Edition HMD

Sara D'Onofrio
Andreas Meier *Hrsg.*

Big Data Analytics

Grundlagen, Fallbeispiele und Nutzungspotenziale

Praxis der Wirtschaftsinformatik

HMD

EBOOK INSIDE

Springer Vieweg

Edition HMD

Reihe herausgegeben von

Sara D'Onofrio, IT Business Integration, Genossenschaft Migros Zürich, Zürich, Schweiz

Hans-Peter Fröschle, i.t-consult GmbH, Stuttgart, Deutschland

Josephine Hofmann, Fraunhofer IAO, Stuttgart, Deutschland

Matthias Knoll, FB Wirtschaft, Hochschule Darmstadt, Darmstadt, Deutschland

Stefan Meinhardt, SAP Deutschland SE & Co KG, Walldorf, Deutschland

Stefan Reinheimer, BIK GmbH, Nürnberg, Deutschland

Susanne Robra-Bissantz, Inst. Wirtschaftsinformatik, TU Braunschweig, Braunschweig, Deutschland

Susanne Strahringer, Fakultät Wirtschaftswissenschaften, TU Dresden, Dresden, Deutschland

Die Fachbuchreihe „Edition HMD“ wird herausgegeben von Dr. Sara D'Onofrio, Hans-Peter Fröschle, Dr. Josephine Hofmann, Prof. Dr. Matthias Knoll, Stefan Meinhardt, Dr. Stefan Reinheimer, Prof. Dr. Susanne Robra-Bissantz und Prof. Dr. Susanne Strahringer.

Seit über 50 Jahren erscheint die Fachzeitschrift „HMD – Praxis der Wirtschaftsinformatik“ mit Schwerpunktausgaben zu aktuellen Themen. Erhältlich sind diese Publikationen im elektronischen Einzelbezug über SpringerLink und Springer Professional sowie in gedruckter Form im Abonnement. Die Reihe „Edition HMD“ greift ausgewählte Themen auf, bündelt passende Fachbeiträge aus den HMD-Schwerpunktausgaben und macht sie allen interessierten Lesern über online- und offline-Vertriebskanäle zugänglich. Jede Ausgabe eröffnet mit einem Geleitwort der Herausgeber, die eine Orientierung im Themenfeld geben und den Bogen über alle Beiträge spannen. Die ausgewählten Beiträge aus den HMD-Schwerpunktausgaben werden nach thematischen Gesichtspunkten neu zusammengestellt. Sie werden von den Autoren im Vorfeld überarbeitet, aktualisiert und bei Bedarf inhaltlich ergänzt, um den Anforderungen der rasanten fachlichen und technischen Entwicklung der Branche Rechnung zu tragen.

Weitere Bände in dieser Reihe http://www.springer.com/series/13850

Sara D'Onofrio • Andreas Meier
Hrsg.

Big Data Analytics

Grundlagen, Fallbeispiele und Nutzungspotenziale

Springer Vieweg

Hrsg.
Sara D'Onofrio
IT Business Integration
Genossenschaft Migros Zürich
Zürich, Schweiz

Andreas Meier
Universität Fribourg
Fribourg, Schweiz

Das Herausgeberwerk basiert auf vollständig neuen Kapiteln und auf Beiträgen der Zeitschrift HMD – Praxis der Wirtschaftsinformatik, die entweder unverändert übernommen oder durch die Beitragsautoren überarbeitet wurden.

ISSN 2366-1127 ISSN 2366-1135 (electronic)
Edition HMD
ISBN 978-3-658-32235-9 ISBN 978-3-658-32236-6 (eBook)
https://doi.org/10.1007/978-3-658-32236-6

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

Springer Vieweg
© 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 des Verlags. 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: 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

Vorwort

Von einfachen Geldautomaten über sensorausgestattete Strassen zu Voice Assistants – in unserer Gesellschaft sind Technologien allgegenwärtig und haben sich in unserer Lebens- und Arbeitsweise verankert. Täglich werden durch die Nutzung digitaler Angebote (Soziale Medien, digitale Services) Unmengen an Daten (Big Data) generiert; ihr Wachstum ist exponentiell. Big Data ist kein Buzzword mehr, denn die effiziente und effektive Handhabung solcher Daten gewinnen zunehmend an Bedeutung. Daher beschäftigen sich immer mehr Unternehmen, Behörden und andere Organisationen mit der Frage, wie die innerhalb und ausserhalb der Organisation verfügbaren Daten gesammelt, zielgerichtet analysiert und genutzt werden können.

Big Data Analytics umfasst Methoden der Analyse, des Reportings und der Visualisierung von großen Datenmengen mit dem Ziel, für die Organisation relevante Informationen (z. B. über Kundenpräferenzen und -verhalten) zu extrahieren und auf eine verständliche Weise zu visualisieren. Mit dem heutigen Fortschritt der Technologie sowie den sinkenden Kosten für deren Einsatz können Daten in Echtzeit für potenzielle Szenarien ausgewertet werden, wodurch Organisationen in der Lage sind, sich nicht nur reaktiv, sondern proaktiv zu verhalten. Die Fähigkeit, sich schneller den Umweltfaktoren anzupassen, verschafft den Organisationen einen Wettbewerbsvorteil. Im Vordergrund steht das Verständnis von Daten und deren Beziehungen untereinander, um Datenanalysen automatisiert durchzuführen. Unterschiedliche Methoden unterstützen diesen Prozess, meistens auf Empfehlung von ausgewiesenen Data Scientists.

Die Edition HMD über Big Data Analytics gibt einen Einblick über die Vielfalt der Methoden und zeigt anhand konkreter Praxisfälle auf, wie diese genutzt werden. Das Herausgeberwerk besteht aus fünf Teilen: Grundlagen (Teil I), Textanalyse (II), Machine Learning (III), Prädiktive Modelle (IV) und Trendforschung (V).

Teil I des Herausgeberwerks widmet sich den Hard- und Soft-Methoden von Big Data Analytics. Ergänzt werden diese Grundlagen mit einer aktuellen Marktstudie über Digital Analytics, welche Aufschluss über dessen Reifegrad und künftige Entwicklungen gibt.

Die Teile II bis V geben aufschlussreiche Fallstudien aus unterschiedlichen Anwendungsgebieten. Dabei werden folgende Themenbereiche adressiert:

- Linguistische Analyse für Compliance
- Textanalyse als Entscheidungsunterstützung im Online-Handel
- Einsatzoptionen von Machine Learning im Handel
- Automatisierte Qualitätssicherung in der Produktion via Image Mining und Computer Vision
- Deep Learning zur Unterstützung von Winzertätigkeiten
- Data Pipelines als Instrument für die politökonomische Forschung
- Plattformen für Self-Service Data Mining
- Datenanalyse zur Vorhersage des Einflusses von Covid-19 auf Wertschöpfungsketten
- Intelligente Bots für die Trendforschung

An dieser Stelle richte ich meinen Dank an die Autorinnen und Autoren, die ihr Expertenwissen, ihre Erfahrungen und wertvollen Erkenntnisse aus Forschung und Praxis in interessanten Kapiteln diskutieren. Ein weiterer Dank geht an die Gutachterinnen und Gutachter für ihre kritischen und konstruktiven Feedbacks, die zur Verbesserung der Qualität und Kohärenz der Inhalte geführt haben. Zudem möchte ich mich beim HMD-Herausgeberteam vom Springer-Verlag für die Unterstützung bedanken. Ein besonderer Dank geht an Andreas Meier, emeritierter Professor der Universität Fribourg (Schweiz), der als Gastherausgeber mitgeholfen hat, diese Edition über Big Data Analytics zu gestalten.

Liebe Leserinnen und Leser, ich wünsche Ihnen eine interessante Lektüre und neue Erkenntnisse. Tauchen Sie in die Welt von Big Data Analytics ein und lassen Sie sich von den Fallstudien inspirieren.

Herzliche Grüsse aus der Schweiz,
Sara D'Onofrio

Zürich, Schweiz Sara D'Onofrio

Einwurf

Es ist immer Zeit, an die Zukunft zu denken[1]

Martin Lauber, Die Schweizerische Post, Bern, Schweiz; martin.lauber@post.ch

Machen Sie sich Gedanken über die Zukunft? Planen Sie grob oder wie manch einer gar akribisch wie Ihr nächster Tag aussehen soll, wo Sie sich in fünf oder zehn Jahren sehen? Selbst wenn Sie für sich beanspruchen eine Person zu sein, die soweit möglich im Jetzt lebt, so werden Sie Ihren Alltag dennoch nur dank antizipativem Handeln meistern können. Sei es, wenn Sie sich auf Ihr Urteil verlassen, an welcher der offenen Kassen im Supermarkt Sie zuerst bedient werden oder wenn Sie als Fussgänger versuchen abzuschätzen, ob das herannahende Fahrzeug rechtzeitig halten wird, um ihnen den Vortritt zum Überqueren der Strasse zu gewähren. Man braucht nicht viel über Evolutionstheorie zu wissen, um zu erkennen, dass vorausschauendes Handeln von der natürlichen Selektion seit jeher gefördert worden ist. Stellen Sie sich zwei Menschen vor, welche vor, sagen wir, 2.3 Millionen Jahren, während der frühesten Epoche der Menschheitsgeschichte gelebt haben. Einer von ihnen ist stets bemüht lauernde Gefahren zu entdecken, während der andere selbst an unübersichtlichen Stellen geradeausstampft. Was denken Sie, welcher der beiden mit grösserer Wahrscheinlichkeit Ihr Vorfahre ist und wessen Gene nicht mehr die Chance erhalten haben zwei zu eins weitergegeben zu werden?

Als Data Scientist ist es ein wesentlicher Bestandteil meines Jobs aufgrund bestehender und neuer Daten die unmittelbare Zukunft abzuschätzen. Mithilfe computergestützter statistischer Methoden und Modellen des maschinellen Lernens („machine learning“) versuche ich aus Daten Informationen zu extrahieren und daraus für den gegebenen Anwendungsfall Muster zu erkennen oder automatisierte Vorhersagen zu ermöglichen. „*Ein Beruf der Zukunft hat“* – höre ich bisweilen ziemlich oft und tatsächlich ist die Dringlichkeit sich zu einer Wissensorganisation zu entwickeln oder zumindest die Möglichkeiten des Datenzeitalters nicht zu verpassen von den meisten Unternehmen erkannt worden. Was für ein zukunftsträchtiges Metier ich mir doch ausgesucht habe! Zum Glück nicht primär aufgrund der Spekulation, dass der Arbeitsmarkt diese Entscheidung belohnen wird, denn von all den Jobs, die sich grösstenteils automatisieren lassen, steht dieser ganz oben auf der Liste. Ob

[1] Dieser Einwurf beruht auf einer Aktualisierung des Beitrags von Lauber M (2019) Es ist immer Zeit, an die Zukunft zu denken. HMD – Praxis der Wirtschaftsinformatik, Heft 329, 56(5): 881–884.

und wie dieser Beruf von entsprechenden Studien eingeschätzt wird, ist mir nicht bekannt. Aber lassen Sie mich meine Einschätzung mit Ihnen teilen.

Data Scientists wollen möglichst schnell erste Resultate vor sich haben und darauf aufbauend ihr Modell dann in weiteren Iterationen verfeinern. Bei den iterativen Anpassungen und Versuchen wollen sie dabei so wenig wie möglich an ihren Skripten überarbeiten müssen. Der Code soll schlank und effizient sein. Zu diesem Zweck bauen sie sich eine Pipeline auf. Sie ermöglicht es, für die einzelnen Funktionen des Programms einen Ablauf zu organisieren und mit einem einzigen Abruf auszulösen. Im Idealfall kann sich ein Data Scientist aus vorbereiteten Libraries bedienen, wenn er ein anderes Modell ausprobieren möchte und braucht bei entsprechender Vorarbeit bloss noch den zuvor gewählten Parameter in seiner Pipeline-Funktion mit dem neuen zu ersetzen. Ein Doppelklick und die Eingabe einer einzigen Zeichenkette später kann er sogleich die neu generierte *Confusion-Matrix*[2] und die visualisierte Performance des Modells begutachten. Wenn das Skript nun auch die Evaluation für den Data Scientist übernehmen könnte, indem dieser definiert aufgrund welcher Werte er sich für oder gegen ein Modell entscheidet, bräuchte er nur noch die in Frage kommenden Modelle gleich allesamt in die Pipeline zu verbauen und das Skript gibt – aufgrund von automatisiertem *Grid Search* und *Cross Validation*[3] – ohne Zutun das optimale Modell aus. Vielleicht könnte es auch einfach direkt den erwünschten Output zurückgeben. Aber dafür bräuchte es wohl noch etwas mehr Rechenpower. Tatsächlich steht diese bereits zur Verfügung und nicht ganz unbekannte IT-Unternehmen und Start-ups stehen mit Rundumlösungen in den Startlöchern beziehungsweise ihre Vertreter in den Eingangsbereichen der Firmen mit Analytics-Abteilungen.

Meine Arbeit – zumindest alles, was mit Modeling zu tun hat – kann im Prinzip bereits jetzt weitgehend automatisiert werden. Also, wozu braucht es mich, den Data Scientist, dann noch? Vielleicht, um neue Anwendungsfälle zu finden, die Technologie zu erklären oder weiterzuentwickeln, Blackboxes transparenter zu machen oder schlicht Daten aufzubereiten. Sie sehen, das Automatisieren einer Aufgabe ist nicht gleichzustellen mit der Rationalisierung von *Full Time Equivalents.*[4] Denn wer bringt bessere Voraussetzungen mit, sich den genannten, verbleibenden und sich neu entwickelnden Aufgaben in diesem Kontext anzunehmen?

Aber zurück zu dem Entscheid Data Scientist zu werden. Nein, nicht die Hoffnung auf eine langwährende Berufsbezeichnung hat mich dazu bewogen und auch nicht ausschliesslich die Faszination für kurzfristige datenbasierte Vorhersagen, die

[2] Die Confusion-Matrix wird bei der Evaluation von Modellen mit überwachtem Lernen („supervised learning“) eingesetzt und gibt an, wie viele Zielwerte das Modell richtigerweise bzw. fälschlicherweise als positiv oder negativ klassifiziert hat.

[3] Grid Search und Cross Validation werden verwendet, um für eine konkrete Anwendung eines Modells des maschinellen Lernens die Parameter zu optimieren. Das Modell wird immer wieder von neuem trainiert und dabei werden immer neue Parameterwerte ausprobiert. Am Ende werden die Resultate verglichen und die besten Einstellungen ausgewählt.

[4] Full Time Equivalent (FTE), zu Deutsch Vollzeitäquivalent, ist eine Messgrösse zur Bestimmung der in Vollzeitstellen ausgedrückten Anzahl Mitarbeiter. Damit kann die Anzahl Mitarbeiter unabhängig von Teilzeitpensen angegeben werden.

viele der Modelle zum Ziel haben und einen erahnen lassen zu welchen Teilen die Welt deterministisch funktioniert und zu welchen Teilen sie Zufällen unterliegt. Vielmehr möchte ich meinen Beitrag zum technologischen Fortschritt leisten. Denn was mich wirklich animiert, ist der Blick in eine etwas fernere Zukunft. Eine Zukunft, die sich massgebend von der Gegenwart unterscheidet und das – so zumindest meine Auffassung – massgeblich aufgrund der Cutting-Edge Technologien, die aus dem Analytics Umfeld erwachsen und in immer besser werdenden Anwendungen künstlicher Intelligenz münden. Niemand kann die Zukunft umfassend vorhersagen und doch gleicht die zukünftige Realität vergangenen Vorstellungen und wird gleichsam von ihnen geprägt. Haben Sie sich auch schon gefragt, ob Science-Fiction Ideen von den anbahnenden Trends in der Zeit, in der sie entstehen, definiert werden oder ob umgekehrt, der aktuelle Zustand der Welt von ebenjenen populären Ideen von der Zukunft, mitgestaltet wurde?

So bewegen mich jene Gedanken, die sich darum drehen, was sein wird. Derzeitige Trends schlicht zu extrapolieren ist zu diesem Zweck unzureichend. Zwar bilden die vorhandenen Faktoren und deren Dynamiken einen wichtigen Bestandteil für Prognosen, dennoch sind es die hinzukommenden disruptiven Entwicklungen oder auch unvorhergesehenen Gegenbewegungen, die die Zukunft mindestens ebenso umfangreich beeinflussen.

Ein Beispiel: Wenn die Anzahl Menschen im Jahre 2050 prognostiziert werden soll, mag es verlockend sein, sich die aktuellen Entwicklungen anzuschauen und auf das Jahr 2050 hochzurechnen. Aber wo bleiben in dieser Prognose die Jahrhundertereignisse oder die nicht zuvor dagewesenen Entwicklungen in der Medizin, die die Lebenserwartung drastisch verändern könnten? Was für einen Einfluss wird im Gegenzug die Verbreitung eines breitabgestützten Mittelstands auf der Welt und die damit einhergehenden Veränderungen des durchschnittlichen Bildungsstands und Familienplanung in weiten Teilen auf die Bevölkerungsentwicklung haben? Aktuelle Trends dürften noch nicht geahnte Gegenbewegungen auslösen. Es ist also durchaus Fantasie gefragt und es kann trotz aller Unsicherheiten schon mal vorkommen, dass einem eine Vorstellung derart plausibel erscheint, dass es einem schwerfällt, sie nicht als noch nicht geschehene Wahrheit, sondern als eine Idee davon einzuordnen. Der Zukunftsforscher und KI-Experte Nick Bostrom schildert in seinem Paper „*Ethical Issues in Advanced Artificial Intelligence*“,[5] derart plausibel, was ein superintelligentes System mit sich bringen würde und was bei der Kreation dessen im Sinne der Menschheit beachtet werden sollte, dass Vorstellungen einer ferneren Zukunft, die nicht zum Grossteil von dieser einen Erfindung bestimmt werden, in meiner Weltsicht kaum noch Platz finden. Darin beschreibt er die Superintelligenz als einen Intellekt, der jenen der Menschen in jeder Hinsicht übersteigt. Auch soziale Fähigkeiten, Kreativität und Weisheit sind damit gemeint. Künstliche Intelligenz vermag bereits heute menschliche Topleistungen zu übertreffen, allerdings immer nur in sehr spezifischen Aufgaben. Bostrom schreibt, die Superintelligenz dürfte die letzte Erfindung des Menschen sein. Dies, weil danach die

[5] Bostrom N (2003) Ethical issues in advanced artificial intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence, 277–284.

wissenschaftlichen Leistungen der Menschen nicht mehr gefragt wären und auch nicht mehr mithalten könnten. Der technologische Fortschritt würde noch einmal stark beschleunigt werden und ein solches System würde sich selbstständig rasant weiterentwickeln. Für Bostrom ist zentral, dass die Ziele, die ein solch mächtiges System verfolgen soll, mit grosser Sorgfalt gewählt werden müssen. Einfach gefragt: Was wünscht sich die Menschheit? Und bedenken Sie, Wünsche die unveränderbar erfüllt werden, sind nicht ganz ohne. Wenn alles zu Gold wird, was man berührt, endet es bekanntlich nicht wie erhofft. Ich denke nicht, dass die Lösung darin liegt, dass sich einige schlaue Köpfe zusammensetzen, um geeignete Ziele zu definieren. Das Formulieren optimaler Direktiven kann nicht dem – im Vergleich zu einem superintelligenten System – inferioren menschlichen Geist entspringen. Vielmehr wird das System selbst, aufgrund vorhandener Daten, Muster erkennen und Erkenntnisse gewinnen, dazu, was den Menschen insgesamt aber auch ganz individuell ausmacht und für ihn von Bedeutung ist. Vorzugeben brauchen wir lediglich, dass es diese Erkenntnisse berücksichtigen soll. In diesem vertieften Verständnis, welches das System von uns allen haben wird, liegt, meiner Ansicht nach, der Kern aller Hoffnung. Nur mit dieser Vorbedingung ist für mich eine Welt denkbar, in der ein superintelligentes System mit den Menschen koexistieren kann und dessen unumkehrbare Inbetriebnahme nicht zu Reue, sondern Dankbarkeit führen wird. Ob nun geschaffen nach dem Ebenbild Gottes oder nicht, wird der Mensch selbst etwas Übermenschliches geschaffen haben und nach, für einige Zeit währender Koexistenz, womöglich gar damit verschmelzen. Dieses System kollektiver Intelligenz, Materie und Energie ist das, was ich mir unter der Singularität vorstelle. Hier wird alles eines und eines alles werden.

Wenn immer ich mich also daran störe, dass ein belangloser oder für mein Empfinden nicht adäquater und/oder intimer Moment auf einer digitalen Plattform geteilt wird, bin ich zugleich beruhigt. Das superintelligente System wird uns bis ins Detail kennen, die Menschheit hat also Zukunft.

Inhaltsverzeichnis

Über die Autoren

Sara D'Onofrio ist IT Business Partner Manager eines der größten Detailhandelsunternehmen der Schweiz, Autorin und Herausgeberin der Zeitschrift HMD - Praxis der Wirtschaftsinformatik bei Springer, Gastdozentin an Hochschulen und Mitglied der Stiftung FMsquare, welche die Anwendung von Fuzzy-Logik zur Lösung von wirtschaftlichen und sozialen Problemen fördert. Sie hat Betriebswirtschaft und Wirtschaftsinformatik studiert und in Informatik promoviert.

Daniel Badura ist als Consultant bei der valantic Business Analytics GmbH tätig und veröffentlicht regelmäßig wissenschaftliche Artikel im Bereich Self-Service Data Science.

Freimut Bodendorf ist seit 1989 Leiter des Lehrstuhls für Wirtschaftsinformatik II und Leiter des Instituts für Wirtschaftsinformatik. Nach Abschluss seines Studiums der Informatik im Jahr 1977 promovierte er auf dem Gebiet der Wirtschaftsinformatik und arbeitete an mehreren Universitäten in Deutschland und der Schweiz. Vor seiner Professur in Nürnberg war er Lehrstuhlinhaber am Institut für Informatik an der Universität Freiburg (Schweiz). Er ist Mitglied zahlreicher internationaler Forschungsorganisationen. Hauptforschungsgebiete sind Service- und Prozessmanagement.

Peter Chamoni war seit 1995 Inhaber des Lehrstuhls für Wirtschaftsinformatik, insb. 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. Seit dem Wintersemester 2019/2020 ist er im Ruhestand.

Lukas Breithaupt hat sein Studium der Diplom-Wirtschaftsinformatik mit den Schwerpunkten Business Intelligence, Data Science und Datenbanksysteme im Jahr 2020 an der TU Dresden abgeschlossen. Seine Forschungsarbeiten befassen sich insbesondere mit dem Einsatz von Deep-Learning-basierten Entscheidungsunterstützungssystemen. Aktuell befasst er sich bei der Aareon Deutschland GmbH mit dem Aufbau eines unternehmensweiten Data Warehouses sowie mit dem Einsatz und der Entwicklung von Entscheidungsunterstützungssystemen auf Basis von Machine Learning für die Wohnungswirtschaft.

Carolin Durst lehrt an der Hochschule Ansbach in den Gebieten Digital Business, Digital Marketing und Digital Transformation. In ihrer Forschung beschäftigt sie sich mit innovativen Technologien und deren Einsatzmöglichkeiten in der Praxis. Als Scientific Director der ITONICS GmbH begleitet Carolin Durst die Methoden- und Produktentwicklung der Software Suite für strategisches Innovationsmanagement.

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.

Henry Goecke geboren 1982 in Dortmund. Studium der Volkswirtschaftslehre an der Technischen Universität (TU) Dortmund und der Strathclyde University Glasgow sowie Promotion an der TU Dortmund. Seit 2012 im Institut der deutschen Wirtschaft, seit 2017 Leiter der Forschungsgruppe Big Data Analytics. In seinen Forschungsarbeiten befasst er sich mit Methoden zur Sammlung und Analyse großer, unstrukturierter Datensätze sowie inhaltlich vor allem mit den Themen der Datenökonomie und der Künstlichen Intelligenz.

René Götz studierte von 2011 bis 2017 Wirtschaftsinformatik an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Seit 2017 ist er wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschaftsinformatik in Dienstleistungsbereich der FAU, welcher von Prof. Dr. Freimut Bodendorf geleitet wird. Sein Forschungsschwerpunkt ist das Thema Produktempfehlungen im Online-Handel mit Bezug auf die Produktwahrnehmung aus Kundensicht.

Michael Grottke ist Principal Data Scientist bei der GfK SE und Apl. Professor an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Er studierte Betriebswirtschaftslehre an der FAU und Economics an der Wayne State University in Detroit, USA. Nach seiner Promotion am Lehrstuhl für Statistik und Ökonometrie der FAU verbrachte er drei Jahre als Research Associate und Assistant Research Professor an der Duke University in Durham, USA. Seine Forschungsarbeiten zu

Themen der stochastischen Modellierung, der statistischen Datenanalyse und des maschinellen Lernens wurden u. a. von dem Bundesministerium für Bildung und Forschung, der Europäischen Kommission sowie dem Office of Safety and Mission Assurance der NASA gefördert.

Kai Heinrich ist Dozent, Forscher und Post-Doc an der TU Dresden. Er lehrt und forscht am Lehrstuhl für Wirtschaftsinformatik, insbesondere Intelligente Systeme und Dienste. Seine Forschung dreht sich um KI-basierte Entscheidungsunterstützungssysteme. Dabei liegen die Schwerpunkte auf dem Design von KI-basierten Systemen sowie der Interaktion dieser Systeme mit dem menschlichen Umfeld. Weiterhin lehrt er in den Themenfeldern allgemeine Wirtschaftsinformatik, intelligente Systeme sowie im Gebiet Data Science.

Urs Hengartner, geboren 1955, ist Dozent am Digital Humanities Lab und der Wirtschaftswissenschaftlichen Fakultät der Uni Basel im Bereich Information Retrieval und Software Engineering. Nach der Matura am Realgymnasium in Basel im Jahre 1976 arbeitete er in einer namhaften Schweizer Versicherung als Analytiker- und System-Programmierer. Nach Studien an der ETH Zürich und Universität Zürich erlangte er 1990 das Diplom in Wirtschaftsinformatik an der Rechts- und Staatswissenschaftlichen Fakultät der Universität Zürich. Er promovierte im Jahr 1996 mit der Dissertation „Entwurf eines integrierten Informations-, Verwaltungs- und Retrieval Systems für textuelle Daten“. Als Mitgründer der Canoo Engineering AG in Basel war er über 18 Jahre als Consultant und Projektleiter in umfangreichen Projekten tätig.

Mohamed Kari ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschaftsinformatik und integrierte Informationssysteme von Professor Reinhard Schütte an der Universität Duisburg-Essen. Er forscht an der Schnittstelle von Machine Learning und Mixed Reality.

Michael Klaas ist Leiter der Fachstelle für digitales Marketing und Senior-Dozent an der Zürcher Hochschule für Angewandte Wissenschaften in den Themenfeldern Marketing, digitales Marketing und Service Design. Neben seiner Forschungstätigkeit gemeinsam mit Unternehmenspartnern leitet er verschiedene Weiterbildungsprodukte, u. a. im Bereich Digital Marketing, Marketing Analytics, KI und Industrie 4.0. An der Universität St. Gallen ist er als Dozent im Bereich Design Thinking tätig.

Laura Kölbl studierte Statistik an der Ludwig-Maximilians-Universität München und arbeitet derzeit am Lehrstuhl für Statistik und Ökonometrie an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Im Rahmen eines vom Bundesministerium für Bildung und Forschung geförderten Forschungsprojektes konzentriert sich ihre Forschung auf die natürliche Sprachverarbeitung und maschinelle Lernverfahren mit besonderem Interesse für mögliche Anwendungen bei der automatischen Erkennung von Trends.

Johannes Maresch studierte an der TU Dresden, wo er im Jahr 2019 sein Diplom im Bereich Wirtschaftsinformatik absolvierte. Neben seinen Schwerpunkten Business Intelligence und Systemarchitektur fokussierte er sich außerdem auf die Konzeption von Anwendungen zur Datenanalyse in verteilten Systemen. Aktuell arbeitet er als Data Engineer bei der LOVOO GmbH. Hier konzipiert und entwickelt er Machine-Learning-basierte Systemkomponenten, welche zur Erkennung und Vorbeugung von Spam und Ad-Fraud eingesetzt werden.

Ulrich Matter ist Assistenzprofessor für Volkswirtschaftslehre an der Universität St. Gallen, wo er in den Bereichen Big Data Analytics, Data Handling und Web Mining unterrichtet. Er studierte Wirtschaftswissenschaften an der Fachhochschule Nordwestschweiz und an der Universität Basel und promovierte an der Universität Basel zu politischer Ökonomie. 2016–2017 war er Gastforscher am Berkman Klein Center for Internet & Society, an der Harvard University. Seine Forschungsinteressen liegen in den Bereichen quantitative politische Ökonomie, Medienökonomik und Data Science.

Andreas Meier hat Musik an der Musikakademie in Wien und Mathematik an der Eidgenössisch-Technischen Hochschule (ETH) in Zürich studiert, wo er doktorierte und habilitierte. Er arbeitete u. a. bei IBM Oesterreich und IBM Schweiz in diversen Positionen, gehörte zum Direktionskader der internationalen Bank SBV und trug Mitverantwortung in der Geschäftsleitung des Versicherers CSS. In der Forschung war er am IBM Research Lab in Kalifornien tätig und gründete das International Research Center Fuzzy Management Methods an der Universität Fribourg in der Schweiz.

Björn Möller hat sein Diplomstudium der Wirtschaftsinformatik mit den Schwerpunkten Business Intelligence und Systemarchitektur an der TU Dresden abgeschlossen. Dabei hat er an Projekten aus den Bereichen Data Science und Machine Learning mitgewirkt und insbesondere Modelle zur Verarbeitung räumlich strukturierter Daten untersucht. Aktuelle Forschungsinteressen sind Self-Supervised Learning und die Erklärbarkeit von Machine-Learning-Modellen.

Christian Mühlroth studierte Betriebswirtschaftslehre und internationale Wirtschaftsinformatik an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) und forscht derzeit am Lehrstuhl für Statistik und Ökonometrie an der FAU. Im Rahmen eines vom Bundesministerium für Bildung und Forschung geförderten Forschungsprojektes konzentriert sich seine Forschung auf die Anwendung der künstlichen Intelligenz in der strategischen Vorausschau und im Innovationsmanagement, um Unternehmen dabei zu unterstützen, zukünftige Chancen und Risiken datengetrieben und frühzeitig zu erkennen. Als CCO der ITONICS GmbH begleitet er globale Innovationsführer dabei, ganzheitliche Innovationssysteme zu implementieren und digitale, KI-gestützte Innovationsplattformen nachhaltig zu etablieren.

Alexander Ossa verfolgt als Softwareentwickler bei der Gruner + Jahr GmbH neueste Trends im Data-Science-Bereich und bastelt in seiner Freizeit gerne am Smart Home.

Alexander Piazza studierte Computational Engineering und Wirtschaftsinformatik an der FAU Erlangen-Nürnberg und promovierte im Anschluss am Lehrstuhl für Wirtschaftsinformatik, insbes. im Dienstleistungsbereich über Produktempfehlungssysteme in der Modebranche. Sein Forschungsinteresse liegt speziell in der Analyse von unstrukturierten Daten und deren Nutzung für die personalisierte Kundenansprache.

Andreas Roth hat im November 2018 sein Diplomstudium an der TU Dresden in der Fachrichtung Wirtschaftsinformatik mit Auszeichnung abgeschlossen. Heute ist er Lead Developer bei esveo, wo er die Konzeption und Umsetzung verschiedener Unternehmensanwendungen auf Basis modernster Webtechnologien überwacht und durchführt. Zudem vermittelt er sein Wissen auf diesem Bereich mit der esveo Academy in Workshops, Konferenzvorträgen und Schulungen an andere Entwickler.

Christian Schieder ist Professor für Wirtschaftsinformatik an der Weiden Business School der Ostbayerischen Technischen Hochschule (OTH) Amberg-Weiden. Seine Forschungsschwerpunkte liegen in der Konzeption und Anwendung analytischer Informationssysteme zur Umsetzung datenbasierter Geschäftsmodelle. Als unabhängiger Berater unterstützt der Diplom-Wirtschaftsinformatiker Unternehmen im Umfeld Digital Business beim Aufbau datengetriebener Entscheidungskulturen. Zuvor war er als Chief Digital Officer beim bayerischen Maschinen- und Anlagenbauer BHS Corrugated für digitale Transformation und Business Development im Bereich industrieller digitaler Lösungen (IoT-, Edge- und Cloud-Services) verantwortlich.

Reinhard Schütte hat den Lehrstuhl für Wirtschaftsinformatik und integrierte Informationssysteme an der Universität Duisburg-Essen inne. Seine Forschungsinteressen sind prozessorientiert und reichen von der Wirkung von Systemen und deren Akzeptanz über das Management von Anwendungssystemen bis hin zur ganzheitlichen Transformation von Unternehmen im Zuge der Digitalisierung. Alle Forschungsbereiche konzentrieren sich auf den Bereich Handel. Neben seiner akademischen Laufbahn war Herr Schütte Mitglied des Vorstands und des Aufsichtsrats der größten deutschen Handelsunternehmen und verantwortlich für eines der bedeutendsten Transformationsprojekte im Handel, eine der größten SAP-Implementierungen weltweit. Derzeit ist er Mitglied des wissenschaftlichen Beirats von Deutschlands zweitgrößtem Softwarekonzern, der Software AG.

Michael Schulz hält eine Professur für Wirtschaftsinformatik, insb. analytische Informationssysteme, an der NORDAKADEMIE – Hochschule der Wirtschaft in Elmshorn. Zudem ist er als Projektmanager bei der valantic Business Analytics

GmbH tätig. Seine Interessenschwerpunkte in Lehre, Forschung und Praxisprojekten liegen in der Business Intelligence und der Data Science.

Sebastian Trinks ist wissenschaftlicher Mitarbeiter und Doktorand am Institut für Wirtschaftsinformatik an der TU Bergakademie Freiberg. Seine Forschungsinteressen sowie der Schwerpunkt seiner Dissertation liegen im Spannungsfeld der Industrie 4.0 sowie der Smart Factory. Herr Trinks forscht in diesem Kontext zu Themen aus den Bereichen Real Time Analytics, Edge Computing sowie Image Processing und Image Mining.

Felix Weber ist Forscher an der Universität Duisburg-Essen und Leiter des Retail Artificial Intelligence Lab am am Lehrstuhl für Wirtschaftsinformatik und integrierte Informationssysteme von Professor Reinhard Schütte mit den Schwerpunkten Digitalisierung, künstliche Intelligenz, Preis-, Promotions- und Sortimentsmanagement sowie Transformationsmanagement. Gleichzeitig ist er Senior Consultant für SAP Systeme im Handel.

Jan Wendt, M.Sc., geboren 1995 in Troisdorf. Studium der Wirtschaftsinformatik an der FH Münster. Seit 2019 im Institut der deutschen Wirtschaft als Data Scientist in der Forschungsgruppe Big Data Analytics. Seine Schwerpunkte liegen in der Forschung und Anwendung im Bereich der Generierung, Analyse (EDA), Auswahl, Bereinigung, Aufbereitung, Konstruktion und Formatierung von Daten sowie der Modellerstellung, -evaluation und -bereitstellung unter Verwendung von Machine Learning und insbesondere Deep Learning.

Fabian Wiser war wissenschaftlicher Mitarbeiter am Lehrstuhl für Informationssysteme der TU Braunschweig. Sein Forschungsschwerpunkt ist die Analyse soziotechnischer Systeme im Innovationsmanagement. Derzeit arbeitet er als IT-Berater bei einem internationalen IT-Dienstleister.

Andrea Zelic absolvierte den Master in Business Administration mit Vertiefung Marketing an der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW). Im Rahmen ihrer Master Thesis beschäftigte sie sich mit der Entwicklung und den Trends im Bereich Digital Analytics. Seit 2020 ist sie als Head of Marketing & Design bei der Unternehmung Manthano GmbH im Bereich AI tätig. Zuvor arbeitete sie als Social Media & Marketing Strategin bei Daneco AG in Fehraltorf.

Patrick Zschech ist Juniorprofessur für Intelligent Information Systems an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Zuvor arbeitete er als wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschaftsinformatik, insbesondere Intelligente Systeme und Dienste, an der TU Dresden. Zudem war er als Projektmitarbeiter und Dozent für die Robotron Datenbank-Software GmbH tätig. Er beschäftigt sich in seiner Forschung mit der Anwendung datengetriebener Verfahren zur Entwicklung analytischer Informationssysteme. Seine Hauptinteressen liegen in den Bereichen Machine Learning, Computer Vision, Process Mining, Industrie 4.0 und Data-Science-Befähigung.

Darius Zumstein ist seit Oktober 2018 Dozent und Senior Researcher am Institut für Marketing Management IMM der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW). Er doziert und forscht zu Digital Commerce, Digital Marketing und Digital Analytics. Zuvor arbeitete er fünf Jahre an der Hochschule Luzern und bei der Raiffeisen Schweiz. Von 2013 bis 2016 leitete er das Team Digital Analytics & Data Management bei der Sanitas Krankenversicherung. Davor beriet er Unternehmen wie BMW, Scout24 und Kabel Deutschland. Bis 2011 war er Assistent der Information Systems Research Group der Universität Fribourg, wo er bei Prof. Dr. Andreas Meier zu Web Analytics promovierte.

Teil I

Grundlagen

Rundgang Big Data Analytics – Hard & Soft Data Mining

1

Andreas Meier

Zusammenfassung

Das Einführungskapitel definiert und charakterisiert verschiedene Facetten des Big Data Analytics und zeigt auf, welche Nutzenpotenziale sich für Wirtschaft, öffentliche Verwaltung und Gesellschaft ergeben. Nach der Klärung wichtiger Begriffe wird der Prozess zum Schürfen nach wertvollen Informationen und Mustern in den Datenbeständen erläutert. Danach werden Methodenansätze des Hard Computing basierend auf klassischer Logik mit den beiden Wahrheitswerten wahr und falsch sowie des Soft Computing mit unendlich vielen Wahrheitswerten der unscharfen Logik vorgestellt. Anhand der digitalen Wertschöpfungskette elektronischer Geschäfte werden Anwendungsoptionen für Hard wie Soft Data Mining diskutiert und entsprechende Nutzenpotenziale fürs Big Data Analytics herausgearbeitet. Der Ausblick fordert auf, einen Paradigmenwechsel zu vollziehen und sowohl Methoden des Hard Data Mining wie des Soft Data Mining für Big Data Analytics gleichermaßen zu prüfen und bei Erfolg umzusetzen.

Schlüsselwörter

Big Data Analytics · Data Science · Fuzzy Logic · Hard Data Mining · Knowledge Discovery in Databases · Paradigmenwechsel · Soft Data Mining

Dieses Kapitel beruht auf einer Erweiterung und Aktualisierung des Beitrags von Meier A. (2019) Überblick Analytics: Methoden und Potenziale. HMD – Praxis der Wirtschaftsinformatik, Heft 329, 56(5): 885–899.

A. Meier (✉)
Universität Fribourg, Fribourg, Schweiz
E-Mail: andreas.meier@unifr.ch

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

S. D'Onofrio, A. Meier (Hrsg.), *Big Data Analytics*, Edition HMD,
https://doi.org/10.1007/978-3-658-32236-6_1

1.1 Motivation und Begriffseinordnung

Wissenschaft, Wirtschaft, öffentliche Verwaltung und Gesellschaft befinden sich in einer Umbruchphase, die als digitaler Transformationsprozess bezeichnet wird. Dabei wird das wirtschaftliche, öffentliche wie private Leben von Informations- und Kommunikationstechnologien getrieben. Zu jeder Zeit und an jedem Ort entstehen Datenspuren: Postings aus sozialen Medien, elektronische Briefe, Anfrageverhalten in Suchmaschinen, Bewertungen von Produkten und Dienstleistungen, Geo-Daten, Messdaten des Haushalts (Smart Meter), Aufzeichnungen von Monitoring-Systemen, Daten aus eHealth-Anwendungen, Prozessdaten aus der Produktion, Kennzahlen von Webplattformen, um nur einige Beispiele zu nennen.

Der Wandel von der Industrie- zur Informations- und Wissensgesellschaft spiegelt sich in der Bewertung der Information als Produktionsfaktor wider. Information hat im Gegensatz zu materiellen Wirtschaftsgütern folgende Eigenschaften:

- *Darstellung*: Information wird durch Zeichen, Signale, Nachrichten oder Sprachelemente spezifiziert.
- *Verarbeitung*: Information kann mit Hilfe von Algorithmen (Berechnungsvorschriften) übermittelt, gespeichert, klassifiziert, aufgefunden und in andere Darstellungsformen transformiert werden.
- *Quelle*: Die Herkunft einzelner Informationskomponenten ist kaum nachweisbar. Manipulationen sind jederzeit möglich. Information ist beliebig kopierbar und kennt per se keine Originale.[1]
- *Kombination*: Information ist beliebig kombinierbar.
- *Alter*: Information unterliegt keinem physikalischen Alterungsprozess. Hingegen spielt die Zeitachse bezüglich Aktualität der Information eine Rolle.
- *Vagheit*: Information ist unscharf (vgl. Abschn. 1.2.2), das heißt sie ist oft unpräzis und hat unterschiedliche Aussagekraft (Qualität).
- *Träger*: Information benötigt keinen fixierten Träger; sie ist unabhängig vom Herkunftsort.

Diese Eigenschaften belegen, dass sich digitale Güter (Information, Software, Multimedia, etc.) in Handhabung sowie in ökonomischer, rechtlicher und sozialer Wertung von materiellen Gütern stark unterscheiden. Beispielsweise verlieren physische Produkte durch Nutzung meistens an Wert, gegenseitige Nutzung von Information hingegen kann einem Wertzuwachs dienen. Ein weiterer Unterschied besteht darin, dass materielle Güter mit kalkulierbaren Kosten hergestellt werden können, die Erzeugung digitaler Produkte jedoch schwierig kalkulierbar bleibt. Allerdings ist Vervielfältigung von Informationen gegenüber materiellen Gütern einfach und dank Moore's Law[2] kostengünstig (Rechenaufwand, Material des Infor-

[1] In Einzelfällen wird versucht, z. B. mit digitalen Wasserzeichen die Urheberschaft kenntlich zu machen und vor Missbrauch zu schützen.

[2] Moore's Law ist eine Faustregel und sagt aus, dass sich die Komplexität integrierter Schaltungen bei gleichbleibenden Kosten innerhalb von ein bis zwei Jahren regelmäßig verdoppelt.

mationsträgers). Zudem bleiben bei Informationsobjekten die Eigentumsrechte und Besitzverhältnisse schwer bestimmbar, obwohl digitale Wasserzeichen und andere Datenschutz- und Sicherheitsmechanismen zur Verfügung stehen (Meier und Stormer 2012).

Das Sammeln, Speichern und Verarbeiten digitaler Information ist zum Alltag geworden und wichtige Dienstleistungen sind davon abhängig; man denke dabei an die digitalen Kontaktdaten. Dies nicht nur bei kommerziellen Anwendungen, sondern auch im öffentlichen Leben. Die wichtigsten Herausforderungen lauten: Wie bewältigen wir diesen Information Overload? Wie können wir die Qualität der heterogenen Daten gewährleisten? Wann können wir den Auswertungen und Empfehlungen trauen? Wie sichern wir unsere Entscheidungen ab?

Die Heterogenität umfangreicher Datensammlungen und die Vielfalt von Auswertungsmethoden rücken Big Data Analytics in den Fokus vieler Entscheidungsträger in Politik, Wirtschaft, öffentlicher Verwaltung und Gesellschaft. Die Herangehensweise zu erfolgversprechenden Auswertungsstrategien ist nicht von vornherein klar erkenntlich und muss eventuell iterativ in Abklärungsschritten erarbeitet werden. Wichtig bleibt, Begriffe und Vorgehensweisen betreffend Big Data Analytics im Vorfeld zu klären, einzuordnen und allen Anspruchsgruppen zu kommunizieren.

1.1.1 Was heißt Big Data?

Seit einigen Jahren sind Unternehmen, Organisationen, Forschungseinrichtungen und Citizens mit Big Data konfrontiert (Fasel und Meier 2016), das heißt mit der Bewältigung umfangreicher Daten aus unterschiedlichen Datenquellen. Die Herkunft der Daten sowie deren Struktur sind vielfältig. Aus diesem Grunde werden die digitalen Daten oft mit dem Begriff Multimedia gemäß Abb. 1.1 charakterisiert.

Big Data Analytics kann mit Hilfe von V's näher gefasst werden (Fasel und Meier 2016; Meier und Kaufmann 2016):

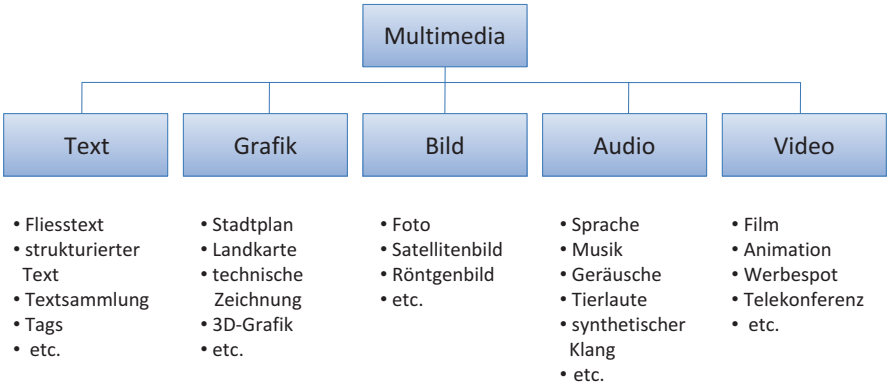

Abb. 1.1 Vielfalt der Multimedia-Daten beim Big Data Analytics, angelehnt an Meier (2018)

- *Volume*: Der Datenbestand ist umfangreich und liegt im Tera- bis Zettabytebereich (Megabyte = 10^6 Byte, Gigabyte = 10^9 Byte, Terabyte = 10^{12} Byte, Petabyte = 10^{15} Byte, Exabyte = 10^{18} Byte, Zettabyte = 10^{21} Byte).
- *Variety*: Unter Vielfalt versteht man bei Big Data Analytics die Verarbeitung von strukturierten, semi-strukturierten und unstrukturierten Multimedia-Daten (Text, Grafik, Bilder, Audio und Video gemäß Abb. 1.1).
- *Velocity*: Der Begriff bedeutet Geschwindigkeit und verlangt, dass im Extremfall Datenströme (Data Streams) in Echtzeit ausgewertet und analysiert werden können.
- *Value*: Big Data Analytics soll den Wert des Unternehmens oder der Organisation steigern. Investitionen in Personal und technische Infrastruktur werden dort gemacht, wo eine Hebelwirkung besteht respektive ein Mehrwert generiert werden kann.
- *Veracity*: Da viele Daten vage oder ungenau sind, müssen spezifische Algorithmen zur Bewertung der Aussagekraft respektive zur Qualitätseinschätzung der Resultate verwendet werden (vgl. Soft Computing in Abschn. 1.2.2). Umfangreiche Datenbestände garantieren nicht per se eine bessere Auswertungsqualität.

Veracity bedeutet in der deutschen Übersetzung Aufrichtigkeit oder Wahrhaftigkeit. Im Zusammenhang mit Big Data Analytics wird damit ausgedrückt, dass Datenbestände in unterschiedlicher Datenqualität vorliegen und dass dies bei Auswertungen berücksichtigt werden muss. Neben statistischen Verfahren und Data Mining existieren unscharfe Methoden des Soft Computing, die einem Resultat oder einer Aussage Wahrheitswerte zwischen wahr und falsch zuordnen (vgl. Ausführungen zum Soft Computing in Abschn. 1.2.2 resp. zum Fuzzy Portfolio in Abschn. 1.3.2).

Big Data ist nicht nur eine Herausforderung für profitorientierte Unternehmen im elektronischen Geschäft, sondern auch für das Aufgabenspektrum von Regierungen, öffentlichen Verwaltungen, NGO's (Non Governmental Organizations) und NPO's (Non Profit Organizations).

Als Beispiel seien die Programme für Smart City oder Ubiquitous City erwähnt, das heißt die Nutzung von Big-Data-Technologien in Städten, Agglomerationen und ländlichen Regionen. Ziel dabei ist, den sozialen und ökologischen Lebensraum nachhaltig zu entwickeln. Dazu zählen zum Beispiel Projekte zur Verbesserung der Mobilität, Nutzung intelligenter Systeme für Wasser- und Energieversorgung, Förderung sozialer Netzwerke, Erweiterung politischer Partizipation, Ausbau von Entrepreneurship, Schutz der Umwelt oder Erhöhung von Sicherheit und Lebensqualität.

1.1.2 Relevanz von Datenspeichersystemen

Relationale Datenbanksysteme, oft SQL-Datenbanksysteme genannt, organisieren die Datenbestände in Tabellen (Relationen) und verwenden als Abfrage- und Manipulationssprache die international standardisierte Sprache SQL (Structured Query Language; Meier und Kaufmann 2016).

Relationale Datenbanksysteme sind zurzeit in den meisten Unternehmen, Organisationen und vor allem in KMU's (Kleinere und Mittlere Unternehmen) im Einsatz. Bei massiv verteilten Anwendungen im Web hingegen oder bei Big-Data-Anwendungen muss die relationale Datenbanktechnologie oft mit NoSQL[3]-Technologien ergänzt werden, um Webdienste rund um die Uhr und weltweit anbieten zu können.

Ein NoSQL-Datenbanksystem unterliegt einer massiv verteilten Datenhaltungsarchitektur. Die Daten selber werden je nach Typ der NoSQL-Datenbank entweder als Schlüssel-Wertpaare („key/value store"), in Spalten oder Spaltenfamilien („column store"), in Dokumentspeichern („document store") oder in Graphen („graph database") gehalten (vgl. Abb. 1.2).

Um hohe Verfügbarkeit zu gewähren und das NoSQL-Datenbanksystem gegen Ausfälle zu schützen, werden unterschiedliche Replikationskonzepte unterstützt. Zudem wird mit dem sogenannten Map/Reduce-Verfahren hohe Parallelität und Effizienz für die Datenverarbeitung gewährleistet. Beim Map/Reduce-Verfahren werden Teilaufgaben an diverse Rechnerknoten verteilt und einfache Schlüssel-Wertpaare extrahiert („map") bevor die Teilresultate zusammengefasst und ausgegeben werden („reduce").

In Abb. 1.2 ist ein elektronischer Shop als Beispiel für die Vielfalt von analytischen Optionen schematisch dargestellt:

- *Key/Value Store*: Um eine hohe Verfügbarkeit und Ausfalltoleranz zu garantieren, wird ein Key/Value-Speichersystem für die Session-Verwaltung sowie für den Betrieb der Einkaufswagen eingesetzt. Die Analyse von Kundenbesuchen

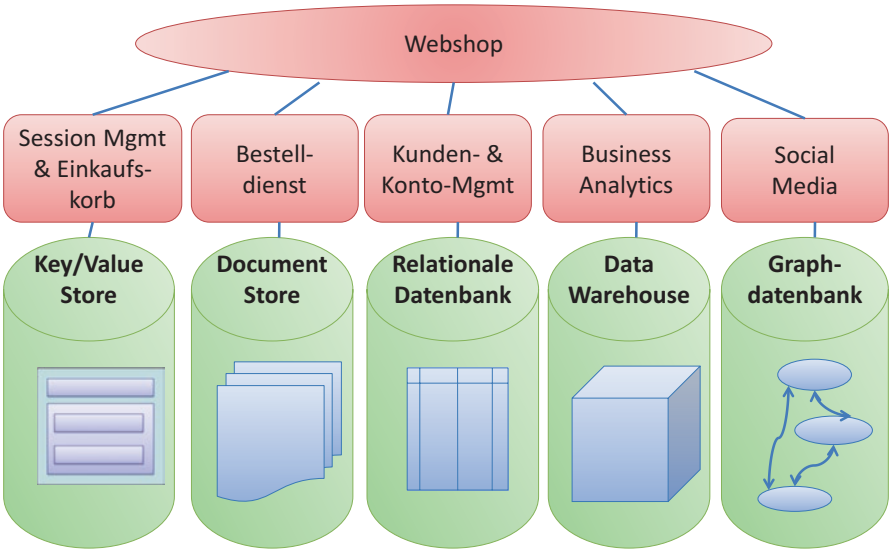

Abb. 1.2 Nutzung von SQL- und NoSQL-Datenbanken im Webshop, angelehnt an Meier (2018)

[3] NoSQL bedeutet ‚Not only SQL'.

respektive die Auswertung der Einkaufswagen kann direkt im Key/Value Store oder im Data Warehouse (siehe unten) vorgenommen werden.
- *Document Store*: Die Kundenbestellungen selber werden im Dokumentspeicher abgelegt. Aktuelle Bestellungen lassen sich direkt im Document Store analysieren. Zeitreihenvergleiche oder differenzierte Auswertungen und Prognosen werden im Data Warehouse (z. B. mit Descriptive oder Predictive Analytics gemäß Abschn. 1.1.3) vorgenommen.
- *Relationales Datenbanksystem*: Kunden- und Kontoverwaltung erfolgt mit einem relationalen Datenbanksystem. Dieses klassische Datenbanksystem garantiert jederzeit Konsistenz und ist unter anderem für lückenlose Buchhaltung und verlässliches Finanzmanagement relevant. Entsprechende Auswertungen wichtiger Finanzkennzahlen erfolgen hier oder im Data Warehouse.
- *Data Warehouse*: Bedeutend für den erfolgreichen Betrieb eines Webshops ist das Performance Measurement. Mit Hilfe von Web Analytics werden wichtige Kenngrößen („key performance indicators“, KPIs) der Inhalte wie der Webbesucher in einem Data Warehouse aufbewahrt. Spezifische Werkzeuge (Data Mining, Predictive Business Analysis) werten Geschäftsziele wie Erfolg der getroffenen Maßnahmen regelmäßig aus. Da die Analysearbeiten auf dem mehrdimensionalen Datenwürfel („datacube“) zeitaufwendig sind, wird dieser InMemory[4] gehalten.
- *Graphdatenbank*: Falls die Beziehungen unterschiedlicher Anspruchsgruppen analysiert werden sollen, drängt sich der Einsatz von Graphdatenbanken auf. Diese erlauben, Geschäftsbeziehungen, soziale Interaktionen, Meinungsäusserungen, Bewertungen von Produkten oder Dienstleistungen, Kritik und Wünsche etc. für die Kundenbindung zu nutzen und auszuwerten.

Die Verknüpfung eines Webshops mit sozialen Medien ist für ein Unternehmen oder eine Organisation zukunftsweisend. Neben der Ankündigung von Produkten und Dienstleistungen kann analysiert werden, ob und wie die Angebote bei den Nutzern ankommen. Bei Schwierigkeiten oder Problemfällen wird mit gezielter Kommunikation und geeigneten Maßnahmen versucht, einen möglichen Schaden abzuwenden oder zu begrenzen. Darüber hinaus hilft die Analyse von Weblogs oder die Verfolgung aufschlussreicher Diskussionen in sozialen Netzen, Trends oder Innovationen für das eigene Geschäft zu erkennen.

1.1.3 Facetten des Big Data Analytics

Unter Analytics versteht man das Analysieren und Interpretieren umfassender, oft heterogener Datenbestände, um Muster und Zusammenhänge in den Daten aufzu-

[4] Ein InMemory-Datenbanksystem nutzt den Arbeitsspeicher des Rechners als Speicher und muss die Daten bei der Verarbeitung nicht auf einem externen Medium (z. B. Festplatte) ein- und auslagern, was zu Effizienzsteigerungen beim Analytics führt.

decken und Entscheidungsgrundlagen für betriebliche wie gesellschaftliche Abläufe oder für private Zwecke zu erhalten. Der Begriff Analytics hat unterschiedliche Ausprägungen, wie Abb. 1.3 aufzeigt.

Ziel des Big Data Analytics ist das Erfassen und Beschreiben relevanter Merkmale oder Attribute zum Erhalt eines Beschreibungsmodells, Analyse- und Empfehlungsmodells zur Erreichung der Ziele des Unternehmens respektive der Organisation. Im Kern stehen Descriptive Analytics, Diagnostic Analytics, Predictive Analytics sowie Prescriptive Analytics:

- *Descriptive Analytics*: Werkzeuge erläutern den Entscheidungsträgern von Unternehmen und Organisationen aufgrund gesammelter Daten den Verlauf der Geschäfts- und Kundenbeziehungen und ermöglichen den Vergleich in Zeitreihen. Spezifische Visualisierungstechniken und Infografiken erlauben, die Veränderungen der Indikatoren (Kennzahlen) darzustellen.
- *Diagnostic Analytics*: Diese Werkzeuge sind darauf ausgelegt, die Hintergründe der Entwicklung des Geschäfts respektive der Beziehungen mit den Anspruchsgruppen zu erklären. Spezifische Werkzeuge zur Berichterstattung extrahieren zudem die Gründe für die zeitliche Entwicklung und bereiten sie in Grafiken auf.
- *Predictive Analytics:* Hier werden künftige Ereignisse und Entwicklungen aufgrund von historischen Daten prognostiziert. Zudem helfen Algorithmen der künstlichen Intelligenz und des maschinellen Lernens aufzuzeigen, welche Maßnahmen welche Wirkungen in Zukunft erzielen könnten (Erklärungsmodell).
- *Prescriptive Analytics*: Mit diesen Werkzeugen werden nicht nur künftige Entwicklungen evaluiert, sondern konkrete Empfehlungsoptionen zur Entscheidungsfindung sowie für Zukunftsszenarien eines erfolgreichen Geschäftsverlaufs generiert. Die Werkzeuge zielen darauf ab, über die reine Vorhersage hinaus Handlungsoptionen zu erhalten, um deren Auswirkungen abschätzen zu können (Entscheidungsmodell).

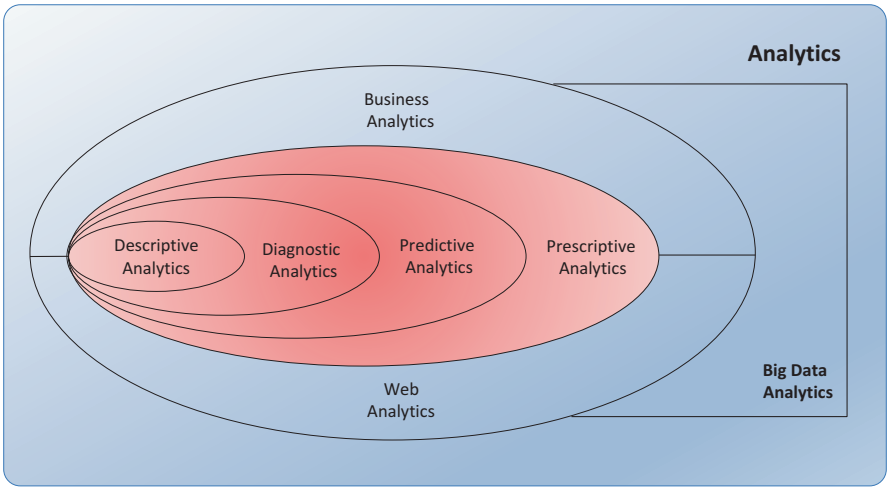

Abb. 1.3 Begriffseinordnung, angelehnt an Gluchowski (2016) und erweitert von Meier (2019)