

Inhaltsverzeichnis

Über die Autorin	7
Vorwort	21
Einleitung	23
Über dieses Buch	23
Törichte Annahmen	24
In diesem Buch verwendete Symbole	24
Wo Sie starten	24
Teil I	
Mit Data Science beginnen	25
Kapitel 1	
Bei Data Science durchblicken	27
Schauen, wer von Data Science Gebrauch machen kann	27
Die Teile des Data-Science-Puzzles betrachten	29
Daten sammeln, abfragen und bearbeiten	29
Von der Mathematik und Statistik Gebrauch machen	30
Programmierung: Teil des Spiels	32
Data Science in Ihrem Fachgebiet anwenden	32
Einblicke kommunizieren	33
Die Grundlagen schaffen	33
Mögliche Vorgehensweisen klären	34
Die offensichtlichen Gewinne ermitteln	35
Kapitel 2	
Data Engineering: Infrastruktur erkunden	37
Big Data definieren	37
Mit dem Datenvolumen ringen	38
Die Datengeschwindigkeit handhaben	38
Die Datenvielfalt behandeln	38
Den Datenwert erzeugen	39
Big-Data-Datenquellen bestimmen	39
Den Unterschied zwischen Data Science und Data Engineering verstehen	40
Data Science definieren	40
Data Engineering definieren	41
Ein Vergleich zwischen Data Scientists und Data Engineers	42
Datensätze mit MapReduce und Hadoop reduzieren	43
In MapReduce einarbeiten	43
Hadoop verstehen	45

Alternative Lösungen für Probleme mit Big Data betrachten	47
Die Echtzeitverarbeitung einführen	47
Massenparallelrechner verwenden	48
In NoSQL-Datenbanken einarbeiten	48
Data Engineering im Einsatz – Eine Fallstudie	49
Die Geschäftsherausforderung erkennen	49
Geschäftsprobleme mit Data Engineering lösen	51
Erfolge aufweisen	51

Kapitel 3

Data Science in Unternehmen und Industrie einsetzen **53**

Datengesteuerte Einblicke in die Geschäftsabläufe einbinden	53
Von geschäftsbezogener Data Science profitieren	54
Datenanalyse zur Umwandlung von Rohdaten in umsetzbare Einsichten	54
Etwas mit den Geschäftsdaten unternehmen	57
Business Intelligence und Data Science unterscheiden	58
Business Intelligence definieren	58
Geschäftsbezogene Data Science definieren	60
Die Hauptunterschiede zwischen BI und geschäftsbezogener	
Data Science zusammenfassen	62
Wissen, wen man holt, um die Arbeit zu erledigen	63
Data Science im Geschäftsleben: Eine datengesteuerte Erfolgsgeschichte	64

Teil II

Die Bedeutung Ihrer Daten mit Data Science erkennen **67**

Kapitel 4

Wahrscheinlichkeit und Statistik einführen **69**

Die grundlegenden Konzepte der Wahrscheinlichkeit vorstellen	69
Die Beziehung zwischen Wahrscheinlichkeit und induktiver Statistik	70
Zufallsvariablen, Wahrscheinlichkeitsverteilungen und Erwartungswerte	
verstehen	71
Gängige Wahrscheinlichkeitsverteilungen kennenlernen	73
Die lineare Regression einführen	75
Einfache Modelle zur linearen Regression	75
Lernen, eine angepasste Regressionsgerade zu erstellen	77
Die Methode der kleinsten Quadrate	79
Simulationen	80
Mit Simulationen Eigenschaften von Testgrößen beurteilen	84
Mit der Monte-Carlo-Simulation einen Schätzwert beurteilen	86

Die Zeitreihenanalyse einführen	88
Muster in Zeitreihen verstehen	88
Die univariate Varianzanalyse modellieren	88

Kapitel 5

Clustering-Verfahren und Klassifikation 93

Die Grundlagen von Cluster-Verfahren und Klassifikation einführen	93
Clustering-Algorithmen kennenlernen	94
Klassifikations-Algorithmen kennenlernen	96
Metriken kennenlernen	99
Cluster in Ihren Daten erkennen	99
Mit dem k -Means-Algorithmus Clusteranalyse betreiben	100
Cluster mit der Kerndichteschätzung abschätzen	101
Hierarchische Algorithmen und Algorithmen zur sortierten Nachbarschaft einsetzen	101
Daten mit Entscheidungsbäumen und Random-Forest-Algorithmen kategorisieren	104

Kapitel 6

Clusteranalyse und Klassifikation mit Nearest-Neighbor-Verfahren 107

Die Bedeutung der Daten mit Nearest-Neighbor-Analysen erkennen	107
Die Wichtigkeit der Clusteranalyse und der Klassifikation erkennen	108
Daten mit Gemittelter-Nearest-Neighbor-Algorithmen klassifizieren	109
Verstehen, wie der Gemittelter-Nearest-Neighbor-Algorithmus funktioniert	110
Die Klassifikation mit k -Nearest-Neighbor-Algorithmen	113
Die Arbeitsweise des k -Nearest-Neighbor-Verfahrens verstehen	114
Wissen, wann man den k -Nearest-Neighbor-Algorithmus einsetzt	115
Gängige Anwendungen von k -Nearest-Neighbor-Algorithmen erkunden	116
Mit den Abständen der nächsten Nachbarn Schlussfolgerungen aus Datenpunkt-Mustern ziehen	116
Probleme der realen Welt mit Nearest-Neighbor-Algorithmen lösen	117
k -Nearest-Neighbor-Algorithmen im Einsatz erleben	117
Gemittelter-Nearest-Neighbor-Algorithmen im Einsatz erleben	118

Kapitel 7

Mathematische Modellierung in der Datenwissenschaft 119

Die multikriterielle Entscheidungsanalyse (MCA) einführen	119
Die MCA im Einsatz betrachten und verstehen	120
Die Fuzzy-MCA anwenden	121
Wissen, wann und wie man die multikriterielle Entscheidungsanalyse einsetzt	123

Die Verwendung von numerischen Methoden in der Data Science	126
Über Taylorpolynome	127
Mit der Bisektion Funktionen halbieren	128
Mathematische Modellierung mit Markov-Ketten und stochastischen Methoden	130

Kapitel 8

Mit der Statistik Geodaten modellieren **133**

Oberflächen aus Raumpunktdaten vorhersagen	133
Die Parameter x , y und z bei Geodaten verstehen	134
Kriging einführen	135
Automatische Kriging-Interpolationen	136
Modelle zur explizit definierten Kriging-Interpolation wählen und verwenden	136
Sich intensiver mit dem Kriging beschäftigen	138
Das beste Schätzverfahren beim Kriging wählen	142
Zur Bestimmung des besten Modells das Residuum untersuchen	143
Ihre Wahlmöglichkeiten bei Kriging-Verfahren kennen	146
Trendanalyse von Oberflächen	146

Teil III

Datenvisualisierungen mit klaren Aussagen **147**

Kapitel 9

Den Prinzipien der Datenvisualisierung entsprechen **149**

Die Arten der Visualisierung verstehen	149
Entscheidungsträgern die Geschichte hinter den Daten erzählen	150
Daten für Analytiker zur Geltung bringen	150
Datenkunst für Aktivisten	150
Ihre Zielgruppe in den Blick nehmen	151
Schritt 1: Ideenfindung für Steffi	151
Schritt 2: Bestimmen Sie Ihr Ziel	152
Schritt 3: Die für Ihr Ziel zweckmäßigste Visualisierungsart wählen	153
Das zweckmäßigste Design wählen	153
Mit dem Design eine analysierende und präzise Reaktion hervorrufen	154
Mit dem Design eine stark emotionale Reaktion hervorrufen	154
Wissen, wann man einen Zusammenhang darstellen muss	156
Daten verwenden, um Zusammenhänge herzustellen	156
Sinnzusammenhänge über Beschriftung herstellen	156
Grafische Elemente zur Herstellung von Sinnzusammenhängen verwenden	157
Wissen, wann man überzeugen muss	157

Eine passende Art von Grafik wählen	158
Standarddiagramme erklären	159
Vergleichende Grafiken erkunden	161
Statistische Diagramme erkunden	165
Topologische Strukturen erkunden	167
Räumliche Darstellungen und Karten erkunden	169
Ihre Grafik auswählen	171
Betrachten der Fragen	172
Berücksichtigung der Nutzer und der Medien	172
Einen letzten Blick auf die Arbeit werfen	172
Kapitel 10	
D3.js zur Visualisierung von Daten verwenden	173
Einführung in die Bibliothek D3.js	173
Wissen, wann man D3.js verwenden sollte (und wann nicht)	174
Der Einstieg in D3.js	175
HTML und DOM einführen	176
JavaScript und SVG einführen	177
Cascading Style Sheets (CSS) einführen	178
Webserver und PHP einführen	178
Fortgeschrittene Konzepte und Methoden in D3.js verstehen	179
Kettensyntax kennenlernen	182
Skalen kennenlernen	184
Übergänge und Interaktionen kennenlernen	185
Kapitel 11	
Webbasierte Anwendungen zur Daten-Visualisierung	187
Kollaborativ genutzte Visualisierungsplattformen	188
Mit Watson Analytics von IBM arbeiten	188
Visualisieren und Kollaborieren mit Plotly	190
Geodaten mit geografischen Tools visualisieren	192
Schöne Karten mit OpenHeatMap herstellen	194
Das Erstellen von Karten und die Untersuchung von Geodaten mit CartoDB	195
Webbasierte Open-Source-Plattformen zur Datenvisualisierung	196
Mit Google Fusion Tables schöne Grafiken erstellen	197
iCharts zur webbasierten Visualisierung verwenden	198
RAW zur webbasierten Visualisierung verwenden	198
Wissen, wann man Infografiken verwendet	200
Mit Infogr.am fetzige Infografiken erstellen	201
Fetzige Grafiken mit Piktochart erstellen	202

Kapitel 12

Die besten Techniken zum Erstellen eines Dashboards 205

Sich an der Zielgruppe orientieren	206
Mit dem großen Ganzen beginnen	206
Die Einzelheiten gut hinbekommen	207
Ihren Entwurf testen	209

Kapitel 13

Aus Geodaten Karten erstellen 211

In die Grundlagen von GIS einsteigen	211
Geodatenbanken verstehen	213
Dateiformate in GIS verstehen	213
Kartennetzentwürfe und Koordinatensysteme verstehen	217
Geodaten analysieren	218
Geodaten abfragen	219
Buffering und Nachbarschaftsfunktionen	220
Analysen basierend auf der Überlagerung einzelner Layer	220
Reklassifikation von Geodaten	222
Mit der Open-Source-Software QGIS arbeiten	222
Die Benutzeroberfläche von QGIS kennenlernen	222
In QGIS einen Vektorlayer hinzufügen	223
Anzeige der Daten in QGIS	225

Teil IV

Programmieren und Data Science 231

Kapitel 14

Python für Data Science verwenden 233

Die grundlegenden Konzepte von Python verstehen	233
Datentypen in Python	235
Schleifen in Python verwenden	237
Funktionen und Klassen kennenlernen	238
Enge Bekanntschaft mit einigen nützlichen Python-Bibliotheken schließen	241
Die Bibliothek NumPy	242
Mit SciPy vertraut werden	244
Zur Visualisierung von Daten Matplotlib einbinden	245
Die Verwendung von Python zur Analyse von Daten – ein Beispiel	247
Python auf Mac OS und Windows installieren	247
CSV-Dateien laden	248
Einen gewichteten Mittelwert berechnen	249
Trendlinien zeichnen	252

Kapitel 15

Das frei zugängliche R in der Data Science benutzen **255**

Die grundlegenden Konzepte einführen	255
Die grundlegenden Begriffe in R kennenlernen	255
Tiefer in Funktionen und Operatoren eintauchen	258
Iterieren in R	262
Beobachten, wie Objekte arbeiten	264
Vorschau auf die Pakete von R	266
Einige gefragte Pakete zur statistischen Analyse	266
Visualisierung, Kartierung und grafische Darstellung in R	267

Kapitel 16

SQL in Data Science verwenden **271**

Mit SQL beginnen	271
Relationale Datenbanken und SQL in den Griff bekommen	271
Datenbanken entwerfen	275
SQL und seine Funktionen in Data Science verwenden	278
SQL, R, Python und Excel in Ihre Data-Science-Strategie integrieren	278
SQL-Funktionen in Data Science verwenden	279

Kapitel 17

Anwendungssoftware für Data Science **285**

Das Leben mit Excel vereinfachen	285
Mit Excel die Daten schnell kennenlernen	286
Umformatieren und Zusammenfassen mit Pivot-Tabellen	290
Aufgaben von Excel mit Makros automatisieren	291
KNIME zur fortgeschrittenen Analyse von Daten verwenden	293
Die Kundenabwanderung mit KNIME verringern	294
Das Beste aus Daten sozialer Netzwerke machen	294
KNIME für eine ökologisch gute Verwaltung verwenden	294

Teil V

Probleme aus der Praxis mit Data Science lösen **295**

Kapitel 18

Data Science im Journalismus verwenden **297**

Die sechs Ws erklären	298
Überprüfen, wer	298
Überlegen, warum Ihr Artikel von Bedeutung ist	300
Zu dem kommen, was Sie sagen wollen	301
Wann ist der richtige Zeitpunkt?	302

Überlegen, wo Ihre Geschichte eine Rolle spielt	303
Überlegen, wie Sie Ihre Reportage entwickeln, formulieren und präsentieren	304
Daten für Ihre Reportage sammeln	305
Screen Scraping für Ihre Reportage nutzen	305
Alert-Dienste einsetzen	306
Die Geschichte hinter den Daten entdecken und erzählen	307
Außergewöhnliche Trends und Ausreißer entdecken	307
Den Kontext untersuchen, um die Signifikanz der Daten zu verstehen	309
Die Geschichte durch Ihre Visualisierung unterstreichen	310
Fesselnde und klar umrissene Reportagen erstellen	311
Den Datenjournalismus lebendig werden lassen: Der Artikel »Schwarze Kassen« in der Washington Post	311

Kapitel 19

Data Science und die Umwelt miteinander verbinden **313**

Modellierung der Wechselwirkung zwischen Mensch und Umwelt anhand ökologischer Intelligenz	313
Die zu lösenden Probleme betrachten	314
Ökologische Intelligenz definieren	315
Wichtige Organisationen kennenlernen, die im Bereich der ökologischen Intelligenz arbeiten	316
Mit ökologischer Intelligenz positiven Einfluss ausüben	317
Natürliche Ressourcen im Urzustand modellieren	318
Die Modellierung von natürlichen Ressourcen erkunden	318
Sich an Data Science versuchen	319
Modellierung natürlicher Ressourcen zur Lösung von Umweltproblemen	319
Mit der Geostatistik Umweltbedingungen abhängig vom Raum vorhersagen	320
Mit der vorhersagenden Geoanalyse Umweltfragen behandeln	321
Den Anteil der Data Science erläutern	321
Die Geostatistik zur Behandlung von Umweltthemen verwenden	322

Kapitel 20

Mit Data Science das Wachstum des E-Commerce vorantreiben **323**

Daten verstehen und für das Wachstum des E-Commerce einsetzen	325
Optimierung der beim Internethandel verwendeten Systeme	326
Analysemethoden kennenlernen	327
Ihre Strategien überprüfen	331
Segmentierung und Zielgruppenansprache tragen zum Erfolg bei	334

Kapitel 21

**Data Science zur Beschreibung und Vorhersage
krimineller Aktivitäten einsetzen**

339

Zeitliche Analyse zur Vorhersage und Verfolgung von Verbrechen	340
Räumliche Analyse zur Vorhersage und Verfolgung von Verbrechen	340
Die Kartografierung von Verbrechen mit GIS-Technologien	341
Einen Schritt weitergehen: Die Standortvorhersage	341
Komplexe räumliche Statistik zum besseren Verständnis von Verbrechen verwenden	342
Die Probleme untersuchen, die mit der Verwendung von Data Science zur Analyse von Verbrechen verbunden sind	345
Die Grundrechte berücksichtigen	345
Gegen technische Probleme kämpfen	346

Teil VI

Der Top-Ten-Teil

349

Kapitel 22

Zehn fantastische frei zugängliche Datenquellen

351

Sich in Data.gov vertiefen	352
Die frei zugänglichen Daten in Kanada ausprobieren	353
Die Webseite data.gov.uk untersuchen	354
Das Datenportal für Deutschland kennenlernen	354
Daten der NASA kennenlernen	355
Auf die Daten der Weltbank zugreifen	356
Sich mit Knoema Data bekannt machen	357
Sich bei Quandl Data in die Schlange stellen	358
Die Exversion-Daten erkunden	359
OpenStreetMap zur Kartierung verwenden	360

Kapitel 23

Etwa zehn freie Tools und Anwendungen zur Data Science

363

Das Erstellen individualisierter webbasierter Visualisierungen mit freien R-Paketen	363
Mit RStudio glänzen	364
rCharts zum Visualisieren verwenden	365
Mit rMaps kartieren	365
Weitere Tools zum Auslesen, Sammeln und Verarbeiten von Daten	366
Daten mit import.io extrahieren	366
Mit ImageQuilts Bilder sammeln	367
Sich Daten mit DataWrangler beschaffen	368

Weitere Tools zum Untersuchen von Daten testen	368
Über Tableau Public reden	368
Mit Gephi vorankommen	369
Maschinelles Lernen mit WEKA	371
Weitere webbasierte Visualisierungstools testen	372
Mit Weave arbeiten	372
Die Visualisierungsangebote von Knoema testen	373

Stichwortverzeichnis	377
-----------------------------	------------