

Stefan Papp
Wolfgang Weidinger
Mario Meir-Huber
Bernhard Ortner
Georg Langs
Rania Wazir

Handbuch Data Science

Mit Datenanalyse und Machine Learning
Wert aus Daten generieren

HANSER

Inhalt

Vorwort	XIII
Kapitelübersicht	XV
1 Einleitung	1
<i>Bernhard Ortner, Stefan Papp, Mario Meir-Huber</i>	
1.1 Strategie	1
1.1.1 Wertschöpfungskettendigitalisierung (Value Chain Digitalization) ..	1
1.1.2 Marketing Segment Analytics	2
1.1.3 360° View of the Customer	3
1.1.4 Zusammenfassung	3
1.2 Umsetzung einer Digitalisierungsstrategie	4
1.2.1 Daten	5
1.2.2 Modellierung und Analyse	5
1.3 Teams	10
1.3.1 Data Scientist	10
1.3.2 Business Analyst	11
1.3.3 Data Architect	12
1.3.4 Data Engineer	12
1.3.5 DevOps	13
1.3.6 Weitere Rollen	13
1.3.7 Team Building	13
2 Grundlagen Datenplattformen	15
<i>Stefan Papp</i>	
2.1 Anforderungen	18
2.2 Systems Engineering	20
2.2.1 Hardware-Topologie	21
2.2.2 Cloud	29
2.2.3 Linux-Grundlagen	35
2.3 Datenplattform	51
2.3.1 Überblick	51

2.3.2	Verarbeitungskonzepte	54
2.3.3	Microservices	55
2.3.4	DFS	58
2.3.5	DWH	67
2.3.6	Object Storage	73
2.4	Data Engineering	73
2.4.1	DevOps	74
2.4.2	Programmierung	76
2.4.3	Design Patterns	83
2.4.4	ELT	87
2.4.5	Load mit Kafka	91
2.4.6	Transform mit Spark	93
2.5	Fazit	100
3	Datenarchitekturen	103
	<i>Mario Meir-Huber, Stefan Papp, Bernhard Ortner</i>	
3.1	Technologische Layer im Big Data Stack	103
3.1.1	Big Data Management	104
3.1.2	Big Data Platforms	105
3.1.3	Big Data Analytics	105
3.1.4	Big Data Utilization	106
3.2	Lambda und Kappa als Architekturparadigmen	107
3.2.1	Lambda-Architektur	107
3.2.2	Kappa-Architektur	110
3.2.3	Vergleich der beiden Architekturen	112
3.3	Operationalisierung des Data Lakes	113
3.3.1	Data Science Lab	114
3.4	Data Governance	115
3.4.1	Datenqualität	116
3.4.2	Datenkatalog	118
3.4.3	Business-Glossar	118
3.5	Metadatenmanagement	120
3.5.1	Stammdatenmanagement	120
3.6	Informationssicherheit	121
3.7	Fazit	122
4	Data Pipelines	125
	<i>Bernhard Ortner</i>	
4.1	Data Pipelines im Big-Data-Zeitalter	125
4.2	Anforderungen an eine Data Pipeline	127
4.3	Sechs Stufen der Data Pipeline	129
4.4	Automatisierung der Stufen	132

4.4.1	Datenerhebung	133
4.4.2	Datenbereinigung	133
4.5	AnalyticsOps und DataOps	134
4.6	Auslieferung	136
4.6.1	Containerization und Kubernetes	136
4.6.2	Modell-Update	137
4.6.3	Modell- oder Parameter-Update	138
4.6.4	Modell-Skalierung	139
4.7	Feedback in die operationalen Prozesse	139
4.8	Fazit	140
4.9	Weiterführende Literatur	140
5	Statistik-Grundlagen	141
	<i>Rania Wazir, Georg Langs</i>	
5.1	Daten	143
5.2	Aus Daten lernen – Grundlagen	144
5.2.1	Überwachtes Lernen	145
5.2.2	Bestärkendes Lernen	145
5.2.3	Unüberwachtes Lernen	146
5.3	Lineare Regression	146
5.3.1	Einfache lineare Regression	146
5.3.2	Multilineare Regression	154
5.4	Logistische Regression	158
5.5	Der Satz von Bayes	167
5.6	Wie gut ist der Algorithmus?	174
6	Machine Learning	177
	<i>Georg Langs, Rania Wazir</i>	
6.1	Grundlagen: Merkmale in Räumen	178
6.2	Übersicht über Klassifikationsmodelle	182
6.2.1	K-Nearest-Neighbor-Klassifikator	182
6.2.2	Support Vector Machines	183
6.2.3	Entscheidungsbäume	184
6.3	Ensemblemethoden	185
6.3.1	Bias und Varianz	186
6.3.2	Random Forests	188
6.3.3	Neuronale Netze und das Perzeptron	191
6.4	Gemeinsame Konzepte	193
6.5	In die Tiefe – Deep Learning	193
6.5.1	Convolutional Neural Networks	194
6.5.2	Recurrent Neural Networks	195
6.5.3	Long-term short-term memory	196
6.5.4	Andere Architekturen und Lernstrategien	197
6.6	Fazit	198

7	Rechtliche Grundlagen	199
	<i>Bernhard Ortner</i>	
7.1	Rechtliche Datenkategorien	199
7.2	Datenschutzgrundverordnung	200
7.2.1	Grundsätze der Datenschutzgrundverordnung	201
7.2.2	Einwilligungserklärung	202
7.2.3	Risikofolgeabschätzung	204
7.2.4	Anonymisierung und Pseudo-Anonymisierung	205
7.2.5	Arten der Anonymisierung	206
7.2.6	Rechtmäßigkeit, Transparenz und Verarbeitung	208
7.2.7	Recht auf Datenlöschung und Korrektur	209
7.2.8	Privacy by Design	210
7.2.9	Privacy by Default	210
7.3	ePrivacy-Verordnung	211
7.4	Datenschutzbeauftragter	211
7.4.1	Internationaler Datenexport in Drittländern	211
7.5	Sicherheitsmaßnahmen	212
7.5.1	Datensicherheit	213
7.6	Fazit	213
7.7	Weiterführende Literatur	214
8	Data Driven Enterprises	215
	<i>Mario Meir-Huber, Stefan Papp</i>	
8.1	Daten als Entrepreneurship-Thema	215
8.1.1	Digitalisierung – Freund oder Feind?	216
8.1.2	Das Team	217
8.1.3	Unternehmerische Datenreife	220
8.1.4	Startpunkt: die Roadmap	222
8.2	Die Plattform aus Business-Sicht	224
8.2.1	Mythos Open Source	224
8.2.2	Cloud	224
8.2.3	Vendorenauswahl	225
8.2.4	Data Lake aus Business-Sicht	226
8.2.5	Die Rolle der IT	226
8.2.6	Data Science Labs	227
8.3	Analytische Vorgehensmodelle	228
8.3.1	Analytical Use Case-Umsetzung	228
8.3.2	MEIR-Modell	228
8.3.3	Self Service Analytics	230
8.4	Fazit	230

9	AI in verschiedenen Branchen	231
	<i>Stefan Papp, Mario Meir-Huber, Wolfgang Weidinger, Thomas Tremel, Marek Danis</i>	
9.1	Automotive	235
9.1.1	Vision	236
9.1.2	Daten	236
9.1.3	Anwendungsfälle	237
9.1.4	Herausforderungen	238
9.2	Aviation	239
9.2.1	Vision	240
9.2.2	Daten	240
9.2.3	Anwendungsfälle	241
9.2.4	Herausforderungen	242
9.3	Energie	242
9.3.1	Vision	243
9.3.2	Daten	244
9.3.3	Anwendungsfälle	244
9.3.4	Herausforderungen	245
9.4	Finanzen	245
9.4.1	Vision	245
9.4.2	Daten	246
9.4.3	Anwendungsfälle	246
9.4.4	Herausforderungen	248
9.5	Gesundheit	248
9.5.1	Vision	249
9.5.2	Daten	250
9.5.3	Anwendungsfälle	250
9.5.4	Herausforderungen	251
9.6	Government	251
9.6.1	Vision	251
9.6.2	Daten	252
9.6.3	Anwendungsfälle	252
9.6.4	Herausforderungen	256
9.7	Kunst	256
9.7.1	Vision	257
9.7.2	Daten	257
9.7.3	Anwendungsfälle	257
9.7.4	Herausforderungen	258
9.8	Manufacturing	258
9.8.1	Vision	259
9.8.2	Daten	259
9.8.3	Anwendungsfälle	260
9.8.4	Herausforderungen	260

9.9	Öl und Gas	261
9.9.1	Vision	261
9.9.2	Daten	261
9.9.3	Anwendungsfälle	262
9.9.4	Herausforderungen	263
9.10	Sicherheit am Arbeitsplatz	264
9.10.1	Vision	264
9.10.2	Daten	265
9.10.3	Anwendungsfälle	265
9.10.4	Herausforderungen	266
9.11	Retail	267
9.11.1	Vision	267
9.11.2	Daten	267
9.11.3	Anwendungsfälle	268
9.11.4	Herausforderungen	268
9.12	Telekommunikationsanbieter	269
9.12.1	Vision	269
9.12.2	Daten	270
9.12.3	Anwendungsfälle	270
9.12.4	Herausforderungen	272
9.13	Transport	272
9.13.1	Vision	272
9.13.2	Daten	273
9.13.3	Anwendungsfälle	273
9.13.4	Herausforderungen	274
9.14	Unterricht und Ausbildung	274
9.14.1	Vision	274
9.14.2	Daten	275
9.14.3	Anwendungsfälle	275
9.14.4	Herausforderungen	276
9.15	Die digitale Gesellschaft	276
10	Mindset und Community	279
	<i>Stefan Papp</i>	
10.1	Data Driven Mindset	279
10.2	Data-Science-Kultur	281
10.2.1	Start-ups gegen Konzerne	281
10.2.2	Agile Softwareentwicklung	282
10.2.3	Firmen- und Arbeitskultur	283
10.3	Antipatterns	285
10.3.1	Abwertung der Domänenexpertise	286
10.3.2	Die IT wird es schon richten	287
10.3.3	Das war schon immer so	287

10.3.4	„Know it all“-Mentalität	288
10.3.5	Schwarzmalerei	288
10.3.6	Pfennigfuchseriei	289
10.3.7	Angstkultur	289
10.3.8	Kontrolle über die Ressourcen	289
10.3.9	Blindes Vertrauen in die Ressourcen	291
10.3.10	Over-Engineering	291
10.4	Fazit	292
11	Literatur	293
12	Die Autoren	297
Index	299