

Inhalt

Geleitwort	XV
Vorwort	XIX
1 Einführung	1
1.1 Was sind Data Science, Machine Learning und Künstliche Intelligenz?	2
1.2 Datenstrategie	9
1.3 Von der Strategie zu den Anwendungsfällen	11
1.3.1 Datenteams	11
1.3.2 Daten und Plattformen	18
1.3.3 Modellierung und Analyse	19
1.4 Implementierung von Anwendungsfällen	19
1.4.1 Iterative Erkundung von Anwendungsfällen	20
1.4.2 End-to-End-Datenverarbeitung	23
1.4.3 Datenprodukte	23
1.5 Beispiele für reale Anwendungsfälle	24
1.5.1 Digitalisierung der Wertschöpfungskette	24
1.5.2 Marketing-Segment-Analyse	25
1.5.3 360°-Sicht auf den Kunden	25
1.5.4 Anwendungsfälle für NGOs und Nachhaltigkeit	26
1.6 Ergebnisse liefern	27
1.7 Kurz und bündig	30
2 Infrastruktur	31
2.1 Einführung	31
2.2 Hardware	33
2.2.1 Verteilte Systeme	36
2.2.2 Hardware für KI-Anwendungen	39
2.3 Linux Essentials für Datenexperten	41
2.4 Terraform	59
2.5 Cloud	63
2.5.1 Basisdienste	66
2.5.2 Cloud-native Lösungen	70
2.6 Kurz und bündig	73

3	Datenarchitektur	75
3.1	Übersicht	75
3.1.1	Maslowsche Bedürfnishierarchie für Daten	76
3.1.2	Anforderungen an die Datenarchitektur	77
3.1.3	Die Struktur einer typischen Datenarchitektur	78
3.1.4	ETL (Extrahieren, Transformieren, Laden)	78
3.1.5	ELT (Extrahieren, Laden, Transformieren)	79
3.1.6	ETLT	80
3.2	Datenerfassung und -integration	80
3.2.1	Datenquellen	81
3.2.2	Traditionelle Dateiformate	82
3.2.3	Moderne Dateiformate	84
3.2.4	Zusammenfassung	86
3.3	Data Warehouses, Data Lakes und Lakehouses	87
3.3.1	Data Warehouses	87
3.3.2	Data Lakes und das Lakehouse	91
3.3.3	Zusammenfassung: Vergleich zwischen Data Warehouses und Lakehouses	93
3.4	Datenverarbeitung und -umwandlung	94
3.4.1	Big Data und Apache Spark	94
3.4.2	Databricks	101
3.5	Workflow-Orchestrierung	103
3.6	Ein Datenarchitektur-Use-Case	105
3.7	Kurz und bündig	110
4	Data Engineering	112
4.1	Integration von Daten	113
4.1.1	Daten-Pipelines	113
4.1.2	Entwerfen von Data Pipelines	120
4.1.3	CI/CD	123
4.1.4	Programmiersprachen	124
4.1.5	Kafka als Referenz-ETL-Werkzeug	127
4.1.6	Entwurfsmuster	132
4.1.7	Automatisierung der Stufen	133
4.1.8	Sechs Bausteine der Data Pipeline	134
4.2	Verwaltung analytischer Modelle	139
4.2.1	Modelllieferung	140
4.2.2	Modell-Update	141
4.2.3	Modell- oder Parameter-Update	142
4.2.4	Modellskalierung	143
4.3	Feedback in die operationalen Prozesse	143
4.4	Kurz und bündig	144

5	Datenmanagement	145
5.1	Datenmanagement	147
5.1.1	Datenkatalog	149
5.1.2	Data Discovery	151
5.1.3	Datenqualität	154
5.1.4	Verwaltung von Stammdaten	156
5.1.5	Gemeinsame Nutzung von Daten	157
5.2	Informationssicherheit	158
5.2.1	Datenklassifizierung	159
5.2.2	Schutz der Privatsphäre	161
5.2.3	Verschlüsselung	163
5.2.4	Secrets Management	165
5.2.5	Defense in Depth	166
5.3	Kurz und bündig	167
6	Mathematik	168
6.1	Lineare Algebra	169
6.1.1	Vektoren und Matrizen	169
6.1.2	Operationen zwischen Vektoren und Matrizen	173
6.1.3	Lineare Transformationen	175
6.1.4	Eigenwerte, Eigenvektoren und Eigendekomposition	176
6.1.5	Andere Matrixzerlegungen	178
6.2	Kalkulus und Optimierung	180
6.2.1	Ableitung	180
6.2.2	Gradient und Hessian	182
6.2.3	Gradientenabstieg	184
6.2.4	Eingeschränkte Optimierung	186
6.3	Wahrscheinlichkeitsrechnung	187
6.3.1	Diskrete und kontinuierliche Zufallsvariablen	188
6.3.2	Erwartungswert, Varianz und Kovarianz	192
6.3.3	Unabhängigkeit, bedingte Verteilungen und Bayes-Theorem	193
6.4	Kurz und bündig	195
7	Statistik – Grundlagen	196
7.1	Daten	197
7.2	Einfache lineare Regression	198
7.3	Multiple lineare Regression	206
7.4	Logistische Regression	209
7.5	Wie gut ist unser Modell?	217
7.6	Kurz und bündig	218

8	Maschinelles Lernen	220
8.1	Einführung	220
8.2	Grundlegendes: Feature Spaces	222
8.3	Klassifizierungsmodelle	226
8.3.1	K-Nearest-Neighbor-Klassifikator	226
8.3.2	Support Vector Machine	227
8.3.3	Entscheidungsbaum	228
8.4	Ensemble-Methoden	230
8.4.1	Bias und Varianz	230
8.4.2	Bagging: Random Forests	232
8.4.3	Boosten: AdaBoost	235
8.5	Künstliche neuronale Netze und das Perceptron	236
8.6	Lernen ohne Label – Struktur finden	239
8.6.1	Clustering	239
8.6.2	Lernen von Mannigfaltigkeiten	240
8.6.3	Generative Modelle	241
8.7	Reinforcement Learning	242
8.8	Übergreifende Konzepte	245
8.9	In die Tiefe gehen – Deep Learning	246
8.9.1	Convolutional Neural Networks	246
8.9.2	Training von Convolutional Neural Networks	248
8.9.3	Recurrent Neural Networks	250
8.9.4	Long Short-Term Memory	251
8.9.5	Autoencoder und U-Netze	253
8.9.6	Adversarial-Trainingsansätze	254
8.9.7	Generative Adversarial Networks	255
8.9.8	Cycle GANs und Style GANs	257
8.9.9	Andere Architekturen und Lernstrategien	258
8.10	Validierungsstrategien für maschinelle Lerntechniken	259
8.11	Schlussfolgerung	260
8.12	Kurz und bündig	261
9	Großartige künstliche Intelligenz erschaffen	262
9.1	Wie KI mit Data Science und maschinellem Lernen zusammenhängt	262
9.2	Eine kurze Geschichte der KI	266
9.3	Fünf Empfehlungen für die Entwicklung einer KI-Lösung	268
9.3.1	Empfehlung Nr. 1: Seien Sie pragmatisch	268
9.3.2	Empfehlung Nr. 2: Erleichtern Sie Maschinen das Lernen – schaffen Sie induktive Verzerrungen	271
9.3.3	Empfehlung Nr. 3: Analysen durchführen	277
9.3.4	Empfehlung Nr. 4: Hüten Sie sich vor der Skalierungsfalle	279
9.3.5	Empfehlung Nr. 5: Hüten Sie sich vor der Verallgemeinerungsfalle (so etwas wie ein kostenloses Mittagessen gibt es nicht)	289

9.4	Intelligenz auf menschlicher Ebene	294
9.5	Kurz und bündig	297
10	Natural Language Processing (NLP)	299
10.1	Was ist NLP, und warum ist es so wertvoll?	299
10.2	NLP-Datenaufbereitungstechniken	301
10.2.1	Die NLP-Pipeline	301
10.2.2	Konvertierung des Eingabeformats für maschinelles Lernen	308
10.3	NLP-Aufgaben und -Methoden	310
10.3.1	Regelbasiert (symbolisch) NLP	311
10.3.2	Ansätze des statistischen maschinellen Lernens	314
10.3.3	Neurales NLP	323
10.3.4	Transferlernen	329
10.4	Auf dem neuesten Stand: Aktuelle Forschungsschwerpunkte für NLP	342
10.5	Kurz und bündig	345
11	Computer Vision	348
11.1	Was ist Computer Vision?	348
11.2	Ein Bild sagt mehr als tausend Worte	350
11.2.1	Das menschliche Auge	350
11.2.2	Das Bildaufnahmeprinzip	352
11.2.3	Digitale Dateiformate	357
11.2.4	Bildkomprimierung	359
11.3	Ich sehe was, was du nicht siehst	360
11.3.1	Computergestützte Fotografie und Bildmanipulation	363
11.4	Computer-Vision-Anwendungen und zukünftige Richtungen	366
11.4.1	Image-Retrieval-Systeme	367
11.4.2	Objekterkennung, Klassifizierung und Verfolgung	369
11.4.3	Medizinische Computer Vision	371
11.5	Menschen sehen lassen	375
11.6	Kurz und bündig	377
12	Modellierung und Simulation – Erstellen Sie Ihre eigenen Modelle	379
12.1	Einführung	380
12.2	Allgemeine Aspekte	381
12.3	Modellierung zur Beantwortung von Fragen	382
12.4	Reproduzierbarkeit und Lebenszyklus des Modells	384
12.4.1	Der Lebenszyklus einer Modellierungs- und Simulationsfrage	386
12.4.2	Parameter- und Output-Definition	387
12.4.3	Dokumentation	390
12.4.4	Verifizierung und Validierung	391

12.5	Methoden	395
12.5.1	Gewöhnliche Differentialgleichungen (ODEs)	396
12.5.2	Systemdynamik (SD)	397
12.5.3	Diskrete Ereignissimulation	400
12.5.4	Agentenbasierte Modellierung	403
12.6	Beispiele für Modellierung und Simulation	406
12.6.1	Dynamische Modellierung von Eisenbahnnetzen zur optimalen Wegfindung mit agentenbasierten Methoden und Reinforcement Learning	406
12.6.2	Strategien zur agentenbasierten Covid-Modellierung	409
12.6.3	Deep-Reinforcement-Learning-Ansatz für eine optimale Nachschubpolitik in einer VMI-Umgebung	414
12.7	Zusammenfassung und Lessons Learned	417
12.8	Kurz und bündig	418
13	Visualisierung von Daten	422
13.1	Geschichte	423
13.2	Welche Tools Sie verwenden sollten	429
13.3	Arten von Datenvisualisierungen	431
13.3.1	Streudiagramm	432
13.3.2	Liniendiagramm	432
13.3.3	Säulen- und Balkendiagramme	433
13.3.4	Histogramm	434
13.3.5	Tortendiagramm	435
13.3.6	Box Plot	436
13.3.7	Heat Map	436
13.3.8	Baumdiagramm	437
13.3.9	Andere Arten von Visualisierungen	438
13.4	Wählen Sie die richtige Datenvisualisierung	438
13.5	Tipps und Tricks	441
13.6	Präsentation der Datenvisualisierung	446
13.7	Kurz und bündig	447
14	Datengetriebene Unternehmen	449
14.1	Die drei Ebenen eines datengesteuerten Unternehmens	450
14.2	Kultur	450
14.2.1	Unternehmensstrategie für Daten	451
14.2.2	Die Analyse des aktuellen Stands	453
14.2.3	Unternehmenskultur und Organisation einer erfolgreichen Datenorganisation	455
14.2.4	Kernproblem: der Fachkräftemangel	463
14.3	Technologie	465
14.3.1	Die Auswirkungen von Open Source	465
14.3.2	Cloud	466

14.3.3	Auswahl des Anbieters	466
14.3.4	Data Lake aus der Unternehmensperspektive	467
14.3.5	Die Rolle der IT	468
14.3.6	Data Science Labs	468
14.3.7	Revolution in der Architektur: das Data Mesh	469
14.4	Business	470
14.4.1	Daten kaufen und teilen	470
14.4.2	Implementierung des analytischen Anwendungsfalls	472
14.4.3	Self-Service Analytics	472
14.5	Kurz und bündig	473
15	Rechtliche Grundlagen	474
15.1	Einführung	474
15.2	Rechtliche Datenkategorien	475
15.3	Datenschutzgrundverordnung	476
15.3.1	Grundsätze der Datenschutzgrundverordnung	477
15.3.2	Einwilligungserklärung	478
15.3.3	Risikofolgeabschätzung	480
15.3.4	Anonymisierung und Pseudo-Anonymisierung	481
15.3.5	Arten der Anonymisierung	481
15.3.6	Rechtmäßigkeit, Transparenz und Verarbeitung	484
15.3.7	Recht auf Datenlöschung und Korrektur	485
15.3.8	Privacy by Design	486
15.3.9	Privacy by Default	486
15.4	ePrivacy-Verordnung	487
15.5	Datenschutzbeauftragter	487
15.5.1	Internationaler Datenexport in Drittländern	488
15.6	Sicherheitsmaßnahmen	488
15.6.1	Datensicherheit	489
15.7	Datenschutz in Kalifornien im Vergleich zur DSGVO	489
15.7.1	Territoriale Gültigkeit	490
15.7.2	Opt-in versus Opt-out	490
15.7.3	Recht auf Datenexport	491
15.7.4	Das Recht, nicht diskriminiert zu werden	491
15.8	Kurz und bündig	492
15.9	Weiterführende Literatur	493
16	AI in verschiedenen Branchen	494
16.1	Automobilindustrie	498
16.1.1	Vision	499
16.1.2	Daten	499
16.1.3	Anwendungsfälle	500
16.1.4	Herausforderungen	501

16.2	Luftfahrt	502
16.2.1	Vision	503
16.2.2	Daten	504
16.2.3	Anwendungsfälle	504
16.2.4	Herausforderungen	505
16.3	Energie	506
16.3.1	Vision	506
16.3.2	Daten	507
16.3.3	Anwendungsfälle	507
16.3.4	Herausforderungen	508
16.4	Finanzen	509
16.4.1	Vision	509
16.4.2	Daten	509
16.4.3	Anwendungsfälle	510
16.4.4	Herausforderungen	512
16.5	Gesundheit	512
16.5.1	Vision	513
16.5.2	Daten	514
16.5.3	Anwendungsfälle	514
16.5.4	Herausforderungen	515
16.6	Regierung	515
16.6.1	Vision	515
16.6.2	Daten	516
16.6.3	Anwendungsfälle	516
16.6.4	Herausforderungen	520
16.7	Kunst	520
16.7.1	Vision	521
16.7.2	Daten	521
16.7.3	Anwendungsfälle	522
16.7.4	Herausforderungen	522
16.8	Produktion	523
16.8.1	Vision	523
16.8.2	Daten	523
16.8.3	Anwendungsfälle	524
16.8.4	Herausforderungen	525
16.9	Öl und Gas	525
16.9.1	Vision	526
16.9.2	Daten	526
16.9.3	Anwendungsfälle	527
16.9.4	Herausforderungen	528
16.10	Sicherheit am Arbeitsplatz	529
16.10.1	Vision	529
16.10.2	Daten	530
16.10.3	Anwendungsfälle	530
16.10.4	Herausforderungen	531

16.11 Einzelhandel	532
16.11.1 Vision	532
16.11.2 Daten	533
16.11.3 Anwendungsfälle	533
16.11.4 Herausforderungen	534
16.12 Anbieter von Telekommunikation	534
16.12.1 Vision	535
16.12.2 Daten	535
16.12.3 Anwendungsfälle	535
16.12.4 Herausforderungen	537
16.13 Transport	538
16.13.1 Vision	538
16.13.2 Daten	539
16.13.3 Anwendungsfälle	539
16.13.4 Herausforderungen	539
16.14 Lehre und Ausbildung	540
16.14.1 Vision	540
16.14.2 Daten	541
16.14.3 Anwendungsfälle	542
16.14.4 Herausforderungen	542
16.15 Die digitale Gesellschaft	543
16.16 Kurz und bündig	545
17 Mindset und Community	546
17.1 Data Driven Mindset	546
17.2 Data-Science-Kultur	549
17.2.1 Start-up oder Beratungsunternehmen?	549
17.2.2 Labs statt Konzernpolitik	550
17.2.3 Keiretsu statt Einzelkämpfertum	551
17.2.4 Agile Softwareentwicklung	552
17.2.5 Firmen- und Arbeitskultur	553
17.3 Antipatterns	556
17.3.1 Abwertung von Fachwissen	556
17.3.2 Die IT wird es schon richten	557
17.3.3 Widerstand gegen Veränderungen	558
17.3.4 Besserwisser-Mentalität	558
17.3.5 Schwarzmalerei	559
17.3.6 Pfennigfuchseriei	560
17.3.7 Angstkultur	560
17.3.8 Kontrolle über die Ressourcen	561
17.3.9 Blindes Vertrauen in die Ressourcen	561
17.3.10 Das Schweizer Taschenmesser	562
17.3.11 Over-Engineering	563
17.4 Kurz und bündig	564

18	Vertrauenswürdige KI	565
18.1	Rechtlicher und Soft-Law-Rahmen	566
18.1.1	Normen	568
18.1.2	Verordnungen	569
18.2	KI-Stakeholder	571
18.3	Fairness in der KI	572
18.3.1	Bias	573
18.3.2	Fairness-Metriken	576
18.3.3	Unerwünschten Bias in KI-Systemen reduzieren	580
18.4	Transparenz von KI-Systemen	581
18.4.1	Dokumentieren der Daten	582
18.4.2	Dokumentieren des Modells	584
18.4.3	Explainability (Erklärbarkeit)	585
18.5	Schlussfolgerung	587
18.6	Kurz und bündig	587
19	Die Autor:innen	588
	Index	593