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?

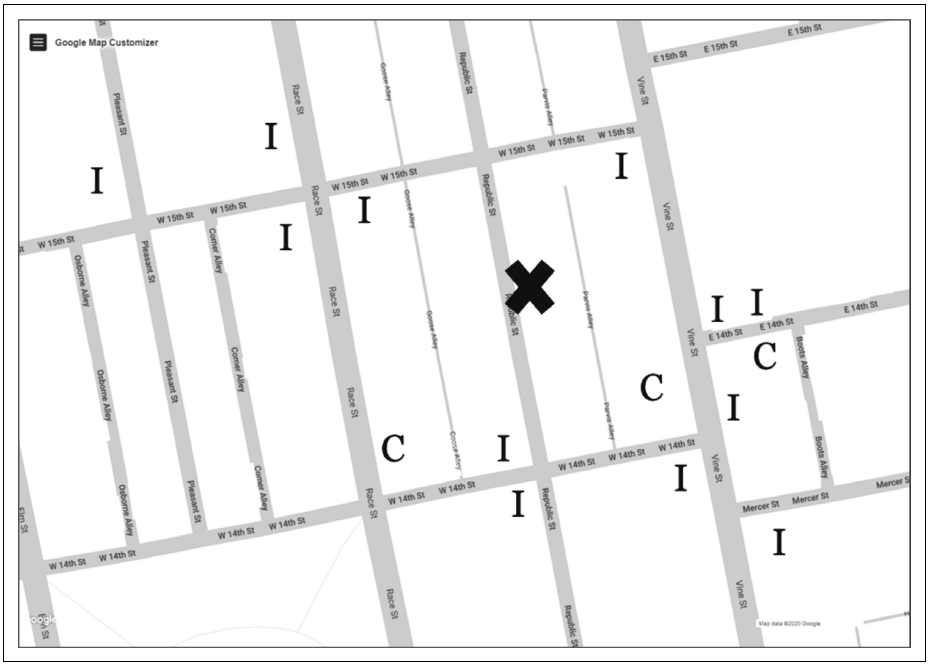

Abbildung E-1: Das Stadtviertel Over the Rhine in Cincinnati, Ohio

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.

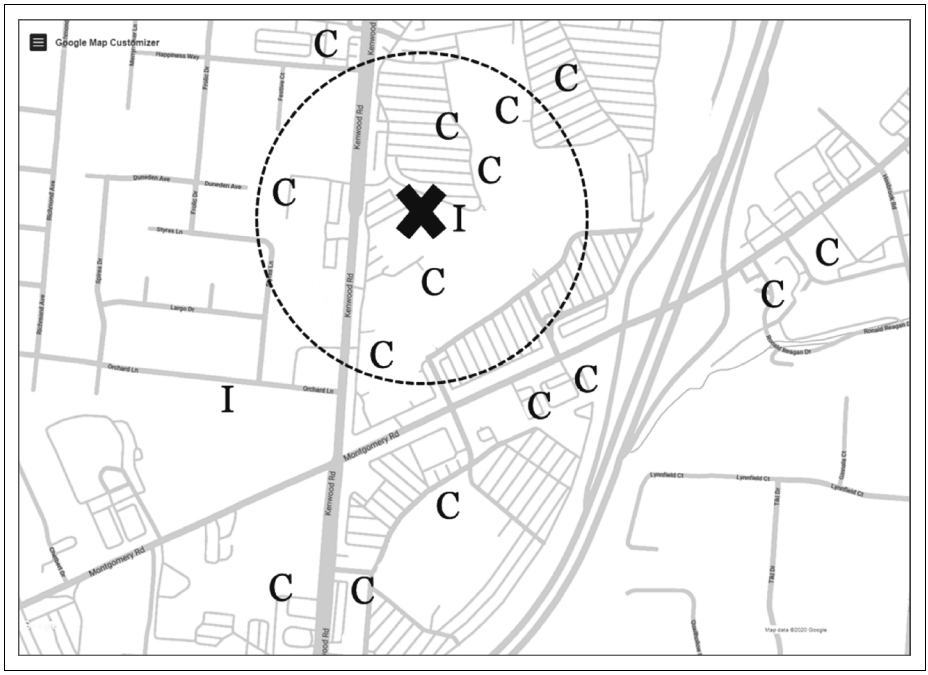

Abbildung E-2: Kenwood Towne Centre, Cincinnati, Ohio

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.«

Die meisten Leute sehen sich allerdings mehrere Restaurants in der Nachbarschaft an. Das zweite Bild zeigt einen Kreis um das neue Restaurant und die sieben nächsten Nachbarn. Vielleicht wählen Sie eine andere Zahl, bei uns waren es sieben. Sechs dieser sieben waren C-Restaurants. Daher lautet unsere Vorhersage, dass das neue Restaurant auch zu einer großen Kette gehören wird.

Ja und?

Wenn Sie das Restaurant-Beispiel verstanden haben, sind Sie bereits auf einem guten Weg, ein Data Head zu werden. In der folgenden Liste zeigen wir Ihnen Schritt für Schritt, was Sie bereits alles gelernt haben:

- Sie haben eine *Kassifikation* vorgenommen, indem Sie das *Label* (Kette oder unabhängig) eines neuen Restaurants vorhergesagt haben. Hierfür haben Sie einen *Algorithmus* anhand eines Datensatzes (Standorte der Restaurants und deren Labels [Kette oder unabhängig]) *trainiert*.
- *Genau das ist Machine Learning!* Nur haben Sie den Algorithmus nicht auf einem Computer erstellt, sondern dafür Ihren Kopf benutzt.
- Genauer gesagt, haben wir hier eine Art maschinelles Lernen angewandt, die als *überwachtes Lernen* (*Supervised Learning*) bezeichnet wird. Das Lernen war »überwacht«, weil Sie vorher wussten, dass die schon vorhandenen Restaurants entweder einer Kette (C) angehörten oder unabhängig waren (I). Die *Labels* haben Ihr Denken dahin gehend gelenkt (d. h. »überwacht«), inwiefern der Standort eines Restaurants damit zusammenhängt, ob es zu einer Kette gehört oder unabhängig betrieben wird.
- Noch genauer gesagt, haben Sie einen *Klassifikationsalgorithmus des überwachten Lernens* verwendet, der als *k-nächste-Nachbarn*[4] bezeichnet wird: Wenn K = 1, verwende das Label des nächsten Restaurants als deine Vorhersage. Wenn K = 7, verwendete die Mehrheit der Labels der nächsten sieben Restaurants als deine Vorhersage. Das ist ein intuitiver und mächtiger Algorithmus – aber sicher keine Magie.
- Sie haben außerdem gelernt, dass Sie Daten brauchen, um fundierte Entscheidungen treffen zu können. Tatsächlich benötigen Sie allerdings noch mehr als das. Schließlich geht es in diesem Buch um kritisches Denken. Wir wollen zeigen, wie die Dinge funktionieren, aber auch, warum sie scheitern. Wir haben Sie gebeten, auf Basis der Daten in den Abbildungen in dieser Einleitung Vorhersagen zu treffen. Ob das neue Restaurant kinderfreundlich sein wird, lässt sich dagegen nicht beantworten. Um fundierte Entscheidungen treffen zu können, reicht es nicht, einfach irgendwelche Daten zu verwenden. Sie brauchen akkurate, relevante und eine ausreichende Menge an Daten.

4 *K-nächste-Nachbarn* kann auch zur Vorhersage von Zahlen anstelle von Klassen verwendet werden. Diese werden als *Regressionsprobleme* bezeichnet und später in diesem Buch behandelt.

- Erinnern Sie sich an die Fachtermini von vorhin? »*... überwachte Lernanalyse auf Basis einer binären Antwortvariablen ...*«? Herzlichen Glückwunsch! Sie haben gerade eine überwachte Lernanalyse einer binären Antwortvariablen durchgeführt. *Antwortvariable* ist einfach nur ein anderes Wort für *Label*. Das Label ist binär, weil es nur zwei mögliche Optionen gibt: Ein Restaurant gehört einer Kette an oder ist unabhängig.

Sie haben in diesem Abschnitt eine Menge gelernt und haben es nicht einmal gemerkt.

Für wen dieses Buch geschrieben wurde

Wie bereits zu Beginn gesagt, beeinflussen Daten das Leben vieler Menschen in der heutigen Zeit. Wir haben folgende *Avatare* für alle diejenigen gefunden, die davon profitieren können, ein *Data Head* zu werden:

- **Michelle** ist Marketingexpertin, die eng mit einem Datenanalysten zusammenarbeitet. Sie entwickelt Marketinginitiativen, wobei ihr Data-Science-Mitarbeiter Daten sammelt, um die Auswirkungen der Initiativen zu messen. Michelle ist der Meinung, die Arbeit könne insgesamt innovativer sein. Ihr fehlt aber die Fähigkeit, ihrem Mitarbeiter effektiv zu erklären, welche Daten und Analysen sie wirklich braucht. Die Kommunikation zwischen beiden ist eine Herausforderung. Sie hat einige der aktuellen Buzzwords gegoogelt (»Machine Learning« und »prädiktive Analyse«). Die meisten gefundenen Artikel sind jedoch zu technisch formuliert, enthalten unentwirrbaren Computercode oder waren Werbeanzeigen für Analysesoftware oder Beraterdienstleistungen. Nach ihrer Suche ist sie noch besorgter und stärker verunsichert als zuvor.
- **Doug** hat einen Doktortitel in Biologie und arbeitet in der Forschungs- und Entwicklungsabteilung eines Großunternehmens. Von Natur aus ein Skeptiker, fragt er sich, ob die neuesten Trends in der Data Science nur einen leeren Hype darstellen. Bei der Arbeit hält Doug sich jedoch mit seiner Skepsis zurück, besonders wenn sein neuer Chef in der Nähe ist, der gern mit einem »Daten sind der neue Speck«-T-Shirt herumläuft. Doug möchte nicht als Datenverweigerer rüberkommen. Gleichzeitig fühlt er sich ausgeschlossen und entscheidet sich, herauszufinden, ob an all dem Gerede wirklich etwas dran ist.
- **Regina** ist Führungskraft in der Chefetage und ist sich der neuesten Entwicklungen in der Data Science bewusst. Sie leitet die neue Data-Science-Abteilung und kommuniziert regelmäßig mit den leitenden Data Scientists. Regina vertraut ihren Data Scientists und setzt sich für deren Arbeit ein. Gleichzeitig wünscht sie für sich selbst ein besseres Verständnis der Arbeit ihres Teams, weil sie diese häufig vor dem Vorstand des Unternehmens präsentieren und verteidigen muss. Außerdem hat sie die Aufgabe, neue Softwaretechnologien für das Unternehmen zu bewerten. Dabei hat sie allerdings den Verdacht, dass die Versprechen einiger Hersteller zur »künstlichen Intelligenz« etwas zu voll-

mundig sind, um wahr zu sein. Aus diesem Grund möchte sie sich mit zusätzlichem technischem Wissen wappnen, um die Marketingversprechen besser von der Realität unterscheiden zu können.

- **Nelson** leitet in seinem neuen Job ein Team aus drei Data Scientists. Als ausgebildeter Informatiker weiß Nelson, wie man programmiert und mit Daten arbeitet. Statistik (abgesehen von einem Kurs, den er am College belegt hat) und Machine Learning sind ihm aber noch fremd. Wegen seines eher technischen Hintergrunds möchte und kann er die Details lernen, findet aber einfach nicht die Zeit dafür. Seine Vorgesetzten drängen außerdem darauf, dass sein Team »mehr mit Machine Learning« macht. Im Moment erscheint ihm das Thema aber noch wie ein Buch mit sieben Siegeln. Nelson sucht nach Material, das ihm hilft, in seinem Team mehr Glaubwürdigkeit zu bekommen und zu erkennen, welche Probleme tatsächlich mit Machine Learning gelöst werden können und welche nicht.

Hoffentlich können Sie sich mit einer der drei vorgestellten Personas identifizieren. Das wiederkehrende Thema bei allen dreien – und hoffentlich auch bei Ihnen – ist der Wunsch, ein besserer »Konsument« der Daten und Analysen zu werden, die Ihnen begegnen werden.

Zusätzlich haben wir noch einen Avatar geschaffen, der für Personen steht, die dieses Buch lesen sollten, es aber sehr wahrscheinlich nicht tun (schließlich braucht jede gute Geschichte auch einen Bösewicht):

- **George**, als Manager der mittleren Ebene, liest viele Artikel in Businessmagazinen über künstliche Intelligenz. Seine Lieblingsartikel leitet er als Beweis für seinen technischen Sachverstand an die oberen Führungsebenen weiter. Auf Vorstandssitzungen ist er dagegen stolz, auf sein »Bauchgefühl« zu hören. George möchte, dass ihm »seine« Data Scientists die Zahlen in maximal ein bis zwei Folien vorlegen. Stimmt die Analyse mit dem überein, was er (auch aufgrund seines Bauchgefühls) bereits vor Durchführung der Studie entschieden hat, leitet er sie weiter nach oben und kann vor seinen Kollegen damit angeben, wie er und sein Bauch es geschafft haben, ein »KI-Unternehmen« auf die Beine zu stellen. Passen Analyse und Bauchgefühl dagegen nicht zusammen, »verhört« er seine Data Scientists mit einer Reihe nebulöser Fragen und schickt sie dann auf eine aussichtslose Jagd nach den von ihm benötigten »Beweisen«, damit er das Projekt weiterzuführen kann.

Seien Sie nicht wie George. Wenn Sie einen dieser Georges kennen, empfehlen Sie ihm dieses Buch und sagen ihm, er erinnere Sie an Regina.

Warum wir dieses Buch geschrieben haben

Wir glauben, dass viele Menschen – wie unsere Avatare – mehr über Daten erfahren möchten, aber einfach nicht wissen, wo sie anfangen sollen. Vorhandene Bücher zu Data Science und Statistik decken ein weites Spektrum ab. Auf der einen

Seite dieses Spektrums gibt es die nicht technischen Bücher, die die Vorteile und Versprechen der Daten anpreisen. Einige sind besser als andere, aber viele wurden von Journalisten geschrieben, die versuchten, die »aufziehende Datendämmerung« zu dramatisieren.

Diese Bücher beschreiben, wie bestimmte Geschäftsprobleme gelöst wurden, indem man sie aus Sicht der Daten betrachtete. Vermutlich kommen sogar Begriffe wie künstliche Intelligenz, Machine Learning und so weiter darin vor. Verstehen Sie uns bitte nicht falsch, diese Bücher schaffen Aufmerksamkeit. Allerdings gehen sie meist nicht sehr in die Tiefe, sondern betrachten das Problem und seine Lösung eher allgemein.

Auf der anderen Seite des Spektrums finden wir hoch technische Bücher. Diese 500-Seiten-Hardcover-Wälzer sind sowohl physisch also auch inhaltlich einschüchternd.

An beiden Enden des Spektrums gibt es eine Vielzahl von Büchern, was auch hier bezeichnend für die oben beschriebene Kommunikationslücke ist. Die meisten Menschen lesen entweder die businessorientierten oder aber die technischen Bücher.

Glücklicherweise gibt es in der Schlucht zwischen beiden Extremen dann doch eine Handvoll ausgezeichneter Bücher. Zwei unserer Favoriten sind:

- *Data Science für Unternehmen: Data Mining und datenanalytisches Denken praktisch anwenden* von Foster Provost und Tom Fawcett (ursprünglich O'Reilly Media 2013, deutsch mitp 2017)
- *Data Smart: Using Data Science to Transform Information into Insight* von John W. Foreman (Wiley 2013)

Mit diesem Buch möchten wir diese Liste erweitern. Sie sollen es entspannt lesen können, ohne hierfür einen Computer oder einen Schreibblock in der Nähe haben zu müssen. Wenn Ihnen unser Buch gefällt, empfehlen wir Ihnen dringend, auch den nächsten Schritt zu gehen und eines der beiden oben genannten Bücher zu lesen, um Ihr Verständnis weiter zu festigen. Sie werden es nicht bereuen.

Außerdem lieben wir dieses Thema. Wenn wir Ihnen das vermitteln und Sie motivieren können, mehr über Daten und Analytik zu lernen – und Sie dazu inspirieren können, mehr lernen zu *wollen* –, dann ist dieses Buch in unseren Augen ein Erfolg.

Was Sie lernen werden

Dieses Buch wird Ihnen helfen, ein mentales Modell von Data Science, Statistik und Machine Learning zu konstruieren. Was ist ein mentales Modell? Eine »vereinfachte Darstellung der wichtigsten Teile eines Problembereichs, die gut genug ist, um die Problemlösung zu ermöglichen«[5]. Stellen Sie sich ein mentales Modell als neuen Speicherplatz in Ihrem Gehirn vor, in dem Sie Information ablegen können.

Einige Bücher und Artikel stellen erst mal eine Liste mit Definitionen an den Anfang: »Machine Learning ist ...«, »Deep Learning bedeutet ...« etc. Eine Liste mit technischen Definitionen ohne ein mentales Modell, in das man die Informationen einordnen kann, ist so, als würde jemand Kisten mit Kleidung abliefern, ohne dass Sie einen Schrank haben, in den sie hineinpasst. Früher oder später endet alles in der Mülltonne.

Mit dem neu konstruierten mentalen Modell werden Sie dagegen lernen, in Daten zu denken, über sie zu sprechen und sie zu verstehen. Sie werden ein *Data Head*.

Durch das Lesen dieses Buchs lernen Sie vor allem:

- Statistisch zu denken und zu verstehen, welche Rolle Variation in Ihrem Leben und beim Treffen Ihrer Entscheidungen spielt.
- Sich mit Daten auszukennen, fundiert darüber zu sprechen und die richtigen Fragen zu Statistiken und Ergebnissen zu stellen, die Ihnen bei der Arbeit begegnen.
- Zu verstehen, was es wirklich mit Machine Learning, Textanalyse, Deep Learning und künstlicher Intelligenz auf sich hat.
- Häufige Fallstricke zu vermeiden, die bei der Arbeit mit und der Interpretation von Daten auftreten können.

Wie dieses Buch strukturiert ist

Data Heads sind Menschen, die unabhängig von ihrer offiziellen Rolle wissen, wie man kritisch über Daten nachdenkt. Ein Data Head kann der Analytiker sein, der seine Arbeit an der Tastatur verrichtet, oder die Person am Kopf des Konferenztischs, die die Arbeit anderer bewertet. In diesem Buch werden Sie, der Data Head, an verschiedenen Stellen in verschiedene Rollen schlüpfen.

Während die »Geschichte« dieses Buchs chronologisch verläuft, sind die Kapitel selbst jeweils eigenständige Lektionen, die Sie auch einzeln lesen können, ohne eine feste Reihenfolge einhalten zu müssen. Dennoch empfehlen wir Ihnen, das Buch von Anfang bis zum Ende zu lesen. Dies wird Ihnen helfen, Ihr mentales Modell beginnend mit den Grundlagen in Richtung Deep Learning immer mehr zu vertiefen.

Das Buch besteht aus vier Teilen:

Teil I: Denken wie ein Data Head

In diesem Teil lernen Sie, wie ein Data Head zu denken – kritisch zu denken und die richtigen Fragen zu geplanten Datenprojekten Ihres Unternehmens zu

5 Diese Idee wird in diesem erstaunlich hilfreichen Buch behandelt: Greg Wilson, *Teaching Tech Together*. CRC Press 2019.

stellen. Sie lernen, was Daten sind, welche Fachbegriffe wichtig sind und wie man die Welt aus statistischer Sicht betrachtet.

Teil II: Sprechen wie ein Data Head

Data Heads nehmen aktiv an wichtigen Konversationen über Daten teil. In diesem Teil lernen Sie, mit Daten zu argumentieren, und Sie erfahren, welche Fragen Sie stellen müssen, damit die Statistiken, die Ihnen begegnen, auch einen Sinn ergeben. Wir werden Sie mit den Grundlagen der Statistik und Wahrscheinlichkeitsrechnung vertraut machen, die Sie brauchen, um Ergebnisse, die Ihnen vorgelegt werden, zu verstehen und kritisch zu hinterfragen.

Teil III: Den Werkzeugkasten des Data Scientist verstehen

Data Heads verstehen die Grundkonzepte und die Funktionsweise von statistischen und Machine-Learning-Modellen. Sie entwickeln ein intuitives Verständnis von unüberwachtem Lernen, Regression, Klassifikation, Textanalyse und Deep Learning.

Teil IV: Den Erfolg sichern

Data Heads kennen die häufigen Fehler und Fallstricke, die bei der Arbeit mit Daten auftreten können. Sie lernen die technischen Probleme kennen, an denen Projekte scheitern, und Sie lernen die Menschen und Persönlichkeiten kennen, die an Datenprojekten beteiligt sind. Zum Abschluss geben wir Ihnen eine Reihe von Richtlinien an die Hand, die Ihnen zeigen, wie man als Data Head erfolgreich ist.

Ein letzter Punkt, bevor es wirklich losgeht

Wir haben festgestellt, dass das Feld der Data Science schneller wächst, als wir die Probleme und Möglichkeiten, die sich daraus ergeben, formulieren können. Wir haben auch erkannt, dass unsere Vergangenheit (sowohl die der Autoren als auch die der Gesellschaft) voll von Datenfehlschlägen ist. Und nur durch das Verstehen der Vergangenheit können wir die Zukunft verstehen. Die ersten Schritte auf diesem Weg sind wir gegangen, indem wir mit dem Restaurant-Beispiel eine Reihe wichtiger Konzepte eingeführt haben.

Um Daten auf einer tieferen Ebene zu verstehen, müssen Sie lernen, das Grundrauschen zu durchdringen, kritisch über Datenprobleme nachzudenken und effektiv mit Datenanalysten zu kommunizieren. Wir sind sicher, dass Sie mit diesem Wissen auf einem guten Weg sind.

Sind Sie bereit? Unsere Reise, ein Data Head zu werden, beginnt auf den folgenden Seiten.