# *Preface*

It is hard to resist starting this book with a phrase like "analyzing data is fun... ." Anyone who does a lot of data analysis knows that this is of course only partly true because the preparation of the data we want to analyze is often far less fun. Another limiting factor during a data analysis might also be the tools at hand. Interactive graphical tools address both these points — they are fun to use and efficient when it comes to understanding where you still have to clean up your data.

Most important, interactive graphical tools are the most powerful means for exploratory data analysis (EDA), which was postulated by John W. Tukey almost half a century ago. Being a visionary, Tukey talked about things he could only envision at that time. It took more than a decade before he could actually implement a first prototype. The PRIM-9 system belongs to history today, but what was important about Tukey's work was the philosophy of data analysis he brought into being. Some of his academic descendants carried this spirit on, and some even created new tools to support this new kind of data analysis. Paul Velleman's DataDesk™ was the earliest benchmark for all of us.

To illustrate that it takes more than new tools to perform a more modern kind of data analysis, we want to quote an academic grandchild of John Tukey. Jay Emerson, student of John Hartigan, gives a short description of the course "Advanced Data Analysis" he taught at Yale:

> [...] I don't want you to leave the course feeling that you have learned about a limited set of tools, allowing you to do only certain types of analyses. I want you to feel prepared to face the unexpected, equipped with a set of skills enabling you to adapt to the inevitable surprises of data analysis. When faced with a fresh challenge, I want you to think, "I may not know the answer, but I bet I can figure it out." Someday, I want you to think, "that was one of the most practically useful courses I had at Yale."
>
> [...] You should be able to think critically about data, use graphical and numerical summaries, apply standard statistical inference procedures (when appropriate) and draw conclusions from such analyses. But most importantly, you should be willing to break out of the box and conduct new, innova-

> tive analyses of problems when standard analyses may not be appropriate.
>
> This course will be computationally intensive, and there is no substitute for getting your hands dirty. I expect to make my share of mistakes this semester (some intentional, some not), and we'll learn from them together. In data analysis, I believe you learn as much (and sometimes more) when things "don't work" than when they go as planned. You have succeeded when you can figure out why something doesn't work (or why some analysis isn't appropriate) and deduce an appropriate course of action as a result. You must be willing to try out new things and to make mistakes — you can't break the computer (at least, it won't go up in smoke), and the sky won't fall. Seek to understand the mistakes, and move onward.

Once you understand the differences and interactions between traditional statistics and data analysis, the usefulness of this book in learning and understanding data analysis by graphical means will be quite obvious.

It takes a radical skepticism against traditional statistical methods and a deep confidence into new approaches to move on to a new discipline. Having learned from and worked with Antony Unwin for many years, both of us started to appreciate this radicalism which allows us to distinguish between things more clearly.

We want to mention some of those who have had a sustained impact on our view of statistics and data analysis. To name just a few, we want to thank (in alphabetical order) Rick Becker, Axel Benner, Adrian Bowman, Andreas Buja, Dianne Cook, Jason Dykes, Fred Eicker, Michael Friendly, Wolfgang Härdle, Heike Hofmann, Kurt Hornik, Al Inselberg, Fritz Leisch, Junji Nakano, Balasubramanian Narasimhan (better known as Naras), Daryl Pregibon, Brian Ripley, Matt Schonlau, John Sammis, Günther Sawitzki, Robert Spence, Debby Swayne, Luke Tierney, Paul Velleman, Bill Venables, Chris Volinsky, Ed Wegman, Hadley Wickham, Rick Wicklin, Adi Wilhelm, Alan Wilks, Graham Wills and Achim Zeileis for their valuable inspirations.

We are also especially grateful to everybody who helped us find interesting and yet easy-to-understand and also easy-to-relate-to real-world datasets. These real life problems are vital input for the research and software development of interactive statistical graphics. The amount of time spent hunting for such datasets is often underestimated. Whereas the task in traditional parametric statistics is to find or simulate a (single) data set that works for a newly invented model class, the best way to move forward in interactive graphics research is to look for data that cannot be analyzed efficiently with the tools we have at hand so far.

Only a few of the datasets we have collected over the last decades were selected for this book. We especially want to thank Robert Erber for the collaboration on the medieval data on the city of Augsburg, the 110 students who took the Probability Exam in 2006, Antony Unwin for looking at the *Titanic* data from Dawson's *JSE* article in mosaic plots (shortly after initial skepticism on what I was doing with these funny plots), Rick Wicklin for sharing the data on the *Titanic* boats with us, Dianne Cook for pointing us to the Olive Oil data and the Tipping data and finally Annerose Zeis, who found a wonderful application of parallel coordinate plots with the Tour de France data (not to forget Sergej Potapov who made collecting the Tour de France data so much easier!).

We also want to thank Rob Calver for his considerate support at all stages of this book project and the reviewers who gave fruitful input that significantly improved the book.

Our final thanks go to our families for their infinite patience and understanding during the (too long) course of getting this book together. Without their support we would never have finished this project.

Martin Theus
Simon Urbanek