Chapter 1
Introduction: Why Do We Visualize Data and What Is This Book About?

> *Restate my assumptions:*
> *One, mathematics is the language of nature.*
> *Two, everything around us can be represented and understood through numbers.*
> *Three, if you graph the numbers of any system, patterns emerge.*
>
> Sean Gullette as Maximilian Cohen in the movie π (1998).

The goal of this book is simple: We would like to show how mortality dynamics can be visualized in the so-called Lexis diagram. To appeal to as many potential readers as possible, we do not require any specialist knowledge. This approach may be disappointing: Demographers may have liked more information about the mathematical underpinnings of population dynamics on the Lexis surface as demonstrated, for instance, by Arthur and Vaupel in 1984. Statisticians would have probably preferred more information about the underlying smoothing methods that were used. Epidemiologists likewise might miss discussions about the etiology of diseases. Sociologists would have probably expected that our results were more embedded into theoretical frameworks....

We are aware of those potential shortcomings but believe that the current format can, nevertheless, provide interesting insights into mortality dynamics, and we hope our book can serve as a starting point to visualize data on the Lexis plane for those who have not used those techniques yet.

Visualizing data has become increasingly popular in recent years.[1] But why do we visualize data at all? Countless books on *how* to visualize data — often with a specific software tool in mind — are published every year. Maybe it seems to be too

[1] This trend is probably best demonstrated by visualizing the popularity of the term "visualizing data" over time, for instance, via Google's *Ngram viewer*. Google Books Ngram Viewer displays the relative frequency of a search term in a corpus of books during a given time frame. Please see, for example: https://books.google.com/ngrams/graph?content=visualizing+data+&year_start=1960&year_end=2008

R. Rau et al., *Visualizing Mortality Dynamics in the Lexis Diagram*,
The Springer Series on Demographic Methods and Population Analysis 44,
DOI 10.1007/978-3-319-64820-0_1

obvious, but only a few of those publications address the question of *why* one should visualize data at all. According to the ones covering the topic, the purpose of data visualization can be narrowed down to three reasons (e.g., Tukey 1977; Schumann and Müller 2000; Few 2014):

1. *Exploration:* John Tukey stresses that exploratory data analysis "can never be the whole story, but nothing else can serve as the foundation—as the first step" (Tukey 1977, p. 3). He uses the expression of "graphical detective work" by trying to uncover as many important details about the underlying data as possible. If one explores data only with preconceived notions and theories, it is likely that essential characteristics remain undiscovered.
2. *Confirmation:* It could be argued that the mere exploration of data without any hypotheses is a misguided endeavor. Exploration needs to be firmly distinguished from confirmatory analysis, though. While the exploration is comparable to the work of the police, this step can be seen as the task of a judge or the jury. Both are important to advance science, the first step is to gather the facts whereas the second step is of judgmental nature: Can the "facts" be interpreted to support the theory? Or do certain findings exclude some hypotheses? In this sense, confirmatory analysis represents the core of scientific progress in Popper's sense, namely by falsifying theories.
3. *Presentation:* Presenting and communicating the findings from the data analysis to the reader, or more appropriately, to the viewer, represents the third pillar of why data visualization is important. Mixing up confirmatory analysis with the presentation of the findings is probably one of the root causes for poor scientific communication. It is a common occurrence at scientific conferences that researchers use the same graphical tools to present their results to others as they used to obtain their findings in the first place. As pointed out by Schumann and Müller (2000, p. 6), this step requires careful thought that third parties are able to understand the findings without any unnecessary difficulties.

Maps and diagrams were already known in ancient Egypt but also communicating scientific results via visualization is at least 400 years old when Galileo Galilei (1613) and others published their observations of sunspots and other celestial bodies (Friendly 2008). But why is data visualization only becoming increasingly popular during the last 15–20 years? We argue that the key reason is the trend towards virtually ubiquitous access to electronic computing resources, enabling more and more people to participate in this endeavor. One could call it even a democratization of computing. In our opinion we can distinguish three key developments that played a crucial role since the 1980s and especially the 1990s. They are not listed in order of importance nor can they be considered in isolation from each other.

Hardware: The introduction of the predecessor of all modern PCs, the IBM personal computer, in 1981 as well as of microcomputers (e.g., the "C-64") in the same era triggered a shift away from the so-called minicomputers of the 1970s[2] to

[2] As noted at https://en.wikipedia.org/wiki/Minicomputer#cite_note-Smith_1970-4 (last accessed on 13 June 2017), the New York Times wrote in 1970 that minicomputers were computers that cost less than US-$ 25,000.

computers that could be purchased by households of average income. The speed of the processors was too slow and the size of computer memory was too small to process data as conveniently as we can nowadays, though. The first PC had an upper limit for working memory (RAM) of 256 kB, that is about 0.000778% of the first author's current desktop workstation. If we disregard developments in cache technology, parallel processing, etc., the pure clock speed of processors is now three orders of magnitude higher than in the early 1980s. Only 20 years ago, the typical size of total RAM was about as large as the size of a *single* digital photo today. But even if there was enough RAM and sufficient clock speed of the CPU, data storage was another limiting factor. The first hard disk with a capacity of more than one gigabyte was introduced in 1980 and cost at least US-$ 97,000.[3] One thousand times the storage capacity is available now at less than US-$ 100. This trend allowed the collection of massive data sets. To illustrate current capabilities: If we were interested in creating a data set, which contains about 1000 alphabetic characters (more than enough for the name, birth date and current residence) of any person alive, we would have to invest less than US-$ 400.[4] But, once again, even if we had the affordable computer storage of today, communicating results graphically was hindered by the low resolution combined with relatively few colors of early graphics standards such as CGA and EGA. Only with the introduction and the extension of the VGA standard, high resolution displays have become feasible.

Software: Having hardware in terms of processing speed, working memory and hard disk capacity to process graphics coincided with a revolution in software in the 1990s: Similar to the introduction of home computers that gave access to almost everyone, the emergence of *free software*, also called open source software, allowed anyone to use software without the costs and other restrictions often imposed by software products. Examples for this development can be found in the area of

- general programming languages (e.g., Python, Perl) as well as
- languages tailored or at least particularly suited for statistical programming and data analysis. The invention of the S language, started in the 1970s, was instrumental.[5] The most prominent example today is probably R (Ihaka 1998), but also other languages such as the now almost completely abandoned XLISP-STAT (de Leeuw 2005) facilitate(d) the visualization of data.[6]
- Lastly, in the area of efficient data storage, especially with the advent of "big data". Although it might be one of the most abused buzzwords currently, data

[3] See: https://www-03.ibm.com/ibm/history/exhibits/storage/storage_3380.html, last accessed on 13 June 2017.

[4] Assuming a world population of less than eight billion, a price for a 2TB hard disk of less than US-$S 100 and one byte per alphabetic letter.

[5] Please see Appendix A in Chambers (2008) for some notes on the history of S.

[6] It should be mentioned, though, that Matlab (Mathworks 2017), which is not published under a free/open-source license, was and is also key for the analysis and visualization of data.

sets in the gigabyte and terabyte range, partly in non-rectangular formats, have become ubiquitous. Those data can be handled by relational and non-relational database systems that are also available under free and open source licenses (e.g., SQLite, MySQL, Postgresql, Cassandra).

Connectivity: While the internet existed already for more than 20 years, the introduction and rising popularity of the world wide web (WWW) was a catalyst for the exchange of information via electronic networks. This technology allows now billions of people on earth to have almost instant access to data. The speed of the internet connection, which is crucial for the exchange of information such as downloading large data sets, has also increased by at least two orders of magnitude since the middle of the 1990s when 56 kbit/s modems were the standard.