Vorwort zur 2. Auflage

Ich bin außerordentlich stolz auf die erste Auflage von *Einführung in Data Science*. Es wurde zu dem Buch, das ich mir vorgestellt hatte. Aber viele Jahre in der Data-Science-Entwicklung, der Fortschritt des Python-Ökosystems und meine persönliche Weiterentwicklung als Programmierer und Lehrer haben meine Vorstellung davon verändert, wie ein erstes Buch über Data Science aussehen sollte.

Im Leben gibt es keine zweiten Versuche, aber beim Schreiben gibt es zweite Auflagen.

Daher habe ich den gesamten Code und die Beispiele für Python 3.6 (und viele seiner neuen Features, wie zum Beispiel Type Annotations) neu geschrieben. Im gesamten Buch habe ich einen Schwerpunkt auf das Schreiben sauberen Codes gelegt. Einige der Spiel-Beispiele aus der ersten Auflage habe ich durch realistischere mit »richtigen« Daten ersetzt. Ich habe neues Material zu Themen wie Deep Learning, Statistik und linguistischer Datenverarbeitung ergänzt – Dinge, mit denen die Data Scientists von heute eher arbeiten werden. (Und ich habe einiges davon entfernt, was mir weniger relevant erschien.) Auch sonst habe ich mir das Buch nochmals genau angeschaut und Fehler behoben, Erläuterungen umgeschrieben, die weniger klar waren, als sie hätten sein können, und ein paar der Witze in die Gegenwart transportiert.

Die erste Auflage war ein tolles Buch, und diese ist noch besser. Viel Spaß!

Joel Grus, Seattle, Washington, 2019

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen werden in diesem Buch verwendet:

Kursiv
: Weist auf neue Begriffe, Webadressen, E-Mail-Adressen, Dateinamen und Dateierweiterungen hin.

`Nichtproportionalschrift`
: Wird sowohl für Programmcode verwendet als auch für Hinweise auf Elemente eines Programms im Text, etwa Namen von Variablen und Funktionen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter.

`Nichtproportionalschrift fett`
: Markiert Befehle oder anderen Text, den der Benutzer wörtlich eingeben sollte.

`Nichtproportionalschrift kursiv`
: Markiert Text, den der Benutzer durch eigene Daten ersetzen sollte, oder Werte, die sich aus dem Kontext ergeben.

Dieses Symbol markiert einen Tipp oder eine Empfehlung.

Dieses Symbol markiert einen allgemeinen Hinweis.

Dieses Symbol markiert eine Warnung oder mahnt zur Vorsicht.

Verwenden von Codebeispielen

Zusätzliches Material (Codebeispiele, Übungen usw.) kann von der Adresse *https://github.com/joelgrus/data-science-from-scratch* heruntergeladen werden.

Dieses Buch ist dazu da, Ihnen beim Erledigen Ihrer Arbeit zu helfen. Im Allgemeinen dürfen Sie die Codebeispiele aus diesem Buch in Ihren eigenen Programmen und der dazugehörigen Dokumentation verwenden. Sie müssen uns nicht um Erlaubnis fragen, solange Sie nicht einen beträchtlichen Teil des Codes reproduzieren. Beispielsweise benötigen Sie keine Erlaubnis, um ein Programm zu schreiben, in dem mehrere Codefragmente aus diesem Buch vorkommen. Wollen Sie dagegen eine CD-ROM mit Beispielen aus Büchern von O'Reilly verkaufen oder verteilen, brauchen Sie eine Erlaubnis. Eine Frage zu beantworten, indem Sie aus diesem Buch zitieren und ein Codebeispiel wiedergeben, benötigt keine Erlaubnis. Eine größere Menge Beispielcode aus diesem Buch in die Dokumentation Ihres Produkts aufzunehmen, bedarf hingegen einer Erlaubnis.

Wir freuen uns über Literaturverweise, verlangen sie aber nicht. Ein Literaturverweis besteht für gewöhnlich aus Titel, Autor, Herausgeber und ISBN, zum Beispiel: »*Einführung in Data Science* von Joel Grus, 2. Auflage, O'Reilly 2020, ISBN 978-3-96009-123-3«.

Wenn Sie das Gefühl haben, zu viel Beispielcode zu verwenden oder die oben genannten Befugnisse zu überschreiten, können Sie sich gern an uns wenden unter der Adresse *kommentar@oreilly.de*.

Danksagungen

Als Erstes möchte ich Mike Loukides dafür danken, dass er meinen Vorschlag für dieses Buch angenommen hat (und darauf bestand, dass ich es auf eine zumutbare Größe eindampfe). Es wäre ihm ein Leichtes gewesen, zu sagen: »Wer ist dieser Mensch, der mir ständig Probekapitel zuschickt, und wie werde ich ihn wieder los?« Ich bin dankbar, dass er das nicht gesagt hat. Ich möchte außerdem meinen Lektorinnen Michele Cronin und Marie Beaugureau für ihre Begleitung während des Publikationsprozesses danken und dafür, dass sich das Buch nun in einem viel besseren Zustand befindet, als ich es jemals selbst hinbekommen hätte.

Ich hätte dieses Buch nicht schreiben können, wenn ich nicht Data Science gelernt hätte, und ich hätte Data Science nicht ohne den Einfluss von Dave Hsu, Igor Tatarinov, John Rauser und die übrige Farecast-Gang gelernt. (Das ist so lange her, dass man es damals nicht einmal Data Science nannte!) Die lieben Leute von Coursera und DataTau verdienen ebenfalls eine Menge Anerkennung.

Ich bin auch meinen Testlesern und Gutachtern dankbar. Jay Fundling hat eine Menge Fehler gefunden und mich auf viele unklare Erläuterungen hingewiesen. Das Buch ist durch ihn viel besser (und sehr viel korrekter) geworden. Debashis Ghosh ist ein Held, weil er sämtliche meiner Statistiken auf Richtigkeit geprüft hat. Andrew Musselman empfahl mir, die Aussage »Leute, die R Python gegenüber bevorzugen, sind moralisch verkommen.« abzuschwächen, was sich als sehr guter Ratschlag entpuppt hat. Trey Causey, Ryan Matthew Balfanz, Loris Mularoni, Núria Pujol, Rob Jefferson, Mary Pat Campbell, Zach Geary Denise Mauldin, Jimmy O'Donnell und Wendy Grus lieferten wertvolle Rückmeldungen. Vielen Dank an alle, die die erste Auflage des Buchs gelesen haben und mir dabei halfen, dieses Buch besser zu machen. Die Verantwortung für sämtliche verbliebenen Fehler liegt selbstverständlich bei mir.

Ich verdanke der #datascience-Gemeinde auf Twitter eine Menge – von der Auseinandersetzung mit neuen Begriffen über das Kennenlernen einer Menge großartiger Menschen bis dahin, dass ich mich wie eine Niete fühlte, sodass ich zur Kompensation dieses Buch schrieb. Besonderer Dank gebührt (abermals) Trey Causey für die (unbeabsichtigte) Ermahnung, ein Kapitel der linearen Algebra zu widmen, und Sean J. Taylor für die (unbeabsichtigten) Hinweise auf einige gigantische Lücken im Kapitel »Arbeiten mit Daten«.

Vor allem aber schulde ich Ganga und Madeline besonderen Dank. Das Einzige, was schwieriger ist, als ein Buch zu schreiben, ist, mit jemandem, der ein Buch schreibt, zusammenzuleben, und ohne ihre Hilfe hätte ich es nicht bis zum Ende durchgestanden.

Vorwort zur 1. Auflage

Data Science

Data Scientist wurde bereits als der »sexiest Job des 21. Jahrhunderts« (*https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*) bezeichnet, vermutlich von jemandem, der nie eine Feuerwache besucht hat. Nichtsdestotrotz ist Data Science ein aktuelles und wachsendes Feld, und man muss kein Meisterdetektiv sein, um zu prognostizieren, dass wir in den nächsten zehn Jahren Millionen und Abermillionen mehr Data Scientists benötigen werden, als es zurzeit gibt.

Aber was ist Data Science eigentlich? Schließlich können wir keine Data Scientists ausbilden, wenn wir Data Science nicht einmal definieren können. Laut einem in der Branche bekannten Venn-Diagramm (*http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram*) setzt sich Data Science zusammen aus:

- der Fähigkeit zu hacken,
- dem Wissen über Mathematik und Statistik sowie
- substanziellem Expertenwissen.

Obwohl ich ursprünglich ein Buch über alle drei Dinge schreiben wollte, wurde mir schnell klar, dass allein eine gründliche Abhandlung über das »substanzielle Expertenwissen« Zehntausende Seiten benötigen würde. Daher beschloss ich, mich auf die ersten beiden Punkte zu beschränken. Mein Ziel ist es, Ihre Fähigkeit zu hacken so zu entwickeln, dass Sie gleich damit beginnen können, Data Science praktisch anzuwenden. Mein Ziel ist es ebenfalls, Sie mit Mathematik und Statistik im Zentrum von Data Science vertraut zu machen.

Das ist ein recht ambitioniertes Ziel für ein Buch. Der beste Weg, hacken zu lernen, ist, Dinge zu hacken. Beim Lesen dieses Buchs werden Sie einen guten Einblick darin bekommen, auf welchem Weg ich Dinge hacke. Das muss nicht zwangsläufig der beste Weg für Sie sein, Dinge zu hacken. Sie werden Kenntnisse über einige von mir genutzte Werkzeuge erlangen, die nicht unbedingt die bestmöglichen Werkzeuge für Sie sind. Sie werden kennenlernen, wie ich mich Datenproblemen nähere, für Sie gibt es aber vielleicht bessere Ansätze. Meine Absicht (und Hoff-

nung) ist, dass meine Beispiele Sie beflügeln werden, Dinge selbst auf Ihre eigene Weise auszuprobieren. Sämtlicher Code und alle Daten zu diesem Buch sind auf GitHub (*https://github.com/joelgrus/data-science-from-scratch*) verfügbar, sodass Sie gleich beginnen können.

Analog dazu besteht der beste Weg, Mathematik zu lernen, darin, Mathematik zu betreiben. Dieses Buch ist aus Rücksicht auf den Leser kein Mathematikbuch geworden, und die meiste Zeit werden wir keine »Mathematik betreiben«. Allerdings können Sie sich ohne *Grundkenntnisse* in Wahrscheinlichkeit, Statistik und linearer Algebra nicht ernsthaft mit Data Science auseinandersetzen. Daher werden wir an angemessener Stelle in mathematische Formeln, Denkweisen, Axiome und in die Zeichentrickversionen großer mathematischer Konzepte eintauchen. Ich hoffe, Sie fürchten sich nicht davor, mit mir hineinzuspringen.

Im Verlauf der Kapitel hoffe ich, Ihnen ein Gefühl für den Spaß am Spielen mit Daten zu vermitteln, weil das Spielen mit Daten eben Spaß macht! (Besonders im Vergleich zu einigen Alternativen wie dem Vorbereiten der Steuererklärung oder dem Kohlebergbau.)

Bei null starten

Es gibt etliche Programmbibliotheken, Frameworks, Module und Werkzeugsammlungen, die die verbreitetsten (und auch die exotischsten) Algorithmen und Techniken für Data Science beinhalten. Sobald Sie ein Data Scientist geworden sind, haben Sie eine innige Freundschaft mit NumPy, scikit-learn, pandas und einer Palette weiterer Bibliotheken geschlossen. Diese eignen sich gut, um Data Science zu betreiben. Sie sind aber auch hilfreich, wenn es darum geht, mit Data Science zu beginnen, ohne überhaupt etwas davon zu verstehen.

In diesem Buch werden wir uns Data Science von Grund auf nähern. Das bedeutet, wir werden uns Werkzeuge selbst bauen und Algorithmen von Hand implementieren, um sie besser zu verstehen. Ich habe viel über klare, gut kommentierte und verständliche Implementierungen und Beispiele nachgedacht. In den meisten Fällen werden unsere selbst gebauten Werkzeuge erhellend, aber unpraktisch sein. Sie werden für kleine Sandkastendatensätze gut funktionieren, aber an solchen mit »Internetausmaßen« kläglich scheitern.

Im Verlauf des Buchs werde ich Sie auf Bibliotheken hinweisen, mit denen Sie diese Techniken auf größere Datensätze anwenden können. Wir werden diese hier aber nicht verwenden.

Es gibt eine gesunde Diskussion darüber, welche die beste Programmiersprache ist, um Data Science zu lernen. Viele meinen, dies sei die Statistikprogrammiersprache R. (Wir nennen sie die »Leute auf dem Holzweg«.) Einige andere empfehlen Java oder Scala. Meiner Meinung nach ist jedoch Python die erste Wahl.

Python besitzt mehrere Eigenschaften, die sie zu einer gut geeigneten Sprache zum Lernen (und Betreiben) von Data Science machen:

- Python ist kostenlos.
- Es ist relativ einfach, darin zu programmieren (und insbesondere den Code zu verstehen).
- Es gibt zahlreiche nützliche Bibliotheken für Data Science in Python.

Ich zögere, Python meine Lieblingsprogrammiersprache zu nennen. Es gibt andere Sprachen, die ich angenehmer oder besser entworfen finde oder bei denen es mir einfach mehr Spaß macht, Code zu schreiben. Dennoch lande ich bei jedem neuen Data-Science-Projekt wieder bei Python. Jedes Mal, wenn ich schnell einen lauffähigen Prototyp schreiben muss, lande ich bei Python. Und jedes Mal, wenn ich Konzepte von Data Science klar und verständlich demonstrieren möchte, lande ich bei Python. Deshalb verwendet dieses Buch Python.

Das Ziel dieses Buchs ist nicht, Ihnen Python beizubringen (aber es ist so gut wie sicher, dass Sie beim Lesen etwas Python lernen werden). Ich werde Sie begleiten durch einen Crashkurs von der Länge eines Kapitels, der die für unsere Zwecke wichtigsten Eigenschaften hervorhebt. Sollten Sie aber gar nichts über das Programmieren in Python (oder über das Programmieren im Allgemeinen) wissen, benötigen Sie möglicherweise begleitend zum Buch ein Tutorial wie etwa »Programmieren lernen mit Python« von Allen B. Downey aus dem O'Reilly Verlag.

Der Rest unserer Einführung in Data Science wird genau diesen Ansatz wählen – dort in die Details gehen, wo es unausweichlich oder erhellend ist, und ansonsten Ihnen die Details zur eigenen Erkundung überlassen (oder zum Nachschlagen auf Wikipedia).

Im Verlauf der Jahre habe ich so manchen Data Scientist ausgebildet. Auch wenn nicht alle von ihnen weltumwälzende Daten-Ninja-Rockstars geworden sind, habe ich alle als bessere Data Scientists entlassen, als ich sie ursprünglich vorfand. Dabei habe ich den Glauben gewonnen, dass jeder mit etwas mathematischer Begabung und ein paar Programmierfähigkeiten sämtliche Grundvoraussetzungen zum Betreiben von Data Science erfüllt. Notwendig sind ein aufgeschlossener Geist, die Bereitschaft zu harter Arbeit und dieses Buch. Darum dieses Buch.