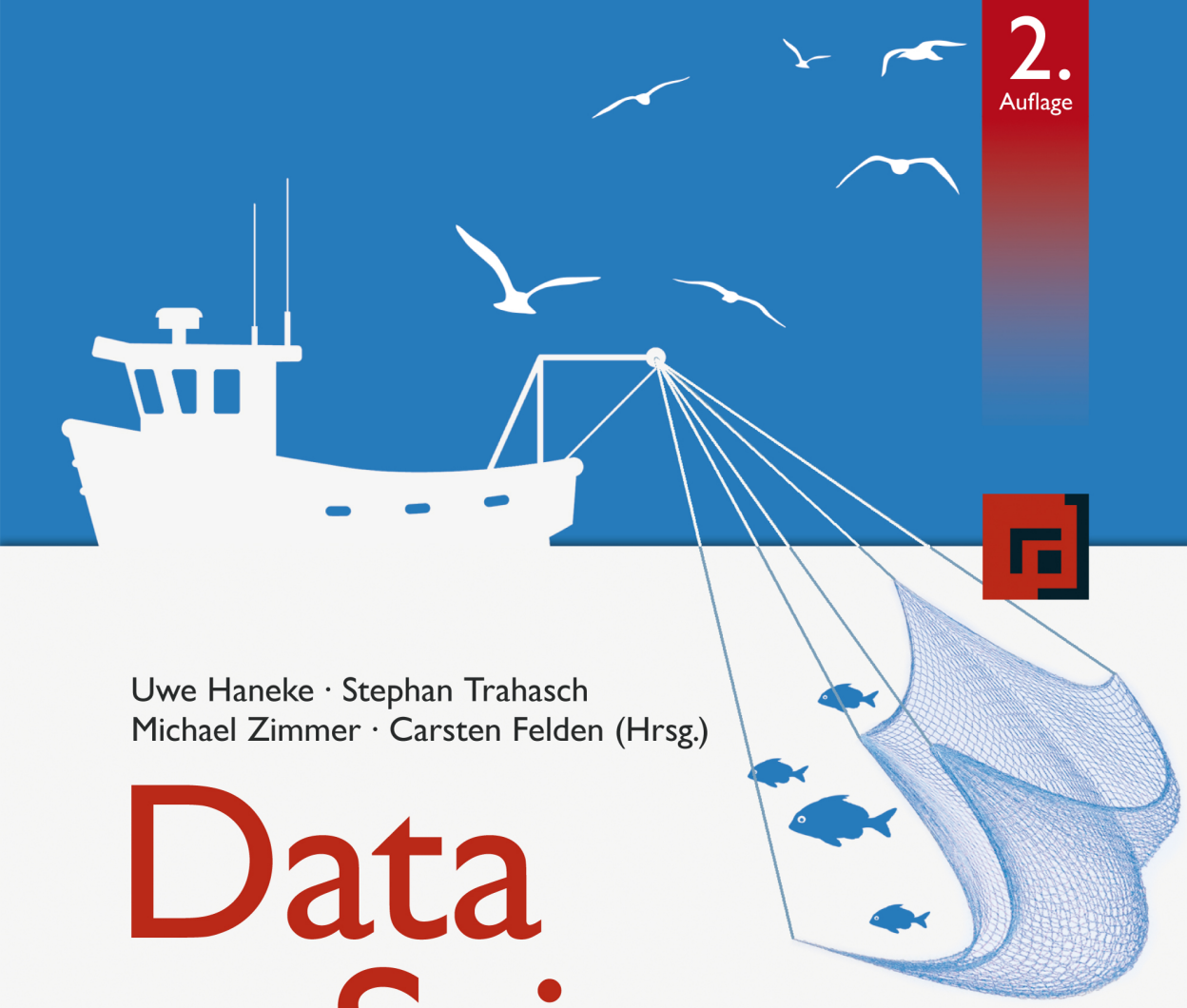


2. Auflage

Uwe Haneke · Stephan Trahasch
Michael Zimmer · Carsten Felden (Hrsg.)

Data Science

Grundlagen, Architekturen und Anwendungen

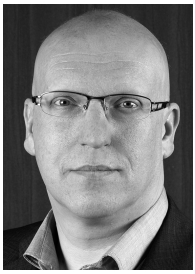
tdwi WISSEN

dpunkt.verlag

Prof. Dr. Uwe Haneke ist seit 2003 Professor für Betriebswirtschaftslehre und betriebliche Informationssysteme an der Hochschule Karlsruhe – Technik und Wirtschaft. Dort vertritt er u. a. die Bereiche Business Intelligence und Geschäftsprozessmanagement im Fachgebiet Informatik. Seine Publikationen beschäftigen sich mit den Themen Open Source Business Intelligence, Self-Service-BI und Analytics.

Prof. Dr. Stephan Trahasch ist Professor für betriebliche Kommunikationssysteme und IT-Sicherheit an der Hochschule Offenburg. Seine Forschungsschwerpunkte liegen in den Bereichen Data Mining, Big Data und Agile Business Intelligence. In Forschungsprojekten beschäftigt er sich mit der praktischen Anwendung von Data Mining und Big-Data-Technologien und deren Herausforderungen in Unternehmen. Er ist Leiter des Institute for Machine Learning and Analytics und Mitglied der Forschungsgruppe Analytics und Data Science an der Hochschule Offenburg.

Dr. Michael Zimmer verantwortet bei der Zurich Gruppe Deutschland das Thema künstliche Intelligenz. Hierbei beschäftigt er sich sparten- und ressortübergreifend mit der Identifikation, Entwicklung, Produktivsetzung und Industrialisierung von KI-Anwendungsfällen. Er hat über Data & Analytics Governance promoviert, ist Autor und Herausgeber diverser Publikationen und TDWI Fellow. Vor seiner Zeit bei der Zurich Deutschland war er fast 14 Jahre in der Beratung tätig und beschäftigte sich mit dem Aufbau komplexer Data-, Analytics- und KI-Architekturen sowie der Einführung und Konzeption zugehöriger Governance-Strukturen.



Prof. Dr. Carsten Felden ist Direktor des Instituts für Wirtschaftsinformatik an der TU Bergakademie Freiberg (Sachsen). Er hat dort die Professur für ABWL, insbes. Informationswirtschaft/Wirtschaftsinformatik inne und vertritt in der Lehre die Themen der Wirtschaftsinformatik mit dem Fokus auf Business Analytics (BA). Zentrale Forschungsthemen sind neben Business Analytics Data Warehousing, eXtensible Business Reporting Language (XBRL) und IT-Reifegradmodelle sowie Digitalisierung im Kontext der Business Intelligence. Er ist Vorstandsvorsitzender des TDWI e.V. und war Vorstandsmitglied des XBRL Deutschland e.V. Er veröffentlichte zahlreiche Artikel sowohl auf internationalen Konferenzen als auch in wissenschaftlichen und praxisorientierten Zeitschriften. Im Weiteren ist er häufig Program Chair bei internationalen Konferenzen wie WI, ECIS oder AMCIS. In Kooperation mit anderen Autoren verfasst er regelmäßig Bücher zu Themen der analytischen Ansätze im betrieblichen Umfeld.

Uwe Haneke · Stephan Trahasch · Michael Zimmer · Carsten Felden (Hrsg.)

Data Science

Grundlagen, Architekturen und Anwendungen

2., überarbeitete und erweiterte Auflage

Edition TDWI

tdwi WISSEN

dpunkt.verlag

Uwe Haneke
uwe.haneke@hs-karlsruhe.de

Stephan Trahasch
stephan.trahasch@hs-offenburg.de

Michael Zimmer
michael.zimmer2@zurich.com

Carsten Felden
carsten.felden@bwl.tu-freiberg.de

Lektorat: Christa Preisendanz
Copy-Editing: Ursula Zimpfer, Herrenberg
Satz: Birgit Bäuerlein, Frank Heidt
Herstellung: Stefanie Weidner
Umschlaggestaltung: Helmut Kraus, *www.exclam.de*
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

Fachliche Beratung und Herausgabe von dpunkt.büchern in der Edition TDWI:
Prof. Dr. Peter Gluchowski · peter.gluchowski@wirtschaft.tu-chemnitz.de

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

ISBN:
Print 978-3-86490-822-4
PDF 978-3-96910-152-0
ePub 978-3-96910-153-7
mobi 978-3-96910-154-4

2. Auflage 2021

Wieblinger Weg 17
69123 Heidelberg

Hinweis:
Dieses Buch wurde auf PEFC-zertifiziertem Papier aus nachhaltiger Waldwirtschaft gedruckt. Der Umwelt zuliebe verzichten wir zusätzlich auf die Einschweißfolie.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: *hallo@dpunkt.de*.

5 4 3 2 1 0

Vorwort zur 2. Auflage

Data Science findet in den unterschiedlichsten Wirtschaftsbereichen eine immer weitere Verbreitung. Verstärkt setzen Unternehmen heute auf die Nutzung von Data Science, um ihre Wettbewerbsfähigkeit zu verbessern. Neue Werkzeuge erlauben es auch Anwendern aus Fachabteilungen, die keine Data-Science-Experten sind, erste datengetriebene Analysen durchzuführen und *Proof of Concepts* zu entwickeln. Dieser Trend wird auch dadurch verstärkt, dass Data-Science-Werkzeuge zunehmend in der Cloud betrieben werden und daher weniger hohe IT-Hürden darstellen. Seit der Veröffentlichung der ersten Auflage im Mai 2019 hat sich aber nicht nur die Verbreitung von Data Science insgesamt geändert. Auch der Grad der organisatorischen Einbindung von Data Science hat sich weiterentwickelt. Data Science ist keine Spielwiese mehr, auf der Data-Science-Teams, abgekoppelt von der sonstigen Organisation, große Datenbestände analysieren, um zu neuen Erkenntnissen über Kunden, Produkte, Wartung, Preisgestaltung etc. zu gelangen, die dann wiederum zum Teil recht mühsam, wenn überhaupt, in den Wertschöpfungsprozess des Unternehmens einfließen. Data Science ist heute schon oft ein direkter Bestandteil der Wertschöpfungskette. Die hier gemeinsam mit den Fachabteilungen erzielten Erkenntnisse fließen direkt in das Produkt, das Produktportfolio, die Produktion ein und macht *Industrialized Data Science* vielfach bereits zur Realität. Die entdeckten Muster, Möglichkeiten und potenziellen Produkte finden nun schneller Eingang in die Wertschöpfung und stärken damit die Position des Unternehmens.

Verstärkt lassen sich nun darüber hinaus nicht nur Fort- und Weiterbildungsangebote finden. In den vergangenen 18 Monaten sind auch an Hochschulen spezialisierte Studiengänge (Bachelor und Master) entwickelt worden, um die Nachfrage nach qualifiziertem Personal zu befriedigen. Ein Ende dieser Entwicklung, in der Data Science zunehmend an Bedeutung gewinnt, ist nach wie vor nicht absehbar. Aus diesem Grund freut es uns, dass wir in der vorliegenden zweiten Auflage des Buches mit erweiterten und neuen Grundlagenkapiteln sowie Fallstudien dies weiter ausbauen konnten.

Das Kapitel »Feature Selection« diskutiert diesen wichtigen Aspekt im Data-Science-Prozess und ergänzt damit den bestehenden Grundlagenteil des Buches.

Vor dem Hintergrund immer weiter zunehmender Datenquellen und -mengen ist Feature Selection mittlerweile zu einer Stellschraube geworden, will man die Güte und die Laufzeiten der Modelle und damit nicht zuletzt die damit verbundenen Kosten im Blick behalten.

Das Kapitel »Deep Learning« wurde um den Aspekt »Deep Reinforcement Learning« erweitert. Hier lässt sich sehen, wie sich methodische Ansätze in der Data Science weiterentwickeln, um den Anforderungen der Praxis gerecht zu werden. Dieser neu integrierte Part zeigt die Verbindung von Deep Learning mit einer Reinforcement-Learning-Strategie an einem konkreten Anwendungsfall.

Die Kapitel »Von einer BI-Landschaft zum Data & Analytics-Ökosystem« und »Self-Service und Governance im Data-Science-Umfeld: der emanzipierte Anwender« wurden für die neue Auflage überarbeitet und aktualisiert. Unter anderem wurde dabei dem Bereich Data Governance mehr Gewicht verliehen. Darüber hinaus lassen sich gerade bei den Self-Service-Angeboten neue Trends und Tendenzen feststellen, die hier vorgestellt werden. Ebenso wurde das Kapitel zu Analytics-Ökosystemen um weitere Aspekte der Industrialisierung und Erfahrungen aus der Praxis beim Aufbau der zugrunde liegenden Architekturen ergänzt.

Eine weitere Fallstudie baut unseren Praxisteil aus. Im Kapitel »Künstliche Intelligenz bei der Zurich Versicherung« wird die Nutzung von KI und Data Science in der Versicherungsbranche beispielhaft präsentiert. Im Vordergrund steht hier, unterschiedliche KI-Anwendungsfälle vorzustellen und aufzuzeigen, dass auch Versicherungen künstliche Intelligenz sehr erfolgreich im operativen Geschäft einsetzen.

Wir möchten uns an dieser Stelle auch beim dpunkt.verlag (hier insbesondere bei Christa Preisendanz) und beim TDWI für das uns entgegengebrachte Vertrauen und natürlich für die Unterstützung bedanken! Ohne diesen Einsatz wäre auch die zweite Auflage nicht möglich gewesen.

Auch bei Ihnen, liebe Leserinnen und Leser, möchten wir uns bedanken. Einerseits natürlich dafür, dass Sie unser Buch aus dem mittlerweile sehr großen Portfolio der Fachliteratur zum Thema Data Science ausgewählt haben. Andererseits aber auch für die Rückmeldungen, die wir erhalten haben. Diese haben uns auf unserem Weg bestärkt, den wir mit diesem Buch eingeschlagen haben. Konstruktive Hinweise und Vorschläge haben wir versucht, so weit wie möglich in der neuen Auflage zu berücksichtigen.

Wir wünschen Ihnen viel Spaß bei der Lektüre dieser zweiten Auflage und hoffen, dass Sie das Buch gut auf Ihrer Reise in und durch die Welt der Data Science begleitet. »Nichts entwickelt die Intelligenz wie das Reisen«, bemerkte einst Émile Zola. Was gibt es Schöneres, als Interessantes und Nützliches zu verbinden!

Uwe Haneke, Stephan Trahasch, Michael Zimmer, Carsten Felden
Karlsruhe, Offenburg, Köln, Freiberg im Dezember 2020

Vorwort

Data Science, Machine Learning und auch künstliche Intelligenz sind in aller Munde und sorgen gerade im Rahmen der Digitalisierung für viel Gesprächsstoff. Von den angesprochenen Begriffen sind nicht nur die Produktionsprozesse, die bearbeiteten Geschäftsfelder, das Produktportfolio und die damit verbundene Wertschöpfung der Unternehmen betroffen. Auch für die bereits bestehende analytische Landschaft innerhalb eines Unternehmens ergeben sich neue Chancen, die sowohl in der zunehmenden Datenbereitstellung und Datennutzung als auch in den Möglichkeiten der neuen Technologien selbst zu verorten sind. Gerade die oftmals über viele Jahre hinweg aufgebauten analyseorientierten Systeme lassen sich durch Data Science und Big Data ergänzen und qualitativ verbessern, was dem wachsenden Analyseverständnis in Unternehmen Rechnung trägt. Dazu hat sich neben der Business Intelligence (BI) und den typischen BI-Analysen mit der Data Science eine weitere Analysewelt entwickelt, die sich mit intensiven Datenauswertungen und insbesondere auch mit prädiktiven Analysen auseinandersetzt. Während in der BI überwiegend historische Daten und teilweise auch Echtzeitdaten ausgewertet und entsprechende Kennzahlen ermittelt werden, ermöglicht Data Science mit prädiktiven Methoden unter anderem auch eine Vorhersage von Kennzahlen. Unternehmen, die Business Intelligence und Data Science verwenden, werden somit in die Lage versetzt, umfangreiche Analysen durchzuführen, die einen Blick in die Vergangenheit, auf den aktuellen Zustand und mit Data Science auch in die »nahe« Zukunft ermöglichen. Bei BI und Data Science handelt es sich nicht um unabhängige Systemwelten – die Schnittstelle zwischen Data Science und Business Intelligence ist durchaus bidirektional zu sehen. Ebenso wie die konsolidierte Datenbasis im etablierten Data Warehouse neben Roh- und weiteren Detaildaten eine wichtige Quelle für tiefergehende Analysen ist, so können auch die Ergebnisse der Data Science wieder zurück in die BI fließen.

Die aus der Zunahme des datengetriebenen Handelns in Unternehmen kombinierte Nutzung von Business Intelligence auf der einen und Data Science auf der anderen Seite bietet Potenziale, die den Unternehmen sowohl neue Entwicklungschancen, effizientere Entscheidungs- und Steuerungssysteme als auch die Verbesserung ihrer Wettbewerbsfähigkeit ermöglichen. Für Unternehmen, die heute bereits

über ein entwickeltes und etabliertes analytisches System, wie eben BI, verfügen, verbindet sich dies mit unterschiedlichen Fragestellungen: Wie sind Data Science und ihre Möglichkeiten fruchtbringend einzusetzen? Was gibt es zu beachten? Welche Konsequenzen hätte ein solcher Schritt? Ist die Nutzung von Data Science überhaupt sinnvoll für das Unternehmen?

Das vorliegende Buch richtet sich in erster Linie an Leserinnen und Leser aus Studium und Praxis, die bereits Erfahrung mit analytischen Systemen haben. Business-Intelligence-Manager gehören ebenso zu der Zielgruppe des Buches wie auch Daten- und Informationsverantwortliche im Unternehmen sowie Projektleiter aus dem BI- und Analytics-Bereich. Das Buch möchte die Anwenderinnen und Anwender in ihrem bestehenden Umfeld abholen, ihnen Schritt für Schritt die Welt der Data Science nahebringen und Möglichkeiten für die Nutzung von Data Science in ihrem Unternehmen aufzeigen. Dies soll über die Einteilung in einen Grundlagen- und einen praxisnahen Anwendungsteil geleistet werden. Dabei werden zunächst die verschiedenen Facetten der Data Science von der Herkunft über die Möglichkeiten der Nutzung und die dazu notwendigen Architekturen bis hin zu ethischen Aspekten und der Frage der Data Privacy diskutiert. Im Anschluss arbeiten Fallstudien aus der Praxis verschiedene Aspekte heraus, die es bei der Implementierung und beim Einsatz von Data Science im Zusammenspiel mit bestehenden BI-Systemen im Unternehmen zu beachten gilt. Die einzelnen Kapitel können dabei auch in einer anderen als der hier vorliegenden Reihenfolge oder auch selektiv gelesen werden.

Wir als Herausgeber sind sehr froh, dass uns der TDWI und der dpunkt.verlag bei diesem Projekt unterstützt und immer wieder auf unserem Weg bestärkt haben. In erster Linie gilt unser Dank natürlich den beteiligten Autorinnen und Autoren, ohne die uns die Umsetzung des Buches nicht gelungen wäre. Wir freuen uns, dass wir bemerkenswerte Spezialisten für die unterschiedlichen Aspekte von Data Science und ihrer Nutzung im Unternehmen gewinnen konnten, was zu einem ganzheitlichen und runden Bild geführt hat, in dem die Praxis nicht zu kurz kommt.

Beim dpunkt.verlag möchten wir uns speziell bei Christa Preisendanz bedanken, die uns mit ihrer Erfahrung stets auf die Fertigstellung des Buches fokussiert gehalten hat.

Wir hoffen, dass Ihnen, liebe Leserinnen und Leser, die Lektüre des vorliegenden Buches gefällt und Sie viel Nutzen daraus ziehen können, wenn Sie sich für den Einsatz von Data Science interessieren. Da sich der Prozess der Verknüpfung von Data Science und der bestehenden analytischen Landschaft erst am Anfang befindet, wird es spannend werden, zu beobachten, wie schnell und in welchem Umfang die Entwicklung in den Unternehmen voranschreiten wird und welche Transformationen sich daraus ergeben werden.

Uwe Haneke, *Stephan Trahasch*, *Michael Zimmer*, *Carsten Felden*
Karlsruhe, Offenburg, Stuttgart, Freiberg im Januar 2019

Inhaltsübersicht

Fallstudien 211

Anhang 331

Inhaltsverzeichnis

Fallstudien 211

1 Einleitung

Uwe Haneke · Stephan Trahasch · Michael Zimmer · Carsten Felden

1.1 Von Business Intelligence zu Data Science

Seit dem Jahr 2015 hat sich die Welt der Business Intelligence (BI) schnell und signifikant verändert. Big Data und die damit zusammenhängenden Entwicklungen im Bereich der Data Science haben auch die Business Intelligence nicht unberührt gelassen. Und so sehen wir aktuell eine Erweiterung der bisherigen BI-Systeme und Architekturen, die die betrieblichen Informationssysteme agiler, schneller, mächtiger und passgenauer machen. Die neue BI-Welt enthält heute eine integrierte analytische Komponente, die weit über das hinausgeht, was man bis vor Kurzem noch kannte.

Dabei ist es nicht so, dass Analytics etwas grundlegend Neues in der Business Intelligence wäre. Allerdings vermochte es Data Science mit ihrem Hintergrund auf der wissenschaftlich, technischen Ebene, einen Innovationsschub auszulösen, dessen Ende noch nicht absehbar zu sein scheint. Die nachfolgenden Ausarbeitungen stellen daher zunächst dar, wie die bisherige BI-Entwicklung beginnend in den 1960er-Jahren bis heute verlief. Dabei wird ein besonderes Augenmerk auf die Business Analytics gelegt, die sich im Grunde genommen als das Pendent der Data Science in der Business Intelligence interpretieren lässt. Stubbs sieht dabei Business Analytics wie folgt [Stubbs 2013]:

> »The cornerstone of business analytics is pure analytics. Although it is a very broad definition, analytics can be considered any data-driven process that provides insight. It may report on historical information or it may provide predictions about future events; the end goal of analytics is to add value through insight and turn data into information.«

Stubbs Definition und unser Verständnis von Data Science, das wir in diesem Buch zugrunde legen wollen, überlappen sich damit großteils. Im Folgenden wird im Buch der Begriff Business Analytics zwar zugunsten von Data Science (vgl. Abschnitt 1.2) aufgegeben, der für die datenanalytischen Methoden und Vorgehensweisen stehen soll. Zum besseren Verständnis und um nicht zuletzt die Ähn-

lichkeiten im Vorgehen zu veranschaulichen, erfolgt aber zunächst eine Herleitung des Begriffs Business Analytics.

Was aber ist das Ziel der Business Analytics und inwieweit wird sich die Rolle von Business Analytics durch Methoden und Technologien aus dem Bereich Big Data und Data Science verändern? Haben die Unternehmen mit Business Analytics nicht auch Data Mining betrieben? Diese Fragen lassen sich beliebig erweitern. Leider stehen den Fragen nur wenige präzise Antworten gegenüber. Wenn man versucht, sich diesem Thema von einer fachlichen Seite zu nähern, stellt man schnell fest, dass die Datenorientierung im betriebswirtschaftlichen Handeln zugenommen hat. Diese Zunahme entsteht auch durch die wachsende Integration unterschiedlicher unternehmensinterner und -externer Systeme. Basierend auf entstehenden Datensammlungen werden im Unternehmen schon von jeher Entscheidungen getroffen. Aktuell ist jedoch eine deutliche Zunahme der Datenorientierung bei Entscheidungen auf allen Unternehmensebenen zu verzeichnen. Dabei gerät nun auch zunehmend die technische und methodische Unterstützung bei der Entscheidungsfindung in die Diskussion – und im BI-Umfeld finden wir diese Diskussion unter der Überschrift *Business Analytics*.

Unter Business Analytics wird die kontinuierliche Erforschung und Untersuchung von vergangenheitsorientierten Geschäftsdaten verstanden, um darin Erkenntnisse sowohl über die abgelaufene als auch die kommende Geschäftstätigkeit zu erlangen, die wiederum in die einzelnen zu planenden Geschäftsaktivitäten einfließen [Felden 2012]. Die Kontinuität entsteht durch die regelmäßige Ausführung von Analysetätigkeiten, die sich entsprechend in einer Ablauforganisation implementieren lassen. Iterativ sind derartige Aktivitäten, weil im Analyseprozess häufig eher neue Fragen als abschließende Antworten entstehen, die letztlich zu untersuchen sind. So kann die bisherige Geschäftstätigkeit nachvollzogen werden, um Verbesserungen bei neuen Handlungen zu ermöglichen.

Letztlich ist Business Analytics ein Prozess, der aus den in der folgenden Abbildung gezeigten Schritten besteht und eng an das in Abschnitt 1.3 vorgestellte CRISP-DM angelehnt ist.

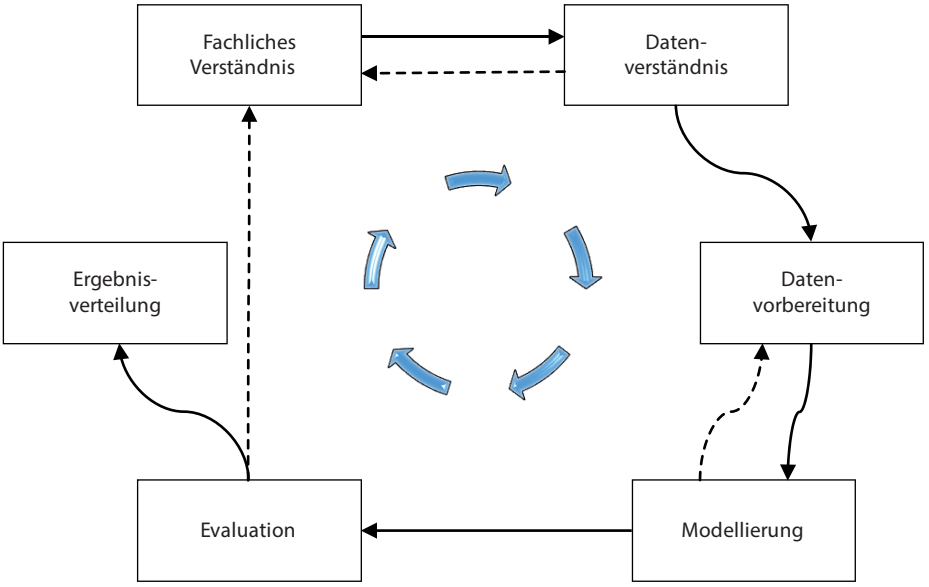


Abb. 1–1 *Schritte der Business Analytics*

Das fachliche Verständnis bestimmt die Auswahl der Daten, wobei dabei oftmals Rückfragen bzw. Nachbesserungen erforderlich sind, sodass fachliches Verständnis und Datenverständnis interdependent sind. Die vorliegenden Daten werden entsprechend aufbereitet in ein Modell überführt. Dabei bringt es die Modellbildung mit sich, dass die Aufbereitung neuerlich durchzuführen ist, da beispielsweise ein anderer Algorithmus als ursprünglich geplant genutzt wird. Die erzeugten Modelle sind zu evaluieren und deren Ergebnisse zur Nutzung an die jeweiligen Anwender weiterzuleiten. Die Erkenntnisse aus deren Nutzung fließen wieder als fachliches Verständnis in einen neuen Durchlauf ein.

Bereits seit Ende der 1990er-Jahre ist der KDD-Prozess (KDD = Knowledge Discovery in Databases) mit seinen Schritten der Datenauswahl, Vorverarbeitung, Transformation, Data Mining und der Ergebnisinterpretation theoretische Grundlage marktgängiger Software. Letztlich basiert auch die Business Analytics auf diesen Vorgehensschritten und erweitert diesen KDD-Prozess um eine Quellenbewirtschaftung zu Beginn und fachliche Handlung im Sinne einer zu treffenden Entscheidung und deren Durchsetzung am Ende des Prozesses. Im Weiteren wird die KDD um den Evaluationsschritt ergänzt, er dient dem Vergleich der erzeugten Modellvarianten anhand eines sogenannten Gütemaßes.

Somit liegt nun ein Prozess vor, der eine Langfristigkeit und damit eine strategische Komponente inhärent in sich birgt, da die Ergebnisse Entscheidungsgrundlage für das unternehmerische Handeln darstellen. Fachliche Analyseanforderungen und technische Komponenten zur zielgruppen- und aufgabenadäquaten Unterstützung sind in diesem Prozess gemeinschaftlich zu betrachten, um im

Rahmen der Informationslogistik, also die Daten zur richtigen Zeit dem richtigen Empfänger in der richtigen Qualität zur Verfügung zu stellen [Dinter & Winter 2008], eine sinnhafte Vollautomation zu erzeugen. Das informationslogistische Verständnis der Business Intelligence, also des Prozesses, Daten zu sammeln, aufzubereiten und zur Entscheidungsfindung zur Verfügung zu stellen [Chamoni & Gluchowski 2006], mündet in der praktischen Umsetzung eher in eine Standardorientierung mit konsistenten Kennzahlen (Metriken) und Analysen. Sie ist Dashboard-basiert mit vordefinierten Berichtsstrukturen zur Beantwortung vorab definierter Fragestellungen, sodass ein indirekter Zugriff auf die multidimensionalen Strukturen, Berichte und aggregierte Daten stattfindet, was jedoch auch zu einem Exception Reporting, also dem Triggern von automatisierten Informationsbereitstellungen bei Schwellenwertüberschreitungen [Felden & Buder 2012, S. 17 ff.], weitergedacht werden kann. Business Analytics ergänzt das Business-Intelligence-Verständnis um weitere Analysen von z. B. Geschäftsaktivitäten und richtet dabei den Fokus auf die Unterstützung von interaktiven und erforschenden Analysen durch Endanwender. Das Ziel ist die Sammlung neuer Erkenntnisse und damit eine Verständnisgewinnung über vergangene Aktivitäten zur Entdeckung unbekannter Muster/Strukturen in den Datenbeständen. Dabei basiert Business Analytics auf Detaildaten, um einzelne Aktivitäten entsprechend betrachten und analysieren zu können.

Daten bzw. bereits daraus generierte Informationen zu besitzen, ist in den Unternehmen nicht mehr ein Wert an sich, vielmehr besteht der Wert darin, die Möglichkeit und Fähigkeit zu haben, Informationen aus unübersichtlichen Mengen von Daten und deren heterogenen Strukturen zu identifizieren und Entscheidungsträgern als Grundlage für unternehmerische Entscheidungen zur Verfügung zu stellen. Mit dem Fokus auf eine Datenauswertung ist dabei zwangsläufig das Thema der Business Analytics zunehmend in den Mittelpunkt gerückt. Auf dieser Basis verbindet Business Analytics moderne Verfahren der Auswertung von großen Datenvorräten, vor allem Data Mining, und maschinelles Lernen auf Grundlage der künstlichen Intelligenz und statistischer Methoden. Mittlerweile kombiniert Business Analytics einzelne Komponenten wie Kennzahlenkonzepte, Active/Realtime Warehousing, Data und Text Mining, User-Interface-Konzepte oder Systemintegration. Hierin liegt der eigentliche Nutzen; die Zusammenführung einzelner Komponenten bringt es mit sich, dass der Entscheider heute viel schneller auf Veränderungen in seinem Unternehmen oder der Unternehmensumwelt reagieren kann. Der strategische Mehrwert von Business Analytics wird damit deutlich. Entwicklungen der letzten Jahre haben das Image und den Agitationsrahmen von Business Analytics erweitert: Stichworte wie Systemintegration, Geschäftsprozessorientierung oder Benutzeroberflächendesign werden mit Business Analytics in Verbindung gebracht [Olson & Delen 2008, S. 151 ff.].

Sowohl Business Intelligence (BI) wie auch Business Analytics (BA) sind Begriffe, die am Ende einer langen Entwicklungsgeschichte der Managementun-

terstützungssysteme (MUS) stehen (siehe Abb. 1–2). Chronologisch wird die Genese der MUS in unterschiedliche Phasen eingeteilt, die jeweils vor dem Hintergrund der verfügbaren IT-Ressourcen zu sehen sind. Allen Phasen gemeinsam ist, dass nach Werkzeugen für eine adäquate Informationsversorgung für das Management gesucht wird. Vorrangig steht dabei die Unterstützung des Managements in der Entscheidungssituation an. Die folgende zeitliche Zuordnung ist nicht trennscharf, da sich die jeweiligen Konzepte überlagern und teilweise latent existieren. Es wird lediglich die dominante Begriffsprägung einer Epoche zugewiesen. Insgesamt stellt der Komplex MUS als Sammelbegriff aller Strömungen ein Kontinuum dar.

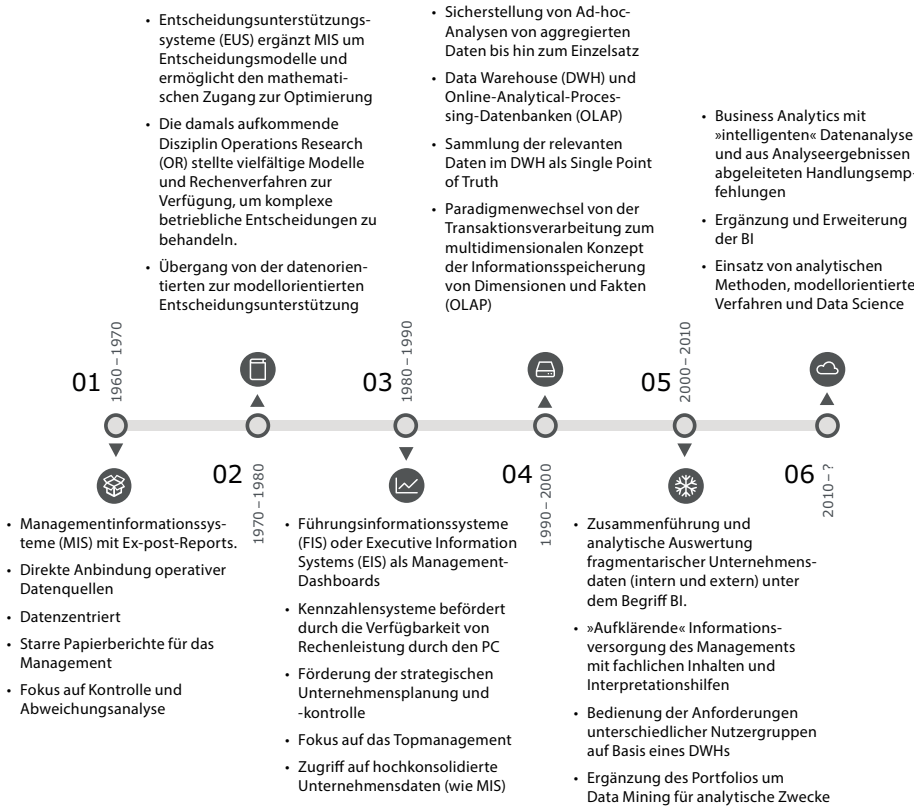


Abb. 1–2 *Die Phasen von MIS (Phase 01) zu Business Analytics (Phase 06)*

Der Begriffswandel in Business Analytics verspricht einen intensiveren Einsatz von »intelligenten« Datenanalysen, verbunden mit direkten Handlungsempfehlungen, die aus den Analyseergebnissen abgeleitet werden. Dabei wird BI nicht diskreditiert, sondern eher in den Kontext der performanten Informationslieferung und aktiven Analyse gesetzt. Hingegen verspricht Business Analytics eine Aufklärung mittels Algorithmen über bestmögliche zukünftige Handlungen. Womit

bekannte Prognoseverfahren und Optimierungsrechnung (siehe Phase 2) erneut in den Fokus rücken. Die neue Qualität von Business Analytics wird in der sinnvollen Kombination von Methoden der Datenanalyse und Modellen liegen, die vor allem dem Umfeld der Data Science zuzurechnen sind. Die Konvergenz von datenorientierten und modellorientierten Verfahren scheint daher naheliegend und bringt tatsächlich neue Aspekte in die Betrachtung von MUS auf dem Zeitstrahl. Vergleichbar der Phase 2 treten Algorithmen in den Vordergrund, die automatisierte Entscheidungsprozesse ermöglichen, die auf großen polystrukturierten Datenbeständen (Big Data) in Realzeit Empfehlungen für bestmögliche Entscheidungen geben oder selbst entscheiden.

1.2 Data Science und angrenzende Gebiete

In der aktuellen Diskussion rund um die neuen Entwicklungen im Bereich der Informations- und Entscheidungssysteme kann man eine polyphone Stimmenvielfalt feststellen, die so manchen Betrachter verwirrt und manchmal sogar ratlos zurücklässt. Dabei stehen gerade die Abgrenzungen der Begriffe künstliche Intelligenz (aka KI, AI oder Artificial Intelligence), Data Science und Machine Learning im Fokus.

Historisch betrachtet wurde zunächst der Begriff künstliche Intelligenz geschaffen. Im Sommer 1956 fand am Dartmouth College in den Vereinigten Staaten eine von John McCarthy organisierte Konferenz zum Thema »Artificial Intelligence« statt. Im Laufe der nächsten Jahre wurden verschiedene Konzepte im Bereich der KI-Forschung verfolgt und zum Teil heftige Dispute über die Ausrichtung der KI und die zu verwendenden Werkzeuge ausgetragen.[1] Nachdem verschiedene Forschungsansätze auf konzeptionelle, zunächst unüberwindlich erscheinende Probleme gestoßen waren, folgte der sogenannte »AI-Winter« in den 1980er-Jahren. Neue Forschungsansätze (z.B. mehrschichtige neuronale Netze, der Backpropagation-Algorithmus oder rekurrente neuronale Netze), stark verbesserte Technologien in Form von Rechenleistung sowie das aufkommende Big-Data-Phänomen mit der damit einhergehenden Flut an zur Verfügung stehenden Daten führten nicht nur zu einem Revival der KI, sondern dazu, dass KI heute als die wichtigste und möglicherweise entscheidende Kompetenz für die wirtschaftliche Entwicklung eines Landes gesehen wird.[2]

---

1. Erwähnt sei an dieser Stelle nur der bekannte Disput zwischen Marvin Minsky, der an der oben erwähnten Konferenz am Dartmouth College teilgenommen hatte und der über 50 Jahre hinweg das Forschungsgebiet künstliche Intelligenz am MIT vorangebracht hat, und Frank Rosenblatt, der das Konzept des Perzeptrons eingeführt hat.
2. Man betrachte hier nur die massiven Investitionen Chinas in die KI-Forschung oder auch die KI-Initiative der Deutschen Bundesregierung (*https://www.bmbf.de/files/180718%20Eckpunkte_KI-Strategie%20final%20Layout.pdf*).

Das Gebiet künstliche Intelligenz ist extrem facettenreich und stark interdisziplinär geprägt. Hier liegt auch der Grund, warum eine Definition von KI so schwer ist. Nach Winston lässt sich formulieren:

> »Künstliche Intelligenz ist die Untersuchung von Berechnungsverfahren, die es ermöglichen, wahrzunehmen, zu schlussfolgern und zu handeln.«[3]

Damit versucht die KI-Forschung die menschlichen Wahrnehmungs- und Verstandesleistungen zu operationalisieren. Folgt man Görz, Schmid und Wachsmuth [Görz et al. 2013], kann man vereinfacht feststellen, dass es das Ziel der KI ist, Computerprogramme für Problembereiche zu entwickeln, die bislang nur von Menschen lösbar sind. Für sie ist KI als Teil der Informatik eine Ingenieurwissenschaft und als Teil der Kognitionswissenschaft auch Erkenntniswissenschaft. Entsprechend lassen sich zwei Ausprägungen unterscheiden: die starke KI und die schwache KI. Während die starke KI das Ziel hat, menschliche Problemlösungskreativität, Selbstbewusstsein und Emotionen abzubilden, fokussiert die schwache KI auf die Lösung konkreter Anwendungsprobleme durch Simulation von Intelligenz durch Methoden der Informatik, der Statistik und der Mathematik.

Hinsichtlich dieses hohen Maßes an Interdisziplinarität gibt es eine große Überlappung zur Data Science. Der Ursprung dieses noch recht jungen Zweigs wird zeitlich unterschiedlich verortet. Gehen Kelleher und Tierney [Kelleher & Tierney 2018] und andere häufig von Jeff Wus [Wus 1997] gehaltener Vorlesung »Statistics = Data Science?« aus, so führt Cao den Namen auf die Nennung des Begriffs im Vorwort eines 1974 publizierten Buches zu Berechnungsmethoden zurück, in dem es heißt, Data Science sei »the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences« [Cao 2017, S. 3]. Noch weiter zurück geht Donoho, der erste Ansätze bereits Mitte der 1950er-Jahre sieht [Donoho 2015, S. 1]. Bei Donoho findet sich auch die folgende Definition für Data Science:

> »This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary applications.«

3. Winston nach [Görz et al. 2013]

Neben der Interdisziplinarität der Data Science rückt Donoho damit auch die Verknüpfung von wissenschaftlicher Entdeckung und Praxis in den Vordergrund. Die Data Science Association sieht ihre Wissenschaft wie folgt:

> »›Data Science‹ means the scientific study of the creation, validation and transformation of data to create meaning. […] Data science uses scientific principles to get meaning from data and uses machine learning and algorithms to manage and extract actionable, valuable intelligence from large data sets.«[4]

Entsprechend ist der Data Scientist »[…] a professional who uses scientific methods to liberate and create meaning from raw data […] The data scientist has a solid foundation in machine learning, algorithms, modeling, statistics, analytics, math and strong business acumen […].«

Damit wird deutlich, dass Machine Learning oder maschinelles Lernen eine der Methoden ist, die neben zahlreichen anderen in der Data Science zum Einsatz kommt. Maschinelles Lernen ist nach Wrobel, Joachims und Mrozik:

> »[...] ein Forschungsgebiet, das sich mit der computergestützten Modellierung und Realisierung von Lernphänomenen beschäftigt« [Wrobel et al. 2013, S. 406].

Bei den eingesetzten Lernverfahren unterscheidet man das überwachte Lernen (supervised learning), das unüberwachte Lernen (unsupervised learning) sowie das Verstärkungslernen (reinforcement learning). Vielfach kommen hier neuronale Netze zum Einsatz, doch werden je nach Kontext und Fragestellung auch andere Verfahren genutzt. Die Autoren sehen Machine Learning, Data Mining und die »Knowledge Discovery in Databases« (KDD) als Teilgebiete der KI, die in den vergangenen Jahren zunehmend Eingang in praktische Anwendungen in Industrie und Wirtschaft gefunden haben. Die klassische Definition von KDD stammt von Fayyad, Piatetsky-Shapiro und Smyth:

> »Knowledge Discovery in Databases describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data« [Fayyad et al. 1996].

Data Mining ist dabei als der Teilschritt dieses Prozesses zu sehen, der sich mit der Analyse beschäftigt. Im kommerziellen Bereich verschwimmt die Unterscheidung zwischen KDD und Data Mining jedoch häufig.

---

4. *http://www.datascienceassn.org/about-data-science*

Die Entwicklungen rund um Data Science fußen nicht zuletzt auf der enormen Menge an Daten, die Wissenschaftlern, Regierungen und natürlich auch den Unternehmen heute zur Verfügung stehen. Unter dem Schlagwort Big Data wird diese Entwicklung zusammengefasst. Big Data umfasst Methoden und Technologien für die hochskalierbare Integration, Speicherung und Analyse polystrukturierter Daten. Dabei bezieht man sich häufig auf die sogenannten 3Vs (Volume, Velocity und Variaty), die zum Teil durch weitere Vs, wie etwa für Value, ergänzt werden (vgl. [Cai & Zhu 2015, S. 2]). Skalierbarkeit bezieht sich insbesondere auf die in der Regel hohen Datenvolumina (Data Volume), das schnelle Anfallen der Daten und die dafür notwendige hohe Datenverarbeitungs- und analysegeschwindigkeit (Data Velocity) sowie eine breite Quellen- und Datenvielfalt (Data Variety) (vgl. [Dittmar 2016, S. 56 f.]).

1.3 Vorgehen in Data-Science-Projekten

Bei Data-Science-Projekten hat sich ein iteratives, agiles Vorgehen bewährt, das sich in der Regel an dem Vorgehensmodell CRoss-Industry Standard Process for Data Mining, kurz CRISP-DM, orientiert (siehe Abb. 1–3).

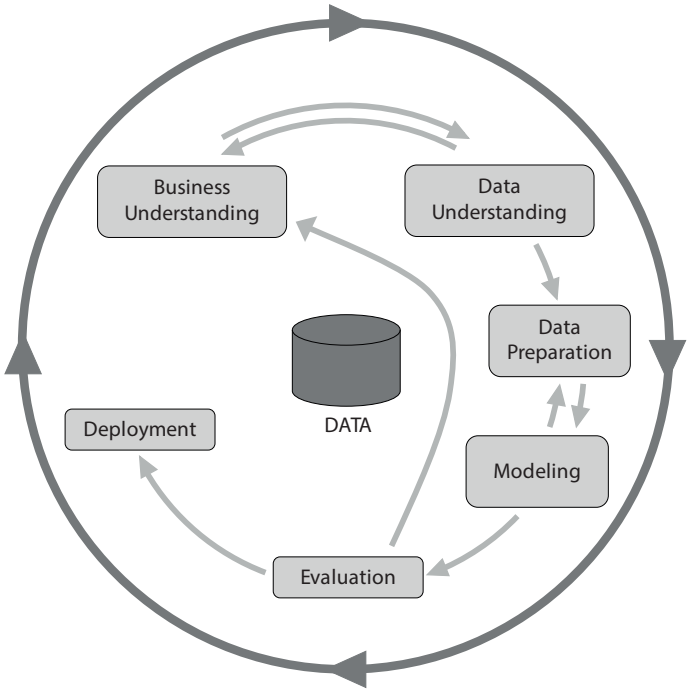


Abb. 1–3 *CRISP-DM mit sechs Phasen*

CRISP-DM besteht aus sechs Phasen, die als zyklischer Prozess zu verstehen sind. Das Business Understanding (fachliches Verständnis) umfasst die Bestimmung der Geschäftsziele, die Beurteilung der aktuellen Situation sowie die konkreten fachlichen Ziele des Data-Science-Projekts und – verbunden damit – die Planung der weiteren Aktivitäten. Im Data Understanding (Verständnis der Daten) werden die Daten und Datenquellen identifiziert, die zur Beantwortung der analytischen Fragestellung notwendig sind. Dieser Schritt enthält auch eine erste Datenerfassung, Datenbeschreibung und die Überprüfung der Datenqualität. Sind die Datenquellen identifiziert und die Daten zusammengestellt, erfolgt eine explorative Datenanalyse, um erste erkennbare Muster zu sichten. Neben der visuellen Analyse und den deskriptiven statistischen Verfahren können auch BI-typische Datenaufbereitungen und -navigationen hilfreich sein, um erste Erkenntnisse über den vorliegenden Datenbestand zu gewinnen. Grundsätzlich folgen solche Analysen einem Prozess, um einen zielorientierten und nachvollziehbaren Ablauf der jeweiligen Datenanalyse zu ermöglichen. Bereits die Business Intelligence liefert hier einen allgemeinen Ablauf, der mit der Datenextraktion, der Transformation und dem Laden in das Data Warehouse beginnt und im weiteren Vorgehen vorab definierte Auswertungen mit einem entsprechenden Analysewerkzeug ermöglicht.

Im Rahmen der Data Preparation (Datenvorbereitung) sind die Daten so aufzubereiten, dass diese im nächsten Schritt für das Training der Modelle verwendet werden können. Modeling (Modellierung) benennt die Parametrisierung und das eigentliche Lernen eines Modells mithilfe von Data-Mining-Algorithmen zur Lösung der Aufgabenstellung. Diese können Regressionsanalyse, Assoziationsanalyse, Klassifikations- oder Clusteranalysen sein. Die Evaluierung erfolgt einerseits bezogen auf die Ergebnisqualität des gelernten Modells und andererseits gegen das Ziel der fachlichen Aufgabenstellung sowie der betriebswirtschaftlichen Bewertung. Die Gewinnung des Geschäftsverständnisses ist ein iteratives Prozedere, in dem die Ergebnisse durch unterschiedliche Algorithmen und Visualisierungen ausgewertet werden, um ein tieferes Verständnis über die erzielten Ergebnisse zu erhalten. Das abschließende Deployment ist die Übertragung der Ergebnisse in die organisationalen Operationen, seien es Vorhersagen zu Marketingaktivitäten oder zu Wartungszyklen der Maschinen in der Fertigung. Zu einem Deployment gehört allerdings auch, dass diese Modelle auf Veränderungen der Betriebsbedingungen zu überwachen sind, da sich Bedingungslagen und Strukturen ändern können, sodass die Gültigkeit von Ergebnissen nicht mehr vorliegt und ein neues Verfahren zu initiieren ist.

Neben CRISP-DM gibt es alternative Ansätze wie beispielsweise der KDD-Prozess nach Fayyad oder SEMMA. Der fayyadsche Ansatz kennzeichnet sich durch die expliziten Phasen Datenauswahl, Datentransformation, Data Mining und die darauffolgende Interpretation (vgl. Abb. 1–4). Implizit wird dabei auch davon ausgegangen, dass Schritte iterativ ausgeführt werden.

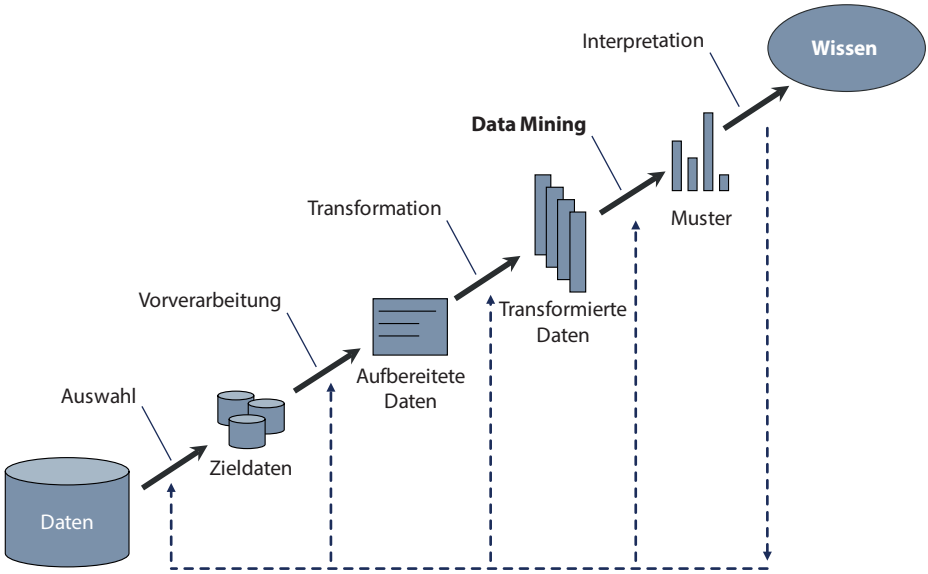


Abb. 1–4 *Überblick über den KDD-Prozess (nach [Fayyad et al. 1996])*

SEMMA, ein früher herstellernaher Ansatz, geht auch phasenorientiert vor, wobei hier von Datenauswahl (Sampling), Datenverständnis (Explore), Modifikation, Algorithmusanwendung (Model) und Ergebnisevaluation (Assess) gesprochen wird.

Die Vorgehensweise ist in fast jedem Data-Science-Projekt iterativ und die Phasen werden mehrmals durchlaufen. Dies bedingt, dass die Nachvollziehbarkeit der einzelnen Schritte wie Datenauswahl, Transformationen etc. und auch das Training in den verschiedenen Phasen ein wesentlicher Punkt ist, der von Projektbeginn an berücksichtigt werden muss. Nur wenn die Nachvollziehbarkeit der Analyse sichergestellt ist, sind eine fundierte Bewertung der Ergebnisse und die Reproduktion der Analyse in der Produktivumgebung und damit das Deployment möglich.

1.4 Struktur des Buches

Das vorliegende Werk ist in einen Grundlagenteil und einem Praxisteil mit Fallstudien gegliedert. Im Grundlagenteil werden verschiedene Aspekte von Data Science erläutert und im zweiten Teil des Buches werden die Grundlagen anhand von konkreten Fallstudien aus Data-Science-Projekten mit deren spezifischen praktischen Problemstellungen und Lösungsansätzen dargestellt. Die Projektberichte nehmen Bezug auf die Grundlagen des ersten Teils, sind in sich jedoch geschlossen und können in einer frei wählbaren Reihenfolge gelesen werden.

In Kapitel 2 diskutiert Uwe Haneke, ob Analytics wirklich das neue BI ist und welche Erkenntnisse die Unternehmen daraus ziehen können. Er geht der Frage nach, warum sich Data Science gerade jetzt so rasant verbreitet und in den Unternehmen Fuß fasst. Im Anschluss wird erläutert, warum dieser Entwicklung eine so große Bedeutung zukommt und wie sich eine mögliche Fusion der alten BI-Welt mit der neuen, erweiterten Analytics-Welt in den Informationssystemen der Unternehmen darstellen könnte.

In Kapitel 3 zeigen die Autoren Marc Beierschoder, Benjamin Diemann und Michael Zimmer anhand eines konkreten Beispiels, unter welchen Rahmenbedingungen der Einsatz von Data Science im Allgemeinen und KI im Speziellen zum Erfolg in einem Unternehmen führen kann.

Anschließend stellt Christoph Tempich in Kapitel 4 vor, wie die Konzeption und die Entwicklung von Data-driven Products erfolgen kann und auf welche Punkte dabei geachtet werden muss. Unter anderem werden die Aspekte Ideenfindung, Value Propostion Design und Zielgrößen näher untersucht und die Messung der Qualität eines Datenprodukts mithilfe einer Feedbackschleife vorgeschlagen.

In Kapitel 5 stellen Stephan Trahasch und Carsten Felden im Überblick grundlegende Methoden der Data Science vor, die in den Phasen Data Unterstanding, Data Preparation, Modeling und Evaluation Verwendung finden.

Angesichts weiter zunehmender zur Verfügung stehender Daten kommt der Feature Selection eine immer größere Bedeutung zu. Diesem wichtigen Aspekt wird in Kapitel 6 von Bianca Huber Rechnung getragen.

Klaus Dorer führt in Kapitel 7 in die Grundlagen neuronaler Netzwerke ein und erläutert anhand von Deep Convolutional Neural Networks für die Objekterkennung in Bildern, wie Deep Learning funktioniert. Neben einigen praktischen Anwendungen gibt das Kapitel auch einen Überblick über die zahlreichen verfügbaren Frameworks und Standarddatensätze für Deep Learning.

Nur mit geeigneten Datenarchitekturen als Grundlage können Unternehmen zukünftig Data Science und Artificial-Intelligence-basierte Anwendungsfälle abbilden. Wie solch eine Datenarchitektur aussehen kann, erläutern Michael Zimmer, Benjamin Diemann und Andreas Holzhammer in Kapitel 8.

Self-Service und Befähigung der Anwender sind in der BI ein aktuelles Thema. In Kapitel 9 stellen Uwe Haneke und Michael Zimmer vor, warum gerade Self-Service-Szenarien in Data Science wichtig sind, um im Unternehmen die analytische Sichtweise zu verankern. Daneben stellen die Autoren ein Konzept für eine differenzierte Data & Analytics Governance vor, da das Thema Governance im Data-Science-Umfeld immer mehr an Bedeutung gewinnt.

In Kapitel 10 diskutieren Victoria Kayser und Damir Zubovic die Rolle von Data Privacy für Analytics und Big Data. Neben der rechtlichen und technischen Ausgestaltung von Data Privacy im Unternehmen diskutieren die Autoren auch, wie die Unternehmen mit der Herausforderung umgehen, Analytics und KI in ihre Prozesse zu integrieren.

Anschließend führen Matthias Haun und Pfarrer Gernot Meier in Kapitel 11 ein Gespräch zur digitalen Ethik, geben einen Einblick in die Vielgestaltigkeit der Diskussion und zeigen auf, welche Fragestellungen auf uns zukommen.

Mit Kapitel 12 beginnt der Praxisteil des Buches. In der ersten Fallstudie stellt Shirin Glander dar, wie mit Methoden der Data Science Vorhersagen zum Churn-Verhalten von Kunden getroffen werden können. Zur prädiktiven Analyse wird ein neuronales Netz mit Keras und TensorFlow trainiert und dies mit einem Stacked-Ensemble-Modell auf Basis von H20 verglichen.

In Kapitel 13 gibt Nicolas March einen Einblick in die Erfahrungen mit Data Science und in die Wirtschaftlichkeitsbetrachtungen bei der Auswahl und Entwicklung von Data-Science-Anwendungen im Online-Lebensmittelhandel.

Mikio Braun stellt in Kapitel 14 vor, wie Zalando die Grundlagen für Analytics, BI und Data Science zum unternehmensweiten Einsatz geschaffen hat und welche Herausforderungen das Unternehmen zu meistern hatte.

Predictive Maintenance hat für die industrielle Produktion eine großes Potenzial. Marco Huber erläutert in Kapitel 15 die verschiedenen Strategien der Instandhaltung und wie die Nutzung von unterschiedlichen Daten, die während der Produktion anfallen, für die vorausschauende Instandhaltung unter Einsatz von Verfahren der Statistik und des maschinellen Lernens erfolgen kann.

Caroline Kleist und Olaf Pier beschreiben in Kapitel 16, wie Scrum in Data-Science-Projekten bei der Volkswagen Financial Services AG erfolgreich eingesetzt wird und mit welchen Herausforderungen sie konfrontiert wurden, und geben Empfehlungen zum Einsatz von Scrum für Data-Science-Teams.

In Kapitel 17 zeigt Matthias Meyer, wie durch die Konzeption und Pilotierung zusätzlicher datenbasierter Serviceangebote für einen Betreiber eines Kundenkartenprogramms ein Mehrwert geschaffen werden konnte.

Abschließend beschäftigt sich Kapitel 18 mit dem Einsatz von KI und Data Science in der Versicherungsbranche. Am Beispiel der Zurich Versicherung zeigen die Autoren anschaulich, vor welchen Herausforderungen das Unternehmen stand und wie KI im Wertschöpfungsprozess heute in verschiedenen Anwendungsfällen in der Versicherung zum Einsatz kommt.

2 (Advanced) Analytics is the new BI?

Uwe Haneke

Die Analytics-Welle, die derzeit durch Unternehmen rollt, erinnert zuweilen an die 1990er-Jahre, in denen das Data Warehouse und Business Intelligence Eingang in die Informations- und Steuerungssysteme fanden. Im folgenden Beitrag wird diskutiert, ob Analytics wirklich das neue BI, also der nächste konsequente und folgerichtige Schritt ist, und welche Schlüsse die Unternehmen aus dieser Erkenntnis ziehen können. Zunächst werden die Parallelen beim Aufkommen der beiden Konzepte dargestellt, bevor der Frage nachgegangen wird, warum sich Data Science, manchmal auch als Advanced Analytics bezeichnet, und Analytics gerade jetzt so rasant verbreiten und in den Unternehmen Fuß fassen. Im Anschluss wird erläutert, warum dieser Entwicklung eine so große Bedeutung zukommt und wie sich eine mögliche Fusion der alten BI-Welt mit der neuen, erweiterten Analytics-Welt in den Informationssystemen der Unternehmen darstellen könnte.

2.1 Geschichte wiederholt sich?

Die aktuelle Entwicklung, die seit einigen Jahren in den Unternehmen zu beobachten ist, erinnert zuweilen an die Anfänge des Data Warehousing in der ersten Hälfte der 1990er-Jahre. Um die Parallelen aufzuzeigen und in einem zweiten Schritt auch Schlüsse für die heutige Situation ziehen zu können, sollen kurz die Herausforderungen und Rahmenbedingungen betrachtet werden, denen die Unternehmen damals gegenüberstanden. Dies betrifft nicht nur die fachlichen und technologischen Aspekte, sondern darüber hinaus auch Fragen der Organisation. Bereits Hans Peter Luhn, der lange vor Howard Dresner den Begriff Business Intelligence prägte, hatte erkannt, dass ein solches Informationssystem nur im Einklang mit entsprechenden organisatorischen Regelungen effizient genutzt werden kann [Luhn 1958].

Als der Data-Warehouse-Gedanke, vor allem getrieben durch die Arbeiten von Kimball und Inmon in den frühen 1990er-Jahren, seinen Siegeszug in der Welt der Unternehmen antrat, sorgte dies für eine grundlegend neue Qualität der betrieblichen Informationssysteme. Bis dato dominierten die sogenannten OLTP-Systeme, deren Hauptaugenmerk in der effizienten Unterstützung von betrieb-

lichen Geschäftsprozessen lag. Waren zunächst in der Regel Insellösungen für die verschiedenen Fachabteilungen zu finden, traten Anfang der 1990er-Jahre verstärkt integrierte Standardsoftwarelösungen auf Client-Server-Basis, allen voran SAPs R/3, auf den Plan. Die neuen ERP-Systeme waren in der Lage, Geschäftsprozesse end-to-end auf einer Plattform abzubilden. Da der Fokus auf der effizienten Unterstützung der Prozesse lag, zeigten die OLTP-Lösungen häufig Schwächen im Bereich des Reportings. Diese Schwächen betrafen unter anderem Zeitreihenanalysen, die Verknüpfung von Daten aus unterschiedlichen OLTP-Anwendungen oder Fachdomänen und die Performance. Data Warehousing und OLAP sollten diese Schwächen nachhaltig überwinden.

Die Idee einer Entkopplung des Informationssystems von den operativen Systemen verbunden mit den neuen Konzepten für die Datenmodellierung und ihrem Fokus auf die Anforderungen der Informationsnachfrager führte letztlich dazu, dass mit dem Data Warehouse vieles von dem umgesetzt werden konnte, was konzeptionell schon lange an- und vorgedacht worden war. Bereits seit den 1960er-Jahren waren immer wieder entsprechende Ideen entwickelt worden, die jedoch zumeist an den technologischen Voraussetzungen scheiterten. Eine interessante historische Übersicht zur Entwicklung von Entscheidungsunterstützungssystemen, die zeigt, wie vielschichtig die Entwicklungen in den letzten 50 Jahren waren, findet sich bei Power [Power 2007]. In ihrem Standardwerk zu Data-Warehouse-Systemen schreiben Bauer und Günzel [Bauer & Günzel 2013] auch entsprechend:

> *»Was sich im Laufe der MIS-Bemühungen als Utopie abzeichnete […] erhält durch den Fortschritt in der Informationstechnologie im Gewand des Data Warehousing eine Renaissance.«*

Die neuen OLAP-Systeme setzten sich nach und nach durch, wobei im Folgenden verschiedene Aspekte vor allem bei ihrer Einführung angesprochen werden sollen, die offensichtliche Parallelen zu heute aufweisen.

Gut Ding will Weile haben

Sowohl BI als auch Data Science benötigten einen langen Atem, bevor sie letztlich Eingang in die Unternehmen fanden und sich dort etablierten. Im Fall von BI wurde gerade schon dargestellt, dass es ein langer Weg mit zahlreichen unterschiedlichen Konzepten war, bevor sich BI tatsächlich als wichtiges Werkzeug der Unternehmenssteuerung etablieren konnte. Data Science wiederum vereint unterschiedliche Ansätze und Konzepte, die ebenfalls über Jahrzehnte hinweg diskutiert und entwickelt wurden, sich jedoch bis dato nicht flächendeckend hatten durchsetzen können. Erst seit Mitte/Ende der 2000er-Jahre hat sich Data Science rasant verbreitet und ist auf dem Weg, für viele Unternehmen zu einem wichtigen Baustein der Unternehmenssteuerung zu werden. Der Begriff Data Science taucht, folgt man Kelleher und Tierney, 1997 zum ersten Mal in einer Vorlesung von Jeff Wu

mit dem Titel »Statistics = Data Science« auf. Die Erweiterung des Fokus über die Statistik hinaus in Richtung Machine Learning und das Aufkommen von Big Data hat aber letztlich erst zu dem Verständnis von Data Science geführt, wie man es heute kennt [Kelleher & Tierney 2018, S. 17 ff.].

Data Science ist dabei so vielschichtig und facettenreich, dass auch die Anforderungen an einen Data Scientist kaum durch eine Person allein abdeckbar zu sein scheinen. Von Machine Learning über Storytelling und Datenbanken gehen diese Anforderungen bis hin zu Domänen-Know-how. Daneben muss der Data Scientist selbstverständlich programmieren können, sich mit NoSQL und verteilten Systemen auskennen und sehr gute Kenntnisse in den Bereichen Statistik und Wahrscheinlichkeitsrechnung aufweisen. Das »skill-set desideratum« für einen Data Scientist ist in Abbildung 2–1 dargestellt.

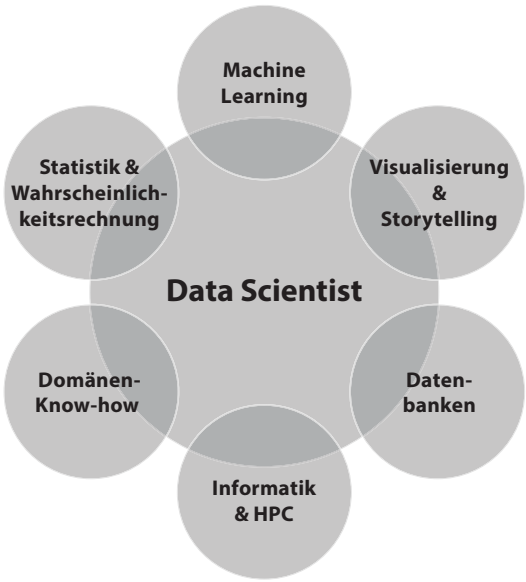


Abb. 2–1 *Das »skill set desideratum« für einen Data Scientist*

Angesichts dieses Profils galt die Suche nach geeignetem und qualifiziertem Personal, um die Data Science im Unternehmen aufzubauen, daher lange als limitierender Faktor. Doch erst durch die Verknüpfung der unterschiedlichen Aspekte, Disziplinen und Kompetenzen der hier zusammenkommenden Teilbereiche hat es Data Science geschafft, den gewünschten und erhofften Mehrwert in den Unternehmen zu erbringen. Für nicht wenige Unternehmen und Geschäftsideen bedeutete Data Science und die daraus gewonnenen Erkenntnisse einen Quantensprung in der Unternehmenssteuerung mit der Möglichkeit, neue Services und Produkte erfolgreich am Markt zu platzieren.[1]

1. Interessante Beispiele hierzu finden sich bei [Mcafee & Brynjolfsson 2018].

Die Technologie muss bereitstehen

Warum aber gerade jetzt? Was hat sich im Vergleich zum Ende der 1990er-Jahre verändert? In Bereichen wie Machine Learning oder Data Mining, abgesehen von der Statistik, wurde seit Jahrzehnten geforscht und gearbeitet, ohne jemals diese Durchschlagskraft zu erreichen. Viele Autoren sind sich einig, dass es zum einen der technologischen Entwicklung geschuldet ist, die das Durchführen komplexer Rechenoperationen in Clustern auf sogenannter »commodity hardware« oder mittlerweile auch in der Cloud für eine breite Masse an Unternehmen ermöglicht hat. Die notwendige Software steht in vielen Fällen als Open Source zur Verfügung, sodass die Unternehmen nicht nur erste Schritte ohne großen Aufwand machen können. Auch die Skalierbarkeit ist durch das Cluster sichergestellt.

Den zweiten wichtigen Faktor stellt sicherlich Big Data dar. Auch wenn man für ein Data-Science-Projekt nicht notwendigerweise Big Data benötigt[2], stellt die Tatsache, dass wir heute über einen enormen Fundus an Daten verfügen, einen wichtigen Faktor für den Erfolg und die Verbreitung von Data Science in der Wirtschaft dar. Ob es sich um Sensordaten, um Logfiles, um Daten aus dem eigenen ERP-System oder um Open Data handelt: Unternehmen verfügen heute über einen sehr großen Datenpool, mit dem sie arbeiten können.

Neben den Ideen und Konzepten müssen auch die geeigneten Technologien vorhanden sein: Wie oben für BI gezeigt, machten es erst die technologischen Fortschritte möglich, die zuvor entwickelten Ideen und Utopien tatsächlich umzusetzen. Ähnliches erleben wir heute im Bereich Data Science. Damals wie heute sind es die technischen Fortschritte, die lang erarbeitete Ideen und Konzepte endlich auch realisierbar machen.[3]

»Garbage in, garbage out«

Im Data Warehousing war der ETL-Prozess lange Zeit ein unterschätzter Faktor. Dabei kommt gerade diesem Teil des Data-Warehouse-Prozesses aus verschiedenen Gründen eine zentrale Rolle zu. Zum einen ist die Auswahl geeigneter Datenquellen von entscheidender Bedeutung. Nur auf der Basis qualitativ hochwertiger Daten kann auch ein qualitativ hochwertiges Ergebnis im Rahmen der bereitgestellten Analysen erwartet werden. Wird dies von den Entwicklern zu wenig beachtet, können die am Ende zur Verfügung gestellten Berichte noch so schön sein, es gilt weiterhin die altbewährte Erkenntnis: »Garbage in, garbage out.«[4]

---

2. Die Definition des Begriffs »Big Data« an sich gestaltet sich bereits schwierig. Mit einem Augenzwinkern sei an dieser Stelle auf David Taylors Blogeintrag »Battle of the Data Science Venn Diagrams« verwiesen [Taylor 2016].
3. Wie weit diese auch in anderen Unternehmensbereichen wirkenden Veränderungen gehen können und mit welcher Geschwindigkeit sie ablaufen, verdeutlicht Pratt. Er spricht im Bereich der Robotik sogar von einer kambrischen Explosion [Pratt 2015]. Diese wird nach Mcafee und Brynjolfsson angetrieben von DANCE (Daten, Algorithmen, Netzwerken, der Cloud und den exponentiellen Verbesserungen der Hardware) [Mcafee & Brynjolfsson 2018, S. 114 ff.].
4. Zur Bedeutung der Datenqualität im Data-Warehouse-Prozess vgl. [Bauer & Günzel 2013].

Zum anderen hat sich immer wieder gezeigt, dass der *Workload*, der mit dem ETL-Prozess verbunden ist, tendenziell unterschätzt wird. Auch wenn mittlerweile mächtige Tools für die Datenbereitstellung genutzt werden, ist vor allem die Bearbeitung von Daten mit Qualitätsmängeln nach wie vor aufwendig.

Gerade diesem Phänomen begegnet man auch im Zusammenhang mit Data Science wieder. Auch hier ist zu beobachten, dass die grundlegende Bedeutung der Data Preparation zu Beginn eines Projekts oder wenn ein Unternehmen plant, Data Science einzuführen, nicht erkannt und oftmals der damit zusammenhängende Arbeitsaufwand unterschätzt wird. Statistiken zufolge verwenden Data Scientists in der Praxis bis zu 80 % ihrer Zeit für das Vorbereiten der Daten, also das Sammeln, Bereinigen und Organisieren der Daten. Kelleher und Tierney stellen dazu treffenderweise fest:

> *»But the simple truth is that no matter how good your data analysis is, it won't identify useful patterns unless it is applied to the right data.«*
>
> [Kelleher & Tierney 2018, S. 67]

Auf die Möglichkeiten, wie man der Data Science die notwendigen Daten oder Datenzugriffe im Unternehmen ermöglicht, wird später in Kapitel 9 näher eingegangen.

Don't be too fast

Benutzerfreundliche Tools mit grafischen Oberflächen ermöglichen es heute auch Einsteigern, relativ schnell erste Erfahrungen im Bereich Data Science zu sammeln und Modelle zu erstellen. Dieser leichte Zugang zu den Möglichkeiten der Data Science ist Segen und Fluch zugleich. Einerseits werden Berührungsängste mit der durchaus komplexen neuen Materie für viele potenzielle Nutzer abgebaut. Andererseits benötigt man eine hohe Fachkompetenz, um den richtigen Algorithmus für den jeweiligen Use Case auszuwählen, die Ergebnisse zu interpretieren und das geeignete Data Set zu erstellen. Die Feststellung »In fact, it has never been easier to do data science badly«[5] ist daher ohne Zweifel richtig. Ein Modell zu erstellen ist mit den heutigen Werkzeugen nicht schwer. Schwierig hingegen ist es, die Güte des Modells zu bewerten und Verbesserungspotenziale zu erkennen.

Ähnliches kennt man aus der Business Intelligence. Die Kunst, die Daten so abzulegen, dass auch bei einem zunehmenden Datenbestand immer noch performant die Anfragen abgearbeitet werden können und dabei die Informationsbedürfnisse der Nutzer befriedigt werden, ist wichtiger als ein schönes buntes Dashboard, das den Datenzugriff erleichtert. Gerade angesichts der, wie sich herausstellte, mangelhaften Agilität der klassischen Modellierungskonzepte war es umso wichtiger, ein nachhaltig tragfähiges Modell zu entwickeln und nicht in einem ersten Wurf einfach ein paar Datenwürfel für Pilotanwender bereitzustellen. In

5. Vgl. [Kelleher & Tierney 2018, S. 36].