DE GRUYTER
OLDENBOURG

PRAXISHANDBUCH

Andreas Wierse, Till Riedel

SMART DATA ANALYTICS

ZUSAMMENHÄNGE ERKENNEN
POTENTIALE NUTZEN
BIG DATA VERSTEHEN

DE G

Andreas Wierse, Till Riedel

Smart Data Analytics

Zusammenhänge erkennen
Potentiale nutzen
Big Data verstehen

DE GRUYTER
OLDENBOURG

Autoren:
Dr. Andreas Wierse
SICOS BW GmbH
Nobelstr. 19
70569 Stuttgart
autoren@smart-data-analytics.de

Dr. Till Riedel
Karlsruher Institut für Technologie
TECO
Vincenz-Prießnitz-Str. 1
76131 Karlsruhe
autoren@smart-data-analytics.de

ISBN 978-3-11-046184-8
e-ISBN (PDF) 978-3-11-046395-8
e-ISBN (EPUB) 978-3-11-046191-6
Set-ISBN 978-3-11-046396-5

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Einbandabbildung: John Lund/Blend Images/gettyimages
Druck und Bindung: CPI books GmbH, Leck
♾Printed on acid-free paper
Printed in Germany

www.degruyter.com

Andreas Wierse, Till Riedel
Smart Data Analytics

Weitere empfehlenswerte Titel

Data Mining, 2. Auflage
J. Cleve, U. Lämmel, 2016
ISBN 978-3-11-045675-2, e-ISBN 978-3-11-045677-6,
e-ISBN (EPUB) 978-3-11-045690-5

Vernetzte Organisation
A. Richter (Hrsg.), 2014
ISBN 978-3-486-74728-7, e-ISBN 978-3-486-74731-7,
e-ISBN (EPUB) 978-3-486-98956-4, Set-ISBN 978-3-486-98957-1

Industrial Software Applications
R. Geisler, 2015
ISBN 978-3-11-037098-0, e-ISBN 978-3-11-037099-7,
e-ISBN (EPUB) 978-3-11-039678-2

Advanced Data Management
L. Wiese, 2016
ISBN 978-3-11-044140-6, e-ISBN 978-3-11-044141-3,
e-ISBN (EPUB) 978-3-11-043307-4

Vorwort der Autoren

Liebe Leserinnen und Leser,

als wir Anfang 2016 mit dem Verlag über dieses Buch sprachen, lautete sein Arbeitstitel „Big Data Praxishandbuch". Schon im Sommer desselben Jahres wurde daraus unter Berücksichtigung des „Hype-Zyklus" ein „Smart Data Praxishandbuch". Und im Herbst schlugen wir dem Verlag dann den endgültigen Titel „Smart Data Analytics Praxishandbuch" vor. Warum thematisieren wir das im Vorwort? Die Informationstechnologie ist doch bekannt für ihre *Buzzwords* (das Leo Online-Wörterbuch übersetzt das nicht nur mit „Modewort" sonder auch mit „leeres Schlagwort" oder „abgedroschene Phrase"!).

Big Data

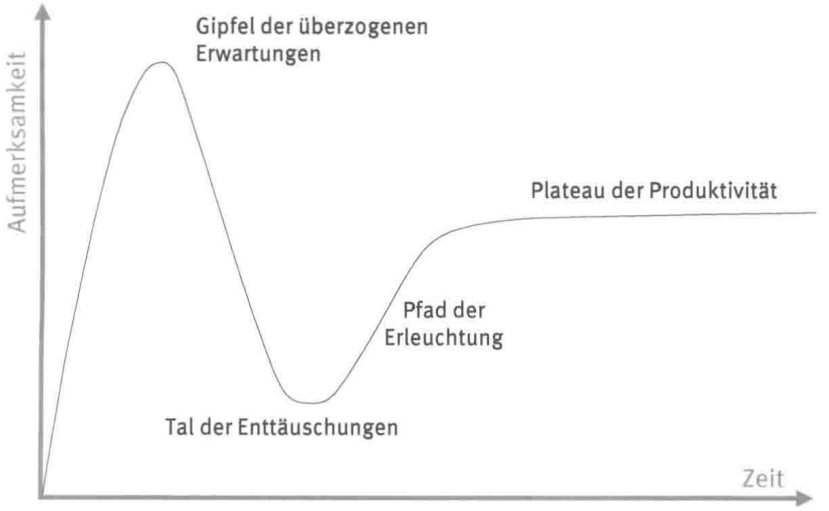

Abb. 1: Der Hype-Zyklus nach Gartner

Nun, selbst in der mit kurzen Hype-Zyklen wahrlich nicht unterversorgten Informationstechnologie ist der Hype, oder besser auf Deutsch: Medienrummel rund um die großen Datenmengen und was man mit ihnen anfangen kann, schon ein besonderer. Das gilt insbesondere vor dem Hintergrund, dass Datenmengen, die in Relation zu den jeweiligen technischen Möglichkeiten groß sind, schon immer von Interesse waren und auch mit den zur Verfügung stehenden Werkzeugen bearbeitet wurden. Es handelt sich also eigentlich um etwas, das im Kern gar nicht neu ist. Und schaut

Hype-Zyklus

zu beschreiben als auch anzudeuten, wo die Reise hingehen wird, soweit das auf dem aktuellen Informationsstand überhaupt möglich ist. Besonders dankbar sind wir, dass ihr das nicht nur gelungen ist, sondern dass sie es auch in einer Art und Weise geschafft hat, die auch dem Nichtjuristen verständlich ist, ohne das komplexe Thema unzulässig zu vereinfachen.

Korrekturlesen

Ebenfalls sehr wichtig für die Qualität dieses Buches sind die kritischen Korrekturleser. Allen voran Markus Klietmann, der uns mit seiner langjährigen einschlägigen IT-Verlagserfahrung zielsicher auf Schwachstellen, sowohl im Sprachlichen, aber insbesondere auch im logischen Aufbau hingewiesen hat. Wir sind sicher, dass seine Kommentare und Vorschläge die Lesbarkeit dieses Buches ganz wesentlich verbessert haben; sein kritischer Blick hat uns mehr als einmal dabei geholfen, Sie als Leser im Auge zu behalten und uns nicht in den Tiefen der IT-Nomenklatur zu verlaufen.

Unterstützt haben uns dabei auch unsere Frauen, Monika und Anna. Deutlich weniger IT-affin als wir Autoren haben sie uns immer wieder auf den Boden der Tatsachen geholt und in der ihnen eigenen Art einen distanzierten Blick auf unseren Text geworfen. Mindestens ebenso wichtig waren allerdings auch ihre Ermunterungen und ihr Verständnis dafür, dass im Laufe des Schreibens immer mehr Zeit in dieses Buch geflossen ist.

Dank gebührt auch unseren Ansprechpartnern beim Verlag, Leonardo Milla und Nancy Christ, die immer schnell, kompetent und hilfreich reagiert haben und vor allem sehr geduldig mit uns waren.

Unterstützung

Abschließend möchten wir all jenen danken, mit denen wir überhaupt erst die Erfahrungen gesammelt haben, die wir in dieses Buch einfließen lassen konnten. Das sind zunächst unsere Partner im Smart Data Solution Center Baden-Württemberg: Prof. Michael Beigl, Prof. Bernhard Neumair, Dr. Nico Schlitter, Andreas Meier und alle weiteren Projektkollegen. Aber auch Peter Castellaz und Dr. Katrin Behaghel vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, ohne deren Unterstützung es das SDSC-BW in der heutigen Form nicht gäbe.

Die wichtigsten Partner, die wir an dieser Stelle aber nicht namentlich aufführen möchten, sind die Mitarbeiter der Unternehmen, mit denen wir im Rahmen des SDSC-BW in den Potentialanalysen zusammenarbeiten durften. Gerade diese Zusammenarbeit hat uns bei vielen Gelegenheiten deutlich gemacht, wie Forschung und Anwendung voneinander profitieren können und worauf es beim praktischen Einsatz der *Smart Data Analyitcs* wirklich ankommt.

Andreas Wierse und Till Riedel
Leonberg und Karlsruhe im Juni 2017

Inhalt

Vorwort der Autoren — V

1 Einleitung — 1
1.1 Ein motivierendes Beispiel — 1
1.2 Für wen ist dieses Buch und wie kann man es lesen? — 8
1.3 Smart Data Solutions statt Big Data — 10
1.4 Das Smart Data Solution Center Baden-Württemberg — 14
1.4.1 Warum ein Smart Data Solution Center? — 15
1.4.2 Ablauf einer Potentialanalyse — 16
1.4.3 Drei Beispiele — 17
1.4.4 Die Partner — 21
1.4.5 Das Smart Data Innovation Lab — 23

2 Grundlagen — 25
2.1 Smart Data vs. Big Data — 25
2.1.1 Die 3Vs: Volume, Velocity, Variety — 26
2.1.2 Veracity, Validity, Value — 27
2.1.3 Variability, Venue, Vocabulary — 29
2.1.4 Das verbliebene V: Vagueness — 30
2.1.5 Smart Data — 31
2.2 Datengetriebene Innovation — 33
2.2.1 Business Intelligence und Verbesserungsprozesse — 35
2.2.2 Operative Geschäftsdaten für Innovation nutzen — 36
2.2.3 Vom eingebetteten System zum Datensee — 37
2.2.4 Kontextsensitive Systeme — 39
2.3 Data Analytics und Maschinelles Lernen — 46
2.3.1 Business Analytics — 49
2.3.2 Klassifikation eines Merkmalsraums — 50
2.3.3 Supervised Learning — 53
2.3.4 Prädiktion und prädiktive Analyse — 57
2.4 Die Bewertung von Vorhersagen — 60
2.4.1 Fehlermaße als Bewertungsfunktion — 60
2.4.2 Validierungsschema — 66
2.4.3 Automatische Verbesserung von Klassifikatoren — 73
2.5 Merkmale und Datentypen — 77
2.5.1 Automatische Merkmalsselektion und -bewertung — 80
2.5.2 Lernen von Merkmalen — 83

8.1.5 Datenschnittstellen — **374**
8.1.6 Datenaufbereitung — **374**
8.1.7 Prozessanbindung — **377**
8.1.8 Mitarbeiter — **378**
8.1.9 Mitarbeiterschulung/-weiterbildung — **381**
8.1.10 Unterstützung durch Dienstleister — **382**
8.2 Cloud vs. On-Premise — 383
8.2.1 Wesentliche Charakteristika — **383**
8.2.2 Service-Modelle — **385**
8.2.3 Einsatzmodelle — **388**
8.2.4 Abwägung: Cloud vs. On-Premise — **395**
8.3 Return on Investment — 400
8.3.1 Das Problem der Skalierung — **401**
8.3.2 Vorgehensweise — **403**
8.3.3 Von anderen lernen — **404**

9 Epilog — 407

Stichwortverzeichnis — 423

1 Einleitung

In diesem Einleitungskapitel möchten wir Ihnen Lust auf die Smart Data Analytics *machen. Ein bisschen Historie, ein interessantes Beispiel, eine kleine Anleitung für das Buch, bevor wir Ihnen ein wenig unseren Hintergrund und die Quelle unserer Erfahrung sowie vieler Informationen in diesem Buch vorstellen.*

1.1 Ein motivierendes Beispiel

Der Begriff *„Big Data"* ist erst seit wenigen Jahren in aller Munde; fast scheint es so, als ginge es bei den (wörtlich übersetzt) großen Daten um ein ganz neues Thema. Das offenbar unaufhaltsame Wachstum der Festplattenkapazitäten, die kontinuierliche Beschleunigung der Datenübertragung und die Allgegenwart des Internets lassen den Eindruck entstehen, dass große Datenmengen etwas ganz Neues sind; ein Problem, das die Generationen vor uns noch gar nicht kannten.

Big Data – ein alter Hut

Aber das stimmt so nicht. Der Umgang mit großen Datenmengen ist nicht erst seit der Verfügbarkeit von Terabyte-Festplatten eine Herausforderung. Lassen Sie uns dazu Matthew Fontaine Maury betrachten, einen Seeoffizier der US-Navy, der im 19. Jahrhundert lebte (siehe Abbildung 1.1). Er diente in den dreißiger Jahren als Seeoffizier, brach sich aber bei einem Sturz die rechte Hüfte und trug eine Knieverletzung davon, die nicht mehr richtig heilte. Aus diesem Grund musste er seinen aktiven Dienst zur See beenden, konnte allerdings bei der Navy bleiben.

Als Matrose hatte er beobachtet, dass Überseekapitäne ihre Beobachtungen über Wetterverhältnisse, gegenläufige Winde, Strömungen und andere Besonderheiten von Wetter und Seegang in ihren Logbüchern verzeichneten. Allerdings wurden diese Informationen praktisch von niemand anderem wahrgenommen und gerieten in Vergessenheit. Als Maury 1842 Direktor des Archivs der Seekarten wurde, fand er dort Unmengen alter Logbücher und Seekarten, die bis ins 18. Jahrhundert zurückreichten und von der Navy zwar nicht entsorgt, aber ohne weitere Verwendung abseits gelagert worden waren. Diese Logbücher nahm er sich vor und untersuchte sie ausführlich.

Bereits im Jahr 1843 waren die von ihm daraus gewonnen Erkenntnisse so erhellend, dass er einen Artikel schrieb mit dem Titel *„Blank Charts on Board Public Cruisers"* (übersetzt in etwa „Unbeschriebene Karten an Bord öffentlicher Kreuzer"). Er schlug vor, dass die unbeschriebenen Seekarten mit Längen- und Breitengraden ausgestattet werden sollten und die Kapi-

unbeschriebene Seekarten zur Datenerfassung

DOI 10.1515/9783110463958-001

täne auf ihren Fahrten dort nicht nur den täglich zurückgelegten Weg einzeichnen, sondern auch alle weiteren Informationen, die für die Navigation in der befahrenen Route von seefahrerischer Bedeutung waren: Windstärken und -richtungen, Strömungen und mehr. Er machte deutlich, dass kurze Reisezeiten auf See nicht einfach nur auf Glück zurückzuführen sind, sondern dass mithilfe dieser Information der Steuermann jederzeit den besten Weg finden könne.

Abb. 1.1: Lt. Matthew Fontaine Maury, Quelle: wikipedia[1]

Im Jahr 1847 veröffentlicht er die „*Wind and Current Chart of the North Atlantic*", also Karten für Wind und Strömung im Nordatlantik. Diese Veröffentlichung erlaubte es den Kapitänen und Steuerleuten, ihren Weg erheblich besser an die jeweils herrschenden Wind- und Strömungsbedingungen anzupassen und führte zu signifikanten Verkürzungen der Reisezeiten.

ein früher Daten-Analyst

Lieutenant Maury war der Archetyp des Daten-Analysten im 19. Jahrhundert, allerdings unter ganz anderen Bedingungen, als wir sie heute kennen. Er hat sich praktisch durch die gesamten verfügbaren Daten (Logbücher) gewühlt und die darin enthaltenen Informationen sortiert und klas-

1 https://de.wikipedia.org/wiki/Matthew_Fontaine_Maury

sifiziert. Er hat Muster gesucht und gefunden, diese in Beziehung zu Reisezeiten und Orten gesetzt und daraus übergreifende Strukturen abgeleitet. Er hat Regeln definiert, wie diese Datenbasis verbessert werden kann, hat die Art der Information definiert, die dafür benötigt wird (Datenschnittstellen). Und aus diesem Prozess sind Erkenntnisse gewonnen worden, die sich ganz deutlich ökonomisch positiv auswirkten, bis heute.

Zweifellos spielen Daten in der Menschheitsgeschichte schon lange eine sehr wichtige Rolle, man denke nur an die Bibliothek in Alexandria im zweiten vorchristlichen Jahrhundert mit geschätzt rund einer halben Million Schriftrollen. Im Jahr 1944 machte sich Fremont Rider, ein Bibliothekar der Wesleyan University, Gedanken über das Wachstum amerikanischer Bibliotheken: er schätzte, dass sich ihr Umfang etwa alle 16 Jahre verdoppeln würde und dass die Yale Bibliothek im Jahr 2040 etwa 200 Millionen Bände haben müsse, die auf rund 10 Millionen Regalmetern stünden und von 6.000 Bibliotheksmitarbeitern betreut werden müssten. Das sind selbst aus heutiger Perspektive beeindruckende Zahlen. 1961 schätzte Derek Price, dass sich die Zahl der wissenschaftlichen Veröffentlichungen alle 15 Jahre verdoppelt, in einem halben Jahrhundert verzehnfacht; er bezieht sich explizit auf das exponentielle Wachstum des Wissens[2].

exponentielles Wachstum des Wissens

Nun ist es allerdings so, dass das Stapeln von Büchern oder wissenschaftlichen Veröffentlichungen bei diesen Dimensionen zwar eine körperlich anstrengende Arbeit sein dürfte, das eigentliche Problem aber ganz woanders liegt: wie finde ich das, was ich suche, am schnellsten? Der Hinweis von Fremont Rider auf die Bibliotheksmitarbeiter zeigt, dass es nicht alleine darum geht, die Bücher irgendwo abzulegen, vielleicht alphabetisch nach dem Autor und dem Titel sortiert. Hier ist bereits die Herausforderung zu erkennen, Struktur in das Ganze zu bringen und die Daten so aufzubereiten, dass die Nutzer (in diesem Fall die Leser) auch gut damit arbeiten können. Und der Hinweis auf das exponentielle Wachstum der wissenschaftlichen Veröffentlichungen bringt uns noch einen entscheidenden Schritt weiter: dieses Wachstum lebt davon, dass Wissen weiterentwickelt, miteinander verknüpft wird.

Damit wird klar, dass im Zentrum dieses Themas nicht die Daten stehen, sondern Information. Die Bezeichnung *Big Data* ist schlicht irreführend, denn im Kern besteht das Problem gar nicht darin, dass es viele Daten gibt (die gab es relativ zu den bestehenden Möglichkeiten, damit umzugehen, schon immer). Es geht vielmehr darum, die Information, die in diesen Daten steckt, zu finden und zu nutzen. Das ist es auch, was viele, die sich für *Big Data* interessieren, eigentlich antreibt: sie haben eine gewisse Men-

nicht Daten, sondern Information

2 http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

ge an Daten und vermuten, dass sich darin Informationen verbergen, die es zu finden gilt, um sie anschließend zum eigenen Vorteil nutzen zu können.

Wir möchten Ihnen das an einem Beispiel deutlich machen, das gleich auch das enorme Potential zeigt, welches sich in dieser Technologie verbirgt und das wir als Nutzer dieser Technologie zu erschließen hoffen. Auch dieses Beispiel, von dem Charles Duhigg Anfang 2012 in der New York Times berichtet[3], ist schon etwas älter. Es fällt in eine Zeit, zu der der Begriff *Big Data* noch nicht geprägt war: Andrew Pole hatte im Jahr 2002 gerade seine Stelle als Statistiker bei der Supermarktkette Target angetreten, als zwei Kollegen aus der Marketing-Abteilung mit einer sehr ungewöhnlichen Frage auf ihn zukamen: „Wenn wir herausfinden wollten, ob eine Kundin schwanger ist, könnten Sie uns dabei helfen? Selbst wenn die Kundin nicht möchte, dass wir es erfahren?“[4].

attraktive Zielgruppe

Der Hintergrund für diese eigentlich sehr intime Frage besteht darin, dass werdende und junge Eltern für einen Supermarkt eine attraktive Zielgruppe darstellen, weil man sich eine langanhaltende und damit einträgliche Kundenbeziehung erhofft. Die meisten Kunden haben recht feste Einkaufsgewohnheiten, d.h. sie kaufen verschiedene Produkte immer wieder in denselben Geschäften: das Brot beim Bäcker, frisches Gemüse beim Gemüsehändler, Spielsachen beim Spielwarenhändler, Holz und Werkzeug im Baumarkt, einen MP3-Player im Elektronikmarkt etc.. Zu Target kommen sie nur, wenn es um Dinge geht, die sie mit Target in Verbindung bringen wie z.B. Toilettenpapier oder Socken. Target ist allerdings ein Voll-Sortiment-Supermarkt, die Kunden könnten dort auch Milch, Teddybären oder Gartenstühle kaufen. Aber es ist schwer, diese Botschaft an die Kunden heranzutragen und noch schwerer, ihre Einkaufsgewohnheiten zu ändern.

Im Leben eines Menschen gibt es jedoch einige wenige Phasen, in denen sich Grundlegendes ändert und auch die Einkaufsgewohnheiten beeinflussbar sind; und die Geburt insbesondere des ersten Kindes ist vielleicht der Zeitpunkt, zu dem sich am meisten ändert. Allerdings werden junge Eltern von dem Moment an, in dem das Kind geboren ist und das auch öffentlich bekannt ist, mit Angeboten aller Art überschüttet. Für einen Supermarkt wie Target wäre es also entscheidend, wenn er vorher bereits wüsste, wann ein Kind zur Welt kommen wird; idealerweise im letzten Drittel der Schwangerschaft. Wenn die jungen Eltern erst einmal anfangen, ihre Windeln bei Target zu kaufen, dann ist die Chance sehr groß, dass sie auch viele

3 http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

4 Wir haben das frei übersetzt; im Original des Artikels heißt es: *„If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?“*

der anderen Einkäufe dort tätigen, denn die Zeit, die einem ein kleines Kind noch zum Einkaufen lässt, ist in der Regel sehr begrenzt.

Erfassung von Kaufverhalten

Nun war es auch schon im Jahr 2002 üblich, Kundinnen möglichst eindeutig zu identifizieren, um ihr Kaufverhalten erfassen zu können. Target hatte für fast jeden Kunden eine eindeutige „Guest ID" vergeben und versuchte möglichst alle Kundenaktivitäten vom Bezahlen des Einkaufs über die Teilnahme an Umfragen, Anrufe bei der Hotline oder den Austausch von E-Mails damit zu verknüpfen. Zusätzlich wurden alle verfügbaren Informationen über die Kundin gespeichert: Wohnort, Fahrzeit zum Supermarkt, Familienstand, Kinder, Kreditkarten, geschätztes Einkommen, welche Webseiten besucht werden, etc.. Darüber hinaus konnte Target eine Menge Daten kaufen: Finanz-Score (bei uns: Schufa), Ethnizität, beruflicher Werdegang, Ausbildung, Jahr des Haus(ver)kaufs, Zeitschriften-Abos, Zahl der Autos, etc..

Muster in den Daten erkennen

Alle diese Daten mögen für sich genommen zwar interessant sein, aber sie führen nicht automatisch zu mehr Umsatz. Hier kommen Andrew Pole und seine Kollegen vom Marketing ins Spiel. Ihre Aufgabe bestand nun darin, aus all diesen Daten Muster zu extrahieren, anhand derer sich schwangere Kundinnen identifizieren lassen. Allerdings ist das keine einfache Aufgabe, verglichen damit, den Familien mit Kindern ein paar Wochen vor Weihnachten Werbung für Spielsachen zu schicken. Erleichtert wurde die Arbeit dadurch, dass sich in den Datenbanken Kundinnen identifizieren ließen, die in den letzten Jahren bereits Kinder bekommen hatten. Mit Hilfe statistischer Werkzeuge dauerte es dann nicht lange, bis aus den vorhandenen Einkaufsdaten dieser Mütter Muster extrahiert werden konnten, die sich eindeutig der Schwangerschaftsphase zuordnen ließen.

Zum Beispiel Lotion: einem Kollegen von Andrew Pole fiel auf, dass werdende Mütter zu Beginn des zweiten Drittels der Schwangerschaft mehr Lotion kaufen, die nicht parfümiert ist. Ein anderer Analyst entdeckte, dass in den ersten 20 Wochen manche Schwangere sehr viele Ergänzungsmittel wie Calcium, Magnesium oder Zink kaufen. Viele Kundinnen erwerben Wattebäusche und Seife, aber wenn sie auf einmal mehr extra große Wattebäusche und unparfümierte Seife kaufen, vielleicht sogar zusammen mit Hand-Desinfektionsmittel und Waschlappen, dann ist das ein Zeichen, dass sie möglicherweise kurz vor dem Geburtstermin stehen. Es gelang Pole, etwa 25 Produkte zu identifizieren, die es, wenn man sie zusammen analysierte, ermöglichten einen „Schwangerschaftsvorhersage-Wert" zu bestimmen; wichtiger noch: er konnte sogar ein relativ kleines Fenster für den Geburtstermin errechnen, was es Target ermöglichte, seinen Kundinnen sehr zielgenau Werbung und Gutscheine für die entsprechenden Schwangerschaftsphasen zu schicken.

Einer der Target-Mitarbeiter erläuterte das Verfahren anhand eines fiktiven Beispiels: Nennen wir die Kundin Jenny Ward, sie ist 23 und wohnt in Atlanta. Im März kauft sie Kakao-Butter Lotion, eine kleine Handtasche in der auch eine Windel Platz findet, Zink und Magnesium-Tabletten sowie eine kräfttg-blaue Wolldecke. Dann könnte eine 87-prozentige Wahrscheinlichkeit bestehen, dass sie schwanger ist und ihr Geburtstermin Ende August liegt. Hinzu kommt nun, dass Target aufgrund der Informationen rund um die Guest ID weiß, dass sie auf E-Mail Gutscheine meist mit einem Online-Kauf reagiert. Wenn die Gutscheine freitags per Post kommen, nutzt sie sie oft bei einem Besuch im Supermarkt am Wochenende; und wenn Target dem Kassenbon einen kostenlosen Kaffee bei Starbucks beilegt, dann löst sie ihn bei ihrem nächsten Besuch ein. Normalerweise ist dieses Wissen nicht sehr hilfreich, aber in dieser Lebensphase kann Target es nutzen, um eine ganze Reihe von Produkten zu bewerben, die eine Frau in dieser Schwangerschaftsphase gerne kauft.

Vorhersage des Geburtstermins

Target veröffentlicht keine detaillierten Geschäftszahlen, die es erlauben würden, einen Zusammenhang zwischen der Arbeit von Andrew Pole und dem Umsatz von Produkten für Mütter und Babys herzustellen. Allerdings sprach Target-Präsident Gregg Steinhafel im Jahr 2005 zu einer Runde von Investoren von einem „erhöhten Fokus auf Produkte und Kategorien, die für spezifische Kunden interessant sind, wie zum Beispiel Mutter und Baby". Und es gibt Indizien, dass der Umsatz mit Produkten für Mütter und ihre Babys bei Target tatsächlich in diesen Jahren besonders deutlich gestiegen ist; dass also die Analyse von großen Datenmengen einen signifikanten Beitrag zur Umsatzsteigerung geleistet hat.

Das bringt uns zum Kern des großen Interesses, das es seit einiger Zeit rund um *Big Data* gibt: nahezu jeder, der diese Technologie einsetzt oder über ihren Einsatz nachdenkt, erhofft sich einen Vorteil, der sich ohne sie nicht oder nur sehr schwer erzielen lässt. Besonders spannend dabei ist, dass heute fast überall schon Daten vorhanden sind, deren Besitzer sich erhoffen, wichtige Erkenntnisse daraus ableiten zu können. Das Öl des 21. Jahrhunderts zu fördern heißt, Erkenntnisse aus großen Datenmengen zu extrahieren.

Vorteile aus Big Data

Dabei geht das Potential weit über die Umsatzsteigerung von Supermärkten mit werdenden Müttern hinaus: Neben weiteren kommerziellen Bereichen wie etwa „Predictive Maintenance", also die Möglichkeit die Wartung von Maschinen so zu planen, dass sie stattfindet, bevor ein Teil tatsächlich kaputt geht, birgt der Einsatz von *Big Data* zum Beispiel auch im Gesundheitswesen enormes Potential: wenn sich aus Gesundheitsdaten Schlüsse ziehen ließen, die es erlauben, Krankheiten schon in einer sehr frühen Phase zu identifizieren, könnte die Erfolgswahrscheinlichkeit von Therapien wohl deutlich erhöht werden. Die Verknüpfung von Krank-

enormes Potential

heitsdaten mit Informationen über die Bewegung von Menschen (z.B. aus Mobilfunkdaten) könnte die Vorausberechnung der Ausbreitung von Epidemien deutlich verbessern. Die Auswertung von Daten rund um den Straßenverkehr würde es erlauben, den Verkehrsfluss zu optimieren und damit möglicherweise auch die Umweltbelastung zu reduzieren.

Im Rahmen dieses Buches möchten wir Ihnen einen Eindruck davon vermitteln, wo diese Technologie heute steht, welche Möglichkeiten sie bietet und vor allem, wie Sie herausfinden können, was *Big Data* Ihnen bietet und wie Sie es zu einem leistungsfähigen Werkzeug in Ihrem Umfeld machen können. Wir werden dabei auch die Randbereiche des Themas, zum Beispiel die rechtlichen Aspekte und auch den Einfluss auf die Privatsphäre und damit verbundene Datenschutzfragen adressieren.

Das bringt uns zum Ende dieses Kapitels noch einmal zurück zur Supermarktkette Target: nachdem entsprechende Werbemaßnahmen umgesetzt worden waren, sah sich ein Filialleiter eines Tages einem wütenden Vater gegenüber. Er wedelte mit einer Handvoll Werbebroschüren und beschwerte sich darüber, dass man seiner noch minderjährigen Tochter Werbung für Baby-Kleidung und Kinderbetten geschickt habe; ob man sie zu einer Schwangerschaft ermutigen wolle? Der Filialleiter, der nichts von den Aktivitäten um Andrew Pole wusste, war völlig überrascht und konnte nichts weiter tun, als sich dafür zu entschuldigen. Diese Angelegenheit ließ ihm allerdings keine Ruhe und ein paar Tage später rief er den Vater noch einmal an, um ihn noch einmal um Verzeihung zu bitten. Dieser war am Telefon allerdings sehr kleinlaut und entschuldigte sich seinerseits: er habe mit seiner Tochter gesprochen und dabei herausgefunden, dass er wohl nicht über alles, was in seinem Haus ablief, informiert war. Der Geburtstermin sei im August.

2004: Kundendaten heute: Facebook?

Wenn es bereits im Jahr 2004 möglich war, aus Kundendaten eine Schwangerschaft zu errechnen, von der noch nicht einmal die engsten Familienmitglieder wussten, kann man sich unschwer vorstellen, welche Möglichkeiten die heute verfügbare Technologie bietet. Und gleichzeitig wird klar, wie sensibel wir mit diesem Werkzeug umgehen müssen, um den Nutzen nicht durch einen umso größeren Schaden wieder in Frage zu stellen.

Nun werden Sie vielleicht einwenden, dass die Vorschläge, die Ihnen ein großer Online-Händler unterbreitet („andere Kunden die dieses Produkt gekauft haben, haben sich auch für folgende Produkte interessiert") oft nicht von großer „künstlicher" Intelligenz zeugen. Oder dass die Idee, jemandem, der im Internet nach schönen Plätzen für einen Camping-Urlaub gesucht hat, immer wieder Werbung für Zelte und Campingkocher anzubieten, keine besondere geistige Leistung darstellt.

Lassen Sie sich davon nicht täuschen, in der Regel steckt deutlich mehr Intelligenz dahinter, als Sie denken (auch wenn das oft erst einmal menschliche ist, die in Algorithmen gegossen wird). Welche Qualität die Werbeanbieter inzwischen mit ihren Analysen erreicht haben, zeigt der Bericht der australischen Zeitung *Australian* über ein internes Dokument von Facebook[5]. In diesem Dokument wird dargelegt, wie Facebook die Stimmung seiner jugendlichen und jungen Nutzer analysiert, um genau dazu passend die richtige Werbung effektiv platzieren zu können.

Wir werden am Ende des Buchs noch einmal darauf zurück kommen, welche Schlüsse Target aus dem Vorfall um die Schwangerschaft der minderjährigen Tochter gezogen hat.

1.2 Für wen ist dieses Buch und wie kann man es lesen?

Big Data ist ein Begriff, der seit einiger Zeit fast omnipräsent ist und seinen Weg auch in die Massenmedien gefunden hat, also weit außerhalb der Fachwelt sichtbar ist. Diese große Sichtbarkeit ist aber nicht gleichbedeutend mit einem großen Verständnis in der Allgemeinheit. Während es für weite Teile der Öffentlichkeit den Anschein hat, als wäre *Big Data* etwas ganz Neues und Modernes, arbeiten viele Wissenschaftler aber auch Entwickler aus dem kommerziellen Umfeld seit langem, eigentlich sogar seit Jahrzehnten mit großen Datenmengen (relativ gesehen zum jeweils aktuellen technischen Umfeld).

So arbeitet das Karlsruher Institut für Technologie bereits seit vielen Jahren mit dem CERN zusammen, um die großen Datenmengen, die bei den Messungen der dortigen Teilchenbeschleuniger anfallen, aufzubereiten und den Teilchenphysikern in der ganzen Welt für ihre Arbeit zur Verfügung zu stellen; damals wurde dieses Thema mit dem Begriff LSDF bezeichnet, die Abkürzung für „Large Scale Data Facility".

Large Scale Data Facility

Seit Anfang 2015 arbeiten die Autoren gemeinsam im Rahmen eines vom baden-württembergischen Ministerium für Wissenschaft, Forschung und Kunst geförderten Projektes mit dem Namen Smart Data Solution Center Baden-Württemberg daran, kleinen und mittelständischen Unternehmen (kurz: KMU) im Land Baden-Württemberg den Einstieg in die *Big Data*-Technologie zu erleichtern (zum Unterschied zwischen *Big Data* und *Smart Data* kommen wir in Kapitel 2.1). Gerade im Dialog mit diesen Unterneh-

5 http://www.theaustralian.com.au/business/media/digital/facebook-targets-insecure-young-people-to-sell-ads/news-story/a89949ad016eee7d7a61c3c30c909fa6?nk=b7bac4079848d1f0708bfadb65681860-1493734909