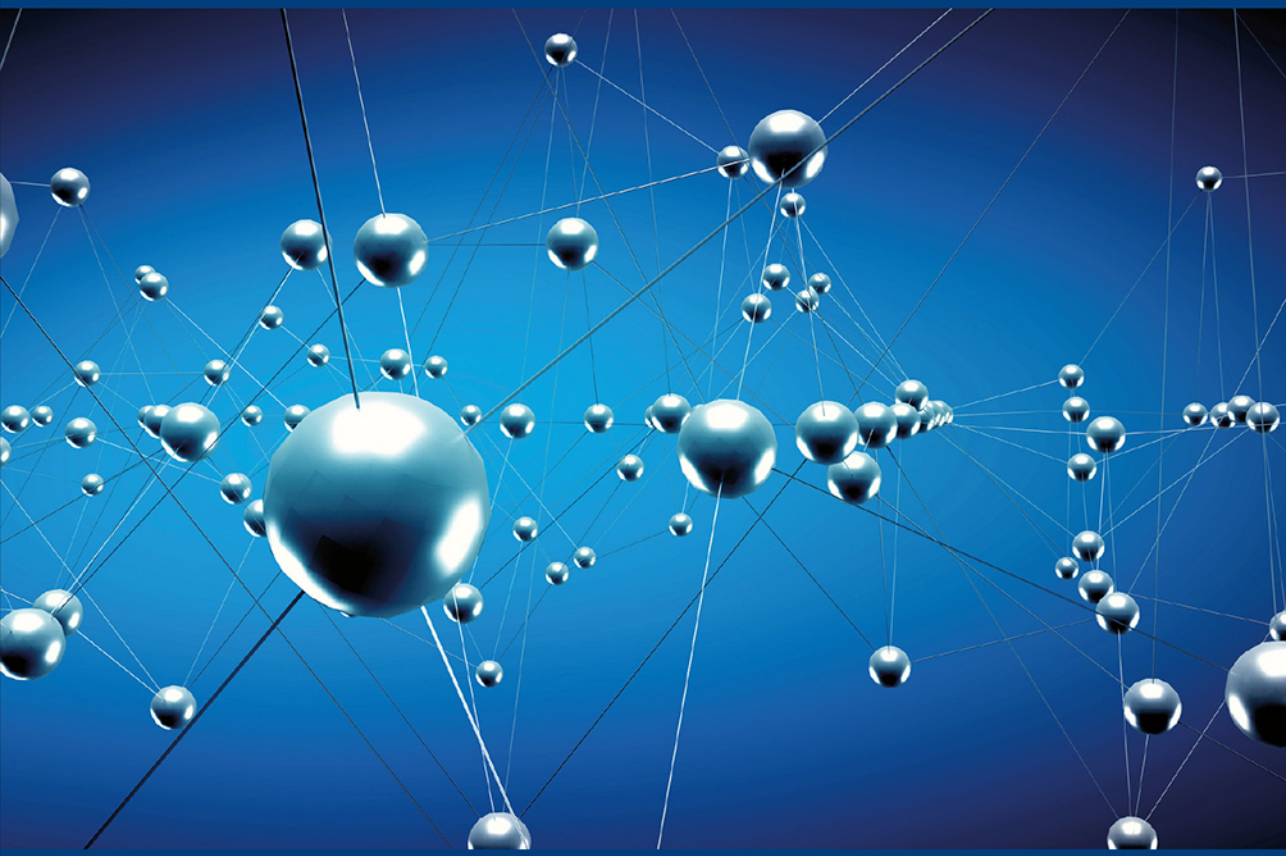
FOSTER PROVOST | TOM FAWCETT



# DATA SCIENCE für UNTERNEHMEN

## Data Mining und datenanalytisches Denken praktisch anwenden

mitp

Foster Provost, Tom Fawcett

# Data Science für Unternehmen

## Data Mining und datenanalytisches Denken praktisch anwenden

Übersetzung aus dem Amerikanischen
von Knut Lorenzen

mitp

**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 978-3-95845-547-4
1. Auflage 2017

www.mitp.de
E-Mail: mitp-verlag@sigloch.de
Telefon: +49 7953 / 7189 - 079
Telefax: +49 7953 / 7189 - 082



Authorized German translation of the English edition of *Data Science for Business*
ISBN 9781449361327 © 2015 Foster Provost and Tom Fawcett
This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Lektorat: Sabine Schulz
Sprachkorrektorat: Simone Fischer
Coverbild: © sdecoret / fotolia.com
Satz: III-Satz, Husby, www.drei-satz.de

# Inhaltsverzeichnis

* Abschnitte, die im Inhaltsverzeichnis mit einem * versehen sind, enthalten mathematische oder technische Details (siehe Seite 17)

*Für unsere Väter*

# Einleitung

»Data Science im Unternehmen« ist für verschiedene Lesergruppen geeignet:

- Führungskräfte und Projektmanager, die mit Data Scientists zusammenarbeiten, Data-Science-orientierte Projekte managen oder in solche Projekte investieren
- Entwickler, die Data-Science-Lösungen implementieren
- angehende Data Scientists

Dies ist weder ein Buch über Algorithmen, noch ist es ein Ersatz für ein solches Buch. Wir vermeiden ganz bewusst einen Ansatz, der sich auf Algorithmen konzentriert, denn wir sind der Meinung, dass es nur einiger weniger grundlegender Konzepte oder Prinzipien bedarf, um aus Daten nützliche Erkenntnisse zu gewinnen. Diese Konzepte dienen als *Grundlage* vieler wohlbekannter Data-Mining-Algorithmen. Sie bilden das Fundament, auf dem die Analyse datenzentrierter, unternehmensrelevanter Probleme, das Erstellen und Bewerten von Data-Science-Lösungen und die Beurteilung allgemeiner Strategien und Lösungsansätze der Data Science beruhen. Dementsprechend orientiert sich die Darstellung an diesen allgemeinen Prinzipien, nicht an bestimmten Algorithmen. Wenn es erforderlich ist, Verfahrensvorschriften detailliert zu beschreiben, verwenden wir statt einer Liste ausführlicher algorithmischer Schritte eine Kombination aus Text und Diagrammen, die unserer Ansicht nach leichter zugänglich ist.

Das Buch setzt keine besonderen mathematischen Kenntnisse voraus. Der Inhalt ist jedoch naturgemäß etwas technisch – Ziel ist es, ein echtes Verständnis von Data Science zu vermitteln, nicht nur einen generellen Überblick zu geben. Wir haben versucht, die Mathematik auf ein Minimum zu beschränken und die Darstellung so »konzeptionell« wie möglich zu gestalten.

Den Aussagen von Branchenkollegen zufolge ist das Buch von unschätzbarem Wert, um eine gute Verständigung zwischen den Managern eines Unternehmens, den Mitarbeitern in Technik/Entwicklung und den Data-Science-Teams zu erzielen. Allerdings stammt diese Beobachtung nur von einer kleinen Gruppe, daher sind wir gespannt, als wie allgemeingültig sich diese Beurteilung tatsächlich erweisen wird (siehe Kapitel 5). Unsere Idealvorstellung sieht so aus, dass jeder Data Scientist seinen Teamkollegen im geschäftlichen Bereich und in der Entwicklung dieses Buch gibt und damit gewissermaßen sagt: Wenn wir wirklich erstklassige Data-Science-Lösungen für unternehmensrelevante Probleme entwickeln und im-

plementieren wollen, dann müssen wir zu einem einheitlichen Verständnis dieses Themas gelangen.

Die Kollegen haben uns außerdem mitgeteilt, dass sich das Buch noch in einem ganz unvorhergesehenen Bereich als nützlich erwiesen hat: für die Vorbereitung auf Bewerbungsgespräche mit Data Scientists. Die Nachfrage nach Data-Science-Experten auf dem Arbeitsmarkt ist hoch und nimmt weiter zu. Aus diesem Grund geben sich immer mehr Stellensuchende als Data Scientists aus. Ein Bewerber für einen solchen Job sollte die Grundlagen der Data Science, die in diesem Buch präsentiert werden, unbedingt beherrschen. (Die Branchenkollegen waren erstaunt, bei wie vielen Kandidaten das nicht der Fall ist. Halb im Scherz, halb im Ernst haben wir sogar erwogen, eine Arbeit mit dem Titel »Anmerkungen zu Bewerbungsgesprächen mit Data Scientists« zu veröffentlichen.)

## Unser konzeptioneller Zugang zu Data Science

Wir stellen in diesem Buch die wichtigsten grundlegenden Konzepte der Data Science vor. Einige davon dienen als Überschriften für entsprechende Abschnitte, andere ergeben sich bei der Erörterung ganz einfach aus dem Zusammenhang (und sind daher nicht unbedingt als grundlegende Konzepte zu betrachten). Diese Konzepte umfassen die Beschreibung der eigentlichen Aufgabe, den Einsatz von Data Science und die Anwendung der Ergebnisse zur Verbesserung von Entscheidungsfindungen. Sie untermauern außerdem eine Vielzahl anderer geschäftsanalytischer Methoden und Verfahren.

Die Konzepte lassen sich in drei allgemeine Kategorien unterteilen:

1. Konzepte, die zeigen, wie Data Science an die Organisation und die Wettbewerbslandschaft angepasst werden kann, inklusive verschiedener Methoden, Data-Science-Teams aufzubauen, zu strukturieren und zu fördern; wie Data Science zu Wettbewerbsvorteilen führen kann und taktische Konzepte zur praktischen Handhabung von Data-Science-Projekten.
2. Allgemeine Konzepte der Datenanalyse, die dabei helfen, geeignete Daten und angemessene Erfassungsmethoden zu erkennen. Diese Konzepte umfassen den *Data-Mining-Prozess* sowie eine Reihe verschiedener *Aufgaben des High-Level-Data-Minings*.
3. Allgemeine Konzepte zur Wissensextraktion aus Daten, die umfangreiche Data-Science-Verfahren und ihre Algorithmen unterstützen.

Eines der fundamentalen Konzepte ist beispielsweise die Erkennung der Ähnlichkeit zweier Objekte, die durch Daten beschrieben werden. Diese Fähigkeit bildet die Grundlage für verschiedene spezielle Aufgaben. Sie kann etwa direkt dazu genutzt werden, Kunden zu *finden*, die einem vorgegebenen Kunden ähnlich sind. Sie bildet den Kern verschiedener *Vorhersage*-Algorithmen, die einen Zielwert

abschätzen, wie z.B. der zu erwartende Ressourcenverbrauch eines Kunden oder die Wahrscheinlichkeit, mit der ein Kunde ein Angebot akzeptiert. Sie bildet außerdem die Grundlage für *Clustering*-Verfahren, bei denen Objekte anhand gemeinsamer Merkmale gruppiert werden, ohne dabei ein festes Ziel zu verfolgen. Ähnlichkeit ist die Grundlage der *Informationsgewinnung*, bei der für eine Suchanfrage relevante Dokumente oder Webseiten abgerufen werden. Und schließlich liegt sie auch vielen gängigen *Empfehlungs*-Algorithmen zugrunde. Ein auf Algorithmen konzentriertes Buch würde all diese Aufgaben womöglich in jeweils eigenen Kapiteln abhandeln, mit unterschiedlichen Bezeichnungen hantieren und die übereinstimmenden Aspekte in den Details von Algorithmen oder mathematischen Sätzen vergraben. In diesem Buch fokussieren wir uns stattdessen auf die vereinheitlichenden Konzepte und stellen bestimmte Aufgaben und Algorithmen als deren natürliche Erscheinungsform vor.

Ein weiteres Beispiel, das bei der Beurteilung der Nützlichkeit eines Musters eine wichtige Rolle spielt, ist der sogenannte *Lift*, der in der Data Science immer wieder auftritt – ein Maß dafür, wie viel verbreiteter ein Muster ist, als man vielleicht erwarten würde. Er dient dazu, völlig verschiedene Muster in unterschiedlichen Kontexten zu beurteilen. Algorithmen für gezielte Werbung werden ausgewertet, indem man den Lift berechnet, den man für die anvisierte Zielgruppe erhält. Der Lift dient zur Beurteilung der Gewichtung von Hinweisen, die für oder gegen eine Schlussfolgerung sprechen. Er gestattet es, zu ermitteln, ob ein gleichzeitiges Auftreten (eine Assoziation) von Daten wirklich von Interesse oder einfach nur auf häufiges Vorkommen zurückzuführen ist.

Wir sind der Ansicht, dass die Erklärung von Data Science anhand dieser grundlegenden Konzepte nicht nur hilfreich für den Leser ist, sondern auch die Kommunikation zwischen geschäftlichen Interessengruppen und Data Scientists vereinfacht. Sie stellt eine gemeinsame Sprache bereit und erleichtert es beiden Seiten, einander besser zu verstehen. Die gemeinsamen Konzepte führen zu intensiveren Diskussionen, die wichtige Themen aufdecken, die anderenfalls vielleicht übersehen würden.

## Hinweise für Dozenten

Dieses Buch wurde erfolgreich als Lehrbuch für ein breites Spektrum von Data Science-Lehrgängen eingesetzt. Es entstand ursprünglich durch die Entwicklung von Fosters fachübergreifenden Data-Science-Kursen an der Stern School der New York University (NYU) im Herbst 2005.[1] Der Kurs wurde eigentlich für Betriebswirtschaftler und Wirtschaftsinformatiker konzipiert, wurde aber auch von Studenten vieler anderer Fächer besucht. Es ist kaum erwähnenswert, dass dieser Kurs bei

1 Natürlich ist jeder der beiden Autoren der Meinung, dass er den Großteil der Arbeit an diesem Buch geleistet hat.

Betriebswirtschaftlern und Wirtschaftsinformatikern gut ankam, da er ja eigentlich für sie gedacht war. Wirklich interessant ist, dass auch Studenten, deren Fächer sich mit Machine Learning und anderen technischen Disziplinen befassten, ihn belegten und als sehr wertvoll ansahen. Der Grund dafür scheint zumindest teilweise darin zu liegen, dass in ihren Lehrplänen außer Algorithmen andere fundamentale Prinzipien und weitere diesbezügliche Themen nicht vorhandenen waren.

An der NYU nutzen wir dieses Buch inzwischen für eine Reihe von Kursen, die in irgendeinem Zusammenhang mit Data Science stehen: den ursprünglichen Kursen für Betriebswirtschaftler und Wirtschaftsinformatiker, Grundkursen für Geschäftsanalyse im Grundstudium, den neuen Vorlesungen über Geschäftsanalyse im Hauptstudium und als Einführung für den neuen Studiengang Data Science der NYU. Darüber hinaus wird das Buch von mehr als zwanzig weiteren Universitäten in neun Ländern an Wirtschaftshochschulen, in Informatikkursen und für allgemeine Einführungen in Data Science eingesetzt (das geschah auch schon vor der eigentlichen Veröffentlichung).

**Hinweis**

Wir führen eine aktuelle Liste der Institute, die das Buch nutzen. Besuchen Sie `http://www.data-science-for-biz.com` und klicken Sie oben auf WHO'S USING IT.

## Weitere Kenntnisse und Konzepte

Es gibt eine Vielzahl weiterer Konzepte und Kenntnisse, die einem praktisch tätigen Data Scientist neben den grundlegenden Prinzipien der Data Science bekannt sein sollten. Diese werden in den Kapiteln 1 und 2 vorgestellt. Wir empfehlen dem interessierten Leser, auch die englische Website zum Buch (`http://www.data-science-for-biz.com`) zu besuchen und sich die dortigen Hinweise zu diesen Konzepten und Kenntnissen näher anzusehen. (Dazu gehören z.B. Python-Skripte, Verarbeitung auf der Unix-Kommandozeile, Datendateien, gängige Datenformate, Datenbanken und Datenbankabfragen, Big-Data-Architekturen und Systeme wie MapReduce oder Hadoop, Datenvisualisierung und andere verwandte Themen.) Auf der Website finden Sie außerdem Lehrmaterialien wie Vorlesungsfolien, Themen für mögliche Hausaufgaben, auf dem Buch aufbauende Beispielprojekte, Prüfungsfragen und vieles andere.

## Konventionen dieses Buchs

Neben vereinzelten Fußnoten enthält das Buch verschiedene Kästen. Dabei handelt es sich im Wesentlichen um erweiterte Fußnoten, die Material enthalten, das

wir für interessant und erwähnenswert halten, das für eine Fußnote jedoch zu umfangreich und für den Fließtext zu abschweifend ist.

### Technische Details – Anmerkung zu mit einem Stern gekennzeichneten Abschnitten

Die gelegentlich erscheinenden mathematischen Details werden in separaten Abschnitten erläutert, die mit einem Stern gekennzeichnet sind. Den Überschriften dieser Abschnitte ist ein * vorangestellt, und sie werden durch einen Kasten wie diesen eingeleitet. Sie enthalten ausführlichere mathematische oder technische Details als der übrige Text, und der einleitende Kasten erläutert den Zweck. Das Buch ist so aufgebaut, dass Sie diese Abschnitte überspringen können, ohne den Faden zu verlieren. An einigen wenigen Stellen verweisen wir den Leser jedoch darauf, dass dort wichtige Details zu finden sind.

Angaben wie (Hinz und Kunz, 2003) sind ein Verweis auf einen Eintrag im Quellenverzeichnis (in diesem Fall auf einen Artikel oder ein Buch von Hinz und Kunz aus dem Jahr 2003). »Hinz und Kunz (2003)« bedeutet dasselbe. Das Quellenverzeichnis für das gesamte Buch finden Sie im Anhang.

Wir versuchen, in diesem Buch mit so wenig Mathematik wie möglich auszukommen, und die vorhandene haben wir weitgehend vereinfacht, ohne dass es zu Missverständnissen kommen kann. Für Leser mit technischen Fachkenntnissen sind einige Anmerkungen bezüglich der von uns vorgenommenen Vereinfachungen angebracht:

1. Wir verzichten auf die Sigma-($\Sigma$) und Pi-($\Pi$) Notation, die in Lehrbüchern üblicherweise für Summen und Produkte benutzt wird. Stattdessen verwenden wir einfache Gleichungen mit Auslassungspunkten wie diese:

   $$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

   In den technischen, mit Sternen gekennzeichneten Abschnitten verwenden wir mitunter die Sigma- und Pi-Notation, wenn Auslassungspunkte einfach zu umständlich wären. Wir gehen davon aus, dass die Leser dieser Abschnitte mit der mathematischen Notation besser vertraut sind und dadurch nicht verwirrt werden.
2. In Statistikbüchern wird sorgfältig zwischen einem Wert und seiner Schätzung unterschieden, indem Variablen, die Abschätzungen sind, mit einem Zirkumflex versehen werden. Die tatsächliche Wahrscheinlichkeit wird typischerweise mit $p$ und die Abschätzung mit $\hat{p}$ gekennzeichnet. Da in diesem Buch fast ausschließlich von Abschätzungen die Rede ist, verzichten wir auf diese Notation, da sie die Gleichungen nur verkomplizieren würde. Sie können davon ausge-

hen, dass es sich immer um Abschätzungen handelt, sofern wir nicht ausdrücklich auf etwas anderes hinweisen.

3. Wir lassen überflüssige Variablen weg und vereinfachen die Notation, wenn ihre Bedeutung aus dem Kontext heraus klar ist. Wenn wir beispielsweise Klassifizierer mathematisch betrachten, haben wir es technisch gesehen mit Entscheidungsprädikaten und Merkmalsvektoren zu tun. Formal würde das zu einer Gleichung wie dieser führen:

   $$\hat{f}_R(\mathbf{x}) = x_{\text{Alter}} \times \ -1 + 0.7 \times x_{\text{Saldo}} + 60$$

   Stattdessen verwenden wir die besser lesbare Gleichung:

   $$f(\mathbf{x}) = \text{Alter} \times \ -1 + 0.7 \times \text{Saldo} + 60$$

   Hier ist **x** ein Vektor und `Alter` und `Saldo` sind dessen Komponenten.

Wir haben uns um einheitliche Typografie bemüht und verwenden für Merkmale und Schlüsselwörter `nicht-proportionale` Schrift. Im Kapitel über Textmining bezieht sich *Ausdruck* beispielsweise auf ein Wort im Dokument, `Ausdruck` hingegen bezeichnet das entsprechende Token in den Daten.

Es gelten die folgenden typografischen Konventionen:

Neue Begriffe, Dateinamen und -erweiterungen werden *kursiv* dargestellt.

Programm-Listings sowie im Fließtext erscheinende Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter werden in `nicht-proportionaler` Schrift gedruckt.

Vom Benutzer einzugebender Text oder kontextabhängige Werte werden in *`kursiver nicht-proportionaler`* Schrift gedruckt.

Auf Websites oder an anderen Stellen auswählbare oder anklickbare Bezeichnungen, wie z.B. Menüpunkte oder Schaltflächen, werden in der Schriftart KAPITÄLCHEN gedruckt.

Im gesamten Buch finden sich Kästen mit Hinweisen und Warnungen, die womöglich unterschiedlich aussehen, je nachdem, ob sie ein auf Papier gedrucktes Buch, ein PDF oder ein E-Book lesen. Sie haben folgende Bedeutung:

**Tipp**

So werden Tipps und Vorschläge dargestellt.

**Hinweis**

Dies ist ein allgemeiner Hinweis.

**Vorsicht**

Warnungen erscheinen in einem Kasten wie diesem. Sie sind wichtiger als Tipps und Hinweise und werden nur sparsam eingesetzt.

## Aus diesem Buch zitieren

Das Buch soll nicht nur eine Einführung in Data Science bieten, sondern auch bei der alltäglichen Arbeit nützlich sein. Die Verwendung von Zitaten oder Beispielen aus diesem Buch unter Angabe der Quelle bedarf keiner besonderen Genehmigung. Üblich sind Nennung von Titel, Autor(en), Verlag, Erscheinungsjahr und ISBN, also beispielweise *Data Science im Unternehmen* von Foster Provost und Tom Fawcett (mitp-Verlag 2017), ISBN 978-3-95845-546-7.

## Danksagungen

Wir möchten den vielen Kollegen und den Menschen danken, die uns in zahlreichen Diskussionen und bei der Durchsicht der Entwürfe zu diesem Buch wertvolle Ideen, Feedback, Kritik und Vorschläge lieferten und die uns stets ermuntert haben. Trotz des Risikos, jemanden zu vergessen, möchten wird insbesondere den folgenden Personen danken:

Panos Adamopoulos, Manuel Arriaga, Josh Attenberg, Solon Barocas, Ron Bekkerman, Josh Blumenstock, Ohad Brazilay, Aaron Brick, Jessica Clark, Nitesh Chawla, Peter Devito, Vasant Dhar, Jan Ehmke, Theos Evgeniou, Justin Gapper, Tomer Geva, Daniel Gillick, Shawndra Hill, Nidhi Kathuria, Ronny Kohavi, Marios Kokkodis, Tom Lee, Philipp Marek, David Martens, Sophie Mohin, Lauren Moores, Alan Murray, Nick Nishimura, Balaji Padmanabhan, Jason Pan, Claudia Perlich, Gregory Piatetsky-Shapiro, Tom Phillips, Kevin Reilly, Maytal Saar-Tsechansky, Evan Sadler, Galit Shmueli, Roger Stein, Nick Street, Kiril Tsemekhman, Craig Vaughan, Chris Volinsky, Wally Wang, Geoff Webb, Debbie Yuster und Rong Zheng.

Darüber hinaus möchten wir den Studenten danken, die an Fosters Kursen *Data Mining for Business Analytics*, *Practical Data Science*, *Introduction to Data Science* und dem Data-Science-Forschungsseminar teilgenommen haben. Ihre Anregungen und Fragen, die bei der Durchsicht der ersten Entwürfe dieses Manuskripts aufgekommen sind, haben wertvolle Hinweise zur Verbesserung des Buchs geliefert.

Wir danken allen Kollegen, die uns all die Jahre Wissen über Data Science vermittelt haben und ebenso, wie man dieses Wissen weitergeben kann. Besonders möchten wir Maytal Saar-Tsechansky und Claudia Perlich danken. Vor einigen Jahren hat

Maytal Foster freundlicherweise die Notizen zu ihrem Data-Mining-Kurs zur Verfügung gestellt. Das Beispiel für den Klassifizierungsbaum in Kapitel 3 (vielen Dank besonders für die Visualisierung durch »Körper«) beruht vornehmlich auf ihren Ideen, die auch den Anstoß gaben für die Visualisierung des Vergleichs der Aufteilung des Hypothesenraums durch Bäume und lineare Diskriminanzfunktionen in Kapitel 4. Auch das Beispiel »Wird David das Angebot annehmen?« in Kapitel 6 und vermutlich weitere, längst vergessene Dinge, basieren auf ihrer Arbeit. Claudia hat in den vergangenen Jahren gemeinsam mit Foster begleitende Kurse zu *Data Mining for Business Analytics* und *Introduction to Data Science* geleitet und ihn dabei vieles über Data Science (und darüber hinaus) gelehrt.

Dank an David Stillwell, Thore Graepel und Michal Kosinski für die Bereitstellung der Facebook-Like-Daten für einige der Beispiele. Dank an Nick Street für die Bereitstellung der Zellkerndaten und des Zellkernbilds in Kapitel 4. Dank an David Martens für die Hilfe bei der Visualisierung der Aufenthaltsorte. Dank an Chris Volinsky für die Bereitstellung der Daten seiner Arbeit über den Netflix-Wettbewerb. Dank an Sonny Tambe für den Zugang zu seiner Arbeit über Big-Data-Technologien und Produktivität. Dank an Patrick Perry für den Hinweis auf das Bank-Callcenter (Kapitel 12). Dank an Geoff Webb, dass wir das Assoziationsanalysesystem *Magnus Opus* benutzen durften.

Vor allem danken wir unseren Familien für ihre Zuneigung, Geduld und Ermutigung.

Beim Verfassen des Buchs kam eine Vielzahl von Open-Source-Software zum Einsatz. Die Autoren möchten den Entwicklern und Mitarbeitern folgender Projekte danken:

- Python und Perl
- Scipy, Numpy, Matplotlib und Scikit-Learn
- Weka
- dem *Machine Learning Repository* der University of California in Irvine (Bache und Lichmann, 2013)

Abschließend möchten wir die Leser noch einmal einladen, unsere englische Website `http://www.data-science-for-biz.com` zu besuchen, die Aktualisierungen, Neues, Errata und Ergänzungen zu dem im Buch vorgestellten Material enthält.

Foster Provost und Tom Fawcett

# Über die Autoren

**Foster Provost** ist Professor und Fakultätsmitglied an der New York University (NYU) Stern School of Business, an der er Business Analytics und Data Science lehrt und Vorlesungen über Betriebswirtschaftslehre hält. Seine preisgekrönten Forschungsarbeiten sind weltweit bekannt und werden häufig zitiert. Bevor er zur NYU wechselte, war er fünf Jahre lang als Data Scientist bei dem Unternehmen tätig, aus dem schließlich Verizon, der größte amerikanische Mobilfunkbetreiber, hervorging. In den letzten zehn Jahren hat Professor Provost verschiedene erfolgreiche Unternehmen mitbegründet, die schwerpunktmäßig Data Science einsetzen.

**Tom Fawcett** hat einen Doktortitel für Machine Learning und war mehr als zwei Jahrzehnte in verschiedenen Branchen (GTE Laboratories, NYNEX/Verizon Labs, HP Labs usw.) in der Forschung und Entwicklung tätig. Die von ihm veröffentlichten Arbeiten zur Methodologie (wie etwa die Beurteilung von Ergebnissen des Data Minings) und Anwendung von Data Science (z.B. Erkennung von Betrugsfällen und Spamfilter) sind zu Standardwerken geworden.

Kapitel 1

# Einführung: Datenanalytisches Denken

*Träume keine kleinen Träume, denn sie haben keine Kraft, die Herzen der Menschen zu bewegen.*

Johann Wolfgang von Goethe

In den vergangenen fünfzehn Jahren haben Industrie und Wirtschaft umfassend in ihre Infrastruktur investiert, um die Möglichkeiten zur Datensammlung innerhalb der Unternehmen zu verbessern. Praktisch das gesamte Wirtschaftsleben steht heute dem Sammeln von Daten offen, und in den folgenden Bereichen findet es schon statt: grundlegender Geschäftsbetrieb, Produktion, Lieferkettenmanagement, Kundenverhalten, Erfolg von Marketingkampagnen, Arbeitsabläufe usw. Und auch über externe Faktoren wie Markttrends, Branchennews und Vorstöße der Konkurrenz sind weitreichende Informationen verfügbar. Dank dieser guten Verfügbarkeit von Daten ist das Interesse an Methoden, mit denen sich nützliche Informationen und Wissen aus diesen Daten gewinnen lassen, gestiegen – das Reich der Data Science.

## 1.1 Allgegenwärtige Datenerfassungsmöglichkeiten

Aufgrund der Verfügbarkeit großer Datenmengen sind Unternehmen fast aller Branchen bestrebt, diese Daten zu nutzen, um Wettbewerbsvorteile zu erzielen. Früher konnten Unternehmen Statistiker, Entwickler und Analysten einsetzen, um die Daten manuell zu untersuchen. Der enorm große Umfang und die Vielfalt der Daten machen heute jedoch eine Fortführung der manuellen Auswertungen schier unmöglich. Doch wir verfügen mittlerweile über immer leistungsfähigere Computer, Netzwerke sind inzwischen allgegenwärtig, und es wurden Algorithmen zur Verknüpfung von Datensätzen entwickelt, die umfassendere und gründlichere Analysen als zuvor ermöglichen. Diese Begebenheiten haben dazu geführt, dass die Anwendung von Data-Science- und Data-Mining-Techniken in Unternehmen enorm gestiegen ist.

Am weitesten verbreitet ist der Einsatz von Data-Mining-Techniken im Marketing, etwa bei der Auswertung von Zielgruppenansprache, Onlinewerbung und Empfehlungssystemen. Data Mining wird im Rahmen des allgemeinen Customer Relation-

ship Managements zur Analyse des Kundenverhaltens eingesetzt, um Kundenschwund zu verhindern und um den Kundenwert zu maximieren. Die Finanzbranche setzt Data Mining zur Bonitätsbewertung, beim Handel mit Krediten, bei der Betrugsermittlung und bei der Personalplanung ein. Und große Einzelhändler wie Walmart oder Amazon nutzen Data Mining in allen Unternehmensbereichen, vom Marketing bis zum Lieferkettenmanagement. Viele Firmen haben sich mit dem Einsatz von Data Science einen strategischen Vorsprung erarbeitet und sind teilweise zu regelrechten Data-Mining-Unternehmen geworden.

Dieses Buch möchte Ihnen dabei helfen, Aufgaben und Herausforderungen im Unternehmen aus der Perspektive der Datenanalyse zu betrachten und die Prinzipien zu verstehen, mit denen Sie diese Daten auswerten und für sich nutzen können. Die datenanalytische Denkweise basiert auf einer fundamentalen Struktur und elementaren Prinzipien, die man erst einmal verstehen muss. Oftmals sind Intuition, Kreativität, gesunder Menschenverstand und Fachwissen unverzichtbar. Eine »Datenperspektive« bietet Ihnen Struktur und Prinzipien und somit ein Grundgerüst für die systematische Analyse von Aufgaben und Problemen. Wenn Sie in dieser datenanalytischen Denkweise geübter sind, werden Sie ein Gespür dafür entwickeln, wie und wo Kreativität und Fachwissen einzusetzen sind.

In den ersten beiden Kapiteln des Buchs werden wir verschiedene Themen und Techniken der Data Science und des Data Minings erörtern. Die Begriffe »Data Science« und »Data Mining« werden oft synonym gebraucht und Ersterer hat eine Art Eigenleben entwickelt, weil viele Personen und Unternehmen versuchen, aus dem derzeitigen Hype Profit zu schlagen. Allgemein gesagt ist Data Science eine Sammlung grundlegender Prinzipien, die die Wissensextraktion aus Daten beschreiben. Data Mining wiederum bezeichnet diese Wissensextraktion aus Daten mithilfe von Verfahren, die eben jene Prinzipien berücksichtigen. Der Begriff »Data Science« wird oft in einem weiteren Sinn gebraucht als der Begriff des traditionellen »Data Minings«. Data-Mining-Verfahren liefern aber einige der besten Beispiele für die Prinzipien der Data Science.

### Hinweis

Es ist wichtig, Data Science zu verstehen, auch wenn Sie nicht beabsichtigen, sie selbst anzuwenden. Die datenanalytische Denkweise ermöglicht es Ihnen, Vorschläge für Data-Mining-Projekte zu beurteilen. Wenn Ihnen beispielsweise ein Angestellter, ein Berater oder ein potenzieller Investitionsempfänger vorschlägt, einen bestimmten Unternehmensbereich durch Wissensextraktion aus Daten zu verbessern, sollten Sie in der Lage sein, diesen Vorschlag systematisch zu beurteilen und zu entscheiden, ob er vernünftig oder fehlerhaft ist. Das soll nicht heißen, dass Sie beurteilen können, ob er tatsächlich Erfolg haben wird – dazu sind bei Data-Mining-Projekten oft Tests erforderlich –, aber Sie sollten offensichtliche Fehler, unrealistische Annahmen und Unvollständigkeiten erkennen.

Im weiteren Verlauf des Buchs werden wir einige grundlegende Prinzipien der Data Science beschreiben. Jede dieser Prinzipien erläutern wir näher anhand einer Data-Mining-Technik, die mit diesem Prinzip arbeitet. Für jede dieser Prinzipien finden sich für gewöhnlich viele verschiedene Verfahren, die dafür eingesetzt werden können, doch in diesem Buch konzentrieren wir uns auf die grundlegenden Prinzipien und legen den Schwerpunkt ganz gezielt nicht auf spezielle Techniken. Wir werden daher nicht zwischen Data Science und Data Mining unterscheiden – es sei denn, es ist für das Verständnis des eigentlichen Begriffs von entscheidender Bedeutung.

Betrachten wir kurz zwei Fallstudien der Datenanalyse zum Erkennen von Vorhersagemustern.

## 1.2 Beispiel: Hurrikan Frances

Aus einem 2004 in der *New York Times* erschienenen Artikel:

> *Hurrikan Frances war unterwegs, raste durch die Karibik und drohte, direkt auf Floridas Atlantikküste zu treffen. Die Anwohner suchten höher gelegenes Gelände auf, um sich in Sicherheit zu bringen. Weit davon entfernt, in Bentonville (Arkansas), beschloss die Geschäftsführung der Walmart-Kette, dass diese Situation ihnen eine großartige Gelegenheit bot, ihr neuestes datengestütztes Instrument einzusetzen: Vorhersagetechnologie.*
>
> *Eine Woche bevor der Hurrikan auf Land traf, hatte Linda M. Dillman, Walmarts IT-Managerin, ihre Mitarbeiter aufgefordert, anhand der Vorkommnisse, die einige Wochen vorher beim Hurrikan Charley eingetreten waren, Vorhersagen zu treffen. In Anbetracht der Billionen von Bytes über das Einkaufsverhalten, die in Walmarts Datenbänken gespeichert waren, kam sie zu dem Schluss, dass ihre Firma versuchen sollte, »vorherzusagen, was geschehen wird, statt darauf zu warten, dass es geschieht,« wie sie sagte. (Hays, 2004)*

Warum wären datengestützte Vorhersagen in diesem Szenario nützlich? Man könnte vielleicht prognostizieren, dass die Menschen in den vom Hurrikan betroffenen Gebieten mehr in Flaschen abgefülltes Wasser kaufen. Gut, das liegt eigentlich auf der Hand, aber wieso bräuchten wir Data Science, um das aufzudecken? Man könnte den vom Hurrikan verursachten *Anstieg der Verkäufe* vorhersagen, um zu gewährleisten, dass die örtlichen Walmarts ausreichend bevorratet sind. Vielleicht würde eine Untersuchung der Daten auch ergeben, dass eine bestimmte DVD in den vom Hurrikan betroffenen Gebieten ausverkauft ist – aber womöglich war sie in der fraglichen Woche landesweit ausverkauft, nicht nur in den vom Hurrikan bedrohten Gebieten. Diese Vorhersage könnte durchaus nützlich sein, ist aber wohl viel allgemeiner als Dillman beabsichtigte.

Von größerem Nutzen wäre es, vom Hurrikan tatsächlich verursachte Verhaltensmuster zu entdecken, die nicht offensichtlich sind. Dazu müssten Analysten die von Walmart in ähnlichen Situationen (wie bei Hurrikan Charley) gesammelten Daten untersuchen, um eine *ungewöhnliche* Nachfrage nach bestimmten Produkten aufzuspüren. Anhand dieser Muster wäre das Unternehmen in der Lage, außergewöhnlich hohe Nachfragen nach bestimmten Produkten vorauszusehen und die Läden entsprechend zu bevorraten, bevor der Hurrikan auf die Küste trifft.

Tatsächlich geschah das auch. Die *New York Times* (Hayes, 2004) schrieb: *»... die Experten untersuchten die Daten und stellten fest, dass die Ladengeschäfte bestimmte Produkte tatsächlich vermehrt benötigen würden – und zwar nicht nur die üblichen Taschenlampen. 'Wir wussten vorher nicht, dass vor einem Hurrikan der Verkauf von im Toaster aufbackbarem Fertiggebäck mit Erdbeergeschmack um den Faktor sieben steigt,' so Dillman in einem Interview. ›Und am besten verkaufte sich Bier.‹«*[1]

## 1.3 Beispiel: Vorhersage der Kundenfluktuation

Wie werden solche Datenanalysen durchgeführt? Sehen Sie sich dazu ein zweites, etwas typischeres Szenario an und überlegen Sie, wie man es aus Sicht der Datenanalyse handhaben würde. Es wird uns als ständiges Beispiel dienen, das viele der in diesem Buch aufgeworfenen Fragen beantwortet und einen gemeinsamen Bezugsrahmen bietet.

Stellen Sie sich vor, Sie arbeiten als Analytiker bei MegaTelCo, einem der größten Telekommunikationsunternehmen der USA. Es gibt ein größeres Problem mit der Kundenbindung im Mobiltelefongeschäft. In den Mittelatlantikstaaten wie New York, New Jersey und Pennsylvania wandern 20 Prozent der Mobiltelefonkunden ab, wenn ihr Vertrag ausläuft, und es wird zunehmend schwieriger, Neukunden zu gewinnen. Seit der Sättigung des Mobiltelefonmarkts ist auch das Wachstum abgeflaut. Die Telekommunikationsunternehmen versuchen, die Kunden bei ihren Konkurrenten abzuwerben und die eigenen bei der Stange zu halten. Den Wechsel von einem Anbieter zum anderen bezeichnen wir als *Abwanderung*, und diese ist besonders kostspielig: Ein Unternehmen muss Geld für Anreize ausgeben, um Kunden anzulocken, ein anderes verliert Umsätze, wenn ein Kunde abwandert.

Nun liegt es an Ihnen, zum Verständnis des Problems beizutragen und eine Lösung zu ersinnen. Neukunden zu gewinnen ist erheblich teurer als Bestandskunden zu halten, daher dient ein Großteil des Marketingbudgets dazu, das Abwandern von Kunden zu verhindern. Die Marketingabteilung hat schon ein entsprechendes Angebot entworfen. Ihre Aufgabe besteht darin, einen genauen, schrittweisen Plan zu entwickeln, wie das Data-Science-Team die riesigen Daten-

1 Natürlich! Was passt besser zu aufgebackenem Fertiggebäck als ein kühles Bier?

ressourcen von MegaTelCo nutzen kann, um zu entscheiden, welchen Kunden vor dem Ablauf ihrer Verträge das neue Angebot unterbreitet werden soll.

Überlegen Sie sorgfältig, welche Daten Sie dazu verwenden könnten und wie Sie diese einsetzen. Stellen Sie sich insbesondere die Frage, wie MegaTelCo die Kunden auswählen soll, die das Angebot für eine Vertragsverlängerung erhalten, damit die Kundenabwanderung so weit wie möglich verhindert und dabei das Budget eingehalten werden kann. Die Beantwortung dieser Frage ist erheblich komplizierter, als man auf den ersten Blick denkt. Wir werden im Verlauf des Buchs wiederholt darauf zurückkommen und die Lösung allmählich verbessern, während wir ein Verständnis für die fundamentalen Konzepte der Data Science entwickeln.

**Hinweis**

Kundenbindung ist tatsächlich eines der Hauptanwendungsgebiete für Data-Mining-Technologien – insbesondere in der Telekommunikations- und Finanzbranche. Diese beiden Branchen haben aus Gründen, auf die wir noch zu sprechen kommen, als erste und am umfassendsten Data-Mining-Technologien eingesetzt.

## 1.4 Data Science, Engineering und datengestützte Entscheidungsfindung

Zur Data Science gehören Prinzipien, Prozesse und Verfahrensweisen, die durch (automatisierte) Datenanalyse zum Verständnis bestimmter Phänomene beitragen. In diesem Buch sehen wir das oberste Ziel der Data Science in der Verbesserung der Entscheidungsfindung, weil diese gemeinhin für Unternehmen von unmittelbarem Interesse ist.

Abbildung 1.1 zeigt Data Science im Kontext verschiedener anderer eng verwandter und mit Daten verbundener Prozesse im Unternehmen. Data Science ist von anderen Aspekten der Datenverarbeitung zu unterscheiden, die zunehmend an Bedeutung gewinnen. Fangen wir oben an.

Die datengestützte Entscheidungsfindung (engl. *Data-Driven Decision-Making*, kurz DDD) beschreibt das Vorgehen, Entscheidungen von der Datenanalyse abhängig zu machen, statt nur der Intuition zu folgen. Beispielsweise könnte eine Marketingfachfrau die Anzeigen, die sie schaltet, allein anhand ihrer langjährigen Erfahrung und ihrer Einschätzung dessen, was gut funktioniert, auswählen. Oder aber sie zieht eine Datenanalyse, die auswertet, wie Kunden auf verschiedene Anzeigen reagieren, zur Entscheidungsfindung heran. Ebenso ist es möglich, eine Kombination beider Ansätze zu verwenden. DDD ist keine Frage von »Alles oder Nichts«, daher setzen verschiedene Firmen DDD in unterschiedlichem Umfang ein.

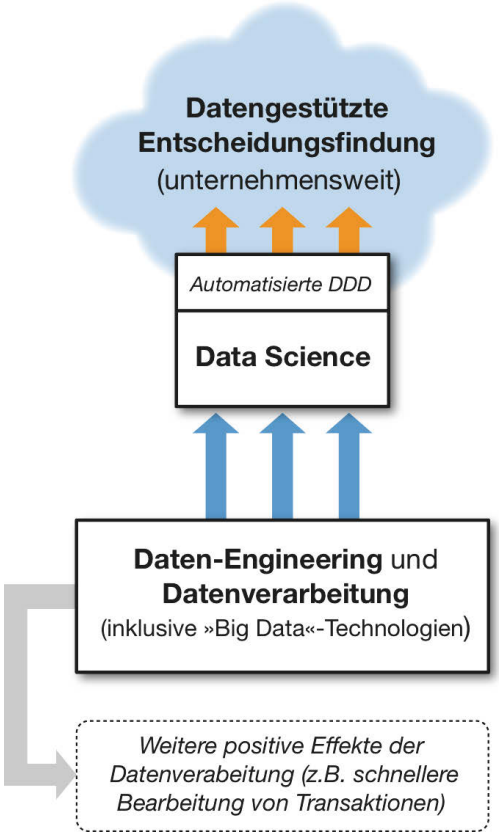



**Abb. 1.1:** Data Science im Kontext verschiedener mit Daten verbundener Prozesse im Unternehmen

Die Vorteile der datengestützten Entscheidungsfindung sind schlüssig dargelegt worden. Der Wirtschaftswissenschaftler Erik Brynjolfsson und seine Kollegen vom MIT sowie der Wharton School der Universität von Pennsylvania haben eine Studie durchgeführt, die zeigt, wie DDD die Unternehmensleistung beeinflusst. (Brynjolfsson, Hitt und Kim, 2011). Sie entwickelten eine Kennzahl für DDD, die Firmen danach beurteilt, in welchem Maß sie auf Daten zurückgreifen, um unternehmensrelevante Entscheidungen zu treffen. Anhand dieser Analyse konnten sie zeigen, dass eine Firma statistisch gesehen umso produktiver ist, je stärker sie Daten nutzt – selbst wenn sie dabei mit einem breiten Spektrums von Störfaktoren konfrontiert ist. Und die Unterschiede sind keineswegs gering. Ein um eine Standardabweichung höherer Wert auf der DDD-Skala ist mit einer Erhöhung der Produktivität von 4 bis 6 Prozent verbunden. Darüber hinaus korreliert DDD auch mit Anlagenrendite, Eigenkapitalrendite, Anlagennutzung sowie Börsenwert – und der Zusammenhang scheint ursächlich zu sein.

Die Art von Entscheidungen, an denen wir interessiert sind, können in zwei Kategorien unterteilt werden:

1. Entscheidungen, bei denen es erforderlich ist, in den Daten etwas zu entdecken und
2. Entscheidungen, die sich wiederholen, besonders solche, die sehr häufig wiederholt getroffen werden müssen. Hier profitiert die Entscheidungsfindung schon von kleinen Verbesserungen, die dadurch bewirkt werden, dass mithilfe von Datenanalysen die Entscheidungsfindung exakter wird.

Das obige Walmart-Beispiel ist vom ersten Typ: Linda Dillman wollte Wissen aufspüren, das Walmart bei der Vorbereitung auf den bevorstehenden Hurrikan Frances helfen sollte.

2012 machte Walmarts Konkurrent Target mit einem Fall datengestützter Entscheidungsfindung Schlagzeilen, ebenfalls ein Beispiel für Typ 1 (Duhigg, 2012). Wie die meisten Einzelhändler ist Target nicht nur an den Einkaufsgewohnheiten seiner Kunden interessiert, sondern auch daran, warum diese etwas kaufen und wie man sie beeinflussen kann. Kunden neigen dazu, ihre Gewohnheiten beizubehalten, und es ist ziemlich schwierig, sie davon abzubringen. Den Entscheidungsträgern bei Target war bewusst, dass die Geburt eines Babys die Einkaufsgewohnheiten einer Familie beträchtlich verändert. Oder wie es ein Analyst von Target formulierte: »Sobald wir eine Familie dazu bewegen können, Windeln bei uns zu kaufen, wird sie bald auch alles andere bei uns einkaufen.« Diese Tatsache ist den meisten Einzelhändlern bekannt und sie konkurrieren daher miteinander darum, Babyartikel an junge Eltern zu verkaufen. Da die meisten Geburtsregister in den USA öffentlich zugänglich sind, beschaffen sich die Einzelhändler diese Informationen und senden den neuen Eltern spezielle Angebote.

Target jedoch wollte der Konkurrenz voraus sein. Ihr Ziel war es, *vorherzusagen*, dass jemand *ein Baby erwartet*. Wenn das gelänge, hätte Target einen Wettbewerbsvorteil gegenüber der Konkurrenz, weil sie ihr Angebot eher als die Wettbewerber unterbreiten könnten. Unter Verwendung von Data-Science-Techniken analysierte Target ältere Daten von Kundinnen, von denen sie wussten, dass sie schwanger geworden waren. Schwangere Frauen stellen z.B. oft ihre Ernährungsweise um, tragen andere Garderobe, nehmen Vitaminpräparate ein usw. Diese Indikatoren könnten den älteren Daten entnommen, zur Entwicklung eines Vorhersagemodells genutzt und in einer Marketingkampagne zum Einsatz gebracht werden. Wie werden Vorhersagemodelle im weiteren Verlauf des Buchs noch sehr ausführlich erörtern. Fürs Erste ist es ausreichend, zu verstehen, dass ein Vorhersagemodell den größten Teil der Komplexität unserer Welt ausblendet und sich auf bestimmte Indikatoren konzentriert, die irgendwie mit einer relevanten Kennzahl zusammenhängen (wer wird kündigen, wer wird kaufen, wer ist schwanger usw.). Entscheidend ist hier, dass sowohl im Walmart- als auch im Target-Beispiel die

Datenanalyse nicht einfach eine Hypothese untersuchte. Stattdessen wurden die Daten in der Hoffnung untersucht, etwas Nützliches zu entdecken.[2]

Bei unserem Beispiel der Kundenabwanderung geht es um ein DDD-Problem des Typs 2. MegaTelCo hat Hunderte von Millionen Kunden, und jeder davon ist ein Abwanderungskandidat. Jeden Monat laufen die Verträge von Millionen Kunden aus, bei denen eine Abwanderung in naher Zukunft somit sehr wahrscheinlich wird. Wenn wir besser abschätzen könnten, wie profitabel es wäre, sich auf einen bestimmten Kunden zu konzentrieren, könnten wir daraus möglicherweise großen Nutzen ziehen, indem wir diese Fähigkeit auf die Millionen von Kunden in der Bevölkerung anwenden würden. Die gleiche Logik ist auch auf viele andere Bereiche anwendbar, in denen Data Science und Data Mining intensiv eingesetzt werden: Direktmarketing, Onlinewerbung, Bonitätsbeurteilung, Finanzhandel, Management von Beratungsstellen, Betrugserkennung, Suchmaschinenplatzierung, Produktempfehlungen usw.

Das Diagramm in Abbildung 1.1 zeigt, dass Data Science die Basis für datengestützte Entscheidungsfindung ist, aber auch, dass es eine Überschneidung zwischen beiden gibt. Hierdurch wird die oft übersehene Tatsache betont, dass geschäftliche Entscheidungen zunehmend *automatisch* von Computersystemen getroffen werden. Verschiedene Branchen haben automatisierte Entscheidungsfindungen eingeführt, einige früher, andere später. Die Finanz- und Telekommunikationsbranche gehörten zu den Ersten, größtenteils deshalb, weil sie schon frühzeitig Datennetze und Computer einsetzten, die eine Zusammenführung und Modellbildung von Daten in großem Maßstab sowie die Anwendung der daraus resultierenden Entscheidungsfindungsmodelle ermöglichten.

In den 1990er-Jahren veränderten sich die Banken- und Verbraucherkreditbranche drastisch aufgrund der automatisierten Entscheidungsfindung. Zur selben Zeit implementierten Banken und Telekommunikationsunternehmen umfassende Computersysteme, die mittels datengestützter Entscheidungsfindungen Betrügereien verhindern sollten. Auch die Handelssysteme wurden zunehmend computergesteuert, und es fand eine Automatisierung von Merchandising-Entscheidungen statt. Bekannte Beispiele sind die Belohnungsprogramme von Harrah's Casinos und die automatisierten Empfehlungen von Amazon oder Netflix. Derzeit erleben wir eine Revolution in der Anzeigenwerbung, vor allem aufgrund der stark zunehmenden Zeit, die Kunden online verbringen und der Möglichkeit, buchstäblich in Sekundenbruchteilen Anzeigen schalten zu können.

2 Target war damit so erfolgreich, dass die Ethik dieser Vorgehensweise infrage gestellt wurde. Ethische Bedenken und Datenschutzfragen sind interessant und sehr wichtig, aber nicht Gegenstand dieses Buchs.