

Table of Contents

Foreword	XV
Preface	XVIII
Acknowledgments	XX
1 Introduction	1
1.1 What are Data Science, Machine Learning and Artificial Intelligence?	2
1.2 Data Strategy	8
1.3 From Strategy to Use Cases	10
1.3.1 Data Teams	11
1.3.2 Data and Platforms	16
1.3.3 Modeling and Analysis	17
1.4 Use Case Implementation	18
1.4.1 Iterative Exploration of Use Cases	19
1.4.2 End-to-End Data Processing	21
1.4.3 Data Products	22
1.5 Real-Life Use Case Examples	22
1.5.1 Value Chain Digitization (VCD)	22
1.5.2 Marketing Segment Analytics	23
1.5.3 360° View of the Customer	23
1.5.4 NGO and Sustainability Use Cases	24
1.6 Delivering Results	25
1.7 In a Nutshell	27
2 Infrastructure	29
<i>Stefan Papp</i>	
2.1 Introduction	29
2.2 Hardware	31
2.2.1 Distributed Systems	34
2.2.2 Hardware for AI Applications	37
2.3 Linux Essentials for Data Professionals	38

2.4	Terraform	54
2.5	Cloud	58
2.5.1	Basic Services	61
2.5.2	Cloud-native Solutions	65
2.6	In a Nutshell	68
3	Data Architecture	69
	<i>Zoltan C. Toth</i>	
3.1	Overview	69
3.1.1	Maslow's Hierarchy of Needs for Data	69
3.1.2	Data Architecture Requirements	71
3.1.3	The Structure of a Typical Data Architecture	71
3.1.4	ETL (Extract, Transform, Load)	72
3.1.5	ELT (Extract, Load, Transform)	73
3.1.6	ETLT	73
3.2	Data Ingestion and Integration	74
3.2.1	Data Sources	74
3.2.2	Traditional File Formats	75
3.2.3	Modern File Formats	77
3.2.4	Summary	79
3.3	Data Warehouses, Data Lakes, and Lakehouses	79
3.3.1	Data Warehouses	79
3.3.2	Data Lakes and the Lakehouse	83
3.3.3	Summary: Comparing Data Warehouses to Lakehouses	85
3.4	Data Processing and Transformation	86
3.4.1	Big Data & Apache Spark	86
3.4.2	Databricks	93
3.5	Workflow Orchestration	94
3.6	A Data Architecture Use Case	96
3.7	In a Nutshell	100
4	Data Engineering	101
	<i>Stefan Papp, Bernhard Ortner</i>	
4.1	Data Integration	102
4.1.1	Data Pipelines	102
4.1.2	Designing Data Pipelines	108
4.1.3	CI/CD	110
4.1.4	Programming Languages	112
4.1.5	Kafka as Reference ETL Tool	115
4.1.6	Design Patterns	119
4.1.7	Automation of the Stages	120
4.1.8	Six Building Blocks of the Data Pipeline	120

4.2	Managing Analytical Models	125
4.2.1	Model Delivery	126
4.2.2	Model Update	127
4.2.3	Model or Parameter Update	128
4.2.4	Model Scaling	128
4.2.5	Feedback into the Operational Processes	129
4.3	In a Nutshell	130
5	Data Management	131
	<i>Stefan Papp, Bernhard Ortner</i>	
5.1	Data Governance	133
5.1.1	Data Catalog	134
5.1.2	Data Discovery	136
5.1.3	Data Quality	140
5.1.4	Master Data Management	141
5.1.5	Data Sharing	142
5.2	Information Security	143
5.2.1	Data Classification	144
5.2.2	Privacy Protection	145
5.2.3	Encryption	147
5.2.4	Secrets Management	149
5.2.5	Defense in Depth	150
5.3	In a Nutshell	151
6	Mathematics	153
	<i>Annalisa Cadonna</i>	
6.1	Linear Algebra	154
6.1.1	Vectors and Matrices	154
6.1.2	Operations between Vectors and Matrices	157
6.1.3	Linear Transformations	160
6.1.4	Eigenvalues, Eigenvectors, and Eigendecomposition	161
6.1.5	Other Matrix Decompositions	162
6.2	Calculus and Optimization	163
6.2.1	Derivatives	164
6.2.2	Gradient and Hessian	166
6.2.3	Gradient Descent	167
6.2.4	Constrained Optimization	169
6.3	Probability Theory	170
6.3.1	Discrete and Continuous Random Variables	171
6.3.2	Expected Value, Variance, and Covariance	174
6.3.3	Independence, Conditional Distributions, and Bayes' Theorem	176
6.4	In a Nutshell	177

7	Statistics – Basics	179
	<i>Rania Wazir, Georg Langs, Annalisa Cadonna</i>	
7.1	Data	180
7.2	Simple Linear Regression	181
7.3	Multiple Linear Regression	189
7.4	Logistic Regression	191
7.5	How Good is Our Model?	198
7.6	In a Nutshell	199
8	Machine Learning	201
	<i>Georg Langs, Katherine Munro, Rania Wazir</i>	
8.1	Introduction	201
8.2	Basics: Feature Spaces	203
8.3	Classification Models	206
8.3.1	K-Nearest-Neighbor-Classifier	206
8.3.2	Support Vector Machine	207
8.3.3	Decision Tree	208
8.4	Ensemble Methods	209
8.4.1	Bias and Variance	210
8.4.2	Bagging: Random Forests	211
8.4.3	Boosting: AdaBoost	215
8.5	Artificial Neural Networks and the Perceptron	215
8.6	Learning without Labels – Finding Structure	218
8.6.1	Clustering	218
8.6.2	Manifold Learning	219
8.6.3	Generative Models	220
8.7	Reinforcement Learning	221
8.8	Overarching Concepts	223
8.9	Into the Depth – Deep Learning	224
8.9.1	Convolutional Neural Networks	224
8.9.2	Training Convolutional Neural Networks	225
8.9.3	Recurrent Neural Networks	227
8.9.4	Long Short-Term Memory	228
8.9.5	Autoencoders and U-Nets	230
8.9.6	Adversarial Training Approaches	231
8.9.7	Generative Adversarial Networks	232
8.9.8	Cycle GANs and Style GANs	234
8.9.9	Other Architectures and Learning Strategies	235
8.10	Validation Strategies for Machine Learning Techniques	235
8.11	Conclusion	237
8.12	In a Nutshell	237

9 Building Great Artificial Intelligence 239

Danko Nikolić

9.1	How AI Relates to Data Science and Machine Learning	239
9.2	A Brief History of AI	243
9.3	Five Recommendations for Designing an AI Solution	245
9.3.1	Recommendation No. 1: Be pragmatic	245
9.3.2	Recommendation No. 2: Make it easier for machines to learn - create inductive biases	247
9.3.3	Recommendation No. 3: Perform analytics	252
9.3.4	Recommendation No. 4: Beware of the scaling trap	254
9.3.5	Recommendation No. 5: Beware of the generality trap (there is no such a thing as free lunch)	263
9.4	Human-level Intelligence	268
9.5	In a Nutshell	270

10 Natural Language Processing (NLP) 273

Katherine Munro

10.1	What is NLP and Why is it so Valuable?	273
10.2	NLP Data Preparation Techniques	275
10.2.1	The NLP Pipeline	275
10.2.2	Converting the Input Format for Machine Learning	281
10.3	NLP Tasks and Methods	283
10.3.1	Rule-Based (Symbolic) NLP	284
10.3.2	Statistical Machine Learning Approaches	287
10.3.3	Neural NLP	295
10.3.4	Transfer Learning	301
10.4	At the Cutting Edge: Current Research Focuses for NLP	312
10.5	In a Nutshell	314

11 Computer Vision 317

Roxane Licandro

11.1	What is Computer Vision?	317
11.2	A Picture Paints a Thousand Words	319
11.2.1	The Human Eye	319
11.2.2	Image Acquisition Principle	321
11.2.3	Digital File Formats	326
11.2.4	Image Compression	327
11.3	I Spy With My Little Eye Something That Is... ..	328
11.3.1	Computational Photography and Image Manipulation	330
11.4	Computer Vision Applications & Future Directions	334
11.4.1	Image Retrieval Systems	334
11.4.2	Object Detection, Classification and Tracking	337
11.4.3	Medical Computer Vision	338



11.5	Making Humans See	341
11.6	In a Nutshell	343
12	Modelling and Simulation – Create your own Models	347
	<i>Günther Zauner, Wolfgang Weidinger</i>	
12.1	Introduction	347
12.2	General Aspects	349
12.3	Modelling to Answer Questions	349
12.4	Reproducibility and Model Lifecycle	351
12.4.1	The Lifecycle of a Modelling and Simulation Question	352
12.4.2	Parameter and Output Definition	354
12.4.3	Documentation	357
12.4.4	Verification and Validation	357
12.5	Methods	361
12.5.1	Ordinary Differential Equations (ODEs)	361
12.5.2	System Dynamics (SD)	362
12.5.3	Discrete Event Simulation	365
12.5.4	Agent-Based Modelling	368
12.6	Modelling and Simulation Examples	371
12.6.1	Dynamic Modelling of Railway Networks for Optimal Pathfinding Using Agent-based Methods and Reinforcement Learning	371
12.6.2	Agent-Based Covid Modelling Strategies	373
12.6.3	Deep Reinforcement Learning Approach for Optimal Replenishment Policy in a VMI Setting	378
12.7	Summary and Lessons Learned	381
12.8	In a Nutshell	381
13	Data Visualization	385
	<i>Barbora Vesela</i>	
13.1	History	386
13.2	Which Tools to Use	391
13.3	Types of Data Visualizations	393
13.3.1	Scatter Plot	394
13.3.2	Line Chart	394
13.3.3	Column and Bar Charts	395
13.3.4	Histogram	396
13.3.5	Pie Chart	397
13.3.6	Box Plot	398
13.3.7	Heat Map	398
13.3.8	Tree Diagram	399
13.3.9	Other Types of Visualizations	400
13.4	Select the right Data Visualization	400

13.5	Tips and Tricks	402
13.6	Presentation of Data Visualization	407
13.7	In a Nutshell	407
14	Data Driven Enterprises	411
	<i>Mario Meir-Huber, Stefan Papp</i>	
14.1	The three Levels of a Data Driven Enterprise	412
14.2	Culture	412
14.2.1	Corporate Strategy for Data	413
14.2.2	The Current State Analysis	415
14.2.3	Culture and Organization of a Successful Data Organisation	417
14.2.4	Core Problem: The Skills Gap	424
14.3	Technology	426
14.3.1	The Impact of Open Source	426
14.3.2	Cloud	426
14.3.3	Vendor Selection	427
14.3.4	Data Lake from a Business Perspective	427
14.3.5	The Role of IT	428
14.3.6	Data Science Labs	428
14.3.7	Revolution in Architecture: The Data Mesh	429
14.4	Business	431
14.4.1	Buy and Share Data	431
14.4.2	Analytical Use Case Implementation	432
14.4.3	Self-service Analytics	433
14.5	In a Nutshell	433
15	Legal foundation of Data Science	435
	<i>Bernhard Ortner</i>	
15.1	Introduction	435
15.2	Categories of Data	436
15.3	General Data Protection Regulation	437
15.3.1	Fundamental Rights of GDPR	437
15.3.2	Declaration of Consent	438
15.3.3	Risk-assessment	440
15.3.4	Anonymization und Pseudo-anonymization	441
15.3.5	Types of Anonymization	442
15.3.6	Lawful and Transparent Data Processing	444
15.3.7	Right to Data Deletion and Correction	445
15.3.8	Privacy by Design	446
15.3.9	Privacy by Default	446
15.4	ePrivacy-Regulation	446
15.5	Data Protection Officer	447
15.5.1	International Data Export in Foreign Countries	447

15.6	Security Measures	448
15.6.1	Data Encryption	449
15.7	CCPA compared to GDPR	449
15.7.1	Territorial Scope	450
15.7.2	Opt-in vs. Opt-out	450
15.7.3	Right of Data Export	450
15.7.4	Right Not to be Discriminated Against	451
15.8	In a Nutshell	451
16	AI in Different Industries	453
	<i>Stefan Papp, Mario Meir-Huber, Wolfgang Weidinger, Thomas Tremel, Marek Danis</i>	
16.1	Automotive	456
16.1.1	Vision	457
16.1.2	Data	458
16.1.3	Use Cases	458
16.1.4	Challenges	459
16.2	Aviation	461
16.2.1	Vision	461
16.2.2	Data	462
16.2.3	Use cases	462
16.2.4	Challenges	463
16.3	Energy	463
16.3.1	Vision	464
16.3.2	Data	464
16.3.3	Use Cases	465
16.3.4	Challenges	466
16.4	Finance	466
16.4.1	Vision	466
16.4.2	Data	467
16.4.3	Use Cases	467
16.4.4	Challenges	469
16.5	Health	469
16.5.1	Vision	470
16.5.2	Data	471
16.5.3	Use Cases	471
16.5.4	Challenges	471
16.6	Government	472
16.6.1	Vision	472
16.6.2	Data	473
16.6.3	Use Cases	473
16.6.4	Challenges	476

16.7	Art	476
16.7.1	Vision	477
16.7.2	Data	477
16.7.3	Use cases	477
16.7.4	Challenges	478
16.8	Manufacturing	478
16.8.1	Vision	479
16.8.2	Data	479
16.8.3	Use Cases	479
16.8.4	Challenges	480
16.9	Oil and Gas	481
16.9.1	Vision	481
16.9.2	Data	481
16.9.3	Use Cases	482
16.9.4	Challenges	484
16.10	Safety at Work	484
16.10.1	Vision	484
16.10.2	Data	485
16.10.3	Use Cases	485
16.10.4	Challenges	486
16.11	Retail	487
16.11.1	Vision	487
16.11.2	Data	487
16.11.3	Use Cases	488
16.11.4	Challenges	488
16.12	Telecommunications Provider	489
16.12.1	Vision	489
16.12.2	Data	490
16.12.3	Use Cases	490
16.12.4	Challenges	492
16.13	Transport	492
16.13.1	Vision	492
16.13.2	Data	493
16.13.3	Use Cases	493
16.13.4	Challenges	494
16.14	Teaching and Training	494
16.14.1	Vision	495
16.14.2	Data	496
16.14.3	Use Cases	496
16.14.4	Challenges	497
16.15	The Digital Society	497
16.16	In a Nutshell	499

17 Mindset and Community	501
<i>Stefan Papp</i>	
17.1 Data-Driven Mindset	501
17.2 Data Science Culture	504
17.2.1 Start-up or Consulting Firm?	504
17.2.2 Labs Instead of Corporate Policy	505
17.2.3 Keiretsu Instead of Lone Wolf	505
17.2.4 Agile Software Development	507
17.2.5 Company and Work Culture	507
17.3 Antipatterns	510
17.3.1 Devaluation of Domain Expertise	510
17.3.2 IT Will Take Care of It	511
17.3.3 Resistance to Change	511
17.3.4 Know-it-all Mentality	512
17.3.5 Doom and Gloom	513
17.3.6 Penny-pinching	513
17.3.7 Fear Culture	514
17.3.8 Control over Resources	514
17.3.9 Blind Faith in Resources	515
17.3.10 The Swiss Army Knife	516
17.3.11 Over-Engineering	516
17.4 In a Nutshell	517
18 Trustworthy AI	519
<i>Rania Wazir</i>	
18.1 Legal and Soft-Law Framework	520
18.1.1 Standards	522
18.1.2 Regulations	522
18.2 AI Stakeholders	524
18.3 Fairness in AI	525
18.3.1 Bias	526
18.3.2 Fairness Metrics	529
18.3.3 Mitigating Unwanted Bias in AI Systems	532
18.4 Transparency of AI Systems	533
18.4.1 Documenting the Data	534
18.4.2 Documenting the Model	535
18.4.3 Explainability	536
18.5 Conclusion	538
18.6 In a Nutshell	538
19 The authors	539
Index	545