Foreword

A lot has happened since the first edition of this book was published in 2014. There is hardly a day where there is no news on data science, machine learning, or artificial intelligence in the media. It is interesting that many of those news articles have a skeptical, if not an even negative tone. All this underlines two things: data science and machine learning are finally becoming mainstream. And people know shockingly little about it. Readers of this book will certainly do better in this regard. It continues to be a valuable resource to not only educate about how to use data science in practice, but also how the fundamental concepts work.

Data science and machine learning are fast-moving fields which is why this second edition reflects a lot of the changes in our field. While we used to talk a lot about "data mining" and "predictive analytics" only a couple of years ago, we have now settled on the term "data science" for the broader field. And even more importantly: it is now commonly understood that machine learning is at the core of many current technological breakthroughs. These are truly exciting times for all the people working in our field then!

I have seen situations where data science and machine learning had an incredible impact. But I have also seen situations where this was not the case. What was the difference? In most cases where organizations fail with data science and machine learning is, they had used those techniques in the wrong context. Data science models are not very helpful if you only have one big decision you need to make. Analytics can still help you in such cases by giving you easier access to the data you need to make this decision. Or by presenting the data in a consumable fashion. But at the end of the day, those single big decisions are often strategic. Building a machine learning model to help you make this decision is not worth doing. And often they also do not yield better results than just making the decision on your own.

Here is where data science and machine learning can truly help: these advanced models deliver the most value whenever you need to make lots of similar decisions quickly. Good examples for this are:

- Defining the price of a product in markets with rapidly changing demands.
- Making offers for cross-selling in an E-Commerce platform.
- Approving credit or not.
- Detecting customers with a high risk for churn.
- Stopping fraudulent transactions.
- And many others.

You can see that a human being who would have access to all relevant data could make those decisions in a matter of seconds or minutes. Only that they can't without data science, since they would need to make this type of decision millions of times, every day. Consider sifting through your customer base of 50 million clients every day to identify those with a high churn risk. Impossible for any human being. But no problem at all for a machine learning model.

So, the biggest value of artificial intelligence and machine learning is not to support us with those big strategic decisions. Machine learning delivers most value when we operationalize models and automate millions of decisions. One of the shortest descriptions of this phenomenon comes from Andrew Ng, who is a well-known researcher in the field of AI. Andrew describes what AI can do as follows: "If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."

I agree with him on this characterization. And I like that Andrew puts the emphasis on automation and operationalization of those models—because this is where the biggest value is. The only thing I disagree with is the time unit he chose. It is safe to already go with a minute instead of a second.

However, the quick pace of changes as well as the ubiquity of data science also underlines the importance of laying the right foundations. Keep in mind that machine learning is not completely new. It has been an active field of research since the 1950s. Some of the algorithms used today have even been around for more than 200 years now. And the first deep learning models were developed in the 1960s with the term "deep learning" being coined in 1984. Those algorithms are well understood now. And understanding their basic concepts will help you to pick the right algorithm for the right task.

To support you with this, some additional chapters on deep learning and recommendation systems have been added to the book. Another focus area is using text analytics and natural language processing. It became clear in the past years that the most successful predictive models have been using unstructured input data in addition to the more traditional tabular formats. Finally, expansion of Time Series Forecasting should get you started on one of the most widely applied data science techniques in the business.

More algorithms could mean that there is a risk of increased complexity. But thanks to the simplicity of the RapidMiner platform and the many practical examples throughout the book this is not the case here. We continue our journey towards the democratization of data science and machine learning. This journey continues until data science and machine learning are as ubiquitous as data visualization or Excel. Of course, we cannot magically transform everybody into a data scientist overnight, but we can give people the tools to help them on their personal path of development. This book is the only tour guide you need on this journey.

Ingo Mierswa
Founder
RapidMiner Inc.
Massachusetts, USA