

Andreas Wierse, Till Riedel
Smart Data Analytics

Zusammenhänge erkennen
Potentiale nutzen
Big Data verstehen

DE GRUYTER
OLDENBOURG

Inhalt

Vorwort der Autoren — V

1 Einleitung — 1

- 1.1 Ein motivierendes Beispiel — 1
- 1.2 Für wen ist dieses Buch und wie kann man es lesen? — 8
- 1.3 Smart Data Solutions statt Big Data — 10
- 1.4 Das Smart Data Solution Center Baden-Württemberg — 14
 - 1.4.1 Warum ein Smart Data Solution Center? — 15
 - 1.4.2 Ablauf einer Potentialanalyse — 16
 - 1.4.3 Drei Beispiele — 17
 - 1.4.4 Die Partner — 21
 - 1.4.5 Das Smart Data Innovation Lab — 23

2 Grundlagen — 25

- 2.1 Smart Data vs. Big Data — 25
 - 2.1.1 Die 3Vs: Volume, Velocity, Variety — 26
 - 2.1.2 Veracity, Validity, Value — 27
 - 2.1.3 Variability, Venue, Vocabulary — 29
 - 2.1.4 Das verbliebene V: Vagueness — 30
 - 2.1.5 Smart Data — 31
- 2.2 Datengetriebene Innovation — 33
 - 2.2.1 Business Intelligence und Verbesserungsprozesse — 35
 - 2.2.2 Operative Geschäftsdaten für Innovation nutzen — 36
 - 2.2.3 Vom eingebetteten System zum Datensee — 37
 - 2.2.4 Kontextsensitive Systeme — 39
- 2.3 Data Analytics und Maschinelles Lernen — 46
 - 2.3.1 Business Analytics — 49
 - 2.3.2 Klassifikation eines Merkmalsraums — 50
 - 2.3.3 Supervised Learning — 53
 - 2.3.4 Prädiktion und prädiktive Analyse — 57
- 2.4 Die Bewertung von Vorhersagen — 60
 - 2.4.1 Fehlermaße als Bewertungsfunktion — 60
 - 2.4.2 Validierungsschema — 66
 - 2.4.3 Automatische Verbesserung von Klassifikatoren — 73
- 2.5 Merkmale und Datentypen — 77
 - 2.5.1 Automatische Merkmalsselektion und -bewertung — 80
 - 2.5.2 Lernen von Merkmalen — 83

2.5.3	Zeitreihen und Sensordaten —	93
2.5.4	Texte —	99
2.5.5	Graphen, Linked Data, geographische Daten —	101
2.5.6	Geographische Daten —	103
3	Visualisierung und Interpretation —	106
3.1	Der menschliche Wahrnehmungsapparat —	108
3.2	Übersicht gebräuchlicher Visualisierungsmethoden —	113
3.2.1	Balken- und Säulendiagramm —	113
3.2.2	Histogramm —	116
3.2.3	Tortendiagramm —	119
3.2.4	Netzdiagramm —	120
3.2.5	Kalender-Heat Map —	123
3.2.6	Raumbezogene Geodaten —	126
3.2.7	Linendiagramm —	127
3.2.8	Mindmap —	128
3.2.9	Social Media-Netzwerkdiagramm —	129
3.2.10	Graph —	131
3.2.11	Kissendiagramm —	135
3.2.12	Sehendiagramm —	137
3.2.13	Box-Plot —	138
3.2.14	Punkt- oder Streudiagramm, Scatter-Plot —	141
3.2.15	Dichte-Plot —	143
3.3	Interaktive Visualisierung —	145
3.3.1	Das bewegte Bild —	145
3.3.2	Erkundung des Hypothesenraums —	147
3.3.3	Höhere Dimensionen —	148
3.3.4	Exploration —	150
3.4	Interpretation —	151
4	Praxisbeispiele —	158
4.1	Voraussage der Auftragsbearbeitungszeit —	158
4.1.1	Daten —	159
4.1.2	Analyse —	164
4.1.3	Bewertung —	167
4.2	Zustandsbasierte Wartung —	171
4.2.1	Daten —	172
4.2.2	Analyse —	174
4.2.3	Bewertung —	177
4.3	Fehler in Protokollen vorhersagen —	180

- 4.3.1 **Daten — 181**
- 4.3.2 **Analyse — 182**
- 4.3.3 **Bewertung — 185**
- 4.4 **Fehlerursachen lokalisieren — 187**
- 4.4.1 **Daten — 188**
- 4.4.2 **Analyse — 189**
- 4.4.3 **Bewertung — 192**
- 4.5 **Materialnutzung optimieren — 194**
- 4.5.1 **Daten — 195**
- 4.5.2 **Analyse — 196**
- 4.5.3 **Bewertung — 200**
- 4.6 **Energieverbrauch auf die Schliche kommen — 201**
- 4.6.1 **Daten — 202**
- 4.6.2 **Analyse — 204**
- 4.6.3 **Bewertung — 205**
- 4.7 **Qualitätsschwankungen verstehen — 208**
- 4.8 **Schneller den Kunden- oder Partnerpool erweitern — 211**
- 4.9 **Kündigungen verhindern und Kunden binden — 213**
- 4.10 **Mehrfachmeldungen zusammenfassen — 214**
- 4.11 **Den perfekten Moment abpassen — 217**
- 5 Organisatorische Anforderungen — 221**
- 5.1 **Prozesse — 221**
- 5.1.1 **Ein einfacher Prozess — 222**
- 5.1.2 **Eine andere Sicht auf den Prozess — 225**
- 5.1.3 **Cross Industry Standard Process for Data Mining (CRISP-DM) — 228**
- 5.2 **Teams — 235**
- 5.2.1 **Ausschließlich externer *Smart Data Analytics*-Partner — 236**
- 5.2.2 **Zusammenarbeit interner und externer Smart Data Analytics-Experten — 239**
- 5.2.3 **Rein internes Smart Data Analytics Team — 240**
- 5.3 **Geschäftsmodelle — 243**
- 5.3.1 **Vom Kauf zur Abrechnung nach Nutzung — 243**
- 5.3.2 **Wertschöpfung aus den Daten — 244**
- 5.3.3 **Smart Data als eigenes Geschäftsfeld — 246**
- 5.4 **Fallstricke und Gefahren — 248**
- 5.4.1 **Smart Data Analytics ist anders — 248**
- 5.4.2 **Alles in den Data Lake? — 250**

5.4.3	Hidden Biases —	252
5.4.4	Big Data - Inklusion oder Exklusion? —	254
5.4.5	Seien Sie skeptisch —	255
5.4.6	Maßnahmen —	258
6	Datenschutz und Schutzrechte —	263
6.1	Datenschutz gilt bei personenbezogenen Daten —	264
6.1.1	Welche Regelungen zum Datenschutz gibt es? —	266
6.1.2	Der Schutz der Daten schützt Menschen —	267
6.1.3	Grundsätze des Datenschutzrechts —	268
6.1.4	Datenschutzrecht greift nur bei persönlichen Daten und bestimmten Handlungen —	269
6.1.5	Die Einwilligung im Datenschutzrecht —	270
6.1.6	Welche Rechte hat die Betroffene? —	271
6.1.7	Pflichten für Unternehmen —	272
6.1.8	Sanktionen bei Verstößen gegen das Datenschutzrecht —	273
6.2	Arbeitnehmerdatenschutz/Beschäftigtendatenschutz —	277
6.2.1	Rechtliche Grundlagen —	277
6.2.2	„Erlaubnis“ bei Begründung oder Durchführung des Beschäftigungsverhältnisses —	278
6.2.3	Neue europarechtliche Regelungen —	279
6.2.4	Fallkonstellationen des Beschäftigtendatenschutzes —	280
6.2.5	Mitbestimmung —	282
6.2.6	Die Einwilligung in die Erhebung, Verarbeitung und Nutzung von Beschäftigtendaten —	284
6.3	Der Datenschutzbeauftragte im Unternehmen —	284
6.3.1	Wann ist ein Datenschutzbeauftragter zu bestellen? —	284
6.3.2	Welche Aufgaben hat der Datenschutzbeauftragte? —	285
6.3.3	Bestellung und Eingliederung —	286
6.3.4	Die neuen europarechtlichen Vorgaben —	287
6.4	Auftragsdatenverarbeitung —	288
6.4.1	Was ist Auftragsdatenverarbeitung? —	288
6.4.2	Neue europarechtliche Vorgaben —	290
6.5	Der Schutz des Datenbankherstellers —	291
6.5.1	Was ist eine Datenbank und unter welchen Voraussetzungen ist sie geschützt? —	292
6.5.2	Welche Rechte hat der Datenbankhersteller? —	294
6.5.3	Der Datenbankhersteller wird nicht schrankenlos geschützt —	295

6.5.4	Grenzen der Vertragsfreiheit — 295
7	Technologie — 297
7.1	Von Lambda und Kappa Architekturen — 297
7.1.1	Batchverarbeitung — 298
7.1.2	Echtzeitverarbeitung — 299
7.1.3	Lambda-Architektur: das Beste beider Welten — 301
7.1.4	Scoring von maschinellem Lernen — 302
7.1.5	Kappa-Architektur — 305
7.2	Skalierung mit Apache Hadoop — 306
7.2.1	Verteilte Dateisysteme — 306
7.2.2	Verteilte Berechnung — 309
7.2.3	Spark — 316
7.3	Big Data und Datenbanken — 318
7.3.1	NewSQL, BeyondSQL, SAP HANA — 320
7.3.2	Analytics Datenbanken — 321
7.3.3	Zeitreihen- und Log-Datenbanken — 323
7.4	Streaming — 325
7.4.1	Complex Event Processing — 325
7.4.2	Distributed Streaming System — 327
7.5	Plattformunabhängigkeit und GPU-Frameworks — 330
7.6	Analyseumgebungen — 332
7.6.1	Excel — 333
7.6.2	SPSS Modeler — 336
7.6.3	RapidMiner — 338
7.6.4	KNIME — 339
7.6.5	Orange, Weka — 339
7.6.6	Spezialisierte Werkzeuge zum unüberwachten Lernen — 341
7.7	Programmiersprachen und Notebooks — 342
7.7.1	R — 342
7.7.2	Python und <i>scikitLearn</i> — 345
7.7.3	Interaktive Notebooks — 349
7.7.4	Mehr Programmiersprachen und Beispiel-Code — 351
8	Wirtschaftliche Betrachtung — 353
8.1	Kosten — 356
8.1.1	Software — 356
8.1.2	Hardware — 366
8.1.3	Infrastruktur — 371
8.1.4	Installation und Inbetriebnahme sowie Wartung — 373

8.1.5	Datenschnittstellen —	374
8.1.6	Datenaufbereitung —	374
8.1.7	Prozessanbindung —	377
8.1.8	Mitarbeiter —	378
8.1.9	Mitarbeiterschulung/-weiterbildung —	381
8.1.10	Unterstützung durch Dienstleister —	382
8.2	Cloud vs. On-Premise —	383
8.2.1	Wesentliche Charakteristika —	383
8.2.2	Service-Modelle —	385
8.2.3	Einsatzmodelle —	388
8.2.4	Abwägung: Cloud vs. On-Premise —	395
8.3	Return on Investment —	400
8.3.1	Das Problem der Skalierung —	401
8.3.2	Vorgehensweise —	403
8.3.3	Von anderen lernen —	404
9	Epilog —	407
	Stichwortverzeichnis —	423