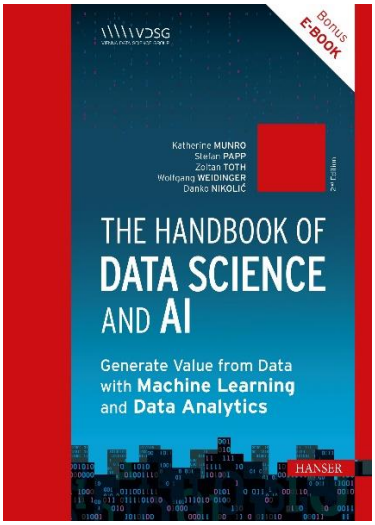
HANSER



Sample Pages

Thermal Analysis in Practice

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

Print ISBN: 978-1-56990-934-8

E-Book ISBN: 978-1-56990-235-6

ePub ISBN: 978-1-56990-411-4

For further information and order see

www.hanserpublications.com (in the Americas)

www.hanser-fachbuch.de (outside the Americas)

Table of Contents

Preface

This preface was NOT written by ChatGPT (or similar).

As I make this statement, I'm wondering how often it will remain true for text or even other forms of media in the future. Over the last two years, this AI-powered tool has risen to enormous popularity, and has given Data Science and AI an incredible awareness boost. As a result, the expectations for Artificial Intelligence have grown seemingly exponentially, and reached such heights that one might ask, if they can ever be achieved.

AI is following the well-known hype cycle. Some of these high expectations are well-deserved: this powerful technology will change the way we live and work in many ways. To name one example: some universities are considering not to ask their students for seminar papers any longer, as it's not possible to check if it was written by an AI tool.

But we also must brace ourselves for some disappointment in the future, as AI inevitably fails to live up to certain people's inflated expectations.

Even when the vision is reasonable, often the timelines these people and organizations have in mind for implementing AI projects is not. This leads to further disappointment, when the hoped-for impact and value fail to materialize within the desired timeframe.

We're already seeing the beginning of this, with ChatGPT and similar tools generating plenty of eloquent and coherent – yet completely inaccurate – information. This isn't helped by the new wave of 'AI experts', who are making ever more outlandish promises about tools invented by themselves or their companies; promises which will be very hard to keep. They are, essentially, selling digital 'snake oil'.

All of this puts even more pressure on data scientists to deal with these expectations, while continuing to deliver on the same goal they've had for decades:

generating understandable answers to questions, using data.

This is what makes neutral organizations such as the Vienna Data Science Group (VDSG [*www.vdsg.at*]) – which fosters interdisciplinary and international knowledge exchange between data experts – so necessary and important. We are still highly dedicated to the development of the entire Data Science and AI ecosystem (education, certification, standardization, societal impact study, and so on), across Europe and beyond. This book represents just one of our efforts towards this goal. Because despite all the hype and hyperbole in the AI and data landscape, Data Science remains the same: an interdisciplinary science gathering a very heterogeneous crowd of specialists. It is made up of three major streams, and we are proud to have expert members in each of them:

- Computer Science and IT
- Mathematics and Statistics
- Domain expertise in the industry or field in which Data Science and AI is applied.

As a matter of fact, the VDSG [*www.vdsg.at*] has always taken a holistic approach to data science, and this book is no different: Starting at Chapter 1 we introduce a fictional company who wants to become more data driven, and we check in with them throughout the book, right up to the end of their data transformation in Chapter 28. Along the way we cover many challenges in their journey, thus providing you with practical insights which were only possible thanks to vibrant exchange among our vast Data Science and AI community.

The result is a greatly expanded edition of our Data Science & AI Handbook, with 10 new chapters covering topics like Building AI solutions (Chapter 13), Foundation Models (Chapter 15), Large Language Models and Generative AI (Chapter 16) and Climate Change and AI (Chapter 25). This is complemented by also tackling the fundamental topics of Data Architecture, Engineering and Governance (Chapters 4, 5 and 6) and topping it off with Machine Learning Operations (MLOps, Chapter 7), which has become a very important discipline in itself.

To provide a firm foundation to help you understand all this, we've again included an introduction to the underlying Mathematics (Chapter 9) and Statistics (Chapter 10) used in Data Science, as well as chapters on the theory behind Machine Learning, Signal Processing and Computer Vision (Chapters 12, 14 and 18). We've also covered topics related to generating value from data, such as Business Intelligence (Chapter 11) and Data Driven Enterprises (Chapter 21), as well as vital information to help you use data safely, including chapters on the new EU AI Act (Chapter 23) and Trustworthy AI (Chapter 27).

This vast expansion of VDSG's Magnum Opus serves one core purpose:

to give a realistic and holistic picture of Data Science and AI.

Data Science and AI is developing at an incredibly quick pace at the moment and so is its impact on society. This means that responsibilities put on the shoulders of data scientists have grown as well, and so has the need for organizations like VDSG [*www.vdsg.at*] to get involved and tackle these challenges too.

Let's go for it!

Summer 2024

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 lot 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

Thinking about 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 again during the last months while 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

Stefan Papp

"I want to be CDO instead of the CDO."
Iznogoud (adjusted)

Questions Answered in this Chapter:

- How could we describe a fictional company before its journey to becoming data-driven?
- What challenges might such a company need to resolve to become data-driven?
- How will the chapters in this book help you, the reader, to recognize and address such challenges in your own organization?

1.1 About this Book

This book takes a practical, experience-led look into various aspects of data science and artificial intelligence. In this, our third edition, the authors also deeply dive into some of the most exciting and rapidly developing topics of our time, including large language models and generative AI.

The authors' primary goal is to give the reader a holistic approach to the field. For this reason, this book is not purely technical: Data science and AI maturity depends as much on work culture, particularly critical thinking and evidence-based decision-making, as it does on knowledge in mathematics, neural networks, AI frameworks, and data platforms.

In recent years, most experts have come to agree that artificial intelligence will change how we work and live. For a holistic view, we must also look at the status quo, if we want to understand what needs to be done to meet our diverse ambitions with the help of AI. One useful frame for doing this is to explore how people deal with data transformation challenges from an organizational perspective. For this reason, we will shortly introduce the reader to a fictional company at the beginning of its journey to integrate evidence-based decision-making into its corporate identity. We'll use this fictional company, in which most things could be more data-oriented but aren't yet, as a model for outlining possible challenges organiza-

tions may encounter when aiming to become more data-driven. By the end of this book, our hypothetical company will also serve as a model of how a data-driven company could look. In the chapters in between, we'll address many of these challenges and provide practical advice on how to tackle them.

Suppose you, as a reader, would rather not read prose about an invented company in order to learn about such typical organizational challenges. In that case, we encourage you to skip this chapter and start with one that fits your interests. As a holistic book on this field, the authors discuss artificial intelligence, machine learning, generative AI, modeling, natural language processing, computer vision, and other relevant areas. We cover engineering-related topics such as data architecture and data pipelines, which are essential for getting data-driven projects into production. Lastly, we also address critical social and legal issues surrounding the use of data. Each author goes into a lot of detail for their specific field, so there's plenty for you to learn from.

We kindly ask readers to contact us directly to provide feedback on how we can do better to achieve our ambitious goal of becoming the standard literature providing a holistic approach to this field. If you feel some new content should be covered in one of the subsequent editions, you can find the authors on professional networks such as LinkedIn.

And with that said, let's get started.

1.2 The Halford Group

Bob entered the office building of the Halford Group, a manufacturer of consumer products, including their best-selling rubber duck. After crossing the office doors, he felt he was thrown back into the eighties. Visitors having to register at the entrance, filling out forms to declare themselves liable in case of an accident, and promising not to take photos, was only the first step. As Bob entered the elevator, with its brass buttons and glossy, mahogany decor, he could have sworn he'd entered the setting of the movie "The Wolf of Wall Street."

The executive office was similar. The brownish carpets showed their age, and the wallpapers looked like they'd inhaled the smoke of many an eighties Marlboro Man. The worn leather couches and the looming wooden desk (mahogany, again), seemed a memory of a great but distant past. Bob could imagine his dad—a man who had always been proud of being in sales and following the teachings of Zig Ziglar—doing business with this company in his younger years.

This image in Bob's imagination was immediately disrupted when a young woman entered the room, and Bob was immediately thrown back into the present time. With an air of determination, she strode forward to reach for Bob's hand. Somewhat taken aback, he took in the shock of platinum blonde hair, and the tattoos that had not been entirely hidden by her tailored suit, and raised his hand in response. The woman smiled.

1.2.1 Alice Halford – Chairwoman

"I'm Alice Halford," she said, "I am the granddaughter of Big Harry Halford, the founder of this group. He built his empire from the ground up."

Bob had read all the legends about the old Halford boss. Every article about him made it clear he did not listen to many people. Instead, "Big Harry" was a proud, determined captain; one who set the course and demanded absolute obedience from his team. Business magazines were yet to write much about Alice, as far as Bob knew. However, he had read one article in preparation for this meeting. Alice was different from the grand old family patriarch, it had said. She had won the succession in a fierce battle against three ambitious brothers, and been selected by the board as chairwoman, thanks to her big plans to transition the company into a modern enterprise that could meet the Zeitgeist of the 21st century.

"Although successful, today's generation would call my granddad a dinosaur who just wanted to leave enough footprints to let the next generation know he had been there," Alice said. "Especially in his last years, he was skeptical about changes. Many principal consultants from respectable companies came with heads high to our offices, explaining that our long-term existence would depend on becoming a data-driven company. However, my granddad always had a saying: The moment a computer decides, instead of a founder who knows their stuff and follows their gut, it's over. All the once proud consultants and their supporters from within the company thought they could convince every executive to buy into their ideas of a modern company, but ultimately, they walked out with their tails between their legs."

Alice smiled at Bob and continued, "my granddad's retirement was long overdue, but, finally, his exotic Cuban cigars and his habit of drinking expensive whiskey forced him to end his work life. I took over as a chairwoman of the board. I want to eliminate all the smells of the last century. When I joined, I found parts of the company were highly toxic. My strategic consultants advised me that every large organization has some organizational arrogance and inefficiency. They also cautioned me to keep my expectations low. While many enthusiasts claim that AI will change the world forever, every large organization is like a living organism with many different subdivisions and characteristics. Changing a company's culture is a long process, and many companies face similar challenges. Ultimately, every company is run by people, and nobody can change people over night. Some might be okay with changes, a few may even want them to happen too fast, but most people will resist changes in one way or another.

At the same time, I understand that we are running out of time. We learned that our main competitors are ahead of us, and if we do not catch up, we will eventually go out of business. Our current CEO has a background in Finance and, therefore, needs support from a data strategist. Bob, you have been recommended as the most outstanding expert to transform a company into a data-driven enterprise that disrupts traditional business models. You can talk with everyone; you have all the freedom you need. After that, I am curious about your ideas to change the company from the ground up."

Bob nodded enthusiastically. "I love challenges. Your secretary already told me I shouldn't have any other appointments in the afternoon. Can you introduce me to your team? I would love to learn more about how they work, and their requirements."

"I thought you'd want to do that. First, you will meet David and Anna, the analysts. Then you'll meet Tom, the sales director. It would be best if you also talked with the IT manager, Peter—" Alice stopped herself, sighed, and continued. "Lastly, I arranged a meeting for you with our production leader, the complaints department, our Head of Security, and finally with our HR. I will introduce our new CEO, who is flying in today to discuss details at dinner. I booked a table in a good restaurant close by. But it makes sense if you first talk to all the other stakeholders. I had my colleagues each arrange a one-on-one with you. You're in for a busy afternoon, Bob."

1.2.2 Analysts

As Alice swept out of the room, a bespeckled man apparently in his mid-forties, and a woman of about the same age, appeared in the doorway. It must have been the analysts, David and Anna. When neither appeared willing to enter the room first, Bob beckoned them inside. He was reminded of an empowerment seminar he'd attended some years ago: The trainer had been hell bent on turning everyone in the workshop into strong leaders, but warned that only the energetic would dominate the world. These analysts seemed to be the exact opposite. David laughed nervously as he entered, and Anna kept her eyes lowered as she headed to the nearest seat. Neither seemed too thrilled to be there; Bob didn't even want to imagine how they would have performed in that seminar's "primal scream" test.

David and Anna sat down, and Bob tried to break the ice with questions about their work. It took him a while, but finally, they started to talk.

"Well, we create reports for management," David said. "We aim to keep things accurate, and we try to hand in our reports on time. It's become something of a reputation," he added with a weak chuckle.

Bob realized that if he was going to make them talk, he'd need to give his famous speech, summarized as, "your job in this meeting is to talk about your problem. Mine is to listen." After all, he needed to transform Halford company into a data-driven company, and they were ones working closest with the company's data.

Bob finished his speech with gusto, but Anna merely shrugged. "The management wants to know a lot, but our possibilities are limited."

Bob tried his best to look both in the eyes, though Anna turned quickly away. "But what is it that prevents you from doing your work without any limits?"

"Our biggest challenge is the batch process from hell," David spoke up suddenly. "This notorious daily job runs overnight and extracts all data from the operational databases. It is hugely complex. I lost count of how often this job failed over time."

Got them, Bob thought, nodding in encouragement.

"And nobody knows why this job fails," Anna jumped in. "But when it does, we don't know if the data is accurate. So far, there has never been a problem if we handed in a report with questionable figures. But that's probably because most managers ignore the facts and figures we provide anyway."

"Exactly!" David threw up his hands. Bob started to worry he had stirred up a hornet's nest.

"When a job fails, it's me who has to go to IT," David said. "I just can't hear anymore that these nerds ran out of disk space and that some DevSecOps closed a firewall port again. All I want is the data to create my reports. I also fight often with our security department. Sometimes, their processes are so strict that they come close to sabotaging innovation. Occasionally, I get the impression they cut access to data sources on purpose to annoy us."

"Often, we are asked if we want something more sophisticated," Anna said, shaking her head in frustration. "It is always the same pattern. A manager visits a seminar and comes to us to ask us if we can 'do AI'. If you ask me honestly, I would love to do something more sophisticated, but we are afraid that the whole system will break apart if we change something. So, I am just happy if we can provide the management with the data from the day before."

Don't get us wrong, ML and AI would be amazing. But our company must still master the basics. I believe most of our managers have no clue what AI does and what we could do with it. But will they admit it? Not a chance."

Anna sat back in a huff. Bob did not need to ask them to know that both were applying at other companies for jobs.

1.2.3 "CDO"

At lunch break, a skinny man in a black turtleneck sweater hurled into the office. He seemed nervous, as if someone was chasing him. His eyes darted around the room, avoiding eye contact. His whole body was fidgeting, and he could not keep his hands still.

"I am the CDO. My name is Cesario Antonio Ramirez Sanchez; call me Cesar," he introduced himself with a Spanish accent.

Bob was surprised that this meeting had not been announced. Meanwhile, his unexpected visitor kept approaching a chair and moving away from it again as if he could not decide whether to sit down or not.

"CDO? I have not seen this position in the org chart," Bob answered calmly, "I have seen a Cesario Antonio Rami ..."

"No no no ... It's not my official title. It is what I am *doing*," Cesar said dramatically. "I am changing the company bottom up, you know? Like guerilla warfare. Without people like me, this company would still be in the Stone Age, you see?"

"I am interested in everyone's view," Bob replied, "but I report to Alice, and I cannot participate in any black ops work."

"No, no, no ..., everything is simple. Lots of imbeciles are running around in this company—" Cesar raised his finger and took a sharp breath, nodded twice, and continued. "I know ... HR always tells me to be friendly with people and not to say bad words. But we have only data warehouses in this company. Not even a data lake. Catastrófica! Its the 21st century, and these dinosaurs work like in Latin America hace veinte años. Increíble!"

He took another breath, and then continued. "Let's modernize! Everything! Start from zero. So much to do. First, we must toss these old devices into the garbage, you know? And replace them with streaming-enabled PLCs. Then, modern edge computing services streams

everything with Kafka to different data stores. All problems solved. And then we'll have a real-time analytics layer on top of a data mesh."

Bob stared at his counterpart, who seemed unable to keep his eyes or his body still for more than a moment. "I am sorry, I do not understand."

"You are an expert, you have a Ph.D., no? You should understand: modern factory, IoT, Industry 4.0, Factory of the Future."

Bob decided not to answer. Instead, he kept his eyebrows raised as he waited for what Cesar would say next.

"So much potential," Cesar went on. "And all is wasted. Why is HR always talking about people's feelings? Everything is so easy. This old company needs to get modern. We don't need artists, we need people with brains. If I want art, I listen to Mariachi in Cancun. If current people are imbeciles, hire new people. Smart people, with Ph.D. and experience. My old bosses in Latin America, you cannot imagine, they would have fired everyone, including HR. Let's talk later; I'm in the IT department en la cava."

Bob had no time to answer. Cesar left the room as fast as he had entered it.

1.2.4 Sales

A tall, slim, grey-haired man entered the room, took a place at the end of the table, leaned back and presented to Bob a salesman grin for which Colgate would have paid millions.

"I am Tom Jenkins. My friends call me 'the Avalanche'. That's because if I take the phone, nobody can stop me anymore. Back in the nineties, I made four sales in a single day. Can you imagine this?"

I get it; you are a hero. Bob thought. *Let's turn it down a bit.*

"My name is Bob. I am a consultant who has been hired to help this company become more data-oriented."

Tom's winning smile vanished when Bob mentioned 'data.'

"I have heard too much of the data talk," Tom said. "No analysis can beat gut feeling and experience. Don't get me wrong. I love accurate data about my sales records, but you should trust an experienced man to make his own decisions. No computer will ever tell me which potential client I should call. When I sit at my desk, I know which baby will fly."

"With all due respect. I can show you a lot of examples of how an evidence-based approach has helped clients to make more revenue."

"Did you hear yourself just now?" Tom answered, "Evidence-based. You do not win sales with brainy talks. You need to work on people's emotions and relationships. No computer will ever do better sales than a salesman with a winning smile. I'll give you an example: One day, our sales data showed that we sold fewer products in our prime region. Some data analysts told me something about demographic changes. What a nonsense!

So, I went out and talked to the people. I know my folks up there. They are all great people. All amazing guys! Very smart and very hands-on. I love this. We had some steaks and beers, then I pitched our new product line. Guess who was salesman of the month after that?

No computer needs to tell me how to approach my clients. So, as long as we get the sales reports right and we can calculate the commission, all is good. It is the salesman, not the computer, who closes a deal."

With that, The Avalanche was on his feet. He invited Bob to a fantastic restaurant—"I know the owner and trust me, he makes the best steaks you'll ever taste!"—and was gone.

1.2.5 IT

Ten minutes past the planned meeting start time, Bob was still waiting for the team member he had heard most about upfront: the IT leader, Peter. His name had been mentioned by various people multiple times, but whenever Bob had asked to know more about him, people were reluctant to answer, or simply sighed and told him, "you'll see."

Finally, Peter stormed into the room, breathless and sweating. "This trip from my office in the cellar to this floor is a nightmare," he said between gasps. "You meet so many people in the elevator who want something. I am constantly under so much stress, you cannot imagine! Here, I brought us some sandwiches. I have a little side business in gastronomy. You need a hobby like this to survive here. Without a hobby in this business, you go mad."

Peter was a squat, red-faced man, who'd been with Halford since he was a lot younger, and had a lot more hair. He sank a little too comfortably in his chair, with the confidence of a man who'd been around so long, he was practically part of the furniture.

He doesn't lack confidence, that's for sure, Bob thought. *I wonder how many dirty secrets this man has learned over the years that only he knows.*

"Okay, let's talk about IT then," Peter sighed after Bob turned down the sandwiches. "My colleagues from the board and the executives still don't get what it is they're asking of me daily. When they invite me to meetings, I often do not show up anymore. We are a huge company, but nobody wants to invest in IT. I am understaffed; we hardly manage to keep the company running. Want to go for a cigarette?"

"No, thank you," Bob said, but Peter was already crumpled pack from his trouser pocket. He rambled all the way to the smoker's chamber, bouncing around from one topic to another. Bob learned everything about Peter, from his favorite food over his private home to his hernia, which was apparently only getting worse. Once Peter got first cigarette into his mouth, he went back to the topic Bob was really interested in.

"The suits want things without knowing the implications. On the one hand, they want everything to be secure, but then again, they want modern data solutions. Often, they ask me for one thing one day, and then the very next, they prioritize something else. To be blunt, I had my share of talks with these external consultants. If I allowed them to do what they asked me to do, I could immediately put all our data on a file server and invite hackers to download it with the same result. To keep things working, you need to firewall the whole company," Peter stubbed out his cigarette, and reached for another.

Bob leaped at the chance to interject. "Can you tell me more about your IT department? I was looking for some documentation of the IT landscape. I have not found much information on your internal file shares. Which cloud provider are you currently using?"

Peter laughed and then started coughing. Tears in his eyes, he answered. “I told you, I’m understaffed. Do you really think I have time to document?” He pointed to his head. “Don’t worry, everything is stored in the grey cells up here. And we have a no-cloud strategy. Cloud is just a marketing thing if you ask me. When we build by ourselves, it is safer, and we have everything under control.

If I just had more people ... Did you meet one of my guys, Cesar? He is also okay when he does not talk, which unfortunately doesn’t happen often. I don’t like when people think they are smarter than me. He doesn’t know Peter’s two rules yet. Rule Number 1: Do not get on your boss’s nerves. Rule Number 2. Follow rule number 1.”

Peter laughed, flicked the second cigarette on the ground, and retrieved a bag from his other pocket. It was full of caramels: Peter popped one into his mouth and continued, chewing loudly. “Alice asked me if I could introduce you to Bill, my lead engineer, but I declined. This guy has the brains of a fox but the communication skills of a donkey. He also gets nervous when you look him straight in the eyes. I am always worried that he might wet his pants— Or am I being too politically incorrect again? Our HR keeps telling me that I should be more friendly. But in this looney bin, you learn to let our your stress by saying what you think. So, please excuse my sarcasm. I am the last person standing between chaos and a running IT landscape, the management keeps getting on my nerves with stupid requests, and last but not least, the HR department is more concerned about how I communicate than about finding the people who could help me keep our company running.”

It took a couple of attempts until Bob could finally break free from Peter’s complaining to head to his next meeting. Even as he was leaving, Peter repeatedly called on Bob to visit his food business sometime, where they could have a drink in private, and Peter could share his Halford ‘war stories’ more openly.

1.2.6 Security

While waiting for the HR representative, Bob received a voice message from Suzie Wong, the head of data security. When Bob played it, he heard traffic sounds in the background.

“Apologies for not showing up. School called me in as one of my kids got sick. I hope a voice message is fine. I am Suzie Wong. I have been with Halford for years. They call me the human firewall against innovation. I take this as a compliment because, in some way, it means I am doing my job well. Could any company be happy with a Head of Security who takes her job easy? My predecessor was more laid back than I am. He was in his fifties and got a little too comfortable, thinking he would retire in a secure job. And then one day ... there was this security breach. His kid’s still in private school, he’s suddenly without a job and, well, I’ll spare you the details.

People often think I’m only around to sign off on their intentions to use data, but my real job is protecting our client’s privacy. Data scientists must prove to me that our client’s data is safe when they want to work with it. Unfortunately, too many take that too lightly.

If the requestor follows the process, a privacy impact assessment could be done within a week. I will send you a link to our security portal later so you can review it. You’ll see for yourself that we do not ask for anything impossible.

I am the last line of defense, ensuring that we do not pay hefty fines because someone thought it was just data they were playing around with. Some people also jokingly call me 'Mrs. No,' because this is my common answer if you cannot express why I should grant you security exceptions or provide access to data containing clients' private information. Some people complain that this way, it may take months to get security approval. But so long as engineers and data scientists still don't get how to address security matters correctly, I don't care if it takes years before I give my final OK.

Anyway, excuse me now, I'm at the school ..."

1.2.7 Production Leader

Bob had some time before his next meeting and looked up his next meeting partner online. He discovered a middle-aged man with a long history on social media, including some questionable photos of his younger self in a Che Guevara t-shirt. Bob chuckled. That young man could be happy that their interview wasn't taking place during the times of the Cold War.

Finally, Bob's interviewee entered the room. He was muscular, and his bushy black beard showed the first signs of greying.

"My name is Hank. Pleased to meet you," he said with a deep voice.

"I heard you are new in your position," Bob said.

"Yes. Alice fired my predecessor because he was a tyrant. I am now one of the first of what she calls 'the new generation.' I accepted because I can change things here now. Let me get to the point: What are you planning to do?"

Bob smiled and said, "the idea in factories is often to use machine learning for automation. Think of processes where people check the quality of an item manually. Imagine that you can automate all this. A camera screens every piece, and defective items — which we call 'rejects' — are filtered out automatically."

Hank stiffened. "My job is to protect jobs, not support removing them. Some of our factories are often in villages, where they are the only source of work."

"Almost every country goes through demographic changes. Can you guarantee that you will be able to maintain a strong enough workforce to keep the factories running? How about doing the same with fewer people?"

"But if you remove a few people, they can end up out of work," Hank said. "What if you don't need workers at all in a few years? I don't want to open the door to a system that makes the bourgeoisie richer and put the ordinary proletarian out of work."

"That is very unlikely," Bob said.

"I see you are solidary with your employees, Hank. Did you consider exploring use cases to protect them? We can use computer vision to see if factory workers wear helmets, for example."

Hank looked deeply into Bob's eyes. Bob couldn't quite tell if it was a good or bad sign, be he did realize something: this was not a man he'd like to meet on a dark, empty street.

"I understand that there might be benefits for my colleagues," Hank said. "I just want to open up a trojan horse: I get one IT system in to prevent accidents, and the next one makes the workers obsolete. But I promised Alice I'd support her. She is a good person. I will talk with my colleagues. I need to get them on board, but one thing is not negotiable: We will never tolerate any system that completely replaces people who need the job they have."

1.2.8 Customer Service

The next interviewee, an elderly woman with perfectly glossy, silver hair, entered the room. She sat down and carefully ran her fingers over classic French bun, ensuring not a hair was out of place.

"I am Annie from the complaints department," she said with something of an aristocratic tone. She seemed more interested in her neatly manicured nails than Bob as she went on. "I honestly do not know why you want to talk to me."

"Well, part of a data-driven enterprise is often also a customer-first strategy. We can measure customer churn and other metrics through data. Most of my clients want to use data to maximize success. They even renamed their departments to 'Customer Satisfaction Department' to underline this."

"Aha," Annie said. There was an uncomfortable silence as she polished the face of her antique watch with her other sleeve.

Bob cleared his throat, anxious to get her attention. "Would you be interested to learn more about your customers through data?"

"Why should I?"

"To serve them better?"

"We have sturdy products. Most complaints have no base. We believe the less money we spend on confused customers, the more we have left to improve our products. This is what I call the real customer value we provide."

Ah-hah. Bob recognized the famous argument against investing in any domain that doesn't directly create revenue. *She probably gets a bonus for keeping yearly costs low*, he thought, seeing an opportunity.

"And how do you keep costs small at the moment?"

"We have an offshore call center. They handle about 80 % of calls, although a lot of those customers just give up, for some reason. The remaining 20 % are forwarded to a small team of more advanced customer support employees. I know it sounds harsh, but you cannot imagine how many confused people try to call us without having a problem at all. Some – it seems – call us just to talk."

"Right. And have you thought of the possibility to reduce costs by building chatbots backed by generative AI? There are also many ways to use data science to filter customer complaints. If properly trained, your clients get better support, and you reduce costs."

"Would it be good enough to shut down the offshore center?

Gotcha. "If done right, yes."

For what felt like the first time, Annie looked at Bob directly. “How much would it cost?”

“At the moment, it is still difficult to estimate.”

Annie thought a while, then stood up to leave. At the door, she paused. “Once you know, call me immediately.”

1.2.9 HR

“I’m, I’m Pratima,” came a woman’s voice at the door. She approached Bob, looked up at him with a welcoming smile and asked, “how can I help you, Bob?”

“Hi, Pratima. Let’s take a seat. As you know, I’m here to transform this company into a more data-oriented one. I saw on LinkedIn that you have previously worked for very modern companies with a strong data culture. How is it now to work for a company at the beginning of its journey?”

“Alice asked me to be open to you. I took this job as a career step to advance to leadership. However, the Wheel of Fortune led me to more challenges than expected.

In my previous job, we had the vibes to attract new talent. It was an environment primed for excellence: fancy office spaces, a modern work culture with flat hierarchies, cool products to work on, and many talented, diverse colleagues. Recruiting was easy because new candidates felt it the spirit of our community.”

Pratima sighed.

“In this company, though, we cannot hide that we are at the beginning of our transition. Applicants usually have many offers to choose from. Sometimes, we have to watch perfect candidates walk away because we do not yet provide a warm and welcoming environment for data professionals.

When managers discuss AI and data transition, some might oversee the human aspect. What if you create the perfect data strategy but cannot attract enough talent? Many companies face this problem, and an elephant is always in the room. To become a data-driven company, you have to create an environment that attracts people who think differently, and this means changing your culture.

“Do you believe management is scared to promote too much change because it is afraid to lose everything?”

“I understand that some seasoned employees might get disappointed and even resign if their comfortable environment starts to modernize. But at the same time, if you do not change at all, you are stuck in the mud, and your competition will make you obsolete. The Dalai Lama says we should be the change we wish to be.”

“Right. And I believe it was Seneca who once said, ‘It’s not because things are difficult that we dare not venture. It’s because we dare not venture that they are difficult.’”

“True! But I have to go now. I am looking forward to continuing our talks.”

1.2.10 CEO

Alice and Bob met at a fusion restaurant downtown in the evening. Alice introduced Bob to Santiago, the long-time CFO turned new CEO. After an excellent meal, they ordered some famous Armenian cognac, and got down to the real discussion.

"I'll be honest with you, Bob," Santiago began. "All your ideas to transform Halford sound fantastic, but as an economist and a numbers person, my first question is, how much will this all cost?"

Oh boy. Bob was prepared for the question, but he knew Santiago wouldn't like the answer. "It depends," he said, and Santiago looked about as dissatisfied as Bob would have expected.

"I understand that everyone looks at the costs," Bob continued, "but history is full of companies that failed to innovate and went bankrupt as their competition moved forward. If you see the full spectrum of artificial intelligence, hardly any company will eventually operate as before."

"Some companies recommend that we start with data literacy workshops to enable leaders to interpret data and numbers efficiently. Literacy sounds as if they want to teach us to read and write again—and for a huge amount of money, of course. Don't get me wrong, please. I understand that we need to innovate, but if I approve everything consultants suggest to me, we will soon be broke."

"But if your leadership team cannot 'think in data'," Bob said, making air quotes as he spoke, "how do they expect to attend our planned strategy workshop on exploring specific data science options for our business goals?"

"What is the difference?"

"In the data literacy workshops, we aim to create an understanding of how to interpret data. In the strategy workshop, we'll create a list of use cases to improve processes in your company, and prioritize them, to integrate new data solutions gradually."

"I understand that we have some tough nuts to crack. Some of our employees do not believe in becoming data-driven, and we may need to invest hugely in Enablement. We once asked external companies to help us modernize our IT. No consulting company gave me a quote with a fixed price for a transition project. They always said we were facing a hole without a bottom."

"Leadership is the only way to move forward. If the executive team is convinced and aligned, this culture can spread.

Your operational IT will need to mature and modernize gradually. However, be aware that an analytical layer can be built outside of corporate IT. One risk is to make data transition to an IT problem; IT is part of it, but becoming a data-driven company is far more than giving some engineers a job to build platforms."

"For me, it's clear," Alice said. "Either we modernize, or we gradually fade out of existence. Bob, what do you need to help us?"

Bob looked from one to the other, carefully considering his next words. "Becoming data-driven does not mean hiring a bunch of data scientists who do a bit of magic, and suddenly the company makes tons of money using AI. As I said, the first step is to align the stakeholders. For me, this is the alpha and omega of AI: creating a data culture based on critical thinking and evidence-based decisions. "

"Great," answered Alice. "Let's get started with that."

1.3 In a Nutshell

Expectation Management

Most companies see the need to become data-driven, as they understand that those organizations that ignore technical evolution mostly fail.

Some employees might have unrealistic expectations about how fast a transition can go. We highlight that changing to a data-driven company is not just a change of practices and processes, it is often a cultural overhaul of how the company does its business.

Many employees fear having to give up some of their autonomy, or even losing their jobs to computers entirely, if AI is introduced at their company. An organization that transitions to become data-driven must address this.

Technology Focus and Missing Strategy

Some companies try to find a silver bullet that solves all problems. "We'll just use this technology, just apply AI in this or that way, and all our problems are resolved," they think. Being too technology-focused, however, is an anti-pattern that can hinder a company's evolution to becoming data-driven.

Data Science and AI are about more than just Understanding Frameworks and Methods

While it is essential to have a team of skilled data scientists and AI engineers to pick the right AI frameworks and build complex AI systems, for large organizations, there are many other considerations to watch for. Not being able to understand the needs of an organization and where AI can make a difference is a risk. With the wrong target, every strategy will fail.

Collaboration between Analysts and IT

In some companies, IT provides the platforms that analysts have to use. If these platforms are error-prone or old, it can get frustrating for analysts. In modern environments, not all analytical platforms must be managed by one central IT department. This can give data teams more freedom to operate on their own.

IT

Many IT teams lack the resources to build the data pipelines needed for data science platforms. Often there is a gap between business users and engineers, making it hard for them to communicate with each other.

IT, especially operations, is often focused on preventing problems. As a result, many strive to protect their systems from change. They want to make it difficult to access data in order to keep platforms secure. Data scientists, however, would like to access data easily, to make progress quickly. This can lead to friction between both teams.

Costs

Introducing machine learning, data science and artificial intelligence can be expensive. It is rarely easy to say how much data science will impact the company's results: the relationship between inputs (such as time, effort, and resources) and outputs is anything but deterministic. The alternative, however, is even more grim. A company that is not ready to invest in innovation, will eventually lose its competitiveness and risk bankruptcy.

Data and Privacy Protection

Data and Privacy Protect may slow down some projects and make them more bureaucratic, but they are absolutely necessary. In addition, it's most likely that nobody wants to live in a system where privacy is not respected. In a day-to-day job, privacy protection is process-driven: Making these processes transparent and efficient.

Hiring

Introducing data science may require more significant changes in the corporate structure or culture, which could reveal hidden conflicts and challenges nobody wants to talk about.

Data professionals are a rare breed, and as there are few of them on the job market, how can a company even think about change without the required skills? Attracting engineers and scientists often requires an offer that goes beyond free fruit in the office.

16 Generative AI and Large Language Models

Katherine Munro, Gerald Hahn, Danko Nikolić

Questions Answered in this Chapter:

- What is Generative AI and how does it relate to other AI and Machine Learning techniques?
- Which characteristics define and differentiate Large Language Models?
- How can you use prompt engineering to get the best out of Generative AI models?
- Which design patterns are useful for building LLM-based applications?
- How can you customize an LLM to improve performance on your specific use case?
- Which vulnerabilities and limitations of Generative AI should you be aware of?
- How can you build robust, reliable, and effective Generative AI-powered applications?

16.1 Introduction to "Gen AI"

The first machine learning models were made to be predictive. They were designed to take an input and generate an output, often expressed as a category or numeric value. For example, we might ask a model to identify the category of an object in an image, or to predict the most likely price of a stock a week from now. The goal, ultimately, was to assign new information to some existing piece of data.

Generative models, by contrast, are designed to generate entirely new data. They can create rich outputs in various modalities, including text, images, and video. So, instead of asking a **predictive model** to classify an image as containing a cat or a dog, for example, we can ask a generative model to create an image depicting a cat, or a dog, or just about anything else we can imagine.

Generally, creating such complex outputs is a much more difficult task, due to the great number of intricate dependencies between even the minutest components of each output. To generate an image, for example, a model cannot simply spit out each of the millions of required pixels independently. Instead, the content generated in one part of the image will depend on what has been generated everywhere else. For example, in real-life photographs of building interiors, paintings tend to hang on the walls; they do not lay on floors, nor are they attached to ceilings. Such "rules" are obvious to us, of course, because we've learned them through a lifetime of observing different interiors for ourselves. A **Generative AI** model needs to learn such rules in a similar way: by being exposed to thousands of images depicting different types of buildings and rooms.

Learning such a huge number of complex dependencies requires a huge amount of input examples. Hence, a generative model usually requires a lot more training data than a predictive one. The "rules," or dependencies, can be understood as multi-dimensional distributions that need to be approximated. These distributions indicate the likelihood that a certain combination of features occurs in the data. For example, pixels representing a photo frame will occur more often in combination with pictures representing a wallpapered wall than they do with pixels representing a carpeted floor. Once an image generation model has learned such dependencies between all sorts of possible combinations of all sorts of possible pixel features, it can generate images which obey them. And using this principle, we can create different modalities of generative models by exposing the right machine learning algorithms to different kinds of inputs: texts, programming code, music, biological sequences, and more.

With enough data and enough computational power, it is possible to achieve impressive results with Generative AI (hereafter referred to interchangeably with "Gen AI"). The learned approximations of feature distributions are usually not perfect, of course, but neither is **predictive AI**. Much like predictive AI may mistake a chihuahua for a muffin in an image classification task, for example, a generative model may create images that would be impossible in real life. When generating images of humans, for instance, Gen AI has a hard time sticking to five fingers on each hand, often drawing four or six fingers instead. But despite such glitches, Gen AI has turned out to be immensely useful, and has vastly widened the general applicability of AI in everyday life.

16.2 Generative AI Modalities

As we just mentioned, generative AI algorithms can be used to train all kinds of "content generators", provided that enough data are available. For example, in the audio domain, we can generate spoken voice or music. In chemistry and biology, AI can generate molecular

structures and protein sequences. Time series and graphs are yet two more interesting possibilities, and so the list goes on. There is no limit to the type of modality to which Gen AI algorithms can be applied.

When we talk of "modality", we don't just mean the type of output that an AI model generates. The inputs that trigger the generation are equally relevant, and can be equally varied. The simplest models, known as **uni-modal models**, use the same mode of data for both inputs and outputs. For example, in Large Language Models, which we'll discuss in detail later, text input is used to generate text outputs.

We can also combine modalities to create **multi-modal models**. Textual inputs can be used to generate images, or, conversely, we can generate text outputs (descriptions) based on image inputs. Information from a piece of text can be converted into a graph, and graphs can be described in natural language. Even the inputs and outputs themselves can be multi-modal. For example, a model may be fed an image and some instructions describing what needs to be done with that image, and it will output an image with the required changes. It could even produce audio or visual content to match the input: it all just depends on how the model was trained.

Much like generative models need to be trained separately for different modalities, it is often useful to separately train them for specific domains within one specific modality. For example, computer code is nothing but text. Nevertheless, we're likely to get a better code generation model if we train a model specifically for that task, and feed it only inputs that include code (perhaps along with explanations and documentations of code). Similarly, it is a good idea to separately create generative models for videos and images, even though a video is technically nothing more than a series of images.

The landscape of common tools and publicly available models is changing quickly. Below we list a few popular examples (at the time of writing) for the various modalities we've just discussed:

- **Text:** ChatGPT, Bing Copilot, Gemini, LLaMA, Claude
- **Images:** Imagen, Stable Diffusion, Midjourney, DALL-E
- **Music:** MusicLM, Soundraw.io, Amper Music, Humtap, Stable Audio
- **Video:** D-ID, Gen-2 from Runway, Pictory, Synthesia, Fliki, Sora
- **Code:** GitHub Copilot, Codey, Tabnine, Polycoder, DeepCode

One may wonder how combining all these diverse modalities into single models is possible. The secret lies in the **encoder-decoder architecture** of multi-modal foundation models (discussed in the previous chapter). The encoder takes the input and transforms it into an intermediate representation, known as an "embedding", which is then used as input to the decoder. A nice property of this intermediate embedding representation is that it is independent of the modality. The embedding is said to describe the semantics of the input: that is, its meaning. For example, you could provide a multi-modal model with either the text, "a horse is passing near an oak tree," or with an image depicting just that. In either case, the algorithm may generate very similar embeddings, under the hood. This is how AI abstracts the inputs from different modalities. It uses two different encoder models, one for text and another for images, with both encoders being trained to work with the same embedding space. Similarly, the same AI may pass these intermediate representations to different decoder models, each trained for a different modality. One may generate sounds, the other images,

and yet another one, texts, but all are working in the same embedding space. Therefore, the trick of multi-modal generative AI lies in the encoder-decoder structure.

16.2.1 Methods for Training Generative Models

While we hope to have made the general concept of training generative models clear, we cannot possibly squeeze in a discussion of all types of architectures and algorithms within this chapter: such a task would fill a book in itself! Those who wish to understand more about machine learning foundations can visit Chapter 12, which even deep dives into two types of models commonly used for image generation: Generative Adversarial Networks (12.6.10) and Autoencoders (12.6.8). Large Language Models, including how to train and fine-tune them, will be discussed next. So, as you read on, remember that virtually anything which can be treated like a text sequence can potentially be modelled via LLMs: this includes text, code, protein sequences, and much more.

16.3 Large Language Models

16.3.1 What are "LLMs"?

Since the launch of OpenAI's **ChatGPT** at the end of 2022, the landscape of Artificial Intelligence text generation has undergone a remarkable transformation. This groundbreaking development marked the rise of accessible AI-powered text generation, captivating the public's imagination and sparking widespread interest. Built upon the foundation of **Large Language Models** (LLMs), ChatGPT represented a significant leap in the capabilities of AI-generated text. While earlier models (such as ELMo and BERT, which will be discussed in Chapter 17) had certainly exhibited much promise, they were primarily of interest to academics and specialists; ChatGPT, on the other hand, rapidly became a tool for anyone to use and explore.

Since its beginning, the field of **LLMs** has evolved rapidly, with companies racing to develop ever more sophisticated and powerful models. This fast pace of innovation has driven the technology forward at an unprecedented rate, pushing the boundaries of what was once thought possible in the domain of AI-driven text generation.

Conceptually, a language model is an Artificial Intelligence system designed to predict the next word in a sequence based on the preceding words or context. In their early implementations, these models made their predictions based on only a few preceding words; Now, thanks to advancements in machine learning and, in particular, neural network algorithms, language models can consider vast sequences of words, leading to more accurate predictions.

Mathematically, the prediction of the next word is framed as calculating conditional probabilities. This involves assessing the likelihood of a specific word occurring given the context provided by the preceding words. The word with the highest probability within the model's

vocabulary is then chosen, in a process known as **sampling**. The chosen word is then appended to the existing text, and the process repeats iteratively until a predetermined stopping condition is met. This **autoregressive mechanism**, also known as **causal generation**, ensures that the model only considers previously generated words and not those predicted afterward, as would be the case in bidirectional approaches.

Over time, language model methodologies have undergone significant changes. Initial approaches relied on simple n-gram models before transitioning to **neural network architectures** such as feedforward networks [1], convolutional [2] and recurrent neural networks [3], including variants like long-short term memory networks [4]. However, the most significant advancement came with the introduction of **Transformer models** [5], which revolutionized the field with their self-attention mechanism. Their architecture allows models to consider a broad context within a sentence, significantly enhancing their predictive capabilities. Chapter 17, Natural Language Processing (NLP), presents all these algorithms in detail, showing how each new development built upon previous successes to bring us to the revolutionary moment NLP is enjoying today.

More recently, newer architectures like **Eagle** [6] and **Mamba** [7] have demonstrated competitive performance without relying on attention mechanisms or employing state-based models. These models promise comparable performance to traditional Transformer-based LLMs of similar size, while being more computationally efficient and enabling faster inference.

The ability of language models to predict the next words and generate coherent text stems from their training on vast amounts of data sourced from the internet. Through learning the statistical probabilities inherent in language patterns, these models become adept at generating meaningful responses to queries and crafting creative pieces of text, such as emails or stories. With recent advancements, they've also gained limited capability to reason through complex tasks (see Section 16.3.3.3).

Large Language Models vary in several key aspects, which contribute to differences in their capabilities:

Performance on benchmark tasks: LLMs are often evaluated based on their performance on benchmark tasks, which serve as standardized tests to measure their effectiveness. Examples of benchmark tasks include language understanding tasks like question answering and text classification. LLMs may excel in certain tasks while performing less optimally in others, depending on their design and training.

Open source vs. closed source: LLMs can be further classified as either open source or closed source. Open-source models provide access to their architecture and parameters, allowing researchers and developers to modify and fine-tune them for specific applications. Closed-source models, on the other hand, restrict access to their internal teams and are typically only provided to the public as pretrained models through APIs or licensed software.

Number of parameters: LLMs come in different sizes, usually quantified by the number of parameters they possess. Small LLMs ("SLMs") might contain a few billion parameters, whereas larger models can encompass hundreds of billions. Generally, the parameter count correlates with the model's complexity and capacity to capture nuances in language (though much research is attempting to test the limits of this relationship).

Algorithms: LLMs leverage different algorithms for text generation, with the Transformer architecture being the most prevalent. Newly developed variations of the original attention mechanism (such as [8]) can enhance a model's ability to capture long-range dependencies and contextual information. Additionally, some LLMs may utilize alternative architectures, such as recurrent neural networks or state space models.

Training data: The size and quality of the training data significantly impact the performance of LLMs. Models trained on larger and more diverse datasets tend to exhibit superior performance due to their exposure to a broader range of linguistic patterns and contexts. Even models with the same parameter count can demonstrate substantial differences in performance based on the quality and quantity of their training data. Fine-tuning existing models on specialized datasets (see Section 16.3.4.2) can further enhance their performance for specific tasks, boosting it beyond that of larger, general-purpose models.

16.3.2 How is Something like ChatGPT Trained?

Pre-training

Off-the-shelf Large Language Models undergo a comprehensive training process before they are made available to users through APIs or downloadable via platforms like HuggingFace. This process starts with an initial **pre-training** phase, where the goal is to train the model to predict words in a sequence. Often this is done by exposing the model to an extensive amount of text data scraped from the internet, randomly masking out words, and having the model fill the gaps. Initially, the model's predictions will be quite random. But as training progresses, guided by an objective to minimize errors in the model's next-word predictions, the model's parameters are iteratively adjusted until it has learned to capture the intricate syntactic and semantic relationships in the text. In other words, through exposure to natural language data, the model gradually gains a statistical "understanding" of which words make sense in different contexts, and in combination with which other words.

Such an approach is known as **self-supervised learning** [9] because it enables the model to learn from data without requiring external labels, relying instead on the inherent structure of the text itself. Unlike in supervised learning, no explicit labels are provided during this process, as the missing words are already known from the text. The quantity and quality of this data are paramount, however, as they directly impact the model's ability to learn and generalize from the information provided.

Instruction Fine-tuning

Pretraining LLMs is a fundamental step in their development, providing them with solid foundational knowledge (hence the term "Foundation Models", the title of the previous chapter). However, simply predicting missing words in a text doesn't necessarily serve much purpose in the real world, which is why we next conduct a supervised learning step, known as **instruction fine-tuning** [10]. The purpose of this phase is to train the model to follow human instructions more precisely. This is achieved by providing the model with labeled data consisting of examples that illustrate how it should respond to specific questions or instructions. Through exposure to explicit examples provided by humans, the model learns to better understand and execute tasks according to human expectations.

Reinforcement Learning from Human Feedback

This phase, though effective, still leaves one problem unsolved: the potential for LLMs to generate outputs that violate human values, resulting in toxic, harmful, or biased responses. To mitigate this problem, reinforcement learning techniques are used in combination with human feedback [11]. In this approach, known as **Reinforcement Learning from Human Feedback** (RLHF, see also Section 16.3.4.2.3), the model receives feedback on the appropriateness of its responses, and adjusts its behavior during training accordingly. This helps ensure that the model's outputs align with societal norms and values.

This combination of pre-training and fine-tuning via instructions and human feedback have been crucial to the success of models like ChatGPT and its successors. The first stage provides the model with broad, general language knowledge, and the latter two help make it particularly adept at enacting human requests. If you're thinking that all these training stages sound complex and resource intensive, however, you're right. Pretraining demands significant computational resources and cutting-edge hardware infrastructure, including high-performance computing clusters and specialized processing units optimized for deep learning tasks. As a result, only a handful of companies worldwide have the infrastructure and expertise necessary to pretrain LLMs. Instruction fine-tuning and RLHF are more feasible for smaller organizations, but nevertheless require carefully curated datasets and significant manual effort. Fortunately, many off-the-shelf LLMs have already been pretrained and fine-tuned, making them useful for many applications straight out of the box.

In the following section, we'll discuss ways to use LLMs directly. Then, in Section 16.3.4, we'll discover how to further customize them for your specific needs.

16.3.3 Methods for Using LLMs Directly

For many people, their first interaction with LLMs was directly through a chat interface, such as ChatGPT. Even development teams may start an LLM initiative by testing their idea directly with a publicly available chatbot, which, given the right instructions, can already achieve an impressive number and variety of tasks. When more is required of the LLM, however, two key design patterns often come into play: augmenting the LLM with additional information to help broaden its knowledge base, and providing it with access to tools, with which to execute more complex tasks. The following three subsections explore at each of these options in turn.

16.3.3.1 Direct Interaction via Prompting

16.3.3.1.1 Zero- to Few-Shot-Inference

The public success of LLMs is not solely attributed to their performance on NLP benchmarks, but also, to the way they enable human interaction. With traditional machine learning models, programmers wrote the code for training the model and directing it to perform tasks. Consequently, proficiency in programming languages was essential. However, LLMs operate differently. They allow direct interaction for everyone. These AI systems can learn from and respond to instructions, known as "**prompts**," given in everyday language, whether spoken or written. Their responses, called "completions," are also given in natural language,

as if the user and the "bot" were engaging in ordinary human conversation. This significantly improves accessibility, reaching a broad audience beyond just experts.

Pretraining the model serves not only to help it comprehend natural language instructions, but also to tackle problems out-of-the-box, without requiring the user to worry about model training. For instance, the model could be prompted to categorize a message as positive, negative, or neutral, without the need for explicit training examples. This capability, called **zero-shot inference**, essentially relies on the model's pre-training knowledge alone to fulfill the assigned task.

When zero-shot inference is not enough, such as with smaller language models and/or more intricate tasks, users can provide the model with a few examples illustrating how to tackle a task. This method has multiple names, including **few-shot learning**, **few-shot inference** [12], and **in-context learning**. To instruct the model to classify customer enquiries by service line, for example, one might present sentences such as, "I would like to extend my payment deadline", "I need to reset my password", and "I want to make a purchase", along with their corresponding labels: "billing", "assistance" and "sales". Through these examples, the model acquires knowledge, enabling it to generalize and address similar problems. Note that the term "learning" here is somewhat ambiguous since the model's internal parameters — its weights — remain unchanged, and once the examples are removed from the prompt, the model will forget them.

While zero- and few-shot-inference can be very powerful, examples alone are not guaranteed to get the best out of an LLM. The way the examples and any other instructions and context are provided to the model are also vitally important, and bring us to the art of "**prompt engineering**".

16.3.3.1.2 Effective Prompt Engineering

The discovery that language models can be instructed using everyday language has led to much excitement, hearsay, and research, on the most effective ways to craft such instructions. Finding "universal rules" to suit all LLMs is a challenge: a highly effective prompt for one LLM might yield subpar results from another. Consequently, a specialized role has emerged: the **prompt engineer**. They have the expertise needed to write effective prompts, honed through experience and a deep understanding of LLM behavior.

If you dream of becoming a prompt engineer, or just want to improve your results with LLMs, we recommend you practice the following techniques, using a variety of models, and stay on the lookout for newly emerging advice in this field.

Ask the question you actually want answered:

It's easy to just start chatting with an LLM, making requests quite spontaneously, and getting outputs that don't quite meet your needs. Perhaps you framed the question badly, or perhaps you asked the wrong thing entirely: you may realize you wanted help with a different problem altogether.

The following prompt engineering best practices can help with this, as they each relate to how you frame the problem the LLM should solve.

- First and foremost, always **be clear and concise**. Read your prompt back to yourself before you submit it, ensuring that related ideas are close to one another and there is no

redundant or duplicated information. Rethink and rephrase, if necessary. This will force you to get specific about the problem you need solved. It will also help the LLM pay attention to the core issue, and not get sidetracked by unnecessary details. And note that a concise prompt need not be short. It should be as thorough and detailed as is needed to clearly convey your requirements.

- Another best practice is to **simplify the problem**. Try turning open-ended requests for information into closed classification questions: Instead of *"How should I set up a machine learning monitoring tech stack?"*, ask, *"Which of the following tools are most appropriate for machine learning monitoring, given that I currently use Kubernetes and Google Cloud Platform: Tool X, or Tool Y?"* You can also have the LLM choose from a provided list of possible answers, instead of coming up with its own interpretations: Instead of *"Which topics can you identify in this customer inquiry email?"*, try, *"Which of the following products and services are mentioned in the provided customer inquiry email? Products & services: {}. Customer Email: []."* Of course, here you would insert your products or services within the {} placeholder, and the customer email within the [].
- **Provide constraints**, to ensure an LLM's output is actually useful to you. This is particularly relevant when using it as an ideation tool or brainstorming partner. Say you want help drafting product descriptions for a new advertising campaign: specify the desired text length, provide certain banned keywords (such as those associated with your competitors), and tell the model to only use product features you'll provide within the prompt.
- Another tip is to **specify the context and target audience**. One way to do this is implicitly, simply by changing the style of your own prompt, and letting the model adapt its style to match. Thus, a question like *"What are the benefits of taking cold showers?"* could result in an academic, yet impersonal response, whereas asking *"Why and how should I incorporate cold showers into my daily routine?"* will likely generate a much more casual, personal output. Instead of changing your own tone, you can also simply state your desired context or target audience directly. This is particularly convenient for documents with well-defined and well-known formats, as you don't need to spend a lot of time specifying those formatting rules: Ask an LLM to create an Instagram Post, LinkedIn newsletter or Google Ads text, for example, and it will adapt its language, output length, and use of hashtags and emojis accordingly. Adding the target audience will also help ensure that the resulting text is appropriate and appealing for the desired final reader.
- This leads us to the next best practice, which is to **describe the input format** so that the LLM knows how to handle it, such as being able to differentiate between instructions and additional context information. For example, try something like, *"You will be provided with a company document, denoted in hashtags, and an employee question, denoted by angular brackets. Use information from the document to answer the question. The document: ##. The question: <>."* Again, here you would insert the document between the ## symbols, and the question between the <>. As this example shows, it can also help to use special characters to structure the prompt more clearly for the model.
- Similarly, **describe the output format** exactly as you need it, keeping in mind the downstream task you need the LLM's output for. Turn unstructured documents into structured data, for example, by having an LLM extract certain expected entities into a JSON string featuring {Entity A: Value, Entity B: Value}. Such an output can easily be fed to code tools or programs, or saved as a table.

Guide the model through the task:

While the above tips will help you frame your problem and desired output clearly and accurately, it can also be very helpful to instruct the LLM on how it should work through your request. These additional best practices can help you do just that:

- First, **give the model examples** showing how to complete the task (as we discussed in the previous section). This may even be quicker, easier, and just as effective as writing a detailed prompt: provided your examples are good, and the task can easily be inferred from them. For example, if your problem involves taking a lot of similar inputs and systematically driving some result from them, you might not need any instruction at all: You can simply provide a few input-output pairs, and then start providing inputs only, for the LLM to complete.
- Given the nature of LLMs to people-please and hallucinate (more on that, later), it's also important to **give the model an escape route**, should it need one. That is, tell it what to do if it can't solve a problem: it could defer to other tools (assuming it's part of an agent system), return a consistent null value such as "NA" or "not specified", or simply state that it doesn't know.
- Speaking of challenging problems: you can also **tell the model to work step-by-step**. This has been shown to help LLMs execute logical and mathematical problems more effectively (though they're still no master mathematicians, yet!). You can also describe the control flow that the overall interaction should take. For example, when crafting a chatbot assistant application, you could specify the order in which the model should ask for specific information from the user, including constraints so that it only attempts to solve the task once all required pieces of information have been requested and received.
- Another simple yet effective practice is to **ask for multiples and variations**. Inexperienced users may write a prompt like, *"Generate a title for a blog post about subject X,"* get disappointed by the result, and give up. A far better strategy is to ask for *N* different variations, pick the best one(s), and then iterate, explaining to the model what worked well and what it should try more of.
- Finally, you can **use structured prompts, or templates**. This is an effective way to use LLMs inside larger applications, but even everyday users who predominantly interact with public chatbots can use this technique. Imagine you've just spent a reasonable amount of time conversing with a chatbot, refining and rephrasing your prompt, and adding additional requirements and constraints as you observed the shortcomings of the model's answers. Once you've gotten the output you really needed, ask yourself, will you ever have to solve a similar task again? If so, it can be wise to pack the most effective parts of your conversation into a new prompt, try it out, and if it works, save it somewhere (popular chatbots such as ChatGPT and Google's Gemini include a history function for this purpose). You can also share with colleagues, so that everyone benefits from your fantastic prompt engineering skills!

While these prompt engineering techniques can help you get the best out of a pretrained model, for some tasks, more sophisticated techniques can be used. One option, which we'll discuss next, is to provide the model with additional documents from which it can retrieve information required for a given task. Another is to fine-tune the model with customized data, which we'll cover in Section 16.3.4.2.

16.3.3.2 Retrieval-Augmented Generation

16.3.3.2.1 Introduction to RAG Systems

Relying only on prompting to solve a task can cause the LLM to give incorrect or made-up answers, known as "hallucinations" or "confabulations" [13] (see Section 16.4.3). This can happen because the model does not have enough specific knowledge for those tasks, or because its knowledge has a date cutoff based on whenever its pre-training data was collected. Another concern is that even if an answer is correct, we might not know where it came from because the information it used is hidden within the pre-training data. But sometimes we want or even need to know the source of a model's responses, either to help us understand and improve it, or to be certain we are building explainable, trustworthy AI.

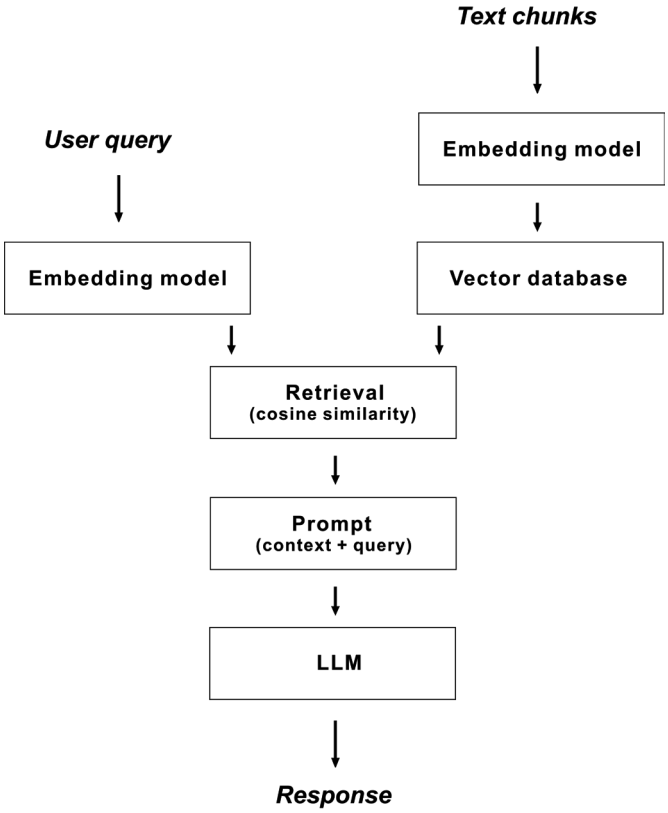


Figure 16.1 Flow of retrieval-augmented generation. User queries and text chunks from relevant documents are embedded using an embedding model. Relevant chunks are retrieved from a vector database based on a similarity measure between the query and chunks. These relevant chunks form the context, which is then added to the prompt, together with the query. The prompt is subsequently fed to the LLM, resulting in the generation of a response.

Theoretically, we could tackle both issues by giving the LLM a lot of information directly in the prompt, such as all of our company's internal documentation. This would provide it with the knowledge it needed, and make its responses more transparent. However, in practice, an LLM's small context window might not fit all the needed documents. Even if it could, we have the token-based processing costs to think about. Finally, this may also lead to **positional bias**: When you load the prompt with lots of information, the LLM tends to focus more on the beginning and end of the window, possibly missing important details in-between [14].

Retrieval-Augmented Generation (RAG) [15] aims to solve these challenges by only adding additional *task-related* information from an external database into the prompt, rather than trying to fit all the needed knowledge into the limited context window (Figure 16.1). A RAG system works like a search engine, finding and retrieving relevant documents to help the LLM with the task at hand. A notable advantage of RAG is that it can easily adapt to changes in the external database. If the database is updated, the LLM can use the new information without needing to retrain its weights. This flexibility allows the LLM to quickly adapt to new data. Also, by focusing on the information in the context window, RAG can reduce problems like hallucinations. Moreover, users can see what information the LLM used to make its decisions, which improves transparency and interpretability.

The RAG process starts when a user asks a question (writes a query) and ends when the LLM gives an answer or solution. This process happens in a few steps: indexing, retrieval, and generation.

Indexing: During indexing, documents are collected, parsed, and stored in a database. However, for a specific question, only a few paragraphs of a specific document might be relevant. Thus, documents are segmented into chunks, usually with some degree of overlap. Next, the text chunks are transformed into embeddings using an embedding model. These are then stored in a vector database, also called index, designed specifically to store and efficiently query embeddings.

Retrieval: The user question is transformed into an embedding with the same model used to embed the text chunks during the indexing stage. The resulting embedding is then compared with all embeddings in the database using similarity measures such as cosine similarity. This method is also known as "semantic search" because it considers the meaning and context of words, unlike a basic search using keywords only. The top-k documents with the highest similarity scores, typically about 3 to 5 text chunks, are selected.

Generation: The chosen top-k retrieved chunks, which we call the "context", are integrated into a structured prompt, together with the original query. The LLM uses this context to generate an answer which is based less on its own internal knowledge than would have been, and instead, is more grounded in the information in the retrieved texts.

29 The Authors

Katherine Munro is a Data Scientist, Data Science Ambassador and Computational Linguist, conducting research and development and corporate training in AI, Natural Language Processing and Data Science. Katherine began her tech career specializing in user interfaces and Natural Language Understanding, with roles at Mercedes-Benz and the Fraunhofer Institute. She then transitioned to data science in the e-commerce and insurance domains, before landing her current role building smart conversational AI systems using NLP techniques and Large Language Models. In her free time, Katherine is an avid tech writer on X and her own newsletters, and is a volunteer for diverse initiatives helping women and girls start their own tech careers.

Stefan Papp is an entrepreneur who works with Fortune 500 companies to build data platforms and helps them to become more data-driven. Living with his family in Armenia, he is also involved in the Armenian startup ecosystem, and he acts there as an advisor and investor.

Although he has a background in computer science, he strongly believes that the transition to AI and data science is more about culture than about technology. As a libertarian, he sees the benefits and opportunities of data to address current global challenges such as climate change or demographic change.

Zoltan C. Toth is a data engineering architect, lecturer and entrepreneur. With a background in Computer Science and Mathematics, he has taught data architectures, big data technologies and machine learning operations to Fortune 500 companies worldwide. In the past two decades he has worked with several large enterprises as a Solutions Architect, implementing data analytics infrastructures and scaling them up to processing petabytes of data. He is also a lecturer at the Central European University. He founded Datapao, a data engineering consultancy that became Databricks's European professional services center and a Microsoft Gold Partner in Data Science.

Wolfgang Weidinger is a Data Scientist and AI professional. He has worked in a wide variety of industries and sectors such as start-ups, finance, consulting, wholesale and insurance. There he led Data Science & AI teams and drove their role as spearheads in digital and data-driven transformation.

He is President of the Vienna Data Science Group (www.vdsg.at), a non-profit association of and for Data Scientists and all other Data & AI professionals. This brings together both research and practice across a wide range of industries. The VDSG is a rapidly growing international community whose goal is to educate about Data Science and its subfields such as Machine Learning and Artificial Intelligence, as well as their impact on society.

Wolfgang is particularly interested in the societal impact of Data Science and AI, as well as the establishment of interdisciplinary Data Science & AI teams in companies and their disruptive impact on business models. He enjoys lecturing and developing Data Science & AI solutions end2end.

Dr. Danko Nikolić is an expert in both brain research and AI. For many years he has run an electrophysiology lab at the Max-Planck Institute for Brain Research. Also, he is an AI and machine learning professional heading a Data Science team and developing commercial solutions based on AI technology. He invented AI Kindergarten–a concept for training AI of the future for achieving near human-level intelligence. He also pioneered using machine learning to read "minds" from the brain's electrical signals; he and his team were able to reconstruct what an animal was seeing solely by analyzing the brain signals. He introduced the concept of ideasthesia ("sensing concepts") into neuroscience and is the author of a theory called practopoiesis describing how biological systems achieve intelligence. He has a degree in Psychology and Civil Engineering from the University of Zagreb, Croatia and a PhD from the University of Oklahoma, USA. He was an honorary professor at the university of Zagreb from 2014 to 2019.

Barbora Antosova Vesela is a data scientist and software engineer working at Frequentis, which operates in a safety critical communication and information environment. Her background is a study of biophysics at Masaryk University in Brno and biomedical engineering both at FH Technikum Wien and Brno University of Technology. She is interested in various topics combining data science and signal and image processing applied in multiple environments, such as medicine, research and air traffic management.

Dr. Karin Bruckmüller studied law and is a criminal lawyer at the Sigmund Freud Private University in Vienna and the Johannes Kepler University in Linz. In both research and teaching, she focuses on medical and nursing criminal law in connection with ethics. She is a regular speaker at relevant nursing congresses and conferences such as the Nursing Congress 2019 in Vienna.

Dr. Annalisa Cadonna is a statistician and data science consultant. She received her Ph. D. in Applied Mathematics and Statistics from University of California, Santa Cruz. Annalisa has applied statistical and machine learning metho ds to deliver projects in the financial, energy and medical industries. Currently, her professional goal is bridging the gap between time series research and industry applications, by using probabilistic programming and cloud technologies. Annalisa strives to use statistics and machine learning as means for the achievement of the Sustainable Development Goals and to be active in the development of tools and frameworks for responsible artificial intelligence. She is also one of the organizers of R-ladies Vienna.

Dr. Jana Eder is an expert in machine learning and information technology. She earned her Ph. D. in AI and medical imaging through a collaborative effort between Paracelsus Medical University and ETH Zurich, exploring the cutting-edge intersection of AI and healthcare. Following her Ph.D., she completed a habilitation on the use of AI in imaging diagnostics for osteoarthritis. Jana worked as a lecturer at Paracelsus Medical University, Paris Lodron University Salzburg, and the University of Vienna. Currently show holds a position as Senior Key Expert for AI focusing on data efficient learning and multi-source learning at Siemens Technology.

Dr. Jeannette Gorzala, BSc is an attorney at law and AI governance expert working with renowned startups, companies and public organizations to implement and design trustworthy AI systems and business models. As an AI literacy coach, she helps to build up AI know-how and design roles and processes allowing organizations to safely and successfully unlock the potential of AI. She has a background in law and international business administration with more than ten years of experience in investment banking and international law firms with a focus on highly regulated industries and the technology sector.

Jeannette is dedicated to positively shaping the AI ecosystem via her involvement in the independent think tank AI Austria and the European AI Forum, via which she brought together nine international AI associations to give European founders a voice in law and policy making.

As Vice President of the European AI Forum, Jeannette represented more than 2,000 AI entrepreneurs in the legislative process for the Artificial Intelligence Act, advocating for a collaborative and innovation-friendly approach while adequately addressing risks and challenges. Further, Jeannette acts as the Deputy Chair of the AI Advisory Board to the Austrian Government

Dr. Gerald Hahn is proficient in Python, setting up and training deep learning models for NLP and computer vision, pytorch, transformers, classic machine learning algorithms, supervised and non-supervised methods, Bayesian approaches, implementing state of the art literature in machine learning, +10 years of experience in analyzing and modeling neuroscience data.

Dr. Georg Langs is a full professor for Machine Learning in Medical Imaging at the Medical University of Vienna, where he heads the Computational Imaging Research Lab at the Department of Biomedical Imaging and Image-guided Therapy. He is co-founder and chief scientist of the spin-off contextflow GmbH, which develops software for AI-based image search. Georg Langs studied mathematics at the Vienna University of Technology and computer science at the Graz University of Technology, and was a Research Scientist at MIT's Computer Science and Artificial Intelligence Lab, where he is still a Research Affiliate.

Dr. Roxane Licandro is a postdoctoral research associate at the Medical University of Vienna and a research fellow at the Massachusetts General Hospital and Harvard Medical School. She graduated from her medical computer science studies at TU Wien, where she worked as a university assistant at the Computer Vision Lab. She was awarded a Marie Skłodowska-Curie Fellowship and completed research stays at Charité Berlin, Children's Hospital Zurich and University College London. She worked at the Kunsthistorisches Museum Wien and at Agfa Healthcare. Her research focus lies on finding new ways to computationally model and predict dynamic processes in space and over time, paediatric and fetal brain development, statistical pattern analysis in cancer research and geometric shape analysis of anatomical and cultural objects.

Christian Mata is a Business intelligence & Analytics Consultant with more than two decades of BI & reporting project experience for well-known international companies. He leads the development of information solutions and acts as a business analyst bridging business and IT.

He is passionate about leveraging data in organizations and empowering business people to make relevant data visible and accessible. As a trainer, he teaches practical data analysis, visualization and data literacy skills to close the gap between data and decisions.

Christian Mata founded Matadata e. U. in 2014 to strengthen the value-added use of data in organizations. He supports decision-makers at all levels in the cultural shift towards data-based management and in making informed decisions with confidence. Accelerating the transition to a sustainable economy and the implementation of the mobility and energy transition is his current focus.

Sean McIntyre is a Solutions Architect helping teams across Europe to deliver reliable analytics faster using the Modern Data Stack. Over his career, he has worked with software and data in the finance, life sciences, manufacturing, and retail industries. A former International Collegiate Programming Contest world finalist, he has been involved in organizing programming contests and coaching university programming teams for 10 years. He holds a B.Sc. degree in Computer Science and Pure Mathematics from the University of Calgary, and a Masters degree from the Interactive Telecommunications Program at New York University.

Mario Meir-Huber is an experienced senior executive working in different large international organizations. He leads data teams and fosters the cultural change for digitalization and data in these organizations. Before that, he worked in solution architecture roles in leading technology providers such as Microsoft and Teradata. In addition to his job, he is a featured speaker at various international events such as GITEX, WeAreDevelopers or London Tech Week and teaches MBA students on Data Strategy. Mario has already published several books on the topic of the Cloud and (Big) Data.

Gyorgi Mora is a data scientist with strong engineering skills. He designed and built large scale data infrastructures, search systems and machine learning applications.

Manuel Pasieka is an AI Solution Architect and Entrepreneur that helps companies make the right AI strategy and technology decisions, leveraging AI in a way that works for them.

He has a background in distributed and high-performance computing and has been working in several technology startups and research institutes in Austria and Spain.

Since 2021 he has been working as an independent AI consultant and in 2024 started his own AI consultancy and development company.

In addition, he is the host of the “Austrian Artificial Intelligence Podcast” that has the goal to highlight the Austrian AI landscape and its protagonists from academia and private industry.

Manuel is specially interested in the cyber security aspects of machine learning methods and their application in cybersecurity.

Victoria Rugli is a data consultant with a focus on analytics. She received her Master's degree in Data Science and Digital Marketing from Emlyon University in Paris, where she was a student research fellow at the Institute for artificial intelligence in management (AIM) as well as a bachelor's degree in International Business at WU in Vienna. Victoria has delivered projects in the financial, logistics and real estate industries. Currently, her professional focus is data governance and artificial intelligence strategy within cloud technologies. Victoria strives to accelerate the maturity of companies' data by improving the quality and use of data by developing and implementing frameworks for a sustainable data governance implementation. She co-hosts a podcast for artificial intelligence called AI Literacy.

Dr. Rania Wazir is a mathematician and data scientist focussing on Trustworthy AI, Natural Language Processing and Social Media Monitoring. She is a vice chair of Austria's Standards Committee on AI, and Austrian delegate to the ISO working group on Trustworthy AI; she is also coordinator of the VDSG's data4good initiative, which works with non-profits on data-based projects. She lead a consortium of machine learning, legal, and social science experts that recently completed an investigation into bias in algorithms for the EU Fundamental Rights Agency, and is currently tech lead in a three-year project to create a fair by design AI development process, funded by the Austrian Research Agency. Dr. Wazir is co-founder, together with open innovation expert Dr. Gertraud Leimüller, of the recent start-up leiwand.ai, whose goal is to provide companies and organizations involved in the development or use of AI systems with the tools and know-how necessary to ensure their systems are trustworthy.

Günther Zauner is a long-time employee at dwh GmbH, a mathematician and expert in the field of modeling and simulation, parametrization and forecast modeling. He is working on industrial projects as well as on research projects (e.g. EU FP7 CEPHOS-LINK, Horizon 2020 RheumaBuddy). He specializes in the development of modeling concepts, integration of routine data and population behavior. He is a member of VDSG, Society of Medical Decision Making (SMDM) and a member of the board of International Society for Pharmacoeconomics and Outcomes Research Austria (ISPOR Austria). Furthermore, he is reviewer of several journals, and he is doing a PhD study in the field of Public Health under the lead of Professor Majdan at the University of Trnava.

Index

B

C

D

E

F

G

H

I

N

O

P

Q

R

S

T

U

V