open source PRESS

Datendesign mit R

100 Visualisierungsbeispiele

professional reference

Thomas Rahlf

Thomas Rahlf

Datendesign mit R

100 Visualisierungsbeispiele

1. Auflage

Open Source Press

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Gesamtlektorat: Dr. Markus Wirtz
Satz: Open Source Press & Thomas Schraitle
Umschlaggestaltung: Olga Saborov, Open Source Press
Gesamtherstellung: Kösel, Krugzell

ISBN: 9783955390952 (E-Book PDF)
ISBN: 9783955390945 (gedruckte Ausgabe)

http://www.opensourcepress.de

Inhaltsverzeichnis

Vorwort 17

1 Daten für alle 21

1.1 Datenvisualisierung zwischen Wissenschaft und Journalismus. 21

1.2 Warum R?. 27

1.3 Das Konzept des Datendesigns. 29

I Grundlagen und Technik 33

2 Aufbau und technische Voraussetzungen 35

2.1 Begriffe und Elemente. 36

2.2 Gestaltungsraster. 36

2.3 Perzeption. 41

2.4 Schriften 48
2.4.1 Fonts 52
2.4.2 Freie Schriften 54
2.5 Symbole 56
2.5.1 Symbolfonts 61
2.5.2 Symbole im SVG-Format 62
2.6 Farbe 64
2.6.1 Farbmodelle 65
2.6.2 Farbe in statistischen Abbildungen 69

3 Umsetzung in R 73

3.1 Installation 74
3.2 Grundkonzepte in R 75
3.2.1 Datenstrukturen 77
3.2.2 Import von Daten 82
3.3 Grafikkonzepte in R 93
3.3.1 Paper-Pencil-Prinzip des Base Graphics System: High-Level- und Low-Level-Funktionen 101
3.3.2 Einstellung von Grafikparametern 106
3.3.3 Randeinstellungen von Abbildungen und Grafiken 117
3.3.4 Mehrfachgrafiken: Panels mit mfrow und mfcol 120
3.3.5 Komplexere Anordnungen mit layout 121

3.3.6 Schrifteinbindung. 125
3.3.7 Ausgabe mit cairo_pdf. 127
3.3.8 Unicode in Abbildungen. 129
3.3.9 Farbeinstellungen. 135
3.4 R-Pakete und -Funktionen in diesem Buch. 138
3.4.1 Pakete. 139
3.4.2 Funktionen. 146
3.4.3 Schematische Vorgehensweise. 161

4 Über R hinaus 163

4.1 Ergänzungen mit LaTeX. 163
4.2 Manuelle Nachbearbeitung und Symbolfonterstellung in Inkscape. 173
4.2.1 Nachbearbeitung. 173
4.2.2 Symbolfonterstellung. 175

5 Zu den Beispielen 183

5.1 Versuch einer Systematik. 183
5.2 Die Skripte zum Laufen bringen. 187

II Beispiele 189

6 Kategoriale Daten 191

6.1 Balken- und Säulendiagramme. 191
6.1.1 Balkendiagramm einfach. 192
6.1.2 Balkendiagramm für Mehrfachantworten – die ersten beiden Antwortkategorien. 200
6.1.3 Balkendiagramm für Mehrfachantworten – alle Antwortkategorien. 207
6.1.4 Balkendiagramm für Mehrfachantworten – alle Antwortkategorien, Variante. 210
6.1.5 Balkendiagramm für Mehrfachantworten – alle Antwortkategorien (Panel). 214
6.1.6 Balkendiagramm für Mehrfachantworten – Symbole für Individuen. 218
6.1.7 Balkendiagramm für Mehrfachantworten – alle Antwortkategorien, gruppiert. 222
6.1.8 Säulendiagramm mit zweizeiliger Beschriftung. 231
6.1.9 Säulendiagramm mit 45-Grad-Beschriftung. 234
6.1.10 Profildiagramm für Mehrfachantworten – Mittelwerte der Antworten. 236
6.1.11 Dotchart für drei Variablen. 240
6.1.12 Säulendiagramm mit Anteilen. 245
6.2 Kreis- und Radialdiagramme. 248
6.2.1 Einfaches Kreisdiagramm. 250
6.2.2 Kreisdiagramme, Beschriftung innen (Panel). 252

6.2.3 Sitzverteilung (Panel). 255
6.2.4 Spie Chart. 260
6.2.5 Radialpolygone (Panel). 264
6.2.6 Radialpolygone (Panel) – andere Spaltenanordnung. 267
6.2.7 Radialpolygone übereinander. 269
6.3 Grafiktabellen. 271
6.3.1 Vereinfachtes Gantt-Diagramm. 273
6.3.2 Vereinfachtes Gantt-Diagramm – Farben nach Personen. 279
6.3.3 Bumpchart. 281
6.3.4 Heatmap. 286
6.3.5 Mosaikplot (Panel). 290
6.3.6 Ballonplot. 294
6.3.7 Treemap. 297
6.3.8 Treemaps für zwei Ebenen (Panel). 301

7 Verteilungen 307

7.1 Histogramme und Boxplots. 307
7.1.1 Histogramme übereinander. 307
7.1.2 Säulendiagramme mit Colorbrewer gefärbt (Panel). 310
7.1.3 Histogramme (Panel). 315

7.1.4 Boxplots für Gruppen – absteigend sortiert. 319

7.1.5 Boxplots für Gruppen – absteigend sortiert, Vergleich zweier Erhebungen. 324

7.2 (Bevölkerungs-)Pyramiden. 331

7.2.1 Pyramide mit mehreren Farben. 333

7.2.2 Pyramiden – Betonung der äußeren Bereiche (Panel). 337

7.2.3 Pyramiden – Betonung der inneren Bereiche (Panel). 342

7.2.4 Pyramiden mit eingezeichneter Linie (Panel). 347

7.2.5 Pyramide mit Zusammenfassungen. 348

7.2.6 Balkendiagramme als Pyramiden (Panel). 352

7.3 Ungleichheit. 356

7.3.1 Einfache Lorenzkurve. 357

7.3.2 Lorenzkurven übereinander. 360

7.3.3 Lorenzkurven (Panel). 364

7.3.4 Vergleich von Einkommensanteilen mit Balkendiagramm (Quintile). 367

7.3.5 Vergleich von Einkommensanteilen mit Balkendiagramm (Dezile). 371

7.3.6 Vergleich von Einkommensanteilen mit Panel-Balkendiagramm (Quintile). 374

8 Zeitreihen 379

8.1 Kurze Zeitreihen. 379
8.1.1 Säulendiagramm für Entwicklungen. 379
8.1.2 Säulendiagramm mit Anteilen für Wachstumsentwicklungen. 384
8.1.3 Quartalswerte als Säulen. 388
8.1.4 Quartalswerte als Linien mit Werte-Beschriftungen. . . . 391
8.1.5 Kurze Zeitreihen übereinander. 394
8.2 Flächen unter und zwischen Zeitreihen. 397
8.2.1 Flächen zwischen zwei Zeitreihen. 397
8.2.2 Fläche als Korridor mit Zeitreihen (Panel). 400
8.2.3 Prognoseintervalle (Panel). 404
8.2.4 Prognoseintervalle Index (Panel). 408
8.2.5 Zeitreihen mit gestapelten Flächen. 412
8.2.6 Flächen unterhalb einer Zeitreihe. 415
8.2.7 Zeitreihen mit Trend (Panel). 419
8.3 Darstellung von Tages-, Wochen- und Monatswerten. 424
8.3.1 Tagseswerte mit Beschriftungen. 424
8.3.2 Tageswerte mit Beschriftungen und Wochensymbolen (Panel). 428
8.3.3 Tageswerte mit Monatsbeschriftung. 434
8.3.4 Zeitreihen aus Wochenwerten (Panel). 437
8.3.5 Monatswerte (Panel). 441

8.3.6 Monatswerte mit Monatsbeschriftung. 444
8.3.7 Monatswerte mit Monatsbeschriftung (Layout). 448
8.4 Sonderfälle und Spezielles. 452
8.4.1 Zeitreihen als Streudiagramm (Panel). 452
8.4.2 Zeitreihen mit fehlenden Werten. 456
8.4.3 Saisonspannweiten (Panel). 461
8.4.4 Saisonspannweiten übereinander. 464
8.4.5 Saisonfigur (Seasonal Subseries Plot) mit Datentabelle. 467
8.4.6 Zeitliche Spannweiten. 472

9 Streudiagramme 475

9.1 Varianten. 478
9.1.1 Streudiagramm Variante 1: Vier Quadranten farblich unterschieden. 478
9.1.2 Streudiagramm Variante 2: Ausreißer farblich hervorgehoben. 483
9.1.3 Streudiagramm Variante 3: Bereiche farblich hervorgehoben. 487
9.1.4 Streudiagramm Variante 4: Eingezeichnete Ellipse. 491
9.1.5 Streudiagramm Variante 5: Verbundene Punkte. 494
9.2 Sonderfälle und Spezielles. 498
9.2.1 Streudiagramm mit wenigen Punkten. 498

9.2.2 Streudiagramm mit selbst definierten Symbolen. 501
9.2.3 Karte von Deutschland als Streudiagramm. 505

10 Karten 509

10.1 Einführende Beispiele. 510
10.1.1 Karten von Deutschland: Ortsnetzbereiche und Postleitzahlengebiete. 510
10.1.2 Gefilterte Postleitzahlenkarte. 513
10.1.3 Europakarte Nuts 2006 (Ausschnitt). 517
10.2 Punkte, Diagramme und Symbole in Karten. 519
10.2.1 Karte von Deutschland mit ausgewählten Orten und Umriss (Panel). 519
10.2.2 Karte von Deutschland mit ausgewählten Orten (Kreisdiagramme) und Umriss. 523
10.2.3 Karte von Deutschland mit ausgewählten Orten (Säulen) und Umriss. 527
10.2.4 Karte von Deutschland als dreidimensionales Streudiagramm. 531
10.2.5 Karte von Nordrhein-Westfalen mit ausgewählten Orten (Symbole) und Umriss. 537
10.2.6 Karte von Tunesien mit selbst definierten Symbolen. 541
10.3 Choroplethenkarten. 545

10.3.1 Choroplethenkarte von Deutschland auf Kreisebene. 546
10.3.2 Choroplethenkarte von Deutschland auf Kreisebene (Panel). 550
10.3.3 Choroplethenkarte von Europa auf Länderebene. 557
10.3.4 Choroplethenkarte von Europa auf Länderebene (Panel). 561
10.3.5 Weltchoroplethenkarte: Regionen. 566
10.4 Sonderfälle und Spezielles. 569
10.4.1 Weltkarte mit Orthodromen. 570
10.4.2 Stadtkarten mit OpenStreetMap-Daten (Panel). 573

11 Illustratives 581

11.1 Tabelle mit Symbolen der Schrift„Symbol Signs". 581
11.2 Radialsäulendiagramme mit Beschriftung (Panel). 584
11.3 Radialsäulendiagramme ohne Beschriftung (Panel). 594
11.4 Radialsäulendiagramm (Poster). 598
11.5 Nacht-Karte von Deutschland als Streudiagramm. 604
11.6 Streudiagramm Gapminder. 608
11.7 Karte von Napoleons Rußlandfeldzug von 1812/13 von Charles Joseph Minard, 1869. 616

III Anhang 621

A Verwendete Daten 623

A.1 ZA2391: Politbarometer 1977-2011 (Partielle Kumulation). 623

A.2 ZA4753: European Values Study 2008: Germany (EVS 2008). 624

A.3 ZA4804: European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). 625

A.4 ZA4972: Eurobarometer 71.2 (May-Jun 2009). 625

A.5 BetterLifeIndex_Data_2011V6.xls. 626

A.6 weltenergiemix.xlsx. 627

A.7 personen.xlsx. 627

A.8 v_frauen_maenner. 627

A.9 gadm.org. 628

A.10 NUTS-Karten. 629

B Literatur 631

B.1 Bücher und Artikel. 631

B.2 Websites zum Einstieg. 637

Index 639

Vorwort

Als ich vor fast zwanzig Jahren eine Reihe von Büchern zur statistischen Grafik und grafisch gestützten Datenanalyse rezensiert habe, war alles noch ganz anders: Formate waren proprietär, Betriebssysteme und ihre Zeichensätze inkompatibel, Grafik- und Statistiksoftware teuer. Seit der Jahrtausendwende änderte sich die Lage grundlegend: Das Internet war den Kinderschuhen entwachsen, Open-Source-Projekte gewannen unter diesem neuen Begriff immer mehr Anhänger und eine Handvoll Enthusiasten stellte Version 1.0 der freien Statistik-Programmierspache R zur Verfügung. Viele Entwickler ließen sich für eine Mitarbeit an diesem Projekt begeistern. 2013 hat R die Versionsnummer 3 erreicht, neben der Basis-Software gibt es aktuell über 4.000 frei verfügbare Erweiterungspakete. Firmen und Organisationen wie Google, Facebook oder die CIA verwenden R zur Datenanalyse. Als besondere Stärke werden immmer wieder die Grafikfähigkeiten hervorgehoben. Praktisch alle für die Datenvisualisierung relevanten Technologien werden zeitnah in R integriert. Man kann mit zahlreichen Funktionen jede nur erdenkliche Abbildung detailgenau konstruieren, Karten erstellen und vieles mehr.

Man muss nur wissen, wie — und dazu möchte dieses Buch einen Beitrag leisten.

Was dieses Buch sein möchte — und was nicht

Das vorliegende Buch ist *keine* Einführung, die systematisch die Garfikwerkzeuge von R erläutert. Es möchte vielmehr anhand von 100 vollständigen Skript-Beispielen in die Grundlagen der Gestaltung von Präsentationsgrafiken einführen und zeigen, wie Balken-und Säulendiagramme, Bevölkerungspyramiden, Lorenzkurven, Boxplots, Streudiagramme, Zeitreihendarstellungen, Radialpolygone, Gantt-Diagramme, Heatmaps, Bumpcharts, Mosaik- und Ballonplots sowie eine Reihe verschiedener thematischer Kartentypen mit dem *Base Graphics System* von R erstellt werden.[1] Jedes Beispiel verwendet reale Daten und erläutert die Abbildung und deren Programmierung Schritt für Schritt. Die Auswahl orientiert sich an meinem persönlichen Erfahrungsschatz – sicher wird der ein oder andere die eine oder andere Abbildung vermissen, anderes als zu ausführlich empfinden. Dennoch sollte ein großer Anwendungsbereich abgedeckt sein.

Das Buch richtet sich an

R-Kenner: Für Sie sind insbesondere die Beispiele nützlich, besonders der Code. Teil I können Sie vermutlich überspringen.

Leserinnen und Leser, die von R schon gehört, es auch schon einmal ausprobiert und keine Angst vor dem Programmieren haben; Sie profitieren von beiden Teilen.

Anfänger: Ihnen helfen vor allem die fertigen und hier abgebildeten Grafiken. Sie sehen, was mit R möglich ist. Oder anders gesagt: Sie

[1] Für die anderen in R verfügbaren Grafikansätze wie `grid` und, darauf aufbauend, `lattice` und `ggplot2` sei auf bereits vorhandene Einführungen verwiesen.

sehen, dass es R überhaupt gibt und dass sich damit Grafiken erzeugen lassen, die Sie schon lange einmal erzeugen wollten – Sie wussten bloß nicht, wie. Der Code ist Ihnen zwar zu kompliziert, aber Sie können evtl. andere damit beauftragen, Grafiken für Sie in R zu programmieren.

Windows, Mac und Linux
: Alle Skripte und Bearbeitungsschritte führen unter Windows , Mac OS X und Linux zu identischen Ergebnissen. Die Beispiele wurden mit Mac OS X erstellt und anschließend unter Ubuntu 12.04 sowie einer Evaluierungskopie von Windows 8.1 getestet.

Danksagung

Für Hinweise, Kommentare, Rückmeldungen, Daten, Diskussionsgelegenheiten oder Hilfe danke ich Gregor Aisch, Insa Bechert, Evelyn Brislinger, Giuseppe Casalicchio, Arnulf Christl, Katja Diederichs, Günter Faes, Mira Hassan, Mark Heckmann, Daniel Hienert, Bruno Hopp, Duncan Temple Lang, Uwe Ligges, Lorenz Matzat, Meinhard Moschner, Stefan Müller, Paul Murrell, David Phillips, Martijn Tennekes, Patrick R. Schmid, Thomas Schraitle, Valentin Schröder, Torsten Steiner, Michael Terwey, Katrin Weller, Bernd Weiss, Nils Windisch, Benjamin Zapilko und Lisa Zhang.

Ganz besonders hat das Manuskript vom Austausch mit einem Infografiker und einem Datenjournalisten profitiert. Stefan Fichtel hat alle Abbildungen angesehen und kritisch kommentiert. Für einzelne Abbildungen hat er eigene Vorschläge gestaltet. Das war mir eine unschätzbare Hilfe. Nicht in allen Fällen waren wir einer Meinung, und an der ein oder anderen Stelle habe ich mich über seinen Rat hinweggesetzt. Verbliebene Fehler und Unzulänglichkeiten gehen daher zu meinen Lasten.

Björn Schwentker hat sich die Mühe gemacht und große Teile des Manuskripts gründlich gegengelesen. Ihm verdanke ich wertvolle Hinweise, die den Text an einigen Stellen mit Sicherheit klarer und lesbarer gemacht haben.

Schließlich gilt mein Dank Markus Wirtz, dass er das Experiment gewagt hat, aus all dem am Ende doch ein Buch zu drucken.

Im Internet

Die Abbildungen sind für unterschiedlichste Endausgaben konzipiert. Insbesondere bei den Karten und Radialsäulendiagrammen ist manches aufgrund des Buchformates grenzwertig klein. Gerade für solche Fälle sei auf die Website des Buches verwiesen, auf der alle Abbildungen in hoher Auflösung bzw. als Vektorgrafik im PDF-Format bereitgestellt werden:

`http://www.datendesign-r.de`

Dort finden Sie auch die verwendeten Daten, sofern sie frei verfügbar gemacht werden können. Für alle anderen Daten sind die Quellen jeweils angegeben. Die Über- und Unterschriften in den Beispielabbildungen wurden für den Druck abgekürzt.

Mit dem vorn im Buch abgedruckten Zugangscode können Sie sich auf der Verlagswebsite registrieren[2] und erhalten dann kostenfrei eine E-Book-Ausgabe dieses Buches sowie eine ZIP-Datei mit allen R-Skripten.

[2] `http://www.opensourcepress.de/voucher`

Kapitel 1

Daten für alle

1.1 Datenvisualisierung zwischen Wissenschaft und Journalismus

Art und Umfang von Daten, unsere Einstellung zu ihnen sowie ihre Verfügbarkeit haben sich in den vergangenen Jahren grundlegend gewandelt. Noch nie gab es so viele Daten wie heute. Noch nie waren sie so leicht verfügbar. Und noch nie waren die Möglichkeiten der Analyse, Aufbereitung und Präsentation größer.

Manche Wissenschaftler, wie etwa der Mathematiker Stephen Wolfram, glauben, dass man den Prozess der Datenanalyse weitgehend automatisieren kann, und sprechen in diesem Zusammenhang sogar von einer Demokratisierung der Wissenschaft. Andere, wie Googles Chefökonom Hal Varian, meinen hingegen, dass dafür mehrere Fähigkeiten erlernt

werden müssen und diese zukünftig zentrale Schlüsselqualifikationen darstellen: „The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades (…)".[1]

In den letzten Jahren ist eine Fülle von Websites, Büchern und anderen Publikationen entstanden, die sich der Visualisierung von Daten widmen. Dabei steht deren erzählende, nicht die explorative Visualisierung im Vordergrund. Eines der bekanntesten Beispiele ist die Mission von Hans Rosling, dem Autor und Erfinder von GAPMINDER, Statistiken zu weltweiten gesellschaftlichen Entwicklungen einem breiten Publikum eingängig zu veranschaulichen. Hans Rosling wurde 2012 vom *Time Magazine* zu den „100 Most Influential People in the World" gezählt. Nahezu in Vergessenheit geratene Sozialwissenschaftler, die sich mit der didaktischen Visualisierung von statistischen, gesellschaftlichen Zusammenhängen befasst haben, allen voran Otto Neurath, werden wiederentdeckt.[2]

Dabei ist es nicht so, dass das Rad neu erfunden wurde. In der Wissenschaft haben Datenvisualisierungen seit jeher und kontinuierlich eine wichtige Rolle gespielt. Bildgebende Verfahren gehören zum festen Bestandteil vieler Analysen in der Medizin, praktisch alle Naturwissenschaften nutzen bildliche Darstellungen von Daten zur visuellen Kommunikation von Ergebnissen. Die Zeitschrift *Nature* bietet Interessenten im Internet als Kaufanreiz für ihre Artikel neben einem Abstract kleine Voransichten der enthaltenden Abbildungen („Figures at a glance").

Im Rahmen der statistischen Methodik haben eine Reihe von Wissenschaftlern schon vor vielen Jahren Grundlagenforschung zur statistischen Grafik

1 `http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers`

2 Eve, Matthew / Burke, Christopher (Hrsg.) / Otto Neurath (2010): From Hieroglyphics to Isotype: A Visual Autobiography. London: Hyphen Press.

betrieben: Bahnbrechend war neben den Arbeiten von William S. Cleveland das Buch von Edward Tufte, *The Visual Display of Quantitative Information*. Das Buch erschien 1983 und erlebte bereits in der ersten Auflage sechzehn Nachdrucke. Zusammen mit zwei in der Folge erschienenen Werken, *Envisioning Information* und *Visual Explanations*, hat Edward Tufte damit den Maßstab für das Thema auf eine sehr genuine Weise definiert.

Auch in der Wirtschaft gibt es eine lange Tradition der Präsentation von Daten. Seit vielen Jahren werden in Unternehmen für interne Zwecke nicht nur Daten gesammelt und ausgewertet, sondern auch in Abbildungen umgesetzt. Nach außen werden in besonders aufbereiteten Publikationen Präsentationsgrafiken in Geschäftsberichten möglichst eindrucksvoll zur Schau gestellt.[3]

Schließlich bemüht sich die amtliche Statistik seit vielen Jahren erfolgreich, ihre Ergebnisse nicht nur in tabellarischer Form bereitzustellen, sondern auch grafisch aufzubereiten. Hier kann man sowohl national als auch international eine nahezu von Jahr zu Jahr fortschreitende Tendenz zur stärkeren Visualisierung des offiziellen Datenmaterials feststellen.

Die Flut von Daten, die auf uns einströmt und uns ihre Auswertung aufdrängt, hat einen Nebeneffekt: Mit ihrer neuen, potentiellen Verfügbarkeit und Offenheit geht ein Umdenken in Bezug auf die Nutzungsrechte und Einsichtmöglichkeiten einher. Zunehmend wird die Offenheit nicht nur von amtlichen, sondern auch von Unternehmensdaten gefordert. Umwelt- und Wetteraufzeichnungen, Verbrauchsdaten oder solche aus den Bereichen Gesundheit oder Bildung, Abstimmungen in Landtagen, Gesetzestexte, Daten zur Verkehrslage oder Fahrpläne sollen frei und offen

[3] In ihrer schönsten Form zusammengestellt bei Rädeker, Jochen / Dietz, Kirsten (2011): Reporting, Unternehmenskommunikation als Imageträger - ausgesuchte Finanz- und Nachhaltigkeitsberichte weltweit. Mainz: Hermann Schmidt Verlag.

zugänglich sein. Gegenüber den USA, Großbritannien oder auch der Schweiz hat Deutschland hier noch Nachholbedarf.[4]

Big Data und Open Data erfordern neue Methoden und neue Herangehensweisen. Eine innovative Variante, die sich die Bezeichnung *Data Science* zu eigen gemacht hat, versteht darunter eine Kombination aus Programmierfähigkeiten, mathematisch-statistischen Kenntnissen und substanzwissenschaftlicher Expertise. Drew Conway hat diese Kombination in Form eines Venn-Diagramms dargestellt, das uns auch sehr anschaulich die Schnittmengen veranschaulicht:

[4] `https://index.okfn.org`

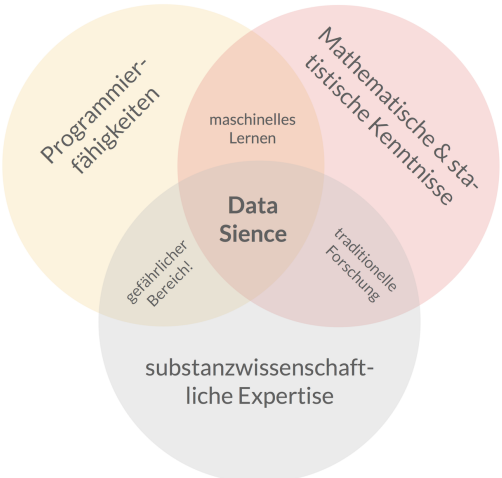

Abbildung 1.1: Data Science Venn-Diagramm nach D. Conway

Diese *Data Science* ist in aller Regel hochmathematisch und elaboriert. Aber auch der journalistische Bereich zeigt ein stark gewachsenes Interesse an Daten. Die allen voran von der *New York Times* und dem *Guardian*, in Deutschland von der *ZEIT* und anderen Medien angebotenen Recherchen und Visualisierungen sind unter dem Begriff *Datenjournalismus* im Aufwind.

So genannte Infografiken, häufig auch animiert und interaktiv, verbreiten sich geradezu explosionsartig im Internet. Seriöse und Maßstäbe setzende Angebote basieren dabei auf der Arbeit umfangreicher Experten-Teams und werden selbst Gegenstand der Forschung.

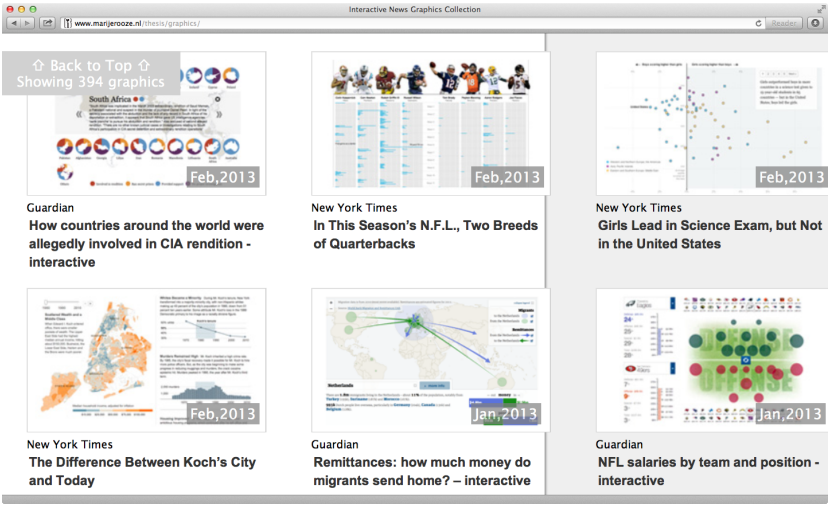

Abbildung 1.2: Infografiken der New York Times und des Guardian (nach M. Rooze)

Daneben erfreuen sich individuelle Angebote von „Information Designern" wie Catherine Mulbrandon, Stephen Few, Robert Kosara, Ben Fry oder Nathan Yau großer Beliebtheit, die eigene Datenvisualisierungssoftware entwickeln, Consulting-Firmen gründen, weltweit Workshops anbieten oder Blogs mit zigtausenden registrierten Nutzern aufbauen.[5]

Aus Sicht der eher „traditionellen" statistischen Grafik schießt das eine oder andere dabei über das Ziel hinaus: So manches wird nicht nur als zu bunt, zu verspielt oder zu überladen empfunden, sondern auch als verwir-

5 `http://visualizingeconomics.com`; `http://www.perceptualedge.com`; `http://kosara.net`; `http://benfry.com`; `http://flowingdata.com`.

rend oder gar verfälschend. Hier ist in jüngster Zeit eine Diskussion entstanden, von der am Ende sicher beide Seiten profitieren werden.[6]

1.2 Warum R?

In diesem Buch werden sämtliche Daten mit der freien Statistik-Software R visualisiert. Unter Wissenschaftlern ist die Programmiersprache inzwischen weit verbreitet und sehr beliebt. Doch jenseits der Forschung ist ihr Potenzial, maßgeschneiderte Grafiken zu produzieren, wenig bekannt. Das ist kein Wunder, denn Grafiker oder Journalisten tun sich mit dem Programmieren bekanntlich schwer. Es wäre sicherlich auch falsch zu behaupten, man lerne R so schnell, dass man in wenigen Minuten die erste ansprechende Grafik erstellt.

Andererseits: Der Einstieg ist leichter als in viele andere Programmiersprachen, weil R speziell für Daten und Statistik gemacht ist – und damit auch für deren Visualisierung. Es bietet einige Vorteile, die auch für Redakteure oder Datendesigner Gold wert sein können und die eine Software wie Excel nicht bietet:

- Alle Grafiken lassen sich im Vektorformat speichern (z.B. PDF, EPS oder SVG) und mit gängigen Vektorgrafikprogrammen wie Adobe Illustrator oder dem freien Inkscape sofort weiterverarbeiten, so dass jedes Grafikelement einzeln anpassbar ist.

[6] Gelman, Andrew / Unwin, Antony, Infovis and Statistical Graphics: Different Goals, Different Looks, in: Journal of Computational and Graphical Statistics 22/1 (2013), S. 2-28. Diskussionsbeiträge von Robert Kosara, Paul Murrell, Hadley Wickham S. 29-44, Antwort S. 45-49.

- Jedes Element der Grafik lässt sich durch R fast beliebig in Farbe oder Form verändern. Es lassen sich nach belieben Text, Symbole, Pfeile oder ganze Zeichnungen hinzufügen oder verschiedene Diagramme kombinieren.

- Die Grundformen der wichtigste Diagrammtypen, wie Säulen-, Linien- oder Kreisdiagramme, lassen sich für einen ersten Eindruck oft schnell durch einen einzigen Befehl erzeugen.

- R beherrscht auch Karten und lässt so beliebige Geo-Visualisierungen zu. Das Kartenmaterial dafür kann zum Beispiel im geläufigen Format von Shape-Dateien eingeladen werden.

- Da Grafiken in R komplett programmiert sind, lässt sich jeder Schritt nachvollziehen, jeder Fehler finden und Änderungen sind leicht möglich. Dies ermöglicht auch eine Qualitätskontrolle durch Dritte und eine Offenlegung des Grafik-Sourcecodes im Sinne maximaler Open-Data-Transparenz.

- R ist kostenlos.

- R ist offen.

- R ist durch viele Programm-Module (Packages) erweiterbar, um besondere Grafiktypen darzustellen oder fortgeschrittene Datenanalysen vorzuschalten. Eine wachsende internationale Community stellt im Internet immer mehr Eweiterungen zur Verfügung

- R-Grafiken können auch als Grundlage für interaktive Online-Grafiken dienen, indem beispielsweise den als SVG gespeicherten Diagrammelementen mit einem JavaScript-Paket wie D3.js interaktives Leben

eingehaucht wird. Alternativ gibt es inzwischen ein kompletes JavaScript-Paket namens *Shiny*[7], mit dem sich interaktive Datenanwendungen im Netz direkt in R schreiben lassen.

1.3 Das Konzept des Datendesigns

Das Buch verfolgt einen 100-Prozent-Ansatz: Alle Beispiele zeigen die vollständige Gestaltung einer konkreten Abbildung. Es wird immer vom Ergebnis ausgegangen: Die Ausgangsfragen waren jeweils: Wie muss eine bestimmte Grafik aussehen oder wie können vorhandene Daten am ehesten visualisiert werden? Dabei wurde unabhängig von einer konkreten Software stets mit einer Skizze begonnen. Erst der nächste Schritt bestand dann darin, sich nach den dafür benötigten Werkzeugen (Paketen und Funktionen) umzusehen und diese anzuwenden.

[7] `http://www.rstudio.com/shiny/`

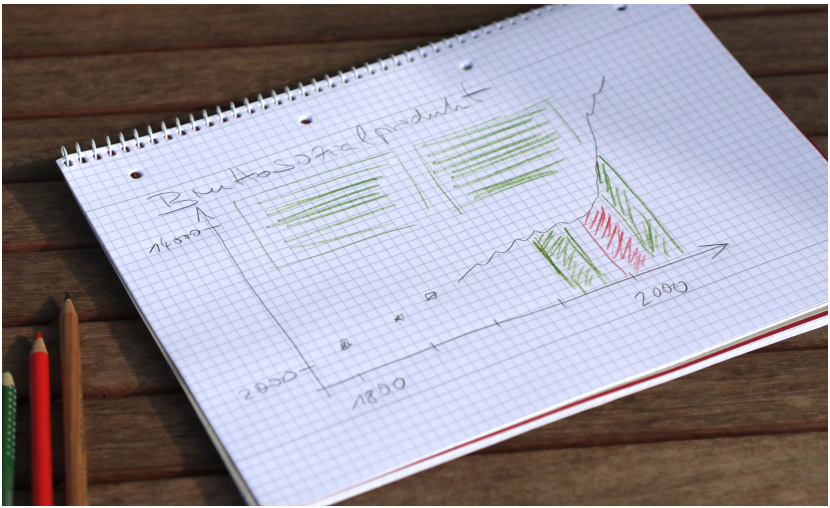

Abbildung 1.3: Skizze einer Abbildung

Die verwendeten Daten stammen ganz überwiegend aus der Sozialwissenschaft und der amtlichen Statistik, einige aus der Betriebswirtschaft, der Makroökonomie, der Politik, der Medizin, der Meteorologie oder den sozialen Medien. Mein Bestreben war, für alle ausgewählten Darstellungsformen geeignete Daten zu finden. Das ist sicher mal mehr, mal weniger gelungen. Die Daten wurden aber nicht „vorfrisiert", sondern in der Form verwendet, in der sie zur Verfügung standen. Dadurch ist zwar der Skriptumfang manchmal etwas größer als unter Laborbedingungen mit jeweils für die Aufgabe schon optimal aufbereiteten Daten. Andererseite ist das lebensnäher und kann Ihnen bei dem ein oder anderen Ihrer Daten-Fallstricke nützlich sein.

Alle Abbildungen sind als PDF-Datei konzipiert, so dass sie möglichst verlustfrei und flexibel weiterzuverwenden sind.

Im Durchschnitt waren für die Erstellung des Ergebnisses 40 Zeilen Code nötig. Von der ersten Idee bis zur fertigen Umsetzung verging pro Abbildung in der Regel ein Tag, manchmal eine Woche. Wenn Sie mit Ihren Daten etwas kommunizieren möchten, lohnt es sich meiner Ansicht nach, diese Zeit zu investieren.