VDSG
VIENNA DATA SCIENCE GROUP

Papp · Weidinger · Munro ·
Ortner · Cadonna · Langs ·
Licandro · Meir-Huber · Nikolić ·
Toth · Vesela · Wazir · Zauner

THE HANDBOOK OF DATA SCIENCE AND AI

Generate Value from Data with **Machine Learning** and **Data Analytics**

HANSER

Papp/Weidinger/Munro/Ortner/Cadonna/Langs/
Licandro/Meir-Huber/Nikolić/Toth/Vesela/Wazir/Zauner

The Handbook of Data Science and AI

Stefan Papp, Wolfgang Weidinger,
Katherine Munro, Bernhard Ortner,
Annalisa Cadonna, Georg Langs,
Roxane Licandro, Mario Meir-Huber,
Danko Nikolić, Zoltan Toth, Barbora Vesela,
Rania Wazir, Günther Zauner

The Handbook of Data Science and AI

Generate Value from Data with Machine Learning and Data Analytics

HANSER
Hanser Publishers, Munich

Distributed by:
Carl Hanser Verlag
Postfach 86 04 20, 81631 Munich, Germany
Fax: +49 (89) 98 48 09
www.hanserpublications.com
www.hanser-fachbuch.de

Coverconcept: Marc Müller-Bremer, www.rebranding.de, Munich
Coverdesign: Max Kostopoulos
Cover image: © gettyimages.de/ValeryBrozhinsky
Typesetting: Eberl & Kœsel Studio GmbH, Altusried-Krugzell, Germany
Printed and bound by Hubert & Co. GmbH und Co. KG BuchPartner, Göttingen, Germany
Printed in Germany

Print ISBN: 978-1-56990-886-0
E-Book ISBN: 978-1-56990-887-7
ePub ISBN: 978-1-56990-888-4

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**

Stefan Papp

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
Zoltan C. Toth

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
Stefan Papp, Bernhard Ortner

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
Stefan Papp, Bernhard Ortner

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
Annalisa Cadonna

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
Rania Wazir, Georg Langs, Annalisa Cadonna

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
Georg Langs, Katherine Munro, Rania Wazir

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
Günther Zauner, Wolfgang Weidinger

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
Barbora Vesela

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
Mario Meir-Huber, Stefan Papp

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
Bernhard Ortner

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
Stefan Papp, Mario Meir-Huber, Wolfgang Weidinger, Thomas Treml, Marek Danis

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
Stefan Papp

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
Rania Wazir

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

Foreword

"Mathematical science shows what is. It is the language of unseen relations between things. But to use and apply that language, we must be able to fully appreciate, to feel, to seize the unseen, the unconscious." – Ada Lovelace

As Computer Literacy over a generation ago represented a new set of foundational skills to be acquired, Artificial Intelligence (AI) Literacy represents the same for our current generations and beyond. Over the last two decades Data Science has come to encompass the mathematical architecture and corresponding language with which we build and interact with systems that extend our senses and decision-making abilities. Thus, it's no longer sufficient to be able to send point-and-click commands into computers, but rather it's vitally important to be able to interpret and interact with AI-enabled recommendations coming out of computers. Currently, machines, as in computers coupled with sensors (in the broadest sense), are processing an increasingly wide array of data including text, images, video, audio, network graphs and a multitude of information from the web, private industry, and public sector sources. Considering diversity of data, the authors of this book approach Data Science as a key underlying topic for society and do so with great insight, from multiple key vantage points, and in enjoyable style that resonates with novices and experts alike.

To gain value from data is arguably the unifying objective of the 21st century knowledge worker. Even professional areas thought of as classically distant from data such as sales and art, now have data-driven sub-areas such as marketing automation and computational design. For the benefit of readers, the authors bring to bear first-hand experiences and diligent research to provide a compelling narrative on how we all have a role to play when attempting to leverage data for better outcomes. Indeed, the breadth conveyed in this work is impressive, spanning that of hardware performance considerations (e.g. CPU, Network, Memory, I/O, GPU) to that of different team member roles when building machines that can find patterns in data. Moreover, the authors provide important coverage on the ways that machines can now see and read, namely, Computer Vision and Natural Language Processing, with implications across nearly every industry area being profound.

As you read this book, I encourage you to be curious and have on top of mind a set of questions on how your professional journey and society as you see it is currently being impacted by increasingly advanced machines: from the capabilities available on your smartphone to that of how jobs are being refashioned in the marketplace with automation tools. Here are some questions to help you get started:

- How does the ratio of what tasks you spend your time on shift with the emergence of increasingly advanced machines in your job area?
- What are the implications of having machines that have perceptive abilities analogous to your own, as in to see, hear, smell, taste, touch and beyond?

- How as society do we grapple with bias in and trust around data?
- How do we make the building and the use of machines that learn more inclusive?
- What distinctly human abilities can you accentuate to help organizations that you care about to be more competitive and sustainable?

I've been cautious not to use the term thinking machines, or artificial general intelligence, as to be wary about overstatements. What I would like to focus your attention on is the wide applicability of what we're seeing coming out of research surrounding machines that have learning capabilities. From my time in laboratories at Columbia and Cornell Universities, to that of the Princeton Plasma Physics Laboratory, the American University of Armenia and NASA-backed TRISH (Translational Research Institute for Space Health) which is collaborating with TrialX, it's clear that machines can find patterns in data across a tremendously wide range of domains and alert humans in both regular and mission critical contexts. Thus, the impacts to human experience are multi-faceted and Data Scientists have an important role in supporting the design of systems where human interaction with machine output is positive sum. I can't underscore this enough that a zero-sum approach to automation in sub-optimal. Entrepreneurs though tend to find a way toward maximum sum.

With colleagues and through my work at the BAJ Accelerator and Covenant Venture Capital, I support startups to engage in a type of tandem learning: how a rapidly growing company can transform an industry by spotting market gaps to that of how a company's invention can learn and provide new capabilities for customers. For example, in the powerful technology area of Computer Vision that is a mainstay in Data Science, three companies stand out as trailblazing in three very different industry areas: Embodied, Scylla and cognaize in healthcare, security and finance, respectively.

- Embodied's flagship product, Moxie, is a robot that supports the emotional well-being and social development of children. To do so, Moxie must see and communicate with family members in a compelling way, understanding visually as well as via other cues the emotional state of people it's interacting with as to engage in meaningful dialogue. Thus, healthcare providers have a new robotic team member to collaborate with. Embodied has been on the cover of TIME Magazine.
- Scylla enables an organization's security team to be proactive in improving safety. With real-time detection capabilities, camera networks no longer need to be passive and can be transformed to being proactive. Applications are numerous from detecting slip-and-falls in hospitals and stadiums as they happen to improve health outcomes to that of making intruder alerts at manufacturing facilities and office buildings to better protect staff. Scylla has been featured in Forbes.
- cognaize supports financial institutions and insurance organizations process a tremendous amount of unstructured data when making risk determinations. A key insight is considering documents not only as text, but rather also considering visual information: style, tables, structure. In addition, cognaize has a human-in-the-loop whereby colleagues and the system overall continually learn. cognaize has been featured on the NASDAQ screen in Times Square.

In the above three examples of rising unicorn startups, Data Scientists work in close collaboration with engineers, analysts, designers, content creators, domain specialists and customers to build machines that learn and interact with humans in nuanced ways. The

result is a transformation in the nature of work: humans are alerted to the most important documents or moments in time and human experience is learned from to improve quality. This is representative of a new shift requiring AI Literacy, where jobs in nearly every facet of the economy will have aspects requiring machine interaction: humans making corrections, learning new skills, reacting to and interpreting alerts, and having a faster response time in helping other humans leveraging machines in support. In the years ahead, I'm excited about the role of Data Science in interface research, new algorithms and how humans can have a force multiplication on their work.

As I co-wrote the first edition of *The Field Guide to Data Science* nearly a decade ago, it's remarkable how much the discipline has advanced both in terms of what has been technically achieved and in an aspirational sense on what is yet to be. The Handbook of Data Science and AI advances the discipline along both of those dimensions and carries the torch forward.

Read on.

Fall 2021

Armen R. Kherlopian, Ph.D.

Preface

"The job of the data scientist is to ask the right questions."
Hillary Mason

Reading the foreword written for our first publication two years ago, I couldn't shake the feeling that some trends essentially stayed the same while others emerged all of a sudden and hit society and companies like an avalanche.

Starting with the changes, that struck society profoundly, it is obvious that the pandemic is one of them. Setting aside the myriad of consequences it had and continues to have on our lives, I want to focus on the facets which relate to the subject of this book: Data Science and AI.

Put simply, the impact there was that entire societies and our whole way of living became data driven in an instant. Key performance indicators like the seven-day incidence rate or forecasts based on pandemic simulations steered our daily life and temporarily even altered basic rights, like the right to leave our homes. This led to discussions and questions, which every Data Scientist with some experience is familiar with and has encountered repeatedly during their working life:

- Can we trust these models and their predictions?
- Is the chosen KPI really the right one for this purpose?
- Is the underlying data quantity and quality good enough?

and so on.

All of these are valid questions and are, just as they were two years ago, fueled by another trend: Digitization. The engine for this is data. On top of that, Data Scientists are still following the same goal:

Giving understandable answers to questions by using data.

Despite all trends, this purpose stays the same and always will be one of the central pillars of doing Data Science.

But this is not the only trend which has remained or become even stronger. The most important, continuing phenomenon is the still massive hype caused by phrases like "Artificial Intelligence" and "Data Science". While these fields are incredibly valuable and powerful, discussions around them unfortunately often evoke false promises and skewed expectations, which in turn lead to disappointment. Some companies already started large, ambitious initiatives in the past, which led to underwhelming results, because expectations were set too high and timelines too short. For example, fully autonomous driving is one particularly challenging problem to solve.

Nevertheless, Artificial Intelligence remains the hope for many companies. Investors perceive it as a general purpose, technology that can be applied almost anywhere. The situation is comparable with the development during the nineties when all things related to the 'Internet' surged. Suddenly, every company needed a web page and significant investments were made to train web programmers. Nowadays, a similar thing happens with everything AI related. Again the investments into AI are enormous and we have a rush of courses on the topic. In the end, the development concerning the 'Internet' led to a vast ecosystem of companies and applications which influence the lives of billions of people in a profound way and it seems that AI follows a similar path.

This explains at least partly another noticeable trend: the further specialization of data science roles with names like "data translator" or "machine learning engineer." It is a somehow natural development as this is a sign that the field is getting more mature, but it also raises the risk of data science responsibilities being scattered across poorly coordinated organizations, and thus, not reaching its full potential. Chapter 14 and 17 go into this in further detail.

Finally, "Trustworthy AI" is emerging as another, highly important movement within Data Science. This is the field of research, which aims to tackle some previously unmet needs, like explainability or fairness. It is therefore included as one of the new chapters in this book (Chapter 18).

Given all these trends in Data Science, one of the reasons for founding the Vienna Data Science Group (VDSG) has become even more important over the last two years: to create a neutral place where interdisciplinary exchange of knowledge between all involved experts can take place internationally. We are still very much dedicated to the development of the entire Data Science ecosystem (education, certification, standardization, societal impact study, and so on), both across Europe and beyond.

A product of the exchange in our community can be found in the 2nd edition of this book, which has been vastly expanded to cover topics like AI (Chapter 9), Machine Learning (Chapter 8), Natural Language Processing (Chapter 10), Computer Vision (Chapter 11) or Modelling and Simulation (Chapter 12) in more depth. To follow our goal to educate society about Data Science and its impacts, a very relevant use case was included in Chapter 12: An agent-based COVID-19 model, which aims to give ideas about the potential impact of certain policies and their combination on the spread of the disease.

To provide our readers with a firm foundation, an introduction to the underlying mathematics (Chapter 6) and statistics (Chapter 7) used in Data Science has been included, and finished with a visualization section (Chapter 13).

Although a lot of content has been added, the goal of this book stays the same and has become even more relevant: to give a realistic picture of Data Science.

Because despite all trends, data science remains the same as well: an interdisciplinary science gathering a very heterogeneous crowd of specialists, which is made up of three major streams:

- Computer Science/IT
- Mathematics/Statistics
- Domain expertise in the industry in which Data Science is applied.

Science aims to generate new knowledge, and this is still used to

- improve existing business processes in a given company (Chapter 16)
- enable completely new business models

Data Science is here to stay and its direct and indirect impact on society is growing at a fast pace, as can be seen during the pandemic. In some areas a bit of disillusionment has set in, but this can be seen as a healthy development to counter the hype. Data Science team roles are becoming more differentiated, and more companies are putting Data Science projects into production.

So, Data Science has grown up and is entering a new era.

Fall 2021
Wolfgang Weidinger

Acknowledgments

We, the authors, would like to take this opportunity to express our sincere gratitude to our families and friends, who helped us to express our thoughts and insights in this book. Without their support and patience, this work would not have been possible.

A special thanks from all the authors goes to Katherine Munro, who contributed a chapter to this book and spent a tremendous amount of time and effort editing our manuscripts.

For my parents, who always said I could do anything. We never expected it would be a thing like this.
Katherine Munro

I'd like to thank my wife and the Vienna Data Science Group for their continuous support through my professional journey.
Zoltan C. Toth

When I think of the people who supported me most, I want to thank my parents, who have always believed in me no matter what and my partner Verena, who was very patient during the last months when I worked on this book. In addition I'm very grateful for the support and motivation I got from the people I met through the Vienna Data Science Group.
Wolfgang Weidinger

1 Introduction

"Data really powers everything that we do."
Jeff Weiner

Questions Answered in this Chapter:

- What makes Data Science, ML, AI and everything else closely connected to generate value out of data so fascinating?
- Why do organisations need a strategy to become data driven?
- What are some everyday use cases in the B2B or NGO world?
- How are data projects structured?
- What is the composition of a data team?

Data Science and related technologies have been the center of attention since 2010. Various changes in the ecosystem triggered this trend, such as

- significant advancements in processing a vast amount of unstructured data,
- substantial cost reduction of disk storage,
- the emergence of new data sources such as social media and sensor data.

The HBR called the data scientist the sexiest job of the 21st century while quoting Hal Varian from Google.[1] Strategy consultants declared data to be the new oil, and there have been occasional "data rushes" where "enthusiasts in data fever" mined new data sources for yet unknown treasures. This book explores data science and incorporates various views on the discipline.

[1] *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century*

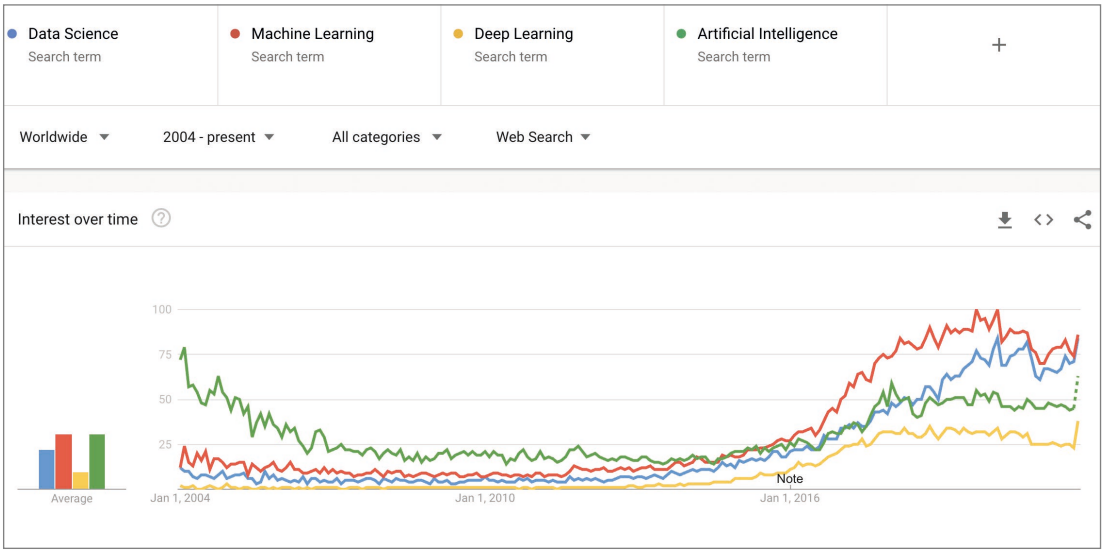


Figure 1.1 Data Science and related technologies on trends.google.com[2]

1.1 What are Data Science, Machine Learning and Artificial Intelligence?

There are many views on data science, and stakeholders in data science projects may give different answers to what they consider data science to be. Representatives address various aspects and may use different vocabulary since businesses and NGOs, for example, pursue different insights from data science applications. Perhaps the one common denominator is this: Everyone expects data science to deliver some value, which was not there before, with the help of data.

Table 1.1 Various views on Data Science

View	Description
Definition from Wikipedia	Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.[3]

[2] Data source: Google Trends *(https://www.google.com/trends)*
[3] *https://en.wikipedia.org/wiki/Data_science#*

View	Description
Application-centered view	We collect data and put this into pandas-data frames or data frames in R Studio. We also use tools such as TensorFlow or Keras. Our goal is to use these tools to explore the data.
Platform-oriented view	We create value from the data that we loaded on our SaaS platform in the cloud. Then, depending on the provided data and its structures, we store them in different storage containers, such as blob storage and distributed databases.
Evangelist-oriented view	Data science was the next big thing in 2015. Now, you should look at more specific applications. Looking at the Gartner charts, invest your time exploring cutting-edge trends such as neuromorphic hardware or augmented intelligence.
Management-oriented view	These are the ways of working to bring our company into the 21st century as a data-driven enterprise. During and after our transition, we will penetrate new markets and monetize data as a service.
Career-oriented view	As a senior data scientist at a major company, I can earn a six-digit yearly salary and explore interesting fields in corporate labs.
Use case-oriented view	Tell me your business problem, and we will tell you how we solved it for another customer. From fraud detection to customer retention to social network analysis, feel free to check out our catalog of possible analytics applications.
Entrepreneurial/ Optimistic view	Data Science is one way to change the world. Using Data Science, we can prevent climate change and fight poverty and hunger on a global scale.
Pessimist view	Data Science is one way to change the world. But, unfortunately, power-hungry people will use it to spy on us and suppress us. So Big Brother will be watching you.
Statistician's view	Data Science is just a buzzword. It is just another word for statistics. We might call it statistics on steroids, maybe. But in the end, it's just another marketing hype to create another buzzword to sell services to someone.

The essentials of data science lay in mathematics. Data scientists apply statistics to generate new knowledge from data. Besides using algorithms on data, a data scientist must understand the scientific process of exploring data, such as creating reproducible experiments and interpreting the results.

There are many different terms related to data science. For example, professionals talk about Artificial Intelligence, machine learning, or deep learning. Sometimes experts also talk about related terms such as analytics or business intelligence and simulation. In the following chapters, we will detail and highlight how we distinguish between analytics and data science. We will also highlight various data science applications, such as gaining insights into a text through Natural Language Processing or extracting objects from images via object recognition or modeling railway networks for optimal pathfinding.

Data Science as Part of a Cultural Shift

Suppose you apply for a job as a data scientist in a company. Imagine that, although it is unlikely you will get such an answer, the HR of this company rejects you because your astrology chart based on the data you have provided in your CV does not match the position.

Humans decide on what they believe is right. But, unfortunately, human judgment is flawed through bias[4], and we have mechanisms, such as confirmation bias, which assure us that we cannot err. For example, some people believe in the flat Earth theory or hollow Earth theory, which shows how powerful mechanisms such as confirmation bias can be.

For many of us, it would be disastrous to realize that a comfortable binary view of the world divided into black and white, good and evil, and right and wrong often does not work out. Modern sociological ideas such as constructivism[5] are more connected to data science than many think. The idea is that everyone constructs a reality based on their experience. Within the framework of "our reality", including its rules and conventions, we make decisions. According to studies, it is not uncommon that we are deeply convinced that we are right even if our choices are questionable to others. For example, suppose we have created mental models for ourselves in which we are confident that astrology must be correct. In that case, it is logical to assume zodiacs for personnel decisions will improve the hiring process. At the same time, people with strong religious beliefs might run into conflicts if they ignore what they might call signs or messages from God. Thanks to the biases mentioned above, our belief systems are often hammered into stone.

Data Science is not just a method to extract value from data; it also has the potential to be a method for making decisions that avoids or reduces human bias in the process. However, as will be shown in Chapter 18 on Trustworthy AI, data alone cannot solve the problem, because historical data and the model building process itself are often imbued with the very same biases. With that, business leaders can integrate data science and transparent and non-discriminatory practices, into corporate culture, and this will substantially impact the company's DNA. For example, a bias-aware company will adjust processes. Hiring a new employee is a good example. Many companies enlarge hiring teams that decide on the outcome of the candidate interviews in order to ensure that the bias of a single interviewer will not affect a hiring decision too much. In modern hiring processes, data science can be used to generate predictionsabout candidates to assist the decision-making process. If done with care, these model predictions can help to minimise biases in employment decisions.

[4] *https://www.weforum.org/agenda/2018/12/24-cognitive-biases-that-are-warping-your-perception-of-reality*
[5] *https://www.buffalo.edu/catt/develop/theory/constructivism.html*

In the beginning, every judgment is a theory. A theory is neither right nor wrong but inconclusive until it is proven or disproven.

Therefore, the positive effect of hiring personnel using astrological zodiacs would be nothing more than a theory. As long as we cannot prove that an astrological assessment would benefit a hiring process, the statement is inconclusive and, therefore, not recommended to use. Calling astrology inclusive rather than wrong might also make the discussion with believers in astrology less emotional.

Investigating the possible effects of astrology using data science is a perfect introduction to the environment we face in data science projects. Astrology claims to divine information about human affairs and terrestrial events by studying celestial objects' movements and relative positions. In a simplified version, astrology reduces everything to the sun sign, depending on birthdays. Using the simplified model, we could collect data on existing data scientists to determine a correlation between astrological signs and professions. In addition, we could collect the birthdate of a large pool of data scientists. As we need only a birth date and no other personal data, it would even be perfectly legal to collect these datasets from LinkedIn or any other data source containing data scientists' birthdates. Most of the analysis will consist of finding appropriate data sources, collecting the data from the data source, anonymizing it, and preparing it for examination.

Mathematics on the collected data will not leave much room for interpretation of results. Nevertheless, based on analysis, we would conclude a correlation between professions and astrological signs.

There is, however, a more complex form of astrology. Astrological charts include all planets and other celestial objects such as Lilith, the black moon, which does not exist in astronomy. In addition, many constellations are contradictory. An astrologer might call a person impulsive because of Venus or Lilith in Aries or passive because of Mars in Cancer. Finally, an astrologer might claim that readings require intuitive interpretations, which are, of course, not measurable.

Many data science projects might end with the assessment that there is insufficient data for a definite answer, and being unable to prove or disprove a theory might be unsatisfactory for many stakeholders. Yet, exploring data often helps bring clarity to the stakeholders, as at least many learn that achieving objective truth is not as easy as it seems. Therefore, we should be free to differ in personal or subjective beliefs and be cautious about things we cannot verify objectively. Of course, in the end, there is a good chance that we are right with our personal views if we have spent a lot of time exploring a specific field, even if we cannot prove it. Still, as long as we do not have enough data to prove something one way or another, it is a question of academic politeness to highlight inclusive outcomes because of insufficient data when talking with others.

Already as early as 2014, the New York Times[6] wrote about the 80/20 rule. This rule means that the team spends 80 % of their time finding and preparing data for data science projects and only 20 % on analytics. This number may vary enormously by industry. In addition to data modeling, we will also address the preparation and management of data in the chapters to follow. We aim to provide a compact introduction to data platforms and engineering.

In the second part of this book, we assume all the data is prepared and ready and focus on analytics. We will present several ways to generate value from data and cover essential topics such as neural networks and machine learning. We will also cover basics such as statistics.

The third and last part of the book is about the application of data science. Here we cover business topics and also address the subject of data protection.

Machine Learning and Deep Learning

Starting from Chapter 6, we will detail the differences between these frequently buzzed-about concepts. Still, as using these terms related to data science often creates confusion, we would like to outline them for you here.

In recent years, many companies prioritized processing vast amounts of data. Consequently, scientific processing, such as formulating the working hypothesis, was pushed into the background. Big Data tries to solve problems with a sufficiently large amount of computer power and data. This fact creates a productivity paradox: More data and better algorithms do not make us more productive; instead, the opposite is often true, as it becomes increasingly difficult to distinguish the signal from the noise. The signal is the information relevant to a question and thus contributes to answering it, while the noise is the irrelevant information.

We attempt to make these signals measurable in the scientific field by measuring he signal detection accuracy and how often algorithms find the signal. The quotient of both measurements expresses the algorithms' accuracy. We describe it as a percentage. A high F1 score means a precise answer, while values around 50 % represent a random result. So if an algorithm has an accuracy of, say, 90 %, it means that 90 % of all information is processed correctly.

This number may sound like a lot; however, data with a large volume is the norm in Big Data. For example, imagine we want to classify comments to find hate speech in social media. Let's say that 510,000 comments were posted per second on Facebook in 2018. Assuming that 10 % were classified incorrectly, we might fail to detect hate speech in 51,000 posts.

To avoid such a situation, deep learning, a group of machine learning algorithms based on neural networks, is currently being applied as an abstract solution to many problems. The advantage of deep learning over classical machine learning is that the former usually scales better with the amount of data and thus provides more accurate results and can be applied to various problems.

[6] *https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html*

The disadvantage of some methods in machine learning is that it can be challenging to interpret a prediction because the solution path is not immediately comprehensible. Furthermore, a statistically generated prediction may or may not be correct, as most models usually have less than 100 % accuracy. Additionally, we cannot use statistical forecasts to predict new data that has not been adequately analyzed or has limited usage. This statement may seem trivial, but it is essential since statistical analysis primarily depends on the input data and thus on the modeling skills of the data scientist. It is, therefore, necessary to interpret the result correctly and not to take it as truth.

An excellent example of this is numerical weather forecasts such as the weather report. We know fundamental physical laws in differential equations, but false predictions repeatedly occur due to non-existent or incorrect data or a simplified model. For example, a result of the solved differential equation can be: "Tomorrow the probability of rain is 10 %". Statistically, this means that we have created an analytical model based on historical data and based on all the data we have analysed, in 10 % of the cases with matching input data, it had rained. So 10 % can be a lot or very little; the important thing is to have an appropriate reference amount and relate it to the quantity obtained. In this case, it means that it is quite possible, although not likely, that it will rain tomorrow.

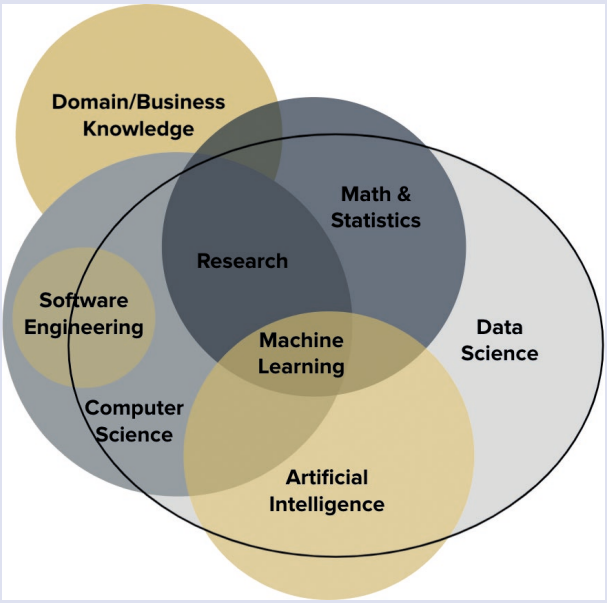


Figure 1.2 Differences *(https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5)*

Artificial Intelligence

When the common people think of AI, they might think of computers taking over the world such as in Terminator.

Artificial Intelligence is the simulation of human intelligence processes by machines, especially computer systems. There is an overlap between Machine Learning and Data Science, but AI can be seen still separated from both disciplines.

In Chapter 9, we explore Artificial Intelligence in detail. We explain the relation to Data Science and give a brief overview of the history of AI. We also discuss the problems that one may encounter when using Data Science skills to develop AI. In particular, we provide five pieces of advice: Be pragmatic, Make it easier for machines to learn through inductive biases, Perform analytics before creating AI architecture, Watch for the intelligence scaling trap and Watch for the generality trap. In this chapter you will have a chance to learn how to avoid mistakes and how to effectively use your Data Science tools to create AI solutions. After reading this chapter you will understand well where the limitations of AI technology are today and how to cope with those limitations.

1.2 Data Strategy

Some experts say that only companies with a data strategy have a future. We might agree or disagree with this assessment. However, everyone will admit that not every company feels the pressure to become data-driven. Many departments work mainly with pen and paper, in monopolies with no pressure to evolve or optimize processes. The figure below is just one of many models that can be found with web research to highlight different stages of a transformation from a non-data company to an entirely data-driven one.

As data maturity depends a lot on external pressure, companies often migrate in phases. When the competition gets more fierce, the market forces companies to innovate. However, the luxury to resist change due to market pressure can also lead to different forms of stress. Some monopolies face the problem sometimes that no vendor supports the legacy software they used for decades.

Introducing data science in organizations, no matter if it is a business, NGO, or governmental institute, starts in most cases with a mission statement. For example, for a global car manufacturer, a strategy could be formulated as follows:

> *"Our company aims to be the cost leader in the global supply chain by 2025. This measure enables us to bring electric mobility to the mass market with less cost than our competitors. For us to achieve this, we have to cut our supply chain costs by 20 %."*

Other companies simplify the strategy inspired by John F. Kennedy's speech on landing a man on the moon and returning him safely to the earth within a decade.

> *"Before this decade is out, all of our manufactured vehicles will be driverless."*

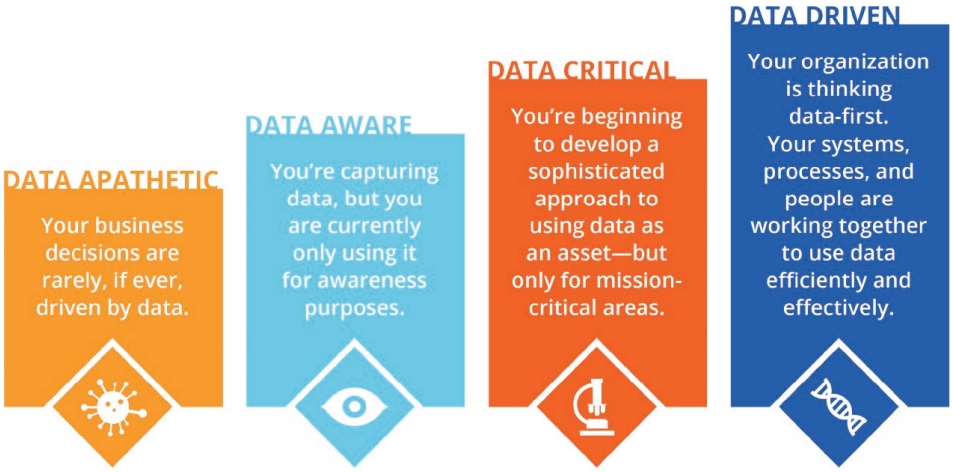


Figure 1.3 Data Maturity Model[7]

An NGO might have less profit-oriented but no less ambitious goals.

> *"With the help of our donors, we will use satellite images to explore dry areas in countries to find water points. Using that technology, we hope to be able to decrease the pain of gaining access to water in developing countries."*

The recommended practice for companies is to have an owner for data topics. Commonly this is the role of a Chief Data Officer, who needs to ensure that the company can realize its vision with the help of gaining insights from data.

Many companies have established processes to explore the past through business intelligence. For example, in maybe the most classic reference case, retail companies analyze how many products they have sold in the past. As a result, they can learn about which stores did a better or worse job. Based on the insights, leadership can then make changes such as replacing key personnel in poorly performing areas or creating additional incentives for growth in other areas.

Many companies have already reached a high level of optimization through traditional analytics. And it often seems as if conventional methods are at their limits.

Data science often helps to generate new knowledge. In other words, instead of using data science to sell more products, companies often use it to create new products. For example, while traditional analytical methods improve numbers, you get new numbers to work with through data science.

Once a CDO has proposed a strategy to meet the corporate goals, the board will approve the plan and allocate a budget. Using that strategy, the CDO then pools together the various department heads to realize the objective. Then, after a fit/gap analysis of the current situation, they will create hiring plans and plan projects to achieve their goals.

[7] *https://www.svds.com/thought-leadership/data-maturity-assessment/*

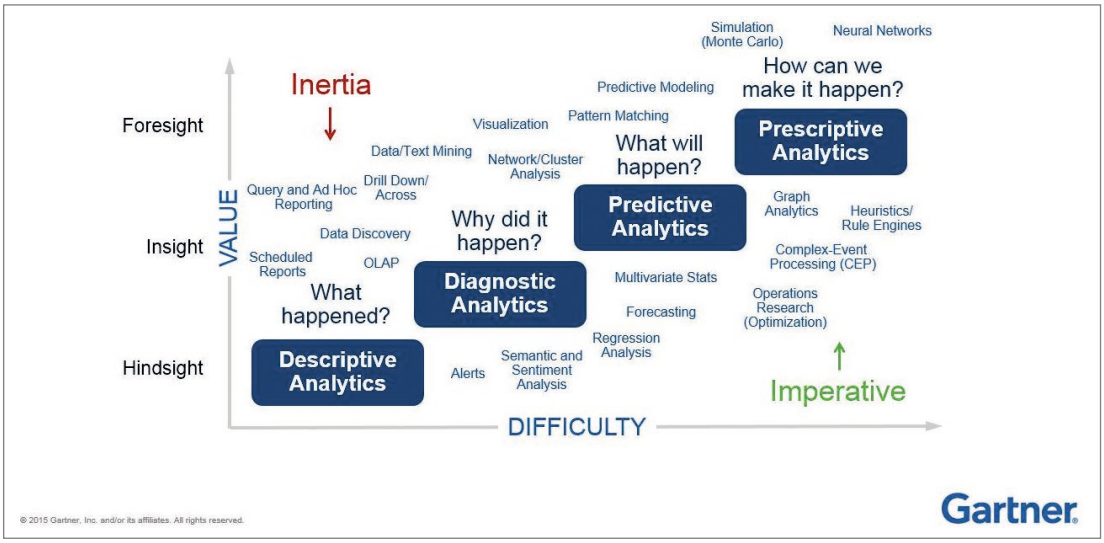


Figure 1.4 The Gartner Analytic Continuum (Source: *https://twitter.com/Doug_Laney/status/611172882882916352/photo/1*)

This position of business also clarifies the role of IT. The CIO is in charge of providing the necessary platforms to enable the teams of the CDO, but IT does not own the data science topic itself. Therefore, the CIO has to assess whether the current IT infrastructure meets the demand of the data strategy and if not, they must come up with a plan to create the required platforms.

1.3 From Strategy to Use Cases

Implementing a strategy defines how a business interprets data and the modeling based on it. Based on the strategy, a company can decide which questions the data scientists must answer. Based on these questions, solution architects can design platforms to host the data and data engineers can determine from which data sources they have to extract data.

Most companies have cross-functional teams for data science projects. They work in an agile team to explore new use cases and methods to apply data science.

Without qualified professionals, a company cannot even begin to implement its ambitious plans. Therefore, we want to look first at how data teams could appear from the project's view. In a corporate world, many of these team members would report to a different department.

1.3.1 Data Teams

We need data professionals to implement a data science strategy or to build up a data-driven start-up. There are two groups of experts in the data world that have evolved.

The first group, people with a statistical background, usually have academic experience and create models to answer the departments' questions. The second group consists of people with an engineering background. They are responsible for fully automating data loading onto the platform and continuously running the developed models' data in the production environment.

In organizations, these two groups have different reporting lines: Business and IT. In most companies, data agendas are a part of the top management board. Therefore, data is associated with the business. Some companies establish the role of a CDO, who directly reports to the CEO and the board. Others create a position, such as Head of Data or Head of Data Science. The authors of this book believe that data should be part of the board. Therefore, we refer to the CDO as the ultimate leader of all data agendas, whereas we refer to the CIO as the position accountable for all IT agendas.

In Figure 1.5 we have many models to describe the different roles per activity and department. Please be aware that we do not cover all roles in detail in this chapter. We cover this topic in more detail in Chapter 14.

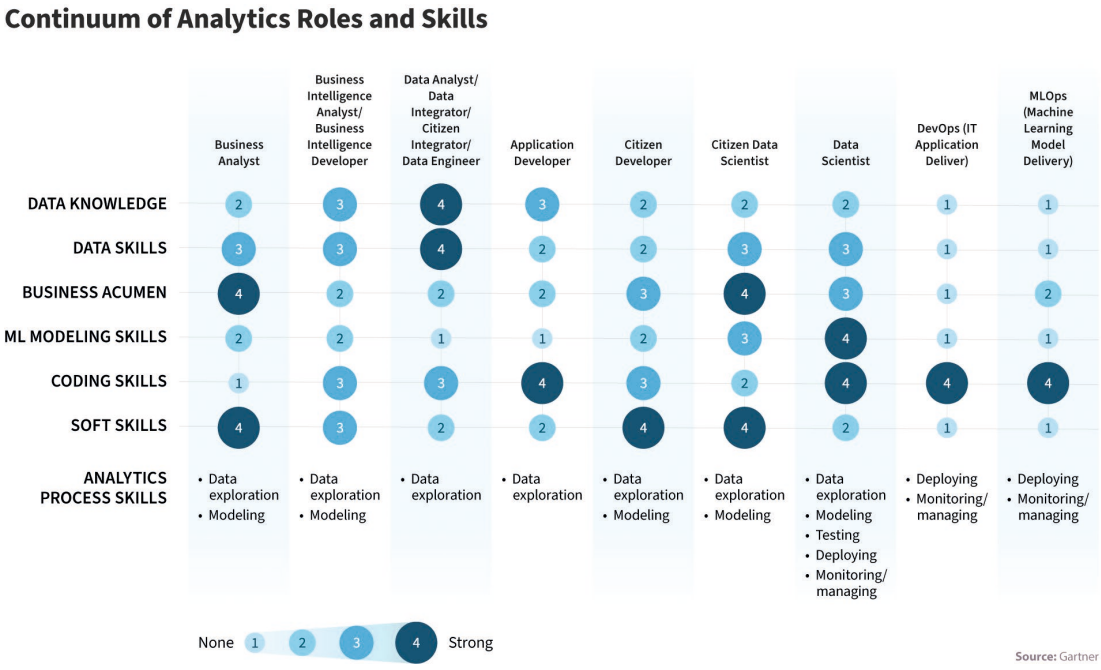


Figure 1.5 Role distribution in data programs (Source: *https://nix-united.com/blog/data-science-team-structure-roles-and-responsibilities*)

1.3.1.1 Subject Matter Expert (Domain Expert)

The SME is an essential person for a data project. Still, this person is often not shown in data teams. A subject matter expert understands, from the inside out, how the company provides its services to its clients inwards-out. They are often also referred to as a domain expert.

An SME is someone who has been performing a day-to-day job for a long time. For example, in a retail organization, a perfect SME might be the person who has been working in a supermarket in different roles for multiple years. They have seen almost every possible imaginable scenario and have a good gut feeling about what clients want. They might also find potential side effects to changes that no one without experience in the field could see.

In some industries, the role of an SME overlaps with an analyst. Finance is a good example. A credit analyst takes all data from a client who applies for credit and calculates the credit risk using a given formula. Unlike data scientists, analysts do not generate new knowledge. However, analysts work with numbers and have a deeper understanding than other types of SMEs.

In an NGO, an SME might be a development worker who fights poverty and plagues in developing countries or works in refugee camps. Therefore, an NGO SME might have a completely different view of what is missing on-site and feasible than those who watch the situation remotely.

SMEs are also often natural authorities in their fields due to their long-term experience. If, for example, a company wants to install a new IT system or new processes on-site, the support of SMEs can be crucial for its successful deployment, as less experienced employees in the field often look up to them.

The actual duties of the SME, therefore, depend on the area of operation but generally include the following activities:

- Provide insights on the existing challenges
- Provide access to possible data sources
- Help formulate goals
- Assist in the release of products and verify their successful outcome
- Guide users towards adopting the new system.

1.3.1.2 Business Analyst

Many projects need a business analyst who acts as a bridge between SMEs and data scientists. The critical skill of a business analyst is to ask the right questions. His job is to find out which activities make sense from a business perspective.

In start-ups, a business analyst helps to formulate the business plan and the value proposition. First, he needs to underline how the business can make profits and measure if we are successful.

Business analysts, therefore, are dedicated to their time to the following activities.

- Write business plans
- Analyze business requirements
- Translate business requirements into work packages for the data team

1.3.1.3 Data Scientist

There is a debate about how much statistics a data scientist should understand. Purists claim that you can be only a "real data scientist" if you have a Ph. D. and are acquainted with scientific methods and statistics in and out. They sometimes call everyone else "fake data scientists."

Many modern views differ and see a data scientist as an expert who puts the data into use and creates something new. For example, she can discover a new relationship in the data and build models. It is essential to highlight that good communication and programming skills are helpful to achieve this.

Data scientists should be as versatile as the data they are working with and open to learn about new domains and to collaborate with experts from different fields. For example, working with and analysing imaging data requires specific knowledge in Computer Vision, image processing, machine learning and also specific domain knowledge of differential geometry or medicine. It is important to understand how data are acquired, which false interpretations are possible and also if an expert is required to create a baseline or to evaluate the designed models (for example annotations of a specific tumor tissue in an computer tomography scan by a medical doctor). In Chapter 11 you will get a deeper insight into the field of Computer Vision and how to work with imaging data as a data scientist.

All in all, every data scientist will have some form of understanding of science and statistics. But similar to many examples of autodidactic programmers, who have not studied Software Engineering, many things can be self-taught. It is often the case that a data scientist team consists of people with a diverse skill set. While some of the members are top-notch mathematicians, others complement them with more communication or programming skills but still contribute as much to the outcome as others.

Mathematics and Statistics

Mathematics and statistics are still the basis of everything we do. Therefore, we dedicate Chapters 5 and 6 to the topics to recap the basics of probability theory, explain a confidence interval, and say that one idea is correct or not mathematically.

The main tasks of data scientists are exciting, sometimes challenging, and highly diverse.

- First, we must prepare our data, often liaising with other departments, such as information systems, and harmonizing various data sources. In many organizations, this is the job of the Data Engineer, especially if these steps need to be automated and have strong SLA requirements.
- Then we engage in exploratory statistical analyses, interpret the results, and use these to gain domain knowledge and conduct further preliminary data investigations.
- Based on these findings, we curate a data set and feed this to a machine learning algorithm, such as those mentioned above, to build a model for a specific task.
- The trained model is tested and fine-tuned to the point where we can use it productively: its outputs, which usually take the form of predictions of a particular output given an unseen test case, will be acted upon by the data science team and other stakeholders in the company.

Of course, this process is not a one-time effort. Data and models must be continuously monitored (and often, continuously retrained) to ensure performance remains at an acceptable level. New research projects must be undertaken based on the company's innovation roadmaps, triggering this process to begin again. We can answer business questions through data, and progress and results must be communicated to various departments, often in sophisticated visualizations and presentations (see Chapter 13, 'Visualisation').

We will describe a lot more about the job of data scientists throughout this book. Data scientists play an essential role in the development of AI solutions (see Chapter 9), but also in the domain of modeling and simulation (see Chapter 12)

1.3.1.4 Data Engineer

Data engineers build and optimize data platforms so that data scientists and analysts have access to the appropriate data. In addition, they load data into the data platform according to the policy set by the architect.

Data engineers implement this activity using data pipelines, load data from third-party systems, transform the data and then store it on the platform. A data pipeline must scale with increasing data volumes and be robust. Therefore, the pipeline must have corresponding fault tolerance. It thus forms the foundation that data scientists and analysts can use to generate knowledge.

Unlike other team members, data engineers must have solid programming skills. Most importantly, a data engineer needs to understand the principles of distributed computation and how to write code that can scale. Thus, the data engineer has a fundamental role in every data science team.

Core activities include:

- Building various interfaces to enable the reading and writing of data
- Integrating internal or external data into existing pipelines
- Applying data transformations to create analytical datasets
- Monitoring and optimization to ensure the continuous quality of the system (and to improve it if necessary)
- Developing a loading framework to load data efficiently

1.3.1.5 DevOps

DevOps is a role that requires a mixture of developer and operational skills. Their task is to operate the data platform upon which the data engineers and data scientists work.

DevOps implement the architectural design for a project or system and address the change requests made by the Data Engineers. With the emergence of cloud systems, DevOps engineers gained popularity and have become a scarce resource in many projects.

Their activities include:

- The scaling of data platforms
- Identification of performance problems in the software
- Automating redeployments

- Monitoring and logging applications
- Identifying resource bottlenecks and problems
- Remediation of issues that occur due to system operations

1.3.1.6 Solution Architect

In the end, someone has to be accountable for everything running smoothly. Only then can the data scientists do their job, and the users can create business value by using the applications developed during the data strategy implementation. In large organizations, this is the solution architect.

Someone must ensure that the proper hardware infrastructure is in place, that the appropriate data management, selecting processing software, can protect data against misuse and theft, and finally, that data scientists and end-users of a system can do their work.

Many organizations have multiple roles for that:

- A **data architect** focuses on data and how data is stored. In addition, she cares about metadata management and the definition of processes to load data into data management software such as databases or object stores.
- A **systems or infrastructure architect** focuses on servers and hardware and ensures the hardware is available. If the company hosts the solution in the cloud, they refer to this role as a 'cloud architect.'
- A **data steward** or **data manager** is responsible for ensuring that the project follows the appropriate corporate policies.
- A **security architect** protects the system against hackers and other intrusion attempts.

In reality, it is hard to isolate those various engineering roles. A data platform must serve multiple purposes and meet multiple functional and non-functional requirements. Without knowing the software, one cannot make a hardware decision, and numerous data platforms have specific hardware requirements. Therefore, there needs to be a generalist who understands everything and can lead other architects to make cost-effective, scalable, robust, and fast solutions.

In large companies, a CIO leads all streams to create standards for every project. Large companies have their frameworks or business units to provide platforms to other departments. A solution architect must often also consider corporate politics as another factor in building the best platform for their project. There are usually fewer restrictions in small companies and more chances to fail with a wrong strategy. Chapter 17, 'Mindset and Community', also explores a risk known as the 'swiss army knife', that might apply to a solutions architect in a small company: Many small companies end up with one person being the single expert for multiple engineering domains.

In many organizations, it often boils down to a situation in which one person with a diverse skill set and broad knowledge is fully accountable for realizing the solution. Although they might be able to delegate responsibilities, depending on the size of a project or company, they ultimately still have to cover multiple roles in other scenarios and thus become a bottleneck.

Typical tasks of a solution architect are:

- As an accountable person for the solution, decide about all parameters or lead the decision-making process. All parameters include, among other things, hardware, operating systems, data management software, data processing, user experience, scalability, and cost-effectiveness.
- Ensure that the project meets all requirements, and the project team has all requirements to build the solution for the ultimate end-users.
- Lead other architects and engineers to implement the solution.
- Ensure that all solutions meet corporate standards for all projects, such as data protection standards.

1.3.1.7 Other Roles

We have not covered BI engineers and Business data owners here. Often in agile teams, we add a Scrum Master to the team.

We will outline in Chapter 16 that data teams might face quite different requirements in different industries. Also, small companies or start-ups have other needs than large enterprises. This diversity means there is no unique definition of how a data team has to be structured. Various roles will exist in one team but not in others.

Data teams in large organizations, especially with regulatory requirements, will incorporate roles such as data managers, security experts, data stewards and more.

1.3.1.8 Team Building

The structure of the team and the operating model depends much on the company's data maturity level. In many cases, some team members have to clear out old legacy systems before creating something new. In some companies, leaders assign individuals to multiple teams.

The success of teams also depends a lot on the corporate culture. We will go more into details on this in Chapter 17, 'Mindset and Community.' Setting up a data-driven organization is the focus of Chapter 14.

1.3.2 Data and Platforms

Company data currently exists in most companies horizontally, in different departments, or vertically, which is fragmented and coupled to various functions and silos. In addition, the proportion of critical information generated outside the usual processes is growing. So then, part of a data strategy is to create a process that can handle various data formats and convert them into a structured and processable format. In this process, we can explore four different properties:

- **Volume:** Describes the amount of data collected through organizations through daily business processes. Volume is an order of magnitudes, such as gigabyte, terabyte, or petabyte.

- **Velocity:** Describes the speed of the data generated during a session or transaction. Sensor data typically has a very high velocity since it often has to be processed immediately. For example, if you detect problems in production with sensors, you want to react within a few minutes.
- **Veracity:** This value describes the trustworthiness or accuracy of the data. For example, we can use data lineage to trace the individual data processing steps and streams. The corresponding signature mechanisms can be confidence-increasing measures. In addition, we can include a watermark to identify which user opened the documents the last time.
- **Variety:** Describes the diverse data formats and data types on a platform. For example, an entire platform must process different data, such as voice data or text data. In addition, it must also have standard connectors to the individual interfaces used in the company to provide the required data efficiently.

Data and Platforms also contain best practices on automating the build of platforms in the cloud or on-premise. There are many non-functional requirements such as durability or availability. One key responsibility of an architect is to find a solution that meets all those requirements.

Chapters 2 to 5 cover infrastructure and data architecture, data engineering, and data management. In Chapter 2, we will look at topologies and hardware. This investigation also includes DevOps-related aspects on how to build data platforms.

In Chapter 3, we look at data architectures in general. We look at solutions on how to store data so that we can access them efficiently. While in Chapter 2, we looked at which hardware we need and how we can automate the creation of platforms, Chapter 3 is about which application platforms provide software to process data in the best possible way.

Chapter 4 covers essential aspects of how to engineer data. Or, more specifically, how to extract data from sources and load it on the platform.

Chapter 11 outlines the necessary routines on how to acquire, store, compress, reconstruct and to process imaging data and their incorporation into Computer Vision systems.

The engineering-related part of this book closes with coverage of data governance in Chapter 5, in which we learn how to set up corporate policies on how to manage data.

1.3.3 Modeling and Analysis

Raw data itself does not yet generate value. Instead, analytical models or algorithms are needed to extract value from the data. Typical use cases are optimization tasks, predictions for the next accounting period, or risk classifications. Thus, leaders must plan how analytical resources, such as personnel or hardware, can be used most efficiently and effectively and thereby have the most potential to generate value.

This initiative must also include automated machine learning model retraining and adaptation to new values still unknown to the system. A possible solution would be, for example, to consider any problems already during the system design or to react to them appropriately during operation.

It is essential to keep the entire end-to-end process in mind, including the users, the complexity of the solution, and the management of the models and the data itself.

As modeling is central to the success of a data strategy, we will explore this already here in more detail.

We will briefly discuss explainable AI strategies on how to make complex computational models interpretable and their decisions and estimations understandable for users. In Chapter 11 explainable AI strategies in Computer Vision for image based models are introduced and it is briefly explained how these can help us humans see what computers and underlying neural networks see.

1.4 Use Case Implementation

Implementing a data science strategy consists of various phases. How the stages look in detail depends on the circumstances. For example, some companies have invested in data platforms, and they have built a scalable data platform that allows the company to deploy new use cases quickly. Other companies have not invested in infrastructure at all and start entirely from scratch.

Figure 1.6 outlines a holistic process that explores the creation of a platform and investigating use cases.

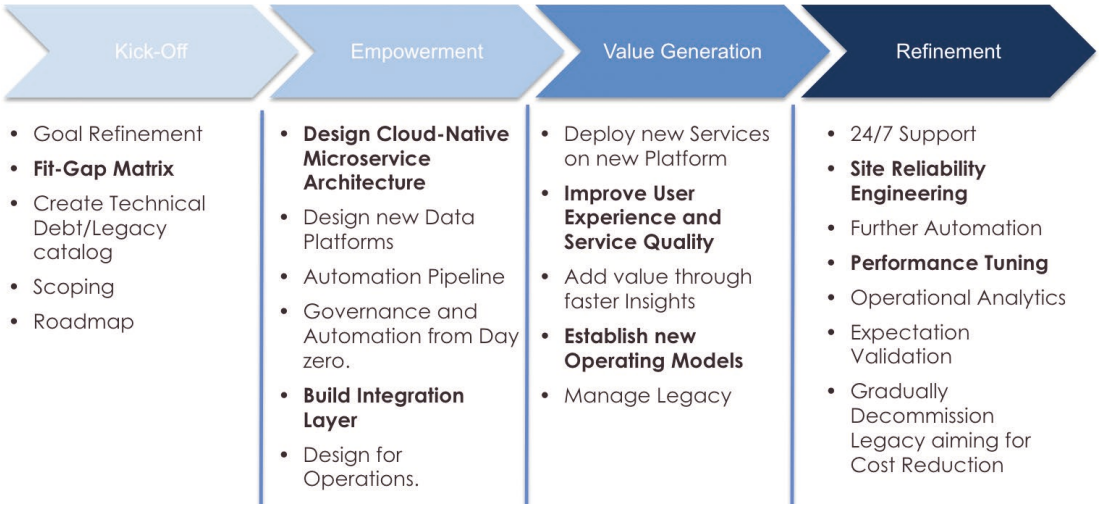


Figure 1.6 Data platform (provided by *AlphaZetta.ai*)

1.4.1 Iterative Exploration of Use Cases

Data teams can combine two approaches to explore which use cases are most beneficial for business goals.

In the **conceptual approach**, the team aims to find as many new use cases as possible to exploit. Open innovation or design thinking can keep the process as flexible as possible and promote creativity: Design thinking describes finding solutions from the end user's perspective, while open innovation significantly increases the solution space.

In this phase, the team enters a cycle of validation and verification to rule out possible dead ends and non-actionable goals at an early stage.

In the **data-driven approach**, the team explores existing data sources gradually to generate new value. For example, a team can organize workshops such as hackathons to expand the understanding of data. This measure helps with an initial analysis of data sources but is less suitable as knowledge of the data increases.

Independent of the chosen approach, the use case must also undergo a profitability check to determine whether it contributes business value. One way to achieve this is that, after teams digitize existing processes at the beginning of a project, they calculate the costs saved through this measure.

The process starts with a workshop where data and domain experts from different departments and business units determine the potential for optimization. The result of the workshop is a **use case list** that gives an overview of possible explorations. Finally, each item in the list is quantified to determine the feasibility of each use case.

Agile methods such as Scrum are an advantage here. They allow an iterative approach to changing the goal within the development, for example, to align it with new business goals.

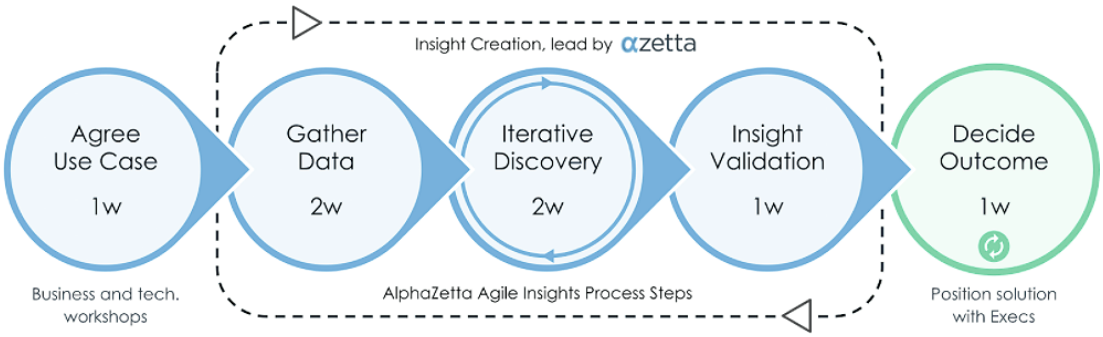


Figure 1.7 Analytical process (provided by *AlphaZetta.ai*)

Figure 1.7 presents an analytical process. Here we can assume that the necessary platform already exists. The analytical process runs as follows:

Table 1.2 Data Science Process

Phase	Description
Agree on Use Case	In this phase, the analytics team agrees on a use case. As a rule, we involve an SME, as she knows which topic is most important for the company.
Gather Data	The team spends two weeks collecting data and reviewing various sources. At this point, the topic of data governance is also essential.
Iterative Discovery	In this phase, data scientists analyze data and gain new insights.
Insight Validation	Once the results are in, you may go back to the Agree on Use Case phase because you came across new insights and data sources during the analysis.
Decide Outcome	In the last step, data scientists evaluate the result, and the team decides whether the project is likely to be a success or failure.

In a positive evaluation, a new phase can follow in which the team brings a use case into production. Of course, the effort required to accomplish this may vary based on the use case complexity. However, the experiences during the discovery phase should indicate what to expect in that step.

Inconclusive Results

The science in Data Science highlights that we deal with a research-oriented environment. If there were no risk to fail, it is not science. Many organizations setting up data science will also have to learn that there are many reasons why some questions will stay at least partially unanswered.

One reason can be legal problems. For example, sometimes, the team learns that the data they need to solve a problem exists, but they do not have the right to use it. Besides data ownership, privacy protection is a huge topic. If the analysis of various data violated regulations such as GDPR, teams have to stop their investigations. We deal with legal issues in Chapter 15.

Legal problems often go hand in hand with internal misalignments. Especially in a large organization, dealing with governance might delay many projects as settling who owns what might take a while. Discussions on a management level about data projects may result in, for outsiders, intransparent corporate politics. Data professionals often end up waiting for management decisions while the real issues the decision-makers are arguing about are unclear. In historically grown organizations, it often might be just a feud between two or more managers. According to the law of triviality, managers are likely to hack their skulls over unessential details.[8] Chapter 17 addresses how to deal with political topics that may be to a particular extent, also called dealing with irrational behavior.

[8] *https://docs.freebsd.org/en/books/faq/#misc*

In many cases, data itself is the problem. The data is available, but the data quality is not sufficient. At other times, the teams might learn that other companies could apply a use case, but that data does not exist for different organizations. Sometimes, an organization also has to realize that the data coming from a supplier is not detailed enough. Negotiating with the supplier to make the data more precise can also be a question of costs. There might also be independent third-party providers of data who sell data at high prices.

Costs, in general, can be a huge issue. For example, when exploring a use case, the team might realize that they need far more computation power or maybe far more engineers to maintain a platform than expected. In addition, managers always have to assess over time if answering specific questions with data is still justified. So, if we learned that it costs, for example, ten times more than expected to solve a problem with data, the value we would gain might not match up.

In general, it is essential to document results in a corporate knowledge base such as Confluence or any other extensive documentation system. An unresolvable problem that exists today might disappear in the future. If the results are not properly documented, a future team might not know that it would make sense to continue past work.

1.4.2 End-to-End Data Processing

In an analysis phase, the data scientists present results to decision-makers and stakeholders and recommend further analysis of additional data sources. In most cases, it is essential that end users can use the analysis results directly, such as in self-service BI tools.

How many results users will leverage also depends on non-technical parameters, such as comprehensibility and usability of the user interface. Some end users might stop using the systems if the data is there but difficult to access. The old UI/UX adage “don’t make me think” is still valid, and the more intuitive the handling of applications is, the better.

We can also connect User Experience to the performance and availability of applications. Chapter 3 explains how to design scalable and cost-efficient systems and address various aspects such as when to use on-premise solutions vs. cloud-based solutions.

The implementation of a data strategy includes the following considerations:

- Every business process is a service. The aggregation and analysis of data from all services is the basis for business decisions. These insights help to optimize the service itself and to eliminate weaknesses.
- Companies use a digital channel to communicate with end-users, other services, or partners. In addition, a digital channel allows the application of analytical models.
- Through a data platform, processes and services are managed automatically and are thus always available to the relevant end-user or business unit.

1.4.3 Data Products

Data Pipelines collect data from data sources such as sensors on the physical product (IoT) or social media applications. In the process, analysts can gain insights into how end-users consume a service and use a product. To increase this efficiency of data collection, engineers can transform a product into a "smart product", which can describe its physical state by being equipped with sensors.

Smart products provide additional value to their original value. For example, a smart fridge will still cool the goods inside, but at the same time, this device can notify its owner that certain products are about to run out or expire. As another example, a forklift truck with sensors can autonomously navigate specific locations, react to its environment via sensor data, and avoid accidents.

Product designers can learn how user groups handle a product by integrating end-users and optimizing future versions for various user profiles.

1.5 Real-Life Use Case Examples

After looking at strategy, it is time to explore what companies can concretely do with data science. Chapter 16, 'AI in Different Industries', explores how we can realize applications in specific industries. In that spirit, the following sections describe how a strategy for optimizing the value chain within a manufacturing company might look.

1.5.1 Value Chain Digitization (VCD)

Analytical methods offer application possibilities in production. Examples are supply chain optimization and predictive maintenance.

Supply Chain Optimization

A supply chain of an organization is usually globalized. Different goods and preliminary products are manufactured, produced, or assembled in various locations. By distributing them across several countries, companies can reduce their costs, but at the same time, the number of critical dependencies, and the risk of supply chain interruption, increase. For example, a manufacturer cannot meet their desired production quantities if a supplier fails or if there is a shortage of raw material.

Efficient supply chain management offers a strategic advantage that can accelerate a company's growth. Those who produce products better and faster than their competitors will survive in the market.

Machine learning applications can help to identify problems proactively. For example, if we have access to historical data, forecasting models use neural networks to predict labor strikes or weather problems. Likewise, by simulating a supply chain, various risks can be

counteracted, which will subsequently fuel just-in-time production. There is a practical example later in Chapter 16.

Predictive Maintenance

Predictive maintenance is about attempting to predict and minimize machine downtime by analyzing machine data proactively. As a prerequisite for this, sensors are needed that can transmit corresponding data. Data pipelines collect this data primarily from machine sensors, like heat or pressure sensors. For example, one possible result may be the realization that a group of components tends to have a higher defect rate above 100 °C in production. Thus, a production manager can minimize scrap by regulating the temperature. However, even external data, such as air pressure or humidity in the area, are also interesting, as data scientists can incorporate them into the analytical model.

The goal of predictive maintenance is also to send the service employee to fix a fault before it even occurs. This measure leads to reduced costs and increases the quality and availability of services.

How to efficiently collect and analyze mass sensor data is also described later in Chapter 16.

1.5.2 Marketing Segment Analytics

This category measures the effectiveness of marketing efforts and relates them to competitors. The goal is to determine the optimal mix of marketing measures and incentives for specific customer segments to retain and win existing and new customers.

Data Scientists can group data into clusters, such as per geo regions or micro and macro trends. In addition, to learn more about historical data, a good practice is to store data on unsuccessful marketing attempts or processes. In total, a data-driven approach increases the efficiency per marketing campaign and documents which marketing tools, in which situations, have delivered the most success. Scenario Supposelanning, performance, and social media optimization will add even further value.

Suppose campaign managers contrast this data with data collected on competitors' campaigns. In that case, new insights may emerge that can be recycled in future marketing efforts to highlight new and unique selling points.

As a real-life example, obtaining and processing data from various commerce channels through an omnichannel strategy is presented later in Chapter 16.

1.5.3 360° View of the Customer

In this discipline, data scientists analyze customer behavior to understand their needs better. Identifying a "*customer journey*" involves examining historical data and offers insights into factors such as which products customers have consumed, which services they have obtained, and where there were complaints. In automated processes, we can use a *recommender engine* to make a customer the right offer for a service or product at the right time.

Another step is integrating daily data, such as the interaction in the "buy & sell" cycle, and determining the customer's satisfaction with the company by analyzing reviews, inquiries and complaints, and even social media information such as Twitter tweets about the company (see Chapter 10 for a deep dive into how Natural Language Processing makes this possible). Understanding how customers are contacting the company and what they are saying makes it possible to improve services, generate ideas for new products that customers are demanding, and predict possible future interactions and upsell and cross-sell opportunities.

The top priority is to retain the existing customer base, that is, to maximize *customer retention*, and at the same time to advance into new segments. The basis for this can be internal (CRM, ERP …) and external (social media) data, which are stored in different data silos on a central platform and evaluated with analyses.

We can use *predictive analytics* to help to identify consumer behavior patterns and evaluate them. For example, it is possible to determine when a customer leaves the sales process and why. This insight is the basis for fine-tuning marketing efforts and optimizing all stages of the consumer journey, for example, from improving the in-store shopping experience to revamping the company's online shop.

Also in the medical sector a 360° view of the customer (for example a patient) is highly demanded. The analysis, structured and secure storage of daily routine medical imaging and record data are the profound baseline for medical analysis, patient specific modelling and the prediction of a patient's outcome. The more information we have on a patient, the more it can contribute to solving the puzzle of finding the right diagnosis, assessing the therapy response and possible adaptations.

1.5.4 NGO and Sustainability Use Cases

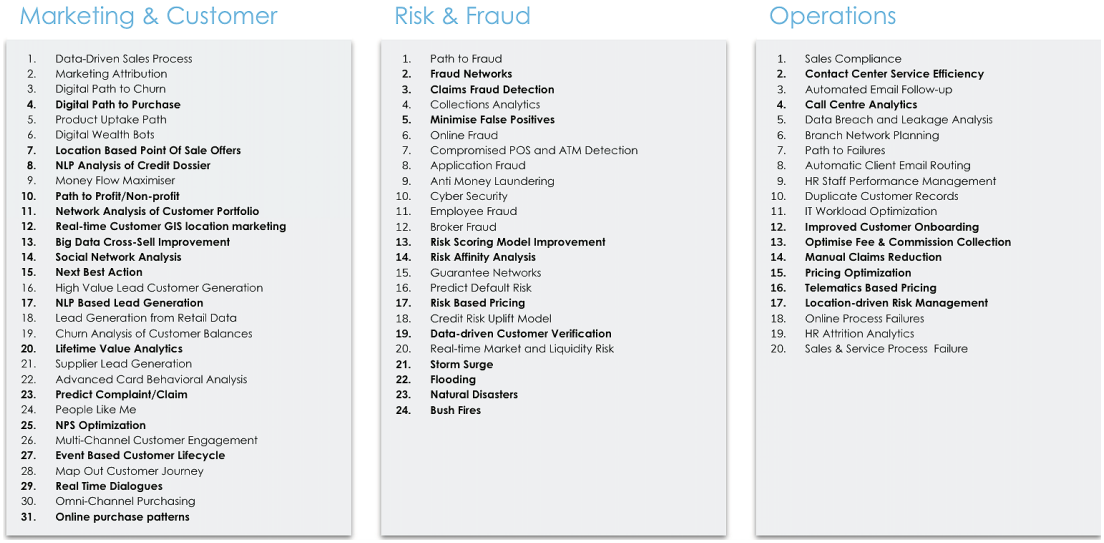
The aforementioned use cases come from the business world and may give the wrong impression that data science is mostly about using data to reduce costs or increase revenues. However, we can apply data science to support humanitarian causes.

Optimizing a supply chain might also help to improve the logistics of an NGO. And the work of many philanthropists and endowments for NGOs is built upon using data for the greater good. For example, NGOs may use satellite images and machine learning algorithms to detect possible water points in dry areas. Or, they may use data to forecast people's possible movements and predict the number of refugees trying to leave their homelands.

The Paris Agreement signed in 2016 highlights various sustainability goals for the world. However, we are still far behind our plan to reduce carbon emissions. There are many ways to use analytical models to forecast CO2 emissions and explore ways to reduce them. Chapter 16 explores some of these in discussing the future of data science and AI in the energy sector.

1.6 Delivering Results

The following image is part of a use case library, representing possible use cases that data scientists could explore within a company.



Marketing & Customer

1. Data-Driven Sales Process
2. Marketing Attribution
3. Digital Path to Churn
4. **Digital Path to Purchase**
5. Product Uptake Path
6. Digital Wealth Bots
7. **Location Based Point Of Sale Offers**
8. **NLP Analysis of Credit Dossier**
9. Money Flow Maximiser
10. **Path to Profit/Non-profit**
11. **Network Analysis of Customer Portfolio**
12. **Real-time Customer GIS location marketing**
13. **Big Data Cross-Sell Improvement**
14. **Social Network Analysis**
15. **Next Best Action**
16. High Value Lead Customer Generation
17. **NLP Based Lead Generation**
18. Lead Generation from Retail Data
19. Churn Analysis of Customer Balances
20. **Lifetime Value Analytics**
21. Supplier Lead Generation
22. Advanced Card Behavioral Analysis
23. **Predict Complaint/Claim**
24. People Like Me
25. **NPS Optimization**
26. Multi-Channel Customer Engagement
27. **Event Based Customer Lifecycle**
28. Map Out Customer Journey
29. **Real Time Dialogues**
30. Omni-Channel Purchasing
31. **Online purchase patterns**

Risk & Fraud

1. Path to Fraud
2. **Fraud Networks**
3. **Claims Fraud Detection**
4. Collections Analytics
5. **Minimise False Positives**
6. Online Fraud
7. Compromised POS and ATM Detection
8. Application Fraud
9. Anti Money Laundering
10. Cyber Security
11. Employee Fraud
12. Broker Fraud
13. **Risk Scoring Model Improvement**
14. **Risk Affinity Analysis**
15. Guarantee Networks
16. Predict Default Risk
17. **Risk Based Pricing**
18. Credit Risk Uplift Model
19. **Data-driven Customer Verification**
20. Real-time Market and Liquidity Risk
21. **Storm Surge**
22. **Flooding**
23. **Natural Disasters**
24. **Bush Fires**

Operations

1. Sales Compliance
2. **Contact Center Service Efficiency**
3. Automated Email Follow-up
4. **Call Centre Analytics**
5. Data Breach and Leakage Analysis
6. Branch Network Planning
7. Path to Failures
8. Automatic Client Email Routing
9. HR Staff Performance Management
10. Duplicate Customer Records
11. IT Workload Optimization
12. **Improved Customer Onboarding**
13. **Optimise Fee & Commission Collection**
14. **Manual Claims Reduction**
15. **Pricing Optimization**
16. **Telematics Based Pricing**
17. **Location-driven Risk Management**
18. Online Process Failures
19. HR Attrition Analytics
20. Sales & Service Process Failure

Figure 1.8 Index of a Use Case Library (Source: provided by *AlphaZetta.ai*)

Based on each item's title, we can approximately assume what a specific use case does. For example, in 'Digital Path to Purchase', we explore how people use an e-commerce system and interact with users who bought products versus those who did not. Through this, we might discover issues such as a confusing order page on our website, which caused users to abandon their purchases.

We can validate the success of some of the methods listed above, using specific metrics like an increased *return on investment* and a reduced *time to market*. But things get tricky if no such measure exists for our new data science project. How do we measure success, then?

Especially in the beginning, it is almost impossible to measure benefits in numbers. For example, let us assume a company implements its first data science use case. It has to hire a team of people and build up infrastructure, which often involves a steep learning curve. Maybe the investment costs exceed the possible revenues of that use case. But what if the team starts working on a second project afterward and uses all the artifacts created in the first use case?

Departments generate value from their data by solving everyday problems. This measure leads to various prototypes or initiatives that we can turn into a "*digital investment*." A corresponding governance strategy is also necessary to document the data processing steps required to generate the maximum benefit from data and thus the value chain and make the applied actions traceable. In addition, governance increases the level of trustworthiness of the platform, as it is now apparent where the data comes from and where it is stored. Therefore, data governance is also the subject of Chapter 5 in this book.

A Blackbox View

Looking at Data Science from the outside, success depends on asking the right questions and giving the answers that satisfy the audience. Asking the right questions is often not easy. It is often also a question of granularity. However, whether you are a CDO for a government, an enterprise, or an NGO, the question will always be about bringing more value to your clients or improving while you generate value.

Some might oversee that answering questions is a unique field as well. If we have all the perfect answers, but if we cannot communicate them appropriately to our audience, all our work might be in vain in the worst case. Presenting results is a lot about personal development and the ability to communicate. Speech trainers help to accentuate and understand how to say the right things at the right time. Especially people with a technical background are not always aware of the potential benefits of a speaker's training. However, once more data professionals explore how much mirror neurons[9] affect the outcomes of speeches, this might change in a future edition.

While all the questions of how we give a good presentation related to personal development are out of the scope of this book, we dedicate Chapter 13 to Visualisation. Visualization is the core of delivering results to the audience because if we present our results in the best possible way, the results speak for themselves. Visualization is a powerful way of processing data. It helps us to better and faster understand the observed situation. It brings us new information, allowing us to see hidden relationships and representations. It forces us to ask better questions and helps us make better decisions. It tells us a story of yesterday and today, and predicts a story for tomorrow.

As a final point, it is essential to highlight that sometimes the answer in experimental environments is to have learned that there is no satisfactory answer to our questions.

To sum everything up: Becoming a data-driven enterprise is a new form of business in which "data is the new oil." We can use data to generate income through subscriptions or to reduce costs by solving problems. Physical products are now just tools to achieve a goal or address a customer problem,[10] and we aim to generate a continuous cash flow rather than only a one-time income.

Compared to the traditional software industry, the data industry has the following differences:

- A data platform is centralized (data lake or data warehouse) or decentralized (data mesh).
- Customers are co-producers of the solutions or products.
- New solutions and products can be continuously optimized and scaled.

[9] *https://blogs.scientificamerican.com/guest-blog/whats-so-special-about-mirror-neurons/*
[10] *https://www.ibsolution.com/academy/blog_en/the-four-steps-on-the-way-to-digitalization*

- There are increased requirements for automation and quality assurance due to the amount of data that is processed.
- The organizational form of the people who use Big Data is agile and continuously adapts to the problem. It also allows for making inevitable mistakes while always following the previously defined use cases or the vision.

To summarize this introduction, we will present the topics that a general digitization strategy usually covers for developing a business field.

1.7 In a Nutshell

From Strategy to Implementation

As a rule, a data strategy is derived from the corporate strategy and orchestrated by a CDO. Implementation is usually top-down.

From Questions to Answers

The key to success with data is always to ask the right questions that data should answer. Unfortunately, finding the right questions is not always trivial and might include workshops with stakeholders. On the other end, for an audience, the results of teams exploring data are as good as the team that presents them.

From Use Cases to Values

There are numerous use cases through which companies try to generate value via data science. In data projects, companies explore the possibilities and realize use cases in agile processes.

From Chaos to Platforms

Central for professional data exploration is also to build a platform that matches the requirements of all the data science use cases. For example, various studies show that teams spend 80 % of the time on data preparation in the beginning. A good platform might reduce these efforts for later use cases.

From Individuals to Teams

Many professionals with diverse skillsets work in a data team. To successfully implement a data project, you need people skilled in hardware, networks, operating systems, programming languages, CI/CD, data processing, data analysis, machine learning, reporting, and more. It is almost impossible to find one person that fits into all of these roles.

2 Infrastructure

Stefan Papp

"Perfection is not achieved when there is nothing left to add, but when there is nothing left to leave out."
Antoine de Saint-Exupéry

Questions Answered in this Chapter:

- Which system environments are necessary for data science projects?
- Why are cloud platforms ideal for experimentation-driven data science environments?
- What is the importance of GPUs and other hardware components for data science projects?
- How are platforms built dynamically with Infrastructure as Code and managed with version management tools?
- What distinguishes a microservice architecture from a monolithic architecture?
- How can Linux systems be used efficiently for Data Science?

2.1 Introduction

The goal of analytical processes in organizations is to gain new knowledge and deeper insights. As a rule, the benefit lies in

- reducing costs,
- making faster decisions and
- penetrating new markets.

There are sometimes decision-makers in companies who hope for results in minimal time. "Quick wins," i.e., situations in which data scientists can already deliver usable insights when exploring the data for the first time – without investing in the preparation of the data – are the exception.

Most companies explored a great deal of innovation using conventional analytical methods. Over the years, perhaps even decades, analysts have already racked their brains over how

to optimize revenues and reduce costs. As a result, the use of tried-and-true analytical techniques is no longer likely to yield groundbreaking new insights for many firms. Instead, the hope of many managers lies in newer methods such as **Artificial Intelligence**, **Machine Learning**, and **Deep Learning**. In later chapters the authors will explain the necessary details about Analytics.

In data-driven approaches, data professionals explore enormous amounts of data – up to the petabyte range – structured in particular topologies such as neural networks. Hardware and operating systems are necessary to host and process this data. As a result, we need a scalable and well-performing infrastructure that meets the business requirements for the data science use cases. Figure 2.1. outlines common layers for data architectures.

The term infrastructure is defined by NIST as follows:[1]

> *The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).*

The infrastructure for a data platform must meet the following requirements:

- Robust design and redundancy must ensure the lowest probability of data loss (**durability**).
- Security mechanisms must protect data from unauthorized access (**physical security**).
- The data platform must comply with the applicable data protection guidelines (**data protection**).
- The data platform must return results in a reasonable time (**performance**).
- Users must be able to access the platform and its data at any time (**availability**).

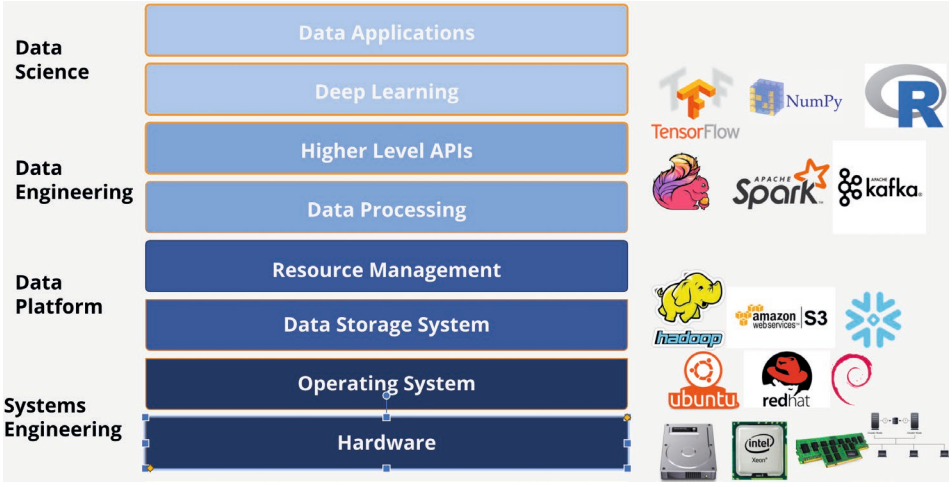


Figure 2.1 The layers of a (monolithic) data platform

[1] *https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf*

Few IT leaders have ever doubted that cloud providers, such as Microsoft, Amazon, and Google, can deliver a powerful, robust and secure infrastructure. However, there was skepticism about whether companies should store sensitive data on servers belonging to outside companies. Surveys confirm that today more and more companies are accepting the cloud as the future data platform.[2]

Experts perceive the cloud as a business model[3] and not as a new technology: the goal is to no longer own IT resources but to rent them. In addition, the cloud is also a perfect experimentation lab for data scientists, as they can generate required resources and decommission them dynamically.

Perhaps the most significant benefit of the cloud becomes apparent when business leaders consider the number of experts needed to operate a data platform in a dedicated data center:

- System architects design the solution based on the requirements.
- Operations engineers replace defective hardware.
- Network engineers build the network, including routers and cabling.
- Operating system experts install and configure operating systems.
- Facility managers take care of systems like air conditioning and fire protection.
- Security personnel secure access to the Data Center against unauthorized entry.

Depending on the requirements, we can subdivide individual tasks can even further. In order to guarantee fail-safety, each expert must also have a substitute. In addition, there is a shortage of IT specialists, which makes it challenging to fill vacancies adequately.[4] The cloud significantly reduces the company's personnel expenses.

2.2 Hardware

Big data platform infrastructures are complex, and mistakes in hardware configuration or procurement can be costly. Below are five real-world examples that illustrate the consequences of the wrong strategy in an on-premises environment.

Example 1: Hardware – CPU

A large international company built a mobile measurement system that collects terabytes of sensor data. One task was to offload the collected data to a central target platform as quickly as possible after a test drive. The transfer to the target platform took place via Ethernet.

[2] *https://www.flexera.com/blog/industry-trends/trend-of-cloud-computing-2020*
[3] *https://medium.com/@storjproject/there-is-no-cloud-it-s-just-someone-else-s-computer-6ecc37cdcfe5*
[4] *https://employer.it-talents.de/blog/it-fachkraeftemangel*

The architects chose the fastest available SSDs for the target platform because they assumed that if several measurement systems write to a target platform in parallel, the physical storage there would become the bottleneck.

When unloading data from the car, users complained that the transfer speed did not meet their expectations. It turned out that all four CPU cores of a measuring system, i.e., the source, were thoroughly utilized during unloading due to the serialization of data. The expensive SSDs of the target platform, on the other hand, had minimum utilization, as multiple data streams wrote in parallel.

Since the client had produced the measuring systems in series, it was impossible to equip the existing measuring systems with more processor cores. Therefore, it was necessary to wait for a new generation that provided more CPU power.

Example 2: Network

In a Big Data system for transferring data, the network turned out to be a bottleneck. Experts analyzing the problem did their best to solve it via software configurations. One variant they investigated was to enable jumbo packets to increase throughput.

Finally, the team investigated the physical cabling of the network. It turned out that the cabling did not support the expected transmission speeds. Completely rewiring the cluster was expensive and time-consuming. The project was delayed.

Example 3: Memory

A company promised the customer a Hadoop-based solution for handling data science projects. The team that had to implement the task assumed that only a single query would run in the system. Nobody expected to have to support a high-concurrency environment where multiple users would run queries simultaneously.

When the team presented a prototype to the customer, multiple users were testing in parallel, and crashes occurred because the Hadoop cluster ran out of RAM while running queries. The team could upgrade memory, but it soon became clear that the customer would have to upgrade significantly more than they were willing to spend money on for proper multi-user operation. As a result, the client canceled the project.

Example 4: I/O

In a proof of concept (POC), the client provided the supplier with a cloud system. The client's engineers had installed a Hadoop system that read data from blob storage on this platform. The goal of the POC was to prepare data for an analytical use case, and the managers expected to receive results promptly. However, a complex query with numerous joins took eight hours using Apache Spark as the execution engine. Waiting eight hours for intermediate results meant it was impossible to present and finish the project on time. After tough negotiations with the platform maintainer, the team was able to perform some optimizations. They replaced the file system and switched to Apache Parquet instead of CSV as the file format. In the end, the team could reduce the query time from eight hours to one.

Example 5: GPU

GPUs are expensive compared to other hardware components. Therefore, to save budget, one company bought only one server with GPUs. However, with multiple data scientists working on projects, delays occurred when too many wanted to use the server simultaneously.

One user suggested dynamically spinning up instances in the cloud for queries and shutting them down after a run. The leadership team rejected this decision as the company had a no-cloud policy. Many data scientists were unhappy with the situation because they could not work efficiently in the existing environment.

The basic knowledge of hardware components of a **Von Neumann Architecture**[5], as displayed in Figure 2.2, is not sufficient for those who set up network clusters themselves to distribute loads to several nodes. The selection of suitable hardware, the positioning of the servers in racks, the cabling, and the operation take up a lot of resources.

Finding the best hardware topology to meet the requirements of business problems is not always easy. In a project, stakeholders may identify new functional and non-functional requirements that the selected topology may not be adequate to support. In addition, the wrong infrastructure strategy can be costly to a business. Some IT departments have to allocate vast amounts of resources to **legacy systems** to keep them alive.

In addition, legacy systems can also affect employee satisfaction in an organization. If there is no opportunity to implement something new in IT, this can also lead to higher staff turnover. Employees are not always willing to work in an environment where there is no room for innovation, and the necessary resources are not available to permanently solve the problems of frustrated users.

The cloud offers companies more flexibility. For example, instead of purchasing hardware themselves, they rent it. Moreover, wrong decisions in the configuration of systems are re-

[5] *https://www.computerscience.gcse.guru/theory/von-neumann-architecture*

versible in the cloud. Research-oriented environments and environments that experience a lot of change thus benefit from the flexibility of the cloud.

Exercises

- Research on the Internet which GPUs are available for data science applications and how they can affect model building.
- Imagine you are preparing data for analytical modeling. The process takes several hours. Your boss asks you for ways to speed it up. Where do you start?
- Become CompTIA Network+ certified, or if you prefer to work with Cloud environments, pick a network certification of your preferred cloud provider or read books recommended to achieve the certification to learn all you can about various protocol stacks.
- Read the PolarFS paper[6] and understand how PolarFS differs from distributed file systems like the Hadoop Distributed File System (HDFS).

2.2.1 Distributed Systems

Two principles, originally from agile software development, are essential for building data science platforms:

- KISS (Keep it simple, stupid) recommends keeping systems as simple as possible, in line with the quote at the beginning of the chapter. Keeping all building blocks as simple as possible leads to a scalable solution.
- YAGNI (You ain't gonna need it) warns against building something on the suspicion that you might need it later.

Many data science projects failed in the past because the infrastructure did not fit the project requirements. Enthusiasts hope to solve all problems with new technologies. Consequently, engineers often perceive complex technology as an opportunity rather than a threat. Suppose a technology is introduced into the company only because it is new and innovative and not because a preliminary project has established that it solves the current business problems. In that case, this can cause the ambitious project to fail.

The demand for more CPU power is not the only issue; other hardware components must also become more powerful to meet the new requirements. There are two ways to scale when existing hardware is too slow. First, scale up improves the system's performance by replacing one hardware configuration with a more powerful one. However, above a particular load, even the best hardware will not solve a problem. Second, scale out means to distribute the load across more hardware. Here, experts also speak of 'distributed processing' or 'parallelization.'

[6] *http://www.vldb.org/pvldb/vol11/p1849-cao.pdf*

At the physical level, the distribution of loads (scale out) can take two forms. In the first variant, operation engineers install multiple components of the same type and a controller within a computer system. Examples of this are RAIDs or multi-core processors.

The operating system ensures that applications can handle the duplication of hardware modules. Those applications can then use threading and multiprocessing to distribute the load internally to several processor cores.

A **Von Neumann architecture** is the basis of a modern computer, as shown in Figure 2.2, and consists of a system that includes a CPU, memory, and an I/O component interconnected by a BUS. Thus, a cluster of computers (commonly called 'nodes') in a server room usually consists of several Von Neumann systems connected via a network.

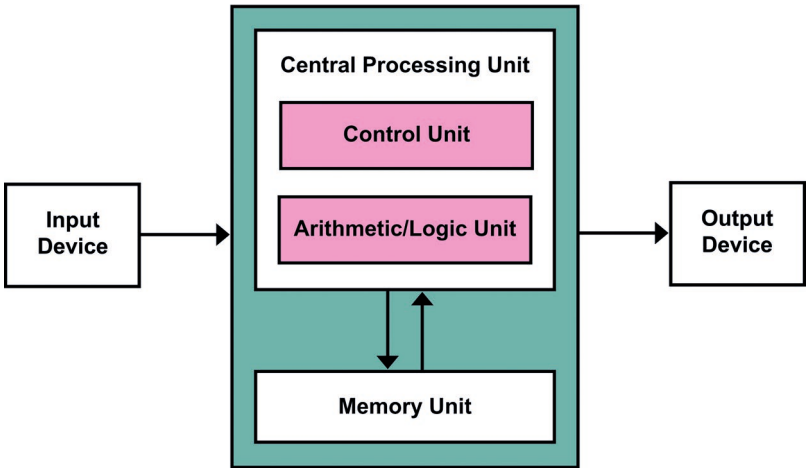


Figure 2.2 Von-Neumann-Architecture[7]

In addition to the hardware modules in a computer system, several computer systems can also be scaled out as nodes in a cluster as a second form of scale-out.

Also, in a cluster, you need software (perhaps the more accurate term would be an 'operating system,' but hardly any engineer uses this term in that context), which orchestrates data processing on multiple machines. This mechanism of managing different devices also includes synchronizing intermediate results and statuses between these processing components. Another task for the software is to react to disturbances, such as a node failure or delayed synchronization.

To illustrate the overhead incurred by parallelization, consider the following example, in which we will contrast local data processing with distributed data processing via Apache Hadoop. When data is processed locally on a single computer, a process loads data into memory. Then, an algorithm is applied to the loaded data to modify it or calculate results from it. These results can be displayed on the screen or saved as a file. If the amount of data to be loaded is larger than the available RAM, delays may occur as the data processing engine needs to swap intermediate results to the local hard disk during processing.

[7] Copyright: Kapooht *(https://en.wikipedia.org/wiki/File:Von_Neumann_Architecture.svg)*

Figure 2.3 indicates the additional overhead, using a simplified representation of distributed processing in Hadoop. We have to imagine that in a distributed system, we store all data on different nodes. This mechanism means each node needs to process as much as possible until a framework can collect reduced data on a few nodes. In a local environment, all this preprocessing first on separate nodes and joining later is not necessary.

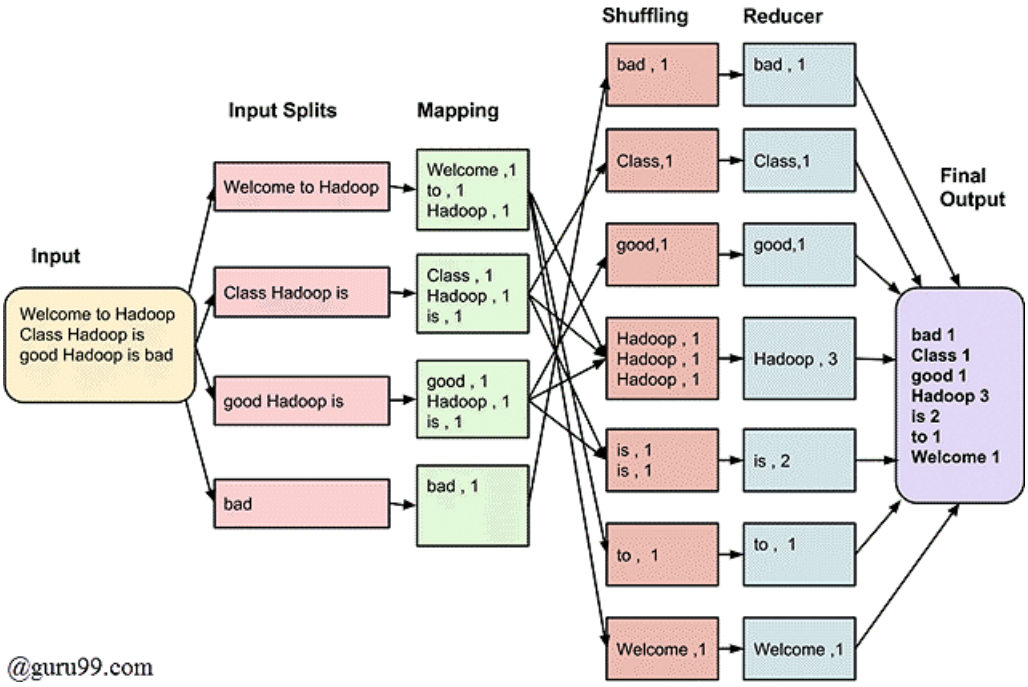


Figure 2.3 Processing with MapReduce[8]

Developed in Java for the distributed data processing, Apache Hadoop is a software that needs to be installed on your cluster nodes to complete such a task.

In data processing via Hadoop, algorithms are executed first on the data stored on the so-called 'data nodes.' Then, analogous to local processing, processes running on the data nodes load the data into RAM, apply the algorithms, and cache the results. Finally, the intermediate outcomes on the individual nodes are merged centrally in a second step. The details of this process, called **MapReduce**, are described in Chapter 3 and on the Hadoop page.[9] In practical applications, layers abstract this type of processing and provide a unified API. Chapter 4, "Data Engineering," details how frameworks like Apache Spark optimize this flow.

Every engineer should be aware that local processing has a lower overhead. It is always faster if the amount of data is small enough to be processed locally.

[8] *https://www.guru99.com/introduction-to-mapreduce.html*

[9] *https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html*