

Handbuch Data Science und KI

Mit Machine Learning und Datenanalyse

Wert aus Daten generieren

DAS INHALTS- VERZEICHNIS

» Hier geht's
direkt
zum Buch

Inhalt

Vorwort	XXV
1 Einführung	1
<i>Stefan Papp</i>	
1.1 Über dieses Buch	1
1.2 Die Halford Group	2
1.2.1 Alice Halford – Vorsitzende	3
1.2.2 Analysten	5
1.2.3 „CDO“	6
1.2.4 Vertrieb	7
1.2.5 IT	8
1.2.6 Sicherheit	10
1.2.7 Leiter der Produktion	11
1.2.8 Kundenbetreuung	12
1.2.9 HR	13
1.2.10 CEO	14
1.3 Kurz und bündig	16
2 Das A und O der KI	19
<i>Stefan Papp</i>	
2.1 Die Datenverwendungszwecke	20
2.1.1 Bias	20
2.1.2 Datenkompetenz	23
2.2 Kulturschock	25

- 2.3 Ideenfindung 29
- 2.4 Entwurfsprozessmodelle 31
 - 2.4.1 Design Thinking 32
 - 2.4.2 Double Diamond 32
 - 2.4.3 Durchführung von Workshops 34
- 2.5 Kurz und bündig 41
- 3 Cloud-Dienste 43**
Stefan Papp
 - 3.1 Einführung 43
 - 3.2 Cloud-Essentials 44
 - 3.2.1 XaaS 46
 - 3.2.2 Cloud-Anbieter 48
 - 3.2.3 Native Cloud-Dienste 49
 - 3.2.4 Cloud-native Paradigmen 53
 - 3.3 Infrastructure as a Service 54
 - 3.3.1 Hardware 55
 - 3.3.2 Verteilte Systeme 57
 - 3.3.3 Linux Essentials für Datenexperten 60
 - 3.3.4 Infrastructure as Code 67
 - 3.4 Platform as a Service 72
 - 3.4.1 Cloud Native PaaS-Lösungen 73
 - 3.4.2 Externe Lösungen 78
 - 3.5 Software as a Service 81
 - 3.6 Kurz und bündig 82
- 4 Datenarchitektur 83**
Zoltan C. Toth, Sean McIntyre
 - 4.1 Übersicht 83
 - 4.1.1 Maslowsche Bedürfnishierarchie für Daten 84
 - 4.1.2 Anforderungen an die Datenarchitektur 85
 - 4.1.3 Die Struktur einer typischen Datenarchitektur 86
 - 4.1.4 ETL (Extrahieren, Transformieren, Laden) 91
 - 4.1.5 ELT (Extrahieren, Laden, Transformieren) 92
 - 4.1.6 ETLT 93

4.2	Datenerfassung und -integration	93
4.2.1	Datenquellen	93
4.2.2	Traditionelle Dateiformate	95
4.2.3	Moderne Dateiformate	98
4.2.4	Welche Speicheroption soll ich wählen?	100
4.3	Data Warehouses, Data Lakes und Lakehouses	101
4.3.1	Data Warehouses	101
4.3.2	Data Lakes und Cloud-Datenplattformen	106
4.4	Data Transformation	109
4.4.1	SQL	111
4.4.2	Big Data & Apache Spark	122
4.4.3	Cloud-Datenplattformen für Apache Spark	131
4.5	Workflow-Orchestrierung	133
4.5.1	Dagster und der Modern Data Stack	135
4.6	Ein Anwendungsfall und seine Datenarchitektur	137
4.7	Kurz und bündig	142
5	Data Engineering	145
	<i>Stefan Papp</i>	
5.1	Abgrenzung zum Software Engineering	146
5.2	Programmiersprachen	148
5.2.1	Code oder kein Code?	148
5.2.2	Auswahl der Programmiersprache	149
5.2.3	Python	151
5.2.4	Scala	155
5.3	Software-Engineering-Prozesse für Daten	157
5.3.1	Konfigurationsmanagement	158
5.3.2	CI/CD	159
5.4	Datenpipelines	161
5.4.1	Gemeinsame Merkmale einer Datenpipeline	161
5.4.2	Datenpipelines in der Unified Data Architecture	162
5.5	Speicheroptionen	166
5.5.1	Datei-Ära	166
5.5.2	Datenbank-Ära	167
5.5.3	Data-Lake-Ära	169

- 5.5.4 Serverless-Ära 170
 - 5.5.5 Polyglotte Speicherung 171
 - 5.5.6 Data-Mesh-Ära 172
- 5.6 Tooling 173
 - 5.6.1 Batch: Airflow 174
 - 5.6.2 Streaming: Kafka 175
 - 5.6.3 Transformation: Databricks Notebooks 181
- 5.7 Gemeinsame Herausforderungen 182
 - 5.7.1 Datenqualität und unterschiedliche Standards 183
 - 5.7.2 Skewed Data 184
 - 5.7.3 Überlastete operationelle Systeme 185
 - 5.7.4 Operative Legacy-Systeme 187
 - 5.7.5 Plattform- und Informationssicherheit 187
- 5.8 Kurz und bündig 187
- 6 Data Governance..... 189**
Victoria Rugli, Mario Meir-Huber
- 6.1 Warum brauchen wir Data Governance? 189
 - 6.1.1 Beispiel 1: Mit Data Governance Klarheit schaffen 190
 - 6.1.2 Beispiel 2: Die (negativen) Auswirkungen einer mangelhaften Data Governance 191
- 6.2 Die Bausteine der Data Governance 192
 - 6.2.1 Data Governance erklärt 194
- 6.3 Menschen 196
 - 6.3.1 Data Ownership 197
 - 6.3.2 Data Stewards 200
 - 6.3.3 Data Governance Board 202
 - 6.3.4 Change Management 203
- 6.4 Prozess 205
 - 6.4.1 Verwaltung von Metadaten 206
 - 6.4.2 Management der Datenqualität 210
 - 6.4.3 Datensicherheit und Datenschutz 215
 - 6.4.4 Stammdatenmanagement 219
 - 6.4.5 Datenzugang und Suche 223

6.5	Technologie (Datenmanagement-Tools)	226
6.5.1	Open-Source-Tools	226
6.5.2	Cloud-basierte Data-Governance-Tools	233
6.6	Kurz und bündig	238
7	Machine Learning Operations (ML Ops)	239
	<i>Zoltan C. Toth, György Móra</i>	
7.1	Übersicht	239
7.1.1	Umfang von MLOps	240
7.1.2	Datenerhebung und Exploration	241
7.1.3	Feature Engineering	241
7.1.4	Modelltraining	242
7.1.5	In der Produktion eingesetzte Modelle	242
7.1.6	Bewertung des Modells	243
7.1.7	Model Understanding	244
7.1.8	Modellversionierung	244
7.1.9	Modellüberwachung	245
7.2	MLOps in einer Organisation	245
7.2.1	Die wichtigsten Vorteile von MLOps	246
7.2.2	Erforderliche Fähigkeiten für MLOps	246
7.3	Verschiedene gängige Szenarien im MLOps-Bereich	246
7.3.1	Integration von Notebooks	247
7.3.2	Features in der Produktion	248
7.3.3	Modelleinsatz	251
7.3.4	Modellformate	251
7.4	MLOps-Werkzeuge und MLflow	253
7.4.1	MLflow	254
7.5	Kurz und bündig	257
8	Cybersicherheit von Machine-Learning-Systemen	259
	<i>Manuel Pasieka</i>	
8.1	Einführung in die Cybersicherheit	260
8.2	Angriffsfläche	261
8.3	Angriffsmethoden	263
8.3.1	Model Stealing	263
8.3.2	Datenextraktion	265

8.3.3	Data Poisoning	268
8.3.4	Adversariale Angriffe	271
8.3.5	Backdoor-Angriffe	273
8.4	Sicherheit von großen Sprachmodellen durch maschinelles Lernen	276
8.4.1	Datenextraktion	277
8.4.2	Jailbreaking	278
8.4.3	Prompt Injection	280
8.5	KI-Bedrohungsmodellierung	283
8.6	Verordnungen	285
8.7	Wie geht es jetzt weiter?	287
8.8	Zusammenfassung	288
8.9	Kurz und bündig	289
9	Mathematik.....	291
	<i>Annalisa Cadonna</i>	
9.1	Lineare Algebra	292
9.1.1	Vektoren und Matrizen	293
9.1.2	Operationen zwischen Vektoren und Matrizen	297
9.1.3	Lineare Transformationen	299
9.1.4	Eigenwerte, Eigenvektoren und Eigendekomposition	300
9.1.5	Andere Matrixzerlegungen	302
9.2	Kalkulus und Optimierung	304
9.2.1	Ableitung	305
9.2.2	Gradient und Hessian	307
9.2.3	Gradientenabstieg	309
9.2.4	Eingeschränkte Optimierung	311
9.3	Wahrscheinlichkeitsrechnung	312
9.3.1	Diskrete und kontinuierliche Zufallsvariablen	313
9.3.2	Erwartungswert, Varianz und Kovarianz	317
9.3.3	Unabhängigkeit, bedingte Verteilungen und Bayes-Theorem ...	319
9.4	Kurz und bündig	321
10	Statistik – Grundlagen.....	323
	<i>Rania Wazir, Georg Langs, Annalisa Cadonna</i>	
10.1	Daten	324
10.2	Einfache lineare Regression	326

10.3	Multiple lineare Regression	334
10.4	Logistische Regression	337
10.5	Wie gut ist unser Modell?	345
10.6	Kurz und bündig	347
11	Business Intelligence (BI)	349
	<i>Christian Mata</i>	
11.1	Einführung in Business Intelligence	352
11.1.1	Definition von Business Intelligence	352
11.1.2	Rolle in Organisationen	352
11.1.3	Entwicklung von Business Intelligence	353
11.1.4	Data Science und KI im Kontext von BI	356
11.1.5	Daten für die Entscheidungsfindung	359
11.1.6	Verstehen des geschäftlichen Kontextes	361
11.1.7	Business-Intelligence-Aktivitäten	363
11.2	Grundlagen des Datenmanagements	366
11.2.1	Was sind Datenmanagement, Datenintegration und Data Warehousing?	366
11.2.2	Datenbeladung – Der Fall von ETL oder ELT	368
11.2.3	Datenmodellierung	370
11.3	Reporting und Datenanalyse	378
11.3.1	Reporting	378
11.3.2	Berichtsarten	382
11.3.3	Datenanalyse	383
11.3.4	Visuelle Datenanalyse	385
11.3.5	Trends in Reporting und Datenanalyse	387
11.4	BI-Technologien und Werkzeuge	388
11.4.1	Relevante BI-Technologien	389
11.4.2	Verbreitete BI-Werkzeuge (BI-Tools)	394
11.5	BI und Data Science: Ergänzende Disziplinen	397
11.5.1	Unterschiede zwischen BI und DS	398
11.5.2	Gemeinsamkeiten von BI und DS	398
11.5.3	Synergien bei BI und DS	399
11.6	Ausblick für Business Intelligence	400
11.6.1	Erwartungen an die Entwicklung von BI	401
11.7	Kurz und bündig	403

12	Maschinelles Lernen	405
	<i>Georg Langs, Katherine Munro, Rania Wazir</i>	
12.1	Einführung	405
12.2	Grundlegendes: Feature Spaces	407
12.3	Klassifizierungsmodelle	411
12.3.1	K-Nearest-Neighbor-Klassifikator	411
12.3.2	Support Vector Machine	412
12.3.3	Entscheidungsbäume	413
12.4	Ensemble-Methoden	415
12.4.1	Bias und Varianz	416
12.4.2	Bagging: Random Forests	418
12.4.3	Boosten: AdaBoost	422
12.4.4	Die Grenzen der Merkmalskonstruktion und -auswahl	422
12.5	Unüberwachtes Lernen: Lernen ohne Etiketten	423
12.5.1	Clustering	423
12.5.2	Manifold Learning	424
12.5.3	Generative Modelle	425
12.6	Künstliche neuronale Netze und Deep Learning	427
12.6.1	Das Perzeptron	427
12.6.2	Künstliche neuronale Netze	428
12.6.3	Deep Learning	430
12.6.4	Convolutional Neural Networks	430
12.6.5	Training von Convolutional Neural Networks	432
12.6.6	Rekurrente neuronale Netze	434
12.6.7	Long Short-Term Memory Networks	436
12.6.8	Autoencoder und U-Nets	437
12.6.9	Adversariales Training	439
12.6.10	Generative Adversarial Networks	440
12.6.11	Cycle-GANs und Style-GANs	443
12.7	Transformer-Modelle und Aufmerksamkeitsmechanismen	444
12.7.1	Die Transformer-Architektur	444
12.7.2	Was der Aufmerksamkeitsmechanismus leistet	446
12.7.3	Anwendungen von Transformer-Modellen	447
12.8	Reinforcement Learning	448
12.9	Andere Architekturen und Lernstrategien	452

12.10	Validierungsstrategien für Techniken des maschinellen Lernens	452
12.11	Schlussfolgerung	454
12.12	Kurz und bündig	455
13	Großartige künstliche Intelligenz erschaffen	457
	<i>Danko Nikolić</i>	
13.1	Wie KI mit Data Science und maschinellern Lernen zusammenhängt	458
13.2	Eine kurze Geschichte der KI	462
13.3	Fünf Empfehlungen für die Entwicklung einer KI-Lösung	464
13.3.1	Empfehlung Nummer eins: Seien Sie pragmatisch	465
13.3.2	Empfehlung Nummer zwei: Erleichtern Sie Maschinen das Lernen – schaffen Sie induktive Verzerrungen	467
13.3.3	Empfehlung Nummer drei: Analysen durchführen	474
13.3.4	Empfehlung Nummer vier: Hüten Sie sich vor der Skalierungsfalle	477
13.3.5	Empfehlung Nummer fünf: Hüten Sie sich vor der Verallgemeine- rungsfalle (so etwas wie ein kostenloses Mittagessen gibt es nicht)	488
13.4	Intelligenz auf menschlicher Ebene	493
13.5	Kurz und bündig	496
14	Signalverarbeitung	499
	<i>Jana Eder</i>	
14.1	Einführung	500
14.2	Abtastung und Quantisierung	501
14.3	Frequenzbereichsanalyse	504
14.3.1	Fourier-Transformation	505
14.4	Rauschunterdrückung und Filtertechniken	512
14.4.1	Rauschunterdrückung mit einem Gaußschen Low-pass-Filter ..	513
14.5	Analyse des Zeitbereichs	515
14.5.1	Signalnormierung und Standardisierung	515
14.5.2	Signaltransformation und Merkmalsextraktion	516
14.5.3	Techniken zur Zerlegung von Zeitreihen	519
14.5.4	Autokorrelation: Verstehen der Signalähnlichkeit über die Zeit ..	522
14.6	Analyse des Zeit-/Frequenzbereichs	524
14.6.1	Kurzzeit-Fourier-Transformation und Spektrogramm	525
14.6.2	Diskrete Wavelet-Transformation	525
14.6.3	Gramian Angular Field	526

- 14.7 Die Beziehung zwischen Signalverarbeitung und maschinellem Lernen .. 529
 - 14.7.1 Techniken für das Feature Engineering 529
 - 14.7.2 Vorbereitung auf maschinelles Lernen 530
- 14.8 Praktische Anwendungen 531
- 14.9 Kurz und bündig 533
- 15 Basismodelle 535**
 - Danko Nikolić*
 - 15.1 Die Idee eines Basismodells 535
 - 15.2 Wie trainiert man ein Basismodell? 539
 - 15.3 Wie verwenden wir Basismodelle? 541
 - 15.4 Ein Durchbruch: Das Lernen hat kein Ende 550
 - 15.5 Kurz und bündig 551
- 16 Generative KI und große Sprachmodelle 553**
 - Katherine Munro, Gerald A. Hahn, Danko Nikolić*
 - 16.1 Einführung in die „Gen-KI“ 553
 - 16.2 Generative KI-Modalitäten 555
 - 16.2.1 Methoden für das Training generativer Modelle 556
 - 16.3 Große Sprachmodelle 557
 - 16.3.1 Was sind LLMs? 557
 - 16.3.2 Wie wird so etwas wie ChatGPT trainiert? 559
 - 16.3.3 Methoden zur direkten Verwendung von LLMs 561
 - 16.3.4 Methoden zur Anpassung von LLMs 574
 - 16.4 Schwachstellen und Grenzen von Gen-KI-Modellen 584
 - 16.4.1 Einführung 584
 - 16.4.2 Prompt Injection und Jailbreaking-Angriffe 585
 - 16.4.3 Halluzinationen, Konfabulationen und Begründungsirrtümer .. 589
 - 16.4.4 Urheberrechtliche Bedenken 590
 - 16.4.5 Bias 595
 - 16.5 Erstellen robuster, effektiver Gen-KI-Anwendungen 598
 - 16.5.1 Kontrollstrategien während der gesamten Entwicklung und Nutzung 598
 - 16.5.2 Guardrails 600
 - 16.5.3 Sicherer und erfolgreicher Einsatz generativer KI 601
 - 16.6 Kurz und bündig 603

17	KI-Agenten	609
	<i>Stefan Papp, Zoltan Toth</i>	
17.1	Argumente für KI-Agenten	609
17.2	Agentische Entwurfsmuster	611
17.2.1	Reflexion	612
17.2.2	Tool-Nutzung	613
17.2.3	Planung	615
17.2.4	Multi-Agent-Muster	616
17.2.5	Skalierung der Komplexität	618
17.3	Implementierung von Agenten	619
17.3.1	Frameworks und Plattformen	619
17.3.2	Erweiterte Datengenerierung	622
17.3.3	Perspektiven	626
17.4	Fallstudie	627
17.4.1	Ein einfacher Agent mit OpenAI	627
17.4.2	Tool-Aufruf	628
17.4.3	MCP in OpenAI-Agenten	628
17.4.4	Mehrere Agenten	628
17.5	Die Zukunft der Agenten	629
17.6	Zusammenfassung	630
18	Vibe Coding	633
	<i>Stefan Papp</i>	
18.1	Was ist Vibe Coding?	633
18.2	Anwendungsfälle für Vibe Coding	636
18.3	Werkzeuge	637
18.3.1	Cursor	639
18.3.2	Claude Code	642
18.3.3	Bolt	645
18.3.4	Was soll ich verwenden?	646
18.4	Vibe-Coding-Beispiel	647
18.4.1	Lösung mit Cursor	648
18.4.2	Google Colab	654
18.5	Konsequenzen	657
18.6	Zusammenfassung	659