O'REILLY®

Werde ein Data Head

Data Science, Machine Learning und Statistik verstehen und datenintensive Jobs meistern

Alex J. Gutman, Jordan Goldmeier

Übersetzung von Jørgen W. Lang

Papier plus PDF.

Zu diesem Buch – sowie zu vielen weiteren O'Reilly-Büchern – können Sie auch das entsprechende E-Book im PDF-Format herunterladen. Werden Sie dazu einfach Mitglied bei oreilly.plus+:

www.oreilly.plus

Werde ein Data Head

Data Science, Machine Learning und Statistik verstehen und datenintensive Jobs meistern

Alex J. Gutman, Jordan Goldmeier

Deutsche Übersetzung von
Jørgen W. Lang

O'REILLY®

Alex J. Gutman, Jordan Goldmeier

Lektorat: Alexandra Follenius
Übersetzung: Jørgen W. Lang
Fachgutachten: Marcus Fraaß
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Michael Oréal, *www.oreal.de*, unter Verwendung der iStock-Illustration ID 1173117448 von Vertigo3d/Getty Images
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

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

ISBN:
Print 978-3-96009-191-2
PDF 978-3-96010-667-8
ePub 978-3-96010-668-5
mobi 978-3-96010-669-2

1. Auflage 2022
Translation Copyright für die deutschsprachige Ausgabe © 2022 dpunkt.verlag GmbH
Wieblinger Weg 17
69123 Heidelberg

Authorized German translation of the English edition of *Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning* by Alex J. Gutman and Jordan Goldmeier, ISBN 9781119741749 © 2021 John Wiley & Sons, Inc., Indianapolis, Indiana. All rights reserved.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

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

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

Die vorliegende Publikation ist urheberrechtlich geschützt. Alle Rechte vorbehalten. Die Verwendung der Texte und Abbildungen, auch auszugsweise, ist ohne die schriftliche Zustimmung des Verlags urheberrechtswidrig und daher strafbar. Dies gilt insbesondere für die Vervielfältigung, Übersetzung oder die Verwendung in elektronischen Systemen.
Es wird darauf hingewiesen, dass die im Buch verwendeten Soft- und Hardware-Bezeichnungen sowie Markennamen und Produktbezeichnungen der jeweiligen Firmen im Allgemeinen warenzeichen-, marken- oder patentrechtlichem Schutz unterliegen.
Alle Angaben und Programme in diesem Buch wurden mit größter Sorgfalt kontrolliert. Weder Autor noch Verlag noch Übersetzer können jedoch für Schäden haftbar gemacht werden, die in Zusammenhang mit der Verwendung dieses Buches stehen.

5 4 3 2 1 0

Inhalt

Vorwort **15**

Einleitung **19**
Die Data-Science-Industrie 19
Warum uns das Thema so wichtig ist 20
Die Krise auf dem US-amerikanischen Subprime-Hypothekenmarkt 20
Die US-Präsidentschaftswahl von 2016 22
Unsere Hypothese 23
Daten am Arbeitsplatz 23
Die berühmte Sitzungssaal-Szene 24
Sie können das große Ganze verstehen 25
Restaurants klassifizieren 25
Ja und? 28
Für wen dieses Buch geschrieben wurde 29
Warum wir dieses Buch geschrieben haben 30
Was Sie lernen werden 31
Wie dieses Buch strukturiert ist 32
Ein letzter Punkt, bevor es wirklich losgeht 33

Teil I Denken wie ein Data Head

1 Was ist das Problem? **37**
Fragen, die ein Data Head stellen sollte 38
Warum ist das Problem wichtig? 38
Wen betrifft das Problem? 40
Was ist, wenn wir nicht die richtigen Daten haben? 41
Wann ist das Projekt zu Ende? 41
Was tun wir, wenn uns die Ergebnisse nicht gefallen? 42

Verstehen, warum Datenprojekte scheitern 42
Szenario: Kundenwahrnehmung 43
Diskussion 44
An den wichtigen Problemen arbeiten 45
Zusammenfassung 46

2 Was sind Daten? 47
Daten oder Informationen? 47
Ein Beispiel-Datensatz 47
Datentypen 49
Wie Daten gesammelt und strukturiert werden 50
Beobachtungsbasierte versus experimentelle Daten 50
Strukturierte versus unstrukturierte Daten 51
Die Basics der zusammenfassenden Statistik 52
Zusammenfassung 53

3 Vorbereitungen für das statistische Denken 55
Stellen Sie Fragen! 56
In allen Dingen ist Variation 57
Szenario: Kundenwahrnehmung (die Fortsetzung) 59
Fallstudie: Nierenkrebsraten 61
Wahrscheinlichkeitsrechnung und Statistik 63
Wahrscheinlichkeit oder Intuition 64
Entdeckungen mit Statistiken 66
Zusammenfassung 68

Teil II Sprechen wie ein Data Head

4 Daten infrage stellen 71
Was würden Sie tun? 72
Katastrophe durch fehlende Daten 74
Erzählen Sie mir die Herkunftsgeschichte der Daten 78
Wer hat die Daten gesammelt? 78
Wie wurden die Daten gesammelt? 79
Sind die Daten repräsentativ? 80
Gibt es eine Stichprobenverzerrung? 80
Wie wurde mit Ausreißern umgegangen? 81
Welche Daten sehe ich nicht? 81
Wie gehen Sie mit fehlenden Werten um? 82
Können die Daten abbilden, was Sie mit ihnen messen wollen? . . . 82
Stellen Sie Daten infrage, egal wie groß die Datenmenge ist 83
Zusammenfassung 83

5 Daten erkunden . . . 85
Ihre Rolle in der explorativen Datenanalyse . . . 86
Wie ein Forscher denken . . . 86
Leitfragen . . . 87
Der Versuchsaufbau . . . 87
Können die Daten Ihre Frage beantworten? . . . 88
Legen Sie Erwartungen fest und benutzen Sie Ihren gesunden Menschenverstand . . . 88
Ergeben die Werte intuitiv einen Sinn? . . . 88
Achtung: Ausreißer und fehlende Werte . . . 92
Sind Ihnen irgendwelche Beziehungen aufgefallen? . . . 93
Korrelation verstehen . . . 93
Achtung: Korrelation falsch interpretieren . . . 94
Achtung: Korrelation bedeutet nicht Kausalität . . . 96
Haben Sie in den Daten neue Einsatzmöglichkeiten oder unentdeckte Potenziale gefunden? . . . 97
Zusammenfassung . . . 97

6 Wahrscheinlichkeiten untersuchen . . . 99
Raten Sie mal . . . 100
Die Spielregeln . . . 101
Schreibweise . . . 101
Bedingte Wahrscheinlichkeit und unabhängige Ereignisse . . . 103
Die Wahrscheinlichkeit mehrfacher Ereignisse . . . 104
Gedankenexperiment zur Wahrscheinlichkeit . . . 107
Die nächsten Schritte . . . 108
Seien Sie vorsichtig bei der Annahme von Abhängigkeiten . . . 109
Fallen Sie nicht auf den Spieler-Fehlschluss herein . . . 110
Alle Wahrscheinlichkeiten unterliegen bestimmten Bedingungen . . . 110
Vertauschen Sie Abhängigkeiten nicht . . . 111
Der Satz von Bayes . . . 112
Stellen Sie sicher, dass die Wahrscheinlichkeiten einen Sinn ergeben . . . 115
Kalibrierung . . . 115
Seltene Ereignisse können und werden eintreffen . . . 116
Zusammenfassung . . . 117

7 Hinterfragen Sie Statistiken . . . 119
Kleine Einführung in die statistische Inferenz . . . 119
Schaffen Sie sich etwas Spielraum . . . 120
Mehr Daten, mehr Evidenz . . . 121
Hinterfragen Sie den Status quo . . . 121
Beweise für das Gegenteil (Evidenz) . . . 122
Entscheidungsfehler ausgleichen . . . 124

Die Vorgehensweise der statistischen Inferenz 125
Die Fragen, die Sie stellen sollten, um Statistiken zu hinterfragen 126
Was ist der Kontext für diese Statistik? 127
Wie groß ist der Stichprobenumfang? 127
Was testen Sie? 128
Wie lautet die Nullhypothese? 128
Wie hoch ist das Signifikanzniveau? 130
Wie viele Tests führen Sie durch? 131
Kann ich bitte die Konfidenzintervalle sehen? 131
Ist dies von praktischer Bedeutung? 132
Gehen Sie von einer Kausalität aus? 133
Zusammenfassung 133

Teil III Den Werkzeugkasten des Data Scientist verstehen

8 Nach versteckten Gruppen suchen 137
Unüberwachtes Lernen 138
Dimensionsreduktion 138
Zusammengefasste Features erstellen 139
Hauptkomponentenanalyse 141
Beispiel: HKA für die sportliche Leistungsfähigkeit 141
Zusammenfassung zur HKA 144
Mögliche Fallen 145
Clustering 146
Clustering mit dem k-Means-Algorithmus 147
Beispiel: Clustering von Verkaufsfilialen 147
Mögliche Fallen 149
Zusammenfassung 151

9 Das Regressionsmodell verstehen 153
Überwachtes Lernen 153
Was macht die lineare Regression? 155
Kleinste-Quadrate-Regression: mehr als nur ein hübscher Name 156
Vorteile der linearen Regression 159
Auf mehrere Features erweitern 160
Probleme und Fallstricke der linearen Regression 161
Unberücksichtigte Variablen 162
Multikollinearität 162
Data Leakage 163

Extrapolationsfehler . . . 164
Viele Beziehungen sind nicht linear . . . 165
Erklärst du noch, oder machst du schon Vorhersagen? . . . 165
Leistungsfähigkeit der Regression . . . 166
Andere Regressionsmodelle . . . 167
Zusammenfassung . . . 168

10 Das Klassifikationsmodell verstehen . . . 169
Einführung in die Klassifikation . . . 169
Was Sie lernen werden . . . 170
Klassifikationsproblem: Versuchsaufbau . . . 171
Logistische Regression . . . 171
Logistische Regression: Na und? . . . 174
Entscheidungsbäume . . . 175
Ensemblemethoden . . . 179
Zufallswälder . . . 179
Gradientenverstärkte Bäume . . . 181
Interpretierbarkeit von Ensemblemethoden . . . 181
Achten Sie auf Fallstricke . . . 182
Falsche Anwendung des Problems . . . 182
Data Leakage . . . 182
Keine Aufteilung der Daten . . . 183
Den richtigen Cut-off-Wert wählen . . . 183
Falsch verstandene Genauigkeit . . . 184
Konfusionsmatrizen . . . 185
Zusammenfassung . . . 187

11 Textanalyse verstehen . . . 189
Erwartungen an die Textanalyse . . . 189
Wie aus Text Zahlen werden . . . 191
Ein großer Sack voll Wörter . . . 191
N-Gramme . . . 194
Worteinbettungen . . . 195
Topic Modeling . . . 198
Textklassifikation . . . 200
Naive Bayes . . . 201
Sentimentanalyse . . . 204
Praktische Überlegungen bei der Arbeit mit Text . . . 204
Die großen Technologiekonzerne haben die Oberhand . . . 205
Zusammenfassung . . . 207

12 Konzepte des Deep Learning 209
Neuronale Netzwerke 210
Worin besteht die Ähnlichkeit zwischen neuronalen Netzwerken und dem Gehirn? 210
Ein einfaches neuronales Netzwerk 211
Wie ein neuronales Netzwerk lernt 213
Ein etwas komplexeres neuronales Netzwerk 214
Anwendungen des Deep Learning 216
Die Vorteile des Deep Learning 218
Wie Computer Bilder »sehen« 219
Neuronale Konvolutionsnetze 220
Deep Learning für Sprache und Wortsequenzen 222
Deep Learning in der Praxis 224
Haben Sie Daten? 224
Sind Ihre Daten strukturiert? 225
Wie wird das Netzwerk aussehen? 225
Die künstliche Intelligenz und Sie 226
Die großen Technologiekonzerne haben die Oberhand 227
Ethik im Deep Learning 228
Zusammenfassung 229

Teil IV Den Erfolg sichern

13 Achten Sie auf Fallstricke 233
Bias und seltsame Datenphänomene 233
Survivorship Bias 234
Regression zur Mitte 235
Das Simpson-Paradoxon 235
Confirmation Bias 237
Effort Bias 237
Algorithmischer Bias 238
Weitere Formen von Bias 239
Die große Liste möglicher Fallstricke 239
Fallstricke der Statistik und des Machine Learning 239
Projektbezogene Fallstricke 241
Zusammenfassung 242

14 Menschen und Persönlichkeiten kennen 243
Sieben Szenarien typischer Kommunikationspannen 243
Das Postmortem 244
Märchenstunde 245
Stille Post 246
Verzettelt 246
Der Realitätsabgleich 247
Die Übernahme 247
Der Angeber 248
Datenpersönlichkeiten 248
Datenenthusiasten 249
Datenzyniker 249
Data Heads 249
Zusammenfassung 250

15 Was kommt danach? 251

Danksagungen 255

Index 257

Für meine Kinder Allie, William und Ellen

Allie war gerade drei Jahre alt, als sie herausfand, dass ihr Vater ein »Doktor« ist. Etwas irritiert sah sie mich an und sagte: »Aber du hilfst den Menschen doch gar nicht …« In diesem Sinne widme ich dieses Buch auch Ihnen, den Leserinnen und Lesern.

Ich hoffe, dass es Ihnen hilft.
– Alex

Für Stephen und Melissa
– Jordan

Vorwort

Werde ein Data Head kommt angesichts der aktuellen Situation der Daten und Analysen in vielen Organisationen genau zur richtigen Zeit. Werfen wir einen kurzen Blick auf die nähere Vergangenheit. Einige wenige führende Unternehmen setzen seit mehreren Jahrzehnten, genauer gesagt seit den 1970er-Jahren, Daten und Analysen effektiv als Orientierungshilfe für ihre Entscheidungen ein. Die meisten anderen Unternehmen haben diese wertvolle Ressource dagegen schlicht ignoriert oder versteckten sie in irgendeinem Hinterzimmer, ohne ihr viel Beachtung zu schenken.

Das begann sich Anfang der 2000er-Jahre zu ändern. Unternehmen begeisterten sich für die Möglichkeiten, ihr Geschäft auf Basis von Daten und Analysemethoden neu aufzustellen. In den frühen 2010er-Jahren verschob sich diese Begeisterung in Richtung *Big Data*, einem Begriff, der ursprünglich von Internetunternehmen stammt, sich aber schnell auf alle fortschrittlichen Wirtschaftszweige ausbreitete. Um mit der ständig steigenden Menge und Komplexität der Daten umgehen zu können, entstand in vielen Unternehmen die Rolle des *Data Scientist* – auch hier zuerst im Silicon Valley und dann überall.

Aber gerade als viele Unternehmen begannen, sich mit Big Data auseinanderzusetzen, verschob sich zwischen 2015 und 2018 das Hauptaugenmerk erneut, und zwar auf künstliche Intelligenz. Das Sammeln, Speichern und Analysieren großer Datenmengen musste Machine Learning, natürlicher Sprachverarbeitung (engl. *Natural Language Processing*, NLP) und Automatisierung weichen.

Eingebettet in diese sehr schnellen Verschiebungen der Aufmerksamkeit, entstand eine Reihe von Annahmen über Daten und Datenanalyse in Unternehmen. Ich bin froh, dass *Werde ein Data Head* viele dieser Annahmen über den Haufen wirft, denn das ist schon lange fällig. Viele, die mit diesen Trends arbeiten oder sie genau beobachten, geben langsam zu, dass sie durch diese Annahmen immer unproduktiver wurden. Im Rest dieses Vorworts werde ich daher fünf miteinander verbundene Annahmen beschreiben und zeigen, auf welche Weise die Ideen in diesem Buch ihnen zu Recht widersprechen.

Annahme 1: Datenanalyse, Big Data und KI sind vollkommen unterschiedliche Phänomene.

Außenstehende gehen oft davon aus, dass »traditionelle« Analysemethoden, Big Data und KI vollkommen eigenständige und unterschiedliche Themenbereiche sind. *Werde ein Data Head* zeigt dagegen, dass diese Themen sogar sehr eng miteinander verknüpft sind. Bei allen geht es um statistisches Denken. Traditionelle Analysemethoden wie die Regressionsanalyse kommen ebenfalls in allen drei Bereichen zum Einsatz und auch in Techniken zur Datenvisualisierung. Die prädiktive Analyse ist im Grunde nichts anderes als überwachtes Machine Learning, und die meisten Techniken zur Datenanalyse funktionieren auf Datensätzen beliebiger Größe. Kurz gesagt: Ein guter Data Head bewegt sich effektiv in allen drei Bereichen und weiß, dass es wenig produktiv ist, sich zu sehr mit den Unterschieden zu beschäftigen.

Annahme 2: Data Scientists sind die einzigen Personen, die in diesem Sandkasten spielen können.

Wir haben Data Scientists über den grünen Klee gelobt und sind oft davon ausgegangen, sie seien die einzigen Menschen, die effektiv mit Daten und Analysen arbeiten können. Tatsächlich findet aktuell eine überaus wichtige Bewegung in Richtung Demokratisierung dieser Ideen statt. Immer mehr Unternehmen setzen auf sogenannte »Laien-Data-Scientists«. Automatische Werkzeuge für das Machine Learning erleichtern die Erstellung hervorragender Vorhersagemodelle. Natürlich gibt es auch weiterhin Bedarf an professionellen Data Scientists, die neue Algorithmen entwickeln und die Arbeit der Laien überwachen, besonders wenn komplexe Analysen durchgeführt werden. Doch Unternehmen, die Analysen und Data Science demokratisieren und »Laien-Data-Heads« einsetzen, können die gesamte Nutzung dieser wichtigen Fähigkeiten deutlich steigern.

Annahme 3: Ein Data Scientist ist eine »Eier legende Wollmilchsau«, die alle Fähigkeiten besitzt, die für diese Aufgaben nötig sind.

Oft wird davon ausgegangen, dass Data Scientists, also Personen, deren Ausbildungs- und Arbeitsschwerpunkt auf der Entwicklung und Programmierung von Modellen liegt, auch alle anderen Aufgaben ausführen können, die für eine vollständige Implementierung dieser Modelle nötig sind. Anders gesagt: Wir stellen sie uns als eine Art Eier legende Wollmilchsau vor, also als wahre Alleskönner. Aber diese Alleskönner gibt es nicht, oder sie sind nur sehr selten anzutreffen. Data Heads, die nicht nur die Feinheiten der Data Science, sondern auch den geschäftlichen Teil kennen, können effektiv Projekte leiten, haben zudem ausgezeichnete Fähigkeiten im Aufbau von Geschäftsbeziehungen und sind eine sehr wertvolle Ressource in Data-Science-Projekten. Sie können produktive Mitglieder von Data-Science-Teams sein und erhöhen die Wahrscheinlichkeit, dass Data-Science-Projekte den Geschäftswert steigern.

Annahme 4: Sie brauchen einen wirklich hohen Intelligenzquotienten und eine Menge Training, um mit Daten und Analysen erfolgreich zu sein.

Eine weitere verwandte Annahme besagt, dass man sehr gut in Data Science ausgebildet sein muss, um in diesem Bereich zu bestehen, und dass ein Data Head sehr gut mit Zahlen umgehen können muss. Training im Umgang mit Zahlen und Sachverstand sind sicher eine Hilfe. *Werde ein Data Head* vertritt allerdings die Meinung (der ich übrigens zustimme), dass man mit etwas Ehrgeiz durchaus in der Lage ist, sich die nötigen Fähigkeiten zu Daten und Analysen anzueignen, um in Data-Science-Projekten nützlich zu sein. Das liegt zum Teil daran, dass die Grundprinzipien statistischer Analyse durchaus keine Raketenwissenschaften sind. Man muss sich nicht einmal extrem gut mit Daten und Analysen auskennen, um in Data-Science-Projekten »nützlich zu sein«. Für die Arbeit mit ausgebildeten Data Scientists oder automatisierten KI-Programmen muss man nur die Fähigkeit und die Neugier besitzen, gute Fragen zu stellen und die Verbindungen zwischen geschäftlichen Themen und quantitativen Ergebnissen herzustellen – ohne dabei auf zweifelhafte Annahmen hereinzufallen.

Annahme 5: Wenn Sie keine quantitativen Fächer (Algebra, Statistik etc.) studiert haben, ist es schon zu spät, sich das für die Arbeit mit Daten und Analysen nötige Wissen anzueignen.

Diese Annahme wird von Umfragedaten gestützt. In einer Umfrage der Internetplattform Splunk aus dem Jahr 2019 unter rund 1.300 Führungskräften weltweit gaben praktisch alle Befragten (98 %) an, dass Fähigkeiten im Umgang mit Daten für die Arbeitsplätze von morgen eine wichtige Rolle spielen.[1] 81 % der Führungskräfte waren außerdem der Ansicht, dass Fähigkeiten im Umgang mit Daten nötig sein werden, um höhere Führungspositionen einzunehmen. 85 % waren sich einig, dass diese Kenntnisse für ihre Unternehmen immer wertvoller werden. Trotzdem gaben 67 % an, sich nicht wohl dabei zu fühlen, selbst auf Daten zuzugreifen oder diese zu nutzen. 73 % glaubten, dass Kenntnisse im Umgang mit Daten schwerer erlernbar sind als andere Geschäftsfähigkeiten. 53 % waren der Meinung, sie wären zu alt, um Fähigkeiten im Umgang mit Daten noch zu erlernen. Dieser »Daten-Defätismus« ist schädlich für Einzelpersonen wie für Unternehmen. Weder die Autoren dieses Buchs noch ich selbst halte ihn für gerechtfertigt. Sehen Sie sich die Seiten nach diesem Vorwort an, und Sie werden feststellen, dass wirklich keine Raketenwissenschaft nötig ist.

Vergessen Sie also diese falschen Annahmen und werden Sie zum Data Head. Ihr Wert als Mitarbeiterin oder Mitarbeiter wird steigen, und Sie werden Ihr Unternehmen erfolgreicher machen. Das ist der Lauf der Welt – es ist an der Zeit, sich damit

1 Splunk Inc., »The State of Dark Data«, 2019, *www.splunk.com/en_us/form/thestate-of-dark-data.html*

vertraut zu machen und mehr über Daten und Analysen zu lernen. Ich bin überzeugt, Sie werden den Prozess und die Lektüre von *Werde ein Data Head* als lohnender und angenehmer empfinden, als Sie es sich vorstellen können.

Thomas H. Davenport
Distinguished Professor, Babson College
Visiting Professor, Oxford Saïd Business School
Research Fellow, MIT Initiative on the Digital Economy
Autor von *Competing on Analytics, Big Data @ Work* und *The AI Advantage*

Einleitung

Ob Sie wollen oder nicht: Daten sind wahrscheinlich der wichtigste Aspekt Ihrer Arbeit. Und sehr wahrscheinlich lesen Sie dieses Buch, um verstehen zu können, worum es überhaupt geht.

Zu Beginn lohnt es sich, noch einmal auszusprechen, was fast schon ein Klischee ist: Wir erzeugen und konsumieren mehr Informationen als jemals zuvor. Wir befinden uns ohne Zweifel im Zeitalter der Daten. Und dieses Zeitalter hat einen ganz eigenen Wirtschaftszweig mit Versprechen, Buzzwords und Produkten hervorgebracht, die Sie, Ihre Vorgesetzten, Ihre Kolleginnen und Kollegen sowie Ihre Mitarbeitenden benutzen oder benutzen werden. Aber trotz aller Behauptungen und weitverbreiteten Datenversprechen und -produkten schlagen Data-Science-Projekte mit alarmierender Häufigkeit fehl.[1]

Damit wollen wir nicht sagen, dass alle Datenversprechen leer und alle Produkte furchtbar sind. Es geht eher darum, dass Sie eine grundsätzliche Wahrheit erkennen müssen, um das Thema wirklich verstehen zu können: Dieses Zeug ist wirklich komplex. Bei der Arbeit mit Daten geht es um Zahlen, feine Unterschiede und Unsicherheit. Sicher, Daten sind wichtig, aber selten einfach. Und trotzdem gibt es eine ganze Branche, die versucht, uns etwas anderes zu erzählen. Eine Branche, die uns Sicherheit in einer unsicheren Welt verspricht und mit der Angst der Unternehmen spielt, etwas zu verpassen. Wir, die Autoren, nennen dies die Data-Science-Industrie.

Die Data-Science-Industrie

Dieses Problem betrifft alle Beteiligten. Unternehmen suchen ständig nach Produkten, die ihnen das Denken abnehmen. Manager stellen Analyseprofis ein, die in Wirklichkeit keine sind. Data Scientists werden von Unternehmen angeheuert, die eigentlich noch gar nicht dafür bereit sind. Führungskräfte werden gezwungen,

1 Venture Beat. »87% of data science projects failing«: *https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production*

sich technologisches Fachchinesisch anzuhören und so zu tun, als verstünden sie alles Gesagte. Projekte geraten in Stocken, Geld wird verschwendet.

Gleichzeitig spuckt die Data-Science-Industrie schneller neue Konzepte aus, als wir in der Lage sind, die neu geschaffenen Möglichkeiten (und Probleme) zu erfassen und auf den Punkt zu bringen. Ein Augenblick – und schon ist wieder eine Chance verpasst. Als die Autoren ihre Zusammenarbeit begannen, war *Big Data* das große Zauberwort. Im Laufe der Zeit wurde dann *Data Science* das neue Thema. Mittlerweile liegt das Hauptaugenmerk auf Dingen wie *Machine Learning*, *Deep Learning* und *künstlicher Intelligenz*.

Für die neugierigen und kritischen Denker unter uns scheint hier irgendetwas nicht zu stimmen. Sind diese Problemstellungen wirklich neu? Oder sind die neuen Begriffe nur alter Wein in neuen Schläuchen?

Die Antwort lautet für beide Fragen natürlich: Ja.

Die größere und wichtigere Frage, die Sie sich hoffentlich stellen, lautet allerdings: *Wie kann ich kritisch über Daten denken und sprechen?*

Genau das wollen wir Ihnen hier beibringen.

Mit diesem Buch geben wir Ihnen die Werkzeuge, Fachbegriffe und Denkweisen an die Hand, die nötig sind, um sich in der Data-Science-Branche zu orientieren und die gesteckten Ziele zu erreichen. Sie werden ein tieferes Verständnis für Daten und ihre Herausforderungen entwickeln. Sie werden lernen, kritisch über Daten und die gefundenen Ergebnisse zu denken, und Sie werden in der Lage sein, informiert und klug über alles zu sprechen, was mit Daten zu tun hat.

Kurz gesagt, Sie werden ein *Data Head*.

Warum uns das Thema so wichtig ist

Bevor wir uns mit den Details befassen, ist es sinnvoll, zu verstehen, warum Ihren Autoren Alex und Jordan dieses Thema so sehr am Herzen liegt. In diesem Abschnitt zeigen wir Ihnen zwei wichtige Beispiele dafür, wie Daten Einfluss auf große Teile der Gesellschaft und uns persönlich genommen haben.

Die Krise auf dem US-amerikanischen Subprime-Hypothekenmarkt

Wir kamen gerade frisch vom College, als die Subprime-Hypothekenkrise über uns hereinbrach. 2009, in einer Zeit, in der es schwer war, überhaupt einen Job zu bekommen, schafften wir es beide, Arbeit bei der Air Force zu bekommen. Wir hatten beide Glück, weil wir eine sehr gefragte Fähigkeit besaßen: Wir konnten mit Daten umgehen. Tagein, tagaus arbeiteten wir mit Daten, um die Forschung von Air-Force-Analysten und -Wissenschaftlern in Produkte zu verwandeln, mit denen die Regierung etwas anfangen konnte. Unsere Anstellung sollte zu einem Vorboten

der Aufmerksamkeit werden, die das ganze Land bald den von uns ausgefüllten Rollen widmen sollte. Als zwei Datenanalysten betrachteten wir die Hypothekenkrise mit Interesse und Neugier.

Zum Entstehen der Subprime-Hypothekenkrise trug eine Reihe verschiedener Faktoren bei.[2] In unserem Versuch, sie als Beispiel zu verwenden, wollen wir weitere Faktoren nicht ignorieren. Dennoch sehen wir, vereinfacht gesagt, die Krise als einen großen Datenfehler. Banken und Investoren erstellten Modelle, um den Wert von hypothekarisch abgesicherten Schuldverschreibungen (engl. *Mortgage-backed Collateralized Debt Obligations*, CDOs) zu verstehen. Vielleicht erinnern Sie sich, dass genau dieses Investitionsmodell für den Zusammenbruch der Märkte in den Vereinigten Staaten verantwortlich war.

CDOs wurden als sichere Investition angesehen, weil das Kreditausfallrisiko auf mehrere Investitionseinheiten verteilt wird. Der Gedanke war, dass in einem Portfolio von Hypotheken der Ausfall einiger weniger Hypotheken keine wesentlichen Auswirkungen auf den zugrunde liegenden Wert des gesamten Portfolios haben würde.

Und trotzdem wissen wir mittlerweile, dass einige grundlegende Annahmen falsch waren. Am schwersten wog die Fehleinschätzung, dass Kreditausfälle voneinander unabhängige Ereignisse waren. Wenn Person A ihren Kredit nicht zurückzahlen kann, hat das keinen Einfluss auf Person B – dachte man. Wenig später mussten wir lernen, dass Kreditausfälle eher wie Dominosteine funktionieren, bei denen ein vorheriger Ausfall ein Anzeichen für weitere Ausfälle ist. Sobald eine Hypothek geplatzt war, sanken in der Folge die Immobilienpreise in der Umgebung, und das Risiko für weitere Ausfälle in dieser Wohngegend stieg. Durch den Kreditausfall wurden die benachbarten Häuser mit in den Abgrund gerissen.

Von Unabhängigkeit auszugehen, wenn die Ereignisse tatsächlich einen Zusammenhang haben, ist ein häufig anzutreffender Fehler in der Statistik.

Aber tauchen wir noch etwas tiefer in die Geschichte ein. Investmentbanken hatten ein Modell geschaffen, das Investitionen überbewertete. Ein Modell ist ein absichtlich stark vereinfachtes Abbild einer realen Situation. Es basiert auf Annahmen über die echte Welt, um bestimmte Phänomene besser zu verstehen und Vorhersagen darüber zu treffen. Auf Modelle werden wir weiter unten im Buch noch genauer eingehen.

Und wer waren die Leute, die dieses Modell erstellt und verstanden haben? Das waren genau diejenigen, die die Grundlagen für ein Berufsbild geschaffen haben, das wir heute als *Data Scientist* bezeichnen. Leute wie wir. Statistiker, Ökonomen, Physiker – Leute, die sich mit Machine Learning, künstlicher Intelligenz und Statistik befassen. Sie arbeiteten mit Daten. Sie waren schlau. Superschlau.

Und trotzdem ging etwas schief. Haben sie nicht die richtigen Fragen zu ihrer Arbeit gestellt? Gingen die Risikoeinschätzungen bei einer Runde »Stille Post« in den

2 *www.brookings.edu/wp-content/uploads/2016/06/11_origins_crisis_baily_litan.pdf*

Telefonaten zwischen Analysten und Entscheidungsträgern verloren? Wurde die Unsicherheit in jeder Runde des Spiels immer weiter zur Seite geschoben, bis der Eindruck eines perfekt vorhersagbaren Wohnungsmarkts entstand? Oder haben die Beteiligten über die tatsächlichen Ereignisse einfach gelogen?

Für uns persönlich ist die Frage viel wichtiger, wie wir ähnliche Fehler bei unserer eigenen Arbeit vermeiden können.

Wir hatten viele Fragen und konnten über die Antworten nur spekulieren. Eine Sache aber war klar: Hier geschah eine flächendeckende Datenkatastrophe. Und es würde nicht die letzte sein.

Die US-Präsidentschaftswahl von 2016

Am 8. November 2016 gewann der republikanische Kandidat Donald J. Trump die Präsidentschaftswahl in den USA gegen die vermeintliche Spitzenkandidatin und demokratische Herausforderin Hillary Clinton. Für die politischen Meinungsforscher war das ein Schock. Ihre Modelle hatten seinen Sieg nicht vorhergesagt. Und ausgerechnet das sollte das Jahr der Wahlvorhersagen sein.

Im Jahr 2008 gelang dem Blog *FiveThirtyEight* von Nate Silver – damals noch Teil der New York Times – eine erstaunlich genaue Vorhersage von Barack Obamas Wahlgewinn. Zu der Zeit waren die Experten noch skeptisch, dennoch sagte Silvers Algorithmus das Wahlergebnis korrekt voraus. 2012 stand Silver erneut im Rampenlicht, weil er einen weiteren Sieg für Barack Obama richtig vorhergesagt hatte.

Zu dieser Zeit begann die Geschäftswelt, Daten als wichtig anzusehen und Data Scientists einzustellen. Die erfolgreiche Vorhersage der Wiederwahl von Barack Obama durch Nate Silver verstärkte noch die Bedeutung der fast orakelhaften Fähigkeiten datenbasierter Vorhersagen. Artikel in Businessmagazinen warnten Führungskräfte vor der Gefahr, von Mitbewerbern geschluckt zu werden, wenn diese ihr Geschäft datenbasiert betrieben, das eigene Unternehmen aber nicht. Die Data-Science-Industrie nahm richtig Fahrt auf.

Bis zum Jahr 2016 hatte jede größere Nachrichtenagentur in Vorhersagealgorithmen investiert, um das Ergebnis der nächsten Präsidentschaftswahlen vorauszuberechnen. Die allergrößte Mehrheit der Modelle sah einen überwältigenden Sieg der demokratischen Kandidatin Hillary Clinton voraus. Oh, wie falsch sie lagen!

Vergleichen wir das mit der Subprime-Hypothekenkrise. Man sollte davon ausgehen, dass man viel aus der Vergangenheit hätte lernen können. Das Interesse an Data Science hätte dazu führen müssen, dass Fehler vermieden werden. Und das stimmt auch: Seit 2008 und 2012 haben Nachrichtenagenturen Data Scientists eingestellt, in Umfrageforschung investiert, Datenteams geschaffen und mehr Geld für gute Daten ausgegeben.

Das führt uns nun zu der Frage: Was ist trotz dieses Einsatzes an Zeit, Geld, Aufwand und Ausbildung denn nun wirklich passiert?[3]

Unsere Hypothese

Warum gibt es Datenprobleme wie diese? Wir sehen drei Gründe: schwer zu lösende Probleme, Mangel an kritischem Denken und schlechte Kommunikation.

Erstens, wie bereits gesagt: *Dieses Zeug ist komplex*. Viele Datenprobleme sind äußerst schwer zu lösen – selbst mit einer Menge Daten und den richtigen Werkzeugen. Auch mit den besten Vorgehensweisen und den schlausten Analysten treten Fehler auf. Vorhersagen können und werden danebenliegen. Das ist einfach so.

Zweitens haben einige Analysten und Entscheider aufgehört, kritisch über Datenprobleme nachzudenken. Die Data-Science-Industrie zeichnete in ihrer Selbstüberschätzung ein Bild von Sicherheit und Einfachheit, und einige Menschen nahmen einfach alles für bare Münze. Vielleicht ist es auch nur menschlich, nicht zugeben zu wollen, dass man keine Ahnung davon hat, was gerade wirklich passiert. Dabei darf man sich nichts vormachen: Beim Nachdenken über Daten und deren Einsatz kann es auch zu falschen Entscheidungen kommen. Das bedeutet, Risiken und Unwägbarkeiten müssen klar kommuniziert werden. Aus irgendeinem Grund ist diese Nachricht wohl untergegangen. Obwohl wir eigentlich gehofft hatten, dass der enorme Fortschritt bei der Erforschung und Anwendung von Datenanalysen das kritische Denken aller schärft, hat es bei einigen eher zu einer kompletten Abschaltung geführt.

Der dritte Grund, warum Datenprobleme unserer Meinung nach auftreten, ist schlechte Kommunikation zwischen Data Scientists und Entscheidern. Trotz bester Absichten gehen Ergebnisse oft auf dem Weg der Übersetzung verloren. Nur selten sprechen Entscheider die Sprache der Data Scientists, weil sich niemand die Arbeit gemacht hat, sie ihnen beizubringen. Und ganz ehrlich: Datenanalysten sind nicht unbedingt gut darin, Dinge zu erklären. Hier gibt es eine klare Kommunikationslücke.

Daten am Arbeitsplatz

Ihre Datenprobleme werden vielleicht nicht die Weltwirtschaft zum Einsturz bringen oder den nächsten Präsidenten der Vereinigten Staaten falsch vorhersagen. Dennoch ist der Kontext dieser Geschichten wichtig. Wenn schlecht kommuniziert wird, wenn Missverständnisse und Versäumnisse beim kritischen Denken auftreten, während die Welt zusieht, dann passiert das sehr wahrscheinlich auch an Ihrem Arbeitsplatz. In den meisten Fällen sind diese Fehlschläge nur winzig. Dennoch fördern sie eine Kultur mangelnder Datenkompetenz.

3 Nate Silver hat eine Reihe von Artikeln verfasst, in denen er dies sehr detailliert beschreibt (*https://fivethirtyeight.com/tag/the-real-story-of-2016*). Ein Fehler war, dass die Meinungsforscher fälschlicherweise von Unabhängigkeit ausgingen, genau wie bei der Hypothekenkrise.

Das ist auch an unserem Arbeitsplatz schon passiert, und es war teilweise unsere eigene Schuld.

Die berühmte Sitzungssaal-Szene

Fans von Science-Fiction- und Abenteuerfilmen kennen diese Szene nur zu gut: Der Held muss eine scheinbar unlösbare Aufgabe bewältigen, also kommen die weltweit führenden Politiker und Wissenschaftler zusammen, um die Situation zu diskutieren. Ein besonders verschrobener Wissenschaftler breitet in einem Schwall unverständlicher Fachbegriffe einen Vorschlag aus, worauf der General bellt: »Sprechen Sie Englisch!« An dieser Stelle erhält der Zuschauer eine Erklärung dessen, was tatsächlich gemeint ist. Die Idee hinter dieser typischen Szene ist, die missionskritischen Informationen in etwas zu übersetzen, das nicht nur unser Held, sondern auch der Zuschauer verstehen kann.

Diese typische Filmszene haben wir in unserer Rolle als Forscher für die US-Regierung oft diskutiert. Warum? Weil sie nie auf diese Weise stattgefunden hat. In der Tat war das, was wir zu Beginn unserer Laufbahn erlebten, oft das Gegenteil dieses Filmmoments.

Die Reaktionen auf unsere Arbeitsergebnisse waren leere Blicke, unmotiviertes Kopfnicken und vereinzelte schwere Augenlider. Wir konnten beobachten, wie ein verwirrtes Publikum das von uns Gesagte ohne jede Rückfrage akzeptierte. Die Zuhörer waren entweder von unserer Schlauheit beeindruckt oder gelangweilt, weil sich nichts verstanden. Niemand forderte uns auf, das Gesagte in allgemein verständlicher Sprache zu wiederholen. Stattdessen unterschied sich die Situation davon dramatisch. Oft begann es wie folgt:

> *Wir: »Basierend auf unserer überwachten Lernanalyse der binären Antwortvariablen unter Verwendung multipler logistischer Regression konnten wir eine Out-of-Sample-Performance mit einer Spezifität von 0,76 und mehrere statistisch signifikante unabhängige Variablen auf Basis eines 95-prozentigen Signifikanzniveaus feststellen.«*
>
> *Geschäftsleute: *betretenes Schweigen**
>
> *Wir: »Haben Sie das verstanden?«*
>
> *Geschäftsleute: *mehr betretenes Schweigen**
>
> *Wir: »Haben Sie irgendwelche Fragen?«*
>
> *Geschäftsleute: »Im Moment keine Fragen.«*
>
> *Geschäftsleute (interner Monolog): »Was zur Hölle erzählen die da?«*

Würden Sie sich diese Szene in einem Film ansehen, könnten Sie denken: »*Moment, noch mal zurückspulen, vielleicht habe ich etwas übersehen …*« Im wahren Leben, wenn Entscheidungen zu Erfolg oder Misserfolg einer Mission führen können, passiert das jedoch nur selten. Wir spulen nicht zurück. Wir bitten nicht um eine Erklärung.

Im Nachhinein betrachtet, waren unsere Präsentationen zu technisch. Einer der Gründe dafür war reine Sturheit. Wie wir lernen mussten, wurden technische Details vor der Hypothekenkrise zu stark vereinfacht. Analysten wurden engagiert, um den Entscheidern zu sagen, was sie hören wollten. Da wollten wir nicht mitspielen. Unser Publikum würde uns zuhören *müssen*.

Tatsächlich haben wir zu stark gegengesteuert. Unsere Zuhörer setzen sich nicht kritisch mit unserer Arbeit auseinander, weil sie das Gesagte einfach nicht verstanden.

Wir dachten, es müsse einen besseren Weg geben. Wir wollten mit unserer Arbeit etwas verändern. Also übten wir, uns gegenseitig und anderen Zuhörern komplexe statistische Konzepte zu erklären. Und wir begannen zu erforschen, was andere von unseren Erklärungen hielten.

Wir haben eine gemeinsame Ebene zwischen Datenanalysten und Geschäftsleuten entdeckt, auf der ehrliche Diskussionen über Daten geführt werden können, ohne zu technisch oder zu stark vereinfachend zu formulieren. Hierfür müssen beide Seite Datenprobleme kritischer betrachten, unabhängig von ihrer Größe. Und genau darum geht es in diesem Buch.

Sie können das große Ganze verstehen

Um Daten und die Arbeit damit besser zu verstehen, müssen Sie bereit sein, augenscheinlich komplizierte Data-Science-Konzepte zu lernen. Und wenn Sie diese Konzepte schon kennen, bringen wir Ihnen bei, wie Sie sie für Ihr Publikum aus Entscheidern und Geschäftsleuten übersetzen können.

Hierfür müssen Sie sich mit einem Aspekt der Daten auseinandersetzen, über den eher selten gesprochen wird: warum sie in vielen Unternehmen weitgehend versagen. Sie werden Intuition, Wertschätzung und eine gesunde Skepsis gegenüber den Zahlen und Begriffen entwickeln, die Ihnen begegnen werden. Auf den ersten Blick kann das ziemlich einschüchternd wirken. Trotzdem werden wir Ihnen in diesem Buch zeigen, wie das funktioniert. Und dafür müssen Sie weder programmieren, noch brauchen Sie einen Doktortitel.

Mit klaren Erklärungen, Denkübungen und Analogien helfen wir Ihnen beim Aufbau eines mentalen Grundgerüsts aus Data Science, Statistik und Machine Learning.

Genau das tun wir im folgenden Beispiel.

Restaurants klassifizieren

Stellen Sie sich vor, Sie gehen spazieren und kommen an einem leeren Ladenlokal vorbei mit dem Schild: »Restaurant, Neueröffnung demnächst«. Sie sind es leid, bei großen Restaurantketten zu essen, und halten daher die Augen offen nach neuen

Restaurants mit lokalen Eigentümern. Daher stellen Sie sich die Frage: »Wird hier ein neues lokales Restaurant eröffnet?«

Lassen Sie uns die Frage etwas formaler stellen: Können Sie vorhersagen, ob das neue Restaurant zu einer großen Kette gehört oder unabhängig betrieben wird?

Raten Sie mal. (Im Ernst, raten Sie, bevor Sie weiterlesen.)

Im wahren Leben hätten Sie in Sekundenbruchteilen eine ziemlich verlässliche Ahnung. Gingen Sie in einem trendigen Kiez mit Kneipen, Bistros und Restaurants spazieren, würden Sie eher auf ein unabhängiges Restaurant tippen. Befänden Sie sich direkt neben der Umgehungsstraße und in der Nähe eines großen Einkaufszentrums, würden Sie eher mit dem Restaurant einer Kette rechnen.

Dennoch haben Sie gezögert, als wir die Frage stellten. Sie dachten: »*Die haben mir nicht genug Informationen gegeben.*« Und Sie hatten recht. Wir hatten Ihnen nicht genug Daten gegeben, um eine Entscheidung zu treffen.

Die Schlussfolgerung: Fundierte Entscheidungen brauchen Daten.

Und jetzt sehen Sie sich die Daten in Abbildung E-1 an. Das neue Restaurant ist mit einem X markiert, die Cs bezeichnen Kettenrestaurants, die Is unabhängige lokale Gastronomie. Wie würden Sie diesmal entscheiden?

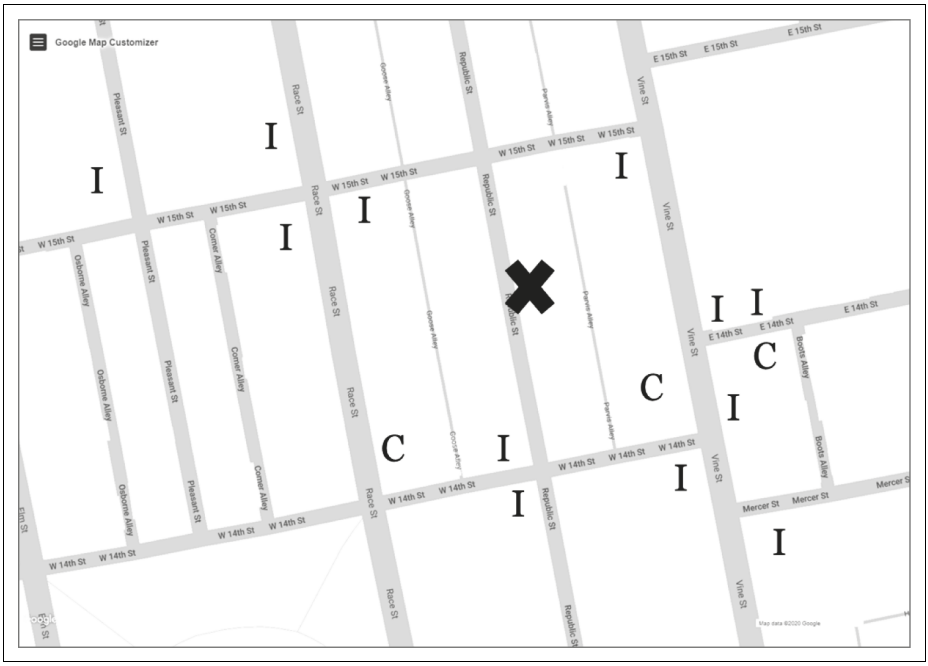

Abbildung E-1: Das Stadtviertel Over the Rhine in Cincinnati, Ohio

Die meisten Menschen tippen hier auf (I), weil die meisten Restaurants in der Umgebung ebenfalls unabhängig (I) sind. Das gilt aber nicht für alle gastronomischen

Angebote in der Umgebung. Wenn wir Sie bitten, auf einer Skala von 0 bis 100 anzugeben, wie sicher Sie sich mit Ihrer Vorhersage sind, würden wir einen ziemlich hohen Wert, aber nicht 100 erwarten. Es ist durchaus möglich, dass sich ein weiteres Kettenrestaurant im Stadtviertel ansiedelt.

Schlussfolgerung: Vorhersagen sollten nie mit hundertprozentiger Sicherheit getroffen werden.

Jetzt sehen Sie sich die Daten in Abbildung E-2 an. In dieser Gegend gibt es ein großes Einkaufszentrum, die meisten Restaurants in der Umgebung werden von großen Ketten betrieben. Als wir hier nach einer Vorhersage fragten, tippte die Mehrheit auf ein weiteres Kettenrestaurant (C). Dennoch freuen wir uns, wenn sich jemand bei der Frage für (I) entscheidet, weil es mehrere wichtige Erkenntnisse aufzeigt.

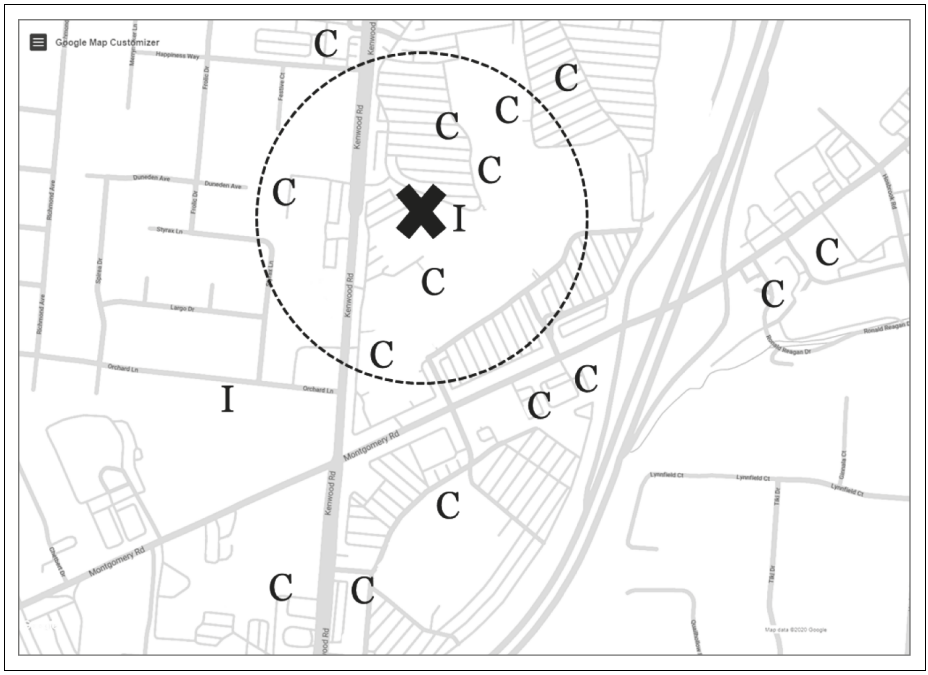

Abbildung E-2: Kenwood Towne Centre, Cincinnati, Ohio

Bei diesem Gedankenexperiment erstellt sich jeder einen etwas anderen *Algorithmus* im Kopf. Natürlich betrachten alle die Markierungen, die unseren Punkt X umgeben, um das Stadtviertel besser zu verstehen. Irgendwann müssen Sie aber entscheiden, wann ein Restaurant zu weit entfernt ist, um Einfluss auf Ihre Entscheidung zu haben. In einem Extremfall (und der ist tatsächlich schon passiert) sieht sich jemand nur den nächsten Nachbarn des neuen Restaurants an, in diesem Fall ein unabhängiges Restaurant, und trifft allein auf dieser Basis eine Vorhersage: »Der nächste Nachbar von X ist ein I, daher ist meine Vorhersage auch ein I.«