

Katherine MUNRO
Stefan PAPP
Zoltan TOTH
Wolfgang WEIDINGER
Danko NIKOLIĆ

3. Auflage

HANDBUCH DATA SCIENCE UND KI

Mit Machine Learning
und Datenanalyse
Wert aus Daten generieren

HANSER

Munro / Papp / Toth / Weidinger / Nikolić
Antosova Vesela / Bruckmüller / Cadonna / Eder / Gorzala / Hahn / Langs /
Licandro / Mata / McIntyre / Meir-Huber / Móra / Pasieska / Rugli / Wazir / Zauner

Handbuch Data Science und KI



bleiben Sie auf dem Laufenden!

Der Hanser Computerbuch-Newsletter informiert Sie regelmäßig über neue Bücher und Termine aus den verschiedenen Bereichen der IT. Profitieren Sie auch von Gewinnspielen und exklusiven Leseproben. Gleich anmelden unter

www.hanser-fachbuch.de/newsletter

Katherine Munro, Stefan Papp, Zoltan C. Toth,
Wolfgang Weidinger, Danko Nikolić,
Barbora Antosova Vesela, Karin Bruckmüller,
Annalisa Cadonna, Jana Eder, Jeannette Gorzala,
Gerald A. Hahn, Georg Langs, Roxane Licandro,
Christian Mata, Sean McIntyre, Mario Meir-Huber,
György Móra, Manuel Paseska, Victoria Rugli,
Rania Wazir, Günther Zauner

Handbuch Data Science und KI

Mit Machine Learning und Datenanalyse Wert aus
Daten generieren

3., aktualisierte und erweiterte Auflage

HANSER



Print-ISBN: 978-3-446-47937-1

E-Book-ISBN: 978-3-446-48072-8

Epub-ISBN: 978-3-446-48357-6

Alle in diesem Werk enthaltenen Informationen, Verfahren und Darstellungen wurden zum Zeitpunkt der Veröffentlichung nach bestem Wissen zusammengestellt. Dennoch sind Fehler nicht ganz auszuschließen. Aus diesem Grund sind die im vorliegenden Werk enthaltenen Informationen für Autor:innen, Herausgeber:innen und Verlag mit keiner Verpflichtung oder Garantie irgendeiner Art verbunden. Autor:innen, Herausgeber:innen und Verlag übernehmen infolgedessen keine Verantwortung und werden keine daraus folgende oder sonstige Haftung übernehmen, die auf irgendeine Weise aus der Benutzung dieser Informationen – oder Teilen davon – entsteht. Ebenso wenig übernehmen Autor:innen, Herausgeber:innen und Verlag die Gewähr dafür, dass die beschriebenen Verfahren usw. frei von Schutzrechten Dritter sind. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt also auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

Die endgültige Entscheidung über die Eignung der Informationen für die vorgesehene Verwendung in einer bestimmten Anwendung liegt in der alleinigen Verantwortung des Nutzers.

Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter <http://dnb.d-nb.de> abrufbar.

Dieses Werk ist urheberrechtlich geschützt.

Alle Rechte, auch die der Übersetzung, des Nachdruckes und der Vervielfältigung des Werkes, oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Einwilligung des Verlages in irgendeiner Form (Fotokopie, Mikrofilm oder einem anderen Verfahren), auch nicht für Zwecke der Unterrichtsgestaltung – mit Ausnahme der in den §§ 53, 54 UrhG genannten Sonderfälle –, reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Wir behalten uns auch eine Nutzung des Werks für Zwecke des Text- und Data Mining nach § 44b UrhG ausdrücklich vor.

© 2025 Carl Hanser Verlag GmbH & Co. KG, München
Kolbergerstraße 22 | 81679 München | info@hanser.de
www.hanser-fachbuch.de

Lektorat: Sylvia Hasselbach

Copy editing: Jürgen Dubau, Freiburg/Elbe

Coverkonzept: Marc Müller-Bremer, www.rebranding.de, München

Covergestaltung: Tom West

Titelmotiv: [gettyimages/ValeryBrozhinsky](https://www.gettyimages.com)

Herstellung: le-tex publishing services GmbH, Leipzig

Satz: Eberl & Koesel Studio, Kempten

Druck: CPI books GmbH, Leck

Printed in Germany

Inhalt

Vorwort	XXIII
Danksagungen	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	8
1.2.5 IT	9
1.2.6 Sicherheit	11
1.2.7 Leiter der Produktion	11
1.2.8 Kundenbetreuung	13
1.2.9 HR	14
1.2.10 CEO	15
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	32
2.4.1	Design Thinking	32
2.4.2	Double Diamond	33
2.4.3	Durchführung von Workshops	34
3	Cloud-Dienste	44
	<i>Stefan Papp</i>	
3.1	Einführung	45
3.2	Cloud-Essentials	45
3.2.1	XaaS	48
3.2.2	Cloud-Anbieter	49
3.2.3	Native Cloud-Dienste	51
3.2.4	Cloud-native Paradigmen	54
3.3	Infrastructure as a Service	56
3.3.1	Hardware	57
3.3.2	Verteilte Systeme	59
3.3.3	Linux Essentials für Datenexperten	63
3.3.4	Infrastructure as Code	70
3.4	Platform as a Service	75
3.4.1	Cloud Native PaaS-Lösungen	76
3.4.2	Externe Lösungen	81
3.5	Software as a Service	84
3.6	Kurz und bündig	85
4	Datenarchitektur	87
	<i>Zoltan C. Toth und Sean McIntyre</i>	
4.1	Übersicht	87
4.1.1	Maslowsche Bedürfnishierarchie für Daten	88
4.1.2	Anforderungen an die Datenarchitektur	89
4.1.3	Die Struktur einer typischen Datenarchitektur	90
4.1.4	ETL (Extrahieren, Transformieren, Laden)	95
4.1.5	ELT (Extrahieren, Laden, Transformieren)	96
4.1.6	ETLT	97
4.2	Datenerfassung und -integration	97

4.2.1	Datenquellen	98
4.2.2	Traditionelle Dateiformate	100
4.2.3	Moderne Dateiformate	102
4.2.4	Welche Speicheroption soll ich wählen?	105
4.3	Data Warehouses, Data Lakes und Lakehouses	105
4.3.1	Data Warehouses	106
4.3.2	Data Lakes und Cloud-Datenplattformen	110
4.4	Data Transformation	114
4.4.1	SQL	116
4.4.2	Big Data & Apache Spark	127
4.4.3	Cloud-Datenplattformen für Apache Spark	136
4.5	Workflow-Orchestrierung	138
4.5.1	Dagster und der Modern Data Stack	141
4.6	Ein Anwendungsfall und seine Datenarchitektur	142
4.7	Kurz und bündig	148
5	Data Engineering	150
	<i>Stefan Papp</i>	
5.1	Abgrenzung zum Software Engineering	151
5.2	Programmiersprachen	153
5.2.1	Code oder kein Code?	153
5.2.2	Auswahl der Programmiersprache	155
5.2.3	Python	156
5.2.4	Scala	160
5.3	Software-Engineering-Prozesse für Daten	162
5.3.1	Konfigurationsmanagement	163
5.3.2	CI/CD	164
5.4	Datenpipelines	166
5.4.1	Gemeinsame Merkmale einer Datenpipeline	167
5.4.2	Datenpipelines in der Unified Data Architecture	167
5.5	Speicheroptionen	172
5.5.1	Datei-Ära	172
5.5.2	Datenbank-Ära	173
5.5.3	Data-Lake-Ära	175
5.5.4	Serverless-Ära	176

5.5.5	Polyglotte Speicherung	177
5.5.6	Data-Mesh-Ära	178
5.6	Tooling	180
5.6.1	Batch: Airflow	180
5.6.2	Streaming: Kafka	182
5.6.3	Transformation: Databricks Notebooks	187
5.7	Gemeinsame Herausforderungen	189
5.7.1	Datenqualität und unterschiedliche Standards	189
5.7.2	Skewed Data	191
5.7.3	Überlastete operationelle Systeme	192
5.7.4	Operative Legacy-Systeme	193
5.7.5	Plattform- und Informationssicherheit	193
5.8	Kurz und bündig	194
6	Data Governance	195
	<i>Victoria Rugli, Mario Meir-Huber</i>	
6.1	Warum brauchen wir Data Governance?	195
6.1.1	Beispiel 1: Mit Data Governance Klarheit schaffen	197
6.1.2	Beispiel 2: Die (negativen) Auswirkungen einer mangelhaften Data Governance	198
6.2	Die Bausteine der Data Governance	199
6.2.1	Data Governance erklärt	200
6.3	Menschen	202
6.3.1	Data Ownership	203
6.3.2	Data Stewards	206
6.3.3	Data Governance Board	208
6.3.4	Change Management	210
6.4	Prozess	212
6.4.1	Verwaltung von Metadaten	213
6.4.2	Management der Datenqualität	217
6.4.3	Datensicherheit und Datenschutz	221
6.4.4	Stammdatenmanagement	225
6.4.5	Datenzugang und Suche	229
6.5	Technologie (Datenmanagement-Tools)	232
6.5.1	Open-Source-Tools	232

6.5.2	Cloud-basierte Data-Governance-Tools	239
6.6	Kurz und bündig	244
7	Machine Learning Operations (ML Ops)	245
	<i>Zoltan C. Toth, György Móra</i>	
7.1	Übersicht	245
7.1.1	Umfang von MLOps	246
7.1.2	Datenerhebung und Exploration	247
7.1.3	Feature Engineering	247
7.1.4	Modelltraining	248
7.1.5	In der Produktion eingesetzte Modelle	249
7.1.6	Bewertung des Modells	249
7.1.7	Model Understanding	250
7.1.8	Modellversionierung	250
7.1.9	Modellüberwachung	251
7.2	MLOps in einer Organisation	251
7.2.1	Die wichtigsten Vorteile von MLOps	252
7.2.2	Erforderliche Fähigkeiten für MLOps	252
7.3	Verschiedene gängige Szenarien im MLOps-Bereich	253
7.3.1	Integration von Notebooks	253
7.3.2	Features in der Produktion	255
7.3.3	Modelleinsatz	257
7.3.4	Modellformate	258
7.4	MLOps-Werkzeuge und MLflow	259
7.4.1	MLflow	260
7.5	Kurz und bündig	263
8	Cybersicherheit von Machine-Learning-Systemen	264
	<i>Manuel Pasiëka</i>	
8.1	Einführung in die Cybersicherheit	265
8.2	Angriffsfläche	267
8.3	Angriffsmethoden	268
8.3.1	Model Stealing	268
8.3.2	Datenextraktion	271
8.3.3	Data Poisoning	273
8.3.4	Adversariale Angriffe	276

8.3.5	Backdoor-Angriffe	279
8.4	Sicherheit von großen Sprachmodellen durch maschinelles Lernen	282
8.4.1	Datenextraktion	283
8.4.2	Jailbreaking	284
8.4.3	Prompt Injection	286
8.5	KI-Bedrohungsmodellierung	289
8.6	Verordnungen	291
8.7	Wie geht es jetzt weiter?	293
8.8	Zusammenfassung	295
8.9	Kurz und bündig	296
9	Mathematik	297
	<i>Annalisa Cadonna</i>	
9.1	Lineare Algebra	298
9.1.1	Vektoren und Matrizen	299
9.1.2	Operationen zwischen Vektoren und Matrizen	303
9.1.3	Lineare Transformationen	305
9.1.4	Eigenwerte, Eigenvektoren und Eigendekomposition	306
9.1.5	Andere Matrixzerlegungen	308
9.2	Kalkulus und Optimierung	310
9.2.1	Ableitung	311
9.2.2	Gradient und Hessian	313
9.2.3	Gradientenabstieg	315
9.2.4	Eingeschränkte Optimierung	317
9.3	Wahrscheinlichkeitsrechnung	318
9.3.1	Diskrete und kontinuierliche Zufallsvariablen	319
9.3.2	Erwartungswert, Varianz und Kovarianz	323
9.3.3	Unabhängigkeit, bedingte Verteilungen und Bayes-Theorem	325
9.4	Kurz und bündig	327
10	Statistik – Grundlagen	329
	<i>Rania Wazir, Georg Langs, Annalisa Cadonna</i>	
10.1	Daten	331
10.2	Einfache lineare Regression	332
10.3	Multiple lineare Regression	341
10.4	Logistische Regression	343

10.5	Wie gut ist unser Modell?	352
10.6	Kurz und bündig	353
11	Business Intelligence (BI)	355
	<i>Christian Mata</i>	
11.1	Einführung in Business Intelligence	358
11.1.1	Definition von Business Intelligence	358
11.1.2	Rolle in Organisationen	359
11.1.3	Entwicklung von Business Intelligence	360
11.1.4	Data Science und KI im Kontext von BI	362
11.1.5	Daten für die Entscheidungsfindung	366
11.1.6	Verstehen des geschäftlichen Kontextes	368
11.1.7	Business-Intelligence-Aktivitäten	371
11.2	Grundlagen des Datenmanagements	373
11.2.1	Was sind Datenmanagement, Datenintegration und Data Warehousing?	374
11.2.2	Datenbeladung – Der Fall von ETL oder ELT	375
11.2.3	Datenmodellierung	377
11.3	Reporting und Datenanalyse	385
11.3.1	Reporting	385
11.3.2	Berichtsarten	389
11.3.3	Datenanalyse	390
11.3.4	Visuelle Datenanalyse	392
11.3.5	Trends in Reporting und Datenanalyse	394
11.4	BI-Technologien und Werkzeuge	396
11.4.1	Relevante BI-Technologien	396
11.4.2	Verbreitete BI-Werkzeuge (BI-Tools)	401
11.5	BI und Data Science: Ergänzende Disziplinen	405
11.5.1	Unterschiede zwischen BI und DS	405
11.5.2	Gemeinsamkeiten von BI und DS	406
11.5.3	Synergien bei BI und DS	406
11.6	Ausblick für Business Intelligence	408
11.6.1	Erwartungen an die Entwicklung von BI	409
11.7	Kurz und bündig	411

12	Maschinelles Lernen	413
	<i>Georg Langs, Katherine Munro, Rania Wazir</i>	
12.1	Einführung	413
12.2	Grundlegendes: Feature Spaces	415
12.3	Klassifizierungsmodelle	419
12.3.1	K-Nearest-Neighbor-Klassifikator	419
12.3.2	Support Vector Machine	420
12.3.3	Entscheidungsbäume	421
12.4	Ensemble-Methoden	423
12.4.1	Bias und Varianz	424
12.4.2	Bagging: Random Forests	426
12.4.3	Boosten: AdaBoost	430
12.4.4	Die Grenzen der Merkmalskonstruktion und -auswahl	431
12.5	Unüberwachtes Lernen: Lernen ohne Etiketten	432
12.5.1	Clustering	432
12.5.2	Manifold Learning	433
12.5.3	Generative Modelle	434
12.6	Künstliche neuronale Netze und Deep Learning	436
12.6.1	Das Perzeptron	436
12.6.2	Künstliche neuronale Netze	437
12.6.3	Deep Learning	439
12.6.4	Convolutional Neural Networks	440
12.6.5	Training von Convolutional Neural Networks	441
12.6.6	Rekurrente neuronale Netze	444
12.6.7	Long Short-Term Memory Networks	446
12.6.8	Autoencoder und U-Nets	447
12.6.9	Adversariales Training	449
12.6.10	Generative Adversarial Networks	451
12.6.11	Cycle-GANs und Style-GANs	453
12.7	Transformer-Modelle und Aufmerksamkeitsmechanismen	455
12.7.1	Die Transformer-Architektur	455
12.7.2	Was der Aufmerksamkeitsmechanismus leistet	457
12.7.3	Anwendungen von Transformer-Modellen	458
12.8	Reinforcement Learning	459

12.9	Andere Architekturen und Lernstrategien	463
12.10	Validierungsstrategien für Techniken des maschinellen Lernens	463
12.11	Schlussfolgerung	465
12.12	Kurz und bündig	466
13	Großartige künstliche Intelligenz erschaffen	467
	<i>Danko Nikolić</i>	
13.1	Wie KI mit Data Science und maschinellem Lernen zusammenhängt	468
13.2	Eine kurze Geschichte der KI	472
13.3	Fünf Empfehlungen für die Entwicklung einer KI-Lösung	475
13.3.1	Empfehlung Nummer eins: Seien Sie pragmatisch	475
13.3.2	Empfehlung Nummer zwei: Erleichtern Sie Maschinen das Lernen – schaffen Sie induktive Verzerrungen	478
13.3.3	Empfehlung Nummer drei: Analysen durchführen	484
13.3.4	Empfehlung Nummer vier: Hüten Sie sich vor der Skalierungsfalle	487
13.3.5	Empfehlung Nummer fünf: Hüten Sie sich vor der Verallgemeinerungsfalle (so etwas wie ein kostenloses Mittagessen gibt es nicht)	499
13.4	Intelligenz auf menschlicher Ebene	505
13.5	Kurz und bündig	508
14	Signalverarbeitung	510
	<i>Jana Eder</i>	
14.1	Einführung	511
14.2	Abtastung und Quantisierung	512
14.3	Frequenzbereichsanalyse	515
14.3.1	Fourier-Transformation	517
14.4	Rauschunterdrückung und Filtertechniken	523
14.4.1	Rauschunterdrückung mit einem Gaußschen Low-pass-Filter ...	525
14.5	Analyse des Zeitbereichs	527
14.5.1	Signalnormierung und Standardisierung	527
14.5.2	Signaltransformation und Merkmalsextraktion	528
14.5.3	Techniken zur Zerlegung von Zeitreihen	531
14.5.4	Autokorrelation: Verstehen der Signalähnlichkeit über die Zeit ..	534
14.6	Analyse des Zeit-/Frequenzbereichs	537

14.6.1	Kurzzeit-Fourier-Transformation und Spektrogramm	537
14.6.2	Diskrete Wavelet-Transformation	538
14.6.3	Gramian Angular Field	539
14.7	Die Beziehung zwischen Signalverarbeitung und maschinellem Lernen	542
14.7.1	Techniken für das Feature Engineering	542
14.7.2	Vorbereitung auf maschinelles Lernen	543
14.8	Praktische Anwendungen	544
14.9	Kurz und bündig	546
15	Basismodelle	547
	<i>Danko Nikolić</i>	
15.1	Die Idee eines Basismodells	547
15.2	Wie trainiert man ein Basismodell?	551
15.3	Wie verwenden wir Basismodelle?	553
15.4	Ein Durchbruch: Das Lernen hat kein Ende	562
15.5	Kurz und bündig	564
16	Generative KI und große Sprachmodelle	566
	<i>Katherine Munro, Gerald A. Hahn, Danko Nikolić</i>	
16.1	Einführung in die „Gen-KI“	566
16.2	Generative KI-Modalitäten	568
16.2.1	Methoden für das Training generativer Modelle	569
16.3	Große Sprachmodelle	570
16.3.1	Was sind LLMs?	570
16.3.2	Wie wird so etwas wie ChatGPT trainiert?	572
16.3.3	Methoden zur direkten Verwendung von LLMs	574
16.3.4	Methoden zur Anpassung von LLMs	587
16.4	Schwachstellen und Grenzen von Gen-KI-Modellen	598
16.4.1	Einführung	598
16.4.2	Prompt Injection und Jailbreaking-Angriffe	598
16.4.3	Halluzinationen, Konfabulationen und Begründungsirrtümer ..	603
16.4.4	Urheberrechtliche Bedenken	605
16.4.5	Bias	609
16.5	Erstellen robuster, effektiver Gen-KI-Anwendungen	612
16.5.1	Kontrollstrategien während der gesamten Entwicklung und Nutzung	612

16.5.2	Guardrails	615
16.5.3	Sicherer und erfolgreicher Einsatz generativer KI	615
16.6	Kurz und bündig	618
17	Natürliche Sprachverarbeitung (NLP)	622
	<i>Katherine Munro</i>	
17.1	Was ist NLP und warum ist es so wertvoll?	623
17.2	Warum „traditionelles“ NLP im „Zeitalter der großen Sprachmodelle“ lernen?	624
17.3	NLP-Datenaufbereitungstechniken	626
17.3.1	Die NLP-Pipeline	626
17.3.2	Konvertierung des Eingabeformats für Machine Learning	634
17.4	NLP-Aufgaben und -Methoden	636
17.4.1	Regelbasiertes (symbolisches) NLP	637
17.4.2	Machine-Learning-Ansätze	641
17.4.3	Neurales NLP	651
17.4.4	Transfer Learning	659
17.5	Das Wichtigste in Kürze	673
18	Computer Vision	676
	<i>Roxane Licandro</i>	
18.1	Was ist Computer Vision?	676
18.2	Ein Bild sagt mehr als tausend Worte	678
18.2.1	Das menschliche Auge	679
18.2.2	Das Bildaufnahmeprinzip	681
18.2.3	Digitale Dateiformate	687
18.2.4	Bildkomprimierung	688
18.3	Ich sehe was, was du nicht siehst	690
18.3.1	Computergestützte Fotografie und Bildmanipulation	693
18.4	Computer-Vision-Anwendungen und zukünftige Richtungen	697
18.4.1	Image-Retrieval-Systeme	698
18.4.2	Objekterkennung, Klassifizierung und Verfolgung	701
18.4.3	Medizinische Computer Vision	702
18.5	Menschen sehen lassen	707
18.6	Kurz und bündig	709

19	Modellierung und Simulation – Erstellen Sie Ihre eigenen Modelle	711
	<i>Günther Zauner, Wolfgang Weidinger</i>	
19.1	Einführung	712
19.2	Allgemeine Aspekte	713
19.3	Modellierung zur Beantwortung von Fragen	714
19.4	Reproduzierbarkeit und Lebenszyklus des Modells	716
19.4.1	Der Lebenszyklus einer Modellierungs- und Simulationsfrage ..	718
19.4.2	Parameter- und Output-Definition	720
19.4.3	Dokumentation	723
19.4.4	Verifizierung und Validierung	724
19.5	Methoden	729
19.5.1	Gewöhnliche Differentialgleichungen (ODEs)	729
19.5.2	Systemdynamik (SD)	731
19.5.3	Diskrete Ereignissimulation	734
19.5.4	Agentenbasierte Modellierung	738
19.6	Beispiele für Modellierung und Simulation	741
19.6.1	Dynamische Modellierung von Eisenbahnnetzen zur optimalen Wegfindung mit agentenbasierten Methoden und Reinforcement Learning	742
19.6.2	Strategien zur agentenbasierten Covid-Modellierung	744
19.6.3	Deep-Reinforcement-Learning-Ansatz für eine optimale Nachschubpolitik in einer VMI-Umgebung	751
19.6.4	Valide Lösungen für ein ressourcenbeschränktes Projektplanungsproblem mithilfe von bestärkendem Lernen und Bewertung der Robustheit mit diskreter Ereignissimulation	754
19.6.5	Zusammenfassung und Lessons Learned	759
19.7	Kurz und bündig	760
20	Visualisierung von Daten	764
	<i>Barbora Antosova Vesela</i>	
20.1	Geschichte	766
20.2	Welche Tools Sie verwenden sollten	772
20.3	Arten von Datenvisualisierungen	775
20.3.1	Streudiagramm	775
20.3.2	Liniendiagramm	776

20.3.3	Säulen- und Balkendiagramme	777
20.3.4	Histogramm	778
20.3.5	Tortendiagramm	779
20.3.6	Box Plot	780
20.3.7	Heat Map	780
20.3.8	Baumdiagramm	781
20.3.9	Andere Arten von Visualisierungen	782
20.4	Wählen Sie die richtige Datenvisualisierung	783
20.5	Tipps und Tricks	786
20.6	Präsentation der Datenvisualisierung	791
20.7	Kurz und bündig	791
21	Datengetriebene Unternehmen	794
	<i>Mario Meir-Huber, Stefan Papp</i>	
21.1	Die drei Ebenen eines datengesteuerten Unternehmens	795
21.2	Kultur	795
21.2.1	Unternehmensstrategie für Daten	796
21.2.2	Die Analyse des aktuellen Stands	799
21.2.3	Unternehmenskultur und Organisation einer erfolgreichen Datenorganisation	801
21.2.4	Kernproblem: der Fachkräftemangel	810
21.3	Technologie	812
21.3.1	Die Auswirkungen von Open Source	813
21.3.2	Cloud	813
21.3.3	Auswahl des Anbieters	814
21.3.4	Data Lake aus der Unternehmensperspektive	814
21.3.5	Die Rolle der IT	815
21.3.6	Data Science Labs	816
21.3.7	Revolution in der Architektur: das Data Mesh	817
21.4	Business	818
21.4.1	Daten kaufen und teilen	819
21.4.2	Implementierung des analytischen Anwendungsfalls	820
21.4.3	Self-Service Analytics	821
21.5	Kurz und bündig	821

22	Leistungsstarke Teams schaffen	822
	<i>Stefan Papp</i>	
22.1	Neue Teams, neues Glück	823
22.2	Storming	823
	22.2.1 Szenario: 50 Shades of Red	823
	22.2.2 Szenario: Retrospektive	828
22.3	Norming	831
	22.3.1 Change Management und Transition	831
	22.3.2 RACI-Matrix	834
	22.3.3 SMART	836
	22.3.4 Agile Prozesse	838
	22.3.5 Kommunikationskultur	840
	22.3.6 DataOps	842
22.4	Forming	847
	22.4.1 Szenario: Eine neue Morgendämmerung	847
	22.4.2 Wachstumsgedanken	849
22.5	Kurz und bündig	852
23	Gesetz über künstliche Intelligenz	854
	<i>Jeannette Gorzala, Karin Bruckmüller</i>	
23.1	Einführung	855
23.2	Definition von KI-Systemen	857
23.3	Anwendungsbereich und Zweck des KI-Gesetzes	859
	23.3.1 Der risikobasierte Ansatz	860
	23.3.2 Unannehmbare Risiken und verbotene KI-Praktiken	862
	23.3.3 Hochriskante KI-Systeme und Compliance	865
	23.3.4 Mittleres Risiko und Transparenzverpflichtungen	867
	23.3.5 Geringes Risiko und freiwillige Selbstverpflichtungen	868
23.4	KI-Modelle mit allgemeinem Verwendungszweck	869
23.5	Zeitplan und Anwendbarkeit	872
23.6	Sanktionen	872
23.7	KI und zivilrechtliche Haftung	873
23.8	KI und strafrechtliche Haftung	873
23.9	Kurz und bündig	877

24	AI in verschiedenen Branchen	878
	<i>Stefan Papp, Mario Meir-Huber, Wolfgang Weidinger, Thomas Tremel, Marek Danis</i>	
24.1	Automobilindustrie	883
24.1.1	Vision	884
24.1.2	Daten	885
24.1.3	Anwendungsfälle	885
24.1.4	Herausforderungen	886
24.2	Luftfahrt	888
24.2.1	Vision	889
24.2.2	Daten	889
24.2.3	Anwendungsfälle	890
24.2.4	Herausforderungen	891
24.3	Energie	891
24.3.1	Vision	892
24.3.2	Daten	893
24.3.3	Anwendungsfälle	893
24.3.4	Herausforderungen	894
24.4	Finanzen	895
24.4.1	Vision	895
24.4.2	Daten	896
24.4.3	Anwendungsfälle	896
24.4.4	Herausforderungen	898
24.5	Gesundheit	899
24.5.1	Vision	899
24.5.2	Daten	900
24.5.3	Anwendungsfälle	901
24.5.4	Herausforderungen	901
24.6	Regierung	902
24.6.1	Vision	903
24.6.2	Daten	903
24.6.3	Anwendungsfälle	904
24.6.4	Herausforderungen	907
24.7	Kunst	908
24.7.1	Vision	909

24.7.2	Daten	909
24.7.3	Anwendungsfälle	910
24.7.4	Herausforderungen	910
24.8	Produktion	911
24.8.1	Vision	911
24.8.2	Daten	912
24.8.3	Anwendungsfälle	912
24.8.4	Herausforderungen	913
24.9	Öl und Gas	914
24.9.1	Vision	914
24.9.2	Daten	915
24.9.3	Anwendungsfälle	916
24.9.4	Herausforderungen	917
24.10	Einzelhandel	918
24.10.1	Vision	918
24.10.2	Daten	919
24.10.3	Anwendungsfälle	919
24.10.4	Herausforderungen	920
24.11	Anbieter von Telekommunikation	921
24.11.1	Vision	921
24.11.2	Daten	922
24.11.3	Anwendungsfälle	922
24.11.4	Herausforderungen	924
24.12	Transport	925
24.12.1	Vision	925
24.12.2	Daten	926
24.12.3	Anwendungsfälle	926
24.12.4	Herausforderungen	926
24.13	Lehre und Ausbildung	927
24.13.1	Vision	928
24.13.2	Daten	928
24.13.3	Anwendungsfälle	929
24.13.4	Herausforderungen	930
24.14	Die digitale Gesellschaft	930
24.15	Kurz und bündig	932

25	Klimawandel und KI	933
	<i>Stefan Papp</i>	
25.1	Einführung	933
25.2	KI – ein Klimaretter?	935
25.3	Messung und Verringerung von Emissionen	936
25.3.1	Basislinie	936
25.3.2	Datenanwendungsfälle	938
25.4	Sequestrierung	940
25.4.1	Biologische Sequestrierung	941
25.4.2	Geologische Sequestrierung	943
25.5	Vorbereiten auf die Auswirkungen	945
25.6	Geoengineering	946
25.7	Greenwashing	948
25.8	Ausblick	950
25.9	Kurz und bündig	952
26	Mindset und Community	953
	<i>Stefan Papp</i>	
26.1	Data Driven Mindset	954
26.2	Data-Science-Kultur	957
26.2.1	Start-up oder Beratungsunternehmen?	957
26.2.2	Labs statt Konzernpolitik	958
26.2.3	Keiretsu statt Einzelkämpfertum	958
26.2.4	Agile Softwareentwicklung	960
26.2.5	Firmen- und Arbeitskultur	961
26.3	Antipatterns	964
26.3.1	Abwertung von Fachwissen	964
26.3.2	Die IT wird es schon richten	966
26.3.3	Widerstand gegen Veränderungen	966
26.3.4	Besserwisser-Mentalität	967
26.3.5	Schwarzmalerei	968
26.3.6	Pfennigfuchseriei	969
26.3.7	Angstkultur	970
26.3.8	Kontrolle über die Ressourcen	970
26.3.9	Blindes Vertrauen in die Ressourcen	971

26.3.10	Das Schweizer Taschenmesser	972
26.3.11	Over-Engineering	973
26.4	Kurz und bündig	973
27	Vertrauenswürdige KI	974
	<i>Rania Wazir</i>	
27.1	Rechtlicher und Soft-Law-Rahmen	975
27.1.1	Normen	978
27.1.2	Verordnungen	979
27.2	KI-Stakeholder	981
27.3	Fairness in der KI	982
27.3.1	Bias	984
27.3.2	Fairness-Metriken	987
27.3.3	Unerwünschten Bias in KI-Systemen reduzieren	992
27.4	Transparenz von KI-Systemen	993
27.4.1	Dokumentieren der Daten	994
27.4.2	Dokumentieren des Modells	996
27.4.3	Explainability (Erklärbarkeit)	997
27.5	Schlussfolgerung	1000
27.6	Kurz und bündig	1000
28	Epilog	1001
	<i>Stefan Papp</i>	
28.1	Halford 2.0	1001
28.1.1	Umwelt, Soziales und Governance	1002
28.1.2	HR	1003
28.1.3	Kundenzufriedenheit	1005
28.1.4	Produktion	1007
28.1.5	IT	1008
28.1.6	Strategie	1010
28.2	Letzte Worte	1012
28.3	Kurz und bündig	1013
29	Die Autor:innen	1014
	Index	1023

Vorwort

Dieses Vorwort wurde NICHT von ChatGPT (oder Ähnlichem) geschrieben.

Während ich diese Aussage treffe, frage ich mich, wie oft sie in Zukunft für Texte oder andere Medienformen gelten wird. In den letzten zwei Jahren hat dieses KI-gestützte Tool enorme Popularität erlangt und Data Science und KI einen unglaublichen Bekanntheitsgrad verschafft. Infolgedessen sind die Erwartungen an Künstliche Intelligenz exponentiell gestiegen und haben solche Höhen erreicht, dass man sich fragen könnte, ob sie jemals erreicht werden können.

Das Thema KI folgt dem bekannten Hype-Zyklus. Einige dieser hohen Erwartungen sind wohlverdient: Diese leistungsstarke Technologie wird die Art und Weise, wie wir leben und arbeiten, in vielerlei Hinsicht verändern. Um ein Beispiel zu nennen: Einige Universitäten erwägen, von ihren Studenten keine Seminararbeiten mehr zu verlangen, da es nicht möglich ist zu überprüfen, ob sie von einem KI-Tool geschrieben wurden.

Aber wir müssen uns auch auf einige Enttäuschungen in der Zukunft gefasst machen, da die KI unweigerlich die überzogenen Erwartungen mancher Leute nicht erfüllen kann. Selbst wenn die Vorstellungen vernünftig sind, ist der Zeitrahmen, den diese Menschen und Organisationen für die Umsetzung von KI-Projekten im Sinn haben, oft nicht realistisch. Dies führt zu weiteren Enttäuschungen, wenn die erhoffte Wirkung und der erhoffte Wert nicht innerhalb des gewünschten Zeitrahmens erreicht werden können.

Die ersten Anzeichen dafür sind bereits zu erkennen, denn ChatGPT und ähnliche Tools liefern eine Fülle von wortgewandten und kohärenten – aber nicht korrekten – Informationen. Die neue Welle von „KI-Experten“, die immer haarsträubendere Versprechungen über die von ihnen oder ihren Unternehmen erfundenen Tools machen, die nur schwer zu halten sein werden, trägt nicht zu mehr Vertrauen bei. Im Grunde genommen verkaufen sie digitales „Schlangenöl“.

All dies erhöht den Druck auf die Data Scientists, mit diesen Erwartungen umzugehen und gleichzeitig weiterhin das gleiche Ziel zu erreichen, das sie seit Jahrzehnten verfolgen:

verständliche Antworten auf Fragen anhand von Daten zu geben.

Aus diesem Grund sind neutrale Organisationen wie die Vienna Data Science Group (VDSG, www.vdsg.at), die den interdisziplinären und internationalen Wissensaustausch zwischen Datenexperten fördert, so notwendig und wichtig. Wir engagieren uns nach wie vor stark für die Entwicklung des gesamten Data-Science- und KI-Ökosystems (Ausbildung, Zertifizierung, Standardisierung, Studien zu den gesellschaftlichen Auswirkungen usw.) in Europa und darüber hinaus. Dieses Buch ist nur eine unserer Bemühungen, um dieses Ziel zu erreichen. Denn trotz all des Hypes und der Übertreibungen in der KI- und Datenlandschaft bleibt Data Science dasselbe: eine interdisziplinäre Wissenschaft, die eine sehr heterogene Gruppe von Spezialisten versammelt. Sie setzt sich aus drei großen Strömungen zusammen, und wir sind stolz darauf, dass wir in jeder von ihnen Experten haben:

- Informatik und IT
- Mathematik und Statistik
- Fachwissen in der Branche oder dem Bereich, in dem Data Science und künstliche Intelligenz angewendet werden.

Die VDSG (www.vdsg.at) hat schon immer einen ganzheitlichen Ansatz für Data Science verfolgt, und das ist auch in diesem Buch nicht anders: Ab Kapitel 1 stellen wir ein fiktives Unternehmen vor, das datengetriebener werden möchte, und begleiten es im Laufe des Buches bis zum Ende seiner Datentransformation in Kapitel 28. Auf dem Weg dorthin gehen wir auf viele Herausforderungen ein und bieten Ihnen so praktische Einblicke, die nur dank des regen Austauschs in unserer großen Data-Science- und KI-Community möglich waren.

Das Ergebnis ist eine stark erweiterte Ausgabe unseres Data Science & AI-Handbuchs mit zehn neuen Kapiteln zu Themen wie Aufbau von KI-Lösungen (Kapitel 13), Foundation Models (Kapitel 15), Large Language Models und generative KI (Kapitel 16) sowie Klimawandel und KI (Kapitel 25). Ergänzend dazu werden auch die grundlegenden Themen Datenarchitektur, Engineering und Governance (Kapitel 4, 5 und 6) behandelt und mit Machine Learning Operations (MLOps, Kapitel 7) abgerundet, das sich zu einer eigenen, sehr wichtigen Disziplin entwickelt hat.

Um eine solide Grundlage zu schaffen, die Ihnen hilft, all dies zu verstehen, haben wir wieder eine Einführung in die zugrunde liegende Mathematik (Kapitel 9) und Statistik (Kapitel 10), die in Data Science verwendet werden, sowie Kapitel über die Theorie hinter Machine Learning, der Signalverarbeitung und der Computer Vision (Kapitel 12, 14 und 18) aufgenommen. Wir haben auch Themen behandelt, die mit der Wertschöpfung aus Daten zu tun haben, wie z. B. Business Intelligence (Kapitel 11) und Data Driven Enterprises (Kapitel 21), sowie wichtige Informationen, die Ihnen helfen,

Daten sicher zu nutzen, einschließlich Kapiteln über das neue EU-KI-Gesetz (Kapitel 23) und vertrauenswürdige KI (Kapitel 27).

Diese umfangreiche Erweiterung des Opus Magnum der VDSG dient vor allem einem Zweck:

ein realistisches und ganzheitliches Bild von Data Science und KI zu vermitteln.

Data Science und KI entwickeln sich derzeit in einem unglaublich schnellen Tempo, und das gilt auch für ihre Auswirkungen auf die Gesellschaft. Das bedeutet, dass die Verantwortung, die auf den Schultern der Data Scientists lastet, ebenfalls gewachsen ist, und damit auch die Notwendigkeit für Organisationen wie die VDSG, sich zu engagieren und diese Herausforderungen zu bewältigen.

Packen wir's an!

Sommer 2024

Wolfgang Weidinger

Danksagungen

Wir, die Autoren, möchten diese Gelegenheit nutzen, um unseren Familien und Freunden, die uns geholfen haben, unsere Gedanken und Einsichten in diesem Buch auszudrücken, unseren aufrichtigen Dank auszusprechen. Ohne ihre Unterstützung und Geduld wäre diese Arbeit nicht möglich gewesen.

Ein besonderer Dank aller Autoren geht an Katherine Munro, die viel zu diesem Buch beigetragen und viel Zeit und Mühe in die Bearbeitung unserer Manuskripte investiert hat.

Für meine Eltern, die immer gesagt haben, dass ich alles schaffen kann. Wir hätten nie erwartet, dass es so etwas sein würde.

Katherine Munro

Ich möchte mich bei meiner Frau und der Vienna Data Science Group für ihre kontinuierliche Unterstützung auf meinem beruflichen Weg bedanken.

Zoltan C. Toth

Wenn ich an die Menschen denke, die mich am meisten unterstützt haben, möchte ich mich bei meinen Eltern bedanken, die immer an mich geglaubt haben, egal, was passiert ist, und bei meiner Partnerin Verena, die in den letzten Monaten wieder sehr geduldig war, während ich an diesem Buch gearbeitet habe.

Darüber hinaus bin ich sehr dankbar für die Unterstützung und Motivation, die ich von den Menschen erhalten habe, die ich durch die Vienna Data Science Group kennengelernt habe.

Wolfgang Weidinger

1

Einführung

Stefan Papp

„Ich möchte CDO werden anstelle des CDOs.“

Iznogoud (angepasst)



Fragen, die in diesem Kapitel beantwortet werden:

- Wie könnte ein fiktives Unternehmen aussehen, das vor seiner Transformation in ein datengesteuertes Unternehmen steht?
- Welche Herausforderungen muss ein Unternehmen bewältigen, um datengesteuert zu werden?
- Wie werden die Kapitel in diesem Buch Ihnen, dem Leser, helfen, solche Herausforderungen in Ihrer eigenen Organisation zu erkennen und zu bewältigen?

1.1 Über dieses Buch

Dieses Buch bietet einen praktischen, erfahrungsbasierten Einblick in verschiedene Aspekte von Data Science und künstlicher Intelligenz. In dieser, unserer dritten, Auflage tauchen die Autoren auch tief in einige der aufregendsten und sich schnell entwickelnden Themen unserer Zeit ein, darunter Large Language Models und generative KI.

Das vorrangige Ziel der Autoren ist es, dem Leser einen ganzheitlichen Zugang zu diesem Bereich zu vermitteln. Aus diesem Grund ist dieses Buch nicht rein technisch: Die Reife von Data Science und KI hängt ebenso sehr von der Arbeitskultur ab, insbesondere vom kritischen Denken und evidenzbasierter Entscheidungsfindung, wie von den Kenntnissen in Mathematik, neuronalen Netzen, KI-Frameworks und Datenplattformen.

In den letzten Jahren sind sich die meisten Experten einig geworden, dass künstliche Intelligenz unsere Arbeits- und Lebensweise verändern wird. Für eine ganzheitliche Betrachtung müssen wir auch den Status quo betrachten, wenn wir verstehen wollen, was getan werden muss, um unsere vielfältigen Ambitionen mithilfe von KI zu erfüllen. Ein nützlicher Rahmen dafür ist es zu untersuchen, wie Menschen mit den Herausforderungen der Datentransformation aus einer organisatorischen Perspektive umgehen. Aus diesem Grund werden wir dem Leser ein fiktives Unternehmen vorstellen, das am Anfang seiner Reise steht, evidenzbasierte Entscheidungsfindung in seine Unternehmensidentität zu integrieren. Wir werden dieses fiktive Unternehmen, in dem vieles datenorientierter sein könnte, als Modell verwenden, um mögliche Herausforderungen zu skizzieren, denen Organisationen begegnen können, wenn sie datenorientierter werden wollen. Am Ende dieses Buches wird unser hypothetisches Unternehmen auch als Modell dafür dienen, wie ein datengesteuertes Unternehmen aussehen könnte. In den dazwischen liegenden Kapiteln gehen wir auf viele dieser Herausforderungen ein und geben praktische Ratschläge, wie man sie bewältigen kann.

Falls Sie als Leser lieber keine Prosa über ein fiktives Unternehmen lesen möchten, um etwas über solche typischen organisatorischen Herausforderungen zu erfahren, empfehlen wir Ihnen, dieses Kapitel zu überspringen und mit einem Kapitel zu beginnen, das Ihren Interessen entspricht. Als ganzheitliches Buch über dieses Gebiet behandeln die Autoren künstliche Intelligenz, maschinelles Lernen, generative KI, Modellierung, Verarbeitung natürlicher Sprache, Computer Vision und andere relevante Bereiche. Wir behandeln Engineer-Themen wie Datenarchitektur und Datenpipelines, die für die Umsetzung datengetriebener Projekte in die Produktion unerlässlich sind. Schließlich gehen wir auch auf kritische soziale und rechtliche Fragen im Zusammenhang mit der Nutzung von Daten ein. Jeder Autor geht sehr detailliert auf sein Fachgebiet ein, damit Sie einen großen Nutzen daraus ziehen können.

Wir bitten die Leser, sich direkt mit uns in Verbindung zu setzen und uns mitzuteilen, wie wir unser ehrgeiziges Ziel erreichen können, die Standardliteratur für einen ganzheitlichen Ansatz in diesem Bereich zu werden. Wenn Sie der Meinung sind, dass einige neue Inhalte in einer der nächsten Ausgaben behandelt werden sollten, können Sie die Autoren über berufliche Netzwerke wie LinkedIn finden.

In diesem Sinne, fangen wir an.

1.2 Die Halford Group

Bob betrat das Bürogebäude der Halford Group, einem Hersteller von Konsumgütern, zu denen auch die meistverkaufte Gummiente gehörte. Nachdem er die Bürotüren passiert hatte, fühlte er sich in die Achtzigerjahre zurückversetzt. Die Besucher mussten sich am Eingang anmelden, Formulare ausfüllen, um im Falle eines Unfalls haft-