Preface

This is a book for people who want to learn about the Julia language with a view to using it for data science. Some effort has gone into making this book suitable for someone who has familiarity with the R software and wants to learn about Julia. However, prior knowledge of R is not a requirement. While this book is not intended as a textbook for a course, some may find it a useful book to follow for a course that introduces statistics or data science students to Julia. It is our sincere hope that students, researchers and data scientists in general, who wish to learn Julia, will find this book beneficial.

More than twenty years have passed since the term data science was described by Dr. Chikio Hayashi in response to a question at a meeting of the International Federation of Classification Societies (Hayashi, 1998). Indeed, while the term data science has only gained notoriety over the past few years, much of the work it describes has been practiced for far longer. Furthermore, whatever the precise meaning of the term, there is no doubt that data science is important across virtually all areas of endeavour. This book is born out of a mixture of experiences all of which led to the conclusion that the use of Julia, as a language for data science, should be encouraged.

First, part of the motivation to write this book came from experience gained trying to teach material in data science without the benefit of a relatively easily understood base language that is effective for actually writing code. Secondly, there is the practical, and related, matter of writing efficient code while also having access to excellent code written by other researchers. This, of course, is the major advantage of R, where many researchers have contributed packages — sometimes based on code written in another language such as C or Fortran — for a wide variety of statistics and data science tasks. As we illustrate in this book, it is straightforward to call R from Julia and to thereby access whatever R packages are needed. Access to R packages and a growing selection of Julia

packages, together with an accessible, intuitive, and highly efficient base language, makes Julia a formidable platform for data science.

This book is not intended as an exhaustive introduction to data science. In fact, this book is far from an exhaustive introduction to data science. There are many very good books that one can consult to learn about different aspects of data science (e.g., Bishop, 2006; Hastie et al., 2009; Schutt, 2013; White, 2015; Efron and Hastie, 2016), but this book is primarily about Julia. Nevertheless, several important topics in data science are covered. These include data visualization, supervised learning, and unsupervised learning. When discussing supervised learning, we place some focus on gradient boosting — a machine learning technique — because we have found this approach very effective in applications. However, for unsupervised learning, we take a more statistical approach and place some focus on the use of probabilistic principal components analyzers and a mixture thereof.

This monograph is laid out to progress naturally. In Chapter 1, we discuss data science and provide some historical context. Julia is also introduced as well as details of the packages and datasets used herein. Chapters 2 and 3 cover the basics of the Julia language was well as how to work with data in Julia. After that (Chapter 4), a crucially important topic in data science is discussed: visualization. The book continues with selected techniques in supervised (Chapter 5) and unsupervised learning (Chapter 6), before concluding with details of how to call R functions from within Julia (Chapter 7). This last chapter also provides further examples of mixture model-based clustering as well as an example that uses random forests. Some appendices are included to provide readers with some relevant mathematics, Julia performance tips and a list of useful linear algebra functions in Julia.

There is a large volume of Julia code throughout this book, which is intended to help the reader gain familiarity with the language. We strongly encourage readers to run the code for themselves and play around with it. To make the code as easy as possible to work with, we have interlaced it with comments. As readers begin to get to grips with Julia, we encourage them to supplement or replace our comments with their own. For the reader's convenience, all of the code from this book is available on GitHub: `github.com/paTait/dswj`.

We are most grateful to David Grubbs of the Taylor & Francis Group for his support in this endeavour. His geniality and professionalism are always very much appreciated. Special thanks to

Professor Charles Bouveyron for kindly agreeing to lend his expertise in the form of a wonderful Foreword to this book. Thanks also to Dr. Joseph Kang and an anonymous reviewer for their very helpful comments and suggestions. McNicholas is thankful to Eamonn Mullins and Dr. Myra O'Regan for providing him with a solid foundation for data science during his time as an undergraduate student. Dr. Sharon McNicholas read a draft of this book and provided some very helpful feedback for which we are most grateful.

A final word of thanks goes to our respective families; without their patience and support, this book would not have come to fruition.

Paul D. McNicholas and Peter A. Tait
Hamilton, Ontatio