Data Visualization and Statistical Graphics in Big Data Analysis

Dianne Cook, Department of Econometrics and Business Statistics, Monash University
Eun-Kyung Lee, Department of Statistics, Ewha Womans University
Mahbubul Majumder, University of Nebraska-Omaha

August 24, 2015

Abstract

This article discusses the role of data visualization in the process of analyzing big data. We describe the historical origins of statistical graphics through the birth of exploratory data analysis and how this impacts practice today. Examples of contemporary data visualizations developed in the process of exploring airline traffic, global standardized test scores, monitoring elections, wikipedia edits, the housing crisis as observed in San Francisco and mining credit card databases. A review of recent literature is provided. Good data visualization yields better models and predictions, and allows for the discovery of the unexpected.

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

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 and 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 such widely visible systems such as GapMinder (`http://www.gapminder.org`) and BabyNameVoyager (`http://www.babynamewizard.com/`) owe some credit to interactive graphics that arose in the first years of EDA research.

It's surprising that the stem-and-leaf plot has persisted in the classroom to present day. The pencil paper methods were essential to the 1970s because access to the technology to do 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 very much about harnessing good computer-generated plots of data. Tukey was fortunate enough to have access to state-of-the-art technology, and was also an early advocate of harnessing technology for data analysis. Forty-five years ago he foresaw today's technological big data world and just how important computational tools would be 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 entirely arise in a vacuum. Applied statistical practice has always utilized data plots prior to modeling to check assumptions and for post-model assessment of the fit. Crowder and Hand (1990) make this very clear:

> *"The first thing to do with data is to look at them.... 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."*

Big data provides new challenges for data visualization. Being able, and knowing how to, make good data plots is an indispensable component of wrangling with big data. The term "big data" means something different depending who you listen to, read, or chat with. The working definition for this article is not entirely to do with size of the data in terms of variables, or samples, but also the complexity. It may be data that has hundreds of variables, stored in many related tables, a large database that has been collected and perhaps neglected for many years, repositories of emails, health records from machines in almost every doctor's office that automatically files information, communications networks from social media, scores from tests administered to our youth across the globe, humongous quantities of data simulated from global climate models to assess the impact of climate change, or new business data being collected and stored in new systems like hadoop. Big data potentially informs us about our world, to learn how to be more efficient in business operations or in delivering health care, what statistics we need improve to get our tennis game to the level of a champion, or who are the key people that bind a social network into a cohesive group.

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

In classical statistical inference and experimental studies, the role of data is primarily to prove or disprove the conjecture. In this case, the distribution of a test statistic is preferably derived theoretically. Data is used to obtain an empirical distribution when theory is not possible. In any case, the role of data is passive and has very specific purpose. We can illustrate this by an example. In a situation where researchers want to test the population mean being equal to a certain value, they don't need the data to find the distribution of

test statistic as long as some assumptions are met. Data is 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 is only needed to finally make the 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 in mind, or testing methods. The challenge is to explore data and discover hidden value in the data which, later, may lead to more formal hypotheses and classical methods to test them. Big data changes the role of data into one that is more active. This essentially requires adopting EDA and visualization as very early steps for big data analysis.

This paper has two components: (1) Contemporary illustrations of the usefulness of data visualization for understanding data, and (2) 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 upfront if we do not name all the people and work that have made important contributions, but 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 more interesting developments.

2 Illustrations of Visualization for Understanding Data

2.1 How to Win a Data Mining Challenge!

Two stories from 2014 point to the use of graphics from successful data mining teams: the key to their model winning the competition was the data pre-processing involving a lot of data plots that helped them to understand what they were working with, and problems with the data that needed to be addressed before being able to make effective models. This is important for big data, how to effectively clean, transform and pre-process large complex data sets. Visualization plays a key role. This is just as important for big data as it was for John Tukey's days, and 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 "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 combined forces with another team to win the final prize of $500,000. This competition is an example of big data challenges of today: large amounts of data on hospital admissions being used to develop models to improve the efficiency of healthcare spending. In interviews post-prize, 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." `http://www.anotherdataminingblog.blogspot.co.uk/2011/12/whats-going-on-here.html`

There is a concrete example on Dr. Brierley's blog that illustrates how a plot indicated to him that the challenge was not worth entering. It is 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" – yes, a really simple time series plot – from which he observed:

> *"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."*

2.1.2 2014 Data Mining Cup Winners' Strategies

Each year Prudsys AG challenges students with the Data Mining Cup competition. In 2014, the problem was announced to student team on April 2, and students needed to have their final entry by 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 (Guillermo Basulto-Elias (statistics), Fan Cao (statistics), Xiaoyue Cheng (statistics), Marius Dragomiroiu (computer science), Jessica Hicks (bioinformatics and computational biology), Cory Lanker (statistics), Ian Mouzon (statistics), Lanfeng Pan (statistics) and Xin Yin (bioinformatics and computational biology/statistics) from Iowa State University was the first north American team to win. A key component of that win was the pre-processing of the data, that utilized substantial graphics to learn about their data, and inform their modeling.

Figure 1 shows one plot used early by the ISU DMC team, to examine return rates by time and product. Yellow indicates the ordered items were kept, blue means they were returned and pink are items to be predicted. Yes, it is a really ugly plot!!! However, it is very informative: Two major structures are immediately visible, new product introductions in July 2012 and January 2013. The most important feature, though, is that new data to be predicted was in the third season of the time period, and this information was crucial to construct good training and test sets for model building.

Figure 1: One of the preliminary plots made by the ISU DMC team. Item ID is plotted against order date, colored by return rate. The students learned that the data to be predicted (pink) did not look like the training data provided for building a model (blue/yellow).

2.2 Visualization is NOT Prediction .... and this is ok

So much of statistics is preoccupied with predictive models which is very 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 published work are shown that illustrate visualization being used in the process of analyzing big data sets. Generally what can be learned 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 OECD PISA

The Programme for International Student Assessment (PISA) is a triennial survey conducted by the Organization for Economic Cooperation and Development (OECD) with a rotating emphasis on one of mathematics, reading, or science. In 2012, the emphasis was on mathematics. The data was 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://beta.icm.edu.pl/PISAcontest/.

This data is big because there is a lot of information collected in addition to the test scores. In the student table there are records for about 500,000 students from 65 different countries, and 635 variables. The variables include information about gender, language, household possessions, attitude to math, use of the internet, many different aspects of their lives. The parent table has 143 variables from 100,000 parent-completed surveys providing information about the students households, such as if both parents are in the home or if its a foster home, parents occupations, how the child's school was selected. The school table contains survey results completed by 18,000 school principals producing 291 variables. These items include information about numbers of teachers, supply shortages, teacher turnover, educational background of teachers, streaming of classes. There are many different questions that we might try to answer with this 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 expectations of what associations we might see in the data.

One issue that we were interested in was about 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 the 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. Sample weights were utilized in calculating the averages. The result is shown in Figure 2. The absence of a universal math gap runs counter to the popular press. This data represents an observational study, and so it can only inform us about association. To understand some potential reasons why the gap does not exist should involve additional investigations into the samples used in each country. One quick

check reveals that it cannot be explained by differing proportions of boys and girls being tested: these 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.

This data is abounding with information ripe for exploration. We can learn about many associations between demographic factors and educational achievement about countries across the globe. These could be mined to form the basis for follow-up experimental studies. Visualization provides an excellent way to mine these associations, across the different categorical levels.

2.2.2 Elections

In the 2008 US Presidential election cycle a young man called Nate Silver burst onto the world stage with an accurate prediction of an Obama win. His web site `http://fivethirtyeight.com/` (538) has 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 in Mosley et al. (2010). Percentage difference in percentage for Obama vs McCain is plotted against time the poll was released, as were made available on the web site `http://www.electoral-vote.com/`. Each dot represents one poll result, and color indicates pollster. The grey line and ribbon represents a loess smoothing (Cleveland et al., 1992) across all the poll results to indicate trend. There is a lot of variation in polls, even when conducted in very similar time frames. The variation in results can be as high as 10 percentage points.

Differences between polls produced by different pollsters can be seen. DailyKos light blue) is consistently higher than the trend, and consistently produced the most pro-Obama results. Rasmussen (pink) tended to be fairly close to the trend or below it (pro-McCain). Hotline (yellow) was varied early on in the season but closer to election day was near the average of all other polls. Gallup (orange) is noticeably varied, it has some of the most pro-McCain results as well as the most pro-Obama results. Gallup is a legacy American pollster who dates back to the early 1900s, and we would have expected that they would be providing more reliable polling numbers than observed. Interestingly the 538 web site now has detailed ratings of the major pollsters operating in the USA, and Gallup scores a poor C+. The plot of the national trend polls allows us to see the variability and the bias' of polling organizations. The pollster DailyKos, also known as Research 200, is a community action organization with political leanings towards to Democratic party. Actually it is one of the pollsters currently "banned" by 538! Rasmussen on the other hand 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 number of electoral votes, roughly based on the size of the population. For example, in

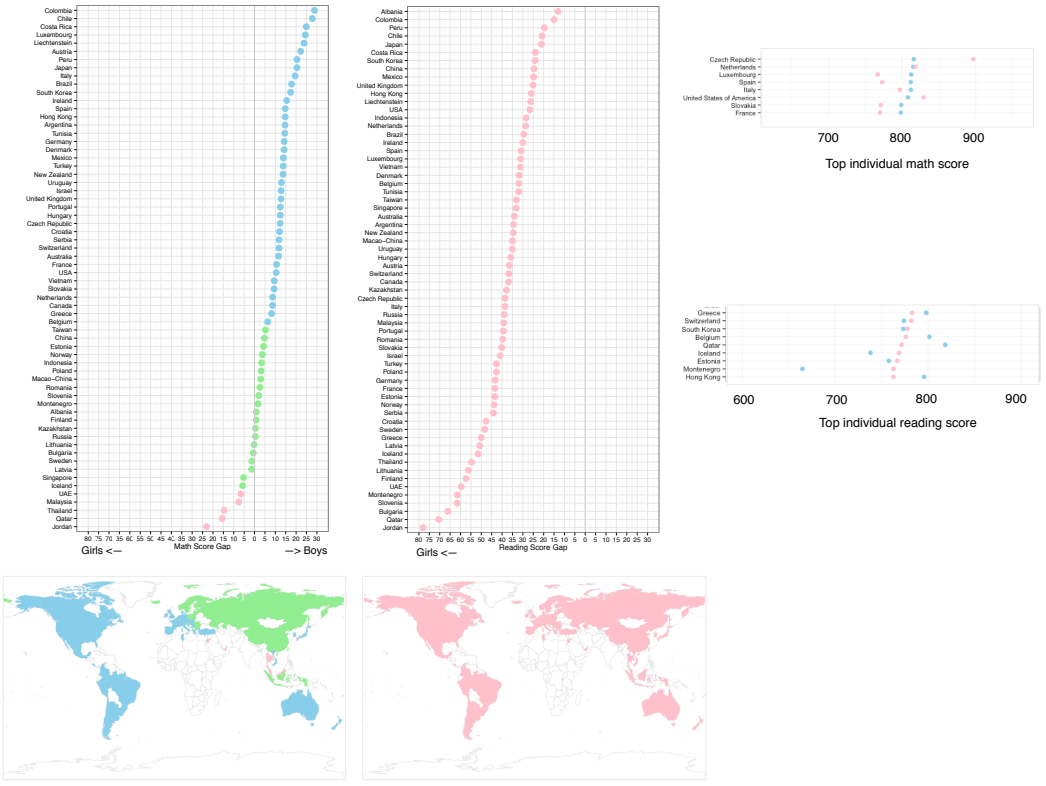

Figure 2: Examining the gender gap in math and reading by country. Dotplots of mean difference shown by country, along with maps. Color roughly indicates magnitude of the gap, blue more than 5 points in favor of boys, green within 5 points, and pink more than 5 points in favor of girls. Surprisingly, this data indicates that the gender gap in math is not universal, many countries do not have a gaps, and a few countries show a gap in favor of girls. On the other hand, 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) score in a few countries, the story is different. Even in countries with a big gender gap in math, e.g. the USA 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 in the data from the maps, is the lack of information from the continent of Africa.

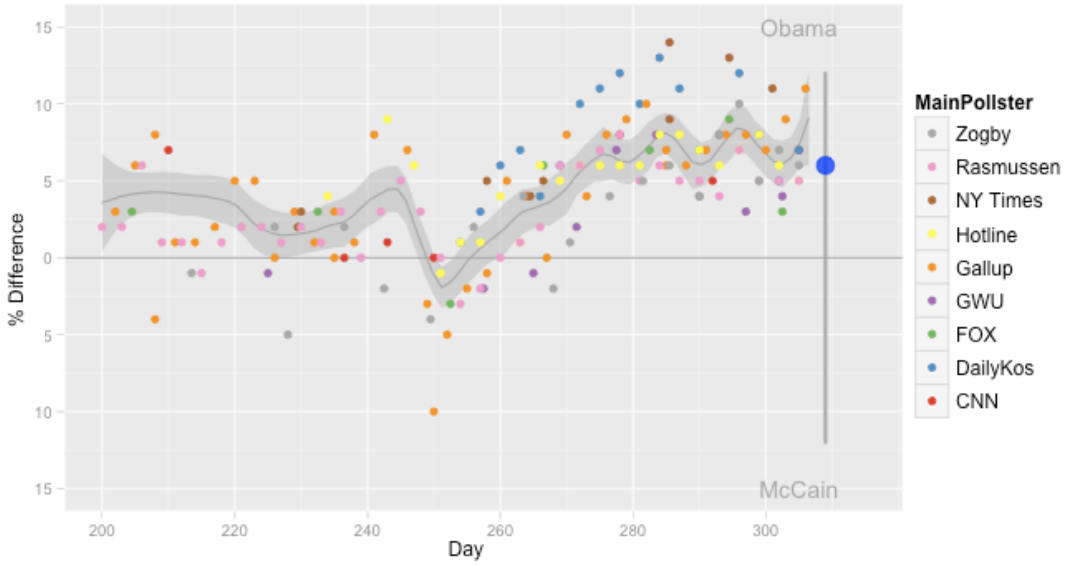

Figure 3: Tracking polls and final popular vote leading up to the 2008 US Presidential election. Percentage difference by time of poll release. Color represents pollster, gray strip shows poll average and the large blue dot indicates final election margin.

2008 Iowa was worth 7 votes, whereas New York was worth 31. The electoral votes for most states as assigned based on winner take all: Obama won New York with 57% in favor, so he earned all the 31 points. A candidate needs to tally up 270 or more points (of the possible 538, hence Nate Silver's web site name) to win the presidency. Figure 4 illustrates the state by state variation in polls. State, top to bottom most support for McCain to most support for Obama, is plotted against percentage difference. Each point indicates some result: the red/blue represents the final election result, black indicates the median of all polls, grey the median of the previous week's polls and white indicates all polls in that state. The yellow strip straddles a 5% margin, with results in this range being states that are too close to call. States in these position tend to have a lot of pollsters operating, so there are more white dots visible, and for the most part it can be seen that the final result closely matched the poll results. There are a couple of exceptions: Montana was predicting 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, we produced these plots and animated them from the previous week, which allowed obtaining a sense of temporal shifts in attitude, and the variability leading up to the actual vote.

In the 2012 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). And the 538 site is expanded to an inde-

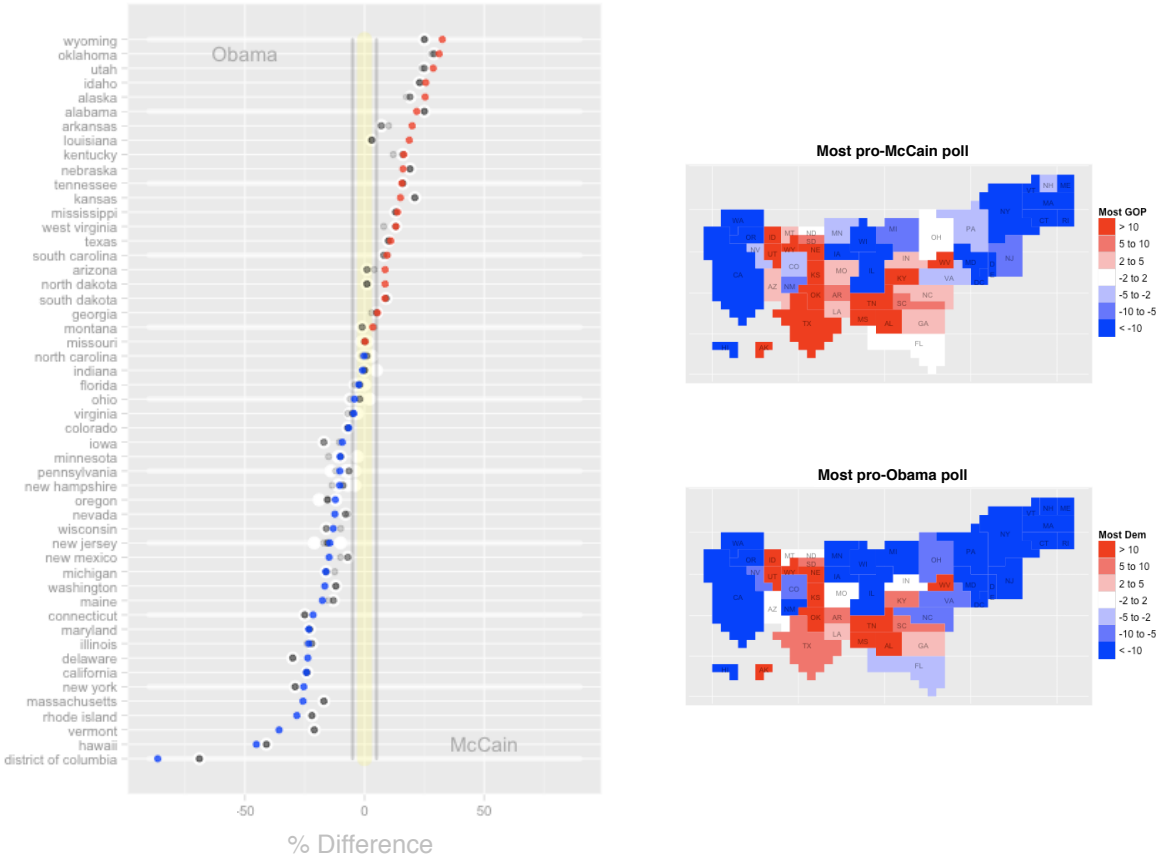

Figure 4: Exploring variability in polls. (Left) Percentage difference by state, top to bottom ordered by McCain to Obama advantage. Color indicates final election result. Black is the median of all polls, grey is the median of the previous week's polls, and white shows all polls. (Right) Block cartograms, size of state represents electoral votes, colored by the most favorable poll result for each candidate, just prior to election day, in each state. In the best case scenario for McCain, the country still looks predominantly blue.

pendent news site and is an exemplary location to browse to read examples of numerically and visually analyzing large data du jour.

2.2.3 Airline Traffic

Every few years, the graphics and computing sections of the American Statistical Association provides a data challenge, the Data Expo (`http://stat-computing.org/dataexpo/`), and encourages students, faculty, industry statisticians to make a visual analysis. In 2009 the data consisted *"of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed."*.

There were 9 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 air travel in the USA: calendar view summaries show differences between years, months and days of the week, time series display volume of traffic and weekly cycles, and a heatmap is used to compare carriers over the time period. This was a lot of data 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 origin to destination show which carriers operate on a hub system and which don't. Time series of volume at major airports show effects of events like the 9/11 tragedy. Side-by-side boxplots were utilized to display delays by airport, revealing the problematic EWR, SFO, ORD, LGA and the efficient operations of DTW, MSP and DFW (Figure 5). Facetted scatterplots with overlaid loess fits show trends in delays by carrier. This group also looked for "gaps" in the data, where planes are last seen at one airport, and then magically appear at another airport. These gaps correspond to ghost flights, planes that fly passengerless, in order to get a vehicle into a location that it is needed. It represents inefficiency in operations. Most carriers have been reducing this costly operation, but Northwest airlines had an increase in the latter few years of this data. Delta, which merged with Northwest in 2008, saw improvement in efficiency. Interactive graphics was employed to examine different chunks of the data, like the relationship between fuel consumption, distance flown by carrier and year (Figure 6). Basic plots revealed that many problems with the data, 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. Stringing delays together with weather patterns revealed the problems that strong cross-winds cause at airports. And using some interpolation between geographic locations of airports, the departure and arrival times, allowed animating air traffic flow across the nation (`https://vimeo.com/119233995`).

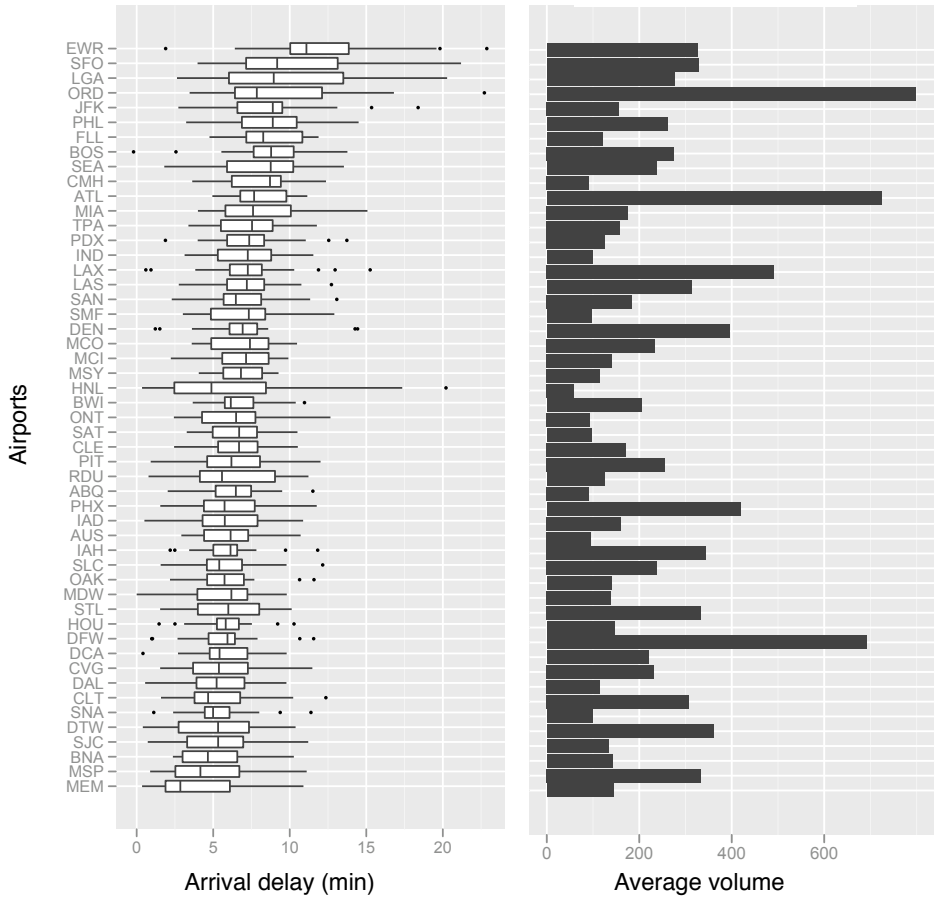

Figure 5: Arrival delays by airport, sorted top to bottom from worst average delay to least, and volume of traffic. Only the top 50 airports based on delay are shown. Over this time period EWR (Newark, NJ) had the worst record in delays even though their traffic volume was not large compared to other airports. ORD (Chicago, IL) had a bad delay record, but they also have the highest volume. DFW (Dallas-Fort Worth, TX) compared favorably with low average delays and very high volume.

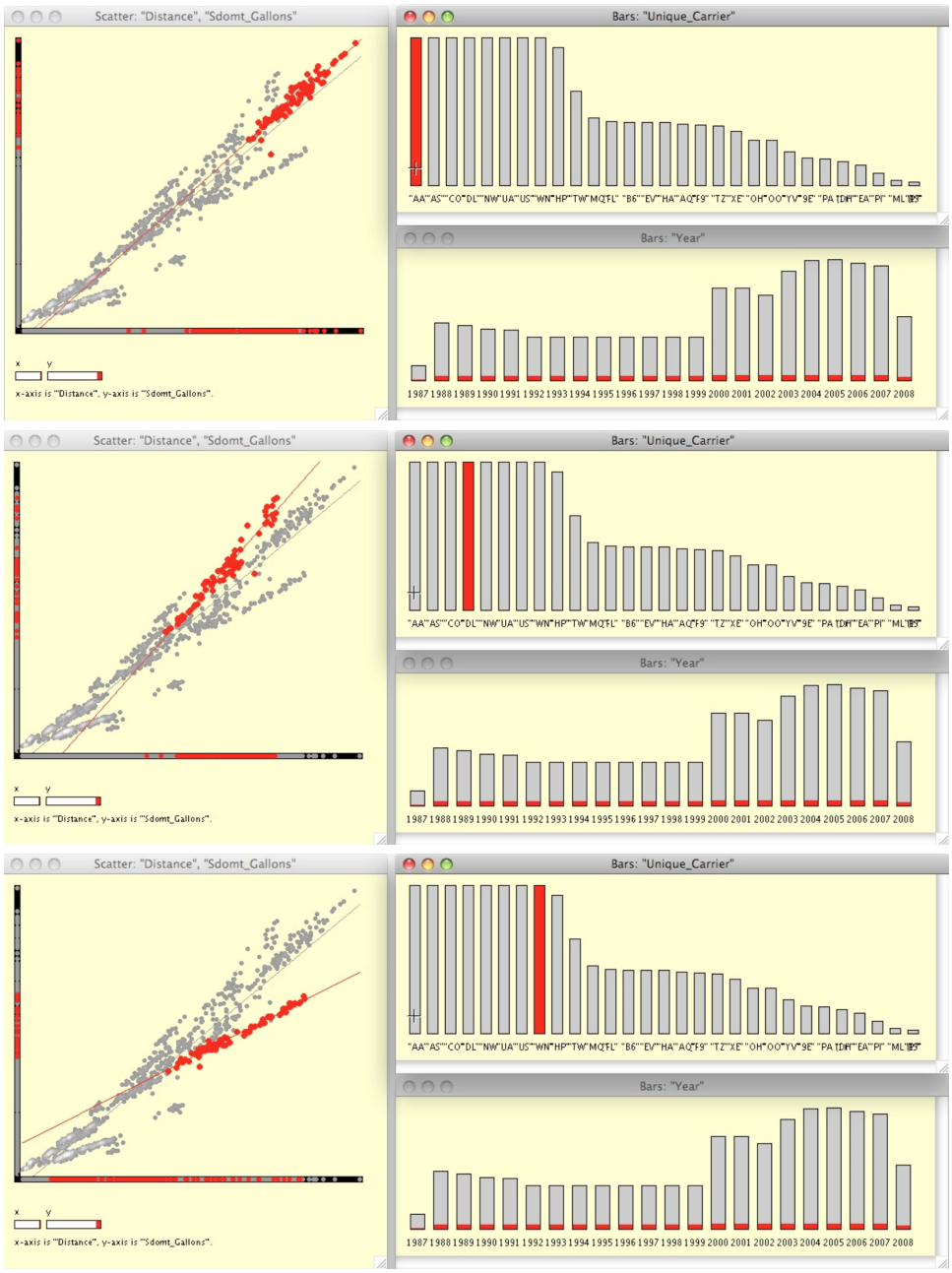

Figure 6: Three snapshots of interactive graphics using the software MANET on a processed chunk of data to explore carrier efficiency. Fuel consumption (vertical) is plotted against distance flown (horizontal) in the scatterplot, and bar charts show carrier, and year. Three airlines are highlighted (red) from top to bottom: American Airlines (AA), Delta (DL) and Southwest (WN). American Airlines is at the top of the pack: big carrier, big consumer. Over this period of time, Delta is relatively inefficient, having relatively higher fuel consumption for the same distance flown. In recent years, which can't be seen in this data, they have improved substantially. Southwest is a big carrier but has substantially more efficient fuel consumption than their competitors.

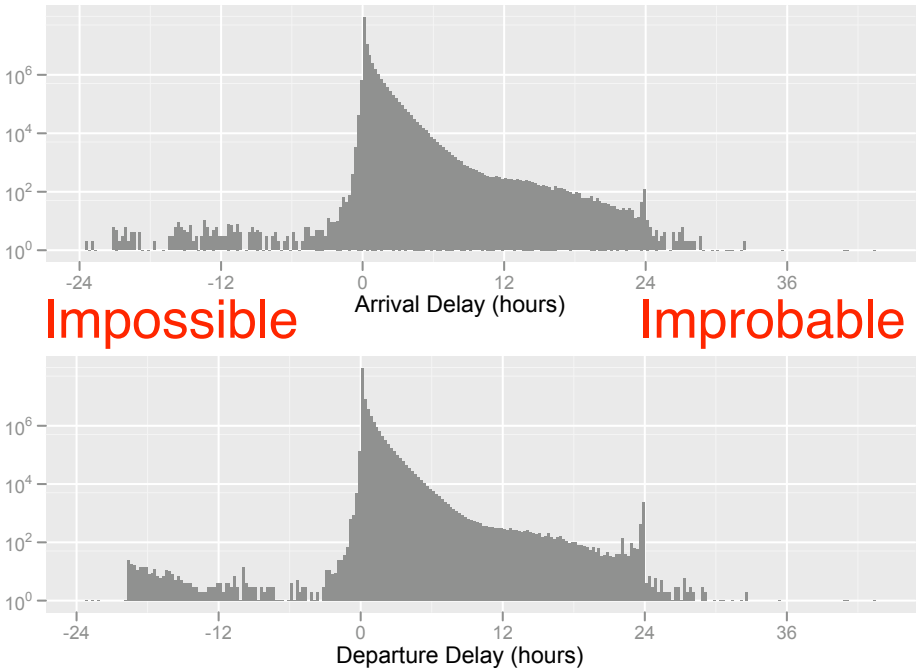

Figure 7: Histograms showing arrival and departure delays across carriers, airports and years. Delays aggregated in 15 minute bin widths. Arrival delay is calculated from the difference between the reported actual and scheduled departures. Arrival delays of -24 hours would be a plane that left a day early, which we would expect is impossible.

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 and Viégas (2010), with associated web sites `http://fernandaviegas.com/wikipedia.html`, `http://www.bewitched.com/historyflow.html` where you can see some of the visualizations. The book chapter describes the process of pulling the data, pre-processing and making visualizations of different aspects. Their endeavor began in 2003, early days for the encyclopedia. As an example of the magnitude of the data, the page on "Microsoft" had 198 edits generating 6.3 megabytes of text, and the page on "Cat" had just 54 edits.

The edits data is huge. To tackle is Wattenberg and Viega, initially created an interactive visualization for single pages. They use a modified parallel coordinate plot (Inselberg, 1985; Wegman, 1990), which might also be considered to be a variation of a hammock plot (Schonlau, 2003; Hofmann and Vendettuoli, 2013b), 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. (Anonymous editors are grey.) The overall height of the plot indicates the length of the article. There are two prominent examples shown: "chocolate" and "abortion". The displays enable some information to be immediate: if there is ownership of a page by a few editors then there are a few differently colored stripes. The page for "abortion" has a big empty patch indicating tha the entire page was erased temporarily, probably by a malicious editor. Tug-or-wars between editors can be seen in some politically or emotionally controversial subjects. Interaction is provided 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 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 enable viewing how an editor has contributed to wikipedia. These both would be considered to be drill downs into the data, because each shows a very narrow slice of the data. There are now 4,714,447 pages, to visit each one would take some time. Ideally, approaching a problem as large as this would also provide some larger visual overviews: number of pages over time, number of authors over time, how many pages different authors edit, ... and provide some comparative views, pairs of pages for example, or hierarchical topic lists, or how pages link to each other. The data implores the analyst to display it in many more different ways. The web site `http://infodisiac.com/Wikimedia/Visualizations/` illustrates ways that many other people have tackled visualizing wikipedia.

Raw information in the form of text provides new challenges for visualization. Universally, people have 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 dif-

ferent facets of text can be found in 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 and Shirley, 2014) provides ways to interactively visualize the data.

2.2.5 The San Francisco Housing Crisis

The Wikipedia example is published in a book called "Beautiful Visualization". Another book in the Beautiful series is "Beautiful Data" and in that volume you will find an article by Wickham et al. (2009) that illustrates visually exploring the housing crisis with an early version of the R software `ggplot2` (Wickham and Chang, 2014). This package is now very widely used for plotting data. This book chapter is is an extraordinary example of pulling data from the web, cleaning and displaying it in different ways to learn about events that affect our lives. The explanation of the process is superb.

The story of the housing crisis begins by studying the temporal scale: average prices and number of sales from 2003-2009. We can see the average house price double over this period, and then drop by half starting summer 2007. Sales tend to cycle some, showing some seasonality, but clear decline starting in 2006. Interestingly, sales tick up again early in 2008 after sharp declines in average price. The authors then examine economic conditions, and compare inflation adjusted, alongside unadjusted, average price 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 regions shows that there are some areas of the city 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 city experienced more turnover in houses. Comparing price decline and demographic factors revealed that higher income areas, and higher percentage of college grads saw less decline, while areas where residents had longer commutes saw bigger hose price declines.

This article 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 transactions database as the example. The reason to read this article is see how the humble histogram can be priceless for exploring big data. The histogram, with carefully crafted £1 binwidth, is used to display petrol (gas) purchases, and department store spending. We might expect that in department stores there to 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 some drivers, a lot of drivers, like to spend whole nice round whole pound amounts when purchasing petrol. Beyond 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 of small 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 visualization of large data in the most likely candidate statistical journals, e.g. Journal of Computational and Graphics Statistics, Computational Statistics, Annals of Applied Statistics. Graphics papers in these journals describe methods useful with relatively small amounts of data, or for special statistics-related purposes, rather than providing solutions to visualizing large amounts of data. 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.

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 address.

- 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, putting them in a loop to animate the display of all pairs is infeasible when the number of variables is really large. Wilkinson et al. (2005) resurrected Tukey's ideas on scagnostics (Tukey and 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 the scatter plot - outlying, skewed, clumpy, sparse, striated, convex, skinny, stringy, and monotonic - based on proximity in graphs. The methods are implemented in an R package `scagnostics` (Wilkinson et al., 2012), and a standalone software, ScagExplorer (Dang and Wilkinson, 2014).

When the number of cases is large, but the number of variables is small, reading distributions from 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) 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.

A third issue arises with big data, that the many variables may be of different types, categorical or temporal in addition to numeric. The work described in Emerson et al. (2013) and Friendly (2014) on the generalized pairs plots, and accompanying R packages (`gpairs` in `YaleToolkit` and `ggpairs` in `GGally`) adapt the scatterplot matrix ideas for heterogeneous variable types.

That statisticians frequently use scatterplots for examining association between pairs of variables, despite the existence of many other graphical forms, 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 adapt the method to big data of today.

- 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 boxplots are the optimal way to make comparisons between groups. Histograms can also serve the purpose, but they provide more complex summaries of the distribution than a boxplot renders. Table plots came to prominence with the tableau software (Stolte et al., 2003), but they were recommended originally by Carr (1995) (also shown in Carr and 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, according to one of the variables in the collection or by another external criteria. 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 and De Jonge, 2014). It displays the sales records for houses sold in Ames, Iowa from 2008-2010. The variables are sorted by the sales price, the top represents the 1% most expensive houses, which have a mean value of about $500,000, and bottom represents the cheapest 1%, having a mean value of about $50,000. There are 1,615 houses in the data set, and these are grouped into 100 bins of equal size by the sales price. The mean values the other variables for houses in each these bins is shown in the other plots. The exception is the house

style, which is categorical, and so proportions of the different styles are shown in stacked bar charts. We learn from this display rough associations:

- Sales price is NOT closely associated with number of bedrooms, which is a little surprising. The pattern in the plot of bedrooms is fairly uniform, evenly distributed across all house prices, which leads to the conclusion that there is no association. On the other hand, there is a slight association with number of baths, because the average number of baths drops from 2 to 1 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, the higher priced homes are more commonly two story style (green) than the lower priced homes.

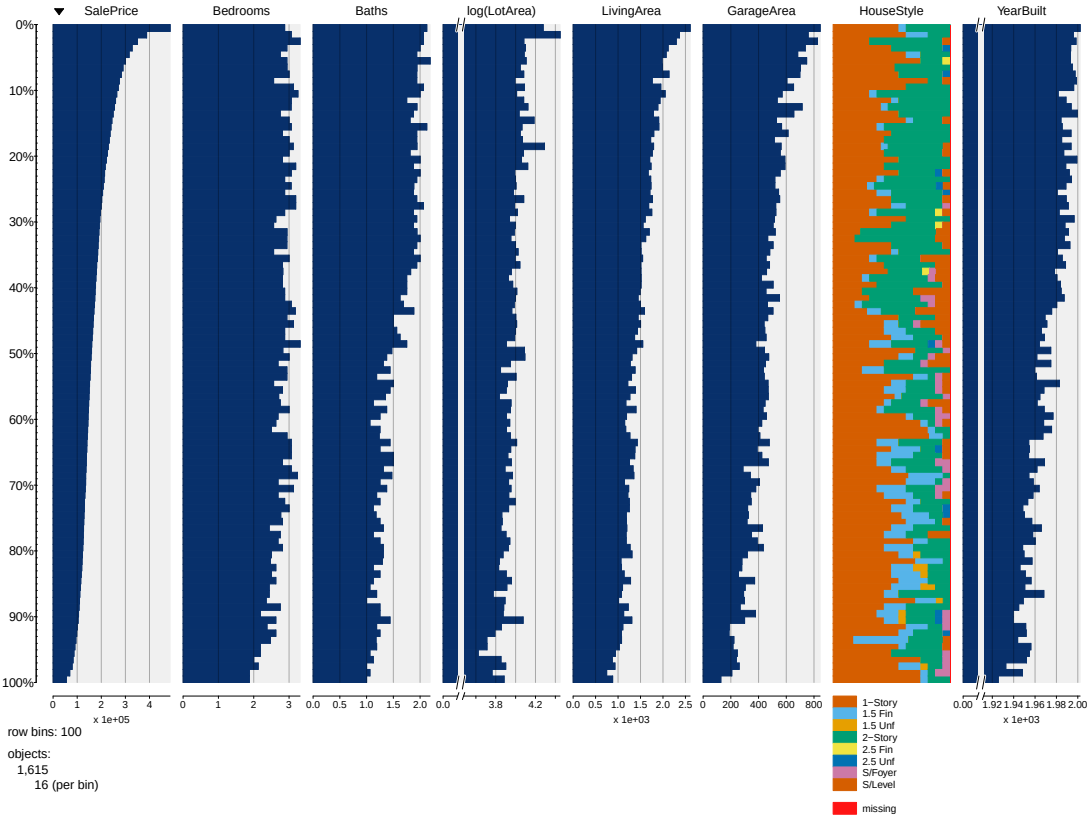

Figure 8: Table plot of Ames, Iowa, housing sales produced using the R package `tabplot`. Variables that have a positive association have similarly shaped distribution.

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, 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. to get a rough sense of the associations it may be reasonable to use, but it is ripe for re-designing 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 and Tennekes, 2013). It uses the `d3` software (Bostock et al., 2011) which is a javascript graphics library useful for creating interactive web visualizations.

- 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 and Oldford (2011a) initially begin working on graphs as a way to organize the axes of a parallel coordinate plot, but the work has extended to provide an algorithm for navigating high-dimensional spaces. Their R package, `PairsViz` (Hurley and Oldford, 2011b), implements the parallel coordinate plot style visualization. Figure 9 shows examples for data on nutritional value of chocolates: (a) is the traditional, and (b) is the eulerian parallel coordinate plot. In the Eulerian adaptation variables are repeated so that all pairs can be examined. Line color indicates type of chocolate (orange=milk, brown=dark). In both, the histogram displays a scagnostic value, the inverse Wilks Λ MANOVA statistic which indicates how well the two variables split the two groups, the higher the bar the more separated. Hofmann and Vendettuoli (2013a) describe variations of parallel coordinate plots for categorical data and their R package, `ggparallel` (Hofmann and Vendettuoli, 2013b), enables others to use the methods (Figure 10). Moustafa et al. (2011) provide envelope methods for parallel coordinate plots for large data.

- Rearranging, summarizing and plotting data elegantly

  When data becomes 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`, `fstread`, `bigrquery`, `ggmap`, `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.,

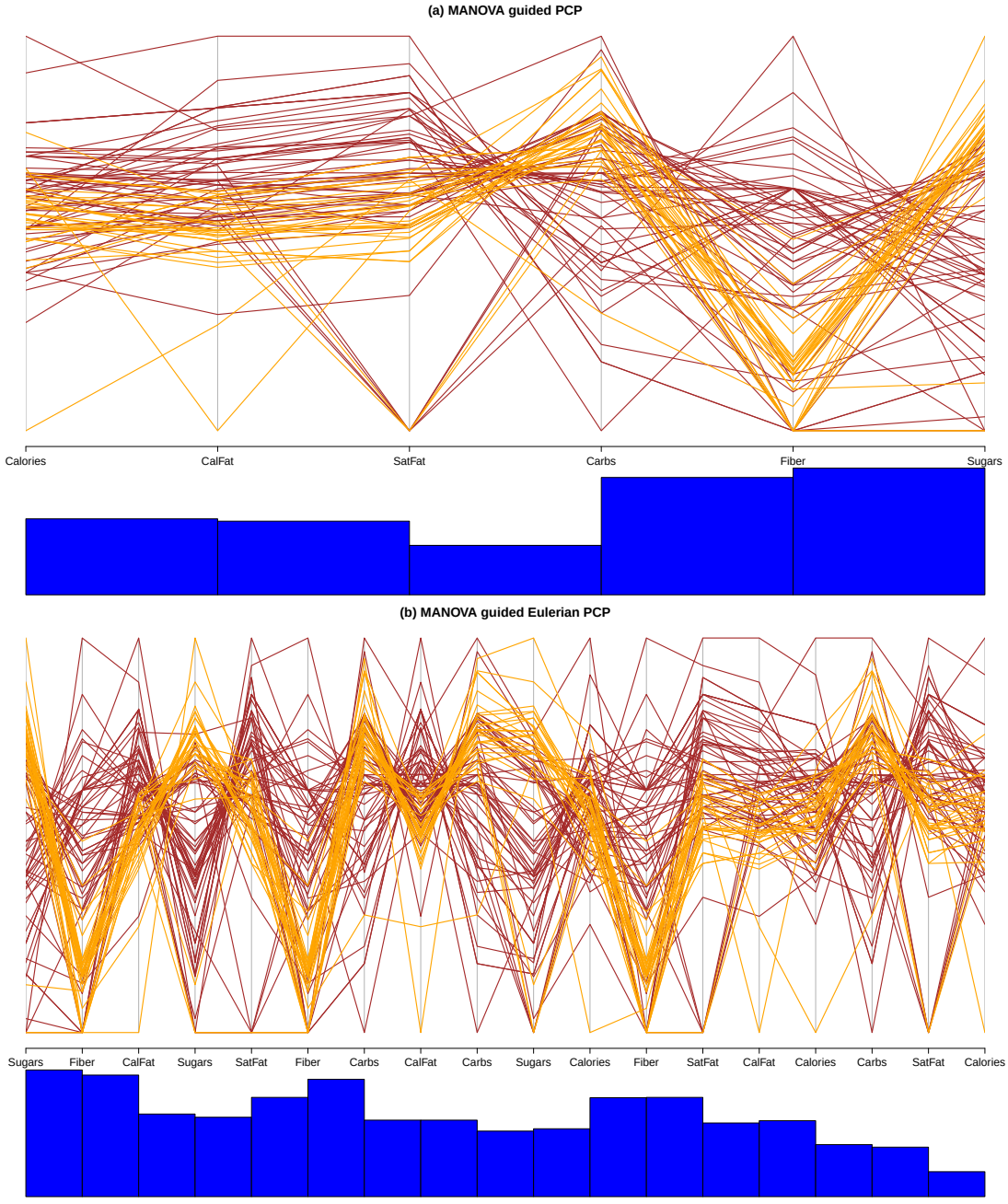

Figure 9: Classical (a) and eulerian (b) parallel coordinate plots of nutritional measurements on chocolates. The eulerian parallel coordinate plot more strongly indicates a difference between the two types of chocolate (orange=milk, brown=dark), and from the scagnostic describing separation displayed by the histogram, we can learn that higher values occur when Fiber is one of the two variables, indicating the importance of this variable.

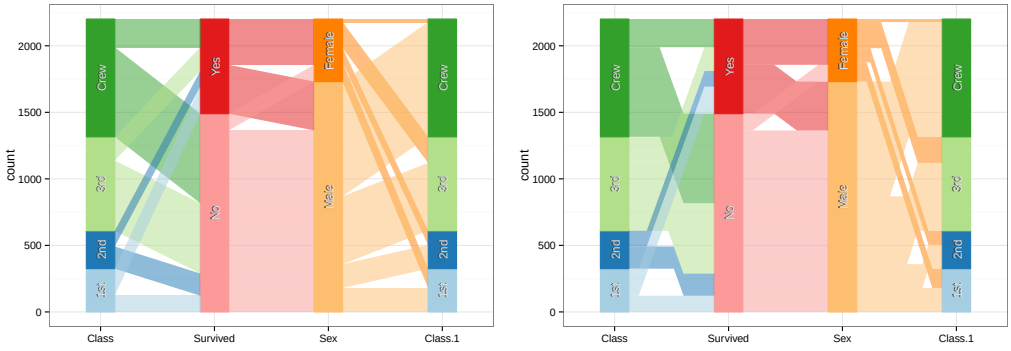

Figure 10: Variations on parallel coordinates for categorical variables. If the ribbons peel off from the bars being initially flat (right side plot) provides for more accurate reading of the proportions in each category.

2015b) packages is the split-apply-combine strategy, meaning that the data 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 with by 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 data. Chunks of data are placed in different locations and indexed using keys. Accessing this data requires operating on chunks and combining the results. The Tessera project (Hafen and 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. Another unpublished R package, `iotools` (Urbanek, 2015) is available for pulling data from hadoop storage units.

- Interactive graphics

  This area 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. Bostock et al. (2011)'s `D3`) available for interactive graphics, but close connections between these and statistical modeling, of the sort achieved in XLispStat (Tierney, 1991), is lagging. Some early explorations with interfaces in R were seen in `iplots` (Urbanek and Theus, 2003) which interfaced R with Java, and `rggobi` (Wickham et al., 2008) which interfaced R with C and gtk. Recent developments include `gridSVG` (Murrell and Potter, 2015), `animint` (Hocking et al., 2015), `cranvas` (Xie et al., 2014),

ggvis (Chang and Wickham, 2015) and shiny (RStudio, 2015). For python programmers, the library bokeh (Avila et al., 2015) provides interactive web graphics using javascript.

Both gridSVG and animint are relatively light-weight interactive graphics, designed to enhance existing static graphics, and add several interactive elements. These would be considered to be primarily for communication purposes, where the information in the data is known, and simply needs to be presented. The interactive graphics provides a level of sophistication and allows the user to change a little, but not much, of the information that 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 thoughout the plot. Video at https://vimeo.com/137049134 illustrates this done using gridSVG. Animint works similarly, taking ggplot2 plots and adding interaction using javascript.

Cranvas, and potentially ggvis, provide more substantial tools for exploring data generally with interactive graphics. Cranvas generates interactive graphics inside R by using wrappers to qt libraries (https://www.qt.io), and theoretically can be used for really large data sets. The video at https://vimeo.com/137042766 shows cranvas being used to explore a clustering of statistics department graduate programs based on the 2012 National Research Council ranking data. Ggvis provides web graphics 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/.

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

3.2 Statistics, Infovis, Biovis

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 exploratory data analysis using graphics and inferential statistics. Hofert and Mächler (2014) provides a graphical goodness-to-fit test for dependence models in higher dimensions, and implements these 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. The R package, spatstat, is

extended to incorporate these diagnostics. Rainbow plots and bagplots (Hyndman and Shang, 2010) were developed for viewing a large number of functional data, smooth curves or surfaces, and has an accompanying R package. Van Long and 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 2D or 3D layout using an optimized star coordinates layout. It allows to explore the distribution of clusters interactively and helps the user to understand the relationship between the clusters and the original data space. Newell et al. (2013) describe methods for finding clusters in populations based on genetic markers, 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 really important 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 if what we see is really there. As a statistician, strictly adhering to the rigid assumptions required by classical hypothesis testing, runs the risk of non-discovery, failing 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) bridges statistics and infovis providing a formal protocol for determining if one type of display better communicates information than another.

In the infovis community there have been many developments for working with large data sets. OnSet (Sadana et al., 2014) is a technique to visualize large-scale binary set data. One observation is represented in one layer of plot and one pixel represents one elements in the plot. For the elements in this observation, pixels are highlighted. With these type of representation, it is easy to compare data sets using boolean operations. Lins et al. (2013) makes it easy to slice and dice large spatiotemporal data sets for viewing in various ways.

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, closely linked with analysis tools available in R from the bioconductor suite.

4 Synopsis

Data visualization and statistical graphics are pursuits which sit on the edge of several disciplines. This can be unnerving when trying to organize thoughts and ideas, provoking uncertainty as to what "box" should the work reside in? Gelman and Unwin (2013), a featured article in the Journal of Computational and Graphical Statistics, with invited

commentary, wrestles with the relative roles and purposes of statistical graphics and information visualization. These are just two of the terms that describe overlapping topics. It would be useful for scientists to have a taxonomy of nomenclature related to data visualization. Gelman and Unwin (2013) falls short of a full discussion on the relative emphases of different pursuits, and simplistically reduces the two domains into a dichotomy of appearance vs functionality. The invited commentaries do a nice job of defusing the incense provoked by the shallow interpretation.

There are many different names associated with data visualization: statistical graphics, information visualization, visual analytics, infographics. (Broadening the list to incorporate pursuits in scientific visualization could obfuscate the 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 does have a primary focus of visualization of data associated with understanding variability. It can be elegant, interactive, and beautifully crafted. A lot of the time, though, it is ephemeral, designed to support the activity of exploring data. Information visualization is a vast area of endeavor to represent data abstractly to reinforce human cognition. There is a stronger association with cognitive perception. It also was driven by the database community, with an inclination to handle large amounts of data. Infographics are designed for mass consumption, and as such typically utilizes very simplistic data representations. Visual analytics arose from a need to support decision systems, provide dash boards for the company executives, that can assist of 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, the role of data changes from passive support following pre-determined conjectures, to an active primal position where there isn't necessarily 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 these concerns, and an effort to broadly build knowledge about randomness among users and developers of data graphics. Understanding that making a hundred plots might result in a few that exhibit interesting patterns (fishing exhibition) by chance, 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 is either seen or it isn't (data snooping). We cannot revert to pretending that there was a probability of observing the pattern, much the same as 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 didn't. Once we have made a plot, we can't 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 (1986) spoke many years ago:

"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."

are germane for big data analysis. These issues will need to be faced for successful big data visualization.

References

Asimov, D. (1985), "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Scientific and Statistical Computing*, 6, 128–143.

Avila, D., Cottam, J., Dodia, K., Doig, C., Paprocki, M., Shi, H., Van de Ven, B., and Wang, P. (2015), "`Bokeh`: Python Interactive Visualization Library," `http://bokeh.pydata.org/en/latest/index.html`.

Baddeley, A., Chang, Y.-M., Song, Y., and Turner, R. (2013), "Residual Diagnostics for Covariate Effects in Spatial Point Process Models," *Journal of Computational and Graphical Statistics*, 22, 886–905.

Bostock, M., Ogievetsky, V., and Heer, J. (2011), "D3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, 17, 2301–2309.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., and Wickham, H. (2009), "Statistical Inference for Exploratory Data Analysis and Model Diagnostics," *Royal Society Philosophical Transactions A*, 367, 4361–4383.

Carr, D. B. (1995), "Using Gray in Plots," *Statistical Computing & Statistical Graphics Newsletter*, 5, 11–14.

Carr, D. B., Lewin-Koh, N., and Maechler, M. (2014), "`hexbin`: Hexagonal Binning Routines," `http://cran.r-project.org/web/packages/hexbin/index.html`, maintained by Pedesma, E.

Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. (1987), "Scatterplot Matrix Techniques for Large N," *Journal of the American Statistical Association*, 82, 424–436.

Carr, D. B. and Nusser, S. (1996), "Converting Tables To Plots: A Challenge From Iowa State," *Statistical Computing & Statistical Graphics Newsletter*, 6, 11–18.

Chang, W. and Wickham, H. (2015), "`ggvis`: Interactive Web Graphics with R," `http://ggvis.rstudio.com/`.

Chelaru, F., Smith, L., Goldstein, N., and Bravo, H. C. (2014), "Epivizr: Interactive Visual Analytics for Functional Genomics Data," *Nature Methods*, 11, 938–940.

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local Regression Models," in *Statistical Models in S*, eds. Chambers, J. M. and Hastie, T., New York: Chapman and Hall, pp. 309–376.

Crowder, M. J. and Hand, D. J. (1990), *Analysis of Repeated Measures*, London: Chapman and Hall.

Dang, T. N. and Wilkinson, L. (2014), "ScagExplorer: Exploring Scatterplots by Their Scagnostics," in *Pacific Visualization Symposium (PacificVis), 2014 IEEE*, IEEE, pp. 73–80.

De Jonge, E. and Tennekes, M. (2013), "`tabplotd3`: Interactive Inspection of Large Data," `http://cran.r-project.org/web/packages/tabplotd3/index.html`, maintained by De Jonge, E.

Dey, T., Phillips, D. J., and Steele, P. (2011), "A Graphical Tool to Visualize Predicted Minimum Delay Flights," *Journal of Computational and Graphical Statistics*, 20, 294–297.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2013), "The Generalized Pairs Plot," *Journal of Computational and Graphical Statistics*, 22, 79–91.

Feinberg, J. (2010), "Wordle," in *Beautiful Visualization: Looking at Data through the Eyes of Experts*, eds. Steele, J. and Iliinsky, N., Sebastopol, CA: O'Reilly Media, pp. 37–58.

Friedman, J. H. and Stuetzle, W. (2002), "John W. Tukey's Work on Interactive Graphics," *The Annals of Statistics*, 30, 1629–1639.

Friendly, M. (2014), "Comment on The Generalized Pairs Plot," *Journal of Computational and Graphical Statistics*, 23, 290–291.

Gelman, A. and Unwin, A. (2013), "Infovis and Statistical Graphics: Different Goals, Different Looks," *Journal of Computational and Graphical Statistics*, 22, 2–28.

Guha, P. K., Hafen, R. P., and Cleveland, W. S. (2009), "Visualization Databases for the Analysis of Large Complex Datasets," in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, eds. van Dyk, D. and Welling, M., vol. 5, pp. 193–200.

Hafen, R. P. and Cleveland, W. S. (2015), "Tessera," http://tessera.io/.

Hand, D. J., Blunt, G., Kelly, M. G., and Adams, N. M. (2000), "Data Mining for Fun and Profit," *Statistical Science*, 15, 111–131.

Hartigan, J. A. (1975), "Printer Graphics for Clustering," *Journal of Statistical Computation and Simulation*, 4, 187–213.

Hocking, T. D., VanderPlas, S., and Sievert, C. (2015), *animint: Interactive animations*, r package version 2015.06.16.

Hofert, M. and Mächler, M. (2014), "A Graphical Goodness-of-fit Test for Dependence Models in Higher Dimensions," *Journal of Computational and Graphical Statistics*, 23, 700–716.

Hofmann, H., Cook, D., Kielion, C., Schloerke, B., Hobbs, J., Loy, A., Mosley, L., Rockoff, D., Huang, Y., Wrolstad, D., and Yin, T. (2011), "Delayed, Canceled, on Time, Boarding Flying in the USA," *Journal of Computational and Graphical Statistics*, 20, 287–290.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012), "Graphical Tests for Power Comparison of Competing Designs," *IEEE Transactions on Visualization and Computer Graphics*, 18, 2441–2448.

Hofmann, H. and Vendettuoli, M. (2013a), "Common Angle Plots as Perception-True Visualizations of Categorical Associations," *IEEE Trans. Vis. Comput. Graph.*, 19, 2297–2305.

— (2013b), "`ggparallel`: Variations of Parallel Coordinate Plots for Categorical Data," `http://CRAN.R-project.org/package=ggparallel`, maintained by Hofmann, H.

Hurley, C. and Oldford, R. (2011a), "Eulerian tour algorithms for data visualization and the PairViz package," *Computational Statistics*, 26, 613–633.

— (2011b), "`PairViz`: Visualization using Eulerian tours and Hamiltonian decompositions," `http://cran.r-project.org/web/packages/PairViz/index.html`, maintained by Hurley, C.B.

Hyndman, R. J. and Shang, H. L. (2010), "Rainbow Plots, Bagplots and Boxplots for Functional Data," *Journal of Computational and Graphical Statistics*, 19, 29–45.

Inselberg, A. (1985), "The Plane with Parallel Coordinates," *The Visual Computer*, 1, 69–91.

Jockers, M. L. (2014), *Text Analysis with R for Students of Literature*, New York, NY: Springer.

Kaplan, A., Hare, E., Hofmann, H., and 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, J. T., and Scheidegger, C. (2013), "Nanocubes for Real-Time Exploration of Spatiotemporal Datasets," *IEEE Transactions on Visualization and Computer Graphics*, 19, 2456–2465.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of American Statistical Association*, 108:503, 942 – 956.

Mosley, L., Cook, D., Hofmann, H., Kielion, C., and Schloerke, B. (2010), "Monitoring the Election Visually," *Chance*, 23, `http://chance.amstat.org/files/2010/12/Visually.pdf`.

Moustafa, R. E., Hadia, A. S., and Symanzik, J. (2011), "Multi-Class Data Exploration Using Space Transformed Visualization Plots," *Journal of Computational and Graphical Statistics*, 20, 298–315.

Murrell, P. and Potter, S. (2015), *gridSVG: Export grid Graphics as SVG*, r package version 1.4-3.

Newell, M., Cook, D., Hofmann, H., and Jannink, J.-L. (2013), "An Algorithm for Deciding the Number of Clusters and Validation using Simulated Data with Application to Exploring Crop Population Structure," *Annals of Applied Statistics*, 7, 1898–1916, Supplementary material, including videos available http://projecteuclid.org/euclid.aoas/1387823303.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

RStudio (2015), "`Shiny`: A Web Application Framework for R," `http://shiny.rstudio.com/`.

Sadana, R., Major, T., Dove, A., and Stasko, J. (2014), "OnSet: A Visualization Technique for Large-scale Binary Set Data," *IEEE Transactions on Visualization and Computer Graphics*, 20, 1993–2002.

Schonlau, M. (2003), "Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots," `http://www.schonlau.net/publication/03jsm_hammockplot.pdf`.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010), "The Hadoop Distributed File System," in *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, Washington, DC, USA: IEEE Computer Society, MSST '10, pp. 1–10.

Sievert, C. and Shirley, K. E. (2014), "`LDAvis`: A Method for Visualizing and Interpreting Topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA, June 27, 2014.*, pp. 63–70.

Stolte, C., Chabot, C., and Hanrahan, P. (2003), "Tableau," `http://www.tableau.com/`.

Tennekes, M. and De Jonge, E. (2014), "`tabplot`: Tableplot, a Visualization of Large Datasets," `http://cran.r-project.org/web/packages/tabplot/index.html`, maintained by Tennekes, M.

Tierney, L. (1991), *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*, New York, NY: Wiley.

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

Tukey, J. W. (1986), "Sunset Salvo," *The American Statistician*, 40, 72–76.

Tukey, J. W. and Tukey, P. A. (1985), "Computer Graphics and Exploratory Data Analysis: An Introduction," in *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics '85*, Fairfax, VA: National Computer Graphics Association, vol. 3, pp. 773–785.

Urbanek, S. (2015), "`ioplots`: High-performance I/O Tools to Run Distributed R Jobs Seamlessly on Hadoop," `https://github.com/s-u/iotools`.

Urbanek, S. and Theus, M. (2003), "`iPlots`: High Interaction Graphics for R," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Van Long, T. and Linsen, L. (2011), "Visualizing High Density Clusters in Multidimensional Data using Optimized Star Coordinates," *Computational Statistics*, 26, 655–678.

Wattenberg, M. and Viégas, F. (2010), "Beautiful History: Visualizing Wikipedia," in *Beautiful Visualization: Looking at Data through the Eyes of Experts*, eds. Steele, J. and Iliinsky, N., Sebastopol, CA: O'Reilly Media, pp. 175–191.

Wattenberg, M., Viégas, F. B., and Hollenbach, K. (2007), "Visualizing Activity on Wikipedia with Chromograms," in *Human-Computer Interaction–INTERACT 2007*, Springer, pp. 272–287.

Wegman, E. (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of American Statistics Association*, 85, 664–675.

Wickham, C. (2011a), "A Tale of Two Airports: Exploring Flight Traffic at SFO and OAK," *Journal of Computational and Graphical Statistics*, 20, 291–293.

Wickham, H. (2011b), "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software*, 40, 1–29.

— (2013), "Bin-summarise-smooth: A Framework for Visualising Large Data," Tech. rep., had.co.nz.

Wickham, H. and Chang, W. (2014), "ggplot2: An Implementation of the Grammar of Graphics," http://cran.r-project.org/web/packages/ggplot2/index.html, maintained by Wickham, H.

Wickham, H., Chowdhury, N. R., and Cook, D. (2014), "nullabor: Tools for Graphical Inference," http://cran.r-project.org/web/packages/nullabor/index.html, maintained by Cook, D.

Wickham, H., Cook, D., and Hofmann, H. (2015a), "Visualizing statistical models: Removing the blindfold," *Statistical Analysis and Data Mining*, 8, 203–225.

Wickham, H., Francois, R., and Rstudio (2015b), "dplyr: A Grammar of Data Manipulation," http://cran.r-project.org/web/packages/dplyr/index.html, maintained by Wickham, H.

Wickham, H., Lawrence, M., Lang, D. T., and Swayne, D. F. (2008), "An Introduction to RGGobi," *R-news*, 8, 37.

Wickham, H., Poole, D., and Swayne, D. F. (2009), "Bay Area Blues: the Effect of the Housing Crisis," in *Beautiful Data: The Stories Behind Elegant Data Solutions*, eds. Segaran, T. and Hammerbacher, J., Sebastopol, CA: O'Reilly Media, pp. 303–319.

Wicklin, R. (2011), "Visualizing Airline Delays and Cancelations," *Journal of Computational and Graphical Statistics*, 20, 284–286.

Wilkinson, L., Anand, A., and Grossman, R. L. (2005), "Graph-Theoretic Scagnostics." in *INFOVIS*, vol. 5, p. 21.

Wilkinson, L., Anand, A., and Urbanek, S. (2012), "scagnostics: Compute Scagnostics - Scatterplot Diagnostics," http://cran.r-project.org/web/packages/scagnostics/index.html, maintained by Urbanek, S.

Xie, Y., Hofmann, H., and Cheng, X. (2014), "Reactive Programming for Interactive Graphics," *Statistical Science*, 29, 201–213.