Preface

Our goal is to introduce you to *Data Science*.

We will provide you with a survey of the fundamental data science concepts as well as step-by-step guidance on practical implementations—enough to get you started on this exciting journey.

WHY DATA SCIENCE?

We have run out of adjectives and superlatives to describe the growth trends of data. The technology revolution has brought about the need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways. *However, the value of the stored data is zero unless it is acted upon*. The scale of data volume and variety places new demands on organizations to quickly uncover hidden relationships and patterns. This is where data science techniques have proven to be extremely useful. They are increasingly finding their way into the everyday activities of many business and government functions, whether in identifying which customers are likely to take their business elsewhere, or mapping flu pandemic using social media signals.

Data science is a compilation of techniques that extract value from data. Some of the techniques used in data science have a long history and trace their roots to applied statistics, machine learning, visualization, logic, and computer science. Some techniques have just reached the popularity it deserves. Most emerging technologies go through what is termed the "hype cycle." This is a way of contrasting the amount of hyperbole or hype versus the productivity that is engendered by the emerging technology. The hype cycle has three main phases: peak of inflated expectation, trough of disillusionment, and plateau of productivity. The third phase refers to the mature and value-generating phase of any technology. The hype cycle for data science indicates that it is in this mature phase. Does this imply that data science has stopped growing or has reached a saturation point? Not at all. On the contrary, this discipline has grown beyond the scope of its initial

applications in marketing and has advanced to applications in technology, internet-based fields, health care, government, finance, and manufacturing.

WHY THIS BOOK?

The objective of this book is two-fold: to help clarify the basic *concepts* behind many data science techniques in an easy-to-follow manner; and to prepare anyone with a basic grasp of mathematics to *implement* these techniques in their organizations without the need to write any lines of programming code.

Beyond its practical value, we wanted to show you that the data science learning algorithms are elegant, beautiful, and incredibly effective. You will never look at data the same way once you learn the concepts of the learning algorithms.

To make the concepts stick, you will have to build data science models. While there are many data science tools available to execute algorithms and develop applications, the approaches to solving a data science problem are similar among these tools. We wanted to pick a fully functional, open source, free to use, graphical user interface-based data science tool so readers can follow the concepts and implement the data science algorithms. RapidMiner, a leading data science platform, fit the bill and, thus, we used it as a companion tool to implement the data science algorithms introduced in every chapter.

WHO CAN USE THIS BOOK?

The concepts and implementations described in this book are geared towards business, analytics, and technical professionals who use data everyday. You, the reader of the book will get a comprehensive understanding of the different data science techniques that can be used for prediction and for discovering patterns, be prepared to select the right technique for a given data problem, and you will be able to create a general-purpose analytics process.

We have tried to follow a process to describe this body of knowledge. Our focus has been on introducing about 30 key algorithms that are in widespread use today. We present these algorithms in the framework of:

1. A high-level practical use case for each algorithm.
2. An explanation of how the algorithm works in plain language. Many algorithms have a strong foundation in statistics and/or computer science. In our descriptions, we have tried to strike a balance between being accessible to a wider audience and being academically rigorous.

3. A detailed review of implementation using RapidMiner, by describing the commonly used setup and parameter options using a sample data set. You can download the processes from the companion website www.IntroDataScience.com and we recommend you follow-along by building an actual data science process.

Analysts, finance, engineering, marketing, and business professionals, or anyone who analyzes data, most likely will use data science techniques in their job either now or in the near future. For business managers who are one step removed from the actual data science process, it is important to know what is possible and not possible with these techniques so they can ask the right questions and set proper expectations. While basic spreadsheet analysis, slicing, and dicing of data through standard business intelligence tools will continue to form the foundations of data exploration in business, data science techniques are necessary to establish the full edifice of analytics in the organizations.

Vijay Kotu
California, USA
Bala Deshpande, PhD
Michigan, USA