

# Inhaltsverzeichnis

Über die Autoren	9
Luca Massarons Widmung	9
John Muellers Widmung	10
Luca Massarons Danksagung	10
John Muellers Danksagung	10
<b>Einleitung</b>	<b>23</b>
Über dieses Buch	23
Törichte Annahmen	24
Im Buch verwendete Symbole	25
Über das Buch hinaus	25
Wie es weitergeht	26
<b>Teil I</b>	
<b>Erste Schritte mit Python für Data Science</b>	<b>27</b>
<b>Kapitel 1</b>	
<b>Wie Data Science und Python zusammenpassen</b>	<b>29</b>
Die Definition des geilsten Jobs des 21. Jahrhunderts	31
Die Entstehung von Data Science	31
Umriss der Kernkompetenzen eines Data Scientists	32
Die Verbindung von Data Science und Big Data	32
Das Verständnis der Rolle der Programmierung	33
Die Entwicklung einer Data-Science-Pipeline	33
Vorbereitung der Daten	33
Darstellung der beschreibenden Datenanalyse	34
Von den Daten lernen	34
Visualisierung	34
Erkenntnisse und Ergebnisse	34
Die Rolle von Python in Data Science	35
Das sich wandelnde Profil eines Data Scientists	35
Die Arbeit mit einer vielseitigen, einfachen und effizienten Sprache	36
Der schnelle Einstieg in Python	36
Daten laden	38
Ein Modell ableiten	38
Anzeige eines Ergebnisses	39

## **Kapitel 2**

### ***Einführung in Pythons Fähigkeiten und Möglichkeiten***

**41**

Warum Python?	42
Verständnis der Kernphilosophie Pythons	43
Gegenwärtige und zukünftige Entwicklungsziele entdecken	43
Arbeiten mit Python	43
Ein Vorgeschmack auf die Sprache	43
Die Notwendigkeit von Einrückungen verstehen	44
Arbeiten mit der Kommandozeile oder IDE	44
Schnelles Prototyping und Experimentieren	49
Die Geschwindigkeit der Ausführung	51
Die Kraft der Visualisierung	52
Das Python-Ökosystem für Data Science	54
Mit SciPy auf wissenschaftliche Werkzeuge zugreifen	54
Grundlagen des wissenschaftlichen Rechnens mit NumPy	54
Datenanalyse mit Pandas	54
Implementierung des maschinellen Lernens mit Scikit-learn	55
Plotten mit Matplotlib	55
Syntaxanalyse von HTML-Dokumenten mit BeautifulSoup	55

## **Kapitel 3**

### ***Einrichtung von Python für Data Science***

**57**

Betrachtung der üblichen wissenschaftlichen Distributionen	58
Continuum Analytics Anaconda	58
Enthought Canopy Express	59
Pythonxy	60
WinPython	60
Installation von Anaconda auf Windows	60
Installation von Anaconda auf Linux	64
Installation von Anaconda auf Mac OS X	65
Download der Datensätze und des Beispielcodes	66
Die Nutzung von IPython Notebook	67
Festlegung des Code-Archivs	68
Verständnis der in diesem Buch verwendeten Datensätze	74

## **Kapitel 4**

### ***Die Grundlagen von Python***

**77**

Arbeiten mit Zahlen und Logik	78
Zuordnung von Variablen	79
Arithmetik	80
Vergleichen von Daten mit booleschen Ausdrücken	82
Erstellung und Nutzung von Zeichenketten	84
Interaktionen mit einer Zeitangabe	85

Erstellung und Verwendung von Funktionen	87
Entwicklung wiederverwendbarer Funktionen	87
Der Aufruf einer Funktion auf unterschiedliche Arten	89
Verwendung von bedingten und iterativen Anweisungen	92
Entscheidungsfindung mit der if-Anweisung	92
Die Wahl zwischen mehreren Optionen mit verschachtelten Entscheidungen	93
Ausführung sich wiederholender Aufgaben mit dem for-Kommando	94
Verwendung der while-Anweisung	95
Daten mit Mengen, Listen und Tupeln speichern	96
Operationen mit Mengen	96
Die Arbeit mit Listen	97
Erstellung und Verwendung von Tupeln	98
Definition nützlicher Iteratoren	100
Indizierung von Daten mit Dictionaries	101

## **Teil II**

### **Mit Daten arbeiten**

**103**

#### **Kapitel 5**

##### **Arbeiten mit richtigen Daten**

**105**

Upload, Streaming und Auswahl von Daten	106
Laden kleiner Datenmengen in den Speicher	106
Laden großer Datenmengen in den Speicher	108
Auswahl von Daten	108
Daten in strukturierter Flatfile-Form	110
Aus einer Textdatei lesen	110
Lesen des CSV-Formats	111
Lesen von Excel- oder anderen Microsoft-Dateien	114
Laden von Daten aus unstrukturierten Dateien	115
Verwaltung von Daten aus relationalen Datenbanken	118
Interaktion mit Daten einer NoSQL-Datenbank	120
Verwendung von Daten aus dem Internet	120

#### **Kapitel 6**

##### **Konditionierung der Daten**

**125**

Zwischen NumPy und Pandas hin- und herjonglieren	126
Wann man NumPy verwendet	126
Wann man Pandas verwendet	126
Validierung der Daten	127
Herausfinden, was in Ihren Daten steckt	127
Duplikate entfernen	129
Erstellung einer Datenkarte und eines Datenplans	129

Manipulation kategorialer Variablen	131
Erstellung kategorialer Variablen	132
Umbenennen der Ebenen	134
Die Kombination von Ebenen	134
Der Umgang mit Zeitangaben in Ihren Daten	136
Formatierung von Datums- und Zeitangaben	136
Die richtige Zeittransformation	137
Umgang mit fehlenden Daten	138
Fehlende Daten finden	138
Codierung fehlender Daten	139
Einspeisung fehlender Daten	140
Schneiden und Vereinzeln: Filtern und Auswählen von Daten	141
Zeilen schneiden	141
Spalten schneiden	142
Vereinzlung	143
Verkettung und Transformation	143
Neue Fälle und Variablen hinzufügen	144
Entfernen von Daten	145
Sortieren und Mischen	146
Aggregation von Daten auf einer Ebene	147

## ***Kapitel 7***

### ***Daten in Form bringen***

**149**

Arbeiten mit HTML-Seiten	149
Parsen von XML und HTML	150
Benutzung von XPath für die Extraktion von Daten	151
Die Arbeit mit reinem Text	152
Die Arbeit mit Unicode	152
Stemming und Entfernen von Stopwörtern	153
Einführung in reguläre Ausdrücke	155
Verwendung des Bag-of-Words-Modells und anderer Modelle	158
Funktionsweise des Bag-of-Words-Modells	158
Arbeiten mit N-Grammen	160
Implementierung von TF-IDF Transformationen	161
Arbeiten mit Graphdaten	163
Die Adjazenzmatrix	163
Grundlagen in NetworkX	163

## ***Kapitel 8***

### ***Das, was Sie schon wissen, in die Tat umsetzen***

**167**

Kontextualisierung von Problemen und Daten	168
Auswertung eines Data-Science-Problems	169
Erforschung von Lösungen	169
Formulierung einer Hypothese	170
Vorbereitung Ihrer Daten	170

Betrachtung der Erstellung von Merkmalen	171
Definition der Merkmalserstellung	171
Kombination von Variablen	172
Klasseneinteilung und Diskretisierung	172
Verwendung von Indikatorvariablen	173
Umwandlung von Verteilungen	173
Operationen mit Arrays	174
Vektorisierung	174
Einfache Arithmetik mit Vektoren und Matrizen	175
Matrix-Vektor-Multiplikation	175
Matrix-Multiplikation	176

## **Teil III**

### ***Visualisierung des Unsichtbaren*** **177**

#### **Kapitel 9**

##### ***Ein Crashkurs in Matplotlib*** **179**

Mit einem Graphen beginnen	180
Definition eines Plots	180
Zeichnen mehrerer Linien und Plots	181
Speichern Sie Ihre Arbeit	182
Einstellen der Achsen, Intervalle und Gitternetzlinien	183
Die Achsen	183
Formatierung der Achsen	183
Hinzufügen von Gitternetzen	185
Das Erscheinungsbild von Linien festlegen	186
Die Arbeit mit Linienstilen	186
Verwendung von Farben	187
Marker hinzufügen	188
Labels, Anmerkungen und Legenden	190
Hinzufügen von Labels	190
Hinzufügen von Anmerkungen zum Diagramm	191
Erstellen einer Legende	192

#### **Kapitel 10**

##### ***Visualisierung von Daten*** **195**

Die Wahl der richtigen Grafik	196
Darstellung von Teilen eines Ganzen mit Kreisdiagrammen	196
Darstellung von Vergleichen mit Balkendiagrammen	197
Darstellung von Vergleichen mit Histogrammen	199
Darstellung von Gruppen mit Boxplots	200
Sehen von Datenmustern mit Streudiagrammen	202

Erstellung erweiterter Streudiagramme	203
Darstellung von Gruppen	203
Darstellung von Korrelationen	204
Plotten von Zeitreihen	206
Abbildung der Zeit auf den Achsen	206
Plotten von Trends über einen bestimmten Zeitraum	208
Plotten geografischer Daten	210
Visualisierung mit Graphen	212
Erstellung ungerichteter Graphen	212
Erstellung gerichteter Graphen	214

## ***Kapitel 11***

### ***Die Tools verstehen***

**217**

Arbeiten mit der IPython-Konsole	217
Arbeiten mit Bildschirmtext	218
Wechseln der Fensteranzeige	220
Die Python-Hilfe	221
Die IPython-Hilfe	223
Nutzung der magischen Funktionen	224
Objekte untersuchen	225
Das IPython Notebook	226
Arbeiten mit Formatvorlagen	226
Neustarten des Kernels	228
Wiederherstellung eines Checkpoints	229
Multimedia- und Grafikintegration	230
Einbetten von Plots und anderen Bildern	230
Laden von Beispielen aus Webseiten	232
Erhalt von Online-Grafiken und Multimedia	232

## ***Teil IV***

### ***Daten handhabbar machen***

**235**

#### ***Kapitel 12***

##### ***Pythons Möglichkeiten erweitern***

**237**

Mit Scikit-learn spielen	238
Klassen in Scikit-learn verstehen	238
Anwendungen für Data Science erkennen	239
Den Hashing-Trick durchführen	242
Hash-Funktionen nutzen	242
Hash-Tricks demonstrieren	243
Mit deterministischer Selektion arbeiten	245

Zeit und Performance berücksichtigen	246
Benchmarking mit timeit	247
Mit dem Speicher-Profiler arbeiten	249
Parallele Verarbeitung	251
Mehrern-Verarbeitung durchführen	252
Mehrern-Verarbeitung demonstrieren	252

## ***Kapitel 13***

### ***Datenanalyse erforschen***

**255**

Der EDA-Ansatz	256
Beschreibende Statistik für numerische Daten	257
Lagemaße bestimmen	258
Messung von Varianz und Spannweite	258
Arbeiten mit Perzentilen	259
Normalitätsmaße	260
Zählen von kategorialen Daten	261
Häufigkeiten verstehen	262
Kontingenztafeln erstellen	263
Angewandte Visualisierung für EDA	263
Boxplots untersuchen	264
T-Test nach dem Boxplot durchführen	265
Parallele Koordinaten beobachten	266
Grafische Darstellung von Verteilungen	267
Streudiagramme zeichnen	268
Korrelation verstehen	270
Kovarianz und Korrelation nutzen	270
Nichtparametrische Korrelation nutzen	273
Chi-Quadrat für Tabellen betrachten	273
Datenverteilungen modifizieren	274
Die Normalverteilung nutzen	274
Eine Z-Score-Standardisierung erstellen	275
Andere beachtenswerte Verteilungen transformieren	275

## ***Kapitel 14***

### ***Dimensionalität verringern***

**277**

SVD verstehen	278
Auf der Suche nach Dimensionalitätsverringern	278
SVD nutzen, um das Unsichtbare zu messen	280
Faktor- und Hauptkomponentenanalyse durchführen	281
Das psychometrische Modell berücksichtigen	282
Nach versteckten Faktoren suchen	282
Komponenten nutzen, nicht Faktoren	283
Dimensionalitätsverringern erreichen	283

Einige Anwendungen verstehen	284
Gesichter erkennen mit PCA	284
Themen mit NMF extrahieren	287
Filme empfehlen	289
<b>Kapitel 15</b>	
<b>Clustering</b>	<b>293</b>
Mit K-means clustern	294
K-means-Algorithmen verstehen	295
Ein Beispiel mit Bilddaten	297
Nach optimalen Lösungen suchen	298
Big Data clustern	301
Hierarchisches Clustering durchführen	302
Jenseits von runden Clustern: DBScan	305
<b>Kapitel 16</b>	
<b>Ausreißer in Daten aufspüren</b>	<b>309</b>
Das Aufspüren von Ausreißern in Betracht ziehen	310
Weitere Dinge finden, die schiefgehen können	311
Anomalien bei neuen Daten verstehen	312
Eine einfache univariate Methode untersuchen	312
Auf die Gauß-Verteilung zählen	314
Annahmen machen und überprüfen	315
Einen multivariaten Ansatz entwickeln	316
Hauptkomponentenanalyse nutzen	316
Cluster-Analyse nutzen	318
Ausreißer mit SVM automatisch erkennen	318
<b>Teil V</b>	
<b>Aus Daten lernen</b>	<b>321</b>
<b>Kapitel 17</b>	
<b>Vier einfache und effektive Algorithmen erkunden</b>	<b>323</b>
Die Zahl schätzen: Lineare Regression	323
Die Familie der linearen Modelle definieren	324
Mehr Variablen nutzen	324
Limitierungen und Probleme verstehen	326
Zur logistischen Regression wechseln	327
Logistische Regression anwenden	327
Betrachtung, wenn es mehrere Klassen sind	328

Die Dinge einfach machen – Naiver Bayes	329
Herausfinden, dass naiver Bayes nicht so naiv ist	331
Textklassifizierungen vorhersagen	332
Faul lernen mit der Nearest-Neighbors-Methode	334
Vorhersagen nach der Beobachtung von Nachbarn	334
Wählen Sie Ihren k-Parameter geschickt	336

## ***Kapitel 18***

### ***Kreuzvalidierung, Selektion und Optimierung durchführen*** **337**

Über das Problem der Anpassung eines Modells nachdenken	338
Trend und Varianz verstehen	339
Eine Strategie zur Modellauswahl definieren	340
Zwischen Trainings- und Testsatz trennen	343
Kreuzvalidierung	345
Kreuzvalidierung auf k Teilmengen anwenden	346
Probenschichtung für komplexe Daten	347
Variablen wie ein Profi auswählen	348
Durch univariate Maße selektieren	348
Eine Greedy-Suche nutzen	350
Ihre Hyperparameter aufbessern	351
Eine Rastersuche implementieren	352
Eine Zufallssuche versuchen	356

## ***Kapitel 19***

### ***Steigerung der Komplexität mit linearen und nichtlinearen Tricks*** **357**

Nichtlineare Transformationen nutzen	357
Variablentransformation ausüben	358
Interaktionen zwischen Variablen erstellen	360
Lineare Modelle regularisieren	364
Sich auf die Kamm-Regression (L2) verlassen	365
Das Lasso (L1) nutzen	365
Nutzung der Regularisierung	366
Elasticnet: L1 & L2 kombinieren	366
Kampf mit Big Data Stück für Stück	367
Bestimmen, ob es zu viele Daten sind	367
Implementierung des stochastischen Gradientenabstiegs	367
Support Vector Machines verstehen	370
Auf ein Berechnungsverfahren verlassen	371
Viele neue Parameter festlegen	373
Mit SVC klassifizieren	375
Nichtlinear arbeiten ist einfach	380
Regression mittels SVR ausführen	382
Stochastische Lösungen mit einer SVM erstellen	383

<b>Kapitel 20</b>	
<b>Die Macht der Vielen verstehen</b>	<b>387</b>
Mit einfachen Entscheidungsbäumen anfangen	388
Einen Entscheidungsbaum verstehen	388
Klassifikations- und Regressionsbäume erstellen	390
Maschinelles Lernen zugänglich machen	392
Mit einem Random Forest Classifier arbeiten	394
Mit einem Random-Forest-Regressor arbeiten	395
Einen Random Forest optimieren	396
Vorhersagen stärken	397
Wissen, dass viele schwache Prädiktoren gewinnen	397
Einen Gradient-Boosting-Klassifikator erstellen	398
Einen Gradient-Boosting-Regressor erstellen	399
GBM-Hyperparameter nutzen	400

**Teil VI**

**Der Top-Ten-Teil** **401**

<b>Kapitel 21</b>	
<b>Zehn wichtige Data-Science-Ressourcensammlungen</b>	<b>403</b>
Einblicke mit Data Science Weekly erhalten	404
Eine Ressourcenliste bei U-Climb-Higher erhalten	404
Einen guten Start mit KDnuggets	404
Auf die lange Liste von Ressourcen auf Data-Science-Central zugreifen	405
Die Fakten über Open-Source-Data-Science von Meistern erhalten	406
Gratis-Lernressourcen mit Quora aufspüren	406
Hilfe zu fortgeschrittenen Themen auf Conductrics erhalten	406
Neue Tricks vom Aspirational Data Scientist lernen	407
Data-Intelligence- und Analytics-Quellen auf AnalyticBridge finden	408
Mit Jonathan Bower die Ressourcen der Entwickler entdecken	408

<b>Kapitel 22</b>	
<b>Zehn Datenherausforderungen, die Sie annehmen sollten</b>	<b>409</b>
Der Data-Science-London + Scikit-learn-Herausforderung begegnen	410
Das Überleben auf der Titanic vorhersagen	410
Einen Kaggle-Wettbewerb finden, der Ihren Bedürfnissen entspricht	411
An Ihren Überanpassungsstrategien feilen	411
Durch den MovieLens-Datensatz gehen	412
Spam-E-Mails loswerden	413
Mit handgeschriebenen Informationen arbeiten	413
Mit Bildern arbeiten	414
Amazon.com-Reviews analysieren	415
Mit einem riesigen Graphen interagieren	416

<b>Stichwortverzeichnis</b>	<b>417</b>
-----------------------------	------------