et al. 2010b))

- Validierung: Prüfung auf Verwendbarkeit und Separierung fehlerhafter Daten.
- Standardisierung: Formatvereinheitlichung und Normierung (Abbildung auf vorgegebene Werteliste).
- Bereinigung (Data Cleansing):
 - Entfernen oder Ersetzen fehlerhafter Daten.
 - Entfernung von Duplikaten.
 - Aufspalten zusammengefasster Daten zu unterschiedlichen Objekten.
 - Anreicherung: Kombination mit zusätzlichen (z. B. geografischen) Daten zur Steigerung des Informationswertes.
- Data Monitoring: Analyse und laufende Überwachung der Entwicklung der Datenqualität im Data Warehouse.

Sehr zentral ist dabei das sogenannte Data Profiling, also den weitgehend automatisierten Prozess zur Analyse vorhandener Datenbestände durch unterschiedliche Analysetechniken, welches in der Abb. 6.9 als iterativer Prozess dargestellt ist.

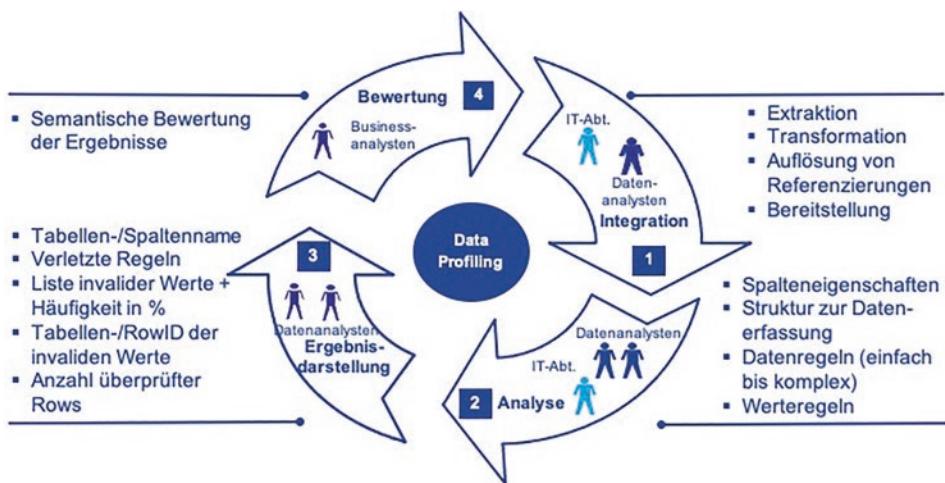


Abb. 6.9 Data-Profiling-Analyse als iterativer Prozess, in Anlehnung an Apel und Behme 2010a, S. 117

6.4 Fazit

Die sogenannten „Megatrends“ Big Data und Industrie 4.0 bestimmen die aktuelle Forschung im Bereich der Data Governance. Die Anzahl der Datenquellen nimmt zu, wie auch die Menge der Daten und insbesondere deren unterschiedlicher Strukturierungsgrad. Im Fokus von Big Data liegen diese Eigenschaften in besonders hohem Maße vor und führen dazu, dass sich mit den klassischen Werkzeugen für Datenhaltung und -verarbeitung ein anforderungsgerechter Umgang mit den verfügbaren Datenmengen an seine Grenzen stößt (vgl. (Brüning et al. 2017), S. 11).

Die Manager in den Unternehmen verstehen Daten inzwischen als wertvolles Unternehmensgut und teilen sie deswegen mit den Business Analysts und Data Scientists in den Fachbereichen – und manchmal auch über die Organisationsgrenzen hinaus. Heute bestimmen konkrete und geschäftsrelevante Anwendungsfälle das Big-Data-Analytics-Geschehen. Die Projekte werden agil und fachlich vorangetrieben. Ohne Data Governance ist jedoch der Wildwuchs in den Data Lakes der Unternehmen zu groß. Es ist dringend notwendig einen Ordnungsrahmen durch Data Governance zu schaffen (vgl. (Eberlein 2016)).

Literatur

Apel, D., Behme, W.: Datenintegration – Ein Prozess zur Verbesserung der Datenqualität, In: Chamoni, P. G. (Hrsg.), Analytische Informationssysteme : Business-Intelligence-Technologien und -Anwendungen (4. Aufl. Ausg., S. 115 - 130). Springer, Berlin (2010a)

- Apel, D., Behme, W., Eberlein, R., Merighi, C.: Datenqualität erfolgreich steuern, 2. Aufl. Ausg. Carl Hanser, München (2010b)
- Brecht, U.: BWL für Führungskräfte. Was Entscheider im Unternehmen wissen müssen 2. Aufl. Ausg. Springer Gabler, Wiesbaden (2012)
- Brüning, A., Gluchowski, P., Kaiser, A.: Data Governance – Einordnung, Konzepte und aktuelle Herausforderungen. Chemnitz Economic Papers, No. 015 (2017)
- Corporate Governance. Wikipedia. https://de.wikipedia.org/wiki/Corporate_Governance (kein Datum). Zugegriffen: 28. Dez. 2020
- Data Lineage.: Wikipedia. <https://de.wikipedia.org/wiki/Data-Lineage>. (kein Datum). Zugegriffen: 28. Dez. 2020
- DEMAND.: Data Economy. Status quo der deutschen Wirtschaft & Handlungsfelder in der Data Economy. <https://www.demand-projekt.de/> (kein Datum). Zugegriffen: 04. Juni 2019
- Dittmar, C., Fürber, C.: Data Governance als Wegbereiter der Digitalisierung. In: Gluchowski, P. (Hrsg.), Data Governance. Grundlagen, Konzepte und Anwendungen, S. 13–31. Dpunkt, Heidelberg (2020)
- Eberlein: Big Data Governance in der Praxis. TDWI-Tagung 2017. München (2017)
- Eberlein, R.: Wildwuchs gefährdet Geschäftserfolg: Ohne Big Data Governance droht das Chaos. CIO.de: <https://www.cio.de/a/ohne-big-data-governance-droht-das-chaos,3260737> (2016). Zugegriffen: 28. Dez. 2020
- Gansor, T., Totok, A.: Von der Strategie zum Business-Intelligence- Competency-Center (BICC): Konzeption, Betrieb, Praxis, 2. Aufl. Ausg. Dpunkt, Heidelberg (2015)
- Gluchowski, P.: Data Governance – Einführung und Überblick. In Gluchowski, P. (Hrsg.), Data Governance. Grundlagen, Konzepte und Anwendungen, S. 3–12. Dpunkt, Heidelberg (2020a)
- Gluchowski, P.: Data Governance. Grundlagen, Konzepte und Anwendungen. Dpunkt, Heidelberg (2020b)
- Governance.: Wikipedia. <https://de.wikipedia.org/wiki/Governance> (kein Datum). Zugegriffen: 28. 12 2020
- Iffert, L.: Komponenten für zufriedenstellende Datenqualität und Stammdaten. In P. Gluchowski (Hrsg.), Data Governance. Grundlagen, Konzepte und Anwendungen, S. 85–98. Dpunkt, Heidelberg (2020)
- Iffert, L., Görlich, O., Grosser, T.: Treiber für Data Governance. BARC-Track: Trends im Data Warehousing und Datenmanagement, TDWI-Tagung 2013. München (2013)
- Krotova, A., Eppelsheimer, J.: Was bedeutet Data Governance? Eine Clusteranalyse der wissenschaftlichen Literatur zu Data Governance. Institut der deutschen Wirtschaft Köln e.V, Köln (2019)
- o.A.: The Economist. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> (2017). Zugegriffen: 28. Dez. 2020
- O’Neal, K.: Big Data: Governance ist the Critical Starting Point. https://www.b-eye-network.com/blogs/oneal/archives/2012/11/big_data_govern.php (2012). Zugegriffen: 03. März 2020
- Seiner, R.: Real-World Data Governance: BI Governance and the Governance of BI Data. <https://de.slideshare.net/Dataversity/realworld-data-governance-bi-governance-and-the-governance-of-bi-data-14889552> (2012). Zugegriffen: 28. Dez. 2020
- The Data Governance Institute (DGI).: How to Use the DGI Data Governance Framework to Configure Your Program. http://datagovernance.com/wp-content/uploads/2014/wp_how_to_use_the_dgi_data_governance_framework.pdf (2014). Abgerufen am 22. 01 2020
- Villar, M., Kushner, T., Wells, D.: Data Governance Fundamentals. Retrieved 12 28, 2020, from https://ecm.elearningcurve.com/Online_Data_Governance_Fundamentals_Course_p/dg-01-a.htm (2018)
- Weill, P., Ross, J.W.: IT Governance: How Top Performers Manage IT Decision Rights for Superior Results. Harvard Business Press, Brighton (2004)



Einsatz von In-Memory Technologien

7

Uwe Schmitz

Zusammenfassung

Viele Unternehmen stehen aktuell vor der Herausforderung, ein stetig steigendes Datenvolumen verarbeiten zu müssen. Die richtigen Informationen sollen zum richtigen Zeitpunkt in adäquater Form, Menge und Qualität für die jeweiligen Anwender im Unternehmen verfügbar sein – und das in nur wenigen Sekunden. Traditionelle relationale Datenbanksysteme stoßen bei diesen Anforderungen häufig an ihre Grenzen. Moderne Datenbankmodelle legen ihre Daten nicht auf herkömmlichen Festplattenspeichern ab, sondern nutzen hierfür direkt den Arbeitsspeicher. Dadurch lassen sich wesentlich höhere Zugriffsgeschwindigkeiten realisieren und in der Konsequenz die betrieblichen Entscheidungsprozesse verkürzen und effizient unterstützen

7.1 Einleitung

Die In-Memory Technologie ist eine bedeutende technologische Grundlage sowohl bei analytischen Informationssystemen im Kontext von Big Data und Business Intelligence Lösungen als auch zunehmend bei operativen Informationssystemen. Es sind bestimmte Anforderungen mit dem Einsatz einer In-Memory-Technologie verknüpft.

Viele Unternehmen stehen vor der Herausforderung eine stetig steigende Menge an Informationen verarbeiten zu müssen, um letztlich die richtigen Informationen zum richtigen Zeitpunkt in der richtigen Form, Menge und Qualität dem richtigen Anwender

U. Schmitz (✉)

FB Wirtschaft, Fachhochschule Dortmund, Dortmund, Deutschland

E-Mail: uwe.schmitz@fh-dortmund.de

zur Verfügung zu stellen (vgl. Schmitz 2006, Konzeption, S. 170 ff.). Damit verbunden können betriebliche Entscheidungsprozesse durch entsprechende IT-Technologien, die diese Menge an Informationen verwalten, verkürzt und effizient unterstützt werden. Trotz des steigenden Datenvolumens besteht ein Anspruch darin, sämtliche Unternehmensdaten vollständig, korrekt, konsistent und mit kurzen Antwortzeiten (in wenigen Sekunden) auswerten zu können. Die traditionellen relationalen Datenbanksysteme stoßen bei diesen Anforderungen häufig an Ihre Grenzen.

Neue Quellen für Unternehmensdaten bilden beispielsweise das Smart Grid (Versorger, Energie), die Kennzeichnung von Waren mit Radio Frequency Identification (Handel oder Transport), die Datenerfassung über die Gesundheitskarte (Gesundheit) oder auch die Erfassung und Überwachung von Serviceprozessen.

Während die reine Erhebung und Speicherung der Daten mit verfügbaren Technologien umsetzbar ist, gerät die Verarbeitung, Aufbereitung (Veredelung), Analyse und Verbreitung derartiger Datenbestände zur Herausforderung. Bisherige Ansätze stoßen dabei an ihre Grenzen, was sich in niedriger Performance oder hohen Systemkosten niederschlägt. Der Umgang mit dem anwachsenden Datenvolumen stellt eine zunehmende Herausforderung für die Unternehmen dar.

Neue Technologien, die u. a. Einzug in Anwendungen wie die der Business Intelligence finden, sollen hier unterstützen (vgl. Schmitz 2010, S. 227 ff.). Beispiele für diese Technologien sind steigende Hauptspeicherkapazitäten und ein damit verbundenes In-Memory Data Management, spaltenorientierte Datenbanken, steigende Rechenkapazitäten durch Cloud Computing und Multicore Architekturen oder Appliances (Integrierte abgestimmte Hardware/Softwarelösungen) sowie Service-orientierte Architekturen.

Diese Technologien ermöglichen neue Potenziale im Umgang mit Informationen. Die bis dato existierende Trennung zwischen operativen (OLTP) und analytischen (OLAP) Datenbeständen soll hierdurch aufgehoben werden und rechenintensive fortgeschrittene Prognose-, Simulations- und Optimierungs-verfahren zur Analyse und robusten Entscheidungsfindung werden ermöglicht. Eine Sicherstellung der Verfügbarkeit der Business Informationen erfolgt nutzergerecht durch Methoden des Service Managements auf neuen, ggf. kooperativen Plattformen.

Als Beispiele für solche Anwendungen können technische und betriebswirtschaftliche Planungssysteme (z. B. zur strategischen Investitionsplanung für Energieanlagen), Echtzeitbereitstellung von entscheidungsrelevanter Information (Real-Time Business Intelligence), Kooperationsplattformen zur Verbreitung (CSCW- und Wissensmanagementsysteme), Simulation von IT-Servicemanagementprozessen als Einführungs-, Optimierungs- und Trainingshilfsmittel, Configuration Management Systeme (CMS) mit einer Configuration Management Database (CMDB) und weiteren Datenbanken als zentralem Informationssystem aller internen und externen IT-Services genannt werden. Damit verbunden ist die Methodik zum Aufbau derartiger Systeme, die sich im Spannungsfeld zwischen den Geschäftsprozessen des Unternehmens und der verfügbaren IT-Technologie bewegt.

Im Folgenden soll der Einsatz von In-Memory Technologien durch Nutzung großer Hauptspeicherkapazitäten in Verbindung mit spaltenorientierten Datenbanken und Multiprozessoren untersucht werden.

7.2 Definition und Abgrenzung In-Memory Technologien

Unter In-Memory-Technologie (auch In-Memory-Computing) versteht man im Allgemeinen das Speichern und Verwalten der Daten im Hauptspeicher. Bei traditionellen Datenbanksystemen werden die Daten ausschließlich auf der Festplatte gespeichert und bei Abfragen von dort gelesen. Dies bedeutet, dass nur ein Bruchteil der Daten im Hauptspeicher verwaltet wird. Mit dem Einsatz der In-Memory-Technologie werden die gesamten Daten entweder direkt im Hauptspeicher gehalten oder beim Programmstart von der Festplatte komplett in den Hauptspeicher geladen. Somit können lesende Zugriffe weitaus schneller erfolgen als bei traditionellen Datenbanksystemen, da keine I/O-Zugriffe (Input/Output) auf die Festplatte erfolgen.

Die traditionellen relationalen Datenbanken (RDBMS) unterscheiden sich in vielen Aspekten von In-Memory-Lösungen. So verwenden diese Datenbanksysteme eine „Disk“ als Datenspeicher. Für eine Datenverarbeitung werden die Daten von der Disk in den Hauptspeicher geladen. Diese Operationen stellen meist einen Engpass im System dar. Aus diesem Grund sind die RDBMS bei der Analyse großer Datenmengen häufig nicht in der Lage, benötigte Abfragen schnell durchzuführen. Die Zugriffslücke zwischen Hauptspeicher und Festplatte ist zwar weggefallen, aber es besteht immer noch eine Barriere zwischen den Prozessoren mit ihrem lokalen CPU-Cache und dem Hauptspeicher. Diese Blockade wird als „Memory Wall“ bezeichnet (vgl. Abb. 7.1).

Die Verwendung eines großen Hauptspeichers im Rahmen der In-Memory-Technologie bietet gegenüber der Festplatte große Performance Vorteile hinsichtlich der Zugriffszeiten und Datentransfer. Der Hauptspeicher ist aktuell der schnellste und geeignete Speichertyp, der sehr große Datenmengen halten kann. Daten im Hauptspeicher können ca. 100.000 Mal schneller verarbeitet werden als auf der Festplatte.

Jedoch sind genügend verfügbare Hauptspeicherkapazitäten und die Bereitstellung schneller Prozessoren nicht alleine ausreichend, um die Datenverarbeitung so zu beschleunigen, dass Anwender tatsächlich Informationen in Echtzeit erhalten. So empfiehlt sich neben der Datenhaltung im Hauptspeicher, um den Datentransfer zu verkürzen eine Minimierung der Datenbewegung sowohl innerhalb der Datenbank als auch zwischen der Datenbank und der Anwendung durch Nutzung einer spaltenorientierten Speicherung und der Verwendung von Kompressionsmethoden. Weitere Geschwindigkeitsvorteile können durch Techniken wie Parallelisierung und Partition erreicht werden. Diese Aspekte sollen im Folgenden näher untersucht werden.

Durch die Haltung der Daten im Hauptspeicher entstehen zwar große Performance Vorteile hinsichtlich der Zugriffszeiten, jedoch besteht eine erhöhte Gefahr für einen möglichen Datenverlust bei einem Stromausfall. Daher ist eine Datenpersistenz

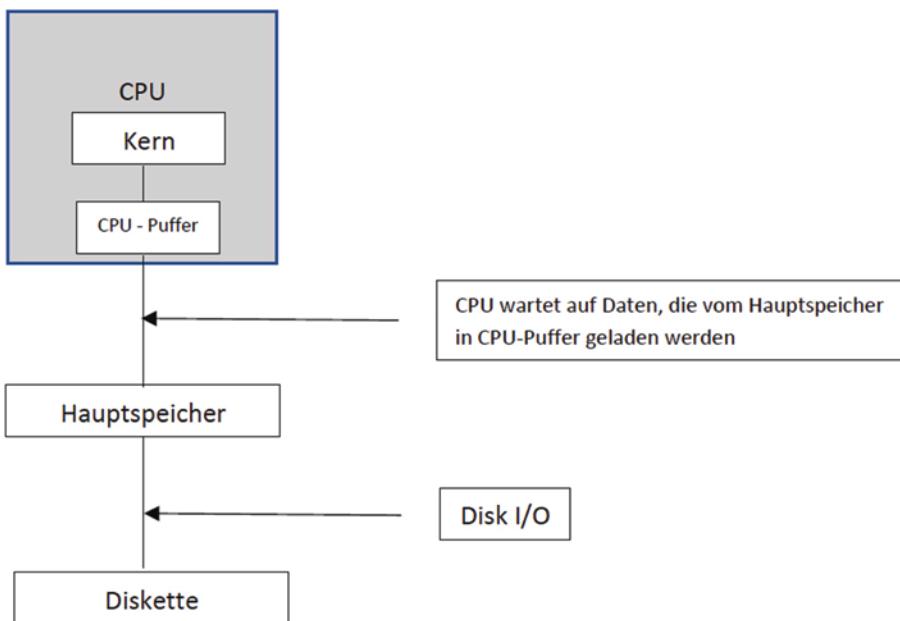


Abb. 7.1 Technische Architektur der RDBMS. (Eigene Darstellung)

sicherzustellen, die neben Atomarität und Konsistenz ein bedeutsames Kriterium für den Einsatz einer In-Memory Technologie darstellt. So kann der Hauptspeicher allein die Datenpersistenz in einer In-Memory Datenbank (IMDB) nicht gewährleisten. Im Fall eines Energieverlustes würden die gesamten Dateninhalte im Hauptspeicher verloren gehen. Um die Datenpersistenz zu gewährleisten ist zusätzlich zum Hauptspeicher ein nichtflüchtiger Speicher, wie zum Beispiel eine Festplatte, notwendig (vgl. Vey et al. 2014, S. 9).

Für die Gewährleistung der Datensicherung in einer IMDB können SavePoints (Snapshots) und Logs genutzt werden. So kann aktuelle Zustand der Datenbank im SavePoint gespeichert und in regelmäßigen Zeitabständen in den nichtflüchtigen Speicher geschrieben werden. Sämtliche zwischenzeitlichen Änderungen werden als Logs in einer Logdatei im nichtflüchtigen Hauptspeicher gespeichert. Nach einem Stromausfall kann durch den SavePoint der aktuellste Zustand der Datenbank wiederhergestellt werden. Alle Änderungen, die zwischen dem Zeitpunkt des Stromausfalls und der letzten SavePoint-Sicherung entstanden sind, werden in einer Logdatei gesichert und können bei Bedarf wiederhergestellt werden.

Die Abb. 7.2 verdeutlicht diesem Sachverhalt und Zeilenorientierte Datenbanken

Andere mögliche Techniken zur Lösung dieses Problems sind beispielsweise die Nutzung von Batterien bzw. Akkus bei einem Stromausfall, um den Speicher weiterhin mit Strom versorgen zu können. Eventuelle Hardwarefehler lassen sich durch

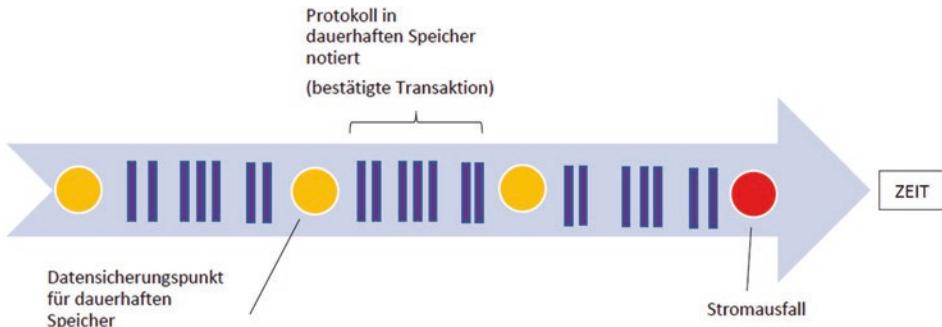


Abb. 7.2 Beispiele für Logs und Savepoints. (Eigene Darstellung)

Fehlererkennungs- und Korrekturmaßnahmen sowie den Einsatz von Redundanzen eliminieren. Allerdings müssen auch bei dem Einsatz solcher Techniken Sicherheitskopien erstellt werden.

Der nächste Faktor zur Beschleunigung der Datenverarbeitung ist die Minimierung der Datenbewegungen sowohl innerhalb der Datenbank als auch zwischen der Datenbank und der Anwendung. Um dies zu erreichen können Techniken wie Kompression oder Spaltenorientierte Speicherung eingesetzt werden.

Der Einsatz von großen Hauptspeicherkapazitäten ist i. d. R. sehr kostenintensiv. Durch den Einsatz von Kompressionsmethoden lässt sich der Speicherplatzbedarf reduzieren und die damit verbundenen Kosten für die benötigte Hauptspeicherkapazität können reduziert werden. Gleichzeitig können durch die Kompression Geschwindigkeitsvorteile erzielt werden (vgl. Plattner und Zeier 2011 S. 64).

Bei der Anwendung von Kompressionsmethoden können in Verbindung mit der Nutzung von sog. „Bibliotheken“ Textwerte in der Spalte einer Tabelle durch Positionsschlüssel (Zahlen) der Bibliothek ersetzt werden (vgl. Vey et al. 2014, S. 10). Praktisch wird jeder Wert einer Spalte in der Bibliothek einmal eingefügt, sodass kein Wert mehrfach auftritt. Bei jeder Verwendung dieses Wertes wird der Positionsschlüssel in der Spalte eingetragen. Die Verwendung von Bibliotheken für Textwerte verringert sich so der Speicherbedarf für eine Tabelle, weil jeder Textwert der Tabelle nur einmal gespeichert werden muss und bei jedem weiteren Ereignis nur ein Verweis auf die Positionsschlüssel in der Bibliothek nötig ist. Die Effizienz von solchen Kompressionstechniken ist bei Datenmengen mit wenigen unterschiedlichen Attributwerten besonders sinnvoll. Im Falle von Datenmengen mit vielen unterschiedlichen Attributwerten sind solche Kompressionsmethoden dagegen nicht sehr effektiv (vgl. Vey et al. 2014, S. 10). In diesem Fall werden die Bibliotheken für die Textwerte sehr groß und der Kompressionseffekt sehr gering.

Die Nutzung einer Spaltenorientierten Kompression ist dann hilfreich, wenn alle Daten innerhalb einer Spalte denselben Datentyp haben und eine ähnliche Semantik aufweisen. Bei einer zeilenorientierten Speicherung hingegen lässt sich eine Kompression

schwerer durchführen, weil in einem Datensatz i. d. R. Werte vieler verschiedener Typen gespeichert sind.

So ist festzustellen, dass die Kompressionsrate vom Datentyp, Anzahl unterschiedlicher Werte einer Spalte und Anzahl von NULL-Werten abhängig ist. Allgemein kann festgehalten werden, dass Kompressionsfaktoren von 8 bis 10 durchaus realistisch sind und in Einzelfällen deutlich höhere Raten möglich sind. Entscheidend bei Kompression ist aber, dass kein Datenverlust entsteht und eine Balance zwischen Kompression und Dekompression besteht. So sollte die gewonnene Rechengeschwindigkeit durch die Kompression deutlich höher sein als die aufgewendete Rechenleistung für die Dekompression der Daten bei einer Analyse (vgl. Velt et al. 2012, S. 68).

Im Gegensatz zu klassischen zeilenorientierten Datenbanken (zeilenorientierte Speicherung), wo die Daten zeilenweise in Tupeln gespeichert werden, werden die Daten in spaltenorientierten Datenbanken (spaltenorientierte Speicherung) tupelübergreifend in separaten Spalten gespeichert. Beim Durchsuchen einer Tabelle nach bestimmten Attributwerten wäre beispielsweise in einer zeilenorientierten Datenbank das Folge-Attribut das nächste Attribut in der nächsten Spalte, während in einer spaltenorientierten Datenbank das nächste Attribut in derselben Spalte ist. Die Abb. 7.3 zeigt einen Vergleich.

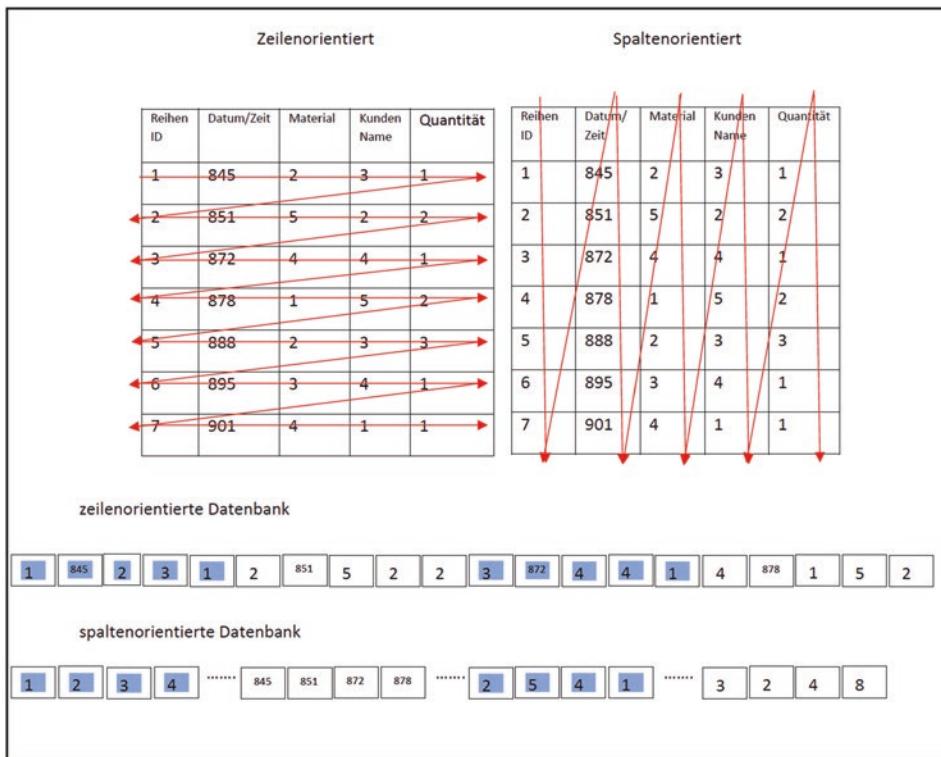


Abb. 7.3 Spalten- und Zeilenorientierte Datenbanken. (Eigene Darstellung)

Bei einer transaktionalen Verarbeitung in einem operativen System ist eine zeilenorientierte Datenbank vorteilhaft, da dort einzelne Datensätze mit wenigen Zeilen bearbeitet werden. Eine spaltenorientierte Datenbank ist besonderes bei analytischen Applikationen (OLAP) vorteilhaft, wenn aus sehr großen Datenmengen nur wenige Attribute abgefragt werden, da für die Anfrage unnötige Spalten nicht geladen werden müssen. Der Vorteil einer spaltenorientierten Speicherung ist, dass nicht jedes einzelne Attribut eines Datensatzes gelesen wird bis das gewünschte Attribut in der Zielspalte erreicht ist, sondern dass nur einmal auf die Zielspalte zugegriffen wird, um die gewünschten Attribute zu lesen. Allein durch die Reduzierung der zu lesenden Datenmenge kann so ein großer Geschwindigkeitsvorteil erreicht werden (vgl. Plattner und Zeier 2011, S. 72).

Für Unternehmen, die sowohl spaltenorientierte als auch zeilenorientierte Speicherung für ihr Tagesgeschäft benötigen, ist der Einsatz von hybriden Datenbanken sinnvoll. In hybriden Datenbanken können die Daten sowohl zeilenorientiert als auch spaltenorientiert gespeichert werden. Somit können Unternehmen von den Vorteilen beider Speicherarten profitieren.

Im Folgenden sollen nun die Techniken Parallelisierung und Partition zur Beschleunigung der Datenverarbeitung kurz untersucht werden.

Unter Parallelisierung im Kontext von Datenbanksystemen versteht man die Nutzung der Verarbeitungskapazität zahlreicher Prozessoren zur Leistungssteigerung. Ein Ziel besteht darin die Bearbeitungszeit von Transaktionen und Abfragen (sog. Queries) zu verkürzen, um inakzeptable Antwortzeiten durch eine sequenzielle Bearbeitung zu vermeiden.

Die Leistung eines Prozessors bzw. dessen Taktfrequenz hängt entscheidend von der Anzahl seiner Kerne (Core) bzw. Verarbeitungseinheiten ab. Die Performance eines Prozessors kann i. d. R. nur durch eine parallele Verarbeitung (parallel processing) gesteigert werden, weil die Performance eines Kernes immer unverändert bleibt. So kann beispielweise bei der Durchsuchung einer Datenbanktabelle nach bestimmten Werten die Tabelle in Untertabellen unterteilt und auf die Kerne des Prozessors partitioniert werden, die dann die Suchabfrage parallel ausführen. Im Vergleich zur Bearbeitung der Abfrage in einem Single-Kern Prozessor kann damit die Bearbeitungszeit um einen Faktor reduziert werden, welcher zu der Anzahl der Kerne des Prozessors nahezu äquivalent ist. Dieses Prinzip gilt ebenfalls für ein Multi-Prozessor System. So könnte ein System mit Zehn 8-Kerne Prozessoren die Bearbeitungszeit bspw. nahezu um den Faktor 80 reduzieren. Die Möglichkeit einer Parallelisierung kann mit verschiedenen Architekturen umgesetzt werden.

7.3 Anforderungen an den Einsatz einer In-Memory-Technologie

Die Anforderungen an den Einsatz einer In-Memory-Technologie können in verschiedene Kategorien strukturiert werden, z. B. betriebswirtschaftliche Anforderungen (u. a. einmalige oder laufende Kosten) oder technische Anforderungen (Performance

etc.). Für die Untersuchung der Anforderungen an den Einsatz einer In-Memory-Technologie und den damit verbundenen Erfüllungsgrad sind Kriterien zu definieren. Folgend werden einige Kriterien beispielhaft erläutert.

Die Echtzeit-Analyse ist die Fähigkeit eines Systems, strukturierte sowie unstrukturierte Daten aus unterschiedlichen Quellen in Sekundenschnelle zu analysieren und die richtigen Daten zur Verfügung stellen (vgl. Plattner und Zeier 2011, S. 10). Die Fähigkeit zur Echtzeit-Analyse ist ein wichtiges Kriterium bei der Bewertung einer IMDB. Damit verbunden ist die Möglichkeit, aus den gegebenen Informationen neue Daten zu erzeugen, auszuwerten oder komplexere Berechnungen durchzuführen.

Die Reaktionszeit ist die Zeit, die eine Lösung benötigt, um die Antwort auf eine Abfrage zu geben oder auf eine Anfrage zu reagieren. Die kurzen Reaktionszeiten einer IMDB resultieren überwiegend aus der Tatsache, dass die Daten bereits im Hauptspeicher vorhanden sind. Die Reaktionszeiten einer In-Memory-Lösung liegen i. d. R. im Bereich weniger Sekunden, teilweise sind Differenzen verschiedener In-memory Lösungen nur im Bereich von Millisekunden feststellbar (vgl. Plattner und Zeier 2011, S. 9).

Die Verfügbarkeit ist eine Kernfunktion einer IMDB. Ihre Aufgabe ist, eine schnelle Bereitstellung und die Sicherheit der Datenbank im Fall eines Systemausfalls, eines Fehlers oder eines Schadens zu gewährleisten. Das Ziel besteht in einer vollautomatischen Hochverfügbarkeit einer IMDB, dessen Erfüllungsgrad entscheidend von den zur Verfügung stehenden Datenbankreplikationsmechanismen (s. u.) bestimmt wird. Des Weiteren kann auch eine redundante Bereitstellung von Ressourcen (Hardware, Software) dieses Ziel unterstützen.

Die Skalierbarkeit beschreibt, wie die Leistung eines Systems durch das Hinzufügen von Ressourcen (z. B. weiteren Rechnern) in einem definierten Bereich zunimmt. Es existieren verschiedene Skalierungsarten wie scale-up oder scale-out Verfahren. Bei einem scale-up Verfahren wird der Arbeitsspeicher des Rechners erhöht während bei einem scale-out Verfahren mehrere Rechner in einem System-verbund kombiniert werden, um so eine Lastverteilung zu erreichen. Dabei können einzelne Datenbanktabellen spezifischen Rechnern zugeordnet werden.

Der Administrationsaufwand bezieht sich auf die Verwaltung von Zugriffen auf ein System, die Wartung, die Installation und die Überwachung einer Lösung. Der Administrationsaufwand ist u. a. abhängig von der Komplexität der Systemlandschaft und kann durch die Bereitstellung verschiedener Administrationsinstrumente reduziert werden. Dazu gehören bspw. Instrumente zur Installationsunterstützung und Konfiguration des Systems, zur Verwaltung der Datenbanktabellen und Indizes oder zur Benutzerverwaltung und Verwaltung der Zugriffskontrolle.

Mithilfe einer Datenbankreplikation können Daten aus verschiedenen Datenbanken miteinander abgeglichen werden. Zur Vermeidung eines Datenverlustes, bspw. bei einem Systemabsturz, unterstützen In-Memory-Datenbanken Funktionen, wie Replikationen, Transaktionslog oder SavePoints. So werden geänderte Daten in regelmäßigen Abständen mit dem persistenten Speicher synchronisiert (vgl. Stirnimann und Ott 2012).

Zu unterscheiden sind eine synchrone und asynchrone Replikation. Bei einer synchronen Replikation wartet das primäre System auf das Ende einer Transaktion und der damit verbundenen Antwort bspw. in Form eines Protokolls aus dem sekundären System. Dieser Modus garantiert eine sofortige Konsistenz zwischen beiden Systemen. Jedoch kann eine Verzögerung durch die benötigte Zeit für die Datenübertragung und Persistenz im sekundären System auftreten. Sollte die Verbindung mit dem sekundären System getrennt werden oder das sekundäre System abstürzen, kann das primäre System nach einem kurzen, konfigurierbaren Timeout den Betrieb wieder aufnehmen. Das sekundäre System bleibt zwar bestehen, es wird aber nicht sofort das empfangene Protokoll übermitteln. Um eine wachsende Liste von Protokollen zu vermeiden, werden inkrementelle Daten-Schnappschüsse asynchron in bestimmten Zeitintervallen aus dem primären System an das sekundäre System übertragen. Abhängig von der verwendeten IMDB können weitere Varianten unterschieden werden.

7.4 Bewertung

Der Einsatz einer IMDB kann zu deutlichen Vorteilen im Bereich der Real-Time-Analyse führen, wenn ausreichende Systemressourcen verfügbar sind. Ein Vorteil der In-Memory-Technologie im Vergleich zu einer RDBMS ist, dass durch die permanente Datenhaltung im Hauptspeicher die Antwortzeiten bei Berichtsauswertungen erheblich verkürzt werden können. Dies gilt insbesondere bei der Auswertung großer Datens Mengen. So können z. B. 10 000 Anfragen pro Stunde gegen eine Datens Menge von 1,3 Terabyte mit Antwortzeiten von weniger als einer Sekunde bearbeitet werden (vgl. Plattner und Zeier 2011 S. 208). Um diesen Vorteil ohne Einschränkung nutzen zu können, ist ein genügend großer und auf die zu erwartende Datens Menge abgestimmter Hauptspeicher obligatorisch. Bei der In-Memory-Technologie wird somit die Kapazität der Datenbank auf die Hauptspeicherkapazität beschränkt.

Ein Vergleich der RDBMS mit einer IMDB zeigt, dass die für eine RDBMS wichtigen Operationen, wie Aggregationen, nicht mehr benötigt werden. Dies reduziert die Komplexität der Analyse, da alle Aggregations-Abfragen aus einer spaltenorientierten In-Memory DBMS beantworten werden können. Dadurch vereinfacht sich u. a. die Datenanalyse erheblich.

Beim Einsatz einer In-Memory-Technologie besteht ein wesentlich höheres Risiko eines Datenverlustes bspw. bei einem „Servercrash“, da die Daten im Hauptspeicher nur „flüchtig“ gehalten werden. Eine Ausnahme bietet die Verwendung eines NVRAM (Non-volatile RAM) zu verwenden, bei dem der Dateninhalt ohne externe Energieversorgung erhalten bleibt. Bei diskbasierten Datenbanken ist dagegen die Gefahr eines Datenverlustes bei einer Abschaltung der Stromversorgung sehr gering.

Zur Erhaltung einer hohen Verfügbarkeit einer IMDB – besonders im Falle von Störungen – können diverse Datenreplikationsmechanismen (s. o.) genutzt werden. Dazu werden zusätzliche Ressourcen benötigt, um Backups sichern zu können und somit die

Persistenz der Daten und Transaktionssicherheit zu gewährleisten. Am Markt verfügbare In-Memory-Datenbanksysteme unterstützen in ihrer Architektur verschiedene Backup-, Datenwiederherstellungs- und Ausfallsicherungssysteme, um Datenverluste zu vermeiden.

Zudem können Hybrid-Lösungen eingesetzt werden. Diese Art von Datenbanken wird als sogenannte hybride In-Memory-Datenbank bezeichnet. Derartige Datenbanksysteme können Daten sowohl im Hauptspeicher als auch auf Festplatten speichern.

In der Literatur werden einige ökonomische Effekte beim Einsatz einer IMDB hinsichtlich ihrer Realisierungswahrscheinlichkeit und Zurechenbarkeit diskutiert. Folgend sind einige Beispiele für mögliche positive ökonomische Effekte aufgeführt (vgl. Meier, M.C.; Scheffler, A.: Ökonomisch sinnhafte Bewertung):

- Wegfall von Ausgaben für Datenqualitätssicherung bei Extraktions- Transformations- Lade Prozessen (ETL),
- eine schnellere Verfügbarkeit analytischer Informationen
 - am Point of Sale, z. B. Call-Center oder Flughafen -Gate und so höhere Aus- schöpfung von Kundenwertpotenzial,
 - bei der Reklamationsbearbeitung und somit eine Erhöhung der Kundenbindung bei Kunden mit einem hohen Wertbeitragspotenzial,
- schnellere Sperrung von Konten bei Missbrauchsverdacht oder Verdacht auf Forderungsausfall und somit eine Verringerung von Einzahlungsverlusten,
- schnellere Reaktion auf Preisschwankungen an Finanz- und Rohstoffmärkten, Ver- ringerung von Out-of-Shelf-Problemen im Handel (in Kombination mit RFID),
- schnellere Reaktion auf Gerüchte in sozialen Netzwerken,
- eine höhere Motivation der Nutzer durch geringere Wartezeiten sowie
- ein positives Image durch schnellere Verarbeitung von Kundenanfragen.

Neben den o.g. möglichen Vorteilen kann der Einsatz von In-Memory-Datenbanken auch Nachteile beinhalten. So bedingt die Integration von In-Memory-Lösungen in ein Unternehmen ggf. erhebliche Migrationskosten. Weiterhin entstehen Kosten für den Erwerb der notwendigen Hardware (insbesondere für den Hauptspeicher) sowie Kosten für System- und Datenmigrationen. Ein wichtiger Kostenfaktor sind auch die teilweise hohen Lizenzkosten.

Dazu kommen die für IT-Projekte typischen Kosten für

- Implementierung,
- Aufbau neuer Schnittstellen,
- Konfiguration,
- Testing,
- Schulungen sowie für
- Supportleistungen.

Der Einsatz einer IMDB führt eventuell auch zu einer Zunahme der Komplexität der Systemadministration und bedingt ggf. auch Anpassungen im Rechenzentrum zur Klimatisierung der neuen Hardware (vgl. Meier, M.C.; Scheffler, A.: Ökonomisch sinnhafte Bewertung). Die Quantifizierbarkeit der Lagerumschlagshäufigkeit zeigt sich darin, dass sie auf einem metrischen Niveau gemessen wird.

7.5 Fazit

Der wesentliche Vorteil für den Einsatz von In-Memory-Datenbanken ist die schnelle Verarbeitung großer Datenmengen. Obwohl Abfragen mit kurzer Antwortzeit möglich sind, leiden In-Memory-Datenbanken oft unter einer eingeschränkten Skalierbarkeit, hohen Hardwareanforderungen und dem Risiko, Daten bei einer Störung zu verlieren.

Jedoch erfordert das exponentielle Datenwachstum in vielen Unternehmen den Einsatz leistungsstarker Datenbanksysteme. Insbesondere durch die zunehmende Sammlung von unstrukturierten Daten und Informationen in den Unternehmen wachsen die Herausforderungen zur Bewältigung der großen Datenmengen, welche aktuell unter dem Begriff „Big Data“ diskutiert werden. Solche enormen Datenmengen können von herkömmlichen Speichersystemen oder Datenbanksystemen nicht mehr ausreichend schnell verarbeitet und analysiert werden. Der Einsatz der relativ neuen „In-Memory Technologie“ bietet Ansätze zur Lösung dieses Problems.

Literatur

- Meier, M.C., Scheffler, A.: Ökonomisch sinnhafte Bewertung von „In- Memory- basierten betrieblichen Informationssystemen. <https://dl.gi.de/handle/20.500.12116/18385> (2020) Zugriffen: 23. Juni 2020
- Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)
- Schmitz, U.: Konzepte für eine Real-time Information Supply Chain. In: Mönchengladbacher Schriften zur wirtschaftswissenschaftlichen Praxis, Jahresband 2010/11. S. 227–240. (2010)
- Schmitz, U.: Konzeption eines wertorientierten Führungsinformationssystems, Chemnitz. (2006)
- Stirnimann R., Ott, J.: In- Memory Datenbanken im Vergleich, https://www.trivadis-training.com/sites/default/files/downloads/pr/120116_In-Memory-Datenbanken_im_Vergleich.pdf (2012). Zugriffen: 23.Juni 2020
- Vey, G., Bachmeier, M., Krutov, I.: SAP HANA on IBM eX5 Systems. <https://www.redbooks.ibm.com/abstracts/sg248086.html> (2014). Zugriffen: 03. Febr. 2014
- Veit, K., Saake, G., Sattler, K.: Data Warehouse Technologien. MITP, Heidelberg (2012)



Big-Data-Technologien

8

Christoph Quix

Zusammenfassung

Big Data stellt sowohl auf technischer als auch auf organisatorischer Ebene enorme Herausforderungen. Zur Bearbeitung der großen, heterogenen Datenmengen wurden in den letzten Jahren verschiedene Systeme und Konzepte entwickelt, die zu einer enormen Vielfalt an Big-Data-Lösungen geführt hat. In diesem Kapitel geben wir einen Überblick über die Technologien, die im Big-Data-Bereich entwickelt wurden. Zunächst betrachten wir die Aspekte Skalierbarkeit und Fehlertoleranz, die in allen Big-Data-Systemen aufgrund der verteilten, parallelen Verarbeitung gewährleistet sein müssen. In den folgenden Abschnitten diskutieren wir dann die Technologien, die zur Bewältigung der drei Herausforderungen Volume, Velocity und Variety genutzt werden können.

8.1 Einleitung

Im Kapitel Data Engineering hatten wir bereits diskutiert, dass sich die Daten-Management-Systeme in Unternehmen in den letzten Jahrzehnten zu komplexen Öko-Systemen weiterentwickelt haben, in denen es eine zunehmende Vernetzung und Abhängigkeiten zwischen den Systemen gibt. Damit einhergehend ist auch eine Zunahme der Datenmenge, der Heterogenität der Daten und der Geschwindigkeit, in der Daten erzeugt werden. Diese Merkmale können durch drei Vs beschrieben werden, die oft benutzt werden, um Big-Data-Probleme zu charakterisieren:

C. Quix (✉)

FB Elektrotechnik/Informatik, Hochschule Niederrhein, Krefeld, Deutschland
E-Mail: christoph.quix@hs-niederrhein.de

- *Volume*: Big-Data-Probleme sind zwar nicht nur durch die Datenmenge gekennzeichnet, jedoch ist dies der wichtigste Faktor.
- *Velocity*: Wenn Daten mit einer hohen Frequenz in kurzer Zeit verarbeitet werden müssen, liegt auch ein Big-Data-Problem vor.
- *Variety*: Daten liegen in vielen verschiedenen Strukturen, verschiedenen Datenformaten, verschiedenen Systemen oder auch unstrukturiert vor.

Diese Aspekte zielen also auf die technischen Eigenschaften von Big Data ab. Für diese Merkmale kann man keine harten Kriterien nennen, die Big Data von „normalen“ Daten abgrenzen. Generell definiert man Big Data als Datensätze (bzw. Fragestellungen, die man mit bestimmten Datensätzen bearbeiten will), die nicht mit traditionellen Datenverarbeitungssystemen verarbeitet werden können. „Traditionell“ bezieht sich hier auf dem Zeitraum vor dem Jahr 2000, also klassische relationale Datenbank-Technologie mit entsprechender Anwendungssoftware, die auf Single-Server-Systemen arbeiten. Anders ausgedrückt handelt es sich bei Big Data um Datensätze und Aufgaben, die neue Datenverarbeitungssysteme erfordern, die massive parallele Verarbeitung in einem Cluster mit vielen Rechnerknoten nutzen.

Zusätzlich zu den oben genannten Merkmalen werden häufig noch weitere V-Begriffe wie Veracity, Validity oder Value im Kontext von Big Data benutzt. Diese Begriffe beziehen sich eher auf die organisatorischen Aspekte von Big Data, also zum Beispiel die Wahrhaftigkeit oder den Wert der Daten. Daten gewinnen für die Wertschöpfung in betriebswirtschaftlichen Prozessen immer mehr an Bedeutung, jedoch entsteht dieser Mehrwert nur dann, wenn die Daten in ausreichender Qualität vorliegen. Dies sind Aspekte, die bei Data Governance berücksichtigt werden müssen, da sie eng mit dem Datenqualitätsmanagement und dem Management der Daten als Wirtschaftsgut zusammenhängen.

8.2 Skalierbarkeit und Fehlertoleranz

Aufgrund der Komplexität der Big-Data-Anwendungen müssen die Aufgaben in einem verteilten System von mehreren Rechnersystemen bearbeitet werden. Meist ist vorher nicht genau abzusehen, wie viele Ressourcen (Prozessoren, Hauptspeicher, Festplattspeicher usw.) zur Verarbeitung der Datenmengen benötigt werden. Daher ist es wichtig, dass man das Big-Data-System nach Bedarf *skalieren* kann.

Bei der Skalierung von Rechnersystemen unterscheidet man zwei Typen: horizontale und vertikale Skalierung. *Vertikale Skalierung* (scale up) bezeichnet die Erweiterung eines einzelnen Rechners mit mehr oder leistungsfähigeren Ressourcen, wie zum Beispiel Prozessoren oder Prozessorkerne, Hauptspeicher oder Festplatten. Da die Ressourcen eines Systems nicht beliebig erweitert werden können, sind dieser Art der Skalierung relativ enge Grenzen gesetzt. Der Vorteil bei der vertikalen Skalierung ist

allerdings, dass die Software meist nicht geändert werden muss, um von den zusätzlichen Ressourcen zu profitieren.

Horizontale Skalierung (*scale out*) dagegen ist die Erweiterung des Gesamtsystems durch das Hinzufügen von einzelnen Rechnersystemen in einem Cluster. Dabei hat man keine theoretischen Grenzen, da man zu einem Netzwerk immer einen Rechner hinzufügen kann (bzw. die Grenze ist sehr hoch). Jedoch gibt es natürlich praktische Grenzen wie Budget, Gebäude, Energie oder die erforderliche Kühlung für die Systeme.

Das Big-Data-System sollte solche Erweiterungen oder Änderungen der Infrastruktur für den Nutzer transparent machen, d. h. es sollte nicht notwendig sein, den Anwendungscode oder die Algorithmen zu ändern, wenn mehr (oder weniger) Ressourcen im Cluster zur Verfügung stehen. Der Anwendungscode muss zwar nicht bei weiterer horizontaler Skalierung angeglichen werden, jedoch müssen die Algorithmen zunächst für eine parallele Verarbeitung angepasst werden. Dies ist eine Herausforderung, da die meisten Algorithmen nur für die klassische sequentielle Verarbeitung entwickelt wurden. Mittlerweile wurden aber für viele Algorithmen, insbesondere im Bereich Datenverarbeitung und Datenanalyse, parallele Varianten entwickelt und in entsprechenden Software-Rahmenwerken implementiert, die direkt in verteilten Big-Data-Systemen eingesetzt werden können.

Im Kapitel Data Engineering hatten wir bereits besprochen, dass NoSQL-Datenbank-Management-Systeme schon für horizontale Skalierung vorgesehen sind, genau das gleiche ist für Big-Data-Systeme der Fall. Daher sind die beiden Systemkategorien „NoSQL“ und „Big Data“ auch nicht strikt trennbar, und es gibt viele Überlappungen.

Ein Big-Data-System besteht also aus einem verteilten System mit vielen einzelnen Rechnerknoten. Dies erhöht zwar die Performance des Gesamtsystems, aber auch gleichzeitig die Ausfallwahrscheinlichkeit. Wenn zum Beispiel im Durchschnitt ein Rechner ein Tag pro Jahr aufgrund von Systemfehlern oder daraus folgenden Wartungsarbeiten nicht verfügbar ist, dann heißt das für einen Cluster mit 1000 Rechnerknoten, dass dieser nur an etwa 23 Tagen¹ im Jahr fehlerfrei arbeiten kann. Je mehr Rechner im Einsatz sind, desto größer ist die Wahrscheinlichkeit, dass irgendeiner dieser Rechner ausfällt. Daher muss ein Big-Data-System auch fehlertolerant sein, d. h. wenn einzelne Knoten oder Netzwerkverbindungen ausfallen, soll das System weiterhin arbeiten können.

Dies wird in der Regel durch den Einsatz von drei Prinzipien erreicht. Zum einen wird eine *Shared-Nothing-Architektur* bei den Big-Data-Systemen eingesetzt. Das heißt, dass die einzelnen Rechnerknoten sich keine Ressourcen teilen (außer Netzwerk) und somit der Ausfall eines einzelnen Knotens nicht zu einem weitergehenden Ausfall des Gesamtsystems führt. Jeder Knoten kann für sich alleine bei einem Ausfall anderer Knoten weiterarbeiten.

Damit trotzdem die Daten des Big-Data-Systems verfügbar sind und in Berechnungen genutzt werden können, müssen die gleichen Daten auf mehreren Rechnern verfügbar

¹(364/365)¹⁰⁰⁰ × 365 ≈ 23,48.

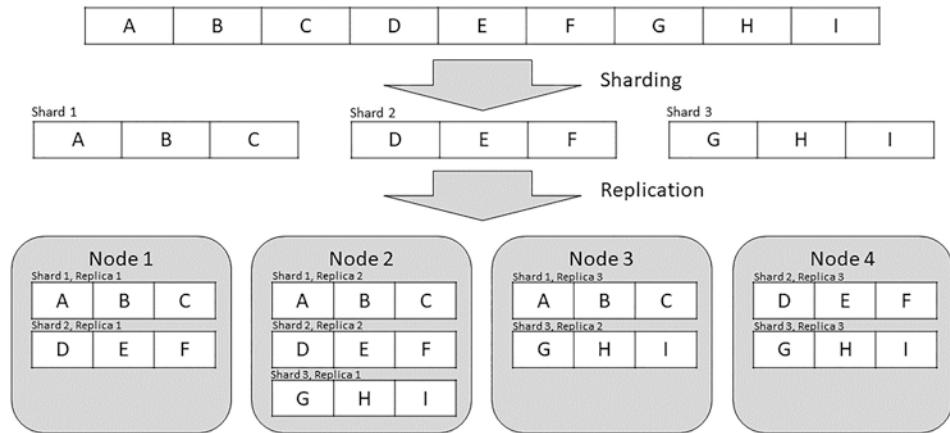


Abb. 8.1 Beispiel für Sharding und Replikation mit Replikationsfaktor 3

sein. Dies wird durch das Prinzip der *Replikation* erreicht, d. h. ein Datensatz wird auf mehrere Knoten kopiert. Dies hat zwei Vorteile. Damit kann einerseits die gewünschte Fehlertoleranz erreicht werden. Andererseits kann dadurch auch die Performance gesteigert werden, da die gleichen Daten parallel auf verschiedenen Rechnerknoten verarbeitet werden können (z. B. für unterschiedliche Anfragen). Die meisten Big-Data-Systeme nutzen gewöhnlich den Replikationsfaktor drei, d. h. die Datensätze sind an drei verschiedenen Knoten verfügbar.

Die Datensätze werden üblicherweise nicht als Ganzes zwischen den Knoten repliziert, sondern nur in einzelnen Teilen. Dafür wird das Prinzip des *Sharding* eingesetzt. Bei verteilten, relationalen Datenbanken würde man von horizontaler Fragmentierung sprechen. Dies bedeutet, dass die Zeilen einer Tabelle in mehrere Teile unterteilt wird und auf verschiedenen Knoten gespeichert wird. Idealerweise sollten diese etwa gleich groß sein, damit die Last zwischen den Knoten gleich verteilt wird. Ein Vorteil des Sharding ist auch wieder ein Performance-Gewinn durch die mögliche parallele Verarbeitung von Anfragen. Statt Sharding wird in einigen Systemen auch der Begriff *Partitioning* benutzt, was prinzipiell gleichbedeutend mit Sharding ist und sich nur im Detail unterscheidet.

Abb. 8.1 zeigt ein Beispiel für Sharding und Replikation für einen Datensatz mit neun Elementen und Replikationsfaktor 3. Der gesamte Datensatz wird zunächst in drei Shards unterteilt, dann werden die Shards auf die Knoten repliziert, sodass jeder Shard dreimal verfügbar ist. Die Strategie zur Aufteilung des Datenbestands in einzelne Shards ist auch entscheidend für die Performance der Datenverarbeitung. Zwei grundlegende Methoden werden unterschieden: Hashed und Ranged Sharding bzw. Partitioning. Beim Hashed Sharding wird für das Datenelement ein Hash-Wert berechnet und dann mit einer

Modulo-Operation der Shard bestimmt. Bei Verwendung einer guten Hash-Funktion erfolgt in der Regel eine gute, gleichmäßige Verteilung der Elemente auf die Shards. Das Ranged Sharding teilt die Daten anhand der Werte eines Attributs auf. Zum Beispiel könnten Personen anhand des Tags des Geburtsdatums aufgeteilt werden (zum Beispiel Tag 1–10 geht in Shard 1, Tag 11–20 in Shard 2 usw.). Beim Ranged Sharding muss man darauf achten, dass die Daten auf die Wertebereiche gleichmäßig verteilt sind. Bei Anfragen nach bestimmten Wertebereichen funktioniert diese Methode meist besser als Hashed Sharding, weil dann nur Daten aus einem Shard gelesen werden müssen. Neben Hashed und Ranged Sharding bieten einige Systeme noch „Custom“-Sharding-Methoden an, bei denen der Entwickler spezielle Methoden zur Datenaufteilung definieren kann.

Fällt im Beispiel von Abb. 8.1 ein Knoten aus, so existieren immer noch zwei weitere Knoten, die den gleichen Shard zur Verfügung stellen können. Big-Data-Systeme arbeiten in der Regel auch in einem Fail-Over-Modus, so dass bei einem längeren Ausfall eines Knotens dann bei anderen Knoten zusätzliche Replikas gebildet werden und die im Replikationsfaktor geforderte Anzahl von Replikas wieder für alle Shards verfügbar ist.

8.3 Volume – Management von großen Datenmengen

Für die Verwaltung von sehr großen Datenmengen (z. B. mehrere Petabytes) muss ein verteiltes System genutzt werden, da die erforderlichen Ressourcen für die Datenverarbeitung und Datenanalyse nicht in einem einzelnen Rechnersystem zur Verfügung gestellt werden können. Auch schon bei kleineren Datenmengen kann es sich lohnen in eine verteilte Dateninfrastruktur zu investieren, da man mit einem verteilten System eine höhere Performance erreichen kann.

Apache Hadoop (<https://hadoop.apache.org/>) wurde als eins der ersten Systeme entwickelt, dass die Anforderungen von Big-Data-Anwendungen erfüllen konnte, indem es die im vorherigen Abschnitt besprochen Funktionen zur Replikation und zum Sharding nutzt. Hadoop basiert auf den Ideen zum Google File System, das um das Jahr 2000 bei Google zur Verwaltung der großen Datenmengen entwickelt wurde, aber nur intern bei Google verfügbar war (Ghemawat et al. 2003). Die Entwicklungsarbeiten zu Hadoop als Open-Source-Projekt begannen etwa im Jahr 2005, eine erste Version wurde 2006 veröffentlicht. Bereits vor der Veröffentlichung der Version 1.0 im Jahr 2012 hatte Hadoop eine enorme Verbreitung erreicht, die sich zum Beispiel in den jährlichen Entwicklerkonferenzen mit mehr als 1000 Teilnehmern zeigte. Mittlerweile ist Hadoop ein fester Bestandteil der meisten Big-Data-Plattformen, auf Cloud-Computing-Plattformen ist es zum Beispiel ein bereits vorkonfiguriertes Software-Paket, dass einfach auf den gewünschten Rechnern zum Einsatz gebracht werden kann.

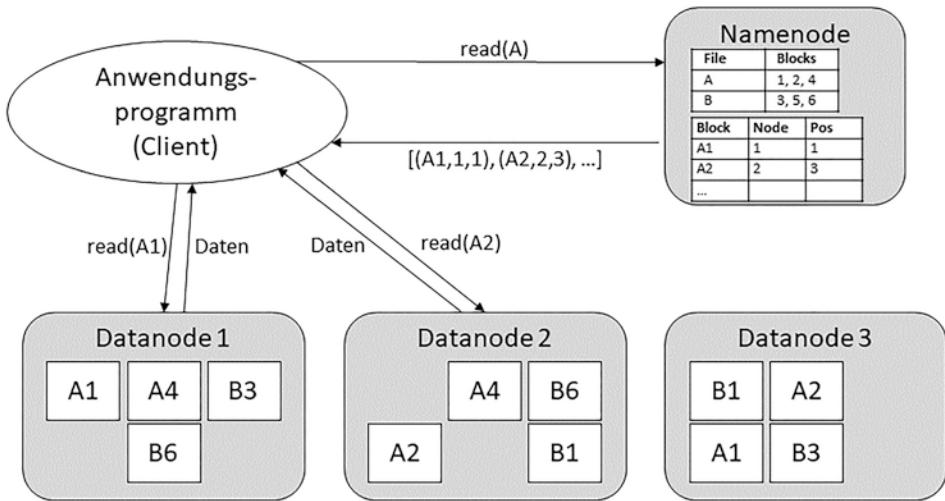


Abb. 8.2 Namenode und Datanodes in HDFS und Ablauf einer Leseoperation

Das Software-Paket Hadoop besteht aus drei Komponenten:

- *Hadoop Distributed File System (HDFS)* ist das verteilte Dateisystem, dass die Ablage von beliebigen Dateien in einem Cluster ermöglicht.
- *Map-Reduce* ist ein Programmiermodell zur parallelen Datenverarbeitung auf Basis von HDFS.
- *YARN (Yet Another Resource Negotiator)* verwaltet die Ressourcen im Cluster und steuert die Ausführung von Applikationen.

Das HDFS wird durch zwei Arten von Knoten in einem verteilten System realisiert. Ein *Namenode* agiert dabei als Metadaten-Verwalter und speichert die Informationen über die im System vorhandenen Dateien. *Datanodes* verwaltet die Dateien in mehreren Blöcken. Der Namenode ist der zentrale Anlaufpunkt für Anwendungen, die auf Daten im Hadoop-Cluster zugreifen wollen. Dieser Knoten kann bei vielen Zugriffen (insbesondere auf viele kleine Dateien) zum Flaschenhals werden. Auch ein Ausfall wäre sehr kritisch, daher wird der Fail-Over zu einem anderen Namenode unterstützt.

In Abb. 8.2 wird die Architektur von Namenodes und Datanodes und der Ablauf einer Leseoperation in Hadoop skizziert. Ein Client kontaktiert zunächst den Namenode, um für die gewünschte Datei eine Liste der Blöcke mit Adressen der Datanodes zu bekommen. Danach läuft der Datentransfer nur noch zwischen den Datanodes und dem Client. Ein Schreiboperation läuft im Prinzip ähnlich ab, am Schluss wird noch von den Datanodes über den Client eine Bestätigung an den Namenode geschickt, dass die Schreiboperation vollständig abgeschlossen ist.

Die Standardgröße für einen Block im HDFS ist 128 MB, bei regulären Dateisystemen in Betriebssystemen wie Windows oder Linux ist die übliche Blockgröße lediglich 4 oder 8 KB. Das zeigt auch, dass das HDFS für große Dateien ausgelegt ist. Bei kleinen Dateien würde viel Speicherplatz verschwendet und die zusätzlichen Interaktionen mit dem Namenode würden einen Datenzugriff relativ langsam machen. Zudem wäre der einzelne Namenode stärker ausgelastet als die diversen Datanodes.

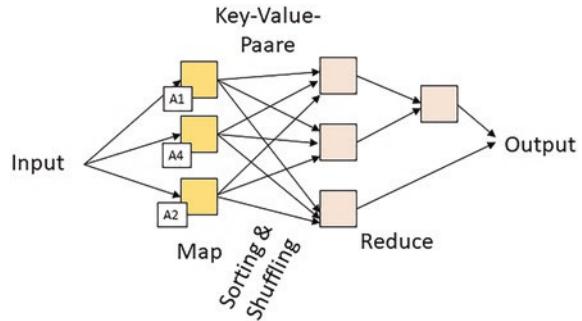
Die Dateien im HDFS sind unveränderbar bzw. es ist nur das Anhängen von neuen Datensätzen am Ende der Datei möglich, d. h. es sind keine Editieroperationen wie bei Dateien in einem lokalen Dateisystem möglich. Das Ändern der Dateien wäre durch Löschen und wiederholtes Einfügen simulierbar. Der Grund für die Unveränderbarkeit von Dateien ist die Datenintegrität, bei Editieroperationen wäre die Gefahr zu groß, dass Fehler bzw. durch parallele Editieroperationen Konflikte auftreten könnten. Für Anwendungen, die Änderungen an einem System im HDFS ablegen wollen, muss die Datenerfassung in ein Protokollierungssystem geändert werden. Statt zum Beispiel einen Wert X in einer Datei zu überschreiben, soll protokolliert werden, dass eine Operation zum Setzen eines neuen Werts von X ausgeführt werden soll. Dies erfordert ein Umdenken in der Verarbeitungslogik, vereinfacht aber das Daten-Management enorm, da nur noch sequentiell auf die Dateien zugegriffen wird und führt daher auch zu einer besseren Performance.

Das Lesen und Schreiben von Dateien sind natürlich sehr rudimentäre Operationen für ein Big-Data-System. In den meisten Big-Data-Anwendungen wollen wir mit den Detaildaten, die in den Dateien enthalten sind, bestimmte Berechnungen durchführen, zum Beispiel für eine Log-Datei eines Webservers die Anzahl der Zugriffe pro Seite berechnen. Wenn die Log-Datei sehr groß ist, kann diese Aufgabe sinnvoll parallelisiert werden. In Hadoop ist dafür das Programmiermodell Map-Reduce verfügbar, dass die parallele Verarbeitung in ein einfaches Rahmenwerk fasst. Im Wesentlichen müssen nur zwei Funktionen definiert werden:

- *Map(data) (key,value)*: Die Map-Funktion gibt für jedes Datenobjekt der Eingabemenge ein Key-Value-Paar aus. Dabei müssen die Keys nicht eindeutig sein, weil im nächsten Schritt über diese Schlüssel gruppiert wird. Die Values können beliebige Datenobjekte sein.
- *Reduce(key, values) (key, values)*: Bevor die Reduce-Funktion für jeden Schlüsselwert aufgerufen, werden die Value-Objekte für diesen Schlüsselwert in einer Liste zusammengefasst. Die Reduce-Funktion kann diese Werteliste noch weiter reduzieren, in dem zum Beispiel die Summe aller Werte dieser Liste berechnet wird.

Das Map-Reduce-Rahmenwerk übernimmt im Zusammenspiel mit dem Resource-Manager YARN die Koordination der Ausführung der Map- und Reduce-Funktionen, insbesondere die Verteilung der Ausführung dieser Funktionen auf den Datanodes. Das primäre Ziel bei der Verteilung der Jobs ist Datenlokalität, d. h. die Jobs sollen möglichst dort ausgeführt werden, wo auch die Daten vorliegen, um Netzwerktransfers zu

Abb. 8.3 Schematischer Ablauf eines Map-Reduce-Jobs



vermeiden. Um die Kosten für den Netzwerktransfer von Daten abschätzen zu können, muss Hadoop auch die Netzwerktopologie des Clusters kennen, d. h. die Anordnung von Knoten, Racks und Data Centers. Mit diesen Informationen, der aktuellen Auslastung der Datanodes und die Verteilung der Datenblöcke berechnet Hadoop dann den optimalen Ausführungsplan. Wichtig dabei ist, dass die einzelnen Knoten ihre Arbeit zunächst unabhängig voneinander durchführen können. Aus diesem Grund kann bei einem Fehler eines Datanodes ein anderer Datanode die Ausführung übernehmen (wenn Daten und Programmcode an diesen Knoten transferiert wurden).

Der Ablauf eines Map-Reduce-Jobs ist in Abb. 8.3 dargestellt. Angenommen, es soll das vorher genannte Beispiel berechnet werden: Anzahl der Zugriffe pro Webseite anhand der Daten einer Log-Datei. In der Log-Datei sind die Einträge zeilenweise in der Form (Zeitstempel, Webseite, IP-Adresse, ...). Wie in Abb. 8.2 dargestellt, besteht die Log-Datei aus drei Datenblöcken A1, A2 und A4, die auf verschiedenen Datanodes vorliegen. Zunächst wird die Bearbeitung eines Datenblocks einem Datanode zugewiesen, zum Beispiel wird A1 auf Knoten 1, A4 auf Knoten 2 und A2 auf Knoten berechnet. Die Map-Funktion wird nun für jede Zeile in diesen Datenblöcken aufgerufen. Als Ergebnis gibt die Map-Funktion den Namen der Webseite als Key zurück und als Wert 1 (für einen Zugriff), also zum Beispiel „./index.html“, 1). Bevor die Reduce-Funktion aufgerufen wird, werden die Ergebnisse der Map-Funktion gruppiert, indem alle Werte eines Schlüssels in einer Liste zusammengefasst werden, zum Beispiel „./index.html“, [1,1,1,1,1,1,1,1,1], wenn die Seite 10-mal aufgerufen wurde. Um diese Gruppierung zu erstellen ist einerseits eine Sortierung der Daten nach den Schlüsselwerten notwendig, andererseits müssen die Daten zwischen den verschiedenen Datanodes ausgetauscht werden, damit ein Gesamtergebnis berechnet werden kann. Diesen letzten Schritt nennt man Shuffling und ist üblicherweise ein aufwendiger Schritt bei Map-Reduce, da viele Daten über das Netzwerk versendet und anschließend zwischengespeichert werden müssen. Die Reduce-Funktion ist anschließend sehr einfach, da nur die Summe bzw. die Anzahl der Werte berechnet werden muss. Diese Funktion kann auch wieder unabhängig voneinander für einzelne Schlüssel auf verschiedenen Knoten berechnet werden. Wie in

Abb. 8.3 angedeutet, kann die Reduce-Funktion auch mehrmals nacheinander aufgerufen werden, wenn nicht alle Werte für einen Schlüssel in einem Schritt reduziert werden können.

Dies ist ein sehr einfaches Beispiel, das mit 5–10 Zeilen Java-Programmcode implementiert werden kann. Hadoop ist in Java implementiert, daher ist Java die primäre Sprache für die Map-Reduce-Funktionen, es können aber auch andere Programmiersprachen benutzt werden. Das Map-Reduce-Programmiermodell ist auch in anderen Systemen vorhanden (z. B. MongoDB) und nicht auf Hadoop beschränkt. Das Map-Reduce-Rahmenwerk nimmt dem Programmierer also viel von der sonst erforderlichen Koordinierungsarbeit bei der verteilten, parallelen Programmierung ab. Allerdings ist eine deutliche Anpassung der meisten Algorithmen notwendig, die einem iterativen Programmiermodell folgen, um sie für das Map-Reduce-Modell anwendbar zu machen.

Map-Reduce liefert für sehr große Datenmengen und gut parallelisierbare Berechnungen gute Ergebnisse. Bei kleineren Datenmengen, wo der Vorteil der Parallelisierung nicht so zum Tragen kommt, oder bei iterativen Algorithmen, die mehrere Map-Reduce-Durchläufe brauchen, ist ein Vorteil eventuell nicht mehr zu erkennen. Daher sollte man immer sorgfältig abwägen, ob der vorliegende Anwendungsfall wirklich für Hadoop und das Map-Reduce-Modell geeignet sind, oder ob nicht andere Datenverarbeitungsmethoden besser geeignet sind. Das oben genannte Beispiel hätte man sicherlich auch bei großen Datenmengen und einer geeigneten physikalischen Datenstruktur auch auf einem relationalen Datenbanksystem mit guter Performance berechnen können. Dafür wäre in SQL sogar nur ein Einzeiler notwendig gewesen:

```
SELECT page, COUNT(*) FROM log GROUP BY page.
```

Um die Nutzung des Hadoop-Systems zu vereinfachen, wurden mittlerweile diverse Rahmenwerke entwickelt, die die Nutzung von deklarativen Sprachen wie SQL ermöglichen und nicht die Programmierung von Map-Reduce-Funktionen in Java verlangen. Apache Hive ist ein solches Beispiel. Es ermöglicht die Erstellung eines relationalen Schemas auf Basis der in Hadoop vorliegenden Daten. Dies ist auch ein Beispiel für das Schema-on-Read-Konzept. Wenn die Tabellen definiert sind, können darauf SQL-Anweisungen ausgeführt werden. Einfache Abfragen können von Hive direkt beim Auslesen der Daten ausgewertet werden, komplexere Abfragen mit Join- oder Gruppierungsoperationen werden in Map-Reduce-Anweisungen übersetzt. Auch hierbei gilt, dass sich der Aufwand der Code-Generierung für einen Map-Reduce-Job nur lohnt, wenn sehr großen Datenmengen verarbeitet werden müssen.

In Abschn. 8.5 werden wir Apache Spark vorstellen, dass einige der Schwächen von Hadoop adressiert und ein effizienteres, hauptspeicherorientiertes Programmiermodell bietet. Zuvor diskutieren wir aber zunächst die Verarbeitung von Datenströmen.

8.4 Velocity – Kontinuierliche Verarbeitung von Datenströmen

Durch die zunehmende Digitalisierung gibt es viele Anwendungen, in denen kontinuierliche Datenströme produziert bzw. genutzt werden. Im Straßenverkehr werden durch stationäre Messstellen oder auch durch die Bewegungsdaten von Mobilfunkteilnehmern kontinuierlich Daten generiert, die zur Bestimmung der Verkehrslage oder zur Erkennung von Gefahrensituationen genutzt werden kann (Geisler et al. 2012). In der Produktion werden durch verschiedene Sensoren in den Maschinen oder Produktionshallen ebenfalls fortwährend Daten über den Produktionsfortschritt, den Status der Maschinen oder die Position von einzelnen Werkstücken generiert (Pennekamp et al. 2019). In der Medizin werden zum Beispiel auf Intensivstationen durch die verschiedenen Geräte und Sensoren durchgehend Daten zum Gesundheitszustand des Patienten generiert, aus denen man eine kritische Situation erkennen kann.

Gemeinsame Eigenschaft dieser Anwendungsbeispiele ist, dass die Daten nicht nur kontinuierlich generiert werden, sondern auch durchgehend ausgewertet und analysiert werden müssen. Je nach Anwendung gibt es unterschiedliche Anforderungen an die Verarbeitungszeiten bzw. Echtzeitanforderungen. Zur Gefahrenerkennung im Straßenverkehr oder in der Medizin sind Antwortzeiten unter einer Sekunde erforderlich, bei der Erkennung der Verkehrslage sind auch Verzögerungen von mehreren Minuten noch akzeptabel.

Eine weitere Herausforderung ist, dass man die Daten aufgrund der Datenmengen und geforderten Antwortzeiten zur Auswertung nicht in einem Datenbanksystem zwischen-speichern kann. Dennoch sollte die Auswertung der Datenströme ähnlich erfolgen wie bei statischen Datensätzen: durch deklarative Anfragen wie zum Beispiel in SQL oder durch Machine-Learning-Algorithmen zur Erkennung von komplexen Mustern in den Daten. Das zugrunde liegende Datenmodell kann bei Datenströmen ähnlich heterogen sein wie bei normalen Daten:

- relationale Daten, die zeilenweise im Datenstrom eintreffen,
- semi-strukturierte Daten wie JSON-Objekte, die objektweise im Datenstrom verarbeitet werden, oder auch
- unstrukturierte Daten, die in irgendeinem Datenformat eintreffen, zum Beispiel Twitter-Nachrichten, Audio- oder Video-Streams.

Neben den eigentlichen Datenobjekten werden bei den Datenströmen noch verschiedene Zeitstempel betrachtet, zum Beispiel, wann ein Datenobjekt erstellt wurde oder wann es im System eingetroffen ist.

Um das Ergebnis von Anfragen berechnen zu können, ist für einige Operationen eine begrenzte Datenmenge erforderlich. Datenströme sind kontinuierlich und daher zunächst nicht begrenzt. Für die Berechnung eines Joins zwischen zwei Datenströmen müssen aber endliche Teilmengen des Datenstroms betrachtet werden. Dafür können sogenannte Windows (Zeitfenster) definiert werden.

- Ein *Tumbling Window* ist ein Zeitfenster, das für einen bestimmten Zeitraum alle Datensätze enthält, deren Zeitstempel in diesen Zeitraum fällt. Ein Tumbling Window für 30 s zum Beispiel enthält Datensätze, die innerhalb eines 30-s-Intervalls eingetroffen sind und wird alle 30 s aktualisiert.
- Ein *Sliding Window* enthält ebenfalls die Datensätze eines bestimmten Zeitintervalls, kann aber häufiger aktualisiert werden. Zum Beispiel enthält ein Sliding Window mit der Größe 30 s und der Schrittweite 5 s, die Datensätze eines 30-s-Intervalls, wird aber alle 5 s aktualisiert.

Die Verarbeitung von Datenströmen und damit auch die Definition von Windows und Anfragen erfolgt in einem Data-Stream-Management-System. Nicht nur der Name, sondern auch die Funktionalität eines Data-Stream-Management-Systems ist ähnlich zu der eines Datenbank-Management-Systems. Der wesentliche Unterschied liegt in der Anfrageverarbeitung und dem Datenzugriff. Während in einem Datenbank-Management-System die Anfragen zu einem bestimmten Zeitpunkt ausgeführt werden, werden die Anfragen in einem Data-Stream-Management-System kontinuierlich berechnet. Ein weiterer Punkt ist der Datenzugriff: in einem Datenbank-Management-System kann man beliebig auf die Daten zugreifen, bei Datenströmen ist nur ein sequentieller Zugriff entsprechend der Reihenfolge der Daten im Datenstrom möglich.

Im Kontext der Verarbeitung von Datenströmen gibt es noch zwei weitere Systemarten, die relevant sein könnten. Time-Series-Datenbank-Systeme fokussieren auf die effiziente Speicherung von großen Zeitreihen, als quasi einem Ausschnitt aus einem Datenstrom. Dabei ist vor allem die Komprimierung der Daten wichtig. Complex-Event-Processing-Systeme sind sehr ähnlich zu Data-Stream-Management-Systemen, verarbeiten aber komplexe Ereignisse mit komplexen Datenstrukturen, wohingegen Data-Stream-Management-Systems eher bei Datenströmen einfachen, gleichförmigen Daten eingesetzt. Die Ereignisse erfolgen auch weniger regelmäßig als Datensätze in einem Sensordatenstrom. Die Übergänge zwischen Data-Stream-Management-Systemen und Complex-Event-Processing-Systemen sind aber fließend, daher sind die Systeme schwer klar einer Kategorie zuzuordnen.

Ein Beispielsystem für die Verarbeitung von Datenströmen ist Apache Kafka (<https://kafka.apache.org/>), welches, wie andere Big-Data-Systeme auch, auf einer verteilten Architektur aufbaut. Kafka war ursprünglich als ein Message-Broker-System geplant, hat sich aber mit der Zeit in eine effiziente Plattform für die Verarbeitung von Datenströmen weiterentwickelt. In Abb. 8.4 ist die Architektur eines Kafka-Systems dargestellt. Den Kern bildet der Kafka-Cluster, der aus mehreren Broker-Systemen besteht. In Kafka werden die Daten in Topics unterteilt, und diese wiederum in Partitionen, gemäß dem Sharding-Konzept. Ein Topic enthält in der Regel Daten der gleichen Art, kann also mit einer Relation in einem relationalen Datenbanksystem verglichen werden. Die Datenobjekte selbst können eine beliebige Struktur haben, Kafka fügt den Datenobjekten noch verschiedene Zeitstempel hinzu, worüber dann die Zeitfenster definiert werden können.

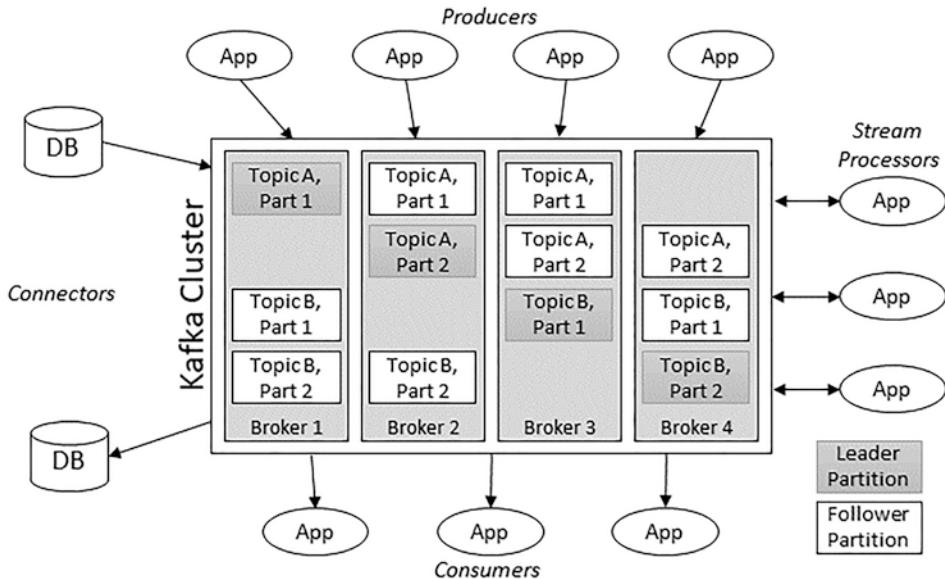


Abb. 8.4 Architektur eines Kafka-Systems

Die Daten werden über Producer in die Topics geschrieben und über Consumer ausgelesen. Ein Consumer kann sich für ein bestimmtes Topic anmelden und bekommt dann einen kontinuierlichen Datenstrom der Elemente in diesem Topic. Die Stream-Processor-Anwendungen können die Daten in einem Topic auslesen, verarbeiten und das Ergebnis wieder in ein anderes Topic schreiben. Daten können auch aus Datenbanken gelesen bzw. geschrieben werden.

Die Skalierbarkeit und Fehlertoleranz werden durch mehrere Broker-Systeme erzielt, die den Kafka-Cluster bilden. Die Partitionen der Topics werden auf verschiedene Broker verteilt, ähnlich wie die Verteilung von Datenblöcken auf Datanodes bei Hadoop. In Kafka wird dabei eine Partition als „Leader“ ausgewählt, in der die Daten zunächst geschrieben werden, weitere Partitionen können „Follower“ sein. Fällt ein Broker aus, der eine Leader-Partition verwaltet, dann kann eine andere Follower-Partition zum Leader ernannt werden.

Neben verschiedenen Programmierschnittstellen bietet Kafka auch eine Anfragesprache KSQL, mit der kontinuierliche Anfragen in einer Syntax ähnlich wie SQL definiert werden können. Die Anfragesprache ist in der Komponente ksqlDB implementiert, die darüber auch die Definition von komplexen Verarbeitungsprozessen realisieren, für die man sonst verschiedene Producer, Stream-Processor-Anwendungen oder Consumer implementieren müsste.

Kafka ist nur ein Beispiel für ein Data-Stream-Management-System, es gibt viele weitere Beispiele, die ähnliche Konzepte unterstützen wie Kafka. Wie in vielen anderen Bereichen von Big-Data-Systemen entwickeln sich auch die Systeme im Bereich der

Datenstromverarbeitung rapide weiter. Da jedes System bestimmte Stärken hat, werden oft auch mehrere Systeme in Kombination eingesetzt. Im nächsten Abschnitt betrachten wir mit Apache Spark ein Big-Data-System, das verschiedene Arten der Big-Data-Verarbeitung in einem einheitlichen Rahmenwerk vereint und damit vor allem das Problem der Heterogenität auf verschiedenen Ebenen löst.

8.5 Variety – Unterstützung für die Zusammenführung von heterogenen Daten

In den vorhergehenden Abschnitten haben wir gesehen, dass nicht nur die Daten bei Big-Data-Systemen sehr heterogen sind, sondern dass sich auch die Systemlandschaft bei Big Data aus mehreren komplexen Systemen zusammensetzt. Das macht die Konfiguration und Wartung eines Big-Data-Clusters zu einer enormen Herausforderung. Neben dem Problem der Konfiguration der einzelnen Knoten und Systeme, muss dafür auch das Zusammenspiel der verschiedenen Komponenten konfiguriert werden. Da sich die einzelnen Systeme mehr oder weniger schnell weiterentwickeln, kommt es dabei auch immer zu Inkompatibilitäten aufgrund von nicht passenden Versionen. Dazu kommen noch die verschiedenen Programmiermodelle, -sprachen und -schnittstellen, die von den einzelnen Systemen genutzt werden. Das macht die Entwicklung einer Big-Data-Anwendung zu einer vielseitigen Herausforderung.

In den letzten Jahren hat sich Apache Spark (Zaharia et al. 2016, <https://spark.apache.org/>) als eine Big-Data-Plattform herausgestellt, die zu einer Standardkomponente in vielen Big-Data-Systemen geworden ist, weil sie viele Aspekte der verschiedenen Systeme in einem einheitlichen Modell mit einfachen Schnittstellen zusammenführt. Die Nutzung von Apache Spark löst zwar nicht das Problem des Zusammenspiels zwischen den verschiedenen Komponenten, vereinfacht aber die Entwicklung einer Big-Data-Anwendung enorm.

Ein weiterer Grund für den Erfolg von Spark ist die Performance des Systems. Das Ziel der Datenverarbeitung in Spark ist es, einen möglichst großen Teil der Operationen alleine durch Zugriffe auf den Hauptspeicher durchzuführen und Zugriffe auf den Sekundärspeicher (d. h. Festplatten) zu vermeiden. Dieses Prinzip gibt es in der relationalen Datenbanktechnologie schon seit den 1970er Jahren, weil Festplattenzugriffe damals wie heute (trotz des technischen Fortschritts) deutlich langsamer sind als Hauptspeicherzugriffe. Der hohe Aufwand für Festplattenzugriffe ist auch ein Grund für die eventuell mangelnde Performance von Map-Reduce-Programmen auf Hadoop. Zwischen den einzelnen Map- und Reduce-Schritten müssen die Zwischenergebnisse jeweils in das HDFS geschrieben werden, damit sie für einen nachfolgenden Schritt, der eventuell auf einem anderen Knoten durchgeführt werden muss, verfügbar sind. Hadoop kann dabei den Gesamtprozess nicht optimieren, da die Verarbeitungslogik im Java-Code versteckt ist und nicht für eine effizientere Ausführung der Datenoperationen optimiert werden kann.

Spark verfolgt dabei einen anderen Weg. Die grundlegende Datenstruktur von Spark sind Resilient Distributed Datasets (RDDs), in der beliebige Datenobjekte in einer verteilten Speicherstruktur in einem Cluster verwaltet werden können. RDDs werden prinzipiell im Hauptspeicher gehalten, bei Bedarf (wenn der Hauptspeicher nicht ausreicht) können die Daten auf Festplatte ausgelagert werden. Spark übernimmt dabei auch die Kontrolle, welche Daten auf Festplatte ausgelagert werden und welche im Hauptspeicher bleiben. Im Prinzip stellt Spark damit für die Rechnerknoten in einem Cluster eine Art *Shared Memory* zur Verfügung, dass für die Verarbeitung von großen Datensätzen genutzt werden kann. Die Verteilung der Daten auf die einzelnen Knoten erfolgt dabei wie bei anderen Big-Data-Systemen nach dem vorher beschriebenen Konzept der Partitionierung (z. B. mit Hashed oder Ranged Partitioning). Im Gegensatz zu Hadoop orientiert sich die Größe und Anzahl der Partitionen nicht an der Datengröße, sondern an der Anzahl der verfügbaren Prozessorkerne im Cluster. Die Idee dabei ist, dass jede Partition durch einen Prozessorkern verarbeitet werden kann, um so die maximale Parallelität zu erreichen. Sind für einen Spark-Job zehn Prozessorkerne verfügbar, so würden für den Datensatz (unabhängig von der Größe) mindestens 10 Partitionen gebildet (als Faustregel wird empfohlen, etwa 2–3-mal so viele Partitionen zu nutzen wie Prozessorkerne zur Verfügung stehen). Das Erstellen der Partitionen erfolgt beim Laden des Datensatzes in den Hauptspeicher.

Um den gesamten Verarbeitungsprozess zu optimieren, muss Spark die Verarbeitungslogik kennen und entsprechend den physischen Gegebenheiten (Anzahl der verfügbaren Knoten, Anzahl der Prozessorkerne, verfügbarer Hauptspeicher usw.) den Ausführungsprozess optimieren. Spark-Anwendungen werden zwar auch in einer beliebigen Programmiersprache implementiert (z. B. Java, Scala, Python), die Datenverarbeitungsoperationen (bzw. die Anfrage) werden entweder durch deklarative SQL-Anweisungen in SparkSQL definiert oder durch entsprechende logische Operationen in der Spark-Programmierschnittstelle. Die Optimierung der Anfrage erfolgt dann wieder nach dem gleichen Prinzip wie bei Datenbank-Systemen:

- Die logische Darstellung der Anfrage wird zunächst auf Redundanzen überprüft und in eine optimale logische Struktur überführt.
- Aus der optimalen logischen Darstellung erstellt Spark dann mehrere physische Ausführungspläne.
- Die physischen Pläne werden mit einem Kostenmodell bewertet, der beste Plan wird zur Ausführung ausgewählt und dann schließlich ausgeführt.

Die Definition der Anfragen erfolgt allerdings nicht direkt auf der RDD-Datenstruktur, sondern auf einer abstrakteren Darstellung, den DataFrames. DataFrames in Spark sind vergleichbar mit relationalen Tabellen oder auch den DataFrames in Python bzw. Pandas. Neben tabellarischen Daten können in den DataFrames aber auch semi-strukturierte Daten, also JSON- und XML-Dokumente dargestellt werden. SparkSQL erlaubt entsprechende Operationen auf den verschachtelten Strukturen, zum Beispiel auch das

Aufbrechen der verschachtelten Struktur oder das Gruppieren von relationalen Daten in eine verschachtelte Struktur.

Durch diese einheitliche Darstellung von verschiedenen heterogenen Datenstrukturen können Daten aus unterschiedlichen Datenbank-Systemen (relational oder NoSQL) oder Datenformaten in einem einheitlichen Rahmenwerk verarbeitet werden. Damit eignet sich Spark sehr gut zur Datenintegration, aber auch zur Datenaufbereitung. Zusammen mit den Eigenschaften zur parallelen Verarbeitung von großen Datenmengen im Cluster, ist Spark daher eine ideale Basis für das Transformation Layer in der Data-Lake-Architektur. Spark hat keine eigene persistente Speicherfunktion, aber aufgrund der Popularität wurden mittlerweile eine Vielzahl von Konnektoren zu diversen Datenbank-Systemen entwickelt. Eine übliche Kombination ist Spark für die Datenverarbeitung und Hadoop für die Datenspeicherung.

Das folgende Beispiel stellt den erforderlichen Scala-Code dar, der zur Berechnung des vorherigen Szenarios (Anzahl der Besuche pro Webseite) erforderlich ist. Im Beispiel wird eine CSV-Datei gelesen, für das Lesen eines anderen Dateiformats oder einer Relation aus einer Datenbank wären nur kleine Anpassungen bei der read-Anweisung notwendig. Die nachfolgende Bearbeitung im DataFrame wäre identisch.

```
val logDataFrame=spark.read.option("inferSchema", "true") .  
    .option("header","true").csv("/path/to/file/server.log") .  
val sqlFrame="SELECT page, COUNT(*) FROM logDataFrame GROUP BY page"  
val apiFrame=logDataFrame.groupBy("page").count()
```

Die beiden DataFrames sqlFrame und apiFrame illustrieren die unterschiedlichen Möglichkeiten einen DataFrame zu definieren. Beide Wege führen zum gleichen Ergebnis und werden auf die gleiche Weise ausgeführt. Komplexere Abfragen mit Joins von verschiedenen DataFrames sind auch möglich.

Spark stellt nicht nur diese grundlegenden Daten-Management-Operationen bereit, sondern bietet im Basissystem auch Funktionen zum Machine Learning, zur Arbeit mit Graphen und zur Verarbeitung von Datenströmen an. Da alle Komponenten aus einer Hand kommen, ist die Integration auch besser als bei der Kombination von separaten Big-Data-Systemen. Vor allem die durchgehende Nutzung der DataFrame-Datenstruktur vereinfacht die Entwicklung von Big-Data-Anwendungen, da jeweils die gleichen Operationen zur Verfügung stehen. Zum Beispiel kann ein DataFrame für einen Datenstrom erstellt werden, dazu kann ein Zeitfenster definiert werden, das wiederum in einem DataFrame dargestellt wird. Dieser DataFrame könnte dann in einer Join-Operation mit einem DataFrame aus einem NoSQL-System genutzt werden, dessen Ergebnis auch wieder ein DataFrame ist, der als Eingabe für einen Machine-Learning-Algorithmus benutzt werden kann. Auch hier ist das Ergebnis wiederum ein DataFrame, der dann in ein Datenbank-System geschrieben werden könnte. Diese Anwendung könnte auf einem einzelnen Rechner unter Nutzung mehrerer Prozessorkerne ausgeführt werden oder auch

in einem Cluster mit hunderten Rechnerknoten. Der Programmcode müsste dafür nicht angepasst werden, da auch hier die verteilte Verarbeitung vollkommen transparent für den Anwendungsentwickler ist.

8.6 Fazit

Der Aufbau eines Big-Data-Plattform in einem Unternehmen erfordert die Kombination von mehreren separaten Big-Data-Systemen, da es nicht ein einzelnes System gibt, das alle Anforderungen abdeckt. Apache Spark hat zwar einen relativ großen Funktionsumfang, aber für die persistente Datenspeicherung müssen andere Systeme genutzt werden. Hadoop ist stark bei der Datenspeicherung und der parallelen Verarbeitung mit Map-Reduce, für Machine Learning oder Datenstromverarbeitung müssen wiederum andere Komponenten genutzt und mit Hadoop integriert werden. Replikation und Partitionierung von Daten ist eine Gemeinsamkeit in vielen Big-Data-Systemen, um die gewünschte Skalierbarkeit und Fehlertoleranz zu erreichen.

Das Thema Big Data ist selbst für die schnelllebige Informatik immer noch ein junges Forschungs- und Entwicklungsgebiet. Nur wenige Systeme können auf eine Historie von mehr als zehn Jahren verweisen. Apache Spark hat seine Popularität in den letzten vier bis fünf Jahren gewonnen, seitdem in Version 2.0 die DataFrame-Funktionen deutlich erweitert wurden. Im Bereich der Data-Stream-Management-Systeme sind die Entwicklungen noch rapide. Aus diesem Grund muss die Entwicklung der existierenden Systeme und von neuen Systemen immer aufmerksam beobachten, um für einen neuen Anwendungsfall die jeweils aktuell beste Lösung auswählen zu können. Für die Zukunft ist zu hoffen, dass sich das Feld der Big-Data-Systeme etwas weniger dynamisch und individuell weiterentwickelt, sondern vor allem auf die Interoperabilität von verschiedenen Systemen geachtet wird.

Literatur

- Ghemawat S., Gobioff H., Leung S.-T.: The Google File System. In:Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP). Bolton Landing, NY (2003)
- Geisler, S., Quix, C., Schiffer, S., Jarke, M.: An Evaluation Framework for Traffic Information Systems Based on Data Streams. Transp. Res. Part C **23**, 29–55 (2012)
- Pennekamp J., Glebke R., Henze M., Meisen T., Quix C., Hai R., Gleim L.C., Niemietz P., Rudack M., Knappe S., Epple A., Trauth D., Vroomen U., Bergs T., Brecher C., Bührig-Polaczek A., Jarke M., Wehrle K.: Towards an Infrastructure Enabling the Internet of Production. In Proc. IEEE International Conference on Industrial Cyber Physical Systems (ICPS), S. 31–37. Taipei, Taiwan (2019). DOI: <https://doi.org/10.1109/ICPHYS.2019.8780276>
- Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I.: Apache Spark: a unified engine for big data processing. Commun. ACM **59**(11), 56–65 (2016). <https://doi.org/10.1145/2934664>



Information Data Models: Das Fundament einer guten Information Strategy

Christian Rupert Maierhofer

Zusammenfassung

Dieser Beitrag beschäftigt sich mit den Voraussetzungen und Leitlinien einer nachhaltig erfolgreichen Data Science Strategie. Sie soll angehenden Data Scientists und Data Architects einen Einblick in die aktuelle Wirtschaftssituation vermitteln, erklären warum aktuelle Verhältnisse normativ wirken und deren Ursprung erklären. Es wird erarbeitet, dass die Hürden zu einer nativ-homogenen Informations-Architektur systemisch sind und die Möglichkeit des Überspringens dieser Hürden maßgeblich im Bereich der Bereitschaft zur Veränderung der handelnden Individuen begründet liegen.

Weiter wird aufgezeigt, in welcher technologischen Phase der Informations-Architektur die Mehrzahl der Akteure eingeordnet werden können, der „Application Centric“ Ära. Nach einer Definition der Begrifflichkeiten Application Centric, Data Driven und Data Centric Architecture wird wiederlegt, dass die „Data Driven“ Ära eine Übergangsphase zu Data Centric Architecture darstellt und nahtlos ist. Aufgezeigt wird ebenfalls, dass eine gewisse Größe von Strukturen und deren gewachsene Organisationen ein großes Hemmnis darstellen und warum sich Startups oft mit diesen Problemen nicht so massiv konfrontiert sehen. Dem Leser wird ebenfalls vermittelt, wie Quick-Wins erzeugt werden um ein entsprechend positiveres Mindset zu erzeugen. Abschließend wird dargelegt, weshalb die Public Cloud ein Weg aus der Zwickmühle sein kann.

Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind – Albert Einstein (vgl. Einstein o. J.)

C. R. Maierhofer (✉)

A/V Software Solutions 360°, IT-Systemhaus Bonn/Köln Bechtle GmbH & Co. KG, Bonn, Deutschland

E-Mail: christian.maierhofer@bechtle.com

9.1 Drei Thesen aus Sicht eines Praktikers

Die nachfolgenden Thesen basieren auf mehr als 30 Jahren Berufserfahrung des Verfassers im Bereich „IT Business Architecture“ und könnten als axiomatisch betrachtet werden:

- 1 Die Informations-Architekturen großer Unternehmen sind nicht selten das reinste Chaos. Solange der Ursprung dieses Problems nicht erkannt und sich der Wille zur Beseitigung formt, wird sich die Gesamtlage weiter verschlimmern und schlussendlich in einem Informationschaos enden.
- 2 Informations-Architekturen spiegeln in einem großen Maße die gewachsene Organisationsstruktur eines Unternehmens wieder. Daher nehmen Hierarchien (Rollen & Rechte) viel Raum ein – sie werden durch Applikationen und deren Benutzbar- und Verfügbarkeit repräsentiert. Diese schränken Möglichkeiten der automatischen barrierefreien Verarbeitung ein.
- 3 Gewachsene Systeme unterliegen, wie der Mensch auch, einer Selbst-Entwicklung in mehreren Stufen. Mitglieder des Systems, die einen existenziellen Wert in Ihrer Systemrolle erkennen, reagieren selbstbeschützend auf Veränderungen, die Ihren Status Quo bedrohen.

Der unreflektierte Selbstschutz entspricht der Reaktion eines Menschen bis zum zwölften Lebensjahr (vgl. Cook-Greuter 2008). Allerdings sind die Entwicklungsschritte einer Organisation nicht zwanghaft linear und können durch den direkten Einfluss des Managements sprunghaft sein.

Der evolutionär-psychologisch anmutende Ansatz mag im ersten Augenblick kontext-fremd wirken – ist er aber de facto nicht. Denn er ist die Basis für das Verhalten, welches in einer Organisation vorherrscht – ein Growth oder Fixed Mindset (vgl. Abb. 9.1).

Das richtige Mindset wird entscheidend sein für den Erfolg einer nativen Data Science Strategie (vgl. Abb. 9.1) da sie eklatante Veränderungsprozesse voraussetzt und zugleich auslösen wird.

Da sich das Innovations-Karussell der IT sehr schnell dreht, erleiden Begrifflichkeiten zügig diffusen Charakter und werden noch lange benutzt obwohl sich der Kern der Sache schon längst verändert hat. Daher ist eine Definition hier verwendeter Begriffe notwendig, um von derselben Sache zu sprechen und um klare Abgrenzungen ziehen zu können.

Für diesen Beitrag werden die folgenden Definitionen für „Big Data“ und „Data Mining“ zugrunde gelegt:

- Big Data befasst sich mit der Extraktion größerer Datenmengen aus existierenden Applikationen, sowie herkömmlichen Methoden und Tools, die sich nicht effizient und in einem akzeptablen zeitlichen Rahmen verarbeiten lassen.

Growth Mindset

Fixed Mindset



Abb. 9.1 Mindset

- Data Mining ist der Vorgang der Analyse von Daten in Bezug auf Relationen und Erkenntnissen. Daraus leitet sich der Beruf des Datenwissenschaftlers ab – der Data Scientist.

Während Big Data die geeignete technische Plattform für eine mögliche Verarbeitung zur Verfügung stellt, ist Data Mining für den eigentlichen Vorgang der Gewinnung von Erkenntnissen aus den vorliegenden Daten zuständig und erzeugt damit den Business Value – also den Unternehmensmehrwert.

Dies mag sich trivial anhören, ist aber in einem Unternehmen ab mittlerer Größe aufwärts eine Herausforderung. Schlussendlich ist Big Data zur Zeit nichts anderes als ein Workaround der unnötig wäre, wenn die existierenden Applikationen benötigte Informationen und oder Relationen erzeugen könnten sowie ausreichend resilient und performant wären. Existierende Applikationen müssten eigenständig in der Lage sein, Data Mining zu prozessieren oder als Minimalanforderung Daten technologisch barrierefrei zu präsentieren. Um dieses zu erreichen ist ein eklatanter Schritt notwendig: **Erlöst die Daten aus der Knechtschaft der Applikationen und gewährt Ihnen eine autonome Existenz!**

Nun sprechen wir von einer „nativen Data Science Strategie“ in einer datenzentrischen Enterprise Architektur. Eine Umsetzung dieser Art bringt fundamentale Veränderungen mit sich und hat Auswirkungen auf alle Bereiche eines Unternehmens. Diese disruptive Veränderung bewältigt man nur mit einem „Willen zur Veränderung“ oder gar nicht. Beispiele einer immigranten Data Science Strategie – also einer Beipass-Lösung, bei suboptimal-organisatorischen Voraussetzungen oder Fixed Mindest, sehen wir im späteren Verlauf.

9.2 It's all about the information

Wie Richard Cobden schon sagte: „Der Erfolg hat viele Väter – der Misserfolg ist ein Waisenkind“. (vgl. Cobden o. J.) So halten sich Grafikkarten-Hersteller für den Vater des Erfolges zum Thema autonomen Fahrens, da durch die Entwicklung der GPUs KI Echtzeitberechnungen erst möglich wurden. Tesla hingegen ist stolz auf Ihre Fahrzeuge und dort implementierte Software aber in Wahrheit sind es die Abermillionen Entscheidungen von Kunden, die diese Fahrzeuge kauften und im alltäglichen Leben in Echtzeit Algorithmen mit Ihrem Fahrverhalten weiterentwickelten. Denn ein Tesla Fahrzeug fährt parallel immer autonom und vergleicht die Fahrentscheidungen des Fahrers mit seinem prognostizierten und so ist die Qualität kommender und existierender Autopiloten im großen Umfang, die Summe aller Fahrentscheidungen von konventionellen Fahrern in der pre-autonomen Phase. Dieses Beispiel zeigt plakativ, das Informationen, egal in welchem Geschäftsmodell, den mit Abstand wichtigsten Wirtschaftsfaktor darstellt. Schlussfolgernd sollte jedes Unternehmen der Gewinnung von Informationen und deren barrierefreien automatischen Verarbeitung ohne Umwege höchste Priorität einräumen.

9.3 Das Heute und seine Hürden

Schauen wir uns in der Wirtschaft um, nehmen Informationen und deren Qualität nicht den höchsten Stellenwert ein. Allerdings differieren hier Selbst-Wahrnehmung und Realität. Viele Unternehmen sind der Auffassung, dass Sie innerhalb Ihrer Informations-Architektur hervorragend aufgestellt sind. Oft bekommt man Aussagen wie: “Wir haben alle Zahlen die benötigt werden – sagen Sie mir was sie brauchen und man wird es Ihnen liefern!“. Die Wahrheit sieht allerdings gravierend anders aus.

Denn die Wahrheit ist, dass die meisten Unternehmen lediglich in den regulierten Bereichen (z. B. Finanzen) akzeptabel aufgestellt sind. Durch entsprechende Gesetze ist jedes Unternehmen verpflichtet gewisse Zahlen, Daten und Fakten nachvollziehbar zur Verfügung zu halten. Aber selbst hier sind Ausreißer möglich, wie der Skandal um die Wirecard AG gezeigt hat. Wenn man in der Lage ist mit einer Handvoll Insidern zwei Milliarden Euro über Jahre hinweg vorzugaukeln, sollte klar werden, wie angreifbar und filigran diese Informationskonzepte sind. Nachvollziehbarkeit erzeugt man durch Informationsketten, die sich selbst authentifizieren. Ein praktisches Konzept welches sich die digitalen Kryptowährungen zunutze machen. Diese trotzen schon seit längerem in einem unreguliertem Raum einer Maschinerie krimineller Energie.

Zurück zur Aussage: “Wir haben alle Zahlen, die benötigt werden“ – hat man sich ein Bild gemacht welche spezifischen KPIs sinnvoll benötigt werden, erkennt man schnell, dass alle Performance Indicators die nicht auf der Basis „Währung“ verarbeitet werden Mangelware sind. Weiter beziehen sich die meisten Informationen, die einfach abrufbar sind auf rein interne erzeugte Prozessinformationen. Simpel dargestellt: „Was sind

meine Kostenkennzahlen und was meine Verkaufserlöse.“ Alle weiteren konsumierbaren Informationen sind eigentlich nur gefilterte Derivate der eben genannten Kennzahlen in numerischen Währungseinheiten.

Frage man sich allerdings, wie sieht meine Kostenstruktur gespiegelt auf spezifische Produktionsfaktoren aus, scheitert es oft schon an der fehlenden internen Leistungsverrechnung und wenn diese doch vorhanden ist, dient sie meistens dazu sich existenziell gegenüber dem Controlling zu verteidigen und weniger um mehrwertorientiertes Data Mining zu ermöglichen.

Lässt man, in seiner Bestrebung an Informationen zu kommen, nicht nach und hat nötige Hürden übersprungen folgt nun eine Arie von Ansprechpartner aus Fachabteilungen und dort verantwortlichen Controllern. Hier werden nun Aussagen folgender Natur getroffen: „Das wissen wir auch nicht so genau“, „Das können wir gar nicht sagen, da fehlen uns Informationen der anderen Abteilung“, oder „Warum wollen Sie das wissen?“ und „Dürfen Sie diese Informationen überhaupt haben und verarbeiten – daraus könnte man leistungsbezogene Auswertungen ableiten“. Ein Evergreen ist immer noch: „Das steckt bei uns im System – ich kann Ihnen keinen Zugriff gewähren – Sie haben weder eine Lizenz noch die Berechtigung“. Diese Begründung wird im Verlauf dieser Abhandlung noch eine gravierende Rolle spielen.

Nun ist man im Dschungel gewachsener Organisationen angekommen, die auf Selbstverteidigung umschalten, wenn eine Gefahr für das System vermutet wird. Ein gutes Beispiel wie man von der höchsten Selbst-Entwicklungsstufe 6 integriert und unitive auf die Stufe 2 selbst-beschützend bzw. konformistisch zurückfällt. Wenn man allerdings Glück hat erscheinen die Informationsanforderungen die man kundtut, so unverfänglich, dass die Antwort mit einer Petabyte großen Pivot Tabelle beantwortet wird, deren Datenmodell so kryptisch erscheint wie das der nächsten Marsmission und wenig bis gar nicht selbsterklärend und dokumentiert ist.

Diese Situation setzt man nun mit der Echtzeitanalyse von autonomen Fahrverhalten in komplexen Verkehrssituationen in Beziehung und man sollte erkennen, dass es sehr wohl Entwicklungspotential im Bereich der Informationsarchitektur von Unternehmen gibt.

Weiter sprechen wir zurzeit lediglich von Prozessinformationen, die rein intern entstehen also respektive „kehren vor der eigenen Haustüre“. Modelle in denen Informationssysteme mit denen anderer externer Systeme, wie zum Beispiel dem allgemeinen oder spezifischen Markt, einem Trend oder gar zur Verfügung stehenden Informationen von Marktbegleitern, Lieferanten oder Kunden gepaart werden sollen, erscheinen gar unrealisierbar in derzeitigen Informations-Architekturen.

9.4 Wie es dazu gekommen ist

Die Antwort liegt auf der Hand und ist recht einfach, denn der Zustand ist aus der Not heraus geboren. Als es lediglich darum ging, die Schreibmaschinen durch Drucker zu ersetzen und diese zu warten, befanden sich noch alle Akteure in ihrer Komfortzone.

Problematisch wurde es erst, als höherwertige Anforderungen an die IT Abteilung gestellt wurden, d. h. das Geschäftsmodell des Unternehmens strategisch zu unterstützen. Es wurden komplizierte Problemstellungen aus dem Fachbereich an die IT weitergeleitet, die damit hoffnungslos überfordert war. Diese Situation erkannten die Hersteller von Software sehr schnell und so war nicht mehr die IT-Abteilung ihr Kunde, sondern die verantwortlichen Fachbereiche.

Aufgrund der in der Praxis weit verbreiteten Auffassung innerhalb der IT-Experten „Never touch a running system“ und der daraus fehlenden Innovationskraft, verlagerten sich Budgets aus der IT in die Fachbereiche. Die IT wurde zum Erfüllungsgehilfen der Fachbereiche degradiert und hatte deren Wünsche umzusetzen.

Um bei schon verwendeten Analogien zu bleiben, die IT baute die Straßen und Tankstellen, die Fachbereiche die Automobile. Ungünstig ist eine solche Lage, wenn man feststellen muss, dass die Zeit der Elektromobilität angebrochen ist und die existierende Infrastruktur nicht zu aktuellen Anforderungen passt. Nun ist die Informations Technologie virtuell gestaltbar und löst keine Revolutionen aus, wie z. B. Umweltaktivisten, die befürchten, dass eine Fabrik einem ganzen Bundesland das Wasser abgräbt. Aber Veränderungen werden innerhalb von Unternehmen genauso emotional und mit Einsatz aller Kräfte diskutiert, sie sind selbstbeschützende Systeme. Wie sieht nun ein Ausweg aus dem Dilemma aus und wie kommt man zu seiner nativen Data Science Strategie?

9.5 Die Enterprise Architektur

Man kann annehmen, dass eine native Data Science Strategie eine gewinnbringende Ableitung einer modernen Enterprise Architektur ist. Diese sollte auf das aktuelle Geschäftsmodell und eine mögliche Kontraktion bzw. Expansion und Akquisitionen andersartiger Informations-Architekturen ausgelegt sein.

Im weiteren Verlauf dieser Ausführungen wird kein technologischer Diskurs über die Vorteile einer Kappa gegenüber einer Lambda Architektur geführt. Beim Erscheinen dieses Buches wären sie schon veraltet. Es soll deutlich werden, dass aktuelle Big Data Lösungen nur Symptome lindern, die aufgrund einer applikationszentrischen Architektur entstehen.

Es sollte erkennbar sein, dass an der „Basis“ etwas falsch läuft. Applikationen „besitzen“ Informationen, die durch das Unternehmen kopiert und verändert werden müssen um an Erkenntnisse zu gelangen.

9.6 Drei Formen der Informations-Architektur und deren Auswirkungen

9.6.1 Das Gestern und leider noch das Heute. Der anwendungszentrierte Ansatz (The Application Centric Approach)

Durch die Verlagerung der Spezifizierung von IT Systemen in die Fachabteilungen der Unternehmen wurde ein Problemfeld ausgelöst: „Was muss eine Anwendung für **meinen** Fachbereich leisten?“. Dies wird in einer Sprache beschrieben, die eine Fachabteilung beherrscht und zwar in Form von Geschäftsprozessen und oder Arbeitsabläufen, im Idealfall modelliert mit einer Standardmodellierungssprache wie z. B. BPMN. Die Frage hingegen: „Wie muss eine Anwendung architekturell gebaut sein, um sich in meine Enterprise Architecture nahtlos zu integrieren?“ spielt dabei ein nachgeordnete bis gar keine Rolle – die Wurzel des Bösen wird genährt.

Das althergebrachte EVA-Prinzip (Eingabe / Verarbeitung / Ausgabe) ist noch gültig. Es sollte niemanden wundern, denn vor der Nutzung von Computern wurde Arbeit in Formularen organisiert, welche klassisch Arbeitsabläufe in Gang gesetzt oder gar vollständig beschrieben haben. Die human-konsumierbare Ausgabe von Ergebnissen ist in erster Linie auf das User Interface der Anwendung selbst oder gar auf den Drucker begrenzt. Aufgrund dieser Tatsache beschäftigen sich viele Unternehmen mit der Interpretation von Informationen aus zahlreichen Applikationen, um aus der Korrelation verschiedenster Metadaten sinnvolle Entscheidungen ableiten zu können. Zusätzlich werden diese auf dem Weg zur organisatorischen Entscheidungsebene individuell und opportunistisch angepasst. In regulierten Bereichen (z. B. Finanzen oder Datenschutz) in welchen „Anpassungen“ strafrechtliche Konsequenzen haben oder die Menge an Daten manuell nicht mehr beherrschbar sind, haben sich Unternehmen die „Zauberwaffe“ der Application Programming Interfaces (API) zunutze gemacht und transportieren große Mengen von Daten von einer Applikation zur anderen und häufen redundante Datenberge in ihren Applikationsspeichern an. Einen tatsächlichen Business Value erzeugen sie damit nicht. Es wird lediglich ein beschleunigter und fehlerresistenterer Verarbeitungsprozess unterstützt.

In der Realität lässt sich eine Trennung der Ansätze Application Centric und Data Driven meistens an der Bidirektionalität der Informationsflüsse und dem zahlreichen Vorhandensein von Informationen ausmachen. Denn in einem Application Centric Approach ist der „Point of Truth“ immer die Stammapplikation und es wird im Allgemeinen ein automatisches Zurückschreiben von Informationen unterbunden. Das gegenseitige iterative Anreichern von Metadaten zwischen den Applikationen bedingt ein flexibles Datenmodell der einzelnen Anwendung. Kein Hersteller von Software ist begeistert wenn Hand an die Kronjuwelen gelegt wird. Aspekte der Wartung, Support und Haftung sind Gründe die unternehmerischen Individualismus unterbinden wollen. Die Unterschiede im Sinne eines iterativen Verbesserungsprozess wird im Vergleich der Abb. 9.2 und der Abb. 9.3 sichtbar.

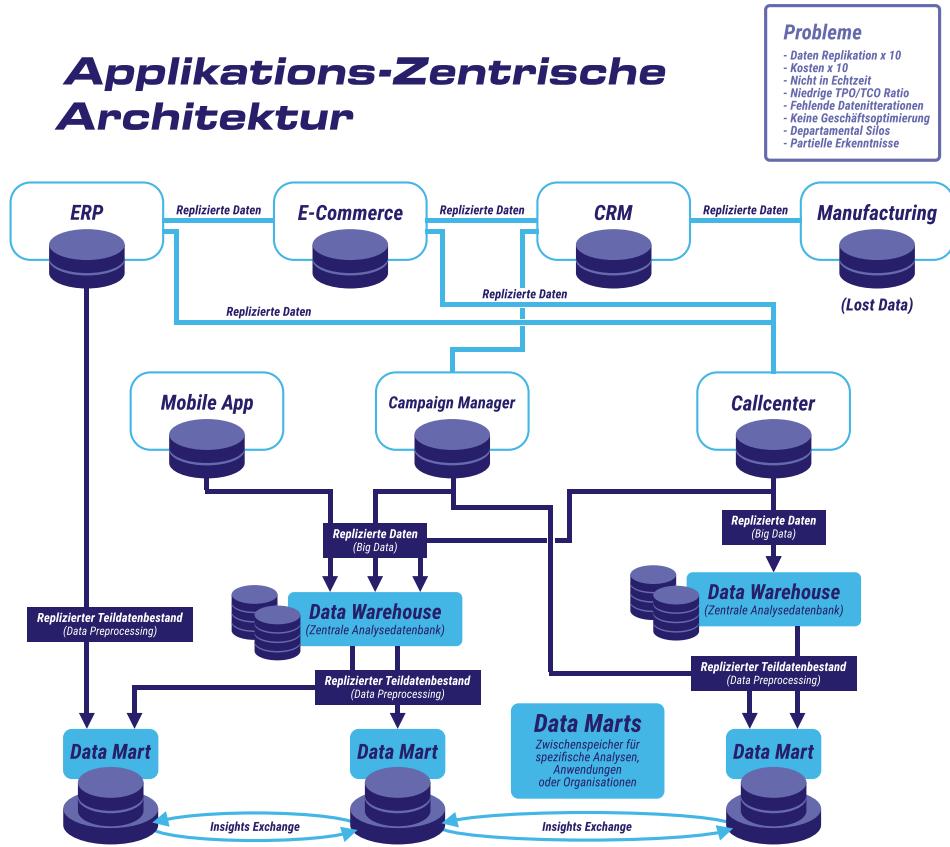


Abb. 9.2 Application Centric

9.6.2 Das Heute und die Morgendämmerung, der datengesteuerte Ansatz (The Data Driven Approach)

Dieser Ansatz hatte seine Geburtsstunde im Marketing. Es gibt kaum einen vergleichbaren Geschäftsbereich, in welchem Aktion und Reaktion respektive Investition und Gewinn so eklatant abgekoppelt sind. Wenn Apple die 101ste Marketing Aktion am Ende eines Monats im deutschen Fernsehen platziert und Mitte des nächsten Monats die Verkaufszahlen anspringen, kann keiner belastbar nachvollziehen wie sich der Einfluss dieser „einen“ Marketing Aktion im Sinne von Investition und Gewinn ausgewirkt hat. Aus diesem Grund ist das heutige digitale Marketing mit wesentlich mehr Sensorik ausgestattet als zu Zeiten der Print- und Fernsehmedien als Testfamilien noch Fragebögen über Ihre Programmwahl ausfüllen sollten um später eine interpolierte Einschaltquote zu ermitteln.

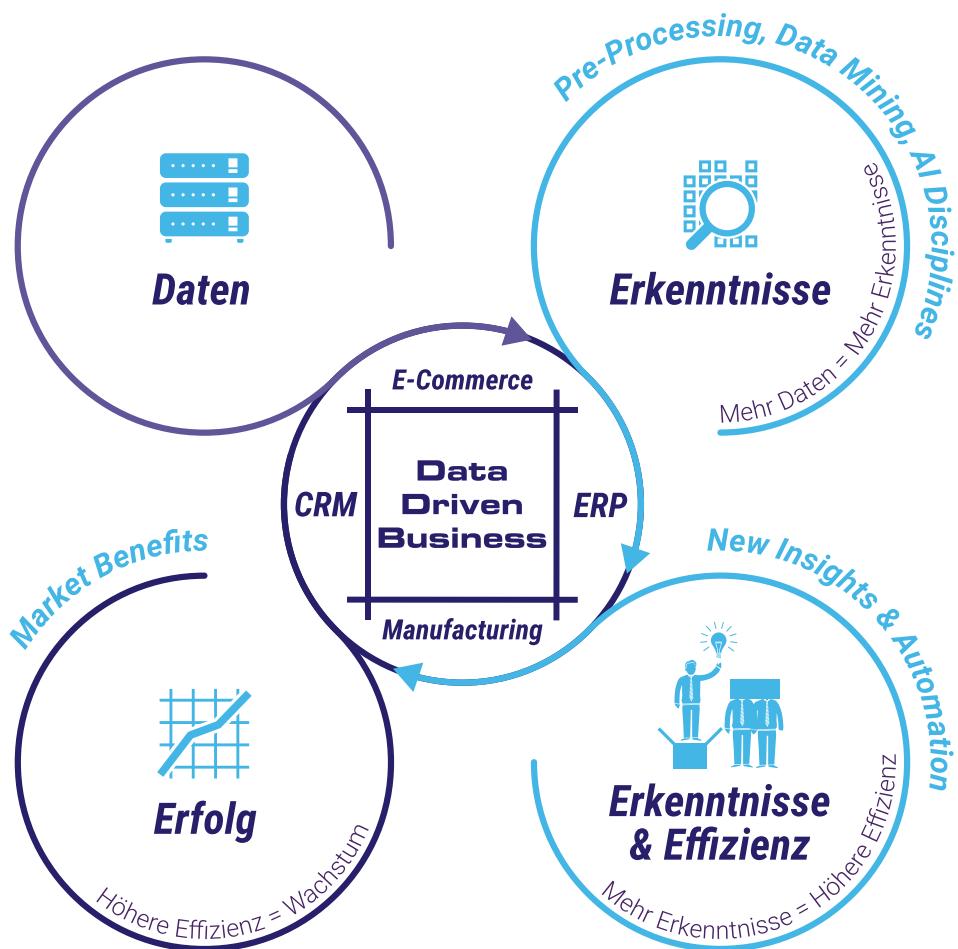


Abb. 9.3 Data Driven Business

Diese Abkopplung findet sich allerdings auch in originären Informationsarchitekturen heutiger Unternehmen wieder. Nicht selten „vermutet“ man Korrelationen zwischen Daten. Hierzu ein kleines Beispiel: „Ich habe weniger Autos produziert – also auch weniger Material verbraucht.“ Es ist allerdings ungünstig, wenn der geringere Output aufgrund von Qualitätsproblemen entstanden ist und sich der Ausschuss auf der Deponie wiederfindet.

Der datengesteuerte Ansatz hat zum Anspruch Unternehmensdaten miteinander in Bezug zu setzen. Im Idealfall sind dies alle Unternehmensdaten und verfügbare Daten aus externen Systemen. Externe Systeme sind hier Akteure die einen signifikanten Einfluss auf den Erfolg des Geschäftsmodells des Unternehmens haben, aber kein aktiver Teil der Informationsarchitektur sind. Beispiele hierfür sind Trends, Kunden,

Lieferanten, Marktbegleiter, Politik, Gesetzte und viele mehr. Ein gutes Beispiel ist hier das Unternehmen Kodak. Der unangefochtene Marktführer hat den Trend der digitalen Fotografie verpasst und stand vor dem aus – ein geplanter Umbau zum Pharmazulieferer ließ die Aktie um 2.000 % anspringen (vgl. Höfler 2020). Da hat es eine Pandemie wie SARS-CoV-2 benötigt, um Trends und Marktpotentiale mithilfe der Politik zu entdecken.

Ein heutiges Unternehmen, das von sich behaupten kann erfolgreich „Data Driven“ zu sein, hat eine hohe digitale Transparenz im Bereich der eigenen Unternehmensdaten geschaffen und diese in Korrelation zu externen Systemen gebracht. Dies geschieht aufgrund der existierende Application Centric Architecture mit Hilfe von Big Data. Wie schon erläutert, beschäftigt sich Big Data damit, große Datenmengen aus existierenden Datenquellen zu extrahieren und einer Prozessierung – dem Data Mining – zuzuführen. Nun erkennt man faktisch das Big Data ein Workaround ist der notwendig wird, weil existierende Applikationen nicht in der Lage sind diese Datenmengen zu speichern und oder zu prozessieren – resultierend aus einem applikations-architekturellem Dilemma.

Also beschäftigen sich Unternehmen heute mit der technologischen Reintegration Ihrer Fachabteilungen und Homogenisierung der Enterprise Architektur – falls überhaupt vorhanden. Sie treffen aufgrund von organisatorischem Unwillen (Mindset) und sicherheitstechnischen Auflagen (z. B. EU-DSGVO) auf eine schier unendliche Anzahl von Gegenargumenten warum diese Maßnahmen den Erfolg des Unternehmens gefährden und daher zu vermeiden sind.

Mit Hilfe von Big Data werden Data Lakes (Datenseen) oder Data Warehouses (zentrales Analyse-Datenbanksystem) sowie Data Cubes (OLAP Würfel) geschaffen auf welchen Data Mining betrieben werden kann. Legt man z. B. die gewachsene Globalität von Unternehmen zugrunde, wird klar, das Kopien von Daten, im Sinne der Echtzeitverarbeitung, so Ihre Tücken haben. Zeitverschiebung, Transportwege, Fehlertoleranzen sind echte Herausforderungen, wenn man die Verkaufszahlen aller Apple Stores weltweit in den ersten 60 min der Filialöffnung am Montagmorgen auswerten wollen würde.

Nichts desto trotz ist der Weg des Data Driven Ansatzes notwendig, um die Wahrnehmung in puncto „Informationen sind das neue Gold des 21 Jahrhunderts“ mit unternehmerischen Quick-Wins zu untermauern. Somit ist der Data Driven Approach eine gute Beipass-Lösung um die verknöcherten Strukturen des Application Centric Approachs aufzubrechen.

Durch die Erhöhung der Transparenz innerhalb der Retrospektive ergeben sich prädiktive Erkenntnisse und es erzeugen sich neue Metadaten die ebenfalls zur Umsetzung neuer Automatisierungen genutzt werden können. Das Ziel eines Unternehmens (4) durch die Iteration zwischen Geschäftsmodell, BigData (1), prädiktiven Handlungen (2) sowie Möglichkeiten der Automatisierung (3) zeigt die Abb. 9.3 schematisch

Wir sprechen von einer immigranten Data Science Strategie – also einer Strategie im Sinne der Vorbereitung. Diese Vorbereitung bezieht sich allerdings mehr auf eine iterative und agile Arbeitskultur sowie der positiven Beeinflussung des Mindsets durch Quick-Wins und weniger auf die Wiederverwendbarkeit architekturellen Komponenten.

Denn der Übergang zur datenzentrischen Architektur ist disruptiv und annihielt bisherige Applikationskonzepte.

9.6.3 Das überfällige Übermorgen, die datenzentrische Architektur (The Data Centric Architecture)

Vorweggenommen und eingestanden ist diese Informationsarchitektur in der Wirtschaft nicht im Übermaß zu finden. Es gibt wenige Beispiele größerer Unternehmen die in der Lage waren dieses Paradigma durchzuhalten. Amazon, Airbnb und Uber sind Vorreiter und erfreuen sich einer äußerst dominierenden Marktmacht. Analysiert man diese Unternehmen entdeckt man schnell, dass die Daten selbst der Business Value sind. Insofern sprechen wir von DaaS (Data as a Service) als Geschäftsmodell. Wohlweislich das die Gründer dieser Unternehmen die katastrophalen Auswirkungen des Application Centric Approach erkannt haben und oft aus der Branche Informations-Technologie selbst kommen. Jeff Bezos, Nathan Blecharczyk und Travis Kalanick haben keine Unternehmen gegründet die sich mit Büchern, Ferienwohnungen oder Taxifahren beschäftigen – sie haben INFORMATIONS-IMPERIEN gegründet.

Einen guten Überblick was unter Data Centric verstanden werden kann bietet das „Data-Centric Manifesto“ (vgl. Data Centric-Manifesto 2020). Das Leitbild dieser Architektur ist die Kernaussage, dass Informationen den Mittelpunkt von allem ausmachen und Applikationen vergänglich sind.

1. Daten sind der wichtigste Aktivposten jeder Organisation,
2. Daten sind selbstbeschreibend und benötigen keine Applikation zur Interpretation und ihrer Bedeutung,
3. Daten werden in offenen nicht prioritären Formaten präsentiert,
4. Die Erlaubnis auf Daten zuzugreifen und deren Sicherheit obliegt der Datenzugriffs-schicht und wird nicht durch Applikationen verwaltet,
5. Anwendungen dürfen Daten benutzen, ihren Zweck erfüllen und die daraus resultierenden Ergebnisse zurückschreiben damit sie von allen genutzt werden können.

Um diese Aussagen zu verifizieren werden innerhalb des Manifests folgende Erhebungen publiziert:

6. Nur eins von drei Projekten in der IT wird als erfolgreich wahrgenommen,
7. Die als erfolgreich wahrgenommenen Projekte haben nur in 20 % der Fälle einen Return of Invest,
8. In 2013 brach Kalifornien sein 208 Mio. teures Straßenverkehrsamt Projekt ohne Ergebnis ab,

9. 2012 brach die Air Force ihr 1 Mrd. teures Projekt aus 2006 ab weil Sie realisiert haben, dass es eine weitere Milliarde kostet und sie nur 20 % der erwarteten Vorteile erzielen würden,
10. Mitte 2014 wurde das Gesundheitswesen Projekt in den USA abgeschlossen mit 800 Mio. Kosten – die realisierten Funktionalitäten entsprachen 1/100stel des Wertes,
11. Datenintegration verschlingt bis zu 65 % des IT Budgets der Unternehmen (vgl. Data Centric-Manifesto 2020).

Die Darstellung in der Abb. 9.4 zeigt den schematischen Aufbau einer datenzentrische Architektur auf.

Es lässt sich leicht erkennen das alle Arten von Informationen den Kern der Architektur ausmachen. Alles wird um die Daten herum konzipiert und das was wir heute als Applikation verstehen geht in den unabhängigen äußeren 4 Ringen auf.

Eine solche Architektur in einem existierenden Unternehmen zu etablieren scheint fast unmöglich. Nicht dass es hier technologisch unüberwindbare Hürden geben würde – aber es egalisiert in großem Maße im Unternehmen existierende Software, reißt interne Führungsbereiche ein und fordert ein hohes Maß an informations-technologischem Knowhow und Eigenverantwortung. Die Zeiten, in denen man einem Hersteller vor-

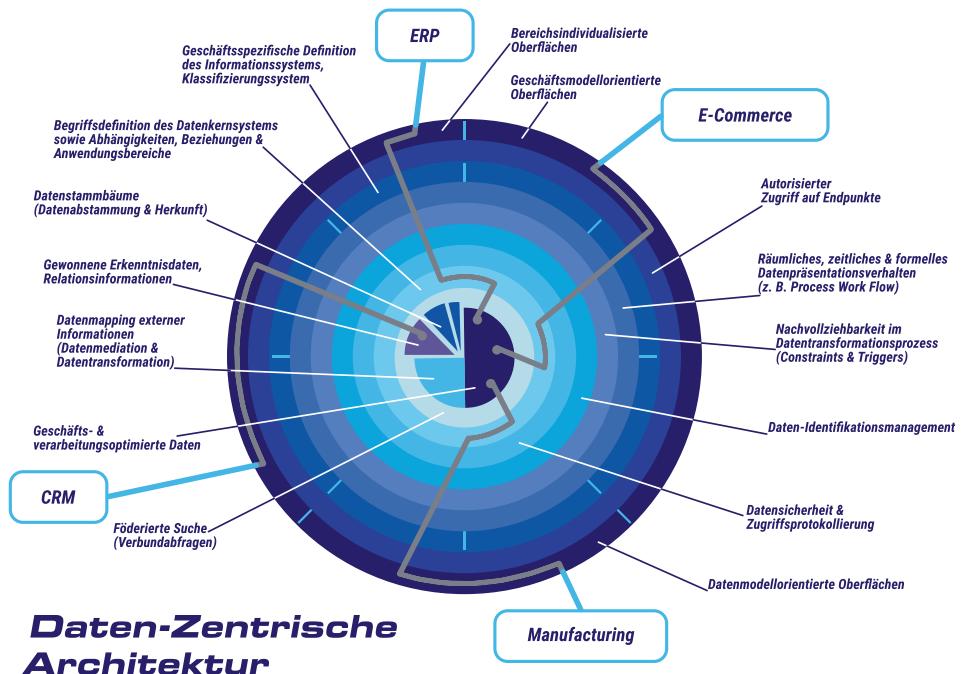


Abb. 9.4 Data Centric Architecture

werfen kann, seine Software würde Fehler produzieren sind vorbei und die Eigenverantwortung erreicht 100 %. Zu einer Zeit in welcher Experten der IT Mangelware sind und gefühlt die IT Abteilung zu groß und schon viel zu viel Budget verbraucht – obwohl man kein Digitalprodukt herstellt.

In einem Umfeld von gefühlter Sicherheit sind das keine Aussichten die Euphorie auslösen würde. Auch kein Softwarehersteller wird sich dafür begeistern können. Jeglicher Vendor-Lock-in (Herstellerabhängigkeit) ist dahin und sich automatisch verlängerte Softwarewartungen wären Schnee von gestern. Regelmäßige planbare Einnahmen sind nebenbei erwähnt einer der wichtigsten Bewertungskriterien von Softwarehersteller – es wird gebundenes Vertragsgeschäft genannt.

Zusammenfassend: die Unternehmen könnten es nur unter erschwerten Umständen, die Mitarbeiter wollen und oder können es nicht und die Softwarehersteller wollen es absolut nicht. Hier sollte man zu dem Schluss gelangen – wenn es keinen Sponsor gibt – gibt es kein Projekt. Das wäre auch der Weisheit letzter Schluss, wenn man folgende Technologien für zukünftige Geschäftsmodelle als unwichtig klassifiziert: Deep & Machine Learning, Predictive Analytics, Cognitiv Computing, Natural Language Processing, Business und Artificial Intelligence.

All diese Technologien greifen ineinander und haben fließende Übergänge. Das was sie alle verbindet ist die Notwendigkeit folgender Existenzen: maximale Qualität und Quantität von Daten – optimaler Weise in Echtzeit, verlust- und barrierefreier Zugang und hochskalierende Rechenleistung. In solchen Umgebungen können Data Scientists ihre ganzen Potenziale ausnutzen und einen echten Business Value erzeugen. Sie werden zum Vor- und Wegbereiter unzähliger Verbesserungs- und Veränderungsprozessen – sie werden die neuen Königsmacher.

„Das steckt bei uns im Systeme (1) – ich kann Ihnen keinen Zugriff gewähren (2) – Sie haben weder eine Lizenz (3) noch die Berechtigung (4)“ gehören der Vergangenheit an – denn alle vier genannten Faktoren treffen nicht mehr zu.

Disruptivität lebt von der Handlungslethargie existierender Modelle. Sind Modelle agil – unterwerfen sie sich dem notwendigen Veränderungsprozess. Startups haben geschätzt einen Agilitätsvorteil von 100:1 zu etablierten Unternehmen. Innerhalb von Startups stehen zwei Ziele hoch im Kurs A) die Finanzierung und B) der durchschlagende Erfolg der Unternehmung der die Akteure eng verbindet. Startups haben keine organisatorisch-strukturelle Erbschuld und meistens flache Hierarchien. Keine Verpflichtungen im Sinne der Kontinuität gegenüber Bestandskunden – Startups werden gegründet um gänzlich neue Geschäftsfelder zu erzeugen oder existierende Geschäftsmodelle zu erobern. Durch ein hohes Maß an Kosteneffizienz und dem immer wichtiger werdenden Faktor „Time to market“ findet man den Großteil von Startups mit ihren IT Architekturen mittlerweile in der Public Cloud wieder. Plattformanbieter wie Amazon, Google oder Salesforce bieten ganze Architektur -Ökosphären an und realisieren damit enorme Skalierungsmöglichkeiten zum einen und wirken mit Ihren – Data- und Domain Modellen normierend zum anderen.

Applikationen gehen auf in ihrem ursprünglichen EVA-Prinzip auf und Funktionen wie Integrity, Identity, Security, Federated Query & Search werden allgemeine Services die von Applikationen genutzt werden können, um Ihre Aufgaben zu erfüllen. Vorbei sind die Zeiten in denen jede Applikation ihre eigene Search Engine mitbrachte. In einem mittelständigen Unternehmen existieren durchschnittlich bis zu 42 autarke Suchmaschinen. Lustiger Weise werden diese zu 70 % von der führenden Open Source Searchengine „Elasticsearch“ konsolidiert. Auch jeder Nichtmathematiker sollte sich nun fragen warum fast 30mal dasselbe installiert und betrieben wird, wenn einmal völlig ausreichend wäre.

Damit diese Services allerdings die Anforderungen der neuen Technologie und einem Data Mining genügen, müssen diese hochskalierend und global verfügbar gemacht werden. Sie sollten cloud-native entwickelt und perfekt auf die Bedürfnisse des Geschäftsmodells angepasst worden sein (Domain Driven Design).

Eruiert ein fiktiver CxO die Erkenntnisse der architekturellen Bedingtheiten und die Notwendigkeit zur Veränderung um in der Zukunft handlungsfähig zu bleiben – kombiniert er dieses mit den hochskalierenden Anforderungen der neuen Shared-Services und einem vorherrschenden firmeninternen Fixed Mindset sowie Unwillen zur Erneuerung – dann liest man vielleicht bald ein Interview von ihm mit der Überschrift: WIR GEHEN IN DIE CLOUD – OHNE WENN UND ABER! Denn eines sollte klar sein – ähnlich wie sich in heutigen Unternehmen die Systeme vor Veränderung schützen, schützt sich der Cloud-Provider vor anwendungszentrischen Architekturen. Sie sind ineffizient, schlecht wartbar und skalieren ineffektiv. Ein Systemwechsel in die Public Cloud und Shared Services lässt sich mit der Verlegung des Firmensitzes in ein unbekanntes Land vergleichen. Es gelten ab sofort die neuen Gesetze und Verstöße werden präventiv unterbunden oder mit Geldstrafe belegt. Um im Beispiel zu bleiben: Sollte ein Unternehmen das Ziel haben ohne Veränderung der Informations-Architektur in die Public Cloud zu migrieren z.B. mit Hilfe von virtuellen Maschinen, dann sprechen wir A) nicht von cloud-native sondern vom Hosting der 90er Jahre und B) wird die Geldstrafe in Form von Consumtion (Verbrauch von Leistungen in der Cloud) bisherige Kosten in den Schatten stellen. Dieser Weg findet allerdings immer öfter Gebrauch, da Stake- und Shareholder erkannt haben, dass ein Wandel de te ipso fast unmöglich ist. Steinig und schwer wird er werden dieser Weg, allerdings in dem Bewusstsein, dass die normative Kraft der Cloud-Plattform stärker sein wird als der Unwille zur Erneuerung.

Literatur

Cobden, R.: Der Erfolg hat viele Väter. Der Mißerfolg ist ein Waisenkind. In: Aphorismen.de, <https://www.aphorismen.de/zitat/8299> (o. J.). Zugegriffen: 24. Aug. 2020

Cook-Greuter, S.: Selbst-Entwicklung -neun Stufen des zunehmenden Erfassens, Havixbeck, <https://www.rofflutterbeck.de/files/0814sd2-CookGreuter.pdf> (2008). Zugegriffen: 24. Aug. 2020

Data Centric-Manifesto: The Data-Centric Manifesto, The Need for this Change, <https://www.datacentricmanifesto.org/need/> (2020).Zugegriffen: 24. Aug. 2020

Einstein, A.: Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind – Albert Einstein. In: Poteus, Zitate für Freunde <https://www.poeteus.de/autor/Albert-Einstein/2> (o. J.). Zugegriffen: 24. Aug. 2020

Höfler, N.: Auffällige Aktienkäufe, US-Regierung setzt Gespräche über 765-Millionen-Kredit für Kodak nach Kritik aus, in Handelsblatt, 06.08.2020, Update 09.08.2020, <https://www.handelsblatt.com/finanzen/maerkte/aktien/auffaellige-aktienkaeufe-us-regierung-setzt-gespraeche-ueber-765-millionen-kredit-fuer-kodak-nach-kritik-aus/26068384.html?ticket=ST-6460115-eWCE2wibUWaQnv6D4j6f-ap2> (2020). Zugegriffen: 13.Oct. 2020

Teil III

Data Analyst: Auswerten, Präsentieren, Entscheiden – Systematische Datenanalyse im Unternehmen

- Detlev Frick; Birgit Lankes: Reporting Multidimensionaler Daten und Kennzahlen
Jens Kaufmann: Fundamentale Analyse- und Visualisierungstechniken für große Datenmengen
Jens Kaufmann: Ausgewählte Verfahren zur Analyse und Datenexploration im Big-Data-Kontext
Daniel Retkowitz: Datenbasierte Algorithmen zur Unterstützung von Entscheidungen mittels künstlicher neuronaler Netze
Peter Gluchowski, Tom Kühne, Anja Tetzner; Melanie Pfoh: Künstliche Neuronale Netze – Grundlagen, Aufbau und Gestaltungsvarianten
Thomas Neifer, Andreas Schmidt, Dennis Lawo, Lukas Böhm, Özge Tetik: Bayesian Thinking in Machine Learning



Reporting multidimensionaler Daten und Kennzahlen

10

Detlev Frick und Birgit Lankes

Zusammenfassung

Business Intelligence ist in den Unternehmen zu einem wichtigen Werkzeug im Wettbewerb geworden. Damit sind betriebswirtschaftliche Kennzahlen wieder in Fokus gerückt. Welche Kennzahlen sollten gewählt und wie können die Kennzahlen effizient ermittelt werden? Antworten dazu wurden im 1. Abschnitt zusammengefasst. Dazu sollen natürlich operative Daten in BI-Systeme transferiert werden und entsprechend aufbereitet werden. Welche Punkte dazu zu beachten sind wurden im 2. Abschnitt behandelt. Abschließend wurden noch im 3. Abschnitt die Berichtsformen und die Anforderungen an Berichte behandelt.

10.1 Betriebswirtschaftliche Motivation

Komplexer werdende Geschäftsprozesse und zunehmende Globalisierung schaffen für Unternehmen einen verschärften Wettbewerb. Um hier zu bestehen, müssen sie ihre Prozesse, Mitarbeitenden sowie die finanzielle Situation zeitnah bewerten. Neben sogenannten weichen Kriterien, werden vor allem (Kenn-)Zahlen zur Bewertung genutzt. Eben diese dienen nicht allein den jeweiligen Fachabteilungen und dem Management,

D. Frick (✉)
Duisburg, Deutschland
E-Mail: detlev.frick@hs-niederrhein.de

B. Lankes
Hückelhoven, Deutschland
E-Mail: Birgit.Lankes@hs-niederrhein.de

auch externe Investoren und Kapitalgeber legen zunehmend Wert auf eine solide Berichterstattung (vgl. Freidank 2010, S. 12).

Die Unternehmen sind herausgefordert die geeigneten Kennzahlen festzulegen – es gilt nicht nur die richtigen Kennzahlen auszuwählen, sondern auch ihre Anzahl und Berechnung festzulegen. Dabei ist zu beachten, dass nicht nur einzelne Faktoren wie z. B. der internationale Wettbewerb fokussiert werden. Die Strategieausrichtung eines Unternehmens wird durch vielfältig Einflussfaktoren bestimmt. So können u. a. Unternehmenszusammenschlüsse und Umstrukturierungen zu organisatorischen Veränderungen führen, die sich im Reporting wiederfinden müssen (vgl. Freidank 2010, S. 12).

Auch haben Unternehmen zunehmend erkannt, dass nicht allein interne Daten für eine strategische Planung ausreichen, Markt- und Kundendaten, die z. T. als unstrukturierte Daten zur Verfügung stehen, stellen eine wichtige Quelle für die Generierung entscheidungsunterstützender Informationen dar (vgl. Horváth et al. 2020, S. 189 f.). Die Auswahl der zu nutzenden Informationen und ihre Verwendung ist eine zentrale Koordinierungsaufgabe der Controller (vgl. Horváth et al. 2020, S. 189 f.). Diese Aufgabe ist durch die fortschreitende Digitalisierung geprägt und erfordert zunehmend Kompetenzen im Bereich Data Science (vgl. Abée et al. 2020, S. 46).

10.1.1 Kennzahlen und ihre Anwendung

Die unterschiedlichen Unternehmensbereiche benötigen jeweils auf ihre Bedürfnisse zugeschnittene Kennzahlen, z. B. Bilanzkennzahlen aber auch Kennzahlen für die Produktion, den Vertrieb, den Einkauf etc. Die bereichsspezifischen Ziele bilden den Ausgangspunkt für die Auswahl der jeweiligen Kennzahlen hinzukommen übergeordnete Kennzahlen, die das Management in der Entscheidungs- und Strategiefindung unterstützen. Die Nutzung und Auswertung der Kennzahlen erfolgt nach dem PDCA-Zyklus (Plan-Do-Check-Act). Nachdem eine Strategie unter Zuhilfenahme von Kennzahlen geplant wurde, erfolgt die Umsetzung im Unternehmen. Diese muss wiederum regelmäßig überprüft werden und die aus dieser Überprüfung gewonnen Erkenntnisse führen zu einer Anpassung der Strategie.

Wesentlich ist hier die Unterscheidung in die verschiedenen Planungshorizonte. So ist laut Schön die operative Planung auf einen Zeithorizont bis zu einem Jahr ausgerichtet, die Mittelfristplanung auf zwei bis drei Jahre und die strategische Planung auf etwa 5 Jahre. (vgl. Schön 2018, S. 9).

Eine Verknüpfung von Planung, Steuerung und Kontrolle gelingt i. d. R. mittels Balanced Scorecard, die die Perspektiven Finanzen, Kunden, Prozesse und Lernen und Entwicklung hinsichtlich der Ziele Kennzahlen, Vorgaben Maßnahmen darstellt. (vgl. Schön 2018, S. 27).

10.1.2 Auswahl von Kennzahlen

Um Kennzahlen für strategische Entscheidungen nutzbar zu machen, ist eine sorgfältig durchdachte Strukturierung und Organisation erforderlich. Horváth, Gleich und Seiter sprechen von der Gestaltung eines Planungs- und Kontrollsystems (vgl. Horváth et al. 2020, S. 95). Dabei sind laut Freidank abhängig vom Unternehmen und der relevanten Branche für das Reporting vorrangig die Ziele für die Aspekte Zeit, Qualität und Kosten festzulegen (vgl. Freidank 2020, S. 18–23).

Die Auswahl der zu nutzenden und nutzenstiftenden Kennzahlen ist folglich auch aus der Kostenperspektive zu betrachten, wobei es sich um die Kosten der Ermittlung eben jener – ausgewählter – Kennzahlen handelt. In der betrieblichen Praxis ist die Auswahl und Definition relevanter Kennzahlen häufig der Historie geschuldet. Einmal ausgewählte Kennzahlen werden selten auf ihre Nutzenstiftung hin überprüft. Ossola-Haring, Schlageter und Schöning empfehlen den Einsatz eines Angebots-/Nachfrage-/Bedarfs-Prüfstands (vgl. Ossola-Haring et al. 2019, S. 93 f.). Dieser ermöglicht den Abgleich der technisch verfügbaren Informationen/Kennzahlen (Angebot) mit vom Management nachgefragten Kennzahlen (Nachfrage) sowie jenen, die einen wesentlichen Beitrag zur Steuerung eines Unternehmens leisten (Bedarf). Eine Darstellung dieser Kennzahlen in Form von Schnittmengen verdeutlicht den Diskussionsbedarf mit der Unternehmensleitung, der zu einer Festlegung hinsichtlich der zu nutzenden Kennzahlen führt. Eine zyklische Wiederholung des Angebots-/Nachfrage-/Bedarfs-Prüfstands ermöglicht eine Korrektur, Entwicklung und Neuausrichtung des Kennzahlensystems (vgl. Ossola-Haring et al. 2019, S. 95).

Hinsichtlich der Kosten für Kennzahlen sind neben den ermittelten Kosten für den Arbeitsaufwand (vgl. Freidank 2020, S. 23) zur Gewinnung, Generierung und Zusammenstellung der Kennzahlen die Kosten für extern bezogene Daten relevant. Ergänzt werden diese um die Kosten für Arbeitsunterbrechungen im operativen Geschäft durch Latenzzeiten und den Arbeitsaufwand für die Analyse/Ansicht der ausgewählten und präsentierten Kennzahlen. Die zu berücksichtigen Kosten stehen häufig in direktem Zusammenhang mit dem Aspekt Zeit, so verursacht beispielsweise eine hohe Latenzzeit entsprechend hohe Kosten. Ebenso sind aufwendige Datenbereinigungen – die den Aspekt der Qualität bedienen – in der Regel sowohl zeit- als auch kostenintensiv.

Neben der Festlegung der relevanten Kennzahlen ist eine unternehmenseinheitliche Berechnung festzulegen. Immer noch werden in verschiedenen Abteilungen unterschiedliche Eingangsgrößen für die Berechnung von Kennzahlen verwendet, was ein Benchmarking sowohl innerhalb des Unternehmens als auch innerhalb einer Branche nahezu unmöglich macht.

Jede Kennzahl sollte dementsprechend detailliert und umfassend dokumentiert werden, um die Zielsetzung der Kennzahl, die Eingangsgrößen, die Berechnung der Kennzahl und die Verantwortlichkeiten eindeutig zu definieren. Für die Dokumentation von Kennzahlen schlägt Kütz (vgl. Kütz 2011, S. 44ff.) Kennzahlensteckbriefe vor, die

alle wichtigen Angaben zu den Kennzahlen in einem Formular enthalten. Neben dem Berechnungsweg enthalten solche Kennzahlensteckbriefe Angaben zu den Datenquellen, Zielwerten und Eskalationsregeln.

10.2 Daten und Business Intelligence

Im Zusammenhang mit der Gewinnung von Kennzahlen und Nutzung der im Unternehmen vorhandenen Daten fällt häufig der Begriff Business Intelligence (BI). Wie einige andere Errungenschaften im Bereich der Informationstechnologie findet die BI ihren Ursprung in einer militärischen Aufgabenstellung. So soll die Ressourcenknappheit der britischen Armee im zweiten Weltkrieg der Auslöser gewesen sein, ein Intelligence-Systeme zu etablieren, dass die Entscheider nicht nur mit den relevanten Informationen versorgte, sondern dies auch rechtzeitig erledigte (vgl. Schieder 2016, S. 15). Die Anwendung dieser Idee im Unternehmenskontext und die Schaffung des Begriffs Business Intelligence im Jahr 1958 wird dem IBM-Forscher Hans Peter Luhn zugeschrieben (vgl. Schieder 2016, S. 17). In den darauffolgenden Jahren verdrängten Decision Support Systeme und Management-Information-Systeme, die heute als Vorfürher-Systeme zur Business Intelligence betrachtet werden, denn Begriff BI weitgehend. Schieder (vgl. Schieder 2016, S. 18) spricht sogar davon, dass erst im Jahr 1986 der Begriff Business Intelligence wiederauftauchte.

Mertens hat 2002(vgl. Mertens 2002, S. 4) in einem Arbeitspapier die unterschiedlichen Begriffsbestimmungen zu BI festgehalten. So wird BI u. a. als Filter für die Informationsflut, als Frühwarnsystem, als Informations- und Wissensspeicherung oder auch als Prozess verstanden. Eine umfassende Definition bietet Gluchowski:

„.... Business Intelligence (lässt sich) als begriffliche Klammer verstehen, die unterschiedliche Technologien und Konzepte im Umfeld der entscheidungsunterstützenden Systeme zusammenführt und dabei eine entscheidungsorientierte Sammlung und Aufbereitung von Daten über das Unternehmen und dessen Umwelt sowie deren Darstellung in Form von geschäftsrelevanten Informationen für Analyse-, Planungs- und Steuerungszwecke zum Gegenstand hat.“(Gluchowski 2001, S. 390)

Nahezu unabhängig von der verwendeten Definition wird zur Darstellung und zum Aufbau eines BI-Systems eine Datawarehouse Architektur zugrunde gelegt. So verweist Hahne allein auf sieben traditionelle BI-Architekturen (vgl. Hahne 2016, S. 149–158) darunter befindet sich das Core Data Warehouse und die Hub-and-Spoke-Architektur. Im Core Data Warehouse liegen nach dem ETL-Prozess alle Daten zur Auswertung vor, was ersichtlich werden lässt, dass direkte Auswertungen, die auf eben diesen großen Datenmengen stattfinden, nur bedingt umsetzbar sind. Im Hinblick auf die möglichen Antwortzeiten – also die Dauer von der Anforderung des Anwenders bis zur Anzeige der Ergebnisse – besteht die Gefahr, dass diese unverhältnismäßig lang werden. Anwender gehen bei langen Antwortzeiten davon aus, dass die gestellte Anforderung

nicht ausgeführt wurde. In der Konsequenz wiederholen sie die Anforderung, was zu einer weiteren Belastung des Systems und einer erneuten Verzögerung der Bearbeitung der ursprünglichen Anforderung führt, oder sie versuchen ohne die angeforderten Informationen zu arbeiten, was das Ergebnis negativ beeinflussen kann. Weiterhin wird das Antwortzeitverhalten durch die Anzahl der Benutzer beeinflusst (vgl. Hahne 2016, S. 152), d. h. je mehr Anwender auf die Daten zugreifen, desto länger muss der einzelne Anwender auf die angeforderte Auswertung warten. Eine zentrale Datenhaltung ist nur dann für unterschiedlichste Auswertungszwecke sinnvoll nutzbar, wenn es sich um möglichst homogene Geschäftsfelder mit nahezu identischen Geschäftsprozessen handelt (vgl. Hahne 2016, S. 152). Um Informationen und Kennzahlen einer Vielzahl von Geschäftsfeldern mit sehr unterschiedlichen Geschäftsprozessen – wie sie in Konzernen oder spartenorientierten Unternehmen vorkommen – auszuwerten, schlägt Hahne als Alternative den Einsatz mehrerer autarker Core Data Warehouse vor, deren jeweiliger Fokus auf einer strategischen Einheit liegt (vgl. Hahne 2016, S. 152).

Die Verwendung der Hub-and-Spoke-Architektur (oder auch abhängige Data Marts) ist auf die jeweiligen betriebswirtschaftlichen Anwendungen bzw. Abteilungen konzentriert. Dabei erhalten die Anwender der jeweiligen Abteilungen Zugriff auf die korrespondierenden Data Marts, die unter Verwendung geeigneter Transformations- und Aggregationsprozesse aus dem Core Data Warehouse generiert werden. Somit bildet das Core Data Warehouse die Basis für alle Data Marts und verantwortet zentral die Transformation der Daten sowie die Qualitätssicherung (vgl. Hahne 2016, S. 154).

10.2.1 Datenmodellierung

Die Entscheidung für eine Datawarehouse Architektur muss sorgfältig erfolgen und wird durch eine dreistufige Modellierung unterstützt. In der ersten Stufe werden zunächst die fachlichen Anforderungen aufgenommen und über semantische Modelle abgebildet, sie bildet die Basis für das logische Modell der zweiten Stufen, was seinerseits wiederum als Grundlage für das physische Modell der dritten Stufe dient.

Die Erfassung und Modellierung der Fachanforderungen erfolgt unabhängig von der später verwendeten Technologie und wird häufig über das Entity-Relationship-Modell (ERM) dokumentiert. Die Darstellung der betriebswirtschaftlichen Zusammenhänge kann zusätzlich über ADAPT-Modellierung erfolgen, allerdings ist diese nicht der Lage das ERM vollumfänglich zu ersetzen (vgl. Kemper et al. 2010, S. 59). Mit der Modellierung der Fachanforderung wird die Basis für logische Modelle, die den Einsatz spezieller Datenhaltungssysteme unterstellen, geschaffen. Die bekannteste Modellierung erfolgt über relationale Datenmodelle (vgl. Kemper et al. 2010, S. 59). Dabei werden die Relationen als zweidimensionale Tabellen abgebildet. In einem relationalen Datenmodell werden aus den Entitätstypen sowie aus jedem komplexem Beziehungstype des ERM jeweils Relationen. Um redundanzfreie Strukturen zu schaffen, werden Relationenmodelle normalisiert (vgl. Kemper et al. 2010, S. 63).

Im Hinblick auf die Analyseanforderung wird zwischen Fakten, Dimensionen und Hierarchisierungen unterschieden, dabei sind Fakten eben jene numerischen Werte, die als betriebswirtschaftliche Kennzahlen zu verstehen sind. Sie geben Informationen in verdichteter Form wieder. Die Dimensionen ermöglichen eine spezielle Sicht (z. B. Filiale, Produkt etc.) auf die Fakten, eine Verdichtung kann in Form von Hierarchien (z. B. Filiale-Region-Land) erfolgen. (vgl. Kemper et al. 2010, S. 66). Die bekanntesten logischen Modellierungen stellen das Star-Schemata, das Fact-Constellation-Schema, das Galaxy-Schema, sowie das Snowflake-Schema dar. (vgl. Kemper et al. 2010, S. 67–70). Im einfachen Star-Schemata werden die Dimensionen um die in der stehenden Faktentabelle (sternförmig) angeordnet, der Übergang zum Snowflake-Schema ist fließend. Letzteres verzichtet zugunsten einer Performanceoptimierung auf eine vollständige Normalisierung (vgl. Kemper et al. 2010, S. 70).

Damit die Daten in der gewünschten, modellierten Weise vorliegen, ist in einem letzten Schritt die techniknahe und damit physische Modellierung vorzunehmen. Diese steht in unmittelbarem Zusammenhang mit der gewählten technischen Lösung.

10.2.2 Datensicherung

Um Daten nachvollziehbar und überprüfbar zu machen, aber auch um Vergleiche zu Vergangenheitsdaten herstellen zu können, ist eine Datensicherung erforderlich. Nach Kemper, Baars und Mehanna wird dabei zwischen Archivierung (zur Wiederherstellung bei fachlichem Bedarf), dem Backup (zur Wiederherstellung bei technischen Ausfällen) und der Historisierung (zur Auswertung unterschiedlicher fachlicher Zustände mittels der Dokumentation von Änderungen) unterschieden (vgl. Kemper et al. 2010, S. 71).

Sowohl Archivierung als auch Backup unterliegen rechtlichen aber auch operativen Anforderungen, die gewählten Historisierungsverfahren beeinflussen die Auswertbarkeit von Daten im Rahmen der BI und damit die für die Unternehmenssteuerung erforderlichen Informationen. Als Historisierungsverfahren steht das Update-Verfahren, das ein Überschreiben der Daten und somit einen Verzicht auf eine Historisierung bedeutet, die Snapshot-Historisierung, die mit einem enormen Datenwachstum der Dimensions-tabellen einhergeht, weil stets alle geänderten und unverändert geblieben Datensätze mit einem Zeitstempel versehen an die Tabelle angehängt werden, und die Delta-Historisierung mit einigen Varianten zur Verfügung (vgl. Kemper et al. 2010, S. 73 f.). Die gängigsten Varianten der Delta-Historisierung sind die Current-Flag-Variante, bei der der gültige Datensatz gekennzeichnet wird, die Historisierung mit Gültigkeitsfeldern, bei der den einzelnen Datensätzen das jeweilige Gültigkeitsintervall zugewiesen wird, sowie die Delta-Historisierung mit künstlicher Schlüsselerweiterung, bei der jeder Datensatz einen künstlichen Teilschlüssel erhält, der um einen Zählerschritt inkrementiert und in der Faktentabelle verwendet wird (vgl. Kemper et al. 2010, S. 75).

Branchenabhängig kann auch eine bitemporale (fachliche und technische Gültigkeit getrennt) oder tritemporale (fachliche, technische und operative Gültigkeit) Historisierung zum Einsatz kommen. (Hahne 2016, S. 161)

10.2.3 Harmonisierung

Aus den unterschiedlichen Quellsystemen sind die Daten so zu übernehmen, dass sie aussagefähige Informationen bilden und sich ergänzen bzw. gemeinsam auswertbar sind. Dazu sind die zugrunde liegenden Datenmodelle in Einklang zu bringen. Es sind mögliche Konflikte zu identifizieren und abschließend zu eliminieren (vgl. Kemper und Finger 2016, S. 137)

Konflikte sind z. B. strukturelle bzw. kodierungsbedingte Konflikte, die auftreten, wenn in einer Quelle ein Informationsobjekt als Relation vorliegt, in einer anderen eben dieses Informationsobjekt als Attribut und in einer dritten als Wert.

Die Beseitigung struktureller Konflikte erfordert nach dem Aufdecken eben dieser das Aufstellen von Transformationsregeln und deren Umsetzung.

Neben strukturellen Konflikten sind Beschreibungskonflikte möglich. So können unterschiedliche Bezeichnungen für identische Objekte (Synonyme) oder auch gleiche Bezeichnungen für unterschiedliche Objekte (Homonyme) vorliegen. Auftreten können z. B. Skalierungskonflikte, die sich in unterschiedlichen Maßeinheiten manifestieren, oder Genauigkeitskonflikte, die z. B. über die Anzahl der Nachkommastellen zu identifizieren sind. (vgl. Kemper und Finger 2016, S. 138).

Zur Behebung dieser Datenkonflikte können z. B. Ähnlichkeitsmaße bei Sprachvarianten eingeführt oder z. B. zur Behandlung von Homonymen und Synonymen Wörterbücher, Thesauri oder Ontologien eingesetzt werden.

Laut Kemper und Finger sind zudem Schlüsseldisharmonien zu lösen, die in Praxis auftreten, wenn unterschiedliche operative Systeme zusammengeführt bzw. Daten aus unterschiedlichen operativen Systemen für eine Auswertung genutzt werden sollen. Die Forderung nach einem grundsätzlichen Redesign der operativen Systeme ist in der Praxis jedoch nicht immer umsetzbar und wird oft durch auf die Auswertungsbedarfe zugeschnittene Workarounds ersetzt (vgl. Kemper und Finger 2016, S. 138 f.).

10.2.4 Daten-/Informationsqualität

Die gewonnenen bzw. generierten Informationen stehen in direktem Zusammenhang mit der Qualität der zur Verfügung stehenden Daten. Je besser die Datenqualität, desto verlässlicher sind die ermittelten Kennzahlen und Informationen. Wolf unterscheidet mit Prozessfehlern, Anwenderfehlern, Programmierfehlern und Kundenfehlern vier Arten von Datenqualitätsmängeln, die es zu verringern bzw. eliminieren gilt (vgl. Wolf 2018, S. 236). Er empfiehlt dies über insgesamt fünf verschiedene Strategien zu realisieren,

wobei die Umsetzung nur einer einzelnen Strategie aufgrund von vernetzen Abhängigkeiten nicht möglich ist (vgl. Wolf 2018, S. 239).

Einen anderen Ansatz verfolgen Rohweder et al. Für sie erfolgt die Bestimmung der Informationsqualität (IQ) über 15 IQ-Dimensionen, wobei die jeweiligen Dimensionen beliebig viele Ausprägungen annehmen können. Eine exakte Trennung der IQ-Dimensionen wird in der Praxis aufgrund des umgangssprachlichen Verständnissen der Bezeichnung erschwert (vgl. Rohweder et al. 2018, S. 25). Da alle 15 IQ-Dimensionen als gleichwertig anzusehen sind, wurden sie zum besseren Verständnis in vier IQ-Kategorien eingeteilt, die sich jeweils auf einen Untersuchungsgegenstand fokussieren.

So ist der zu betrachtende Untersuchungsgegenstand der systemunterstützenden IQ-Kategorie das System. Dieser Kategorie zugeordnet sind die IQ-Dimensionen Zugänglichkeit und Bearbeitbarkeit. Die Kategorie der inhärenten Merkmale fokussieren den Untersuchungsgegenstand Inhalt und enthält die IQ-Dimensionen hohes Ansehen, Fehlerfreiheit, Objektivität und Glaubwürdigkeit. In der darstellungsbezogenen Kategorie, die Darstellung betrachtet, finden sich die Verständlichkeit, die Übersichtlichkeit, die einheitliche Darstellung und die eindeutige Auslegbarkeit. Die letzte als zweckabhängig bezeichnete Kategorie fokussiert die Nutzung und umfasst die IQ-Dimensionen Aktualität, Wertschöpfung, Vollständigkeit, angemessener Umfang und Relevanz. (vgl. Rohweder et al. 2018, S. 28 f.)

Für eine bestmögliche Datenqualität ist es sinnvoll sowohl die Informationsqualität vor der Aufnahme von Daten zu analysieren als auch Datenqualitätsmängel zu verringern.

10.2.5 Datenbereitstellung

Die Datenbereitstellung kann in unterschiedlichen Datenbanken erfolgen. Dabei entscheidet die Anwendung, das Volumen und die Datenformate über die Datenbank. Die klassische relationale Datenbank ist für Auswertungen durchaus kritisch zu sehen (vgl. Bange 2016, S. 108), in-memory Datenbanken eignen sich dagegen besonders für abfrageintensive Anwendungen (vgl. Abts und Mülder 2017, S. 185).

Im Zuge zunehmender Digitalisierung gewinnen polystrukturierte Daten aus Social Media Kanälen oder Sensoren von Maschinen (Internet of Things) an Bedeutung. Eine Speicherung dieser Daten in relationalen Datenbanken ist i. d. R. unwirtschaftlich, sodass verschiedene Speichertechnologien mittel NoSQL-Datenbanken kombiniert werden (vgl. Bange 2016, S. 109). Dabei steht das „No“ für „not only“

10.3 Reporting/Berichtswesen

Um die verfügbaren Daten für eine Entscheidungsunterstützung nutzbar zu machen, reicht es nicht aus, diese lediglich über Business Intelligence Tools verfügbar zu machen. Vielmehr bedarf es einer Strategie, die die Integration von Planung und Reporting

beinhaltet. Eine isolierte Betrachtung birgt laut Schön die Gefahr von Überschneidungen und Lücken, die eine Steuerung des Unternehmens bzw. der Unternehmensbereiche erschweren oder unmöglich machen (vgl. Schön 2018, S. 2). Totok empfiehlt zunächst die Entwicklung einer BI-Strategie, sofern Unternehmen feststellen, dass z. B. die Datenqualität und oder die Antwortzeiten von den Anwendern als unzureichend bewertet werden. Auch das Fehlen einer gezielten Kommunikation sowie unterschiedliche Berichte aus unterschiedlichen Quellen können als Indiz für eine fehlende aber erforderliche BI-Strategie betrachtet werden (vgl. Totok 2016, S. 35). Die zu entwickelnde BI-Strategie muss sich an den Unternehmens- und IT-Zielen ausrichten (vgl. Totok 2016, S. 36). Dabei soll die BI-Strategie als kontinuierlicher Prozess verstanden werden, dessen Bestandteile fortlaufend weiterentwickelt und die in regelmäßigen mehrjährigen Abständen mit der Unternehmensausrichtung abgeglichen wird (vgl. Totok 2016, S. 37).

Die für die Aufbereitung der Informationen zu nutzenden Anwendungen sind durch ihre Komplexität und die Freiheitsgrade für die Anwender bestimbar. So bieten einfache Dashboards, Scorecardings oder Cockpits überwiegend eine visuelle Aufbereitung aggregierter Informationen, der Anwender hat nur begrenzte Möglichkeiten die Darstellung zu variieren, eine Änderung der dargestellten Kennzahlen ist in der Regel nicht vorgesehen. Dagegen bieten Data Mining Tools den Anwendern die Möglichkeiten zur Ermittlung neuer Erkenntnisse, da die benötigten Daten bedarfsorientiert zusammengestellt und visualisiert werden können.

Im Wesentlichen wird die Aufbereitung von Informationen in die Berichtsarten Standardreporting, Exception Reporting, Analysereporting und Ad-hoc-Reporting unterschieden (vgl. Schön 2018, S. 49). Das Standardreporting beinhaltet festgelegte Informationen in definierter Form und wird in vorgegebenen Zyklen an genau bestimmte Empfänger verteilt. Schön empfiehlt hier eine regelmäßige Überprüfung und Anpassung der Berichtsstruktur und Inhalte (vgl. Schön 2018, S. 49). Im Rahmen des Exception Reportings werden Toleranzgrößen, Grenz- und Schwellenwerte festgelegt, bei deren Erreichung automatisch Berichte generiert und den zuvor definierten Empfängern vorgelegt werden. Dabei können die Schwellenwerten sowohl manuell als auch mittels statistischer Funktionen festgelegt werden (vgl. Schön 2018, S. 49 f.). Das Analyse-reporting bietet dem Anwender über einen vordefinierten Einstieg die Möglichkeit zur Durchführung strukturierter Recherchen. Das Ziel ist es, neue Erkenntnisse zu gewinnen. Da die Recherchen die bereitgestellten Daten nutzen, liegt hier die Limitierung in eben dieser Datenbasis (vgl. Schön 2018, S. 50). Ad-hoc-Reporting wird auf Basis individueller Anforderungen durchgeführt. In der Praxis können diese Anforderungen z. T. über das Analysereporting erfüllt werden. Das ist immer dann der Fall, wenn aus der vorliegenden Datenbasis die angeforderten Informationen generiert werden können. In einigen Fällen muss jedoch die Datenbasis um weitere Informationen ergänzt werden, bevor die Abfrage durchgeführt werden kann. (vgl. Schön 2018, S. 50).

10.3.1 Berichtsgrundformen

Innerhalb der Berichtsarten finden sich abhängig von der zu erzielenden Aussage vielfältig gestaltete Berichte. Trotz unternehmensindividueller Ausgestaltung dieser Berichte lassen sich laut Schön (vgl. Schön 2018, S. 50) gewisse Grundstrukturen erkennen. Um beispielsweise Veränderungen gegenüber Vergleichsperioden auszuweisen, eignen sich Ist-Ist-Vergleiche. Hierbei werden Kennzahlen mit ihren Ausprägungen der Vorperioden – seien es Monate, Quartale oder Jahre – verglichen, üblicherweise wird zusätzlich das Delta und die prozentuale Veränderung ausgewiesen. Häufig werden diese Vergleiche mit grafischen bzw. farbigen Elementen angereichert, so dass signifikante Veränderungen auf einen Blick erkennbar sind. Die Einsatzmöglichkeiten dieser Ist-Ist-Vergleiche sind begrenzt. Sie liefern keine aussagekräftigen Informationen, wenn beispielsweise die Daten der Vorperioden fehlen oder das Unternehmen größere organisatorische Umstrukturierungen durch Zu- oder Verkauf von Unternehmensteilen erfahren hat. Ebenfalls sind keine Aussagen hinsichtlich einer Zielerreichung möglich, da keine Plandaten einfließen. (vgl. Schön 2018, S. 52).

Um eine solche Aussage treffen zu können, sind die Soll-Ist- bzw. Plan-Ist-Vergleiche nützlich. Analog zu den Ist-Ist-Vergleichen werden die tatsächlichen Werte einer Periode den erwarteten, geplanten Werten gegenübergestellt. Auch hier werden i. d. R. die absoluten sowie die prozentualen Abweichungen ausgewiesen und durch grafische und/oder farbige Elemente visuell hervorgehoben. Mit dem Einsatz von Soll-Ist-Vergleichen bietet sich Unternehmen die Möglichkeit zeitnah zu erkennen, ob die Zielerreichung möglich ist und eventuell steuernd einzugreifen. Allerdings ist eine sorgfältige Planung erforderlich, die unter Umständen durch aktuelle Ereignisse obsolet wird. (vgl. Schön 2018, S. 53).

Um Prognosen zu berücksichtigen empfiehlt Schön Soll-Wird-Vergleiche, dabei kann der eine Prognose darstellende Wird-Wert auf unterschiedliche Weise ermittelt werden. Auch mit diesen Vergleichen ist – bei entsprechender Aktualisierung der Planung – ein Steuerungsbedarf leicht erkennbar, zudem werden zukünftige Entwicklungen berücksichtigt (vgl. Schön 2018, S. 54).

Zielerreichungsberichte stellen die Ist-Werte den Zielvorgaben für den gesamten zu betrachtenden Zeitraum gegenüber. Unternehmen können so z. B. erkennen, dass sie in den ersten beiden Quartalen lediglich 40 % des geplanten Jahresumsatzes erwirtschaftet haben. Diese Berichte erfordern – wie auch die Plan-Soll- und Plan-Wird-Vergleiche – eine intensive Planung und können durch aktuelle Geschehnisse hinfällig werden. Zudem sind Trends nur in Verbindung mit Zeitreihenanalysen erkennbar. (vgl. Schön 2018, S. 56)

Die Darstellung von Kennzahlen über mehrere Perioden hinweg – die Zeitreihenanalyse – erfolgt häufig grafisch und ermöglicht so eine schnelle Erkennung von Ausreißern und Trends. Die auf das Unternehmen angepasste Wahl der Anzahl der Perioden ist entscheidend für die Aussagefähigkeit einer Zeitreihenanalyse. Ebenso sind zyklische Schwankungen z. B. durch saisonale Einflüsse zu dokumentieren. (vgl. Schön 2018, S. 57)

Schön verweist auf weitere Berichtsgrundformen, wie die ABC-, Top-/Flop- und Klassen-Analyse, die Portfolio-Analyse, die Objekt-und Benchmark-Vergleiche, die Break-Even-Point-Analyse sowie die Scoring- bzw. Nutzwertanalyse (vgl. Schön 2018, S. 58–63).

10.3.2 Anforderungen an Berichte

Zur Steuerung eines Unternehmens bedarf es nicht nur der Auswahl geeigneter Kennzahlen, sondern auch einer angemessenen Darstellung. Zur leichteren Auffindbarkeit von Informationen empfiehlt sich eine einheitliche Struktur, die eine gleiche inhaltliche Reihenfolge aufweist. So sind die relevanten Informationen für die Empfänger der Berichte ohne erhöhten Aufwand auffindbar.

Hinsichtlich der Gestaltung finden sich unterschiedlich detaillierte Empfehlungen. So erscheinen die SUCCESS-Regeln von Hichert im Vergleich zu Schöns Empfehlungen (vgl. Schön 2018, S. 66–112) überschaubar. Beiden gemein ist das Plädoyer für unternehmenseinheitliche, strukturierte und klare Berichte. Damit dies gelingt, gilt es den Erstellern der Berichte klar zu machen, dass nicht der Bericht, sondern die zu entnehmende Information im Fokus stehen muss. Schmückendes Beiwerk in Form von besonderen Effekten, möglichst vielen Farben und Schriftarten und –größen muss als hinderlich für die Informationsgewinnung erkannt werden. Zur leichteren Umsetzung empfehlen sich Vorgaben – wie sie auch für das Corporate Design eines Unternehmens zu nutzen sind – bereitzustellen.

Beide Autoren weisen darauf hin, dass das Berichtswesen im Unternehmen zu etablieren ist.

Literatur

- Abée, S., Andrae, S., Schlemminger, R.B.: Strategisches Controlling 4.0. Wie der digitale Wandel gelingt. Springer, Wiesbaden (2020)
- Abts, C., Mülder, W.: Grundkurs Wirtschaftsinformatik. Eine kompakte und praxisorientierte Einführung, 9. Aufl. Springer, Wiesbaden (2017)
- Bange, C.: Werkzeuge für analytische Informationssysteme. In: Chamoni, P., Gluchowski, P. (Hrsg.) Analytische Informationssysteme. Business Intelligence-Technologien und Anwendungen, 5. Aufl., S. 97–128. Springer Gabler, Berlin (2016)
- Freidank, C.-C.: Herausforderungen im Reporting. In: Panitz, K., Waschkowitz, C. (Hrsg.) Reportingprozesse optimieren Praxislösungen für ein effizientes Rechnungswesen, S. 11–28. Schäffer-Poeschel, Stuttgart (2010)
- Gluchowski, P.: Ansatzpunkte zur Gestaltung einer Business Intelligence-Strategie. In: Götze, U., Lang, R. (Hrsg.) Strategisches Management zwischen Globalisierung und Regionalisierung, S. 381–401. Springer Gabler, Wiesbaden (2001)

- Hahne, M.: Architekturkonzepte und Modellierungsverfahren für BI-Systeme. In: Chamoni, P., Gluchowski, P. (Hrsg.) Analytische Informationssysteme. Business Intelligence-Technologien und-Anwendungen, 5. Aufl., S. 147–186. Springer Gabler, Berlin (2016)
- Horváth, P., Gleich, R., Seiter, M.: Controlling, 14. Aufl. Vahlen, München (2020)
- Kemper, H.-G., Baars, H., Mehanna, W.: Business Intelligence – Grundlagen und praktische Anwendungen. Eine Einführung in die IT-basierte Managementunterstützung, 3. Aufl. Vieweg+ Teubner, Wiesbaden (2010)
- Kemper, H.-G., Finger, R.: In: Chamoni, P., Gluchowski, P. (Hrsg.): Analytische Informationssysteme. Business Intelligence-Technologien und-Anwendungen, 5. Aufl., S. 129–146. Berlin (2016)
- Kleindienst, B.: Performance Measurement und Management. Advance online publication. (2017). <https://doi.org/10.1007/978-3-658-19449-9>
- Knabke, T., Olbrich, S.: Grundlagen und Einsatzpotentiale von In-Memory-Datenbanken. In: Chamoni, P., Gluchowski, P. (Hrsg.) Analytische Informationssysteme. Business Intelligence-Technologien und-Anwendungen, 5. Aufl., S. 187–204. Springer Gabler, Berlin (2016)
- Kütz, M.: Kennzahlen in der IT Werkzeuge für Controlling und Management, 4. Aufl. dpunkt, Heidelberg (2011)
- Losbichler, H.: Controlling 4.0: Muster des Wandels. In: Gleich, R., Losbichler, H., Zierhofer, R. (Hrsg.) Unternehmenssteuerung im Zeitalter von Industrie 4.0, S. 43–61. Freiburg, Haufe-Lexware (2016)
- Mertens, P.: Business Intelligence – ein Überblick, Arbeitspapier an der Universität Erlangen-Nürnberg 2/2002, Nürnberg (2002)
- Ossola-Haring, C., Schlageter, A., Schöning, S.: 11 Irrtümer über Kennzahlen. Mit den richtigen Erkenntnissen führen, 2. Aufl. Springer, Wiesbaden (2019)
- Reichmann, T., Kißler, M., Baumöl, U.: Controlling mit Kennzahlen. Die systemgestützte Controlling-Konzeption, 9. Aufl. Vahlen, München (2017)
- Rohweder, J.P., Kasten, G., Malzahn, D., Piro, A., Schmid, J.: Informationsqualität – Definitionen, Dimensionen und Begriffe. In: Hildebrand, K., Gebauer, M., Hinrichs, H., Mielke, M. (Hrsg.) Daten- und Informationsqualität. Auf dem Weg zur Information Excellence, 4. Aufl., S. 23–46. Vieweg+ Teubner, Wiesbaden (2018)
- Schieder, C.: Historische Fragmente einer Integrationsdisziplin – Beitrag zur Konstruktgeschichte der Business Intelligence. In: Chamoni, P., Gluchowski, P. (Hrsg.) Analytische Informationssysteme. Business Intelligence-Technologien und-Anwendungen, 5. Aufl., S. 13–30. Springer Gabler, Berlin (2016)
- Schön, D.: Planung und Reporting im BI-gestützten Controlling. Grundlagen, Business Intelligence, Mobile BI und Big-Data-Analytics, 3. Aufl. Springer Fachmedien, Wiesbaden (2018)
- Totok, A.: Von der Business-Intelligence-Strategie zum Business Intelligence Competency Center. In: Chamoni, P., Gluchowski, P. (Hrsg.) Analytische Informationssysteme. Business Intelligence-Technologien und-Anwendungen, 5. Aufl., S. 33–54. Springer Gabler, Berlin (2016)
- Wolf, J.: Organisatorische Maßnahmen für gute Datenqualität. In: Hildebrand, K., Gebauer, M., Hinrichs, H., Mielke, M. (Hrsg.) Daten- und Informationsqualität. Auf dem Weg zur Information Excellence, 4. Aufl., S. 235–251. Springer Gabler, Wiesbaden (2018)



Fundamentale Analyse- und Visualisierungstechniken

11

Jens Kaufmann

Zusammenfassung

Professionelle Datenanalyse basiert auf statistischen Methoden und der geschickten, (semi-)automatisierten Anwendung geeigneter Algorithmen auf zuvor aufbereiteten Datenbeständen. Während durch die fortschreitende Digitalisierung immer umfangreichere Datenmengen bereitstehen und hierfür komplexe und hochspezialisierte Werkzeuge zum Einsatz kommen, lassen sich die üblichen Fragen in Unternehmen in der Regel durch den Einsatz fundamentaler Analysetechniken lösen und die Ergebnisse nachvollziehbar grafisch aufbereiten. Der Beitrag stellt die relevanten Methoden mit ihren zugrunde liegenden Konzepten vor, gibt erste Einblicke in die Feinheiten der Anwendung und bietet Leserinnen und Lesern eine Übersicht des Feldes der Datenanalyse. Er vermittelt so die Möglichkeit, die richtigen Methoden für die ersten eigenen Analysen auszuwählen und das passende Vertiefungsgebiet zu finden.

11.1 Einleitung und Begriffswelt

Datenanalyse und Datenvisualisierung sind Kernbestandteile der Tätigkeit moderner Unternehmen. Ein Mehrwert ergibt sich aber erst, wenn die relevanten Daten sinnvoll analysiert werden, oder, um es aus der geschäftlichen Perspektive zu benennen, wenn die richtigen Fragen gestellt und von den richtigen Personen mit dem richtigen Werkzeug bearbeitet werden. Dieser Beitrag beschäftigt sich vor allem mit den Methoden als einem essentiellen Teil des Werkzeugkastens und verfolgt den Ansatz, die Vielfalt möglicher

J. Kaufmann (✉)

FB 08, Hochschule Niederrhein, Mönchengladbach, Deutschland

E-Mail: jens.kaufmann@hs-niederrhein.de

Analysewege aufzuzeigen, um die für das jeweilige Problem passenden auswählen und dort im Anschluss tieferes Fachwissen aufbauen zu können. Zu Beginn wird dafür die Begriffswelt der Datenanalyse kurz vorgestellt.

Einer der umfassendsten Begriffe, der die Datenanalyse mit behandelt, findet sich im Titel dieses Sammelwerks und beschreibt als *Data Science* einen Ansatz, Daten systematisch im gesamten Wertschöpfungsprozess zu betrachten. Dieser Beitrag folgt der Definition von Schulz et al. (2020):

Data Science ist ein interdisziplinäres Fachgebiet, in welchem mithilfe eines wissenschaftlichen Vorgehens, semiautomatisch und unter Anwendung bestehender oder zu entwickelnder Analyseverfahren Erkenntnisse aus teils komplexen Daten extrahiert und unter Berücksichtigung gesellschaftlicher Auswirkungen nutzbar gemacht werden.

Innerhalb der Data Science ist die Datenanalyse nach dieser Definition der Teil des gesamten Arbeitskreislaufs, der voraussetzt, dass bereits Daten erhoben und bereitgestellt, teilweise sogar schon aufbereitet wurden, und auf diese Daten neue und bekannte Methoden und Algorithmen anwendet, die Erkenntnisse bringen. Die betriebliche Verwendung dieser analytischen Erkenntnisse ist dabei nicht mehr originärer Bestandteil der Datenanalyse selbst.

Neben Data Science hat sich in den letzten Jahren *Big Data* als Begriff etabliert. Wenngleich auch hier unterschiedliche Definitionen existieren, besteht in der Fachwelt doch relative Einigkeit, dass sich Big Data durch das Vorhandensein von mindestens drei Kennzeichen definiert: Eine große Menge an Daten, eine hohe Entstehungs- und Verarbeitungsgeschwindigkeit sowie eine hohe Variabilität der Daten. Diese drei Eigenschaften werden üblicherweise als 3V-Definition zusammengefasst (vom Englischen volume, velocity, variety) (Meier 2018).

Im Fokus von *Business Intelligence* (BI) stehen je nach verwendeter Definition ein umfangreiches Berichtswesen und die nutzergetriebenen, eher explorativen Datenbetrachtungen des Online Analytical Processing (OLAP) oder darüberhinausgehende Techniken zur Datenverarbeitung und automatisierten Datenanalyse (Kemper et al. 2010). In diesem Beitrag werden Vorgehensweisen im Bereich BI zwar ebenfalls als analytisch getrieben verstanden, eine automatisierte Datenanalyse in all ihren Facetten jedoch eher als ergänzendes bzw. selbstständiges Themenfeld betrachtet.

Forschungsgegenstand im Bereich *Künstliche Intelligenz* (KI) ist die Suche nach der Möglichkeit, Computer Dinge tun zu lassen, in denen sie derzeit den Menschen noch unterlegen sind (Weber 2020). Als ein Kernbereich der KI gilt das *Maschinelle Lernen* (machine learning), dessen Algorithmen sich in drei grundlegende Klassen einteilen lassen. Beim *überwachten Lernen* liegen sowohl Eingabedaten als auch (korrekte) Ausgabedaten vor. Jedes Modell wird so gestaltet, dass es mit der Zeit durch Veränderung „lernen“ kann, welche Ausgabe bei welcher Eingabe erwartet wird. Im Gegensatz dazu liegen beim *unüberwachten Lernen* nur Eingabedaten vor, ein Ergebnis ist im Vorhinein nicht bekannt. Das *bestärkende Lernen* kann als eine Art Zwischenstufe aufgefasst

werden. Der Algorithmus erhält hier ein verstärkendes oder abschwächendes Signal, wenn er gewünschte oder unerwünschte Zustände erreicht.

Die Methoden und Algorithmen des maschinellen Lernens überlappen sich stark mit denen des *Data Mining*, einer Disziplin, deren Trennung zum maschinellen Lernen nur schwer vorzunehmen ist. Im Sinne dieses Beitrags fokussiert Data Mining die Anwendung der Algorithmen, um Muster in Daten zu entdecken. Dabei lassen sich vier Hauptbereiche unterscheiden, die in den Abschnitten dieses Beitrags mit ihren grundlegenden Ideen, Fragestellungen und Methoden vorgestellt werden: *Klassifikation* (die Zuordnung von Objekten zu festgelegten Klassen), *Clustering* oder Segmentierung (die Einteilung von Objekten in Gruppen), *Prognose* (von numerischen Werten auf Basis schon bekannter numerischer Werte) und *Assoziationsanalyse* (das Erkennen von Zusammengehörigkeit einzelner Elemente und daraus ableitbare Regeln) (Witten et al. 2017).

Datenanalyse ist zu einem großen Teil Anwendung von Statistik. Computer und Algorithmen haben die Analysen komplexer und leistungsfähiger gemacht, im Kern werden aber statistische Verfahren angewendet und jede Datenanalyse sollte damit beginnen, die vorliegenden Daten mithilfe ihrer Basiseigenschaften zu verstehen. Im Folgenden werden einige essenzielle Begriffe kurz aufgegriffen, um ein einheitliches Verständnis zu gewährleisten. Gegeben sei eine Menge von Daten, beispielsweise die Körpergröße aller Personen in einem Raum, gemessen in Zentimetern (cm). Daten dieser Art sind, wie die meisten Zahlen-Daten, *kardinalskaliert*. Die Abstände einzelner Datenpunkte sind bestimmbar (in der Einheit cm), es gibt einen Nullpunkt und es lassen sich zudem Verhältnisse bestimmen. Sind die Datenpunkte, sogenannte *Beobachtungen*, gegeben und wurde ausschließlich die Körpergröße notiert, so liegen die Daten eindimensional vor, da nur eine *Variable* bestimmt wurde. Sollen mit wenigen Kennzahlen Aussagen über eine solche Datenmenge getroffen werden, so können unter anderem das *arithmetische Mittel* (die Summe aller Zahlen geteilt durch deren Anzahl), der *Median* (die Beobachtung, die größer ist als 50 % der Werte und kleiner als die anderen 50 % der Werte), *Quartile* (wie der Median, aber mit einer 25/75- bzw. 75/25-Verteilung) oder darüberhinausgehende Werte wie die Standardabweichung verwendet werden. Um diese Basis-Eigenschaften zu visualisieren, werden *Boxplots* verwendet. Abb. 11.1 zeigt einen

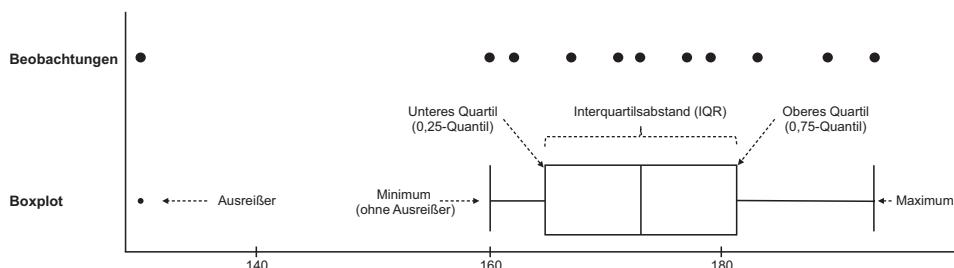


Abb. 11.1 Boxplot von 11 Beobachtungen (Körpergröße von 11 Personen in cm)

solchen Boxplot, bei dem zusätzlich in Form von Punkten über dem Boxplot die 11 Beobachtungen dargestellt sind.

Die eigentliche Box beinhaltet die mittleren 50 % der Datensätze, der Median ist durch den senkrechten Strich dargestellt. Entlang der waagerechten Linien rechts und links der Box liegen die unteren bzw. oberen 25 % der Daten, die senkrechten Endmarkierungen (im Englischen „whiskers“ genannt) stellen Minimum und Maximum der Datensätze dar. Besonders ist hier, dass bereits ein *Ausreißer* markiert wurde. Ausreißer stellen ungewöhnlich große oder kleine Beobachtungen dar und es existiert keine allgemeingültige Definition für diese. Üblich ist, sie z. B. dort zu verorten, wo die Werte eine 1,5-fache Überschreitung des Interquartilsabstandes vom Rand der Box aufweisen (Mittag 2017). Es sei darauf hingewiesen, dass Ausreißer zunächst nur beobachtet werden, eine Interpretation ist immer notwendig. So lässt sich erkennen, ob der Datensatz fehlerhafte Daten enthält (hier im Beispiel könnte ein Kind erfasst sein, obwohl nur Erwachsene erfasst sein sollten) oder ob der Ausreißer auffällig, aber korrekt ist (z. B., weil ein Erwachsener geringer Körpergröße erfasst wurde). Boxplots bieten, insbesondere, wenn sie für den Vergleich mehrerer Datensätze verwendet werden, einen grundlegenden Einstieg in die Visualisierung und Analyse der Daten.

11.2 Lineare Regression

Die lineare Regression ist eines der am häufigsten eingesetzten Werkzeuge in der Datenanalyse, was daran liegen mag, dass sie Bestandteil zahlreicher Grundlagen-Veranstaltungen der Statistik ist, dass sie in Standard-Tabellenkalkulationssoftware integriert ist und dass sie zur Vorhersage von Werten eingesetzt werden kann. Im Rahmen der fundamentalen Analysemethoden deckt sie daher den Bereich Prognose aus dem Data Mining ab.

11.2.1 Basisidee und Begrifflichkeiten

Regressionsanalysen modellieren Zusammenhänge zwischen *unabhängigen* und *abhängigen* Variablen. Unabhängige Variablen gehen als Eingabe in die Analyse, die abhängige Variable soll prognostiziert werden. Die lineare Regression unterstellt dabei einen linearen Zusammenhang zwischen den Variablen. Die Modelle der linearen Regressionsanalyse können auf unterschiedliche Weisen formuliert werden. Dieser Beitrag folgt im Wesentlichen der Notation und den Beschreibungen von James et al. (2013). Sei X eine unabhängige Variable (z. B. das Alter eines Menschen in Jahren) und Y die abhängige Variable (z. B. die Körpergröße in cm). Es lässt sich nun ein vermuteter linearer Zusammenhang der Variablen formulieren:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

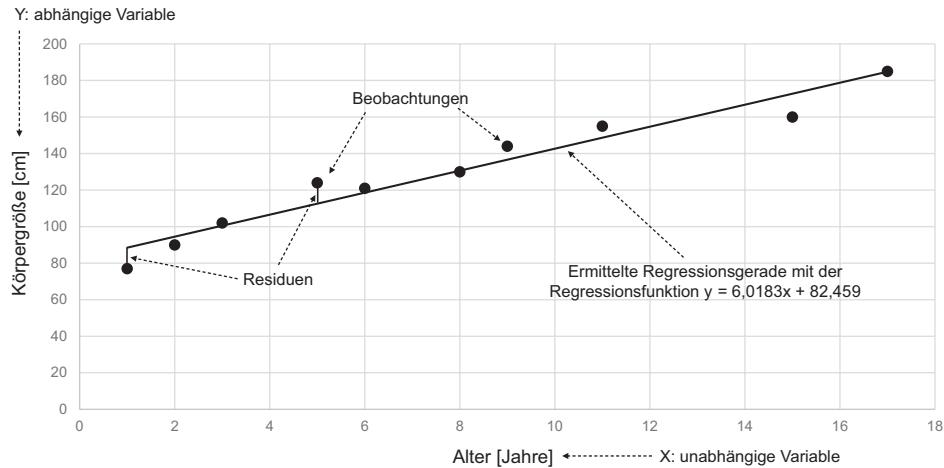


Abb. 11.2 Darstellung einer linearen Regressionsanalyse (Körpergröße in Abhängigkeit vom Alter)

Ziel der Regressionsanalyse ist es, den *Steigungsparameter* β_1 und den *Achsenabschnitt* β_0 so zu bestimmen, dass die prognostizierten Werte möglichst gut die gemessenen Werte abbilden. Das ermittelte Modell kann genutzt werden, um für jeden Wert x_i den zugehörigen prognostizierten Wert y'_i zu ermitteln und mit dem tatsächlichen Wert y_i zu vergleichen. Die Differenzen der Werte sind die *Residuen*, sie stellen den Fehler im Modell dar, das die Wirklichkeit selten vollständig erfasst, weil in der Regel kein vollständig linearer Zusammenhang zwischen zwei Variablen vorliegt, sondern ein Fehlerterm („Epsilon“) in der zugrunde liegenden Beziehung auftaucht. Abb. 11.2 stellt die Begrifflichkeiten und Zusammenhänge grafisch zusammenfassend dar.

Erkennbar ist, dass hier nur eine unabhängige Variable existiert. Sollen mehrere Variablen genutzt werden, um eine abhängige Variable zu prognostizieren (z. B., weil neben dem Alter auch das Geschlecht in die Prognose einfließen soll), muss das Modell entsprechend erweitert werden. Es liegt dann eine *Multiple lineare Regression* (MLR) vor. Das entsprechende Modell ähnelt dem der einfachen linearen Regression und nimmt weiterhin einen linearen Zusammenhang zwischen den unabhängigen und der abhängigen Variable an. Die Anzahl der unabhängigen Variablen wird üblicherweise mit p bezeichnet:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

11.2.2 Beispiel und Ergebnisinterpretation

Für die Ergebnis-Interpretation einer multiplen linearen Regression, wird im Folgenden exemplarisch ein Datenset verwendet, das aus 20 Beobachtungen und drei Variablen

```

Call: lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.503 -6.604 -2.184  8.956 17.109 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9.1591    6.9566   1.317   0.205    
X1          3.4470    0.4086   8.435 1.76e-07 ***  
X2         -0.1034    0.1513  -0.684   0.503    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.25 on 17 degrees of freedom
Multiple R-squared:  0.8221, Adjusted R-squared:  0.8012 
F-statistic: 39.29 on 2 and 17 DF,  p-value: 4.225e-07

```

Abb. 11.3 Ausgabe eines multiplen linearen Regressionsmodells

besteht. Aufgabe der MLR ist es, ein Modell zu entwickeln, das beschreibt, wie sich Y in Abhängigkeit von X1 und X2 verändert. Abb. 11.3 zeigt die Ausgabe eines solchen Modells, das mithilfe der frei lizenzierten Statistiksoftware R erstellt wurde. Weitere Hinweise zu geeigneter Software für Datenanalyse finden sich in Abschn. 11.6.

Die Ausgabe zeigt, welche Formel grundsätzlich für das Modell verwendet wurde, gefolgt von einer Darstellung der Kennwerte zur Verteilung der Residuen (residuals). Im folgenden Teil der Ausgabe stehen zeilenweise die *Koeffizienten* (coefficients) des Modells, im vorliegenden Fall also β_0 , β_1 und β_2 . Sie geben an, wie die Regressionsgerade bzw. im mehrdimensionalen Raum die (Hyper-)Ebene verläuft, auf der die prognostizierten Werte liegen. Der positive Koeffizient von X1 zeigt an, dass je Steigerung von X1 um eine Einheit der Wert von Y um 3,4470 Einheiten ansteigt. Der in der nächsten Spalte aufgeführte *Standardfehler* der Koeffizienten ermöglicht eine Aussage über die Präzision der Vorhersage. Der *p-Wert* (letzte Spalte) ist Teil des Ergebnisses eines Hypothesentests des jeweiligen Koeffizienten. Er gibt an, bei welchem angenommenen Signifikanzniveau der Wert akzeptiert werden würde. Die korrekte Interpretation und Formulierung von „Signifikanz“ sowie des p-Wertes ist für eine gute Datenanalyse entscheidend, entsprechende Ausführungen finden sich z. B. bei Mittag (2017), typische Fehler bei Goodman (2008). Vereinfacht und anwendungsorientiert werden in der Praxis zumeist all diejenigen Koeffizienten als „korrekt“ oder „verwendbar“ betrachtet, die ein Signifikanzniveau von höchstens 5 % (also 0,05) aufweisen. Ihr p-Wert liegt entsprechend zwischen 0,00 und 0,05. Im vorliegenden Beispiel zeigt sich, dass der Koeffizient β_1 einen sehr geringen p-Wert hat ($1,76 \times 10^{-7}$, als niedrig gekennzeichnet durch „***“, siehe Legende unterhalb der Koeffizienten), β_2 hingegen einen sehr hohen p-Wert hat (0,503), was auf das Fehlen der Eignung der Variable X2 für eine Prognose von Y hinweist.

Es lassen sich für MLR-Modelle diverse weitere Kennzahlen ermitteln, unter denen das *Bestimmtheitsmaß*, auch R^2 oder R-squared genannt, eine Sonderstellung einnimmt, da es in der Regel als erster Maßstab für die Güte des Modells genommen wird. R^2 liegt zwischen 0 und 1 und gibt an, wie gut das Modell die Varianz in den Daten erklärt. Bei einem Bestimmtheitsmaß von 1 liegen alle Datenpunkte auf der Regressionsgeraden. Es gibt dabei keine Mindestwerte, die R^2 erreichen muss, was auch daran liegt, dass R^2 steigt, je mehr unabhängige Variablen in ein bestehendes Modell aufgenommen werden. Für umfangreiche Betrachtungen der Güte und für die Auswahl der zu verwendenden Variablen bieten sich daher ein adjustiertes Bestimmtheitsmaß oder weitere Kennzahlen wie z. B. das Akaike-Informations-Kriterium an (James et al. 2013).

11.2.3 Prüfen der Voraussetzungen und Variablentransformation

Um eine (multiple) lineare Regression überhaupt anwenden zu können, müssen die zugrunde liegenden Daten bestimmte Voraussetzungen erfüllen. Dazu gehören unter anderem (James et al. 2013):

- 1 Ein prinzipiell linearer Zusammenhang der Daten
- 2 Die Absenz von Extremwerten
- 3 Nicht miteinander korrelierende (in Beziehung stehende) unabhängige Variablen (d. h. fehlende *Multikollinearität*)
- 4 Gleichverteilte Varianzen der Residuen (*Homoskedastizität*)
- 5 Unabhängigkeit der Residuen (wie sie bei Zeitreihen z. B. häufig nicht gegeben ist)

Diese Voraussetzungen können mithilfe diverser grafischer Abbildungen (plots) nach der Modellerstellung geprüft werden. Sofern sich hier große Auffälligkeiten zeigen, können Daten oder Modell ggf. nicht in der angedachten Form verwendet werden. Eine Darstellung dieser Plots und ihrer Interpretation findet sich bei Kim (2015).

Am Beispiel des linearen Zusammenhangs zeigt sich, wie durch Datentransformation die Datenbasis so verändert werden kann, dass eigentlich die Annahmen verletzende Daten doch verwendet werden können. Abb. 11.4 zeigt die Wertentwicklung eines Sparkontos über 50 Jahre (links). Hier besteht durch Zinseszinseffekte ein exponentieller Zusammenhang zwischen der Laufzeit und dem Wert. Ein lineares Regressionsmodell lässt sich nicht anwenden. Wird aber die abhängige Variable (Kontowert) zuvor logarithmiert, so stellt sich ein fast perfekter linearer Zusammenhang dar (rechts).

Diese *Variablentransformation* ist möglich, weil das Modell einen „technisch“ linearen Zusammenhang zwischen den Variablen voraussetzt, keinen „inhaltlich nachvollziehbaren“. Das Ergebnis ist allerdings komplizierter zu interpretieren, da der Koeffizient der unabhängigen Variable nun nicht mehr die Steigerung in der Einheit des Kontowertes (EUR) angibt, sondern vielmehr eine prozentuale Steigerung. Je nach Transformationsart muss bei der Interpretation daher mit großer Sorgfalt vorgegangen

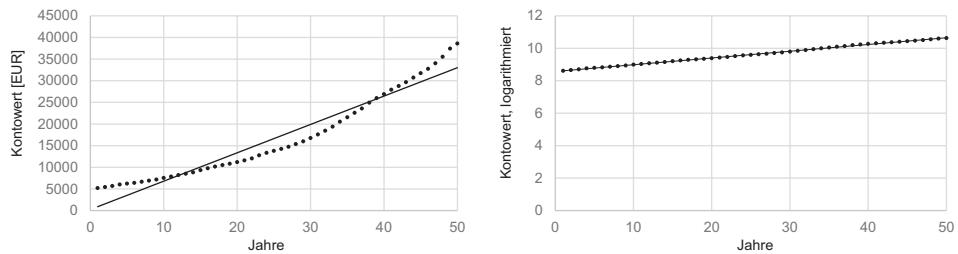


Abb. 11.4 Wertentwicklung eines Kontos (links) und logarithmierte Darstellung (rechts)

werden. Dies gilt umso mehr, wenn neben einzelnen unabhängigen Variablen auch Kombinationen betrachtet werden sollen, sogenannte *Interaktionstermine* (Wooldridge 2015). Dazu kommt, dass im vorliegenden Beispiel eine Zeitreihe analysiert wurde, die wie oben beschrieben die Unabhängigkeit der Residuen in der Regel nicht erfüllt, weshalb korrekte Zeitreihenanalysen im Rahmen der fortgeschrittenen Analysemethoden betrachtet werden. Zusammenfassend ist die MLR ein grundlegendes und mächtiges Werkzeug der Datenanalyse. Ihre Handhabung erfordert aber ein hohes Maß an Aufmerksamkeit und das Wissen, dass eine Prognose insbesondere außerhalb des bereits „bekannten“ Datenraumes mit großer Unsicherheit verbunden sein kann.

11.3 Einfache Klassifikationsverfahren

Klassifikationsverfahren nehmen einen großen Raum innerhalb der Algorithmenvielfalt des Data Mining und maschinellen Lernens ein, da sie es erlauben, automatisiert Entscheidungen vorzubereiten. Standardbeispiele aus der Betriebswirtschaft sind z. B. die Kreditwürdigkeitsprüfung oder die Kundenzuordnung. Beim produzierenden Gewerbe kann beispielsweise eine automatisierte Materialprüfung fehlerhafte Teile erkennen und aussortieren.

11.3.1 k-Nearest-Neighbors

Um Daten zu klassifizieren, müssen zunächst Klassen definiert sein. Im einfachsten Fall sind dies zwei verschiedene (gut/schlecht, kreditwürdig/nicht kreditwürdig, rot/blau, ...). Für eine gegebene Menge an Beobachtungen seien die Eigenschaften der Datenpunkte (die Variablenausprägungen) ebenso bekannt wie die zugehörige Klasse. Sollen nun neu beobachtete Datenpunkte einer Klasse zugeordnet werden, geschieht dies auf Basis der jeweiligen Variablenausprägungen. Das K-Nearest-Neighbors-Verfahren betrachtet die Datenpunkte, deren Variablenausprägungen am nächsten an denen eines neuen Datenpunktes liegen und ordnet den neuen Datenpunkt der Klasse zu, die mehrheitlich bei diesen „Nachbarn“ auftaucht. K ist dabei eine ungerade Zahl, damit eine eindeutige

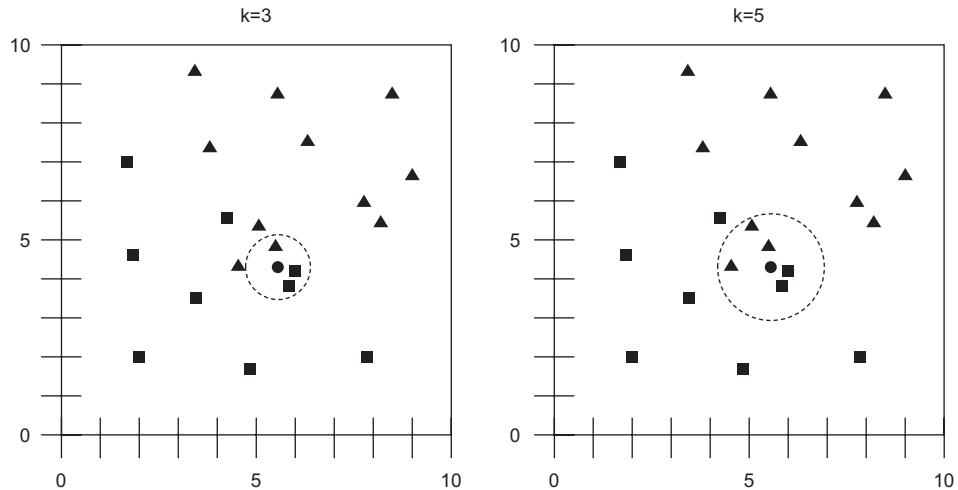


Abb. 11.5 K-Nearest-Neighbors-Klassifikation mit $k = 3$ (links) und $k = 5$ (rechts)

Mehrheit erzielt wird. Abb. 11.5 zeigt dies für $k = 3$ (links) und $k = 5$ (rechts) beispielhaft an einem neuen Datenpunkt (Kreis), der entweder den Quadraten (Klasse 1) oder den Dreiecken (Klasse 2) zugeordnet werden soll. Die Wahl von k kann das Ergebnis der Zuordnung erheblich beeinflussen, je nach Streuung der Werte und der Nähe der neuen Werte zum Rand des Datenraums.

11.3.2 Naive Bayes

Naive Bayes beschreibt eine Gruppe von Klassifikationsverfahren, die auf der („naiven“) Annahme beruhen, dass Eigenschaften, die bei Datenpunkten auftreten, unabhängig voneinander sind. Unter dieser Annahme können die Häufigkeiten des gemeinsamen Auftretens bestimmter Eigenschaften genutzt werden, um Wahrscheinlichkeiten für das Vorliegen einer Klasse unter bestimmten Bedingungen zu ermitteln. Der zugrunde liegende Satz von Bayes gibt den Verfahren den zweiten Teil des Namens. Die vielleicht bekannteste Anwendung von Naive-Bayes-Klassifikation findet sich in Spam-Filtern von E-Mail-Postfächern. Betrachtet man die Häufigkeit von bestimmten Begriffen, die in Spam-Mails auftauchen (also die *bedingte Wahrscheinlichkeit* für das Auftauchen), bspw. „Lotterie“ und „Gewinn“, so kann im Umkehrschluss errechnet werden, wie hoch die bedingte Wahrscheinlichkeit ist, dass eine neu eintreffende Mail als Spam zu klassifizieren ist, wenn die Begriffe „Lotterie“ und „Gewinn“ auftauchen. Naive-Bayes-Verfahren liefern trotz ihrer Einfachheit in vielen Anwendungsfällen gute Ergebnisse, insbesondere, wenn die zu analysierenden Variablen in den Daten gut gewählt werden (Witten et al. 2017). Weitere Ausführungen zum Einsatz von Bayes finden sich auch in Kap. 15.

11.3.3 Entscheidungsbäume

K-Nearest-Neighbor-Verfahren und Naive-Bayes-Verfahren bieten einen einfachen Einstieg in die Klassifikation, ihr Vorgehen ist aber in Teilen auf bestimmte Fragestellungen beschränkt und die Nachvollziehbarkeit der Klassifikation ist nicht immer klar gegeben, da entweder nur ein kleiner Teil der Daten betrachtet wird oder die Zuordnung auf einer umfangreichen Verformelung basiert. Entscheidungsbäume bieten einen alternativen Zugang zur Klassifikation. Sie zeichnen sich durch eine hohe Flexibilität aus und können für die Analyseempfänger verständlich dargestellt werden. Das Ergebnis ähnelt einem Gesprächsleitfaden. Beginnend von einem *Wurzelknoten*, der in der Regel zuoberst dargestellt wird und in dem zunächst alle Daten versammelt sind, verzweigt sich der Datenbestand durch Aufteilungen (*splits*) in feinere Äste und schließlich Blätter, in denen sich möglichst *homogene* Datenbestände finden, also solche, bei denen möglichst viele Elemente einer Klasse angehören. Abb. 11.6 zeigt das Prinzip auf Basis eines zweidimensionalen Datensatzes.

Links sind die Daten zweier Klassen auf zwei Dimensionen dargestellt. Eine erste Aufteilung wird durch einen Schnitt (gestrichelte senkrechte Linie) auf der X_1 -Achse vorgenommen. Alle Objekte, deren X_1 -Wert bei 5,5 oder kleiner liegt, fallen in einen Datenbereich, alle anderen in den zweiten. Die Datenbereiche können dann unabhängig voneinander weiter aufgeteilt werden, dargestellt durch die waagerechten Schnittlinien. Der Entscheidungsbaum auf der rechten Seite spiegelt das Prinzip wider. Auf oberster Ebene liegen 20 Elemente im Wurzelknoten. Soll für einen neuen Datensatz die zuzuordnende Klasse ermittelt werden, so wird dem Entscheidungspfad nach links gefolgt, wenn der X_1 -Wert des neuen Datensatzes kleiner oder gleich 5,5 ist. Bei Ankunft

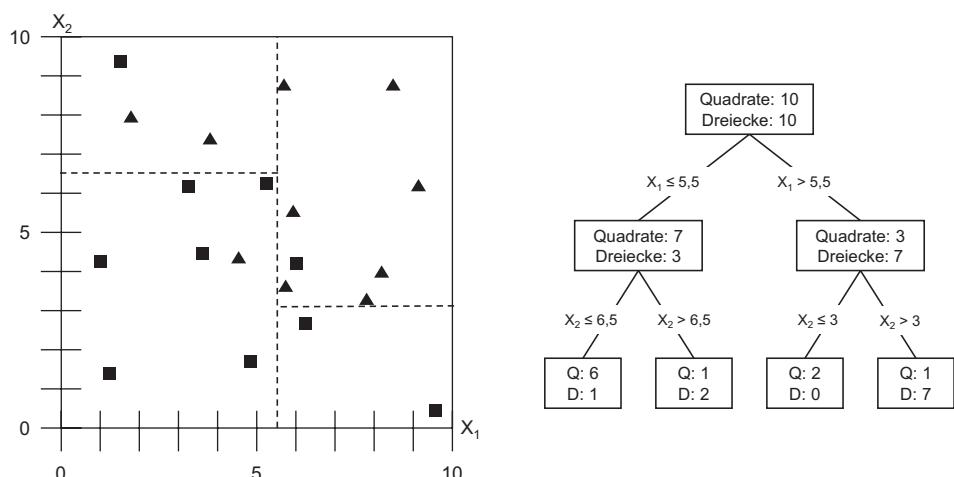


Abb. 11.6 Klassifikation als 2D-Darstellung (links) und als Entscheidungsbau (rechts)

in einem Blattknoten entscheidet in der Regel die Mehrheit der dort befindlichen Klassenmitglieder, wie zugeordnet wird. In einem „guten“ Entscheidungsbaum sind diese Blattknoten daher möglichst homogen, wie es im Beispiel bei dem dritten Blattknoten der Fall ist, in dem alle Elemente der Klasse Quadrat angehören.

Entscheidungsbäume splitten die Daten nach einzelnen Variablen in Abhängigkeit vom erzielten Informationsgewinn, dem Maß, in dem sich die „Unordnung“ reduziert bzw. eine gewollte Ungleichverteilung einstellt. Dabei kommen unterschiedliche Verfahren zum Einsatz, die mit numerischen aber auch kategorialen Eingangsvariablen (z. B. „Verheiratet: Ja/Nein“) umgehen können, binäre oder vielfache Splits erlauben und auf unterschiedliche Weise mit dem Problem des *Overfitting* umgehen. Overfitting beschreibt in der Klassifikation den Fall, dass ein Modell zu gut auf die Daten angepasst wurde, mit denen es erstellt (oder „trainiert“) wurde, sodass es für die allgemeine Nutzung nicht mehr tauglich ist. Auf Entscheidungsbäumen aufbauende, komplexere Verfahren werden im Bereich der fortgeschrittenen Datenanalyse diskutiert.

11.4 Clustering-Verfahren

Während bei der Klassifikation die Klassen bereits vorgegeben sind, sollen mit Hilfe von Clustering-Verfahren überhaupt erst Gruppierungen innerhalb einer Datenmenge gefunden werden. Es gibt daher auch kein „richtig“ oder „falsch“ am Ende einer solchen Analyse, sondern nur die Frage, wie klar die Gruppen getrennt werden können, welche Eigenschaften sie aufweisen und ob sie der Fragestellung dienlich sein können. Dabei werden im Wesentlichen zwei Hauptarten von Clustering-Verfahren unterschieden, die im Folgenden vorgestellt werden.

11.4.1 Hierarchische Verfahren

Hierarchische Verfahren arbeiten entweder *agglomerativ*, d. h. sie führen so lange Datenpunkte immer weiter zu möglichen Gruppen zusammen, bis kein weiteres Zusammenfassen mehr möglich ist, oder *divisiv*, d. h. sie führen den Prozess genau entgegengesetzt aus. Der Einfachheit halber und wegen der höheren Verbreitung in der Praxis werden hier nur agglomerative Verfahren vorgestellt.

Datenpunkte liegen in einem mehrdimensionalen Datenraum. Dimensionen (Variablen) können bei Menschen z. B. Größe, Alter, Gewicht, Anzahl der Kinder, etc. sein, bei Kunden z. B. die Anzahl der Käufe pro Monat oder der Umsatz und bei produzierten Autoreifen z. B. der Durchmesser, der Bremsweg in einem standardisierten Test oder die Laufleistung in Kilometern. Eine Ähnlichkeit zwischen zwei Datenpunkten wird immer dort angenommen, wo deren Abstand gering ist. Ziel des Clusterings ist es, ähnliche Datenpunkte zusammenzufassen und unähnliche in unterschiedliche Gruppen einzuteilen. Daher wird zunächst eine Distanzmatrix bestimmt, die für alle Datenpunkte

angibt, wie weit entfernt sie voneinander im Datenraum liegen, wie unähnlich sie sich also sind. Diese Distanzmatrix kann z. B. mithilfe der *euklidischen Distanz* (Wurzel der summierten quadratischen Abweichungen in den einzelnen Dimensionen zweier Punkte) bestimmt werden. Neben der euklidischen Distanz sind weitere Distanzmaße denkbar, die das Clustering-Ergebnis maßgeblich beeinflussen können, darunter auch Maße, die sich weniger auf die absoluten Werte in den einzelnen Dimensionen beziehen, sondern auf gleichförmige Veränderungen über die Dimensionen hinweg. Um im mehrdimensionalen Raum die gleichartige Behandlung der Daten sicherzustellen, bietet es sich an, diese zu standardisieren (James et al. 2017).

In einem ersten Schritt werden nun die beiden Punkte mit der geringsten Distanz zusammengefasst. Der so entstehende Cluster wird wie ein Datenpunkt behandelt, sodass im nächsten Schritt wieder die zwei nächst beieinander liegenden Datenpunkte/ Cluster verschmolzen werden können. Der Prozess stoppt, wenn nur noch ein großer Cluster vorhanden ist. Entscheidend für das Clustering-Ergebnis ist die Frage, wie die Distanz von einem neu gebildeten Cluster zu den verbleibenden Datenpunkten und Clustern behandelt wird. Linkage-Verfahren betrachten dafür z. B. die minimale (*single linkage*) oder maximale (*complete linkage*) Distanz, die sich bei der Betrachtung der einzelnen Cluster-Punkte ergeben würde. Andere Verfahren treffen die Auswahl der zu verschmelzenden Punkte basierend auf den sich ergebenden Eigenschaften, bspw. den minimalen Varianzen, innerhalb der Cluster (*Ward-Verfahren*). Eine umfassende Darstellung diverser Verfahren und ihrer Eigenschaften bieten Legendre und Legendre (2012). In jedem Fall lässt sich der Clustering-Prozess nachträglich über ein *Dendrogramm* nachvollziehen, das darstellt, welche Punkte in welcher Reihenfolge verschmolzen wurden. Abb. 11.7 zeigt ein solches Dendrogramm (rechts), bei dem die

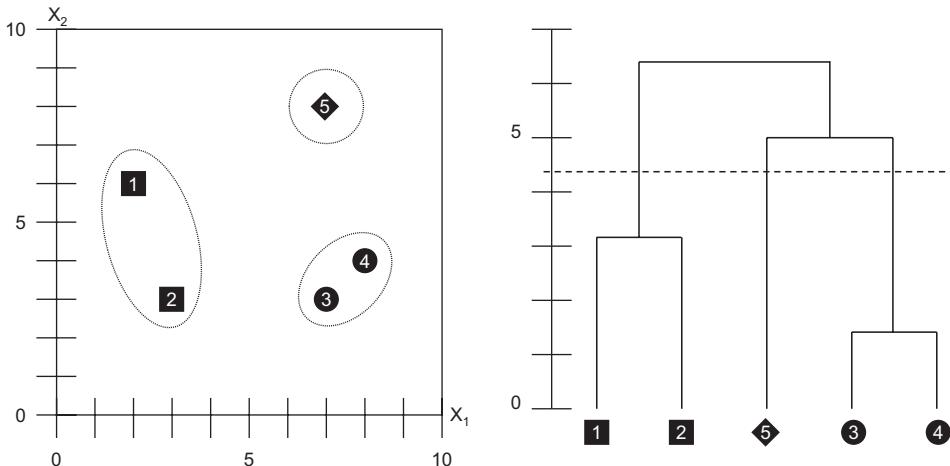


Abb. 11.7 Datenbasis (links) mit identifizierten Cluster-Zugehörigkeiten der Datenpunkte und Dendrogramm (rechts) eines complete-linkage-Clusterings auf diesen Daten

Datenpunkte (waagerechte Achse) zusammengeführt werden. Dort, wo eine waagerechte Linie zu sehen ist, lässt sich an der senkrechten Achse die Distanz ablesen, die überwunden werden musste, um die Punkte zusammenzuführen. Auf Basis dieses Dendrogramms kann eine geeignete Cluster-Zahl bestimmt werden (hier: 3 Stück, mit der gestrichelten Linie als Schnitt durch die Zusammenführungen dargestellt), die eine hohe Heterogenität der Cluster untereinander, aber eine große Homogenität in den Clustern vermuten lässt. Zum Vergleich sind links die zugrunde liegenden Datenpunkte zweidimensional dargestellt, die unterschiedlichen Formen und die Ellipsen repräsentieren die Zugehörigkeit zu den drei identifizierten Clustern.

11.4.2 Partitionierende Verfahren

Partitionierende Clustering-Verfahren betrachten die Daten nicht Punkt für Punkt, sondern weisen der gesamten Datenmenge auf einmal Cluster-Zugehörigkeiten zu. Das bekannteste Verfahren ist das *k-means*-Verfahren, das die Zielzahl an Clustern (k) bereits als Parameter vordefiniert bekommt. Das Vorgehen des Algorithmus ist je nach Implementation leicht unterschiedlich, folgt aber weitgehend folgendem Schema:

1. Ordne jedem Datenpunkt einen (zufälligen) Cluster zu (Startlösung)
2. Berechne den Mittelpunkt aller Cluster
3. Weise jedem Datenpunkt dem Cluster zu, dessen Mittelpunkt er am nächsten liegt
4. Wenn die Lösung sich nicht mehr ändert, ist der Zielzustand erreicht, ansonsten gehe wieder zu Schritt 2

Die Wahl der Startlösung beeinflusst maßgeblich, wie schnell das Clustering-Verfahren zu einem stabilen Ergebnis kommt und wie dieses aussieht. Eine schlechte Startlösung führt zu einem stabilen Zustand, der aber nicht optimal im Sinne der möglichen Homogenität der k Cluster ist oder erst nach vielen Iterationen erreicht wird. Für das Finden guter Startlösungen und effizienter Neuzuweisungen wurde und wird daher in der Data-Science-Forschung einiges an Aufwand betrieben (Fränti und Sieranoja 2019).

11.5 Assoziationsanalyse

Die Hauptfragestellung in der Assoziationsanalyse ist das Finden von Wenn-Dann-Regeln. Im Kontext der Betriebswirtschaft ist das typische Beispiel die Identifikation von Gegenständen, die in einem Supermarkt zusammen gekauft werden, weshalb auch vereinfacht von *Warenkorbanalyse* gesprochen wird, obwohl die Assoziationsanalyse generell weitere Anwendungen erfasst. Die hier vorgestellte Form kann auf alle Daten angewendet werden kann, die in transaktionaler Form aufbereitet werden können, und

der Beitrag folgt in diesem Abschnitt den Ausführungen und Begrifflichkeiten zur Warenkorbanalyse von Bramer (2016).

Eine Transaktion ist eine beobachtete Menge von Gegenständen (*Items*), die zusammen ein *Itemset* bilden, beispielsweise {Bier, Brokkoli, Brot, Butter} oder {Chips, Hammer, Windeln}, und die sich wiederum in Teilmengen aufteilen lassen. Je häufiger eine Kombination von Items beobachtet wird, desto relevanter wird sie für eine Analyse. Ausgedrückt wird dies durch den Support (S), der die Anzahl von Transaktionen, in denen ein Itemset vorkommt ins Verhältnis zu der Gesamtzahl der Transaktionen setzt. Itemsets mit hohem Support können genutzt werden, um Regeln abzuleiten, bspw. „Wer Brot kauft, kauft auch Butter.“, was notiert wird als $\{Brot\} \rightarrow \{Butter\}$ oder allgemein als $L \rightarrow R$, wobei L die „Linke-Hand-Seite“ (*antecedent*) der Regel ist und R analog die „Rechte-Hand-Seite“ (*consequent*). Zur Generierung dieser Regeln wird häufig der *Apriori-Algorithmus* eingesetzt, der die Regeln zunächst basierend auf einem als ausreichend definierten Support erstellt (eine Regel, die sich auf Items bezieht, die nur einmal im Jahr verkauft werden, schafft keinen Mehrwert an Information) und sie abschließend nach Prüfung ihrer *Confidence* beibehält oder verwirft. Die Confidence (C) ist dabei die Zuverlässigkeit einer Regel, also die Angabe, wie häufig sie zutrifft. Eine Confidence nahe 100 % besagt, dass die Regel fast immer zutrifft. Abschließend können die Regeln noch auf ihre Nützlichkeit (gemessen durch den *Lift*) geprüft werden. So können Regeln zwar ggf. immer zutreffen ($C = 100\%$), aber dennoch nutzlos sein. Z. B. ist an einer Tankstelle die Regel „Wer einen Schokoriegel kauft, kauft auch Treibstoff.“ möglicherweise sehr zuverlässig, aber da nahezu jeder Kunde Treibstoff kauft, ist sie für die Organisation des Verkaufsbetriebs nicht nützlich. Mit der Assoziationsanalyse schließt hier der Kanon der vier großen Methodengruppen des Data Mining.

11.6 Ergänzende Überlegungen, Software und Tools

Die vorherigen Abschnitte geben einen kurzen Überblick über die „klassischen“ Methoden des Data Mining, die als Basis für eine Vielzahl an Analysen in Unternehmen dienen können. Zu fast allen vorgestellten Verfahren und zu jeder der Methodengruppen existieren weitergehende Überlegungen und fortgeschrittene Analyseformen. Wie zu Beginn erläutert, bieten die aktuellen Entwicklungen in den Bereichen Maschinelles Lernen, Künstliche Intelligenz, Big Data und den verwandten Feldern auch immer wieder neue oder neu entdeckte Methoden, von denen derzeit z. B. Künstliche Neuronale Netze oder Deep Learning intensiv diskutiert werden. Diese Techniken setzen nicht immer die hier beschriebenen fundamentalen Analysetechniken voraus, sind aber häufig für spezielle Fragestellungen gedacht. Personen, die als Data Analyst arbeiten wollen oder zumindest in diesem Bereich Expertise suchen, sollten aber zunächst einen Überblick gewinnen und können mit den hier vorgestellten Methoden schnell an Probleme des (betrieblichen) Alltags herantreten ohne hochspezialisierte Software oder Infrastrukturen nutzen zu müssen.

Für die Anwendung existieren diverse, häufig frei verfügbare Programmiersprachen, Statistiksprachen oder -pakete, Software-Sammlungen und Online-Ressourcen. Im Bereich der fundamentalen Datenanalyse haben sich in den letzten Jahren R (als Statistik-Programmiersprache und -software) und Python (als Programmiersprache, die durch diverse Pakete mit Statistik-Funktionen versehen werden kann) etabliert. Beide sind frei verfügbar, es existieren für beide umfangreiche Tutorials, ergänzende Funktionen und aktive Online-Communities. Umfragen zeigen wechselnde Beliebtheitswerte, wobei derzeit eine Entwicklung hin zu einer Python-Präferenz auszumachen scheint. Die Gründe dafür sind – genau wie die Interpretationen von Umfragen und deren Datenerhebung – wie immer vielfältig. Einsteiger in die Thematik finden in beiden Varianten exzellente Werkzeuge, deren Erlernen jedoch insbesondere für nicht programmier-affine Personen etwas Zeit brauchen kann. Alternativ bieten namhafte Softwarehäuser Analyse-Programme an, entweder als eigenständige Lösung oder als Bestandteil von Büro- oder Unternehmenssoftware. Ergänzt wird das Angebot durch Cloud-Lösungen und kostenlose grafisch gestützte Software.

Literatur

- Bramer, M.A.: Principles of data mining, 3. Aufl. Springer, London (2016)
- Fräntti, P., Sieranoja, S.: How much can k-means be improved by using better initialization and repeats? *Pattern Recogn.* **93**, 95–112 (2019)
- Goodman, S.: A Dirty Dozen: Twelve P-Value Misconceptions. *Semin. Hematol.* **45**(3), 135–140 (2008)
- James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer, NY (2013)
- Kemper, H.-G., Baars, H., Mehanna, W.: Business Intelligence – Grundlagen und praktische Anwendungen: Eine Einführung in die IT-basierte Managementunterstützung, 3. Aufl. Vieweg+Teubner, Wiesbaden (2010)
- Kim, B.: *Understanding Diagnostic Plots for Linear Regression Analysis*. <https://data.library.virginia.edu/diagnostic-plots/> (2015). Zugegriffen: 29. Juli 2020
- Legendre, P., Legendre, L.: Numerical ecology, 3. Aufl. Elsevier, Amsterdam (2012)
- Meier, A.: Werkzeuge der digitalen Wirtschaft: Big Data, NoSQL & Co. essentials. Springer, Wiesbaden (2018)
- Mittag, H.-J.: Statistik: Eine Einführung mit interaktiven Elementen, 5. Aufl. Springer, Berlin (2017)
- Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Gröschel, A., Brauner, D., Badura, D., Passlick, J., Kloker, S., Gölzer, P., Kerzel, U., Rissler, R., Alekozai, E., Binder, H., Welter, F., Badewitz, W., Felderer, M., Rohde, H., Prothmann, M., Dann, D., Lanquillon, C., Gehrke, N. (2020) DASC-PM v1.0. Ein Vorgehensmodell für Data-Science-Projekte. Hamburg, Elmshorn
- Weber, F.: Künstliche Intelligenz für Business Analytics: Algorithmen. Plattformen und Anwendungsszenarien, Springer Fachmedien, Wiesbaden (2020)
- Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data mining: practical machine learning tools and techniques, 4. Aufl. Morgan Kaufmann, Cambridge (2017)
- Wooldridge, J.M.: Introductory econometrics: a modern approach, 6. Aufl. Cengage, Boston (2015)



Fortgeschrittene Verfahren zur Analyse und Datenexploration, Advanced Analytics und Text Mining

12

Jens Kaufmann

Zusammenfassung

Leistungsfähige Computersysteme und umfassende Datenbestände ermöglichen den Einsatz diverser, teils hochkomplexer Analyseverfahren, die nicht nur numerische Werte, sondern auch Texte und weitere Datenarten auswerten können. Grundstein einer Datenanalyse ist das Verständnis darüber, welche Daten vorliegen und welche Fragestellungen überhaupt beantwortet werden sollen. Der Beitrag zeigt, wie Daten grafisch aufbereitet werden können, um einen schnellen Einblick in deren Struktur zu erhalten und beschreibt unterschiedliche Analyseverfahren, die über die fundamentalen Methoden hinausgehen. Dabei wird auch auf die Bewertung von Entscheidungen und die Güte von einzelnen Verfahren eingegangen sowie ein Ausblick auf spezialisierte Verfahren zur Zeitreihen- und Text-Analyse gegeben. Leserinnen und Leser erhalten einen Einblick in die Möglichkeiten fortgeschrittener Methoden, möglicher Schwerpunkte zur weiteren Vertiefung und Informationen zu weiteren Schritten in der Datenanalyse.

12.1 Einleitung

Fundamentale Verfahren der Datenanalyse lassen sich über die vier Hauptgruppen des Data Mining (Prognose, Klassifikation, Segmentierung und Assoziationsanalyse) vergleichsweise gut einordnen und es fällt leicht, zu jeder der Gruppen wenigstens eine häufig eingesetzte Methode zu identifizieren und zu beschreiben (vgl. Kap. 11). Selbst,

J. Kaufmann (✉)

FB 08, Hochschule Niederrhein, Mönchengladbach, Deutschland

E-Mail: jens.kaufmann@hs-niederrhein.de

wenn die Datenanalyse nur als ein Schritt von Data-Science-Projekten verstanden wird, wie es z. B. beim Data Science Process Model (DASC-PM) der Fall ist (Schulz et al. 2020), ist die Vielfalt an Methoden, Werkzeugen und möglichen Analysen deutlich umfangreicher als es diese vier Gruppen beschreiben können. Dieser Beitrag greift ausgewählte weitere Methoden und Grundlagen der Datenanalyse auf, die im betrieblichen Alltag nützlich sein können. Die Auswahl folgt dabei rein pragmatischen Gesichtspunkten und erhebt keinen Anspruch auf Vollständigkeit oder eine besonders ausgewogene Abdeckung des Feldes. So wird mit der Datenexploration und -darstellung nochmal kurz auf die initiale Betrachtung eines Datensets eingegangen, die es ermöglicht, die vorliegenden Daten und deren Eigenschaften besser zu verstehen, um die richtigen Analysemethoden sowie Datenvorbereitungen auswählen zu können und um den Analysepfad auch adressatengerecht aufzubereiten zu können. Die Principal Component Analysis ergänzt diese Betrachtungen mit der Möglichkeit, sehr komplexe Datenbestände auf einfache Repräsentationen zu reduzieren. Random Forests und Logistische Regression beschreiben weitere Klassifikationsverfahren, da die Klassifikation eine der häufigsten Fragestellungen im Geschäftsbetrieb ist. Die auch zu diesem Bereich gehörenden Künstlichen Neuronalen Netze finden sich hingegen in Kap. 13 und Kap. 14 dieses Buches beschrieben. Überlegungen zur Entscheidungsbewertung, Zeitreihenanalyse und eine Einführung in Text Mining stellen im Anschluss dar, welche Möglichkeiten der Datenanalyse es für ausgewählte spezielle Fälle gibt, bevor der letzte Abschnitt dieses Beitrags aufzeigt, in welche Richtungen sich Datenanalyse-Interessierte noch weiter entwickeln können.

12.2 Datenexploration und -darstellung

Das menschliche Gehirn ist darauf ausgelegt, Muster zu erkennen und auszuwerten (Mattson 2014). Auf Basis dieser Muster und einiger Basisinformationen über die Daten können erste Hypothesen erstellt und passende Analyseverfahren ausgewählt werden. Dabei ist es insbesondere bei großen Datenmengen in der Regel angenehmer und zielführender, eine Grafik zu betrachten als zehntausend Zeilen numerischer Werte in einer Tabelle. Zu Beginn einer Analyse können die Daten also explorativ, aber nicht planlos betrachtet werden. Dazu bieten sich zusammenfassende Darstellungen an, um Muster auf höherer Ebene sichtbar zu machen. Abb. 12.1 stellt eine Datenbasis von 229 Ländern der Erde dar (Beobachtungen), die jeweils über drei Eigenschaften (Variablen) verfügen (Geburten pro 1000 Einwohner, Todesfälle je 1000 Einwohner, Bevölkerungswachstum in Prozent).

Gezeigt wird ein *Matrixplot*, der für jede Variable eine Spalte und eine Zeile vorsieht. In der Diagonalen sind für die Variablen jeweils *Histogramme* gezeigt, die die Daten in Gruppen zusammenfassen (Abszisse / x-Achse) und die Anzahl von Datenpunkten in der jeweiligen Gruppe zählen (Ordinate / y-Achse). Dies erlaubt eine erste Sicht auf die Verteilung der Werte mit Bezug auf die einzelnen Variablen. Die drei *Punktdiagramme* im

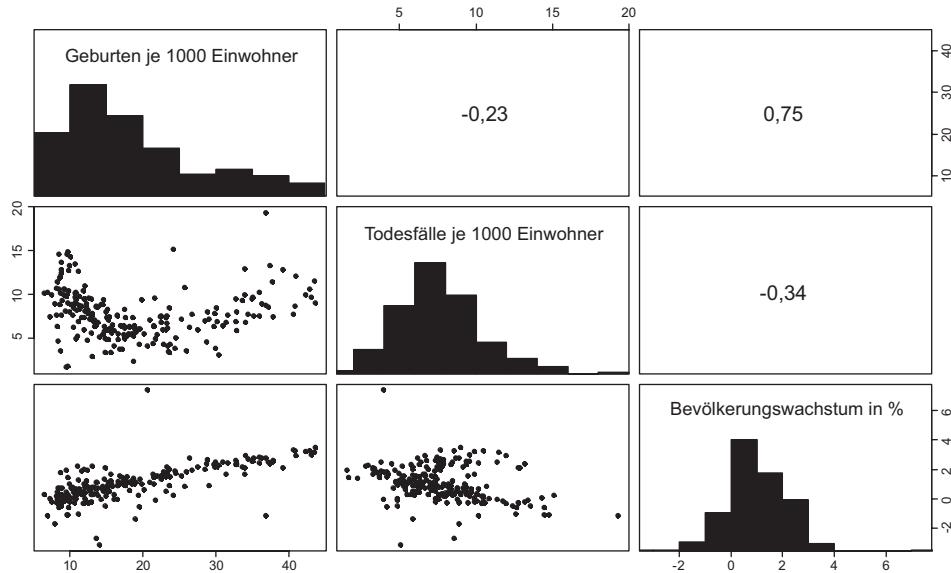


Abb. 12.1 Matrixplot mit Histogrammen und Korrelationskoeffizienten (229 Länder)

linken, unteren Teil zeigen jeweils im Schnittpunkt der Variablen die 2D-Darstellungen der beiden beteiligten Variablen. Die weiteren Variablen werden dabei ignoriert. Im rechten, oberen Teil sind die *Korrelationskoeffizienten* angegeben, die mit einem Wert von -1 über 0 bis +1 einen negativen, nicht vorhandenen oder positiven Zusammenhang der Variablen beschreiben. Im dargestellten Beispiel zeigt sich, dass bei steigender Geburtenrate auch tendenziell ein größeres Bevölkerungswachstum vorliegt (0,75). Geburten und Todesfälle hingegen stehen über alle Länder hinweg nur in einem schwach negativen Zusammenhang. Korrelationsanalysen werden insbesondere bei großen Datensätzen und im Zuge von Big Data häufig durchgeführt. Entscheidend ist aber, dass hier nur ein beobachteter statistischer Zusammenhang ausgedrückt wird und keine Aussage über einen kausalen Zusammenhang getätigt wird. Die Gründe für bestimmte Eigenschaften der Datenbasis müssen in weiteren Analysen aufgearbeitet und im Nachgang mit Fachwissen interpretiert werden.

12.3 Principal Component Analysis

Im vorherigen Abschnitt zeigt sich bereits ein grundsätzliches Problem der Datendarstellung: Je mehr Variablen (oder Dimensionen) die Daten haben, desto problematischer ist es, alle Zusammenhänge auf einmal darzustellen. Mehr als drei Dimensionen sind nur sehr schwer nachvollziehbar in einer Grafik abbildbar. Um die Daten dennoch handhabbar

zu gestalten und insbesondere auch um (Un-)Ähnlichkeiten oder Gruppen in den Daten auszumachen und im zweidimensionalen Raum grafisch darstellen zu können, muss die Anzahl der Dimensionen verringert werden. Ein Standardwerkzeug dafür ist die *Principal Component Analysis* (PCA) oder *Hauptkomponentenanalyse*. Ziel hierbei ist es, die bestehenden Variablen geschickt miteinander zu kombinieren und dadurch neue, künstlich erzeugte Variablen zu erschaffen, die die Varianz in den Daten mit weniger Dimensionen möglichst gut erklären. Mathematisch betrachtet ist jede *Principal Component* (Hauptkomponente) eines Datensets eine normalisierte Linearkombination der Variablen. Für die erste Hauptkomponente gilt:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

Dabei drücken die *loadings* $\varphi_{11}, \dots, \varphi_{p1}$ aus, wie stark jede Variable auf die entsprechende Hauptkomponente wirkt (James et al. 2013). Die Ergebnisse lassen sich dann in einem *Biplot* darstellen, ein Beispiel ist in Abbildung Abb. 12.2 gegeben.

Dargestellt sind die Daten von 85 Profifußballern, die nach vier Variablen beurteilt wurden (Ausdauer, Beschleunigung, Geschwindigkeit, Kopfballstärke). Um diese vierdimensionalen Daten in einer zweidimensionalen Abbildung darstellen zu können, wurde eine PCA durchgeführt und die ersten beiden Hauptkomponenten bilden die Grundlage für die 2D-Darstellung (links). Die Variablen wurden standardisiert, um alle Variablen in gleicher Stärke einfließen zu lassen. Jeder Punkt ist eine Beobachtung. Die vier Pfeile repräsentieren die Variablen und geben die loadings der Hauptkomponenten an. Je weiter ein Pfeil nach rechts zeigt, desto mehr positiven Einfluss hat die Variable auf Hauptkomponente 1, je weiter nach oben er zeigt, desto mehr positiven Einfluss auf Hauptkomponente 2. Pfeile, die nach links oder unten zeigen wirken negativ auf die Hauptkomponenten. Es lässt sich erkennen, dass der „Wert“ eines Fußballers auf Hauptkomponente 1 stark steigt, wenn seine Beschleunigungswerte hoch sind. Die loadings können nun genutzt werden, um die Hauptkomponenten zu interpretieren, was durchaus ein gewisses Maß an Kreativität erfordern kann oder nur sehr schwer möglich ist. Hauptkomponente 1 scheint die „Schnelllauffähigkeit“ abzubilden, Hauptkomponente 2 die „Langlauffähigkeit“. Die Kopfballstärke wirkt negativ auf beide Komponenten und ist nur schwierig in die Interpretation zu integrieren. Möglicherweise wäre hierzu die Betrachtung einer weiteren Hauptkomponente erforderlich. Die ersten beiden Hauptkomponenten (von vier) erklären zusammen aber bereits fast 80 % der Varianz in den Daten (siehe Abbildung rechts). Mit der Principal Component Analysis ist es so möglich, einen großen Teil der Varianz mit einer kleinen Anzahl an Dimensionen abzubilden.

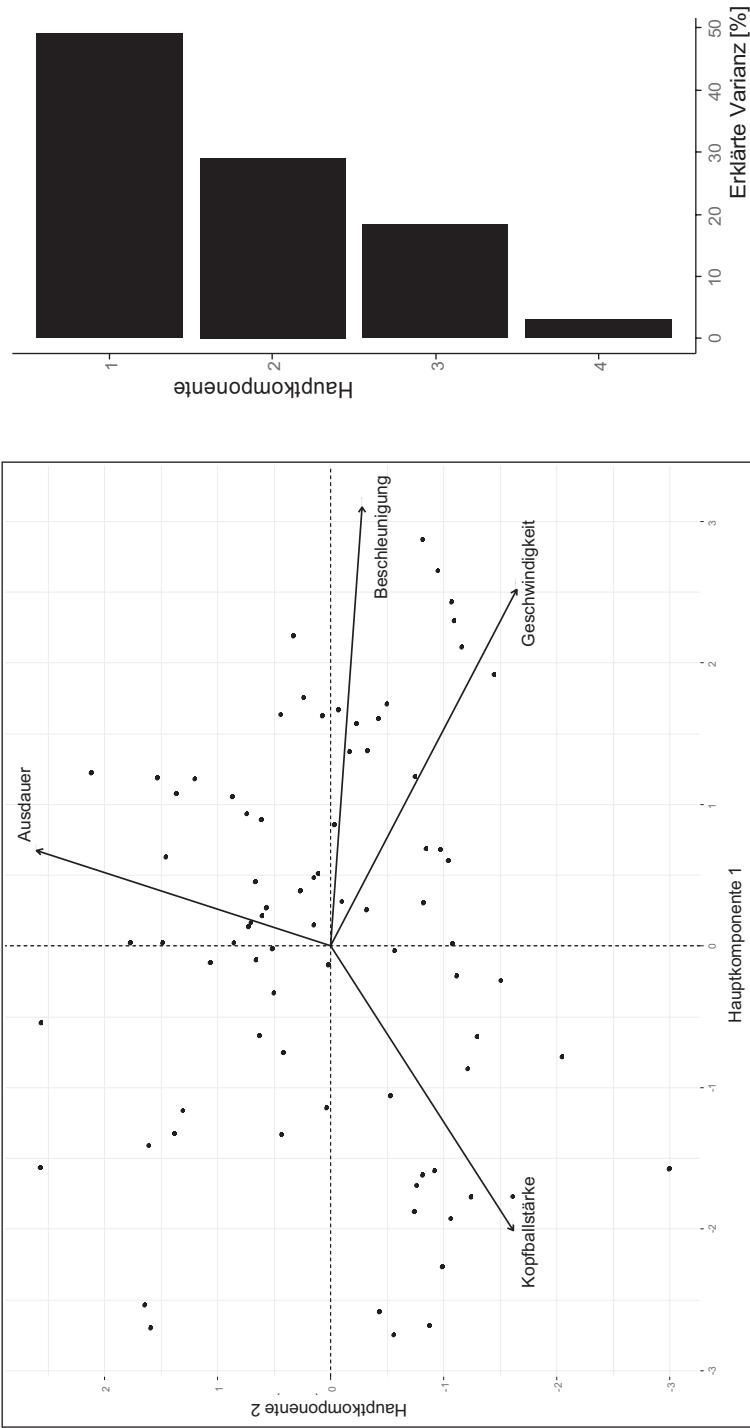


Abb. 12.2 Biplot (links) und Varianzerklärungsanteile (rechts) einer Hauptkomponentenanalyse von Fußballdaten

12.4 Random Forests

Entscheidungsbäume wurden als eine Methode der fundamentalen Datenanalyse beschrieben (vgl. Abschn. 11.3). Wie bei allen Klassifikationsaufgaben zeigt sich die Güte eines Modells aber erst, wenn es auf Daten verwendet wird, die nicht genutzt wurden, um das Modell zu erstellen. Um die Gefahr eines Overfitting, also einer Überanpassung der Daten an ein Modell, die im Nachgang zu schlechten Klassifikationsergebnissen bei Real-Daten führen würde, zu verringern, existieren diverse Methoden, die die Erstellung und Verwendung von Entscheidungsbäumen verbessern oder erweitern. Dabei wird im einfachen Fall die vorliegende Datenmenge vor der Modellerstellung in zwei, besser drei Teilmengen zerlegt. Das *Training Set* wird zur eigentlichen Modell-Erstellung genutzt, das *Validation Set* wird zur Verbesserung der Modellparameter verwendet und das *Test Set* dient als vollständig unabhängiges Datenset zur Modellgütemessung (Chicco 2017). Liegen nicht genug Daten vor, um diese sinnvoll splitten zu können, können Cross-Validation-Ansätze oder Bootstrapping verwendet werden, die mehrfach auf gleiche Weise ein Modell auf den Daten erstellen (dies können leicht hunderte von Varianten oder mehr sein), dabei aber immer nur einzelne Datensätze zur Bewertung (wie ein Mini-Test-Set) auslassen und teilweise sogar Daten mehrfach verwenden. Für Entscheidungsbäume lässt sich dieses Verfahren noch dadurch erweitern, dass auch nicht immer alle möglichen Variablen in das Modell gegeben werden, sodass einzelne Modelle gezwungen sind auch die Variablen zu berücksichtigen, die eigentlich keinen besonders großen Informationsgewinn versprechen und darum in den meisten Modellen nie berücksichtigt würden. Fasst man alle erstellten Modelle (alle Bäume) zusammen, ergibt sich nicht ein großer Baum, sondern vielmehr eine Ansammlung von Bäumen (ein Wald), der auf der Basis von zufälligen Daten- und Variablen-Auswahlen erstellt wurde, ein *Random Forest* also. Bei der Klassifikation bisher unbekannter Werte kommen alle Modelle zum Tragen und das Ergebnis wird z. B. per Mehrheitsentscheid bestimmt. Die Klassifikationsgüte dieser Random-Forest-Modelle übertrifft häufig die von einzelnen Entscheidungsbäumen (James et al. 2013).

12.5 Logistische Regression

In manchen Fällen interessiert nicht nur die Klassifikation einer Beobachtung, sondern vielmehr die Wahrscheinlichkeit, mit der ein Datenpunkt zu einer bestimmten Klasse gehört. Die *logistische Regression* folgt im Grundprinzip dabei der Idee, dass die Daten zunächst einfach unterteilt werden können – so als ob eine Linie durch die Daten gezogen wird und alles, was auf der linken Seite liegt, als „Klasse 1“, alles, was rechts von der Linie liegt, als „Klasse 2“ eingeordnet wird. Wird eine neue Beobachtung zugeordnet, so kann geprüft werden, wie weit diese von der teilenden Linie, der *Entscheidungsgrenze*, entfernt liegt (Provost und Fawcett 2017). Liegt der Wert weit entfernt, sollte die Wahrscheinlichkeit, dass er tatsächlich der vermuteten Klasse zugehörig ist, größer sein als

wenn er nah an dieser Grenze liegt. Da der Datenpunkt prinzipiell beliebig weit weg von der Entscheidungsgrenze liegen kann, Wahrscheinlichkeiten der Zugehörigkeit zu einer Klasse aber zwischen null und eins liegen (also zwischen 0 % und 100 %), muss eine Art von Umrechnungsfunktion gebildet werden. Dazu werden Wahrscheinlichkeiten als *Chancen* ausgedrückt. Eine Wahrscheinlichkeit von 50 % entspricht einer Chance von 50:50, 90 % sind 90:10 (oder einfach 9) und die Chance im Lotto zu gewinnen (sechs Richtige plus Zusatzzahl) liegt bei 1:139.838.160 oder etwa 0,000.007.151 %. Eine ausführliche Beschreibung des Gedankenganges und der dahinterliegenden Rechnung findet sich ebenfalls bei Provost und Fawcett (2017). Durch Logarithmieren dieser Chancen ergibt sich zum einen der passende Wertebereich, zum anderen nimmt die Funktion, die den Abstand in Wahrscheinlichkeiten umrechnet, eine S-förmige Form an (*Sigmoidfunktion*). Dies führt dazu, dass bei sich bei kleinen Distanzen die Zuordnungs-Wahrscheinlichkeiten schnell verändern und dass sie bei großen Distanzen recht stabil nahe 0 oder 1 liegen. Dieses Verhalten ist gewünscht, da wie beschrieben bei einem großen Abstand auch eine recht hohe Sicherheit in der Zuordnung vermutet wird. Gesucht wird nun ein Modell, das die Entscheidungsgrenze so festlegt, dass für bekannte Werte der zwei Klassen im Durchschnitt die höchsten bzw. niedrigsten Werte der Wahrscheinlichkeiten ermittelt werden.

12.6 Entscheidungsbewertung

Wie in Abschn. 12.4 beschrieben, können Datenmengen in unterschiedliche Sets zerlegt werden. Nachdem ein Modell erstellt oder trainiert wurde, werden die Daten des Test Sets verwendet, um es auf seine Güte zu prüfen. In einer Konfusionsmatrix kann nun einander gegenübergestellt werden, wie häufig welche Klasse vorhergesagt wurde und wie häufig welche Klasse tatsächlich vorlag. Im Optimalfall wird eine Genauigkeit (Accuracy) von 100 % erreicht. Am Beispiel eines medizinischen Tests wären dann alle Gesunden als gesund vorhergesagt worden und alle Kranken als krank. Es lassen sich aus den vier Zahlen der Konfusionsmatrix diverse Kennzahlen ableiten, die teilweise diverse Synonyme aufweisen. Für die unterschiedlichen Kennzahlen sei daher auf weitere Literatur verwiesen, z. B. James et al. (2013) oder Witten et al. (2017). Diese Kennzahlen verändern sich aber nicht nur, wenn grundlegend andere Modelle verwendet werden, sondern auch, wenn bestehende Modelle anders eingestellt werden. Beispielsweise wird standardmäßig bei der logistischen Regression ab einer Wahrscheinlichkeit von größer 0,5 (=50 %) der Datenpunkt der entsprechenden Klasse (hier: Test positiv, also Patient krank) zugeordnet. Dies führt in der Realwelt auch teilweise zu falschen Zuordnungen. Je nach Anwendungsgebiet (Medizinische Tests, Produktionsfehlerprüfung, Kreditwürdigkeitsprüfung, ...) kann es daher geboten sein, die Wahrscheinlichkeitsschwelle, den *threshold*, höher oder niedriger anzusetzen. So könnte z. B. gefordert sein, dass ein Patient erst dann als krank eingestuft wird, wenn die Wahrscheinlichkeit bei 80 % liegt. Hohe Schwellwerte verhindern dabei tendenziell „falsch positive“

		Beobachtungen	
		Positiv	Negativ
Vorhersagen	Positiv	72	56
	Negativ	12	60

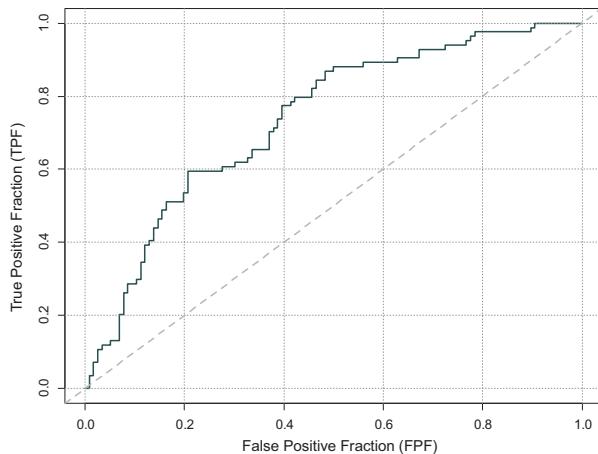


Abb. 12.3 Konfusionsmatrix (links, threshold = 0,5) und ROC-Kurve (rechts) eines fiktiven medizinischen Tests

Zuordnungen, allerdings nimmt auch die Anzahl der „echt positiven“ Zuordnungen ab. Die beiden zugehörigen Kennzahlen *True Positive Fraction* (TPF) und *False Positive Fraction* (FPF) lassen sich in einem Diagramm in Abhängigkeit des gewählten Schwellwertes gegeneinander abtragen und erzeugen eine *Receiver-Operating-Characteristics*-Kurve, kurz *ROC-Kurve*. Abb. 12.3 stellt eine beispielhafte Konfusionsmatrix eines medizinischen Tests (links, Schwellwert 0,5) und eine dazugehörige ROC-Kurve (rechts) dar. Jede „Stufe“ der Kurve steht für eine Veränderung des Schwellwertes. Dort, wo die Kurve den größten Abstand von der Diagonale aufweist, arbeitet das Modell nach diesen Maßstäben möglichst präzise, die Auswahl, wie viele „Fehlalarme“ in Kombination mit wie vielen echt positiven Zuordnungen tragbar sind, muss aber immer problemspezifisch getroffen werden. Grundsätzlich gilt, dass Modelle, die in Summe eine größere Fläche unter sich einschließen als andere (*Area under the Curve*, kurz: AUC), die besseren Modelle sind (Provost und Fawcett 2017).

12.7 Zeitreihenanalyse

Zeitreihen sind Daten, die die Entwicklung einer oder mehrerer Variablen über einen Zeitraum hinweg abbilden, indem für diverse Zeitpunkte der dann jeweils aktuelle Wert der Variablen notiert wird. Dies können bspw. Börsenkurse sein, aber auch Verkäufe oder die Temperatur und Dichte eines Werkstücks im Produktionsprozess. Daten dieser Art werden als Längsschnittdaten bezeichnet – im Gegensatz zu den bisher vorgestellten Querschnittsdaten, die Werte vieler Variablen von vielen Objekten zu einem Stichzeitpunkt beschreiben. Für diesen Beitrag wird angenommen, dass eine Zeitreihe nur die

Werte einer Variable erfasst. Damit gilt, dass folgende Datenmenge aus m Objekten zu n Zeitpunkten vorliegt:

$$\mathbf{Y} = \{\mathbf{y}_{1,1}, \dots, \mathbf{y}_{i,j}, \dots, \mathbf{y}_{m,n}\}$$

Backhaus et al. (2018) beschreiben als eine wesentliche Fragestellung die Untersuchung kausaler Zusammenhänge zwischen mehreren Zeitreihen, als zweite wesentliche Fragestellung die Zeitreihenextrapolation, also die Prognose zukünftiger Werte, für die sie eine fünfstufige Vorgehensweise vorschlagen:

1. Visualisierung der Zeitreihe
2. Formulierung eines Modells
3. Schätzung des Modells
4. Erstellung von Prognosen
5. Prüfung der Prognosegüte

Die Visualisierung wird mithilfe eines Punkt-, Linien- oder Säulendiagramms vorgenommen. Eine beispielhafte Darstellung wird im Verlauf des Abschnitts gezeigt. Die Formulierung des Modells muss unter Berücksichtigung der Eigenschaften von Zeitreihen geschehen. Daher ist es trotz optischer Ähnlichkeit zu Darstellungen einfacher Datenmengen und der fundamentalen Datenanalyse in der Regel nicht möglich oder zumindest nicht sinnvoll, eine einfache lineare Regressionsanalyse auf die Daten durchzuführen und diese als Prognoseinstrument zu benutzen. Zeitreihen werden stattdessen entweder additiv oder multiplikativ in einzelne Komponenten zerlegt. Dazu gehören in der Regel mindestens die Trendkomponente, die langfristige Entwicklungen beschreibt, die saisonale Komponente, die wiederkehrende Muster, z. B. im Jahres- oder Monatsverlauf beschreibt und eine Restgröße oder Zufallskomponente, die den Teil der Datenveränderung aufnimmt, der durch die anderen Komponenten nicht beschrieben werden kann. Abb. 12.4 zeigt eine in dieser Art zerlegte Zeitreihe (Eheschließungen in Deutschland von 1990–2000, Datenquelle: Statistisches Bundesamt (2020)). Dargestellt sind die beobachtete Zeitreihe (oben), die Trend-Komponente (2. Reihe), die Saisonkomponente (3. Reihe) und die Zufallskomponente (unten).

Durch die Dekomposition deutlich zu erkennen ist der zunächst leicht abnehmende Trend und der tendenziell seit ca. 2015 erkennbare Wiederanstieg desselben. Die Auswertung der Saisonkomponente zeigt die erwartbar hohen Zahlen in den Sommermonaten sowie im Dezember. Deutlich wird, dass eine gründliche Dokumentation der Datenbasis essentiell für eine spätere Interpretation ist. Im vorliegenden Fall sind Eheschließungen notiert, die im Dezember aus steuerlichen Gründen ansteigen könnten, nicht Hochzeitsfeiern oder kirchliche Trauungen, die vermutlich weniger hohe Fallzahlen im Dezember aufweisen. Die Zufallskomponente dient insbesondere dazu, auffällig hohe oder niedrige Werte zu identifizieren, wie bspw. hier den August 2008. Auch in solchen Fällen liefert der Algorithmus keine Erklärung („Schnapszahltermin 08.08.08“), sondern markiert nur Auffälligkeiten.

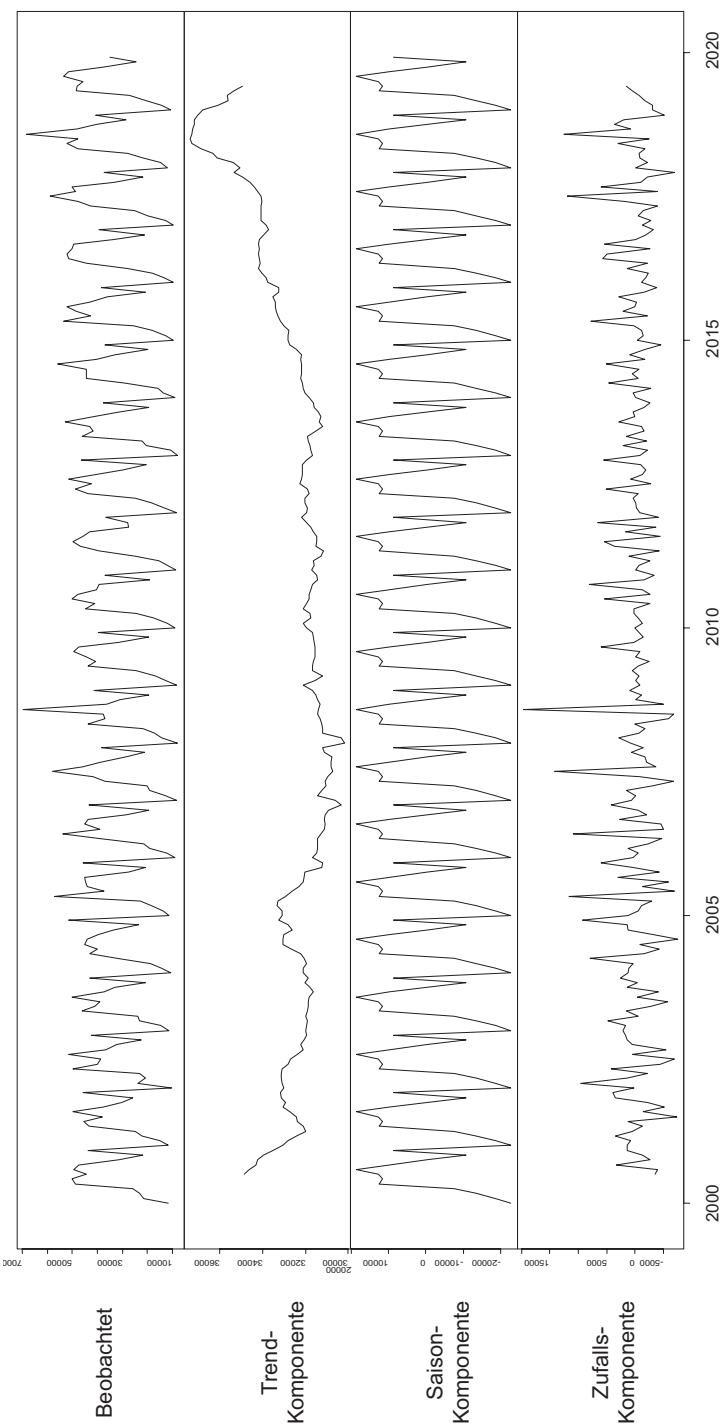


Abb. 12.4 Komponentendarstellung einer zerlegten Zeitreihe (Eheschließungen in Deutschland von 1990 bis 2019) (Datenquelle: Statistisches Bundesamt (Destatis), Genesis-Online, Datenlizenz by-2-0)

Auf Basis eines solchen Modells können jetzt zukünftige Werte geschätzt werden, wobei die Genauigkeit mit der zeitlichen Entfernung zum letzten bekannten Datensatz abnimmt, sodass für Schätzungen in der Regel Konfidenzintervalle angegeben werden, die ausdrücken, zwischen welchen Minimal- und Maximalwerten die Werte zukünftig mit einer Wahrscheinlichkeit von bspw. 95 % liegen werden. Auch hier kann die Prognosegüte ähnlich zu Klassifikationsmodellen durch Trainings- und Test-Daten geprüft werden. Ein weiteres Problem, was auch die bisher vorgestellten Modellvarianten betrifft, die in Teilen auf Regressionsmodellen beruhen, ist das Vorliegen von *Autokorrelation*. Dabei stehen frühere Werte der Zeitreihe mit späteren Werten in direkter Verbindung. So beträgt z. B. die Bevölkerung eines Landes in einem Jahr über 100 Mio. Personen, eben, weil sie zuvor auch schon über 100 Mio. Personen betrug. Hier ist keine Unabhängigkeit gegeben, die für die lineare Regression aber als Voraussetzung gefordert ist. Komplexere Modelle bieten für diese und weitere Besonderheiten der Zeitreihenanalyse umfangreiche Lösungsideen an, wie sie z. B. bei Groß (2010) beschrieben sind.

12.8 Text Mining

Texte sind Daten. Sie werden zwar im Vergleich zu den numerischen Werte, mit denen sich ein Großteil der Datenanalyseverfahren beschäftigt, anders repräsentiert, aber dennoch sind es Daten und als solche können sie ebenfalls analysiert werden. Mit Text Mining können unterschiedliche Bücher, Betriebsanweisungen oder Kundenrezensionen in Gruppen zusammengefasst werden (Clustering), neu hinzukommende Texte können in solche Gruppen eingesortiert werden (Klassifikation) und unabhängig davon können Aussagen über die Texte getroffen werden, die berücksichtigen, dass es sich in der Regel um natürlich-sprachliche Dokumente handelt, die mit einer bestimmten Intention oder einer bestimmten Geisteshaltung verfasst wurden (*Sentiment-Analyse*). Da Text-Mining sprachsensitiv ist, muss sichergestellt werden, dass für die Wörter, Regeln und Besonderheiten der jeweils zu analysierenden Sprache genug Vorinformationen bekannt sind. Für das Englische ist das in der Regel der Fall und auch Deutsch, Französisch oder Spanisch können mit vorgefertigten Informationskatalogen versehen werden. Schwieriger wird es bei Sprachen, die nur von wenigen Menschen gesprochen werden. Spracharten, die einem gänzlich anderen Aufbau und einer anderen Notation folgen (bspw. diverse asiatische Sprachen), benötigen auch angepasste Vorgehensweisen, weshalb sich dieser Abschnitt des Beispiels der deutschen Sprache bedient.

Dokumente, die analysiert werden sollen, müssen zunächst vorbereitet werden, d. h. alle nicht-inhaltlichen Elemente müssen herausgefiltert werden. Das können Seitenzahlen in Büchern, Metainformationen in Wikipedia-Artikeln, aber auch Werbetexte auf Webseiten sein. Als einfache Möglichkeit, den Text für eine Analyse zu strukturieren, bietet sich ein *bag-of-words*-Ansatz an. In einer Tabelle wird für jedes in mehreren zu analysierenden Dokumenten vorkommende Wort eine Spalte angelegt und jedes

Dokument erhält eine Zeile, in der dann pro Spalte notiert wird, wie häufig dieses Wort in dem jeweiligen Dokument vorkommt. Dabei müssen diverse Fehlerquellen eliminiert werden:

1. Groß- und Kleinschreibung müssen vereinheitlicht werden (z. B. wegen Satzanfängen)
2. Sonderzeichen müssen entfernt werden
3. Unterschiedliche Schreibweisen („Kommas“, „Kommata“) müssen berücksichtigt werden
4. Rechtschreibfehler müssen erkannt werden
5. Unterschiedliche Wortformen (Pluralnutzung, Konjugationen, Genitive, ...) müssen vereinheitlicht werden (*Stemming*)
6. Wörter ohne echten Analyse-Nutzen (*Stoppwörter*) müssen entfernt werden (z. B. „der“, „die“, „das“)
7. Feststehende Wort-Kombinationen (*N-Gramme*) sollten wie ein Wort behandelt werden (z. B. „zum Beispiel“)

Sind all diese Verarbeitungen erfolgt, kann die Ähnlichkeit zweier Einträge in der bag-of-words-Tabelle mithilfe der Kosinus-Ähnlichkeit bestimmt werden. Abb. 12.5 zeigt das zugrunde liegende Prinzip. Jedes Dokument wird durch einen Vektor im vieldimensionalen Raum repräsentiert, der Kosinus des Winkels zwischen den Vektoren gibt die Ähnlichkeit an.

Beide Dokumente im Beispiel bestehen nur aus zwei Wörtern („rot“ und „blau“, siehe bag-of-words-Tabelle (links)). Es ergeben sich zwei Vektoren im zweidimensionalen Raum. Der Kosinus des Winkels zwischen ihnen liegt zwischen 0 (bei 90°) und 1 (bei 0° , also bei Überlappung). Durch diese Kosinus-Ähnlichkeit lässt sich eine Distanzmatrix ermitteln, die im Folgenden z. B. von regulären Clustering-Verfahren genutzt werden kann, um die Dokumente in Gruppen einzuteilen.

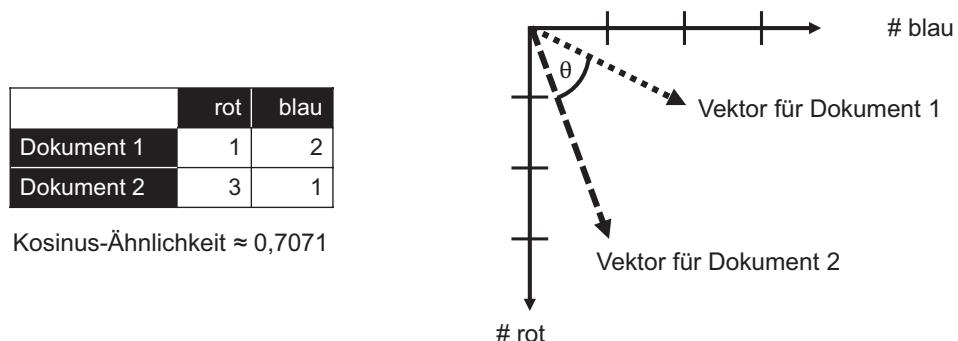


Abb. 12.5 Bag of Words und Darstellung der Berechnung der Kosinus-Ähnlichkeit für zwei Dokumente

Text Mining ist eine komplexe Analyseform, da neben den diversen Vorarbeiten auch viele Besonderheiten der Text-Betrachtung eine Rolle spielen. So ist nicht allein die Anzahl eines Wortes entscheidend, um ein Dokument zu charakterisieren, sondern z. B. auch, wie häufig dieses Wort überhaupt in allen betrachteten Dokumenten (dem *Korpus*) vorkommt. Seltene Worte haben mehr Gewicht, was sich über Kennzahlen wie die *Term Frequency – Inverse Document Frequency* (TFIDF) abbilden lässt (Provost und Fawcett 2017). Die Analyse des Sentiments, also der Stimmung eines Textes erfordert Kenntnisse über die Konnotation von Wörtern in einzelnen Sprachen, eine Zusammenhangsbetrachtung und muss zudem mit Analyse-hinderlichen Konzepten wie Ironie und Sarkasmus umgehen können. Obwohl Text Mining schon länger diskutiert wird, gewinnt es im Kontext von Big Data und aktuellen technologischen Entwicklungen, die auf Sprachunterstützung und/oder Sprachsteuerung beruhen, weiter an Aufmerksamkeit und wird ergänzt oder erweitert durch Konzepte wie z. B. *Natural Language Processing* (NLP), Speech Recognition und Language Understanding.

12.9 Weitere Analysemöglichkeiten

Der vorhergehende Abschnitt hat bereits mit einer Aufzählung von weitergehenden Analysekonzepthen geschlossen. Diese Liste lässt sich nahezu beliebig erweitern, denn sowohl die Forschungsfelder als auch die tatsächlich angewendeten Algorithmen sind vielfältig. Im Bereich des Maschinellen Lernens finden Konzepte Anwendung, die teilweise schon mehrere Jahrzehnte alt sind, aber im Zuge immer leistungsstärkerer Rechner und Netzwerke sowie umfassender Datenmengen wieder in den Fokus rücken. Hier nicht vorgestellt wurden bspw. *Support Vector Machines* zur Klassifikation. Darüber hinaus liegt ein Schwerpunkt in Forschung und Anwendung im Bereich der Datenanalyse in der Nutzung *Künstlicher Neuronaler Netzwerke*, deren Facettenreichtum und vergleichsweise komplexe Basisstruktur in eigenen Beiträgen vorgestellt werden. Letztlich obliegt es den mit einer Analyse betrauten Personen, immer kritisch zu prüfen, welche Daten überhaupt vorliegen, welche Methoden sinnvoll eingesetzt werden und vor allem, welche Fragestellung eigentlich beantwortet werden soll. Auch die fortschrittlichsten und modernsten Analyseverfahren können nicht „auf Daten losgelassen werden“ und im Alleingang für den jeweiligen Anwendungsfall nützliche Ergebnisse identifizieren und hervorbringen. Bei korrekter Anwendung aber liefern die hier vorgestellten fortgeschrittenen Methoden einen über die fundamentale Datenanalyse hinausgehenden Mehrwert für unternehmerische Entscheidungen und betriebliche Fragestellungen im Allgemeinen.

Literatur

- Backhaus, K., Erichson, B., Plinke, W., Weiber, R.: Multivariate Analysemethoden, 15. Aufl. Springer, Berlin (2018)
- Chicco, D.: Ten quick tips for machine learning in computational biology. In: BioData Mining Bd. 10, S. 35 (2017)
- Groß, J.: Grundlegende Statistik mit R: Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R. Vieweg+Teubner, Wiesbaden (2010)
- James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer, NY (2013)
- Mattson, M. P.: Superior pattern processing is the essence of the evolved human brain. In: Frontiers in neuroscience, Bd. 8, S. 265 (2014)
- Provost, F., Fawcett, T.: Data Science für Unternehmen: Data Mining und datenanalytisches Denken praktisch anwenden. MITP, Frechen (2017)
- Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Gröschel, A., Brauner, D., Badura, D., Passlick, J., Kloker, S., Gölzer, P., Kerzel, U., Rissler, R., Alekozai, E., Binder, H., Welter, F., Badewitz, W., Felderer, M., Rohde, H., Prothmann, M., Dann, D., Lanquillon, C., Gehrke, N.: DASC-PM v10 Ein Vorgehensmodell für Data-Science-Projekte. Elmshorn, Hamburg (2020)
- Statistisches Bundesamt (Destatis) Genesis-Online: *Eheschließungen: Bundesländer, Monate*, URL: <https://www-genesis.destatis.de/genesis//online?operation=table&code=12611-0011> (2020). Abruf am 31.07.2020. Lizenziert unter dl-de/by-2-0: www.govdata.de/dl-de/by-2-0
- Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data mining: practical machine learning tools and techniques, 4. Aufl. Morgan Kaufmann, Cambridge (2017)



Datenbasierte Algorithmen zur Unterstützung von Entscheidungen mittels künstlicher neuronaler Netze

13

Daniel Retkowitz

Zusammenfassung

Künstliche neuronale Netze sind eines der aktuell vielversprechendsten Gebiete im Bereich der Künstlichen Intelligenz und des maschinellen Lernens. Sie finden in immer mehr Bereichen Anwendung und versprechen damit, einen wesentlichen Beitrag im Zuge der Digitalisierung in Unternehmen zu leisten. Künstliche neuronale Netze sind dabei ein Ansatz, aus Beispieldaten zu „lernen“ und auf diese Weise maschinelle Entscheidungsfindung und Bewertung in unterschiedlichsten Bereichen zu ermöglichen. Die Bandbreite reicht von unterstützenden Systemen bis hin zu einer vollständigen Automatisierung. Ein wesentlicher Unterschied zu klassischen Softwarelösungen liegt darin, dass zuvor kein spezifisches, schrittweises Lösungsverfahren mit festen Regeln entwickelt werden muss. Künstliche neuronale Netze haben daher einen starken Einfluss auf die Art und Weise wie Software zukünftig entwickelt wird.

13.1 Datenbasierte Algorithmen und maschinelles Lernen

Ein Algorithmus wird üblicherweise als eine eindeutige Beschreibung von Bearbeitungsschritten eines maschinell ausführbaren Ablaufs zur Lösung eines spezifischen Problems verstanden (vgl. Saake und Sattler 2014). Algorithmen verarbeiten Eingabedaten und liefern als Ausgabe das gewünschte Ergebnis. Dazu muss im Allgemeinen zunächst Wissen über das Anwendungsgebiet explizit gemacht und in einen Algorithmus trans-

D. Retkowitz (✉)

FB 08, Hochschule Niederrhein, Mönchengladbach, Deutschland

E-Mail: daniel.retkowitz@hs-niederrhein.de

formiert werden, der anschließend in Form von Software implementiert und maschinell ausgeführt werden kann. Die Voraussetzung expliziten Wissens wird auch am Beispiel klassischer Expertensysteme deutlich. Diese dienen der Unterstützung von Entscheidungen und zeichnen sich dadurch aus, dass die Entscheidungsfindung auf zuvor festgelegten Regeln basiert. Diese stellen das Wissen der Experten eines Fachgebiets explizit und in einer maschinell verarbeitbaren Form dar, mit dem Ziel, dass das System auf Basis dieses Wissens ähnlich gute Ergebnisse wie die Experten selbst liefert. Solche Systeme heißen daher auch wissensbasierte Systeme. Durch immer mehr verfügbare Daten und gestiegene Rechenkapazitäten hat inzwischen jedoch ein anderes Vorgehen immer mehr Verbreitung gefunden, nämlich dass Algorithmen selbst aus großen Datensätzen automatisiert abgeleitet werden. Nicht das Wissen der Verarbeitung wird explizit gemacht, sondern lediglich das Lernvorgehen festgelegt. Aus Beispieldaten wird dann die Verarbeitungslogik automatisiert hergeleitet. Dieses Vorgehen wird maschinelles Lernen genannt.

13.1.1 Maschinelles Lernen

Maschinelles Lernen (ML) kann aus unterschiedlichen Perspektiven betrachtet werden. Im Kontext von Data Science ist ML eine von vielen verschiedenen Methoden, um Daten zu analysieren und Strukturen zu erkennen. Im Kontext der Künstlichen Intelligenz (KI) ist ML eines von vielen Teilgebieten, mit dem Ziel vormals dem Menschen vorbehaltene Fähigkeiten auf technische Systeme zu übertragen. Im Kontext des Software Engineering kann ML als ein alternatives Paradigma der Softwareentwicklung betrachtet werden. Anstatt imperativer Algorithmen werden Daten und Statistik eingesetzt. Die Problemstellung wird durch Daten und Bewertungen beschrieben. Mit statistischen, probabilistischen Methoden wird ML ermöglicht, sodass ein System in die Lage versetzt wird, neue, noch unbekannte Daten zu bewerten.

Abb. 13.1 verdeutlicht die Betrachtung von ML aus Perspektive der Softwareentwicklung. Auf der linken Seite ist das klassische imperative Entwicklungsparadigma dargestellt. Das Wissen über ein Fachgebiet muss zunächst explizit gemacht werden, wie im oberen Teil illustriert. Das ist die Basis von Algorithmen, die in Form eines Programms implementiert werden und anschließend zusammen mit Eingabedaten durch einen Computer zu den gewünschten Ergebnissen verarbeitet werden. Auf der rechten Seite ist die Vorgehensweise im Kontext von ML dargestellt. Hierbei wird der untere Teil aus der linken Hälfte auf die Ebene der Wissensexplikation gehoben, was mit einer „Vertauschung“ von Programm und Ergebnissen einhergeht. Das Programm, welches die Eingabedaten in die gewünschten Ergebnisse transformiert, ist nicht bekannt, sondern muss erst „erlernt“ werden. Dazu dienen Beispieldaten mit zugehörigen Ergebnissen aus denen ein ML-System (das Programm) erstellt wird. Dieses wird dann im unteren Teil eingesetzt, um auf konkrete Eingabedaten angewandt die Ergebnisse zu berechnen.

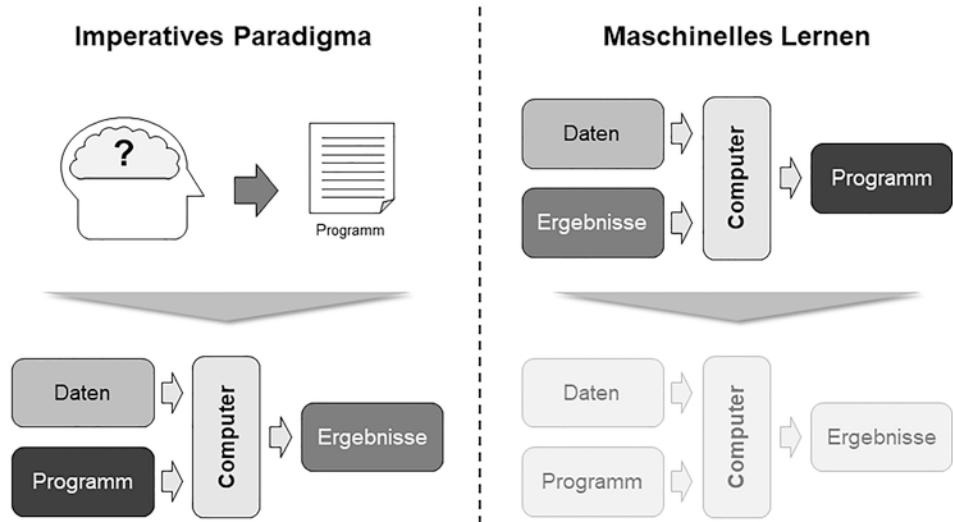


Abb. 13.1 Maschinelles Lernen in der Softwareentwicklung

13.1.2 Lernverfahren

Ein wesentlicher Teil bei der Entwicklung eines ML-Systems ist der Lernprozess. Dabei existieren unterschiedliche Vorgehensweise, die sich in drei Typen von Lernverfahren untergliedern lassen.

Überwachtes Lernen (engl. „*supervised learning*“) basiert auf Beispieldaten, die als vorgegebenes Expertenwissen im Lernprozess eingesetzt werden. Der Lernprozess erfordert somit einen „Lehrer“, den Experten, der Erfahrungswerte in Form von Beispieldaten mit korrekten Ergebnissen bereitstellt und dem anzulernenden System dadurch vermittelt, was richtig und was falsch ist. Das Expertenwissen wird auch als Grundwahrheit (engl. „*ground truth*“) bezeichnet.

Bestärkendes Lernen (engl. „*reinforcement learning*“) setzt keine Beispieldaten mit vorgegebenen Ergebnissen voraus. Welches Ergebnis „richtig“ oder „falsch“ ist, ist im Vorfeld nicht bekannt. Um das System anzulernen, erhält es Belohnungen und Bestrafungen für erwünschtes beziehungsweise unerwünschtes Verhalten. Das Lernen findet also nach dem Prinzip Trial-and-Error statt. Eine Herausforderung beim bestärkenden Lernen ist es, die Belohnungsfunktion geeignet zu definieren.

Unüberwachtes Lernen (engl. „*unsupervised learning*“) kann eingesetzt werden, um bisher unbekannte Strukturen in Daten zu erkennen. Diese Möglichkeit kann eingesetzt werden, wenn eine Bewertung in „richtig“ und „falsch“ oder eine Belohnung des Systems nicht möglich sind.

Zudem kann Transferlernen (engl. „transfer learning“) angewandt werden indem ein bereits angelerntes System auf ein anderes Anwendungsgebiet übertragen wird. Das zuvor Erlernte wird in dem Fall angepasst oder erweitert. Der Vorteil liegt darin, dass das System bereits „Vorwissen“ nutzen kann und der Lernprozess dadurch vereinfacht und beschleunigt werden kann.

13.2 Künstliche neuronale Netze

Eines der bedeutendsten Themenfelder im Zusammenhang mit der fortschreitenden Digitalisierung in Unternehmen und Gesellschaft ist der Bereich künstlicher neuronaler Netze (KNN) und des Deep Learning. In immer mehr Anwendungsgebieten wird ML unter Einsatz von KNN angewandt und es werden damit beachtliche Ergebnisse erreicht. KNN sind Rechenmodelle, die sich aus künstlichen Neuronen zusammensetzen, die in unterschiedlichen Topologien vernetzt sind. Ein künstliches Neuron ist ein Modell, das sich an Neuronen im Gehirn von Lebewesen als Vorbild orientiert und deren Funktionsweise simuliert. Durch die Vernetzung vieler solcher künstlichen Neuronen können auch sehr große Netze aufgebaut werden. Ein detaillierter Einblick in den Aufbau und das Lernverhalten von KNN findet sich in ► Kap. 14 dieses Buches.

Typische Aufgabenstellungen für KNN, insbesondere im Bereich des überwachten Lernens, sind die Klassifikation und die Regression. Bei der Klassifikation wird eine Eingabe einer von mehreren zuvor definierten Klassen zugeordnet. Bei der Regression gibt es anstatt vorgegebener Klassen eine kontinuierliche Zielgröße, für die eine möglichst gute Vorhersage berechnet wird.

13.2.1 Netzarchitekturen

KNN unterscheiden sich in ihrer Struktur und den Elementen, aus denen sie aufgebaut sind. Die typische Topologie eines KNN besteht aus miteinander verbundenen Schichten. Bei Architekturen mit vielen Schichten wird von Deep Learning gesprochen (vgl. LeCun et al. 2015).

Die einfachsten Topologien sind einschichtige und mehrschichtige Feedforward-Netze, auch Perceptron bzw. Multilayer Perceptron (MLP) genannt. Der Begriff Feedforward-Netz bedeutet, dass in solchen Netzen der Signalfluss nur in eine Richtung – nämlich vorwärts – erfolgt, das heißt es gibt keine Zyklen im Netz. Die erste Schicht, die die Eingabedaten aufnimmt, wird auch Input Layer genannt, die letzte Schicht mit den Ausgaben des Netzes wird Output Layer genannt. Alle Schichten dazwischen werden Hidden Layer genannt.

Eine Erweiterung der Feedforward-Netze sind Convolutional Neural Networks (CNN). CNN haben sich als ein Standardverfahren im Bereich der Computer Vision etabliert. Eine Herausforderung bei Bilddaten mit vielen Millionen Bildpunkten und

mehreren Farbkanälen liegt darin, dass die Anzahl der Dimensionen typischerweise zu hoch für ein klassisches Feedforward-Netz ist. Die Anzahl der benötigten Neuronen der Input Layer wäre dadurch zu groß. CNN ermöglichen eine Reduktion der Dimension der Eingabedaten indem zunächst Faltungsschichten, sogenannte Convolution Layers eingesetzt werden.

Die bisher genannten Topologien zeichnen sich dadurch aus, dass sie keine Zyklen enthalten. Dies ist anders bei Recurrent Neural Networks (RNN), auch rekurrente Netze genannt. Solche Netze werden verwendet, um Sequenzen von Daten zu verarbeiten, zum Beispiel bei der Verarbeitung von Texten und Sprache oder allgemein Audio- oder Video-Sequenzen. Zeitreihen kommen in vielen praktischen Problemstellungen vor, immer wenn nicht nur ein einzelner Zeitpunkt, sondern ein zeitlicher Verlauf betrachtet werden soll.

Ein weiterer Typ von Netzen sind Generative Adversarial Networks (GAN), auch gegnerische generative Netze genannt (vgl. Goodfellow et al. 2014). Sie kommen insbesondere beim unüberwachten Lernen zum Einsatz. Der Aufbau besteht aus zwei KNN, die im Trainingsprozess gegeneinander antreten. Das eine Netz wird Generator genannt und lernt Datensätze zu erzeugen, die den vorgegebenen Trainingsdaten möglichst ähnlich sind. Das andere Netz wird Diskriminatator genannt und bewertet die vom Generator erzeugten Datensätze. Dabei versucht der Diskriminatator, die Daten des Generators von den echten Trainingsdaten zu unterscheiden, während der Generator versucht solche Datensätze zu erzeugen, die vom Diskriminatator nicht mehr von den echten Trainingsdaten unterschieden werden können. Mit solchen GAN können zum Beispiel künstlich generierte Portraitfotos (vgl. Karras et al. 2018) erzeugt werden, die nur sehr schwer von Fotos echter Personen zu unterscheiden sind. Da auf diese Weise täuschend echte Ergebnisse erzeugt werden können, können diese auch für böswillige Zwecke oder kriminelle Handlungen eingesetzt werden. Dabei wird auch von „Deep Fakes“ gesprochen, also Fälschungen, die mit Techniken des Deep Learning erzeugt werden.

13.2.2 Grenzen künstlicher neuronaler Netze

Das Trainieren eines KNN basiert (beim überwachten Lernen) ganz wesentlich auf der Qualität der Trainingsdaten. Es kann nur gelernt werden, was an Informationen auch in den Daten enthalten ist. Daten sind im Allgemeinen mit verschiedensten Verzerrungen behaftet und spiegeln so auch Diskriminierungen und vorurteilsbehaftete menschliche Entscheidungen wider. Beim Trainieren eines KNN werden diese übernommen und sogar verstärkt. Dies führt zu diskriminierenden Entscheidungen, die auf Basis eines solchen Systems getroffen werden.

Gefahren bei der Entwicklung von KNN entstehen jedoch nicht nur durch Verzerrungen in den zugrunde liegenden Trainingsdaten. Durch die stark gestiegenen Möglichkeiten, die KNN tatsächlich bieten, eröffnet sich auch ein großer Spielraum an behaupteten Fähigkeiten, die sich bei genauerer Analyse als nicht haltbar herausstellen.

Es existiert beispielsweise eine Vielzahl von Unternehmen, die KI-basierte Software zur Automatisierung von Entscheidungen zur Einstellung von Bewerbern anbieten. Dabei zeigt sich jedoch, dass diese Systeme oft erhebliche Verzerrungen aufweisen und problematische Ergebnisse produzieren (vgl. Raghavan et al. 2019). Insbesondere können solche Verfahren auch direkt oder indirekt auf Korrelationen zurückgreifen, die aus rechtlichen oder ethischen Gründen auszuschließen sind.

Eine weitere Anforderung an KNN ist in vielen Fällen die Erklärung und Begründung von Ergebnissen. Ein System zur Entscheidungsfindung muss also in der Lage sein, zu begründen und zu belegen, wie es zu einem Ergebnis gekommen ist. In vielen Situation müssen auch maschinell getroffene Entscheidungen in jedem Einzelfall nachvollziehbar und begründbar sein. ML hat einerseits starke Verbesserungen in der Anwendbarkeit auf praktische Anwendungsfälle, in denen mit vielen Unschärfen umgegangen werden muss, ermöglicht. Andererseits sind die entstehenden statistischen Modelle oft Black-Box-Systeme, in deren Funktionsweise von außen kein Einblick möglich ist. Um Vertrauen in die Ergebnisse solcher Systeme zu erreichen und diese zu legitimieren, reicht der Nachweis einer „statistischen Wirksamkeit“ oft jedoch nicht aus.

KNN liefern von sich aus kein inhärentes Erklärungsmodell mit, da sie keinerlei inhaltliches Verständnis der jeweiligen Anwendungsdomäne haben und dieses auch nicht im Trainingsprozess erwerben. Insbesondere führt der Trainingsprozess nicht dazu, dass kausale Zusammenhänge identifiziert und im Modell abgebildet werden (vgl. Lake et al. 2017). In vielen Anwendungsfällen muss aber die Möglichkeit bestehen, eine fallbasierte Begründung zu erhalten und auch Entscheidungen anzweifeln oder anfechten zu können.

Das Forschungsgebiet der eXplainable Artificial Intelligence (XAI) beschäftigt sich mit der Frage, wie Systeme mit Erklärungsmodellen entwickelt werden können (vgl. Holzinger 2018). Dazu existieren verschiedene Ansätze, von denen viele so funktionieren, dass sie nachträglich zu einem bereits trainierten KNN ein zusätzliches Erklärungsmodell liefern, mithilfe dessen die Nachvollziehbarkeit und Erklärbarkeit der Ergebnisse erreicht werden soll. Dieser Ansatz wird daher auch Post-hoc-Erklärungsmodell genannt.

13.3 Beispielhafte Anwendungsfelder

Im Zuge der Digitalisierung gewinnt Software in immer mehr Bereichen an Bedeutung. Viele Produkte und Services sind schon heute ohne Software kaum noch denkbar. Zunehmend kommen als Bestandteil der Software auch ML und KNN zum Einsatz, beispielsweise in der Bilderkennung, Spracherkennung oder zur Entwicklung von Chatbots, um nur einige wenige Beispiele zu nennen.

Bildgebende Medizin

Der Bereich mit den vielleicht deutlichsten Fortschritten durch KNN und Deep Learning ist die Computer Vision. Anwendungsfälle sind die Klassifikation von Bildern ent-

sprechend ihrem Inhalt, die Erkennung und Identifikation einzelner Objekte in Bildern oder auch die Segmentierung von Objekten in Bildern, das heißt die genaue Abgrenzung von Objekten vom Hintergrund eines Bildes.

In der Medizin haben sich durch ML neue Möglichkeiten der Bildverarbeitung und der Analyse und Bewertung in bildgebenden Verfahren ergeben. Ein Beispiel ist das Mammographie-Screening zur Brustkrebsfrüherkennung. Die dabei entstehenden Röntgenaufnahmen müssen durch Ärzte ausgewertet werden, um Hinweise auf bösartige Tumoren zu finden. Diese Aufgabe ist herausfordernd, insbesondere da viele falsch positive Diagnosen zu unnötigen Folgeuntersuchungen und Behandlungen führen, die ihrerseits wiederum negative Auswirkungen haben. Für solche Routineuntersuchungen ist es daher erstrebenswert, eine sehr genaue Erkennung und Differenzierung der Fälle zu erreichen.

In verschiedenen Studien wird dazu untersucht, wie gut KNN bei der Auswertung von Mammographie-Aufnahmen im Vergleich zu menschlichen Analysen abschneiden (vgl. Wu et al. 2020). Darin wird gezeigt, dass die Leistung des Systems in etwa mit der von erfahrenen Radiologen vergleichbar ist. Dies ist eine beachtliche Leistung, wenn man bedenkt, dass in der Praxis nicht immer erfahrene Radiologen für die Beurteilung von Aufnahmen zur Verfügung stehen. Es zeigt sich außerdem, dass die besten Ergebnisse in einem hybriden Ansatz erreicht werden, das heißt Ärzte analysieren die Ergebnisse und stellen die Diagnose, werden dabei aber von einem System auf Basis eines KNN unterstützt. Dies ist ein Beispiel dafür, dass ML nicht in erster Linie Experten überflüssig macht, sondern zum Ziel hat, bei der Entscheidungsfindung zu unterstützen und so das Gesamtergebnis zu verbessern.

Chatbots und Konversationssysteme

Der Bereich der Computerlinguistik, im Englischen „Natural Language Processing“ (NLP) genannt, hat sich ebenfalls durch die Fortschritte im Bereich des ML deutlich weiterentwickelt. So kommen beispielsweise Chatbots auf Basis von KNN immer häufiger zum Einsatz. Chatbots sind virtuelle Assistenten, die mit einem Nutzer text- oder sprachbasiert kommunizieren. Zwar gab es bereits in den 1960er Jahren erste Chatbots, so ist jedoch erst heute durch die Entwicklungen im Bereich der KI ein Einsatz auch in komplexeren Anwendungsszenarien möglich. Digitale virtuelle Assistenten wie Apples Siri, der Google Assistant oder Amazon Alexa sind heute bereits im privaten Bereich weit verbreitet. Aufgrund der steigenden Leistungsfähigkeit werden digitale virtuelle Assistenten auch für den Einsatz in Unternehmen immer interessanter.

Sprachmodelle wie OpenAIs GPT-3 (vgl. Brown et al. 2020) erreichen auf Basis von Deep-Learning-Modellen sehr gute sprachliche Fähigkeiten und können Texte generieren, die in vielen Fällen kaum von denen menschlicher Autoren unterscheidbar sind. Es erscheint daher realistisch, dass in näherer Zukunft weitere Entwicklungsstufen in diesem Bereich zu erwarten sind. Allerdings sind bei den derzeitigen Ansätzen auch Limitationen dahingehend zu erkennen, dass die Sprachmodelle keinerlei Verständnis über den Inhalt von Texten erwerben. Sie sind somit zwar in der Lage, muster-

basiert sinnvolle, zusammenhängende Texte zu konstruieren, können aber nicht auf der inhaltlichen Ebene operieren. Werden einem solchen Modell beispielsweise einfache mathematische Aufgaben gestellt, so wird das Modell zwar erkennen, dass als Antwort zum Beispiel eine Zahl erwartet wird, es ist aber nicht fähig, die eigentlich nötige Berechnung durchzuführen und das korrekte Ergebnis zu liefern.

13.4 Entwicklungsprozess

Im Bereich der Data Science existieren bereits seit langem verschiedene Vorgehensmodelle, wie etwa im Data Mining der KDD-Prozess (Knowledge Discovery in Databases) (Fayyad et al. 1996) oder der Cross Industry Standard Process for Data Mining (CRISP-DM) (Wirth und Hipp 2000). Wenngleich diese Prozessmodelle aus dem Data Mining auch für die Entwicklung von KNN Anwendung finden, gehen sie doch nur begrenzt auf Besonderheiten ein und Fokussieren stärker auf die Aspekte der Datenselektion und -vorbereitung. Für die Entwicklung von KNN soll hier eine grobe Betrachtung ausreichen, die in Abb. 13.2 illustriert ist.

Der Prozess zur Entwicklung eines KNN lässt sich in mehrere Phasen einteilen, der Einfachheit halber werden hier nur vier Phasen unterschieden. In einer ersten Phase müssen Daten aufbereitet werden. Diese Phase kann in unterschiedliche Schritte unterteilt werden, auf die hier nicht näher eingegangen wird. Die im Trainingsprozess verwendeten Daten sind eine wesentliche Grundlage für die Qualität und Leistungsfähigkeit des späteren Systems. Häufig auftretende Probleme sind, dass zu wenige Datensätze vorliegen oder dass die Daten nicht repräsentativ sind, das heißt, dass einzelne Konstellationen unausgewogen repräsentiert sind und die Daten Verzerrungen beinhalten.

In der anschließenden Phase der Modellentwicklung muss zunächst ein initiales Modell erstellt werden, welches anschließend trainiert und optimiert werden kann. Dazu muss festgelegt werden, welches Lernverfahren zum Einsatz kommen soll und welche grundlegende Architektur das KNN haben soll. Im weiteren Verlauf der Entwicklung des KNN kann der Aufbau größeren Veränderungen unterliegen. Es gibt eine Vielzahl

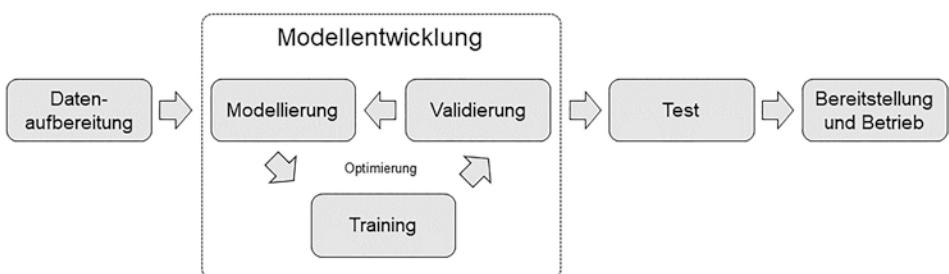


Abb. 13.2 Entwicklungsprozess für künstliche neuronale Netze

von Parametern, die im Rahmen der Entwicklung optimiert werden. Solche Parameter werden Hyperparameter genannt und beeinflussen den Lernprozess und auch den Aufbau sowie die Funktionsweise des KNN.

Um das Modell in der Entwicklung nach jeder Trainingsphase evaluieren zu können, wird der Datenbestand für den weiteren Prozess in Trainingsdaten und Testdaten aufgeteilt. Nur die Trainingsdaten werden für das Training genutzt, die Testdaten dienen hingegen ausschließlich dazu, das trainierte Modell zu evaluieren und festzustellen, wie gut das Modell mit „unbekannten“ Daten umgehen kann. Dies ist ein wichtiger Aspekt, der verhindern hilft, dass ein KNN beispielsweise die Testdaten einfach „auswendig“ lernt. Eine solche Zweiteilung berücksichtigt allerdings noch nicht, dass zur Entwicklung eines Modells neben dem eigentlichen Training auch eine Optimierung der Hyperparameter erfolgt. Daher ist es oft sinnvoll, die Daten in drei Teildatenbestände aufzuteilen, die Trainingsdaten, die Validierungsdaten und die Testdaten (vgl. Chicco 2017). Hierbei werden die Validierungsdaten nach jeder Trainingsphase zur Evaluation eingesetzt, die Testdaten hingegen kommen nur ein einziges Mal, nach abgeschlossener Entwicklung des Modells zum Einsatz. Damit wird sichergestellt, dass diese Daten nicht bereits in die Optimierung der Hyperparameter eingeflossen sind.

Wenn die Modellentwicklung abgeschlossen ist, folgt eine Testphase zur Qualitäts sicherung des Modells bevor dann der Übergang zur Bereitstellung und dem produktiven Betrieb erfolgen kann. Testverfahren für ML und KNN im Speziellen sind Gegenstand der laufenden Forschung. Durch den zunehmenden Praxiseinsatz von KNN entsteht ein steigender Bedarf nach einer umfassenden Qualitätssicherung. Ein ausschließlich „intuitives“ Testen im Rahmen der Entwicklung reicht nicht aus. Vielmehr werden systematische Testverfahren benötigt, wie sie auch in anderen Bereichen der Softwareentwicklung zum Einsatz kommen. Hierzu werden zunehmend neue Ansätze entwickelt, zum Beispiel Verfahren wie das Metamorphic Testing (vgl. Dwarakanath et al. 2018).

13.5 Entwicklungsplattformen und Werkzeuge

Beim Einsatz von KNN in der Praxis besteht eine Herausforderung darin, die Ansätze aus der Forschung möglichst breit und einfach anwendbar zu machen. Das bedeutet, dass die Anforderungen an das nötige Knowhow nicht zu hoch sein dürfen, denn dann können nur wenige spezialisierte Experten und Unternehmen, die ebendiese beschäftigen, mit ML und KNN arbeiten. Ziel ist es daher, möglichst vielen, die beispielsweise Vorerfahrung in der Softwareentwicklung mitbringen, den Zugang zu ML zu ermöglichen. Dies ist auch für Anbieter von Plattformen, Werkzeugen und Cloud-Diensten für ML ein wichtiges Ziel.

13.5.1 TensorFlow und PyTorch

Es gibt eine große Vielfalt an verschiedene Plattformen zur Entwicklung von KNN und zum Betrieb der damit entwickelten Lösungen (vgl. Nguyen et al. 2019). Eine der weitverbreitetsten Plattformen ist TensorFlow, aber auch PyTorch kommt immer häufiger zum Einsatz.

TensorFlow ist eine quelloffene Plattform für die Entwicklung von Software im Bereich des ML mit der Programmiersprache Python. TensorFlow wurde in einer Forschungsabteilung bei Google entwickelt, zunächst für den internen Gebrauch in unterschiedlichen Google-Diensten wie etwa der Übersetzung von Texten, später dann als eigenes Produkt, das auch für Entwickler außerhalb von Google zur Verfügung steht. TensorFlow wurde erstmalig 2015 veröffentlicht und ist spezialisiert auf effiziente numerische Berechnungen und unterstützt dabei auch den Einsatz spezieller Hardware wie Grafikprozessoren (GPUs). Solche numerischen Berechnungen sind wesentlich für das Trainieren von KNN, weshalb sich TensorFlow schnell in diesem Anwendungsfeld verbreitet und etabliert hat. Für die Entwicklung von KNN werden durch TensorFlow mehrere Programmierschnittstellen, sogenannte Application Programming Interfaces (APIs), auf unterschiedlichen Abstraktionsniveaus zur Verfügung gestellt. Je nach Einsatzzweck und der Notwendigkeit mehr oder weniger tief in die Mechanismen eines ML-Modells einzugreifen, kann zwischen maximaler Kontrolle über den Aufbau und die Funktionsweise des Modells und der Einfachheit der Implementierung abgewogen werden.

In TensorFlow stehen seit der Version 2.0 zwei verschiedene API-Levels zur Verfügung. Auf der höchsten Abstraktionsebene stehen die Estimator-API und Keras zur Verfügung. Diese APIs verbergen viele Details von TensorFlow, wodurch eine einfache und flexible Entwicklung von KNN ermöglicht wird. Für viele Standardfälle werden bereits fertige Komponenten angeboten, die parametrisiert und angewandt werden können.

Die Estimator-API bietet sogenannte Estimators an, die vollständige Modelle repräsentieren (vgl. Cheng et al. 2017). Diese können parametrisiert und besonders einfach auf verteilter Hardware trainiert, validiert und dann in produktive Zielumgebungen exportiert werden. TensorFlow liefert bereits eine Reihe von vorgefertigten Estimators mit. Darüber hinaus ist es auch möglich, individuelle Estimators selbst zu entwickeln. Keras ist als eigenständige API auch unabhängig von TensorFlow auf Basis verschiedener anderer Deep-Learning-Bibliotheken verwendbar. In TensorFlow 2.0 ist Keras aber auch als integraler Bestandteil und als High-Level-API in TensorFlow enthalten, was die Nutzung in TensorFlow erleichtert und Probleme durch inkompatible Versionen vermeidet. Keras bietet zwei Unter-APIs an, die Sequential-API und die Functional-API. Mit der Sequential-API erlaubt es Keras, Modelle sequenziell aus verschiedenen parametrisierbaren Schichten aufzubauen. Hierbei können unterschiedliche Arten von Schichten verkettet werden, um auf diese Weise individuelle Architekturen von KNN umzusetzen. Damit sind individuellere Lösungen als mit den Estimators möglich. Die Functional-API von Keras erlaubt eine noch größere Flexibilität. Damit sind

auch Modelle mit einer komplexeren Netzarchitektur umsetzbar, die zum Beispiel nicht aus einer klaren Abfolge einzelner Schichten aufgebaut sind, sondern Verzweigungen oder eine Mehrfachverwendung von Schichten erfordern.

Neben den High-Level-APIs ermöglicht TensorFlow über die Low-Level-API auch einen Zugriff auf seine elementaren Funktionen, die für den Aufbau und die Berechnung von KNN benötigt werden. Die Low-Level-API verkapselt damit die aus Effizienzgründen in C++ implementierten Rechenfunktionen und erlaubt es daraus beliebige Berechnungsgraphen zu erstellen und zu berechnen. Die Implementierung dieser Funktionen ist für verschiedene Plattformen verfügbar, um auf Standardprozessoren (CPUs) ausgeführt zu werden. Darüber hinaus ist mittels der Bibliothek CUDA des Herstellers Nvidia auch die Ausführung auf Grafikprozessoren (GPUs) möglich, wodurch die Berechnung deutlich beschleunigt werden kann.

PyTorch ist wie auch TensorFlow eine auf Python basierende, quelloffene Plattform für ML und wurde von Facebook entwickelt. PyTorch basiert auf Torch, einer älteren ebenfalls quelloffenen ML-Bibliothek, die bereits seit 2002 existiert, heute aber nicht länger weiterentwickelt wird. Während TensorFlow zunächst im Wesentlichen auf die Low-Level-API und einen statischen Berechnungsgraphen ausgerichtet war, lag der Fokus von PyTorch von Beginn an auf einer High-Level-API und einem imperativen Programmiermodell. Beide Plattformen haben sich im Laufe ihrer bisherigen Entwicklung aber einander angenähert und voneinander profitiert. Positive Aspekte der einen Plattform wurden dabei in die jeweils andere übernommen. Dies zeigt sich auch daran, dass sich der Python-Quellcode einer TensorFlow- und einer PyTorch-Lösung nicht gravierend unterscheidet und beide Plattformen eine umfassende Unterstützung für die Entwicklung und den Einsatz von ML bieten.

13.5.2 Ausführungsmodi

Sowohl TensorFlow als auch PyTorch unterstützen unterschiedliche Ausführungsmodi. Die beiden Plattformen haben dort in ihrer Historie unterschiedliche Ansätze verfolgt, sich aber schrittweise einander angeglichen. Zu unterscheiden sind die beiden Ausführungsmodi Lazy Execution und Eager Execution, die jeweils verschiedene Vor- und Nachteile haben, insbesondere hinsichtlich der Implementierung des Quellcodes und der Effizienz und Skalierbarkeit der Berechnungen.

Lazy Execution bedeutet, dass zu berechnende Ausdrücke zunächst in einem statischen Berechnungsgraphen, dem abstrakten Syntaxbaum, abgebildet werden. Die Ausdrücke werden erst später bei Bedarf ausgewertet. In TensorFlow war Lazy Execution anfangs der einzige verfügbare Ausführungsmodus. Ein TensorFlow-Programm baute zunächst den Berechnungsgraphen auf, der dann anschließend in einer sogenannten Session für eine konkrete Berechnung verwendet werden konnte (vgl. Planche und Andres 2019). Der Vorteil dieses Modells liegt darin, dass die Berechnung hinsichtlich der Effizienz

optimiert werden kann. So werden Berechnungen erst dann ausgeführt, wenn die Ergebnisse tatsächlich benötigt werden und es werden gleiche Berechnungen zur Optimierung in einem Cache zwischengespeichert und müssen nicht mehrfach berechnet werden. Der Berechnungsgraph kann auch auf eine parallele Berechnung hin optimiert werden, was für die effiziente Berechnung komplexer ML-Modelle sehr wichtig ist. Der Nachteil ist allerdings, dass die Entwicklung aufwendiger ist, da für jede Berechnung zunächst der Berechnungsgraph und zusätzlich der Code für die Ausführung in einer Session implementiert werden muss. Der Code wird dadurch komplexer und ist schwerer zu debuggen. Um eine einfachere Entwicklung zu ermöglichen, wurde ab TensorFlow 1.5 auch der Ausführungsmodus Eager Execution angeboten, der mit TensorFlow 2.0 zur Standardeinstellung wurde.

Eager Execution bedeutet, dass zu berechnende Ausdrücke sofort ausgewertet werden. Dies entspricht dem Standard in der imperativen Programmierung und ist auch in Python der Normalfall. Eager Execution ist in PyTorch und seit der Version 2.0 auch in TensorFlow die Standardeinstellung, was die Entwicklung von ML-Modellen vereinfacht und für viele Entwickler dem gewohnten Vorgehen entspricht. Damit trotzdem eine Optimierung und effiziente Berechnung auch auf verteilten Ressourcen möglich ist, bieten Plattformen wie TensorFlow und PyTorch entsprechende Mechanismen an. So kann in TensorFlow beispielsweise mittels sogenannter Python-Decorator die Bibliothek AutoGraph verwendet werden, die das automatische Generieren eines Berechnungsgraphen aus dem Python-Code übernimmt. Für die Berechnung von Gradienten im Trainingsprozess müssen Berechnungsfunktionen differenziert werden. Dies ist mit Eager Execution schwieriger, da kein Berechnungsgraph gegeben ist. In TensorFlow wird dies durch ein sogenanntes Gradient Tape gelöst, welches während der Berechnung automatisch die Gradienten aufzeichnet. In PyTorch wurde von Beginn an ebenfalls mit einem solchen Tape-Mechanismus gearbeitet, der in PyTorch Autograd heißt.

Durch die heute in TensorFlow und PyTorch zur Verfügung stehenden Mechanismen, ist die Eager Execution kein Nachteil für die spätere Nutzung eines KNN mehr und hat sich daher etabliert. Dies ist ein wichtiger Schritt auf dem Weg, die Entwicklung von KNN möglichst einfach zu gestalten, so dass auch Softwareentwickler in verschiedenen Bereichen außerhalb des ML einen schnellen und einfachen Zugang erreichen können. Durch die darüber hinaus verfügbaren Mechanismen zur Lazy Execution ist für beide Plattformen auch die effiziente Ausführung in verteilten Umgebungen und im produktiven Betrieb möglich.

13.5.3 Deployment und Betrieb

Neben der Entwicklung von KNN ist eine wesentliche Voraussetzung für den praktischen Einsatz die effiziente Ausführung im Betrieb. Dazu muss das entwickelte Modell in effizient ausführbaren Code übersetzt werden. Für die Bereitstellung und den

produktiven Betrieb eines fertig entwickelten Modells sind verschiedene Aspekte zu beachten, die während der Entwicklung zunächst nicht im Vordergrund stehen. Das KNN muss für den Betrieb in eine Zielumgebung deployt werden. Dabei ist zu unterscheiden, ob dies eine Serverumgebung in einem eigenen Rechenzentrum („on-premise“) ist oder ob dazu Cloud-Dienste genutzt werden sollen. Eine weitere Variante ist das Deployment auf sogenannten Edge Devices. Ein Edge Device ist ein Endgerät am „Rand“ eines Netzwerks, das hier als Zielumgebung verwendet wird. Das können Geräte (Controller, Sensoren etc.) im Internet of Things (IoT) sein, oder auch mobile Geräte wie Smartphones oder Tablets, auf denen das ML-System betrieben werden soll. Abhängig davon ergeben sich unterschiedliche Anforderungen.

TensorFlow unterstützt mit TensorFlow Serving den Betrieb von ML-Modellen in Produktivumgebungen. TensorFlow Serving ist ein Teil von TensorFlow Extended (TFX), das eine Betriebsplattform für TensorFlow-Modelle darstellt, um die Konfiguration von Modellen, das Training und den effizienten Betrieb sowie die Überwachung des Systems zu erleichtern. Damit wird angestrebt, Ansätze aus dem Software Engineering und der agilen Softwareentwicklung wie Continuous Integration und Continuous Delivery auch im Entwicklungsvorgehen von ML-Systemen einzusetzen (vgl. Baylor et al. 2019). TFX sieht sogenannte Continuous Data-Driven Pipelines vor, die dazu dienen, neue Daten für das Training einzulesen und aufzubereiten, den Trainingsprozess durchzuführen und zu steuern und trainierte Modelle anschließend in eine Produktionsumgebung zu deployen, in der TensorFlow Serving den Betrieb übernimmt. Neben dem Betrieb in dedizierten Serverumgebungen wird auch der Einsatz von TensorFlow im Webbrower mit TensorFlow.js und der Betrieb auf Mobilgeräten und Edge Devices mit TensorFlow Lite unterstützt. TensorFlow.js erlaubt es, ML-Modelle mit JavaScript zu entwickeln und auszuführen. Dies kann zum Beispiel direkt im Webbrower geschehen, so dass TensorFlow auch als Ergänzung zu bereits existierenden JavaScript-basierten Webanwendungen eingesetzt werden kann, ohne dass ein Wechsel der Plattform notwendig wird.

Auch für PyTorch steht eine Reihe von Werkzeugen zum Betrieb von ML-Modellen zur Verfügung. Mit der Version 1.5 ist der quelloffene Server TorchServe veröffentlicht worden, der als Produktivumgebung für PyTorch-Modelle eingesetzt werden kann und von Facebook in Zusammenarbeit mit Amazons Cloud-Computing-Tochter Amazon Web Services (AWS) entwickelt wurde. PyTorch adressiert genauso wie TensorFlow neben dedizierten Servern weitere Zielumgebungen wie Webbrower oder Mobilgeräte und Edge Devices. Mit PyTorch Mobile für Android und iOS können PyTorch-Modelle leichter in mobile Anwendungen integriert werden. Um PyTorch-Modelle in effizient ausführbaren Code zu transformieren wurde TorchScript eingeführt, das eine Repräsentation in Form eines statischen Berechnungsgraphen ermöglicht. Auch wenn PyTorch noch nicht solange wie TensorFlow den Betrieb von ML-Systemen in den Fokus genommen hat, ist auch hier ein Angleichen der beiden großen Plattformen sichtbar.

Die nach Marktanteil größten Cloud-Computing-Anbieter Amazon, Microsoft und Google bieten allesamt eine ganze Reihe von ML-Services an. Da ML ein aktuelles

Thema mit hohem Wachstumspotenzial ist, versuchen die Cloud-Anbieter hier möglichst viele Belange zu bedienen und den Zugang zu ML möglichst leicht zu machen. Dies ist ein wesentlicher Aspekt, da die wenigsten Unternehmen eigene Ressourcen im Bereich der Forschung und Entwicklung von ML-Systemen haben oder die Hardware-Ressourcen für die Entwicklung und den späteren Produktivbetrieb besitzen. Die typischen Anwendungsfälle, die im Programm der Cloud-Anbieter zu finden sind, umfassen Empfehlungen, Prognosen, Bildanalyse, Textanalyse, Dokumentanalyse, Text to Speech (TTS), Conversational Agents, Übersetzungen, Spracherkennung oder Betrugserkennung. Als Marktführer im Cloud-Geschäft bietet allein Amazon mit AWS Machine Learning weit über 20 verschiedene ML-Produkte an. Diese werden ständig um neue Anwendungsfälle erweitert.

13.6 Fazit und Ausblick

Der Einsatz von KNN eröffnet eine Vielzahl neuer Anwendungsmöglichkeiten und einen komplementären Ansatz der Softwareentwicklung. Im Zuge der fortschreitenden Digitalisierung in Unternehmen bieten sich dadurch neue Möglichkeiten, Menschen bei Entscheidungen und Bewertungen zu unterstützen. Eine Automatisierung kann so teilweise oder sogar vollständig erreicht werden. Die technischen Möglichkeiten werden durch leistungsfähige Plattformen und High-Level-APIs immer leichter zugänglich. Dadurch ergibt sich auch eine größere Verbreitung in der Unternehmenspraxis. Durch Entwicklungswerzeuge und Services der Cloud-Computing-Anbieter wird der Einsatz weiter vereinfacht. Insbesondere die gezielte und bedarfsgerechte Nutzung von spezieller GPU-Rechenleistung für das Training und den Betrieb ist ein Vorteil des Cloud Computing im Zusammenhang mit ML.

Es ist davon auszugehen, dass sich diese Trends in Zukunft weiter fortsetzen werden. Dabei liegt eine weitere Herausforderung in der Interdisziplinarität. Sowohl in der Forschung als auch in der praxisorientierten Anwendung gilt es daher zukünftig in so unterschiedlichen Disziplinen wie der Informatik, der Mathematik, der Psychologie, den Neurowissenschaften, den Rechtswissenschaften oder der Philosophie noch stärker zusammenzuarbeiten. Für viele Szenarien ist auch ein Zusammenwirken der Systeme selbst erforderlich. Um in der Lage zu sein, ein autonomes Fahrzeug zu steuern, wird beispielsweise nicht nur Objekterkennung, Fahrspurerkennung oder die Auswertung von Sensordaten jeweils einzeln benötigt, sondern in Kombination miteinander, damit sich die Steuerung des Fahrzeugs weiter automatisieren lässt.

Die Entwicklung hin zu einer einfachen Nutzung von ML birgt allerdings auch die Gefahr, dass es zu unreflektierten, fehlerhaften und ethisch problematischen Anwendungen kommt. KNN lassen sich prinzipiell auf Grundlage beliebiger Daten trainieren. Ob die Ergebnisse dann übertragbar sind und ein solches KNN anwendbar ist, ist eine andere Frage. Durch Verzerrungen in den verwendeten Trainingsdaten kann ein unerwünschtes Verhalten des entwickelten Systems resultieren. Dabei kommt es nicht

nur auf eine sehr genaue Auswahl der Trainingsdaten an, sondern auch auf die Netzarchitektur und die verwendeten Hyperparameter. Dazu ist weiterhin eine entsprechende Kenntnis der Wirkmechanismen des ML und der Data Science notwendig. Die Ethik ist in der KI ein zunehmend wichtiges Thema, das in Entwicklungsprojekte einfließen muss. Dies zeigt sich auch daran, dass zunehmend Leitlinien und Standards in diesem Bereich entwickelt werden (vgl. IEEE 2019).

Literatur

- Baylor, D., Haas, K., Katsiapis, K., Leong, S., Liu, R., Menwald, C. et al.: Continuous Training for Production ML in the TensorFlow Extended (TFX) Platform. In: Proceedings of the 2019 USENIX Conference on Operational Machine Learning (OpML '19). Berkeley, CA, USA: The USENIX Association. S. 51–53 (2019)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language Models are Few-Shot Learners. (2020) [https://arxiv.org/pdf/2005.14165](https://arxiv.org/pdf/2005.14165.pdf)
- Cheng, H.-T., Haque, Z., Hong, L., Ispir, M., Mewald, C., Polosukhin, I., et al.: TensorFlow Estimators: Managing Simplicity vs. Flexibility in High-Level Machine Learning Frameworks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). New York, NY, USA: Association for Computing Machinery (ACM). S. 1763–1771 (2017). <https://doi.org/10.1145/3097983.3098171>
- Chicco, D.: Ten quick tips for machine learning in computational biology. *BioData Mining* **10**(1), 35 (2017). London, UK: BioMed Central (BMC). <https://doi.org/10.1186/s13040-017-0155-3>
- Dwarkanath, A., Ahuja, M., Sikand, S., Rao, R. M., Bose, R. J. C., Dubash, N., Podder, S.: Identifying implementation bugs in machine learning based image classifiers using metamorphic testing. In: ISSTA'18. Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis. New York, NY, USA: Association for Computing Machinery (ACM). S. 118–128 (2018)
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **17**(3), 37 (1996). Palo Alto, CA, USA: Association for the Advancement of Artificial Intelligence (AAAI). <https://doi.org/10.1609/aimag.v17i3.1230>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative Adversarial Nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). Cambridge, MA, USA: MIT Press. S. 2672–2680 (2014)
- Holzinger, A.: Explainable AI (ex-AI). *Informatik Spektrum* **41**(2), Heidelberg: Springer. S. 138–143 (2018). <https://doi.org/10.1007/s00287-018-1102-5>
- IEEE: Ethically Aligned Design. Prioritizing Human Wellbeing with Autonomous and Intelligent Systems. New York, NY, USA: Institute of Electrical and Electronics Engineers (IEEE). (2019)
- Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, S. 4396–4405, New York, NY, USA: Institute of Electrical and Electronics Engineers (IEEE). (2019) <https://doi.org/10.1109/CVPR.2019.00453>
- Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. *The Behavioral and brain sciences* **40**, e253 (2017). Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/S0140525X16001837>

- LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015). London, UK: Nature Research. <https://doi.org/10.1038/nature14539>
- Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., et al.: Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artif Intell Rev* **52**(1), 77–124 (2019). Berlin: Springer Nature. <https://doi.org/10.1007/s10462-018-09679-z>
- Planche, B., Andres, E.: Hands-on computer vision with TensorFlow 2. Leverage deep learning to create powerful image processing apps with TensorFlow 2.0 and Keras. Birmingham, UK: Packt Publishing. (2019)
- Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). S. 469–481. New York, NY, USA: Association for Computing Machinery (ACM). (2020)
- Saake, G., Sattler, K.-U.: Algorithmen und Datenstrukturen. Eine Einführung mit Java. 5., überar. Aufl. dpunkt.verlag, Heidelberg (2014)
- Wirth, R., Hipp, J.: CRISP-DM: Towards a Standard Process Model for Data Mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Blackpool, Lancashire, UK: Practical Application Company. S. 29–39 (2000)
- Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., et al.: Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans. Med. Imaging* **39**(4). New York, NY, USA: Institute of Electrical and Electronics Engineers (IEEE). S. 1184–1194 (2020). <https://doi.org/10.1109/TMI.2019.2945514>



Künstliche Neuronale Netze – Aufbau, Funktion und Nutzen

14

Anja Tetzner, Tom Kühne, Peter Gluchowski und Melanie Pfoh

Zusammenfassung

Niemals zuvor wurden derart gewaltige Datenmengen produziert wie in jüngster Zeit. Daraus erwächst die Erwartung, dass sich in den Daten interessante Informationen finden lassen, wenn es nur gelingt, dieses hohe Datenvolumen zielgerichtet auszuwerten. Sowohl in der Wissenschaft als auch zunehmend in der Praxis werden daher Verfahren und Technologien diskutiert, die interessante Muster in umfangreichen Datenbeständen aufdecken und Prognosen über zukünftige Ereignisse und Gegebenheiten ermöglichen.

Zunehmende Bedeutung erlangt in diesem Kontext die Beschäftigung mit künstlichen Neuronen und den zugehörigen künstlichen neuronalen Netzen, welche auf eine insgesamt fast 80-jährige Entwicklung zurückblickt. Zwar hat in dieser Zeit eine stetige Weiterentwicklung der Konzepte stattgefunden, allerdings greifen bis heute die ursprünglich zugrunde gelegten basalen Annahmen und Sichtweisen. Inzwischen finden künstliche neuronale Netze Anwendung für unterschiedlichste Aufgabenstellungen in einem weiten Spektrum ökonomischer und nicht-ökonomischer Einsatzbereiche.

A. Tetzner · T. Kühne · P. Gluchowski (✉) · M. Pfoh

Fakultät für Wirtschaftswissenschaften, Technische Universität Chemnitz, Chemnitz,
Deutschland

E-Mail: peter.gluchowski@wirtschaft.tu-chemnitz.de

A. Tetzner

E-Mail: anja.tetzner@wirtschaft.tu-chemnitz.de

T. Kühne

E-Mail: tom.kuehne@wirtschaft.tu-chemnitz.de

M. Pfoh

E-Mail: melanie.pfoh@wirtschaft.tu-chemnitz.de

Der vorliegende Beitrag beschreibt den Aufbau sowie die Funktionsweise von künstlichen neuronalen Netzen und greift dabei auch neuere Entwicklungen auf. Aus einer betriebswirtschaftlichen Perspektive erweisen sich vor allem die Nutzenpotenziale als relevant, sodass auch hierauf eingegangen wird.

14.1 Einleitung

Künstliche neuronale Netze (KNN, engl. artificial neural networks) gehören zu den gegenwärtig am meisten diskutierten Themen im Bereich Data Science. Ihre Anwendungsmöglichkeiten reichen vom Einsatz im Rahmen von Industrie 4.0, bspw. zur Predictive Maintenance, über die Unterstützung der Diagnostik in der Medizin, bspw. bei der Erkennung von Krebszellen, bis hin zur Verwendung im Konsumentenbereich sowohl aufseiten der Unternehmen durch umfassende Analysen von Kundendaten als auch auf Seiten der Kunden, bspw. in Form von persönlichen Assistenten wie Alexa, Siri und Co.

Als Nachbildung natürlicher menschlicher und tierischer neuronaler Netze ermöglichen künstliche neuronale Netze Maschinen das Lösen komplexer Aufgabenstellungen, welche umfassende kognitive Fähigkeiten erfordern. Sie stellen dadurch einen wesentlichen Bestandteil des maschinellen Lernens (ML, engl. machine learning) und dem damit eng verbundenen Bereich der künstlichen Intelligenz (KI, engl. artificial intelligence) dar (vgl. Patterson und Gibson 2017, S. 4). Eine besondere Bedeutung im Zusammenhang mit dem Einsatz künstlicher neuronaler Netze besitzt das Teilgebiet des Deep Learning, wobei hierunter künstliche neuronale Netze subsummiert sind, welche sich durch komplexe Topologien mit einer Vielzahl von Neuronen und Schichten auszeichnen (vgl. Patterson und Gibson 2017, S. 81). Deep-Learning-Netze besitzen darüber hinaus die Fähigkeit, mittels einfacher, nichtlinearer Funktionen automatisiert aus den gegebenen Rohdaten Features extrahieren zu können. Diese Features (auch als Attribute oder Merkmale bezeichnet) enthalten die zur Lösung der gegebenen Aufgabenstellung notwendigen und zumeist auf das Wesentliche komprimierten Informationen aus den Rohdaten und dienen den Netzen als Eingabedaten (vgl. LeCun et al. 2015, S. 436). Deep-Learning-Netze kommen aufgrund ihrer hohen Problemlösungskompetenz vor allem bei der Verarbeitung komplexer Daten mit hoher Dimensionalität, wie z. B. Bild-, Audio- und Videodaten, zur Anwendung (vgl. Ertel 2016, S. 299).

Der nachfolgende zweite Abschnitt widmet sich dem grundlegenden Aufbau von Neuronen und künstlichen neuronalen Netzen. Anschließend erörtert der dritte Abschnitt, welche unterschiedlichen Verfahren beim Lernen in künstlichen neuronalen Netzen zur Anwendung gelangen. Schließlich beleuchtet der vierte Abschnitt die verschiedenen Einsatzfelder, bevor ein Fazit den Beitrag beschließt.

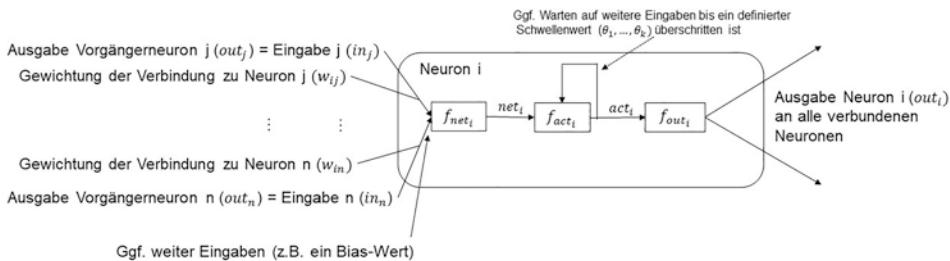


Abb. 14.1 Künstliches Neuron

14.2 Aufbau

Künstliche neuronale Netze bestehen im Kern aus zwei elementaren Bestandteilen: den Neuronen und den Verbindungen zwischen den Neuronen. Auch wenn mit der Entwicklung im Bereich künstlicher neuronaler Netze mittlerweile eine Reihe unterschiedlicher und teilweise sehr komplexer Neuronen und Verbindungsvariationen zwischen diesen entstanden sind, lassen sich die unterschiedlichen Topologien auf einen grundlegenden Aufbau zurückführen.

14.2.1 Künstliches Neuron

Erstmals vorgestellt wurde ein künstliches Neuron 1943 durch McCULLOCH UND PITTS, welche bereits damals zeigten, dass ein künstliches Neuron grundsätzlich dazu in der Lage ist, universelle Berechnungen durchzuführen (vgl. McCulloch und Pitts 1943).

Die nachfolgende Abb. 14.1 zeigt die schematische Darstellung des grundlegenden Aufbaus eines künstlichen Neurons.

Ein künstliches Neuron i ist mit einem oder mehreren Vorgängerneuronen (j, \dots, n) über eingehende Verbindungen verknüpft. Die Ausgaben der Vorgängerneuronen (out_j, \dots, out_n) dienen dem Neuron i gleichzeitig als Eingaben (in_j, \dots, in_n). Die Netzeingabefunktion f_{net_i} nutzt diese Eingaben unter Berücksichtigung einer Gewichtung (w_{ij}, \dots, w_{in})¹, um die Netzeingabe net_i des Neurons i zu berechnen. Neben den Gewichten und den Eingabewerten der Vorgängerneuronen können weitere Eingabewerte in die Berechnung der Netzeingabe einfließen, bspw. ein Bias-Wert, welcher eine

¹ Gewichte dürfen nur den Kanten zugeordnet werden, nicht den Neuronen selbst. Daraus ergibt sich, dass Eingabeneuronen (vgl. Abschn. 14.2.2) ihre Netzeingabe nicht über die Multiplikation der übergebenen Eingabewerte mit den entsprechenden Verbindungsgewichten errechnen, sondern, dass die Netzeingabe von Eingabeneuronen den externen Eingabewerten in ein künstliches neuronales Netz entsprechen (vgl. Kruse et al. 2011, S. 46)

konstante Eingabe in das Neuron generiert. Die Aktivierungsfunktion f_{act_i} nutzt die Netzeingabe, um die Aktivierung act_i des Neurons zu berechnen. Hierbei können Schwellenwerte (θ_i) eine Rolle spielen. Wird ein Schwellenwert nicht überschritten, wartet das Neuron auf weitere Eingabesignale der verbundenen Neuronen und ruft die Netzeingabe- und Aktivierungsfunktion wiederholt bis zum Überschreiten des Schwellenwertes auf. Erst dann erfolgt die Übergabe der Aktivierung act_i an die Ausgabefunktion f_{out_i} . Anhand der übergebenen Aktivierung berechnet die Ausgabefunktion f_{out_i} die Ausgabe des Neurons. Das Neuron übergibt die errechnete Ausgabe out_i an alle verbundenen Neuronen, wodurch sich der beschriebene Ablauf in diesen Neuronen wiederholt. Formal lässt sich die Ausgabe eines künstlichen Neurons folgendermaßen zusammenfassen:

$$out_i = f_{act}(\sum_j^n w_{ij}in_j + b_i) \quad (14.1)$$

wobei b_i einen optionalen Bias-Wert darstellt.

Für die Aktivierungsfunktion künstlicher neuronaler Netze existieren verschiedene Varianten. Der Grundgedanke, welcher bereits durch McCULLOCH UND PITTS (1943) zur Anwendung kam, beruht auf der Nutzung einer Schwellenwertfunktion. Die Schwellenwertfunktion nimmt so lange Eingaben vorgelagerter Neuronen entgegen, bis ein zuvor definierter Schwellenwert erreicht wird. Erst wenn dieser Schwellenwert überschritten ist, leitet das Neuron sein Ausgabesignal an alle nachgelagerten Neuronen weiter. Wird der Schwellenwert nicht überschritten, bleibt das Neuron inaktiv. Eine weitere einfache Form der Aktivierungsfunktion eines künstlichen Neurons ist die Identitäts- bzw. lineare Funktion. Sie sorgt für eine lineare Abbildung der Eingabe eines Neurons auf seine Ausgabe, das bedeutet, ein Neuron wird desto stärker aktiviert je größer die Eingabe ist. Für das effektive Lernen komplexer Aufgabenstellungen ist die Identitätsfunktion allerdings weniger geeignet, aufgrund dessen kommen im Bereich künstlicher neuronaler Netze vordergründig nicht-lineare Aktivierungsfunktionen zum Einsatz. Zu den am weitesten verbreiteten zählen die Sigmoidfunktion, der hyperbolische Tangens (Tanh-Funktion) und die Rectifier-Funktion (ReLU-Funktion). Die Sigmoidfunktion normalisiert alle Eingaben eines Neurons in einem Wertebereich zwischen 0 und 1, vgl. Abb. 14.2. Das bedeutet, die Ausgabe eines Neurons mit einer sigmoiden Aktivierungsfunktion erzeugt einen Ausgabewert, der zwischen 0 und 1 liegt. Im Gegensatz zur Sigmoidfunktion ist die Ausgabe der Tanh-Funktion nullzentriert. Das bedeutet, dass sich der Wertebereich der Ausgabe zwischen -1 und +1 bewegt. Die ReLU-Funktion liefert für alle negativen Eingaben den Wert 0 zurück, für positive Werte entspricht der Ausgabewert der Eingabe in das Neuron.

Neben den genannten Funktionen gewinnt auch die Softmax-Aktivierungsfunktion zunehmend an Bedeutung. Sie berechnet die Wahrscheinlichkeit der Zugehörigkeit eines gegebenen Wertes zu einer Reihe vorgegebener Klassen. Die Klasse, für welche die Softmax-Funktion die höchste Zugehörigkeitswahrscheinlichkeit berechnet, wird als die finale Klasse für einen zu klassifizierenden Eingabewert angenommen. Die Softmax-Aktivierungsfunktion kommt daher vordergründig in Neuronen der Ausgabeschicht (vgl. Abschn. 14.2.2) zum Einsatz.

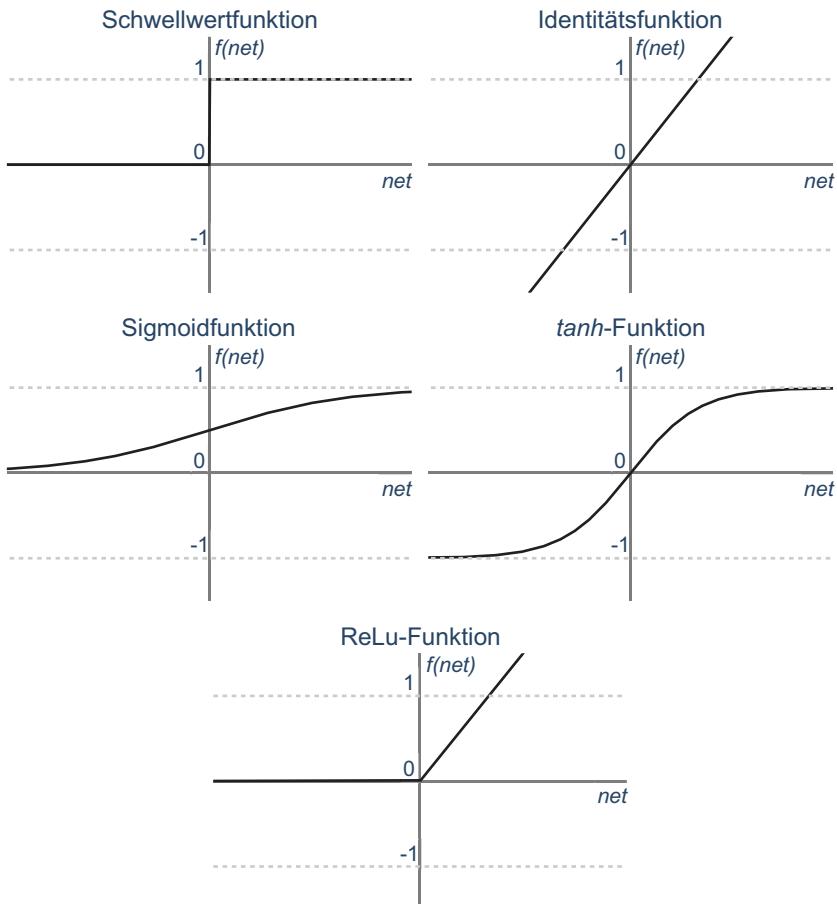


Abb. 14.2 Übersicht Aktivierungsfunktionen

14.2.2 Künstliche neuronale Netze

Im Allgemeinen ist unter einem künstlichen neuronalen Netz ein mathematisches Modell zu verstehen, welches einen gerichteten Graphen, bestehend aus Knoten (Neuronen) und Kanten (den gewichteten Verbindungen zwischen den Neuronen), beinhaltet. Diese Definition bildet die Grundlage für die sogenannte Schichten-Architektur künstlicher neuronaler Netze, nach welcher die Neuronen in unterschiedlichen Schichten organisiert sind (vgl. hierzu auch Abb. 14.3). Neuronen in der Eingabeschicht (engl. input layer) nehmen die Eingabe von Werten von außen in das Netz entgegen. Die Ausgabe der durch ein Netz berechneten Ergebnisse erfolgt über die Neuronen der Ausgabeschicht (engl. output layer). Zwischen Eingabe- und Ausgabeschicht kann eine unterschiedliche Anzahl sogenannter versteckter Schichten (engl. hidden layer) existieren. Versteckte Schichten

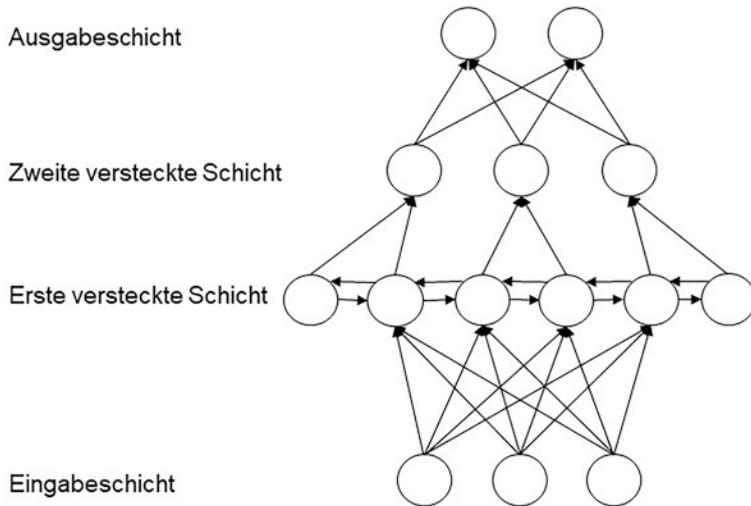


Abb. 14.3 Exemplarischer Aufbau eines künstlichen neuronalen Netzes

besitzen keine Verbindung zur Umgebung des künstlichen neuronalen Netzes. Sie sind damit von außen nicht beeinflussbar.

Zwischen den unterschiedlichen Schichten bzw. den Neuronen in diesen Schichten können verschiedenartige Verbindungen vorkommen. Künstliche neuronale Netze, deren Verbindungen lediglich in eine Richtung von der Eingabeschicht hin zur Ausgabeschicht gerichtet sind und keine Schleifen oder Rückverbindungen zu vorgelagerten Schichten oder zu vorangehenden Neuronen der gleichen Schicht besitzen, gehören zur Klasse der vorwärtsgerichteten künstlichen neuronalen Netze. Künstliche neuronale Netze, welche Schleifen oder Zyklen in den Verbindungen aufweisen bzw. Verbindungen einer Schicht zu einer vorgelagerten Schicht besitzen, zählen zur Klasse der rückgekoppelten künstlichen neuronalen Netze. Eine vollständige Verbindung der Neuronen einer Schicht mit denen einer nachfolgenden Schicht ist möglich, jedoch nicht zwingend erforderlich (vgl. z. B. die im Folgenden beschriebenen Convolutional Neural Networks).

Abb. 14.3 zeigt einen möglichen schematischen Aufbau eines künstlichen neuronalen Netzes mit unterschiedlich gerichteten Verbindungen zwischen den Neuronen.

Eine der grundlegendsten Arten künstlicher neuronaler Netze sind mehrschichtige Perzeptren (MLP entsprechend dem engl. multilayer perceptrons). Mehrschichtige Perzeptren besitzen eine Eingabe- und eine Ausgabeschicht, dazwischen können beliebig viele Schichten verdeckter Neuronen existieren. Verbindungen sind nur zwischen direkt benachbarten Schichten möglich (vgl. Kruse et al. 2011, S. 44). Jede Schicht ist mit ihren benachbarten Schichten vollständig verbunden, das bedeutet, ein Neuron einer Schicht ist mit allen Neuronen der direkt benachbarten Schichten verbunden, ohne dass Zyklen in den Verbindungen bestehen (vgl. Patterson und Gibson 2017, S. 41). Dementsprechend zählen mehrschichtige Perzeptren zu den vorwärtsgerichteten künstlichen neuronalen Netzen.

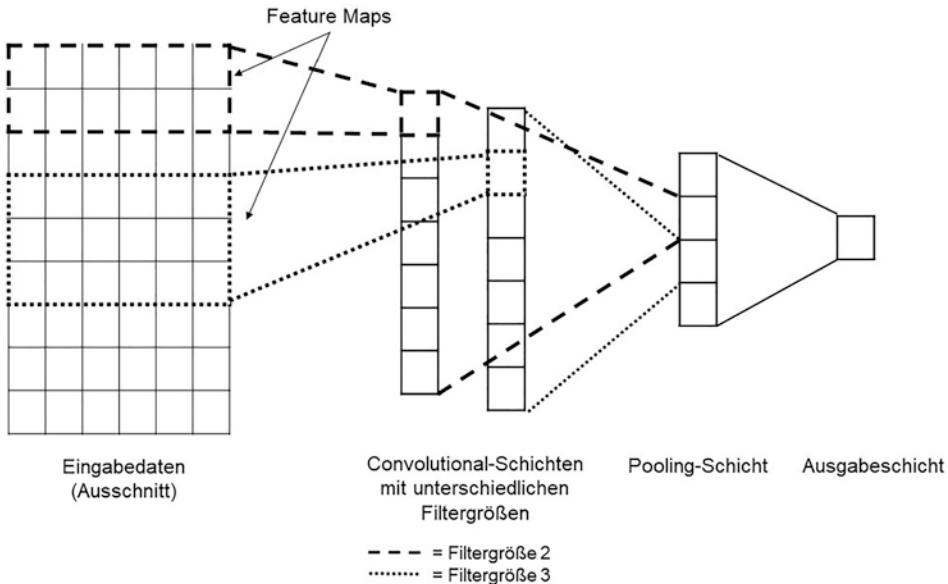


Abb. 14.4 Schematischer Aufbau eines CNN

Weitere Vertreter vorwärtsgerichteter künstlicher neuronaler Netze sind Convolutional Neural Networks (CNN). Sie zeichnen sich durch die Kombination zweier spezifischer Arten von Schichten aus, die Convolutional- oder auch Faltungsschichten (engl. convolutional layer) und die Pooling- bzw. Verbindungsschichten (engl. pooling layer). Abb. 14.4 zeigt einen auszugsweisen schematischen Aufbau eines CNN.

Vereinfacht wenden Convolutional-Schichten einen Filter einer definierten Größe auf einen Ausschnitt der Eingabedaten bzw. der Daten einer vorgelagerten Schicht an und verarbeiten die Informationen innerhalb dieses Filterausschnittes, welcher auch als Feature Map bezeichnet wird. Das Netz extrahiert auf diesem Weg für den Filterausschnitt Informationen auf einem höheren Level als der vorgelagerten Schicht und lernt damit wichtige Informationen zur Lösung der gegebenen Problemstellung. Die Pooling-Schichten verdichten die Informationen eines Filterausschnitts zur Verstärkung des Lerneffektes und erzielen gleichzeitig eine Reduktion der Anzahl der Parameter innerhalb des Netzes, da der Filterausschnitt bzw. das Ergebnis der Untersuchung des Filterausschnitts um einen definierten Wert verkleinert wird (vgl. LeCun et al. 2015, S. 439). CNN kommen vor allem im Bereich der Bildverarbeitung zum Einsatz, eignen sich aber z. B. auch für die Verarbeitung von Textdaten.

Selbstorganisierende Karten (engl. self-organizing maps, kurz SOM) stellen einen weiteren Typ vorwärts gerichteter künstlicher neuronaler Netze dar (vgl. Kohonen 1997, S. 69). Wie in Abb. 14.5 ersichtlich, bestehen sie lediglich aus einer Eingabe- und einer Ausgabeschicht und verfügen über keine weiteren, versteckten Schichten. Selbstorganisierende

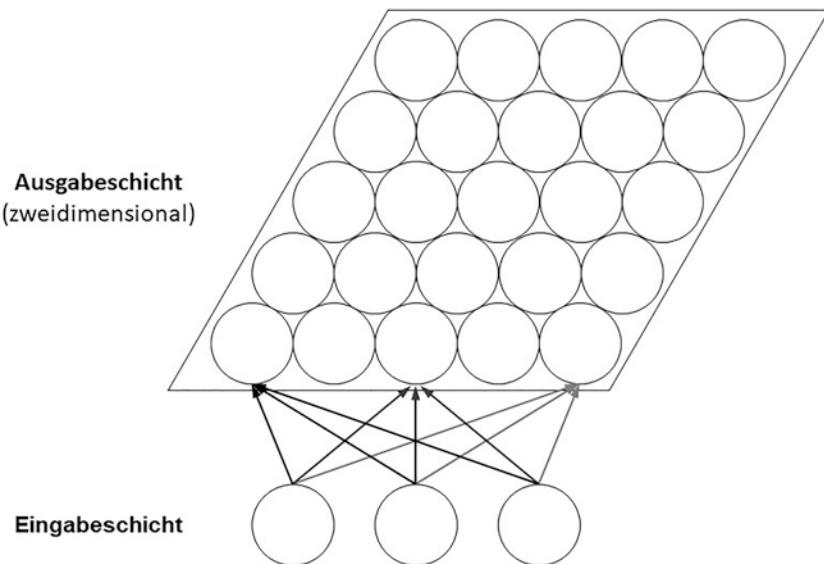


Abb. 14.5 Schematischer Aufbau einer Selbstorganisierenden Karte

Karten zeichnen sich dadurch aus, dass die Neuronen der Ausgabeschicht in einem elastischen Netz angeordnet sind. Die dadurch definierten Nachbarschaftsbeziehungen erlauben die Ausnutzung von Selbstorganisationsmechanismen basierend auf einer speziellen Form des Wettbewerbslernens, auf das in Abschn. 14.3.2 näher eingegangen wird. Das ermöglicht eine Projektion der Daten aus einem n -dimensionalen Eingaberaum auf einen m -dimensionalen Ausgaberaum unter Beibehaltung der Ähnlichkeitsbeziehungen. Daher können selbstorganisierende Karten im Rahmen der Dimensionsreduktion (vgl. bspw. Saraswati et al. 2018; Aghajari und Chandrashekhar 2017; Lin et al. 2000) eingesetzt werden, um hochdimensionale Daten beispielsweise in einer zweidimensionalen Fläche abzubilden. Darüber hinaus finden sie aber auch beim Clustering von Daten Anwendung (vgl. bspw. D’Urso et al. 2020; Isa et al. 2009; Powers und He 2008).

Recurrent Neural Networks (RNN) als Vertreter rückgekoppelter neuronaler Netze eignen sich besonders für die Verarbeitung sequentieller Daten bzw. für die Verarbeitung von Daten, welche eine Zeitabhängigkeit aufweisen. Sie sind damit speziell für die Verarbeitung von Zeitserien, Textdaten sowie Audiodaten geeignet (vgl. Patterson und Gibson 2017, S. 144). RNN weisen rückgekoppelte Verbindungen zwischen den Neuronen einer Schicht auf, die es den Neuronen ermöglichen, Informationen aus einem vorangegangenen oder ggf. auch zukünftigen Sequenzabschnitt in die Verarbeitung des aktuellen Abschnittes einfließen zu lassen. Abb. 14.6 zeigt den vereinfachten Aufbau eines RNN.

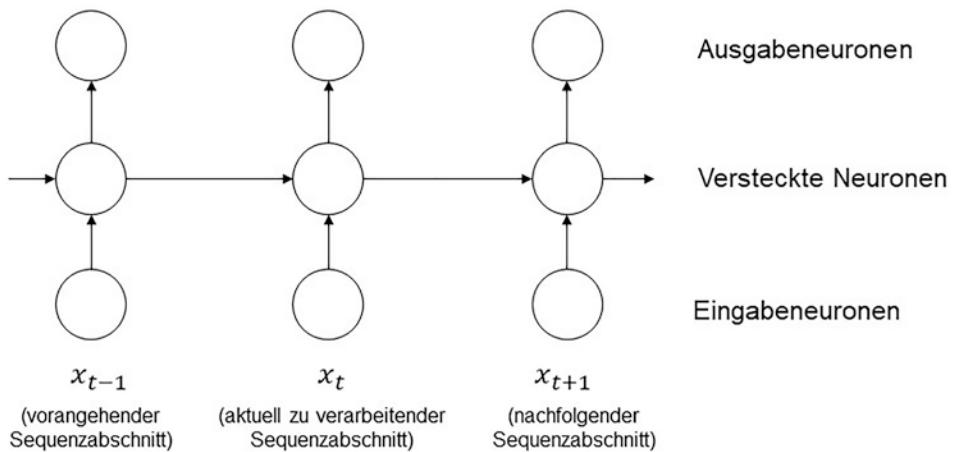


Abb. 14.6 Schematischer Aufbau eines RNN

Ein Neuron einer Schicht leitet seine Ausgabe nicht nur an Neuronen der nachfolgenden Schicht weiter, sondern gleichzeitig auch an nachgelagerte bzw. ggf. vorgelagerte Neuronen der gleichen Schicht. Auf diese Weise ist auch eine bidirektionale Verarbeitung von Datensequenzen möglich, vom Beginn bis zum Ende der Datensequenz und umgekehrt. Als Erweiterung von RNN haben sich Long-Short-Term-Memory-Netze (LSTM) (vgl. Hochreiter und Schmidhuber 1997) und Gated Recurrent Units (GRU) (vgl. Chung, Gulcehre et al. 2014) entwickelt, welche den Informationsfluss innerhalb des Netzes durch spezielle sogenannte Gating-Mechanismen steuern. Die Gating-Mechanismen bestimmen, wie viel der Informationen aus verbundenen Neuronen übernommen und wie viel der im Neuron selbst berechneten Information an alle nachfolgenden Neuronen weitergegeben werden.

14.3 Lernen künstlicher neuronaler Netze

Das Ziel eines künstlichen neuronalen Netzes besteht darin, durch eine Anpassung der Verbindungsgewichte eine nichtlineare Abbildung der Eingabedaten auf die Ausgabedaten zu schaffen und darüber die Lösung einer komplexen Aufgabenstellung zu erlernen. Ein Optimum in diesem Zusammenhang würde bedeuten, dass eine vollständig korrekte Abbildung einer Ausgabe auf eine entsprechende Eingabe erreicht wird und somit (immer) eine fehlerfreie Lösung der gegebenen Aufgabenstellung vorliegt, was in der Praxis allerdings nur schwer zu erreichen ist. Stattdessen wird in der Regel eine näherungsweise Lösung angestrebt (vgl. Kruse et al. 2011, S. 40). Zur Berechnung dieser näherungsweisen Lösung kommen spezifische Lern- bzw. Optimierungsalgorithmen, alternativ auch als Lernverfahren bezeichnet, zum Einsatz. Das Lernen

künstlicher neuronaler Netze kann grundsätzlich in zwei Arten unterschieden werden, das überwachte sowie das unüberwachte Lernen, welche jeweils spezifische Lernalgorithmen nutzen. Welche Art des Lernens zum Einsatz kommt, ist unmittelbar von der gegebenen Aufgabenstellung sowie dem verfügbaren Datenbestand abhängig.

14.3.1 Überwachtes Lernen – Lernen mittels Backpropagation

Das überwachte Lernen kommt vordergründig bei Klassifikations- bzw. Regressionsaufgaben zum Einsatz. Voraussetzung zur Anwendung des überwachten Lernens ist eine Information zur Einordnung der im Datenbestand enthaltenen Datensätze, bspw. in Form von Labels. Labels geben dabei z. B. in kodierter numerischer Form (0, 1, 2, ...) die Klassenzugehörigkeit eines Datensatzes an und dienen dazu, die durch das künstliche neuronale Netz bestimmte Klassenzugehörigkeit mit der tatsächlichen zu vergleichen. Abweichung zwischen der bestimmten und der tatsächlichen Klassenzugehörigkeit bilden die Grundlage für das Lernen des künstlichen neuronalen Netzes.

Die verschiedenen spezifischen Lernalgorithmen im Bereich des überwachten Lernens künstlicher neuronaler Netze beruhen im Kern auf dem Backpropagation-Algorithmus. Dieser durchläuft drei grundsätzliche Schritte zur Anpassung der Verbindungsgewichte und damit verbunden dem Lernen des künstlichen neuronalen Netzes. Zunächst sind die Gewichte des künstlichen neuronalen Netzes durch Zufallswerte zu initialisieren, die Trainingsdatensätze in das Netz einzugeben und die Ausgabe des Netzes zu berechnen (Schritt 1). Die berechnete Ausgabe dient im Vergleich mit der erwarteten Ausgabe anschließend der Ermittlung des Gesamtfilters des Netzes auf Basis einer vorgegebenen Fehlerfunktion (Schritt 2). Der Gesamtfehler wird danach anhand der Berechnung des Gradientenabstiegs zurück durch das Netz propagiert und so auf alle beitragenden Gewichte verteilt, dass der Gesamtfehler minimal ist (vgl. Gardner und Dorling 1998, S. 2629; Patterson und Gibson 2017). Die beschriebenen Schritte sind so lang zu wiederholen, bis das Netz einen zufriedenstellend kleinen Gesamtfehler erreicht (vgl. Gardner und Dorling 1998, S. 2629). Der Backpropagation-Algorithmus liefert allerdings keine Garantie für die Identifikation eines globalen Fehlerminimums und damit dem Erreichen eines Optimums der Anpassung der Verbindungsgewichte (vgl. Patterson und Gibson 2017, S. 58; Rumelhart et al. 1986, S. 535).

Es existieren zwei grundsätzliche Varianten der Anwendung des Backpropagation-Algorithmus: das Online-Lernen und das Batch-Lernen (vgl. Rumelhart et al. 1986, S. 535). Das Online-Lernen entspricht dem ursprünglichen Gedanken des Backpropagation-Algorithmus und passt die Verbindungsgewichte nach jedem Datensatz im Trainingsdatenbestand an. Durch häufige Aktualisierung der Verbindungsgewichte neigt diese Variante allerdings zu einer hohen Fluktuation, was dazu führen kann, dass das globale Fehlerminimum nur schwer zu erreichen ist. Das Batch-Lernen im Gegensatz nimmt eine Anpassung der Verbindungsgewichte aggregiert nach dem Durchlaufen aller im Trainingsdatenbestand enthaltenen Datensätze vor. Der Gradient ist dementsprechend

für den gesamten Trainingsdatenbestand zu berechnen, was zu langen Rechenzeiten führen kann. Das Mini-Batch-Lernen kombiniert beide Varianten, in dem die Anpassung der Gewichte nach n Trainingsdatensätzen² erfolgt (vgl. Ruder 2016, S. 2). Über die Zeit haben sich unterschiedliche Variationen des ursprünglichen Backpropagation-Algorithmus entwickelt, die auf unterschiedliche Art und Weise versuchen, die Schwächen des ursprünglichen Algorithmus, allem voran das zu schnelle Konvergieren in einem lokalen Fehleroptimum oder auch das Überspringen des globalen Fehleroptimums, zu vermeiden. Zu den am weitesten verbreiteten Varianten des Backpropagation-Algorithmus gehören Adadelta, Adagrad, Adam, Nadam und insbesondere RMSprop (vgl. Ruder 2016).

14.3.2 Unüberwachtes Lernen – Lernen mittels Wettbewerbslernen

Im Gegensatz zum überwachten Lernen benötigt das unüberwachte Lernen keine weiteren Informationen zur Einordnung der Datensätze des Datenbestandes. Label, die die Klassenzugehörigkeit eines bestimmten Datensatzes wiedergeben, werden also nicht benötigt. Das Lernergebnis wird somit maßgeblich durch die inhärenten Eigenschaften und Strukturen der bereitgestellten Daten beeinflusst (vgl. Mallot und Hübner 2014, S. 380). Unüberwachte Lernverfahren kommen daher vor allem bei der Datenanalyse zum Einsatz, bspw. beim Clustering. Sie können aber auch für eine unüberwachte Klassifikation genutzt werden.

Grundlage vieler unüberwachter Lernalgorithmen für künstliche neuronale Netze bildet die Hebb'sche Lernregel. Im Gegensatz zum bereits betrachteten Backpropagation-Algorithmus kann die Anpassung der Verbindungsgewichte dabei jedoch nicht auf einer Fehlerfunktion basieren, da kein Label vorliegt, mit dem die berechnete Ausgabe verglichen werden kann. Stattdessen erfolgt die Anpassung der Verbindungsgewichte in Abhängigkeit von der Aktivierung der einzelnen Ausgabeneuronen (vgl. Hassoun 1995, S. 90). D. h. die Verbindungsgewichte zu Neuronen, die durch einen Datensatz eine hohe Aktivierung erfahren, werden stärker angepasst. Somit kommt es zu einem sich selbstverstärkenden Effekt und das neuronale Netz lernt. Auf diesem Prinzip baut auch das Wettbewerbslernen auf. Nach der Aktivierung der Neuronen durch einen Datensatz wird das Neuron ausgewählt, das die stärkste Aktivierung aufweist (das sog. „Gewinner“-Neuron). Nur die Verbindungsgewichte dieses Neurons erfahren eine Anpassung (vgl. Mallot und Hübner 2014, S. 384).

Aufbauend auf diesen grundlegenden Lernregeln gibt es viele weitere Varianten des unüberwachten Lernens für unterschiedliche Typen von künstlichen neuronalen Netzen. Auch der Lernalgorithmus der bereits vorgestellten selbstorganisierenden Karten basiert auf dem Wettbewerbslernen (vgl. Kohonen 1997, S. 86 f.). Bei Eingabe eines Datensatzes

²Das Durchlaufen einer Sequenz von n Trainingsdatensätzen entspricht dabei einem Batch.

wird das Neuron der Ausgabeschicht bestimmt, dessen Verbindungsgewichte die geringste Distanz zu den Eingabedaten aufweisen. Dieses Neuron stellt das „Gewinner“-Neuron dar und seine zugehörigen Verbindungsgewichte werden angepasst. Zusätzlich werden aber auch die Verbindungsgewichte benachbarter Ausgabeneuronen verändert. Dabei wird die Anpassung der Verbindungsgewichte mit steigender Entfernung zum „Gewinner“-Neuron abgeschwächt.

14.4 Nutzenpotenziale und Herausforderungen

Der Einsatz künstlicher neuronaler Netze ist gleichermaßen mit einer Reihe von Potenzialen wie auch Herausforderungen verbunden.

Komplexe künstliche neuronale Netze, insbesondere aus dem Bereich des Deep Learning, sind in der Lage, die zur Analyse von Daten notwendigen Features automatisiert aus den Rohdaten zu lernen. Dies führt zu einer teilweise erheblichen Verkürzung der benötigten Zeit der Datenvorverarbeitung insbesondere im Hinblick auf die Extraktion informativer Features, die als Eingabedaten in das Netz dienen. Besonders in Hinblick auf unstrukturierte Daten, wie bspw. Text-, Bild- oder Audiodaten, sind Deep-Learning-Netze in der Lage, effektiv geeignete Features zur Lösung komplexer Aufgabenstellungen zu lernen, wie z. B. die Analyse von Meinungen in Texten (Sentiment-Analyse) (vgl. bspw. Lai et al. 2015; Santos und Gatti 2014; Tang et al. 2015) oder die Erkennung von Objekten oder Gesichtern in Bildern (vgl. bspw. Krizhevsky et al. 2012; Simonyan und Zisserman, 2015; Szegedy et al. 2015). Durch die automatisierte Feature-Extraktion erfolgt eine Entlastung des Datenanalysten, welcher sich dadurch verstärkt auf die Gestaltung einer zur Lösung der Aufgabenstellung geeigneten Netztopologie konzentrieren kann. Gleichzeitig sind die Netze in der Lage, umfangreiche Datenbestände zu verarbeiten. In Verbindung mit dem stetigen Anstieg der Datenflut ergeben sich damit kontinuierlich neue Anwendungsfelder für den Einsatz künstlicher neuronaler Netze. Teilweise ermöglicht auch erst der Einsatz künstlicher neuronaler Netze die Analyse komplexer Datenbestände, die durch andere Verfahren oder auch den Menschen nicht verarbeitet werden können. Darüber hinausgehend ist eine Verbreitung des Einsatzes kombinierter komplexer Netze zur Verarbeitung multimodaler Datenbestände zu erkennen. Als Beispiel hierfür ist die automatisierte Generierung von Bildbeschreibungen (caption generation) zu nennen (vgl. bspw. Bai und An 2018; Karpathy und Fei-Fei 2017; Mao et al. 2014). Künstliche neuronale Netze aus diesem Bereich kombinieren CNN mit vorwiegend RNN aber auch MLP, um eine gleichzeitige Verarbeitung von Bild- und Textdaten zu ermöglichen. Die CNN-Komponente trainiert mit dem Ziel, Bilddaten hinsichtlich identifizierbarer Objekte, unterschiedlicher Bildebenen, Beziehungen zwischen Bildobjekten etc. erkennen zu können. Die RNN- bzw. MLP-Komponente verarbeiten parallel textuelle Beschreibungen zu den gegebenen Bilddaten und versucht, Zusammenhänge zwischen einzelnen Bildelementen und Textabschnitten zu lernen. Bei Eingabe eines unbekannten Bildes in das trainierte Netz soll es diesem

anhand der gelernten Zusammenhänge möglich sein, automatisiert die Bildbestandteile zu erkennen und diese textuell zu beschreiben.

Zur Verbesserung der Lernergebnisse eines künstlichen neuronalen Netzes wird der Trainingsdatenbestand mehrfach durchlaufen. Es besteht dabei allerdings die Gefahr eines „Auswendiglernens“ der Trainingsdatensätze durch das Netz, dem sogenannten Überanpassen (engl. overfitting). Die Folge einer Überanpassung ist eine schlechte Generalisierbarkeit des Netzes. Als Generalisierbarkeit ist dabei die Fähigkeit eines künstlichen neuronalen Netzes zu verstehen, für unbekannte (Test-)Datensätze gute Ergebnisse zu erzielen. Eine schlechte Generalisierbarkeit bedeutet dementsprechend, dass die Berechnung von Eingabe- zu Ausgabewerten für Datensätze, die das Netz nicht im Rahmen des Trainingsprozesses zur Verfügung gestellt bekommen hat, fehlerhafte bzw. schlechte Ergebnisse liefert (vgl. Svozil et al. 1997, S. 47). Zu wenige Trainingsdurchläufe oder auch zu wenige Neuronen und damit eine schlechte Repräsentation der Eingabe- auf die Ausgabedaten können dagegen zu einer Unteranpassung (engl. underfitting) führen, bei welcher eine Minimierung des Gesamtfehlers und damit die Lösung der gegebenen Aufgabenstellung nur unzureichend oder gar nicht erfolgt.

Die Gestaltung der Topologie eines künstlichen neuronalen Netzes sowie die Wahl der sogenannten Hyperparameter sind stets problemspezifisch. Zu den Hyperparametern zählen all jene Parameter eines künstlichen neuronalen Netzes, welche von außen durch den Entwickler vorgegeben werden. Klassische Beispiele für die Hyperparameter sind die Anzahl der Neuronen und Schichten, die Aktivierungsfunktion, die Wahl des Lernalgorithmus, die Lernrate, die Funktion zur Bestimmung des Fehlers eines Netzes, die Batch-Größe und die Anzahl der Trainingsdurchläufe (Epochen). Je nach konkreter Gestaltung eines Netzes ist die Festlegung weiterer Hyperparameter notwendig, z. B. die Filtergröße für die Convolutional-Schichten eines CNN. Es existiert nicht das „eine“ künstliche neuronale Netz, welches in der Lage ist, alle denkbaren Problemstellungen gleichermaßen zufriedenstellend zu lösen. Die Wahl einer geeigneten Topologie sowie passender Hyperparameter ist dabei eine komplexe Aufgabe, die meist nur durch das Austesten verschiedener Gestaltungsvarianten zu lösen ist, da gegenwärtig effektive Verfahren zur gezielten problemspezifischen Bestimmung fehlen. Der Einsatz von Verfahren aus Wissenschaftsdisziplinen, welche sich der Lösung von Optimierungsproblemen abseits des Einsatzes künstlicher neuronaler Netze widmen zur Bestimmung optimaler Hyperparameterausprägungen ist möglich³, bietet aber ebenfalls keine zweifelsfreie Sicherheit hinsichtlich einer problemlösungsoptimalen Netzgestaltung. Erschwerend kommt hinzu, dass sich die Hyperparameter gegenseitig stark beeinflussen und daher stets auch die Kombinationen verschiedener Hyperparameterausprägungen zu berücksichtigen sind.

Die Lösung spezifischer Aufgabenstellungen erfordert entsprechende Datenbestände. Nur anhand eines repräsentativen Datenbestandes, der den Umfang der Problemstellung

³Denkbar sind an dieser Stelle bspw. genetische Algorithmen.

adäquat abbildet, ist es einem künstlichen neuronalen Netz möglich, effektiv eine Lösung für eine gegebene Aufgabenstellung zu lernen. Gerade im Bereich der Klassifikationsaufgaben ist es jedoch teilweise mit einem erheblichen Aufwand verbunden, derartige Datenbestände bereitzustellen, da die einzelnen Datensätze des Datenbestandes klassifiziert und mit einem Label versehen werden müssen. Dies ist in einer hohen Qualität meist nur durch einen erheblichen manuellen Aufwand möglich.

14.5 Fazit

Künstliche neuronale Netze zählen zu einer der gegenwärtig vielversprechendsten Technologien im Bereich der Analyse großer und insbesondere komplexer, unstrukturierter Datenbestände. Gleichzeitig führt die Verfügbarkeit von immer mehr Daten aus den unterschiedlichsten Interessensgebieten angetrieben durch Themen wie Big Data und der digitalen Transformation sowie die stetigen Weiterentwicklungen auf dem Gebiet der künstlichen neuronalen Netze zu einer ständigen Ausweitung der Anwendungsbereiche bei gleichzeitiger Verbesserung der Problemlösungsqualität. Die Netze sind immer besser in der Lage, komplexe Problemstellungen unterschiedlicher Anwendungsbereiche zufriedenstellend und zuverlässig zu lösen. Bereits jetzt zählen künstliche neuronale Netze zu ständigen Begleitern des Alltags, bspw. in Form von digitalen persönlichen Assistenten, Recommendation Engines in Online Shops sowie Customer Interaction Center in Form von Chat Bots. Die Weiterentwicklung im Bereich künstlicher neuronaler Netze und insbesondere die steigende Zuverlässigkeit der Netze sorgen darüber hinaus für eine fortschreitende Automatisierung bestimmter Aufgaben, wodurch Kapazitäten zur Konzentration auf andere Aufgabenbereiche sowohl im privaten als auch Unternehmensumfeld frei werden.

Literatur

- Agharaji, E., Damayanti, G., Chandrashekhar, D.: Self-Organizing map based extended fuzzy C-Means (SEFC) algorithm for image segmentation. *Applied Soft Computing* **54**, 347–363 (2017)
- Bai, S., An, S.: A survey on automatic image caption generation. *Neurocomputing* **311**, 291–304 (2018)
- Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. NIPS 2014 Workshop on Deep Learning (2014)
- dos Santos, C., Gatti, M.: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 69–78
- D'Urso, P., Giovanni, L.D., Massari, R.: Smoothed self-organizing map for robust clustering. *Inf. Sci.* **512**, 381–401 (2020)
- Ertel, W.: Grundkurs Künstliche Intelligenz, 4. Aufl. Springer, Wiesbaden (2016)

- Gardner, M.W., Dorling, S.R.: Artificial neural networks (the Multilayer Perceptron) - A review of applications in the atmospheric sciences. *Atmos. Environ.* **32**(14–15), 2627–2636 (1998)
- Hassoun, M.H.: Fundamentals of artificial neural networks. The MIT Press, Cambridge (1995)
- Hochreiter, S., Schmidhuber, J.: Long Short-Term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
- Isa, D., Kallimani, V.P., Lee, L.H.: Using self organizing map for clustering of text documents. *Expert Syst. Appl.* **36**(5), 9584–9591 (2009)
- Karpathy, A., Fei-Fei, L.: Deep Visual-Semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 664–676 (2017)
- Kohonen, T.: Self-Organizing Maps, 2. Aufl. Springer, Berlin (1997)
- Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems – Bd. 1, 1097–1105 (2012)
- Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Ruß, G., Steinbrecher, M.: Computational Intelligence, 1. Aufl. Vieweg + Teubner, Wiesbaden (2011)
- Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), 2267–2273. Austin, Texas (2015)
- LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
- Lin, C., Lee, Y., Pu, H.: Satellite sensor image classification using cascaded architecture of neural fuzzy network. *IEEE Trans. Geosci. Remote Sens.* **38**(2), 1033–1043 (2000)
- Mallot, H.A., Hübner, W.: Neuronale Netze. In: Görz, G., Schneeberger, J., Schmid, U. (Hrsg.) Handbuch der Künstlichen Intelligenz, 5. Aufl., S. 357–404. Oldenbourg, München (2014)
- Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A. L.: Explain Images with Multimodal Recurrent Neural Networks. arXiv preprint [arXiv:1410.1090](https://arxiv.org/abs/1410.1090) (2014)
- McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity: The bulletin of mathematical biophysics. *Bulletin of Mathematical Biophysics* **5**(4), 115–133 (1943)
- Patterson, J., Gibson, A.: Deep Leraning: A Practitioner's Approach. O'Reilly Media Inc, CA (2017)
- Powers, S.T., He, J.: A hybrid artificial immune system and Self Organising Map for network intrusion detection. *Inf. Sci.* **178**(15), 3024–3042 (2008)
- Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint [arXiv:1609.04747](https://arxiv.org/abs/1609.04747) (2016)
- Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986)
- Saraswati, A., Nguyen, V.T., Hagenbuchner, M., Tsoi, A.C.: High-resolution Self-Organizing Maps for advanced visualization and dimension reduction. *Neural Networks* **105**, 166–184 (2018)
- Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015)
- Svozil, D., Kvasnieka, V., Pospichal, J.: Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* **39**, 43–62 (1997)
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A.: Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015)
- Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1422–1432. Lisbon, Portugal (2015)



Bayesian Thinking in Machine Learning

15

Wie ein Pfarrer unbewusst die Statistik revolutionierte

Thomas Neifer, Andreas Schmidt, Dennis Lawo, Lukas Böhm und Özge Tetik

Zusammenfassung

Das Bayes-Theorem ermöglicht die Integration von Vorwissen und Erfahrung in die Datenanalyse und schafft dadurch Instrumente, die einen Mehrwert gegenüber klassischen multivariaten Verfahren hinaus gehen. Im Rahmen des maschinellen Lernens tritt es insbesondere in Regressions- und Klassifikationsfragestellungen in den Vordergrund und dient dort u. a. zur Klassifikation und Analyse von Texten, der Erkennung von Spam-Nachrichten oder auch Spracheingaben bei Sprachassistenten. Dieser Beitrag gibt einen Einblick in die Grundprinzipien des Bayes-Theorems, diskutiert seine Rolle in Regressions- und Klassifikationsfragestellungen und zeigt exemplarisch auf, wie er im Rahmen des Naive Bayes Classifiers im maschinellen Lernen zum Einsatz kommt.

T. Neifer (✉) · A. Schmidt · D. Lawo · L. Böhm · Ö. Tetik

Wirtschaftswissenschaften, Hochschule Bonn-Rhein-Sieg, Sankt Augustin, Deutschland

E-Mail: thomas.neifer@h-brs.de

A. Schmidt

E-Mail: andreas.schmidt@h-brs.de

D. Lawo

E-Mail: dennis.lawo@h-brs.de

L. Böhm

E-Mail: lukas.boehm@uni-siegen.de

Ö. Tetik

E-Mail: ozge.tetik@smail.wis.h-brs.de

15.1 Bayesian Thinking

Der Mathematiker, Statistiker, Philosoph und Pfarrer Thomas Bayes wurde 1701 in London geboren. Er ist bekannt für seine Abhandlung „An Essay Towards Solving a Problem in the Doctrine of Chances“, welche den berühmten Satz von Bayes beinhaltet (Bayes 1763). Allerdings zweifelte Bayes so stark an seiner Entdeckung, dass die dem Bayes-Theorem zugrunde liegende Formel erst posthum veröffentlicht wurde.

Das Bayes-Theorem ist ein Satz der Wahrscheinlichkeitstheorie. Im Rahmen der Inferenzstatistik werden Entscheidungen für einen gegebenen Datensatz anhand der Erkenntnisse über die Parameter von statistischen Modellen getroffen. Während klassische Ansätze lediglich die Daten zur Schätzung der Modellparameter betrachten, bezieht die bayes'sche Statistik zusätzlich auch die vorherigen Informationen über das zugrunde liegende Problem (*a priori Informationen*) in die Parameterschätzung mit ein. Es wird dabei auch nicht zwischen Parametern und Beobachtungen unterschieden, diese werden im Modell alle als Zufallsvariablen betrachtet (Smith und Gelfand 1992; Gasparini 1997; Lee 1997; Albert 2009; Kruschke 2014).

Vereinfacht ausgedrückt wird dadurch adressiert, wie sehr neue Informationen das Vertrauen in einen bestimmten, vorherig gefestigten Glauben verändern sollten. Insbesondere unter Unsicherheit bietet das Bayes-Theorem die Möglichkeit, bestehende Theorien zu aktualisieren und dadurch Entscheidungen unter solchen Bedingungen zu treffen, in denen nur wenige Informationen verfügbar sind (Corfield und Williamson 2013).

Nehmen wir dazu an, dass wir das klassische Zufallsexperiment des Münzwurfs betrachten: Zwei Personen werfen eine faire Münze. Person A beobachtet dabei, dass die Münze auf Zahl gelandet ist und demnach Kopf anzeigen muss. Person B bemerkt jedoch nichts. Unter den klassischen Annahmen wird Person B hier erwartungsgemäß schätzen, dass die Wahrscheinlichkeit für Kopf bei 50 % liegt. Person A besitzt allerdings Vorwissen, welches in die Wahrscheinlichkeitsbeurteilung integriert werden sollte. Daher wird die 50 %ige Wahrscheinlichkeit nicht als konstant angenommen, sondern als Parameter θ betrachtet, welcher je nach Vorwissen variiert.

Das nachfolgende Zitat von Daniel J. Levitin (2016) zeigt den grundlegenden Vorteil des bayesianischen Denkens zusammenfassend auf:

“Critical thinking is an active and ongoing process. It requires that we all think like Bayesians, updating our knowledge as new information comes in.”

Der bayes'sche Ansatz nimmt für eine Zufallsvariable Z mit zugehöriger Wahrscheinlichkeitsverteilung $P(Z = z)$ und bedingter Wahrscheinlichkeitsverteilung $P(Z = z|\theta)$, welche vom unbekannten und zu schätzenden Parameter θ abhängt, eine Verteilung sowohl für die Daten als auch für den Parameter an, einschließlich einer Wahrscheinlichkeitsverteilung $P(\theta)$, welche eine *a priori* (vor den Daten) und eine *a posteriori* (nach den Daten) Form besitzt. Die angenommene Verteilung für die Daten und die *a priori*

Verteilung der Parameter werden durch das *Bayes-Theorem* kombiniert, was zu der *a posteriori* Verteilung $P(\theta|Z = z)$ der Parameter führt (Kruschke 2014) (Abb. 15.1):

mit: $P(\cdot)$ = Dichtefunktion,

θ = Vektor der zu schätzenden Parameter,

z = Vektor der Ausprägungen der abhängigen Variable.

$P(z)$ kann dabei ausgedrückt werden als:

$$P(z) = \int P(z|\theta)P(\theta)d\theta$$

Die gegebenen Beobachtungen z führen dazu, dass $P(z)$ konstant ist. Daher lässt sich der verallgemeinerte *Satz von Bayes* ausdrücken als (Kruschke 2014):

$$P(\theta|z) \propto P(z|\theta) \cdot p(\theta)$$

Der verallgemeinerte *Satz von Bayes* setzt sich demnach aus den folgenden drei Elementen zusammen (Gelman et al. 2013):

A priori Verteilung

Die *a priori Wahrscheinlichkeit* $P(\theta)$ beschreibt einen Wahrscheinlichkeitswert, welcher *a priori* aufgrund von Vorwissen festgelegt werden kann. Ist kein weiteres Vorwissen vorhanden, so bietet sich das *Indifferenzprinzip* an. Dieses basiert auf der *Laplace Wahrscheinlichkeit* und definiert die *a priori Wahrscheinlichkeit* über die Annahme einer diskreten Gleichverteilung als $p = 1/n$ (Laplace 1820). Für *a priori Verteilungen* bedeutet dies, dass diese dem Vorwissen über die Verteilung eines unbekannten Parameters θ entspricht, bevor eine Stichprobe gezogen wurde. θ entstammt der Gesamt-population und wird im Anschluss durch Beobachtungen z der Zufallsgröße Z auf Basis der gezogenen Stichprobendaten mittels des *Satzes von Bayes* geschätzt (Gelman et al. 2013; Kruschke 2014). Die *a priori Verteilung* kann daher auch als Dichtefunktion der zu schätzenden Parameter betrachtet werden und ist vom Anwender zu spezifizieren (Gensler 2003).

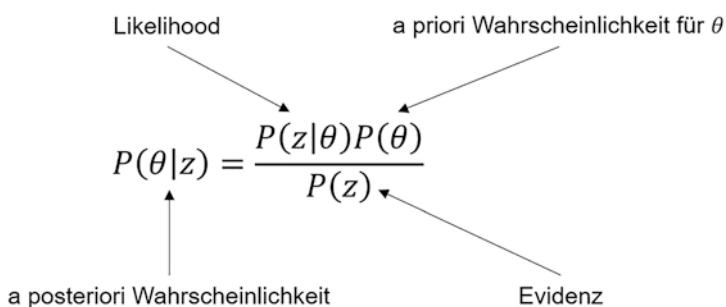


Abb. 15.1 Satz Quelle: Eigene Darstellung in Anlehnung an Lantz (2015)

Likelihood

Die Likelihood $P(z|\theta = \theta_0)$, welche auch als inverse Wahrscheinlichkeit bezeichnet wird, entspricht der Verteilung der Stichprobendaten unter der Bedingung des Vorwissens über den Modellparameter θ (Smith und Gelfand 1992; Lee 1997).

A posteriori Verteilung

Die *a posteriori Wahrscheinlichkeitsverteilung* $P(\theta|z)$ entspricht der Dichtefunktion der Parameter bei gegebenen Beobachtungen. Sie stellt die herzuleitende Größe der Bayes-Statistik dar und ermöglicht eine Aussage über die unbeobachteten Parameter einer Verteilung bei gegebenen Beobachtungen. Die *a priori Verteilung* bestimmt dabei mittels der durch die Daten aktualisierten *Likelihood* die *a posteriori Verteilung* (Gensler 2003; Gelman et al. 2013).

Die *a posteriori Verteilung* der Parameter hat dabei oft keine traktierbare Form und ist somit sowohl nicht vollständig bekannt als auch nicht direkt nutzbar (Smith und Gelfand 1992; Kruschke 2014). Hier kommen Sampling-Verfahren wie MCMC zum Einsatz, welche die Intraktabilität dadurch überwinden, dass sie die Werte von θ aus approximierten bekannten Verteilungen ziehen und diese Ziehungen so lange korrelieren, bis sie näherungsweise an der Zielverteilung $P(\theta|z)$ liegen.

Zur Erklärung der einzelnen Begriffe soll nachfolgend das Beispiel eines Spamfilters verwendet werden. Dieser klassiert eine E-Mail je nach ihrem Inhalt als Spam oder Nicht-Spam (Lantz 2019).

Beispiel

Nehmen wir an, die nachfolgende Tabelle (Tab. 15.1) beinhaltet die bisherigen E-Mail-Nachrichten, welche entsprechend mit dem Label „Spam“ oder „Kein Spam“ versehen sind.

Die *a priori Wahrscheinlichkeit* $P(\theta)$ für eine Spam-E-Mail beschreibt die vorherige Wahrscheinlichkeit, dass es sich bei einer Nachricht ohne Kenntnis ihres Inhalts um Spam handelt. Dies ist bei zwei von fünf Nachrichten der Fall.

$$P(\text{Spam}) = \frac{2}{5}$$

Tab 15.1 Datensatz des Spam-Beispiels. (Quelle: SMS Spam Collection Datensatz (Almeida und Gómez 2011))

Nachricht	Label
Finished where are you	Kein Spam
Wah lucky man. Then can save money	Kein Spam
Today is your lucky day!	Spam
Lucky i havent reply	Kein Spam
Hi, 2nights ur lucky night!	Spam

Die *Evidenz* $P(z)$ für ein spezifisches Wort der Nachricht ergibt sich als Wahrscheinlichkeit, dass ein bestimmtes Wort entweder in einer Spam- oder einer Nicht-Spam-Nachricht vorkam. Da die Evidenz als Divisor auftritt und in beiden Fällen gleich ist, muss sie nicht berücksichtigt werden.

Die *Likelihood* $P(z|\theta)$ beschreibt die Wahrscheinlichkeit, dass ein spezifisches Wort z in einer E-Mail vorkommt, die als Spam klassifiziert ist. „lucky“ tritt zwei Mal in Spam-Nachrichten mit insgesamt zehn Wörtern auf.

$$P(\text{lucky}|\text{Spam}) = \frac{2}{10}$$

Die *a posteriori Wahrscheinlichkeit* $P(\theta|z)$ definiert die Wahrscheinlichkeit für eine Spam-E-Mail im Falle eines spezifisch enthaltenen Wortes, z. B. „lucky“. Diese beträgt hier 8 %.

◀
$$P(\text{lucky}|\text{Spam}) = P(\text{lucky}|\text{Spam}) \cdot P(\text{Spam}) = \frac{2}{10} \cdot \frac{2}{5} = \frac{2}{25} = 0,08$$

Für endlich viele Ereignisse $A_i, i = 1, \dots, N$ lässt sich der Satz von Bayes zur Berechnung der a posteriori Wahrscheinlichkeit $P(A_i|B)$ ausdrücken als (Lantz 2019):

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1}^N P(B|A_j) \cdot P(A_j)}$$

15.2 Bayes in Machine Learning

Das Bayes-Theorem ermöglicht es, das Wissen über ein Problem schrittweise zu aktualisieren, wenn mehr Informationen darüber hinzukommen. Diesen Grundsatz machen sich auch verschiedene Verfahren des maschinellen Lernens (Machine Learning, ML) zunutze, weshalb das Bayes-Theorem auch dort weite Verbreitung findet. So kommt es sowohl im Rahmen von Klassifikations- als auch Regressionsverfahren zum Einsatz (Lantz 2019).

15.2.1 Bayes in Regressionsverfahren

Die Regressionsanalyse zielt darauf ab, Abhängigkeiten zwischen einer abhängigen Variable y sowie einer oder mehreren unabhängigen Variablen $x_i, i = 1, \dots, n$ formal zu modellieren. Die Beziehung zwischen abhängiger und unabhängigen Variablen kann ausgedrückt werden als:

$$y = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_n \cdot x_n +$$

Als Matrixgleichung formuliert, kann das lineare Modell für eine beliebige Anzahl von Prädiktoren weiter verallgemeinert werden:

$$y = \theta^T \cdot X +$$

Die wahren Modellparameter sind hier als θ gekennzeichnet. θ_0 beschreibt dabei den Niveauparameter, die weiteren Gewichte jeweils die Steigungsparameter und damit die Auswirkungen von Veränderungen der zugehörigen Prädiktorvariablen x_i auf die abhängige Variable y . Bei ϵ handelt es sich um eine stochastische Störgröße (Residuum), welche die Abweichungen zwischen geschätzter Regressionsfunktion \hat{y} und den tatsächlichen Datenpunkten y wiedergibt. Eine Voraussetzung im Rahmen der linearen Regressionsanalyse ist dabei, dass ϵ normalverteilt ist (Elster et al. 2015; Natrop 2015).

Klassische lineare Regression: Im Rahmen der klassischen linearen Regression werden Trainingsdaten dazu verwendet, um diejenigen Werte für die Modellparameter θ zu finden, welche die Daten bestmöglich erklären. Die Ermittlung der optimalen Schätzwerte wird zumeist über die Methode der kleinsten Quadrate (Ordinary Least Squares, OLS) oder eine Maximum Likelihood-Schätzung (Maximum Likelihood Estimation, MLE) realisiert. Dabei wird in beiden Fällen die Summe der kleinsten Abweichungsquadrate (Sum of Squared Residuals, SSR) minimiert, was zu den entsprechend der Daten optimalen Modellwerten führt (Elster und Wübbeler 2015).

Bayesianische lineare Regression: Anders als die Verfahren der klassischen Regressionsanalyse, besteht das Ziel der bayesianischen linearen Regression nicht darin, den einen, optimalen Wert der Modellparameter zu finden, sondern die a posteriori Verteilung für die Modellparameter zu bestimmen. Damit greift diese Methode nicht auf Punktschätzungen wie OLS oder MLE zurück, sondern leitet ihre Parameter aus Wahrscheinlichkeitsverteilungen ab (Elster et al. 2015).

Folgt die abhängige Variable y einer Normalverteilung, welche durch den Erwartungswert μ und die Varianz σ^2 beschrieben wird, so kann das Modell definiert werden als:

$$y \sim N(\mu, \sigma^2 I)$$

μ ergibt sich hierbei als transponierte Parametermatrix multipliziert mit der Prädiktormatrix: $\mu = \theta^T X$. Aufgrund des mehrdimensionalen Modells gibt I die Einheitsmatrix an (Deisenroth et al. 2020).

Neben der Tatsache, dass y hier aus einer Verteilung abgeleitet wird, wird im Rahmen der bayesianischen Regression ebenfalls angenommen, dass die Modellparameter auch aus einer solchen Verteilung stammen. Dies wird über den Satz von Bayes realisiert, der die a posteriori Verteilung der Modellparameter bestimmt. Diese hängt von den Trainingsdaten ab und ergibt sich als:

$$P(\text{Modell}|\text{Daten}) = \frac{P(\text{Daten}|\text{Modell}) \cdot P(\text{Modell})}{P(\text{Daten})}$$

Durch diese Modellierung ist es möglich, Vorkenntnisse über die Parameter eines Modells mit in die Berechnung einzubinden, bevor das Modell entsprechend der Daten optimiert wird. $P(\text{Modell})$ repräsentiert hier die a priori Annahme der Verteilung der Modellparameter. Werden die Parameter des Modells nun entsprechend der neuen Daten

über $P(\text{Daten}|\text{Modell})$ aktualisiert, so ergibt sich dadurch die a posteriori Verteilung der Modellparameter $P(\text{Modell}|\text{Daten})$ (Gelman et al. 2013).

Entsprechend ergibt sich für die Regressionsparameter (Deisenroth et al. 2020):

$$P(\theta|y, X) = \frac{P(y|\theta, X) \cdot P(\theta|X)}{P(y|X)}$$

Nehmen wir für unser Modell an, dass unser Satz an Parametern θ normalverteilt ist (a priori Annahme, Prior). Wird das Modell mit zusätzlichen Daten trainiert, so verringert sich die Varianz und die a posteriori Verteilung wird aktualisiert und genauer (vgl. Abb. 15.2) (Smith und Gelfand 1992; Gasparini 1997; Gelman et al. 2013).

Kommen weitere Daten hinzu, so wird die a posteriori Verteilung zu der a priori Verteilung und es startet eine neue Iteration. Ist das Modell ausreichend trainiert, so wird der maximum a posteriori-Schätzer (MAP-Schätzer) $\hat{\theta}_{map}$ als Parameterwert verwendet (Deisenroth et al. 2020).

Allerdings hat die a posteriori Verteilung der Parameter oftmals keine traktierbare Form und ist somit sowohl nicht bekannt als auch nicht direkt zur Approximation eines MAP-Schätzers nutzbar. Um diesem Problem zu begegnen, kommen zumeist Markov

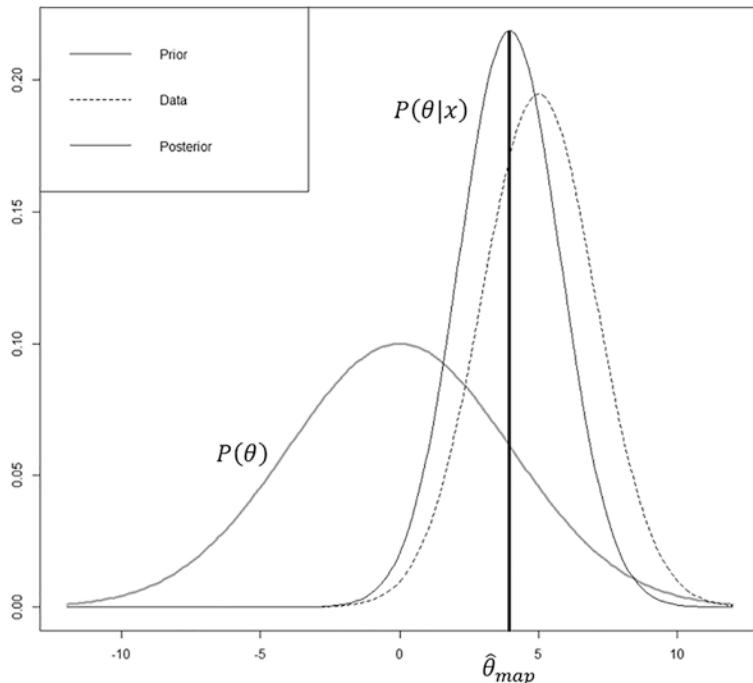


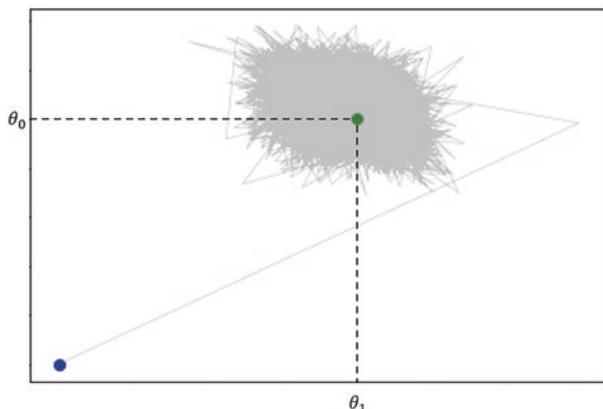
Abb. 15.2 a priori und a posteriori Verteilungen der Parameter θ Quelle: Eigene Darstellung in Anlehnung an Deisenroth et al. (2020)

Chain Monte Carlo-Methoden (MCMC-Methoden) zum Einsatz, welche Zufallsstichproben aus einer Verteilung ziehen, um diese zu approximieren (Gelman et al. 2013). Sie führen dabei eine sequenzielle Simulation durch, bei der die Verteilung jedes simulierten Wertes von dem vorhergehenden abhängt und die Ziehungen somit eine Markov-Kette bilden. Die Markov-Kette stellt hierbei sicher, dass sich die approximativen Verteilungen bei jedem Schritt der Simulation durch Konvergenz zum Zielwert verbessern. Die stationäre Verteilung des Markov-Prozesses entspricht dabei der Zielverteilung $P(\theta|y, X)$. Die Simulation wird so lange durchgeführt, bis die aktuelle Verteilung der stationären Verteilung sehr nahe kommt. Für die Parameter θ_0 und θ_1 ist dies beispielhaft in Form eines Trace Plots dargestellt, welcher die Annäherungen des MCMC-Samplers an die wahren Werte θ_0 und θ_1 aufzeigt (vgl. Abb. 15.3) (Brooks et al. 2011; Gelman et al. 2013).

Klassische versus bayesianische lineare Regression: Um die Vorteile des bayes'schen Ansatzes aufzuzeigen, soll nachfolgend anhand eines Beispiels die Ergebnisse eines OLS- und Bayes-Modells gegenübergestellt werden. Das Beispiel wurde in R programmiert und es wurde der Boston-Datensatz verwendet, welcher Informationen über Gebiete rund um die Stadt Boston und den Median der dortigen Hauspreise beinhaltet. Ziel ist es, die Hauspreise anhand verschiedener Merkmale (z. B. Kriminalitätsrate im Stadtteil, durchschnittliche Raumanzahl, etc.) über eine lineare Regression vorherzusagen. Dazu wurden die Daten eingangs in Trainings- (70 %) und Testdaten (30 %) aufgeteilt. Die Ergebnisse können der nachfolgenden Tabelle (Tab. 15.2) entnommen werden.

Es wird ersichtlich, dass die Gütemaße sich nur unwesentlich unterscheiden. Daher bleibt die berechtigte Frage, welchen Vorteil das Bayes'sche Regressionsmodell bietet. Dazu betrachten wir einen Plot der Konfidenzintervalle der beiden Modelle für die vorhergesagten und tatsächlichen Werte (vgl. Abb. 15.4).

Abb. 15.3 Trace-Plot des MCMC-Samplers



Tab. 15.2 Ergebnisse von OLS- und Bayesianischer Regressionsanalyse. Quelle: Eigene Berechnungen

Modell	Bestimmtheitsmaß R^2	Multipler Korrelationsko-effizient R	Mean Absolute Error MAE
OLS-Modell	0,7525	0,8.183.175	3,261.146
Bayes-Modell	0,7615	0,8.182.854	3,260.804

Dadurch, dass der bayesianische Ansatz die Verteilung der Parameter aus einer a posteriori Verteilung berechnet, welche sich bei neuen Informationen aktualisiert, fällt die Varianz geringer aus. Dies führt dazu, dass die Konfidenzintervalle der Werte kompakter werden und das Vertrauen in jeden einzelnen Wert zunimmt (Elster und Wübbeler 2015; Spyroglou und Rigas 2018; Deisenroth et al. 2020).

Die Vorteile der bayes'schen Modellierung liegen somit in der a priori Annahme, die es im Falle von Vor- oder Domänenwissen ermöglicht, dieses im Modell zu integrieren. Die klassischen Ansätze (OLS, MLE) ignorieren dieses Vorwissen und gehen davon aus, dass das Wissen über die Regressionsparameter nur aus den Daten generiert werden kann. Weiterhin ermöglicht der bayes'sche Ansatz die Quantifizierung der Unsicherheit über das Modell. Je mehr Datenpunkte es gibt, desto weniger stark ist die a posteriori Verteilung der Modellparameter aufgrund der Aktualisierung des Modells verteilt (Corfield und Williamson 2013; Gelman et al. 2013; Elster et al. 2015; Elster und Wübbeler 2015).

15.2.2 Bayes in Klassifikationsverfahren

Im Rahmen des ML ist das Bayes-Theorem insbesondere bei Klassifikationsproblemen weit verbreitet. Klassifikationsverfahren beschäftigen sich damit, die Wahrscheinlichkeit der Zugehörigkeit zu einer Klasse bzw. Kategorie anhand der beobachteten Merkmale zu bestimmen (Lantz 2019). Es wird auch hier zwischen a posteriori und a priori Wahrscheinlichkeiten unterschieden, jedoch handelt es sich hierbei um Klassenwahrscheinlichkeiten. Der Satz von Bayes lässt sich demnach verstehen als:

$$P(\text{Klasse}|\text{Daten}) = \frac{P(\text{Daten}|\text{Klasse}) \cdot P(\text{Klasse})}{P(\text{Daten})}$$

$P(\text{Daten}|\text{Klasse})$ definiert die Wahrscheinlichkeitsdichtefunktion aller Daten, die zur betrachteten Klasse gehören. Sie wird aus den Trainingsdaten berechnet und auch als Wahrscheinlichkeit verstanden, die ein Prädiktor bei gegebener Klasse innehalt. Bei $P(\text{Daten})$ handelt es sich um die Wahrscheinlichkeitsdichtefunktion, die allen zugrunde liegenden Daten gemeinsam ist und $P(\text{Klasse})$ ist die a priori Wahrscheinlichkeit der Klassenzugehörigkeit. $P(\text{Klasse}|\text{Daten})$ beschreibt schließlich die a posteriori Wahrscheinlichkeit einer Klasse bei gegebenen Prädiktoren (Janssen und Laatz 2017).

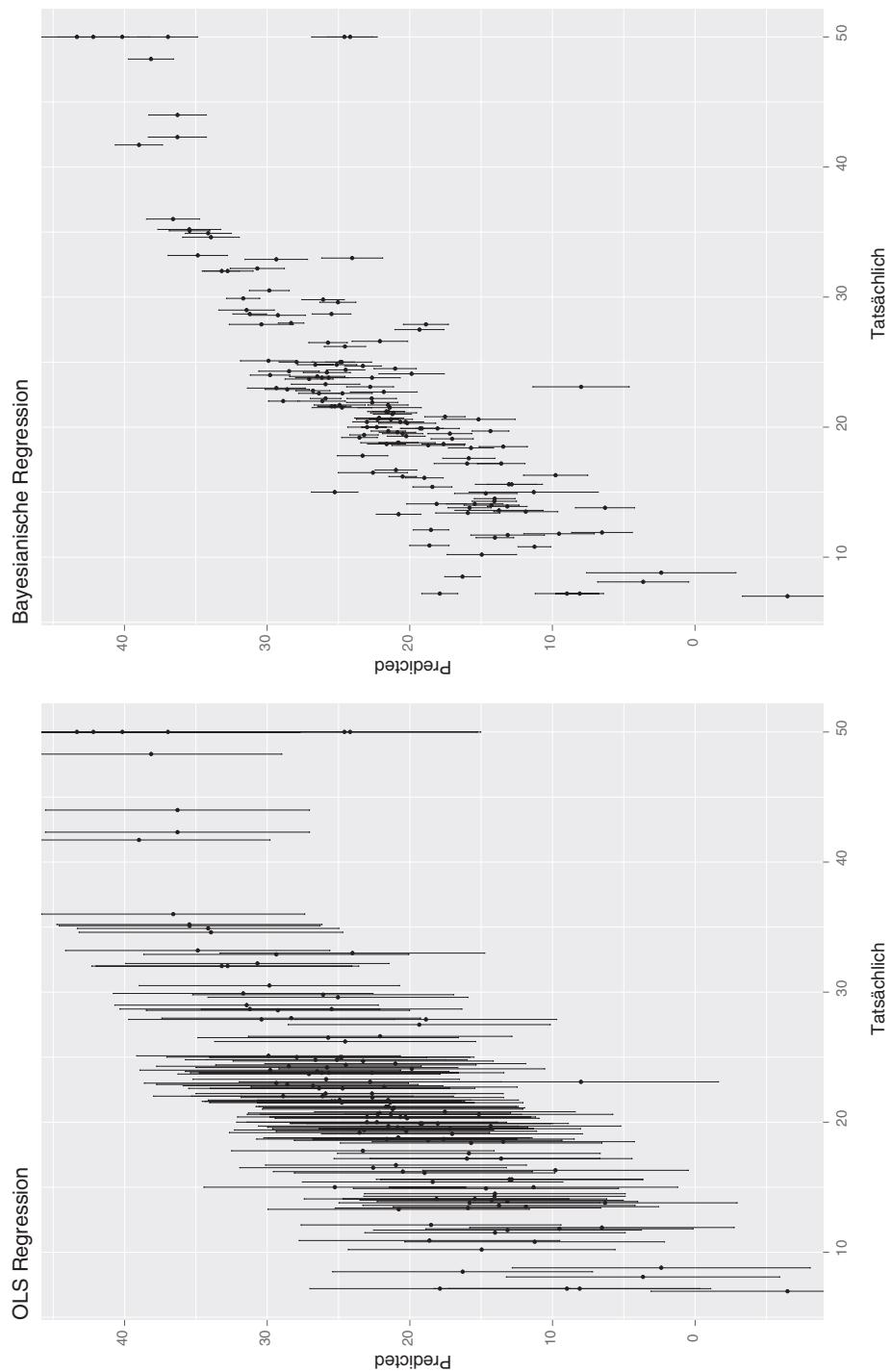


Abb. 15.4 Vergleich der Konfidenzintervalle für tatsächliche und vorhergesagte Werte

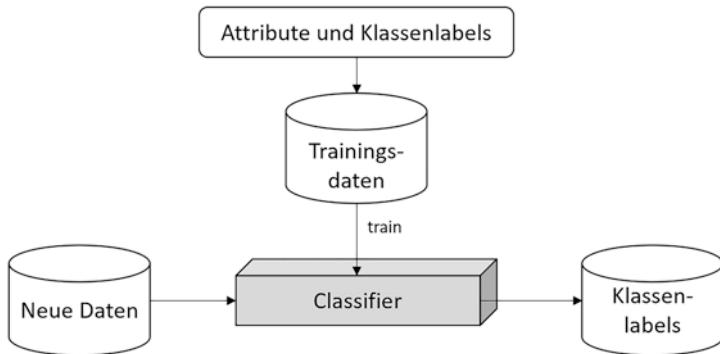


Abb. 15.5 Klassifikationsmodell. Quelle: Eigene Darstellung in Anlehnung an Kirk (2017)

Anhand von Neifer_ Abb. 15.5 wird ersichtlich, wie ein Klassifikator im ML konstruiert wird (Kirk 2017).

Dadurch, dass der Bayes-Gedanke die Integration von Vorwissen in die Klassifikation ermöglicht und die Wahrscheinlichkeiten anhand neuer Daten inkrementell aktualisiert werden, erzielen die bayesianischen Klassifikatoren sehr gute Ergebnisse. Mit jedem neu klassifizierten Datensatz wird sich der Klassifikator verbessern (Noaman et al. 2010; Janssen und Laatz 2017).

Aufgrund seiner hohen praktischen Relevanz soll nachfolgend der Naive Bayes Classifier exemplarisch vorgestellt werden.

15.3 Naive Bayes Classifier

15.3.1 Grundlagen

Der Naive Bayes Classifier (NBC) ist ein Klassifikationsalgorithmus des überwachten maschinellen Lernens, dessen zugrunde liegendes probabilistische Modell auf dem Bayes-Theorem basiert. Dadurch hat er den Vorteil, dass auch bereits kleine Mengen an Trainingsdaten für gute Ergebnisse ausreichen und er dort viele andere Alternativen übertrifft (Dey et al. 2016). Der Algorithmus wird als naiv bezeichnet, da er auf der grundlegenden Annahme basiert, dass alle Merkmale eines Datensatzes gleich relevant und unabhängig voneinander sind. In der Praxis trifft dies jedoch nur in seltenen Fällen zu, dennoch realisiert er auch dort gute Resultate. Insbesondere für die Klassifizierung von Texten erfährt der NBC eine große Aufmerksamkeit (Lantz 2019). So setzen bspw. Spamfilter auf den NBC, um anhand des E-Mail-Inhalts ein Labeling in Spam oder Nicht-Spam vorzunehmen (Ramasubramanian und Singh 2018). Weiterhin können auch Produkt-Reviews klassifiziert, Dokumente kategorisiert, Suchmaschineneingaben komplettiert und die Verarbeitung von Spracheingaben im Rahmen von Voice Assistants optimiert werden (Jansen et al. 2007; Noaman et al. 2010; Haque et al. 2018; Polyakov et al. 2018).

15.3.2 Methodik

Zum besseren Verständnis betrachten wir die Methodik des NBC im Rahmen eines einfachen Beispiels (Oettinger 2020).

Beispiel

Nehmen wir an, wir haben einen Datensatz über verschiedene Früchte. Dieser umfasst 1.000 Trainingsdaten und besteht aus der Fruchtbezeichnung, der Länge der Frucht (kurz oder lang), dem Geschmack der Frucht (süß oder nicht süß) sowie der Farbe (gelb oder nicht gelb). Tabellarisch (Tab. 15.3) können die Merkmale mit ihren Häufigkeiten und der zugehörigen Fruchtbezeichnung wie folgt dargestellt werden:

Die a priori Wahrscheinlichkeit für die jeweiligen Früchte lauten demnach:

$$P(\text{Banane}) = \frac{500}{1.000} = 0,5; P(\text{Orange}) = \frac{300}{1.000} = 0,3 \text{ und } P(\text{Sonstige Frucht}) = \frac{200}{1.000} = 0,2$$

$$P(\text{lang}) = \frac{500}{1.000} = 0,5; P(\text{süß}) = \frac{650}{1.000} = 0,65 \text{ und } P(\text{gelb}) = \frac{800}{1.000} = 0,8$$

Unter der Bedingung des a priori bekannten Vorwissens, besitzen die Stichprobendaten folgende Likelihoods:

$$P(\text{lang} \mid \text{Banane}) = \frac{400}{500} = 0,8; P(\text{lang} \mid \text{Orange}) = \frac{0}{300} = 0,$$

$$P(\text{lang} \mid \text{Sonstige Frucht}) = \frac{100}{200} = 0,5 \text{ und } P(\text{nicht gelb} \mid \text{Sonstige Frucht}) = \frac{150}{200} = 0,75$$

Die a posteriori Wahrscheinlichkeit für eine neu zu klassifizierende Frucht ergibt sich nach dem Satz von Bayes für eine als lang, süß und gelb beschriebene Frucht jeweils als:

Tab. 15.3 Häufigkeitsverteilung des Frucht-Beispiels. Quelle: Oettinger (2020)

Frucht-bezeichnung	Länge		Geschmack		Farbe		Gesamtanzahl der Früchte
	lang	kurz	süß	nicht süß	gelb	nicht gelb	
Banane	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Sonstige Frucht	100	100	150	50	50	150	200
Gesamt	500	500	650	350	800	200	1,000

$$P(\text{Banane}|\text{lang, süß, gelb}) = \frac{P(\text{lang}|\text{Banane}) \cdot P(\text{süß}, |\text{Banane}) \cdot P(\text{gelb}|\text{Banane}) \cdot P(\text{Banane})}{P(\text{lang}) \cdot P(\text{süß}) \cdot P(\text{gelb})}$$

$$P(\text{Banane}|\text{lang, süß, gelb}) = \frac{0,8 \cdot 0,7 \cdot 0,9 \cdot 0,5}{0,5 \cdot 0,65 \cdot 0,8}$$

$$P(\text{Orange}|\text{lang, süß, gelb}) = \frac{P(\text{lang}|\text{Orange}) \cdot P(\text{süß}, |\text{Orange}) \cdot P(\text{gelb}|\text{Orange}) \cdot P(\text{Orange})}{P(\text{lang}) \cdot P(\text{süß}) \cdot P(\text{gelb})}$$

$$P(\text{Orange}|\text{lang, süß, gelb}) = \frac{0 \cdot 0,5 \cdot 1 \cdot 0,3}{0,5 \cdot 0,65 \cdot 0,8}$$

$$P(\text{S.F.}|\text{lang, süß, gelb}) = \frac{P(\text{lang}|\text{S. F.}) \cdot P(\text{süß}, |\text{S. F.}) \cdot P(\text{gelb}|\text{S.F.}) \cdot P(\text{S.F.})}{P(\text{lang}) \cdot P(\text{süß}) \cdot P(\text{gelb})}$$

$$P(\text{S. F.}|\text{lang, süß, gelb}) = \frac{0,5 \cdot 0,75 \cdot 0,25 \cdot 0,2}{0,5 \cdot 0,65 \cdot 0,8}$$

Dabei kann aufgrund dem für alle geltenden konstanten Wert für $P(\text{lang}) \cdot P(\text{süß}) \cdot P(\text{gelb})$ die Wahrscheinlichkeit vereinfacht ermittelt werden als:

$$P(\text{Banane}|\text{lang, süß, gelb}) = 0,8 \cdot 0,7 \cdot 0,9 \cdot 0,5 = 0,252$$

$$P(\text{Orange}|\text{lang, süß, gelb}) = 0 \cdot 0,5 \cdot 1 \cdot 0,3 = 0$$

$$P(\text{SonstigeFrucht}|\text{lang, süß, gelb}) = 0,5 \cdot 0,75 \cdot 0,25 \cdot 0,2 = 0,1875$$

Nun wird der wahrscheinlichste Wert ausgewählt. Dieser liegt bei der Banane, weshalb die Frucht als Banane ($0,252 > 0,1875 > 0$) klassifiziert wird. ◀

Laplace-Korrektur

Die Wahrscheinlichkeit für Orangen unter der Bedingung, dass diese lang, süß und gelb sind war im obigen Beispiel Null. Dies ist darauf zurückzuführen, dass in den Trainingsdaten keine Orangen vorhanden waren, die das Merkmal „lang“ aufwiesen. Dies ist nachvollziehbar, da Orangen generell nicht lang sind. Im Rahmen eines Modells mit vielen Merkmalen ist es jedoch unter Umständen nicht sinnvoll, dass die Wahrscheinlichkeit eines Merkmals den Wert Null annimmt, denn dadurch würde die Gesamtwahrscheinlichkeit ebenfalls Null betragen, auch wenn die anderen Merkmale eine hohe Wahrscheinlichkeit aufweisen würden. Um eine solche Situation zu vermeiden, werden die Variablen mit einer Häufigkeit von Null auf Eins erhöht. Dies wird als Laplace-Korrektur bezeichnet (Maimon und Rokach 2005).

Mit diesem Vorgehen wird nun die Bestimmung der Zugehörigkeitswahrscheinlichkeit von neuen, unbekannten Daten zu einer Klasse ermöglicht. Der NBC ist in verschiedenen Anwendungsbereichen ein wirkungsvolles Instrument zur Klassifikation.

So kommt er auch im Rahmen des Natural Language Processing (NLP) u. a. zur Stimmungsanalyse zum Einsatz. Dabei ist das Vorgehen analog zur obigen grundlegenden Methodik (Dey et al. 2016).

15.4 Fazit

Es wurde aufgezeigt, dass das bayesianische Denken über die Integration von Vorwissen in die Analyse enorme Potenziale bietet (Levitin 2016). Es erfährt im maschinellen Lernen große Aufmerksamkeit und wird insbesondere zur Lösung von Regressions- und Klassifikationsproblemen genutzt.

Der auf dem Bayes-Theorem basierende NBC hat Vor- und Nachteile, die je nach Situation zu berücksichtigen sind. Die Vorteile bestehen insbesondere in der effizienten Klassifikation sowie dem inkrementellen Lernprozess, welcher durch den Satz von Bayes dafür sorgt, dass der Klassifikator mit jedem Datensatz genauer wird. Er besticht hierbei insbesondere durch seine Einfachheit und Schnelligkeit und stellt ein wirkungsvolles Instrument, auch bei hochdimensionalen Problemen dar (Robles et al. 2003). Demgegenüber wird als kritisch erachtet, dass er auf der unrealistischen und vereinfachenden Annahme der Unabhängigkeit der Merkmale untereinander beruht.

Die auf dem Satz von Bayes basierenden Markov Chain Monte Carlo-Methoden stellen darüber hinaus eine Klasse der mächtigsten Methoden zur Verarbeitung von Daten und Wissen dar. Durch die Verwendung von Algorithmen statt Theoremen können dadurch reale Probleme „exakt simuliert“ werden. MCMC wird daher auch als ein Quantensprung der Statistik bezeichnet (McGrayne 2011).

Literatur

- Albert, J.: Bayesian computation with R. Springer, NY (2009)
- Almeida, T. A., José María, G.: SMS Spam Collection. SMS Spam Collection v.1. <https://www.dt.fee.unicamp.br/~tiago/smsspamcollection/> (2011). Zugegriffen: 26. Juli 2020
- Bayes, T.: LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London 370–418 (1763)
- Brooks, S., Gelman, A., Jones, G., Meng, X.L.: Handbook of markov chain monte carlo. CRC Press, NY (2011)
- Corfield, D., Williamson, J.: Foundations of Bayesianism. Springer Science & Business Media, Dordrecht (2013)
- Deisenroth, M. P., Aldo Faisal, A., Ong, C. S.: Mathematics for machine learning. Cambridge University Press, Cambridge (2020)
- Dey, L., Chakraborty, S., Biswas, A., Bose, B., Tiwari, S.: Sentiment analysis of review datasets using naive bayes and k-nn classifier (2016). arXiv preprint [arXiv:1610.09982](https://arxiv.org/abs/1610.09982)
- Elster, C. et al.: A guide to Bayesian inference for regression problems. deliverable of EMRP Project NEW04 “Novel Mathematical and Statistical Approaches to Uncertainty Evaluation,” (2015)

- Elster, C., Wübbeler, G.: Bayesian regression versus application of least squares—an example. *Metrologia* 53, S. 10 (2015)
- Gasparini, M.: Markov chain Monte Carlo in practice. Taylor & Francis, London (1997)
- Gelman, A. et al.: Bayesian data analysis. Chapman and Hall/CRC Press, Taylor & Francis, Boca Raton, London, New York (2013)
- Gensler, S.: Heterogenität in der Präferenzanalyse. Deutscher Universitätsverlag, Wiesbaden (2003)
- Haque, T. U., Saber, N. N., Shah, F. M.: Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 1–6. Bangkok: IEEE <https://ieeexplore.ieee.org/document/8376299/> (2018). Zugegriffen: 18. Juli 2020
- Jansen, B. J., Booth, D. L., Spink, A.: Determining the user intent of web search engine queries. In Proceedings of the 16th international conference on World Wide Web, 1149–1150 (2007)
- Janssen, J., Laatz, W.: Naive Bayes. In Statistische Datenanalyse mit SPSS S. 557–569. Springer, Berlin (2017)
- Kirk, M.: Thoughtful Machine Learning with Python: A Test-Driven Approach. O'Reilly Media, Inc., Sebastopol (2017)
- Kruschke, J.: Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press, London, San Diego, Waltham, Oxford (2014)
- Lantz, B.: Machine learning with R: expert techniques for predictive modeling. Packt Publishing Ltd., Birmingham (2019)
- Laplace, P. S.: Théorie analytique des probabilités. Courcier, Paris (1820)
- Lee, P.M.: Bayesian Statistics, (Arnold). Wiley, New York (1997)
- Levitin, D. J.: A field guide to lies: Critical thinking in the information age. Penguin Random House, New York City (2016)
- Maimon, O., Rokach, L.: Decomposition methodology for knowledge discovery and data mining. In Data mining and knowledge discovery handbook, S. 981–1003. Springer, Boston (2005)
- McGrayne, S.B.: The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy. Yale University Press, UK (2011)
- Natrop, J.: Angewandte Descriptive Statistik: Praxisbezogenes Lehrbuch mit Fallbeispielen. De Gruyter Oldenbourg, Berlin (2015)
- Noaman, H. M., Elmougy, S., Ghoneim, A., Hamza, T.: Naive Bayes classifier based Arabic document categorization. In 2010 the 7th International Conference on Informatics and Systems (INFOS), S. 1–5. IEEE (2010)
- Oettinger, M.: Data Science: Eine praxisorientierte Einführung im Umfeld von Machine Learning, künstlicher Intelligenz und Big Data. 2. erw. Aufl., tredition, Hamburg (2020)
- Polyakov, E. V. et al.: Investigation and development of the intelligent voice assistant for the Internet of Things using machine learning. In 2018 Moscow Workshop on Electronic and Networking Technologies (MWENT), S. 1–5. IEEE (2018)
- Ramasubramanian, K., Singh, A.: Machine Learning Using R: With Time Series and Industry-Based Use Cases in R. Apress, New York (2018)
- Robles, V., Larrañaga, P., Menasalvas, E., Pérez, M. S., Hervés, V.: Improvement of Naive Bayes collaborative filtering using interval estimation. In Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), S. 168–174. IEEE (2003)
- Smith, A.F., Gelfand, A.E.: Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician* 46(2), 84–88 (1992)
- Spyroglou, I.I., Rigas, A.G.: A two-step Bayesian approach for modeling a complex neurophysiological system. *International journal of biology and biomedical engineering* 12, 66–74 (2018)

Teil IV

Anwendungsorientierte Data Science

Patrick Bachmaier: Text Mining: Durchführung einer Sentiment Analysis mit SAP HANA

Christoph Quix: Weiterbildung in Data Science

Alexa Scheffler und Valeria Knoll: Erfolgsfaktoren von Big Data Plattformen aus dem Blickwinkel der Plattform Ökonomie

Markus Eßwein, Domenica Martorana, Martina Reinersmann und Peter Chamoni: Akzeptanz und Nutzung von maschinellem Lernen und Analytics im Rechnungswesen und Controlling

Eva Schoetzau: Durch Big Data zu neuen Geschäftsmodellen – und umgekehrt im Kontext von Car-Sharing

Uwe Rudolf Fingerlos, Alexander Pastwa: Einsatz von Logit- und Probit-Modellen in der Finanzindustrie



Text Mining: Durchführung einer Sentiment Analysis mit SAP HANA

16

Patrick Bachmaier

Zusammenfassung

In diesem Beitrag wird das Vorgehen zur erfolgreichen Durchführung einer Sentiment Analysis mit SAP HANA beschrieben. Es wird einer gekürzten Version eines Vorgehensmodells gefolgt, welches im Rahmen der Masterthesis von Herrn Bachmaier erarbeitet wurde. Im Detail wird beschrieben, welche konkreten Tätigkeiten in SAP HANA umzusetzen sind, um grundlegend eine Stimmungsanalyse auf Basis von Twitterdaten durchzuführen. Dazu gehören in diesem Fall die zunächst notwendige Datenakquise sowie u. a. Tokenization, Stop Word Removal, Stemming und Classification. Abschließend erfolgt die Betrachtung der Ergebnisse der Sentiment Analysis.

16.1 Einleitung

Die äußerst schnell fortschreitende Verbreitung der Social Media sowie die hochfrequente Nutzung von Social Media Plattformen führt dazu, dass die Aufmerksamkeit von Unternehmen immer stärker auf Analysen der daraus resultierenden Daten gelenkt wird. Die Ergebnisse dieser Analysen sind immer öfter für unternehmensstrategische Entscheidungen relevant. (vgl. Böck et al. 2017, S. 1.)

Ein grundsätzliches Kundenfeedback kann durch Product Reviews oder auch durch E-Mail-Nachrichten an das Unternehmen übermittelt werden. Strebt ein Unternehmen einen besseren Überblick über das eigene Kundenfeedback an, so ist hierbei ein vertieftes

P. Bachmaier (✉)
Moers, Deutschland
E-Mail: patrick.bachmaier@web.de

Verständnis von Textdaten notwendig. Das Internet stellt eine riesige Menge von Textdaten zur Verfügung. Diese können u. a. auf Plattformen wie Twitter, Facebook, oder aber auch anderen Internetplattformen abgerufen werden. Textdaten sind praktisch überall vertreten. (vgl. Provost und Fawcett 2013, S. 252.)

Das Aggregieren von Meinungen der Kunden ermöglicht den Zugang zu generellen Informationen bezüglich des Stimmungsbildes der Kundschaft. Die Analyse solcher Daten ermöglicht, die Kunden besser zu verstehen und damit einhergehend ein besseres Verständnis über die Wünsche der Kunden zu erhalten, und kann daher dazu beitragen, Entscheidungen im Unternehmen zu optimieren. (vgl. Zhai und Massung 2016, S. 393.)

Um diese Informationen aus Texten zu erhalten, müssen auf Basis von Textanalysen, unstrukturierte Textdaten in strukturierte Daten transformiert und aufbereitet werden. Semantische sowie linguistische Technologien können u. a. Fakten und auch Kernaussagen extrahieren. (vgl. Kayser und Rath 2015, S. 132.) Speziell die Analyse von Sentiments (zu Deutsch „Stimmungen“), welche in den Textdaten enthalten sein können, setzt deren Identifikation voraus. Würde dies manuell durchgeführt, so würde diese Unternehmung sehr viel Zeit in Anspruch nehmen. Es ist somit eine Automatisierung erforderlich. (vgl. Kaiser 2012, S. 13.)

Im Rahmen dieser Arbeit wird eine Sentiment Analysis, basierend auf deutschen Textdaten, durchgeführt. Als Datenquelle wird Twitter ausgewählt. Der inhaltliche Betrachtungsaspekt, zu dem Textdaten mit Sentiments ermittelt werden, lautet Elektroautos. Es ist möglich, jedes beliebige andere Thema zu wählen. Um die Sentiment Analysis durchzuführen, wird SAP HANA eingesetzt.

16.2 Grundlagen

Text Mining stellt einen Prozess dar, bei welchem unstrukturierte Daten (Text) in einer Form aufbereitet werden, in welcher diese mit weiteren Analysemethoden verarbeitet werden können, um Informationen daraus zu extrahieren. Prinzipiell wird beim Text Mining versucht, Strukturen und Muster in Texten zu ermitteln. (vgl. Gronwald 2017, S. 141.)

Generell kann Text Mining in zwei Ansätze unterteilt werden. Zum einen in den mathematischen Ansatz, bei welchem numerische Methoden angewendet werden, um möglichst viele Informationen aus Texten zu erhalten. Und zum anderen in den linguistischen Ansatz, durch welchen versucht wird, die Struktur und die Bedeutung des Textes durch Grammatik zu ermitteln. (Vgl. Gronwald 2017, S. 142.)

Bei einer Sentiment Analysis wird die Analyse von Emotionen und Meinungen einzelner Personen, z. B. gegenüber Unternehmen oder Produkten, durchgeführt. Solche Meinungen und Emotionen werden beispielsweise in Texten in sozialen Medien kommuniziert. (vgl. Gronwald 2017, S. 55.) Zur Durchführung einer Sentiment Analysis wird textueller Input benötigt, welcher analysiert wird, um verschiedene Sentiments zu extrahieren (vgl. Zhai und Massung 2016, S. 392).

Die Tonalität von Inhalten wird als Sentiment bezeichnet. Diese Tonalität stellt dar, wie über ein gewisses Thema kommuniziert wird. In der Regel erfolgt eine Einordnung von Sentiments in „positive“, „negative“ und „neutrale“ Abstufungen. Hierbei treten häufig Probleme auf. Aus diesem Grund sollte immer auch eine manuelle Prüfung der Klassifikation durchgeführt werden. (vgl. Evertz 2018, S. 150 f.)

Mit SAP HANA ist die Durchführung von Sentiment Analysen realisierbar (vgl. SAP o. Jg). SAP HANA verfügt hierzu über eine Text Analytics Engine, welche es Entwicklern erlaubt, entsprechende Applikationen zur Analyse von Daten zu entwerfen. SAP HANA stellt dabei Natural Language Processing Methoden zur Verfügung. Mit SAP HANA ist die Analyse von verschiedenen Sprachen aus verschiedenen Quellen möglich. (vgl. SAP 2018d, S. 4.) Dabei werden Sprachen wie z. B. Deutsch, Englisch, Französisch und auch Spanisch unterstützt (vgl. SAP o. Jg).

SAP HANA stellt weiterhin mehrere Data Provisioning Adapter zur Verfügung. Diese können genutzt werden, um SAP HANA mit einer Datenquelle zu verknüpfen und Daten in SAP HANA zu überführen. Dabei sind Verbindungen zu anderen Datenbanken, wie z. B. einer DB2-Datenbank, aber auch zu sozialen Plattformen wie z. B. Twitter, möglich. (vgl. SAP o. Ja.)

Weiterhin gibt es die Möglichkeit SAP HANA Studio zu installieren. Dabei handelt es sich um eine Sammlung von Applikationen für SAP HANA. SSAP HANA Studio läuft unter Eclipse. Damit stehen dem Nutzer weitere Möglichkeiten zur Behandlung der Datensätze zur Verfügung. (vgl. SAP 2018b, S. 5.)

Twitter stellt APIs zur Verfügung (vgl. Russell 2011, S. 4). Diese lassen sich generell in zwei Komplexe untergliedern. Mit der sogenannte Search API ist es möglich, historische Daten zu erheben. Über die Streaming API ist es möglich Echtzeitdaten abzufragen. (vgl. Pfaffenberger 2016, S. 22.)

Für die Wahl von Twitter als Datenquelle sprechen die von Twitter bereitgestellten Metadaten. Hier wird neben dem eigentlichen Tweet z. B. die Sprache des Tweets zur Verfügung gestellt. Twitter bietet eine hohe Verfügbarkeit sowie eine standardisierte Struktur dieser Daten. Der Service wird aufgrund dieser Merkmale häufig als Datengrundlage für Forschung genutzt. (vgl. Pfaffenberger 2016, S. 15.)

16.3 Umsetzung

16.3.1 Vorgehensmodell

Das in diesem Kapitel vorgestellte Vorgehensmodell wurde im Rahmen der Masterthesis erarbeitet. Die Ausarbeitung des Modells wird in diesem Kapitel nicht näher erläutert, da der Hauptaspekt die Darstellung der Umsetzung einer Sentiment Analysis mit SAP HANA ist. Das Modell wurde durch Betrachtung der folgenden Werke abgeleitet und definiert (siehe Quellenverzeichnis): Beierle und Kern-Isberner 2014; Gronwald 2017; Jannaschk 2018; Müller und Lenz 2013; Nisbet et al. 2012; Nagarajan und Gandhi 2018;

Weiss et al. 2005; Pfaffenberger 2016; Evertz 2018; Dorschel et al. 2015; Zhai und Massung 2016; Liu 2011; Kaiser 2012; Ignatow und Mihalcea 2018; Grabs et al. 2017. Das verkürzte Vorgehensmodell lautet wie folgt:

1. Datensammlung
2. Datenverarbeitung
 - Sprachen aussortieren
 - Text in kleingeschriebene Buchstaben formatieren
 - Tokenization – inklusive folgender Tätigkeiten:
 - Satzzeichen und Sonderzeichen entfernen
 - Twitter-Konventionen „@“, „RT“ und „#“ behandeln
 - Stop Word Removal
 - Stemming
 - Classification
3. Datenanalyse
 - Ermittlung der Anzahl Sentiments je Sentimentklasse
 - Darstellung durch Tag Clouds

Die **Kleinschreibung** aller Begriffe führt dazu, dass mögliche Kombinationen von groß- und kleingeschriebenen Wörtern entfallen. Dies erleichtert das Abfragen der Daten. (vgl. Pfaffenberger 2016, S. 88.)

Tokenization bezeichnet den Vorgang des „Zerlegens“ eines Textdatenobjekts in einzelne,zählbare Einheiten (Tokens) (vgl. Zhai und Massung 2016, S. 149). Das Objekt wird dabei in sinnvolle Wörter, Phrasen und Symbole unterteilt (vgl. Nagarajan und Gandhi 2018, S. 4). Ein Beispiel: (vgl. Ignatow und Mihalcea 2018, S. 101.)

Ausgangsdatensatz: „Ich mag sehr gerne Elektroautos.“

Tokenisierte Datensätze: „Ich“, „mag“, „sehr“, „gerne“, „Elektroautos“

Twitter weist folgende Funktionsweisen bezüglich der Kommunikation auf dieser Plattform auf. So werden andere Nutzer, durch die Nutzung des „@“-Zeichens, in einem Tweet markiert. Weiterhin können sogenannte Retweets kommuniziert werden. Dabei wird der identische Wortlaut eines anderen Tweets zitiert. Solche Retweets sind durch „RT“ gekennzeichnet. Durch die Verwendung des „#“-Zeichens wird eine Zuordnung zu einem bestimmten Thema vorgenommen. (vgl. Grabs et al. 2017, S. 421 f.)

Stop Word Removing bezeichnet das Entfernen von Wörtern aus einem String, welche keine zu erwartende Bedeutung für die Analyse haben. Beispielhafte Stoppwörter können „ich“, „das“, „mein“, „es“ oder „von“ sein. Nur durch Angabe dieser Wörter könnte kein Rückschluss darauf gezogen werden, welchen inhaltlichen Bezug der Textstring hat. Solche Stoppwörter werden in der Sprache hochfrequent genutzt. Jedoch nicht, weil sie viele Informationen mit sich bringen, sondern da sie aus grammatischen Gründen benötigt werden. Durch die Entfernung dieser Wörter

erfolgt somit eine Reduzierung auf wenige bedeutsame Worte mit einer gleichzeitigen Verkleinerung des Datensatzes. (vgl. Zhai und Massung 2016, S. 66.)

In vielen Sprachen weisen die verschiedenen Wörter einige grammatischen Formen auf, je nach Kontext, in welchem sie verwendet werden. So haben Nomen z. B. Pluralformen und Verben verfügen über Zeitformen. Diese Variationen gehen alle auf eine Ausgangsform des Wortes zurück. Durch das Durchführen von **Stemming** werden die verschiedenen Wörter der Datenbasis zu ihren Wortstämmen überführt. Sind von einem Wort mehrere Formen vorhanden, wie z. B. „gut“ und „gutes“, so würden diese von einer Abfrage auf nur eines dieser Worte nicht ermittelt bzw. berücksichtigt werden. Aus diesem Grund ist das Durchführen von Stemming relevant. (vgl. Liu 2011, S. 228.)

Bei einer **Klassifikation** werden Objekte einem bekannten Klassifikationsschema zugeordnet (vgl. Müller und Lenz 2013, S. 111).

16.3.2 Implementierung

Die SAP HANA Web IDE verfügt maßgeblich über vier Tools. Zwei davon werden für diese Ausarbeitung benötigt. Der **Editor** stellt die Einheit dar, in welcher Repository Objekte z. B. erzeugt und geändert werden können. Im **Catalog** hingegen können die SQL-Artefakte verwaltet werden. (vgl. SAP 2018a, S. 7.)

Um auf die Datenquelle für die Sentiment Analysis dieser Arbeit zugreifen zu können, muss zunächst ein Twitter-Account erstellt werden. Wurde dies umgesetzt, kann auf Basis dieses Accounts eine Twitter App erstellt werden. Über diese werden u. a. App Keys und Access Tokens zur Verfügung gestellt (vgl. SAP o. Jf.), um die Anbindung an SAP HANA sowie den Abruf von Tweets zu ermöglichen (vgl. SAP o. Jf.).

Die in SAP HANA implementierten Text Analytics Funktionen stellen u. a. Funktionalitäten für die linguistische Analyse und die Faktenextraktion dar. Im Bereich der linguistischen Analyse sind dies Text Analyse-Funktionen wie z. B. Stemming. Die Faktenextraktion ermöglicht es Sentiments zu ermitteln. (vgl. SAP o. Jg.)

Die Software verfügt über verschiedene Sprachmodule, welche u. a. Dictionaries für die Textverarbeitung inkludieren. Diese Module nutzen die zuvor beschriebenen Funktionalitäten (linguistische Analyse und Faktenextraktion). Für die deutsche Sprache sind die linguistische Analyse sowie Sentiment Analysis möglich. Es ist weiterhin möglich die Dictionaries und Rules zu nutzen, um die Analyse auf die eigenen Gegebenheiten anzupassen. (vgl. SAP o. Jg.)

Bei der Durchführung einer Sentiment Analysis in HANA werden die einzelnen Stimmungen maßgeblich in fünf Kategorien eingeordnet. Diese lauten „Strong positive sentiment“, „Weak positive sentiment“, „Neutral sentiment“, „Weak negative sentiment“ und „Strong negative sentiment“. (vgl. SAP 2018d, S. 20.)

Im Rahmen dieser Arbeit wird mit der SAP HANA Web-based Development Workbench in der Version 1.00.122.12.1502962396 gearbeitet.

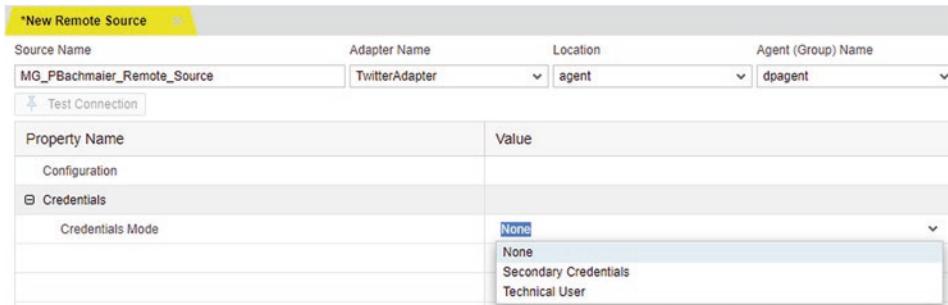


Abb. 16.1 Neue Quelle: Screenshot der SAP HANA Web IDE (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)

16.3.2.1 Datenakquise

Um die Datensammlung, also das Sammeln von Tweets, zu ermöglichen, muss in SAP HANA zunächst eine sogenannte Remote Source erstellt werden. Um eine solche Remote Source zu erstellen, muss im Catalog der Web IDE unter „Provisioning“, per Rechtsklick auf den Ordner „Remote Source“, „New Remote Source“ ausgewählt werden. Der Remote Source muss ein Name hinzugefügt werden. Für diese Arbeit lautet der Name „MG_PBachmaier_Remote_Source“. Unter „Adapter Name“ muss der „TwitterAdapter“ ausgewählt werden. Für den „Credentials Mode“ ist die Option „Technical User“ auszuwählen (s. Abb. 16.1). (vgl. SAP o. Jb.)

Durch Auswahl dieser Option erscheinen neue Felder, welche zur Eingabe des API Keys, API Secrets, Access Tokens und Access Token Secrets benötigt werden (vgl. SAP o. Jb). Diese können in der Twitter Application abgerufen, kopiert (vgl. SAP o. Jf) und in das jeweilige Feld eingefügt werden. Wurden diese Schritte durchgeführt, kann die Remote Source durch Klicken auf das Speicherdiskettensymbol erstellt werden (s. Abb. 16.2) (vgl. SAP o. Jb).

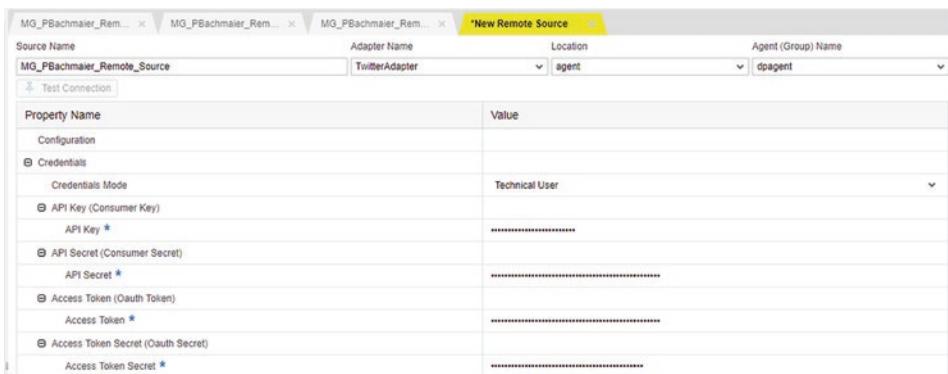
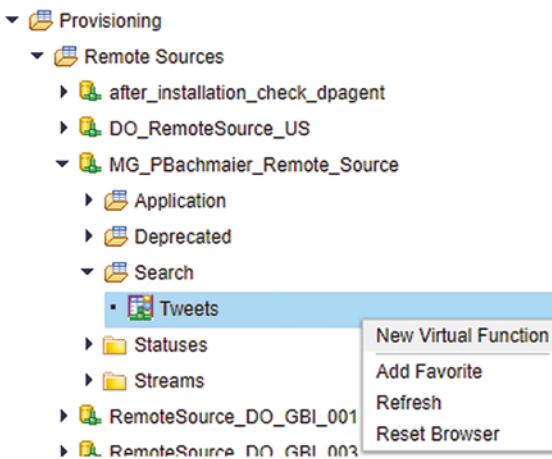


Abb. 16.2 Neue Quelle: Screenshot der SAP HANA Web IDE (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)

Abb. 16.3 Neue Quelle:
Screenshot der SAP HANA
Web IDE (Mit Genehmigung
der SAP SE verwendet.
© 2020. SAP SE)



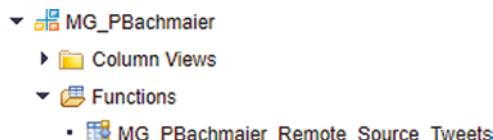
Diese Remote Source dient nun als Ausgangspunkt für das Sammeln der Tweets aus Twitter. Es existieren im Wesentlichen zwei Methoden, um Twitterdaten zu sammeln. Das Verwenden von **Virtual Tables** (Realtime) sowie der Einsatz von **Virtual Functions** (Batch). (vgl. SAP o. Jd.)

Im Rahmen dieses Kapitels wird die Nutzung von **Virtual Functions** beschrieben. Um diese Funktion zu nutzen, gilt es im Catalog unter „/Provisioning/Remote Sources/[Remote Source Name]/Search“ (der Remote Source Name lautet in diesem Fall „MG_PBachmaier_Remote_Source“) das Objekt „Tweets“, mit der rechten Maustaste anzu-klicken. Dann muss „New Virtual Function“ ausgewählt werden (s. Abb. 16.3). (vgl. SAP o. Jc.)

Es erscheint ein weiteres Fenster, in welchem ein Name für die virtuelle Funktion vergeben und das Schema definiert werden muss. Der Name wird auf „MG_PBachmaier_Remote_Source_Tweets“ festgelegt und das Schema „MG_PBachmaier“ wird ausgewählt. Durch Klicken auf „OK“ wird die Funktion erzeugt. Diese wird, wie sich zeigt, im Ordner „Functions“ des Schemas angelegt (s. Abb. 16.4). (vgl. SAP o. Jc.)

Da die Twitterdaten auch abgelegt werden müssen, erfolgt an dieser Stelle zunächst das Erzeugen einer Zieltabelle. Dies kann im Editor der SAP HANA Web IDE erfolgen (vgl. SAP 2018a, S. 7). Der notwendige Aufbau der Tabelle kann in der SAP Dokumentation eingesehen werden (vgl. SAP 2017, S. 276 f.).

Abb. 16.4 Neue Quelle:
Screenshot der SAP HANA
Web IDE (Mit Genehmigung
der SAP SE verwendet.
© 2020. SAP SE)



Anschließend erfolgt die Definition von Suchterminen. Da der inhaltliche Betrachtungsaspekt für die Sentiment Analysis „Elektroautos“ lautet und deutsche Tweets gesammelt werden sollen, werden u. a. folgende Terme definiert: Elektroauto, E-Automobil und Elektromobilität.

Jetzt gilt es, im Editor der Web IDE ein SQL-Statement zu erstellen. Dieses nutzt die soeben erzeugte Funktion zur Abfrage von Tweets über die Twitter API „Search Tweets“. (vgl. SAP o. Je.) Das verkürzte und somit beispielhafte Statement lautet:

```
UPSERT "MG_PBachmaier"."MG_PBachmaier_Tweets_Elektroauto" SELECT *
FROM "MG_PBachmaier"."MG_PBachmaier_Remote_.
Source_Tweets"('Elektroauto OR E-Automobil OR', '100.000', null, null,
null, null, null, null);
```

Durch *UPSERT "MG_PBachmaier"."MG_PBachmaier_Tweets_Elektroauto"* wird ein Update und Insert der Tabelle bewirkt. Mit *SELECT * FROM "MG_PBachmaier"."MG_PBachmaier_Remote_Source_Tweets"* wird über die virtuelle Funktion ein Select vorgenommen. Die in Klammern angegebenen Parameter sind für die Twitter Search API relevant. Es wird mit den Suchterminen begonnen „*Elektroauto OR E-Automobil [...]J*“. Darauf folgt die Zahl *'1500'*. Diese legt fest, wie viele Tweets maximal zurückgegeben werden sollen. Die weiteren Parameter werden auf „Null“ gesetzt, da diese in diesem Fall nicht relevant sind. Hier kann z. B. auf einen „Geocode“ oder auf ein „until“-Datum eingegrenzt werden. (vgl. SAP o. Je.)

Da bei diesem Verfahren immer nur eine einmalige Abfrage von Tweets erfolgt, muss eine regelmäßige Ausführung des Statements getätigten werden. Innerhalb der Dauer von etwa zwei Monaten wurde dieser Vorgang kontinuierlich wiederholt und dokumentiert. Es wurden in 45 Sammelvorgängen insgesamt 32.518 Tweets erhoben. Dies entspricht einer durchschnittlichen täglichen Anzahl von etwa 723 Tweets.

16.3.2.2 Datenverarbeitung

Beim Betrachten der Datensätze (siehe Abb. 16.5) fällt auf, dass trotz der sorgsamen Auswahl von deutschen Suchterminen Tweets in verschiedenen Sprachen vorhanden sind. Dies belegt die Notwendigkeit des Überprüfens und Aussortierens anderer Sprachen (vgl. Pfaffenberger 2016, S. 87 f.). Ebenso fällt eine gewisse Häufigkeit der durch „RT“ (vgl. Grabs et al. 2017, S. 422) gekennzeichneten Retweets auf. Wie bereits im

ID	SCREENNAME	TWEET
		2017 waren 126,4 Millionen Fahrgäste mit den Straßenbahnen der elektrisch mobil. #Elektromobilität

Abb. 16.5 Tabelleninhalt Quelle: Screenshot der SAP HANA Web IDE (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)

definierten Vorgehensmodell beschrieben wurde, muss der Umgang mit diesen Tweets festgelegt werden.

Gemäß dem geplanten Vorgehen gilt es nun zunächst Tweets in von Deutsch abweichenden Sprachen zu ermitteln und aus dem Datensatz zu entfernen. Es wird im Arbeitsverzeichnis, im Editor der Web IDE, gearbeitet. Um die verschiedenen Sprachen abzufragen, wird auf das Attribut „ISOLANGUAGECODE“ der Tabelle zugegriffen. Hierdurch werden alle „ISOLANGUAGECODES“ selektiert, die nicht dem Wert „de“ entsprechen. Es handelt sich dabei um ein Attribut, welches von Twitter mitgeliefert wird (vgl. SAP o. Jh) und den Code der jeweiligen Sprache beinhaltet. Es sind insgesamt 34 andere Sprachen im Datensatz enthalten. Durch ein weiteres SQL-Statement wird die Anzahl aller deutschen Tweets abgefragt. Es ergibt sich, dass insgesamt 26.455 deutsche Datensätze in den 32.518 Tweets vorhanden waren.

Nun wird zunächst eine weitere Tabelle erstellt, welche nur die in Deutsch verfassten Tweets enthalten wird. Dazu wird erneut ein CREATE-Statement, zum Erzeugen einer Tabelle, ausgeführt.

Um die in Deutsch verfassten Tweets zu ermitteln und diese in die neu erzeugte Tabelle einzufügen, wird erneut das Attribut „ISOLANGUAGECODE“ der Tabelle genutzt und auf den Wert „de“ abgefragt. Die Daten werden über ein INSERT-Statement in die neue Tabelle überführt. Durch Ausführung eines SELECT-Statements wird überprüft, dass alle 26.455 Sätze überführt werden.

Der nächste Schritt des Vorgehens ist das Formatieren der Tweets in ausschließlich kleingeschriebene Buchstaben. Analog zur Ermittlung der deutschen Tweets wird erneut eine neue Tabelle erzeugt. Es wird das zuvor genutzte CREATE-Statement ausgeführt. Nun werden, wie bereits zuvor, die deutschen Tweets in die neue Tabelle überführt. Daraufhin müssen sämtliche Tweets in der neuen Tabelle in kleingeschriebene Buchstaben transformiert werden. Dazu wird die „LOWER“-Funktion genutzt.

Nun muss das Behandeln von Retweets erfolgen. Diese zeichnen sich dadurch aus, dass der Datenstring mit „RT @“ beginnt (vgl. Grabs et al. 2017, S. 422). Zunächst erfolgt die Ermittlung der Anzahl dieser Tweets, durch folgendes SQL-Statement:

```
SELECT COUNT(*) AS COUNT.  
FROM "MG_PBachmaier"."MG_PBachmaier_Tweets_".  
Elektroauto_Deutsch" WHERE "TWEET" LIKE 'RT @%';
```

Diese Abfrage wird auf die Tabelle ausgeführt, welche die deutschen Tweets enthält. Es wurden 11.696 Sätze ermittelt. Dabei handelt es sich mit rund 44 % der Gesamtdatenbasis von 26.455 Sätzen um einen recht großen Anteil der gesamten Daten. In dieser Arbeit wird keine Entfernung der Retweets vorgenommen. Die Datenbasis würde durch Entfernung dieser Retweets drastisch verkleinert werden.

Die Daten sind nun soweit vorbereitet, dass die übrigen konzipierten Schritte zur Durchführung einer Sentiment Analysis ausgeführt werden können. Um die Schritte Tokenization, Stemming und Classification durchzuführen, muss in SAP HANA ein

Index für den String, welcher den Tweet enthält, erzeugt werden. Hierbei muss die Konfiguration von SAP HANA, „EXTRACTION_CORE_“

VOICEOFCUSTOMER“, referenziert werden. Wie sich bei der Ausführung des Statements zeigt, wird hierbei auch Stop Word Removal umgesetzt. Das SQL-Statement lautet: (vgl. SAP 2018c, S. 4 ff.)

```
CREATE FULLTEXT INDEX MG_PBACHMAIER_INDEX_TWEETS_
ELEKTROAUTO_DEUTSCH_TO_LOWER ON
"MG_PBachmaier"."MG_PBachmaier_Tweets_Elektroauto_Deutsch_To_Lower"
("TWEET")
CONFIGURATION 'EXTRACTION_CORE_VOICEOFCUSTOMER'
LANGUAGE DETECTION ('EN', 'DE')
TEXT ANALYSIS ON;
```

Durch dieses Statement wird festgelegt, dass der Name der Index-Tabelle „MG_PBachmaier“. „MG_PBachmaier_Tweets_Elektroauto_Deutsch_To_Lower“ lauten soll. Durch Angabe des Attributs „TWEET“ wird angegeben, dass der Index auf die Werte dieses Attributs zu erzeugen ist. Durch „CONFIGURATION 'EXTRACTION_CORE_VOICEOFCUSTOMER'“ wird definiert, dass die Konfiguration „Voice of Customer“ angewendet werden soll. (vgl. SAP 2018c, S. 4 f.)

SAP HANA hat aus den 26.455 Ursprungssätzen insgesamt 145.619 neue Sätze erzeugt. Dies wird durch Anwendung eines COUNT-Statements ermittelt. Durch die Indexierung der Tweets wurden durch HANA beispielhaft folgende Datensätze erzeugt (s. Abb. 16.6):

In der Spalte „TA_TOKEN“ werden die einzelnen extrahierten Tokens aufgeführt. Der „TA_COUNTER“ stellt die Nummerierung der Tokens, aus dem Ursprungstweet, dar. Bei diesem Tweet wurden elf Tokens erzeugt. Die Spalte „TA_LANGUAGE“ stellt die Sprache dar, in welcher der Tweet verfasst wurde und „TA_TYPE“ zeigt die einzelnen Klassen auf, in welche die Tokens einsortiert wurden. (vgl. SAP 2018c, S. 6.) Es ist zu erkennen, dass dieser Tweet zwei schwach positive Sentiments enthält, welche „gut“

ID	TA_RULE	TA_COUNTER	TA_TOKEN	TA_LANGUAGE	TA_TYPE
	Entity Extraction	1		de	SOCIAL_MEDIA/ID_TWITTER
	Entity Extraction	10	#sonneimtank	de	Topic
	Entity Extraction	11	#sonneimtank	de	SOCIAL_MEDIA/TOPIC_TWITTER
	Entity Extraction	2	elektroauto abends	de	NOUN_GROUP
	Entity Extraction	3	schönes pilotprojekt	de	Sentiment
	Entity Extraction	4	schönes	de	WeakPositiveSentiment
	Entity Extraction	5	pilotprojekt	de	Topic
	Entity Extraction	6		de	SOCIAL_MEDIA/ID_TWITTER
	Entity Extraction	7		de	SOCIAL_MEDIA/ID_TWITTER
	Entity Extraction	8	gut zu #sonneimtank	de	Sentiment
	Entity Extraction	9	gut	de	WeakPositiveSentiment

Abb. 16.6 Tabelleninhalt Quelle: Screenshot der SAP HANA Web IDE (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)

und „schönes“ lauten. Ebenso werden Sätze, welche als „Sentiment“ klassifiziert sind, aufgeführt. In diesen wird auch der Kontext der Sentiments dargestellt („schönes Pilotprojekt“). Da sie jedoch nicht im Sinne einer Gewichtung klassifiziert sind (positiv oder negativ), werden diese in der Arbeit nicht weiter berücksichtigt. Zudem werden Stoppwörter, wie z. B. „mit“ und „zu“, entfernt. Es zeigt sich jedoch auch, dass das Stemming nicht einwandfrei funktioniert. Es wäre zu erwarten, dass aus dem Token „schönes“ der Stamm „schön“ gebildet wird. Dies ist jedoch nicht der Fall und muss bei der weiteren Bearbeitung berücksichtigt werden. Wird nicht korrekt gestemmt, so wird dies dazu führen, dass Tokens mit der gleichen Bedeutung (z. B. „schönes“ und „schön“) einzeln aufgeführt werden, obwohl es ausreichend wäre, das Token „schön“ zweimal zu zählen. Dies wird durch weitere SQL-Abfragen bestätigt und führt dazu, dass das Stemming manuell angepasst werden muss.

Weiterhin werden die ermittelten Sentiments durch folgendes Statement abgefragt:

```
SELECT "TA_TOKEN", COUNT(*) COUNT
FROM "MG_PBachmaier"."$TA_MG_PBACHMAIER_INDEX_
TWEETS_ELEKTROAUTO_DEUTSCH_TO_LOWER_ANPASSUNG"
WHERE "TA_TYPE" LIKE 'WeakPositiveSentiment'
GROUP BY "TA_TOKEN"
ORDER BY COUNT DESC;
```

Das Statement bewirkt, dass alle Tokens, welche als schwach positives Sentiment klassifiziert wurden, mit der Anzahl der Nennungen aufgeführt werden. Folgende Tab. 16.1 zeigt die ersten vier Ergebnissätze:

Es zeigt sich, dass auch die Klassifizierungen nicht einwandfrei durchgeführt werden. So repräsentiert „richtige“ keine direkte Stimmung. Es muss somit auch eine manuelle Bereinigung der Klassifizierungen vorgenommen werden.

Nun kann einer von zwei Ansätzen verfolgt werden: eine manuelle Bereinigung der Klassifikationen und des Stemms vor der Index gebildet wird (Anpassung der Dictionaries und Rules (vgl. SAP 2018c, S. 20 f.)), oder die nachträgliche Bereinigung der Indextabelle. Es wird zunächst mit der Betrachtung der proaktiven Möglichkeiten begonnen.

Tab. 16.1 Abfrageergebnisse schwach positiver Sentiments. (Quelle: Basierend auf einem Screenshot der Abfrageergebnisse der SAP HANA Web IDE)

TA_TOKEN	COUNT
richtige	924
gute	178
saubere	155
sauberes	117

Es ist möglich eigene Dictionaries zu definieren und diese in der Configuration zu hinterlegen. In einem Dictionary können Entitätstypen und -namen definiert werden. Ein solches Dictionary wird im XML-Format verfasst und das File muss auf den Suffix „.hdbtextdict“ enden. Wenn das Dictionary mit einem Landesnamen, gefolgt von einem Bindestrich, beginnt (z. B. „german-dictionary“), so wird dieses Dictionary nur für diese Sprache verwendet (anstatt für alle Inputdaten). (vgl. SAP 2018c, S. 20.)

Selbiges gilt auch für Rules. Es können eigene Regeln definiert werden. Diese eignen sich z. B. dafür, komplexere Entitätstypen zu identifizieren. Dafür muss ein File erzeugt werden, das auf „.hdbextrule“ endet. Auch hier gilt, sollte das File mit einem Landesnamen, gefolgt von einem Bindestrich, beginnen (also z. B. „german-rule“), so wird die Regel nur für die jeweilige Sprache angewendet anstatt für sämtlichen Input. (vgl. SAP 2018c, S. 20 f.)

Der Versuch der Bereinigung der Klassifikation und des Stemming vor der Indexbildung führte nicht zum gewünschten Ergebnis. Somit wird mit der manuellen Bereinigung der Tokens fortgefahren.

Zunächst wird eine neue Tabelle erstellt, welche zu Beginn einen identischen Abzug der Datensätze aus der unbehandelten Index-Tabelle enthält. Dazu wird ein CREATE-Statement entwickelt und ausgeführt, welches aus dem Aufbau der Indextabelle aus SAP HANA abgeleitet ist. Die Tabelle hat somit den gleichen Aufbau wie die initiale Index-tabelle in SAP HANA.

Daraufhin erfolgt die Überführung aller Datensätze aus der ursprünglichen Index-Tabelle durch ein INSERT-Statement. Nun kann die Anpassung der Sätze erfolgen. Als erstes wird ein Statement erstellt, welches dazu dient, die nicht korrekt gestemmten Sätze anzupassen. Um dafür alle Sentiments zu ermitteln und zu überprüfen, welche Tokens angepasst werden müssen, wird für jede Sentiment-Klasse eine Abfrage, wie bereits zuvor beschrieben, ausgeführt.

Lediglich die WHERE-Condition muss zur Abfrage jeder weiteren Klasse angepasst werden. Es wird nacheinander ‚StrongPositiveSentiment‘, ‚WeakNegativeSentiment‘, ‚StrongNegativeSentiment‘ und ‚NeutralSentiment‘ für ‚WeakPositiveSentiment‘ eingesetzt. Als Ergebnis wird jeweils eine Liste der Sentiment-Tokens der Klasse, mit Anzahl der Nennungen, aufgeführt. Basierend auf diesen Listen werden Statements zur Anpassung der Stems erstellt. Dieses enthält insgesamt 203 Anpassungen von Stems, die exemplarisch wie folgt aufgebaut sind:

```
UPDATE "MG_PBachmaier"."$TA_MG_PBACHMAIER_INDEX_
TWEETS_ELEKTROAUTO_DEUTSCH_TO_LOWER_ANPASSUNG" SET
"TA_TOKEN" = 'abartig' WHERE "TA_TOKEN" LIKE 'abartig%';
```

Durch Ausführung dieses Statements wird bewirkt, dass alle Variationen des Wortes „abartig“ (z. B. „abartiges“) zum Wortstamm „abartig“ transformiert werden. Weitere Beispiele sind alle Variationen des Wortes „zuverlässig“ (z. B. „zuverlässiger“), welche zu „zuverlässig“ gestemmt werden. Durch Tests wird deutlich, dass dabei mit Bedacht

vorgegangen werden muss, da z. B. Wörter wie „mager“ versehentlich zu einem positiven Sentiment gestemmt werden könnten („mag“).

Hierbei ist zu beachten, dass es sich nicht um eine allumfassende Stem-Liste handelt, welche auf jeden beliebigen Datensatz anwendbar wäre. Es handelt sich lediglich um eine passende Stem-Liste für den in dieser Arbeit behandelten Datensatz, welche noch immer nicht abschließend ist, wie sich in Abschn. 16.3.2.3 zeigt. Die Erstellung von umfassenden Dictionaries nimmt viel Zeit in Anspruch, und es benötigt Experten wie z. B. Linguisten (vgl. Ignatow und Mihalcea 2018, S. 85).

Als nächstes erfolgt die Erstellung und Ausführung eines Statements zur Anpassung der Klassifikationen der Tokens. Ein beispielhafter Auszug dieses Statements sieht wie folgt aus:

```
UPDATE "MG_PBachmaier"."$TA_MG_PBACHMAIER_INDEX_
TWEETS_ELEKTROAUTO_DEUTSCH_TO_LOWER_ANPASSUNG"
SET "TA_TYPE" = 'NOT_A_SENTIMENT' WHERE "TA_TOKEN" LIKE 'absehbar%';
```

Durch dieses Statement wird bewirkt, dass Sätze des Wortes „absehbar“ als „NOT_A_SENTIMENT“ klassifiziert werden. Der Wert für „TA_TYPE“ wird für die einzelnen Klassen angepasst (zu „WeakPositiveSentiment“, etc.) und Stems, welche in den Abfrageergebnissen der Sentimentklassen enthalten sind, und falsch klassifiziert sind, werden einer anderen Klasse zugeordnet. Insgesamt werden 26 Tokens der Klasse „NOT_A_SENTIMENT“ zugeordnet, 29 der Klasse „WeakPositiveSentiment“, 16 der Klasse „WeakNegativeSentiment“, 12 der Klasse „StrongNegativeSentiment“ und ein Token der Klasse „NeutralSentiment“.

16.3.2.3 Datenanalyse

Der Datensatz ist nun so weit vorbereitet, dass die Datenanalyse und -interpretation durchgeführt werden kann. Alle Sentiments sind einheitlich klassifiziert und gestemmt.

Zunächst erfolgt die Ermittlung der Gesamtanzahl verfügbarer Sentiment-Tokens, ohne eine Betrachtung der einzelnen Klassen. Es sind insgesamt 6392 Sentiment-Tokens vorhanden. Dazu erfolgt die Ausführung folgenden Statements:

```
SELECT SUM("COUNT") AS SUM
FROM (SELECT COUNT("TA_TOKEN") AS COUNT
FROM "MG_PBachmaier"."$TA_MG_PBACHMAIER_INDEX_
TWEETS_ELEKTROAUTO_DEUTSCH_TO_LOWER_ANPASSUNG"
WHERE "TA_TYPE" LIKE '%Sentiment%'.
AND "TA_TYPE" NOT LIKE 'Sentiment'.
GROUP BY "TA_TOKEN");
```

Werden die Sentiments nun in ihre Klassen aufgeschlüsselt betrachtet, so kann die Anzahl der Sentiments je Klasse mit folgendem Statement abgerufen werden:

Tab. 16.2 Anzahl Tokens je Sentimentklasse.
Quelle: Basierend auf einem Screenshot der Abfrageergebnisse der SAP HANA Web IDE

TA_TYPE	COUNT
WeakPositiveSentiment	3979
WeakNegativeSentiment	1056
StrongPositiveSentiment	796
StrongNegativeSentiment	385
NeutralSentiment	176

```
SELECT "TA_TYPE", COUNT(*) AS COUNT.
FROM "MG_PBachmaier"."$TA_MG_PBACHMAIER_INDEX_.
TWEETS_ELEKTROAUTO_DEUTSCH_TO_LOWER_ANPASSUNG".
WHERE "TA_TYPE" LIKE '%Sentiment%' AND "TA_TYPE" NOT LIKE 'Sentiment'
GROUP BY "TA_TYPE" ORDER BY COUNT DESC;
```

In Tab. 16.2 ist zu sehen, dass mit Abstand am meisten Sentiments in der Klasse der schwach positiven Stimmungen vorhanden sind. Es sind rund 4000. Werden die positiven (schwach und stark) Sentiments sowie die negativen (schwach und stark) jeweils addiert, so ergibt sich, dass die positiven Sentiments mit 4775 Tokens (~75 %) und die negativen mit 1441 Tokens (~22 %) vertreten sind. Die neutralen Sentiments stellen nur einen kleinen Teil von 176 Tokens (~3 %) dar. Die positive Tendenz ist deutlich ersichtlich.

Um nun Tag Clouds der einzelnen Sentimentklassen zu erstellen, wird mit SAP HANA Studio gearbeitet. Dies gilt es zunächst zu installieren. Die auf diese Weise erstellbaren Tag Clouds werden in den Abb. 16.7 und 16.8 dargestellt. Je größer ein Wort dargestellt wird, desto höher ist die Anzahl der Nennungen dieses Wortes (vgl. Kaiser 2012, S. 32).

Abb. 16.7 zeigt die als schwach positiv klassifizierten Tokens.



Abb. 16.7 Tag Quelle: Screenshot von SAP HANA Studio (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)



Abb. 16.8 Tag Quelle: Screenshot von SAP HANA Studio (Mit Genehmigung der SAP SE verwendet. © 2020. SAP SE)

Abb. 16.8 stellt die schwach negativen Sentiments dar. Wie im vorangegangenen Kapitel angekündigt zeigt sich, dass noch immer nicht alle Tokens korrekt klassifiziert und gestemmt sind. In der Abbildung wird z. B. auch der Begriff „bequemer“ dargestellt. Dieser ist augenscheinlich nicht in der Klasse der schwach negativen Sentiments anzusiedeln.

16.4 Fazit

In diesem Kapitel wurde aufgezeigt, wie eine Sentiment Analysis mit SAP HANA auf Basis von deutschen Textdaten umsetzbar ist. Anhand des Vorgehensmodells wurde dargestellt, welche Schritte notwendig sind und wie diese in SAP HANA umgesetzt werden können, um Stimmungen aus Tweets zu extrahieren.

Sofern noch keine Datenbasis besteht hat zunächst eine Datensammlung zu erfolgen. Hierzu wurde in der Arbeit auf Twitterdaten zurückgegriffen. Daraufhin wurde die Datenverarbeitung umgesetzt. In diesem konkreten Fall musste die Kontrolle und das Aussortieren verschiedener Sprachen erfolgen, welche in der Datenbasis enthalten waren. Daraufhin galt es den Text in kleingeschrieben Buchstaben zu formatieren. Im Anschluss daran wurden Stop Word Removal, Stemming und Classification umgesetzt.

Die Sentiment Analysis konnte erfolgreich umgesetzt werden. Es wurden Sentiments in der Datenbasis ermittelt. Weiterhin konnte dargestellt werden, dass etwa 75 % der ermittelten Sentiments bzgl. des Themas Elektroautos positiver Natur waren. Die in einzelne Klassen eingeteilten Sentiments wurden in Tag Clouds dargestellt.

Literatur

- Beierle, C., Kern-Isbner, G.: Methoden wissensbasierter Systeme Grundlagen Algorithmen Anwendungen, 5. Aufl. Vieweg+ Teubner, Wiesbaden (2014)
- Böck, M., Köbler, F., Anderl, E., Le, L.: Social-Media-Analyse - mehr als nur eine Wordcloud. Springer, Wiesbaden (2017)
- Dorschel, J., Dorschel, W., Föhl, U., van Geenen, W., Hertweck, D., Kinitzki, M., Küller, P., Lanquillon, C., Mallow, H., März, L., Omri, F., Schacht, S., Stremler, A., Theobald, E.: Wirtschaft. In: Dorschel, J. (Hrsg.) Praxishandbuch Big Data Wirtschaft Recht Technik, S. 15–166. Springer Gabler, Wiesbaden (2015)
- Evertz, S.: Analysiere das Web! Wie Sie Marketing und Kommunikation mit Social Media Monitoring verbessern. Haufe-Lexware, Freiburg (2018)
- Grabs, A., Vogl, E., Bannour, K.-P.: Follow me! Erfolgreiches Social Media Marketing mit Facebook Twitter und Co, 4. Aufl. Rheinwerk, Bonn (2017)
- Gronwald, K.-D.: Integrierte Business-Informationssysteme ERP SCM CRM BI Big Data Analytics - Prozesssimulation Rollenspiel Serious Gaming, 2. Aufl. Springer, Berlin (2017)
- Ignatow, G., Mihalcea, R.: An introduction to text mining Research design data collection and analysis. Sage Publications, Thousand Oaks (2018)
- Jannaschk, K.: Infrastruktur für ein Data Mining Design Framework Eine Untersuchung mit Fallbeispielen. Wiesbaden (2018)
- Kaiser, C.: Business Intelligence 2.0 Knowledge-Based Services zur automatisierten Analyse der Meinungsbildung im Web 2.0, Diss. Universität Erlangen-Nürnberg. Springer, Wiesbaden (2012)
- Kayser, S., Rath, H.R.: Marktforschung 2.0 - Authentische Meinungen in Echtzeit erschließen. In: Keller, B., Klein, H.-W., Tuschel, S. (Hrsg.) Zukunft der Marktforschung Entwicklungschancen in Zeiten von Social Media und Big Data, S. 121–134. Springer Gabler, Wiesbaden (2015)
- Liu, B.: Web Data Mining Exploring Hyperlinks Contents and Usage Data, 2. Aufl. Springer, Heidelberg (2011)
- Müller, R.M., Lenz, H.-J.: Business Intelligence. Springer, Berlin (2013)
- Nagarajan, S.M., Gandhi, U.D.: Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Comput. Appl.* **50**(4), 109 (2018)
- Nisbet, R., Fast, A., Waner, A., Delen, D., Miner, G., Thompson, J., Elder, J., Balakrishnan, K., Winters-Miner, L. A., Foley, R., Hill, T.: Practical text mining and statistical analysis for non-structured text data applications. Waltham (2012)
- Pfaffenberger, F.: Twitter als Basis wissenschaftlicher Studien Eine Bewertung gängiger Erhebungs- und Analysemethoden der Twitter-Forschung. Springer, Wiesbaden (2016)
- Provost, F., Fawcett, T.: Data science for business What you need to know about data mining and data-analytic thinking. O'Reilly Media, Sebastopol (2013)
- Russell, M.A.: Mining the social web. O'Reilly Media, Sebastopol (2011)
- SAP: SAP HANA Modeling Guide, https://help.sap.com/doc/aa153c4bd28c47f69e642e43a4b35a46/1.0_SP12/en-US/SAP_HANA_EIM_Configuration_Guide_en.pdf (2017). Zugegriffen: 29. Juli 2020

- SAP.: SAP HANA Developer Guide For SAP HANA Web Workbench, https://help.sap.com/doc/7d957b1b6c1a4791a1212b45e9a797df/1.0.12/en-US/SAP_HANA_Developer_Guide_for_SAP_HANA_Web_Workbench_en.pdf (2018a). Zugegriffen: 31. Aug. 2018
- SAP.: SAP HANA Studio Installation and Update Guide Linux Microsoft Windows Mac OS, https://help.sap.com/doc/e3abfad2b0fd1014944295e9acd75d79/1.0.12/en-US/SAP_HANA_Studio_Installation_Update_Guide_en.pdf (2018b). Zugegriffen: 31. Aug. 2018
- SAP.: SAP HANA Text Analysis Developer Guide, https://help.sap.com/doc/916350b4473746f6bd2d7ee7fbf9b8f3/1.0.12/en-US/SAP_HANA_Text_Analysis_Developer_Guide_en.pdf (2018c). Zugegriffen: 31. Aug. 2018
- SAP.: SAP HANA Text Analysis Extraction Customization Guide, https://help.sap.com/doc/ca7939955fbc425293bbfa36a64fc37/1.0.12/en-US/SAP_HANA_Text_Analysis_Extraction_Customization_Guide_en.pdf (2018d). Zugegriffen: 31. Aug. 2018
- SAP.: Configure Data Provisioning Adapters, [https://help.sap.com/viewer/7952ef28a6914997abc01745fe1745fef1b607/1.0_SPSS12/en-US/6ed502701abd4d1ca94d463d7dc6e99f.html?q=Modeling%20\(o.Ja.\)](https://help.sap.com/viewer/7952ef28a6914997abc01745fe1745fef1b607/1.0_SPSS12/en-US/6ed502701abd4d1ca94d463d7dc6e99f.html?q=Modeling%20(o.Ja.)). Zugegriffen: 01. Sept. 2018
- SAP.: Create a Twitter Remote Source, https://help.sap.com/viewer/7952ef28a6914997abc01745fe1fb607/1.0_SPSS12/en-US/4f7e061e34024d06bb2ea669beaa50f4.html (o. Jb). Zugegriffen: 01. Sept. 2018
- SAP.: Create a Virtual Function, [https://help.sap.com/viewer/71c4a6e6b4dc4a5ab3e17bb1d7e98104/1.0_SPSS12/en-US/4e08a9193967455f93dc2bbd9f588f66.html?q=virtual%20%20\(o.Jc.\)](https://help.sap.com/viewer/71c4a6e6b4dc4a5ab3e17bb1d7e98104/1.0_SPSS12/en-US/4e08a9193967455f93dc2bbd9f588f66.html?q=virtual%20%20(o.Jc.)). Zugegriffen: 02. Sept. 2018
- SAP.: Creating Twitter Virtual Tables and Functions, https://help.sap.com/viewer/71c4a6e6b4dc4a5ab3e17bb1d7e98104/1.0_SPSS12/en-US/f109ece31cf645b0917e38ceb1f8ab9a.html (o. Jd). Zugegriffen: 02. Sept. 2018
- SAP.: Search: Tweets, https://help.sap.com/viewer/71c4a6e6b4dc4a5ab3e17bb1d7e98104/1.0_SPSS12/en-US/c49d70a2fdd64230ae16a6d32ab3fb2.html (o. Je). Zugegriffen: 02. Sept. 2018
- SAP.: Set Up a Twitter Account, https://help.sap.com/viewer/7952ef28a6914997abc01745fe1fb607/1.0_SPSS12/en-US/e2f10649e50f435daa50c6fe398c1aaa.html (o. Jf). Zugegriffen: 01. Sept. 2018
- SAP.: Text Analysis Language Reference Guide Overview of Text Analysis Functionality, <https://help.sap.com/viewer/2eddaf05dd984ce6b1d389a3bd042d11/Cloud/en-US/57b64bb6d6d1014b3fc9283b0e91070.html> (o. Jg). Zugegriffen: 29. Juli 2020
- SAP.: Twitter, https://help.sap.com/viewer/71c4a6e6b4dc4a5ab3e17bb1d7e98104/1.0_SPSS12/en-US/a172971916c84756aace0b046807dc32.html?q=twitter%20 (o. Jh). Zugegriffen: 02. Sept. 2018
- SAP.: Twitter Remote Source Configuration, https://help.sap.com/viewer/7952ef28a6914997abc01745fe1fb607/1.0_SPSS12/en-US/5713d798c11f4b32929ae7702ceba8f1.html (o. Ji). Zugegriffen: 01. Sept. 2018
- Weiss, S.M., Damerau, F.J., Indurkhy, N., Zhang, T.: Text Mining Predictive Methods for Analyzing Unstructured Information. Springer, NY (2005)
- Zhai, C., Massung, S.: Text data management and analysis A practical introduction to information retrieval and text mining. Acm, NY (2016)



Weiterbildung in Data Science

17

Christoph Quix

Zusammenfassung

In den letzten Jahren wurden von verschiedenen Anbietern Lehr- und Weiterbildungskurse im Bereich Data Science entwickelt, die unterschiedliche Zielgruppen ansprechen sollen. Von Hochschulen und Universitäten wurden vor allem Master-Programme entwickelt, die von Studierenden besucht werden können, die schon einen Abschluss mit Informatik-Bezug erreicht haben. Aufgrund der fort schreitenden Digitalisierung gibt es zunehmend Angebote zur beruflichen Weiterbildungen in der Digitalisierung, aber insbesondere auch für Data Science. In immer mehr Fachgebieten sind zumindest Grundlagenkenntnisse der Data Science erforderlich, da Daten erfasst, verwaltet und analysiert werden müssen. Daher werden Weiterbildungen für Data Science allgemein oder für Data Science in bestimmten Anwendungsbereichen verstärkt nachgefragt.

In diesem Kapitel geben wir einen Überblick über verschiedene Weiterbildungsangebote, ausgehend von Kompetenzfeldern, die für Data Science definiert wurden. Basierend auf den Kompetenzfeldern ist ein Vergleich der Studiengänge und Weiterbildungsangebote möglich.

C. Quix (✉)

FB Elektrotechnik/Informatik, Hochschule Niederrhein, Krefeld, Deutschland
E-Mail: christoph.quix@hs-niederrhein.de

17.1 Kompetenz-Rahmenwerke für Data Science

Data Science ist ein sehr umfangreiches und vielseitiges Feld, wie die auch die Themenvielfalt in diesem Buch zeigt. In verschiedenen Publikationen (z. B. Kelleher und Tierney 2018; Costa und Santas 2017) wird dargestellt, dass ein Data Scientist nicht nur Daten-Management- und -Analyse-Kompetenzen besitzen sollte, sondern auch Domänenwissen aus dem Anwendungsbereich, Fähigkeiten zur Visualisierung und Präsentation der Ergebnisse oder auch rechtliche Kenntnisse zum Thema Datenschutz mitbringen sollte. Selbst auf der technischen Ebene sind die Anforderungen an Data-Science-Experten enorm vielfältig. Sie müssen nicht nur die Daten aufbereiten und in effizienten Systemen zur Verfügung stellen (siehe Kapitel Big-Data-Technologien und In-Memory-Datenbanksysteme), sondern auch die verschiedenen Datenanalyse-Algorithmen kennen und sie auf einer verteilten, skalierbaren Big-Data-Plattform so einsetzen können, dass die Ergebnisse in der gewünschten Zeit zur Verfügung stehen. Ist es offensichtlich, dass dieses Kompetenzspektrum nicht in einer Person vereinigt werden kann. Daher sollte ein Data Scientist zwar ein breites Grundlagenwissen haben, aber er muss vor allem ein Team Player sein, damit die notwendigen Detailkenntnisse auf verschiedene Spezialisten im Team verteilt werden können.

Diese Vielfalt stellt auch eine Herausforderung für die Ausbildung von Data Scientists dar, da man schwer alle Aspekte in einem Weiterbildungskurs oder Studienprogramm abdecken kann. Bevor wir uns einige Studien- bzw. Weiterbildungsangebote anschauen, gehen wir aber zunächst auf die Kompetenzfelder genauer ein.

Allgemeine Rahmenwerke für Kompetenzen in der Informations- und Kommunikationstechnologie (IKT), wie SFIA (Skills Framework for the Information Age, <https://sfia-online.org/>) oder das European e-Competence Framework (e-CF, <https://www.ecompetences.eu>), geben nur eine sehr grobe Beschreibung von Kompetenzfeldern. Beide Rahmenwerke werden auf internationaler Ebene entwickelt (SFIA weltweit, e-CF in Europa). e-CF wurde sogar mittlerweile als europäische bzw. deutsche Norm (DIN EN 16.234-1:2016-12) umgesetzt. Zwar lassen sich in diese Rahmenwerke auch Data Scientists einordnen, jedoch finden sich Datenkompetenzen explizit nur an wenigen Stellen wieder. Das e-CF ist beispielsweise in fünf Bereiche „Plan“, „Build“, „Run“, „Enable“ und „Manage“ unterteilt, in die 40 Kompetenzen eingruppiert werden. Data Science wird nicht explizit als Kompetenz erwähnt, einen Bezug zu Data Science findet man in den Kompetenzen „Application Design“, „Systems Engineering“, „Information and Knowledge Management“ und „Forecast Development“.

SFIA ist etwas spezifischer und nennt zum Beispiel die Kompetenzen „Analytics“, „Data Visualization“, „Information Management“, „Data Management“ oder „Data Modeling and Design“. Diese werden dann in bis zu sieben Verantwortungsstufen beschrieben, zum Beispiel „Assist“, „Apply“ oder „Enable“. Aber auch hier fehlt die genaue Beschreibung eines Kompetenzprofils für Data Scientists. Mit einer gewissen Abstraktion könnte man aber die Rahmenwerke nutzen um Kompetenzprofile für Data

Scientists zu konkretisieren. Costa & Santos bestätigen auch, dass Data Scientists in die Kompetenzen von SFIA und e-CF eingeordnet werden können, aber das konkrete Kompetenzen, die für Data Scientists notwendig sind, fehlen (Costa und Santas 2017).

Konkreter für Data Science ist das Data Science Competence Framework (CF-DS) des EDISON-Projekts (<https://edison-project.eu> bzw. <https://github.com/EDISONcommunity/EDSF>) (Demchenko 2018a). Das Projekt hat im Zeitraum von 2015 bis 2017 verschiedene Dokumente veröffentlicht, um das Profil von Data Scientists konkreter zu definieren und um damit die Verbreitung von Data Science als Beruf zu unterstützen. Im Hauptdokument (Demchenko 2018a) sind fünf Kompetenzgruppen definiert, die in Tab. 17.1 aufgeführt sind.

Die ersten drei Gruppen werden als die Kernkompetenzgruppen eines Data Scientist definiert, wohingegen die Kompetenzen in den Bereichen Daten-Management und wissenschaftliche Methoden „nur“ als häufig erforderlich angesehen werden. Statt „Research Methods“ kann im wirtschaftlichen Kontext auch die Kompetenzgruppe „Business Process Management“ relevant sein. In der Definition des CF-DS werden dann die einzelnen Kompetenzen in den fünf Gruppen noch genauer beschrieben. Neben den fünf genannten Hauptgruppen werden auch noch weitere Kompetenzgruppen beschrieben, die sich eher mit konkreten Technologien beschäftigen. Dazu gehören zum Beispiel Kenntnisse von Programmier- und Datenbanksprachen (z. B. Python, R, SQL) und konkreten Systemen (z. B. Hadoop, Spark, scikit-learn).

Auf Basis der Kompetenzgruppen werden dann in weiteren Dokumenten eine Wissensbasis (Body of Knowledge, vgl. Demchenko 2018b), Modell-Curricula (vgl. Demchenko 2018c) und Berufsprofile (vgl. Demchenko 2018d) für Data Science definiert. Für die Modell-Curricula werden zunächst für die Kompetenzen Lernziele auf drei verschiedenen Lernstufen definiert (Kennen, Anwenden, Bewerten). Mit den Lernzielen und den Berufsprofilen werden dann beispielhafte Curricula für Bachelor- und Master-Studiengänge definiert. Hier werden sogar die Inhalte einzelner Kurse detailliert beschrieben, inklusive Angaben von Credits Points im ECTS (European Credit Transfer and Accumulation System).

Auch wenn das EDISON CF-DS aktuell nicht mehr gepflegt wird, stellt es dennoch eine wichtige Grundlage für die Entwicklung von Studien- oder Weiterbildungsangeboten dar. Bei der Entwicklung von Studiengängen oder Weiterbildungs-kursen orientiert man sich oft auch an den Kompetenzen der verfügbaren Lehrenden, wie man an den Beispielen sieht, die in den nächsten Abschnitten vorgestellt werden. Damit aber die verschiedenen Studiengänge und Weiterbildungen und somit auch deren Abschlüsse vergleichbar bleiben, sollte man sich bei der Gestaltung eines Bildungsprogramms für Data Science an Rahmenwerken wie dem EDISON CF-DS orientieren. Fehlende Kompetenzen auf der Seite der Lehrenden können dann zum Beispiel durch entsprechende Berufungen ergänzt werden.

Tab. 17.1 Kompetenzgruppen des EDISON Data Science Competence Framework (Demchenko 2018a)

Kompetenzgruppe	Beschreibung
Data Analytics	Anwenden von geeigneten Datenanalysen und statistischen Techniken auf verfügbare Daten, um neue Zusammenhänge zu entdecken, Einblicke in Forschungsprobleme oder organisatorische Prozesse zu erhalten und die Entscheidungsfindung zu unterstützen
Data Science Engineering	Anwenden von technischen Prinzipien und moderner IT, um neue Datenanalyseanwendungen zu erforschen, zu entwerfen und zu implementieren; Entwicklung von Experimenten, Prozessen, Instrumenten, Systemen und Infrastrukturen zur Unterstützung des Daten-Managements während des gesamten Datenlebenszyklus
Domain Knowledge and Expertise	Anwenden von Domänenwissen (wissenschaftlich oder geschäftlich), um relevante Datenanalyseanwendungen zu entwickeln; Anwendung und Anpassung von allgemeinen Data-Science-Methoden für domänenspezifische Datentypen bzw. -repräsentationen, Daten- und Prozessmodelle, organisatorische Rollen und Beziehungen
Data Management and Governance	Entwicklung und Umsetzung einer Daten-Management-Strategie für die Erfassung, Speicherung, Erhaltung und Verfügbarmachung von Daten für die weitere Verarbeitung
Research Methods and Project Management	Entwicklung von neuen Erkenntnissen und Fähigkeiten, indem wissenschaftliche Methoden (Hypothese, Test/Artefakt, Bewertung) oder ähnliche technische Methoden verwendet werden, um neue Ansätze zu entdecken, um neues Wissen zu schaffen und Forschungs- oder Organisationsziele zu erreichen

17.2 Studiengänge zu Data Science

Studiengänge zu Data Science werden seit etwa 2016 verstärkt an deutschen Universitäten und Hochschulen angeboten. Auf internationaler Ebene wurden entsprechende Studiengänge auch nicht wesentlich früher etabliert. Die New York University hat beispielsweise ihr Center for Data Science (<https://cds.nyu.edu/>) 2013 gegründet und einen Master-Studiengang angeboten. Im Folgenden betrachten wir beispielhaft einige

Studiengänge der TU Dortmund, der RWTH Aachen University und der TU München, deren Übersicht in Tab. 17.2 dargestellt ist.¹

An der TU Dortmund gab es bereits ab 2002 einen Master-Studiengang zu Datenwissenschaften innerhalb der Fakultät für Statistik. Ebenso gab es dort einen Studiengang auf Bachelor-Niveau zu Datenanalyse und Datenmanagement. Entsprechend der Ausrichtung der Fakultät lag der Schwerpunkt bei beiden Studiengängen zunächst auf Statistik und Datenanalyse und weniger auf Daten-Management. Seit 2019 werden beide Studiengänge unter der Bezeichnung „Data Science“ angeboten, die mittlerweile auch mehr Kompetenzen abdecken, die unter dem Bereich „Data Science Engineering“ des EDISON CF-DS fallen. Sowohl im Bachelor- als auch im Master-Programm gibt es eine Reihe von Anwendungsfächern, die aber gemessen an den ECTS-Punkten nur eine untergeordnete Rolle spielen.

An der RWTH Aachen University gibt es seit 2019 einen Master-Studiengang Data Science, der von der Fachgruppe Informatik in der Fakultät Mathematik, Informatik und Naturwissenschaften angeboten wird. Inhaltlich sind wenige Kurse vorgegeben, da viele Wahlmöglichkeiten für Kurse aus den anderen Studiengängen der Fakultät bestehen. Die Mehrzahl der ECTS-Punkte muss aber im Bereich Informatik erbracht werden, was eine sehr große Vielfalt bei den Studienverläufen ermöglicht. Dies ist aber auch schon bei den anderen Studiengängen der Fachgruppe zu sehen, die mit dreizehn Lehrstühlen und ca. vierzig Lehrenden ein sehr breites Profil hat. Insofern können sich die individuellen Abschlüsse stark unterscheiden, d. h. von einem mathematisch-theoretisch Schwerpunkt über eine Big-Data-Spezialisierung bis hin zu einer Vertiefung in Hochleistungsrechnen ist alles möglich. Von anderen Fakultäten (Medizin) werden in Aachen noch die berufsbegleitenden Studiengänge in „Medical Data Science“ und „Data Analytics and Decision Science“ angeboten, die ihre Schwerpunkte in den jeweiligen Anwendungsfächern und dem Kompetenzbereich „Data Analytics“ des EDISON CF-DS haben.

Die TU München hat seit 2016 zwei Master-Studiengänge im Bereich Data Science im Angebot. Zum einen gibt es den Master-Studiengang „Data Engineering and Analytics“ der Fakultät für Informatik, zum anderen gibt es in der Fakultät für Mathematik den Master-Studiengang „Mathematics in Data Science“. Die beiden Studiengänge stehen aber nicht in Konkurrenz zueinander, sondern haben ein gemeinsames Kursprogramm. Verglichen mit den Kompetenzgruppen des EDISON CF-DS liegt der Schwerpunkt bei „Mathematics in Data Science“ im Bereich Data Analytics und im anderen Studiengang dem Namen entsprechend bei „Data Science“

¹Wir können hier nur eine Auswahl von Studiengängen näher betrachten. Der Hochschulkompass der Hochschulrektorenkonferenz (<https://www.hochschulkompass.de/>) listet Ende Oktober 2020 insgesamt 155 Studiengänge unter dem Stichwort „Data Science“ in Deutschland, davon 65 Bachelor- und 90 Master-Studiengänge. 62 dieser Studiengänge werden an Universitäten, 93 an (Fach-)Hochschulen angeboten. Nicht abgefragt wurden Studiengänge mit Data-Science-Schwerpunkt, die aber unter anderem Bezeichnungen angeboten werden.

Tab. 17.2 Master- und Bachelor-Studiengänge zu Data Science

Hochschule	Abschluss	Kernbereiche	Anwendungsbereiche
TU Dortmund	Data Science, B. Sc	Mathematik, Statistik, Datenstrukturen, Software-technik, Datenbanken	Life Science, Wirtschafts- und Ingenieurwissenschaften, Musik, Journalismus
TU Dortmund	Data Science, M. Sc	Statistik, Data Science, Big Data	Life Science, Wirtschafts- und Ingenieurwissenschaften
RWTH Aachen University	Data Science, M. Sc	Informatic, Mathematik, Ethik	Physik, Wirtschafts-, Lebens- und Sozialwissenschaften
TU München	Data Engineering and Analytics, M. Sc. bzw. Mathematics in Data Science, M. Sc	Data Engineering, Data Analytics, Data Analysis	Überfachliche Grundlagen, Social and Political Aspects

Engineering“. In den Studienplänen wird zwischen zwei ähnlich klingenden Themenbereichen unterscheiden: „Data Analysis“ beschäftigt sich eher mit den mathematischen Aspekten und den statistischen Grundlagen bei der Datenanalyse, wohingegen „Data Analytics“ sich auf die Informatik-Seite der Datenanalyse konzentriert (z. B. Data Mining, Machine Learning, Computer Vision). Der dritte Kernbereich „Data Engineering“ befasst sich mit den technischen Aspekten von Big-Data-Systemen und Daten-Management. Es bestehen zwar viele Wahlmöglichkeiten innerhalb der Kernbereiche, thematisch sind die Möglichkeiten aber stärker eingegrenzt als zum Beispiel an der RWTH Aachen University. Das Profil der Studiengänge an der TU München ist demnach klarer auf das Kompetenzprofil „Data Science“ ausgerichtet. Anwendungsfächer spielen dort aber nur eine sehr kleine Rolle.

Bei den Hochschulen für Angewandte Wissenschaften (HAW) gibt es auch ein breites Angebot für Data-Science-Studiengänge. Während bei den Universitäten die Master-Studiengänge überwiegen, halten sich die Angebote Bachelor- und Master-Studiengänge für Data Science bei den HAW ungefähr die Waage. Erwartungsgemäß spielen bei den HAW die mathematischen Fächer eine geringere Rolle und die Kurse mit Praxisbezug nehmen einen stärkeren Anteil ein als bei den Angeboten der Universitäten.

17.3 Berufliche Weiterbildung zu Data Science

Der Markt für berufliche Weiterbildungen ist in den letzten Jahren aufgrund der Popularität des Themas „Data Science“ enorm gewachsen. An einigen Universitäten und Hochschulen gibt es berufsbegleitende Angebote für Data Science, oft auch speziell für bestimmte Anwendungsbereiche, wie im vorherigen Abschnitt dargestellt. In diesem Abschnitt wollen wir uns aber speziell auf Zertifikatskurse in der Weiterbildung konzentrieren, die einzelne Kompetenzbereiche abdecken, aber keine vollständigen Studiengänge sind. In diesen Kursen erhält man also nach erfolgreicher Teilnahme und Prüfung ein Zertifikat und nicht einen Abschluss auf Bachelor- oder Master-Niveau.

Weiterbildungen mit Zertifikatsabschluss sind klar zu trennen von Weiterbildungen, die lediglich Teilnahmebestätigungen oder ähnliches ausstellen. Personenzertifizierungen können in Deutschland nur von akkreditierten Zertifizierungsstellen durchgeführt werden. Dabei werden die Kompetenzen einer Person gemäß eines Zertifizierungsprogramms bewertet. Das Zertifizierungsprogramm definiert die Kompetenzen und Stufen (z. B. Kennen, Anwenden, Beurteilen), die für eine bestimmte Zertifizierung erreicht werden müssen. Die vorhandenen Kompetenzen werden in der Regel durch eine Prüfung verifiziert, die auf verschiedene Arten abgelegt werden kann (z. B. mündlich, schriftlich, Projektarbeit). Für die Durchführung von Personenzertifizierungen und der Arbeit der Zertifizierungsstellen sind enge Rahmen vorgegeben. So dürfen zum Beispiel die Lehrenden in einer Weiterbildung nicht auch gleichzeitig Prüfer sein. Verschiedene Ausschüsse prüfen die Inhalte und Prüfungspläne der Weiterbildungen, die zu einem zertifizierten Abschluss führen sollen.

Zertifizierungsstellen werden von der Deutschen Akkreditierungsstelle (DAkkS) akkreditiert. Beispiele dafür sind der TÜV, die Deutsche Gesellschaft für Qualität oder die Personenzertifizierungsstelle bei der Fraunhofer Gesellschaft (<https://www.personenzertifizierung.fraunhofer.de/>). Das Zertifikat wird von der Zertifizierungsstelle ausgestellt in deren Verantwortung auch die Prüfung durchgeführt wird und trägt auch deren Namen, was also auch ein gewisses Qualitätsmerkmal ist.

Darüber hinaus können auch Hochschulen und Universitäten Zertifikatskurse anbieten, für die aber wie für reguläre Studiengänge Modulhandbücher und Prüfungsordnungen veröffentlicht werden müssen. Seit einigen Jahren wird das aus der Schweiz stammende Modell des „Master of Advanced Studies“ auch an einigen deutschen Hochschulen angeboten. Master of Advanced Studies sind Weiterbildungen auf Hochschulniveau, in der Regel berufsbegleitend, für Personen, die bereits einen Hochschulabschluss und einschlägige Berufserfahrung haben. Die Programme sind mehrstufig aufgebaut:

- Certificate of Advanced Studies (CAS): Ein CAS setzt sich meist aus mehreren Modulen (d. h. einzelnen Kursen) zusammen und umfasst etwa 10 ECTS-Punkte, was einem Arbeitsaufwand von etwa 300 h entspricht.
- Diploma of Advanced Studies (DAS): Das DAS umfasst ca. 30 ECTS-Punkte und setzt sich demnach aus drei CAS-Bausteinen zusammen. Neben den CAS-Modulen kann auch eine separate Leistung erforderlich sein, wie beispielsweise eine Projektarbeit.
- Master of Advanced Studies (MAS): Je nach Hochschule umfasst die MAS-Stufe 60 bis 90 ECTS-Punkte und beinhaltet auch eine abschließende Master-Arbeit, vergleichbar mit den Abschlussarbeiten in regulären Master-Studiengängen.

Für Data Science werden verschiedene Zertifikatskurse angeboten. Beispielhaft stellen wir hier die Zertifikatskurse der Fraunhofer-Allianz „Big Data und Künstliche Intelligenz“, der Hochschule Niederrhein und der Bitkom Akademie vor. An der Erstellung und Durchführung der beiden erstgenannten Weiterbildungsangebote war der Autor beteiligt, das dritte Angebot wurde auf Basis der öffentlich verfügbaren Informationen beschrieben.

17.3.1 Zertifikatsprogramm der Fraunhofer Gesellschaft zu Data Science

Die Fraunhofer-Allianz „Big Data und Künstliche Intelligenz“ (<https://www.bigdata-ai.fraunhofer.de/>) ist ein Zusammenschluss von etwa 25 Fraunhofer-Instituten. Neben verschiedenen Beratungsangeboten im Umfeld von Big Data und Data Science bietet die Allianz auch ein Schulungsprogramm zu Data Science mit einigen Zertifikatskursen an. Bei der Entwicklung des Programms wurde auch das EDISON Data Science

Competence Framework genutzt, dessen Struktur sich aber nicht mehr im Programm wiederfinden lässt. Das Programm ist mehrstufig und beginnt mit der Weiterbildung „Data Scientist – Basic Level“, die seit 2016 als Zertifikatskurs angeboten wird. Der Kurs vermittelt Grundlagenwissen in verschiedenen Kompetenzbereichen. Die Weiterbildung wird in der Regel in Präsenz innerhalb von fünf Tagen in einer Woche durchgeführt, die Prüfung findet am darauffolgenden Tag statt. Durch die Corona-Krise wurde die Durchführung auf Online-Medien umgestellt, die Prüfung kann ebenso online durchgeführt werden. Die Weiterbildung besteht aus mehreren Modulen, die die Themen Big-Data-Systeme, Daten-Management, Datenanalyse, Datenvizualisierung, Datensicherheit und Geschäftsperspektiven umfassen. Aufgrund der Themenbreite kann dabei kaum Detailwissen vermittelt werden. Es werden also viele Kompetenzen des EDISON-Rahmenwerks angesprochen, aber nur eine geringe Kompetenzstufe erreicht. Die Zielgruppen sind daher in erster Linie Führungskräfte und Projektverantwortliche, die Data-Science-Projekte koordinieren oder darin als Anwender involviert sind, d. h. sie müssen selbst nicht implementieren, aber qualifizierte Entscheidungen über den Ablauf eines Projekts treffen können.

Um die erste Stufe des Programms „Foundation Level“ zu erreichen, muss neben dem Basis-Kurs ein weiterer Kurs mit etwas geringerem Umfang (meist drei bis vier Tage) zu einem Spezialthema absolviert werden. Beispiele für Spezialkurse sind „Data Analytics“, „Deep Learning“ oder „Data Management“. Hier werden dann auch höhere Kompetenzstufen erreicht, weil die Kenntnisse in bestimmten Bereichen vertieft werden. Die zweite Stufe „Advanced Level“ erfordert dann noch Berufserfahrung im Data-Science-Bereich und die Erstellung einer Projektarbeit. Für die abschließende Stufe „Senior Level“ ist noch weitere Berufserfahrung und eine Studienarbeit mit Vortrag erforderlich. Die Zertifizierungen werden von der Personenzertifizierungsstelle bei Fraunhofer durchgeführt.

Im Vergleich mit anderen Weiterbildungsangeboten fallen hier das Fraunhofer-Zertifikat, die wissenschaftliche Qualität und die Vielfalt der Themen in den Basis- und Spezialkursen auf. Neben einer fachlichen Vertiefung ist auch eine Vertiefung in einem Anwendungsbereich möglich (z. B. Energie, Industrie 4.0, Gesundheit). Die meisten Teilnehmenden schließen nur den Basic-Level-Kurs ab, einige nehmen auch an einem Spezialkurs teil, um dann mit einer Projektarbeit das „Advanced Level“ zu erreichen. Da die Schulungen ganztägig an mehreren Tagen durchgeführt werden, erfordern sie einen hohen Zeitbedarf, der bei der Zielgruppe aufgrund des meist hohen Arbeitsdrucks nicht erreicht werden kann. Dennoch werden seit 2016 etwa 25 Kurse pro Jahr in den verschiedenen Varianten durchgeführt. Auf Anfrage werden auch Inhouse-Schulungen durchgeführt, diese jedoch meist ohne Zertifikat.

17.3.2 Zertifikatsstudien der Hochschule Niederrhein

Die Zertifikatsstudien der Hochschule Niederrhein zu Data Science wurden seit 2018 im Rahmen des Projekts „Wissenschaftliche Weiterbildung für die digitale Wirtschaft“, u. a.

in Kooperation mit der Hochschule Bonn-Rhein-Sieg und der Fachhochschule Dortmund entwickelt und werden nun vom Zentrum für Weiterbildung an der Hochschule (<https://www.hs-niederrhein.de/weiterbildung/>) fortgeführt.

Das Programm besteht aus drei Zertifikatsstudien auf CAS-Niveau, die jeweils drei bis vier Weiterbildungskurse umfassen. Ein Weiterbildungskurs umfasst etwa drei bis vier Tage und entspricht mit Vor- und Nachbearbeitung auch etwa drei bis vier ECTS-Punkte. Die drei Zertifikatsstudien behandeln die folgenden Themen:

- CAS Data Analyst: Hier wird zunächst das Reporting mit multidimensionalen Kennzahlen in klassischen Data-Warehouse-Systemen betrachtet. Anschließend werden verschiedene Techniken zur Datenanalyse und Datenvisualisierung vertieft (siehe Kapitel *Fundamentale Analyse- und Visualisierungstechniken* und *Fortgeschrittene Verfahren zur Analyse und Datenexploration, Advanced Analytics und Text Mining*).
- CAS Data Architect: In diesem Zertifikatsstudium werden zunächst die Grundlagen des Data Engineering (bzw. Daten-Management) besprochen, dann erfolgt eine Vertiefung in In-Memory- oder Big-Data-Technologien. Ergänzt wird das CAS durch den Kurs Data Governance, der sich mit den organisatorischen und rechtlichen Herausforderungen beim Daten-Management beschäftigt.
- CAS Data Strategist: Auch in diesem Zertifikatsstudium werden organisatorische Fragen des Daten-Managements betrachtet, aber weniger aus der technischen Perspektive wie im CAS Data Architect, sondern eher aus der wirtschaftlichen Perspektive, d. h. mit welchen Geschäftsmodellen lassen sich Data-Science-Methoden gewinnbringend einsetzen.

Inhaltlich decken die Kurse also die Kompetenzbereiche Data Analytics, Data Science Engineering und Data Management des EDISON-Rahmenwerks ab. Geschäftsmodelle bzw. der Wert von Data Science für ein Unternehmen werden in EDISON CF-DS nicht berücksichtigt, wohl auch weil das Rahmenwerk eher auf grundständige Studiengänge ausgerichtet ist und weniger für die berufsbegleitende Weiterbildung konzipiert wurde.

Die meisten Kurse im Weiterbildungsprogramm der Hochschule Niederrhein werden über mehrere Wochen mit einem Schulungstag pro Woche angeboten. Diese zeitliche Planung wurde von den Teilnehmenden in einer ersten Pilotierung begrüßt, da mehrere aufeinanderfolgende Tage mit den sonstigen beruflichen Aufgaben schwer vereinbar sind. Die Prüfungen wurden in der Form von Projektarbeiten und/oder Vorträgen abgenommen.

Die Pilotphase wurde erfolgreich in 2020 trotz erschwerter Bedingungen in der Corona-Krise erfolgreich abgeschlossen. Nach einer leichten Anpassung des Programms ist 2021 eine Fortführung des Zertifikatsprogramms geplant. Neben inhaltlichen Ergänzungen bzw. Korrekturen ist auch die Entwicklung eines DAS Data Science geplant. Für das DAS müssen zwei CAS und eine Studienarbeit abgeschlossen werden. Die Studienarbeit ist vom Umfang her vergleichbar mit einer Bachelor-Arbeit und soll die Anwendung von Data-Science-Methoden auf ein konkretes Anwendungsproblem

darstellen. Die Entwicklung eines MAS wurde auch besprochen, aber aufgrund der zeitlichen Anforderungen hielten es die beteiligten Lehrenden nicht für realistisch, dass sich Teilnehmenden dafür interessieren würden. Es müssten etwa 10 bis 15 Weiterbildungs-kurse in einem begrenzten Zeitraum (drei bis fünf Jahre) erfolgreich absolviert werden.

Im Vergleich zum Zertifikatsprogramm von Fraunhofer haben diese Kurse den Vorteil, dass ihre Inhalte gemäß dem ECTS-Schema definiert sind und so eine Anrechenbarkeit für andere berufsbegleitende oder reguläre Studiengänge möglich wäre. Wie auch bei Fraunhofer spielt auch hier der wirtschaftliche Mehrwert der Data Science für Unternehmen eine wichtige Rolle, da die Teilnehmenden im Berufsleben sich häufig diesen Fragen stellen müssen. In der Pilotphase wurden die Kurse innerhalb eines CAS von den gleichen Teilnehmenden besucht, man konnte also die Kurse inhaltlich aufeinander aufbauen lassen. Das Weiterbildungsprogramm von Fraunhofer ist dagegen offener gestaltet, hier können Spezialkurse auch von Teilnehmenden besucht werden, die vorher nicht am Basic-Level-Kurs teilgenommen haben. Für die Lehrenden erschwert dies aber die Durchführung des Kurses, da man nicht bei allen das gleiche Vorwissen voraussetzen kann. Das unterschiedliche Vorwissen der Teilnehmenden ist eine generelle Herausforderung bei der Weiterbildung und könnte durch den Einsatz von E-Learning-Elementen vor Beginn der Weiterbildung ausgeglichen werden.

17.3.3 Zertifikatslehrgang zum Data Scientist der Bitkom Akademie

Der Zertifikatslehrgang Data Scientist der Bitkom Akademie (<https://www.bitkom-akademie.de/zertifikatslehrgang/ausbildung-data-scientist>) ist eine berufsbegleitende Ausbildung, die in Zusammenarbeit mit der Steinbeis+Akademie angeboten wird. Der Lehrgang umfasst fünf Module, die in einem Zeitraum von vier bis fünf Monaten absolviert werden sollen. Jedes Modul wird an zwei aufeinanderfolgenden Tagen gelehrt. Der Abschluss ist ein Diploma of Advanced Studies im Umfang von 15 ECTS-Punkten, das von der Steinbeis+Akademie verliehen wird. Die Zielgruppe sind wie bei den beiden anderen vorgestellten Zertifikatskursen Entscheidungsträger innerhalb eines Unternehmens.

Ebenso sind die Inhalte vergleichbar mit den anderen Angeboten:

- Data Scientist – Berufsbild der Zukunft: Der Lehrgang beginnt mit einer Einführung und Vorstellung des Potenzials von Data Science in Unternehmen und verschiedenen Branchen.
- Datenhaltung & Data Governance: Hier werden Themen des organisatorischen und technischen Daten-Managements besprochen. Dazu gehören zum Beispiel die Grundlagen von Datenbank-Management-Systemen aber auch die Datenschutzgrundverordnung.
- Datenakquisition & Datenintegration: In diesem Modul wird mit konkreten Systemen gearbeitet und verschiedene Architekturen zum Daten-Management vorgestellt.

- Data-Science-Algorithmen: Verschiedene Datenanalyse-Methoden sind Teil dieses Moduls, wobei es auch um die kritische Bewertung der Ergebnisse geht.
- Generierung von Business Value und Outcome: Neben Datenvisualisierungstechniken werden vor allem die Hürden bzw. die Voraussetzungen für die erfolgreiche Durchführung für Data-Science-Projekte in Unternehmen besprochen.

Die Beschreibung und der Umfang der Themen lassen vermuten, dass auch hier eher in die Breite als in die Tiefe der Data-Science-Themen gearbeitet wird. Daten-Management steht bei diesem Zertifikatslehrgang mehr im Vordergrund als bei den anderen Angeboten. Datenanalyse-Methoden werden lediglich in einem Modul behandelt.

Da die Bitkom Akademie nur den Rahmen für Weiterbildungskurse bereitstellt, die von Lehrenden aus verschiedenen Unternehmen oder Forschungseinrichtungen durchgeführt werden, gibt es zu den anderen Angeboten der Akademie keine Anknüpfungspunkte. Der Zertifikatslehrgang zu Künstliche Intelligenz könnte gut mit diesem Lehrgang verknüpft werden, aber eine Möglichkeit für weitergehende Abschlüsse nach der Teilnahme an mehreren Lehrgängen wird nicht aufgezeigt.

17.4 Fazit

Die Angebote für Studiengänge und Weiterbildungen im Themenfeld Data Science sind sehr vielfältig und werden voraussichtlich auch noch in den nächsten Jahren zunehmen. Es ist davon auszugehen, dass in den Weiterbildungsangeboten das Thema Künstliche Intelligenz in Zukunft eine stärkere Rolle spielen wird. Die Übergänge zwischen Data Science und Künstlicher Intelligenz sind sowieso fließend, vor allem wenn man die Bereiche Machine Learning oder Deep Learning betrachtet.

Bei der Entwicklung eines Data-Science-Programms sollte auf entsprechende Kompetenz-Rahmenwerke zurückgegriffen werden. Das EDISON Data Science Competence Framework stellt eine gute Grundlage dar, die bei der Abdeckung eines breiten Kompetenzspektrums hilft. In der beruflichen Weiterbildung spielen aber weitere Aspekte wie datengetriebene Geschäftsmodelle, Wert von Daten, Mehrwert von Data-Science-Methoden eine größere Rolle, die im Rahmenwerk des EDISON-Projekts nicht adäquat abgebildet sind.

Wird die Weiterbildung nicht nur zur persönlichen Fortbildung, sondern zum Steigern der Karrierechancen und der Chancen auf dem Arbeitsmarkt besucht, sollte darauf geachtet werden, dass die Weiterbildung mit einem Zertifikat einer renommierten Einrichtung abgeschlossen werden kann. Entsprechende Angebote sind etwas preisintensiver und richten sich eher an Entscheidungsträger als an Software-Entwickler von Data-Science-Lösungen. Letztere können sicherlich für bestimmte Detailkenntnisse auch in der Vielzahl der kostenlosen Online-Angebote ein passendes Angebot finden. Die Online-Plattform Coursera (<https://www.coursera.org/>) listet Ende Oktober 2020 über 1800 Angebote zum Suchbegriff „Data Science“, von denen die meisten Inhalte kosten-

los in Anspruch genommen werden können. Prüfungen mit Zertifikaten sind dort in der Regel kostenpflichtig. Umfang, Qualität und Struktur der Angebote variieren dort aber sehr stark.

Bei den Studiengängen sollten Studieninteressierte sich genau den Studienverlaufsplan und das Modulhandbuch anschauen, um zu überprüfen, ob das Studienangebot auch zu den eigenen Erwartungen passt. Der Aufbau der Studiengänge und die Wahlmöglichkeiten bei den Vertiefungsgebieten sind bei den verschiedenen Hochschulen und Universitäten doch sehr unterschiedlich.

Literatur

- Costa, C., Santas, M.Y.: The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. *Int. J. Inf. Manage.* **37**(6), 726–734 (2017). <https://doi.org/10.1016/j.ijinfomgt.2017.07.010>
- Demchenko, Y. (Hrsg.): EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS), Release 3. December 2018. https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_CF-DS-release3-v10.pdf (2018a)
- Demchenko, Y. (Hrsg.): EDISON Data Science Framework: Part 2. Data Science Body of Knowledge (DS-BoK). Release 3. December 2018. https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DS-BoK-release3-v06.pdf (2018b)
- Demchenko, Y. (Hrsg.): EDISON Data Science Framework: Part 3. Data Science Model Curriculum (MC-DS). Release 3. December 2018. https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_MC-DS-release3-v05.pdf (2018c)
- Demchenko, Y. (Hrsg.): EDISON Data Science Framework: Part 4. Data Science Professional Profiles (DSPP). Release 3. December 2018. https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DSPP-release3-v07.pdf (2018d)
- Kelleher, J.D., Tierney, B.: Data Science. MIT Press, Cambridge, Massachusetts; London, England (2018)



Plattformökonomie für Data Plattformen

18

Valeria Knoll und Alexa Scheffler

Zusammenfassung

Es liegt nahe, dass in der Datennutzung von Unternehmen ähnliche Wertschöpfungsprinzipien gelten, wie bei plattformbasierten Geschäftsmodellen wie Amazon und AirBnB. Der Beitrag „Plattformökonomie für Data Plattformen“ untersucht, welche Design-Prinzipien den erfolgreichen Plattformen zugrunde liegen und wie sich diese auf Data Plattformen übertragen lassen. Dafür wird das Konzept einer Plattform auf ein Data-Umfeld übertragen sowie Plattformteilnehmer und Netzwerkeffekte definiert. Durch erste Erkenntnisse zu Design-Prinzipien können Data Plattformen so gesteuert werden, dass Nutzeffekte gezielter herbeigeführt und realisiert werden.

18.1 Motivation

Daten sind „eine Schlüsselressource für gesellschaftlichen Wohlstand und Teilhabe, für eine prosperierende Wirtschaft und den Schutz von Umwelt und Klima, für den wissenschaftlichen Fortschritt und für staatliches Handeln“ (Bundesregierung 2019). So begründet die Bundesregierung ihre Datenstrategie. Die besondere Bedeutung von Daten als wirtschaftlicher Treiber wird auch an weiteren Stellen hervorgehoben. Das Wirtschaftsmagazin Forbes proklamiert „Every company is a data company“ (Bean 2018), „Big Data“ wurde als „Management Revolution“ bezeichnet (McAfee und Brynjolfsson

V. Knoll · A. Scheffler (✉)

Köln, Deutschland

E-Mail: Alexa.Scheffler@axa.de

V. Knoll

E-Mail: valeria.knoll@axa.de

2012). Die datenbezogenen Geschäftsmodelle und Technologien sind jedoch noch recht jung. Demnach existieren noch wenige Erfahrungswerte über den Aufbau von Data Plattformen auf dem Markt. Der renommierte Professor Dan Arielly twitterte provokant: „Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it“ (Arielly 2013). Die Relevanz des Themas und der damit einhergehende Investitionsbedarf scheint aber allgemein anerkannt zu sein. 88 % der befragten Vorstände geben an, dass sie dringenden Bedarf für Investitionen in Daten sehen (NVP 2019, S. 2). Dazu gehört insbesondere auch die Investition in Data Plattformen (McKinsey 2018). Mangelnde Fachkräfte, neue technologische Infrastruktur und der crossfunktionale Charakter sind jedoch für Unternehmen große Herausforderungen bei dem Aufbau einer Data Plattform.

Best-Practices für den Aufbau und das Betriebsmodell von Data Plattformen können anhand der Erfolgsgeschichten anderer Plattformtypen abgeleitet werden, die schon länger präsent sind (z. B. von Dienstleistungs- oder Warenplattformen wie Ebay oder AirBnB). Unternehmen wie Amazon, Spotify oder Uber haben den Erfolg eines plattform-basierten Geschäftsmodells bewiesen und befinden sich inzwischen unter den wertvollsten Marken der Welt (Interbrand 2019). Es lässt sich beobachten, dass Plattformen die Geschäftsmodelle ganzer Industriezweige verändern (Van Alstyne et al. 2016). Im Retail-Sektor verzeichnete der Marktwert von Amazon in den Jahren 2006–2016 fast 2000 % Wachstum, während der von den etablierten amerikanischen Retailern Sears, JC Penney und BestBuy um bis zum 96 % schrumpfte (Yueh 2017, S. 30). Dieser Beitrag untersucht, welche Design-Prinzipien bei diesen Plattformen zugrunde liegen und sich auf Data Plattformen übertragen lassen. Betreiber von Data Plattformen können diese Design-Prinzipien anwenden, um die Monetarisierung besser steuern und planen zu können.

18.2 Begriffshaushalt

Als Einstieg in das Thema klären wir die Begriffe Plattform(ökonomie) und Data Plattform.

18.2.1 Plattformen und Plattformökonomie

Auf einer Plattform treffen sich Konsumenten sowie Anbieter für den wertschöpfenden Austausch von Ressourcen. Dabei nutzen sie die Infrastruktur, die ein Plattformbetreibern zur Verfügung stellt. Das besondere an der Plattformökonomie ist, dass sie die bisher vorherrschende Pipeline-Ökonomie ablöst (Van Alstyne et al. 2016). In der Pipeline-Ökonomie, die der industriellen Wirtschaft des 20. Jahrhunderts zugrunde lag, entsteht der Wertbeitrag eines Unternehmens durch die Kontrolle über eine lineare Aktivitätenabfolge, die sie ausüben, weil sich im Besitz der Ressourcen befinden.

In der Plattformökonomie ist die Wertschöpfung von dem Besitz der Ressourcen getrennt. Ein Plattform-Unternehmen ist kein Eigentümer von Ressourcen, die über die angebotene Plattform ausgetauscht werden. Die Plattform ermöglicht lediglich den Kontakt zwischen den eigentlichen Besitzern von Ressourcen und deren potenziellen Konsumenten (Parker et al. 2017, S. 7–12). Ressourcen oder Werteinheiten, die über eine Plattform ausgetauscht werden, sich von Waren (z. B. Amazon, eBay) und Dienstleistungen (z. B. Uber, Airbnb) bis hin zu Software, Content (z. B. Facebook, Instagram, Tiktok). Aber auch Daten sind eine mögliche Ressource (Schreieck et al. 2016). Einige Ressourcen existieren unabhängig von einer Plattform, durch welche sie ausgetauscht werden, z. B. gibt es auch Bücher, die nicht über Amazon verschickt werden. Andere Ressourcen werden dagegen explizit für eine Plattform und mithilfe einer Plattform geschaffen. Dies ist vor allem für digitale Ressourcen charakteristisch, u. a. Tweets, Facebook-Posts, mobile Applikationen.

Die Nutzung der von Plattformbetreibern zur Verfügung gestellten Infrastruktur konstituiert einen zweiseitigen Markt (Anbieter und Konsumenten). Der Wertbeitrag, der durch die Interaktion auf einem zweiseitigen Markt generiert wird, steigt mit der zunehmenden Anzahl von Nutzern, z. B. mehr Facebook- bzw. Instagram-Nutzer liefern mehr Content, was wiederum ein breiteres Auditorium interessiert und anlockt. Dies wird als Netzwerk-Effekt bezeichnet. Netzwerkeffekte wurden vor langer Zeit erkannt und auch in der Pipeline-Ökonomie ausgenutzt. Jedoch waren sie nie so stark wie heute, da die technologischen Entwicklungen durch das Internet die Eintrittskosten für neue Nutzer sehr stark gesenkt haben. Diese niedrigen Eintrittskosten führen dazu, dass Anbieter und Konsumenten oft austauschbar sind. Über-Fahrer können die Fahrten selbst in Anspruch nehmen, Airbnb-Mieter können selbst Übernachtungsmöglichkeiten anbieten und Verfasser von Tweets lesen wiederum die Beiträge der anderen Verfasser. Aufgabe des Plattform-Betreibers ist es, für ein passendes Verhältnis an Anbietern und Konsumenten zu sorgen und einen Rahmen für die Nutzung zu schaffen. Im Plattform-Kontext spricht man auch von „kuratieren“. Bspw. zieht Ebay durch besondere Maßnahmen in Form von Rabatt-Aktionen an Feiertagen Käufer an. Die Plattform fungiert somit als semi-regulierter Marktplatz, der durch den Plattform-Betreiber teilweise gesteuert wird. Zu den Steuerungsinstrumenten gehören z. B. auch Käuferschutzprogramme für Transaktionen auf Ebay und Absicherung der Zahlungen über Paypal (Jacobides et al. (2018), S. 2260), Garantie- sowie Rückgaberegelungen auf Amazon usw. Andere Instrumente sind Bewertungen, Repots, Likes o.ä., die der Autoregulierung auf einer Plattform dienen.

Ein wesentlicher Bestandteil der Wertschöpfung auf einer Plattform stellt das Matching dar. Das Matching hat zum Ziel, einem Nutzer die für ihn relevanten Ressource anzubieten. Beim Matching unterscheidet man implizite und explizite Filter (Parker et al. 2017, S. 40–41, 47–48). Explizite Filter sind in der Regel Suchfunktionen. So kann bspw. eine bestimmte mobile Applikation gefunden bzw. ein Kanal mit interessanten Inhalten abonniert werden. Implizite Filter hingegen sind Vorschläge, die auf der Nutzungshistorie basieren und Push-Vorschläge machen. Die impliziten

Filter sind nicht immer für den Nutzer als Filter erkennbar. Z.B. werden Nutzern von Dating-Plattformen nur Vorschläge für potenzielle Partner gemacht, die eine ähnliche wahrgenommene Attraktivität haben. Je besser die impliziten Filter einer Plattform sind, desto einfacher ist es für den Nutzer zu den für ihn relevanten Werteinheiten zu kommen. Durch sie besteht aber auch die Gefahr, dass sie dem Nutzer nur Ressourcen vorschlagen, die seinen vermeintlichen Bedürfnissen entsprechen und dadurch isolieren („Filterblase“). Es ist also erforderlich die impliziten Filter für den Nutzer transparent zu machen.

18.2.2 Data Plattform

Um zu analysieren, inwiefern die Best Practices und die Gestaltungsprinzipien des Plattform-Geschäftsmodells sich auf Data-Plattformen übertragen lassen, muss erst geklärt werden, was eine Data-Plattform ist.

Der Begriff Data Plattform wird in der Presse und in der Literatur nicht mit einheitlichem Begriffsverständnis und teilweise mit Synonymen verwendet (z. B. Data Analytics Plattform, Data Lake, Data Factory). Nach unserem Verständnis ist eine Data Plattform ein Ökosystem für die Datenauswertung (Reporting und Dashboards, Planung sowie Data Science und Künstliche Intelligenz). Der Bedarf für Data Plattformen besteht insbesondere durch die neuen Möglichkeiten für KI-Ansätze (wie z. B. Natural Language Processing in der Kundenkommunikation oder Computer Vision bei der Schadenbearbeitung im Versicherungsfall). Daneben spielen aber auch Automatisierungspotenziale für Standard-Reporting und Finanzplanung eine wichtige Rolle.

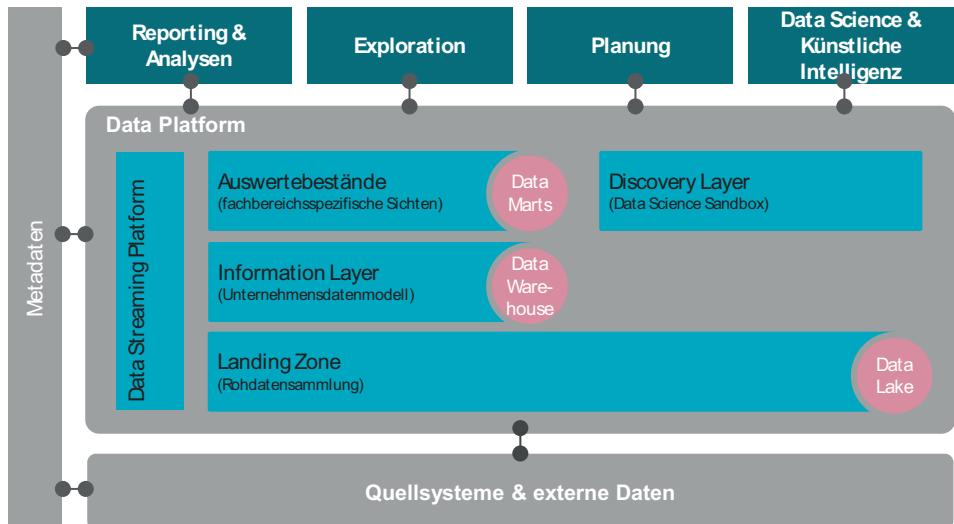
Reporting und Dashboards Periodische Bereitstellung vordefinierter Berichte, oft als Ergebnisnachweis einzelner Geschäftsbereiche, z. B. Segmente, Länder, Tochtergesellschaften usw. Visualisierungen helfen dem Nutzer dabei Daten zu erfassen und dadurch weitere Handlungsbedarfe zu identifizieren. Man unterscheidet vordefinierte Berichte, bei denen die Kennzahlen und Auswertungsdimensionen bereits vorgegeben sind, und ad-hoc Reporting, bei denen der Nutzer anlassbezogen Daten abfragt. Reportings bestehen in der Regel aus Tabellen, auf die der Nutzer verschiedene Filterfunktionen setzen kann. In Dashboards stehen mehr grafische Elemente zur Verfügung (z. B. Landkarten, interaktive Schaltflächen etc.)

Planung Bestimmung der erwarteten relevanten Zukunftswerte mit Hilfe von Modellen und Simulationen, sowie Plan-Ist-Vergleiche.

Data Science und Künstliche Intelligenz Anwendung von statistischen, mathematischen und ökonometrischen Modellen, sowie das Trainieren der KI-Modelle auf Basis des verfügbaren Datenbestands zum Kategorisieren und Entscheidungstreffen.

Tab. 18.1 Funktionen von Data Plattform, Data Warehouse und Data Lake

	Data Warehouse	Data Lake
Datennutzung	deskriptiv	explorativ
Datenarten	strukturiert	unstrukturiert
Struktur	Schreibschema	Leseschema
Technologien	SQL-Abfragen, Reporting- und Visualisierungstools	Programmoberflächen, z. B. Python oder R
Nutzergruppen	Data Analysten	Data Scientisten

**Abb. 18.1** Exemplarische Architektur einer Data Platform

Eine Data Plattform kombiniert die Funktionalitäten eines Data Warehouses mit Funktionalitäten eines Data Lakes. Die unterschiedlichen Merkmale von Data Warehouse und Data Lake sind in Tab. 18.1 abgebildet.

Eine Data Plattform enthält folgende Bausteine (vgl. Abb. 18.1):

Landing Zone (auch Staging Area oder Data Ingestion Schicht) Bezeichnet die Schicht, in der Rohdaten gespeichert werden und in der diese aus Quellsystemen geliefert werden. Die Rohdaten werden nach dem erfolgreichen Weiterreichen gelöscht. In der Landing Zone werden auch unstrukturierte Daten abgelegt. Sie daher für die Verarbeitung großer Datenmengen optimiert.

Information Layer Hier wird ein Unternehmensdatenmodell abgespeichert. In dieser Schicht werden die Daten entsprechend der Business Logik zusammengeführt, bereinigt, harmonisiert und historisiert.

Auswertebestände (auch Data Mart bzw. Information Marts oder Cubes) Erfüllen die spezifischen Anforderungen einzelner Fachbereiche. Hierzu werden die Teilmengen von Daten modelliert, um diese für Reporting und Analysen schnell und effizient zugreifbar zu machen.

Discovery Layer (auch Use Case Layer) Diese integrative Schicht wird sowohl aus der Landing Zone, dem Information Layer als auch aus Auswertebeständen in beliebiger Kombination bedient, um den Informationsbedarf für einen bestimmten Data Science Case zu decken. In dieser Schicht können unstrukturierte und strukturierte Daten analysiert und Modelle trainiert werden.

Metadaten Metadaten sind beschreiben die auf der Data Plattform verfügbaren Daten, z. B. Formate, Herkunft, Gültigkeit, Zeitstempel usw. und dokumentieren den Lebenszyklus der Daten.

18.3 Design-Prinzipien für Data Plattformen

Nach fast zwei Jahrzehnten Erfahrung mit Plattformen lassen sich Muster für erfolgreiche Strategien der öffentlichen Plattformen ableiten und in Form von Designprinzipien auf andere Plattformen übertragen.

18.3.1 Netzwerkeffekte durch gemeinsam genutzte Datenobjekte

Charakteristisch für Plattformen sind Netzwerkeffekte, d. h. wenn die Anzahl der Teilnehmer steigt, steigt auch der Nutzen der Plattform. Bei Data Plattformen manifestiert sich dieser Netzwerkeffekt wie folgt: je mehr Unternehmensbereiche eigene Datenobjekte an die Data Plattform anschließen, desto vollständiger deckt die vorhandene Datenbasis den komplexen betriebswirtschaftlichen Datenhaushalt des Unternehmens ab. Daten, die ohne Data Plattform oft nur mit einem manuellen Aufwand und zahlreichen Medienbrüchen zusammengeführt werden können, stehen in einer einheitlichen Form zur Verfügung – idealerweise mit aufschlussreichen Metadaten für die bereichsübergreifende Anwendung zur Verfügung. Der First Mover auf der Data Plattform hat einen signifikanten Migrationsaufwand, bekommt aber neue Funktionalitäten rund um seinen Datenbestand. Wenn der Second Mover sich der Data Plattform anschließt und eigene Datenobjekte auf die Plattform bringt, können beide Unternehmensbereiche Analysen auf der kombinierten Datengrundlage durchführen (Abb. 18.2). Jeder weitere Unternehmensbereich, der weitere Datenobjekte auf die Data Plattform bringt, erhöht das Nutzungspotenzial. Darüber hinaus können aus der Verknüpfung bisheriger Datenobjekte neue selbstständige Datendomänen entstehen.

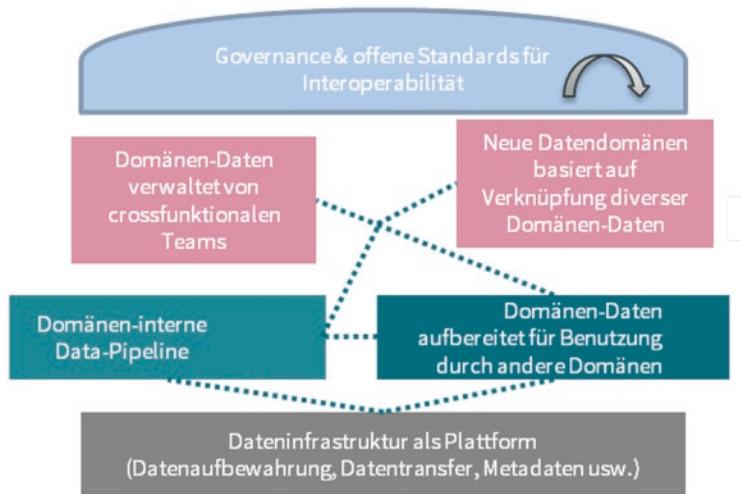


Abb. 18.2 Netzwerkeffekte einer Data Platform (Dehghani 2019)

18.3.2 Strategien für die Aktivierung von Plattformteilnehmern

Plattformteilnehmer Die Rolle des Betreibers kann sowohl eine interne Abteilung als auch ein externer Dienstleister übernehmen. Interner Betreiber ist in der Regel eine IT-Abteilung. Die Entwicklung geht dahin, dass sich zudem die Rolle des Chief Data Officers etabliert, um den Fokus nicht nur auf den technischen Betrieb, sondern auch auf die Business Ownerschaft für die Data Plattform zu stärken (New Vantage Partners 2019, S. 7). Als Multiplikator für eine Datenkultur und die damit verbundene Nutzung der Data Plattform kann ein crossfunktionales Data Lab fungieren (Fountaine et al. 2019; Scheffler und Wirths 2019). Externe Betreiber können Drittanbieter sein wie bspw. Amazon Web Services, Talend oder Snowflake. Die einzelnen Komponenten der Data Plattform können auch in einer Mischform bezogen werden, d. h. Speicherung auf einer externen Cloud-Infrastruktur und Datenaufbereitung innerhalb des Unternehmens. In der Rolle der Konsumenten sind sowohl Mitarbeiter der Fachbereiche als auch bereichsinterne und unternehmensübergreifende BI- und Analyse-Einheiten. Sie nutzen die Plattform, um ihren Bedarf für Datenauswertungen zu decken. In der Rolle des Anbieters sind diejenigen Abteilungen, welche die Daten produzieren und Datenbestände betreiben. Da die Datenerfassung zunehmend automatisiert (ohne personelle Datenpflege) und verteilt stattfindet (Datensätze zu Geschäftsvorfällen, Kundenkommunikation, Webdaten usw.), kann das ganze Unternehmen als Datenproduzent betrachtet werden. Ähnlich wie auf den öffentlichen Plattformen, sind die Rollen von Konsumenten und Anbietern auf einer Data Plattform austauschbar. Wie Fahrer/Beifahrer auf Uber oder ein Gast/Gastgeber auf AirBnB, können einzelne Bereiche im Unternehmen gleichzeitig sowohl als Anbieter und als Konsumenten auftreten.

Anwender werden eine Data Plattform nur dann aufsuchen, wenn sie die gesuchten Daten aufweisen. Aktivitäten zur Generierung von höheren Nutzerzahlen entwickelten sich in der Plattformökonomie nicht selten als intuitive, spontane Reaktion auf bereits passierte oder drohende Misserfolge: Twitter (2007) und AirBnB (2008) erreichten es mit einer offensiven Vermarktungskampagne auf einem Filmfestival, die kritische Nutzeranzahl innerhalb von wenigen Tagen auf ihre Plattform zu bringen (Parker et al. 2017, S. 2, S. 97). Eine Plattform hat nur dann eine breite Datenbasis und damit die gesuchten Daten, wenn die Nutzer sie auch mit Daten speisen und für eine Qualitäts sicherung sorgen. Es gibt verschiedene Strategien, wie sich dieses Henne-Ei-Problem lösen lässt (Parker et al. 2017, S. 96–107).

- 1 *Follow-the-rabbit-Strategie*: Ein erstes Demoprojekt zeigt die Vorteile der Infrastruktur, z. B. durch Mock-Up-Reports mit Demo-Daten um das Look-and-Feel von modernen Reporting-Werkzeugen gegenüber Excel aufzuzeigen.
- 2 *Huckepack-Strategie*: Durch die Replizierung bereits existierender, dezentral betriebener Data Marts auf die zentrale Data Plattform wird der vorhandene Nutzerstamm des Data Marts auf die Plattform geholt. Die migrierten User können dann als Promotoren für andere Nutzergruppen fungieren (*Testimonial Strategie*).
- 3 *Seeding Strategie*: Man erzeugt für mindestens eine Nutzergruppe relevante Datenobjekte als Anschub, d. h. der Plattform-Betreiber bereitet die Daten für eine bestimmte Datendomäne auf.
- 4 *Big-Bang-Strategie*: Konzentriertes Push-Marketing weckt das Interesse potenzieller Nutzer, z. B. über eine breit angelegte Unternehmenskommunikation.

18.3.3 Einfacher Zugang durch Self-Service

Hohe Nutzerzahlen für eine Data Plattform lassen sich dadurch erreichen, dass sie für die Nutzer einfach zu adaptieren ist. D. h. der Umgang mit der Plattform muss einfach zu verstehen und schnell sichtbare Ergebnisse hervorbringen. Die Nutzer benötigen einen einfachen Self-Service, bei dem sie möglichst selbstständig agieren können und bestenfalls spontan ausprobieren können, ob die Plattform für ihren Anwendungszweck nützlich ist und (in Anlehnung an Adoptionsraten von Innovationen nach Rogers 2003, S. 14–15). Ein effizientes Technologie-Onboarding der Nutzer ist also wesentlicher Erfolgsfaktor. Zugangsbarrieren, die einem schnellen Self-Service der Nutzer im Wege stehen, müssen eliminiert werden. Dies könnten bspw. fehlende standardisierte Schnittstellen, langwierige Freigabeprozesse oder Bottlenecks bei der technischen Implementierung sein. Besondere Hindernisse können durch die Einhaltung der anspruchsvollen Auflagen aus der Datenschutzgrundverordnung (DSGVO) sein. Alle Teilnehmer, d. h. sowohl Plattformbetreiber als auch Anbieter und Konsumenten sind für die Einhaltung der DSGVO verantwortlich. Es empfiehlt sich ein standardisierter Workflow für Datenschutzfolgeabschätzungen, der für den Zugriff auf die Daten obligatorisch

ist und somit ein unumgehbarer Schritt beim Design eines Datenflusses ist (Data Privacy by Design).

18.3.4 Effektives Matching durch Metadaten

Ein weiterer wesentlicher Erfolgsfaktor für eine Data Plattform ist das leichte Auffinden von Daten. Eine Studie identifiziert, dass 47 % aller Datenanalysten Probleme dabei haben, Daten zu lokalisieren. Sie zeigt ebenfalls, dass Projekte mit einer guten Daten-dokumentation sehr viel häufiger erfolgreich sind (Dresner 2019).

Zur Dokumentation eignen sich besonders Metadaten. Darunter versteht man Beschreibungen über Daten. Ein Metadaten Repository (auch unter dem Begriff Daten-katalog bekannt) enthält verschiedene Typen von Dokumentationen (vgl. Tab. 18.2).

Die Dokumentationstypen werden bei den verschiedenen Software-Anbietern mit unterschiedlichen Begrifflichkeiten verwendet werden, die Funktionalitäten und Inhalte sind aber weitestgehend gleich. Moderne Systeme zur Metadatenverwaltung können hierbei ein geeignetes Mittel sein, indem sie ein technisches Glossar durch automatisierte Scanner der vorhandenen Datenbanken erfassen. Die Data Lineage kann je nach verwendetem ETL-Tool ebenfalls automatisiert erfasst werden. Besondere Schwierigkeit besteht, bei dem fachlichen Glossar und der Datenherkunft, da sie sich auf die individuell vorhandene Business-Logik eines Unternehmens beziehen. Dies lässt sich nicht automatisiert erfassen, sondern bedarf der personellen Dokumentation und Aktualisierung. Eine weitere wichtige Funktion, die einen hohen Nutzungsgrad einer Data Plattform unterstützt, ist eine komfortable Suchfunktion, die es idealerweise ermöglicht ein Element in allen beschriebenen Dokumentationen zu finden.

18.4 Monetarisierung

Der Wertbeitrag einer Data Plattform besteht darin, dass sie eine bessere Verfügbarkeit von Daten ermöglicht. Sie bietet als „Single Point of Information“ den Datenzugriff über eine einzige Plattform anstatt über viele verteilte Datenbanken. Während es für

Tab. 18.2 Elemente eines Metadaten Repositories

Fachliches Glossar	Dokumentation der fachlichen Interpretation der Felder
Technisches Glossar	Technische Dokumentation der Datenbanken, Tabellen und Felder (Feld-längen, Datentypen etc.)
Data Dictionary	Verknüpfung von fachlichem und technischem Glossar
Date Lineage	Dokumentation der Transformationsschritte der Felder in den Daten-flüssen zur Rückverfolgung der Daten
Datenherkunft	Dokumentation der fachlichen Nutzung von Daten

die Schätzung der Kosten einer Data Plattform etablierte Methoden aus dem IT-Projektmanagement gibt, ist die Bewertung des Nutzens weniger eindeutig. Dies liegt daran, dass sich der Nutzen, insbesondere von IT-Infrastruktur, nur in wenigen Fällen direkt einem Use Case zuordnen lässt.

Nutzen lässt sich in direkt- und indirekt monetär-bewertbaren Nutzen kategorisieren. Direkter Nutzen entsteht u. a. durch Einsparungen bei Lizenzen (durch Open Source Software), geringere Kosten für den Betrieb von Data Marts oder Ablösung von Altsystemen. Indirekter Nutzen entsteht durch geringere Investitionskosten für die Umsetzung neuer Analytik-Projekte oder Return on Investment (ROI), der durch Use Cases auf der Data Plattform erzielt wird. Letzterer ist insofern ein indirekter Nutzen, als dass Data-Use-Cases (z. B. Next-Best-Offer-Engines) auch ohne eine zentrale Data Plattform umgesetzt werden könnten – allerdings mit weniger zur Verfügung stehenden Daten, mit einer schlechten Datenqualität oder verbunden mit höheren Betriebskosten.

Beim Einsatz in einem wirtschaftlich agierenden Unternehmen, gehen wir davon aus, dass sich der ROI von Data Plattformen entlang der Adoptionskurve entwickelt. In Anlehnung an die Innovations-Diffusions-Theorie (Rogers 2003, S. 11) unterstellen wir eine S-Kurve mit den Dimensionen „ROI“ und „Verfügbare Daten“. Die Verteilung der Kurve erklärt sich folgendermaßen: Zunächst ist eine Anfangsinvestition in den Aufbau der Plattform notwendig. Je mehr Daten nutzbar gemacht werden, desto stärker steigt der ROI an. Dies lässt sich durch die Netzwerkeffekte erklären. Ab einer gewissen Menge an verfügbaren Daten kommt es zu einem Break-Even des ROI und er wechselt in den positiven Bereich. Sobald ein Großteil der im Unternehmen produzierten Daten nutzbar gemacht wurde und ggf. mit externen Daten angereichert wurde, unterstellen wir eine gewisse Sättigung des ROI (vgl. Abb. 18.3).

Unter Annahme, dass es auch Datendomänen gibt, die nur bedingte Netzwerkeffekte hervorrufen, wäre auch ein alternativer Verlauf der S-Kurve denkbar (vgl. Abb. 18.4).

18.5 Zusammenfassung und Fazit

Dieser Beitrag stellt Design-Prinzipien vor, die von erfolgreichen Plattform-Unternehmen angewandt wurden und sich auf Data Plattformen übertragen lassen. Betreiber von Data Plattformen können diese Design-Prinzipien anwenden, um die Monetarisierung ihrer Plattform besser steuern und planen zu können. Wesentliche Kennzeichen einer Plattform sind, dass sich Anbieter und Konsumenten zu einer wertschöpfenden Interaktion verbinden und dabei durch einen Plattform-Betreiber unterstützt werden. Der Plattform-Betreiber stellt die Infrastruktur bereit, legt Governance-Regeln für die Plattform fest und regt die Interaktion zwischen den Teilnehmern an. Zu den besonderen Eigenschaften von Plattformen gehört, dass sie auf zweiseitigen Märkten mit Netzwerkeffekten existieren und dass sie eines Matchings bedürfen, um relevante Werteinheiten zum Austausch anzubieten. Data Plattformen sind eine spezifische Ausprägung von Plattformen, welche die Funktionen von Data Lake, Data Warehouse und Data Mart vereinen. Eine Data Platt-

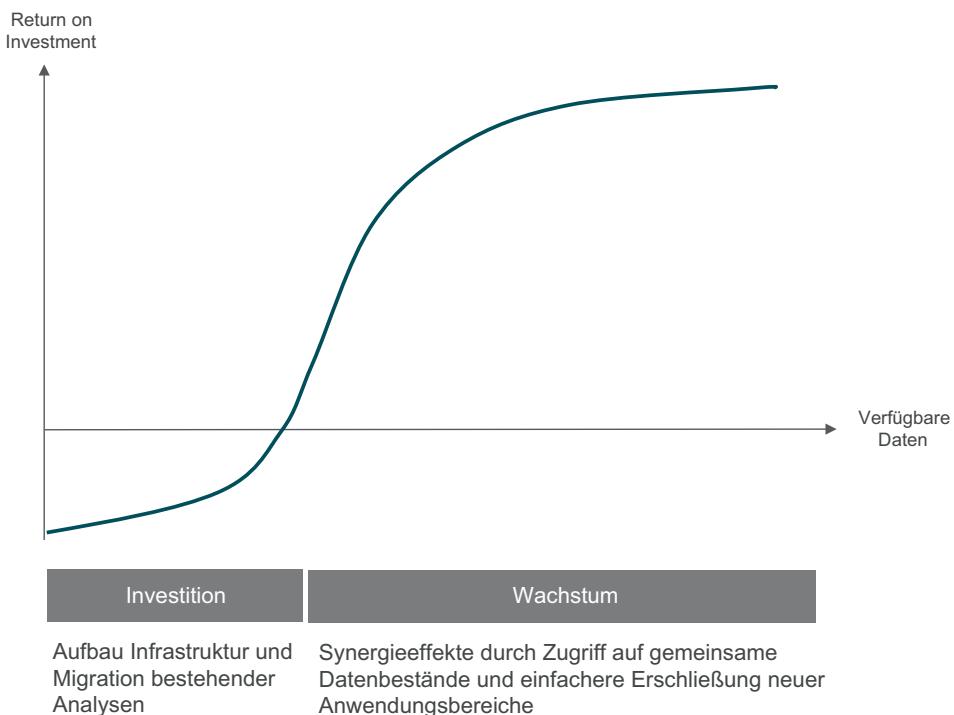


Abb. 18.3 Entwicklung des ROI entlang der Menge der verfügbaren Daten (Szenario 1)

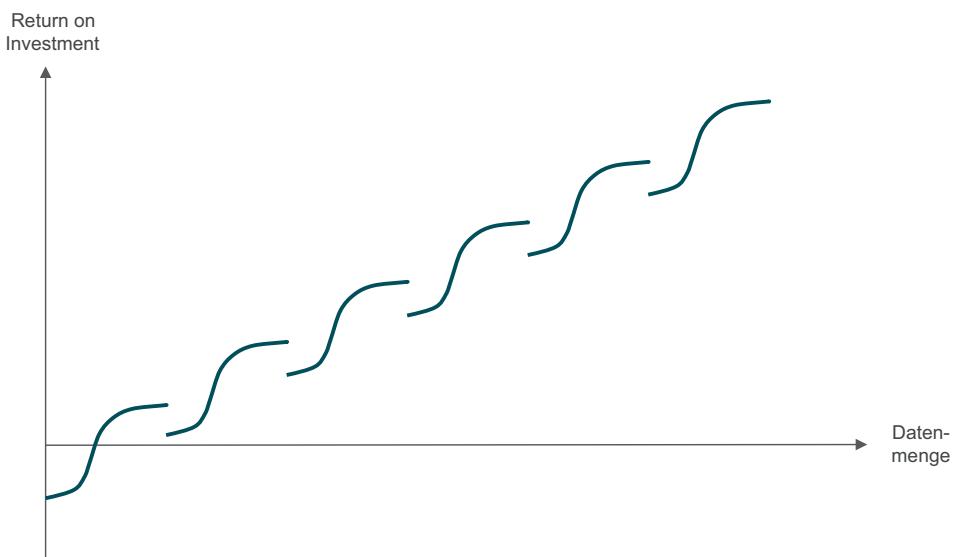


Abb. 18.4 Entwicklung des ROI entlang der Menge der verfügbaren Daten (Szenario 2)

form besteht aus Schichten mit Rohdaten, mit einem konsolidierten Unternehmensdatenmodell und mit fachbereichsspezifischen Sichten. Zudem gibt es eine Data Science Umgebung, die spezifische Anforderungen Datenverarbeitung hat (Verarbeitung großen Mengen unstrukturierter Daten) und ein Metadaten Repository für das Auffinden von Daten und Matching. Einsatzbereiche sind Planungen (z. B. Budgets oder Workforce), Reportings und Dashboards, aber komplexe Data-Science-Modelle und Künstliche Intelligenz. Es gibt vier Design-Prinzipien, die sich aus der Plattformökonomie ableiten lassen und wesentlich auf den Erfolg einer Data Plattform einzahlen. Durch gemeinsam genutzte Datenobjekte entstehen Netzwerkeffekte (Abschn. 14.3). Der Plattform-Betreiber setzt gezielt Strategien ein, um die Anzahl der aktiven Teilnehmer zu erhöhen und somit die Netzwerkeffekte zu erhöhen (Abschn. 14.3.2). Self-Services ermöglichen den Nutzern einen einfachen Zugang zur Plattform und umgehen Bottlenecks durch begrenzte Kapazitäten in IT-Teams. Dies wiederum führt zu einer höheren Anzahl an aktiven Teilnehmern und verstärkt wiederum die Netzwerkeffekte (Abschn. 14.3.2). Mit Hilfe einer umfassender Metadaten-Dokumentation können die Nutzer die gesuchten Daten einfach auffinden und interpretieren (Matching). Die Plattform wird dadurch für die Nutzer attraktiver, was zu einer höheren Adoptionsrate führt, die sich in der Anzahl der aktiven Teilnehmer widerspiegelt (Abschn. 14.3.4).

Wesentlicher Erfolgsfaktor für eine Data Plattform ist also die Anzahl der aktiven Teilnehmer. Der wirtschaftliche Erfolg zeigt sich letztendlich durch einen positiven ROI. Aus der Innovationstheorie weiß man, dass sich die Wertentwicklung neue Technologien als S-Kurve entwickelt.

Eine Studie zeigt, dass zwar viele Unternehmen Aktivitäten für eine stärkere datengetriebene Ausrichtung gestartet haben, die Potenziale daraus jedoch noch nicht monetarisieren (McKinsey 2017). Dies bestätigt, den zunächst negativen Verlauf der S-Kurve wie er in Abb. 14.3 und Abb. 14.4 dargestellt ist.

Die Investitionsstrategie der Bundesregierung zeigt, dass es gerade in wirtschaftlich herausfordernden Zeiten wichtig ist, das Zukunftsthema „Daten“ zu fördern. Das Einwerben von Budget oder größeren Investitionen ist wegen der wirtschaftlich angespannten Situation schwierig. Unternehmen sollten die Zeit nutzen, den Ist-Zustand zu konkretisieren und zumindest konzeptionelle Überlegungen für eine eigene Datenstrategie anzustellen. Eine klar formulierte Datenstrategie auf Basis eines priorisierten Portfolios an agilen Datenprodukten und Use Cases schafft die Voraussetzungen für die Entscheidungen bzgl. monetärer Investitionen und sollte deshalb als erstes angegangen werden.

Literatur

Arielly, D.: <https://twitter.com/danariely/status/287952257926971392?lang=de> (2013).

Zugegriffen: 22. Mai 2020

Bean, R.: Every company is data company <https://www.forbes.com/sites/ciocentral/2018/09/26/every-company-is-a-data-company/#31b3ef005cfc> (2018). Zugegriffen: 04. Juni 2020

- Bundesregierung: Eckpunkte einer Datenstrategie der Bundesregierung. <https://www.bundes-regierung.de/resource/blob/997532/1693626/e617eb58f3464ed13b8ded65c7d3d5a1/2019-11-18-pdf-datenstrategie-data.pdf> (2019) Zugegriffen: 10. Juni 2020
- Dehghanim, Z.: How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, <https://martinfowler.com/articles/data-monolith-to-mesh.html#DataAndSelf-servePlatformDesignConvergence> (2019). Zugegriffen: 23. Juni 2020
- Dresner Advisory Sercives: Data Catalog Study, <https://www.collibra.com/> (2019). Zugegriffen: 10.Juni 2020
- Fountain, T., McCarthy, B., Saleh, T.: Building the AI-Powered Organization. Harvard Business Review **97**(4), 62–73 (2019)
- Interbrand: Best Global Brands <https://www.interbrand.com/best-brands/best-global-brands/2019-ranking/> (2019). Zugegriffen: 22. Mai 2020
- Inmon W.H., Linstedt, D.: Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault, Morgan Kaufmann (2014)
- Jacobides, M.G., Cennamo, C., Gawer, A.: Towards a Theory of Ecosystems. Strateg. Manag. J. **39**, 2255–2276 (2018)
- McAfee, A., Brynjolfsson, E.: Big Data: The Management Revolution. Harvard Business Review **90**(10), 60–68 (2012)
- McKinsey: Fueling growth through data monetization, <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/fueling-growth-through-data-monetization> (2017). Zugegriffen: 04. Juni 2020
- McKinsey: Building a great data platform, <https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/building-a-great-data-platform> (2018). Zugegriffen: 22. Mai. 2020
- New Vantage Partners: Big Data and AI Executive Survey 2019. <https://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf> (2019). Zugegriffen: 22. Mai 2020
- Parker, G. G., Van Alstyne, M. I. W., Choudary, S. P.: Die Plattform-Revolution. Mitp, Frechen (2017)
- Rogers, E.M.: Diffusion of Innovations, 5. Aufl. Free Press, NY (2003)
- Scheffler, A., Wirths, C.: Data Innovation @ AXA Germany: Journey Towards A Data-Driven Insurer. In: Urbach, N., Röglinger, M. Digitalization Cases. Springer, Cham (2019)
- Schreieck, M., Wiesche, M., Krcmar, H.: Design and Governance of Platform Ecosystems – Key Concepts and Issues for Future Research. In: Proceedings of the European Conference on Information Systems (2016)
- Van Alstyne, M.W., Parker, G.G., Choudary, S.P.: Pipelines, Platforms, and the New Rules of Strategy. Harvard Business Review **94**(4), 54–62 (2016)
- Yueh, J.: Disrupt or Die – What the World Needs to Learn from Silicon Valley to Survive the Digital Era. Lioncrest Publishing (2017)



Akzeptanz und Nutzung von maschinellem Lernen und Analytics im Rechnungswesen und Controlling

19

Markus Eßwein, Domenica Martorana, Martina Reinersmann und Peter Chamoni

Zusammenfassung

Rechnungswesen und Controlling wirken wie mögliche Paradebeispiele für den Einsatz von maschinellem Lernen und Analytics. Große Mengen strukturierter Daten werden täglich erzeugt, verdichtet und zur Entscheidungsfindung aufbereitet. Dennoch scheinen die Akzeptanz und Nutzung dieser beiden Technologien weit hinter den Prognosen der letzten Jahre zurückzustehen. Auch fehlt es an Treibermodellen, die beim Verständnis dieser Diskrepanz unterstützen könnten. Vor diesem Hintergrund stellt der folgende Beitrag die Ergebnisse einer Studie zur Akzeptanz und Nutzung von maschinellem Lernen und Analytics vor. Die Ergebnisse zeigen, dass für große Teile der befragten Führungskräfte die Aufgabencharakteristika den stärksten Einfluss auf die eigentliche Nutzung haben. Der Wunsch, die Technologien zunächst parallel zu bestehenden Lösungen zu benutzen, ist ein Indikator für eine geringere

M. Eßwein (✉)

Maxdorf, Deutschland

E-Mail: markus.esswein@uni-due.de

D. Martorana

Düsseldorf, Deutschland

E-Mail: dmartor@gwdg.de

M. Reinersmann · P. Chamoni

Wirtschaftsinformatik, Universität Duisburg-Essen, Duisburg, Deutschland

E-Mail: martina.reinersmann@uni-due.de

P. Chamoni

E-Mail: peter.chamoni@uni-due.de

Nutzung. Darüber hinaus wird die Vermutung bestätigt, dass Führungskräfte in höheren Positionen tendenziell weniger vertraut mit Methoden aus dem Bereich des maschinellen Lernens sind. Der Beitrag schließt mit vier Handlungsempfehlungen zu den Stichworten Intelligenz, Investitionen, Implementierung und Incentivierung.

19.1 Eine Herausforderung für die Finanzfunktion

Bisher nie dagewesene technologiegetriebene Veränderungen sorgen gleichermaßen für Chancen und Risiken über alle Wirtschafts- und Gesellschaftsbereiche hinweg. Dieser Wandel betrifft Unternehmen als Ganzes, aber auch die einzelnen Unternehmensbereiche und zwingt sie, sich anzupassen. Einer dieser Bereiche, die Finanzfunktion und damit das zahlengeriebene Gewissen des Unternehmens, läuft Gefahr, den Anschluss zu verpassen (Halper 2014). Obwohl sie die Finanzfunktion einst zum Vorreiter bei der Einführung neuer Technologien machten, stehen Führungskräfte aktuell vor der Herausforderung, die schnell voranschreitenden Entwicklungen in den Bereichen des maschinellen Lernens („Machine Learning“), der fortschrittlichen Analyseverfahren („Analytics“) und der Datenplattformen („Data Lake“) für die eigene Finanzfunktion nutzbar zu machen. Dabei gilt es jedoch auch und vor allem, die einzelnen Mitarbeiter nicht zu verlieren.

Nimmt man maschinelles Lernen und Analytics (ML&A) als Beispiele, gibt es einige qualitative Forschungsbeiträge (siehe u. a. Brands und Holtzblatt 2015), die aufzeigen, dass Unternehmen das Potenzial dieser Technologien nicht annähernd ausschöpfen. Auf der einen Seite gibt es eine zunehmende Anzahl an Anbietern von Softwarelösungen für spezielle Probleme aus dem Bereich des maschinellen Lernens, auf der anderen Seite stehen Unternehmen, die noch nicht einmal wissen, welche Anwendungsfälle sie auswählen und wie sie diese Projekte angehen sollten. Es fehlt sowohl in der akademischen als auch der praxisnahen Literatur an konkreten Modellen und Fallstudien, die Unternehmen bei der Auswahl passender Anwendungsfälle oder beim Aufbau einer datengetriebenen Unternehmenskultur („Analytics Culture“) unterstützen. Außerdem fehlt es an einem Verständnis der Wirkungszusammenhänge, um die bisher verhaltene Nutzung zu steigern.

Vor diesem Hintergrund stellen sich zwei Leitfragen:

- Wie ist der aktuelle Stand der Akzeptanz und Nutzung von maschinellem Lernen und Analytics?
- Welche Treiber sind die relevantesten für einen erfolgreichen Einsatz?

19.2 Nutzerakzeptanzforschung zu maschinellem Lernen

Einer der Grundpfeiler des aktuellen technologischen Fortschritts ist maschinelles Lernen, welches Maschinen befähigt, eine Aufgabe basierend auf aus bekannten Beispielen abgeleiteten Regeln auszuführen (Simon 1983). Algorithmen aus dem Bereich des maschinellen Lernens können in drei Bereiche unterteilt werden: überwachtes, unüberwachtes und bestärkendes Lernen (Kacprzyk und Pedrycz 2015). Beim überwachten Lernen nutzt der Algorithmus zur Bestimmung der Parameter Daten, die bereits eine definierte Zielvariable haben – zum Beispiel die Umsatzzahlen der vergangenen Jahre. Im Gegensatz dazu stehen beim unüberwachten Lernen nur die Rohdaten zur Verfügung, auf Basis derer zum Beispiel eine Gruppierung stattfinden soll. Bestärkendes Lernen ist ein iterativer Prozess, bei dem in regelmäßigen Abständen, die von der Maschine unüberwacht vorgenommenen Modellanpassungen überprüft und bewertet werden. Diese Bewertung führt dann gegebenenfalls zu weiteren Modellanpassungen. Für weiterführende Grundlagen zu maschinellem Lernen siehe unter anderem (Kacprzyk und Pedrycz 2015).

Mit Wurzeln im Feld der Statistik ist Analytics ein Überbegriff für den Erkenntnisgewinn aus mehr oder weniger großen Datensätzen (Agarwal und Dhar 2014). Seit der Jahrtausendwende wird der Begriff Analytics verwendet, um Informationssysteme zu beschreiben, die auf fortschrittliche Analyseverfahren aus den Bereichen der Statistik und des maschinellen Lernens zurückgreifen mit dem Ziel, Wirkungszusammenhänge herauszustellen oder Prognosen zu erstellen (Baesens et al. 2016). Mit der stark gestiegenen Rechenleistung heutiger IT-Architekturen bietet sich sowohl für Praxisanwender (Lavalle et. al. 2011) als auch Wissenschaftler (Agarwal und Dhar 2014) zunehmend die Möglichkeit, die schiere Masse an Transaktionsdaten aus Enterprise-Resource-Planning (ERP)-Systemen, Sensoren, sozialen Netzwerken oder mobilen Endgeräten zu verarbeiten und zu analysieren. Die drei Hauptfelder der Analyse von solchen Massendaten sind dabei deskriptiv (Beschreibung von Wirkungszusammenhängen), prädiktiv (Prognose von Entwicklungen) und präskriptiv (Handlungsempfehlungen) (Delen und Demirkan 2013).

Selbst die vielversprechendste Technologie versagt jedoch, wenn Anwender deren Vorteile nicht erkennen und nutzen. Aus diesem Grund entwickelte sich bereits in den 1970er Jahren nach einer Welle von gescheiterten Technologieprojekten ein Forschungszweig zur Nutzerakzeptanz von Informationssystemen (Chittur 2009). Unter den dabei entstandenen Modellen ist das Technology-Acceptance-Model (TAM) von Davis (1989) eines der am weitesten verbreiteten. Das Modell postuliert, dass die Einstellung eines Benutzers zur Nutzung eines Informationssystems maßgeblich von der empfundenen Benutzerfreundlichkeit („Perceived Ease of Use“) und der empfundenen Nützlichkeit

keit („Perceived Usefulness“) des Systems abhängt. Über die Jahre wurden diverse Erweiterungen am ursprünglichen Modell vorgeschlagen, teils unter der Kritik, dass diese keinen klaren Beitrag zur Erklärung der eigentlichen Nutzung leisten würden. Insgesamt bleiben die beiden Konstrukte des Ausgangsmodells aber auch heute noch valide (Benbasat und Barki 2007).

Mit einer anderen Perspektive näherten sich Goodhue und Thompson (1995) der Nutzerakzeptanz in ihrem Task-Technology-Fit-Modell (TTF). Zwei Konstrukte, die Aufgabencharakteristika (z. B. Routine oder nicht) und die Technologiecharakteristika (z. B. Zuverlässigkeit) werden in diesem Modell – mediert durch den Task-Technology-Fit – zur Erklärung der Leistungsfähigkeit eines Nutzers herangezogen. Später wurde das Modell um die individuellen Charakteristika (z. B. Netzwerkanbindung) erweitert.

Keines der beiden Modelle ist jedoch unumstritten, weshalb Dishaw und Strong (1999) vorschlugen, diese zu kombinieren, um die jeweiligen Schwachstellen zu kompensieren. Ebenso wurde bisher keines der Modelle auf die Nutzung von maschinellem Lernen und Analytics angewendet. Aus diesen Gründen wurde für den vorliegenden Beitrag eine Kombination der beiden Modelle als Grundlage für eine empirische Untersuchung zur aktuellen Nutzung und Akzeptanz von maschinellem Lernen und Analytics gewählt.

19.3 Befragung von Führungskräften

19.3.1 Strukturgleichungsmodell

Strukturgleichungsmodelle sind statistische Modelle, die zur Schätzung von Korrelationen zwischen verschiedenen Konstrukten herangezogen werden, wenn letztere nicht direkt messbar sind. Für diesen Beitrag wurden daher die neun Konstrukte des Strukturmodells (latente Variablen, erste Spalte in Abb. 19.1) mithilfe der insgesamt 15 Indikatoren des Messmodells (beobachtete Variablen, zweite Spalte in Abb. 19.1) geschätzt. Letztere wurden aus akademischer und „grauer“ Literatur zu maschinellem Lernen und Analytics abgeleitet. Für weitere Informationen zu Strukturgleichungsmodellen, den nötigen Voraussetzungen und Auswertungsmethoden sei an dieser Stelle auf Hair Jr. et al. (2016) verwiesen.

19.3.2 Umfrage

Die Datenerhebung fand unter den Partnerunternehmen des Schmalenbach Arbeitskreises „Digital Finance“ im Zeitraum von Januar bis März 2019 statt. Dabei wurden

Konstrukt	Indikator		Erfassung
Aufgaben-Charakteristika	Anwendungsbereich	Umsätze Kosten Working capital Risiko	Likert Häufigkeit
	Aufgaben	Planung Budgetierung Prognose	Likert Häufigkeit
Technologie-Charakteristika	Datentypen	Strukturiert, intern Unstrukturiert, intern Strukturiert, extern Unstrukturiert, extern	Likert Häufigkeit
	Methoden	Vorhersage Klassifizierung Zeitreihenprognose Assoziation Segmentierung	Likert Häufigkeit
Individuelle Charakteristika	Demographie	Mitarbeiter im Unternehmen Position Abteilung Mitarbeiter im Finanzbereich Rolle	Multiple Choice
Aufgaben-Technologie-Fit	Aufgabe * Technologie	(TTF)	Orthogonale Interaktion
Empfundene Benutzer-freundlichkeit	Familiarity	Deskriptive Statistik Lineare Regression Zeitreihenmodelle Maschinelles Lernen	Likert Ausmaß
Empfundene Nützlichkeit	Grund der Nutzung	Einen Wettbewerbsvorteil Erlangen Unterstützung der Unternehmensziele Leistungssteigerung Bessere Entscheidungen Bessere Entscheidungsprozesse Wissensgenerierung Wert aus Daten gewinnen	Likert Zustimmung
	Grad der Unterstützung	Den Mensch nicht übertreffen Den Mensch bei der Arbeit unterstützen Die meiste Arbeit des Menschen erledigen Menschen bei der Arbeit schneller werden lassen Menschen bei der Arbeit genauer werden lassen	Likert Zustimmung
	Schritte der Entscheidung	Identifikation von Problemen Entwicklung von Lösungen Bewertung von Lösungsalternativen	Likert Ausmaß
Intention zu nutzen	Eigene Arbeit	Bleibe beim Vertrauten Nutze ergänzend Übertrage alles Mögliche	Likert Zustimmung
Tatsächliche Nutzung	Aktuelle Nutzung	Einzelne Prototypen Parallele Nutzung Volumfähiglich	Likert Häufigkeit
	Anwendungsbereich	Rechnungswesen Controlling Treasury	Likert Häufigkeit

Abb. 19.1 Konstrukte und Indikatoren des Modells

Tab. 19.1 Informationen zu den Teilnehmern an der Umfrage

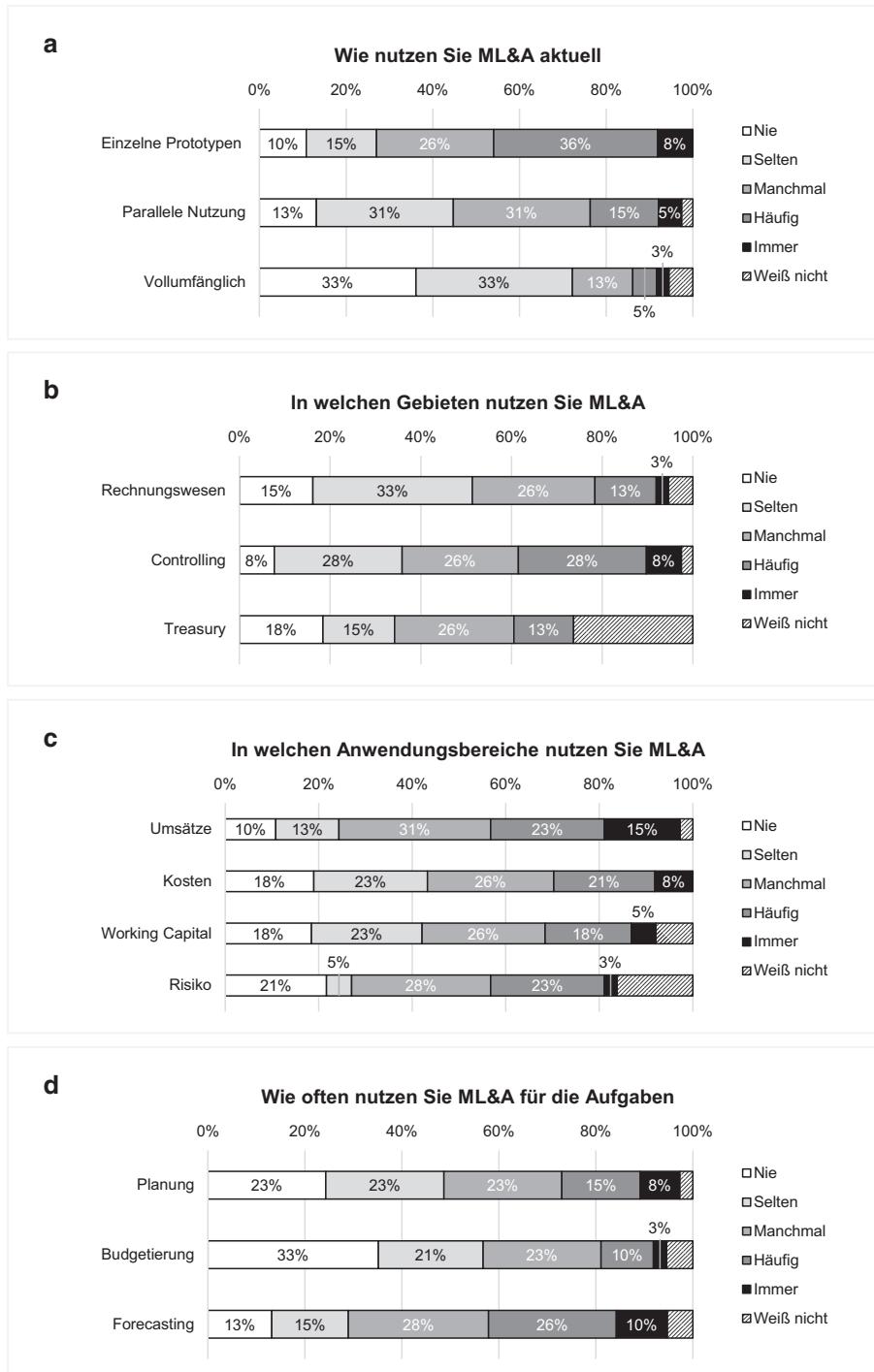
Position	No	%	Abteilung	No	%	Rolle	No	%
Obere Führungs-ebene	5	13	Finance	26	67	Endbenutzer	20	51
Mittlere Führungs-ebene	12	31	Business	4	10	Power-Nutzer	12	31
Untere Führungs-ebene	11	28	IT	5	13	Entwickler	4	10
Ohne Führungsverantwortung	11	28	Other	4	10	Other	3	8
Gesamt	39	100		39	100		39	100
Mitarbeiter im Unternehmen	No	%	Mitarbeiter in der Abteilung	No	%			
<250	0	0	<25	9	23			
250–10,000	3	8	25–1,000	26	67			
10,001–100,000	23	59	>1,000	4	10			
>100,000	13	33						
Gesamt	39	100		39	100			

den insgesamt 39 Teilnehmern 46 Einzelfragen zu den neun Konstrukten des Modells gestellt, welche sie auf einer Likert-Skala mit fünf Skalenelementen beantworteten. Die Beschreibung der Skalenelemente variierte dabei je nach Indikator (siehe letzte Spalte in Abb. 19.1). Beispielsweise wurde für den Anwendungsbereich die Häufigkeit der Nutzung in den Teilbereichen Rechnungswesen, Controlling und Treasury auf einer Skala [1] „nie“, [2] „selten“, [3] „manchmal“, [4] „häufig“, [5] „immer“ abgefragt. Nähere Informationen zu den Hintergründen der Befragten können Tab. 19.1 entnommen werden. Der Fokus auf große und sehr große Unternehmen spiegelt die Zusammensetzung des Arbeitskreises wider.

19.4 Aktuelle Nutzung und Treiber

19.4.1 Ergebnisse der Befragung

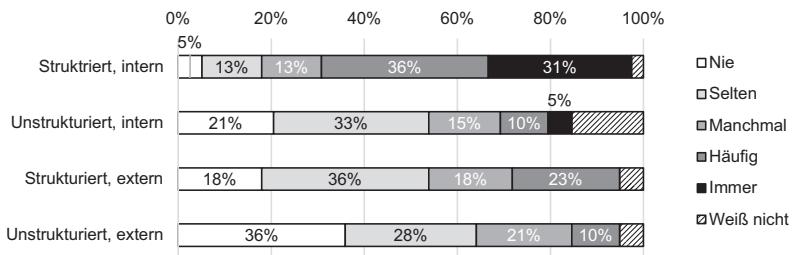
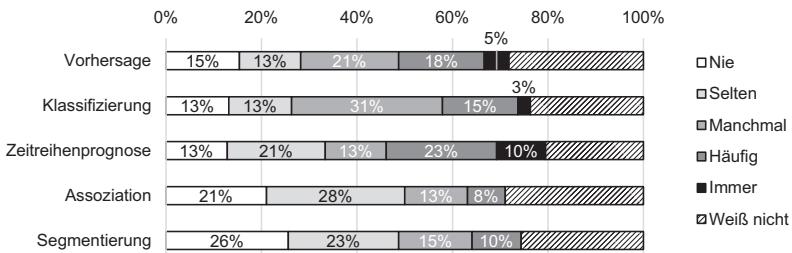
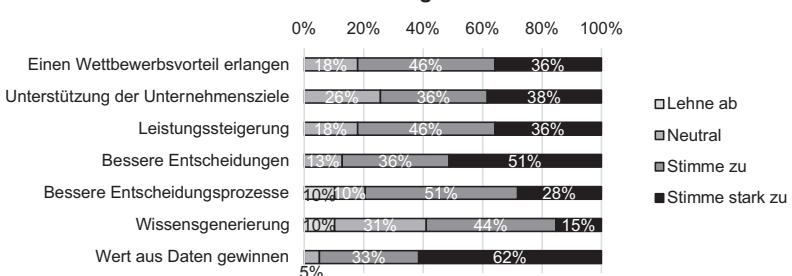
Vier Fragen behandelten die Intensität der Nutzung von ML&A. Die Ergebnisse zeigen eindeutig, dass aktuell die Nutzung in Form von Prototypen am weitesten verbreitet ist. 70 % der Befragten gaben an, diese manchmal, häufig oder immer zu nutzen,

**Abb. 19.2** Tatsächliche Nutzung, Anwendungsgebiete, Anwendungsbereiche und Aufgaben

wohingegen nur 51 % bzw. 21 % angaben, ML&A parallel zu bestehenden Modellen oder vollumfänglich für einzelne Anwendungen zu nutzen (vgl. Abb. 19.2a). Am anderen Ende des Spektrums wiesen nur 10 % der Befragten darauf hin, dass sie bislang noch keine Erfahrung in Form von Prototypen gesammelt haben, wohingegen 33 % noch keine vollumfängliche Nutzung vorweisen können. Schaut man auf die Verteilung dieser Nutzung auf die einzelnen Anwendungsgebiete, zeigt sich Controlling als Vorreiter (vgl. Abb. 19.2b). Vor dem Hintergrund der recht gleichmäßigen Verteilung über die Anwendungsbereiche Umsätze, Kosten, Working Capital und Risiko – alle im Bereich von 50 % „manchmal“ oder „häufig“ (vgl. Abb. 19.2c) – liegt das Controlling als Ansatzpunkt nahe. Abschließend gaben die Befragten in diesem Block an, dass sie ML&A am häufigsten für die Aufgabe der Prognoseerstellung (Forecasting) nutzen, am seltensten für die Budgetierung (vgl. Abb. 19.2d).

Der nächste Fragenblock behandelte die verwendeten Daten und Algorithmen, die unterstützten Schritte bei der Entscheidungsfindung sowie die Motivation zur Nutzung von ML&A (vgl. Abb. 19.3). Die Ergebnisse zu den verarbeiteten Daten zeigen sehr deutlich, dass hauptsächlich strukturierte, interne Daten für ML&A genutzt werden. 67 % der befragten Teilnehmer nutzen diese häufig oder immer. Demgegenüber stehen strukturierte oder unstrukturierte, externe Daten, die von 50 % der Befragten nur selten oder nie genutzt werden. Überwachte Lernverfahren finden etwas häufiger Anwendung, alle drei Aufgabenarten (Prognose, Klassifizierung, Zeitreihenprognose) werden ähnlich häufig genutzt, – von einem geringen Anteil der Befragten – wird die Zeitreihenprognose sogar immer genutzt. Die empfundene Nützlichkeit bei verschiedenen Schritten der Entscheidungsfindung ist ebenfalls recht ausgeglichen. Jeweils etwa 30 % der Befragten empfanden sie als hoch oder sehr hoch und 10 % als nicht vorhanden. Die Ergebnisse zur Motivation der ML&A Nutzung zeigen, dass es in der Tat eine große Vielfalt an Gründen zur Nutzung gibt. Zum Beispiel gaben gut 40 % der Befragten an, dass sie einen Wettbewerbsvorteil oder eine gesteigerte Leistung erwarten. 51 % und 62 % stimmten den Aussagen, dass ML&A zu besseren Entscheidungen respektive zur Gewinnung von Wert aus Daten führen, vollumfänglich zu. Verhaltener war die Zustimmung für die Generierung von Wissen und für bessere Entscheidungsprozesse. Beides wurde von 10 % der Befragten abgelehnt.

Der letzte Fragenblock themisierte die Vertrautheit der Nutzer im Umgang mit verschiedenen Algorithmen, ihre Intention zur Verwendung von ML&A für die eigene Arbeit und suchte nach den Initiatoren der Umsetzung von ML&A in den Unternehmen der Befragten (vgl. Abb. 19.4). Die Vertrautheit mit deskriptiven Modellen ist mit 28 % „sehr“ und 36 % „äußerst“ klar am größten. Es folgen mit abnehmenden Werten lineare Regressionsmodelle, Zeitreihenmodelle und Algorithmen des maschinellen Lernens. Zu den letzten beiden Algorithmenarten gaben mehr als die Hälfte der Führungskräfte an, gar nicht oder nur in geringem Maße vertraut zu sein. Diese teils fehlende Vertraut-

a**Welche Daten nutzen Sie Basis für ML&A****b****Welche ML&A Algorithmen nutzen Sie****c****Wie sehr helfen ML&A für die folgenden Schritte der Entscheidungsfindung****d****Welchen Vorteil bringen Ihnen ML&A****Abb. 19.3** Daten, Algorithmen, Entscheidungsschritte und Vorteile

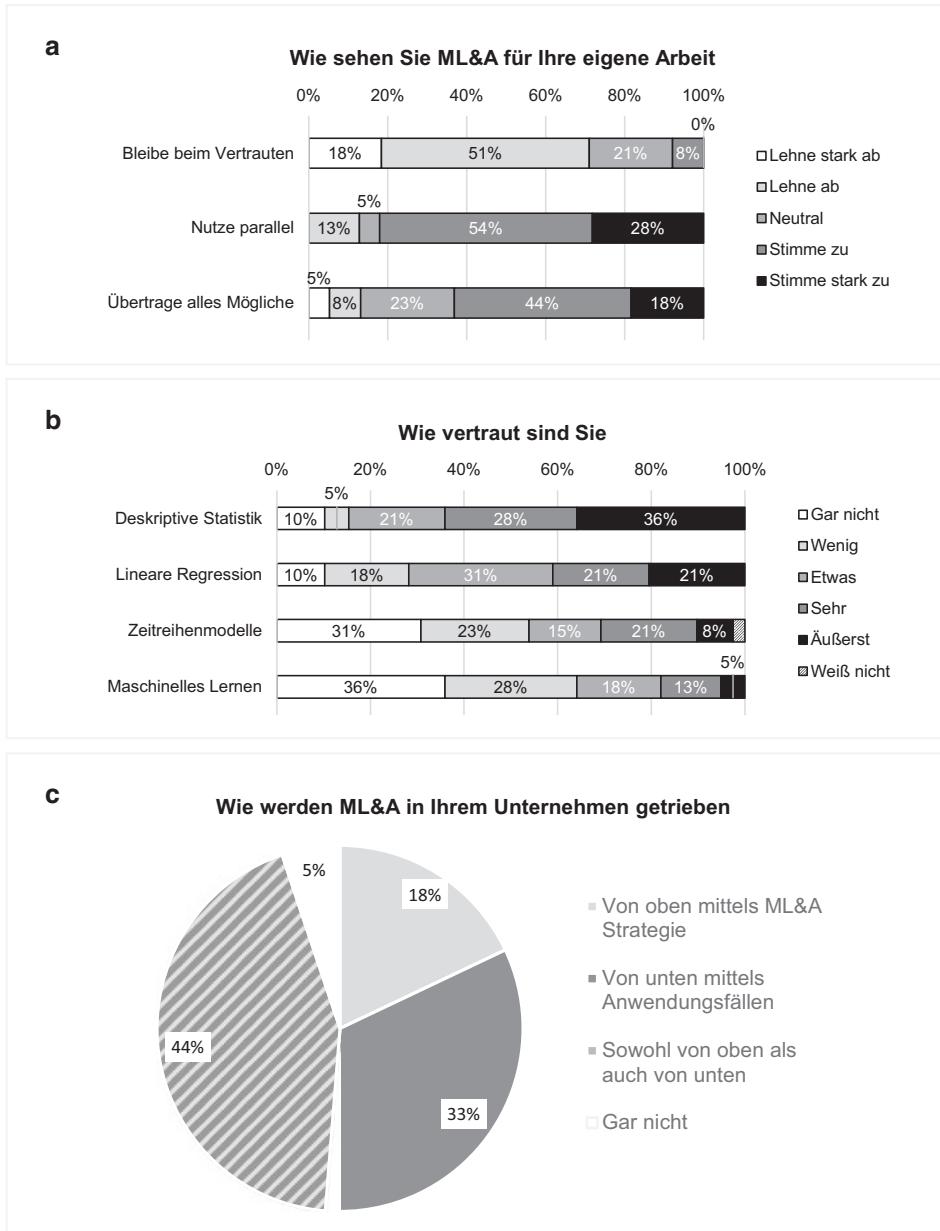


Abb. 19.4 Einstellung, Vertrautheit und Treiber

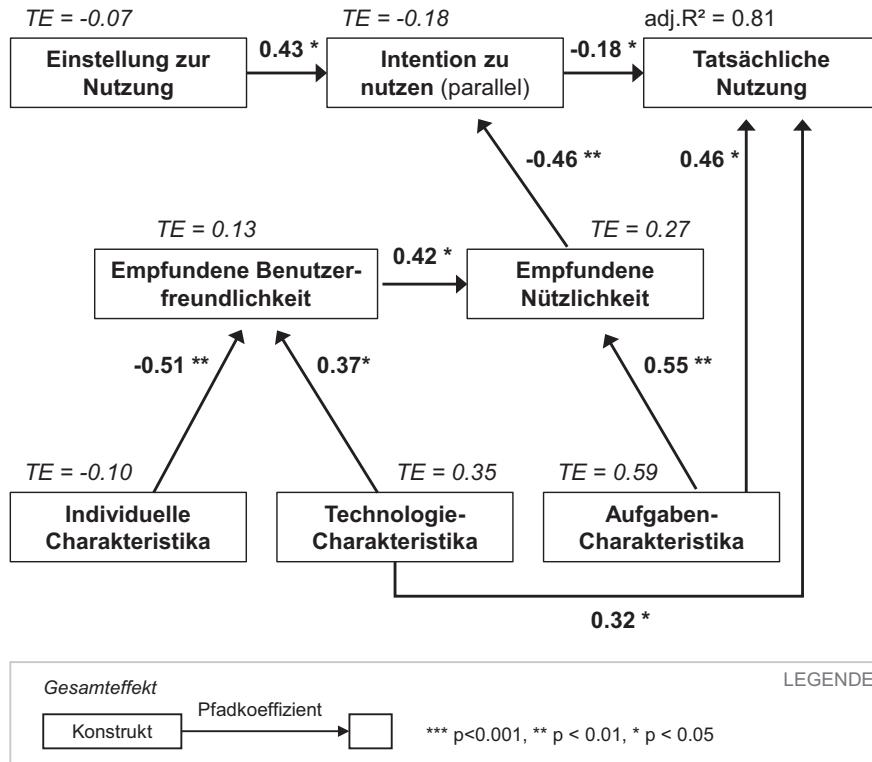


Abb. 19.5 Geschätzte Parameter des Strukturgleichungsmodells

heit mit den Algorithmen erklärt das Antwortverhalten der Teilnehmenden zur Intention, ML&A zu nutzen. 82 % der Befragten gaben an, die Maschine in jedem Fall als Parallel-lösung zu nutzen, während 62 % der Aussage zustimmten, dass sie alles Mögliche auf die Maschine übertragen werden. Nur ein geringer Anteil von 8 % bevorzugt den Verbleib bei den bestehenden Methoden. Ein abschließender Blick auf die Initiatoren einer ML&A-Umsetzung zeigt, dass die Mehrheit der Unternehmen sowohl eine Analytics-Strategie verfolgt als auch über Anwendungsfälle mit bisher unzureichenden Lösungen die Einführung vorantreibt. Insgesamt ist jedoch die Anzahl der Unternehmen, die ein strategisches Vorgehen präferieren, geringer im Vergleich zu den Unternehmen, die einen experimentellen Ansatz mit Anwendungsfällen verfolgen.

19.4.2 Treibermodell zur Nutzung und Akzeptanz

Die Schätzung des Modells erfolgte mithilfe der Statistiksoftware R und dem von Ray und Danks (2019) erstellten Paket SEMinR. Abb. 19.5 zeigt das finale Modell, wobei

Tab. 19.2 Gesamteffekte der Konstrukte

Konstrukt	Effekt
Einstellung zur Nutzung	-0.07
Individuelle Charakteristika	-0.10
Intention zur (parallelen) Nutzung	-0.18
Empfundene Benutzerfreundlichkeit	0.13
Empfundene Nützlichkeit	0.27
Technologiecharakteristika	0.35
Aufgabencharakteristika	0.59

Rechtecke die Konstrukte und Pfeile die Wirkungspfade darstellen. Die über den Pfeilen angeführten Pfadkoeffizienten können als standardisierte Pfadkoeffizienten einer linearen Regression angesehen werden und erlauben entsprechend Aussagen wie „Steigt der ermittelte Wert für die Technologiecharakteristika um eine Standardabweichung, so steigt auch der Wert der tatsächlichen Nutzung um 0,32 Standardabweichungen.“ Umgekehrt führt die Erhöhung des Wertes der individuellen Charakteristika um eine Standardabweichung zu einer Reduzierung des Wertes für die empfundene Benutzerfreundlichkeit um 0,51 Standardabweichungen. Dieser Effekt ist auf dem 1 % Niveau signifikant. Wie aus Abb. 19.5 zu erkennen ist, tragen alle Konstrukte zur Erklärung des eigentlichen Nutzens in direkter oder indirekter Weise bei – von allen führt ein statistisch signifikanter Pfad zur Zielvariable des Modells.

Der aus den Pfadkoeffizienten bestimmbare Gesamteffekt eines Konstrukts (siehe Tab. 19.2) zeigt, dass Aufgabencharakteristika aktuell die mit Abstand stärksten Treiber für die tatsächliche Nutzung von ML&A sind. Dies ist insbesondere vor dem Hintergrund der aktuell häufig noch im Prototypenstatus befindlichen ML&A-Lösungen einleuchtend. Personen, die keine Tätigkeiten ausüben, welche sich mittels ML&A unterstützen lassen, kommen aktuell auch nicht in Berührung mit diesen Lösungen. Auch der recht geringe Beitrag der empfundenen Benutzerfreundlichkeit lässt sich damit erklären, dass zunächst der Anwendungsfall im Vordergrund steht und nicht die möglichst benutzerfreundliche Darstellung. Eine weitere Schlussfolgerung lässt auch der leicht negative Gesamteffekt der individuellen Charakteristika zu, hier unter anderem abgeleitet aus der Hierarchiestufe der Führungskraft. Steigt eine Führungskraft höher in der Hierarchie, nimmt die Nutzung von ML&A geringfügig ab. Abschließend soll an dieser Stelle auch auf die Intention zur parallelen Nutzung eingegangen werden. Der negative Koeffizient lässt den Schluss zu, dass das Ziel, die Maschine (nur) als Ergänzung zu bestehenden Lösungen zu nutzen, die möglichen Anwendungsfälle einschränkt. Die Entscheidung zur vollumfänglichen Nutzung der Maschine erfordert hin-

gegen eine entschiedenere Herangehensweise, was sich am Ende jedoch auch positiv auf die tatsächliche Nutzung auswirkt.

19.5 Handlungsempfehlungen und Ausblick

Die zuvor beschriebenen Ergebnisse zeigen, dass die Umsetzung der perfekten Symbiose aus Mensch und Maschine noch weit in der Zukunft liegt. Dennoch gibt es bereits jetzt einige Stimmen hochrangiger Wissenschaftler und Unternehmensvertreter, die eine frühzeitige Besinnung auf gemeinsame Werte und Regularien fordern (FLI 2015). Bis die Prognosen eintreten und maschinelle Intelligenz die menschliche überschreitet, sollten Unternehmen nicht auf Sicherheit gehen, sondern den Wandel mitgestalten.

Dabei sei an dieser Stelle darauf hingewiesen, dass die Ergebnisse der Studie nur die Einschätzung zur Akzeptanz und Nutzung von ML&A eines kleinen Kreises, vorrangig großer Unternehmen, abbilden und damit die Übertragbarkeit auf andere Bereiche möglicherweise eingeschränkt ist. Insgesamt lassen sich jedoch Handlungsempfehlungen ableiten, die im Folgenden unter vier Stichworten zusammengefasst werden sollen.

Intelligenz: Menschliche Arbeitskräfte werden auf absehbare Zeit nicht einfach von Maschinen ersetzt, zumindest in vielen Bereichen nicht. Aktuell beschränkt sich die Nutzung noch stark auf Prototypen. Eine parallele Nutzung von ML&A ist ein guter Startpunkt, um das Vertrauen der Mitarbeiter in die Korrektheit der Ergebnisse der Maschine zu stärken. Auch kann dadurch der Drang, Neues auszuprobieren und Dinge zu verbessern, gefördert werden. Auf längere Sicht wird eine vollumfängliche Nutzung jedoch unausweichlich für den Erhalt der Wettbewerbsfähigkeit sein. Unternehmen sollten möglichst früh auf dieses Szenario hinarbeiten.

Investitionen: Die Wahl der richtigen Projekte bleibt weiterhin eine große Herausforderung. Aus den verschiedenen Teilbereichen des maschinellen Lernens und der Analytics ergeben sich viele Möglichkeiten, die jedoch nicht allein um der neuen Technologien Willen betrachtet werden sollten. Vielmehr sollte stets der konkrete Anwendungsfall im Vordergrund stehen – die Aufgabencharakteristika sind die aktuell wichtigsten Treiber für eine tatsächliche Nutzung. Es empfiehlt sich also, möglichst früh die Prozessverantwortlichen miteinzubeziehen.

Implementierung: Es zeichnet sich zurzeit eine Kehrtwende zurück zum „Insourcing“ ab. Der Wissensaufbau im eigenen Unternehmen wird als Wettbewerbsvorteil gesehen und verringert die Abhängigkeit von Dritten. Auch in der zahlengetriebenen Finanzfunktion bietet es sich an, Mitarbeitern das Erlernen von Fertigkeiten aus dem Bereich ML&A möglichst einfach zu machen. Aktuell ist die Vertrautheit mit

Algorithmen des maschinellen Lernens noch recht gering, was einer der möglichen Gründe für die bisherige Zurückhaltung ist. Dem sollte mittels interaktiven, bereichsübergreifenden Trainingsprogrammen für Mitarbeiter aller Hierarchiestufen und Tätigkeitsfelder entgegengewirkt werden.

Incentivierung: Um das volle Potenzial ausschöpfen zu können, sollten auch die Anreizsysteme überdacht werden. Oftmals ist der (finanzielle) Gesamtunternehmenserfolg ein wesentlicher Bestandteil der Bonusvergütung einer Führungskraft. Bis zu einer gewissen Hierarchiestufe lässt sich dieser jedoch kaum beeinflussen. Vielmehr sollte die Leistung der Führungskraft an der von ihr zu verantwortenden Komponente des Unternehmenserfolgs gemessen werden. Mithilfe statistischer Verfahren lassen sich Prognosen zum Beispiel in Trend-, Saison- und Restkomponenten aufspalten. Insbesondere letztere liegen im Rahmen der Möglichkeiten einer Führungskraft. Schlussendlich kann diese Anpassung zu einer stärkeren Identifikation mit dem Unternehmen und einer faireren Kompensation führen.

Literatur

- Agarwal, R., Dhar, V.: Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research* **25**(3), 443–448 (2014)
- Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., Zhao, J. L.: Transformational Issues of Big Data And Analytics in Networked Business. *MIS Quarterly* **40**(4), 807–818 (2016)
- Benbasat, I., Barki, H.: Quo vadis TAM? *J. Assoc. Inf. Sys.* **8**(4), 211–218 (2007)
- Brands, K., Holtzblatt, M.: Business Analytics: Transforming the Role of Management Accountants. *Management Accounting Quarterly* **16**(3), 1–12 (2015)
- Chittur, M.: Overview of the Technology Acceptance Model: Origins, Developments and Future Directions. *Working Papers on Information Systems* **9**(37), 9–37 (2009)
- Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. *Manage. Sci.* **35**(8), 982–1003 (1989)
- Delen, D., Demirkan, H. : Data, information and analytics as services. *Decis. Support Syst.* **55**(1), 359–363 (2013)
- Dishaw, M.T., Strong, D.M.: Extending the technology acceptance model with task–technology fit constructs. *Inf. Manage.* **36**(1), 9–21 (1999)
- FLI – Future of Life:. Research Priorities for Robust and Benificial Artificial Intelligence. <https://futureoflife.org/ai-open-letter/?cn-reloaded=1> (2015)
- Goodhue, D.L., Thompson, R.L.: Task-technology fit and individual performance. *MIS Quarterly* **19**(2), 213–236 (1995)
- Hair, J.F., Jr., Hult, G.T.M., Ringle, C., Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). Sage Publications, Thousand Oaks (2016)
- Halper, F.: Predictive analytics for business advantage. TDWI Research. <https://tdwi.org/research/2013/12/best-practices-report-predictive-analytics-for-business-advantage.aspx> (2014)

- Kacprzyk, J., Pedrycz, W.: Springer handbook of computational intelligence. Springer, Dordrecht (2015)
- Lavalle, S., Lesser, E., Shockley, R., Hopkins, M., Kruschwitz, N.: Special report: Analytics and the new path to value. MIT Sloan Manag. Rev. **52**(2), 22–32 (2011)
- Ray, S., Danks, N.: SEMinR. <https://cran.r-project.org/web/packages/seminr/vignettes/SEMinR.html> (2019)
- Simon, H.A.: Why should Machines Learn? In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (Hrsg.) Machine Learning, S. 25–37. Morgan Kaufmann, San Francisco (1983)



Durch Daten zu neuen Geschäftsmodeellen und Prozessoptimierungen – im Kontext von Car-Sharing

20

Im Praxisbeispiel erfahren Sie, wie Sie es besser *nicht* machen sollten

Eva Schoetzau

Zusammenfassung

Durch die optimale Nutzung von Daten können neue Geschäftsmodelle entwickelt werden. Jedoch sind dafür nicht jegliche Arten von Daten geeignet, sondern vornehmlich Daten, die die Interessen und Bedürfnisse von Menschen beziehungsweise potenziellen Kunden widerspiegeln und dadurch für die Neuproduktentwicklung oder die Optimierung bestehender Angebote genutzt werden können. Genauso können durch neue Produkte oder Angebote neue und gegebenenfalls vielfältigere Daten eingeholt und analysiert werden, wodurch Optimierungen möglich werden und neue Kunden angesprochen werden können, die vorher möglicherweise gar nicht im Fokus standen. Des Weiteren können durch dokumentierte Vorgehensweisen bei Projekten sogenannte Lessons Learned durchgeführt werden, um zukünftige Prozesse zu optimieren und aus Fehlern zu lernen.

20.1 Kurze Einführung

Dieser Beitrag soll Erfahrungen hinsichtlich eines Projektes teilen, das sich mit dem Thema Car-Sharing befasst. Dabei soll verdeutlicht werden, wie wichtig Daten sind und wofür man diese beispielsweise nutzen kann und sollte.

Das Praxisbeispiel, auf das an jedem Ende eines Kapitels eingegangen wird, soll darlegen, wie es nicht funktioniert hat. Daraus können Sie, liebe Leserinnen und Leser, gegebenenfalls Schlüsse für Ihr Unternehmen und Ihre Projekte ziehen.

E. Schoetzau (✉)
Grevenbroich, Deutschland
E-Mail: eva-schoetzau@gmx.de

20.2 Durch Daten zu neuen Ideen und Optimierungen

Durch eine sachgerechte Auswertung und Analyse von vornehmlich unstrukturierten Daten besteht die Potenzialität, dass der Bedarf von (End-)Kunden an neuen Produkten eruiert werden kann. Denkt man an strukturierte Daten, die bei Unternehmen entweder in Excel-Tabellen oder in Datenbanken hinterlegt sind, wird der Gedanke an Namen, Adressen, Geburtsdaten oder Ähnliches geweckt. Daten, die Fakten bezüglich persönlicher Angaben enthalten – quasi ein ausführliches Telefonbuch. Doch diese strukturierten Daten helfen Unternehmen nicht dabei, neue Geschäftsmodelle oder -ideen zu entwickeln. Strukturierte Daten enthalten keine Informationen zu persönlichen Interessen, Vorlieben, Bedürfnissen und so weiter.

Unstrukturierte Daten hingegen können ein gutes Indiz für Bedarfe sein, aus denen entweder eine Produktneuentwicklung oder eine Optimierung eines bestehenden Produktes oder einer Dienstleistung resultieren kann. „Doch was ist nun neu an den Datenmengen, die in jüngster Zeit unter der Bezeichnung ‚Big Data‘ so viel Aufmerksamkeit erregen? Natürlich ist die Menge an verfügbaren Daten durch das Internet der Dinge [Industrie 4.0], durch Mobile Devices und Social Media immens gestiegen. Entscheidend ist jedoch, dass durch die zunehmende Ausrichtung von Unternehmens-IT auf den Endkunden und die Digitalisierung der Geschäftsprozesse die Zahl der kundennahen Kontaktpunkte, die sowohl zur Generierung von Daten als auch zum systematischen Aussteuern der Kommunikation genutzt werden können, gestiegen ist.“ (vgl. Gentsch 2018, S. 7 f.).

Stellen Sie sich Folgendes vor: Sie wohnen in einer Großstadt, besitzen kein eigenes Auto und möchten Car-Sharing nutzen. „Beim CarSharing – zu Deutsch ‚Autoteilen‘ – besitzt man das Auto nicht selbst, sondern teilt es sich mit Anderen. Halter des Autos ist in der Regel der CarSharing-Anbieter. Kunden schließen mit dem Anbieter bei der Anmeldung einen Rahmenvertrag.“ (Bundesverband CarSharing, o. J.).

Car-Sharing-Varianten (vgl. Bundesverband CarSharing, o. J.):

- a) Stationsbasiertes Car-Sharing: Das Fahrzeug hat einen festen Stellplatz und wird wieder auf diesem Platz abgestellt wo es auch abgeholt worden ist. Reservierungen können auch Wochen im Voraus erfolgen.
- b) Free-Floating: Das Fahrzeug hat keinen festen Stellplatz, sondern steht irgendwo in der Stadt geparkt und kann nach der Fahrt in auch irgendwo in dem Nutzungsgebiet wieder abgestellt werden. Reservierungen sind nicht oder nur für kurze Zeit möglich.
- c) Peer-to-Peer Car-Sharing: Eine Privatperson bietet ihr eigenes Fahrzeug über eine Internetplattform zum Teilen an.
- d) RideSharing: Nutzer eines privaten Fahrzeuges bieten einzelne Fahrten als Mitfahrglegenheit an.

Hierzu sind Sie bei einer App angemeldet, dessen Herausgeber Free-Floating Car-Sharing anbietet. Die App zeigt an, dass das nächstgelegene freie Fahrzeug 800 m ent-

fernt steht. Ihre U-Bahn-Station liegt lediglich 200 m entfernt und Sie wissen, dass die nächste U-Bahn bald eintreffen wird. Außerdem haben Sie die aktuelle Verkehrslage noch nicht nach möglichen Staus geprüft. Deshalb entscheiden Sie sich für die U-Bahn, da Sie die Wahrscheinlichkeit, pünktlich an Ihrem Zielort anzukommen als höher einschätzen.

Unabhängig vom Umweltaspekt oder weiteren Vor- und Nachteilen der Nutzung des öffentlichen Personennahverkehrs, soll hier lediglich darauf aufmerksam gemacht werden, dass angebotene Produkte und Dienstleistungen für potenzielle Kunden auch dort zur Verfügung stehen müssen, wo sie gebraucht und genutzt werden. Natürlich kann nicht in jeder Straße ein Car-Sharing-Fahrzeug stehen. Nichtsdestotrotz gäbe es die Möglichkeit, dass Sie als potenzieller Kunde das entsprechende Car-Sharing Unternehmen beispielsweise über soziale Medien darauf aufmerksam machen, dass Sie gegebenenfalls schon vermehrt kein entsprechendes Fahrzeug in Ihrer Nähe zur Verfügung stehen hatten. Würde dies häufiger der Fall sein und die Information dem Car-Sharing-Anbieter (als unstrukturierte Daten durch beispielsweise Social-Media-Posts) vorliegen, hätte dieser verschiedene Möglichkeiten, sein Angebot zu erweitern beziehungsweise das Kundenerlebnis zu optimieren:

- a) Seinen Fahrzeugpool in den entsprechenden Bereichen erhöhen.
- b) Die Fahrzeuge umplatzieren, falls vermehrt ungenutzte Fahrzeuge in anderen Bereichen stehen sollten.
- c) Seine App dahingehend erweitern, dass weitere Mobilitätsangebote angezeigt werden, um zu dem nahegelegenen freien Fahrzeug zu gelangen.

Unternehmen können natürlich nicht auf jeden Vorfall eingehen oder eine Lösung dafür anbieten. Dennoch ist es möglich zum Beispiel mit dem oben genannten Punkt c. eine Lösung darzubieten, die gegebenenfalls nicht für alle Kunden optimal ist, die jedoch eine Möglichkeit bietet, das eventuelle Problem zu lösen und dem Kunden zu zeigen, dass sich das Unternehmen für die Bedürfnisse seiner Kunden interessiert.

Im Praxisbeispiel: Bei einem internen Ideen- und Innovationswettbewerb einer Autoherstellerbank, der dem betrieblichen Vorschlagswesen nahekommt, wurde eine Idee eingereicht, die den Ansatz verfolgte, dass es einer Gruppe von Personen möglich sein sollte, zusammen einen Leasingvertrag für ein Fahrzeug abzuschließen. Das Fahrzeug sollte über eine App geöffnet und geschlossen werden können, um lästige Schlüsselübergaben zu vermeiden. Außerdem konnte über eine Kalenderfunktion in der App das Fahrzeug gebucht beziehungsweise reserviert werden und es wurde die Möglichkeit geboten mit den weiteren Fahrern der Gruppe in Kontakt zu treten, damit beispielsweise Fahrgemeinschaften organisiert werden konnten. Neu war nicht der Gedanke dieser App, sondern die Möglichkeit, ein Fahrzeug und die dafür anfallenden Kosten zu teilen und das alles in einem vertraglich festgelegten Rahmen. Eine nahezu einfache Idee, die der Markt bis dato noch nicht hergegeben hatte.

In der Leasingrate sollte neben dem Fahrzeug noch ein Wartungspaket, die Key-Box zum Öffnen des Fahrzeugs via App und die Nutzung der App enthalten sein. Optional konnte eine günstige Versicherung hinzugebucht werden – ohne Einschränkung hinsichtlich des Fahrerkreises, Alter der Fahrer und der KM-Laufleistung. Durch das Leasingangebot mit seiner optionalen Offerte einer Versicherung sollten vor allem junge Fahrer angesprochen und angezogen werden, da in dieser Altersgruppe gerade die Kfz-Versicherung enorme Kosten verursacht.

Zugegebenermaßen entstand diese Idee nicht durch die Analyse und Auswertung von Daten, sondern durch die Anwendung von Kreativitätstechniken. Jedoch ist „(d)ie Ideengewinnung mit ihren Arten Ideensammlung und Ideengenerierung (...) ein wichtiger Bestandteil des Innovationsprozesses. (...) Wesentlich höhere Erfolgsaussichten bietet eine gezielte Ideengenerierung, die Ideen entsprechend den Zielen, der Strategien und den Möglichkeiten des Unternehmens liefert. Dabei sind Kreativitätstechniken ein unverzichtbares Werkzeug zum Erfolg.“ (vgl. Gaubinger et al. 2009, S. 77).

- ▶ „Die meisten Menschen sind es gewohnt, rational und linear zu denken. Probleme werden gelöst wie eine mathematische Gleichung. Kreativitätstechniken unterstützen dagegen chaotisches Denken. Ziel der Kreativitätstechniken ist es, unsere Gedanken auf neue Art zu verknüpfen und dabei aus von gewohnten Denkmustern auszubrechen. (...) Kreativitätstechniken schaffen folglich keine Kreativität, sondern sie unterstützen Kreativität.“ (vgl. Becker et al. 2018, S. 89).

Nach Anwendung der Kreativitätstechnik wurden im Zuge einer Machbarkeitsstudie – auch bekannt als Proof of Concept – Daten erhoben, die durch Umfragen, Recherchen, Statistiken etc. eingeholt werden konnten:

Umfragen: In Kollaboration mit einem externen Dienstleister, der auf Online-Panels zur Durchführung von Online-Umfragen spezialisiert ist, wurde ein adäquater Fragebogen für die Umfrage entwickelt. Durch die gestellten Fragen konnten von den Umfrage-Teilnehmern Informationen bezüglich der Verfügbarkeit, Nutzungshäufigkeit und des Vorhandenseins eines Fahrzeuges im Haushalt eingeholt werden – neben sozio-demographischen Daten wie beispielsweise Alter, Wohnort etc. Außerdem war es essenziell herauszufinden, ob eine Bereitschaft gegeben wäre, ein Fahrzeug mit bekannten Personen zu teilen und aus welchen Gründen diese Bereitschaft einhergeht. Auf Basis dieser Informationen hätten entsprechende Marketingmaßnahmen geplant und durchgeführt werden können, da die ‚Message‘ durch eine gewissenhafte Auswertung der beantworteten Fragen adäquat hätte gewählt werden können – beispielsweise durch einen gesetzten Schwerpunkt auf den günstigen Preis oder die gegebene Flexibilität und Verfügbarkeit. Doch wie Sie im weiteren Verlauf dieses Beitrages noch erfahren werden, wurden keine adäquaten Marketingmaßnahmen ergriffen.

Ohne weiter ins Detail gehen zu wollen, soll die Information genügen, dass die Hälfte der Umfrage-Teilnehmer der Vorstellung, ein Fahrzeug mit einem festgelegten Personen-

kreis zu teilen, positiv gegenüberstand. Somit war die Aussicht eines Absatzes des Produktes hypothetisch gegeben.

Recherchen: Internetrecherchen bezüglich eines ähnlichen Produktes haben kein Ergebnis geliefert. Lediglich die in dem Abschn. 20.2 genannten Car-Sharing-Varianten konnten identifiziert werden. Allerdings konnte eine Vorlage eines privaten Car-Sharing-Vertrages auf der Website des ADAC (siehe <https://www.adac.de/-/media/adac/pdf/jze/carsharing-vertrag.pdf>) sondiert werden, wodurch ein Eindruck gewonnen wurde, auf welche Punkte in einem vertraglichen Rahmen eingegangen werden sollte.

Statistiken: Bereits existierende Statistiken über die Nutzungshäufigkeit von klassischen Car-Sharing-Angeboten aus dem oben genannten „Stellen-Sie-sich-Folgendes-vor“-Beispiel sowie Meinungen zu Car-Sharing als Alternative zum Fahrzeugkauf oder Gründe gegen die Nutzung von Car-Sharing konnten im Internet abgerufen werden. Diese Informationen konnten für die Argumentation für das neue Car-Sharing-Leasing-Modell verwendet werden, da beispielsweise die Unsicherheit bezüglich der Verfügbarkeit oder dass keine fremden Autos gefahren werden wollen, entkräftet werden konnten.

Letztlich konnte durch die Erhebung und Analyse dieser Daten der Bedarf des vertraglich geregelten Car-Sharings festgestellt werden, weshalb die Machbarkeitsstudie ein positives Ergebnis erzielt hat. Somit wurde die Idee weiterverfolgt und sollte in die Tat umgesetzt werden.

20.3 Umdenken im Unternehmen

Ein neues Produkt, eine neue Dienstleistung oder generell ein noch nicht dagewesenes Angebot bringen Veränderungen mit sich. Vor allem disruptive Geschäftsmodelle gehen mit Veränderungen einher, da diese häufig nicht ohne Weiteres in alltägliche und gewohnte Prozesse eingebunden werden können, die seit Jahren und Jahrzehnten im Unternehmen etabliert sind. Neue Ideen bedeuten auch, dass etwas Neues ausprobiert wird – „neu“ im Sinne von anders. „Der neue Weg mag komplexer sein, holpriger. Bei neuen Produkten und Geschäftsmodellen muss man auch mal bei Null anfangen, das Pferd von hinten aufzäumen, träumen – oder seine Kunden fragen, die Wettbewerber beobachten, Rollen tauschen. Und ein wenig Mut beweisen, wenn es an die Umsetzung geht. Hinzu kommt die Bereitschaft, auch mal danebenzugreifen, eine Idee umzusetzen, die sich vielleicht doch nicht verkaufen lässt.“ (vgl. Spancken 2018, S. 64). Die Erkenntnisse, die bei einer vom Kunden gegebenenfalls nicht angenommenen Idee erlangt werden, können in Form von ‚Lessons Learned‘ an Kollegen oder andere Personen weitergegeben werden, damit diese von solchen Erfahrungen profitieren können. Auf das Thema ‚Lessons Learned‘ wird im Abschn. 20.5 eingegangen.

Für Unternehmen, die seit Beginn ihrer Geschäftstätigkeit ihre Umsätze mit dem immerwährenden Angebot generiert haben, heißt das, dass ein Umdenken stattfinden muss. Ein Veränderungsprozess, der häufig nicht einfach von heute auf morgen vonstattengehen kann, weil nicht nur Prozesse und Abläufe sowie Hard- und Software etc.

angepasst, sondern vor allem auch die Mitarbeiter eingebunden und abgeholt werden müssen. Das ‚Neue‘ muss zunächst akzeptiert werden, sodass kein Unmut in der Belegschaft entsteht. Und dann kann benötigtes Know-how aufgebaut oder gegebenenfalls extern eingeholt werden. „Digitales Denken zu etablieren, neue Kundengruppen zu verstehen, innovative Nutzen zu erkennen – Letztere für die Kunden und für das eigene Unternehmen: All das kann zwei, drei Jahre dauern, und diese Zeit haben wir kaum noch.“ (vgl. Spancken 2018, S. 16).

Junge Unternehmen und Startups haben den vermeintlichen Ruf, dass diese flexibel und agil sind, da sie noch keine eingefahrenen Prozesse aufgebaut haben und jüngere Arbeitnehmer offener sind für Neuerungen. Häufig sieht man daher, dass etablierte Unternehmen beispielsweise Tochtergesellschaften gründen, die dafür zuständig sind, sich mit aktuellen und zukunftsweisenden Entwicklungen zu beschäftigen, damit bestehende Angebote und etablierte Prozesse nicht adaptiert werden müssen und weiterhin bestehen können, ohne, dass das Alltagsgeschäft darunter womöglich leidet.

„Schuster, bleib bei deinen Leisten“ hat vermutlich jeder von Ihnen schon einmal gehört. Dies impliziert, dass Ideen und neue Produkte dem Unternehmen und seinen bisherigen Erfahrungen entsprechen und daher nicht allzu innovativ und neu sein sollten, damit sich nicht auf unbekanntes Terrain eingelassen werden muss. Und so sträuben sich Unternehmen vor Neuerungen, vor vermeintlich unwegsamen Entwicklungen, weil die Angst überwiegt, das Neue nicht managen zu können – die Angst, dass das Neue nicht den Erwartungen der Kunden entspricht, nicht dem was das Unternehmen verkörpert. Doch warum immer in Schwarz und Weiß denken? Es gibt nicht nur ‚entweder oder‘. Hybride Modelle, die das Alte mit dem Neuen verbinden, können doch ein Weg in die richtige Richtung sein. Ein Weg in Richtung Innovation und in die Richtung etwas zu wagen. Die wenigsten Unternehmen können und wollen alles von heute auf morgen umstellen, was kein Grund sein sollte, gar nicht erst anzufangen.

So werden beispielsweise altbekannte Automobilhersteller zu Mobilitätsdienstleistern, indem sie eine Car-Sharing-Flotte zur Verfügung stellen, ohne das Kerngeschäft oder etablierte Prozesse umzustellen. Und es werden bewährte Produkte genutzt, um ein neues Angebot darzubieten: Das eigentliche Kernprodukt, nämlich das Automobil, wird dahingehend adaptiert, dass durch eingebaute und zur Verfügung gestellte Hard- und Software, die Dienstleistung des Car-Sharings angeboten werden kann (siehe beispielsweise <https://www.we-share.io/>).

Auch neue Marketing- und Distributionskanäle müssen erschlossen werden, um die Kanäle zu nutzen, in denen sich die potenzielle Zielkundengruppe aufhält. Digitales Marketing ist längst kein Fremdwort mehr, jedoch entwickelt sich dieser Bereich stetig weiter. Allein dadurch, dass neue Social-Media-Kanäle entwickelt werden und in den Markt eintreten, bleibt auch die Entwicklung des Digitalen Marketings nicht stehen. Diese Kanäle können nicht nur für Marketing und Distribution genutzt werden, sondern auch dafür, (unstrukturierte) Daten zu sammeln. Beispielsweise kann Social-Media-Monitoring wichtig sein, um Meinungen und Äußerungen zu aktuellen Produkten und gegebenenfalls Ideen für Produktoptimierungen einzuholen, indem entweder aktiv nach Meinungen gefragt wird oder (einzelne) Posts ausgewertet werden.

Im Praxisbeispiel: Nachdem durch die erwähnten Umfragen festgestellt werden konnte, dass die Bereitschaft gegeben ist, das vertraglich festgelegte Car-Sharing-Angebot zu nutzen, musste nun ein entsprechender Prozess entwickelt werden, der für zu involvierende Abteilungen sowie für zukünftige Kunden, aber auch für die Autohändler, die das neue Produkt letztlich an den Kunden bringen, einfach und verständlich sein sollte.

Dies stellte sich als große Herausforderung dar: Die Entwicklung des Prozesses für das neue Car-Sharing-Produkt hatte Monate gedauert und den Launch ständig nach hinten verschoben, da immer wieder neue Sorgen geäußert worden sind. Jeglicher Versuch, die Diskussionen beispielsweise durch einen ein- oder mehrtägigen Workshop abzukürzen, schlugen fehl.

Die Produktidee hatte durch den internen Innovationswettbewerb auch international und auf der Führungsebene für Aufsehen gesorgt und sollte daher definitiv weiterverfolgt und umgesetzt werden. Doch dies wurde nicht von allen Kollegen positiv begrüßt, da sie durch den neuen Prozess vermeintlich mehr Arbeit hatten und umdenken mussten. Somit gestaltete sich die Eruierung eines geeigneten Prozesses sehr zäh und unflexibel. Jegliche Möglichkeiten den Prozess des Vertragsabschlusses intern und extern zu digitalisieren, schlugen ebenfalls fehl. Warum? Die Angst war zu groß, dass damit auch bestehende Prozesse und Abläufe digitalisiert werden müssten und damit Arbeitsplätze verloren gehen würden. Ein Vorurteil mit dem die Digitalisierung bereits seit Beginn kämpfen muss. „Die Digitalisierung rationalisiert zwar viele Arbeitsplätze weg, allerdings zumeist im Niedriglohnsektor, sodass sie parallel das Bildungsniveau anhebt – wenn wir politisch, gesellschaftlich und wirtschaftlich die richtigen Weichen stellen!“ (vgl. Spancken 2018, S. 13). Die Weiterentwicklung des Unternehmens und den Mehrwert, den dieses Car-Sharing-Leasingprodukt mit sich bringen würde, wie beispielsweise neue Kundengruppen und eine erhöhte Kundenbindung, wurden nicht gesehen. Da jedoch niemand dafür verantwortlich sein wollte, dass das Produkt nicht auf den Markt kommt, wurde ein Prozess implementiert, der für niemanden attraktiv war - weder für die Mitarbeiter, die für die Bearbeitung der neuen Verträge verantwortlich waren noch für die Kunden oder die Autohändler. So wurden nur marginale Anpassungen an internen Systemen vorgenommen, was bei den ‚Ideen-Gegnern‘ auf entsprechende Zustimmung gestoßen war. Ein flexibles, neues System parallel zu entwickeln und zu betreiben, kam für niemanden infrage, obwohl ein solches System gegebenenfalls auch für andere Prozesse hätte Vorteile bringen können.

Es gab jedoch nicht nur Nachteile bei dieser Vorgehensweise, denn beispielsweise waren die Investitionskosten und die Risiken durch die Verwendung bestehender Prozesse und Systeme entsprechend gering. Vorteile, die nicht unerwähnt bleiben sollen, auch wenn diese für ein innovatives Vorankommen nicht dienlich gewesen sind.

Aufgrund der fehlenden Möglichkeit einen solchen Car-Sharing-Leasingvertrag online abzuschließen, konnten nicht viele Wege für einen Vertragsabschluss angeboten werden; genauer gesagt lediglich einer, nämlich klassisch über das Autohaus. In diesem Falle hatten sich die Geschäftsführer von zwei Autohäusern bereit erklärt, dieses Angebot beziehungsweise dieses Produkt zu verkaufen. Somit war die Einschränkung nicht nur wegen des Distributionskanals ‚Autohaus‘ sehr enorm, sondern auch aufgrund der regionalen Begrenzung auf zwei Städte.

Aus Angst, dass das neue Produkt nicht angenommen und zusätzlich öffentliche Kritik diesbezüglich geäußert werden würde, wurden Marketingmaßnahmen auf das Nötigste beschränkt und so das Gegenteil getan, was man bei neuen Produkten unternehmen sollte, nämlich: Nicht darüber sprechen. Es wurden Informationen auf der Website der entsprechenden Herstellerbank und den beiden Autohäusern zur Verfügung gestellt. Somit wurde niemand auf das Produkt aufmerksam gemacht, der nicht regelmäßig einer dieser Websites aufrief. Die Beschränkung auf Owned Media, wobei lediglich die Website und noch nicht einmal Social-Media-Kanäle genutzt worden sind, wurde dem Produkt zum Verhängnis. Dabei war die Zielgruppe im Laufe des Projektes eindeutig definiert worden: Dieses Produkt sollte vor allem junge Menschen ansprechen. Außerdem hätten durch die Analyse der eingeholten Daten via Umfragen, Recherchen und Statistiken entsprechende Werbemittel und Marketingmaßnahmen konzipiert werden können. „Natürlich ist die Datenmenge durch (...) Social Media immens gestiegen – doch dies ist eher ein gradueller Argument. Entscheidend ist, dass durch die Möglichkeiten der IT und die Digitalisierung der Geschäftsprozesse kundennahe Kontaktpunkte sowohl zur Generierung von Daten als auch zum systematischen Aussteuern der Kommunikation gestiegen sind.“ (vgl. Gentsch 2018, S. 8).

Aufgrund der jungen Zielgruppe wäre Social-Media-Marketing höchstwahrscheinlich ein geeigneter Kanal gewesen, um Aufmerksamkeit zu erregen, aber letztlich werden wir nie erfahren, ob das Produkt durch erweiterte Marketing- und Distributionskanäle tatsächlich erfolgreicher geworden wäre.

Das Umdenken in Unternehmen kann auch durch die Zuhilfenahme von externen Unternehmen oder Agenturen erfolgen, die Erfahrung in den Punkten haben, die für das entsprechende Unternehmen gegebenenfalls noch ‚Neuland‘ sind. Auf das Praxisbeispiel bezogen hätte dies eine Agentur oder ein Beratungsunternehmen sein können, das sich beispielsweise auf Social-Media-Marketing und -Monitoring spezialisiert hat, welches hier sicherlich einen Mehrwert bringen oder zumindest entsprechend hätte beraten können.

20.4 Durch ständige Überwachung zur stetigen Anpassung

Es ist wichtig die Rückmeldung der Kunden auf das Angebot stetig im Auge zu behalten, um gegebenenfalls Optimierungen vornehmen zu können, die letztlich auf den Kunden ausgerichtet sind. Rückmeldungen sind jedoch nicht immer so einfach zu erhalten. Es gibt einige ‚Anlaufstellen‘, die gewählt werden können, um Feedback und Daten einzuholen.

a. Kundenservice

Sollte das Unternehmen einen Kundenservice anbieten beziehungsweise haben, sollte dies die erste Anlaufstelle sein, um etwaig benötigtes Feedback zu einem (neuen) Produkt oder einer (neuen) Dienstleistung einzuholen. Im Kundenservice werden

bestehenden Kunden Hilfe und Unterstützung angeboten, weshalb dort Fragen eingehen und Punkte angesprochen werden, die beispielsweise nicht verständlich genug sind oder schlichtweg nicht funktionieren. Daher ist es sinnvoll, ein regelmäßiges Reporting zu führen, welches die wichtigsten und häufigsten Fragen (FAQ) aufzeigt. Hieraus kann schlussendlich Optimierungsbedarf sondiert werden, der direkt von den Kunden kommt.

Des Weiteren kann auch überlegt werden solche Service-Dialoge zu automatisieren. Einfache Anliegen können so durch einen Chatbot erledigt werden, da in den meisten Branchen über 80 % der Anfragen repetitiv sind (vgl. Gentsch 2018, S. 132). Die FAQ dienen als Wissensdatenbank für den Chatbot, der mit dieser entsprechend verbunden werden würde, um ihn mit den vorliegenden Daten „zu füttern“. „Die Kern-Idee dahinter ist, dass die Teilnehmer des Dialoges automatisiert durch den Bot zu Produkten und Services geleitet werden, die in den Dialogen eine Rolle spielen.“ (vgl. Gentsch 2018, S. 133).

b. Social Media

Die Möglichkeiten mit einem Unternehmen in direkten Kontakt zu treten sind heutzutage recht einfach und vielfältig: E-Mail, Kontaktformulare, Präsenz in sozialen Medien, Hotline usw. Besonders auf Social-Media-Kanälen können von Kunden erstellte Inhalte auch von anderen Nutzern eingesehen werden, sodass Unternehmen in diesem Bereich eine schnelle Reaktionsfähigkeit unter Beweis stellen sollten, damit zumindest bei negativer Meldung, kein sogenannter ‚Shitstorm‘ ausgelöst wird, der ein Unternehmen quasi dazu zwingt, zu antworten, um den negativen Feedbacks Einhalt zu gebieten. Der soziale Druck, der so auf ein Unternehmen ausgelöst wird, wenn es beispielsweise auf eine Kundenbeschwerde nicht (zeitnah) antwortet, kann enorm sein. Außerdem kann dadurch schnell der Ruf eines Unternehmens leiden. Kommentare in Social-Media-Kanälen sollten somit immer ernst genommen und schnell beantwortet werden, auch wenn ein Post möglicherweise banal erscheint.

Durch Social-Media-Monitoring kann herausgefunden werden, wo über das Unternehmen und das jeweilige Produkt gesprochen wird. Hier können entsprechende Kampagnen geschaltet und im Laufe solcher auch entsprechend eingegriffen werden. „Digitale Ideen und Tools tun hauptsächlich (...) dies: den Fokus auf den Kunden richten. Angepasst an neue Kommunikations- und Marketingstrategien mit effektiven Kontrollwerkzeugen lässt sich wesentlich gezielter feststellen, was der Kunde will und ob, wie gut und mit wie viel Erfolg er erreicht wurde.“ (vgl. Spancken 2018, S. 18).

c. Umfragen

Durch Umfragen, die an bereits existierende Kunden gerichtet sind, können Informationen eingeholt werden, die wichtig für entsprechende Produktverbesserungen oder -erweiterungen sind. Doch auch potenzielle Kunden können befragt werden, damit beispielsweise die Frage geklärt werden kann, warum das entsprechende Produkt nicht gekauft oder die entsprechende Dienstleistung nicht genutzt wird. Diese Informationen können ebenfalls erheblichen Aufschluss darüber gegeben,

was optimiert werden kann oder eventuell nicht bedacht worden ist. Grundsätzlich kann es von Vorteil sein, eine solche Umfrage zu incentivieren, um einen Anreiz für die Teilnahme zu schaffen.

Solche Umfragen können beispielweise telefonisch, im digitalen Kundenportal oder auch im Social-Media-Kanal durchgeführt werden.

Vor allem durch Social Media und den heutigen Auswertungsmöglichkeiten von solchen Kanälen, aber auch durch anderweitige Marketingkampagnen können unstrukturierte Daten gesammelt und analysiert werden. Diese können Aufschluss darüber geben, was den (potenziellen) Kunden an einem Produkt gefällt und was nicht – hier kann stetig optimiert werden.

Im Praxisbeispiel: Die einzige Komponente, die bei dem Car-Sharing-Produkt ausgewertet worden war, waren abgeschlossene Verträge. Werden keine oder nur wenige Verträge abgeschlossen, wird das Produkt kurzum eingestellt, da es augenscheinlich erfolglos war.

Doch warum nicht auf Daten hören, sondern nur darauf was scheinbar ‚möglich‘ ist? Durch Umfragen konnte festgestellt werden, wo die Nachfrage für ein solches Produkt am höchsten war, da auch demografische Daten eingeholt worden waren. Doch das neue Produkt wurde in anderen Gebieten gelauncht – nämlich dort, wo Autohäuser bereit waren, dieses anzubieten. Auch die Marketingmaßnahmen waren unzureichend. Woher oder worüber sollten potenzielle Kunden von dem Angebot erfahren?

Auf Grund dessen, dass das Produkt keine hohe Akzeptanz in der Belegschaft erfahren hatte, hatte auch kein Interesse darin bestanden, beispielsweise weitere Umfragen durchzuführen oder für das Unternehmen bis dato unbekannte Marketingkanäle wie Social Media zu nutzen, um den Bekanntheitsgrad zu erhöhen. Dies hätte nicht nur die Werbetrommel für das neue, innovative Car-Sharing-Produkt gerührt, sondern auch für das Unternehmen selbst. Somit hätte das angebotene Produkt das Unternehmen beispielsweise dazu befähigen können, neue Kundengruppen anzusprechen. Mit gezielten Marketingmaßnahmen hätten nicht nur junge Fahrer, sondern auch Wohn- oder Hausgemeinschaften, Institutionen, anderweitige Gemeinschaften oder auch Familien, die sich allein kein (Zweit-)Auto hätten leisten können, gezielt angesprochen werden können.

Des Weiteren hätten durch die Analyse und Auswertung von Daten die Funktionen der App sowie zusätzliche Services mit der Zeit entwickelt und schlussendlich angeboten werden können, sodass das Angebot stetig erweitert werden können. Doch die Angst vor zu großer Veränderung und zu viel ‚Mehrarbeit‘ haben quasi zum Stillstand der Innovation geführt. Auch scheinbar hybride Modelle, die das Innovative mit dem bereits Existierenden verbinden, führen nicht automatisch zu Zustimmung und Erfolg und lassen sich auch nicht zwangsläufig einfacher umsetzen. Nichtsdestotrotz kann auch

bei einem Misserfolg optimiert und angepasst werden, sodass die Neuerungen unter Umständen dann den erwarteten Absatz erreichen. Das Produkt oder das Angebot einzustellen, ist nicht zwangsläufig die beste Lösung, gerade dann nicht, wenn viel Arbeit und Geld in das Projekt geflossen sind.

20.5 Mit ‚Lessons Learned‘ zur Optimierung von Geschäftsmodellen und -prozessen

In vielen deutschen Unternehmen herrscht keine ausgeprägte Fehlerkultur, sprich über Niederlagen, Missverständnisse und Fehler wird wenig oder überhaupt nicht gesprochen. Es wird vielmehr danach gestrebt, Probleme möglichst schnell zu beheben – das ist pragmatisch, jedoch nicht hilfreich, um es in Zukunft von vornherein besser zu machen. Doch wie soll es beim nächsten Mal oder bei zukünftigen Projekten besser laufen, wenn nicht gesagt werden kann, warum etwas falsch gelaufen oder nicht angenommen worden ist? „Im Silicon Valley etwa werden Unternehmer, die bereits gescheitert sind, sehr ernst genommen – sie haben sich die blutige Nase schon geholt, aus Fehlern gelernt und den Mumm bewiesen, es noch einmal zu versuchen. (...) Fallen lernen und dann wieder aufstehen.“ (vgl. Spancken 2018, S. 64 f.).

Mit ‚Lessons Learned‘ können gesammelte Erfahrungen und Erkenntnisse dokumentiert und bewertet werden, um diese für kommende Projekte nutzbar zu machen. Dies setzt voraus, dass die Planung, das Projektmanagement, die Zusammenarbeit, ausgewählte beziehungsweise genutzte Tools, die Gründe für Verzögerungen und so weiter dokumentiert werden und sich dessen bewusst gemacht wird, was letztlich zu einem verzögerten Launch, zu einem Misserfolg oder weiteren Punkten geführt hat. Letztlich kommt es darauf an, wo angesetzt werden soll: Die Entstehung der Idee, die Durchführung des Projektes, das Produkt oder die Dienstleistung an sich und so weiter. Bei einem Projekt können insgesamt verschiedene Punkte ‚schieflaufen‘.

Somit kann in Bezug auf ‚Lessons Learned‘ auch wieder von der Nutzung von Daten gesprochen werden, was essenziell ist, um ein Projekt, ein Produkt oder eine Dienstleistung zu optimieren.

a. Entstehungsprozess von Ideen

Welche Kriterien führten zu der Idee des Produktes oder der Dienstleistung: Das Angebot des Wettbewerbs, Kundenanfragen, der Innovationsgedanke oder -druck oder lediglich aus der Laune mancher Führungskräfte heraus? Es gibt verschiedene Gründe, die zu der Entstehung oder Initialzündung eines neuen Produktes führen können. Wichtig ist, dass hier von Beginn an hinterfragt wird, ob ein Mehrwert für das eigene Unternehmen und vor allem für die Kunden generiert werden kann und aus welchem Antrieb heraus das Angebot entsprechend erweitert werden soll. Dies ist

essenziell, da hierdurch schon analysiert werden kann, ob sich das neue Angebot tatsächlich an den Bedürfnissen des Marktes und des Kunden orientiert. Dies ist kein Garant dafür, dass das Angebot einen Absatz findet, jedoch ist die Wahrscheinlichkeit gegeben, da Wettbewerbs- und Innovationsdruck sowie Kundenbedürfnisse fundierte Gründe für die Entwicklung und den Launch eines neuen Produktes oder einer neuen Dienstleistung sind.

Sollte eine Idee aus der Feder eines Mitarbeiters (unabhängig davon welcher Hierarchieebene) entsprungen sein, sollte zunächst eine Machbarkeitsstudie, ein Proof of Concept, durchgeführt werden. Dies ist sinnvoll, um eine Einschätzung zu erhalten, ob zum einen die Idee überhaupt durchführbar wäre und zum anderen kann in dieser Phase auch eine Umfrage und/oder Marktanalyse durchgeführt werden, damit das mögliche Kundeninteresse oder sogar Absatzpotenzial ermittelt werden kann. Ein solches Vorgehen kann natürlich zu jeder Zeit durchgeführt werden. Es soll jedoch verdeutlicht werden, dass Ideen sorgfältig geprüft werden sollten, vor allem, wenn diese ohne die Daten des Marktes also des Wettbewerbs oder ohne fundierte Kundenmeinungen entstehen.

Lessons Learned können hier eine Unterstützung sein: Ist dokumentiert wie die entsprechende Idee auf den Weg gekommen ist und welche Maßnahmen (nicht) getroffen worden sind, um das mögliche Kundeninteresse oder Absatzpotential zu ermitteln, kann zukünftig entsprechend ebenso verfahren oder ein anderer Weg eingeschlagen werden. Somit können ‚Lessons Learned‘ selbst Ideengeber und Anhaltspunkte sein, um das Vorgehen zu gestalten und mögliche Maßnahmen zu konzipieren. Da einige Unternehmen ihre eigenen Regeln für Prozesse und Vorgehen haben, ist es sinnvoll, auf unternehmensinterne Erfahrungen und Informationen zurückzugreifen, damit eventuell existierende Formalitäten bewusst eingehalten werden.

b. Durchführung des Projektes

In den letzten Jahren haben vor allem agile Projektmanagementmethoden den Einzug in deutschen Unternehmen gehalten. Der Begriff ‚Scrum‘ schien und scheint ein Synonym für „das muss ja dann funktionieren“ zu sein. Ein Allerheilmittel, das Probleme löst und Produkte flexibel und agil erscheinen lässt, so lässt zumindest der Hype um diese Projektmanagementmethode vermuten. Unabhängig davon welche Projektmanagementmethode letztlich angewandt wird, ist es nützlich wenigstens im Nachhinein zu hinterfragen, ob dies die richtige Wahl gewesen ist. Dies sollte entsprechend dokumentiert und begründet werden, damit Auswahlkriterien bestimmt werden können, die indizieren, bei welcher Art von Projekt welche Projektmanagementmethode verwendet werden sollte. Die Erfahrung lässt im Idealfall darauf schließen, welche Methode für welches Projekt, aber auch für das entsprechende Unternehmen sinnvoll ist, denn dies kann auch ein entscheidender Faktor sein, der die Wahl beeinflussen sollte.

„Klassische, traditionelle Vorgehensweisen haben wesentliche Beiträge im Projektmanagement geleistet und leisten diese weiterhin. In der Produkt- und Softwareentwicklung hat sich jedoch gezeigt, dass viele Projekte, welche mit einer Wasserfallmethode gemanagt wurden, nicht die gewünschten Resultate brachten oder sogar scheiterten. (...) Agiles Projektmanagement heißt bewegliches, flinkes, prozessorientiertes, reflexives, lernendes Vorgehen.“ (vgl. Kuster, et al. 2019, S. 18 f.).

Es gibt weitere Punkte, die zu der Durchführung eines Projektes gehören, welche an dieser Stelle jedoch keine Erwähnung finden. Letztlich ist es wichtig, dass am Ende durch die Dokumentation und deren Auswertung analysiert werden, was ‚gut gelaufen‘ ist und was optimiert werden könnte. Solche Erkenntnisse sollten unternehmensweit geteilt werden, damit auch Kollegen von den Erfahrungen anderer profitieren können.

C. Marketing und Launch

Bei der Nutzung der Auswertungen von Marketingmaßnahmen können die Ergebnisse dazu führen, dass zukünftig auf genutzte Maßnahmen verzichtet werden kann. Es ist häufig nicht einfach, Rückschlüsse daraus zu ziehen, welche Marketingmaßnahme zum Erfolg geführt hat beziehungsweise wie die einzelnen Aktivitäten ins Gewicht gefallen sind. Nichtsdestotrotz ist die Auswertung der Daten wichtig, um die Maßnahmen entsprechend umzugestalten, zu optimieren oder einzustellen. Für zukünftige Produkte oder Dienstleistungen kann dies ein Indiz dafür sein, welche Kundengruppe auf welche Maßnahme reagiert – in welcher Form auch immer, also ob positiv oder negativ. Hier können ‚Lessons Learned‘ eine Hilfestellung sein, damit die Kosten entsprechend besser auf Marketingkanäle verteilt werden können. Nicht auszuschließen ist, dass man zu der Erkenntnis gelangt, dass externe Unterstützung bei der Auswertung und/oder Durchführung von einschlägigen Marketingkampagnen benötigt wird. Dokumentierte Erkenntnisse sind ein Mittel, um nicht nur Kosten, sondern auch Zeit zu sparen, wenn in diesem Falle direkt auf eine entsprechende externe Unterstützung zurückgegriffen werden würde. Doch auch grundsätzlich können ‚Lessons Learned‘ eine Kosten- und Zeitsparnis darstellen wie in diesem Kapitel dargestellt werden soll.

Auch beim Launch gibt es Punkte, die beachtet werden sollten. Der richtige Zeitpunkt, der richtige Ort etc. können ausschlaggebend dafür sein, ob das neue Produkt angenommen wird. Durch die Auswertung von Daten kann somit ein mögliches Fiasko verhindert werden, indem möglicherweise einfach nur ein oder zwei Wochen später oder früher gelauncht wird. Mögliche Parallelveranstaltungen, Launches von Wettbewerbern, Urlaubszeit etc. können gegebenenfalls die Aufmerksamkeit der Kunden auf andere Dinge lenken, sodass das eigene Produkt womöglich in den Hintergrund gerückt wird. Hier können Erfahrungswerte „Gold wert sein“.

d. Kundenmeinungen einholen

Dass Kundendaten und -meinungen hilfreich sind, steht außer Frage, denn durch solche Daten ist es möglich auf die Bedürfnisse, Wünsche und Vorstellungen der Kunden einzugehen. Jedoch ist es wichtig, diese korrekt auszuwerten und entsprechend einzusetzen.

Wenn beispielsweise ein Prototyp hergestellt werden kann, ist es sinnvoll, diesen einer Probandengruppe vorzustellen, damit zu diesem Zeitpunkt noch mögliche Anpassungen und Optimierungen vorgenommen werden können. Hier kann es entsprechend sinnvoll sein, die bereits erwähnte agile Projektmanagementmethode ‚Scrum‘ anzuwenden, da in diesem Fall die Erweiterungen oder Optimierungen des Produktes in einem festgelegten Zeitraum vorgenommen werden, um das Produkt stetig weiterzuentwickeln und einen Mehrwert (MVP) zu schaffen. Dieses adaptierte Produkt kann dann entweder einer Probandengruppe, also dem Endkunden oder den sogenannten Stakeholdern vorgestellt werden. Denkbar wäre auch, einen Workshop mit Kunden, also unternehmensexternen Personen, durchzuführen.

Ist es nicht möglich mit einem Prototyp zu arbeiten, weil das neue Produkt beispielsweise eine Dienstleistung ist, gibt es die Möglichkeit, Kundenumfragen durchzuführen. Solche Umfragen müssen optimal und adäquat vorbereitet werden, damit letztendlich auch die Fragen beantworten werden, die notwendig sind, um das neue Produkt an die Bedürfnisse der Kunden anpassen zu können.

Umfragen können auch über Social-Media-Kanäle durchgeführt und gegebenenfalls incentiviert werden. Gewinnspiele können so wie das sogenannte ‚betriebliche Vorschlagswesen‘ genutzt werden, nur, dass hier mit Kundenideen auf Social-Media-Kanälen gearbeitet würde; ganz nach dem Motto: Die beste Idee gewinnt. Abstimmen können die anderen Nutzer beziehungsweise Kunden. Die Umsetzungsmöglichkeit der Gewinneridee(n) wird dann unternehmensintern über eine Machbarkeitsstudie analysiert. Der Fortschritt dieser Studie und später auch des Entstehungsprozesses des Produktes wird über die entsprechenden Social-Media-Kanäle, vor allem auf demjenigen Kanal, auf dem das Gewinnspiel gelaufen ist, gepostet werden.

Wie bereits in dem Abschn. 20.4 erwähnt, ist ein vermeintlicher ‚Klassiker‘ die Auswertung von Kunden(an)fragen, die über den eigenen Kundenservice oder generell eine Servicehotline eingehen. Hier werden Fragen gestellt, die dafür genutzt werden können, das eigene Produkt zu verbessern. Denn die Fragen sind ein Indiz dafür, was gegebenenfalls noch nicht eindeutig, einfach oder selbsterklärend ist. Diese Fragen können dafür genutzt werden, entsprechende Anpassungen vorzunehmen.

„Lessons Learned“ können auch hier helfen, umgesetzte Maßnahmen zu analysieren und die gewonnenen Erkenntnisse für Folgeprojekte zu nutzen.

Im Praxisbeispiel: Anstatt das Produkt aufgrund fehlender Vertragsabschlüsse direkt einzustellen, hätte durch ‚Lessons Learned‘ zum einen ein Optimierungsbedarf analysiert werden können, um entsprechende Maßnahmen zur Verbesserung des Produktes oder zur Optimierung der Marketingmaßnahmen einleiten zu können. Zusätzlich hätte auch der

gesamte Umsetzungsprozess der Idee beziehungsweise die Durchführung des Projektes analysiert werden müssen, damit die Umsetzung zukünftiger Geschäftsideen effektiver und effizienter vonstattengeht.

20.6 Fazit

Dieser Beitrag legt dar, dass Daten dafür genutzt werden können neue Ideen für Produkte, Dienstleistungen oder Angebote jeglicher Art zu entwerfen und zu optimieren beziehungsweise zu adaptieren. Daten sind kein Garant für das Bestehen und die Akzeptanz eines Angebotes, jedoch können sie ein sehr gutes Indiz dafür sein, ob die Bedürfnisse von (potenziellen) Kunden aus den daraus folgenden Angeboten befriedigt werden können. Es hilft in jedem Falle, die offene Kommunikation mit Kunden zu suchen, um Wissen über ihre Bedürfnisse und Wünsche zu erlangen. Diese können für Produktneuerungen als auch für Optimierungen verwendet werden, sofern diese korrekt erhoben, analysiert und ausgewertet werden.

Auf der anderen Seite können neue Produkte beispielsweise neue Kundengruppen anziehen, wodurch wieder neue Daten generiert werden können. Somit ist nicht nur die Möglichkeit gegeben, dass durch Daten neue Geschäftsmodelle eruiert werden können, sondern auch umgekehrt. Des Weiteren soll der Beitrag zeigen, dass existierende Daten entsprechend genutzt und die daraus erlangten Erkenntnisse entsprechend korrekt eingesetzt werden müssen, damit ein Vorteil generiert werden kann. Außerdem will darauf aufmerksam gemacht werden, dass es vorteilhaft sein kann, wenn Erkenntnisse und Erfahrungen während eines Projektes dokumentiert werden, damit diese als ‚Lessons Learned‘ für zukünftige oder laufende Projekte eine Hilfestellung sein und vermeintliche Fehler vermieden werden können.

Literatur

- Becker, J.H., Ebert, H., Pastoors, S.: Praxishandbuch berufliche Schlüsselkompetenzen. (2018).
https://doi.org/10.1007/978-3-662-54925-4_11
- Bundesverband CarSharing: Was ist CarSharing?. (o. J.) <https://carsharing.de/alles-ueber-carsharing/ist-carsharing/ist-carsharing>, Zugegriffen: 31. Juli 2020
- Gaubinger, K., Werani, T., Rabl, M.: Praxisorientiertes Innovations- und Produktmanagement: Grundlagen und Fallstudien aus B-to-B-Märkten. Springer, Wiesbaden (2009)
- Gentsch, P.: Künstliche Intelligenz für Sales, Marketing und Service: Mit AI und Bots zu einem Algorithmic Business – Konzepte, Technologien und Best Practices, Springer, Wiesbaden (2018)
- Kuster, J., Bachmann, C., Huber, E., Hubmann, M., Lippmann, R., Schneider, E., Schneider, P., Witschi, U., Wüst, R.: Handbuch Projektmanagement: Agil – Klassisch – Hybrid. Springer, Berlin (2019)
- Spancken, C.: Digital denken statt Umsatz verschenken: Online-Strategien für den Mittelstand. Econ, Berlin (2018)



Einsatz von Logit- und Probit-Modellen in der Finanzindustrie

21

Uwe Rudolf Fingerlos und Alexander Pastwa

Zusammenfassung

Ein häufiges Anwendungsfeld von statistischen Verfahren in der Finanzindustrie ist die Ermittlung von Krediten, die von einem Ausfall betroffen sind. Der vorliegende Beitrag behandelt mit den Logit- und Probit-Modellen zwei Einsatzmöglichkeiten von Verfahren des überwachten statistischen Lernens in der Finanzindustrie und überprüft ihre Eignung zur Modellierung von Kreditausfallwahrscheinlichkeiten. Als Analysewerkzeug wird RStudio verwendet. Die Datengrundlage ist ein fiktiver Datenbestand, dessen Zusammensetzung sich an den Vorarbeiten von Altmann et al. orientiert. Im Vergleich wird deutlich, dass alle vorgestellten Probit- und Logit-Modellvarianten eine ansprechende Trennschärfe und Prognosefähigkeit für den betrachteten Anwendungsfall zeigen und sie darüber hinaus praktisch identische Ergebnisse liefern.

21.1 Einleitung

Der vorliegende Artikel gibt einen stilisierten und vereinfachten Überblick über die Anwendbarkeit von Logit- und Probit-Modellen im Bankenumfeld, wie sie bei der Modellierung der Kreditausfallwahrscheinlichkeit („Probability of Default“, PD) in internen Ratingsystemen im Bereich der Eigenkapitalrichtlinien des Baseler-Regelwerkes zum Einsatz kommen.

U. R. Fingerlos
Sant Quirze del Valles, Spanien

A. Pastwa (✉)
Waltrop, Deutschland

In Abschn. 21.2 erfolgt zunächst eine kurze Darstellung der theoretischen Grundlagen von Logit- und Probit-Modellen. In Abschn. 21.3 wird ein fiktiver Untersuchungsdatenbestand erzeugt, auf dem die Modellierungsschritte in Abschn. 21.4 aufbauen. Nach einer Überprüfung der Modellannahmen in Abschn. 21.5 werden in Abschn. 21.6 die Ergebnisse vorgestellt. Abschn. 21.7 liefert eine Zusammenfassung der Prognosegüte der vorgestellten Modelle für den betrachteten Anwendungsfall.

21.2 Logit- und Probit-Modelle

Der vorliegende Abschnitt widmet sich den statistischen Eigenschaften von Logit- und Probit-Modellen. Bei diesen beiden Verfahren des überwachten statistischen Lernens handelt sich um einfache Klassifikatoren, die im Gegensatz zu fortgeschrittenen Verfahren wie Ensemble-Methoden Vorteile bei der Interpretierbarkeit der Ergebnisse bieten (für einen Überblick vgl. Lessmann et al. (2015). Fingerlos et al. (2020) zeigen vergleichbare Anwendungsmöglichkeiten von Entscheidungsbäumen und Künstlichen Neuronalen Netzen im Bankenumfeld).

Ist ein Kredit i von einem Zahlungsausfall betroffen, dann nimmt der binäre Ausfallindikator D_i den Wert $D_i = 1$ an, andernfalls gilt $D_i = 0$. Dementsprechend lässt sich die individuelle Ausfallwahrscheinlichkeit als $PD_i = Pr(D_i = 1)$ schreiben. Wird mithilfe der Wahrscheinlichkeitsfunktion $Pr(\cdot)$ der Erwartungswert eines Ausfalles in der Form $E[D_i] = 1 \cdot Pr(D_i = 1) + 0 \cdot Pr(D_i = 0)$ definiert, dann führt dies zu einem Erwartungswert von $E[D_i] = PD_i$ – was der Ausfallwahrscheinlichkeit entspricht. Abstrahierend von dieser Einzelbetrachtung bedeutet dies, dass $E[D] = PD$ im Gesamtdatenbestand gilt.

Ein Scoring- oder Ratingmodell zur Beurteilung von Ausfallwahrscheinlichkeiten besteht aus einer Funktion, welche die in den erklärenden Variablen enthaltene Information mit der Ausfallwahrscheinlichkeit PD in Verbindung setzt. Der Begriff „Scoring“ ist im Zusammenhang mit Krediten an natürliche Personen, der Begriff „Rating“ im Zusammenhang mit Krediten an Unternehmen gebräuchlich. Ein gutes Modell soll hohe Ausfallwahrscheinlichkeiten für ausfallende und niedrige Ausfallwahrscheinlichkeiten für nicht ausfallende Kredite prognostizieren. Die binäre erklärte Variable D_i lässt sich wie folgt als lineare Funktion der erklärenden Variablen schreiben, wobei β ein Zeilenvektor mit den zu bestimmenden Gewichten bzw. Koeffizienten (inklusive Interzept β_0) und X_i ein Spaltenvektor (und X_i ein Zeilenvektor) mit den k erklärenden Variablen für Kredit i ist, in dem die 1 wird für die Multiplikation mit dem Interzept benötigt wird (vgl. Löffler und Posch 2011, S. 1 ff.; Baesens et al. 2016, S. 95 ff.; Bellini 2019, S. 38; Rösch und Scheule 2020, S. 204 ff.; Fingerlos et al. 2020, S. 72 ff.):

$$D_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \boldsymbol{\beta} \cdot \mathbf{X}_i$$

$$\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_k]$$

$$X_i = [1 \ X_{i1} \ X_{i2} \ \dots \ X_{ik}]$$

Die Ausfallwahrscheinlichkeiten PD_i über $PD_i = Pr(D_i = 1|X_i) = F(D_i)$ werden mit der in den erklärenden Variablen linearen und additiven Gleichung $D_i = \beta X_i$ in Verbindung gesetzt, indem eine geeignete Transformationsfunktion $F(D_i)$ zur Anwendung kommt, welche die Ausprägungen von D_i in die korrespondierenden Wahrscheinlichkeiten PD_i überführt. Da sich die Wahrscheinlichkeiten definitionsgemäß im Intervall (0; 1) bewegen müssen, kommen für diese Transformation zwei Arten von Verteilungsfunktionen („cumulative distribution functions“, CDF) in Betracht, die entsprechende Eigenschaften aufweisen (vgl. Hill et al. 2018, S. 685 ff., S. 837 ff.; Greene 2018, S. 731 ff.; Wooldridge 2016, S. 525 ff., S. 525 f., S. 685 f.).

Bei einer logistischen Regression kommt die Verteilungsfunktion der logistischen Verteilung $\Lambda(l) = Pr(L \leq l) = (1/(1 + \exp(-l)))$ für eine logistisch verteilte Zufallsvariable L zum Einsatz. Sie gibt die Wahrscheinlichkeit an, dass L eine Ausprägung nicht größer als l annimmt, und führt durch Einsetzen zu $PD_i = Pr(D_i = 1|X_i) = \Lambda(D_i) = (1/(1 + \exp(-\beta X_i)))$. Nach Umformung lässt sich die Logit-Regression als $\ln(PD_i/(1 - PD_i)) = \beta X_i$ schreiben, wobei $\ln(\cdot)$ den natürlichen Logarithmus und der Ausdruck $\ln(PD_i/(1 - PD_i))$ die logarithmierten Odds (Log-Odds) bzw. das sogenannte Logit bezeichnet. Das Logit dient in der Praxis zur Scorecard-Erstellung, indem zum Beispiel mithilfe einer „Weight-of-Evidence“ (WOE)-Kodierung (siehe sogleich) kategorisierte erklärende Variablen zur Modellierung benutzt und anschließend das Logit linear in die Form $S_i = a + b \cdot \ln(PD_i/(1 - PD_i))$ transformiert werden, um derart für jedes PD-Intervall einfach zu interpretierende Scoringpunkte zu erhalten, die die Kreditqualität widerspiegeln (vgl. Siddiqi 2017, S. 240 ff.; Baesens et al. 2016, S. 101 ff.; Bellini 2019, S. 38 ff.; Jopia 2019, S. 14).

Bei einer Probit-Regression wird die Verteilungsfunktion der Standardnormalverteilung $\Phi(z) = Pr(Z \leq z) = \int_{-\infty}^z (1/\sqrt{2\pi}) \exp(-0.5u^2) du$, für eine standardnormalverteilte Zufallsvariable Z verwendet. Sie gibt die Wahrscheinlichkeit an, dass Z einen Wert nicht größer als z annimmt und führt durch Einsetzen zu $PD_i = Pr(D_i = 1|X_i) = \Phi(D_i) = \Phi(\beta' X_i)$. Der Unterschied zur Logit-Regression liegt in der mitunter einfacheren Interpretierbarkeit der Verteilung der z-Werte gegenüber jener der Log-Odds sowie in dem Umstand, dass die Logarithmierung in der Logit-Regression dazu tendiert, sehr große Effekte zu „untertreiben“, während sie sehr kleine Effekte „übertreibt“ (vgl. Greene 2018, S. 742 ff.; Hill et al. 2018, S. 837 ff.; Wooldridge 2016, S. 528 ff., 565 f. und 685 f.; Osborne 2015, S. 297 ff.).

Wie die Steigung der beiden auf der linken vertikalen Achse der folgenden Abb. 21.1 abgetragenen Funktionsgraphen zeigt, ist die Probit-Funktion (kontinuierliche Linie) etwas steiler und weist bei gleichem Wert der erklärenden Variable x höhere Wahrscheinlichkeitswerte aus als die Logit-Funktion (gepunktete Linie). Die aus der Logit-Funktion resultierenden Odds und ihren Zusammenhang mit der Wahrscheinlichkeit zeigt die auf der rechten vertikalen (logarithmisch skalierten) Achse abgetragene gestrichelte Linie. Da es sich um eine logarithmische Skalierung handelt, erscheinen die Odds als

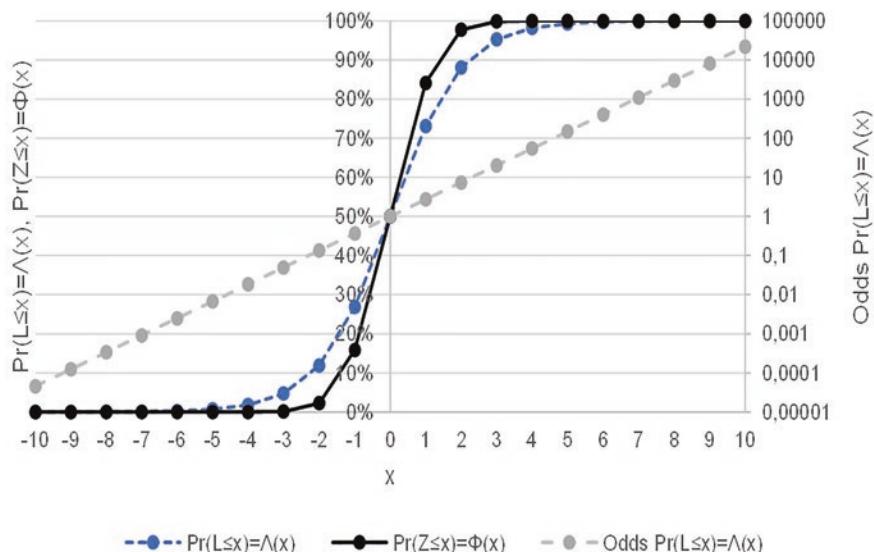


Abb. 21.1 Logit, Probit und Odds

Gerade, die die Logit- und Probit-Funktion auf halber Höhe schneidet. In diesem Punkt entspricht erwartungsgemäß die Ausfallwahrscheinlichkeit (50 %) den Odds.

Der anschließende Abschnitt erläutert die in der Studie verwendete Datengrundlage.

21.3 Datengrundlage

Den folgenden Auswertungen liegt ein fiktiver Datenbestand zugrunde, dessen Charakteristika an eine von Altman et al. (2017) durchgeführte, breit angelegte empirische Überprüfung der Performance des sogenannten Z“-Score-Modells angelehnt sind. Das ursprüngliche Z-Score-Modell wurde von Altman (1968) entwickelt und hat zum Ziel, zukünftige Zahlungsschwierigkeiten von Unternehmen zu prognostizieren, indem grundlegende Bilanzkennzahlen zu einer Diskriminanzfunktion verdichtet werden, deren Funktionswert den Namen Z-Score trägt. Liegt der Wert dieses Z-Scores in einem bestimmten kritischen Bereich, muss mit bevorstehenden Zahlungsschwierigkeiten des betreffenden Unternehmens gerechnet werden (vgl. Altman 1968; Altman et al. 2017). Das Z-Score-Modell wurde weiterentwickelt und hinsichtlich der einfließenden erklärenden Variablen sowie der abgedeckten Industriesektoren erweitert. Das Modell mit dem breitesten Einsatzgebiet ist das Z“-Modell, das sowohl auf börsennotierte und nicht börsennotierte Unternehmen sowie innerhalb und außerhalb des verarbeitenden Gewerbes anwendbar ist (vgl. Altman et al. 2017, S. 136; Altman und Hotchkiss 2006, S. 248).

Die vier im Z"-Modell zur Ausfallprognose verwendeten Bilanzkennzahlen sind *WCTA* (Working Capital / Total Assets), *RETA* (Retained Earnings / Total Assets), *EBITTA* (Earnings Before Interest and Taxes / Total Assets) und *BVETD* (Book Value of Equity / Book Value of Total Liabilities), die zur folgenden Konfiguration der Diskriminanzfunktion führen: $Z'' = 3,25 + 6,56 \cdot WCTA + 3,26 \cdot RETA + 6,72 \cdot EBITTA + 1,05 \cdot BVETD$. Wenn der auf Basis dieser Kennzahlen berechnete Z"-Score eines Unternehmens unter den kritischen Wert von 0 fällt, ist mit einem Ausfall zu rechnen (vgl. Altman und Hotchkiss 2006, S. 248). Dass die hierfür verwendeten erklärenden Variablen zurecht auch in PD-Modelle zur Prognose von Ausfallwahrscheinlichkeiten einfließen, liegt auf der Hand.

Insgesamt besteht der Datenbestand, den Altman et al. (2017) ihrer Validierung des Z"-Score-Modells zugrunde legen, aus 2.602.563 Beobachtungen nicht ausgefallener und 38.215 Beobachtungen ausgefallener Firmen, was einer beobachteten Ausfallquote („Observed Default Date“, *DR*) von $DR = 38.215/(38.215 + 2.602.563) = 0,0145 \approx 1,5\%$ entspricht (vgl. Altman et al. 2017, Tab. 1 und Tab. 2). Altman et al. (2017) weisen darauf hin, dass für *WCTA*, *RETA* und *EBITTA* der Median und der Mittelwert bei nicht ausgefallenen Unternehmen nahe beieinander liegen, was auf symmetrische Verteilungen hinweist. Hingegen übersteigen bei ausgefallenen Unternehmen die Mediane aller drei Variablen die Mittelwerte, was auf rechtssteile (linksschiefe) Verteilungen deutet. Bei der Variablen *BVETD* unterschreiten die Mediane die Mittelwerte sowohl im Fall von nicht ausgefallenen als auch ausgefallenen Unternehmen. Dies bedeutet, dass die Verteilungen hier linkssteil (rechtsschief) sind (vgl. Altman et al. 2017, S. 153).

Mithilfe dieser Informationen lässt sich in RStudio ein Analysedatenbestand mit annahmegemäß 10.000 Beobachtungen wie folgt simulieren:

Auf Basis einer Binomialverteilung mit dem Wahrscheinlichkeitsparameter $p = 0,015$ können den 10.000 Beobachtungen die Ausfälle bzw. die Nichtausfälle zufällig zugeordnet werden, was im konkreten Fall zu einer beobachteten Ausfallquote von 1,47 % (147 Ausfälle) führt. Handelt es sich um einen Ausfall, dann nimmt die als binärer Ausfall-indikator definierte erklärende Variable den Wert $D = 1$ an. Im Fall eines Nichtausfalls gilt $D = 0$. Ob den gezeigten PDs eine sogenannte „Point-in-Time“ (PIT)- oder „Through-the-Cycle“ (TTC)-Ratingphilosophie bzw. Kalibrierungsphilosophie zugrunde liegt, steht nicht im Vordergrund der gezeigten Ausführungen. PIT-PDs orientieren sich eher an der aktuellen wirtschaftlichen Entwicklung und schwanken mit dem Konjunkturzyklus. TTC-PDs orientieren sich an langfristigen Mittelwerten über den Konjunkturzyklus hinweg. Sie tendieren dazu, konjunkturelle Schwankungen der PDs zu vermeiden. Damit haben sie den Vorteil, dass sich die regulatorisch erforderliche Mindestkapitalausstattung, die unter anderem von den PDs als Risikoparameter abhängt, ebenfalls weniger zyklisch verhält (vgl. Bellini 2019, S. 20; Siddiqi 2017, S. 127 f.; Baesens et al. 2016, S. 155 ff.).

Aus Gründen der Vereinfachung werden für die Simulation aller vier erklärenden Variablen $\mathbf{X} = \{WCTA, RETA, EBITTA, BVETD\}$ sowohl in der Gruppe der Ausfälle als auch der Nichtausfälle Normalverteilungen unterstellt, welche die von Altman et al. (2017) in ihrem Datenbestand beobachteten Mittelwerte und Standardabweichungen aufweisen. Konkret werden die Realisierungen der erklärenden Variablen

für die nicht ausgefallenen Kredite (*ND*) aus Normalverteilungen mit den folgenden Mittelwerten und Standardabweichungen gezogen: $WCTA \sim N(0,147, 0,442)$, $RETA \sim N(0,188, 0,509)$, $EBITTA \sim N(0,055, 0,227)$, $BVETD \sim N(3,594, 11,499)$. Für die ausgefallenen Kredite (*D*) bestimmen sich die entsprechenden Variablenwerte wie folgt: $WCTA \sim N(-0,213, 0,604)$, $RETA \sim N(-0,317, 0,767)$, $EBITTA \sim N(-0,108, 0,296)$, $BVETD \sim N(0,703, 5,712)$. Auf die in der Praxis übliche Standardisierung wird verzichtet, weil im Folgenden die Prognose von Ausfallwahrscheinlichkeiten gegenüber der Interpretierbarkeit bzw. Gewichtung einzelner erklärender Variablen den Vorzug erhält (vgl. Osborne 2015, S. 131 ff.).

Abschließend wird der Datenbestand per Zufallsauswahl ohne Zurücklegen in einen anhand der Verteilung des Ausfallindikators *D* stratifizierten Trainingsdatenbestand (70 % der Beobachtungen) und einen Testdatenbestand (30 % der Beobachtungen) aufgeteilt. Die zu erstellenden Modelle werden unter Verwendung des Trainingsdatenbestandes geschätzt und anschließend ihre Performance anhand des Testdatenbestandes evaluiert (Validierung „Out of Sample“).

Dies führt zu der in Tab. 21.1 dargestellten Deskriptivstatistiken für die numerischen Variablen im Trainings- und Testdatenbestand, wobei für alle erklärenden Variablen eine positive Wirkungsrichtung im Hinblick auf die erklärte Variable *D* zu vermuten ist:

Tab. 21.1 Deskriptivstatistiken im Trainings- und Testdatensatz. (Quelle: Eigene Darstellung)

Training					
Variable	D	WCTA	RETA	EBITTA	BVETD
Minimum	0,0000	-16,8530	-19,1570	-0,8633	-40,4960
1. Quartil	0,0000	-0,1449	-0,1769	-0,1032	-3,8470
Median	0,0000	0,1522	0,1694	0,0600	3,8230
Mittelwert	0,0151	0,1510	0,1739	0,0563	3,7070
3. Quartil	0,0000	0,4449	0,5186	0,2162	11,3510
Maximum	1,0000	17,4620	20,8520	0,9236	46,0380
Beobachtungen	7000	7000	7000	7000	7000
Test					
Variable	D	WCTA	RETA	EBITTA	BVETD
Minimum	0,0000	-15,9110	-23,6940	-0,7606	-38,0200
1. Quartil	0,0000	-0,1635	-0,1750	-0,0980	-4,2440
Median	0,0000	0,1407	0,1944	0,0525	3,5890
Mittelwert	0,0137	0,1355	0,1801	0,0527	3,5450
3. Quartil	0,0000	0,4345	0,5359	0,2023	11,1490
Maximum	1,0000	16,3080	18,1140	0,8656	45,8560
Beobachtungen	3000	3000	3000	3000	3000

21.4 Modellierung

Bevor die Logit- und Probit-Modelle auf den Untersuchungsdatenbestand angewendet werden, wird zunächst mit der WOE-Kodierung eine von zahlreichen Transformationsmöglichkeiten in Bezug auf die erklärenden Variablen vorgestellt. Es handelt sich dabei um einen univariaten Ansatz zur Gruppen- bzw. Klassenbildung („Binning“), der in der Praxis zur Aufbereitung der erklärenden Variablen in Logistischen Regressionsmodellen zum Einsatz kommt, welche die statistischen Eigenschaften sowie die Interpretierbarkeit des zu erstellenden Prognosemodells verbessern sollen. Die bei der WOE-Kodierung erzeugten Gewichte sind ein Maß dafür, wie gut eine erklärende Variable im statistischen Modell zwischen ausgefallenen und nicht ausgefallenen Krediten zu differenzieren vermag. Da die Logistische Regression definitionsgemäß einen linearen Zusammenhang zwischen den Log-Odds, $\ln(PD_i/(1 - PD_i))$, und den Modellparametern, β , in der Form $\ln(PD_i/(1 - PD_i)) = \beta'X_i$ voraussetzt, kann eine entsprechende WOE-Transformation dabei helfen, nichtlineare Zusammenhänge zu modellieren. Zudem führt sie zu stabileren Modellen, weil die Klassenbildung den Einfluss von Ausreißern reduziert, und trägt mitunter zu einem besseren Verständnis des Erklärungsbeitrages der erklärenden Variablen im statistischen Modell bei (vgl. Siddiqi 2017, S. 179 ff.; Rösch und Scheule 2020, S. 122; Bellini 2019, S. 39 ff.; Baesens et al. 2016, S. 82 ff.; vgl. aber Osborne 2015, S. 131 ff. und Harrell 2015, S. 18 ff. für eine Diskussion mit Gegenbeispielen).

Allerdings hat die WOE-Kodierung auch Nachteile. Da es sich bei ihr um ein univariates Klassifizierungskonzept handelt, reduziert sich ihre Präzision, wenn multivariate Modelle und Interaktionen zwischen den Variablen vorliegen. Ferner ist die Annahme, wonach beobachtete Ausfallquoten den datengenerierenden Prozess widerspiegeln, bei geringen Beobachtungs- und Ausfallzahlen nicht haltbar. Da die WOE-Kodierung auf den in Logistischen Regressionen verwendeten Log-Odds basiert, ist sie bei der Anwendung anderer Modellierungstechniken weniger genau (vgl. Rösch und Scheule 2020, S. 122).

Für die Berechnung der einzelnen Gewichte WOE_k (wobei $K = 1, \dots, k$ gilt) wird zunächst die Summe aller Nichtausfälle in Klasse k mit ND_k , die Summe aller Ausfälle in Klasse k mit D_k , die Summe aller Nichtausfälle im Gesamtdatenbestand mit $\sum_{k=1}^K ND_k$ und die Summe aller Ausfälle im Gesamtdatenbestand mit $\sum_{k=1}^K D_k$ berechnet. Die Gewichte ergeben sich dann als logarithmiertes Verhältnis zwischen dem Anteil aller nicht ausgefallenen Kredite, die auf Klasse k entfallen ($AND_k = ND_k / \sum_{k=1}^K ND_k$), und dem Anteil aller ausgefallenen Kredite, die auf Klasse k entfallen ($AD_k = D_k / \sum_{k=1}^K D_k$). Somit entspricht das Gewicht $WOE_k = \ln(AND_k/AD_k) = \ln\left(\sum_{k=1}^K ND_k/D_k\right) = \ln(ND_k/D_k) - \ln(\sum_{k=1}^K ND_k / \sum_{k=1}^K D_k)$ den in Klasse k vorliegenden Log-Odds eines Nichtausfalls, während das logistische Regressionsmodell allgemein auf den Log-Odds eines Ausfalles im Gesamtdatenbestand basiert (vgl. Siddiqi 2017, S. 182 ff.; Baesens et al. 2016, S. 101 ff.; Bellini 2019, S. 87; Rösch und Scheule 2020, S. 122, S. 156; Anderson 2007, S. 192).

Negative Werte von WOE_k bedeuten, dass in der betreffenden Klasse ein höherer Anteil ausgefallener als nicht ausgefallener Kredite vorliegt. Zudem zeigt der rechte Ausdruck in der obigen Gleichung, dass $WOE_k = 0$ gilt, wenn die Log-Odds eines Nichtausfalles in Klasse k , $\ln(ND_k / \sum_{k=1}^K ND_k)$, den Log-Odds eines Nichtausfalles im Gesamtdatenbestand, $\ln(\sum_{k=1}^K ND_k / \sum_{k=1}^K D_k)$, entsprechen (vgl. Anderson 2007, S. 192). Aus den derart ermittelten WOE_k lässt sich für jede Klasse k der sogenannte „Information Value“ $IV_k = (AND_k - AD_k) \cdot WOE_k$ ableiten. Durch Summierung wird ein Gesamtwert $IV = \sum_{k=1}^K IV_k$ ermittelt, der unmittelbar auf den gesamten Erklärungsbeitrag (Prädiktivität und Trennschärfe) rückschließen lässt, der mit der betrachteten Variablen erzielt werden kann. In der Praxis stehen (mit Abweichungen je nach Quelle) Werte von $IV < 0,1$ für einen schwachen Erklärungsbeitrag einer erklärenden Variablen, während Werte ab $IV > 0,3$ als starker Erklärungsbeitrag gewertet werden können (Anderson 2007, S. 192; Siddiqi 2017, S. 185; Rösch und Scheule 2020, S. 157; Bellini 2019, S. 88).

Tab. 21.2 zeigt die WOE-Kodierung und Berechnung des IV im Trainingsdatenbestand. Die Klasseneinteilung erfolgt iterativ, in der Praxis beispielsweise automatisiert mit den Prozeduren „PROC HPBIN“ und „PROC HPSPLIT“ in SAS oder mit dem Paket „smbinning“ in RStudio, um jene Gruppierungen bzw. Klassengrenzen zu finden, die nicht nur den IV der betreffenden Variablen maximieren, sondern auch logisch-ökonomisch Sinn machen und darüber hinaus eine möglichst linear aufsteigende oder absteigende Reihe von WOE_k -Werten erzeugen. Hier erfolgte die Klassifizierung in RStudio algorithmisch mithilfe von „Conditional Inference Trees“ im Paket „smbinning“ unter der Einschränkung, dass jede Klasse mindestens 5 % der Beobachtungen beinhalten muss und auf eine einzelne Klasse höchstens 50 % der Beobachtungen entfallen dürfen. In der Praxis werden mehr als die hier gezeigten Klassen verwendet und intuitivere Klassengrenzen gewählt (vgl. Jopia 2019; Hothorn et al. 2020; Siddiqi 2017, S. 186 ff.).

Alle vier erklärenden Variablen weisen einen $IV > 0,4$ auf und sind somit als stark prädiktiv einzustufen. Zudem zeigen die Variablen *WCTA*, *RETA* und *EBITTA* auch den erwarteten annähernd linearen Anstieg der WOE mit ansteigendem Klassenindex. Der Anteil der Nichtausfälle AND_k nimmt mit steigendem Variablenwert zu, während der Anteil der Ausfälle AD_k abnimmt. Eine Ausnahme hiervon stellt die Variable *BVETD* dar, deren WOE einen u-förmigen Verlauf beschreibt. So wäre für die Klasse $k = 1$, wo $BVETD \leq -8,1906$ gilt, ein stärker negatives WOE als in Klasse $k = 2$ zu erwarten. Dieses Muster ist auf die im Vergleich mit den Ausfällen viel weniger stark streuende Verteilung der Nichtausfälle zurückzuführen, das dazu führt, dass Ausfälle in erster Linie in Klasse $k = 2$ konzentriert sind.

An dieser Stelle ist darauf hinzuweisen, dass im empirischen Datensatz von Altman et al. (2017) bei der Variablen *BVETD* die Mediane die Mittelwerte sowohl im Fall von nicht ausgefallenen als auch ausgefallenen Unternehmen unterschreiten und dass die Verteilungen in der Empirie linkssteil (rechtsschief) sind. Ferner unterstreichen Altman et al. (2017), dass in der Realität die Werte von *BVETD* im unteren Quartil in der Gruppe

Tab. 21.2 WOE-Kodierung und Information Value. (Quelle: Eigene Darstellung)

Klasse [k]	Klassen grenze	Beob acht ungen	Anteil Beobachtungen	Nicht- ausfall (AND)	Ausfall (D)	Anteil Ausfall (AD)	Nicht Ausfall quote (non- default rate, NDR)	Ausfall quote (default rate, DR)	Odds Nichtaus- fall = NDR /DR	WOE = AND /AD	IV
Variable: WCTA											
[1]	$\leq -0,5794$	360	0,0514	330	0,0479	30	0,2830	0,9167	0,0833	11,0000	-1,7771
[2]	$\leq -0,2303$	1023	0,1461	995	0,1443	28	0,2642	0,9726	0,0274	35,5357	-0,6044
[3]	$> -0,2303$	5617	0,8024	5569	0,8078	48	0,4528	0,9915	0,0085	116,0208	0,5788
Total	-	7.000	1,0000	6.894	1,0000	106	1,0000	0,9849	0,0151	65,0377	0,0000
Variable: RETA											
[1]	$\leq -0,6525$	353	0,0504	318	0,0461	35	0,3302	0,9008	0,0992	9,0857	-1,9683
[2]	$\leq -0,2824$	949	0,1356	923	0,1339	26	0,2453	0,9726	0,0274	35,5000	-0,6054
[3]	$> -0,2824$	5698	0,8140	5653	0,8200	45	0,4245	0,9921	0,0079	125,6222	0,6583
Total	-	7.000	1,0000	6.894	1,0000	106	1,0000	0,9849	0,0151	65,0377	0,0000
Variable: EBITDA											
[1]	$\leq -0,2859$	518	0,0740	486	0,0705	32	0,3019	0,9382	0,0618	15,1875	-1,4545
[2]	$> -0,2859$	6482	0,9260	6408	0,9295	74	0,6981	0,9886	0,0114	86,5946	0,2863
Total	-	7.000	1,0000	6.894	1,0000	106	1,0000	0,9849	0,0151	65,0377	0,0000
Variable: BVETID											
[1]	$\leq -8,1906$	1041	0,1487	1036	0,1503	5	0,0472	0,9952	0,0048	207,2000	1,1587
[2]	$\leq 8,7774$	3649	0,5213	3552	0,5152	97	0,9151	0,9734	0,0266	36,6186	-0,5744

(Fortsetzung)

Tab. 21.2 (Fortsetzung)

der Ausfälle deutlich geringer sind als jene in der Gruppe der Nichtausfälle. Da dieses Charakteristikum bei der hier gezeigten Simulation nicht berücksichtigt wurde, führt die größere Standardabweichung bei den Nichtausfällen im simulierten Datenbestand zum erwähnten u-förmigen Verlauf der WOE für die Variable *BVETD*.

In der Praxis ist je nach Anwendungserfordernis und Ursache der Nichtlinearität erstens eine manuelle Korrektur des WOE der Klasse $k = 1$, WOE_1 , denkbar, um eine konservativere Berücksichtigung der zu erwartenden höheren Ausfälle im Bereich der niedrigsten Werte von *BVETD* zu erreichen. Zweitens ist es denkbar, die Klassengrenzen manuell anzupassen und Klassen derart zusammenzufassen, dass ein monotoner Verlauf der WOE von Klasse zu Klasse erzielt wird. Dies geht mit einem verminderter IV der erklärenden Variablen einher (vgl. Siddiqi 2017, S. 186 ff.). Im Folgenden werden die Klassen 1 und 2 gemäß der in diesem Abschnitt beschriebenen zweiten Möglichkeit zusammengefasst. Es verbleibt lediglich die Einteilung in die beiden Klassen $BVETD \leq 8,7774$ und $BVETD > 8,7774$ übrig. Dies ist als „*BVETD_c*“ im unteren Bereich der Tabelle kenntlich gemacht.

21.5 Überprüfung der Modellannahmen

Vor der Durchführung der Logistischen oder Probit-Regression ist es zweckmäßig, die Gültigkeit der eingangs beschriebenen Linearitäts- und Additivitätsannahme der Log-Odds bzw. des Probit im Hinblick auf die erklärenden Variablen zu überprüfen, da nichtlineare oder nichtmonotone Zusammenhänge von diesen Modelltypen nicht ohne Weiteres adäquat abgebildet werden können. Eine praktische Möglichkeit zur Beurteilung von Nichtlinearitäten in Logit- und Probit-Modellen gleichermaßen besteht in der direkten Aufnahme quadratischer und kubischer Terme in die Regressionsgleichung (im Hinblick auf die Linearitätsannahme) sowie der Anwendung von Interaktionstermen (im Hinblick auf die Additivitätsannahme) und anschließenden Überprüfung, ob die betreffenden Koeffizienten statistisch signifikant sind. Diese sowie weiterführende Strategien zur Überprüfung und Sicherstellung von Linearität sowie Additivität diskutieren beispielsweise Hayden (2011), Hosmer et al. (2013) und Harrell (2015). Es handelt sich dabei um geglättete Scatterplots, Indikatorvariablen, Polynommodelle und Spline-Funktionen. Ebenso ließe sich jede der k erklärenden Variablen X_k mithilfe sogenannter Box-Tidwell-Transformationen in die Form $V_k = X_k \cdot \ln(X_k)$ überführen, um Linearität sicherzustellen (vgl. Hayden 2011, S. 17 ff., Hosmer et al. 2013, S. 94 ff.; Harrell 2015, S. 18 ff., S. 236 ff.; Box und Tidwell 1962; Osborne 2015, S. 98 ff., S. 208 ff.). Die praktische Implementierung von Polynommodellen zeigen zum Beispiel Ambler und Royston (2001), Sauerbrei et al. (2006), Ambler et al. (2015) und Fingerlos et al. (2020).

Neben der Linearität und Additivität des Logit bzw. Probit ist ebenso zu überprüfen, wie stark die erklärenden Variablen miteinander korreliert sind (Multikollinearität).

Die hier nicht abgebildeten Pearson-Korrelationskoeffizienten im Fall der untransformierten (kontinuierlichen) bzw. die Spearman-Korrelationskoeffizienten im Fall der WOE-transformierten (gruppierten) erklärenden Variablen bestätigen, dass eine Multikollinearität im vorliegenden Datenbestand nicht auftritt (vgl. Wooldridge 2016, S. 83 ff.; Osborne 2015, S. 254 ff.). Der folgende Abschnitt zeigt die Ergebnisse einer Reihe von Logit- und Probit-Modellen, die neben der linearen Originalspezifikation von Altman et al. (2017) auch einige erweiterte Spezifikationen beinhalten.

21.6 Vorstellung der Ergebnisse

Die in diesem Beitrag gezeigten Logit- und Probit-Modellvarianten basieren auf der linearen Originalspezifikation von Altman et al. (2017), auf Spezifikationen unter Einbeziehung quadratischer Terme und Interaktionsterme sowie im Fall der Logit-Regression auf Spezifikationen auf Basis der WOE-Transformation (einerseits Indikatorvariablen und andererseits WOE-Gewichte als Werte der erklärenden Variablen). Tab. 21.2 mit der optimalen WOE-Kodierung gibt den Hinweis, dass insbesondere bei der Variablen *BVETD* von einem stark nichtmonotonen Zusammenhang zwischen Log-Odds und den Variablenwerten auszugehen ist (vgl. Siddiqi 2017, S. 186 ff.; Baesens et al. 2016, S. 101 ff.; Bellini 2019, S. 40 ff.).

Tab. 21.3 bildet eine Reihe von logistischen Modellen (Spezifikation 1 bis 4) sowie Probit-Modellen (Spezifikation 5 und 6) ab, die allesamt auf Basis der Trainingsdaten geschätzt wurden. Spezifikation (1) nimmt die erklärenden Variablen *WCTA*, *RETA*, *EBITTA* und *BVETD* wie in der Originalspezifikation von Altman et al. (2017) ohne jede Transformation auf. Spezifikation (2) inkludiert quadratische Terme sowie Interaktionsterme erster Ordnung. Es wurden auch kubische Terme in eine weitere Spezifikation mit aufgenommen, die allerdings statistisch insignifikant blieben. Spezifikation (3) beinhaltet die erklärenden Variablen in Form von Indikatorvariablen, die mithilfe der optimalen WOE-Transformation gebildet wurden. In Spezifikation (5) fließen die erklärenden Variablen hingegen indirekt mit den zugehörigen WOE-Gewichten aus der WOE-Kodierung ein. Die Spezifikationen (5) und (6) replizieren die Spezifikationen (1) und (2) auf Basis einer Probit-Regression anstelle der Logit-Regression.

Während die Spezifikationen (1) und (5) nichtlineare Zusammenhänge ohne weitere Transformation der erklärenden Variablen nicht korrekt erfassen, werden sie in Spezifikation (3) aufgrund der Kodierung mithilfe von Indikatorvariablen korrekt abgebildet. Spezifikation (3) ist weniger anfällig in Bezug auf Ausreißer, hat allerdings den Nachteil, wesentlich mehr Koeffizienten bestimmen zu müssen und somit mehr Freiheitsgrade zu verbrauchen. Dies kann in Portfolios mit geringer Beobachtungsanzahl mitunter zu einem Nachteil werden. Spezifikation (4) ist wie Spezifikation (1) sparsam im Verbrauch von Freiheitsgraden, weil sie die empirischen Variablenwerte jeder erklärenden Variable durch die zuvor ermittelten WOE-Gewichte ersetzt und somit die Anzahl der zu schätzenden Parameter mit der ursprünglichen Anzahl von erklärenden Variablen

Tab. 21.3 Ergebnisse. (Quelle: Eigene Darstellung)

Erklärende Variablen	Erklärte Variable: D					
	Logit-Modell			Probit-Modell		
(1)	(2)	(3)	(4)	(5)	(6)	
WCTA	-1,607 ***(0,222)	-1,318 *** (0,197)		-0,662 *** (0,096)	-0,618 *** (0,090)	
WCTA_2		1,122 *** (0,246)			0,512 *** (0,118)	
RETA	-1,742 *** (0,196)	-1,240 *** (0,165)		-0,671 *** (0,085)	-0,592 *** (0,074)	
RETA_2		1,203 *** (0,179)			0,599 *** (0,082)	
EBITTA	-2,609 *** (0,435)	-1,958 *** (0,398)		-0,997 *** (0,185)	-0,912 *** (0,172)	
EBITTA_2		4,388 *** (1,018)			2,303 *** (0,467)	
BVETD	-0,031 *** (0,009)	-0,032 * (0,019)		-0,014 *** (0,004)	-0,014 * (0,008)	
BVETD_2		-0,013 *** (0,002)			-0,006 *** (0,001)	
WCTA_- k02 ≤ -0,2303			-0,968 *** (0,302)			
WCTA_- k03 > -0,2303				-2,245 *** (0,269)		
RETA_- k02 ≤ -0,2824				-1,406 *** (0,293)		
RETA_- k03 > -0,2824				-2,676 *** (0,259)		
EBITTA_- k02 > -0,2859				-1,749 *** (0,246)		

(Fortsetzung)

Tab. 21.3 (Fortsetzung)

Erklärende Variablen	Erklärte Variable: D					
	Logit-Modell			Probit-Modell		
(1)	(2)	(3)	(4)	(5)	(6)	
BVETD_ k02 ≤ 8,7774		1,691 *** (0,468)				
BVETD_ k03 > 8,7774		-1,166 * (0,679)				
WOE_WCTA			-0,983 *** (0,110)			
WOE_RETAA			-1,015 *** (0,096)			
WOE_EBITTA			-1,010 *** (0,139)			
WOE_BVETD_c			-1,040 *** (0,203)			
Interzept	-4,285 *** (0,124)	-4,465 *** (0,204)	-0,144 (0,571)	-4,173 *** (0,131)	-2,154 ** (0,046)	-2,315 *** (0,085)
Beobachtungen	7.000	7.000	7.000	7.000	7.000	7.000
Log Likelihood	-446,237	-375,882	-396,782	-407,238	-454,581	-373,026
Akaike Inf. Crit	902,474	769,764	809,564	824,477	919,161	764,051

Anmerkungen: *: p<0,1; **: p<0,05; ***: p<0,01; Standardfehler in Klammern

(plus Interzept) übereinstimmt. Allerdings ist es hier zweckmäßig, die WOE-Gewichte der nachträglich manuell gruppierten Variable *BVETD_c* anstelle der WOE-Gewichte der optimal transformierten Variable *BVETD* zu verwenden, um nicht das nichtmonotone Verhalten letzterer in der Regression zu umgehen.

Die signifikanten Koeffizienten der quadratischen Terme in den Spezifikationen (2) und (6) bestätigen, dass Nichtlinearitäten vorliegen, die in den Spezifikationen (1) und (5) nicht korrekt berücksichtigt werden. Die bessere Anpassungsgüte der Modelle (2) und (6) zeigt sich auch in den geringeren Werten des Akaike-Informationskriteriums. Variablenamen plus „_2“ (z. B. *WCTA_2*) bezeichnen die quadratischen Terme, während Variablenamen plus „_k..“ (z. B. *WCTA_k02*) die Indikatorvariablen aus der WOE-Kodierung bezeichnen. Die jeweils erste Kategorie „_k01“ entfällt, um die sogenannte „Dummy-Variable-Trap“ zu vermeiden, die aus der perfekten Multikollinearität von Dummy-Variablen für alle Kategorien resultieren würde und eine Verunmöglichung der Parameterschätzung zur Folge hätte. Variablenamen, die mit „WOE_“ beginnen (z. B. *WOE_WCTA*), stehen für die WOE-Gewichte, die anstelle der Variablenwerte eingesetzt wurden. Sind quadratische Terme einer erklärenden Variablen signifikant und zugleich nicht quadrierte Terme derselben erklärenden Variablen insignifikant, bleiben auch die letztgenannten im Modell.

21.7 Vergleichende Beurteilung

Der abschließende Vergleich der erstellten PD-Modelle erfolgt mithilfe zweier Indikatoren der Prognosegüte, die auch aus regulatorischer Sicht bei der Modellvalidierung zur Anwendung kommen. Die Indikatoren sind einerseits die Trennschärfe („Discriminatory Power“) und andererseits die Kalibrierung bzw. Prognosefähigkeit („Predictive Ability“) eines PD-Modells. Eine hohe Trennschärfe ist gegeben, wenn in guten Ratingklassen ein geringer Anteil und in schlechten Ratingklassen ein hoher Anteil der insgesamt ausfallenden Kredite auftritt. Ferner liegt eine gute Kalibrierung eines PD-Modells vor, wenn die prognostizierten Ausfallwahrscheinlichkeiten nahe an den später eingetretenen Ausfallquoten liegen (vgl. Deutsche Bundesbank 2003, S. 64). Für beide Bereiche existiert eine Reihe verschiedener Kennzahlen (vgl. beispielsweise Baesens et al. 2016, S. 385 ff.; Castermans et al. 2010, S. 359 ff.; Blochwitz und Hohl 2011, S. 254 ff.; Engelmann 2011, S. 269 ff.; Deutsche Bundesbank 2003, S. 71 ff.).

Da aus regulatorischer Sicht einerseits die Verwendung der Fläche unter der „Receiver Operating Characteristic“ (ROC)-Kurve („Area Under the Curve“, AUC) zur Beurteilung der Trennschärfe und andererseits die Verwendung von Jeffreys-Tests zur Beurteilung der Kalibrierung (Prognosefähigkeit) von internen PD-Modellen präferiert wird, kommen diese beiden Konzepte auch im hier gezeigten Vergleich zur Anwendung (vgl. European Central Bank 2019, S. 14 ff.). Weicht in der Praxis in den periodisch auf Grundlage des neuesten, nicht bei der Modellierung verwendeten Datenbestandes (Anwendungsdaten) durchzuführenden Rückvergleichsanalysen („Backtests“)

insbesondere die Kalibrierung der Modelle von den zuvor festgelegten Mindestanforderungen dauerhaft zu stark ab, hat dies in der Regel langfristig eine Neuschätzung des betreffenden Modells zur Folge (vgl. European Banking Authority 2018, S. 64 ff.; European Central Bank 2019, S. 14 ff.).

Tab. 21.4 enthält einerseits die auf Basis der Testdaten mit dem Paket „pROC“ in RStudio berechneten AUC-Werte der zuvor erstellten Modelle, inklusive beidseitiger 95 %-Konfidenzintervalle, die mithilfe der Methode von DeLong et al. (1988) auf Basis der Testdaten erzeugt wurden (vgl. Robin et al. 2020). Andererseits fasst sie die auf den PD-Prognosen im Testdatenbestand basierenden, beidseitigen $100(1 - \alpha)\%$ Jeffreys-Konfidenzintervalle zusammen, die für ein Signifikanzniveau $\alpha = 0,05$ um die vom betreffenden Modell prognostizierte Kreditausfallwahrscheinlichkeit $\widehat{PD} = x/n$ gelegt werden. Dabei bezeichnet $n = 3000$ die Beobachtungsanzahl im Testdatensatz, $x = n\widehat{PD}$ die Anzahl der prognostizierten Kreditausfälle und $DR = (\sum D)/n$ die Ausfallquote. Für die geschätzte Kreditausfallwahrscheinlichkeit \widehat{PD} lassen sich die Untergrenze UG und die Obergrenze OG des Jeffreys-Konfidenzintervalls, das auf der Quantilsfunktion $B^{-1}(\alpha; a; b)$ einer Beta-Verteilung mit den Parametern a und b basiert ($Beta(a; b)$), über die Formeln $UG = B^{-1}(\alpha/2; x + 1/2; n - x + 1/2)$ und $OG = B^{-1}(1 - \alpha/2; x + 1/2; n - x + 1/2)$ berechnen, wobei $UG(0) = 0$ und $OG(N) = 1$ festgelegt werden (vgl. Brown et al. 2001, S. 108 ff.).

Liegt DR innerhalb des Konfidenzintervalls, dann kann die Nullhypothese $H_0 : DR = \widehat{PD}$ nicht verworfen werden. Das untersuchte PD-Modell ist als ausreichend genau und somit akzeptabel einzustufen. Liegt hingegen DR über der Obergrenze OG des Konfidenzintervalls, dann ist die Nullhypothese $H_0 : DR = \widehat{PD}$ zu verwerfen und die Alternativhypothese $H_1 : DR \neq \widehat{PD}$ zu akzeptieren. In diesem Fall unterschätzt das Modell die DR und es ist eine Neukalibrierung erforderlich. Wenn jedoch die DR so tief liegt, dass sie selbst die Untergrenze des Konfidenzintervalls noch unterschreitet, dann lässt sich das Modell als zu konservativ (zu ungenau) einstufen und gegebenenfalls ebenfalls eine Neukalibrierung anstoßen. Dieser Fall ist insofern weniger problematisch, als sich hier die Schätzung zumindest auf der sicheren Seite befindet (vgl. Castermans et al. 2010, S. 362 ff.; Basel Committee on Banking Supervision 2005, S. 47 ff.; Brown et al. 2001, S. 101 ff.; Blochwitz et al. 2011, S. 293 ff.; Rauhmeier 2011, S. 320 ff.; Baesens et al. 2016, S. 407 f.).

Das hier rund um den PD-Schätzer gezeichnete beidseitige Jeffreys-Konfidenzintervall bietet im Gegensatz zur regulatorischen Vorgabe in European Central Bank (2019) den Vorteil, wie in statistischen Testverfahren üblich das Konfidenzintervall um den Prognosewert \widehat{PD} legen zu können und somit intuitiv besser verständlich zu sein. European Central Bank (2019) fordert in diesem Zusammenhang einen linksseitigen Jeffreys-Test, der auf einer Berechnung von p-Werten für die PD-Schätzung (\widehat{PD}) beruht: $H_0 : \widehat{PD} > DR$, $H_1 : \widehat{PD} \leq DR$. In diesem Fall wird also ein nach rechts (oben) offenes Konfidenzintervall um die Ausfallsquote, DR , gelegt und die Nullhypothese verworfen, wenn \widehat{PD} kleiner als der kritische Wert der DR des Tests ist bzw. der für \widehat{PD} berechnete p-Wert kleiner als das Signifikanzniveau α ist (vgl. European Central Bank 2019, S.

Tab. 21.4 Modellgüte. (Quelle: Eigene Darstellung)

Modell	Trennschärfe („Discriminatory Power“)		Kalibrierung („Predictive Ability“)		$\widehat{PD} - DR$	Jeffreys-Test
	AUC	AUC95%CI	\widehat{PD}	UG		
Logit-Modell (1)	0,8125	0,7370–0,8881	1,113 %	0,783 %	1,537 %	1,367 %
Logit-Modell (2)	0,8854	0,8312–0,9395	1,433 %	1,053 %	1,906 %	1,367 %
Logit-Modell (3)	0,8738	0,8199–0,9277	1,476 %	1,090 %	1,956 %	1,367 %
Logit-Modell (4)	0,8536	0,7984–0,9089	1,482 %	1,095 %	1,962 %	1,367 %
Probit-Modell (5)	0,8124	0,7368–0,8881	1,505 %	1,114 %	1,988 %	1,367 %
Probit-Modell (6)	0,8858	0,8320–0,9396	1,442 %	1,061 %	1,917 %	1,367 %

20). Das Vorgehen von European Central Bank (2019) hat wiederum den Vorteil, dass sich beim Verwerfen der Nullhypothese – in diesem Fall unterschätzt die geschätzte Ausfallwahrscheinlichkeit \widehat{PD} die beobachtete Ausfallquote DR statistisch signifikant – aus der Differenz zwischen dem kritischen Wert der DR und dem Schätzwert \widehat{PD} ein Aufschlag berechnen lässt, der zum Schätzwert \widehat{PD} addiert werden müsste, um in den Nichtablehnungsbereich des Tests zu kommen. Ein solcher Aufschlag wird auch als Sicherheitsspanne („Margin of Conservatism“) bezeichnet, welche aus regulatorischer Sicht die Auswirkungen derartiger Schätzfehler (das Unterschätzen der DR) mildern soll (vgl. European Banking Authority 2018, S. 15 ff.).

Der Vergleich der Modelle veranschaulicht, dass alle im Testdatensatz eine ansprechende Performance (Trennschärfe und Prognosefähigkeit) zeigen und überdies die Logit- und Probit-Modelle insgesamt praktisch identische Ergebnisse liefern.

Werden in Bezug auf die Trennschärfe (AUC) die in der Literatur (vgl. Hosmer et al. 2013, S. 177; Castermans et al. 2010, S. 366) üblichen Standardwerte zugrunde gelegt, dann fallen alle Modelle in die zweitbeste Kategorie: $AUC = 0,5$: keine Trennschärfe; $0,5 < AUC < 0,7$: geringe Trennschärfe; $0,7 \leq AUC < 0,8$: akzeptable Trennschärfe; $0,8 \leq AUC < 0,9$: exzellente Trennschärfe; $AUC \geq 0,9$: außergewöhnliche Trennschärfe. Dies ist in Anbetracht der Einfachheit der Modelle und im Vergleich mit der Modellierungspraxis ein zufriedenstellender Wert. Am besten schneiden das Logit-Modell (2) und das Probit-Modell (6) ab, die beide mit quadratischen Termen sowie (insignifikanten und deshalb im finalen Modell nicht gezeigten) kubischen Termen und Interaktionstermen kalibriert wurden und so im Gegensatz zum Logit-Modell (1) und dem Probit-Modell (5) auch Nichtlinearitäten erfassen können. Knapp dahinter finden sich die Logit-Modelle (3) und (4), die aufgrund der für Logit-Modelle spezifischen WOE-Kodierung sparsamer im Verbrauch von Freiheitsgraden sind als die Modelle mit quadratischen Termen.

Im Hinblick auf die Kalibrierung bzw. Prognosefähigkeit schneiden ebenfalls alle Modelle sehr zufriedenstellend ab. Mit Ausnahme des Logit-Modells (1) sind zudem alle Modelle konservativ: Die Prognosewerte \widehat{PD} fallen im Testdatenbestand höher aus als die Ausfallquoten DR . Lediglich das Logit-Modell (1) unterschätzt die DR geringfügig, bleibt aber ebenfalls innerhalb des Jeffreys-Konfidenzintervalls. Insgesamt ist somit zu konstatieren, dass im vorliegenden vereinfachten Rahmen die fortgeschritteneren Spezifikationen aufgrund ihrer besseren Performance gegenüber der Basisspezifikation von Altman et al. (2017) zu bevorzugen sind.

Literatur

- Altmann, E.I.: Financial ratios discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23, 589–609 (1968)
- Altmann, E.I., Hochkiss, E.: Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed Debt, 3. Aufl. Wiley, Hoboken (2006)

- Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K., Suvas, A.: Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-Score model. *Journal of International Financial Management & Accounting* **28**(2), 131–171 (2017)
- Ambler, G., Royston, P.: Fractional polynomial Model Selection procedures: Investigation of type I Error Rates. *J. Stat. Comput. Simul.* **69**, 89–108 (2001)
- Ambler, G., Benner, A., Luecke, S.: Multivariable fractional polynomials, package ‘mfp’, 9. September 2015, Version 1.5.2. <https://cran.r-project.org/web/packages/mfp/mfp.pdf> (2015). Zugegriffen: 28. Aug. 2020
- Anderson, R.: The Credit Scoring Toolkit: Theory and practice for retail credit risk management and decision automation. Oxford University Press, Oxford (2007)
- Baesens, B., Rösch, D., Scheule, H.: Credit risk analytics: Measurement techniques, applications, and examples in SAS. Wiley, Hoboken (2016)
- Basel Committee on Banking Supervision.: Studies on the validation of internal rating systems. Working Paper No. 14, Revised Version May 2005. https://www.bis.org/publ/bcbs_wp14.pdf (2005). Zugegriffen: 28. Aug. 2020
- Bellini, T.: IFRS 9 and CECL credit risk modelling and validation: A practical guide with examples worked in R and SAS. Academic Press, London (2019)
- Blochwitz, S., Martin, M.R.W., Wehn, C.S.: Statistical approaches to PD Validation. In: Engelmann, B., Rauhmeier, R. (Hrsg.) The basel II risk parameters: Estimation, validation, stress testing – with applications to loan risk management, 2. Aufl., S. 293–309. Springer, Berlin (2011)
- Box, G.E.P., Tidwell, P.W.: Transformation of the independent variables. *Technometrics* **4**(4), 531–550 (1962)
- Brown, L.D., Cai, T., DasGupta, A.: Interval estimation for a binomial proportion. *Statistical Science* **16**(2), 101–117 (2001)
- Castermans, G., Martens, D., Van Gestel, B., Hamers, B., Baesens, B.: An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society* **61**, 359–373 (2010)
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988)
- Deutsche Bundesbank: Validierungsansätze für Interne Ratingsysteme. Monatsbericht September, 55(9), 61–74. Eigenverlag, Frankfurt a. M (2003)
- Engelmann, B.: Measures of a rating's discriminative power: Applications and limitations. In: Engelmann, B., Rauhmeier, R. (Hrsg.) The basel II risk parameters: Estimation, validation, stress testing – with applications to loan risk management, 2. Aufl., S. 269–291. Springer, Berlin (2011)
- European Banking Authority: Leitlinien für die PD-Schätzung, die LGD-Schätzung und die Behandlung von ausgefallenen Risikopositionen, EBA/GL/2017/16, 23/04/2018. https://eba-europa.eu/documents/10180/2192133/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29_DE.pdf (2018). Zugegriffen: 28. Aug. 2020
- European Central Bank: Instructions for reporting the validation results of internal models – IRB Pillar I models for credit risk, February 2019. https://www.bankingsupervision.europa.eu/banking/tasks/internal_models/shared/pdf/instructions_validation_reporting_credit_risk.en.pdf (2019). Zugegriffen: 28. Aug. 2020
- Fingerlos, U.R., Golla, G., Pastwa, A., Gluchowski, P., Gabriel, R.: Risikoreporting in Finanzinstituten: Anforderungen, Konzepte. Prototyping. Springer Gabler, Wiesbaden (2020)
- Greene, W.H.: Econometric analysis, 8. Aufl. Pearson Education, New York (2018)

- Harrell, F.E., Jr.: Regression modeling strategies – With applications to linear models, logistic and ordinal regression, and survival analysis, 2. Aufl. Springer, Cham (2015)
- Hayden, E.: Estimation of a rating model for corporate exposures. In: Engelmann, B., Rauhmeier, R. (Hrsg.) The basel II risk parameters: Estimation, validation, stress Testing – with applications to loan risk management, 2. Aufl., S. 13–24. Springer, Berlin (2011)
- Hill, R.C., Griffiths, W.E., Lim, G.C.: Principles of econometrics, 5. Aufl. Wiley, Hoboken (2018)
- Hosmer, D.W., Jr., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, 3. Aufl. Wiley, Hoboken (2013)
- Hothorn, T., Seibold, H., Zeileis, A.: A toolkit for recursive partytitioning, package ‘partykit’, 10. Juli 2020, Version 1.2–9. <https://cran.r-project.org/web/packages/partykit/partykit.pdf> (2020). Zugriffen: 28. Aug. 2020
- Jopia, H.: Scoring modeling and optimal binning, package ‘smbinning’, 1. April 2019, Version 0.9. <https://cran.r-project.org/web/packages/smbinning/smbinning.pdf> (2019). Zugriffen: 28. Aug. 2020
- Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C.: Benchmarking State-of-the-Art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **247**, 124–136 (2015)
- Löffler, G., Posch, P.N.: Credit risk modeling using excel and VBA, 2. Aufl. Wiley, Chichester (2011)
- Osborne, J.W.: Best practices in logistic regression. Sage Publications, Thousand Oaks (2015)
- Rauhmeier, R.: PD-Validation: Experience from Banking Practice. In: Engelmann, B., Rauhmeier, R. (Hrsg.) The basel II risk parameters: Estimation, validation, stress Testing – with applications to loan risk management, 2. Aufl., S. 311–347. Springer, Berlin (2011)
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Siegert, S., Doering, M.: Display and analyze ROC curves, package ‘pROC’, 19. März 2020, Version 1.16.2. <https://cran.r-project.org/web/packages/pROC/pROC.pdf> (2020). Zugriffen: 28. Aug. 2020
- Rösch, D., Scheule, H.: Deep credit risk: Machine learning with python. AutorenServices, Fulda (2020)
- Sauerbrei, W., Meier-Hirmer, C., Benner, A., Royston, P.: Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Comput. Stat. Data Anal.* **50**, 3464–3485 (2006)
- Siddiqi, N.: Intelligent credit scoring: Building and implementing better credit risk scorecards, 2. Aufl. Wiley, Hoboken (2017)
- Wooldridge, J.M.: Introductory econometrics – A modern approach, 6. Aufl. Cengage Learning, Boston (2016)

Stichwortverzeichnis

A

- Ad-hoc-Reporting, 175
- Akaike-Informations-Kriterium, 185
- Algorithmus, 209
- Analysereporting, 175
- Analytics Culture, 306
- Application Centric, 150
- Arbeitswelt 4.0, 65
- Assoziationsanalyse, 191
- Ausbildung, 278

B

- Balanced Scorecard, 168
- Barcamps, 69
- Bayes
 - Native, 187
 - Thomas, 242
- Bayes-Theorem, 242
- Big Data, 54
 - Definition, 3
 - Hype, 54
- Big-Data-Plattform, 148
- Bitkom Akademie, 287
- Business
 - Analytics (BA), XIII
 - Intelligence (BI), XI, XIII, 170, 180

C

- CarSharing, 322
- CAS Data
 - Analyst, 286
 - Architect, 286
 - Strategist, 286

Chatbots, 215

- Clustering-Verfahren, 191
- Convolutional-Schichten, 231
- Core Data Warehouse, 170
- Corporate Governance, 106
- CRISP-DM, 86
- Crowdworker, 67

D

- Data
 - Application, 34
 - Catalog, 115
 - Catalog Vocabulary (DCAT), 93
 - Cleaning, 95
 - Consulting, 36
 - Engineering, 86
 - Governance, 106
 - Lake, 91, 306
 - Literacy, 28
 - Literacy Project, 29
 - Management, 86
 - Mining, 6, 216
 - Pedigree, 114
 - Profiling, 92
 - Provenance, 114
 - Science, 180
 - Scientist, 60
 - Warehousing, XI
 - Wrangling, 87
- Data-Lake-Systemen, 90
- Data-Lineage, 114
- Datafication, XII
- DataLab, 36
- Datanode, 138

- DataScience Pipeline, 55
Daten
 Management, 86
 unstrukturierte, 322
Daten Lesen Lernen
 Projekt am Campus Göttingen, 36
Datenanalyse, 179
Datenbereinigung, 95
Datendarstellung, 197
Datenkompetenz, 28
Datenmodellierung, 87
Datenvisualisierung, 179
Decision Support System, 170
Design Thinking, 76, 77
DevCamps, 69
Digital
 Immigrants, 69
 Leader, 72
 Leadership, 64
 Natives, 69
Donut-Cloud, 10
- E**
Ebenenmodell, 42, 45
EDISON-Rahmenwerk, 286
Enable-Prozess, 46
Entity-Relationship-Modell, 171
Entscheidungsbaum, 200
ETL, 88
Exception Reporting, 175
- F**
Fakttabelle, 172
Flare-Chart, 10
Fraunhofer, 287
Führung
 agile, 64
 virtuelle, 64
- G**
Geschäftsmodell, 15
Governance, 105
- H**
Hadoop, 7
- Hashed Sharding, 136
Hub-and-Spoke-Architektur, 170
Hype-Thema, XI
- I**
In-Memory-Computing, 123
In-Memory-Technologie, 9, 121
Information Engineering, 86
Informationsmanagement, 41
Institutes der Deutschen Wirtschaft, 53
IT-Governance, 43
IT-Strategie, 43
- K**
Karte, selbstorganisierende, 231
KDD-Prozess, 86
Kennzahl, 168
Kennzahlensteckbrief, 169
Klassifikationsverfahren, 186
Knowledge Discovery in Databases (KDD), XII
Konfusionsmatrix, 202
Künstliche Intelligenz (KI), 180
Künstliche neuronale Netze (KNN), 212, 226
- L**
Lernprozess, 211
Lernverfahren, 211
Logit-Modell, 337
- M**
Machine Learning, 93, 226, 306
Make-Prozess, 44
Management-Information-System, 170
Map-Reduce, 139
Map-Reduce-Anweisung, 141
Markov-Kette, 248
Medizin, bildgebende, 215
Methode, agile, 73
Mission, 109
- N**
Namenode, 138
Neuron, 226

- künstliches, 227
New Work, 64
NoSQL-Datenbank, 6, 87, 174
- O**
OLAP, 88
OLTP, 88
Overfitting, 34
- P**
Partitioning, 136
PDCA-Zyklus, 168
Polizeibehörde, 53
Polyglot Persistence, 88, 102
Principal Component Analysis (PCA), 196, 198
Probit-Modell, 337
Product Owner, 74
Profiling, 96
Python, 219
PyTorch, 219
- Q**
Queries, 127
- R**
Rahmenkonzept, 44
zum Informationsmanagement, 41
Ranged Sharding, 136
RDBMS, 123
Recurrent Neural Networks, 232
Regression, 246
bayesianische, 246
klassische, 246
lineare, 182
Reifegradmessung, 57
Reporting, 168
Reverse Engineering, 87
- S**
SAP HANA, 261
- Scale Out, 135
Scoring, 338
Scrum, 73, 332
Scrum-Master, 74
Sentiment Analysis, 260, 263
Sharding, 136
Shards, 137
Shuffling, 140
Sigmoidfunktion, 228
Sliding Window, 143
Snowflake-Schema, 172
Source-Prozess, 44
Standardreporting, 175
Stemming, 263
- T**
TensorFlow, 219
Text Mining, 9, 205
Tumbling Window, 143
- V**
Value, 5
Variablentransformation, 185
Variety, 4
Velocity, 4
Veracity, 5
Vision, 109
Volume, 3
VOPAModell, 77
- W**
Weiterbildung in Data Science, 278
Weiterbildungsangebot, 278
Word-Cloud, 10
- Z**
Z-Score-Modell, 340
Zahlungsausfall, 338