Data Visualization and Statistical Graphics in Big Data Analysis

Dianne Cook,[1] Eun-Kyung Lee,[2] and Mahbubul Majumder[3]

[1]Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia; email: dicook@monash.edu

[2]Department of Statistics, Ewha Womans University, Seoul 120-750, Korea; email: lee.eunk@ewha.ac.kr

[3]Department of Mathematics, University of Nebraska, Omaha, Nebraska 68182; email: mmajumder@unomaha.edu

Annu. Rev. Stat. Appl. 2016. 3:133–59

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-041715-033420

Keywords

exploratory data analysis, information visualization, visual analytics, high-dimensional data, interactive graphics

Abstract

This article discusses the role of data visualization in the process of analyzing big data. We describe the historical origins of statistical graphics, from the birth of exploratory data analysis to the impacts of statistical graphics on practice today. We present examples of contemporary data visualizations in the process of exploring airline traffic, global standardized test scores, election monitoring, Wikipedia edits, the housing crisis as observed in San Francisco, and the mining of credit card databases. We provide a review of recent literature. Good data visualization yields better models and predictions and allows for the discovery of the unexpected.

1. INTRODUCTION

In the 1970s, J.W. Tukey introduced the world to exploratory data analysis (EDA). Data visualization was a major component of this area, and Tukey made substantial contributions to statistical graphics. (A good summary of his contributions can be found in Friedman & Stuetzle 2002.) His philosophy was that good pictures of data can reveal what we never expected to see. His pencil-and-paper method, the stem-and-leaf plot, is universally taught in introductory statistics, and the fruits of his experiments in plotting high-dimensional data in the PRIM-9 system can be found in today's data visualization software. Even widely visible systems such as GapMinder (**http://www.gapminder.org**) and Baby NameVoyager (**http://www.babynamewizard.com**) owe some credit to interactive graphics that arose in the first years of EDA research.

It is surprising that the stem-and-leaf plot has persisted in the classroom to the present day. The pencil-and-paper methods were essential in the 1970s because access to the technology to create interactive graphics was limited. Today, there is almost universal access to technology for data analysis and little need for hand-sketching numbers. Today's EDA is focused on harnessing good computer-generated plots of data. Tukey was fortunate enough to have access to state-of-the-art technology and was an early advocate of harnessing technology for data analysis. Forty-five years ago, he foresaw today's technological big-data world, the importance of computational tools for statistical analysis, and how good utilization of technology can attract the best young talent to the field.

It is important to realize that EDA did not arise in a vacuum. Applied statistical practice has always utilized data plots before modeling to check assumptions and after modeling to assess the fit. Crowder & Hand (1990) make this clear:

> The first thing to do with data is to look at them . . . . [This] usually means tabulating and plotting the data in many different ways to "see what's going on". With the wide availability of computer packages and graphics nowadays there is no excuse for ducking the labour of this preliminary phase, and it may save some red faces later. (Crowder & Hand 1990, p. 130)

Big data provides new challenges for data visualization. The ability to make good data plots is an indispensable component of wrangling big data, a term that means something different depending on whom you listen to, what you read, or whom you chat with. The working definition of big data for this article is based not only on the size of the data in terms of variables or samples but also on its complexity. Examples include data that have hundreds of variables stored in many related tables, large databases that have been collected and perhaps neglected for many years, repositories of emails, health records from machines in doctors' offices that automatically file information, communications networks from social media, scores from tests administered to youth across the globe, huge quantities of data simulated from global climate models to assess the impact of climate change, and business data being collected and stored in new systems such as Hadoop. Big data potentially informs us about our world, including how to be more efficient in business operations or in the delivery of health care, what statistics we need to improve to get our tennis game to the level of a champion, or who the key people are that bind a social network into a cohesive group.

Tukey's EDA and the tools for plotting data are all around us today. Knowing how to effectively leverage visualization is a fundamental skill in today's society. New methods and software have made this easy and accessible for most people. Big data, open data, and open source software make this a golden age for data visualization.

In classical statistical inference and experimental studies, the role of data is primarily to prove or disprove a conjecture. In this case, the distribution of a test statistic is preferably derived theoretically, and data are used to obtain an empirical distribution when using theory is not possible. In any case, the role of data is passive and has a specific purpose. We can illustrate this by using an example: In a situation in which researchers want to test the population mean being equal to a certain value, they do not need the data to find the distribution of the test statistic if certain assumptions are met. Data are used only to find the evidence for or against the hypothesis. The formation of a hypothesis and the derivation of test statistics do not require data. Data are only needed to make the final decision.

However, the advent of big data has changed the classical way of thinking. A researcher with an enormous amount of data does not necessarily have a well-defined hypothesis or testing methods in mind. The challenge is to explore data and discover hidden value in the data, which, later, may lead to more formal hypotheses that can be tested using classical methods. Big data gives the data a more active role. Thus, big data analysis essentially requires adopting EDA and visualization as early steps.

This article has two components: (*a*) contemporary illustrations of the usefulness of data visualization for understanding data and (*b*) a review of the literature on methodology for big data visualization. Inevitably, the review cannot be entirely comprehensive, and we ask for the reader's leniency up-front if we do not mention all of the important contributions. We hope that our coverage points to a broad selection of advances that will entice the reader to use these as a starting point to dig deeper and independently discover other interesting developments.

2. ILLUSTRATIONS OF VISUALIZATION FOR UNDERSTANDING DATA

2.1. How to Win a Data Mining Challenge

In two examples from 2014, the use of graphics played a key role in data mining teams winning competitions. In both cases, the teams used data preprocessing involving the creation of data plots to help them understand what they were working with and identify problems with the data that needed to be addressed before they could make effective models. Large, complex data sets must be effectively cleaned, transformed, and preprocessed before they can be utilized. Visualization plays a key role. This is just as important for big data as it was in Tukey's day, even though the technology has radically changed.

2.1.1. Kaggle Health Heritage Prize winner's advice. In April 2011, Kaggle posted the details of the Heritage Health Prize Competition, whose slogan was "Improve Healthcare, Win $3,000,000." Dr. Phil Brierley was part of the three-person team that won the first two milestone awards of $230,000, and he combined forces with another team to win the final prize of $500,000. This competition is typical of the big data challenges of today: Large amounts of hospital admissions data were used to develop models to improve the efficiency of health care spending. In Brierley's postprize interviews, there are echoes of Tukey's long-ago words: "In many of the analytics problems I have been involved in, the problem you end up dealing with is not the one you initially were briefed to solve. These new problems are always discovered by visualising the data in some way and spotting curious patterns" (Brierley 2011).

There is a concrete example on Dr. Brierley's blog (Brierley 2011) that illustrates how a plot indicated to him that a different challenge was not worth entering. It was an algorithmic trading

challenge with data from the London Stock Exchange, and he made a simple time series plot showing the timing of liquidity shocks, from which he observes:

> Now it is quite clear there is something going on at 1pm, 2:30pm, after 3:30pm and at 4pm.
>
> Interestingly these spikes are only evident when all commodities are looked at together, they are not as obvious in any individual commodity.
>
> My first question if I was solving a business problem would be to return to the business to get more insight in what was going on here. My initial thoughts were lunch breaks and the opening times of other Stock Exchanges around the world—as 3:30pm London time could be around opening time in New York.
>
> Understanding the cause of these peaks is important as you would expect the reaction to them (the problem to solve) to be a function of the cause.
>
> If we did discover it was the opening times of other exchanges, then I would ask for extra information like the specific dates, so I could calculate when these peaks would occur in the future when the clocks changed. We do not have this information at the current time, or even the day of the week (it can be inferred but not accurately as there will be public holidays when the exchanges are closed).
>
> As it stands any models built could potentially fail on the leaderboard (or real life) data as our model might think 2:30pm is a special time, whereas really it is when another exchange opens, or when people come back from lunch. We need this causal information rather than just dealing with the effect—time differences change—lunch breaks may change.
>
> The current competition data is potentially lacking the full information required to build a model that is as robust as possible over time. (Brierley 2011)

2.1.2. 2014 Data Mining Cup winners' strategies. Each year, Prudsys AG challenges students with the Data Mining Cup (DMC) competition. In 2014, the problem was announced to student teams on April 2, and the deadline for final entries was May 14. The student teams had six weeks to develop a solution for a data mining problem on the topic of optimal return prognosis. More specifically, the goal was to use an online shop's historical purchase data to come up with a model for new orders that would calculate the probability of a purchase leading to a return. In this year's competition, a team of students from Iowa State University (ISU) was the first North American team to win. A key component of that win was the preprocessing of the data using graphics, which helped the team learn about their data and informed their modeling.

Figure 1 shows one plot used early in the competition by the ISU DMC team to examine return rates by time and product. The plot is ugly but informative. Two major structures are immediately visible: new product introductions in July 2012 and January 2013. The most important feature, though, is that the items to be predicted were in the third season of the time period, information that was crucial to construct good training and test sets for model building.

2.2. Visualization Is Not Prediction, and This Is OK

Much of statistics is preoccupied with predictive models, which are important. But an equally important part of working with big data is to develop methodology for helping analysts explore and understand what patterns are present. We might call this "playing in the sandbox." In what follows, several examples from our own work and other published work are shown that illustrate visualization being used in the process of analyzing big data sets. Generally, what can be learned

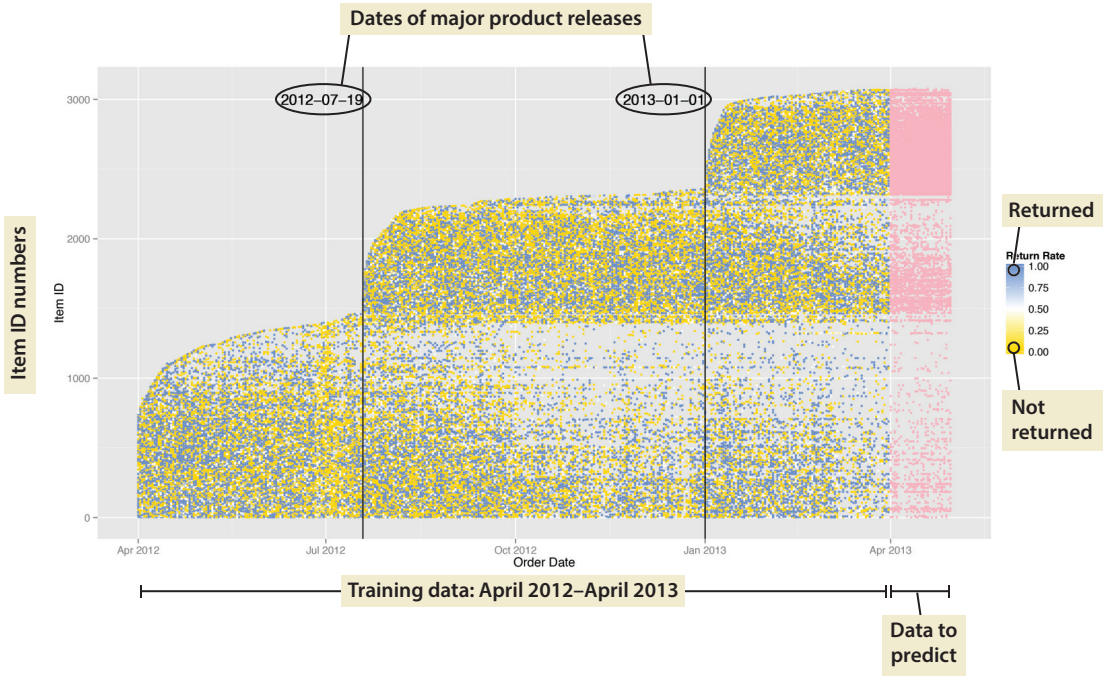


Figure 1

One of the preliminary plots made by the Iowa State University Data Mining Challenge team, with annotations. The item identification (ID) number is plotted against the order date and colored by the return rate. The students noticed that the data to be predicted (*pink*) did not look like the training data provided for building a model (*blue/yellow*). Yellow indicates that the ordered items were kept, blue means that they were returned. Figure adapted with permission from Xiaoyue Cheng.

about the data from plots can be very different from what would be learned by modeling and prediction, which means both types of summarization are equally important.

2.2.1. Organisation for Economic Co-operation and Development Programme for International Student Assessment. The Programme for International Student Assessment is a triennial survey conducted by the Organisation for Economic Co-operation and Development (OECD) with a rotating emphasis on mathematics, reading, or science. In 2012, the emphasis was on mathematics. The data were made available by the OECD as part of a data challenge for the useR! 2014 conference. Entries to the competition can be found at **http://www.oecd.org/pisa/pisaproducts/datavisualizationcontest.htm**.

This qualifies as big data because much information is collected in addition to the test scores. In the student table, there are records for approximately 500,000 students from 65 different countries, along with 635 variables that include information about gender, language, household possessions, attitude toward math, use of the Internet, and many other aspects of the students' lives. The parent table has 143 variables from 100,000 parent-completed surveys providing information about the students' households, such as if both parents were in the home or if it was a foster home, parents' occupations, and how the child's school was selected. The school table contains survey results completed by 18,000 school principals and includes 291 variables. These variables include information about numbers of teachers, supply shortages, teacher turnover, educational background of teachers, and streaming of classes. There are many different questions that we

might try to answer with these data. After the magnitude of the data was determined, by making quick counts of each of the tables provided, and examining the data dictionaries, our group hashed out possible questions and expected associations that we might see in the data (L. Fostvedt, A. Shum, I. Lyttle & D. Cook, manuscript in preparation).

One issue that we were interested in was the gender gap between boys and girls in math. We hear about this in the media frequently, and we were interested to see if evidence of a gap was present in this multinational test data. To examine this question, we calculated the difference between the mean math test scores for boys and girls in each of the countries and plotted it. We used sample weights when calculating the averages. The result is shown in **Figure 2**: The absence of a universal math gap runs counter to the claims of the popular press. These data represent an observational study and can only inform us about association. To understand some potential reasons why the gap does not exist would involve additional investigation of the samples used in each country. One quick check reveals that it cannot be explained by different proportions of boys and girls being tested: Proportions are roughly the same in all countries, so the math gap in favor of girls in some countries is not due to just a few top girls being tested.

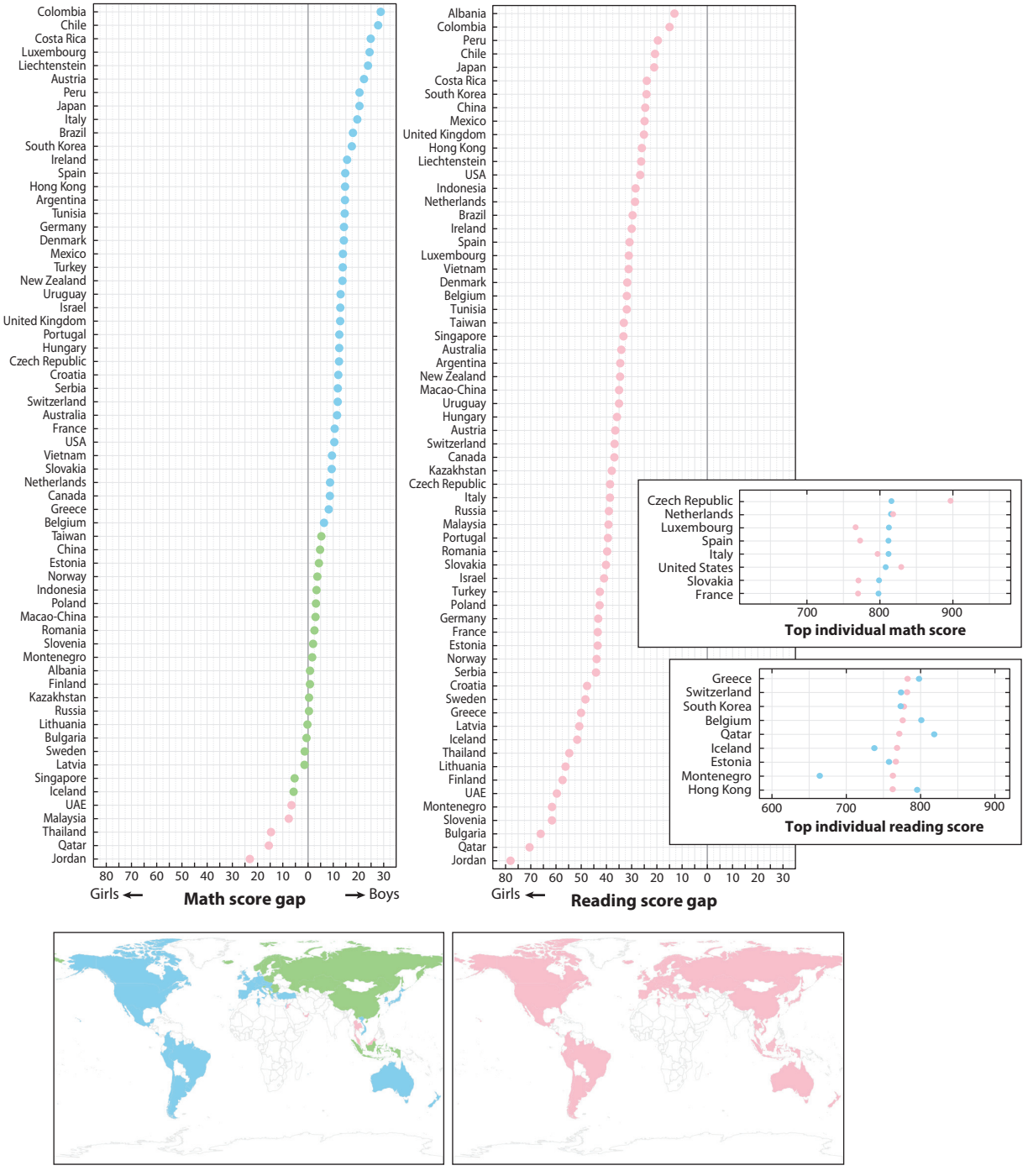
These data abound with information ripe for exploration. We can learn about different associations between demographic factors and educational achievement in countries around the globe. These associations could be mined to form the basis for follow-up experimental studies. Visualization provides an excellent way to mine these associations.

2.2.2. Elections. In the 2008 US presidential election cycle, a young man named Nate Silver burst onto the world stage with an accurate prediction of an Obama win. His website **http://fivethirtyeight.com/** (538) has now expanded from politics to cover data stories in economics, science, life, and sports. To obtain his accurate prediction of the election outcomes, he aggregated polls from different sources, but an important component was to adjust and weight the polls from different pollsters. **Figure 3** shows the polls from major pollsters in the 100 days leading up to the election, as reported by Mosley et al. (2010). The percentage difference in favor of Obama or McCain is plotted against the day each poll was released according to the website **http://www.electoral-vote.com/**. There is much variation in these polls, even when they are conducted in similar time frames—the variation in results can be as high as 10 percentage points.

There are clear differences between polls produced by different pollsters. DailyKos is consistently higher than the trend and consistently produced the most pro-Obama results. Rasmussen tended to be fairly close to the trend or below it (pro-McCain). Hotline was varied early on in the season, but closer to election day, it was near the average of the other polls. Gallup is noticeably varied: It had some of the most pro-McCain results as well as the most pro-Obama results. Gallup is a legacy US pollster that dates back to the early 1900s, and we would have expected more reliable polling numbers than were observed. Interestingly, the 538 site now has detailed ratings of the major pollsters operating in the United States, and Gallup scores a poor C+. The plot of

Figure 2

Examining the gender gap in math and reading by country. Plots of mean difference are shown by country, along with maps. Color indicates the magnitude of the gap, with blue indicating more than 5 points in favor of boys, pink indicating more than 5 points in favor of girls, and green indicating that the genders are within 5 points. Surprisingly, these data indicate that the gender gap in math is not universal—many countries do not have a gap, and a few countries show a gap in favor of girls. In contrast, the reading gap is universally in favor of girls in all of the countries in this study. On an individual scale, the small plots show the top boys' (*blue*) and girls' (*pink*) scores in a few countries, and the story is different. Even in countries with a big gender gap in math (e.g., the United States has a 10-point gap), the top score for that year was attained by a girl. Similarly, for reading, individual boys top the reading score in many countries. One glaringly obvious deficiency shown in the data from the maps is the lack of information from the continent of Africa.



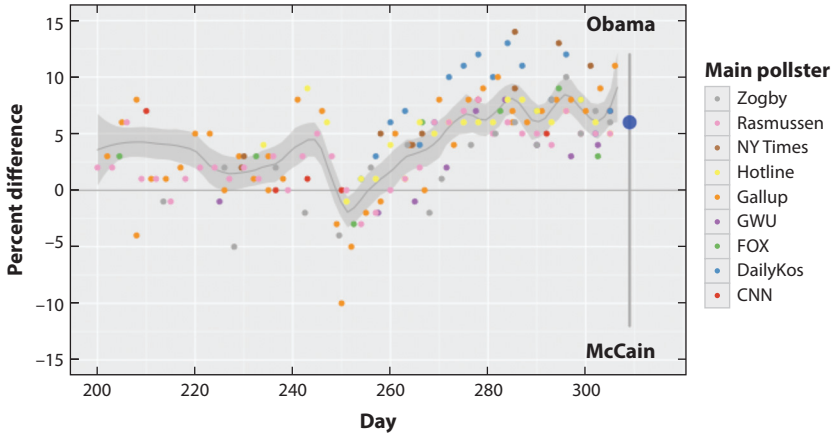


Figure 3

Tracking polls and final popular vote leading up to the 2008 US presidential election. Percentage difference is plotted versus the time of poll release. Colors represent the individual pollsters, with each dot representing one poll result, and the large blue dot indicates the final election margin. The gray line and ribbon represent a loess smoothing (Cleveland et al. 1992) across all of the poll results to indicate the trend.

the national trend polls allows us to see the variability and the biases of polling organizations. The pollster DailyKos, also known as Research 200, is a community action organization with political leanings toward the Democratic party, and it is one of the pollsters currently excluded by 538. Rasmussen, in contrast, has a reputation of leaning to the right and currently has a large adjustment value to correct this on 538. Hotline is fairly neutral.

The US election is not won on the popular vote, though. Each state is assigned a certain number of electoral votes that is roughly based on the size of the population. For example, in 2008, Iowa was worth 7 votes, whereas New York was worth 31. The electoral votes for most states are allocated on a winner-takes-all basis: Obama won New York with 57% in favor, so he received all 31 votes. A candidate needs to tally up 270 or more votes (of the possible 538, hence Nate Silver's website name) to win the presidency. **Figure 4** illustrates the state-by-state variation in polls—the state is plotted against the percentage difference between the two candidates. States in the 5% margin, whose results are considered too close to call, tend to have many pollsters operating, and for the most part the final result closely matches the poll results. There are a couple of exceptions: Montana was predicted to be a toss-up but ended up being more for McCain than expected, and Iowa ended up closer than the latest polls predicted. During the election cycle, Mosley et al. (2010) produced plots similar to **Figure 4** and animated the change from the previous week, which gave a sense of the temporal shifts in attitude and the variability leading up to the actual vote.

In the 2012 election season, this work was expanded to explore the effect of political action committee spending after the 2010 Supreme Court decision enabled unlimited election spending by organizations (Kaplan et al. 2010). The 538 site has now expanded to be an independent news site, and it hosts exemplary numerical and visual analysis of big data questions in the news.

2.2.3. Airline traffic. Every few years, the graphics and computing sections of the American Statistical Association provide a data challenge, the Data Expo (**http://stat-computing.org/dataexpo/**), in which they encourage students, faculty, and industry statisticians to make a visual analysis of a data set. In 2009, the data consisted of flight arrival and departure details for

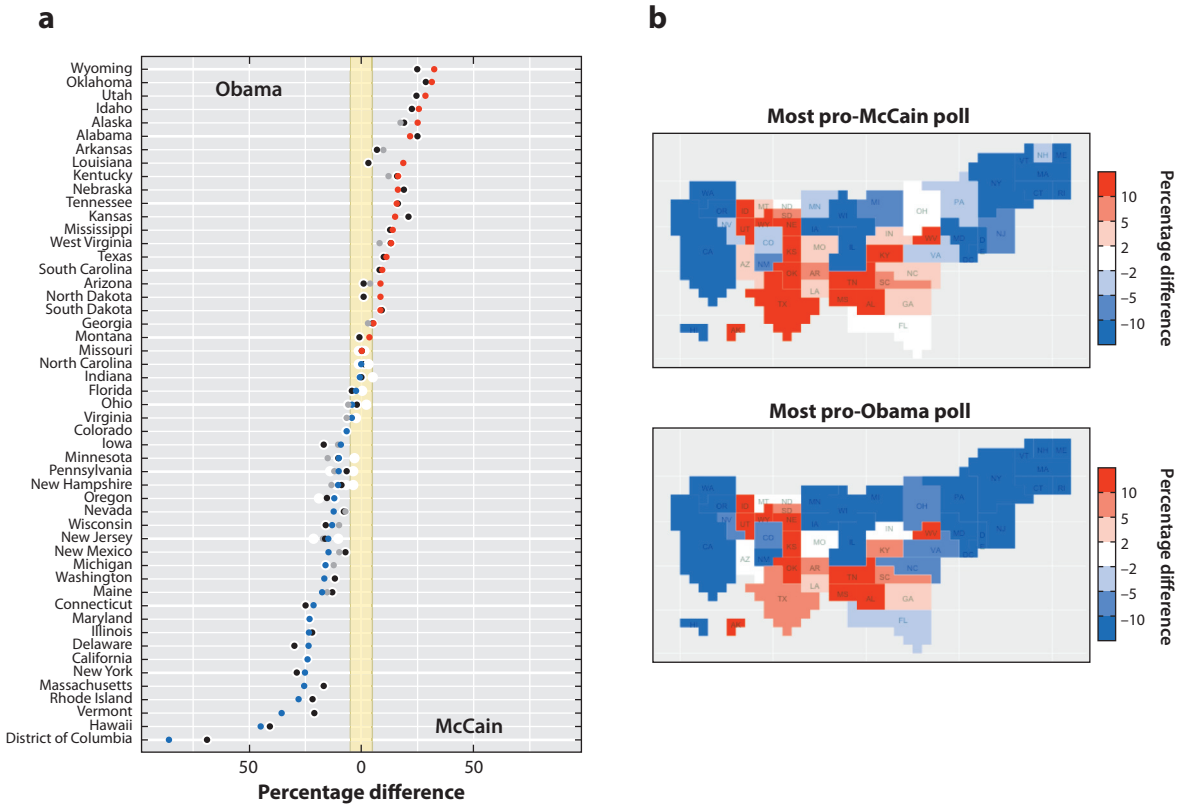


Figure 4

Exploring variability in polls. (*a*) Percentage difference by state, ordered from top to bottom by McCain to Obama advantage. Color (*red* or *blue*) indicates final election result—conventionally, in the United States, blue indicates Democratic and red means Republican. Black is the median of all polls, gray is the median of the previous week's polls, and white shows all of the polls considered. The yellow strip shows a 5% margin of uncertainty above and below 0, indicating polls in which there is no significant difference between candidates. (*b*) Block cartograms in which the size of the state represents electoral votes, colored by the most favorable poll result for each candidate just prior to election day. In the best-case scenario for McCain, the country still looks predominantly blue.

all commercial flights within the United States, from October 1987 to April 2008. This is a large data set: There are nearly 120 million records in total, and they take up 1.6 GB of space when compressed and 12 GB when uncompressed.

There were nine entries, four of which won prizes and are described in short articles. Wicklin (2011) used SAS to produce a number of informative displays about departure delays in US air travel: The calendar view summaries show differences between years, months, and days of the week; time series display volume of traffic and weekly cycles; and a heat map is used to compare carriers over the time period. A large amount of data was displayed very succinctly, providing key details of flight delays. Wickham (2011a) tackled a smaller task, comparing operations at two different airports, and Dey et al. (2011) focused on a model to find the path of least delay between any pair of airports.

Hofmann et al. (2011) explored many aspects of the data. Maps of origins and destinations show which carriers operate on a hub system and which do not. Time series of volume at major airports

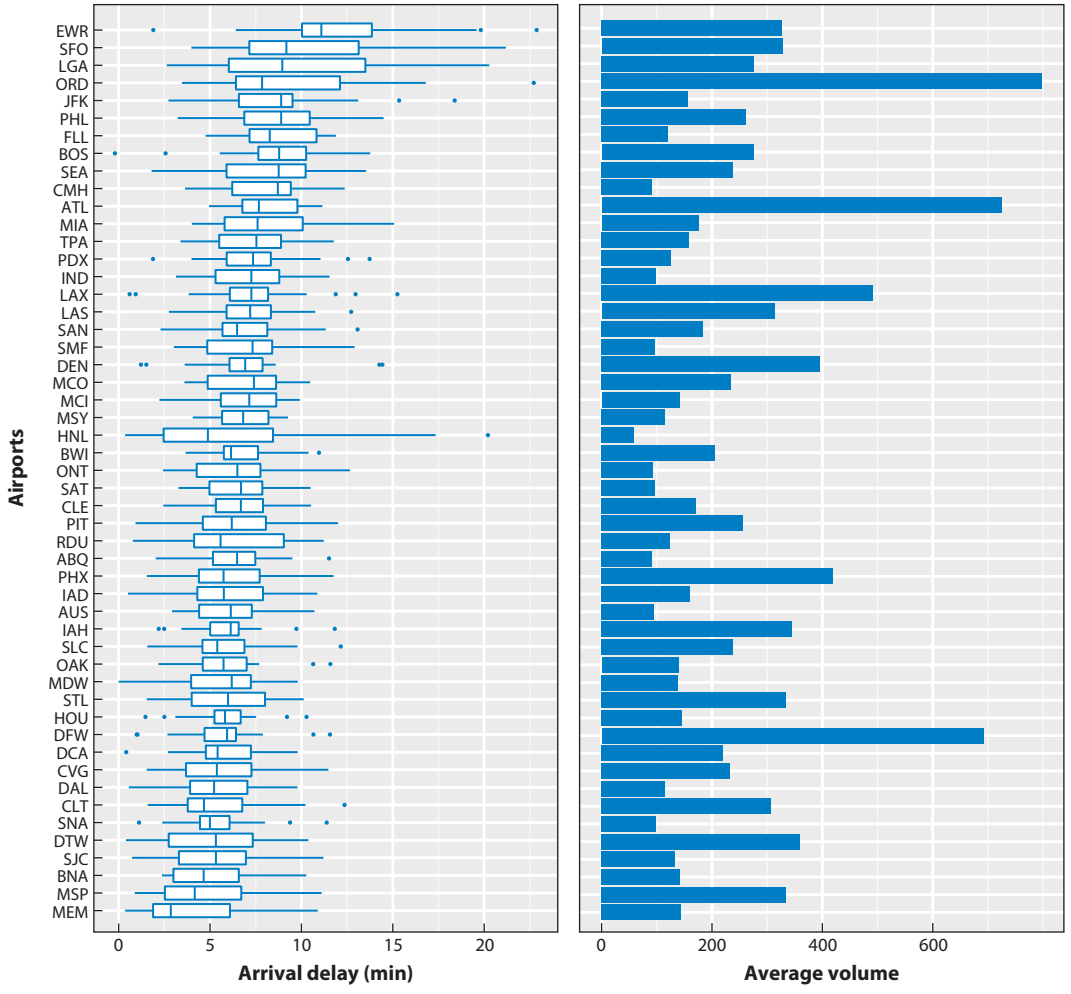


Figure 5

Arrival delays by airport, sorted from largest average delay (*top*) to smallest average delay (*bottom*) and volume of traffic. Only the top 50 airports based on delay are shown. Over this time period, EWR (Newark, NJ) had the worst record in delays even though their traffic volume was not large compared with other airports. ORD (Chicago, IL) had a bad delay record, but it also had the highest volume. DFW (Dallas/Fort Worth, TX) compared favorably, with low average delays and very high volume. Figure adapted with permission from Hofmann et al. (2011).

show effects of events such as the 9/11 tragedy. Side-by-side box plots display delays by airport, revealing the problematic airports (Newark, San Francisco, Chicago O'Hare, LaGuardia) and efficient operations (Detroit, Minneapolis, Dallas/Fort Worth) (**Figure 5**). Faceted scatterplots with overlaid loess fits show trends in delays by carrier. This group also looked for gaps in the data, in which planes are last seen at one airport and then magically appear at another airport. These gaps correspond to ghost flights—planes that fly without passengers to move a vehicle, which represent inefficiency in operations. Most carriers have been reducing this costly operation, but Northwest Airlines had an increase in the latter few years of these data. Delta, which merged with Northwest in 2008, saw reduced delay time after the merger. Interactive graphics were employed to

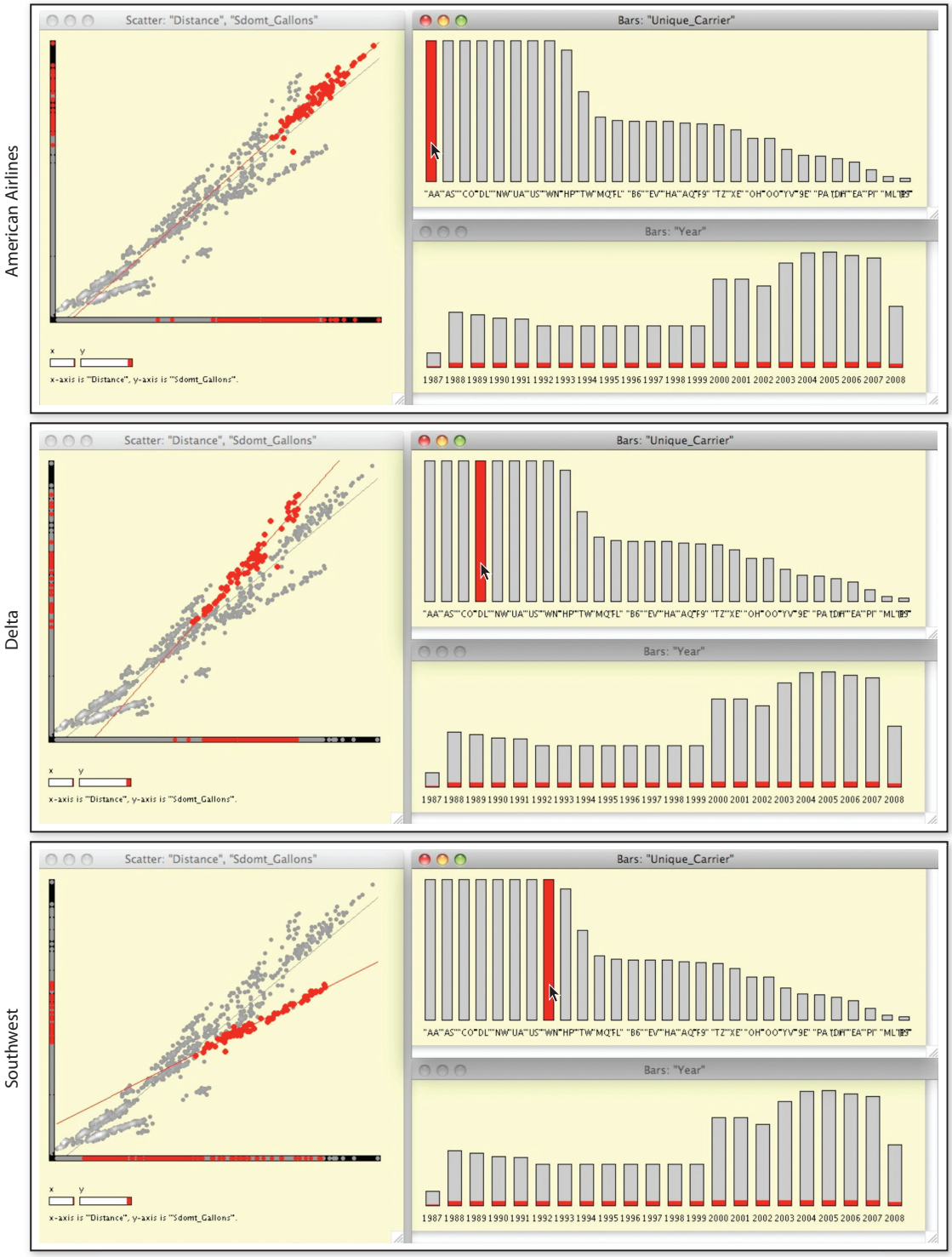
examine different chunks of the data, such as the relationship between fuel consumption, distance flown by carrier, and year (**Figure 6**). Basic plots revealed many problems with the data, including flights leaving 12 hours before they were scheduled (**Figure 7**), several hot air balloons that travel 430 miles per hour, and more than 200,000 flights of less than 50 miles. Correlating delays with weather patterns revealed the problems that strong crosswinds cause at airports, and interpolating between geographic locations of airports based on departure and arrival times allowed the animation of air traffic flow across the nation (**Video 1**).

2.2.4. Wikipedia. Wikipedia (**https://www.wikipedia.org/**) is a collaboratively written encyclopedia that is a huge resource for the general public. Because it is a primary example of mass editing, the flow of edits is potentially interesting. This problem was tackled by Wattenberg & Viégas (2010), with associated websites **http://fernandaviegas.com/wikipedia.html** and **http://www.bewitched.com/historyflow.html** that include some of the visualizations. The book chapter describes the process of pulling the data, preprocessing, and making visualizations of different aspects. Their endeavor began in 2003, which was in the early days of the encyclopedia. As an example of the magnitude of the data, the page on Microsoft had 198 edits generating 6.3 MB of text, whereas the page on Cat had just 54 edits.

The edits data set is huge. To tackle it, Wattenberg & Viégas (2010) initially created an interactive visualization for single pages. They use a modified parallel coordinate plot (Inselberg 1985, Wegman 1990), which might also be considered a variation of a hammock plot (Hofmann & Vendettuoli 2013b, Schonlau 2003), with versions as the variables and text size as the stripe thickness. Stripes are colored by author, so individual contributions to pages can be tracked. The overall height of the plot indicates the length of the article. They show two prominent examples, the pages for chocolate and abortion. The displays enable some information to be immediately apparent: If there is ownership of a page by a few editors, there are a few differently colored stripes. The page for abortion has a large empty patch indicating that the entire page was erased temporarily, probably by a malicious editor. A tug-of-war between editors can be seen in some politically or emotionally controversial subjects. In the same article, Wattenberg & Viégas (2010) provide interactive plots to trace editors' contributions and to see the actual text that was edited.

With big data, we see that there are many different choices in what to plot. Here, the authors chose to tackle the problem by using the page as a basic unit. In a secondary task, their basic unit is an editor, and a new display called a chromogram (Wattenberg et al. 2007) is employed to view an individual editor's contributions to Wikipedia. Both tasks would be considered drilling down into the data because each shows a very narrow slice of the data. There are now 4,714,447 pages, so visiting each one would take some time. Ideally, approaching a problem as large as this would also provide some larger visual overviews, for example, number of pages over time, number of authors over time, or how many pages different authors edit, and provide some comparative views, for example, pairs of pages, hierarchical topic lists, or how pages link to each other. The data implore the analyst to display them in many different ways. The website Wikipedia Visualizations (**http://infodisiac.com/Wikimedia/Visualizations/**) illustrates ways that many others have tackled visualizing Wikipedia.

Raw information in the form of text provides new challenges for visualization. People have universally adopted the use of tag clouds to display word frequency of blocks of text, e.g., Wordle (Feinberg 2010). But there is more to understanding patterns of text than showing frequency. Some good examples of processing text data and visualizing different facets of text are provided by Jockers (2014), who examines British literary history. New tools for grouping text into topic models using latent Dirichlet allocation have been developed, and the R (R Core Team 2014) package `LDAvis` (Sievert & Shirley 2014) provides ways to interactively visualize the data.



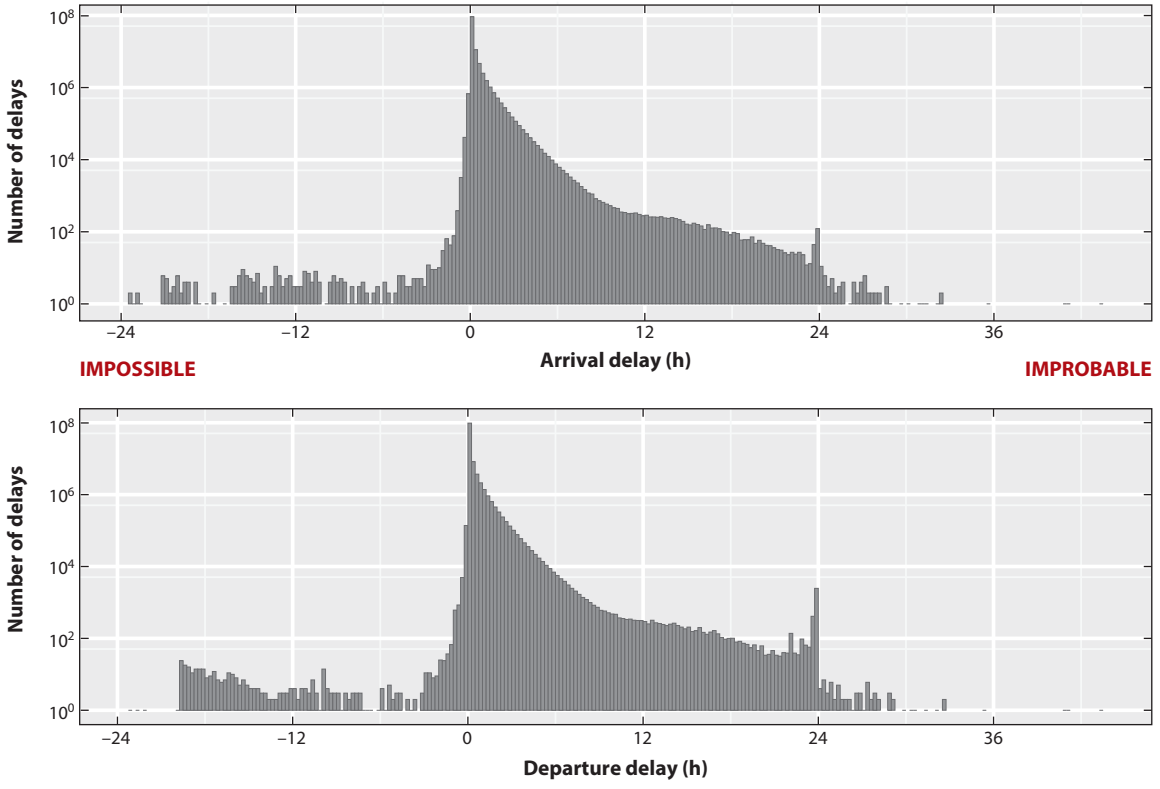


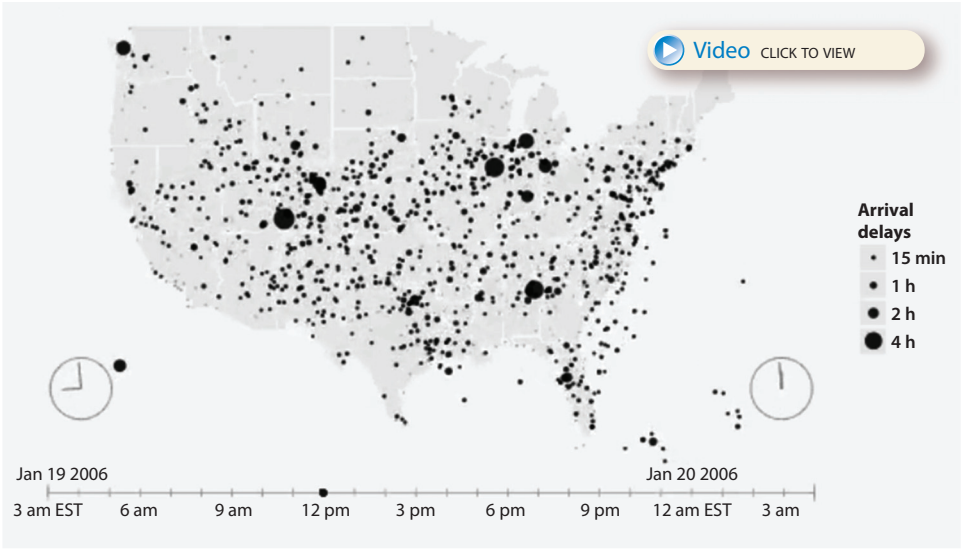
Figure 7

Histograms showing arrival and departure delays across carriers, airports, and years. Delays are aggregated in 15-minute bin widths. Arrival delay is calculated from the difference between the reported and scheduled departures. An arrival delay of −24 hours would mean that a plane left a day early, which we would expect is impossible. Figure adapted with permission from Hofmann et al. (2011).

2.2.5. The San Francisco housing crisis. The Wikipedia example is published in a book called *Beautiful Visualization* (Steele & Iliinsky 2010). Another book in the same series is *Beautiful Data*, in which a chapter by Wickham et al. (2009) presents a visual exploration of the housing crisis using an early version of the R software `ggplot2` (Wickham & Chang 2014). This package is now very widely used for plotting data. This chapter represents an extraordinary example of pulling data from the web, cleaning it, and displaying it in different ways to learn about events that affect our lives. The explanation of the process is superb.

Figure 6

Three snapshots of interactive graphics using the software MANET on a processed chunk of data to explore carrier efficiency. Fuel consumption (vertical axis) is plotted against distance flown (horizontal axis) in the scatterplot, and bar charts show carrier and year, with bar height indicating relative number of flights. Three airlines are highlighted (*red*) from top to bottom: American Airlines (AA), Delta (DL), and Southwest (WN). American Airlines is at the top of the pack as a big carrier and a big consumer. Over this period of time, Delta is relatively inefficient, with relatively higher fuel consumption for the same distance flown. In recent years (not shown in these data), it has improved substantially. Southwest is a major carrier but has substantially more efficient fuel consumption than its competitors. Figure adapted with permission from Hofmann et al. (2011).



Video 1

Still image of video showing plane movements across the United States on a normal day of operations, January 19, 2006. The video includes red-eye planes leaving the West Coast for the East Coast, the East Coast waking up, and sporadic delayed flights. The code (including links to the data) is available at **https://github.com/heike/usflights**. An accompanying video at **https://vimeo.com/119233996** shows operations during a northeastern snow day, March 13, 1993. Used with the permission of Heike Hofmann. To view the video, access this article on the Annual Reviews website at **http://www.annualreviews.org**.

The story of the housing crisis in this chapter begins by studying the temporal scale, namely average prices and number of sales from 2003 to 2009. We can see the average house price double over this period and then drop by half starting in the summer of 2007. Sales tend to cycle some, showing some seasonality, but a clear decline starts in 2006. Interestingly, sales tick up again early in 2008 following sharp declines in the average price. The authors then examine economic conditions by comparing inflation-adjusted and unadjusted average prices to learn that these two measures started diverging mid-2005. The two had not converged by 2009.

Breaking the data into house price deciles and examining these relative to the median house value shows that the disparity in house prices is expanding, supporting a perspective that the more expensive houses are becoming relatively more expensive. Drilling down into each geographic region shows that there are some parts of the San Francisco Bay Area that are more affected than others: San Pablo experienced the full brunt of the boom and bust but Berkeley saw barely a hint of decline. Plotting geographically reveals that the eastern part of the region experienced more turnover in houses. Comparing price decline and demographic factors revealed that higher income areas and areas with a higher percentage of college graduates saw less decline, whereas areas where residents had longer commutes saw bigger house price declines. This chapter illustrates elegantly how visualization can be used to explore data.

2.2.6. Credit card purchases. Hand et al. (2000) is an early article explaining data mining, using a large Barclay's credit card transaction database as the example. This article shows how the humble histogram can be priceless for exploring big data. The histogram, with a carefully crafted

£1 bin width, is used to display petrol (gas) purchases and department store spending. We might expect that in department stores, there would be strong peaks just before the whole pound, and it is exactly what we see. But to see similar patterns in petrol purchases is a surprise. There are large peaks in petrol purchases at £10, £15, and £20, and to a lesser extent £12, £25, and £30. This behavior is driven by the consumer rather than the price points of store products—clearly, many drivers like to spend whole-pound amounts when purchasing petrol. Apart from these peaks, the distribution looks close to bell-shaped, centered at £20.

Working with huge amounts of data can often be done with basic statistical graphics. There are a few cautions. Bars showing a small number of counts get lost easily with big data, which might result in failing to observe rare events. Basic scatterplots may suffer from overplotting. Scaling up of basic statistical graphics does require some care.

3. LITERATURE REVIEW

There is a scarcity of papers on big data visualization in the most likely candidate statistical journals, e.g., the *Journal of Computational and Graphics Statistics*, *Computational Statistics*, and *Annals of Applied Statistics*. Graphics papers in these journals tend to describe methods useful only with relatively small amounts of data or for special statistics-related purposes rather than providing solutions to visualizing large amounts of data. The exceptions are the special papers in *Computational Statistics* and the *Journal of Computational and Graphical Statistics* written to describe approaches taken to visualizing data from the Graphics and Computing Sections Data Expo held every few years at the Joint Statistics Meetings. Dey et al. (2011), Hofmann et al. (2011), Wickham (2011a), and Wicklin (2011), discussed in Section 2.2.3, are examples. The most reliable source of big data visualization examples is *IEEE Transactions on Visualization and Computer Graphics*. The papers presented at the annual InfoVis conference are published in this journal in one of the last two issues of each year. A more visible location to find the latest research from the statistics community is directly on the CRAN archive (**http://www.r-project.org/CRAN**), and occasionally articles describing the methodology behind these packages can be found in the *Journal of Statistical Software* or the *R Journal*. The R packages appearing in this section are listed in **Table 1**.

3.1. Reconditioning Old Favorites

Many of the conventional and useful methods of plotting data need a little renovation for working with big data, which is addressed by some recent papers.

3.1.1. Scatterplots. One of the most useful methods for viewing multivariate data, the scatterplot matrix (Hartigan 1975), is less useful when there are more than a handful of variables. Even plotting pairs of variables separately, using a loop to animate the display of all pairs, is infeasible when the number of variables is sufficiently large. Wilkinson et al. (2005) resurrected Tukey's ideas on scagnostics (Tukey & Tukey 1985) to provide automated ways to extract pairs of variables that might be the most interesting to plot. Their approach is to calculate nine measures for interesting features in a scatterplot—outlying, skewed, clumpy, sparse, striated, convex, skinny, stringy, and monotonic—based on the relationship between the variables. The methods are implemented in an R package, `scagnostics` (Wilkinson et al. 2012), and a stand-alone piece of software, `ScagExplorer` (Dang & Wilkinson 2014).

When the number of cases is large but the number of variables is small, reading distributions from a scatterplot can be nebulous because points will be overplotted. Carr et al. (1987) approached this with a modification employing density plots, for which the recent `hexbin` (Carr et al. 2014)

Table 1 R software packages referenced in Section 3 with current version publication year and brief description

Package name	Year	Description
animint	2015	Tool for producing animated, interactive web graphics
bigrquery	2015	Interface to Google's BigQuery API
bigvis	2013	Exploratory data analysis for large data sets
broom	2015	Converts statistical analysis objects from R into tidy data frames
copula	2015	Calculates multivariate dependence with copulas
cranvas	2014	Interactive statistical graphics with R and Qt
dplyr	2015	Grammar of data manipulation for working with data frame–like objects
epivizr	2014	R interface for interactive visualization of genomic data
GGally	2014	Extension to ggplot2 with additional plot templates
ggmap	2015	Spatial visualization with ggplot2 to visualize data on top of maps
ggpairs	2014	Component of GGally to make a matrix of plots with a given data set
ggscatmat	2014	Component of GGally to make a scatterplot matrix
gpairs	2014	Produces a generalized pairs (gpairs) plot
ggparallel	2015	Parallel coordinate plots for categorical data with ggplot2
ggplot2	2015	An implementation of the grammar of graphics—data plotting
ggvis	2015	Interactive grammar of graphics based on ggplot2
gridSVG	2015	Exports graphics drawn with package grid to interactive SVG format
hexbin	2014	Binning and plotting functions for hexagonal bins
htmlWidgets	2015	Provides web graphics using JavaScript
iotools	2015	Runs distributed R jobs on Hadoop and handles chunk-wise data processing
iplots	2013	Produces interactive graphics using Java
lubridate	2013	Provides functions to simplify performing math with dates and times
nullabor	2014	Tools for graphical/visual inference to examine patterns in data
PairViz	2011	Visualization using Eulerian tours and Hamiltonian decompositions
plyr	2015	Tools for splitting, applying, and combining data
rainbow	2015	Rainbow plots, bagplots, and box plots for functional data display
rbokeh	2015	R interface to the Python library bokeh
RCloud	2015	Web-based platfrom for analytics, visualization, and collaboration using R
readr	2015	Reads tabular data
reshape2	2015	Flexibly reshapes data
rggobi	2014	Interface between R and GGobi, an interactive and dynamic graphics package
tabplot	2014	Visualization of large data sets
tabplotd3	2013	Interactive inspection of large data sets using JavaScript graphics library
tidyr	2015	Data tidying
scagnostics	2012	Calculates graph theoretic scagnostics–scatterplot diagnostics
spatstat	2015	Spatial point pattern analysis, model fitting, simulation, and tests
YaleToolkit	2014	Tools for the graphical exploration of complex multivariate data

package can be used. Another alternative is to utilize the transparency capabilities of today's graphics hardware to roughly produce density displays by layering virtual ink.

Another issue arises with big data: The many variables may be of different types, such as categorical or temporal, in addition to numeric. The work described by Emerson et al. (2013) and Friendly (2014) on the generalized pairs plots, along with the accompanying R packages (gpairs in YaleToolkit and ggpairs or ggscatmat in GGally), adapts the scatterplot matrix ideas for heterogeneous variable types.

Statisticians frequently use scatterplots for examining association between pairs of variables, despite the existence of many other graphical forms, which earns some derision from the InfoVis

community. But scatterplots are the bread-and-butter method to examine joint distributions, something of fundamental importance to statistical thinking, so they are very important for the statistics community. These three additions (scagnostics for large p, binning and transparency for large n, and mixed types of plots for mixed variable types in the scatterplot matrix) adapt the method to the big data of today.

3.1.2. Table plots. Table plots are adapted from side-by-side histograms of different variables. We most often see side-by-side plots used to compare the distributions of subsets of the same data, for example, comparing males and females. Generally, side-by-side box plots are the optimal way to make comparisons between groups. Histograms can also serve this purpose, but they provide more complex summaries of the distribution than a box plot renders. Table plots came to prominence with the `Tableau` software (Stolte et al. 2003), and a similar style of plot was developed by Carr (1995) (also shown in Carr & Nusser 1996) as a way to replace tables with graphs.

The table plot bins the values of different variables and displays the mean value of each bin as a bar. For categorical variables, stacked bars are displayed. Each plot is sorted in the same way, either according to one of the variables in the collection or by another external criterion. The sorting enables a rough assessment of the association between variables—if the two plots have the same shape, then the two variables have positive association. **Figure 8** shows an example produced using the recently released R package `tabplot` (Tennekes & De Jonge 2014). It displays the sales records for houses sold in Ames, Iowa, from 2008 to 2010. The variables are sorted by the sale price; the top of the plot represents the most expensive 1% of houses, which have a mean value of approximately $500,000, and the bottom represents the cheapest 1%, with a mean value of approximately $50,000. There are 1,615 houses in the data set, and these are grouped into 100 bins of equal size by the sale price. The mean values of the other variables for houses in each of these bins are shown in the other plots. The exception is the house style, which is categorical, so proportions of the different styles are shown in stacked bar charts. We learn some rough associations from this display:

- Sale price is not closely associated with the number of bedrooms, which is a little surprising. The pattern in the plot of bedrooms is fairly uniform and evenly distributed across all house prices, which leads to the conclusion that there is no association. Conversely, there is a slight association with the number of bathrooms because the average number of bathrooms drops from two to one for lower-priced homes.
- Price is more closely associated with living area and garage area because fairly strong declines in both coincide with declines in price.
- House style shows a slight association with price: The higher-priced homes are more commonly two-story style than the lower-priced homes.

Because the display uses a bar to represent mean values, there is implicit distortion of the data, a trait that is discouraged by Tufte (1983). From a statistical perspective, we know that means or averages may not satisfactorily represent a set of numbers. Additionally, a mean is a point estimate and is ideally represented by a point on a plot with another graphical element, such as a line corresponding to standard error, displaying the variability associated with the estimate. Hence, the table plot is a gross reduction and distortion of a large data set—it may be reasonable to use to get a rough sense of the associations, but it is ripe for a redesign to reduce data distortions, possibly by using dot displays (Cleveland 1993).

An interactive version of table plots is available with the R package `tabplotd3` (De Jonge & Tennekes 2013). It uses the `d3` software (Bostock et al. 2011), which is a JavaScript graphics library useful for creating interactive web visualizations.

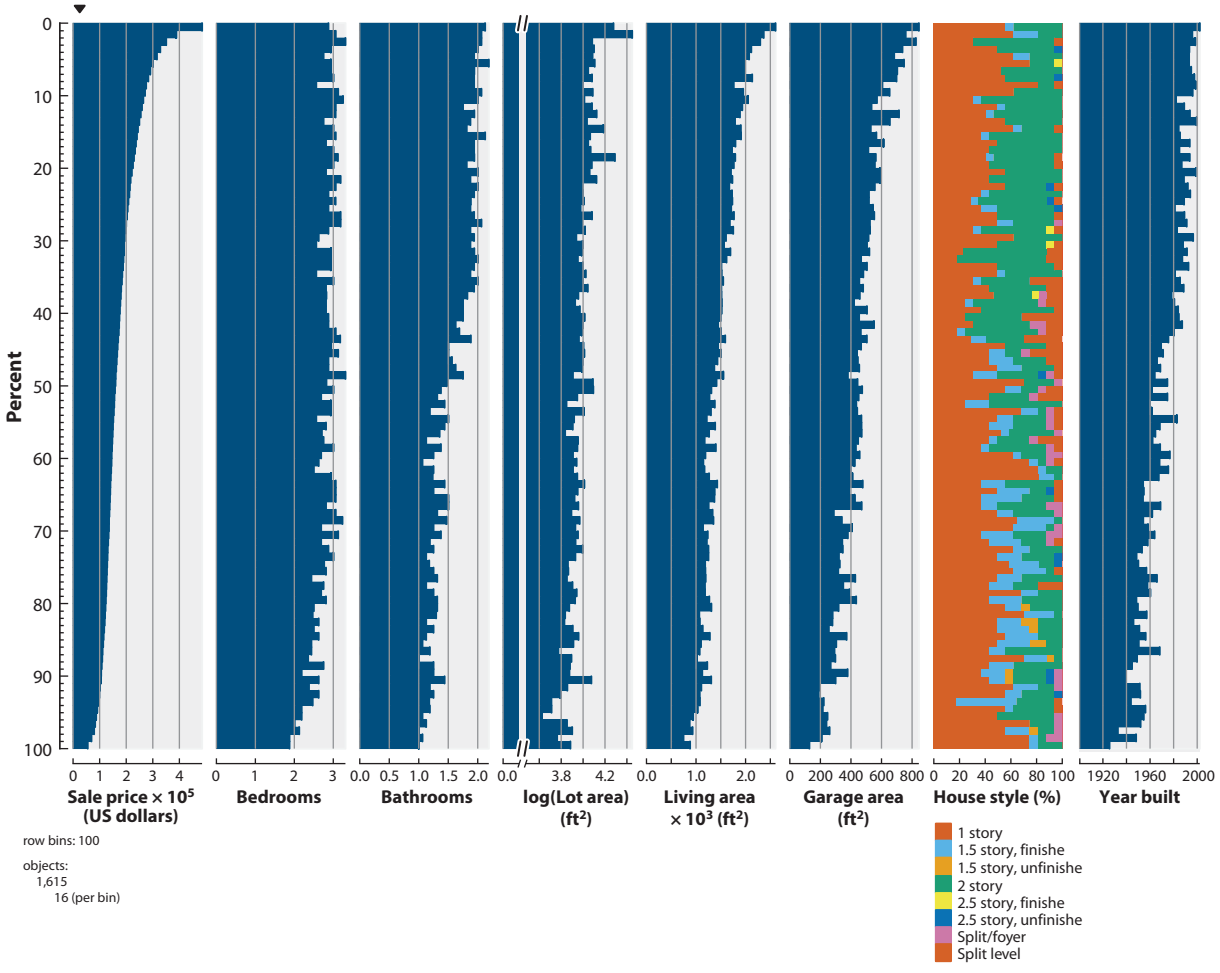


Figure 8

Table plot of Ames, Iowa, housing sales produced using the R package `tabplot`. Similar shapes (e.g., sale price and living area) indicate variables that have a positive association. Houses are binned by sale price into 100 bins, e.g., 0–1% represent the top sale prices and have an average value of almost $500,000.

3.1.3. Parallel coordinates. Another way to display high-dimensional data is the parallel coordinate plot (Inselberg 1985, Wegman 1990). Several developments of these plots have been made in recent years. Hurley & Oldford (2011a) use graph theory to organize the axes of a parallel coordinate plot, and the work has extended to provide an algorithm for navigating high-dimensional spaces. Their R package, `PairViz` (Hurley & Oldford 2011b), implements the parallel coordinate plot style visualization. **Figure 9** shows examples for data on the nutritional value of chocolates in both traditional and Eulerian parallel coordinate plots. In the Eulerian adaptation, variables are repeated so that all pairs can be examined. In both panels, the histogram displays a scagnostic value (1− Wilks's Λ MANOVA statistic), which indicates how well the variables distinguish between the two groups: the higher the bar, the more separated the variables. Hofmann & Vendettuoli (2013a) describe variations of parallel coordinate plots for categorical data, and their R package

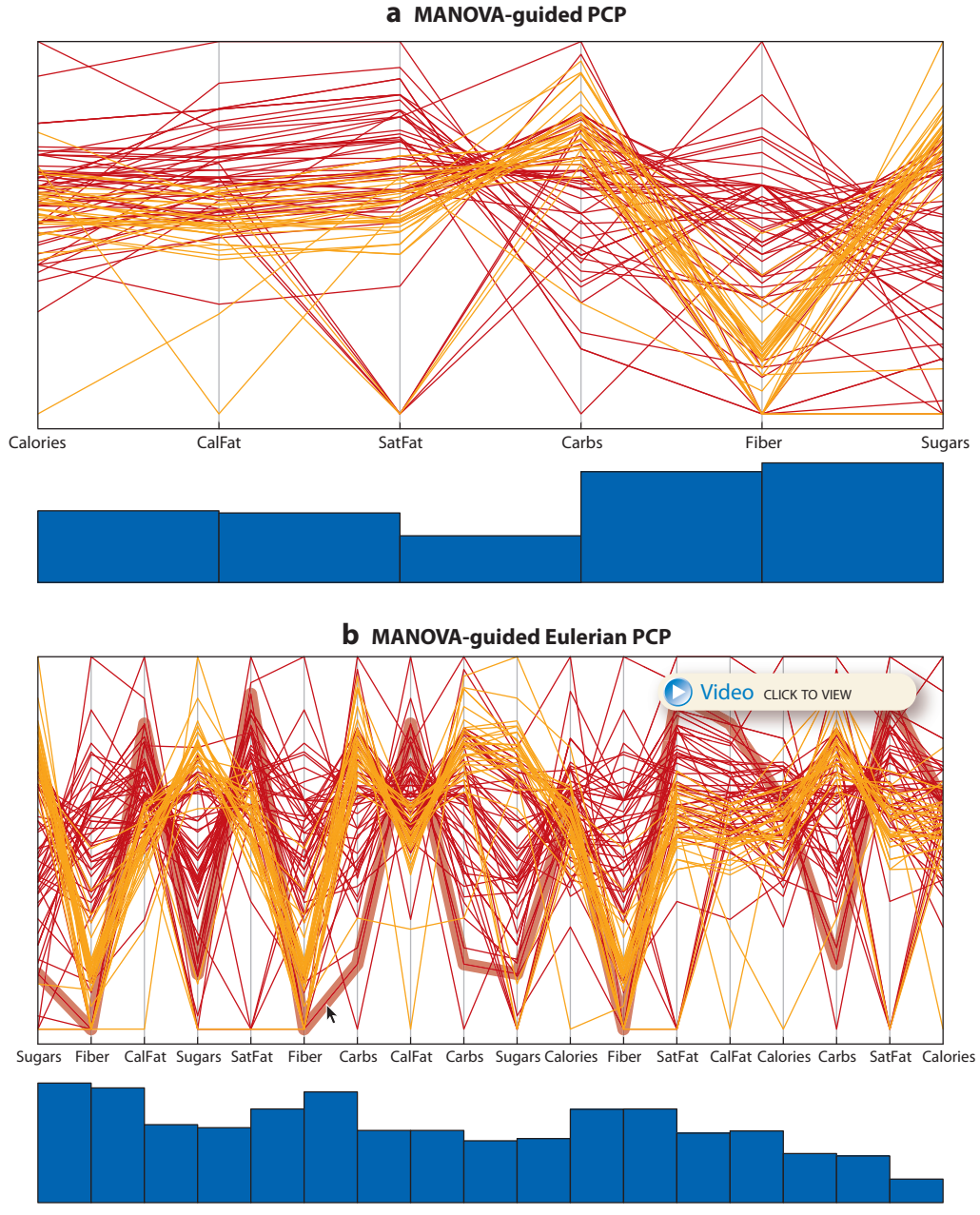


Figure 9

(*a*) Classical and (*b*) Eulerian parallel coordinate plots (PCPs) of nutritional measurements of chocolates. The Eulerian PCP more strongly indicates a difference between the two types of chocolate, milk (*orange*) and dark (*red*), and from the scagnostic describing separation displayed by the histogram (*blue*), we learn that higher values occur when fiber is one of the two variables, indicating its importance as a variable. A simple interaction added to this plot using `gridSVG` highlights the line closest to the mouse cursor so that we can follow it throughout the plot. To view the video, access this article on the Annual Reviews website at **http://www.annualreviews.org**. SVG graphic files are also available for download: Follow the **Supplemental Material link** from the Annual Reviews website.

Supplemental Material

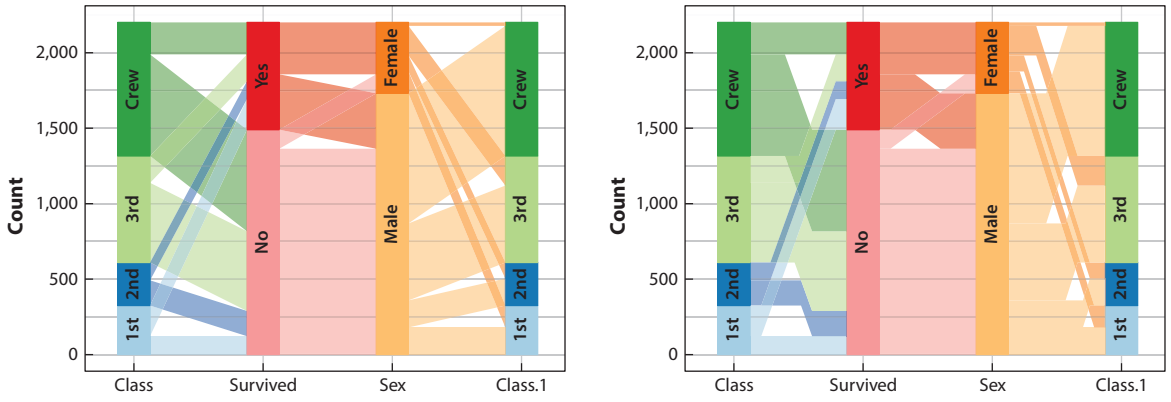


Figure 10

Variations on parallel coordinates for categorical variables. Ribbons peeling horizontally from the bars (*right*) allow for more accurate reading of the proportions in each category.

`ggparallel` (Hofmann & Vendettuoli 2013b) enables others to use the methods (an example is shown in **Figure 10**). These plots are similar to how Wattenberg & Viégas (2010) visually explored the Wikipedia data (Section 2.2.4). Moustafa et al. (2011) provide envelope methods for parallel coordinate plots for large data.

3.1.4. Rearranging, summarizing, and plotting data elegantly. When data sets become very large, an alternative approach can be to make summaries and display these. The `ggplot2` package has revolutionized the display of data for many analysts, and it comes with siblings in the Wickham suite of software that can help process data for viewing in different ways: `tidyr`, `reshape2`, `dplyr`, `lubridate`, `broom`, `readr`, `bigrquery`, `ggmap`, and `bigvis`.

The `bigvis` package (Wickham 2013) summarizes large amounts of data using aggregation and smoothing techniques, and from these summaries, users can make various plots with `ggplot2`. Behind the `plyr` (Wickham 2011b) and `dplyr` (Wickham et al. 2015b) packages is the split-apply-combine strategy, meaning that the data set is divided into chunks, a function is applied, and the results are joined. For example, with climate records from multiple locations, we might want to examine linear models at each location. This is easy to do using split-apply-combine. The `dplyr` package extends the approach with a grammar and off-loads some computations to database operations.

The split-apply-combine strategy resembles the approach that Hadoop (Shvachko et al. 2010) clusters employ for storing humongous amounts of data. Chunks of data are placed in different locations and indexed using keys. Accessing these data requires operating on individual chunks and combining the results. The `Tessera` project (Hafen & Cleveland 2015) provides an R interface to Hadoop-style distributed file systems, and an early project paper (Guha et al. 2009) describes its use for visualizing large data sets. Another R package, `iotools` (Urbanek 2015a), is available for pulling data from Hadoop storage units.

3.1.5. Interactive graphics. The area of interactive graphics is still very much a work in progress despite existing as a field of research since the late 1960s. But there are some exciting developments, driven in part by the availability of new technology. There are many additional software tools (e.g., `d3`, Bostock et al. 2011) available for interactive graphics, but close connections between these and statistical modeling of the sort achieved by Tierney (1991) is lagging. Some early explorations

with interfaces in R were seen in `iplots` (Urbanek & Theus 2003), which interfaced R with Java, and `rggobi` (Wickham et al. 2008), which interfaced R with C and `gtk` (**http://www.gtk.org**). Recent developments include `gridSVG` (Murrell & Potter 2015), `htmlWidgets` (Vaidyanathan et al. 2015), `animint` (Hocking et al. 2015), `cranvas` (Xie et al. 2014), `ggvis` (Chang & Wickham 2015), and `shiny` (RStudio 2015). For Python programmers, the library `bokeh` (Avila et al. 2015) provides interactive web graphics using JavaScript, and `rbokeh` (Hafen 2015) provides an R interface.

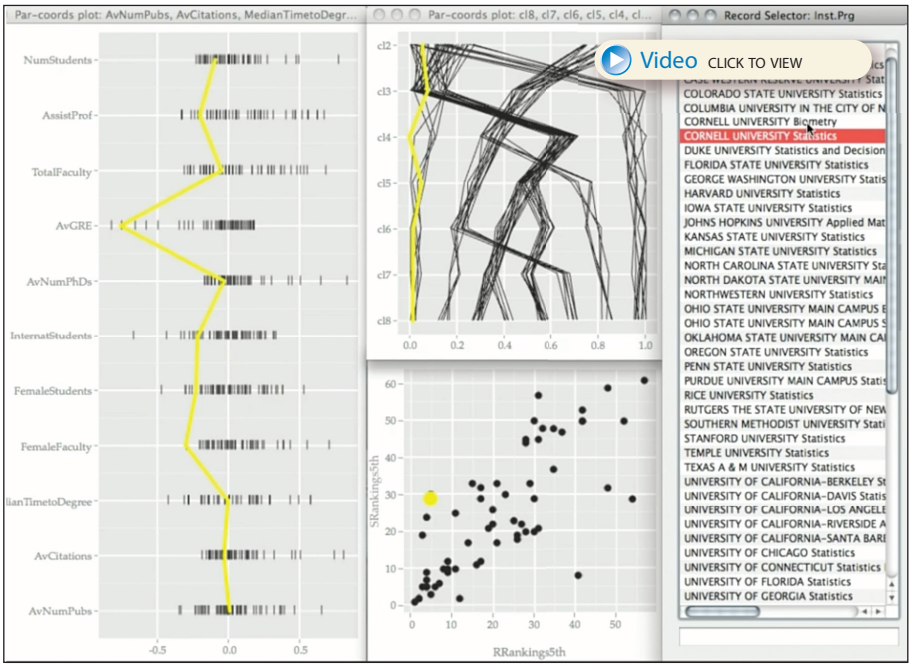
Both `gridSVG` and `animint` are relatively lightweight interactive graphics, designed to enhance existing static graphics and add several interactive elements. The `gridSVG` and `animint` packages would be considered primarily for communication purposes, where the information in the data is known and simply needs to be presented. The interactive graphics provide a level of sophistication and allow the users to change a little, but not much, of the information they are viewing. A simple example of the usefulness is the parallel coordinate plot shown in **Figure 9**. With a static plot, it is hard to trace a line through all of the variables. A simple interaction to add to this plot would be to highlight the line closest to the mouse cursor so that we can follow it throughout the plot. A video at **https://vimeo.com/137049134** illustrates how this is done using `gridSVG`. `Animint` works similarly, taking `ggplot2` plots and adding interaction using JavaScript.

`Cranvas`, and potentially `htmlWidgets` and `ggvis`, provides more substantial tools for exploring data generally with interactive graphics. `Cranvas` generates interactive graphics inside R by using wrappers to the `qt` libraries (**https://www.qt.io**) and theoretically can be used for very large data sets. **Video 2** shows `cranvas` being used to explore a clustering of statistics department graduate programs based on the 2012 National Research Council ranking data. `HtmlWidgets` and `ggvis` provide web graphics in R using JavaScript, and `shiny` provides graphical user interface elements for the web. `Shiny` has been rapidly adopted by the general community, and a showcase of user apps can be viewed at **https://www.rstudio.com/products/shiny/shiny-user-showcase/**. Although no formal publication is available, it may be worth keeping an eye on `RCloud` (Urbanek 2015b), which provides interactive web graphics for massive data.

With the increasing availability of electronic publications, it is also possible to publish interactive graphics as part of a regular journal article. Newell et al. (2013) include supplementary material with videos of tours for viewing the cluster structure of data. Wickham et al. (2015a) include links to videos to illustrate several concepts in advocating for better use of visualization to understand statistical models, particularly for viewing the model in the data space.

3.2. Statistics, Information Visualization, and Biological Visualization

Within the statistical literature, there has been work on providing visual model diagnostics, a new plot type for visualizing clustering of multidimensional data and stronger connections between EDA using graphics and inferential statistics. Hofert & Mächler (2014) provide a graphical goodness-of-fit test for dependence models in higher dimensions, and they implement it in an R package, `copula`. Baddeley et al. (2013) provide new graphical residual diagnostics for covariate effects in spatial point processes, which help assess and improve the fit of these complex models. They extend the R package `spatstat` to incorporate these diagnostics. Rainbow plots and bagplots (the `rainbow` package, Hyndman & Shang 2010) were developed for viewing a large amount of functional data, smooth curves, or surfaces. Van Long & Linsen (2011) propose the visualization method for a hierarchical tree of high-density clusters in high-dimensional data. They project the multidimensional clusters to a two-dimensional or three-dimensional layout using an optimized star coordinates layout. It allows the user to explore the distribution of clusters interactively and helps the user to understand the relationship between the clusters and the original data space.



Video 2

Use of interactive graphics to explore rankings of statistics departments in the United States. The plots show rating variables as side-by-side dotplots (*left*), a cluster analysis (*center*, *top*), and a scatterplot of 5th-percentile rank computed using the S (vertical) and R (horizontal) methods (*center*, *bottom*), and institution name lookup (*right*). Selecting an institution highlights (*yellow*) its values in each of the other plots. Cornell University is highlighted: We can see that its rank by the two methods differs substantially, with a good R rank (around 5) but not such a good S rank (around 30). On the ranking criteria, the department is around the middle of the pack: It is average in terms of number of publications and citations, has few women faculty and students, and accepts students with lower GRE scores than most statistics departments. To view the video, access this article on the Annual Reviews website at **http://www.annualreviews.org**.

Newell et al. (2013) describe methods for finding clusters in populations based on genetic markers, which are very sparse, high-dimensional data, and using the tour (Asimov 1985) to visualize the cluster structure. Buja et al. (2009) and Majumder et al. (2013) detail new protocols for data visualization that would enable statistical inference to be conducted to quantify the significance of structure seen in plots. This is a key development for working with big data sets because we typically do not have a classical inference environment. It is easy to imagine patterns in data, and these protocols provide a way to determine whether what we see is really there. For a statistician, strict adherence to the rigid assumptions required by classical hypothesis testing runs the risk of nondiscovery, failure to see something that is present in the data. This new work makes it possible for statisticians to be both explorers and skeptics. The protocols are implemented in the R package `nullabor` (Wickham et al. 2014). And Hofmann et al. (2012) bridge statistics and information visualization, providing a formal protocol for determining if one type of display better communicates information than another.

From the information visualization community, there are a couple of interesting new approaches to working with large data sets. `OnSet` (Sadana et al. 2014) is a technique for visualizing

large-scale data by representing it as presence/absence displays. It has a web interface that allows data to be uploaded for plotting. Data can be compared using Boolean operations. Lins et al. (2013) make it easy to slice and dice large spatiotemporal data sets for viewing in various ways, with many possible applications, including climate change, housing trends, telecommunications, security, and crime.

The biological community has been grappling with humongous data sets for many years. There is still considerable work to be done to provide better visualization for these volumes of data, but there are some exciting recent developments. The software `epivizr` (Chelaru et al. 2014) provides interactive web graphics produced with JavaScript and is closely linked with analysis tools available in R from the `Bioconductor` suite (**https://www.bioconductor.org/**).

4. SYNOPSIS

Data visualization and statistical graphics are pursuits that sit on the edge of several disciplines. This can be unnerving when trying to organize thoughts and ideas, provoking uncertainty as to what box the work should reside in. In a featured article in the *Journal of Computational and Graphical Statistics* with invited commentary, Gelman & Unwin (2013) wrestle with the respective roles and purposes of statistical graphics and information visualization. These are just two of the terms that describe these overlapping topics. It would be useful for scientists to have a taxonomy of nomenclature related to data visualization. Gelman & Unwin (2013) fall short of a full discussion on the relative emphases of different pursuits and simplistically reduce the two domains to a dichotomy of appearance versus functionality. The invited commentaries do a nice job of defusing the indignation provoked by the shallow interpretation.

There are many different names associated with data visualization: statistical graphics, information visualization, visual analytics, and infographics. (Broadening the list to incorporate pursuits in scientific visualization could obfuscate distinctions further.) Any taxonomy of the nomenclature is imperfect: The roles, purposes, and functionality overlap, and the borders are porous. However, the existence of different terms indicates that there are useful, if not important, distinctions between them. Statistical graphics do have a primary focus on the visualization of data associated with understanding variability. They can be elegant, interactive, and beautifully crafted. Much of the time, though, they are ephemeral, designed to support the activity of exploring data. Information visualization is a vast area of endeavor focused on representing data abstractly to reinforce human cognition. There is a stronger association with cognitive perception. Its development also was driven by the database community, which has an inclination to handle large amounts of data. Infographics are designed for mass consumption and typically utilize very simplistic data representations. Visual analytics arose from a need to support decision systems and provide dashboards for company executives that can assist in making data-driven decisions. This is a convenient way to think about the different pursuits in data visualization research, which in practice have many activities in common.

With big data, one thing is clear: Data sets used to be a means of passive support of predetermined conjectures, but they have now taken an active position even in the absence of a well-defined hypothesis. Data visualization is likely to continue to provoke alarms about fishing expeditions and data snooping. Allaying these fears will require developing an understanding of the associated concerns along with an effort to broadly build knowledge about randomness among users and developers of data graphics. The idea that making 100 plots might result in a few that exhibit interesting patterns by chance (i.e., a fishing exhibition) is an important concept. Prior to making a plot of data, various patterns may have some probability of being observed by chance, but after the plot is made, a pattern either is seen or is not (i.e., data snooping).

Similarly, before flipping a coin, there is randomness to the possibility of observing a head, but after the coin has been flipped, either a head showed or it did not. We cannot revert to pretending that there was a probability of observing the pattern. Once we have made a plot, we cannot put the pattern back in the bag and draw again to learn about the probability of it occurring. Data collection practices also impact data visualization, and the words Tukey wrote many years ago are germane for big data analysis: "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data" (Tukey 1986, pp. 74–74). These issues will need to be faced for successful big data visualization.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank the Editorial Committee and Illustration Editor of the *Annual Review of Statistics and Its Application* for their help in preparing this article.

LITERATURE CITED

Asimov D. 1985. The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.* 6:128–43

Avila D, Cottam J, Dodia K, Doig C, Paprocki M, et al. 2015. Bokeh: Python interactive visualization library. *Data Visualization Software.* **http://bokeh.pydata.org/en/latest/index.html**

Baddeley A, Chang YM, Song Y, Turner R. 2013. Residual diagnostics for covariate effects in spatial point process models. *J. Comput. Graph. Stat.* 22:886–905

Bostock M, Ogievetsky V, Heer J. 2011. D3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.* 17:2301–9

Brierley P. 2011. What's going on here. *Another Data Mining Blog*, Dec. 17. **http://www.anotherdataminingblog.blogspot.co.uk/2011/12/whats-going-on-here.html**

Buja A, Cook D, Hofmann H, Lawrence M, Lee E, et al. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. R. Soc. A* 367:4361–83

Carr DB. 1995. Using gray in plots. *Stat. Comput. Stat. Graph. Newsl.* 5:11–14

Carr DB, Lewin-Koh N, Maechler M. 2014. hexbin: hexagonal binning routines. *R Software Package for Binning and Plotting.* **http://cran.r-project.org/web/packages/hexbin/index.html**

Carr DB, Littlefield RJ, Nicholson WL, Littlefield JS. 1987. Scatterplot matrix techniques for large *N. J. Am. Stat. Assoc.* 82:424–36

Carr DB, Nusser S. 1996. Converting tables to plots: a challenge from Iowa State. *Stat. Comput. Stat. Graph. Newsl.* 6:11–18

Chang W, Wickham H. 2015. ggvis: interactive web graphics with R. *R Software Package for Data Visualization.* **http://ggvis.rstudio.com/**

Chelaru F, Smith L, Goldstein N, Bravo HC. 2014. Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods* 11:938–40

Cleveland WS. 1993. *Visualizing Data*. Summit, NJ: Hobart

Cleveland WS, Grosse E, Shyu WM. 1992. Local regession models. In *Statistical Models in S*, ed. JM Chambers, T Hastie, pp. 309–76. New York: Chapman & Hall

Crowder MJ, Hand DJ. 1990. *Analysis of Repeated Measures*. London: Chapman & Hall

Dang TN, Wilkinson L. 2014. ScagExplorer: exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 73–80. Piscataway, NJ: IEEE

De Jonge E, Tennekes M. 2013. Tabplotd3: interactive inspection of large data. *R Software Package for Data Visualization.* **http://cran.r-project.org/web/packages/tabplotd3/index.html**
Dey T, Phillips DJ, Steele P. 2011. A graphical tool to visualize predicted minimum delay flights. *J. Comput. Graph. Stat.* 20:294–97
Emerson JW, Green WA, Schloerke B, Crowley J, Cook D, et al. 2013. The generalized pairs plot. *J. Comput. Graph. Stat.* 22:79–91
Feinberg J. 2010. Wordle. In *Beautiful Visualization: Looking at Data Through the Eyes of Experts*, ed. J Steele, N Iliinsky, pp. 37–58. Sebastopol, CA: O'Reilly
Friedman JH, Stuetzle W. 2002. John W. Tukey's work on interactive graphics. *Ann. Stat.* 30:1629–39
Friendly M. 2014. Comment on the generalized pairs plot. *J. Comput. Graph. Stat.* 23:290–91
Gelman A, Unwin A. 2013. Infovis and statistical graphics: different goals, different looks. *J. Comput. Graph. Stat.* 22:2–28
Guha PK, Kidwell P, Hafen RP, Cleveland WS. 2009. Visualization databases for the analysis of large complex datasets. *JMLR Workshop Conf. Proc. Vol. 5: Proc. 12th Int. Conf. Artif. Intell. Stat., Clearwater Beach, FL, April 16–18*, ed. D van Dyk, M Welling, pp. 193–200. Berkeley, CA: Microtome
Hafen RP, Cleveland WS. 2015. Tessera. *Data Analysis and Visualization Software.* **http://tessera.io/**
Hafen RP, Russell K, Owen J. 2015. Rbokeh: R interface for Bokeh. *R Software Package for Data Visualization.* R package version 0.2.3.2. **http://hafen.github.io/rbokeh/rd.html**
Hand DJ, Blunt G, Kelly MG, Adams NM. 2000. Data mining for fun and profit. *Stat. Sci.* 15:111–31
Hartigan JA. 1975. Printer graphics for clustering. *J. Stat. Comput. Simul.* 4:187–213
Hocking TD, VanderPlas S, Sievert C. 2015. Animint: interactive animations. *R Software Package for Data Visualization.* **http://github.com/tdhock/animint**
Hofert M, Mächler M. 2014. A graphical goodness-of-fit test for dependence models in higher dimensions. *J. Comput. Graph. Stat.* 23:700–16
Hofmann H, Cook D, Kielion C, Schloerke B, Hobbs J, et al. 2011. Delayed, canceled, on time, boarding... flying in the USA. *J. Comput. Graph. Stat.* 20:287–90
Hofmann H, Follett L, Majumder M, Cook D. 2012. Graphical tests for power comparison of competing designs. *IEEE Trans. Vis. Comput. Graph.* 18:2441–48
Hofmann H, Vendettuoli M. 2013a. Common angle plots as perception-true visualizations of categorical associations. *IEEE Trans. Vis. Comput. Graph.* 19:2297–305
Hofmann H, Vendettuoli M. 2013b. Ggparallel: variations of parallel coordinate plots for categorical data. *R Software Package for Data Visualization.* **http://CRAN.R-project.org/package=ggparallel**
Hurley C, Oldford R. 2011a. Eulerian tour algorithms for data visualization and the PairViz package. *Comput. Stat.* 26:613–33
Hurley C, Oldford R. 2011b. PairViz: visualization using Eulerian tours and Hamiltonian decompositions. *R Software Package for Data Visualization.* **http://cran.r-project.org/web/packages/PairViz/index.html**
Hyndman RJ, Shang HL. 2010. Rainbow plots, bagplots and boxplots for functional data. *J. Comput. Graph. Stat.* 19:29–45
Inselberg A. 1985. The plane with parallel coordinates. *Vis. Comput.* 1:69–91
Jockers ML. 2014. *Text Analysis with R for Students of Literature*. New York: Springer
Kaplan A, Hare E, Hofmann H, Cook D. 2010. Can you buy a president? Politics after the Tillman Act. *Chance* 27. **http://chance.amstat.org/2014/02/president/**
Lins L, Klosowski JT, Scheidegger C. 2013. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Trans. Vis. Comput. Graph.* 19:2456–65
Majumder M, Hofmann H, Cook D. 2013. Validation of visual statistical inference, applied to linear models. *J. Am. Stat. Assoc.* 108(503):942–56
Mosley L, Cook D, Hofmann H, Kielion C, Schloerke B. 2010. Visually monitoring the 2008 election. *Chance* 23
Moustafa RE, Hadia AS, Symanzik J. 2011. Multi-class data exploration using space transformed visualization plots. *J. Comput. Graph. Stat.* 20:298–315
Murrell P, Potter S. 2015. GridSVG: export grid graphics as SVG. *R Software Package for Data Visualization*. **https://cran.r-project.org/web/packages/gridSVG/index.html**

Newell M, Cook D, Hofmann H, Jannink JL. 2013. An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure. *Ann. Appl. Stat.* 7:1898–916

R Development Core Team. 2014. *R: a language and environment for statistical computing*. Vienna: R Found. Stat. Comput.

RStudio. 2015. Shiny. *Web application framework for R*. **http://shiny.rstudio.com/**

Sadana R, Major T, Dove A, Stasko J. 2014. Onset: a visualization technique for large-scale binary set data. *IEEE Trans. Vis. Comput. Graph.* 20:1993–2002

Schonlau M. 2003. Visualizing categorical data arising in the health sciences using hammock plots. *Proc. Sect. Stat. Graph. Am. Stat. Assoc.* **http://www.schonlau.net/publication/03jsm_hammockplot.pdf**

Shvachko K, Kuang H, Radia S, Chansler R. 2010. The Hadoop distributed file system. *Proc. 2010 IEEE 26th Symp. Mass Storage Syst. Technol.*, pp. 1–10. Piscataway, NJ: IEEE

Sievert C, Shirley KE. 2014. LDAvis: a method for visualizing and interpreting topics. *Proc. Workshop Interact. Lang. Learn. Vis. Interfaces, Baltimore, MD, June 27*, pp. 63–70. Stroudsburg, PA: Assoc. Comput. Linguist.

Steele J, Iliinsky N, eds. 2010. *Beautiful Visualization: Looking at Data Through the Eyes of Experts*. Sebastopol, CA: O'Reilly Media

Stolte C, Chabot C, Hanrahan P. 2003. Tableau. *Software for Business Intelligence and Analytics.* **http://www.tableau.com/**

Tennekes M, De Jonge E. 2014. Tabplot: Tableplot, a visualization of large datasets. *R Software Package for Data Visualization.* **http://cran.r-project.org/web/packages/tabplot/index.html**

Tierney L. 1991. *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York: Wiley

Tufte E. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics

Tukey JW. 1986. Sunset salvo. *Am. Stat.* 40:72–76

Tukey JW, Tukey PA. 1985. Computer graphics and exploratory data analysis: an introduction. In *The Collected Works of John W. Tukey: Graphics 1965–1985*, Vol. 5, ed. WS Cleveland, pp. 419–38. New York: Chapman & Hall

Urbanek S. 2015a. Ioplots. *High-performance I/O tools to run distributed R jobs seamlessly on Hadoop.* **https://github.com/s-u/iotools**

Urbanek S. 2015b. RCloud. *Software for Collaboratively Developing and Sharing R Scripts.* **http://stats.research.att.com/RCloud/**

Urbanek S, Theus M. 2003. iPlots—high interaction graphics for R. *Proc. 3rd Int. Workshop Distrib. Stat. Comput., Vienna, March 20–22*, ed. K Hornik, F Leisch, A Zeileis. **https://www.r-project.org/conferences/DSC-2003/Proceedings/UrbanekTheus.pdf**

Vaidyanathan R, Xie Y, Allaire J, Cheng J, Russell K. 2015. HtmlWidgets: html widgets for R. **http://www.htmlwidgets.org**

Van Long T, Linsen L. 2011. Visualizing high density clusters in multidimensional data using optimized star coordinates. *Comput. Stat.* 26:655–78

Wattenberg M, Viégas F. 2010. Beautiful history: visualizing Wikipedia. In *Beautiful Visualization: Looking at Data Through the Eyes of Experts*, ed. J Steele, N Iliinsky, pp. 175–91. Sebastopol, CA: O'Reilly

Wattenberg M, Viegas FB, Hollenbach K. 2007. Visualizing activity on Wikipedia with chromograms. In *Human-Computer Interaction–INTERACT 2007*, pp. 272–87. Berlin: Springer

Wegman E. 1990. Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc.* 85:664–75

Wickham C. 2011a. A tale of two airports: exploring flight traffic at SFO and OAK. *J. Comput. Graph. Stat.* 20:291–93

Wickham H. 2011b. The split-apply-combine strategy for data analysis. *J. Stat. Software* 40:1–29

Wickham H. 2013. *Bin-summarise-smooth: a framework for visualising large data*. Tech. Rep. **http://vita.had.co.nz/papers/bigvis.html**

Wickham H, Chang W. 2014. Ggplot2: an implementation of the grammar of graphics. *R Software Package for Data Visualization*. **http://cran.r-project.org/web/packages/ggplot2/index.html**

Wickham H, Chowdhury NR, Cook D. 2014. Nullabor: tools for graphical inference. *R Software Package for Data Visualization*. **http://cran.r-project.org/web/packages/nullabor/index.html**

Wickham H, Cook D, Hofmann H. 2015a. Visualizing statistical models: removing the blindfold. *Stat. Anal. Data Min.* 8:203–25

Wickham H, Francois R, RStudio. 2015b. Dplyr: a grammar of data manipulation. *R Software Package for Data Manipulation.* **http://cran.r-project.org/web/packages/dplyr/index.html**

Wickham H, Lawrence M, Lang DT, Swayne DF. 2008. An introduction to rggobi. *R-news* 8:3–7

Wickham H, Swayne DF, Poole D. 2009. Bay Area blues: the effect of the housing crisis. In *Beautiful Data: The Stories Behind Elegant Data Solutions*, ed. T Segaran, J Hammerbacher, pp. 303–19. Sebastopol, CA: O'Reilly

Wicklin R. 2011. Visualizing airline delays and cancelations. *J. Comput. Graph. Stat.* 20:284–86

Wilkinson L, Anand A, Grossman RL. 2005. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization (InfoVis 05), Minneapolis, Minn., October 23–25*, pp. 157–64. Piscataway, NJ: IEEE

Wilkinson L, Anand A, Urbanek S. 2012. Scagnostics: compute scagnostics - scatterplot diagnostics. *R Software Package for Data Analysis.* **http://cran.r-project.org/web/packages/scagnostics/index.html**

Xie Y, Hofmann H, Cheng X. 2014. Reactive programming for interactive graphics. *Stat. Sci.* 29:201–13

Annual Review of Statistics and Its Application

Volume 3, 2016

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics