DATA VISUALIZATION MADE SIMPLE

INSIGHTS INTO BECOMING VISUAL

KRISTEN SOSULSKI

ROUTLEDGE

Data Visualization Made Simple

Data Visualization Made Simple is a practical guide to the fundamentals, strategies, and real-world cases for data visualization, an essential skill required in today's information-rich world. With foundations rooted in statistics, psychology, and computer science, data visualization offers practitioners in almost every field a coherent way to share findings from original research, big data, learning analytics, and more.

In nine appealing chapters, the book:

- examines the role of data graphics in decision making, sharing information, sparking discussions, and inspiring future research;
- scrutinizes data graphics, deliberates on the messages they convey, and looks at options for design visualization; and
- includes cases and interviews to provide a contemporary view of how data graphics are used by professionals across industries.

Both novices and seasoned designers in education, business, and other areas can use this book's effective, linear process to develop data visualization literacy and promote exploratory, inquiry-based approaches to visualization problems.

Kristen Sosulski is Associate Professor of Information Systems and the Director of Learning Sciences for the W.R. Berkley Innovation Labs at New York University's Stern School of Business, USA.

Data Visualization Made Simple

Insights into Becoming Visual

Kristen Sosulski

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2019
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this title has been requested

ISBN: 978-1-138-50387-8 (hbk)
ISBN: 978-1-138-50391-5 (pbk)
ISBN: 978-1-315-14609-6 (ebk)

Typeset in Avenir Next
by Apex CoVantage, LLC

For Penn

Contents

Preface

Data visualization is the process of representing information graphically. Relationships, patterns, similarities, and differences are encoded through shape, color, position, and size. These visual representations of data can make your findings and ideas stand out.

Data visualization is an essential skill in our data-driven world. Almost every aspect of our daily routine generates data: the steps we take, the movies we watch, the goods we purchase, and the conversations we have. Much of this data, our digital exhaust, is stored waiting for someone to make sense of it. But why is anyone interested in these quotidian actions?

Imagine you are Nike, Netflix, Amazon, or Twitter. Your data helps these companies better understand you and other users like you. Companies utilize this information to target markets, develop new products, and ultimately outpace their competition by knowing their customers' habits and needs. However, such insights do not just "automagically" happen.

One does not simply transform data into information. It requires several steps: cleaning the data, formatting the data, interrogating the data, analyzing the data, and evaluating the results.

Let's take this a step further. Suppose you identify new markets your company should target. Would you know how to effectively share this information? Could you provide clear evidence that would convince your company to allocate resources to implement your recommendations?

What would you rather present: a spreadsheet with the raw data? Or a graphic that shows the data analyzed in an informative way? I imagine you would want to show your insight so that it could be understood by anyone from interns to executives.

Data visualization can help make access to data equitable. Data graphics with dashboard displays and/or web-based interfaces, can change an organization's culture regarding data use. Access to shared information can promote data-driven decision making throughout the organization.

Clear information presentations that support decision making in your organization can give you a leg up. Understanding data and making it clear for others via data graphics is the art of becoming visual.

The strategies in this book show you how to present clear evidence of your findings to your intended audience and tell engaging data stories through data visualization.

This book is written as a textbook for creatives, educators, entrepreneurs, and business leaders in a variety of industries. The data visualization field is rooted in statistics, psychology, and computer science, which makes it a practice in almost every field that involves data exploration and presentation. Whether you are a seasoned visualization designer or a novice, this book will serve as a primer and reference to becoming visual with data.

As a professor of information systems, my work lies at the intersection of technology, data, and business. I use data graphics in my practice for data exploration and presentation.

I teach executives, full-time MBA students, and train companies in the process of visualizing data. Teaching allows me to stay current with the latest software and challenges me to articulate the key concepts, techniques, and practices needed to become visual. The following chapters embody my data visualization practice and my course curriculum.

This book promotes both an exploratory and an inquiry-based approach to visualization. Data tasks are treated as visualization problems, and they use quantitative techniques from statistics and data mining to detect patterns and trends. You'll learn how to create clear, purposeful, and beautiful displays. Exercises accompany each chapter. This allows you to practice and apply the techniques presented.

How and why do professionals incorporate data visualization into their practice? To answer these questions, I engaged professionals in business analytics, human resources, marketing, research, education, politics, gaming, entrepreneurship, and project management to share their practice through brief case studies and interviews. The cases and interviews illustrate how people and organizations use data visualization to aid in their decision making, data exploration, data modeling, presentation, and reporting. My hope is that these diverse examples motivate you to make data visualization part of your practice.

By the end of this book, you will be able to create data graphics and use them with purpose.

How to Use This Book

This book is intended for use as a textbook on data visualization—the process of creating data graphics. There are five icons that will prompt you to try out a technique, learn more about a practice or topic, and show you how data visualization is used in organizations or one's profession.

Try It

Tutorials and exercises to guide you in becoming visual.

Pro Tip

Short-cuts and best practices from the field.

Sidebar

Additional resources to further your knowledge.

Use Case

An illustration of how data graphics are used in a specific field explained by a practitioner.

Interview With a Practitioner

Interviewer

Interviewee

Interviews with professionals who use data visualization in their work.

I

BECOMING VISUAL

This chapter answers the following questions:

What is data visualization?
Who are the visualization designers and what do they do?
Why use data visualization?
How can I incorporate data visualization into practice?

Data Visualization Made Simple: Insights into Becoming Visual is a contemporary view of how data graphics are used by professionals across industries. The book examines the role of data graphics in decision making, informing processes, sharing information, sparking discussions, and inspiring future research. It scrutinizes data graphics, deliberates on the message they convey, and looks at design visualizations.

Beautiful (and not so beautiful) charts and graphs are everywhere. Visualization of information is a human practice dating back to the Chauvet cave drawings, over 32,000 years ago (Christianson, 2012). The way we view everyday information, such as the weather, fitness progress, and account balances, is through visual interfaces. These interfaces aggregate and display key data points such as the temperature, calories burned, miles run, and personal rates of return. The charts we regularly use to show quantities and change over time, like bar charts and line graphs, were first employed in the late 1700s.

William Playfair (1786) is credited as the pioneer who showed economic data using bar charts. Playfair (1786) also invented the line graph. Playfair's work in the 1700s is paramount to the field of data visualization; it provided the foundation for future statistical data displays.

Forces of Change

Data visualization has gained immense popularity over the last five years. Many forces have contributed to the torrent of data graphics that we see all around us. First, there's a lot more data available in the world; we are living in the era of big data. From individuals to governments, there is a movement toward sharing data for public good. Platforms like Kaggle provide open data sets and a community to explore data, write and share code, and enter Machine Learning competitions. All of the services we employ, from AT&T to American Express, collect, mine, and share our data. Second, software to analyze and visualize data is ubiquitous. Tableau, for example, is designed for the explicit purpose of visualizing data. It's only been available for both Mac and PC users since 2014. Programming languages such as Python and R have packages, such as ggplot2 and plotly, that make the process of data visualization straightforward and manageable, even for non-programmers. Charts are no longer limited to static displays; they are dynamic, interactive, and animated. Third, the cost of hardware is decreasing while computing power is increasing, in line with or perhaps outpacing Moore's Law. Cloud computing has eliminated the barrier to data storage and processing power; it's possible to mine and visualize data without the economic and maintenance burdens. Fourth, education has embraced these technological advances. Top universities have established research centers and launched academic programs in data science, big data, business analytics and other subject-specific variations. These variations include healthcare analytics, learning analytics, sports analytics, and sustainability analytics. Furthermore, in the spirit of knowledge sharing and freemium content, online tutorials on how to do almost anything can be found on YouTube. For example, you can learn how to build data graphics through online tutorials. These resources complement this book, and I encourage you to explore them.

Trends in Data Visualization—Storytelling

The use of data graphics for *storytelling* is a popular technique employed to engage an audience. When well-designed data graphics are used in presentations, they highlight the key insights or points you want to accentuate. Storytelling is not limited to in-person presentations. Stories can be told through video, web narratives, and even through audience-driven interfaces.

How can we use visuals to tell engaging data stories and provide evidence of findings or insights? A picture may be worth a thousand words, but not all pictures are readable, interpretable, meaningful, or relevant. Figure 1.1 is a preview of three images that support data stories about Manhattan.[1]

Stories can begin with a question or line of inquiry.

Highlighting behaviors > Who's hailing a cab when the clocks strike midnight on New Year's Eve? Map A shows the location of taxi cab customer pickups at 12:00am on January 1, 2016.

Revealing similarities and differences > Where do the most motor vehicle accidents occur in Manhattan? Map B is a point map that shows the locations of each accident during the month of January 2016.

Displaying locations > Where can I pick up free Wi-Fi? Map C shows the location of each Wi-Fi hotspot in Manhattan.

In many TED Talks, presenters use charts to lead the audience through a narrative about an important topic or issue. Skilled presenters rarely show a graph on the screen without providing some context or explanation. Rather, they highlight specific data points for audience examination or they walk the audience through the graph by progressively revealing key data points.

Telling stories with data: Viewing Manhattan

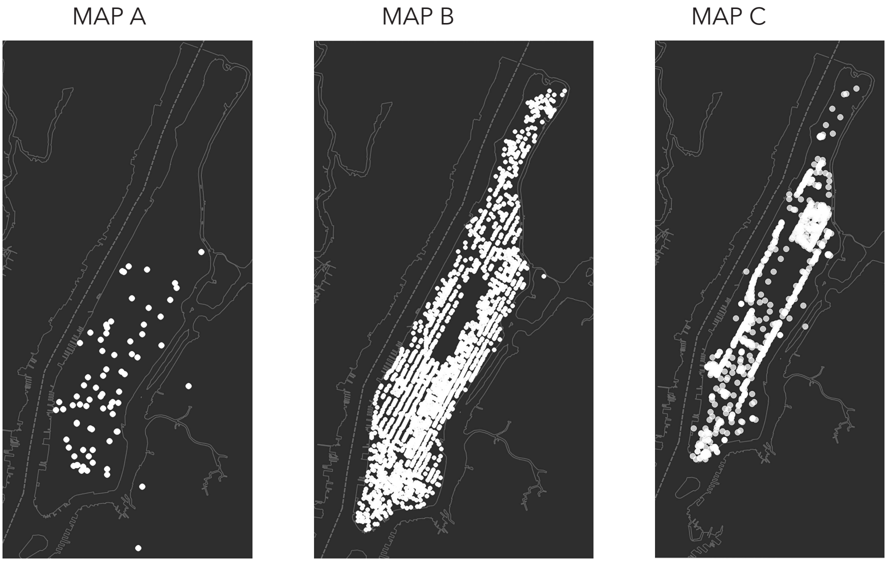
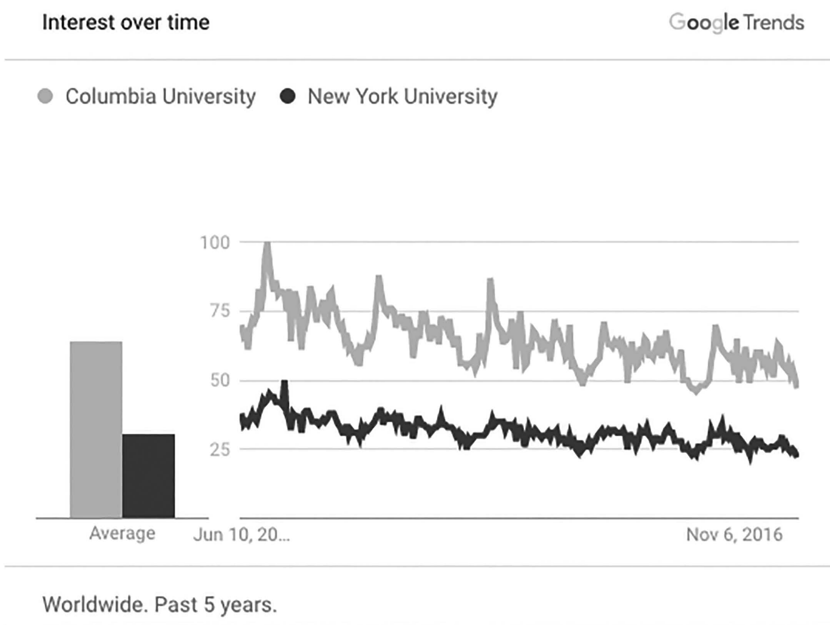


Figure 1.1 Viewing Manhattan through the lens of taxi hails, motor vehicle accidents, and Wi-Fi hotspots.

Examples of presenter-driven storytelling
http://becomingvisual.com/portfolio/presenterdrivenstories

Storytelling does not have to be presenter driven. User-driven storytelling is becoming increasingly popular utilizing data visualizations. For example, the Gapminder Foundation created an interface to view and explore public health data, human development trends and income distribution. The data graphics presented by the *New York Times* allow for rich exploration of the U.S. Census American Time Usage Survey, such as How Different Groups Spend Their Day. Google provides open access to explore Google search trends. With Google Trends, you can compare search volume of different keywords or topics over time. For example, interest in my two alma maters, Columbia University and New York University, is compared over time using a simple line graph. See Figure 1.2.

These are just a few examples of interfaces that are intended to help users build their own stories. Chapters VI—THE AUDIENCE and VII—THE PRESENTATION offer strategies and techniques for delivering presentations and telling stories with data graphics.



Source: Google Trends (www.google.com/trends)

Figure 1.2 Google search trends for New York University and Columbia University

Examples of user-driven storytelling
http://becomingvisual.com/portfolio/userdrivenstories

Trends in Data Visualization—Interactive Graphics

Static charts and graphics are antiquated. Interactive data graphics are the new norm. This has changed the way we interact with data. From media sites to individual blogs, interactive data graphics are used to engage and entice audiences. Users interact with graphics and search for meaning in the visual information presented, in essence creating their own narrative or story.

Data graphics with filters enable the querying or questioning data through a simple click of a button. The simplicity of visual interfaces that overlay data encourage inquiry without sophisticated training in data science or analytics. The ubiquity of these interfaces impels anyone who works with data to consider interactive data graphics as their new standard format.

Let's use a simple example of how interactive graphics have changed the way we engage with information. Let's say you wanted to know the median household income for your neighborhood. Let's assume you live in trendy Williamsburg, Brooklyn, 11211. How would you expect to be presented with the data?

THE MEDIAN HOUSEHOLD INCOME FOR WILLIAMSBURG, BROOKLYN, 11211 IS $50,943

This information is less than satisfying.

This is the middle household income value for all of the households in 11211. What you may really want to know is the distribution of household income in your neighborhood. The map below (see Figure 1.3)

The boundary of the zip code 11211 in Williamsburg, Brooklyn

Source: *Leaflet* | Data, imagery and map information provided by *CartoDB*, *OpenStreetMap*, and contributors, *CC-BY-SA*

Figure 1.3 A choropleth map showing the boundary of Williamsburg (11211) defined by the green line outlines Williamsburg in green. Within this neighborhood, there are many U.S. Census Block Groups[2] that are shaded using a grayscale. The darker the shade, the higher the median household income for that particular block group. This allows for comparisons of one census block to another.

Using the city-data.com website, you can highlight those blocks that have the highest and lowest median income by zooming in and selecting specific Census Block Groups.

Figure 1.4 shows the maximum median income for the area and Figure 1.5 shows the minimum.

These three maps show the median income for Williamsburg, Brooklyn in the context of others, rather a single number. The shading in all three maps in Figures 1.3, 1.4, and 1.5 designates the areas with higher (darker shades) versus lower (lighter shades) median household income.

Source: *Leaflet* | Data, imagery, and map information provided by *CartoDB*, *OpenStreetMap*, and contributors, *CC-BY-SA*

Figure 1.4 A Census Block Group (selected in green) has one of the highest median household incomes ($100,089).

Source: *Leaflet* | Data, imagery, and map information provided by *CartoDB*, *OpenStreetMap*, and contributors, *CC-BY-SA*

Figure 1.5 A Census Block Group (selected in green) has one of the lowest median household incomes ($6,442).

1.1 What Is Data Visualization?

In my experience, everyone defines this term slightly differently. Let's imagine that you are one of my data visualization graduate students. Before the course begins, I ask my students to define data visualization in their own words.

I encourage you to quickly take this survey to assess for yourself what you already know about data visualization at: http://becomingvisual.com/survey. Throughout this book, the examples from the survey will be referenced and explained.

On the first day of class, I display a word cloud of student definitions as shown in Figure 1.6. This image depicts the frequency of the top

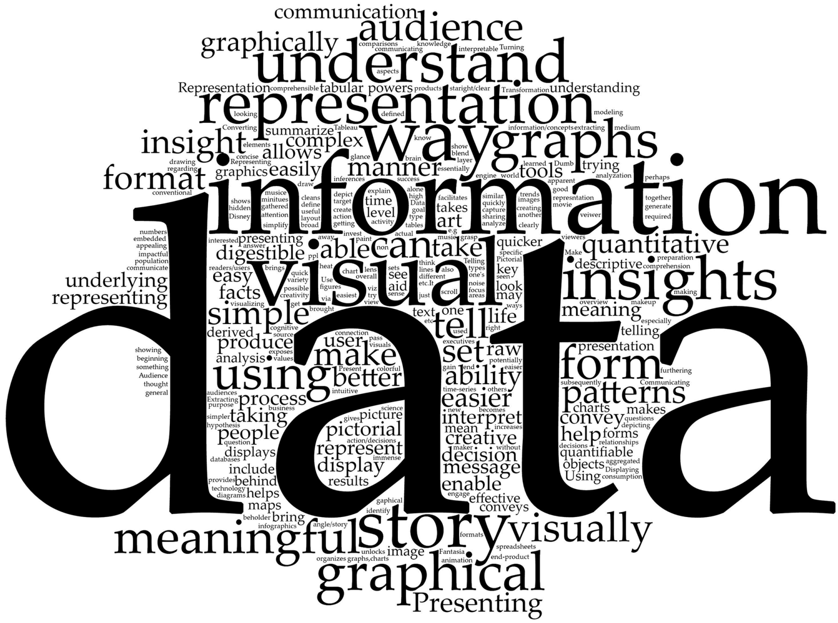


Figure 1.6 A word cloud that shows the frequency of the top 150 words used by students when asked to define data visualization

150 words from their definitions. The larger the word, the more times students used it in a definition. Note: The phrase "data visualization" has been filtered out.

Next, I reduce the list of words from 150 to 40 and re-graph it (see Figure 1.7). The words *data* and *information* stand out as the largest words. Then, we discuss the importance of transforming data into information.

Finally, I reduce the word cloud to the top five words (see Figure 1.8). This brings us to the key words that comprise the definition: **a visual way to tell a story with data and information.** This exercise always leads to an interesting conversation about how visualization is used in practice.

I conclude this exercise by sharing a few simple explanations by experts in the field.

Visualization is a graphical representation of some data or concepts.

–COLIN WARE, 2008, p. 20

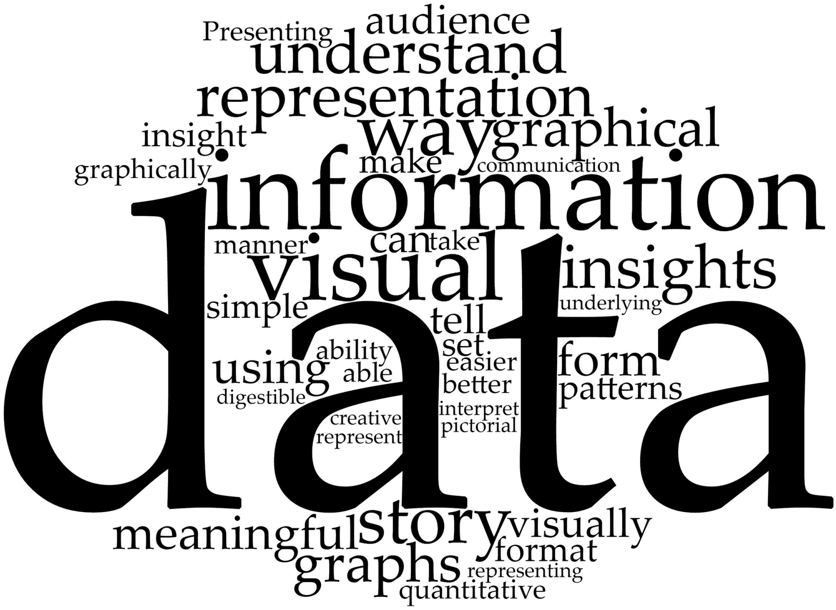


Figure 1.7 A word cloud that shows the frequency of the top 40 words used by students when asked to define data visualization

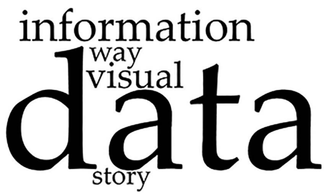


Figure 1.8 A word cloud that shows the frequency of the top five words used by students when asked to define data visualization

When a chart is presented properly, information just flows to the viewer in the clearest and most efficient way. There are no extra layers of colors, no enhancements to distract us from the clarity of the information.

–DONA WONG, 2010, p. 13

Visualization is a kind of narrative, providing a clear answer to a question without extraneous details.

–BEN FRY, 2008, p. 4

Visualization is often framed as a medium for storytelling. The numbers are the source material, and the graphs are how you describe the source.

—NATHAN YAU, 2013, p. 261

While some view data visualization as a technique, I define data visualization as a **process used to create data graphics**.

1.2 Who Are Visualization Designers and What Do They Do?

Anyone who works with data and visualizes it is a visualization designer. To produce a graphical representation of data, the designer engages in a process where the data is the input, the output is a graphic, and in between is a transformation of data into *an information graphic*. The transformation stage involves chart creation and refinement. After the graphic is refined, it becomes a communication device for use with a target audience.

This book will help you master the practice of data visualization design, whether you are just starting out, or have been working at it for a while. Given that you are reading this book, you may already have some visual instincts. For example, you may cringe when you see a slide presentation with a lot of text or become frustrated when you cannot find the information you need on a poorly designed website. Even if you think that you are not a visual person, you can still visualize data.

Becoming visual means you must develop a new habit.

Habit is a fixed tendency or pattern of behavior that is often repeated and is acquired by one's own experience or learning, whereas an instinct tends to be similar in nature to habit, but it is acquired naturally without any formal training, instruction or personal experience.

DIFFERENCE BETWEEN HABIT AND INSTINCT, 2017, para. 1

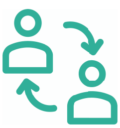
Essentially, this means you must integrate visualization into your workflow, rather than making it an extra step in the exploration, analysis and communication of information.

Developing a visual habit requires practice. This book provides many opportunities for such practice. There are conceptual and hands-on exercises at the end of each chapter. No amount of observing or reading will give you competence in visualizing information. The software available makes the actual creation of charts and graphs easy. However, the software will not fix bad data or provide you with worthwhile insights.

The exercises are designed to build your confidence in visualizing data. In addition, you can find visualization tutorials and real examples at becomingvisual.com.

1.3 Why Use Data Visualization?

Over the years, I've given numerous talks on data visualization to students, executives, and data gurus. In my experience, at first, most people want to learn how to best use the tools (see Chapter II—THE TOOLS). However, there is much more to the practice of visualization. There are several arguments for why data visualization is essential to your practice.



Reason one: to communicate

When data attributes are simplified into a visual language, patterns and trends can reveal themselves for easy comprehension. At the most fundamental level, a table of numbers is useful to look up a single value.

For example, what if you ran a product review site and wanted to know how many daily user reviews were written in a year? Table 1.1 makes it easy to see the total number of reviews by day. On January 4, there were 12 reviews.

How did you read this table of numbers? You probably read each value, individually, one at a time. However, a graph can help us see many values at once. For example, Figure 1.9 shows the number of daily reviews for a single year. You can see how the reviews have fluctuated over time, during each day of each month.

Visual displays combine many values into shapes that we can easily see as a whole, such as the line in the graph that shows the changing number of reviews over time. This enables efficient human information processing because many values can be perceived through a single line (Evelson, 2015), as illustrated in Figure 1.9.

Table 1.1 A table of data that shows the number of user reviews of products by day

Date	Total Reviews
1/2/2018	3
1/3/2018	27
1/4/2018	12
1/5/2018	23
1/6/2018	1
1/7/2018	0
1/8/2018	253
1/9/2018	238
1/10/2018	145

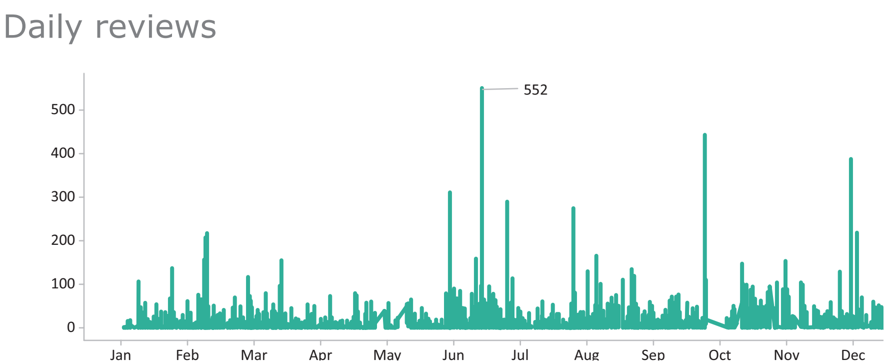


Figure 1.9 A line chart showing daily reviews for a single year

> The goal of visualization is to aid our understanding of data by leveraging the human visual system's highly tuned ability to see patterns, spot trends, and identify outliers.
>
> –HEER, Bostock, & Ogievetsky, 2010, p. 1

The arrangement of the data encodings (dots, lines, bars, shaded areas, bubbles, etc.) can reveal where the obvious correlations, relationships, anomalies, or patterns exist. For example, the chart on the left in Figure 1.10 shows a positive correlation while the chart on the right shows the presence of an outlier in the top right corner.

Reason two: transform data into information

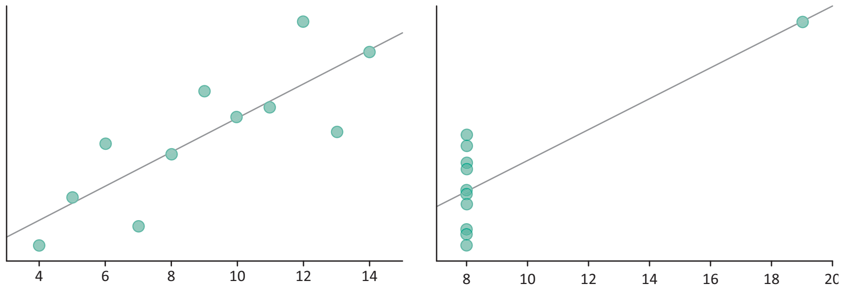


Figure 1.10 A chart showing correlation and outliers based on Anscombe (1973)

In this era of big data, visualization is a powerful way to make sense of the data. Big data is much more than just a lot of data. IBM data scientists break big data into four dimensions: volume, variety, velocity, and veracity.

Data differs with respect to its volume or physical size. This is measured in bytes, the speed in which it is generated (velocity), the forms it takes (variety), and its accuracy (veracity). These differences make data a challenge to work with but provide a terrific opportunity for data exploration.

Learn more about Big Data:
http://becomingvisual.com/portfolio/bigdata

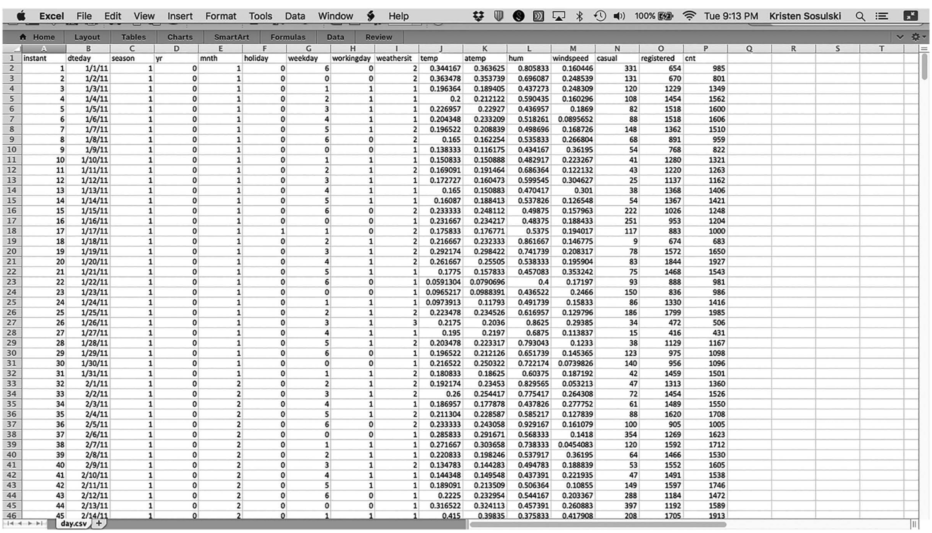
Think about the data you generate every day. For example, when you browse the web, all of your clickstreams and analytics are captured and collected on each page you view. All of your browsing history is saved in your web browser. When you call or text, that history is saved too. Every post, like, view, and click on each online platform from Facebook to Yelp is collected. This collected data is used by companies and researchers to learn more about how people interact (buy, sell, search, communicate, etc.) in online communities.

When it comes to the practical use of data visualization, there is a big difference between using real data to reflect real-world phenomena and the analytical process of modeling to make predictions. In the analysis phase, the data is interrogated to learn more, such as developing an understanding of the particular phenomenon. Then, by identifying a key insight, you can take the data a step further by **transforming a basic information graphic into a knowledge graphic**. To decode the data into usable knowledge requires use of appropriate models, statistics, and data mining techniques for data analysis. Once you make sense of the data insights, you may need to share them with others. This means you must communicate the results in a way that your audience can understand.

> Now, through ADV [Advanced Data Visualization], potential exists for nontraditional and more visually rich approaches, especially in regard to more complex (i.e., thousands of dimensions or attributes) or larger (i.e., billions of rows) data sets, to reveal insights not possible through conventional means.
>
> —EVELSON, 2011, para. 6

The challenge in working with a lot of data is that it can be difficult to view and interpret. For example, on my MacBook Air, I can only view 45 rows of data at any given time with a maximum of 20 attributes (columns) (see Figure 1.11).

Data visualization tools work within the limits of the screen to present data via an interface. The interface may include tools to question, filter, and explore the data visually. With modern software, visualizations can be configured to show deep and broad data sets (see Chapter IV—THE DATA). In addition, they can accommodate data that is dynamic and



	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	1/1/11	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
3	2	1/2/11	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
4	3	1/3/11	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
5	4	1/4/11	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
6	5	1/5/11	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
7	6	1/6/11	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
8	7	1/7/11	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
9	8	1/8/11	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
10	9	1/9/11	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
11	10	1/10/11	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
12	11	1/11/11	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
13	12	1/12/11	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
14	13	1/13/11	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
15	14	1/14/11	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421
16	15	1/15/11	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222	1026	1248
17	16	1/16/11	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251	953	1204
18	17	1/17/11	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117	883	1000
19	18	1/18/11	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9	674	683
20	19	1/19/11	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78	1572	1650
21	20	1/20/11	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83	1844	1927
22	21	1/21/11	1	0	1	0	5	1	1	0.1775	0.157833	0.457083	0.353242	75	1468	1543
23	22	1/22/11	1	0	1	0	6	0	1	0.0591304	0.0790696	0.4	0.17197	93	888	981
24	23	1/23/11	1	0	1	0	0	0	1	0.0965217	0.0988391	0.436522	0.2466	150	836	986
25	24	1/24/11	1	0	1	0	1	1	1	0.0973913	0.11793	0.491739	0.15833	86	1330	1416
26	25	1/25/11	1	0	1	0	2	1	2	0.223478	0.234526	0.616957	0.129796	186	1799	1985
27	26	1/26/11	1	0	1	0	3	1	3	0.2175	0.2036	0.8625	0.29385	34	472	506
28	27	1/27/11	1	0	1	0	4	1	1	0.195	0.2197	0.6875	0.113837	15	416	431
29	28	1/28/11	1	0	1	0	5	1	2	0.203478	0.223317	0.793043	0.1233	38	1129	1167
30	29	1/29/11	1	0	1	0	6	0	1	0.196522	0.212126	0.651739	0.145365	123	975	1098
31	30	1/30/11	1	0	1	0	0	0	1	0.216522	0.250322	0.722174	0.0739826	140	956	1096
32	31	1/31/11	1	0	1	0	1	1	2	0.180833	0.18625	0.60375	0.187192	42	1459	1501
33	32	2/1/11	1	0	2	0	2	1	2	0.192174	0.23453	0.829565	0.053213	47	1313	1360
34	33	2/2/11	1	0	2	0	3	1	2	0.26	0.254417	0.775417	0.264308	72	1454	1526
35	34	2/3/11	1	0	2	0	4	1	1	0.186957	0.177878	0.437826	0.277752	61	1489	1550
36	35	2/4/11	1	0	2	0	5	1	2	0.211304	0.228587	0.585217	0.127839	88	1620	1708
37	36	2/5/11	1	0	2	0	6	0	2	0.233333	0.243058	0.929167	0.161079	100	905	1005
38	37	2/6/11	1	0	2	0	0	0	1	0.285833	0.291671	0.568333	0.1418	354	1269	1623
39	38	2/7/11	1	0	2	0	1	1	1	0.271667	0.303658	0.738333	0.0454083	120	1592	1712
40	39	2/8/11	1	0	2	0	2	1	1	0.220833	0.198246	0.537917	0.36195	64	1466	1530
41	40	2/9/11	1	0	2	0	3	1	2	0.134783	0.144283	0.494783	0.188839	53	1552	1605
42	41	2/10/11	1	0	2	0	4	1	1	0.144348	0.149548	0.437391	0.221935	47	1491	1538
43	42	2/11/11	1	0	2	0	5	1	1	0.189091	0.213509	0.506364	0.10855	149	1597	1746
44	43	2/12/11	1	0	2	0	6	0	1	0.2225	0.232954	0.544167	0.203367	288	1184	1472
45	44	2/13/11	1	0	2	0	0	0	1	0.316522	0.324113	0.457391	0.260883	397	1192	1589
46	45	2/14/11	1	0	2	0	1	1	1	0.415	0.39835	0.375833	0.417908	208	1705	1913

Figure 1.11 An Excel spreadsheet open on a MacBook Air that shows the maximum amount of data that can be viewed at one time on my screen

can work with analysis tools for data interrogation through dashboard interfaces.

Reason three: to show evidence

Data graphics are used to show findings, new insights, or results. The data graphic serves as the visual evidence presented to the audience. The data graphic makes the evidence clear when it shows an interpretable result such as a trend or pattern. Data graphics are only as good as the insight or message communicated.

Using data graphics as evidence are best understood with an example from the field.

Interview with a practitioner

I interviewed Samantha Feldman from Gray Scalable who described how she uses data graphics to support her work.

Kristen Sosulski (KS)

Samantha Feldman (SF)

KS:

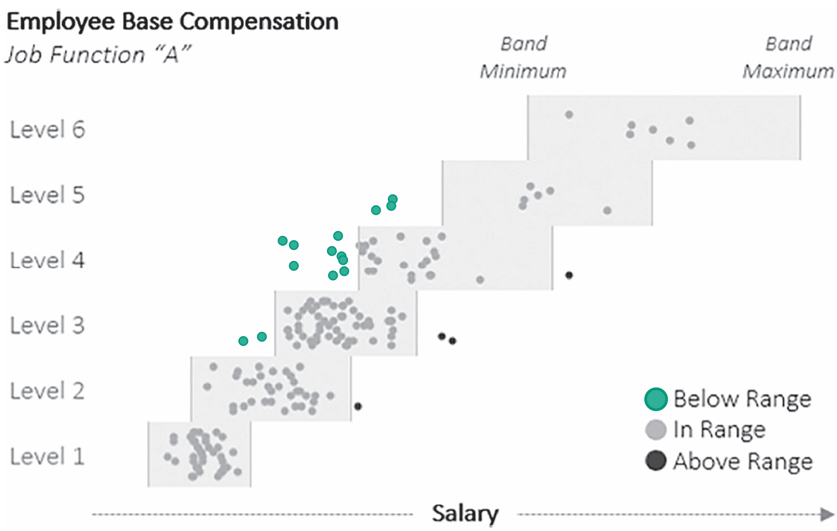
Who are you and what do you do?

SF:

I'm Samantha Feldman, I work for Gray Scalable, a consulting firm that provides human resources consulting for start-ups. Most of our clients have somewhere between 150 and 500 employees, are growing quickly, and hire us to help scale their recruiting practice, train their employees, and provide help with employee relations. I run the reporting and analytics arm of the business. The majority of my client projects center around helping our clients with employee compensation models, recruiting reporting, employee survey analysis, and pretty much any HR practice where numbers are involved.

KS: How do you use data visualization in your practice?

SF: The most common question I get from my clients is "how are we paying our employees relative to market rates?" This requires understanding broader market data, employee seniority, different job functions, and a few other variables. Most of our competitors provide results for each employee in spreadsheet format. Reviewing hundreds of rows of information makes it hard to get a holistic understanding of your current pay practices or see trends among different levels or job functions. I use data visualization to solve that, with what one of our clients named "the dot graph."



KS: What insight is evidenced by the graph provided? What did you do with that insight?

SF: In the graph, every employee is represented by a dot. From this one view, a client (usually the head of operations, finance, or HR) can quickly see that most employees fall within market salary bands.

A visualization like this also helps them spot how large the trouble spots are. In this case, I would point out that employees are being paid within range for the first three levels, but that employees start to fall behind around Levels 4 and 5. These could be employees who have been at the company long enough that their salary increases have not kept pace with the market. They could be underpaid for a number of other reasons (we also look to make sure gender is not a factor). From here, I do a deeper dive with the client to show who those employees are and devise a plan to correct employee compensation where needed.

KS:

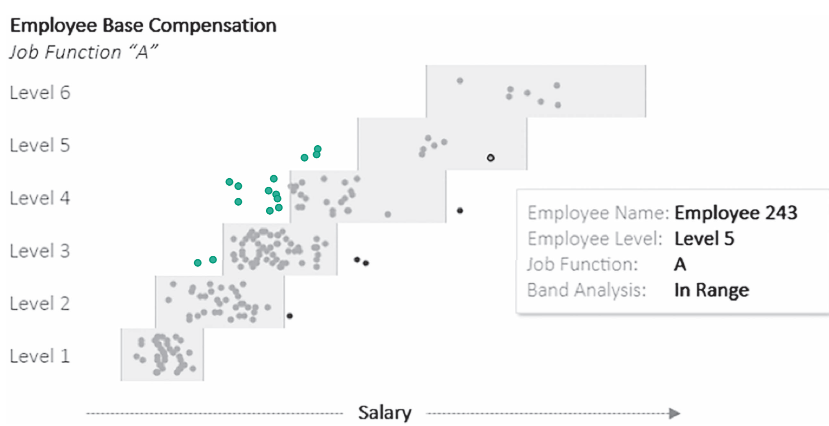
How did you create it? What was that data? What was the software? What would have been the alternative?

SF:

This graph was created with Tableau. It starts as a box and whisker plot—with the box and whisker reference lines removed. In their place, I make a reference band that is unique to each level and job function. One of the more important things I figured out how to do is to add a jitter calculation in Tableau (the reason the dots look scattered within the level). Because a company can have a group of employees that all make the same salary (e.g., 10 account managers who all make $65,000), this keeps the dots from overlapping and allows you to see the true volume of employees.

When I am on site with a client doing this with Tableau, I set up the tooltip so I can easily answer questions about specific employees as well, as shown below.

The alternative would be to view the data as a table by employee:



				Salary range in USD			
Employee name	**Salary**	**Job type**	**Level**	**Low**	**Mid**	**High**	**Band analysis**
Employee 1	$60,000	A	Level 1	$61,600	$75,900	$87,100	Below Range
Employee 2	$81,000	A	Level 1	$61,600	$75,900	$87,000	In Range
Employee 3	$110,000	A	Level 4	$110,000	$115,000	$135,000	In Range
Employee 4	$112,000	A	Level 3	$92,300	$112,300	$128,000	In Range
Employee 5	$60,000	A	Level 2	$49,000	$66,200	$81,400	In Range
Employee 6	$74,000	A	Level 2	$49,000	$66,200	$81,400	In Range
Employee 7	$74,000	A	Level 2	$49,000	$66,200	$81,400	In Range
Employee 8	$104,000	A	Level 2	$74,100	$92,600	$115,000	In Range
Employee 9	$58,000	A	Level 1	$40,400	$49,200	$61,000	In Range

> This allows for a more detailed view. I have used employee-level data when with clients to review outliers and summarize total cost to fix below range employees.

This example shows how data graphics are used in human resources consulting. Having the skills to support decision making in your organization through clear information presentations can give you a leg up. Understanding data and making it clear for others through data graphics is the process of becoming visual.

1.4 How Do You Incorporate the Visualization Process Into Practice?

Becoming visual requires many skills. You need to know how to process and mine data to identify findings, produce presentation quality graphics, and communicate your findings to your target audience.

As visualization designers, we are "melding the skills of computer science, statistics, artistic design, and storytelling."
—KATIE CUKIER, 2010, para 3

Expert Practice

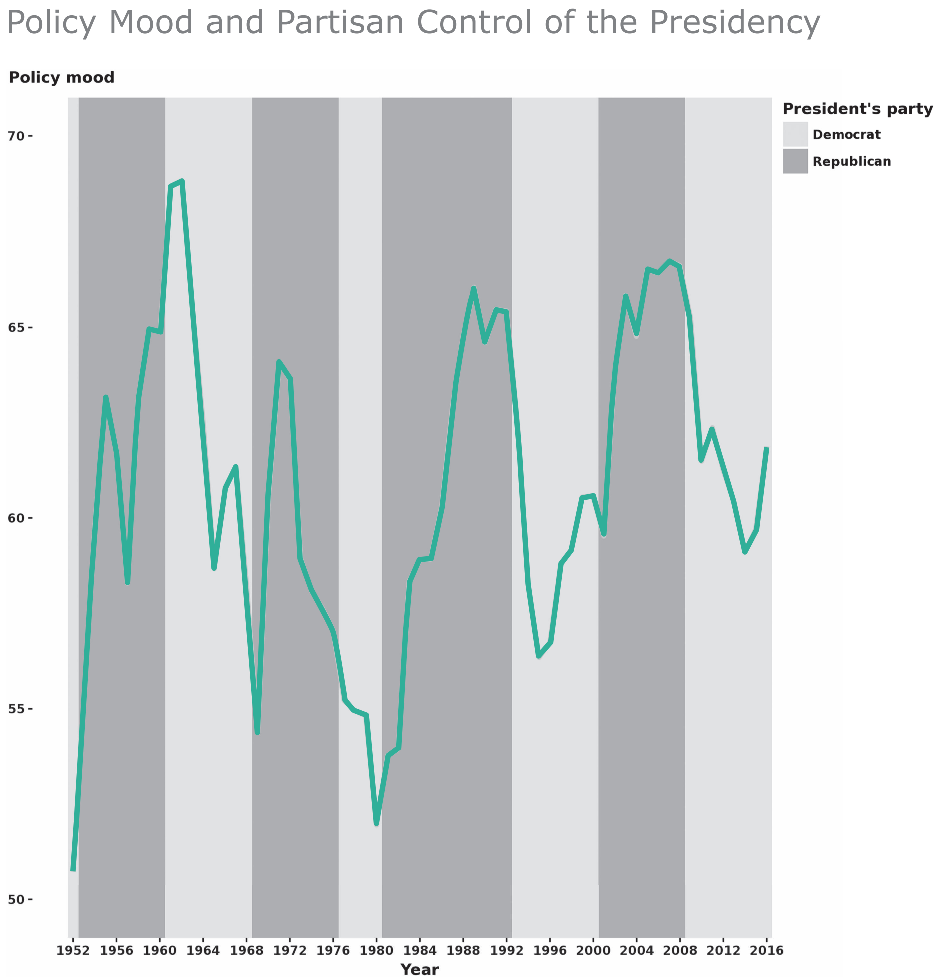
Gregory J. Wawro, a professor of Political Science at Columbia University discussed an example of how data graphics supported his teaching.

At the end of the Fall 2016 semester, I was looking for a visualization to help make an argument about what students should expect in terms of the future course of American politics. Many students were confused—some even distraught—about the results of the 2016 U.S. presidential election. I wanted to show them that the outcome was actually not all that unusual if we look at post-WWII dynamics in public policy preferences and partisan control of the presidency. As a political scientist, part of my job is to find systematic explanations for political phenomena, which has become somewhat more difficult given the unusual twists and turns we have witnessed in American politics recently. One thing that I emphasize to students is that political outcomes are often driven by larger, longer term forces that are difficult for individuals or single events to alter. For the 2016 election, a case can be made that the forces in play favored Republicans winning the White House, despite what just about every poll was predicting.

One such force is cyclicity in public opinion with respect to demand for liberal versus conservative policies. James A. Stimson, in his book *Public Opinion in America: Moods, Cycles, and Swings*, developed the concept of "policy mood" to better understand how demand for public policy works. Policy mood refers to "shared feelings" about issues and policies "that move over time and circumstance" and assumes that publics view issues through general dispositions (p. 20). To measure policy mood, Stimson developed a sophisticated algorithm to produce a general measure that aggregates a broad array of items across numerous surveys concerning opinions about various policies and issues. The algorithm addresses difficult problems with survey data, such as missing cases and variations in question wording, to construct a relatively simple, longitudinal measure that indicates whether the polity prefers more liberal or more conservative policies in a given year.

To visualize movement in public opinion and how it relates to election outcomes and representation, I used the ggplot package for R to plot policy mood against a background indicating which party controlled the presidency (higher values for mood indicate a preference for more liberal policies, lower values indicate a preference for more conservative policies).

There are two striking patterns that appear in the plot. The first is that elections tend to produce outcomes that are consistent with the direction of policy mood. When the public wants more conservative policies, the Republicans usually win the White House. When it wants more liberal policies, the Democrats are usually victorious. The second pattern, however, indicates that once a party wins the presidency, mood shifts in the opposite direction of the kind of policies we would anticipate that party to pursue. When Republicans control the White House, which suggests they are moving policy in a more conservative direction, policy mood generally trends in a more liberal direction. When a Democrat is president, policy mood trends in a more conservative direction. For example, mood moved from 59.5 in the first year of the George W. Bush administration to 66.6 during his last year in



office—approximately a 7-point change. During Barack Obama's first year as president, mood was 65.2, but declined to 61.9 at the end of his presidency. Interestingly, in 2015, mood had returned to approximately where it was at the beginning of the Bush administration. This implies that when policy action moves in a particular ideological direction, the public wants it to go in the other direction (or at least wants it to go less in the direction that it is heading). What is somewhat ironic is that, once the party that policy mood indicates is preferred wins the presidency, mood tends to shift away from the ideological predisposition of that party. Perhaps it is the case that the public experiences a kind of "buyers' remorse" when they give a party control of the White House. Or perhaps the party in power enacts policy that goes farther ideologically than what the public wants. Whatever the mechanism behind the pattern, there appears to be a cyclicity to policy mood that is related to oscillating control of the presidency.

Given that policy mood trended significantly in the conservative direction after the election of Obama in 2008, it would not have been surprising to see a Republican elected in 2016, irrespective of who that candidate was. Mood did tick upward just prior to the election, perhaps due to Republicans gaining control of the U.S. Senate after the 2014 elections. Indeed, some of the movement in mood throughout the series seems to be associated with which party controls Congress. In any case, we would predict, based on the historical dynamics revealed in the plot, that mood will trend in the more liberal direction during the presidency of Donald Trump, and if it trends strongly enough in that direction, it may very well lead to Democrats taking back the White House in the 2020 elections.

This example shows how a data graphic was used in classroom teaching to visualize movement in public opinion and how it relates to election outcomes. Throughout the book, practitioners share their practice with you through interviews. Five in-depth use cases with professionals that show you how data graphics are used in the context of work and research.

The followings chapters will guide you in the process of visualizing data for your practice.

CHAPTER II—**THE TOOLS** describes the popular software, platforms, and programming languages used to visualize data.

CHAPTER III—**THE GRAPHICS** presents over 30 types of charts and the insights that they best portray.

CHAPTER IV—**THE DATA** provides techniques for data preparation including data formatting and cleaning. Visual data exploration methods that aid in data understanding are presented with examples.

CHAPTER V—**THE DESIGN** demonstrates the application of design standards to improve readability, clarity, and accessibility of the data insights through graphics.

CHAPTER VI—**THE AUDIENCE** offers practical tips for telling stories with data that will resonate with your audience.

CHAPTER VII—**THE PRESENTATION** offers tactics for designing and delivering data presentations. The common pitfalls and how to avoid them are explained.

CHAPTER VIII—**THE CASES** illustrates how data graphics are used in practice through five case studies. Each case study showcases a unique approach to using data graphics in different settings.

CHAPTER IX—**THE END** synthesizes the key takeaways from each chapter into a concise roadmap to guide your visualization practice.

1.5 Exercises

1. Describe three ways visualization will be used in your workflow and practice.
2. The late Hans Rosling popularized the use of information graphics in presentations. He was a professor of international health and director of the Gapminder Foundation. Using a tool called Trendalyzer, Rosling runs an animation that shows the changes in poverty by country. Look at this video and answer the following questions: http://becomingvisual.com/portfolio/hansrosling
   a. Which attributes of Hans Rosling's presentation are especially effective? Explain why.
   b. What questions are being addressed by the presentation?
   c. What data is used to create the visualization?
   d. What symbols are used to represent the data?
3. Build three basic charts (using any visualization tool).
   a. Audience: design a chart for an executive to access sales over the past day.
   b. Data: download the data from http://becomingvisual.com/sales.xls
   c. Insight: show age and gender demographic that has the most sales.
   d. Display: select a chart type that best shows your insight.

Notes

1 The data is from NYC OpenData's website: https://data.cityofnewyork.us

2 "A Census Block Group is a geographical unit used by the United States Census Bureau which is between the Census Tract and the Census Block. It is the smallest geographical unit for which the bureau publishes sample data, i.e. data which is only collected from a fraction of all households" (Wikipedia, 2017, para 1—https://en.wikipedia.org/wiki/Census_block_group). Learn more at: www.census.gov/geo/reference/gtc/gtc_bg.html?cssp=SERP

Icons

Analytics by Kamal from the Noun Project
Big data by Eliricon from the Noun Project
Communication by ProSymbols from the Noun Project

Bibliography

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21. Retrieved from www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf

Christianson, S. (2012). *100 diagrams that changed the world*. New York, NY: Plume.

The City of New York. (2017). 2016 green taxi trip data. Retrieved from https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

Cukier, K. (2010). Show me: New ways of visualizing data. Retrieved from www.economist.com/node/15557455

Difference between habit and instinct. (2017). Retrieved from www.differencebetween.info/difference-between-habit-and-instinct

Evelson, B. (2011). What is ADV and why do we need it? Retrieved from http://blogs.forrester.com/boris_evelson/11-11-18-what_is_adv_and_why_do_we_need_it

Evelson, B. (2015). Build more effective data visualizations. Retrieved from http://blogs.forrester.com/boris_evelson/15-10-28-build_more_effective_data_visualizations

The four V's of big data. (2013). Retrieved from www.ibmbigdatahub.com/infographic/four-vs-big-data

Fry, B. (2008). *Visualizing data*. Beijing, China: O'Reilly Media.

Gapminder. Retrieved from www.gapminder.org/

Google trends. Retrieved from https://trends.google.com/trends/

Heer, J., Bostock, M., & Ogievetsky, V. (2010, May 1). A tour through the visualization zoo. *Queue*, *8*, 20–30. doi:10.1145/1794514.1805128

NYC wi-fi hotspot locations map. (2017). Retrieved from https://data.cityofnewyork.us/City-Government/NYC-Wi-Fi-Hotspot-Locations-Map/7agf-bcsq

NYPD motor vehicle collisions. (2017). Retrieved from https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Playfair, W. (1786). *Commercial and political atlas* (1st ed.). Printed for J. Debrett, London. (3rd ed., 1801) Printed for J. Wallis, London.

Rosling, H. (2006). The best stats you've ever seen. Retrieved from www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Ware, C. (2008). *Visual thinking for design*. Burlington, MA: Morgan Kaufmann

Wong, D. M. (2010). *The Wall Street Journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures*. New York, NY: W. W. Norton & Company.

Yau, N. (2011). *Visualize this: The FlowingData guide to design, visualization, and statistics*. Indianapolis, IN: Wiley.

Yau, N. (2013). *Data points: Visualization that means something*. Indianapolis, IN: Wiley.

Becoming Visual

Christianson, S. (2012). 100 diagrams that changed the world. New York, NY: Plume.
Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27(1), 17–21. Retrieved from www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf
The City of New York . (2017). 2016 green taxi trip data. Retrieved from https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb
Cukier, K. (2010). Show me: New ways of visualizing data. Retrieved from www.economist.com/node/15557455
Difference between habit and instinct. (2017). Retrieved from www.differencebetween.info/difference-between-habit-and-instinct
Evelson, B. (2011). What is ADV and why do we need it? Retrieved from http://blogs.forrester.com/boris_evelson/11-11-18-what_is_adv_and_why_do_we_need_it
Evelson, B. (2015). Build more effective data visualizations. Retrieved from http://blogs.forrester.com/boris_evelson/15-10-28-build_more_effective_data_visualizations
The four V's of big data. (2013). Retrieved from www.ibmbigdatahub.com/infographic/four-vs-big-data
Fry, B. (2008). Visualizing data. Beijing, China: O'Reilly Media.
Gapminder . Retrieved from www.gapminder.org/
Google trends . Retrieved from https://trends.google.com/trends/
Heer, J. , Bostock, M. , & Ogievetsky, V. (2010, May 1). A tour through the visualization zoo. Queue, 8, 20–30. doi:10.1145/1794514.1805128
NYC wi-fi hotspot locations map. (2017). Retrieved from https://data.cityofnewyork.us/City-Government/NYC-Wi-Fi-Hotspot-Locations-Map/7agf-bcsq
NYPD motor vehicle collisions. (2017). Retrieved from https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
Playfair, W. (1786). Commercial and political atlas (1st ed.). Printed for J. Debrett , London. (3rd ed., 1801) Printed for J. Wallis , London.
Rosling, H. (2006). The best stats you've ever seen. Retrieved from www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen
Ware, C. (2008). Visual thinking for design. Burlington, MA: Morgan Kaufmann
Wong, D. M. (2010). The Wall Street Journal guide to information graphics: The dos and don'ts of presenting data, facts, and figures. New York, NY: W. W. Norton & Company.
Yau, N. (2011). Visualize this: The FlowingData guide to design, visualization, and statistics. Indianapolis, IN: Wiley.
Yau, N. (2013). Data points: Visualization that means something. Indianapolis, IN: Wiley.

The Tools

Evelson, B. , & Yuhanna, N. (2002). The Forrester Wave: Advanced Data Visualization (ADV) Platforms, Q3 2012.

The Graphics

The City of New York . (2014). Public recycling bins. Retrieved from https://data.cityofnewyork.us/Environment/Public-Recycling-Bins/sxx4-xhzg
Corum, J. (2011). The class of 2011. Retrieved from www.nytimes.com/interactive/2011/06/10/education/commencement-speeches-graphic.html
Ognyanova, K. (2016). Basic and advanced network visualization with R. Retrieved from www.kateto.net/wp-content/uploads/2016/04/Sunbelt%202016%20R%20Network%20Visualization%20Handout.pdf

Ribecca, S. (2013). The data visualisation catalogue. Retrieved from http://datavizcatalogue.com/index.html
Wong, D. M. (2010). The Wall Street Journal guide to information graphics: The dos and don’ts of presenting data, facts, and figures. New York, NY: W. W. Norton & Company.
Yi, J. S. , Kang, Y. , Stasko, J. T. , & Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1224–1231. doi:10.1109/TVCG.2007.70515

The Data

Bernard, M. , Lida, B. , Riley, S. , Hackler, T. , & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best? Retrieved from http://usabilitynews.org/a-comparison-of-popular-online-fonts-which-size-and-type-is-best/
Calkins, K. G. (2005). Definitions, uses, data types, and levels of measurement. Retrieved from www.andrews.edu/~calkins/math/edrm611/edrm01.htm#DATA_TYPE
Chernoff, H. (2017). Casino player dataset.
Fanaee-T, H. , & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 1–15. Retrieved from http://dx.doi.org/10.1007/s13748-013-0040-3
Level of measurement . (2008). In W. Kirch (Ed.), Encyclopedia of public health (pp. 851–852). Dordrecht, The Netherlands: Springer. Retrieved from https://link.springer.com/referenceworkentry/10.1007%2F978-1-4020-5614-7_1971 on April 1, 2018
Nasdaq . Apple Inc. common stock quote and summary data. Retrieved from www.nasdaq.com/symbol/aapl on June 1, 2017
Paradi, D. (2017). Using graphs and tables on presentation slides. Retrieved from www.thinkoutsidetheslide.com/using-graphs-and-tables-on-presentation-slides/
Publication manual of the American Psychological Association. (2010). (6th ed., repr. ed.). Washington, DC: American Psychological Association.
Wheelan, C. J. (2013). Naked statistics (1st ed.). New York, NY: W. W. Norton & Company.
Wickham, H. , & Grolemund, G. (2016). R for data science (1st ed.). O’Reilly Media. Retrieved from http://proquestcombo.safaribooksonline.com/9781491910382
Zimmer, J. (2010). PowerPoint math: The 1–6–6 rule. Retrieved from https://mannerofspeaking.org/2010/03/04/powerpoint-math-the-1-6-6-rule/

The Design

Bureau of Labor Statistics. (2017). Labor force statistics from the current population survey. Retrieved from https://data.bls.gov/timeseries/LNS14000000
Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Burlingame, CA: Analytics Press.
Nasdaq . Apple Inc. common stock quote and summary data. Retrieved from www.nasdaq.com/symbol/aapl on June 1, 2017
Tufte, E. R. (1997). Visual explanations: Images and quantities, evidence and narrative. Cheshire, CT: Graphics Press. Retrieved from http://catalog.hathitrust.org/Record/003159488
Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, CT: Graphics Press.
Wong, D. M. (2010). The Wall Street Journal guide to information graphics: The dos and don’ts of presenting data, facts, and figures. New York, NY: W. W. Norton & Company.
The World Bank. (2017). World development indicators. Retrieved from http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators&preview=on on September 15, 2017

The Audience

Chen, C. (2005). Top 10 unsolved information visualization problems. IEEE Computer Graphics and Applications, 25(4). Retrieved from http://ieeexplore.ieee.org/document/1463074/

Fanaee-T, H. , & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, 2(2), 1–15. doi:10.1007/s13748-013-0040-3

Global Policy Forum. (2017). Comparison of the world’s 25 largest corporations with the GDP of selected countries. Retrieved from www.globalpolicy.org/component/content/article/150-general/50950-comparison-of-the-worlds-25-largest-corporations-with-the-gdp-of-selected-countries.html

Healey, C. G. , & Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. IEEE Transactions on Visualization and Computer Graphics, 18(7), 1170–1188.

Mayer, R. E. (2001). Multimedia learning. Cambridge: Cambridge University Press.

Ware, C. (2013). Information visualization: Perception for design (3rd ed.). Burlington, MA: Morgan Kaufmann.

The Presentation

Akinson, C. , & Mayer, R. (2004). Five ways to reduce PowerPoint overload. Retrieved from http://www.bhertz.nl/wp-content/uploads/2011/07/atkinson_mayer_powerpoint_4_23_04.pdf

Bernard, M. , Lida, B. , Riley, S. , Hackler, T. , & Janzen, K. (2002). A comparison of popular online fonts: Which size and type is best? Retrieved from http://usabilitynews.org/a-comparison-of-popular-online-fonts-which-size-and-type-is-best

Paradi, D. (2017). Using graphs and tables on presentation slides. Retrieved from www.thinkoutsidetheslide.com/using-graphs-and-tables-on-presentation-slides

Publication manual of the American Psychological Association. (2010). (6th ed., repr. ed.). Washington, DC: American Psychological Association.

Zimmer, J. (2010). PowerPoint math: The 1–6–6 rule. Retrieved from https://mannerofspeaking.org/2010/03/04/powerpoint-math-the-1-6-6-rule

The End

Automaton . Retrieved from https://en.wikipedia.org/wiki/Automaton on January 1, 2018

Party favor . Retrieved from https://en.wikipedia.org/wiki/Party_favor on January 1, 2018