

<b>Vorwort</b> .....	<b>15</b>
<b>Einleitung</b> .....	<b>19</b>
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

<b>1 Was ist das Problem?</b> .....	<b>37</b>
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
<b>2 Was sind Daten? . . . . .</b>	<b>47</b>
Daten oder Informationen? . . . . .	47
Ein Beispiel-Datensatz . . . . .	47
Datentypen . . . . .	49
Wie Daten gesammelt und strukturiert werden . . . . .	50
Beobachtungs-basierte versus experimentelle Daten . . . . .	50
Strukturierte versus unstrukturierte Daten . . . . .	51
Die Basics der zusammenfassenden Statistik . . . . .	52
Zusammenfassung . . . . .	53
<b>3 Vorbereitungen für das statistische Denken . . . . .</b>	<b>55</b>
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

<b>4 Daten infrage stellen . . . . .</b>	<b>71</b>
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

<b>5</b>	<b>Daten erkunden</b> .....	<b>85</b>
	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
<b>6</b>	<b>Wahrscheinlichkeiten untersuchen</b> .....	<b>99</b>
	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
<b>7</b>	<b>Hinterfragen Sie Statistiken</b> .....	<b>119</b>
	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

<b>8 Nach versteckten Gruppen suchen . . . . .</b>	<b>137</b>
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
<b>9 Das Regressionsmodell verstehen . . . . .</b>	<b>153</b>
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
<b>10 Das Klassifikationsmodell verstehen . . . . .</b>	<b>169</b>
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
<b>11 Textanalyse verstehen . . . . .</b>	<b>189</b>
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

<b>12</b>	<b>Konzepte des Deep Learning</b>	<b>209</b>
	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

<b>13</b>	<b>Achten Sie auf Fallstricke</b>	<b>233</b>
	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

<b>14 Menschen und Persönlichkeiten kennen</b> .....	<b>243</b>
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
<b>15 Was kommt danach?</b> .....	<b>251</b>
<b>Danksagungen</b> .....	<b>255</b>
<b>Index</b> .....	<b>257</b>