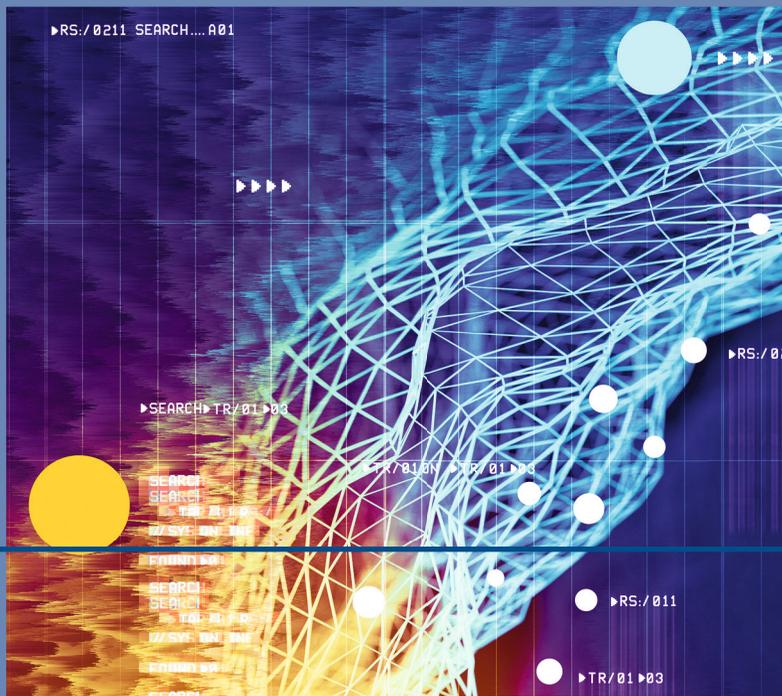


Ralf Otte
Boris Wippermann
Sebastian Schade
Viktor Otte

inkl. Ausblick
auf Small Data
und Mind Data

Von Data Mining bis Big Data

Handbuch für die industrielle Praxis



HANSER

Otte/Wippermann/Schade/Otte

Von Data Mining bis Big Data



Blieben Sie auf dem Laufenden!

Hanser Newsletter informieren Sie regelmäßig über neue Bücher und Termine aus den verschiedenen Bereichen der Technik. Profitieren Sie auch von Gewinnspielen und exklusiven Leseproben. Gleich anmelden unter

www.hanser-fachbuch.de/newsletter

Ralf Otte
Boris Wippermann
Sebastian Schade
Viktor Otte

Von Data Mining bis Big Data

Handbuch für die industrielle Praxis
inklusive Small Data und Mind Data

Mit 204 Bildern und 52 Tabellen

HANSER

Die Autoren:

Prof. Dr.-Ing. Ralf Otte ist Hochschullehrer für Prozessautomatisierung und Künstliche Intelligenz an der Technischen Hochschule Ulm.

Boris Wippermann ist Principal bei der h&z Unternehmensberatung AG in München.

Sebastian Schade, ist Lead Consultant bei der INFOMOTION GmbH Frankfurt.

Prof. Dr.-Ing. habil. Viktor Otte ist emeritierter Hochschullehrer der Universität Wuppertal im Fachgebiet Maschinenbau und langjähriger Berater für Data-Mining-Projekte.



Alle in diesem Buch enthaltenen Informationen wurden nach bestem Wissen zusammengestellt und mit Sorgfalt geprüft und getestet. Dennoch sind Fehler nicht ganz auszuschließen. Aus diesem Grund sind die im vorliegenden Buch enthaltenen Informationen mit keiner Verpflichtung oder Garantie irgendeiner Art verbunden. Autoren und Verlag übernehmen infolgedessen keine Verantwortung und werden keine daraus folgende oder sonstige Haftung übernehmen, die auf irgendeine Art aus der Benutzung dieser Informationen – oder Teilen davon – entsteht.

Ebenso wenig übernehmen Autoren und Verlag die Gewähr dafür, dass beschriebene Verfahren usw. frei von Schutzrechten Dritter sind. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

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.

Dieses Werk ist urheberrechtlich geschützt.

Alle Rechte, auch die der Übersetzung, des Nachdruckes und der Vervielfältigung des Buches, oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Genehmigung des Verlages in irgendeiner Form (Fotokopie, Mikrofilm oder ein anderes Verfahren) – auch nicht für Zwecke der Unterrichtsgestaltung – reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

© 2020 Carl Hanser Verlag München, www.hanser-fachbuch.de

Lektorat: Dipl.-Ing. Volker Herzberg

Herstellung: Björn Gallinge, Carl Hanser Verlag/le-tex publishing services GmbH, Leipzig

Coverkonzept: Marc Müller-Bremer, www.rebranding.de, München

Coverrealisation: Max Kostopoulos

Satz: Kösel Media GmbH, Krugzell

Druck und Bindung: UAB BALTO print, Vilnius (Litauen)

Printed in Lithuania

Print-ISBN: 978-3-446-45550-4

E-Book-ISBN: 978-3-446-45717-1

Vorwort

Das vorliegende Buch entstand im Ergebnis einer über 25-jährigen Beschäftigung der Autoren mit Data-Mining-Projekten im industriellen Umfeld und einer mehrjährigen Tätigkeit im Big-Data-Business im kommerziellen Bereich.

Als 2004, also vor über 15 Jahren, das Buch „Data Mining für die industrielle Praxis“ im HANSER Verlag von zwei der heutigen Autoren erschien [OTT04], reagierte es auf einen zunehmenden Trend in Wirtschaft, Verwaltung, Forschung und Industrie, die Informationen vorliegender Prozessdaten für die Prozessoptimierung besser zu nutzen. „Graben in Daten“ war damals bereits angesagt. Die Effizienz der vorhandenen Rechnersysteme war gestiegen, neue mathematische Verfahren zur statistischen Datenanalyse waren hinreichend erprobt und der Bedarf der potenziellen Anwender dieser neuen Technologie war groß. Schon bis zum Erscheinen des damaligen Buches, aber insbesondere in der darauffolgenden Zeit, entstanden vielfältige Data-Mining-Projekte, die nahezu alle Facetten dieser relativ neuartigen Technologie tangierten. Es musste auch viel gelernt werden. Bei der Projektrealisierung wurden viele handwerkliche Fehler gemacht, insbesondere die Phase der Datenbereitstellung umfangs- und zeitmäßig unzureichend geplant, die Kosten unterschätzt und die Projekteinführung in den industriellen Prozess zu optimistisch prognostiziert. Jeder, der ein derartiges Projekt real begleitet oder sogar selbst verantwortlich geführt hat, weiß davon ein Lied zu singen. Letztlich hat das dem Siegeszug von Data Mining oder Data Science aber nicht geschadet.

Die Aufgaben zur Optimierung vorhandener Prozesse sind heute noch genauso aktuell wie vor 15 Jahren und deshalb haben sich die Autoren des damaligen Buches in Abstimmung mit dem Verlag entschlossen, trotz neuer Entwicklungen und Trends in der Datenverarbeitung wie Big Data oder Small Data dem klassischen Data Mining in diesem Buch erneut hinreichend Platz einzuräumen. Will man Data Mining anwenden, so muss man die dazugehörigen Verfahren bzw. Methoden beherrschen oder zumindest theoretisch verstehen. Obwohl auch dieses Buch kein Lehrbuch sein will, so erschien es uns zweckmäßig, die im Buch von 2004 beschriebenen mathematischen Grundlagen für das vorliegende Buch komplett zu übernehmen, ergänzt durch einige zweckmäßig ausgewählte Beispiele. Der Leser

unseres Vorgängerbuches wird also in den Kapiteln 2 bis 3 nichts wesentlich Neues finden. Es lohnt sich aber auf jeden Fall, sich mit den verschiedenen, problemangepassten Möglichkeiten der Datenselektion und -zusammenführung, der Datenvorverarbeitung, der Datentransformation und der statistischen Datenanalyse wie z.B. multivariaten Verfahren, Künstlichen Neuronalen Netzen (KNN), Support-Vektor-Maschinen, Selbstorganisierenden Merkmalskarten (SOM) oder Clusterverfahren und den daraus ziehbaren Schlussfolgerungen, wie sie z. B. durch verschiedene Regelgenerierungsverfahren oder Visualisierungen hochdimensionaler Datenräume möglich werden, nochmals zu beschäftigen. Zur Klassifizierung von Daten im Sinne von Lernprozessen hat sich die Informatik in der letzten Zeit sehr stark auf Künstliche Neuronale Netze konzentriert, vor allem angeregt durch die neuen Möglichkeiten des Deep Learning mit tiefen Neuronalen Netzen. Dazu wird jedoch Spezialliteratur empfohlen, die Grundlagen der neuronalen Lernprozesse muss man aber verstanden haben. Auch deshalb ist es zweckmäßig, sich nochmals mit der Theorie zu beschäftigen. In Kapitel 4 werden einige Auswertungsmöglichkeiten für praktische Anwendungsfälle dargestellt, die heute wie vor 15 Jahren noch hochaktuell sind.

Viele der hier beschriebenen mathematischen Verfahren findet die Leserin bzw. der Leser auch im KI-Buch „Künstliche Intelligenz für Dummies“ aus dem Jahre 2019 von einem unserer Autoren [OTT19]. Da die Künstliche Intelligenz heute zu einem Großteil auf maschinelles Lernen setzt, sind viele Verfahren und auch Anwendungsbeispiele in beiden Büchern nahezu deckungsgleich. Manche Algorithmen, wie die zu Unrecht fast nirgends erwähnten Selbstorganisierenden Merkmalskarten (SOM), sind in dem hier vorliegenden Buch allerdings detaillierter erläutert, andere, eher übliche KI-Verfahren, wie Entscheidungsbäume, Backpropagation-Netze oder Deep-Learning-Faltungsnetze (CNN) sind in dem KI-Buch näher beschrieben. Der Schwerpunkt des hier vorliegenden Buches liegt eindeutig in der Datenanalyse und nicht so sehr in der Beschreibung der Künstlichen Intelligenz selbst.

In den letzten Jahren hat sich ein gravierender Wandel in den Datenverarbeitungstechnologien ergeben. Während klassische Data-Mining-Projekte immer mit einem gewissen Zeitpolster bearbeitet werden konnten und weiterhin auch können, gibt es zunehmend Aufgaben in allen Bereichen moderner Gesellschaften, bei denen riesige, temporär anfallende Daten (Volume) unterschiedlichster Formate (Variety) in extrem kurzer Zeit (Velocity) verarbeitet werden müssen. Die drei „Vs“ sind zum Charakteristikum einer neuen Form der Datenverarbeitung geworden, die als Big Data bezeichnet wird. Alle Datenverarbeitung bleibt aber nur sinnvoll, wenn den ermittelten Informationen ein inhaltlicher Wert (Value), ein Nutzen, zugeordnet werden kann. Betrachtet man die täglich anfallenden riesigen Datenmengen in sogenannten Sozialen Netzen oder, um ein diametrales Beispiel zu nennen, im Gesundheitswesen, so wird schnell klar, dass die abgeleiteten Informationen bzgl.

ihre Korrektheit oder Zuverlässigkeit sehr unterschiedlich bewertet werden müssen. Hinweise zum Kaufverhalten des Nutzers eines Online-Shops haben durchaus einen anderen Qualitätsanspruch als eine medizinische Diagnose, die aus Vergleichsdaten ähnlicher Fälle computergestützt erstellt wird. Ansprüche an Prädiktion und Wertvorstellungen sind also kontextabhängig. Wir wollen sie hier nicht betrachten, sondern uns stattdessen auf die technologischen Herausforderungen der drei Vs konzentrieren.

Kapitel 5 beschreibt zunächst Architekturkonzepte zum Umgang mit vielen Daten, geht dann auf die Verarbeitungsmöglichkeiten in Form eines deskriptiven oder präskriptiven Datenmodells ein, benennt Möglichkeiten der Hardware- und Datenvirtualisierung und den weiteren Trend zur Entkoppelung von Speicherung und Verarbeitung von Daten. Nach diesen grundsätzlichen Bemerkungen werden in Kapitel 6 die Komponenten der Big-Data-Pionierlösung Hadoop und ihre Weiterentwicklungen dargestellt, auf Apache Spark, eine weitere Evolutionsstufe, eingegangen (die sich gegenwärtig als Standard herausbildet) und anschließend die vielfältigen Modifikationen und Neuentwicklungen beschrieben, die verschiedenen Nutzern den Zugang zu Big-Data-Diensten erleichtern sollen. Dieses Kapitel informiert auch über solche Schlüsseltechnologien wie Apache Kafka oder Kappa, die insbesondere für eine Streamingverarbeitung genutzt werden. Von besonderem Interesse könnte für Leser auch der zunehmende Einsatz von NoSQL-Datenbanken sein, die das übliche relationale Speicherkonzept auch zugunsten von massiv verteilten WEB-Anwendungen aufgegeben haben. Es werden verschiedene Einsatzfälle beschrieben. Hier sind hunderte Open-Source-Anbieter auf dem Markt. Abschließend informiert das Kapitel über die Möglichkeit, für wiederkehrende Aufgaben klar definierte Lösungsmuster, sogenannte Stacks, einzusetzen. Damit sind die Voraussetzungen genannt, um die in den folgenden Abschnitten spezialisierten Ausführungen zu Datenarchitekturen, zum Aufbau von Data Lakes und zum Cloud Computing einordnen zu können. Die Frage, wie Big Data und Künstliche Intelligenz (KI) zusammenwirken, wird abschließend diskutiert. Big Data wird zu einer Quelle für KI und für den Data Scientist, der letztlich für den Mehrwert der Datenanalyse zuständig ist.

Kapitel 7 und 8 schildern Anwendungen. Die gesamten Ausführungen dieser Kapitel sind sehr kompakt gehalten und man bekommt viele Detailinformationen, die auch als Kompass zum Navigieren in diesem äußerst vielseitigen und sich gegenwärtig noch ausdehnenden Terrain dienen. Die Autoren wissen natürlich, dass die dargestellten Informationen nur eine Übersicht sein können und kein Lehrbuch zum Studium einzelner Facetten dieses riesigen Gebietes der Datenverarbeitung. Die angegebenen Quellen, im Schwerpunkt Internetadressen, helfen aber weiter, wenn man die Absicht hat, sich tiefer in die Problematik von Big Data einzuarbeiten. In vielen, in den vergangenen Jahren realisierten Beispielen hat sich gezeigt, dass die ausreichende Vorbereitung und Planung von Data Mining oder Big-Data-

Projekten oft sträflich unterschätzt wird, oftmals sogar bereits die Aufgabenstellung nicht hinreichend an den unternehmerischen Belangen, sondern zu sehr aus einer IT-Perspektive heraus ausgerichtet ist. Hierbei sind also nicht nur die Informatiker und Kenner der inhaltlichen Probleme gefragt, sondern vor allem die Geschäfts- und Prozessverantwortlichen sowie Entscheider. Hier wird auch das Urteil darüber gefällt, ob das Data-Projekt erfolgreich durchgeführt werden wird, ob es beginnt, wie es realisiert oder ob es bereits in einer frühen Phase gestoppt wird. Wir haben uns deshalb entschlossen, diese Problematik sehr ausführlich darzustellen. Die dafür zuständigen Autoren kennen die zu behandelnden Aufgaben genau, da sie sich in ihrem beruflichen Alltag mit allen diesen Fragen der Projektplanung, Projektrealisierung und Projekteinführung täglich beschäftigen müssen. Die gegebenen methodischen Hinweise sind deshalb äußerst praxisnah und oft auch wie ein Kochrezept abzuarbeiten. Dem Leser, der ein Data-Projekt von der Planung bis zur Einführung bearbeiten muss, sind die in diesen Kapiteln gegebenen Hinweise deshalb nachdrücklich zu empfehlen, insbesondere auch deshalb, weil hier verschiedene betriebliche Anwendungsfelder behandelt werden. Einige der Beispiele zu Anwendungsfällen von Data Mining sind wieder dem älteren Buch „Data Mining für die industrielle Praxis“ entnommen, da sie ihre didaktische Funktion, die in Kapitel 2 und 3 dargestellte Theorie am praktischen Beispiel zu erläutern, noch immer gut erfüllen können. Weitere, kleine Ausführungsbeispiele findet man in dem bereits genannten Fachbuch „Künstliche Intelligenz für Dummies“ von einem unserer Autoren.

Verfolgt man die Entwicklung zu Datenverarbeitungstechnologien, so taucht seit einiger Zeit ein neuer Begriff auf, das sogenannte Small Data. Hier wird das Paradigma geändert: Nicht noch immer mehr und immer schneller, sondern zurück zu den Anfängen. Small Data erklärt sich aus der Absicht, nur wenige Daten zu nutzen oder auch nur zur Verfügung zu haben, aber auch aus ihnen wesentliche Informationen zur Lösung eines Problems ableiten zu müssen. In der Philosophie des Ansatzes versteckt sich auch die Sorge, dass extrem viele Daten möglicherweise den Problemerkern verschleiern könnten, man also besser erst einmal versucht, die Dinge zu vereinfachen und mit dem zu arbeiten, was ohnehin zur Verfügung steht. Kapitel 9 widmet sich diesem Thema, wobei schnell klar wird, dass es sich eigentlich nicht um eine neue Technologie, sondern tatsächlich nur um einen Ansatz handelt, der etwas in Vergessenheit geraten ist. Neue Erkenntnisse über die Welt gewinnt man durch Induktion und Deduktion, also durch Extraktion versteckter Inhalte aus riesigen Datenbergen mittels statistischer Verfahren (induktives Lernen), aber auch durch logische (oder kreative) Schlussfolgerung aus oft nur wenigen (oder gar keinen) Daten (deduktives Schließen). Der Mensch kann beides, die modernen Datenverarbeitungstechnologien haben den Akzent seit Data Mining und Big Data aber mehr auf die Statistikvariante gelegt. Es wird also Zeit, sich wieder der Deduktion zu erinnern und zu hinterfragen, wie eigentlich der Mensch mit

seinem Gehirn Small Data und letztlich Deduktion praktiziert. Wie also wird die Zukunft eines technisch optimierten Small Data aussehen?

Hier nun haben die Verfasser dieses Kapitels den bisher beschrittenen Weg der Analyse und Zusammenfassung vorliegender, bewährter Fachinhalte verlassen und – mit dem Risiko „zweifelnder Akzeptanz“ beim Leser – einen vollkommen neuen, hypothetischen Ansatz gewählt. Sie bezeichnen die im menschlichen Gehirn ablaufenden Datenverarbeitungsprozesse als Mind Data und führen aus, dass sich Mind Data nur unter Nutzung von Bewusstsein realisieren lässt. Damit entsteht die grundsätzliche Frage, wie maschinelles, technisches Bewusstsein erzeugt werden kann. Es wird versucht, Antworten darauf aus einem Konzept abzuleiten, welches Bewusstsein als imaginären physikalischen Prozess beschreibt. Dieses Problem tangiert die aktuellen Forschungsarbeiten zur Künstlichen Intelligenz und es wird sofort klar, dass hier auch die neuesten Ansätze der KI hineinspielen, Künstliche Neuronale Netze mit neuromorpher Struktur oder Quantentechnologien aufzubauen.

Damit schließt sich der Kreis dieses Buches, angefangen mit Erläuterungen zu statistischen Verfahren zur Informationsgewinnung im Rahmen von Data Mining, über den Einsatz modernster Datenverarbeitungstechnologien zur Verarbeitung extrem großer Datenmengen in extrem kurzer Zeit durch Big Data bis zur Wiedergeburt der Erkenntnis des Small Data, dass auch in geringen Datenbeständen viel Information enthalten sein kann, die sich mit stark reduzierten Data-Technologien und auch alten, konventionellen Methoden gewinnen lässt. Daraus folgt zwangsläufig der Ansatz eines Mind Data, also einer Datenverarbeitung, die von den zukünftigen Entwicklungen der KI zum Aufbau eines rudimentären, technischen Bewusstseins geprägt sein wird.

Die Autoren hoffen, Ihnen, liebe Leser, damit einen verständlichen und interessanten Überblick über Technologien gegeben zu haben, die im „Zeitalter der Digitalisierung“ schon heute und künftig noch nachhaltiger unser Leben bestimmen werden.

Das vorliegende Buch ist das gemeinschaftliche Werk von vier Autoren. Selbstverständlich gibt es bei den Fachkapiteln dedizierte Verantwortlichkeiten. Insbesondere verantwortlich für die Fachkapitel 2, 3, 4 und 9 und für zahlreiche Industriebeispiele in diesem Buch, unter anderem in den Kapiteln 7 und 8, sind Ralf Otte und Viktor Otte. Das Kapitel 4.1 basiert auf Ausarbeitungen von Christian Rohdanz aus Konstanz, wofür wir uns an dieser Stelle nochmals ganz herzlich bedanken wollen. Unser Autor Sebastian Schade ist als Big-Data-Experte bei den Kapiteln 5 und 6 und bei zahlreichen anderen Praxisbeispielen in verschiedenen Kapiteln federführend gewesen, und Boris Wippermann, als Experte in der Verknüpfung von Digitalisierung und Unternehmensoptimierung, zeichnet hauptverantwortlich für die Anwendungskapitel 7 und 8.

In der Drucklegung des Buches werden leider keine Farben wiedergegeben. Die Autoren weisen extra darauf hin, da bei verschiedenen Bildern die dargestellten Grauwerte nicht immer als unterschiedlich identifiziert werden können.

Wenn Sie Fragen, Anregungen oder auch Verbesserungsvorschläge an einen unserer Autoren haben, freuen wir uns über eine Mail an den Hauptverantwortlichen des Buches unter *ralf.otte@email.de*.

Am Ende des Vorwortes möchten wir allen danken, die zur Erstellung des Buches beigetragen haben, aber ganz besonders auch unseren Familienangehörigen, die durch Verzicht, aber auch Aufmunterung am Gelingen des Projektes beteiligt waren.

Die Autoren

Weinheim, München, Frankfurt/Main und Magdeburg im Frühjahr 2020

Inhalt

Vorwort	V
1 Einführung	1
2 Warum Data Mining? Wozu Big Data?	3
2.1 Definition und Einordnung der Begriffe	6
2.1.1 Was ist Data Mining?	6
2.1.2 Was ist Big Data?	14
2.1.3 Data Mining im Kontext anderer Datenanalyseverfahren	15
2.2 Spezielle Anforderungen der Industrie an die Datenanalyse	22
2.3 Gibt es einen Handlungsbedarf für die Industrie?	29
3 Das theoretische und mathematische Konzept der technischen Datenauswertung	33
3.1 Einführung	33
3.2 Datenselektion und Datenzusammenführung	35
3.2.1 Aufbau einer Datentabelle	35
3.2.2 Denormalisierung von Datentabellen	36
3.2.3 Synchronisierung von Datentabellen	37
3.3 Datenvorverarbeitung	39
3.3.1 Festlegung der Datentypen	39
3.3.2 Diskretisierung von metrischen Daten	41
3.3.3 Statistiken und Tests für metrische Daten	43
3.3.4 Das Problem ungenauer Messungen	48
3.3.5 Behandlung von Datenlücken	51
3.3.6 Behandlung von Ausreißern	53
3.3.7 Behandlung von Mehrdeutigkeiten	55
3.4 Datentransformation	60

3.5	Datenanalyse	64
3.5.1	Visuelle explorative Analysen	64
3.5.2	Überblick über multivariate Verfahren zur Datenanalyse	67
3.5.2.1	Regressionsanalysen	67
3.5.2.2	Varianzanalyse	73
3.5.2.3	Diskriminanzanalyse	76
3.5.2.4	Korrelationsanalyse	79
3.5.2.5	Faktoranalyse	83
3.5.2.6	Clusteranalyse	86
3.5.3	Einführung in Data-Mining-Methoden	93
3.5.4	Data Mining zum Auffinden von Zusammenhängen	97
3.5.4.1	Neuronale Netze	99
3.5.4.2	Support-Vektor-Maschinen	114
3.5.4.3	Gütemaße für Modelle und Klassifikatoren ...	119
3.5.5	Data Mining zum Auffinden von Strukturen	127
3.5.5.1	Fuzzy-Clusterverfahren	128
3.5.5.2	Demographisches Clustern	130
3.5.5.3	Selbstorganisierende Merkmalskarten	131
3.5.5.4	Gütemaße für Clusterverfahren	143
3.5.6	Data Mining zum Generieren von Regeln	145
3.5.6.1	Bayessche Netze	146
3.5.6.2	Entscheidungsbäume	152
3.5.6.3	Assoziationsregeln	162
3.5.6.4	Gütemaße für Regeln	165
3.5.7	Data Mining zum Visualisieren hochdimensionaler Datenräume	166
3.5.7.1	Selbstorganisierende Merkmalskarten für topologieerhaltende Projektionen	166
3.5.7.2	Gütemaße für Projektionen	173
3.5.8	Zusammenfassung der Data-Mining-Verfahren	177
3.6	Interpretation der Ergebnisse	180
3.6.1	Fehlinterpretationen	181
3.6.2	Strittige Interpretationen	187
3.6.3	Konsequenzen	189
4	Hilfreiche Auswertemöglichkeiten für praktische Anwendungsfälle	191
4.1	Text Mining – das Auswerten unstrukturierter Daten	191
4.2	Versuchsplanungen zur Erzeugung von Prozessdaten	197
4.3	Automatische Diskretisierungen	202

4.4	Güte und Sicherheit von Regressionsschätzungen	204
4.5	Auffinden der sensitiven Einflussgrößen	208
4.6	Ausschluss von zufälligen Zusammenhängen	212
4.7	Datenbasierte Optimierungen	216
5	Big Data – die Datenhaltungs- und Verarbeitungskonzepte der Gegenwart	229
5.1	Digitale Transformation und Big Data	230
5.2	Grundprinzipien eines Paradigmenwandels	232
5.2.1	Die drei Vs – und der Wert	232
5.2.2	Scale-up und Scale-out	232
5.2.3	Unabhängige Verarbeitung direkt auf den Daten	233
5.2.4	Schema on Read versus Schema on Write	234
5.2.5	Hardwarevirtualisierung und Containermanagement ..	234
5.2.6	Datenvirtualisierung	235
5.2.7	Entkoppelte Systeme	236
6	Technische Big-Data-Lösungen zur industriellen und kommerziellen Datenanalyse	237
6.1	Datenmanagement im Big-Data-Umfeld	237
6.1.1	Hadoop machte den Anfang	237
6.1.2	Apache Spark – die nächste Evolutionsstufe	240
6.1.3	Abstrahierte Datenverarbeitung und -speicherung	241
6.1.4	Komplexe Eventverarbeitung mit Kafka & Co.	245
6.1.5	Das beste beider Welten – von Lambda und Kappa	246
6.1.6	Big-Data-Plattformen	247
6.1.7	NoSQL-Datenbanken	248
6.1.8	Anwendungsfälle für NoSQL-Datenbanken	249
6.1.9	Technologiestacks	250
6.2	Datenzentrische Architekturen	251
6.2.1	AI-basierte Systeme brauchen IA-basierte Plattformen ..	251
6.2.2	Die logische Architektur	252
6.2.3	Die Softwarearchitektur	252
6.2.4	Die technische Architektur	252
6.3	Der Supervised Data Lake (SDL)	253
6.3.1	Ein Data Lake braucht ein Konzept, damit der See nicht zum Sumpf wird	253
6.3.2	Die unterschiedlichen Bereiche eines SDL	255
6.3.3	Quellen und Ladearten	255

6.3.4	Raw Zone	256
6.3.5	Ingestion Zone	256
6.3.6	Discovery und Sandbox	256
6.3.7	Integration	257
6.3.8	Serving	258
6.3.9	Associated Processes	258
6.3.10	Access und Application	259
6.4	Aufbau eines Data Lakes	259
6.4.1	Think Big – Start Small – Act Now	259
6.4.2	Vision, Ziele und Standortbestimmung	260
6.4.3	Konzeption des Data Lakes	260
6.4.4	Implementierung der Basisumgebung	261
6.4.5	Data Lake Ramp-up – Use Case Driven	261
6.4.6	Industrialisierung – die betriebsfokussierte Datenfabrik	262
6.5	Cloud-Computing und Services	263
6.5.1	Die Cloud-Ausbaustufen – Everything as a Service	264
6.5.2	Offene Ökosysteme	265
6.5.3	Der Data Lake in der Cloud	266
6.6	Big Data, Data Mining und Artificial Intelligence	268
6.6.1	Analytic Data Hub	269
6.6.2	Data-Science- und Data-Mining-Plattformen	270

7	Die Anwendersicht – Systematik für industrielle Anwendungen	279
7.1	Aufgabenstellung und Zielsetzung	279
7.1.1	Datengetriebene Identifikation von Aufgabenstellungen	279
7.1.2	„Produktgetriebene“ Identifikation	280
7.1.3	Geschäftsorientierte Identifikation von Aufgabenstellungen	280
7.1.3.1	Reduktion von Kosten, Verlusten, Verschwendungen	283
7.1.3.2	Erhöhung operativer Performance	284
7.1.3.3	Ergebnisverbesserung funktionaler Prozesse	285
7.2	Vorgehensmethodik	286
7.2.1	Workshop zur Ideenfindung und Datenanalyse	289
7.2.1.1	Design-Thinking-Workshop	289
7.2.1.2	Wertschöpfungsschritte	290
7.2.1.3	Perspektiven	291
7.2.1.4	Schmerzpunkte und Mehrwerte	292
7.2.1.5	Erzeugen des Mehrwertes	292

7.2.1.6	Geschäftsmodell	294
7.2.1.7	Anwendungen und Lösungsansätze identifizieren	296
7.2.2	Hackathons als alternative Möglichkeit der Lösungsfindung und Pilotierung	297
7.2.3	Aufsetzen konkreter Aufgabenstellungen	299
7.2.3.1	Definition der Aufgabenstellung	299
7.2.3.2	Modellauswahl	300
7.2.3.3	Beauftragung von Dienstleistern	301
7.2.4	Explorations- und Umsetzungsphase eines Use Case ...	302
7.2.4.1	Sichtung der Daten	302
7.2.4.2	Bestimmung der sensitiven Eingangsgrößen ..	308
7.2.4.3	Modellierung und Ergebnisbewertung	315
7.2.4.4	Die Königsklasse: Vektorielle Optimierung eines Use Case	316
7.2.5	Auswertung und Detailkonzept, Applikationserstellung und Implementierung	321
8	Die Anwendersicht – typische Anwendungsfelder am konkreten Beispiel	327
8.1	Anwendungen in den Geschäftsfunktionen	330
8.1.1	Forschung und Entwicklung	330
8.1.2	Engineering	333
8.1.3	Produktmanagement	334
8.1.4	Einkauf, Supply Chain Management, Logistik	336
8.1.5	Fertigung und Produktion	338
8.1.6	Qualitätsmanagement	340
8.1.7	Service und Instandhaltung	342
8.1.8	Service und After Market	344
8.1.9	Marketing und Vertrieb	347
8.2	Ausgewählte Data-Mining- und Big-Data-Beispiele	350
8.2.1	Forschung, Entwicklung und Engineering	351
8.2.1.1	Beschleunigung einer Produktentwicklung ...	351
8.2.2	Einkauf	358
8.2.2.1	Spend Cube	360
8.2.2.2	Bündelung	363
8.2.2.3	Spezifikations- und Kostenhebel	366
8.2.3	Produktion, Fertigung und Service	370
8.2.3.1	Störungsanalysen	370
8.2.3.2	Instabilitätsanalysen in einem Klärwerk	372
8.2.3.3	Fehlerdetektion in einem Kraftwerk	381

8.2.3.4	Analyse der Dynamik von chemischen Batchprozessen	390
8.2.4	Instandhaltung und Service	394
8.2.4.1	Aufbau einer Datenbasis für erweiterte Analysen und Monitoring von Industrieanlagen	394
8.2.4.2	Erweiterung eines digitalen Zwillings um Maschinendaten und Strompreisdaten im Bereich Windenergie	396
8.2.5	Marketing und Vertrieb	398
8.2.5.1	Cross-Selling-Effekte mit Data Mining finden ..	398
8.2.5.2	Cross-Selling-Analysen mit Big-Data- Technologien beschleunigen	405
8.2.5.3	Optimale Preisschwellen mit Data Mining aufspüren	407
8.2.6	Data Mining für die strategische Unternehmensführung	412
9	Small Data gehört die Zukunft	421
9.1	Einführung in die Thematik	421
9.2	Charakteristik von Small Data	423
9.3	Machine Learning versus menschlicher Geist – die Mind-Data- Hypothese	428
9.4	Bewusstsein als übergeordnete Ordnungsstruktur neuronaler Systeme	431
9.5	Mind-Data-Auswertungen mit maschinellem Bewusstsein	442
10	Ausblick und mögliche Weiterentwicklungen von Data Mining und Big Data	451
11	Liste der häufig verwendeten Formelzeichen und Symbole ..	457
12	Literaturverzeichnis	461
13	Autoren	471
	Index	473

1

Einführung

Seit vielen Jahren lassen sich im Investitionsverhalten von Industrienationen veränderte Trends erkennen. Während zahlreiche Neuinvestitionen in aufwärtsstrebenden asiatischen Ländern durchgeführt werden, sind Neuinvestitionen vor allem von Großprojekten in den USA, Japan, Europa und insbesondere auch in Deutschland erheblich zurückgegangen. Dieser Trend erscheint zumindest für die nächsten Jahre unumkehrbar zu sein, da aus wirtschaftlichen, aber auch aus sozialpolitischen Überlegungen die Produktionsanlagen möglichst nahe an den Verbrauchermärkten installiert werden und somit gerade in etablierten Industrienationen Neuinvestitionen nur noch in geringem Ausmaß notwendig sind. Deshalb werden hier in den nächsten Jahren der Schwerpunkt und die Herausforderungen darin bestehen, die vorhandenen industriellen Anlagen noch effizienter zu nutzen.

Derart notwendige Effizienzsteigerungen lassen sich durch verfahrenstechnische oder konstruktive Verbesserungen, durch bessere Materialien, bessere Rohstoffe, aber natürlich auch durch eine verbesserte „Fahrweise“ der Anlagen erreichen. Und gerade diese verbesserte Fahrweise ist unter Kosten-Nutzen-Aspekten ein vielversprechender Ansatz.

Auf der anderen Seite sind wir im Zeitalter der Daten angekommen. Daten sind zum Rohstoff geworden, der „gefördert“ und verarbeitet werden muss. Big Data ist in aller Munde, jeden Tag werden tausende Terabyte von Daten auf der Welt erfasst und gespeichert. Im kaufmännischen und technischen Umfeld, aber auch in sozialen Bereichen wird die Nutzung dieser erfassten Daten mit Data Mining und Big-Data-Technologien zum entscheidenden Wettbewerbsvorteil.

Ein Blick in heutige Produktionsbereiche zeigt die Lage: Pro Sekunde entstehen hier Tausende von Informationen, Sensoren messen Geschwindigkeit und Qualität der Fertigung, Steuerungen horten wichtiges Wissen über den Zustand von Maschinen und Anlagen. Ursprünglich haben Unternehmen diese Informationsquellen gezielt für die Automatisierung ihrer Fertigung eingerichtet. Durch die Vernetzung von Maschinen und das Internet der Dinge werden hier inzwischen Datenmengen generiert, die das Volumen jeder anderen Branche übersteigen. Durch intelligente Datenauswertung lassen sich Daten in Erkenntnisse überfüh-

ren, die Entscheidungen unterstützen oder sogar Grundlage für automatisierte Handlungen und Optimierungen sein können. Welches sind die größten Einflussfaktoren auf die Produktivität meiner Fabrik? Wie kann ich Maschinenausfälle vermeiden? Wie kann ich die Qualität meiner Produkte erhöhen? Wie kann ich Energie einsparen?

Die Beratungsfirma Frost & Sullivan leitet von einem Big-Data-Einsatz enorme Optimierungspotenziale ab: eine Steigerung der Produktionseffizienz von zehn Prozent, eine Reduktion der Betriebskosten um fast 20 Prozent und eine Reduktion der Instandhaltungskosten um 50 Prozent. Roland Berger nennt eine Gesamtsumme von 1,25 Billionen Euro bis 2025, die in Europa durch Digitalisierung in der Industrie zusätzlich „gehoben“ werden könnte.

Das Bewusstsein des Wertes der Daten wächst also zunehmend. In der IT-Branche ist dies eine Selbstverständlichkeit. Technologieunternehmen wie Google oder Facebook begründen ihr gesamtes Geschäftsmodell auf der intelligenten Datenauswertung und erzeugen jährliche Gewinne von mehreren Milliarden US-Dollar. Big-Data-Methoden sind hier schon lange etabliert, Rechenleistung und Speicher sind erschwinglich, die Fortschritte im maschinellen Lernen sind enorm und Frameworks zur Nutzung dieser Verfahren frei verfügbar. Doch der Einsatz von Data Mining und Big-Data-Methoden ist im Maschinen- und Anlagenbau, insbesondere bei kleinen und mittleren Unternehmen, nach wie vor verhalten. Laut einer McKinsey-Studie von 2016 verstehen nur 15 Prozent der Betriebe in der industriellen Fertigung, dass Daten als Teil der Wertschöpfung anzusehen sind. In 50 Prozent der Unternehmen bleiben Daten für die Entscheidungsfindung gänzlich ungenutzt.

Dieses Buch beschreibt Möglichkeiten, aufbauend auf vorhandenen Unternehmens- und Prozessdaten, die Ressourcen noch effizienter einzusetzen, die Störungen drastisch zu senken und Ursachen für Unregelmäßigkeiten im Prozess auf eine intelligente und ökonomische Art aufzufinden. Ähnliches gilt für Forschungs- und Entwicklungsprozesse, bis hin zur Konstruktion und experimentellen Erprobung von Objekten oder Verfahren. Nicht alle Unternehmen sind für diese Art der Datennutzung heute schon hinreichend vorbereitet, aber der Trend zur Datenarchivierung ist in keinem Industriezweig mehr aufzuhalten. Damit gewinnt die Auswertung von Daten immer mehr an Bedeutung. Daten stellen Wissen dar und das erzeugt, richtig genutzt, einen entscheidenden Wettbewerbsvorteil.

2

Warum Data Mining? Wozu Big Data?

Die Welt ist in ein neues, ein „digitales Zeitalter“ eingetreten. Jeden Tag werden unzählige neue Daten erzeugt, gespeichert, archiviert, zunehmend auch genutzt. Allerdings hinkt die Nutzung den jeweiligen technischen Möglichkeiten hinterher, da die richtige Auswertung von Daten zur Generierung von Wissen hochkomplex ist. Die Antwort auf dieses Problem ist seit über 20 Jahren das Data Mining, seit fast 10 Jahren sprechen wir sogar von Big Data. Große Verbreitung hat der Big-Data-Ansatz bereits im Marketing & Sales und im Bereich Social Media gefunden. Dafür gibt es im Wesentlichen zwei Gründe: Genau in diesen Bereichen fallen seit Jahren riesige Datenmengen in digitaler Form an, die computergestützt ausgewertet werden können. Aber auch der zweite Grund ist nicht uninteressant. Auswertefehler sind in diesen Anwendungsfeldern nicht so relevant wie in anderen, kritischeren Lebensbereichen. In diesem Buch wollen wir uns verstärkt die Industriebereiche anschauen, da diese aktuell noch sehr viel Nachholbedarf bei Data Mining und Big-Data-Anwendungen besitzen.

In jedem Industriezweig hat sich eine Vielzahl von Spezialisten herausgebildet, die verfahrenstechnisch, konstruktiv, heuristisch oder analytisch den technischen Prozess zu verbessern helfen. Was in der Industrie allerdings immer noch nicht Stand der Technik ist, ist die Anwendung der Data-Mining-Verfahren zur Prozessanalyse und -optimierung. Nun ist die Mathematik bzw. theoretische Informatik den Anwendungen in der Praxis schon seit jeher um viele Jahre voraus, dennoch ist der Zustand unbefriedigend, da durch einen konsequenten Einsatz von modernen Auswertetechniken Einsparungen im Millionenbereich, wenn nicht Milliardenbereich, allein in Deutschlands Industrie erreicht werden könnten.

Data Mining und Big Data treten dabei nicht in Konkurrenz zu klassischen Verfahren der Prozessoptimierung wie Benchmarking, Schwachstellenanalysen, neuen Logistikkonzepten usw., sondern stellen vielmehr eine interessante Ergänzung zur schnellen Effizienzsteigerung und Erfolgsoptimierung dar. Data Mining tritt aber in Konkurrenz zu Prozesssimulationen, dem Bau von analytischen Modellen, dem Erstellen von Reports und dem „Bauchgefühl“ der Anlagenfahrer. Mit Data Mining werden klassische Fragen eines jeden Prozessverantwortlichen aufgegriffen, die

beispielsweise wie folgt lauten könnte: „Wie müssen in diesem oder jenem Augenblick die 100 Einstellgrößen und Prozessparameter festgelegt werden, um mit den geringstmöglichen Kosten und in der geringstmöglichen Zeit die beste Qualität robust zu produzieren, d. h., wie produziert man ein Produkt polyoptimal?“

Diese leicht nachvollziehbare Frage lässt sich in der Praxis aber gar nicht so leicht beantworten, da oftmals nicht bekannt ist, was bei so vielen Parametern als das Optimum zu betrachten ist. Ist es zum Beispiel der geringste Ausschuss bei gleichzeitig maximalem Durchsatz der Anlage? Und wenn ja, dann bleibt als weitere Frage, wie die dazugehörigen Stellgrößen eingestellt werden müssen, um dieses vektorielle Optimum zu erreichen. Die Anforderungen an die moderne Prozessführung sind mittlerweile so komplex, dass der Mensch bei der Durchführung seiner Tätigkeit immer mehr an seine Grenzen stößt.

Mathematisch gesehen kommt weiterhin hinzu, dass es offensichtlich viele parallele und gleichberechtigte Optima eines Prozesses gibt. Die Arbeit in der Praxis zeigt beispielsweise, dass Schichtführer ihre Produktionslinien nach ihren persönlichen Erfahrungen einstellen, die nachfolgenden Schichtführer die gleichen Linien jedoch nach einer anderen Art fahren. In beiden Fällen wird aber oftmals die gleiche Qualität produziert. Das kann übergeordnete Stellen ratlos machen. Doch aus rein mathematischer Sicht kann ein solches Prozessverhalten durchaus normal sein, da ein Produktionsprozess in der Regel gar nicht eindeutig, sondern mehrdeutig ist. Mehrdeutigkeit in dem hier verstandenen Sinne bedeutet, dass es mehrere Kombinationen von Einstellgrößen gibt, die das gleiche Prozessergebnis erzeugen und dass umgekehrt die gleichen Einstellparameter an einem anderen Tage zu völlig verschiedenem Prozessverhalten führen können. Beide Fälle treten auf und man erkennt, dass sie eher die Regel als die Ausnahme darstellen.

Hat man sich aber an den Gedanken gewöhnt, dass jeder Produktionsprozess mehrdeutig ist, erkennt man schnell, dass viele klassische Auswertverfahren, wie z. B. analytische Modellbildungen, scheitern müssen. Nicht, weil sie nicht korrekt wären, im Gegenteil. Sie scheitern an ihrer Korrektheit, weil sie mit der Mehrdeutigkeit der Praxis, d. h. mit der schlechten Datengüte, den ständigen Änderungen in den Anlagen, den Umwelteinflüssen und den subjektiven Einstellungen der Anlagenfahrer nicht umgehen können. Sie sind oftmals die falschen Antworten auf die richtigen Fragen.

Letztendlich kommt es nicht auf die genaue Modellbildung an. Am Ende sollte neben jeder Modellbildung eine konkrete Optimierung und damit eine konkrete Parametervorgabe für die „Anlagenfahrer“ entstehen, damit sie den Prozess nach Maßgabe ihrer Kunden kostengünstiger und besser als vorher fahren können. „Anlagenfahrer“ im erweiterten Sinne sind auch alle diejenigen Bearbeiter, die Prozesse steuern müssen, ganz gleich, ob sie im Einkauf, in der Entwicklung, in der Produktion oder im Vertrieb tätig sind.

Data-Mining-Verfahren versuchen, diese Lücke der komplexen und mehrdeutigen Prozessoptimierung zu schließen, denn mit ihnen werden gerade Problemstellungen der Mehrdeutigkeit, der unvollständigen Daten, der täglichen Änderungen und der enormen Prozesskomplexität gelöst. Data Mining ist damit mehr eine neuartige Herangehensweise an eine komplexe Aufgabenstellung als ein bestimmtes mathematisches Verfahren. Es ist letztlich das Synonym für eine Vielzahl von selbstlernenden, unscharfen Verfahren aus dem maschinellen Lernen und dem Soft Computing.

Was aber ist nun Big Data? Bei [GAB19] finden wir folgende Erklärung: *Mit „Big Data“ werden große Mengen an Daten bezeichnet, die u. a. aus Bereichen wie Internet und Mobilfunk, Finanzindustrie, Energiewirtschaft, Gesundheitswesen und Verkehr und aus Quellen wie intelligenten Agenten, sozialen Medien, Kredit- und Kundenkarten, Smart-Metering-Systemen, Assistenzgeräten, Überwachungskameras sowie Flug- und Fahrzeugen stammen und die mit speziellen Lösungen gespeichert, verarbeitet und ausgewertet werden. Es geht u. a. um Rasterfahndung, (Inter-)Dependenzanalyse, Umfeld- und Trendforschung sowie System- und Produktionssteuerung. Wie im Data Mining ist Wissensentdeckung ein Anliegen. Das weltweite Datenvolumen ist derart angeschwollen, dass bis dato nicht gekannte Möglichkeiten eröffnet werden. Auch die Vernetzung von Datenquellen führt zu neuartigen Nutzungen, zudem zu Risiken für Benutzer und Organisationen. Wichtige Begriffe in diesem Kontext sind „cyber-physische Systeme“ und „Internet der Dinge“, relevante Ansätze angepasste Datenbankkonzepte, Cloud Computing und Smart Grid.*

Was ist all den im Zitat genannten Anwendungen gemeinsam? Denken wir an die drei Vs aus dem Vorwort. Big-Data-Analyse bezeichnet die Auswertung riesiger Datenmengen, die eine sehr schwache oder ständig wechselnde Struktur oder Größe besitzen, so dass eine klassische Datenverarbeitung schwierig, wenn nicht gar – wie bei Streamingprozessen – unmöglich wird.

Data Mining und insbesondere Big-Data-Analysen lassen sich erst durch den massenhaften Zugang zu Rechnern mit enormer Rechenleistung und Speicherkapazität wirtschaftlich umsetzen, da erst nach Schaffung dieser technischen Voraussetzungen viele Milliarden von Datensätzen vergleichbar und in Relationen zueinander gebracht werden können. Es ist daher nicht verwunderlich, dass Big Data eine sehr junge Disziplin ist, die aber durch die teilweise überaktive mediale Berichterstattung einen hohen Bekanntheitsgrad besitzt.

In bestimmten Anwendungen, wie z.B. der Auswertung von Daten aus sozialen Medien, haben Data Mining und Big Data bereits einen negativen Beigeschmack bekommen. Hier müssen ethische Richtlinien geschaffen werden, um eben nicht alles auszuwerten, was technisch auswertbar ist. In den technischen Prozessen ist das Potenzial jedoch noch lange nicht ausgeschöpft.

■ 2.1 Definition und Einordnung der Begriffe

2.1.1 Was ist Data Mining?

Schaut man in die Literatur, so wird man sehr schnell mit einer Vielzahl von Termini konfrontiert. Konnektionistische Systeme, multivariate Analysen, Soft Computing, Knowledge Discovery in Data Bases (KDD), Wissen, Information, explizite und implizite Regeln, Maschinelles Lernen und Datenbanktechniken. All diese Begriffe haben etwas mit Data Mining zu tun, doch all diese Begriffe beschreiben nur einen Teilaspekt von Data Mining. Aber die Thematik ist nicht neu. Es gibt nahezu unendlich verschiedene Sichten auf ein und dieselbe Faktenlage. Jede dieser Sichten hebt gewisse Eigenschaften hervor und unterdrückt andere.

Bevor man sich mit Möglichkeiten der Datenauswertungen befasst, muss geklärt werden, was Daten überhaupt sind. Jeder hat natürlich eine intuitive Vorstellung davon, aber wenn man den Begriff „Daten“ definieren soll, fällt es einem schwer.

In der alten Norm DIN 44300 waren Daten *„Gebilde aus Zeichen oder kontinuierliche Funktionen, die aufgrund bekannter oder unterstellter Abmachungen Informationen darstellen, vorrangig zum Zweck der Verarbeitung und als deren Ergebnis.“*

Diese Definition ist jedoch etwas unglücklich, da man Daten durch Informationen zu erklären versucht, später Informationen jedoch als ganz bestimmte Daten einführt. Zirkelschlüsse sind nie gut, jedenfalls nicht, wenn man sich in ein neues Fachgebiet einzuarbeiten versucht.

Schauen wir zu Wikipedia: *„In der Informatik und Datenverarbeitung versteht man Daten gemeinhin als (maschinen-) lesbare und -bearbeitbare, in der Regel digitale Repräsentation von Information. Ihr Inhalt wird dazu meist zunächst in Zeichen bzw. Zeichenketten kodiert, deren Aufbau strengen Regeln folgt, der sogenannten Syntax. Um aus Daten wieder die Informationen zu abstrahieren, müssen sie in einem Bedeutungskontext interpretiert werden. So kann eine Ziffernfolge wie „123456“ zum Beispiel in Abhängigkeit vom Kontext für eine Telefonnummer, eine Kontonummer oder die Anzahl von Kfz-Neuzulassungen in einem bestimmten Zeitraum stehen. Die betrachtete Zeichenfolge „123456“ oder auch „11110001001000000“ als solche kann nur als Aneinanderreihung von Ziffern erkannt werden; ihre konkrete Bedeutung wird erst im jeweils passenden Kontext (siehe Semantik) klar.“* [Wikipedia/Daten]

Diese Erklärung von Wikipedia erzeugt auch kein besseres Verständnis, daher soll der Begriff einfacher definiert werden. Versuchen wir einen intuitiven Zugang. Was sind also Daten? Stellen wir uns eine grüne Wiese vor, mit endlos vielen Grashalmen. Sind diese Grashalme Daten? Intuitiv gesehen, ja. Und würde man diese Grashalme mit einem Lichtmikroskop betrachten, würden sich verschiedene Oberflächenstrukturen zeigen. Sind diese Oberflächenbilder in dem Mikroskop nun

auch Daten? Wieder ja. Aber was ist, wenn man mit einem Elektronenmikroskop noch tiefer in die Details der Grashalmoberfläche eindringen könnte, würden jetzt neue Daten entstehen? Natürlich. Und wenn man auf atomare Ebene hinabsteigen würde? Wie viele Daten entstünden dann?

Man muss diese Fragen dringend beantworten, wenn man über Big Data nachdenkt, sonst verliert man den Blick für die Realität. Eine wichtige technische Frage wäre nämlich, ob man ein Cloud-System entwerfen kann, das alle Daten, die sich aus einer grünen Wiese generieren lassen, speichern könnte? Die Antwort wird ein klares Nein sein. Niemals lässt sich ein System entwerfen, das alle Daten unserer Umwelt erfassen kann, auch Big Data kann das nicht.

Denn was sind Daten nach unserer Definition? Ein Datum ist jede Abhebung vor einem Hintergrund, oder im Plural:

Daten sind Unterscheidungen eines beliebigen Objektes von seiner Umgebung.

Beispiele: Ein schwarzer Punkt auf einer weißen Wand stellt ein Datum dar. Eine Eins, eingebettet in eine Vielzahl von Nullen, stellt auch ein Datum dar. Eine schwarze Struktur auf einer gelben Wand stellt bereits sehr viele Daten dar. Eine 7-stellige Nummer in einem Telefonbuch stellt Daten dar. Aber auch eine Schwarzfärbung auf einer Telefonbuchseite stellt Daten dar. Auch ein gefundenes Haar auf einer solchen Seite ist ein Datum, zum Beispiel, wenn man Kriminologe ist und einen Fall aufklären muss. Wohin wir auch blicken, überall finden wir Daten.

Damit ist bereits klar, dass es in unserer Umwelt unendlich viele Daten gibt, denn nicht nur die Abweichungen eines Vordergrundes zu einem Hintergrund wären Daten, sondern auch die Ableitungen der Abweichungen, und sogar die Ableitungen der Ableitungen usw. Big Data hat daher, an den Möglichkeiten gemessen, noch nicht sehr viele Daten erfasst. Es ist gerade erst der Anfang.

Und jetzt kommt die entscheidende Einschränkung, um aus der riesigen Vielfalt aller Daten die richtigen herauszufiltern, denn wir wollen gar keine Daten, wir wollen Informationen.

Informationen sind Daten mit potenzieller Bedeutung für einen Nutzer. Wie groß die Bedeutung ist, ist dabei noch unbekannt.

Die Daten auf der grünen Wiese stellen für die meisten von uns zwar Daten, aber keine Informationen dar. Aus dem unendlich großen Vorrat der Natur werden diejenigen Daten herausgefiltert, erfasst und ausgewertet, die für irgendjemanden eine Bedeutung haben oder haben könnten; man denke an die heutige Datenerfassung im Telefonverkehr, die alles sammelt, was möglich ist. Man muss das verstehen, wenn man von den großen Datenmengen hört, die mittlerweile auf unseren Cloud-Systemen abgelegt sind. All diese Daten sind eigentlich Informationen, haben also für irgendjemanden eine Bedeutung, zumindest eine potenzielle, sonst wären sie nicht erfasst worden.

Im Alltag werden diese gespeicherten Informationen (leider) wieder Daten genannt, aber eigentlich ist das falsch. Und ganz verwirrend wird es, wenn man sich fragt, wie groß denn die Bedeutung einer Information sei. Denn seit Shannon [SHA48] wissen wir zwar, wie wir die Informationsmenge, die sogenannte Entropie E , ausrechnen können, die in einem System enthalten ist oder von diesem übertragen werden soll. Wir wissen aber nicht, wie groß die Bedeutung, die semantische Informationsmenge, nun ganz konkret ist. In [OTT19] wurde ein Verfahren vorgeschlagen, um auch die kontextabhängige Bedeutung einer Information auszurechnen. Denn es ist klar, dass sich diese Art der Bedeutung nicht allein aus der Information selbst ergibt. An dieser Stelle soll es uns genügen zu verstehen, dass man die Bedeutung einer Information für einen Empfänger dadurch bestimmen kann, dass man ermittelt, wie groß die Zustandsänderung eines Systems ist, wenn es eine Information (mit der Entropie nach Shannon) empfängt. Dieser Ansatz ist rein intuitiv verständlich, denn er bestimmt eine Bedeutung für die Tragweite, die eine Information für den Empfänger besitzt, und er genügt uns für spätere Diskussionen.

Kommen wir zu dem Begriff der Daten zurück. Im Zeitalter der Digitalisierung versteht man unter Daten diejenigen Informationen, die digital erfasst (da sie eine potenzielle Bedeutung besitzen) und auf den Rechnersystemen der Welt abgelegt wurden. Mit den Daten der oben erwähnten grünen Wiese hat das nicht viel zu tun, die Datenmengen der Umwelt bleiben – wie bereits betont – unendlich groß. Im Folgenden werden wir nur noch von denjenigen Daten als „digitale Daten“ sprechen, die mit digitalen Daten-Technologien ausgewertet werden sollen. Die unmittelbare Auswertung analoger Daten spielt heute kaum noch eine Rolle, wobei das aus Sicht der Autoren falsch ist, denn der Mensch verarbeitet nur analoge Daten und insbesondere im Bereich der Künstlichen Intelligenz wird es eine Rückkehr zu analogen Systemen, zum Beispiel neuromorphen Systemen, geben müssen, siehe Ausblick und [OTT19].

Warum speichert man all diese Daten? Was will man aus den Daten (Informationen) herausfiltern? Die Antwort: Man will Wissen extrahieren und genau dies macht man mit Verfahren des Data Mining.

Data Mining ist ein Prozess der (halb-)automatischen Wissensextraktion aus bereits abgespeicherten (strukturierten) Daten (Informationen).

Und bei Big Data ist dieser Prozess ins 1000-fache überhöht und echtzeitfähig. Um Big Data zu verstehen, muss man sich daher zuerst mit Data Mining befassen, weshalb das Buch gängige Data-Mining-Verfahren ausführlich behandelt. Der Weg zu Big Data ist dann ein leichter, zumindest auf abstraktem Niveau.

Nun wollen wir den Begriff des Data Mining vertiefen. Bei diesem Begriff gibt es verschiedene Betrachtungsstandpunkte. Der mathematisch interessierte Data-Mining-Ingenieur hebt verschiedene Verfahren der Datenanalyse in den Vorder-

grund, der Informatiker stellt interessante Aspekte der effizienten Implementierung von Data-Mining-Algorithmen dar und der betriebswirtschaftliche Anwender erklärt Data Mining mit einer Reihe von sehr interessanten Projekten im Banken-, Versicherungs- oder Handelsbereich. Es verwundert deshalb nicht, dass unterschiedliche Interessenlagen zu unterschiedlichen Sichtweisen auf dieses Thema führen. Und doch gibt es eine Definition mit dem kleinsten gemeinsamen Nenner:

Data Mining ist ein Prozess zum Auffinden von unbekanntem und nicht trivialen Strukturen, Zusammenhängen und Trends in Datenbeständen.

Wir wollen uns dieser allgemeineren Auffassung anschließen, aber noch auf einen Unterschied aufmerksam machen. In der technischen Praxis gibt es sehr häufig Aufgaben, die durch nur wenige Daten charakterisiert sind, im kommerziellen Bereich und bei Social Media liegen dagegen oftmals Hunderttausende oder Millionen von Daten über Kunden und deren Verbrauchsverhalten vor. Daher werden verschiedene Auswertemethoden notwendig sein.

Kommen wir auf die obige Definition zurück: Mit dem Begriff *Prozess* wird verdeutlicht, dass es nicht um Algorithmen und mathematische Verfahren geht, sondern um eine komplexe Betrachtungsweise, die Datenerhebung, Datenselektion, Datenvorverarbeitung, Datenanalyse, Interpretation und Anwendung einschließt. Die Wörter *unbekannt* und *nicht trivial* bedeuten, dass nur dann ein (erfolgreicher) Data-Mining-Prozess vorliegt, wenn neue und hinreichend komplexe Ergebnisse erzielt werden. Aus mathematischer Sicht wäre diese Eingrenzung auf neu und nicht trivial nicht notwendig, aus praktischer Sicht ist sie allerdings sehr wohl entscheidend. Ein triviales Ergebnis wäre beispielsweise die Erkenntnis, dass am Sonntag kein Ausschuss produziert wird, weil eben gar nicht produziert wird. Diese Art von Ergebnissen wird natürlich von den verwendeten Data-Mining-Algorithmen auch geliefert. Es wird damit klar, dass an jede Analyse eine Interpretationsphase angeschlossen sein muss.

Um die Begriffe *Strukturen*, *Zusammenhänge* und *Trends* zu erklären, betrachte man nachfolgende Datentabelle aus einem Produktionsprozess, Bild 2.1:

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Wochentag	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	...	x_i	x_j	y_1	y_2	y_3	y_4
5	1920	0,1	8	8	100	290,00	7,7	192	192	1400	100	1400	20,8	0,1	5	14	
6	8630	0,2	24	24	350	290,00	7,7	3020,5	2828,5	3212,5	400	-80	400	22	0,1	14	9
0	9960	0,3	12	12	580	290,00	6,7	5776,8	2756,3	8989,3	1200	0	1200	23,2	0,5	9	14
1	9770	3,2	120	120	790	290,00	8,3	7718,3	1941,5	16707,6	1200	20	1200	24,4	0,7	14	18
2	9730	0,2	140	140	750	290,00	7,9	7297,5	-420,8	24005,1	1000	-140	1000	25,5	2,5	18	15
3	9730	30,4	52	52	770	290,00	7,4	7492,1	194,6	31497,2	2400	111429	2400	27,1	3,5	15	5
4	9620	29	20	53,71	800	290,00	11,9	5772	-1720,1	37263,2	1400	142857	1286	27,3	1,3	6	9
5	10610	7	380	106,9	600	300,00	9,6	6366	594	43635,2	1300	-22,857	1271	27,5	1,3	9	6
6	10130	4,8	288	144,6	470	303,33	9,1	4761,1	-1604,9	48396,3	2000	-11429	1500	27,8	0,1	6	4
0	9460	1	1170	310	620	306,67	11,8	5865,2	1104,1	54261,5	2000	-11429	1614	28	0,1	4	2
1	9150	0,1	1336	483,7	730	310,00	11,9	6679,5	814,3	60941	2000	-25,714	1729	28,4	0,1	2	8
2	8620	23	588	547,7	680	313,33	11,5	5861,6	-817,9	68802,6	2800	14,2857	1986	28,1	0,4	8	13
3	8220	36	484	609,4	710	316,67	12,2	5836,2	-25,4	72638,8	1400	-20	1843	29	0,8	13	7
4	8990	19	540	683,7	710	320,00	11,2	6382,3	548,7	79021,7	2800	-21429	2043	29,4	0,9	7	10
5	9000	35	704	730	630	300,00	12,1	5670	-712,9	84691,7	2800	8,57143	2257	29,7	0,4	10	9
6	9130	38	648	781,4	490	318,33	11,9	4473,7	-1196,3	89165,4	1400	0	2171	28,6	0,3	9	12
0	10220	35	600	700	650	336,67	11,7	6643	2169,3	95808,4	2000	0	2171	28,5	0,6	12	10
1	8980	28	1240	686,3	650	355,00	10,5	5753	-884	101567	2000	114286	2171	27,8	0,1	10	7
2	9000	22	636	693,1	840	373,33	11,9	7560	1801	109127	2000	-5,7143	2057	27,3	0,1	7	8
3	9460	33	616	712	810	391,67	11,9	7662,6	102,6	116790	1800	114286	2114	27,5	0,1	8	13
4	8860	5,5	240	683,1	830	410,00	9	7353,8	-308,8	124144	2000	5,71429	2000	28,7	0,3	13	12
5	9430	20	496	639,4	710	310,00	12,1	6695,3	-658,5	130839	2400	-22,857	1943	28,5	0,5	12	29
6	9250	29	548	625,1	670	293,33	11,9	6197,5	-497,8	137037	3000	-14,286	2171	28,2	0,5	29	26
0	11260	6,5	276	578,9	1020	276,67	8,8	11485,2	5287,7	148522	3000	-5,7143	2314	28,4	0,1	26	21
1	11390	4,5	416	461,1	390	260,00	10,6	4442,1	-7043,1	152964	2400	-5,7143	2371	27	0,1	21	16
2	10180	14	404	428	480	243,33	11,4	4886,4	444,3	157850	2400	-8,5714	2429	27	1	16	11
3	8920	16	504	412	410	226,67	11,2	3657,2	-1229,2	161508	2400	0	2514	26,7	0,3	11	8
4	10630	25	540	454,9	460	210,00	11,6	4889,8	1232,6	166397	2000	5,71429	2514	26,7	0,2	8	6
5	10240	34	712	485,7	480	230,00	11,9	4915,2	25,4	171313	2000	14,2857	2457	27,5	0,1	6	5
6	11410	33	624	496,6	400	235,00	11,6	4564	-351,2	175877	2000	17,1429	2314	27,9	0,2	5	4
0	11690	13	520	531,4	760	240,00	11,3	8884,4	4320,4	184761	1800	5,71429	2143	28	0,1	4	4
1	12410	23	520	546,3	600	245,00	10,3	7446	-1438,4	192207	2000	5,71429	2086	26,3	0,1	4	5
2	10860	25	1176	656,6	560	250,00	9,6	6081,6	-1364,4	198289	2000	14,2857	2029	26,1	0,1	5	11
3	8870	28	468	651,4	550	255,00	11,5	4878,5	-1203,1	203167	1400	2,85714	1886	27,2	0,2	11	8
4	11360	2	472	641,7	530	260,00	11,3	6020,8	1142,3	209188	1800	8,57143	1857	27,7	0,1	8	10
5	10120	10	400	537,1	590	270,00	9,6	5970,8	-50	215159	1400	0	1771	28,2	0,4	10	11
6	10480	14	420	568	380	268,33	10,3	3982,4	-1988,4	219141	2000	-2,8571	1771	28,2	0,1	11	6

Bild 2.1 Datentabelle aus einem Produktionsprozess

In den Spalten stehen die einzelnen Prozessvariablen x_1, x_2, x_3, \dots und die Zielgrößen y_1, y_2, \dots . In den Zeilen stehen die Ausprägungen dieser Variablen für ihre Abtastzeiten, im Beispiel des Bildes 2.1 ist die Abtastzeit einmal pro Tag, siehe Spalte Wochentag (Sonntag bis Samstag sind mit 0 bis 6 gekennzeichnet).

Man kann sich vorstellen, dass es nun mehrere Möglichkeiten gibt, eine solche Tabelle auszuwerten, nämlich spalten- oder zeilenweise. Die Data-Mining-Anwendungen und Algorithmen unterscheiden deshalb folgende Analysemöglichkeiten:

- Zusammenhangsanalyse: Man sucht nach Zusammenhängen zwischen den Spalten der Tabelle, also nach Zusammenhängen der Form $y = f(x_1, x_2, x_3, \dots)$,
- Strukturanalyse: Man sucht nach Zusammenhängen zwischen den Zeilen der Tabelle, also den Ausprägungen untereinander,
- Trendanalyse: Man sucht Vorhersagemöglichkeiten einer Zielgröße, entweder nur aus den Eingangsgrößen oder aus den Eingangsgrößen und der Zielgröße selbst.

Selbstverständlich gibt es dabei Überschneidungen. Dennoch ist es sehr hilfreich, wenn man am Beginn der Analyse weiß, ob man eher nach Strukturen oder eher

nach Zusammenhängen in den Daten sucht. Um diesen wichtigen Punkt noch besser zu verdeutlichen, sei nachfolgendes einfaches Datenbeispiel aus Bild 2.2 mit zwei Variablen x_1 und y_1 gegeben (Definitionsbereich von x_1 sei 36 bis 45,5; dargestellt sind aus Platzgründen nur Werte von 36 bis 42).

	A	B
1	x_1	y_1
2	36	3700
3	36	5200
4	36,5	2700
5	36,5	4100
6	36,5	1700
7	37	400
8	37	3900
9	37	4600
10	37,5	2200
11	37,5	2400
12	37,5	3500
13	38	4800
14	38	5200
15	38	800
16	38,5	1700
17	38,5	3400
18	38,5	3400
19	39	2600
20	39	4700
21	39	400
22	39,5	3100
23	39,5	2600
24	39,5	4000
25	40	800
26	40	6100
27	40	2500
28	41,5	5700
29	41,5	7200
30	41,5	4700
31	42	6100
32	42	3700

Bild 2.2
Zwei Variablen x_1 und y_1

Durch einen Scatter-Plot (x - y -Plot oder Streudiagramm) nach Bild 2.3 erkennt man bereits rein visuell Zusammenhänge und Strukturen:

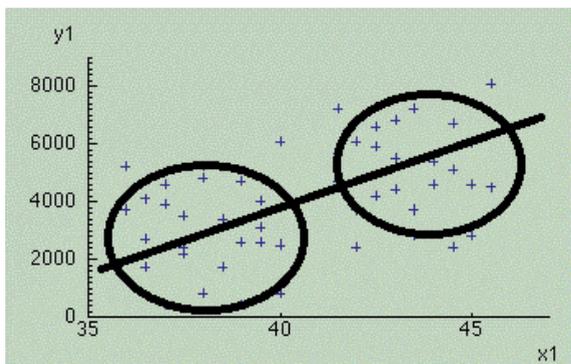


Bild 2.3
Scatter-Plot der Variablen x_1 und y_1 mit eingezeichneten Ellipsen für Strukturen und einer Gerade für Zusammenhänge

Die Gerade in Bild 2.3 zeigt, dass es einen schwachen Zusammenhang zwischen y_1 und x_1 gibt. Die beiden Ellipsen verdeutlichen die beiden Untergruppen des Datensatzes, also eine Struktur.

Ausgehend von diesen rein visuellen Überlegungen lassen sich folgende Datenanalysen durchführen:

- Korrelationsanalyse (ungerichtete Zusammenhangsanalyse)
- Regressionsanalyse (gerichtete Zusammenhangsanalyse)
- Clusteranalyse (Strukturanalyse)

```
File  Generate

Statistics for field : x1
Minimum      =      36
Maximum      =     45.500
Occurrences  =      52
Mean         =     40.750
Standard Deviation =    2.9811
Correlation (Pearson Product-Moment) for field :
  y1         =  0.478 ( Medium positive correlation)

Statistics for field : y1
Minimum      =      400
Maximum      =     8100
Occurrences  =      52
Mean         =    4096.2
Standard Deviation =    1837.2
Correlation (Pearson Product-Moment) for field :
  x1         =  0.478 ( Medium positive correlation)
```

Bild 2.4 Univariate Statistik und Korrelation der Variablen x_1 und y_1

Die Korrelationsanalyse nach Bild 2.4 zeigt eine schwache positive Korrelation zwischen den Variablen x_1 und y_1 von ca. 0,47 (Pearson-Product), d.h., man erkennt folgende, allerdings nur schwach ausgeprägte Regel: Je größer (kleiner) x_1 , desto größer (kleiner) ist y_1 bzw. je größer (kleiner) y_1 , desto größer (kleiner) ist x_1 .¹

Des Weiteren sieht man in Bild 2.5, dass man mit einer Genauigkeit von 85% (bei dieser Angabe bezogen auf den Messbereich $y_{1\max} - y_{1\min}$) den Wert y_1 aus x_1 mit der Technik der neuronalen Netze schätzen kann. Natürlich könnte man den Zusammenhang auch durch eine einfache Regressionsgerade herausfinden.

¹⁾ Korrelationswerte zwischen zwei Variablen, deren absoluter Betrag kleiner als 0,75 ist, werden hier als schwach bezeichnet.

```
File

Neural Network "y1" architecture
Input Layer      : 1 neurons
Hidden Layer #1 : 4 neurons
Output Layer     : 1 neurons

Predicted Accuracy : 85.18%

Relative Importance of Inputs
x1                  : 0.26797
```

Bild 2.5Neuronales Vorhersagemodell $y_1 = f(x_1)$

Weiterhin sind statistisch auch zwei Cluster voneinander separierbar, Bild 2.6.

```
File

Number of inputs = 1
Number of records = 52
Initial number of clusters = 2
Final number of clusters = 2
Max. number of iterations = 20
Error change = 0,000001
Iteration 1, Error = 0,218624
Iteration 2, Error = 0,000000

cluster 1: 26 examples
x1 : 38.0769

DISTANCE FROM :
Cluster 2 : 0,562753
-----
cluster 2: 26 examples
x1 : 43.4231

DISTANCE FROM :
Cluster 1 : 0,562753
-----
```

Bild 2.6Clusteranalyse der Daten x_1 und y_1

Fragt man also nach Zusammenhängen in diesem Datensatz, kann man mittels einer Regression (im Beispiel mit einem neuronalen Modell) erkennen, dass sich y_1 durch x_1 erklären lässt (aber auch umgekehrt). Fragt man nach Strukturen, so sieht man, dass es zwei Untergruppen (Cluster) gibt. Je nach Fragestellung erhält man unterschiedliche Informationen über die in den Daten enthaltenen Zusammenhänge und Strukturen.

Obwohl das dargestellte Beispiel sehr einfach ist, zeigt es prinzipielle Möglichkeiten bei der Auswertung von Datensätzen. Es ist notwendig, vor dem Beginn von Data Mining festzulegen, was man in den Daten suchen möchte. Will man wissen,