

Table of Contents

Preface	XXI
Acknowledgments	XXIII
1 Introduction	1
<i>Stefan Papp</i>	
1.1 About this Book	1
1.2 The Halford Group	2
1.2.1 Alice Halford – Chairwoman	3
1.2.2 Analysts	4
1.2.3 “CDO”	5
1.2.4 Sales	6
1.2.5 IT	7
1.2.6 Security	8
1.2.7 Production Leader	9
1.2.8 Customer Service	10
1.2.9 HR	11
1.2.10 CEO	12
1.3 In a Nutshell	13
2 The Alpha and Omega of AI	15
<i>Stefan Papp</i>	
2.1 The Data Use Cases	16
2.1.1 Bias	16
2.1.2 Data Literacy	19
2.2 Culture Shock	20
2.3 Ideation	24
2.4 Design Process Models	25
2.4.1 Design Thinking	26
2.4.2 Double Diamond	27
2.4.3 Conducting Workshops	28
2.5 In a Nutshell	34

3	Cloud Services	35
	<i>Stefan Papp</i>	
3.1	Introduction	35
3.2	Cloud Essentials	36
3.2.1	XaaS	38
3.2.2	Cloud Providers	39
3.2.3	Native Cloud Services	41
3.2.4	Cloud-native Paradigms	44
3.3	Infrastructure as a Service	45
3.3.1	Hardware	46
3.3.2	Distributed Systems	48
3.3.3	Linux Essentials for Data Professionals	51
3.3.4	Infrastructure as Code	57
3.4	Platform as a Service	61
3.4.1	Cloud Native PaaS Solutions	62
3.4.2	External Solutions	66
3.5	Software as a Service	69
3.6	In a Nutshell	70
4	Data Architecture	71
	<i>Zoltan C. Toth and Sean McIntyre</i>	
4.1	Overview	71
4.1.1	Maslow's Hierarchy of Needs for Data	72
4.1.2	Data Architecture Requirements	73
4.1.3	The Structure of a Typical Data Architecture	74
4.1.4	ETL (Extract, Transform, Load)	78
4.1.5	ELT (Extract, Load, Transform)	79
4.1.6	ETLT	80
4.2	Data Ingestion and Integration	80
4.2.1	Data Sources	80
4.2.2	Traditional File Formats	82
4.2.3	Modern File Formats	84
4.2.4	Which Storage Option to Choose?	86
4.3	Data Warehouses, Data Lakes, and Lakehouses	86
4.3.1	Data Warehouses	86
4.3.2	Data Lakes and Cloud Data Platforms	90
4.4	Data Transformation	93
4.4.1	SQL	95
4.4.2	Big Data & Apache Spark	103
4.4.3	Cloud Data Platforms for Apache Spark	110
4.5	Workflow Orchestration	112
4.5.1	Dagster and the Modern Data Stack	114

4.6	A Data Architecture Use Case	115
4.7	In a Nutshell	119
5	Data Engineering	121
	<i>Stefan Papp</i>	
5.1	Differentiating from Software Engineering	122
5.2	Programming Languages	123
5.2.1	Code or No Code?	123
5.2.2	Language Ecosystem	125
5.2.3	Python	126
5.2.4	Scala	129
5.3	Software Engineering Processes for Data	132
5.3.1	Configuration Management	132
5.3.2	CI/CD	133
5.4	Data Pipelines	135
5.4.1	Common Characteristics of a Data Pipeline	135
5.4.2	Data Pipelines in the Unified Data Architecture	136
5.5	Storage Options	139
5.5.1	File Era	139
5.5.2	Database Era	140
5.5.3	Data Lake Era	142
5.5.4	Serverless Era	143
5.5.5	Polyglot Storage	143
5.5.6	Data Mesh Era	145
5.6	Tooling	146
5.6.1	Batch: Airflow	146
5.6.2	Streaming: Kafka	147
5.6.3	Transformation: Databricks Notebooks	152
5.7	Common challenges	153
5.7.1	Data Quality and Different Standards	154
5.7.2	Skewed Data	155
5.7.3	Stressed Operational Systems	156
5.7.4	Legacy Operational Systems	157
5.7.5	Platform and Information Security	157
5.8	In a Nutshell	157
6	Data Governance	159
	<i>Victoria Rugli, Mario Meir-Huber</i>	
6.1	Why Do We Need Data Governance?	159
6.1.1	Sample 1: Achieving Clarity with Data Governance	160
6.1.2	Sample 2: The (Negative) Impact of Poor Data Governance	161
6.2	The Building Blocks of Data Governance	162
6.2.1	Data Governance Explained	163

- 6.3 People 165
 - 6.3.1 Data Ownership 165
 - 6.3.2 Data Stewardship 168
 - 6.3.3 Data Governance Board 169
 - 6.3.4 Change Management 170
- 6.4 Process 172
 - 6.4.1 Metadata Management 173
 - 6.4.2 Data Quality Management 176
 - 6.4.3 Data Security and Privacy 179
 - 6.4.4 Master Data Management 182
 - 6.4.5 Data Access and Search 185
- 6.5 Technology (Data Governance Tools) 187
 - 6.5.1 Open-Source Tools 187
 - 6.5.2 Cloud-based Data Governance Tools 193
- 6.6 In a Nutshell 197
- 7 Machine Learning Operations (ML Ops) 199**
 - Zoltan C. Toth, György Móra*
 - 7.1 Overview 199
 - 7.1.1 Scope of MLOps 200
 - 7.1.2 Data Collection and Exploration 201
 - 7.1.3 Feature Engineering 201
 - 7.1.4 Model Training 201
 - 7.1.5 Models Deployed to Production 202
 - 7.1.6 Model Evaluation 202
 - 7.1.7 Model Understanding 203
 - 7.1.8 Model Versioning 203
 - 7.1.9 Model Monitoring 204
 - 7.2 MLOps in an Organization 204
 - 7.2.1 Main Benefits of MLOps 204
 - 7.2.2 Capabilities Needed for MLOps 205
 - 7.3 Several Common Scenarios in the MLOps Space 205
 - 7.3.1 Integrating Notebooks 205
 - 7.3.2 Features in Production 207
 - 7.3.3 Model Deployment 209
 - 7.3.4 Model Formats 209
 - 7.4 MLOps Tooling and MLflow 210
 - 7.4.1 MLflow 211
 - 7.5 In a Nutshell 214

8	Machine Learning Security	215
	<i>Manuel Pasieka</i>	
8.1	Introduction to Cybersecurity	216
8.2	Attack Surface	217
8.3	Attack Methods	218
8.3.1	Model Stealing	218
8.3.2	Data Extraction	220
8.3.3	Data Poisoning	222
8.3.4	Adversarial Attack	225
8.3.5	Backdoor Attack	227
8.4	Machine Learning Security of Large Language Models	230
8.4.1	Data Extraction	230
8.4.2	Jailbreaking	232
8.4.3	Prompt Injection	233
8.5	AI Threat Modelling	236
8.6	Regulations	237
8.7	Where to go from here	239
8.8	Conclusion	240
8.9	In a Nutshell	241
9	Mathematics	243
	<i>Annalisa Cadonna</i>	
9.1	Linear Algebra	244
9.1.1	Vectors and Matrices	244
9.1.2	Operations between Vectors and Matrices	247
9.1.3	Linear Transformations	250
9.1.4	Eigenvalues, Eigenvectors, and Eigendecomposition	251
9.1.5	Other Matrix Decompositions	252
9.2	Calculus and Optimization	253
9.2.1	Derivatives	254
9.2.2	Gradient and Hessian	256
9.2.3	Gradient Descent	257
9.2.4	Constrained Optimization	259
9.3	Probability Theory	260
9.3.1	Discrete and Continuous Random Variables	261
9.3.2	Expected Value, Variance, and Covariance	264
9.3.3	Independence, Conditional Distributions, and Bayes' Theorem	266
9.4	In a Nutshell	267

10	Statistics – Basics	269
	<i>Rania Wazir, Georg Langs, Annalisa Cadonna</i>	
10.1	Data	270
10.2	Simple Linear Regression	271
10.3	Multiple Linear Regression	278
10.4	Logistic Regression	280
10.5	How Good is Our Model?	287
10.6	In a Nutshell	289
11	Business Intelligence (BI)	291
	<i>Christian Mata</i>	
11.1	Introduction to Business Intelligence	293
11.1.1	Definition of Business Intelligence	294
11.1.2	Role in Organizations	294
11.1.3	Development of Business Intelligence	295
11.1.4	Data Science and AI in the Context of BI	297
11.1.5	Data for Decision-Making	299
11.1.6	Understanding Business Context	300
11.1.7	Business Intelligence Activities	302
11.2	Data Management Fundamentals	304
11.2.1	What is Data Management, Data Integration and Data Warehousing?	305
11.2.2	Data Load Processes – The Case of ETL or ELT	306
11.2.3	Data Modeling	308
11.3	Reporting and Data Analysis	314
11.3.1	Reporting	314
11.3.2	Types of Reports	317
11.3.3	Data Analysis	318
11.3.4	Visual Analysis	320
11.3.5	Significant Trends	321
11.3.6	Relevant BI Technologies	323
11.3.7	BI Tool Examples	326
11.4	BI and Data Science: Complementary Disciplines	329
11.4.1	Differences	329
11.4.2	Similarities	330
11.4.3	Interdependencies	330
11.5	Outlook for Business Intelligence	331
11.5.1	Expectations for the Evolution of BI	332
11.6	In a Nutshell	333

12	Machine Learning	335
	<i>Georg Langs, Katherine Munro, Rania Wazir</i>	
12.1	Introduction	335
12.2	Basics: Feature Spaces	337
12.3	Classification Models	340
12.3.1	K-Nearest-Neighbor-Classifier	341
12.3.2	Support Vector Machine	342
12.3.3	Decision Trees	342
12.4	Ensemble Methods	344
12.4.1	Bias and Variance	345
12.4.2	Bagging: Random Forests	346
12.4.3	Boosting: AdaBoost	350
12.4.4	The Limitations of Feature Construction and Selection	350
12.5	Unsupervised learning: Learning without labels	351
12.5.1	Clustering	351
12.5.2	Manifold Learning	352
12.5.3	Generative Models	353
12.6	Artificial Neural Networks and Deep Learning	354
12.6.1	The Perceptron	354
12.6.2	Artificial Neural Networks	355
12.6.3	Deep Learning	357
12.6.4	Convolutional Neural Networks	357
12.6.5	Training Convolutional Neural Networks	358
12.6.6	Recurrent Neural Networks	360
12.6.7	Long Short-Term Memory	361
12.6.8	Autoencoders and U-Nets	363
12.6.9	Adversarial Training Approaches	364
12.6.10	Generative Adversarial Networks	365
12.6.11	Cycle GANs and Style GANs	367
12.7	Transformers and Attention Mechanisms	368
12.7.1	The Transformer Architecture	368
12.7.2	What the Attention Mechanism Accomplishes	370
12.7.3	Applications of Transformer Models	370
12.8	Reinforcement Learning	371
12.9	Other Architectures and Learning Strategies	374
12.10	Validation Strategies for Machine Learning Techniques	374
12.11	Conclusion	375
12.12	In a Nutshell	376

13	Building Great Artificial Intelligence	377
	<i>Danko Nikolić</i>	
13.1	How AI Relates to Data Science and Machine Learning	377
13.2	A Brief History of AI	381
13.3	Five Recommendations for Designing an AI Solution	383
13.3.1	Recommendation No. 1: Be Pragmatic	383
13.3.2	Recommendation No. 2: Make it Easier for Machines to Learn – Create Inductive Biases	385
13.3.3	Recommendation No. 3: Perform Analytics	390
13.3.4	Recommendation No. 4: Beware of the Scaling Trap	392
13.3.5	Recommendation No. 5: Beware of the Generality Trap (there is no such a thing as free lunch)	401
13.4	Human-level Intelligence	406
13.5	In a Nutshell	408
14	Signal Processing	411
	<i>Jana Eder</i>	
14.1	Introduction	411
14.2	Sampling and Quantization	413
14.3	Frequency Domain Analysis	416
14.3.1	Fourier Transform	416
14.4	Noise Reduction and Filtering Techniques	422
14.4.1	Denoising Using a Gaussian Low-pass Filter	423
14.5	Time Domain Analysis	425
14.5.1	Signal Normalization and Standardization	425
14.5.2	Signal Transformation and Feature Extraction	425
14.5.3	Time Series Decomposition Techniques	428
14.5.4	Autocorrelation: Understanding Signal Similarity over Time	431
14.6	Time-Frequency Domain Analysis	434
14.6.1	Short Term Fourier Transform and Spectrogram	434
14.6.2	Discrete Wavelet Transform	434
14.6.3	Gramian Angular Field	435
14.7	The Relationship of Signal Processing and Machine Learning	437
14.7.1	Techniques for Feature Engineering	438
14.7.2	Preparing for Machine Learning	438
14.8	Practical Applications	439
14.9	In a Nutshell	441

15	Foundation Models	443
	<i>Danko Nikolić</i>	
15.1	The Idea of a Foundation Model	443
15.2	How to Train a Foundation Model?	446
15.3	How Do we Use Foundation Models?	449
15.4	A Breakthrough: There is no End to Learning	455
15.5	In a Nutshell	456
16	Generative AI and Large Language Models	459
	<i>Katherine Munro, Gerald Hahn, Danko Nikolić</i>	
16.1	Introduction to “Gen AI”	459
16.2	Generative AI Modalities	460
16.2.1	Methods for Training Generative Models	462
16.3	Large Language Models	462
16.3.1	What are “LLMs”?	462
16.3.2	How is Something like ChatGPT Trained?	464
16.3.3	Methods for Using LLMs Directly	465
16.3.4	Methods for Customizing an LLM	475
16.4	Vulnerabilities and Limitations of Gen AI Models	483
16.4.1	Introduction	483
16.4.2	Prompt Injection and Jailbreaking Attacks	484
16.4.3	Hallucinations, Confabulations, and Reasoning Errors	487
16.4.4	Copyright Concerns	488
16.4.5	Bias	491
16.5	Building Robust, Effective Gen AI Applications	494
16.5.1	Control Strategies Throughout Development and Use	494
16.5.2	Guardrails	495
16.5.3	Using Generative AI Safely and Successfully	496
16.6	In a Nutshell	497
17	Natural Language Processing (NLP)	503
	<i>Katherine Munro</i>	
17.1	What is NLP and Why is it so Valuable?	503
17.2	Why Learn “Traditional” NLP in the “Age of Large Language Models”?	505
17.3	NLP Data Preparation Techniques	506
17.3.1	The NLP Pipeline	506
17.3.2	Converting the Input Format for Machine Learning	513
17.4	NLP Tasks and Methods	514
17.4.1	Rule-Based (Symbolic) NLP	515
17.4.2	Statistical Machine Learning Approaches	518
17.4.3	Neural NLP	527
17.4.4	Transfer Learning	532
17.5	In a Nutshell	543

18	Computer Vision	547
	<i>Roxane Licandro</i>	
18.1	What is Computer Vision?	547
18.2	A Picture Paints a Thousand Words	549
18.2.1	The Human Eye	549
18.2.2	Image Acquisition Principle	551
18.2.3	Digital File Formats	556
18.2.4	Image Compression	557
18.3	I Spy With My Little Eye Something That Is...	558
18.3.1	Computational Photography and Image Manipulation	560
18.4	Computer Vision Applications & Future Directions	564
18.4.1	Image Retrieval Systems	564
18.4.2	Object Detection, Classification and Tracking	567
18.4.3	Medical Computer Vision	568
18.5	Making Humans See	571
18.6	In a Nutshell	573
19	Modelling and Simulation – Create your own Models	577
	<i>Günther Zauner, Wolfgang Weidinger, Dominik Brunmeir, Benedikt Spiegel</i>	
19.1	Introduction	578
19.2	General Considerations during Modeling	579
19.3	Modelling to Answer Questions	580
19.4	Reproducibility and Model Lifecycle	581
19.4.1	The Lifecycle of a Modelling and Simulation Question	583
19.4.2	Parameter and Output Definition	584
19.4.3	Documentation	587
19.4.4	Verification and Validation	588
19.5	Methods	591
19.5.1	Ordinary Differential Equations (ODEs)	592
19.5.2	System Dynamics (SD)	593
19.5.3	Discrete Event Simulation	596
19.5.4	Agent-based Modelling	599
19.6	Modelling and Simulation Examples	601
19.6.1	Dynamic Modelling of Railway Networks for Optimal Pathfinding Using Agent-based Methods and Reinforcement Learning	602
19.6.2	Agent-Based Covid Modelling Strategies	604
19.6.3	Deep Reinforcement Learning Approach for Optimal Replenishment Policy in a VMI Setting	609
19.6.4	Finding Feasible Solutions for a Resource-constrained Project Scheduling Problem with Reinforcement Learning and Implementing a Dynamic Planing Scheme with Discrete Event Simulation	612
19.7	Summary and Lessons Learned	616
19.8	In a Nutshell	616

20	Data Visualization	621
	<i>Barbora Antosova Vesela</i>	
20.1	History	622
20.2	Which Tools to Use	627
20.3	Types of Data Visualizations	629
20.3.1	Scatter Plot	630
20.3.2	Line Chart	630
20.3.3	Column and Bar Charts	631
20.3.4	Histogram	632
20.3.5	Pie Chart	633
20.3.6	Box Plot	634
20.3.7	Heat Map	634
20.3.8	Tree Diagram	635
20.3.9	Other Types of Visualizations	636
20.4	Select the right Data Visualization	636
20.5	Tips and Tricks	638
20.6	Presentation of Data Visualization	643
20.7	In a Nutshell	643
21	Data Driven Enterprises	647
	<i>Mario Meir-Huber, Stefan Papp</i>	
21.1	The three Levels of a Data Driven Enterprise	648
21.2	Culture	648
21.2.1	Corporate Strategy for Data	649
21.2.2	The Current State Analysis	651
21.2.3	Culture and Organization of a Successful Data Organisation	653
21.2.4	Core Problem: The Skills Gap	660
21.3	Technology	662
21.3.1	The Impact of Open Source	662
21.3.2	Cloud	662
21.3.3	Vendor Selection	663
21.3.4	Data Lake from a Business Perspective	663
21.3.5	The Role of IT	664
21.3.6	Data Science Labs	664
21.3.7	Revolution in Architecture: The Data Mesh	665
21.4	Business	667
21.4.1	Buy and Share Data	667
21.4.2	Analytical Use Case Implementation	668
21.4.3	Self-service Analytics	669
21.5	In a Nutshell	669

22	Creating High-Performing Teams	671
	<i>Stefan Papp</i>	
22.1	Forming	671
22.2	Storming	672
22.2.1	Scenario: 50 Shades of Red	672
22.2.2	Scenario: Retrospective	676
22.3	Norming	678
22.3.1	Change Management and Transition	678
22.3.2	RACI Matrix	680
22.3.3	SMART	682
22.3.4	Agile Processes	683
22.3.5	Communication Culture	685
22.3.6	DataOps	686
22.4	Performing	690
22.4.1	Scenario: A new Dawn	691
22.4.2	Growth Mindsets	692
22.5	In a Nutshell	695
23	Artificial Intelligence Act	697
	<i>Jeannette Gorzala, Karin Bruckmüller</i>	
23.1	Introduction	698
23.2	Definition of AI Systems	700
23.3	Scope and Purpose of the AI Act	701
23.3.1	The Risk-Based Approach	702
23.3.2	Unacceptable Risk and Prohibited AI Practices	703
23.3.3	High-Risk AI Systems and Compliance	705
23.3.4	Medium Risk and Transparency Obligations	707
23.3.5	Minimal Risk and Voluntary Commitments	708
23.4	General Purpose AI Models	708
23.5	Timeline and Applicability	711
23.6	Penalties	711
23.7	AI and Civil Liability	712
23.8	AI and Criminal Liability	712
23.9	In a Nutshell	715
24	AI in Different Industries	717
	<i>Stefan Papp, Mario Meir-Huber, Wolfgang Weidinger, Thomas Tremel</i>	
24.1	Automotive	720
24.1.1	Vision	721
24.1.2	Data	722
24.1.3	Use Cases	722
24.1.4	Challenges	723

- 24.2 Aviation 725
 - 24.2.1 Vision 725
 - 24.2.2 Data 726
 - 24.2.3 Use Cases 726
 - 24.2.4 Challenges 727
- 24.3 Energy 727
 - 24.3.1 Vision 728
 - 24.3.2 Data 728
 - 24.3.3 Use Cases 729
 - 24.3.4 Challenges 730
- 24.4 Finance 730
 - 24.4.1 Vision 730
 - 24.4.2 Data 731
 - 24.4.3 Use Cases 731
 - 24.4.4 Challenges 733
- 24.5 Health 733
 - 24.5.1 Vision 734
 - 24.5.2 Data 735
 - 24.5.3 Use Cases 735
 - 24.5.4 Challenges 735
- 24.6 Government 736
 - 24.6.1 Vision 736
 - 24.6.2 Data 737
 - 24.6.3 Use Cases 737
 - 24.6.4 Challenges 740
- 24.7 Art 740
 - 24.7.1 Vision 741
 - 24.7.2 Data 741
 - 24.7.3 Use cases 741
 - 24.7.4 Challenges 742
- 24.8 Manufacturing 742
 - 24.8.1 Vision 743
 - 24.8.2 Data 743
 - 24.8.3 Use Cases 743
 - 24.8.4 Challenges 744
- 24.9 Oil and Gas 745
 - 24.9.1 Vision 745
 - 24.9.2 Data 745
 - 24.9.3 Use Cases 746
 - 24.9.4 Challenges 748
- 24.10 Retail 748
 - 24.10.1 Vision 748
 - 24.10.2 Data 749

24.10.3 Use Cases	749
24.10.4 Challenges	750
24.11 Telecommunications Provider	750
24.11.1 Vision	751
24.11.2 Data	751
24.11.3 Use Cases	751
24.11.4 Challenges	753
24.12 Transport	753
24.12.1 Vision	754
24.12.2 Data	754
24.12.3 Use Cases	754
24.12.4 Challenges	755
24.13 Teaching and Training	755
24.13.1 Vision	756
24.13.2 Data	757
24.13.3 Use Cases	757
24.13.4 Challenges	758
24.14 The Digital Society	758
24.15 In a Nutshell	760
25 Climate Change and AI	761
<i>Stefan Papp</i>	
25.1 Introduction	761
25.2 AI – a Climate Saver?	763
25.3 Measuring and Reducing Emissions	763
25.3.1 Baseline	763
25.3.2 Data Use Cases	765
25.4 Sequestration	766
25.4.1 Biological Sequestration	768
25.4.2 Geological Sequestration	769
25.5 Prepare for Impact	770
25.6 Geoengineering	771
25.7 Greenwashing	773
25.8 Outlook	774
25.9 In a Nutshell	776
26 Mindset and Community	777
<i>Stefan Papp</i>	
26.1 Data-Driven Mindset	777
26.2 Data Science Culture	780
26.2.1 Start-up or Consulting Firm?	780
26.2.2 Labs Instead of Corporate Policy	781

26.2.3	Keiretsu Instead of Lone Wolf	781
26.2.4	Agile Software Development	783
26.2.5	Company and Work Culture	783
26.3	Antipatterns	786
26.3.1	Devaluation of Domain Expertise	786
26.3.2	IT Will Take Care of It	787
26.3.3	Resistance to Change	787
26.3.4	Know-it-all Mentality	788
26.3.5	Doom and Gloom	789
26.3.6	Penny-pinching	789
26.3.7	Fear Culture	790
26.3.8	Control over Resources	790
26.3.9	Blind Faith in Resources	791
26.3.10	The Swiss Army Knife	792
26.3.11	Over-Engineering	792
26.4	In a Nutshell	793
27	Trustworthy AI	795
	<i>Rania Wazir</i>	
27.1	Legal and Soft-Law Framework	796
27.1.1	Standards	798
27.1.2	Regulations	798
27.2	AI Stakeholders	800
27.3	Fairness in AI	801
27.3.1	Bias	802
27.3.2	Fairness Metrics	805
27.3.3	Mitigating Unwanted Bias in AI Systems	808
27.4	Transparency of AI Systems	809
27.4.1	Documenting the Data	810
27.4.2	Documenting the Model	811
27.4.3	Explainability	812
27.5	Conclusion	814
27.6	In a Nutshell	814
28	Epilogue	815
	<i>Stefan Papp</i>	
28.1	Halford 2.0	815
28.1.1	Environmental, Social and Governance	816
28.1.2	HR	817
28.1.3	Customer Satisfaction	818
28.1.4	Production	819
28.1.5	IT	820
28.1.6	Strategy	822

28.2	Final Words	823
28.3	In a Nutshell	824
29	The Authors	825
Index		833