Foreword

Everybody can be a data scientist. And everybody should be. This book shows you why everyone should be a data scientist and how you can get there. In today's world, it should be embarrassing to make any complex decision without understanding the available data first. Being a "data-driven organization" is the state of the art and often the best way to improve a business outcome significantly. Consequently we have seen a dramatic change with respect to the tools supporting us to get to this success quickly. It has only been a few years that building a data warehouse and creating reports or dashboards on top of the data warehouse has become the norm in larger organizations. Technological advances have made this process easier than ever and in fact, the existence of data discovery tools have allowed business users to build dashboards themselves without the need for an army of Information Technology consultants supporting them in this endeavor. But now, after we have managed to effectively answer questions based on our data from the past, a new paradigm shift is underway: Wouldn't it be better to answer what is going to happen instead? This is the realm of advanced analytics and data science: moving your interest from the past to the future and optimizing the outcomes of your business proactively.

Here are some examples of this paradigm shift:

- Traditional Business Intelligence (BI) system and program answers: *How many customers did we lose last year?* Although certainly interesting, the answer comes too late: the customers are already gone and there is not much we can do about it. Predictive analytics will show you *who will most likely churn within the next 10 days and what you can do best for each customer to keep them.*
- Traditional BI answers: *What campaign was the most successful in the past?* Although certainly interesting, the answer will only provide limited value to determine what is the best campaign for your upcoming product. Predictive analytics will show you *what will be the next best action to trigger a purchase action for each of your prospects individually.*

- Traditional BI answers: *How often did my production stand still in the past and why?* Although certainly interesting, the answer will not change the fact that profit was decreased due to suboptimal utilization. Predictive analytics will show you exactly *when and why a part of a machine will break and when you should replace the parts instead of backlogging production without control.*

Those are all high-value questions and knowing the answers has the potential to positively impact your business processes like nothing else. And the good news is that this is not science fiction; predicting the future based on data from the past and the inherent patterns living in the data is absolutely possible today. So why isn't every company in the world exploiting this potential all day long? The answer is the data science skills gap.

Performing advanced analytics (predictive analytics, data mining, text analytics, and the necessary data preparation) requires, well, advanced skills. In fact, a data scientist is seen as a superstar programmer with a PhD in statistics who just happens to understand every business problem in the world. Of course people with such a rare skill mix are very rare; in fact McKinsey has predicted a shortage of 1.8 million data scientists by the year 2018 only in the United States. This is a classical dilemma: we have identified the value of future-oriented questions and solving them with data science methods, but at the same time we can't find the answers to those questions since we don't have the people able to do so. *The only way out of this dilemma is a democratization of advanced analytics.* We need to empower more people to do create predictive models: business analysts, Excel power users, data-savvy business managers. We can't transform this group of people magically into data scientists, but we can give them the tools and show them how to use them *to act like a data scientist*. This book can guide you in this direction.

We are in a time of modern analytics with "big data" fueling the explosion for the need of answers. It is important to understand that big data is not just about volume but also about complexity. More data means new and more complex infrastructures. Unstructured data requires new ways of storage and retrieval. And sometimes the data is generated so fast it should not be stored at all, but analyzed directly at the source and the findings stored instead. Real-time analytics, stream mining, and the Internet of Things become a reality now. At the same time, it is also clear that we are in the midst of a sea change: data alone has no value, but the hidden patterns and insights in the data are an extremely valuable asset. Accessing this asset should no longer be an option for experts only but should be given into the hands of analytical practitioners and business managers of all kinds. This democratization of advanced analytics removes the bottleneck of data science and unleashes new business value in an instant.

This transformation comes with a huge advantage for those who are actually data scientists. If business analysts, Excel power users, and data-savvy business managers are empowered to solve 95% of their current advanced analytics problems on their own, it also frees up the scarce data scientist resources. This transition moves what has become analytical table stakes from data scientists to business analytics and leads to better results faster for the business. At the same time it allows data scientists to focus on new challenging tasks where the development of new algorithms is a must instead of reinventing the wheel over and over again.

We created RapidMiner with exactly this purpose in mind: empower nonexperts to get to the same findings as data scientists. Allow users to get to results and value much faster. And make deployment of those findings as easy as a single click. RapidMiner empowers the business analyst as well as the data scientist to discover the hidden patterns and unleash new business value much faster. This unlocks the huge business value potential in the marketplace. I hope that Vijay's and Bala's book will be an important contribution to this change, supporting you to remove the data science bottleneck in your organization, and, last but not least, discovering a complete new field for you that delivers success and a bit of fun while discovering the unexpected.

Ingo Mierswa
CEO and Co-Founder, RapidMiner