Improving Visual Statistics

by

Michael A. Correll

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 08/04/2015

The dissertation is approved by the following members of the Final Oral Committee:

Michael Gleicher, Professor, Computer Sciences
Charles Franklin, Professor, Law and Public Policy (Marquette University)
Bilge Mutlu, Associate Professor, Computer Sciences
Steven Franconeri, Associate Professor, Psychology (Northwestern University)
Robert Roth, Assistant Professor, Geography
Kevin Ponto, Assistant Professor, Design Studies

To those people for whom these sorts of things are usually dedicated.

ACKNOWLEDGMENTS

Personally I never care for fiction or storybooks. What I like to read about are facts and statistics of any kind. If they are only facts about the raising of radishes, they interest me. Just now, for instance, before you came in, I was reading an article about "Mathematics." Perfectly pure mathematics. My own knowledge of mathematics stops at "twelve times twelve," but I enjoyed that article immensely. I didn't understand a word of it; but facts, or what a man believes to be facts, are always delightful. That mathematical fellow believed in his facts. So do I. Get your facts first, and then you can distort 'em as much as you please.

— Mark Twain (quoted in Rudyard Kipling's *From Sea to Sea*)

I would like to thank my committee for providing advice, sounding boards, and encouragement, and generally putting up with me throughout my graduate career. In particular I would like to thank my advisor Michael Gleicher for putting up with me the most.

I would also like to thank collaborators in other fields who have allowed me to gain new perspectives on my work while being patient with me as I deployed new systems or learned new methodologies: Dave O'Connor and his lab in virology, Steve Franconeri and his lab in psychology, and Mike Witmore and the rest of the literature scholars working on the Visualizing English Print project (both before and after it had that name). Specifically for the work presented in this document I would like to thank Michael Gleicher, who collaborated on chapters 2-6 inclusive, Steve Franconeri who collaborated on chapters 2 and 3, Danielle Szafir who collaborated on chapter 2, Mike Witmore who collaborated on chapter 5, and Alper Sarikaya, Adam Bailey, and David O'Connor who collaborated on chapter 6.

I would also like to thank my fellow members of the graphics lab who have been my co-authors, colleagues, friends, and sanity checkers throughout my time at Wisconsin. Finally I thank my family for encouraging me throughout this whole grad school process: in return I will not force them to read this document. Oh, I also want to thank Mary; she's pretty cool.

This work in this dissertation was funded by NSF grant IIS-1162037, NIH award R01 AI077376, and a Mellon Foundation grant.

CONTENTS

LIST OF FIGURES

IMPROVING VISUAL STATISTICS

Michael A. Correll

Under the supervision of Professor Michael Gleicher
At the University of Wisconsin-Madison

In this work, I explore *visual statistics* — the interaction between *statistical* and *visual* techniques for dealing with data. I present experiments and theoretical work showing that, with careful design, humans have a robust capability to extract and make use of statistical information from visualizations. I also explore the limits of these abilities, and how, in order to support different styles of argumentation, and to overcome certain perceptual and cognitive biases, designers may need to radically alter visualizations. Lastly, I present deployed systems which are mindful of the capabilities and limitations of visual statistics in order to support thoughtful and complicated data analysis of statistical patterns.

Michael Gleicher

ABSTRACT

Statistics as a field provides powerful methods for dealing with data. We can summarize data, build models, and make inferences. However, statistics can be complex, esoteric, and rely on many assumptions — aspects that make it difficult to communicate statistical information. Visualization offers another set of powerful techniques for data, relying on the human perceptual system to summarize and structure visual information. I believe that these two approaches can operate in harmony, and that particularly we can trust viewers to perform *visual* analysis that mirrors *statistical* analysis.

In this work, I explore *visual statistics* — the interaction between *statistical* and *visual* techniques for dealing with data. I present experiments and theoretical work showing that, with careful design, humans have a robust capability to extract and make use of statistical information from visualizations. I also explore the limits of these abilities, and how, in order to support different styles of argumentation, and to overcome certain perceptual and cognitive biases, designers may need to radically alter visualizations. Lastly, I present deployed systems which are mindful of the capabilities and limitations of visual statistics in order to support thoughtful and complicated data analysis of statistical patterns.

1 INTRODUCTION

When discussing the ubiquity of data, there is a temptation to treat data as though they are some naturally occurring process, that data trails behind us like the wake of a ship. Data are not self-arising objective facts about the world; they are collected, curated, and circulated by (limited, biased, and flawed) human beings, and these processes are teleological: data are always *for* something (even if that "something" is idle speculation). Data are used to explore, to persuade, and to decide. We must present data in different ways to meet these different goals, and be mindful of the choices we make with our data.

The unavoidable presence of human beings and human thinking in the collection and use of data means that biases and persuasive elements are not uncommon edge cases to be dealt with as they arise, but an integral and unavoidable aspect of the use of data. Our task as designers and scientists in data-driven fields is to both usefully and responsibly present data in a way that is mindful of the needs and capabilities of our intended audience.

Statistics offers a way forward for structuring data for use by humans. Techniques from statistics can summarize the data in useful ways: **descriptive statistics** such as the mean, the mode, variance, &c., can condense enormous amounts of data into just a few numbers. **Inferential statistics** can, when employed correctly, offer a great deal of information about the behavior of datasets, providing both models and comparisons which extrapolate beyond the known and quantify the uncertainty in decision-making and prediction.

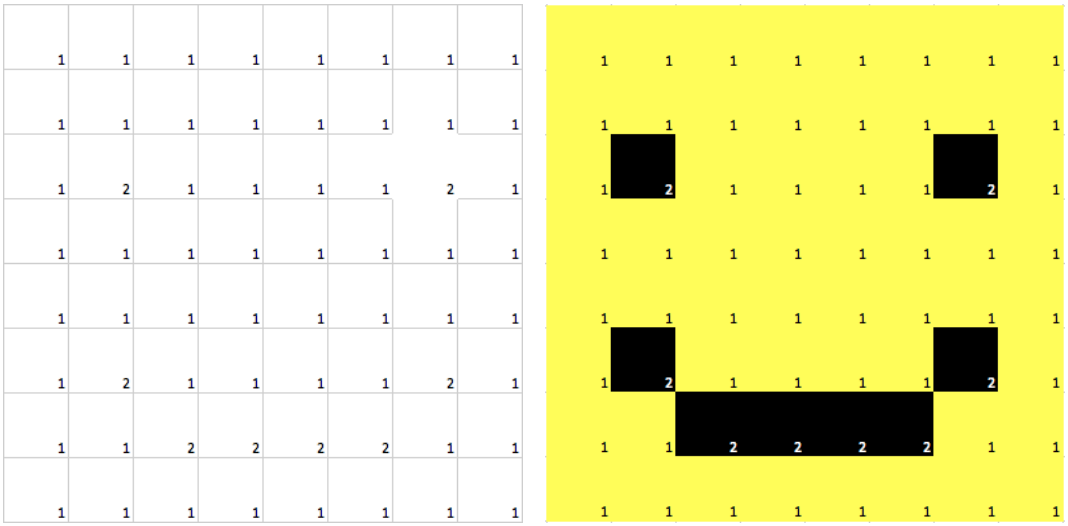
Unfortunately, the use and interpretation of statistics is difficult, often requiring years of study. The mis-use of statistics to deceive is both simple and commonplace [Huff 1993]. Statistical models can be overly complex, overly simplistic, or rely on esoteric knowledge to meaningfully interpret. Statistics are often employed without careful thought, or to an unprepared audience. The naïve presentation of statistical models can hide the flaws and assumptions present in statistical work and therefore mislead and distort. To further complicate matters, human intuitions about risk and uncertainty are frequently at odds with statistical expectations: there are a number of biases and apparent paradoxes in how humans reason about uncertainty [Tversky and Kahneman 1974]. While sometimes these biases create "irrational" behavior

(humans choose outcomes which do not maximize the expected utility in uncertain situations), in other cases differences in risk tolerance, costs of different errors, or additional domain knowledge means that there might be advantages to a human analyst "overruling" relevant statistical measures such as expected value or p-value.

Visualization (and specifically information visualization) is another approach to structuring data for human consumption and use. I refer to **information visualization** as any technique (usually computer-assisted) which creates a visual artifact (a picture, an animation, an object in the real world) that is driven by some set of data. This definition is intentionally broad, including things like maps and artistic visualization as well as more traditional visualizations like charts and graphs. While often visualization is presented as affording a difference in *degree* in data-driven human thinking (hence definitions of visualization such as Card's where we "use visual computing to amplify human cognition with abstract information"[Card 2002]), I believe it is also a difference in *kind* — the visual system affords capabilities like gestalt grouping and pattern recognition that are vastly more difficult to perform when given non-visual presentations of data (as in Fig. 1.1).

Much of the potential of visualization is based off the powerful capabilities of the human visual system. As part of the perceptual machinery connected with how visual information is gathered and processed, human beings are capable of perceptual tasks that are roughly analogous to statistical operations like averaging, extrapolation, and outlier detection. Visual tasks that seem elementary to us (as in Fig. 1.1) might require complex statistical modeling or even machine learning to capture in an automatic way.

However, as with statistics, we must take care in how information visualizations are designed and used. The mere *presence* of a chart can sway human perception [Tal and Wansink 2014], and there are several techniques for misleading viewers which have consistent and measurable impact on how the data are perceived [Pandey et al. 2015]. Just as with statistics, it is easy to create visualizations that mislead the viewer [Rogowitz et al. 1996], skewing the perception of data.



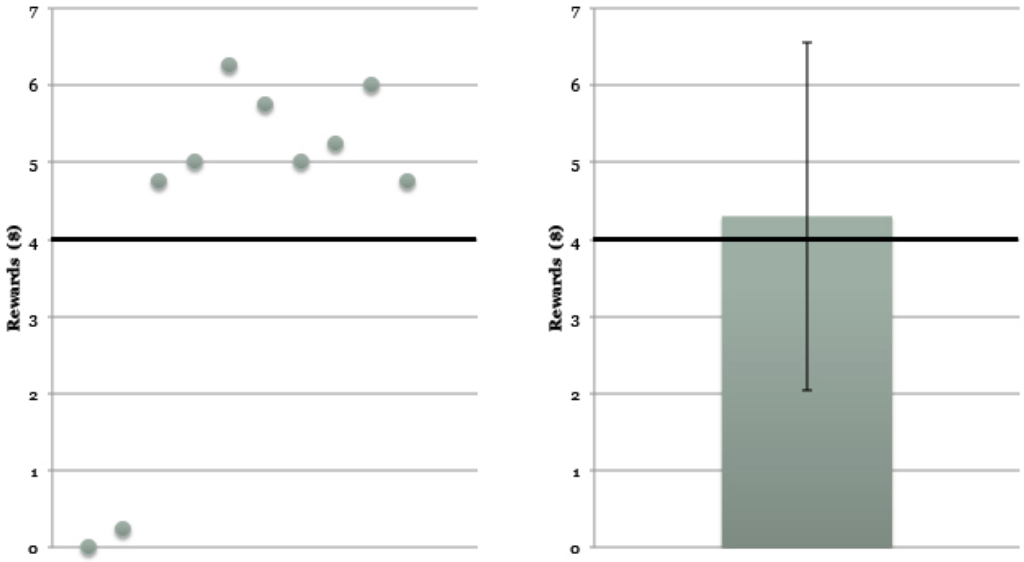
(a) A table of numbers. What patterns are visible in this data?

(b) The same table but with color, highlighting a smiley 2's.

Figure 1.1: Supplementing the presentation of data with visual elements allows viewers to use many sophisticated components of the human perceptual system. In order to find the "smiley face" hidden in this data table, the viewer *groups* elements of common color together, and recognizes common *patterns* from the resulting groups. Humans excel at this type of grouping and pattern recognition: if we can characterize the class of *statistical tasks* which align with these *perceptual tasks*, then we can support statistically savvy decision-making from viewers who may not have a background in statistics.

1.1 Visual Statistics

Both information visualization and statistics offer different approaches to dealing with data, with their own strengths and weaknesses. It is my contention that we can employ a hybrid, complementary approach that borrows good elements from both approaches. Central to this approach is what I call "**visual statistics**." This is a class of techniques for visualization which is oriented towards conveying statistical summaries (such as mean or mode) or statistical models and inferences (error, p-value, goodness of fit,&c.). In some cases, the statistical values of interest are explicitly displayed (either in conjunction with, or in replacement of the data values). In other cases, the data are presented in such a way that the viewer can compute or estimate these values by themselves (see Chapter 2 for more information). In short: **visual statistics is the visual presentation of data in a way that is mindful of the**



(a) Scatterplot of bet outcomes.

(b) Bar chart of bet outcomes with error bars.

Figure 1.2: Should you take this $4 bet? The *expected value* of the bet is greater than $4, so under certain definitions taking the bet is the rational choice. I believe (and work in this thesis confirms) that the way designers visually present information can change how we reason about data, especially in cases of uncertainty. For instance, viewers looking at the scatterplot might see that most outcomes have rewards at least a dollar or two great than the initial $4 bet, and so might consider the two low-paying outlier rewards an acceptable risk. The bar chart hides the full range of outcomes but makes it clear that, on average, the payout is only a few cents higher than $4, which might be an unacceptably low reward.

statistical quantities and tasks required by the viewer. This term is idiosyncratic to my work, a necessary invention given the overly wide scope of "visualization" generally, but the necessity of referring to designs that occur in this space.

Creating visualizations with an eye towards not just how individual values are transmitted, but relevant statistical quantities, is common. Trend lines, error bars, and explicit marks for aggregate statistics of interest (for instance the quartiles of a box plot, or the numerous higher level moments available in boxplot extensions such as Potter et al. [2010]) are all examples of visual statistics. However, not all visualizations using visual statistics are created equal (see Fig. 1.2). The decision of what statistical information to show or hide, and how to visually encode the

information we choose to present, create different patterns of decision-making and performance. For instance, presenting data as bar charts encourages comparison of individual values, whereas the same data in a line graph encourages comparison of trends over time [Zacks and Tversky 1999].

Good visual statistics is a process of balancing and counteracting the strengths and weaknesses of statistical and visual perspectives of the data. For instance, we might need to simplify the communication of a particular statistical model or concept when presenting this information to a general audience, but take care not to oversimplify or mislead the viewer into seeing patterns that are not really there. Among many other potential factors, we need to determine the degree of *trust* we have in our viewers (and in our statistics; in many cases the knowledge of the viewer may "override" the statistical model, no matter how complicated). How much of the statistical complexity should we show or hide? How much can we trust viewers to make the "right" decisions from visualizations? Figure 1.3 shows an example of this continuum: presenting no additional information beyond the raw data values requires that we trust the viewer to perform rather complicated statistical tasks from visual information alone (averaging of values, performing a statistical inference that might take into account variability and stand size). Presenting just the results of a statistical inference might place too little faith in the viewer: there is more involved in human decision-making than a yes/no decision from a t-test.

Visual statistics as they are used in practice have some deficiencies in their design and evaluation. Although there has been extensive empirical investigation of how different visualizations of data perform for extracting and comparing individual data values (such as the seminal work of Cleveland and McGill [1984]), exploring how well viewers can perform tasks related to higher level statistics is understudied (*q.v.* chapter 3). Popular encodings for displaying statistical values are used without consideration of whether or not they are successful at conveying data in a way that aligns with statistical expectations (*q.v.* chapter 4). Design maxims meant to improve the clarity of presenting individual data values are inappropriately applied to situations where communication and persuasion of statistics might require non-standard design choices. My intervention in the space of visual statistics is a combination of empirical and theoretical work designed to address these deficiencies, exploring how humans with or without statistical training extract and compare statistical

values from visualizations, and how to design visualizations for statistical tasks.

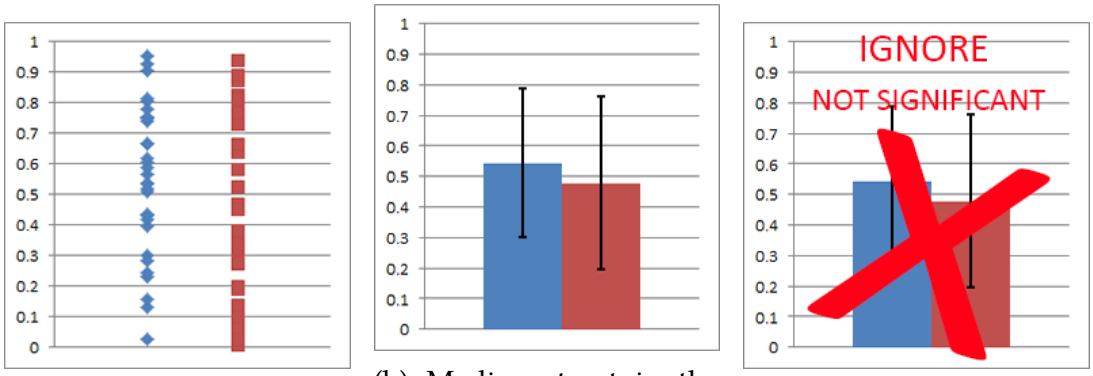
1.2 Thesis Statement

It is my thesis that **we can improve visual statistics by making informed choices about what aggregate information to display, and by creating encodings that are mindful of how humans reason and argue about statistics and uncertainty**.

I believe that designers frequently underestimate the ability of viewers of visualizations to build up complex and accurate conceptions of statistical values and models. Work from perceptual psychology indicates that the visual summaries the brain creates as it parses visual information can be used by viewers to estimate and compare complicated aggregate values (chapter 2 describes this prior work in more detail). Part of the work of this document is *extending* these results to the realm of visualizations, showing that not only does this ability to estimate statistical quantities from visual stimuli extend to complicated displays of data, but that these estimations affect how analysis of data is performed, changing the decisions and confidence of viewers.

Conversely, designers frequently underestimate the impact of perceptual and cognitive biases on how data are interpreted. Many common visualization designs for showing information are not mindful of how human beings use and reason about data. Naïvely presenting "just the data" can result in human inferences that are systematically different from mathematical expectation. Detecting and measuring these biases, and creating designs which avoid or correct for them, is a major component of improving visual statistics, and justifying our relatively high trust in viewers to make statistically justified decisions.

If we can *quantify* how well human beings can reason about statistics from visualizations we can *advocate* for existing (or novel) designs that support this sort of reasoning, and *design* complex systems which employ visual statistics in a grounded way. My thesis work thus contains empirical studies with novel methodologies that allow us to bridge perceptual psychology, visualization, and decision theory, novel techniques and designs for the presentation of statistical information to viewers, and novel systems which build on these foundational results to provide real utility to underserved data domains.



(a) High trust in the viewer. (b) Medium trust in the viewer. (c) Low trust in the viewer.

Figure 1.3: Three examples of visual statistics for the task of inferring whether the means of two groups are different, showing different levels of trust in the statistical sophistication of the viewer. Fig. 1.3a leaves to the viewer the task of averaging the two groups and inferring margins of error, which allows the viewer to bring their own semantic understanding and judgment to the task (for instance they could discard points as well-known outliers), but presumes that the viewer is able to make sophisticated calculations from visual presentation alone. Fig. 1.3c is an extreme case – the viewer is not trusted to interpret means or margins of error correctly, and so they are obscured in order to present the "correct" statistical inference. In practice it is useful to borrow approaches from both extremes – trust the viewer to build up coherent pictures of the data, but also to be mindful that many common derived statistical values (like margin of error, or expected value) are commonly misinterpreted even by experts.

1.3 Thesis Outline and Contributions

The defense of this thesis is organized into three categories of contributions:

- **Theories** about how to create visualizations that use statistics to persuade a wide variety of audiences, with special attention to visualizing both uncertain information, and information in the aggregate.
- **Experiments** that detail the capabilities of general audiences at extracting statistical information from a wide variety of designs, with an eye towards examining designs that are unconventional, but have properties that suggest their utility for conveying high level statistical quantities.
- **Systems** that apply knowledge gained through experimentation and theory-building to assist domain experts in a variety of fields. This systems are

specifically focused on domains where we have to convey potentially complicated statistical information to an audience that has an idiosyncratic way of using and arguing about data.

This document is organized around each of the three contributions presented above. The first chapter presents the key theoretical concept of visual aggregation, the mechanism by which we can create visualizations that communicate aggregate statistical quantities without explicitly encoding the values directly. The next two chapters present crowd-sourced empirical work examining visual aggregation in detail, including cases where it was necessary to "de-bias" the viewer away from incorrect interpretations of statistical data. The next two chapters present specific systems that operate in data domains requiring the sophisticated communications of statistical information. I conclude with lessons learned from the process, including situations where standard design maxims are insufficient to capture important components of making successful visual statistics.

Theoretical Analysis of Visual Statistics

A key contribution of this thesis work is that it is possible to communicate statistical information even to those who lack expertise in statistics, and furthermore that designers may not need to explicitly encode statistical information in visualizations, but trust the viewer to extract these values themselves (as per Figure 1.3a). This assertion, and the theoretical underpinnings behind it, requires both explanation and justification.

To this end, chapter 2 describes **visual aggregation** — the capacity of humans to extract higher order statistics from low-level visualizations of data by exploiting the human visual system. In this chapter I explain situations where we can rely on visual aggregation for comparison of important statistical quantities, and argue for the general utility of visual aggregation in a wide variety of scenarios. I also address some pitfalls of this process, and how designers can intervene to support visual aggregation in their systems.

Crowd-sourced Evaluation of Visual Statistics

I present the results of a number of **crowd-sourced human subjects experiments** that both *explore* the capabilities of human beings to extract statistical information from common visualizations, as well as *validate* alternative designs meant to improve these capabilities. These experiments employ methodologies that allow us to apply results from perceptual psychology to the realm of information visualizations. We provide evidence that human beings are able to perform rather sophisticated statistical tasks from real-world displays of data (which are often larger, messier, and noisier than stimuli from psychology).

Chapter 3 presents a series of experiments that explores visual aggregation over a wide range of statistically-motivated tasks and encodings for time series data. I describe **color weaving**, a technique meant to support aggregate judgments in timeseries at the expense of lower-level point-based tasks. I also discuss potential additions to the color weaving technique meant to address the deficiencies of the basic weaving approach.

Chapter 4 presents a series of experiments showing that a standard encoding for presenting uncertainty in measurements, bar charts with error bars, have properties that make them unsuitable for judgment tasks involving uncertainty. I present and evaluate alternatives to bar charts with error bars, and also present work showing that in some cases we can rely on visual aggregation of unadorned data for these sorts of comparison tasks.

Systems Employing Visual Statistics

I present **systems** which use knowledge gained in our experiments to assist domain experts in a wide variety of fields solve concrete data analysis problems, from helping to understand Shakespeare to exploring natural immunities to AIDS. These systems have been used to advance scholarship in their relevant fields as well as shed light on new ways of presenting detailed statistical information to non-statistical audiences.

Chapter 5 describes **CorpusSeparator and TextViewer**, a tool suite for the analysis of word usage in text corpora for scholars in the digital humanities. I describe the design process, present several use cases, and show how this system allows

literary scholars to make decisions which are both grounded in statistical models but also in tune with their methods of persuasion and argumentation.

Chapter 6 describes the **LayerCake** tool for the visualization of viral population dynamics in next generation sequencing data. I describe how our work on the affordances of visual aggregation in time series data led to both the general design of the tool as well as specific choices of visual encoding.

2 VISUAL AGGREGATION

In this work I draw a distinction between *data* and *statistics*. Statistical operations are performed on data, often to summarize or make inferences about values. Designs that clearly present the data may not be sufficient to clearly present relevant statistics. Statistics may be complex, counterintuitive, or esoteric. Presenting them in a useful way is not simply a matter of e.g., making a bar chart of "skew," "kurtosis," or "p-value" and assuming that such visualizations will successfully communicate the relevant statistical inferences to the viewer. Designers have limited resources to communicate statistical values in addition to data. We are limited in the visual space available to us as well as the visual complexity of our designs. Calculating every potentially interesting model, summarization, or inference requires computational resources; communicating the meaning of each these calculations, and the potentially subtle differences between them, requires pedagogical resources. If we can rely on viewers *themselves* to extract statistical quantities of interest from visualizations, then we can intervene only when necessary.

A term of art used throughout this work is **visual aggregation**. I define "visual aggregation" as the capacity of humans to extract higher order *statistics* (such as mean, variance, or even more complicated statistics such as correlation or statistical inferences) from *visualizations per se*. A person estimating the mean of a point cloud is an example of visual aggregation; a person reading off a p-value from the legend of a chart is not. Tasks requiring aggregation may not necessarily be global ones (requiring the consideration of every data value), but they do occur at a level beyond the the consideration and comparison of individual data values.

Visual aggregation offers a way forward for allowing statistics to guide decision-making without requiring that the people using their data have deep statistical knowledge. We use the capabilities of the human visual system to perform complicated operations such as averaging, filtering, and selection as proxies for the equivalent statistical operations. However, we must be mindful both of the limitations of our perceptual capabilities, and also that certain design choices can help or hinder visual aggregation.

This chapter operates as a brief overview of the literature concerning visual aggregation: why results from perceptual psychology suggest that such sophisticated

statistical judgments are possible, and how these findings apply to the design of techniques for visual statistics.

2.1 A Theory of Perceptual Aggregation

Perceptual psychology offers empirically-based mechanistic insight into how users might perform aggregate judgment tasks in complex displays. Visualizations have leveraged perceptual studies to design encodings that ease cognitive demand in visual search tasks, such as making relevant data points readily "pop-out" to viewers [Healey and Enns 1998], and effectively managing clutter [Rosenholtz et al. 2011; 2005]. However, only recently has the visualization community begun to consider using findings from psychology to design visualizations that boost visual summarization and aggregate comprehension tasks [Albers et al. 2011, Elmqvist and Fekete 2010, Ramanarayanan et al. 2008].

The perception literature has illustrated various stages of the processes by which the visual system collects aggregate information in complex scenes such as visualizations. When initially viewing a scene, the visual system pre-attentively collects approximations of spatially organized features into a series of *ensemble statistics* – clusters of visual information not actively being attended to (e.g. pre-attentive images [Alvarez 2011, Ariely 2001, Greene and Oliva 2009, Treisman 2006, Treisman and Gelade 1980] and objects in the periphery [Freeman and Simoncelli 2011]) summarized according to their aggregate features. Prior work [Franconeri et al. 2009, Healey et al. 1996] has demonstrated that pre-attentively-based encoding choices can support rapid numerical estimation tasks within a scene. More recent work [Rosenholtz et al. 2012, Treisman 2006] suggests that being able to efficiently process the ensemble statistics generated from these encodings supports rapid aggregate comparisons of a scene and can even help guide visual search tasks.

Efficient summarization of unattended regions can be explained by a generalized theory of perceptual averaging. Perceptual averaging suggests that pre-attentive visual parameters, such as size [Ariely 2001, Chong and Treisman 2003], color [Balas et al. 2009, Maule and Franklin 2015, Treisman and Gelade 1980], or orientation [Alvarez and Oliva 2009, Choo et al. 2012, Parkes et al. 2001], can be averaged over a spatial range efficiently by the mechanisms that generate ensemble statistics of

unattended features. This process can be roughly mimicked by perceiving a display while squinting one's eyes – this mimics a spatial blurring of the image (analogous, but different processes occur to "blur" other visual properties such as orientation and size) : the resulting blur is effectively done in early phases of the visual system, where ensemble statistics are collected in parallel across varying levels of spatial frequencies in the scene [Hughes et al. 1996]. These multi-scale representations allow viewers to "see" averages directly, however certain presentations of this data can bias viewers' estimates [Izard and Dehaene 2008, Michael et al. 2014, Price et al. 2014].

The positional encodings used in line graphs may not permit rapid summary phenomena as well as color and other pre-attentive encodings. As a result, line graphs may not be averaged as efficiently as other displays, such as colorfields. Aggregate judgments in line graphs rely on summarizing and comparing different complex shapes generated by the line; however this process happens in higher-level visual areas with far less efficiency than early visual processing mechanisms [Wolfe and Bennett 1997]. While it is possible that comparing the average heights of regions of a line graph is in some sense analogous to comparing the average height of bars in a bar chart (in which case the visual system could make use of existing mechanisms for comparing average size of groups [Ariely 2001, Chong and Treisman 2003]), performing judgments of this sort would require that the visual system efficiently and flexibly segment subregions of a line graph, potentially at multiple scales. Perceptual segmentation mechanisms, by contrast, are known to be generally inflexible [Franconeri et al. 2009, Singh and Hoffman 1997]. Breaking local structure within the line graph by permutation or scatterplot representations may aid segmentation mechanisms for aggregating subregions to some extent by breaking local shape continuity, but it is unclear if this is sufficient to support rapid perceptual averaging tasks. We hypothesize that, by instead exchanging the fidelity of positional encodings for an encoding the better supports rapid summarization, we can design an encoding that better supports aggregate judgments within a series.

Applied Perceptual Aggregation

Consider an example statistical judgment from a visualization: comparing the relative means of two classes of points in a scatterplot (the experimental task from [Gleicher et al. 2013]). There are potentially two steps to this task: *selecting or segmenting* the scatterplot into the relevant data classes, and then performing some sort of *summarization or abstraction* on each class. In order to recommend visual aggregation for tasks such as these, we need evidence that both of these components can be performed rapidly and accurately.

The human visual system can quickly construct many types of abstractions from sets, including numerosity (see [Franconeri et al. 2009], for review), and averages over dimensions like size [Ariely 2001, Chong and Treisman 2005], orientation [Choo et al. 2012], motion direction [Levinthal and Franconeri 2011], spatial frequency [Alvarez and Oliva 2009], and perhaps even more complex properties like facial emotion and gender [Haberman and Whitney 2007] (but see [Marchant et al. 2013], for caveats). Observers can also average spatial position over small sets of objects [Alvarez and Oliva 2008], and can make saccades to the centerpoint of objects made up of large sets of dots that form a rough object contour [Melcher and Kowler 1999].

Efficient segmentation of sets has been studied using tasks such as visual search [Treisman 1985], texture boundary identification [Callaghan 1984; 1989], and number discrimination [Halberda et al. 2006]. Many cues have been shown to afford rapid and accurate set segmentation, including relative differences in hue, orientation, shape, and size [D'Zmura 1991, Treisman and Gormican 1988, Treisman and Gelade 1980], as well as some more complex visual properties such as lighting direction [Enns and Rensink 1990]. Some features are processed more efficiently than others [Treisman and Gormican 1988] - e.g., tasks involving selecting lines with atypical colors in a display are faster and less error prone than tasks involving selecting lines with atypical concavity. Haroz and Whitney [Haroz and Whitney 2012] also explore how the mechanisms of attention limit performance in various visual tasks.

The visual system can create abstractions (e.g., numerosity estimation, mean position, spatial envelope extraction) across the set of visual field locations that are currently selected by attention. Concretely, attentional selection is a relative amplification of visual information that meets certain criteria, such as being in a specific location (e.g., in the upper left of a display), or containing specific feature

values (e.g., red, left tilted, curved, or two-inches-tall). The possible criteria are constrained by the presence of existing feature maps [Franconeri et al. 2013, Serences and Boynton 2007] that index the presence or absence of that feature across the visual field, such that novel or arbitrary criteria are not available. Increasing the weight on one or more of these maps would lead to amplification of the visual information that is spatially correlated with the locations highlighted by that map. The perception literature explores many questions related to how this type of model operates, such as whether we can amplify multiple maps corresponding to values on the same dimension [Huang and Pashler 2007, Levinthal and Franconeri 2011], or whether new maps can be constructed with practice [Buonomano and Merzenich 1998].

Even if humans have the *capability* of creating aggregate feature maps, for the purposes of making statistically-based decisions from visualizations, we need to determine if they can create these feature maps *easily*, and can make good*use* of these feature maps. Most of the related work relies on briefly presented visual search displays, and shows that ignoring particularly salient objects can be difficult in some types of displays, suggesting a default mode where people automatically weight feature maps with unique spots of activation (e.g., [Belopolsky et al. 2007]). Other work shows that instruction or recent experience can alter the weights on these feature maps, leading to increased attentional control over what spatial locations or feature values contribute most to attentional selection (for review see [Egeth et al. 2010]) .

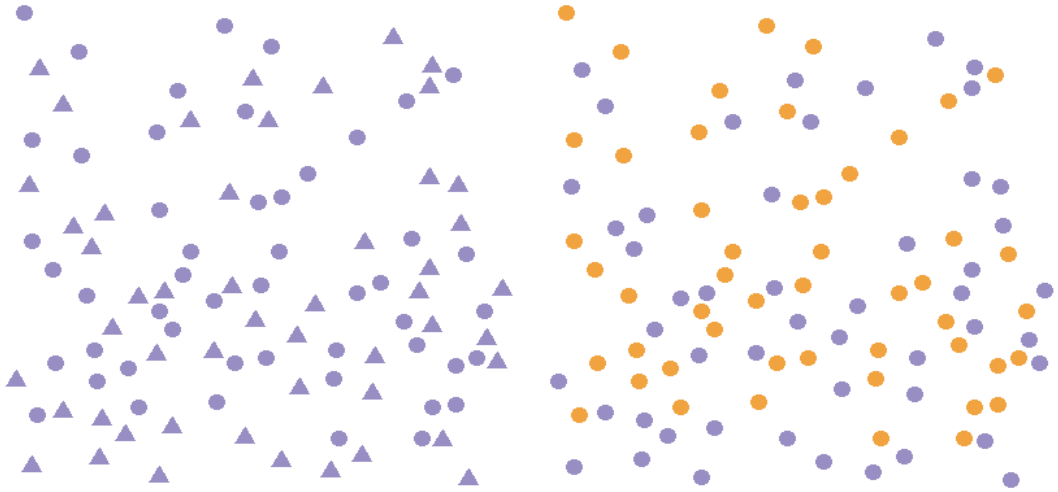
Researchers in visualization and graphics have investigated how known perceptual processes and features interact in more realistic displays. Healey, Booth & Enns varied features for encoding salmon migration data [Healey et al. 1996], finding that participants could successfully perform numerical estimation of items of a particular hue (with task-irrelevant orientation) and of a particular orientation (with task-irrelevant hue) quickly ($<$ 200 ms) and accurately. They also found no effect of interference from the task-irrelevant features, unlike previous studies ([Callaghan 1984; 1989]). Their displays were regular grids, and the values were contiguous regions, and therefore are quite different than scatterplots. More recently, others have investigated the best symbols for data encoding. Li et al. varied lightness and size of symbols in scatterplot displays from which participants performed several visual

analytic tasks [Li et al. 2010a;b]. Participant performance was used to model an optimal discriminability scale with equal perceptual separation between scatterplot symbol lightness and sizes.

2.2 Initial Visual Aggregation Experiments

In our own work, we have conducted several initial experiments that seek to validate that the principles encountered in perceptual psychology hold for the more complex and time-intensive considerations of visualization design. The focus of these experiments have been to verify that a) humans can make use of visual aggregation techniques to perform *useful* statistical decision-making tasks and b) inform design for improved performance at these visual aggregation tasks. Figures 2.1,2.2,2.3 are visual examples of where our empirical findings on visual aggregation tasks have suggested new or non-standard designs for information visualization; supporting different sorts of statistical judgments might require different visual presentations of information. For instance, the scatterplot work suggested that redundancy in encodings may not necessarily help viewers in making judgments about scattered data. Likewise for the tagged text work, the conflation of numeracy and visual area might require explicit de-biasing in order to support better judgments. Likewise (as further detailed in Ch. 3), how time series data is presented has a measurable impact on how easily viewers can extract statistical information.

These initial experiments employ methodology and theory-building from perceptual psychology to concretely inform visualization design. This interdisciplinary approach is reflected by other work in the field. For instance, in Fouriezos et al. [2008b], perceptual estimates of mean height in populations of bar charts are compared to traditional statistics of comparison. There has been also a great deal of work investigating how humans perform at sophisticated tasks like estimating correlation, as in Doherty et al. [2007],Rensink and Baldridge [2010], and most recently in Harrison et al. [2014], which explicitly connects estimation of statistical properties from visualizations with psychophysical "laws." Our work differs from these approaches in that it is not only descriptive of the human capacity to estimate and compare aggregate statistics but also prescriptive, weighing between different design choices, and proposing novel designs meant to support visual aggregation.



(a) Which group of points has the higher *average* value, triangles or circles?

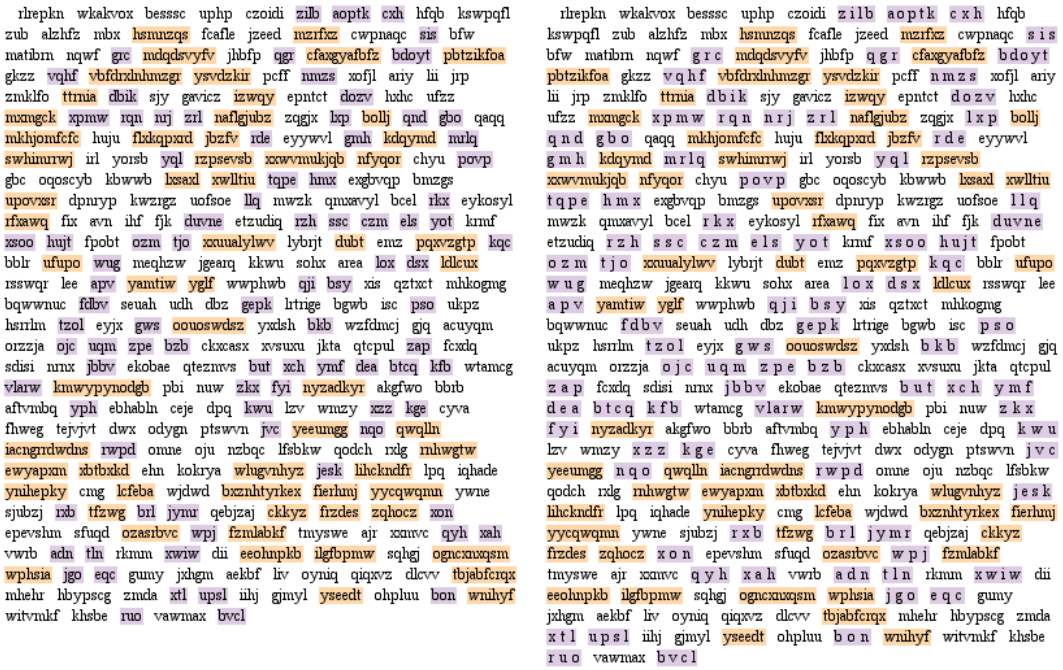
(b) Which group of points has the higher *average* value, orange or purple?

Figure 2.1: Designing for visual aggregation in scatterplots. When asked to compare mean values, *color cues* are associated with better performance than *shape cues*. By using stronger cues we can make visual aggregation easier [Gleicher et al. 2013].

2.3 Conclusion

I believe visual aggregation is a series of mechanisms that provides evidence that designers of visualizations can rely on viewers to make statistically sophisticated judgments without having in depth statistical training.

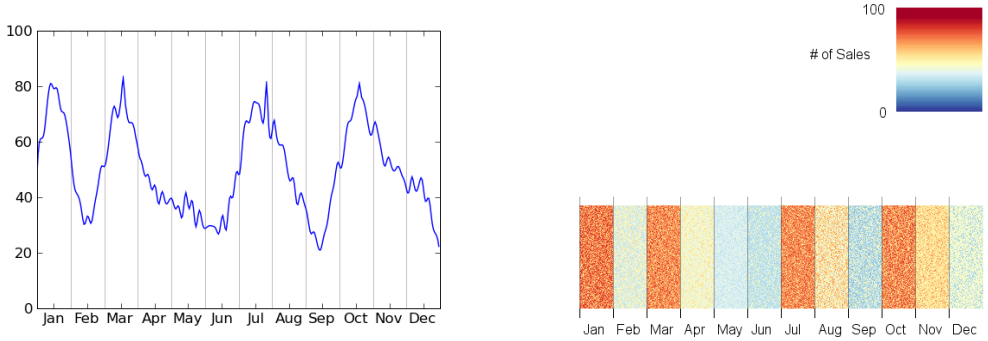
In the following chapters (3 and 4), we employ empirical methodology inspired from perceptual pyschology to analyze how aggregate statistics are perceived and compared in common visual displays, with the ultimate aim of determining to what extent we can rely on viewers to correctly estimate important statistical components of our visualizations, and where we as designers need to intervene.



(a) Standard tagged text.

(b) Tagged text with intraword tracking correction.

Figure 2.2: What percentage of these paragraphs of texts is orange (versus purple)? Here the task requiring visual aggregation is *numerosity estimation*. A known bias in human estimations (the conflation of numerosity and visual area) results in systematic errors in judgment. By correcting for this (in this case by artificially lengthening the purple words, making their visual area more in tune with their numerosity), human performance measurable and significantly improves [Correll et al. 2013].



(a) A standard linegraph display of time series data.

(b) A woven colorfield display of the same data. The pixel values within months are permuted generating noisy but distinct per-month color patterns.

Figure 2.3: Which month has the highest average value? This visual aggregation task requires mean comparison over twelve subunits (see Ch. 3 for more information). By using nonstandard displays of information, we can greatly increase performance at certain classes of visual aggregation tasks — participants were able to correctly find the month with the highest average in 90% of all woven colorfields, but only 65% of all standard linegraphs.

3 VISUAL AGGREGATION IN TIME SERIES DATA

Statistics is less concerned with individual values in datasets; rather, it calculates values based on data in the aggregate. Visualizations that seek to make statistical information readily available to the viewer must support either visual aggregation (as described in Chapter 2), or designers must intervene to explicitly encode relevant statistical values we cannot rely on the viewer to extract unaided. There are costs (in terms of visual complexity, clarity, and in decision-making for cases where the statistics do not tell "the whole story") of performing these kind of interventions. For instance, including a trend line based on a naïve linear model in a scatterplot might cause viewers to trust the model rather than their own predictive judgments. Therefore we need to determine whether visualization affords the estimation of statistical information.

The choice of how to visually encode data can strongly impact how accurately and efficiently viewers can extract aggregate information. Yet, foundational work in information visualization has focused on how viewers interpret single data points rather than aggregate statistics. If we wish to improve visual statistics, then we need a greater understanding on how well viewers can estimate and compare higher level statistics from visualizations. This provides us with an answer on how much we ought to trust viewers to perform their own forms of statistical analysis, and where we as designers need to intervene. Prior work has shown that viewers can reliably estimate a variety of potential statistics of interest. In this chapter, extending experimental results from Correll et al. [2012a] and Albers et al. [2014], we examine this ability from the perspective of design: how can we as designers support viewers in making judgments about aggregate statistics? We specifically focus on time series data, which is both ubiquitous enough to reflect real world visualizations, but flexible enough to afford a variety of different statistical tasks and visual designs. Time series also present a case where the choice of visual design has a noticeable impact on what sort of analysis is done — for instance a bar chart presentation of series data encourages point-to-point comparisons, while a line graph encourages global analysis of trends [Shah et al. 1999].

In this chapter, we examine the design choices that go into creating time series displays meant for visual aggregation. This means both how to maximize the

performance benefits of newer encodings that have a proven ability to support aggregate judgments (but may otherwise be unfamiliar), but also to see if there are ways to modify standard encodings to receive some of these benefits without the associated costs of newer, special purpose encodings. We present the results of a series of crowd-sourced experiments for aggregate tasks (initially mainly a comparison of mean, but later a wider variety of statistical tasks) that explore the design parameters for aggregate encodings. We examine an encoding designed specifically for aggregate rather than point judgments, "color weaving," and explore the design space around color weaving to proposed modifications to fit a wider range of potential tasks and settings.

Visual Aggregation

In chapter 2 as well as our prior work, we show that viewers are capable of extracting a wide variety of statistics of interest from displays [Correll et al. 2012a, Correll and Gleicher 2013, Gleicher et al. 2013], a procedure we term "visual aggregation." Moreover, viewers can perform visual aggregation even without the explicit annotation of these statistics.

In many displays, there is too much data, or too many potential tasks, to explicitly encode every potentially important aggregate statistic (such as regional means, variance, or kurtosis). If viewers are able to accurately and efficiently extract these statistics from the underlying data, then designers both do not need to simultaneously display multiple kinds of statistics in the same display, and can also be more confident that viewers will see the big picture in a way that is in line with the underlying data.

In many cases, there is a tradeoff to be made between designs which support the extraction of individual, low level values (such as the values of particular data points) and designs which support the extraction of higher level statistics (such as the mean of a region of points). By exploring the tradeoff between high- and low-level judgments, we can provide assistance to designers who must tailor visualizations for specific classes of tasks.

3.1 Background

The line graph has long been considered the canonical display for series data. Studies in graphical perception have provided some insight into how different properties of line graphs impact the types of judgments viewers make (e.g., shape [Nourbakhsh and Ottenbacher 1994], slope, curvature [Best et al. 2007], dimensionality[Kumar and Benbasat 2004, Meserth and Hollands 1999], and continuity [Eaton et al. 2005]). However, these works have historically focused on simple high-level judgments of single series line graphs, such as overall trend or value judgments, seldom considering issues such as comparing multiple series or making rapid comparative judgments within a series.

The visualization community has recently begun to explore these issues, specifically considering multiple time series displays. Recent work has examined the efficiency of different representations for comparing multiple series datasets [Javed et al. 2010] and how interaction techniques could be used to support such comparisons [Lam et al. 2007]. These studies have not considered aggregate judgements within series, instead opting to focus on more detail-oriented tasks; however, they have demonstrated the value of using empirical evaluation to understand visualization design and comparison across multiple series. More recent studies in time series visualization have begun to consider more complex aggregate judgments, such as trend identification [Fuchs et al. 2013] and joint analyses of qualitative and quantitative information [Aigner et al. 2012], but do not attempt to illuminate the connection between judgment and encoding.

This practice of using empirical studies to inform visual design provides insight into the tradeoffs of different encoding choices [Cleveland and McGill 1984, Kosara et al. 2003]. By understanding these tradeoffs, visualization designers can make informed decisions about the techniques they use to support the breadth of tasks required of their design, potentially culminating in novel encoding designs. For the specific task of making aggregate judgments within series, empirical studies can illuminate how different properties of an encoding support judgments about specific regions within a series (e.g., mean and variance)[Fouriezos et al. 2008a].

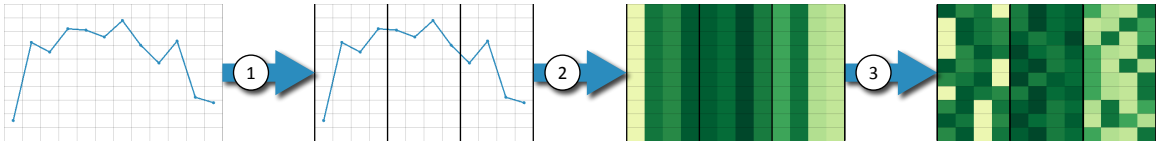


Figure 3.1: The color weaving pipeline, given an input sample time series. The resulting display sacrifices some positional information in order to generate a visualization that affords greater accuracy in extracting high-level statistics from discrete temporal regions of the series.
1. The time series is divided up into b "blocks" of predetermined size.
2. For each block, the data are encoded as a region of color and then divided into r rows.
3. For each row in each block, the regions of color are permuted, generating a "quilt" of interwoven values.

Color Weaving

Recent work has shown that extracting individual points in color displays is easier when sorting is used to introduce visual structure to regions [Haroz and Whitney 2012]. We postulate that the inverse holds for aggregate judgments - by removing visual structure from regions of a color display, we make it easier to compare data in the aggregate. Introducing randomness also provides benefits related to sampling theory – individual subsections of a display are better related to the whole if local patterns are disrupted. We explore color weaving as an example of a technique in this space. Based on prior work [Correll et al. 2012a], we postulate that color weaving takes advantage of our visual system's tendency toward visual summation to efficiently encode dense data values. Inspired by the "weaving" technique evaluated in Hagh-Shenas et al. [Hagh-Shenas et al. 2007] for using a woven display of colors for overlaying multiple datasets, we instead used a pixel-level permutation of values to break local one-dimensional structure and create a woven aggregate block. This block essentially becomes a "quilt" of the values in a region (see Figure 3.1). The above-discussed perceptual literature suggests that by permuting the values within a given region of screen space viewers will be able to quickly extract a series of general statistics about the values in the region, including the average color value, without necessarily sacrificing other information (such as the variance, mode, or heterogeneity of a region).

3.2 Experimental Evaluation

We conducted two sets of crowd-sourced experiments using Amazon's Mechanical Turk to examine design tradeoffs involved in visual averaging. Both focused on color weaving as representative of a broader class of encodings designed to make use of visual aggregation for the extraction of aggregate statistics.

The first experiment set contains three individual experiments that are designed to promote color weaving as a useful technique to solve a specific type of problem. These experiments are meant to determine whether or not we can achieve any of the performance benefits of color weaving without the high cost (both in terms of the relative difficulty of extracting individual values from color encodings, and also the disruption to local structure brought on by permutation). We also seek to determine whether the benefits of color weaving are robust to changes in design.

- Experiment One is a replication of Correll et al. [2012a], but with a design that examines more concretely the permutation aspects of weaving. We examine the difference between *positional* and *color* encodings at performance for comparing means. We also examine the benefits of *permutation per se* — since positional aggregation is a different

- Experiment Two examines alternate choices in *color ramps* for color weaving. Since we postulate that the mean comparison tasks in color encodings relies on the ability of viewers to aggregate regions of color, different choices in color encodings ought to affect this ability.

- Experiment Three examines separation. Color weaving, by choosing discrete bins in which to permute pixels, tacitly creates visual separation between blocks of the time series. We examine whether or not this isolation of important sub-regions, *separation per se*, affects performance.

The second set of experiments is covered in greater detail in Albers et al. [2014], but expands our consideration of visual aggregation to a larger number of tasks, including the comparison of statistics such as range and variance. Color weaving, by suppressing local structure, necessarily makes it difficult to locate and compare individual points. Color weaving's benefits therefore come at a cost. We believe

that many designs for statistical tasks are of this same nature: optimizing for the presentation of particular statistics of interest comes at the expense of making other statistical tasks more difficult. We also examine situations where we explicitly encode particular statistical values, both to generate baseline accuracy for potentially esoteric tasks, but also to show that how these values are calculated and presented results in different patterns of performance.

General Methods

Significant elements of the design were identical amongst all presented experiments. Participants were all recruited using Amazon's Mechanical Turk crowdsourcing platform. Participants were drawn exclusively from the pool of North American Mechanical Turk users. For each experiment, participants were presented with a brief tutorial explaining the model problem. Participants were then exposed to a number of stimuli in sequence. After indicating they were ready, they were exposed to the stimuli for 30 seconds, after which the stimulus image was removed. Users could answer the question at any point after the stimulus had been exposed. Demographics data were collected after all questions had been answered.

We followed established best practices for Mechanical Turk experiments [Heer and Bostock 2010, Kittur et al. 2008, Paolacci et al. 2010], including designing against "click-through" behavior, providing multiple types of input, and creating a priori exclusion criteria. Since many of these experiments relied on counterbalancing, if a participant did not meet our inclusion criteria then more participants were recruited so that an identical number of participants would be included in each condition. Since the task required accurate color perception, in all experiments an initial set of digitally rendered Ishihara plates were used to test for Color Vision Deficiency (CVD) and exclude participants who failed these tests [Clark 1924]. The resulting pool of participants roughly matched the general demographics of North American Turkers as a whole [Ross et al. 2010].

Model Problem

For the these experiments, we focused on comparative judgments of aggregate values within a single time series. This task requires no special domain knowledge

nor any particular familiarity with statistical concepts. Participants were given a time series data representing sales of a product over the course of a year (of 12 "months" of 30 days each, generating a 360 day "year") and asked to select which month had the highest value of a particular statistic (highest average for the first experiment set, highest variance for the second set).

We acknowledge that this task is somewhat artificial — queries of this sort could be aided by explicitly encoding the statistic in question, or the answer provided by a single database query. However, visual averaging has utility because the explicit display of aggregate statistics is not always feasible. For example, it requires knowing ahead of time the specific statistic the viewer is interested in (e.g., if they wish to know the mean rather than some other summary statistic), and the specific range the user wishes to aggregate over. In many cases, there are enough possible candidate statistics to display that choosing the correct subset for the task at hand is difficult or impossible. Choosing irrelevant statistics to display, or statistics calculated at irrelevant scales, impairs ability to extract other, relevant statistics [Albers et al. 2014]. Furthermore, by letting the viewer build up their own picture of the data, they might have more trust or connection to the data in question. However, some statistical tasks might require too much statistical knowledge or expertise to be visualized from the data directly without intervention in the form of explicitly encoding values.

Data Generation

The stimulus for our trials use a time series with 360 steps, divided into 12 "months." We developed a process that generated random series that provided control over which month has the highest average and which month contained the highest single value. Since there is potential ambiguity in our model problem between "find the month with the highest value" and "find the month with the highest *average* value", it was important that our data generation was robust enough to disambiguate these two cases and confirm that participants had understood the model task correctly. By selecting the highest absolute month and highest average months independently, we were able to create stimuli where the correct answer (highest average) occurred in the same month as the the absolute highest value in less than 10% of stimuli (see Figure 3.2). Participants were not informed of this property of the data, to discourage

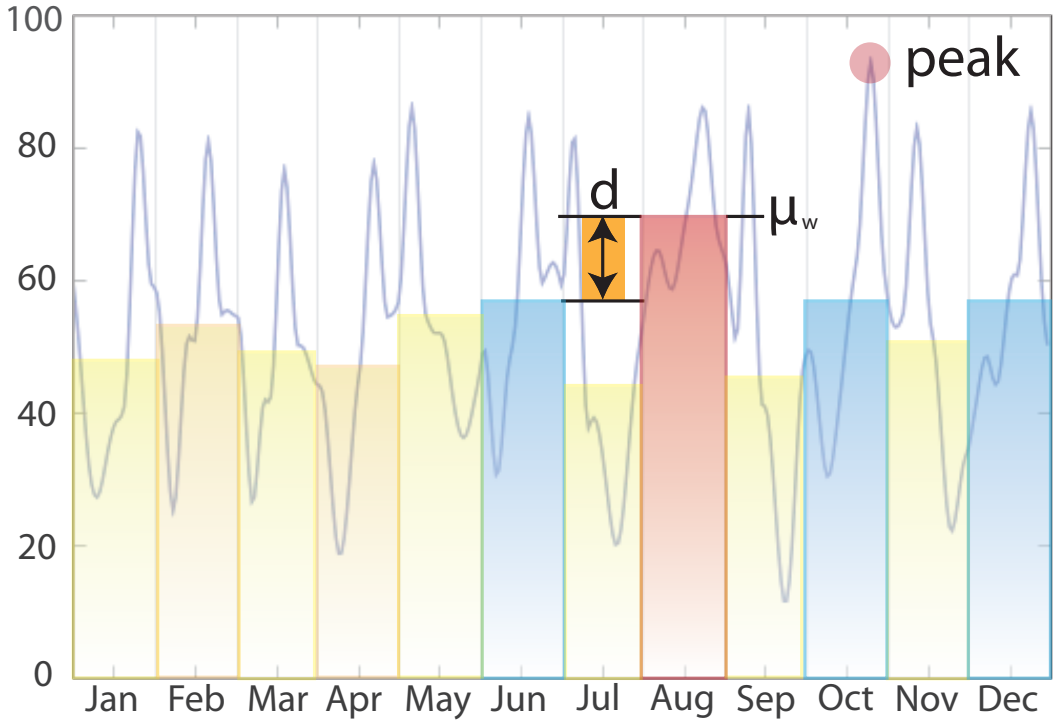


Figure 3.2: Illustrating our parameter *d* on a sample time series. The winning month, August, has an average value of μ_w. June, October, and December are "distracter" months with average value of $\mu_w - d$. As d become smaller it becomes more difficult to disambiguate the winning and distracter months. To discourage "peak finding" as a viable strategy for completing the model task, the highest *absolute* value is seldom in the same month as the month with the highest *average* value.

both peak-finding and peak-avoiding as proxy strategies for the experimental task.

Our stimulus generation also contained two controls over the level of difficulty of an example. First, it allowed a variable level of noisiness in the signal. Second, it created 2-4 "distractor" months that had values close to the winning month, as shown in Figure 2. The value difference between the winning month and the distractors is controlled by a parameter d, which in our experiments was as low as 1 or as high as 5 depending on the granularity of the sampled difficulty levels. Our previous experiments [Correll et al. 2012a] have shown that noise level and d are good controls over the difficulty of the task: larger d and lower noise were highly

correlated with task performance. Another potential factor relating to difficulty, the number of runner-up months, was found in previous experiments and in piloting to not be a significant factor in task difficulty, but was nonetheless balanced across conditions in the subsequent experiments.

For a specific parameter setting of d and noise level, the winning and distractor months were randomly selected. The signal was then created by generating random signals at various scales [Perlin 1985] and performing a least-squares optimization to enforce the constraints of valid signals, such as having the correct monthly averages, having the correct maximum value, and staying within the valid range.

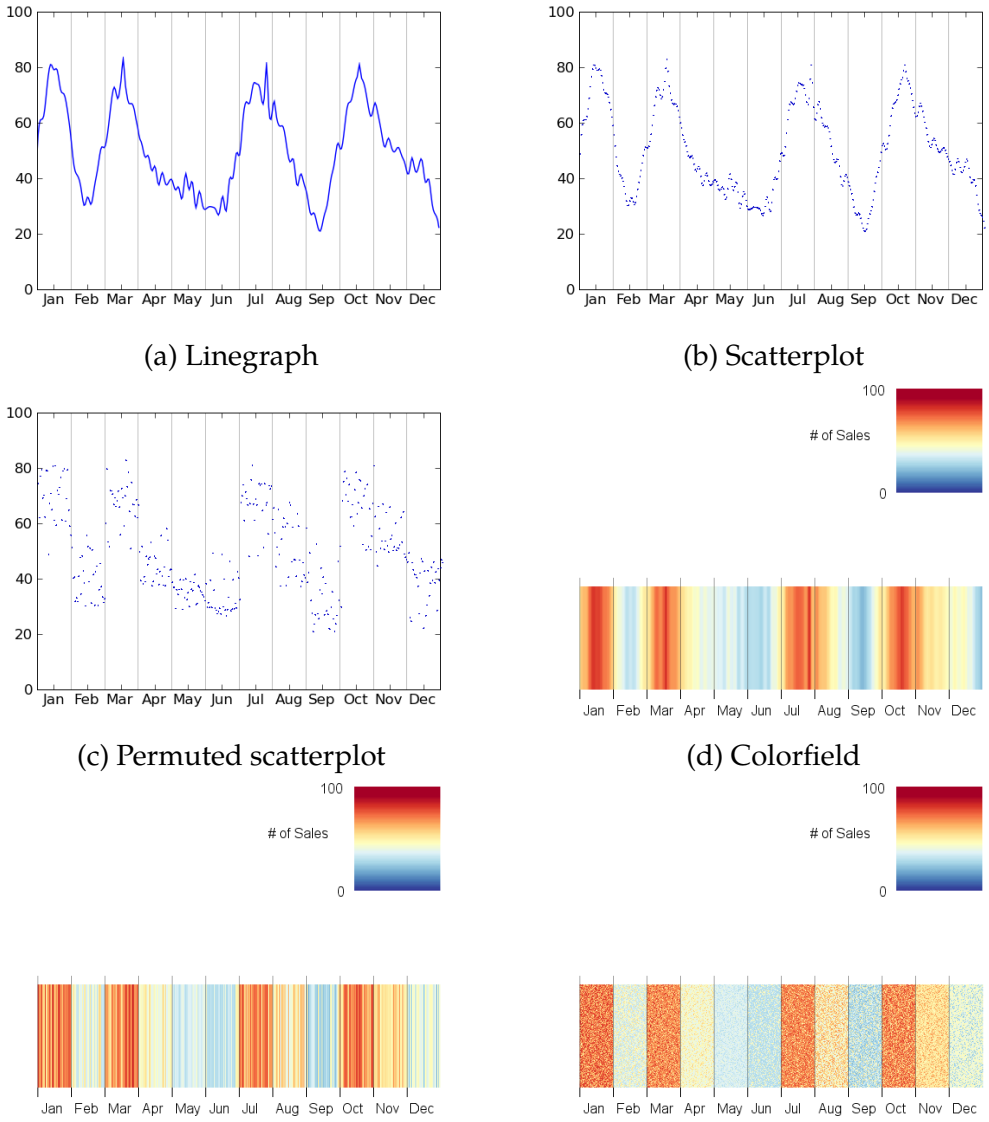
3.3 Experiment Set One: Mean Comparisons

Colorweaving has two components: *permutation* of local regions, and visual *separation* of a time series into discrete bins. While past work has shown the benefits of colorweaving at averaging tasks, the potential cost of color weaving is high (for instance we lose the ability to readily compare individual points in the time series). In this experiment set, we examine whether or not either of these two components to colorweaving, permutation and separation, can improve performance by themselves, reducing the potential cost of such encodings. We also explore how robust colorweaving is to design changes such as different color ramps or separation strategies.

Experiment One: Exploring Color, Position, and Permutation

The experiment detailed in Correll et al. [2012a] was meant to compare performance across traditional time series encodings (where position encodes value, as in line graphs) and color encodings. Since color weaving relies on permutation within regions of interest, we tested whether permutation per se was helpful in making comparative judgments of average value or whether (as we would expect) it was only in the color case that permutation provided a benefit. This experiment built upon the findings of Correll et al. [2012a] by expanding the variety of encodings used in the comparison.

It was our hypothesis that, since perceptual mechanisms exist for the efficient aggregation of color whereas it is unclear if similar mechanisms exist for shape



(a) Linegraph

(b) Scatterplot

(c) Permuted scatterplot

(d) Colorfield

(e) Permuted colorfield

(f) Woven colorfield

Figure 3.3: The six conditions of Mean Experiment One, all encoding the same time series, comparing the effects of permutation on two different kinds of encodings for time series data: encoding value with position or with color.

or position, that color encodings would be superior for the aggregation task. We also suspected that permutation per se would not create improved performance. One dimensional permutation within blocks might remove potentially confusing local structures, but still did not make the task of averaging values any easier from a perceptual standpoint. We believed that two-dimensional permutation would permit the viewer to harness color aggregation to improve accuracy at the mean estimation task.

Design

We explored this hypothesis in a mixed design 2 (color or position encoding) $\times$ 3 (permutation type) $\times$ 2 (noisy or smooth) $\times$ 2 (high or low d) experiment. Choice of encoding was a between subjects factor; all others were within subjects. Color encodings were either presented as a colorfield,as a colorfield with values randomly shuffled within each month (horizontal permutation), or woven at the per-pixel level within each month (permuted both horizontally and vertically). Position encodings were either presented as a line graph, as a scatter plot, or a scatter plot where values were permuted within each month. Since the use of position for encoding prevents the use of permutation along the vertical axis the factor levels are not strictly analogous across conditions: rather, the choice of position-encoded factor levels is meant to capture the contention that it is permutation and, analogously, the removal of visual confounds due to local structure that results in improved performance for woven fields. Figure 3.3 shows the same time series as represented by all six encoding types used in this experiment.

Another set of factors were related to the task difficulty. Participants were given two different levels of our parameter d (difference between the winning month and any runner-up months) and two different levels of relative signal noisiness. Participants say three stimuli at each combination of d and noise, in random order, for a total of 12 stimuli per permutation type. These were presented in 3 blocks (block order counterbalanced across all participants) for a total of 36 stimuli.

We recruited 52 participants (22 females and 30 males, μ age = 35.3). Statistically significant underperformance on questions where d was high was used to determine if participants should be excluded, and 4 such exclusions were made, for a total of 48 total participants.

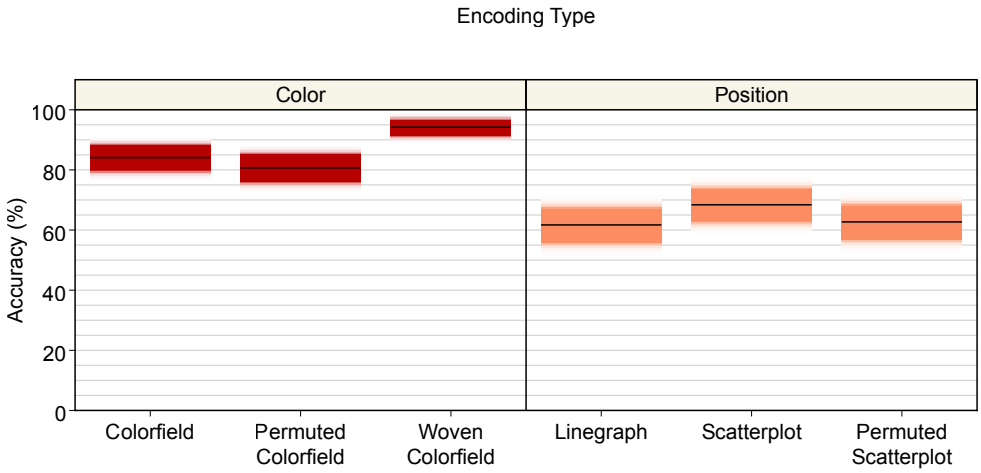


Figure 3.4: The effect of different types of permutation on performance in Experiment One, as gradient plots. Fully opaque colored regions represent a 95% t confidence interval. Participants in conditions where data were encoded using color performed significantly better than in conditions where position was used to encode data. For color, color weaving also significantly improved performance. No type of permutation resulted in significant improvements for position encodings.

Results

We performed a two-way analysis of covariance (ANCOVA) to analyze the effects of different encoding types and permutation types on task performance. In all conditions aggregate performance was significantly better than chance (which was 1/12 assuming that viewers were equally likely to choose any month, or approximately 1/3 if the viewers selected only from between winning and "runner-up" months). We found that there was a significant effect of encoding type ($F(1,1718) = 126$, $p < .0001$), with color encodings significantly outperforming linegraphs (average accuracy of 86.3% in the color case, versus 64.7% in the position case). In addition, the effect of permutation type combined with encoding on accuracy was significant ($F(2,1718) = 7.90$, $p = 0.0004$)). A post-hoc comparison using Tukey's Test of Honest Significant Difference (HSD) confirmed that performance was similar in the position encoding case no matter the choice of permutation, while for color, the woven encoding performed significantly better. Figure 3.4 summarizes these results.

Our choice of parameters associated with task difficulty also proved to significantly influence performance. Our parameter d (difference between the largest and next largest average value) had a significant effect ($F(1,1718) = 96.4$, $p < .0001$), with performance decreasing as d is smaller. Noise behaved similarly ($F(1,1718) = 28.4$, $p < .0001$), with noisier data being more difficult to accurately compare. These results were consistent across all reported mean experiments. Figure 3.5 summarizes these results. Our hardness parameters were monotonic with performance (the noisier the data, and the smaller the average differences to be compared, the worse the performance). Neither d nor noise had interactions with performance across encoding type ($F(1,1718) = 0.410$, $p = 0.522$ and $F(1,1718) = 0.410$, $p = 0.522$ respectively), indicating that these parameters are robust and useful metrics for task difficulty.

There are a number of possible alternate strategies (or misinterpretations of the experimental task) for locating months believed to have high averages – two we considered were selecting the month containing the highest absolute value (the peak month), and selecting the month with the highest range midpoint (the average of the maximum and minimum value for a month). The peak month was explicitly de-correlated from the month with the highest average (see §3.2), but in some cases participants did select the peak month (in 37.1% of incorrect answers). While the range midpoint was not de-correlated from the average, It was rare for participants to be fooled by months with high range midpoints (only 10.7% of incorrect answers). A Student's t-test found that participants were not significantly more likely to choose a peak month incorrectly in the color conditions as opposed to the positional conditions (36.1% for color vs. 37.6% for position, $p = 0.765$). However, participants in the color condition were significantly less likely to choose months with a high range midpoint incorrectly (6.0% of incorrect responses in the color condition vs. 12.7% of incorrect responses in positional encodings, $p = 0.016$). This difference is likely due to the difficultly of viewers from accurately estimating range in colorfield displays [Albers et al. 2014].

In summary, we find that **color encodings outperform positional encodings** for the task of comparing averages within regions of a time series, even against positional encodings which attempt to break local structure, and color weaving outperforms other color encodings.

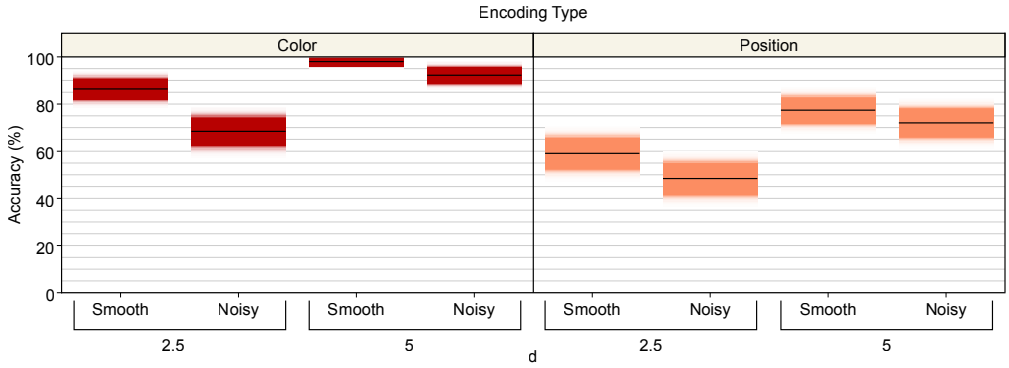
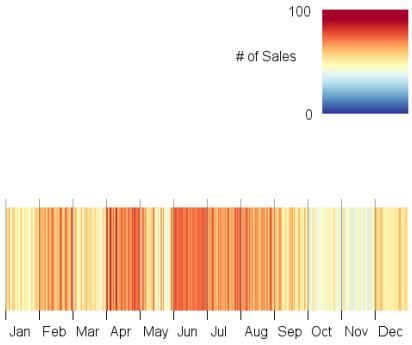


Figure 3.5: Our proposed task difficulty parameters compared to actual performance, as gradient plots. Fully opaque colored regions represent a 95% t confidence interval. The value d represents the difference between the highest average and the next highest value: both lower values of d and noisier signals resulted in lesser overall accuracy.

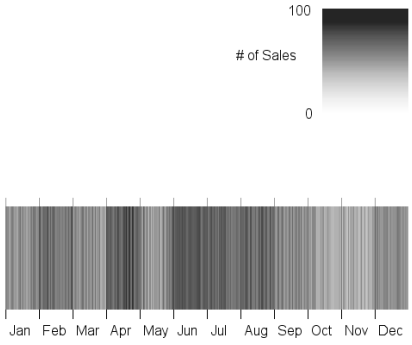
Experiment Two: Exploring Color Ramp Designs

One of the limitations of using color encodings to display is that the fidelity and precision of the encoding is highly sensitive to the choice of color ramp. In this experiment we sought to see if the performance of encodings were robust to different color ramps, and to confirm that the perceptual mechanisms at work in visual averaging are affected by common choices in the design of color ramps.

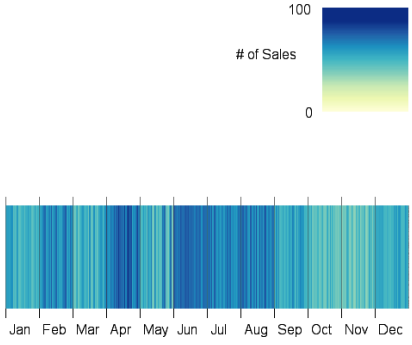
While we initially believed that performance would be roughly identical among color ramps, internal piloting revealed apparent differences. We speculated that our model task would privilege certain color ramps: namely, since participants were asked to estimate the *highest* average month, they would implicitly or explicitly discard months in which low values were present. Thus having more dynamic range among values along the upper end of the color ramp would afford greater discrimination of high average values and so greater performance. We hypothesized that diverging color ramps, being constructed with more dynamic range among the higher values, would have the highest performance.



(a) Diverging



(b) Monochrome Sequential



(c) Multi-hue Sequential

Figure 3.6: The three color ramps used in Mean Experiment Three, all encoding the same time series.

Design

We evaluated choice of color ramp in a 2 (colorfields or color weaving) $\times$ 3 (color ramp) mixed design experiment. Choice of color ramp was a a between participants factor. We selected our color ramps from a set of known ColorBrewer ramps [Harrower and Brewer 2003]. Participants saw either the diverging color ramp presented in the earlier experiments, a white-black color ramp (monochrome sequential), or a sequential color ramp from yellow to blue (multi-hue sequential). The ramps were normalized to the range 0-100, regardless of the values of the actual data series; in practice, as a consequence of the data generation process, most stimuli used almost the full range of the dataspace regardless. Figure 3.6 shows examples of each stimuli. Equal amounts of woven and non-woven data using these ramps were presented as a within-participants condition. Two other within participants conditions were the d parameter, sampled from levels 1-4, and the presence or absence of noise. Participants saw 2 stimuli at each level of d, and noise for a total of 16 stimuli per permutation type. These were once again presented in blocks, with block order determined via a Latin squares design, for a total of 32 stimuli.

We recruited 53 participants (30 females and 23 males, μ age = 29.7). Significant underperformance on all questions was used to determine exclusions. 5 exclusions were made on this basis, for a total of 48 participants.

Results

We performed a two-way ANOVA to analyze the choice of color ramp on task performance. We found a significant effect of color ramp ($F(2,1526) = 16.3$, $p < .0001$). A Tukey HSD confirmed that participants performed better with the two-point diverging color ramp than the sequential ramps (average accuracy of 80.1% versus 68.4% and 66.7% for the monochrome and sequential color ramps respectively). There was no significant effect of choice of ramp across permutation type ($F(2,1526) = 0.485$, $p = 0.616$). Figure 3.7 summarizes these results. Even across a diverse set of color ramps, weaving significantly outperformed standard colorfields ($F(1,1526)=42.8$, $p < .0001$, average accuracy of 78.7% versus 64.8% respectively).

While the benefits of weaving were present even across different color ramps, our results show that choice of color is a tool to control task performance. In

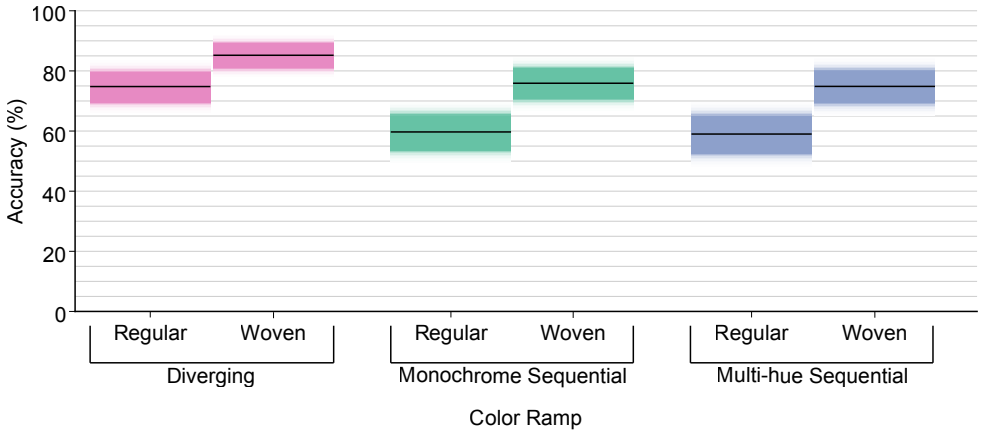


Figure 3.7: The effect of choice of color ramp on task performance. Error bars represent a 95% t confidence interval. The diverging color ramp performed significantly better than the other two ramps. Weaving improved performance across all color ramp conditions.

particular our results seem to indicate that design decisions in selecting color ramps apply equally to the woven and non-woven cases: ramps with categorically greater dynamic range at each extrema better facilitate visual average comparison. However, the choice of a diverging ramp requires the selection of a zero point, which may heavily impact how viewers discriminate between values.

Experiment Three: Exploring Task-relevant Separation

The decision to employ color weaving has clear costs – local structures are completely suppressed, and the resulting analysis is highly sensitive to how the time series is segmented. In our experiments, for example, positional information at scales smaller than one month are completely discarded. The choice of highlighting aggregate values (such as average and variance) also has performance costs when asking about individual values (such as local extrema) [Albers et al. 2014]. By performing permutation of data within semantically meaningful blocks (each month was a separate block in our experimental task), color weaving creates explicit visual boundaries between months. It is possible that this visual segmentation explains the

higher participant performance seen in our previous experiments. We performed an experiment to see if group separation per se could improve performance at extracting average values without having to perform the more disruptive step of permuting color within blocks. We used explicit spatial separators between blocks ("gutters") to create explicit between-block boundaries.

We hypothesized that this separation would be helpful in the non-woven case, but that the implicit separation performed by color weaving would provide sufficient boundaries, so additional boundaries would not improve performance. However, we anticipated that color weaving has performance benefits beyond simply creating block boundaries.

Design

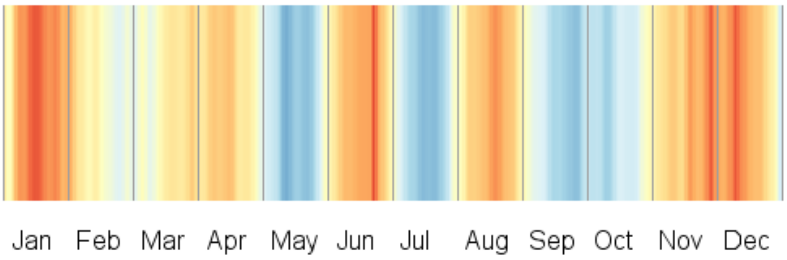
We tested visual segmentation through a 2 (position or color encoding) $\times$ 2 (standard colorfield/linegraph, woven colorfield/permuted scatterplot) $\times$ 3 (segmentation type) mixed design experiment. Choice of main encoding (color or position) was a between subjects factor; all other factors were within subjects. We evaluated three visual segmentation boundaries: a one pixel black line between months, a five pixel black gutter between months, and a ten pixel white gutter between months (the same color as the background of the image, making each month "float" in the foreground). Figure 3.8 shows an example of each of these types of separation.

In total, each participant saw 4 stimuli at each level of hardness and permutation choice for a total of 16 stimuli for each separation type. These were presented in 3 blocks for a total of 48 stimuli. Participant block order was counterbalanced in full.

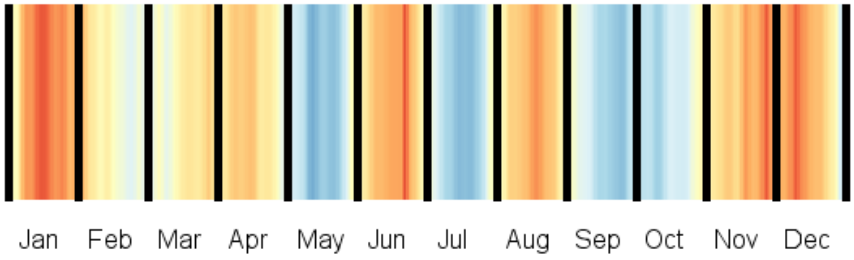
We recruited 48 participants (18 females and 30 males, μ age = 32.6). Statistically significant underperformance on easy questions was used to determine if participants should be excluded, but ultimately all no exclusions were made.

Results

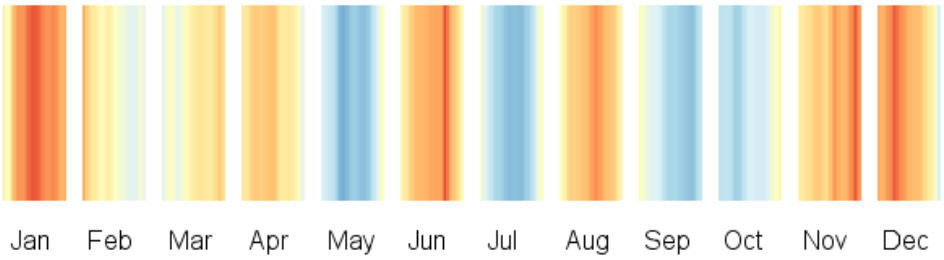
We performed a three-way ANOVA to analyze the effects of encoding type, permuation type, and gutters on task performance. We found a significant effect of type of gutter on performance ($F(2,2292)=4.04$, $p = 0.017$). A Tukey HSD showed that conditions with gutters had significantly higher performance than stimuli with



(a) One pixel gutters



(b) Black gutters



(c) White gutters

Figure 3.8: The three different gutter types in Mean Experiment Two, all encoding the same time series.

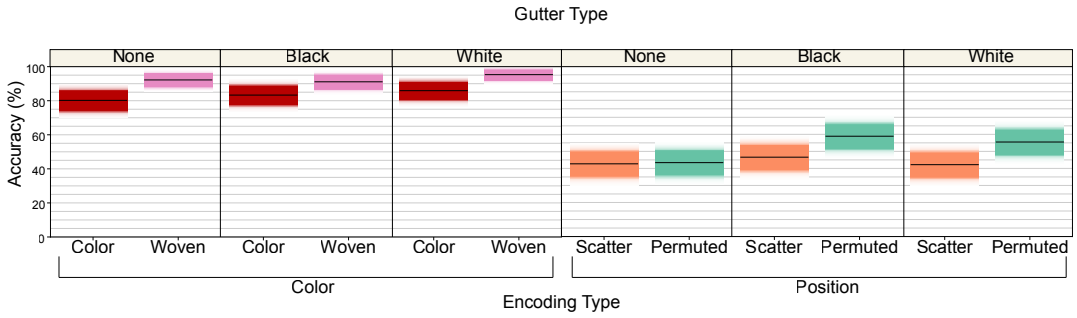


Figure 3.9: The effect of explicit separators on task performance, as gradient plots. Fully opaque colored regions represent a 95% t confidence interval. While in the baseline conditions for colorfield and linegraph there were no significant differences in performance between different choices of gutter, using gutters had an interaction with permutation. For scatterplots specifically, adding separation between months made improved performance (likely by making individual point clouds for months easier to differentiate).

no gutters. There was a marginal interaction between encoding type and gutter type ($F(2,2292)=2.37$, $p = 0.093$). A Tukey's HSD showed that while gutter type did not lead to significantly different performance among color encodings, for position encodings black gutters resulted in significantly higher performance than with no gutters. Figure 3.9 summarizes these results.

Permutation was also a significant main effect ($F(1,2292) = 28.1$, $p < 0.0001$), with the permuted encodings (woven colorfields, permuted scatterplots) performing better than the baseline encodings. There was a marginal interaction between encoding type, permutation type, and gutter type ($F(2,2292) = 2.23$, $p = 0.10$). A Tukey's HSD showed significant difference between permuted and baseline encodings across all gutter types, with the exception of black gutters in the color condition (where the difference was marginal, $p = 0.069$) and where there were no gutters in the position encoding, where there was no significant difference ($p = 0.877$).

These results show that while explicit separation between groups does not alone explain the success of color weaving, having explicit visual separation between groups can improve performance for aggregate tasks, especially for encodings like the scatterplot where point clouds can be difficult to segment.

3.4 Experiment Set Two: Exploring Additional Encodings and Tasks

The prior experiments deal with only one types of statistical task, comparison of mean. However, we do not believe that color encodings generally, and color weaving specifically, are appropriate design choices for a wider array of statistical tasks. To better outline the space of encodings to support statistical judgments, as well as delineate the tasks for which color encodings excel, we conducted an additional series of experiments, one for each of six statistical tasks (§3.4), and compared the performance of viewers asked to make comparative judgments from time series data across the eight different visual encodings (§3.4) The experiments shared some common features across both tasks and encodings that we describe here. The Results section describes the specifics of each experiment along with their results for clarity. Albers et al. [2014] has more information about these experiments, and a theoretical framework for contextualizing the results. We present them here as evidence that the capability of viewers to extract statistical information from displays extends to a wide variety of real statistical tasks, but that visualization design can affect these capabilities.

Each experiment focused on one of a set of statistical tasks. The encoding used to visualize the time series data was a between-subjects factor (see Figure 4.3). Thus the entire experiment set can be thought of as a 6 (statistical task) $\times$ 8 (visual encoding) $\times$ 4 (our d parameter) $\times$ 2 (noise level of the signal) mixed model experimental design. The d parameter is defined as per our previous experiments, the difference in value from the "winning" month to the runner ups. Since the different statistical tasks varied widely in difficulty, appropriate d levels were tuned via piloting.

In piloting, we observed a learning effect. To partially counteract this, we presented participants with an initial set of four stimuli designed to show the heterogeneity of difficulties present in the task and also to help participants develop an initial understanding of the task and encoding. These initial "training" stimuli were excluded from analysis. We also randomly interspersed stimuli that were intentionally "easy" to serve as validation questions to gauge both validity of responses and participant understanding of task. For each task, we determined a minimum acceptable accuracy on validation questions based on piloting. We recruited additional

participants to replace participants failing to reach this level. Validation stimuli were otherwise excluded from analysis. Each participant saw a total of 44 stimuli (4 training, 6 validation, and 32 experimental) and was paid $1.00.

Hypotheses

In our experiments across tasks, we wished to highlight two different distinctions between encodings. The first distinction is between our *color*- and *position*-based encodings. While we have evidence that people are capable of creating mental summaries based on colorfields (see chapter 2) we do not believe that positional encodings afford this aggregation to the same degree (although we do believe that humans are capable of extracting average position from visualizations, as in our scatter plots work [Gleicher et al. 2013]). Thus we believed that, **for aggregate tasks, color encodings will outperform positional encodings**. We therefore considered the color and linegraph encodings from the first experiment set, but also a variety of new encodings from each class.

Another factor spread among our conditions was the decision on what information to explicitly encode in our charts. While naturally, **explicitly encoding task-relevant values will improve performance** over having to calculate these values through visual aggregation, this explicit encoding can have costs. This is a cost both in visual complexity but also potential confusion if the explicitly encoded values are relevant but do not strictly align with the task at hand. We made sure to include both examples where relevant per-month statistics were directly encoded (as with the composite graphs where the per-month average is encoded as a bar chart) but also where relevant statistics did not quite align with tasks (as with modified stock charts, where the moving average, rather than the per-month average, is displayed).

Tasks

Our previous experiments involved comparison of means, which is a relatively simple statistical quantity. Expanding the scope of statistics under consideration required designing for participants who may not have explicit statistical knowledge. For instance, range is the difference between the local minimum and maximum,

whereas spread sounds similar but considers variation amongst all points. There is little existing research on asking lay audiences about outliers and spread (although see Roth and MacEachren [2015]). For our experiments, we needed to determine effective ways of asking about these statistics. We generated candidate wordings by consulting the Simple English Wikipedia and evaluated these candidates in a pilot study on Mechanical Turk, asking participants to asses the comprehensibility and accuracy of several phrasings. We tested the following statistical properties, using the following final wordings:

1. **Maxima**: Which month had the day with the highest sales for the year?
2. **Minima**: Which month had the day with the lowest sales for the year?
3. **Range**: Which month had the largest range of values?
4. **Average**: Which month had the highest average sales for the year?
5. **Spread**: Look at the average sales from each month. Which month had the sales which were the most spread out from their monthly average?
6. **Outliers**: Which month had the most unusual (outlier) sales days?

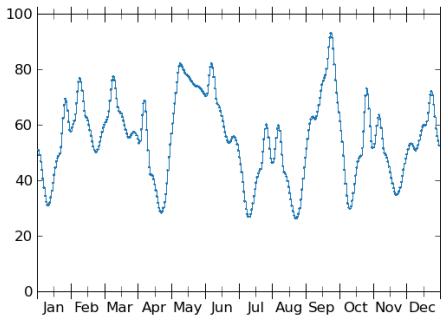
For all of these tasks, since the viewer's specific goal was known to the designer, the answer could have been given directly. However, our goal is to understand how visualizations work in settings where the designer may not know the exact goal of the viewer, or the viewer may have multiple goals.

Experimental Conditions

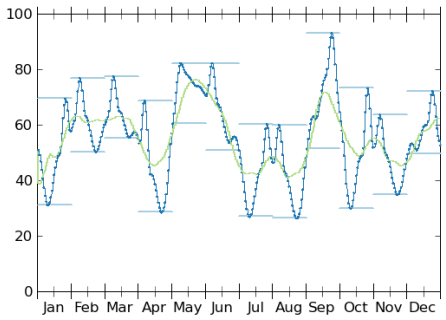
Position-Based Encodings

Line graphs (Figure 3.10a) are the canonical approach for visualizing time series data using position. Position encodings support extracting exact values from a visualization [Cleveland and McGill 1984].

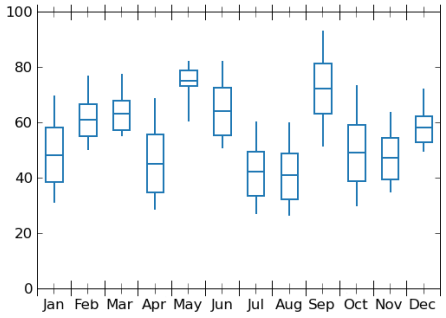
Modified Stock Charts (Figure 3.10b) supplement summary judgments in line graphs by layering a moving average over the original series. Extrema of discrete regions are encoded using range bars. We anticipate that the presence of the moving



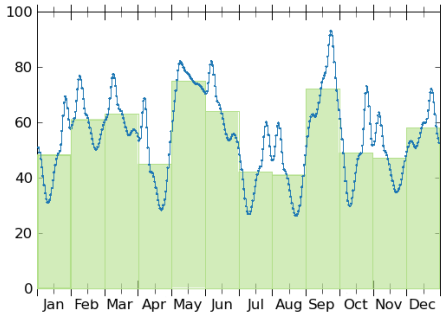
(a) Line graph – vertical position encodes value.



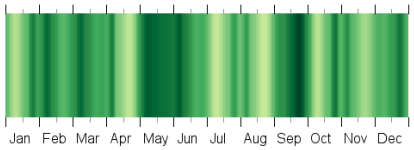
(b) Modified Stock Chart – line graph with monthly highs and lows (horizontal bars) and 30-day moving average (green line).



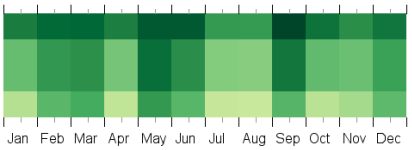
(c) Box Plot – each month shows its interquartile range (box), mean (horizontal line), and extrema (whiskers).



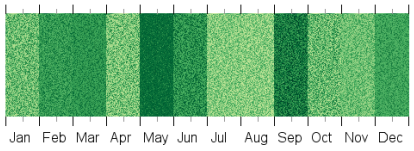
(d) Composite Graph – bar chart encoding the monthly average is overlaid on a line graph of the raw data.



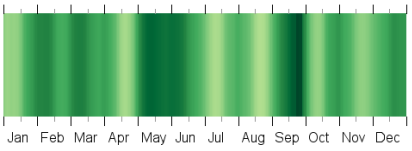
(e) Colorfield – a vertical stripe of color encodes the value of each day (darker green means higher sales).



(f) Color Stock Chart – each month has 3 color blocks: the top encodes the maximum, middle is the mean, bottom is the minimum.



(g) Woven Colorfield – original colorfield's pixels are randomly permuted within each month, creating discrete "blocks" of values.



(h) Event Striping – smoothed data plotted as a colorfield with outliers overlaid as vertical color bands (end of Sep, dark green).

Figure 3.10: Visual designs explored in the Additional Tasks Experiment. The first two rows of encodings use position to encode value; the bottom two rows use color. Conditions 3.10d, 3.10b, 3.10c, 3.10g, 3.10f, and 3.10h calculate and display different statistics at the per-month scale, which requires prior task knowledge (e.g. that the tasks will be performed at the scale of months).

average will help with summary comparisons, albeit the continuous mean aggregation may still limit value extraction from discrete regions. The increased saliency of the extrema as discrete range bars will better afford minimum, maximum, and range comparisons. However, the amount of information encoded by the chart may cause issues of visual clutter.

For some comparison tasks, summary statistics may sufficiently summarize the necessary information in a series. **Box plots** (Figure 3.10c) discretely compute and visualize the range, interquartile range (IQR), and mean of the series for each temporal region. The explicit encoding of these statistics may better afford comparisons of the encoded statistics, but does so at the expense of the raw data.

Composite graphs (Figure 3.10d) layer a line graph over a bar chart representing averages of discrete subregions. By explicitly mapping the mean value aggregated over each month, this approach may enhance the viewer's ability to extract averages from the visualization without inhibiting their ability to extract point-level information from the original series. Visually encoding the average may also provide a benchmark statistic for comparisons requiring average extraction, such as spread (average distance from the average).

Color-Based Encodings

Prior work demonstrates that color encodings, such as those used in **colorfields** (Figure 3.10e), may better support average comparisons than position encodings [Correll et al. 2012b]. Colorfields map each datapoint within a series to a point on a color scale, creating a one-dimensional heatmap. We anticipate that the perceptual system's ability to preattentively summarize color will support summary comparisons; however, we also anticipate that colorfields will be less effective for point comparisons due to the limited perceptual fidelity of color.

Color Stock Charts (Figure 3.10f) explicitly map the local extrema and average of each temporal range using color (average in the center, with top and bottom runners representing local maxima and minima respectively). This approach simplifies the visual computation required to extract point values from a colorfield while preserving some high-level statistics from the series; however, the performance benefit of this mapping may be limited by the ability of the color encoding to

communicate each statistic. Further, encoding only these tasks statistics sacrifices the ability to extract data about local features or other distributional information.

Color weaving [Albers et al. 2011, Correll et al. 2012b] (Figure 3.10g) breaks local structures in a colorfield by randomly permuting data values at the pixel-level within each month. While the previous experiments illustrated the utility of this encoding for mean comparison, the increased difficulty of extracting a particular datapoint may complicate point comparisons using color[Stone 2012].

Event striping [Albers et al. 2011, Correll et al. 2011a] (Figure 3.10h) highlights outliers in the dataset by representing outlier values as broad "stripes" drawn over a smoothed colorfield representation of the original series. Explicitly mapping outlier values within the series visually boosts unusual values while the smoothed colorfield preserves the context of the series. Event striping provides an example of an encoding designed specifically for a given task. Its visual design is very similar to colorfields; however, the design choices made to support outlier identification may influence how well the encoding supports other tasks. This approach may be especially beneficial for noisy data, where outliers may be camouflaged by the variability of the series, and for point comparisons when target values are outliers. However, the visual saliency of striped outliers within the signal may bias the perceptual summarization of the visualization, making summary comparisons more difficult.

Results

In this section, we detail each experiment and its results. Figure 3.11 summarizes our findings. For each experiment, we performed an Analysis of Covariance (ANCOVA) to determine the effect of encoding type on accuracy. The model also tested for interaction effects between encoding type and our hardness parameters (d and noise level). Hardness parameters had generally highly significant effects in the expected direction (noisier signals underperform smoother signals, smaller d are more difficult, etc.), and so we omit these factors from reporting unless unusual. For significant results, we performed Tukey's Test of Honest Significant Difference (HSD) with $\alpha = 0.05$ to extract clusters of performance. We also performed post-hoc mean squared contrast tests to verify significant differences within clusters.

Encoding		Maxima	Minima	Range	Average	Spread	Outliers
Line graph		87.5%	78.9%	74.2%	47.7%	48.8%	36.7%
Modified Stock Chart		88.7%	96.1%	91.8%	56.3%	39.7%	34.0%
Box Plot		75.0%	93.8%	88.5%	68.8%	85.0%	X
Composite Graph		93.0%	88.3%	77.0%	85.9%	53.8%	33.6%
Colorfield		59.4%	56.6%	48.8%	60.5%	57.8%	31.3%
Color Stock Chart		69.9%	73.4%	64.8%	70.3%	X	X
Woven Colorfield		43.0%	45.7%	38.7%	77.7%	71.3%	23.0%
Event Striping		61.7%	59.4%	44.1%	52.3%	42.2%	66.8%

Figure 3.11: A summary of the results of the Additional Tasks Experiment. All measures are in accuracy across all participants. Gray rows indicate position encodings, which use position to encode value; white indicate color encodings, which use color to encode value. Gray columns indicate summary comparison tasks, which involve aggregate statistical information; white columns indicate point comparison tasks, which rely on per-point statistical information. An "X" indicates that the encoding does not afford that task and so no experiment was conducted for this combination of task and encoding (e.g. since the color stock charts only explicitly encoded max, min, and average, it would be impossible for a participants to determine spread from that encoding). Since performance is not strictly comparable across tasks, cell color encodes the number and direction of standard deviations from the task mean: $\leqslant -1$,(-0.5,-1),[0.5,-0.5],(1,0.5),$\geqslant 1$.

Including piloting and the main tasks, we recruited a total of 582 participants, 306 male and 276 female (μ_{age}=31.3, σ_{age}=10.3). A Student's t test showed no significant differences in performance across gender (μ_f=60.1%, μ_m=64.4%, $p = .0938$). For each experiment, 8 participants were recruited per encoding, totalling 64 participants for tasks evaluating all eight encodings, 56 for the spread experiment (which excluded color stock charts), and 48 for the outlier experiment (which excluded box plots and color stock charts), totaling 360 participants for the main experiments. If a participant failed to achieve acceptable performance on validation stimuli, we discarded their data and recruited additional participants for that condition. Across all experiments, 37 additional participants were recruited for this reason. Although accuracy was our performance metric, we tracked response time for each task and found the longer a participant spent on a particular question, the *more* likely they were to be incorrect

($b = -1.6\%$ accuracy/sec, Pearson's $r = 0.83$).

Maxima

For this task, participants were asked to locate the month containing the day with the highest absolute sales. Maxima within the series were created by amplifying the peak in the base series and constraining all remaining values to be at least d less. Especially in the color conditions where detecting individual points is difficult, we considered that picking the month with the highest average sales could be a confounding strategy, so we de-correlated the month with the highest average sales from the month with the highest absolute sales. We sampled evenly across ds of 1,2,3,4 with validation stimuli with $\Delta = 20$.

Encoding had a significant main effect ($F(7, 2016) = 45.8$, $p < .0001$). Generally, position encodings outperformed color encodings, with one exception. Box plots significantly under-performed all other positional encodings ($F(1, 2016) = 24.5$, $p < .0001$), and were not statistically significantly different from the color stock chart ($F(1, 2016) = 1.70$, $p = .1930$). The remaining color encodings performed significantly worse than the color stock charts ($F(1, 2016) = 28.8$, $p < .0001$) and the position encodings as a group.

Position encodings, which afford precise judgment of individual points, outperformed color encodings, which are not as accurate for extracting exact values.

Color stock charts, which were the only color encoding to explicitly encode the maximum value in each month, outperformed other color encodings, while box plots, which were one of two position encodings to explicitly encode maximum values, under-performed compared to the other position encodings. This may be due to biases arising from visual properties of box plots that have been shown to impact the perception of whisker values [Behrens et al. 1990].

Minima

For this task, participants were asked to locate the month containing the day with the lowest absolute sales. This task was functionally identical to the Maxima task – questions about "highest" were changed to "lowest" and the stimuli were derived using the same constraints as the Maxima task. Despite the similarities in the tasks,

prior work [Sanyal et al. 2009] suggests that there are differences in performance between the two and that different encodings may be appropriate.

Encoding had a significant main effect ($F(7, 1984) = 59.1$, $p < .0001$). Within groups, line graphs significantly underperformed the rest of the position encodings ($F(1, 1984) = 25.5$, $p < .0001$), and were only marginally better than color stock charts ($F(1, 1984) = 2.76$, $p = .0966$). The remaining color encodings proved significantly worse than the color stock charts ($F(1, 1984) = 46.1$, $p < .0001$), and also the position encodings as a group. Unlike other experiments (even the similar Maxima experiment), the noisiness of the signal had no significant effect on accuracy ($F(1, 1984) = 0.18$, $p = .6725$).

As in the Maxima experiment, position encodings tended to outperform color encodings.

Box plots and modified stock charts, both of which both explicitly encode monthly minima, outperformed line graphs, which do not. Color stock charts outperformed all other color encodings. These findings align with prior work suggesting that different encodings may be effective for minimum and maximum tasks [Sanyal et al. 2009].

Range

For this task, participants were asked to locate the month with the largest range of sales — the largest gap between the maximum day and the minimum day. Initial piloting showed that participants would frequently confound the range with the maximum. To avoid confounds with the maximum and the related measure of spread, we explicitly decorrelated these three quantities. The task proved more difficult than either of the extrema tasks as it required participants to compare the difference between two points. To avoid floor effects we sampled from ds of 4, 7, 10, and 15, with validation stimuli with $d = 20$.

Encoding had a significant main effect ($F(7, 1984) = 59.3$, $p < .0001$). The color encodings all significantly underperformed the position conditions. Encodings which explicitly encoded extrema performed significantly better than the other encodings of their type: color stock charts outperformed the other color encodings ($F(1, 1984) = 45.8$, $p < .0001$), and box plots and modified stock charts outperformed the other positional encodings ($F(1, 1984) = 28.9$, $p < .0001$).

Position encodings afford greater fidelity in extracting point values than color encodings.

Box plots, modified stock charts, and color stock charts all explicitly encode local extrema values and all outperformed other encodings with equivalent visual variables.

Averaging

For this task, participants were asked to compare means of months. In piloting, the highest *average* value was often confused with the highest *absolute* value, so these values were decorrelated in the stimuli. We sampled Δs of 1,2,3,4, with validation stimuli at $\Delta = 20$.

Encoding had a significant main effect ($F(7, 1984) = 22.6$, $p < .0001$). Encodings which explicitly encoded discrete monthly averages (the composite graph, box plot, and color stock chart) and discretely blocked woven colorfields significantly outperformed the remaining encodings ($F(1, 1984) = 122$, $p < .0001$) Within clusters, there were several pairwise results. In particular, composite charts outperformed woven color fields ($F(1, 1984) = 4.24$, $p = .0395$), and regular colorfields outperformed line graphs ($F(1, 1984) = 11.4$, $p = .0008$).

Colorfields, which support pre-attentive methods of summarization, outperformed line graphs, which do not.

Composite graphs, which explicitly encode mean, outperformed woven colorfields, which do not; however, color stock charts, which also explicitly encode montly averages, did not outperform woven colorfields, which leverage visual aggregation.

All of the encodings which discretely aggregated the data per-month outperformed the other encodings.

Spread

For this task, participants were asked to compare the spread of each month. Since strict control over standard deviation requires complex optimization, we measured spread using the more practical related statistic of absolute deviation ($\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$). Linear scaling about the monthly mean was used to tune the absolute deviation of individual months to fit our constraints. Even so, it is difficult to generate large

differences in variation as each point must remain in the [0,100] interval. As "spread" is an ambiguous term, we decorrelated the month with the highest absolute deviation from the month with the largest range. To avoid floor effects for what was in piloting a difficult task, we sampled Δs of 2,3,4, and 10, with validation stimuli with $\Delta = 15$ - the largest that could be reliably generated in sufficient numbers.

Encoding had a significant main effect ($F(6, 1736) = 36.8$, $p < .0001$). Box plots outperformed color weaving ($F(1, 1736) = 13.7$, $p = .0002$), which in turn outperformed all the remaining encodings ($F(1, 1736) = 50.0$, $p < .0001$). Standard colorfields outperformed both boosted colorfields and modified stock charts ($F(1, 1736) = 17.9$, $p < .0001$). Noise had only a marginal effect on performance ($F(1, 1736) = 3.39$, $p = .0656$), and the number of distractors had no significant effect ($F(3, 1736) = 0.847$, $p = .4679$).

Woven colorfields performed better than nearly all other encodings, as weaving allows for quick visual summarization of the variance of a region despite not explicitly encoding this value.

Only box plots explicitly encoded a statistical variable that was highly correlated with absolute deviation (IQR) and best supported this task.

While the top two encodings both explicitly blocked data together into months, composite graphs were not statistically different from any of the other encodings despite being blocked with respect to a benchmark statistic (average).

Outliers

For this task, participants were asked which month contained the highest number of outliers. The task required both extracting summary statistics and numerosity estimation of points violating these statistics. We generated outliers from existing signals by amplifying days varying largely from the series mean to between 2.25-2.75 standard deviations from the mean. To avoid visual "plateaus" where consecutive outliers appear as on data point, outliers were at least 3 days apart and no month contained more than 8 outliers. Spread can confound outlier count, so we decorrelated the month with the highest absolute deviation from the month with the most outliers by reducing the absolute deviation of the high outlier month. To avoid confounds between the month with the greatest number of outliers and the month with the largest outlier, we de-correlated the largest value from the month with

the most outliers. For this task, d means that if the winning month had x outliers, the other months had at most $x - d$ outliers. We used ds of 1,2,3,4, with $d = 5$ for validation.

Encoding had a significant main effect ($F(5, 1488) = 28.3$, $p < .0001$). A Tukey HSD showed two clusters - event striping outperformed all other displays ($F(1, 1488) = 127$, $p < .0001$). The only other significant difference among conditions was the color woven display, which under-performed all of the remaining conditions ($F(1, 1488) = 11.4$, $p = .0008$).

These results show that, by explicitly devoting space to outliers, event striping outperformed all other encodings.

Results Summary

Across all tasks and encodings there are several interesting trends. Figure 3.11 presents these results in a single chart. In particular, while *explicitly encoding relevant statistics* results in good performance, the affordances of different encodings may dominate even explicitly encoding the values of interest. For instance linegraphs (which do not directly encode the maximum value, except in as much as the viewer can search to find the highest point) outperformed color stock charts (which directly encode extrema) for extrema-finding tasks, due to the strength of positional encodings for displaying individual values. By contrast, color encodings performed better for aggregate tasks (such as comparing average and spread) than positional encodings, unless those positional encodings directly encoded the statistics of interest. Also of interest to the project of visual statistics is that some statistical tasks (outlier detection) were so complex or esoteric that only purpose-built encodings generated passable performance in viewers.

3.5 Extending Color Weaving

Our implementation of the color weaving technique intentionally sacrifices the exact temporal fidelity of displays in order to afford the quick apprehension of distributional information. In the examples in this chapter, these disruptions of local temporal structure occur within discrete, month-long "blocks." In many tasks the metaphor of a block is imperfect – viewers might care about a scale smaller or

larger than the block size, or might need to make judgments across or within blocks. In our current applications employing color weaving, (such as [Albers et al. 2011]) users interactively choose the size of blocks, but this only partially addresses these limitations.

It is possible to extend the color weaving algorithm evaluated in this chapter to cover a larger variety of use cases, where scales of interest are unknown, or viewers must make judgments about regions that are not defined discretely. While none of these extensions are parameter-free, reasonable choices of parameters, in piloting, have generated performance on par with the other colorfield displays explored in this chapter.

Lens Weaving

One alternative is to let the user of a system choose the location and extent of the woven region on the fly. Fig. 3.12 shows an example of this technique - while most of the series is presented normally, an interactively placed "lens" creates a region of interest in which all points are woven. The user determines both the width of the lens, as well as the lens' horizontal (temporal) location. This technique allows the viewer to adapt to new data and new tasks without input from the designer. However, this interactivity comes at a cost – static displays have different affordances than dynamic displays, and forces the burden of navigation of both data space and parameter space upon the viewer. If the user wishes to compare distributional information across different locations, the user must either create multiple lenses (which might require stacking if the regions are overlapping), or move a lens manually and rely on memory.

Icicle Weaving

In many cases a single *a priori* block scale is not appropriate for a particular set of tasks, but there are still *sets* of different block scales that might be useful across tasks. We can use these sets to divide a series into a tree-like structure of blocks of different scales instead of focusing on any single scale. Inspired by icicle plots for displaying hierarchies [Kruskal and Landwehr 1983], icicle weaving stacks woven displays at different temporal "resolutions" atop one another. Fig. 3.13 presents

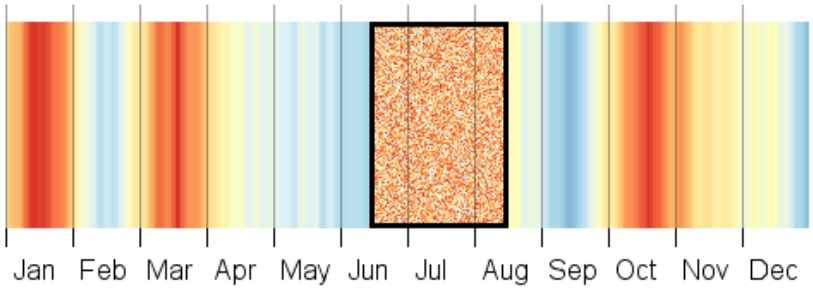


Figure 3.12: Lens weaving technique for interactive investigation of data without pre-computed blocks. Outside of a focus region, the display is a standard colorfield. Inside the region of focus (in this case a the 60-day window centered in July), all pixels are permuted as with a standard woven colorfield. This dynamic selection of focal regions affords quick apprehension of distributional information even when there are not discrete or mutually exclusive areas of interest.

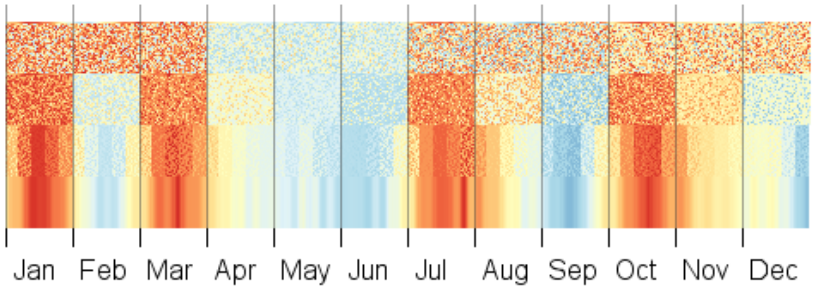


Figure 3.13: Icicle weaving performs color weaving at multiple temporal scales and stacks the resulting graphs atop each other. In this example, there is a layer for quarters of the year, months, weeks, and days. By examining different layers viewers can make distributional judgments across multiple scales. Note how local trends (such as the relatively low values in February) can be erased or reversed by aggregating at different temporal scales.

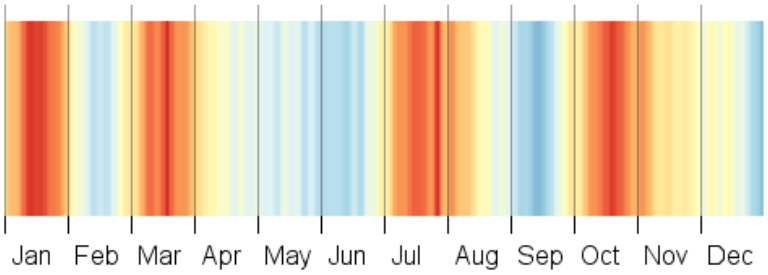
an example where the resolutions are quarters, months, weeks, and days of a year. While this allows comparison across and within different scales, the technique is still reliant on discrete blocks – if no one layer has the "correct" scale or partitioning for a particular task, viewers must build up a picture of the distribution of a region by aggregating regions in the lower layers. This complexity is compounded by the fact that each layer comes at a cost of screen space – we cannot display every potential scale of interest simultaneously without running out of pixels.

Continuous Weaving

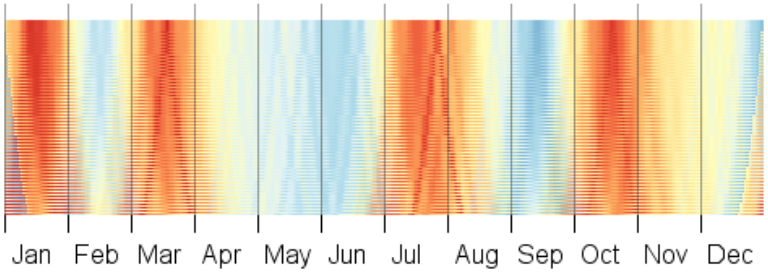
With some modifications to the color weaving algorithm it is possible to dispense with discrete blocks entirely. In blocked weaving, any pixel in the original colorfield image equally likely to end up in any particular location within a block. By using a sampling kernel instead of block boundary, we can weight the probability of where a pixel might move such that most pixels end up close to their original location. For each column of the original colorfield, pixels are drawn from distances as defined by the sampling kernel, and then the entire column is permuted. Fig. 3.14 illustrates this process with a Gaussian sampling kernel. The resulting image resembles a low-pass filtered version of the original image, and yet is a strict permutation of the original image. A continuously woven image breaks local structure and promotes the estimation of aggregates while not depending on a priori knowledge of task scale. A drawback to this technique is that it is sensitive to the choice of kernel; both the type of function (e.g. a binomial, Gaussian, tent, or box kernel) and the parameters of that function (e.g. the width or variance of the kernel). As with lens weaving, interactive control of these parameters by the user might mitigate this sensitivity, but in many cases the choice of these values is not intuitive, especially for a novice user. As a compromise point between fully blocked weaving and the original color series, it is also possible that this technique does not go far enough in creating regions with easy to compare distributional information, especially when meaningful blocks *do* exist — in such cases values from discrete sub regions would "bleed" into their neighbors, complicating the comparison process. Initial piloting has shown only marginal improvement in the monthly mean comparison task for continuously woven fields over standard colorfields.

3.6 Discussion

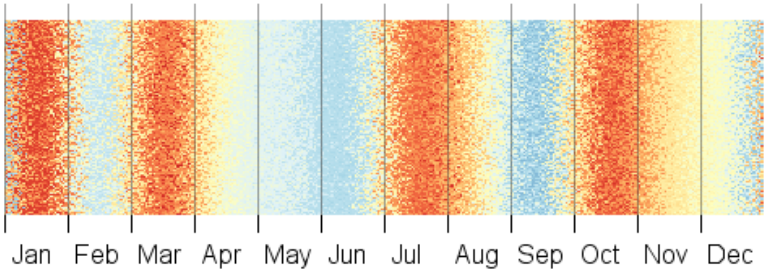
Our experiments provide insight into designing displays that support visual aggregation by investigating how specific design elements influence viewers' abilities to compare averages within time series data. Our results confirm that color encodings better support visual averagings than more conventional approaches. In particular, color weaving provides greater accuracy for average comparison than all other tested



(a) Original Series



(b) Gaussian Sampling



(c) Final Permutation

Figure 3.14: Continuous color weaving technique for generating woven colorfield displays without *a priori* knowledge of task scale. Pixel values for each column of the series (one column per day of the year) are offset with respect to a Gaussian - the modal pixel in a column is unmoved, but some pixels are drawn from farther away in the series, with the number decreasing with distance from the column (Fig. 3.14b shows intermediate sampling). The pixels within a column are then internally permuted (Fig. 3.14c), maintaining the general structure of the original series and highlighting the average color value of regions at the expense of exact positional fidelity.

encodings. This performance is robust across a wide variety of designs and task difficulties. We explored alternate explanations for the success of this encoding (permutation and visual separation) and found that color weaving's success cannot be explained by either of these visual components alone. Viewers were likely leveraging their ability to efficiently average small regions of color, an affordance provided by standard color encodings and enhanced through color weaving.

Color weaving allows viewers to accurately extract mean values from pre-defined subsections of a time series. While this ability is present in existing line graph encodings, it is less precise than in colorfield displays. While the choice of color to encode value over position improves performance, the permutation of data across individual aggregate blocks provides additional performance benefits over our other choices of encoding. However, this permutation relies on a priori knowledge of task scales. We have proposed several techniques that leverage the benefits of color weaving without necessarily knowing the viewer's task in advance. We plan to evaluate these new designs in future work.

Limitations

Our experimental tasks were somewhat artificial, dealing only with the comparison of aggregate values for specific, predefined areas. A more thorough analysis of how the color weaving technique generalizes to higher order tasks and larger scales of data remains future work. An empirical evaluation of color weaving extensions also requires new experimental designs that are not within the scope of this work.

3.7 Conclusion

In this work we show that, more than simply extracting individual values or low-level details from a display, even relatively untrained viewers can extract meaningful high-level patterns from data. We explore an encoding, color weaving, that has been designed and evaluated to harness this ability to perceptually aggregate visual features for making high-level judgments. More than just determining the mean for particular blocks of data, we believe this encoding is an example of a general class of techniques where high level properties are emphasized.

This work highlights the strengths and weaknesses of positional and color-based encodings of series data. In some cases the standard position encodings underperform the more novel approaches, namely in aggregate averaging tasks where the exact location of temporal data is less important than the quick apprehension of general overviews of a particular time series. Our previous research and implementations of the color weaving technique also indicate that weaving is useful in the case of dense ordered series, giving viewers a general overview of the data even if there is insufficient visual space to explicitly encode each sample.

The experimental tasks in this chapter, and Gleicher et al. [2013], could be measured in terms of accuracy — there was a correct decision to be made for each stimulus. For many real decision tasks we are making predictions about future behavior, which may or may not have a correct answer. These decisions must take into account not only the judgments of viewers, but also their confidence in their judgments. In the following chapter we examine a task where, once again, the viewer (who may lack statistical training) is using visualizations to make a statistically sophisticated decision, but in this case we are forced to use certainty and confidence to measure and compare performance.

4 EVALUATING ENCODINGS FOR UNCERTAINTY

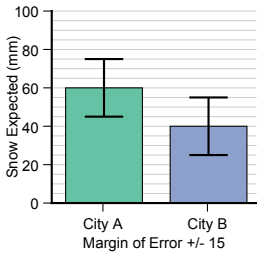
For real world decisions, the uncertainty associated with the data can be as important as the difference in data values. Big differences in data values may not be significant or interesting if there is too much error: for instance too much noise, uncertainty, or spread. Techniques from inferential statistics (including comparison of interval estimates, null hypothesis significance testing, and Bayesian inference) address this issue, but can be complicated, counter-intuitive, or equivocal. Careful design could produce visualizations which convey the general notion of varying levels of error even when the viewer does not have a deep statistical background.

The most common encoding for sample means with associated error is a bar chart with error bars. Despite their ubiquity, many fields (including perceptual psychology, risk analysis, semiotics, and statistics) have suggested severe shortcomings with this encoding, which could result in decisions which are not well-aligned with statistical expectations. While alternate encodings for mean and error have been proposed, to our knowledge none have been rigorously evaluated with respect to these shortcomings.

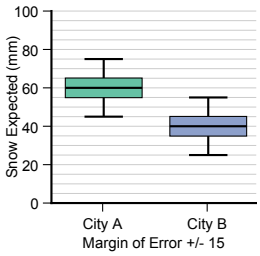
In this chapter we investigate how differences in the presentation of mean and error data result in differing interpretations of viewer confidence and accuracy for judgment tasks. We investigate the drawbacks of the standard encoding for mean and error, bar charts with error bars. We investigate standard practices for depicting mean and errors. We present and evaluate alternative encoding schemes for this data (see Fig. 4.1). Lastly, we present the results of a crowd-sourced series of experiments that show that bar charts with error bars, the standard approach for visualizing mean and error, do not accurately or consistently convey uncertainty, but that changes in design can promote viewer judgments and viewer certainty that is more in line with statistical expectations, even among a general audience.

Work from this chapter was originally published in Correll and Gleicher [2014].

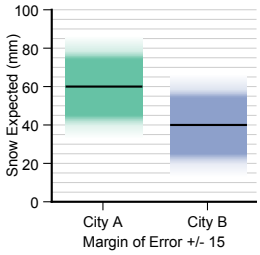
Contributions: We present a series of issues with how the standard encoding for mean and error, bar charts with error bars, are interpreted by the general audience. We adapt established encodings for distributional data — violin plots[Hintze and Nelson 1998] and gradient plots[Jackson 2008] — for tasks in inferential statistics. We validate the performance of these encodings with a series of crowd-sourced



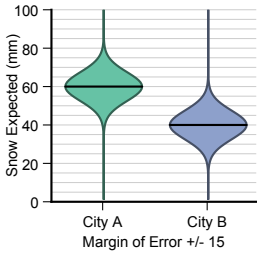
(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.



(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.



(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.



(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

Figure 4.1: Four encodings for mean and error evaluated in this chapter. Each prioritizes a different aspect of mean and uncertainty, and results in different patterns of judgment and comprehension for tasks requiring statistical inferences.

experiments.

4.1 Background

Issues with the presentation of mean and error, especially with bar charts with error bars, have been studied by multiple fields, including psychology, statistics, and visualization. We present a summary of these findings. We provide evidence that, while visualizations of mean and error are valuable, care must be taken in how they are designed and presented, especially to a general audience. We show with an analysis of practices in information visualization and elsewhere that audiences with a wide range of expected statistical backgrounds are nevertheless presented with mean and error data in similar ways. Despite the drawbacks we present, we

confirm that bar charts with error bars are the modal encoding for presentation of this sort of data in the information visualization community.

Visualization of Mean and Error

Mean and error, as in confidence intervals or error bars, has been proposed as a solution for some of the perceived deficiencies in traditional significance testing [Nickerson 2000], both for pedagogy and in analysis [Schmidt 1996]. Unfortunately, while inferential statistics might offer techniques for approaching complex problems, human reasoning (especially in matters of statistics and probability) operates via a series of heuristics that may or may not arrive at the "right" answer. Tversky and Kahneman [Tversky and Kahneman 1974] offer examples of systematic errors these heuristics generate for decision problems based on uncertain data. An example as applied to the information visualization community is the "fallacy of availability" — we remember dramatic or remarkable events with greater ease than ordinary ones, skewing our perception of base rates. For example, a technique which provides good results in most cases but fails catastrophically for a particular case might be seen as more unreliable than a technique that has more frequent, but less severe, failures. Inbar [Inbar 2009] provides evidence that how we visually encode uncertainty and probability can work to "de-bias" data which would ordinarily fall prey to an otherwise inaccurate set of heuristics (by comparison to an outcome maximizing classical statistical view) . Designing visualizations to support decision-making and perform de-biasing is not trivial, and how the task is laid out in text can conflict with attempts to de-bias [Micallef et al. 2012]. Even so, the visual presentation of uncertainty can promote better understanding than textual presentation [Lipkus and Hollands 1999].

Error bars, the common way of encoding uncertainty or error, have a number of additional biases, some in concert with other common encodings types. One is ambiguity — an error bar can encode any number of values, from range to standard error. In many cases the error bars are not explicitly labeled, or are labeled in text that is visually distant from the chart in question. This ambiguity, combined with widespread misconceptions about inferential statistics, means that even experts in fields that frequently use error bars have difficulty perceiving how they are

connected to statistical significance, estimating p values that are incorrect by orders of magnitude [Belia et al. 2005]. For error bars with bar charts, the most common combination of mean and error, since bars are large, graphically salient objects that present the visual metaphor of "containing" values, values visually within the bar are perceived as likelier data points than values outside of the bar [Newman and Scholl 2012]. Lastly, by presenting error bars as discrete visual objects, designers emphasize an "all or nothing" approach to interpretation —values are either within the bar or they are not. By only showing information about one kind of statistical inference, viewers are unable to draw their own conclusions for their own standards of proof, exacerbating existing problems with null-hypothesis significance testing [Cohen 1994, Johnson 1999, Schmidt and Hunter 2013].

Mean and Error in General Practice

Since mean and error are critical for decision-making based on uncertain data, different communities have codified different approaches to communicating these values, while highlighting the importance of communicating both mean and error to audiences. This is true of both the psychology community, where the audience is assumed to have at least a basic understanding of statistical inference, and also in the journalism and mass communication community, where statistical expertise cannot be assumed.

The American Psychological Association recommends that point estimates "should also, where possible, include confidence intervals" or other error estimates. Furthermore, they should allow the reader to "confirm the basic reported analyses" and also to "construct some effect-size estimates and confidence intervals beyond those supplied" [Association 2005]. More recently, the APA has pushed for the greater use and reporting of intervals, as opposed to significance testing [Wilkinson 1999].

The Associated Press also recommends reporting the margin of error in polling data (in practice, the 95% t-confidence interval) [Goldstein 1994]. Since p-values are not common concepts for a general audience, they recommend stating that one candidate is leading if and only if the the lead is greater than twice the margin of error (in practice this is an α value of less than .01). The existence of these guidelines (and the similar reporting and summarization of model and measurement uncertainty in

popular, general audience websites such as `http://fivethirtyeight.com` and `http://www.pollster.com`), indicates that the display and interpretation of inferential statistics is a problem that extends beyond the academic community.

Mean and Error in InfoVis

The information visualization community contains members with heterogeneous backgrounds who have different internal statistical practices but nonetheless must report inferential statistics in a mutually intelligible way. We believed that the visualization of mean and error within the community would offer both an example of statistical communication meant for general audiences, as well as provide a diverse set of potential visual designs for communicating statistics. To that end we analyzed the visual display of sample mean and error in the past proceedings of accepted IEEE VisWeek papers in the InfoVis track, 2010-2013. In the 163 papers available, 46 had some visual display of sample means (usually in the context of evaluating the performance of a new visualization tool). Of these 46 papers, 36 (approx. 78%) used error bars to encode some notion of error or spread. The modal encoding was a bar chart with error bars, which occurred in 26 (approx. 56%) of the papers. Boxplots were also common (7 papers, approx. 15%), as were dot plots with error bars (5 papers, approx. 10%).

There was a heterogeneous use of error bars across papers. In many cases the error bars were unlabeled (22 papers, approx. 48%). This is despite the fact that error bars can be used to represent many different quantities. In the papers we found, error bars were labeled as range, 95% confidence intervals, 80% confidence intervals, standard error, standard deviation, or 1.5$\times$ the interquartile range (IQR). Should one wish to use these error estimates to estimate statistical significance (a practice which is controversial [Schmidt and Hunter 2013]), each of these interpretations of error would necessitate a different heuristic for "inference by eye" — that is, a different way to determine the relative significance of different effects [Cumming and Finch 2005]. Given this ambiguity, a common practice was to denote statistically significant differences with an asterisk; however as the number of sample means increases, the number of glyphs required to explicitly encode all statistical significant pooled sample t-tests increases exponentially. Even if the number of comparisons

is small, the link between graphical overlap of confidence intervals and the results of significance testing decays, and the probability of Type I errors increases (*cf.* techniques such as the Benferroni correction that attempts to correct for the increased likelihood of Type I errors as the number of comparisons increases).

4.2 Alternatives to Bar Charts with Error Bars

There are many potential designs for mean and uncertainty. Potentially any visual channel can be combined with another encoding to unite a "data map" and an "uncertainty map" [MacEachren 1992]. We chose two potential encodings for this data based on current practices for displaying probability distributions, tweaked for the specific use-case of inferential statistics: the gradient plot (which uses α transparency to encode uncertainty), and violin plots (which use width). Neither encoding is particularly common in the information visualization community — only a version of the violin plot, in the form of a vertically-oriented histogram, was found in our search of InfoVis conference papers. We believe this rarity is beneficial for this problem setting, since existing semantic interpretations might interfere with our intended use and meaning of these encodings for this problem (which is similar to, but sufficiently different from the standard visualization problem of visualizing distributions). That is, we do not want viewers to confound visualizations of the distribution of *error* and the distribution of *data*.

Processing code for generating all of the plots seen in the paper is available in the supplementary materials.

Design Goals

The recommendation of style manuals designed for the presentation of results to diverse audiences, combined with the heterogeneity of real world uses of mean and error data, led us to formulate a series of goals for any proposed encoding of mean and error:

- The encoding should clearly present the effect size — that is, accuracy at visualizing error should not come at the expense of clear visualization of the mean.

- The encoding should promote "the right behavior" from viewers (such as refraining from judgment if means are dissimilar but error is very high), even if the viewers lack extensive statistical training. Likewise, viewer confidence in judgments ought to correlate with power of the relevant statistical inferences. These effects should apply to different problem domains and framings.
- The encoding should afford the estimation or comparison of statistical inferences that have not been explicitly supplied.
- The encoding should avoid "all or nothing" binary encodings — the encodings should permit different standards of proof other than (for instance) an α of 0.05. This will likely require encodings which display confidence *continuously*, rather than as discrete levels.
- The encoding should mitigate known biases in the interpretation of error bars (such as within the bar bias, and mis-estimation of error bars due to the presence of central glyphs). This will likely require encodings which are *visually symmetric* about the mean.

To fulfill these goals we adapted two existing encodings (usually used for visualizing distributional information), violin plots and gradient plots, for use in inferential tasks. Box plots, as a standard encoding for distributional data, are discussed as a separate case. We believe that these encodings fulfill the design goals presented above. In addition, since they both adapt general techniques for the visualization of distributions, they can be adapted to many different error statistics beyond the t-confidence intervals presented here.

Gradient Plots

Jackson [Jackson 2008] argues for using color to encode data such as probability distributions functions (pdf). In that technique, and in similar techniques, a sequential color ramp is used to encode likelihood or density, usually varying the α, brightness, or saturation. Low saturation and low α values have a strong semiotic connection with uncertainty [MacEachren et al. 2012], and thus are a commonly used visual metaphor for conveying uncertain data [Gershon 1998]. Recent research

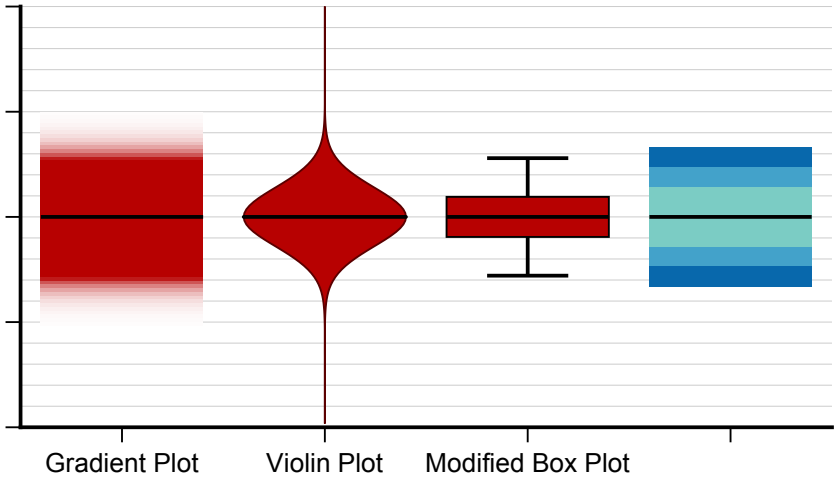


Figure 4.2: The alternate plots we propose for encoding mean and error. From left to right gradient plots, violin plots, and modified box plots. The colored bars on the right are standard error, a 95% t-confidence interval, and a 99% confidence interval, for reference.

has shown that using gradients in this manner affords robust understandings of uncertainty even for general audiences [Toet et al. 2014]. We call this specific technique a "gradient plot."

Our version of the gradient plot differs slightly from the standard approach, which is to take the density trace and map each value to a color. We wished to keep some connection with the discreteness of error bars, and so all values within the 95% two-tailed t-confidence interval are fully opaque. Outside of the margin of error, the α value decays with respect to the cumulative probability for the absolute value of the y coordinate based on an underlying t-distribution. That is, the α value of a particular y coordinate is linearly related to the size of the t-confidence interval needed to reach that value — a 95% confidence interval is fully opaque, and the (fictional) 100% confidence interval would be fully transparent. In practice, since the inverse cumulative probability function decays so rapidly, there is a block of solid color surrounded by "fuzzy" edges. Figure 4.2 shows a sample gradient plot in more detail. Viewers are not very proficient at extracting precise α values, and perhaps can only distinguish only a few different "levels" of transparency [Boukhelifa et al. 2012]. Issues with interpreting α values are exacerbated by the non-standard ways in which tone and transparency are reproduced between displays. Nonetheless, we believe that this imprecision is a "beneficial difficulty" [Hullman et al. 2011] as it

discourages artificially precise comparisons where there is a great deal of uncertainty associated with the data. In general we believe the gradient plot is superior to the standard bar chart with error bars for a number of reasons:

- A visual metaphor that aligns with expected behavior: minimal transparency (and so uncertainty) within the 95% confidence interval, quickly decaying certainty outside of that region. This extends to comparison: if two samples are very statistically similar than their "fuzzy" regions will overlap.
- Use of a continuous but imprecise visual channel provokes a "willingness to critique" [Wood et al. 2012] in a way that discrete but precise encodings or styles do not.
- Visual symmetry about the mean, mitigating "within-the-bar" bias (the tendency to see values visually contained by the bar chart as being likelier than values outside the bar).

Violin Plots

Hintze and Nelson [1998] proposed "violin plots" for displaying distributional data. In the canonical implementation, a density trace is mirrored about the y axis, and then a box plot is displayed inside the region, forming a smooth, violin-like shape with interior glyphs. "Bean plots" replace the interior box plot with lines representing individual observations [Kampstra 2008]. In either case the additional level of detail affords a quick judgment about the general shape of the distribution (*cf.* a unimodal and a bimodal distribution which might have identical box plots but would have vastly different violin plots). Width and height are both positional encodings of distributional data: position as a visual channel has higher precision than color for viewer estimation tasks. Ibrekk & Granger [Ibrekk and Morgan 1987] confirm this inequality for the case of violin plots of probability distributions specifically.

Our version of the violin plot for inferential statistics discards the interior glyphs and encodes the probability density function rather than the sample distribution. We believe that the distribution used to make inferences is more valuable for these tasks than the distribution of the data themselves. Figure 4.2 shows a sample violin plot

of the design used in our study. The pdf is not intrinsically relevant to a significance test, which tends to rely on the cumulative distribution function, or cdf. Initial piloting with the symmetric cdf version of violin plots (where the width of the violin encoded the likelihood that the absolute value of the y position is greater than or equal to the mean) were confusing for the general audience compared to the relatively straightforward pdf violin plots. The general visual metaphor, namely that as we move away from the mean, values become less likely, is maintained even in the pdf version. Additionally, previous work has shown that viewers are capable of aggregating regions of a line graph with some precision [Correll et al. 2012b], affording both cdf- and pdf-reliant judgments. We believe that violin plots used in the way we propose have a number of benefits over standard bar charts with error bars:

- Affordance of comparison of values beyond the discrete “within the margin of error/outside of the margin of error" judgments afforded by bar charts with error bars.
- Use of a strong, high fidelity visual encoding (position) to afford precise readings of the pdf.
- Visual symmetry about the mean, mitigating within-the-bar bias.

Box Plots

There are several classes of visualizations of distributions which are more common than the two we propose, such as box plots, or dot plots with error bars. While we have included box plots in our evaluation, and they meet many of the design goals above, we believe they are unsuitable for our task. The largest problem is that they are both commonly and popularly used to encode the actual distribution of data. For this problem we do not encode the distribution of data, but (in this case) the distribution of a potential population mean given a sample and certain statistical assumptions. Most commonly used probability distributions (including the Student’s t distribution, the normal distribution, &c.) are unimodal and have infinite extent, while the data about which we are making inferences may not. A box plot as commonly used to depict the distribution of the data is thus several analytical

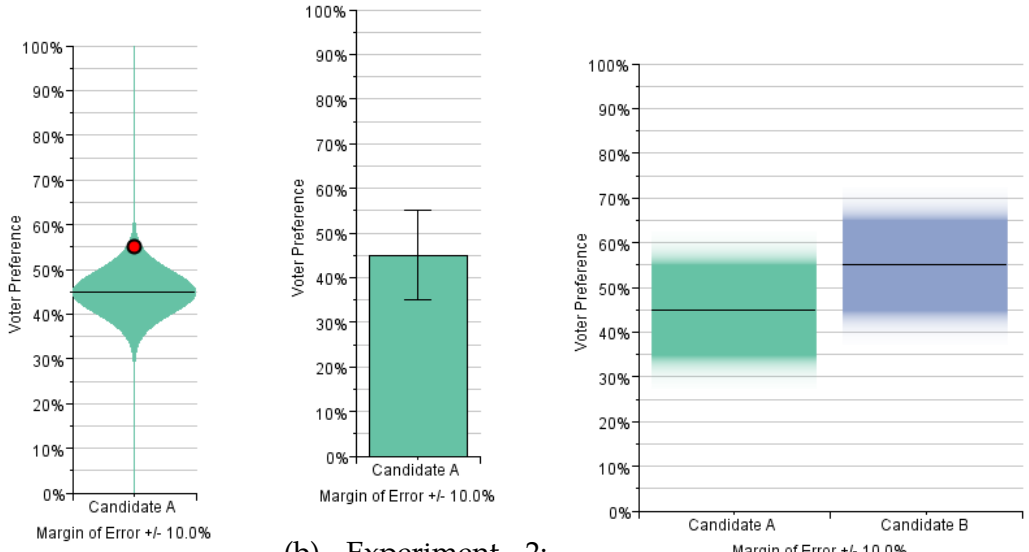
steps removed from a confidence interval. Standard choices in box plots also conflict with our desire to visualize a distribution of population means — whiskers are a form of error bars, but in a box plot whiskers usually denote range or 1.5× the interquartile range (although there are exceptions to this convention, see 4.1). If the first convention is used, then the whiskers of a t-distribution would extend infinitely far along the y-axis. Lastly, there is a perceptual illusion in box plots where large boxes make viewers underestimate the length of error bars, and overestimate the length when boxes are small [Stock and Behrens 1991].

Nonetheless, box plots are a popular encoding for distributional data, with many extensions to show a wide variety of complex and higher-order statistics [Potter et al. 2010]. In order to adapt box plots to an inferential rather than descriptive role we made several modifications. The first is that we chose to visualize the pdf of interest rather than the data. The whiskers are the margins of error, in this case the 95% t-confidence interval. We calculate the extent of the box (normally bound by the first and third quartiles of the data) by calculating the inverse cdf at points 0.25 and 0.75 (i.e. the locations which are equivalent to 25 and 75% of the of the indefinite integral of the pdf, which is analogous to quartile locations). The center line of the box is the mean. Figure 4.1 shows an example of a box plot modified in this fashion. We believe that this modification captures the "spirit" of box plots while still being relevant to the task at hand. We believe that even these modified box plots will have the following advantages over bar charts with error bars:

- Additional levels of comparison — while for bars charts with error a y location is either inside the error bar or is not, for box plots there are three such levels (outside the error bar, inside the error bar, inside the box). A point inside the box is within a 50% t-confidence interval from the sample mean.
- Visual symmetry about the mean, mitigating "within-the-bar" bias.

4.3 Evaluation

The goals of our evaluation were three-fold: to see if general audiences would make decisions that were informed by both mean and error, to assess how certain biases which affect how bar charts with error bars are impacted by our proposed alternate



(a) Experiment 1: How likely is the red outcome?

(b) Experiment 2: How likely is the outcome where candidate A gets 55% of the vote?

(c) Experiment 3: How likely is candidate B to win the election?

Figure 4.3: Example stimuli from our experiments. Each presents tasks which are similar in concept, but deal with different aspects of the visual presentation of statistical inference. The graphs are presented as violin plots, bar charts with error bars, and gradient plots, respectively, but all experiments tested multiple graph types.

encodings, and to assess other strategies for mitigating these biases. Our results confirmed that our proposed encodings offered concrete benefits over bar charts with error bars. We report on three experiment sets here:

- Our experiment with **one-sample judgments** presents participants with a single sample mean, postulates a potential outcome (in the form of a red dot), and asks participants to reason about the relationship of this potential outcome to the sample. Our hypothesis was that bar charts are subject to "within-the-bar" bias (where points contained by the bar are seen as likelier than points outside the bar), even for inferential tasks, but that alternate encodings (violin plots and gradient plots) would mitigate this bias.

- Our experiment with **textual one-sample judgments** evaluates another potential approach to mitigating within-the-bar bias, which is to abstract some of the

information from the bar chart itself into text (that does not have the metaphor of visual containment). Our hypothesis was that this approach would be ineffective, and would introduce unacceptable inaccuracy in comparisons.

- Our final experiment with **two-sample judgments** evaluates our alternative encodings in a setting that resembles how these visualizations are frequently used in practice: to compare samples and make predictive inferences about the differences in mean, given the error. Our hypothesis was that viewers with limited statistical backgrounds would be able to make assessments in a way that resembles statistical expectation, but that our alternate encodings would provide a better pattern of performance.

General Methods

We conducted a series of experiments using Amazon's Mechanical Turk to evaluate the performance of different graphical encodings for inferential tasks. Participants were recruited solely from the North American Turker population. Participants were exposed to a series of different graphs and asked to complete a set of tasks per graph. Since domain knowledge and presuppositions can alter the visual interpretation of graphs [Trafton et al. 2002], another factor was the framing of the problem: samples were represented as either polling data ("Voter preference for Candidate A"), weather forecast data ("Snowfall predicted in City A"), or financial prediction data ("Payout expected by Fund A"). The experiments were a mixed model 4 (bar,violin,gradient or boxplot encoding) $\times$ 3 (polling, finance, or weather framing) $\times$ 6 (mean differences) $\times$ 6 (margins of error) design, where the type of encoding seen and the framing problem were both between-subjects factors, but the distances between means and size of margins of error were within-subjects (participants saw multiple, balanced levels of different sample means and margins of error). In all experiments where we varied the problem framing, it was a significant effect, so it was included as a covariate in our analyses. Including piloting (which includes the results presented in [Correll and Gleicher 2013], which had a slightly different study design) we recruited 368 total participants. A total of 240 participants were involved in the presented experiments, of which 102 (42.5%) were male, 138 female, (average age = 33.3, $\sigma = 10.2$). Of the involved participants, 90 had some college education, 110 were

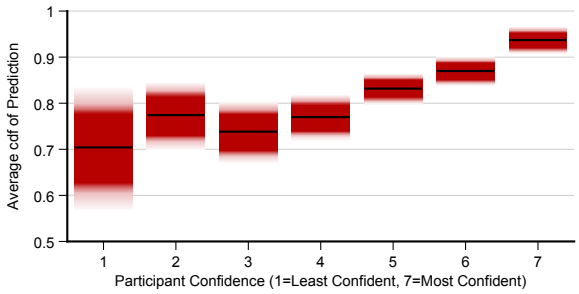
college graduates, and 31 had post-graduate degrees — the remainder were high school graduates with no college experience. Each participant for each experiment saw a total of 36 graphs in sequence. Participants were given no explicit time limit to complete the experimental task, but the median participant took approximately 8 minutes (approx. 14 seconds per graph) to complete the task. We used ColorBrewer [Harrower and Brewer 2003] to select colors for the stimuli. Figure 4.3 shows example stimuli and tasks from each of the three experiment sets we present.

We include data tables, example stimuli, and screenshots of our experimental setup online at `http://graphics.cs.wisc.edu/Vis/ErrorBars`. F and p-values reported in the results sections are from two-way analyses of covariance (ANCOVAs) unless otherwise stated.

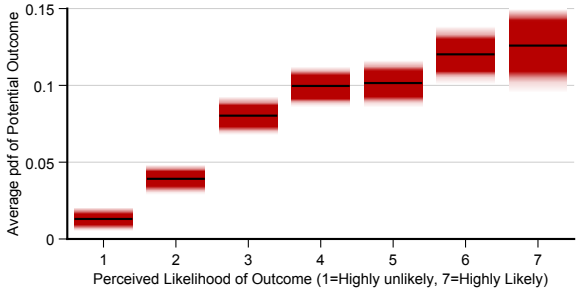
Experiment 1: One-sample judgments

"Within-the-bar" bias as originally proposed is a bias dealing with descriptive statistics: a sample mean is made up of points, points far from the mean are less likely to be members of the sample, but the visual area of a bar in a bar chart creates a region of false certainty. We believed that due to visual metaphor of bar charts, something similar would occur for tasks involving statistical inferences. We believed our alternate encodings, by using a different visual metaphor, would not create this bias.

In this experiment, participants were shown a series of 200x400 pixel graphs, each with one sample value and an associated margin of error. For each graph a red dot was plotted at some set distance from the mean ($\pm$ 5,10, or 15 units in a 100 unit y-axis). The experimental task dealt only with the interaction between the red dot, the difference from the mean, and the margin of error. Piloting confirmed no significant effect of sample mean on task response, so sample means were randomly selected from the set {35,40,45,50,55,60,65}. There were 6 different levels of margin of error {2.5, 5.0, 7.5,10.0,12.5,15.0}. Each participant saw 36 graphs, 6 per margin of error. There were 3 different levels of the between-subjects encoding factor (violin plot, gradient plot, or bar chart). There were also three levels of problem frame (election, weather, or financial data). The wording of task questions were slightly altered to fit the problem frame. The participants had three main task questions.



(a) Aggregate cdf values of all the stimuli that participants associated with a particular prediction confidence level. A dot on the sample mean would have a cdf of 0.5, representing the zero point.



(b) Aggregate pdf value of all stimuli that participants associated with a particular outcome likelihood.

Figure 4.4: Gradient plots of our results from the one-sample judgments experiment (§4.3). Participants were shown a sample mean with error, and a red dot representing a proposed outcome. They were asked to predict whether or not the population outcome was likely to be lower or higher than the red dot, and then asked for their confidence in this prediction. This response is analogous to a question about the cdf of the t-distribution (4.4a). They were also asked how likely the red dot was, given the sample mean. This response is analogous to a question about the pdf of the t-distribution (4.4b).

Verbatim from the election problem frame:

1. How do you think the candidate will perform in the actual election, compared to the red potential outcome? (Fewer votes, more votes)

2. How confident are you about your prediction for question 1, from 1=Least Confident, 7=Most Confident?

3. How likely (or how surprising) do you think the red potential outcome is, given the poll? From 1=Very surprising (not very likely) to 7=Not very surprising

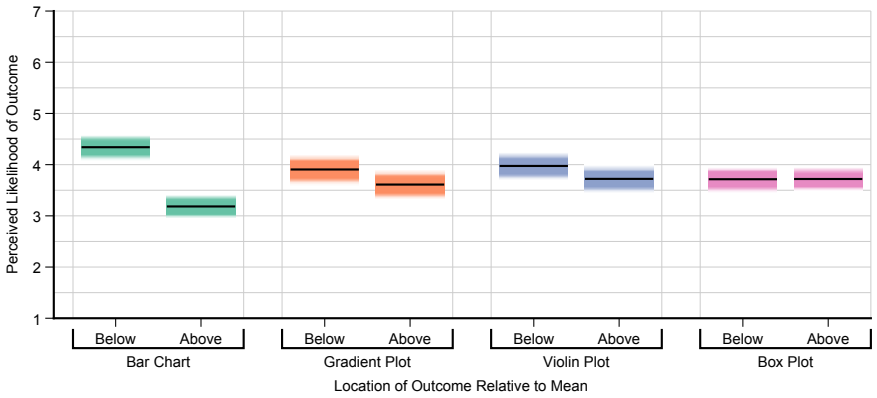


Figure 4.5: A gradient plot of results from our one-sample judgments experiment (§4.3). Participants were shown a red dot representing a potential outcome and judged how likely this outcome was given the sample mean and the margin of error. Statistical expectation is that likelihood would be symmetric about the mean — that is, red dots above the sample mean would be perceived as just as likely as those below the mean. For bar charts this is not the case — points visually contained by the glyph of the bar (below the sample mean)were seen as likelier than those not contained by the bar. Visually symmetric encodings mitigate this issue.

(very likely)

The expected behavior (based on statistical expectations) for question 1 is to predict that the sample mean is an accurate estimate of the actual mean (so if the red dot is above the sample mean, you would expect that candidate A would receive fewer votes in the actual election). If this strategy is followed, then question 2 (which is contingent on the guess for question 1) is somewhat analogous to a question about the cumulative density function: what proportion of the probability space is above (or below, depending on the answer to question 1) the red dot? Question 3 by the same reasoning is somewhat analogous to a question about the probability density function. Our hypotheses were:

H1 Participant responses would generally follow expected behavior. That is, participant responses to question 1 would "follow the sample mean" — if the red dot is above the sample, assume the real election will be lower than the red dot, and vice versa. The answers to question 2 should correlate with the cdf of the t distribution given the data, and the answers to question 3 should correlate with the pdf. Both cdf and pdf are modulated by both the difference in value

between the predicted outcome and the sample mean, and the margin of error of the sample.

H2 The non-symmetric encoding (bar charts) would exhibit within-the-bar bias — proposed outcomes within the bar would be seen as likelier than outcomes outside of the bar. Symmetric encodings (box, violin, and gradient plots) would not have this bias.

H3 The proposed encodings, which encoded the t-distribution in a non-binary way (gradient and violin plots), would provide more accurate and more confident judgments about the t-distribution than the binary encodings (bar charts and box plots).

Results

We recruited 96 participants, 8 for each combination of problem frame and graph type. We determined significance through two sets of two-way ANCOVAs, testing for the effect of different encodings and data values on confidence in estimating cumulative probability, and estimating the probability density. We included whether the red dot was above or below the mean as a factor as well, and its interaction with the graph type, to explicitly test for "within-the-bar" bias. Inter- and intra-participant variance in performance was included as a covariate, as was problem frame.

Our results **generally support H1**: We expected participant answers on question 1 to follow the sample mean, and in general this strategy was followed in 87.1% of trials (but see H3 results below).

We expected participant answers on question 2 (reported confidence) to follow the cdf. That is, the perceived confidence that the election would have results below a certain proposed result would be correlated with the cdf of the t-distribution, and the perceived confidence that the election would have results higher than a certain value would be 1- the cdf at that location. Indeed, the relevant value (the cdf if the participant predicted the real outcome would be less than the proposed outcome, 1-cdf otherwise) was a significant main effect on reported confidence ($F(1,3359) = 55.6$, $p < 0.0001$). Participant's average reported confidence was positively correlated with the relevant value of the cdf ($R^2 = 0.805$, $\beta = 6.78$). Figure 4.4a

shows the relationship between answers on question 2 (how confident are you in your prediction?) and the actual cdf values of responses.

We expected participant answers on question 3 (reported likelihood of the proposed outcome) to follow the pdf. That is, the perceived likelihood of a dot plotted on the graph should correlate to the value of the probability distribution at that point. The value of the pdf was only a marginal effect across all results ($F(1, 3359)=3.05$, $p = 0.081$), but was a significant effect for trials where the participant followed the correct strategy for question 1 ($F(1,3361)=30.2$, $p < 0.0001$). Participant's average judgments about the likelihood of outcomes was positively correlated with the pdf values of the stimuli presented ($R^2 = 0.842$, $\beta = 5.70$). Fig 4.4b shows the relationship between responses on question 3 ("how likely is this proposed outcome?") and the actual pdf values.

Our results **support H2**: We observed a significant interaction between the position of the dot (above or below the mean) and encoding ($F(2,2)=21.3$, $p < 0.0001$) on the perceived likelihood of the dot as an outcome. A Tukey's test of Honest Significant Difference (HSD) confirmed that participants in the bar chart condition considered red dots below the mean (and so within the visual area of the bar) significantly more likely than those above the bar. This effect was not significant for any of the remaining, symmetric encodings. Figure 4.5 summarizes these results.

Our results **generally support H3**: A Tukey's HSD confirmed that participants more consistently followed the expected strategy for question 1 (following the sample mean) with symmetric encodings (violin: 89.2% of trials, gradient: 88.5%, box: 87.4%) than with bar charts (83.2%). Graph type was also a significant main effect on confidence ($F(3,2982)=7.46$, $p < 0.0001$). A Tukey's HSD confirmed that participants were significantly more confident with the alternate encoding types which provided more detail about the probability distribution (gradient: $M = 5.12$, violin: $M = 5.06$) than with the bar charts and box plots ($M = 4.86$ for both encodings).

Discussion

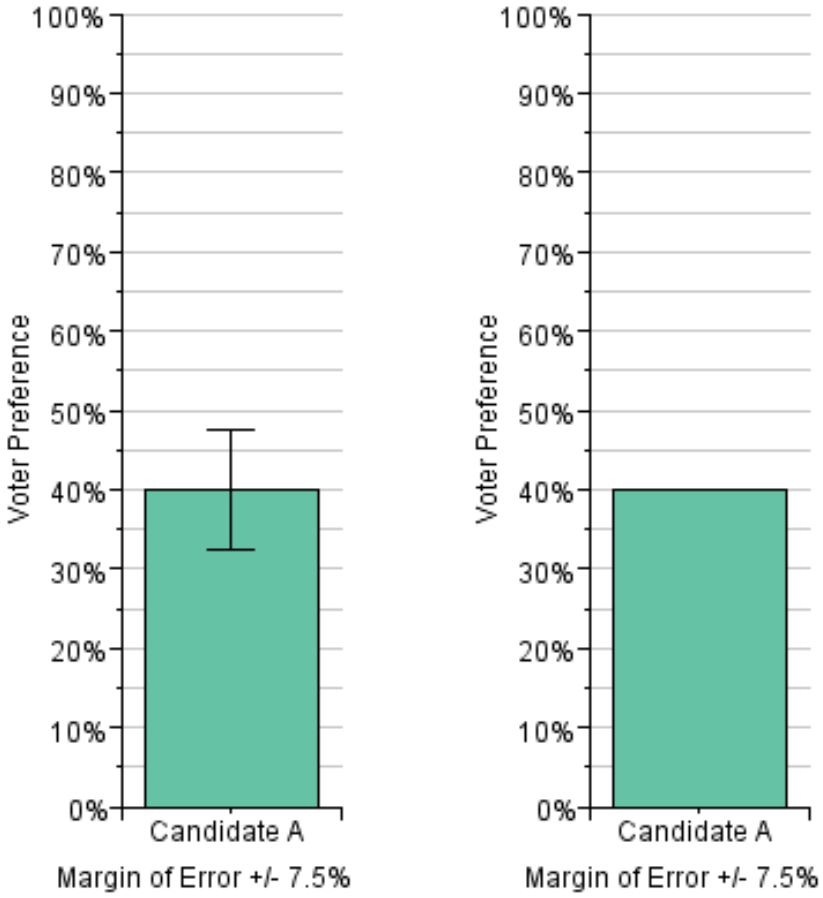
This experiment shows that a lay audience, even exposed to encodings that are unfamiliar, and with no expectation of particular training, can perform judgments that are correlated with inferential statistics: points that are far away from the mean

are seen as more unlikely, but smaller margins of error also reduce the perceived likelihood of distant points. However, this study shows that within the bar bias (where points contained by the visual boundaries of the bar are seen a likelier members of a sample than those outside it) is present even for inferential tasks, and can be severe enough to not just impact the perceived likelihood of different outcomes, but even the *direction* of inference. Our proposed encodings, by virtue of being symmetric about the mean, mitigate this bias, for a pattern of judgment that is better aligned with statistical expectations. The alternate encodings also offer more information about the probability distribution than a bar with errors, allowing viewers to reason more confidently at tasks beyond "this value is within the confidence interval" or "this value is beyond the confidence interval."

Experiment 2: Textual One-sample judgments

If within-the-bar bias is a visual bias (a red dot is visually contained within a bar), then it is possible that simply encouraging comparisons to be done with only partial assistance of the visualization might mitigate the bias. That is, by moving both the potential outcome and the margins of error to text, judgments might be better aligned with statistical accuracy. This scenario also represents how polling data is frequently depicted in practice, with information about poll size and margin of error written in a legend, but the chart itself displaying the sample means. We wished to evaluate this potential solution, as we speculated that it would introduce a great deal of inaccuracy to judgments and comparisons involving sample means (since it seemed likely that viewers would have to mentally project the text values into the space of the graph).

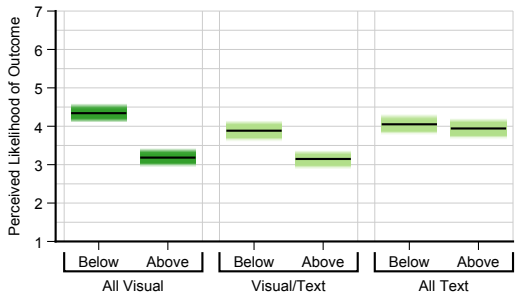
This experiment had the same factor levels and task questions as the previous experiment (and so each participant saw 36 stimuli), with three differences. The first is that instead of plotting a red dot on the graph itself, the red potential outcome was displayed in colored text under the graph. The second is that we presented only two graph types as a between-subjects factor: a bar chart with error bars, and a bar chart without error bars (in both cases the margin of error was displayed in text below the graph). That is, the conditions reflected moving some portion of the information to text from the graph, either the proposed outcome or both the



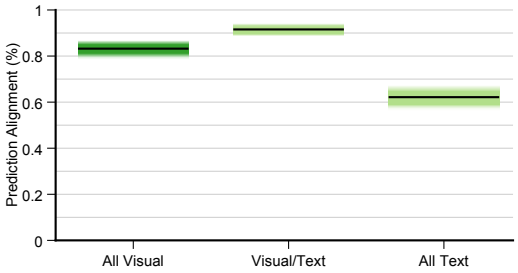
(a) Potential outcome in text, margins of error are visual.

(b) Both potential outcome and margin of error in text.

Figure 4.6: The stimuli for the textual one-sample judgments experiment (§4.3). Unlike in the first experiment, where participants were presented with a red dot representing a potential outcome, here the outcome was presented in text (e.g. "how likely is candidate A to receive 45% of the vote?"). In the second condition the margin of error was also presented textually rather than with explicit error bars.



(a) Within-the-bar bias when information is moved from the graph to text. If only the proposed outcome is moved from graph to text, values within the bar are seen as likelier than values outside the bar. Only when both outcome and margin of error are removed is the bias mitigated.



(b) Changes in adherence to expectation maximizing strategy when information is moved from the graph to text. Removing both margin of error and proposed outcome to text results in a significant drop in participant accuracy.

Figure 4.7: Gradient plots of our results of our textual one sample judgments experiment §4.3). When asked to consider potential outcomes, the expected behavior is that viewers will "trust" the sample mean – if a potential outcome is higher than the sample mean, then the "real" outcome will likely by *lower* than the potential outcome. Participants largely adhered to this strategy throughout experiments. While moving information from the graph to the text does mitigate within the bar bias, it significantly affects alignment with expected strategy. Since viewers must mentally project the potential outcomes and margins of error to the graph space, the relationship between the potential outcome and the sample mean becomes more difficult to analyze.

proposed outcome and the margin of error. This experiment used only one problem frame (the election phrasing).

Our hypotheses were:

H1 Participant responses will be similarly connected with statistical expectation as in the previous experiment — responses to question 1 will align with the direction of the proposed outcome to the sample mean, question 2 will correlate

with the cdf, and question 3 with the pdf.

H2 Removing the proposed outcome from the plot and placing it in text will mitigate within the bar bias, since the visual metaphor of containment is broken.

Results

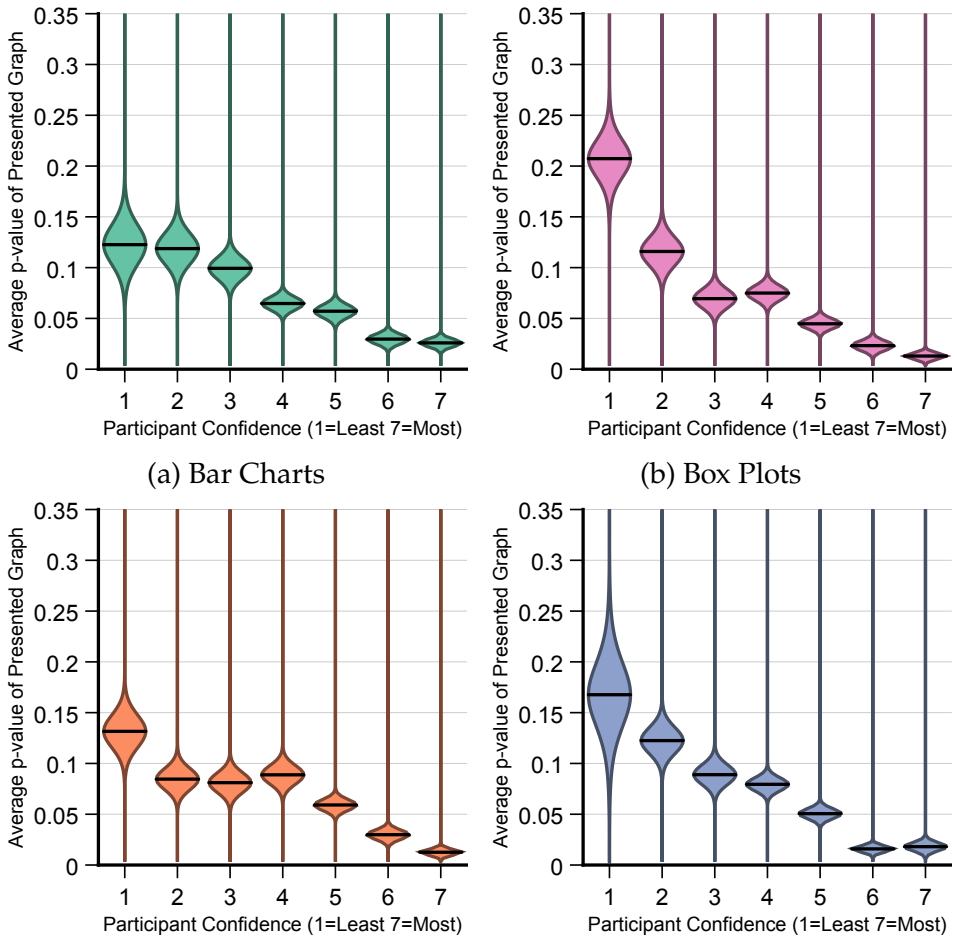
We recruited 48 participants, 24 for each graph type. We conducted similar ANCOVAs as in the previous experiment, testing how different encodings, potential outcome placement, and margin of error affected both cdf and pdf tasks.

Our results **only partially support H1**. Our expected strategy for question 1 was that participants would follow the sample mean. A Student's t-test confirmed that the participants followed the expected strategy significantly more with bar charts with visual error bars (91.6% of trials) than with bar charts with only textual margins of error (62.2%). Figure 4.7 summarizes this result. Despite this poor performance, participants were significantly *more* confident in their judgments with the graphs with no visual error bars than in the standard graphs ($F(1,1717)=64.8$, $p < 0.0001$, $M = 5.4$ with no visual error bars, $M = 4.9$ with visual error bars).

Our results **only partially support H2**. There was a significant interaction between the graph type and whether or not the proposed outcome was above or below the mean ($F(1,1717)=15.3$, $p < 0.0001$). A post-hoc Tukey's HSD confirmed that only for graphs with explicit visual error bars was there a significant difference in confidence between values above or below the mean ($M = 3.9$ and $M = 3.1$ respectively) – that is, within the bar bias was mitigated by moving both margin of error and proposed outcome to text, but not otherwise.

Discussion

This experiment shows that the visual metaphor of the bar is sufficient to create within the bar bias even if the actual values to be considered are conveyed in text rather than plotted. Removing both margins of error and the proposed value from the graph and to text mitigates this bias, but does so at the expense of making the chart sufficiently confusing to interpret that participants are highly inaccurate (or at least unpredictable) even at simple tasks, and additionally they are *unjustifiably* more confident in their incorrect judgments.



(a) Bar Charts

(b) Box Plots

(c) Gradient Plots

(d) Violin Plots

Figure 4.8: Violin plots of the participant's perceived confidence in their judgment between sample means (i.e. "which of two candidates will win the election, given the polling data?"), plotted against the actual average p-value of the relevant 2-sample t-test. While, for across all presented graph types, participants' average confidence was negatively correlated with p-value ($R^2 = 0.66$, $\beta = -8.30$), unlike in statistical practice (where we would reject as not statistically significant differences with p-values of 0.05 or higher), participants in general become gradually more confident on average with decreases in p-value.

Experiment 3: Two-sample judgments

In many real world visualizations of mean and error, the primary task is comparison of multiple groups with uncertain values. In order to recommend our alternate encodings for general use, it was important to both confirm that general audiences could generally perform comparison tasks with patterns of uncertainty that were

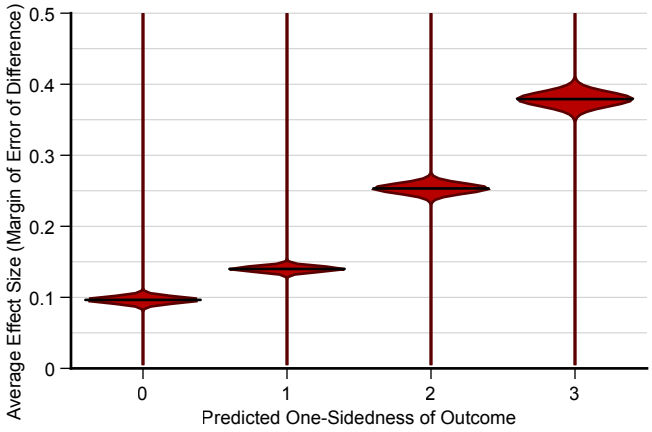


Figure 4.9: A violin plot of results from our two-sample judgments experiment (§4.3). Participants were asked to predict the severity of the outcome based on the sample. For instance, in the election problem frame, they were asked whether the election will be very close or one candidate will win in a landslide. This question is analogous to an estimation of effect size. We display the aggregate effect size (calculated here as the difference between means in terms of the margin of error) for all stimuli that participants associated with a particular level of one-sidedness. Participants' average estimation of one-sidedness were positively correlated with effect size ($R^2 = 0.567$, $\beta = 2.10$).

based on statistical expectation.

In this experiment participants were shown 400x400 graphs depicting sample means from two populations (A and B), and asked to make judgments comparing the likely performance of the two. Sample means were normalized such that A+B = 100 units. There were six different sample means for A, {75,60,55,45,40,25}. As with the first experiment, there were six different different margins of error, {2.5,5.0,7.5,10.0,12.5,15.0} (of which the participant saw a total of six per level, for 36 total stimuli), three different between-subjects graph types (bar with error bars, violin plot, or gradient plot), and three between-subjects problem frames (polling, weather, and financial frames). The participants were presented with three main task questions, with wording slightly altered to fit the problem frame (here from the polling frame):

1. If forced to guess, which candidate do you predict will win the actual election?
2. How confident are you about your prediction for question 1, from 1=Least Confident, 7=Most Confident?

3. Which outcome do you think is the most likely in the actual election, from 1=Outcome will be most in favor of A, 7=Outcome will be most in favor of B? (This was measured internally as a value from -3,3, with the "predicted effect size" being the absolute value of the response to this question.)

The expected strategy based on statistical expectation for question 1 is to choose the group with the highest sample mean. Question 2 is then analogous to a two-sample t-test (or, if it is known or assumed that the sample means will always be 100, a one-sample t-test with the null hypothesis that μ=50). Question 3 is then a question about effect size. Since the prediction task was isomorphic to a t-test, we calculated p-values internally for each sample mean comparison. The median p-value was 0.05 by design, however the p-values were not equally distributed among different margins (i.e. where the margins of error were 2.5 or 5.0, there were no stimuli which would fail a t-test at the $\alpha = 0.05$ level).

Our hypotheses were:

H1 In general, reported confidences and effect sizes will generally follow statistical expectation. That is, participants will "follow the sample" with question 1 — if one candidate is leading in the polls then that candidate will likely lead in the actual election. The participant answers to question 2 should align with p value, and the answers to question 3 ought to align with effect size.

H2 Encodings that encode margin of error in a binary way (bar charts and box plots) will have different patterns of performance than continuous encodings (violin and gradient plots), predicting bigger effects with more (perhaps even unjustified) confidence.

Results

We recruited 96 participants, 8 for each combination of problem frame and graph type. We conducted two sets of one-way ANCOVAs, testing for different encodings, framings, and data values on confidence in predicted "winners," and predicted effect size. Inter- and intra-participant variance in performance was included as a covariate.

Our results **supported H1**: We expected answers to question 1 to generally match statistical expectation, which is that the candidate leading in the sample will also lead in the population. This strategy was followed in 95.4% of trials. A Tukey's HSD showed no significant difference in strategy adherence among different encodings.

We expected the answers to question 2 to correspond to the p-value of the relevant two sample t-test. Large p-values ought to be associated with low confidence in the predictions of winners in the population based on the sample. Indeed, p-value was a main effect on confidence ($F(1,3424)= 49.4$, $p < 0.0001$). Figure 4.8 shows the connection between reported participant confidence in predictions and actual p-value in detail.

We expected the answers to question 3 to correspond to the effect size. We calculated effect size in terms of number of margins of error between the two sample values (a scalar multiple of Cohen's d). Effect size was a significant main effect on predicted magnitude of outcome ($F(1,3424) = 1210$, $p < 0.0001$). Figure 4.9 shows this result in detail.

Our results **partially supported H2**. For the predicted effect size, graph type was a significant main effect ($F(3,3424)=23.1$, $p < 0.0001$). A post-hoc Tukey's HSD confirmed that participants using bar charts predicted outcomes that were significantly larger than with other encodings (bar: $M = 1.65$, box and gradient: $M = 1.54$, violin: $M = 1.43$). This was also the case for confidence in predictions ($F(3,3424)= 3.38$, $p = 0.018$): participants were significantly more confident in predictions made by bar charts ($M = 5.21$) than for other encodings, but confidence in the other three charts was not statistically significantly different (gradient: $M = 5.07$, box and violin: $M = 5.02$). This gap was even more significant for stimuli which fail to pass a t-test at the 0.05 level of significance ($M = 4.42$ for bar charts vs. $M = 4.15$ for other encodings). That is, the elevated participant confidence was in a sense *unjustified*, occurring whether differences were statistically significant or not.

Discussion

Our results show again that the right choice of visualization can allow even a general audience to make decisions that are aligned with statistical expectation, but that these decisions are sensitive to how information is presented. We also show that the alternate encodings, by conveying more detailed information about unlikely

outcomes outside of the margin of error, encourage more appropriate doubt about inferences from samples to populations.

4.4 Summary

Our experiments show that even the general audience is capable of making nuanced statistical inferences from graphical data, taking into account both margin of error and effect size. However, the most common method of visualizing mean and error, bar charts with error bars, have several issues that negatively affect viewer judgments.

Bar charts suffer from:

- Within-the-bar bias: the glyph of a bar provides a false metaphor of containment, where values within the bar are seen as likelier than values outside the bar.
- Binary interpretation: values are within the margins of error, or they are not. This makes it difficult for viewers to confidently make detailed inferences about outcomes, and also makes viewers overestimate effect sizes in comparisons.

We can mitigate these problems by choosing encodings that are *visually symmetric* and *visually continuous*. Gradient plots and violin plots are example solutions. Our experiments confirm that these proposed encodings mitigate the biases above, and that modification of bar charts (for instance by moving margins of error to text rather than graphing them explicitly) address these biases only at the expense of introducing inaccuracy and complexity to inferential tasks.

Our experiments show that the general audience can robustly reason about mean and error. However, the issues we described above do occur in practice, and affect how the general audience reasons about uncertain information. The experiments also suggest that these issues can be mitigated with alternate encodings. Moreover, the cost of using alternate encodings appears to be low: even though the ones tested are unfamiliar, they still offer performance advantages to a general audience. The performance improvements of the alternate encodings are measurable in our experiments, but the practical effect of these differences is difficult to determine. Other

experimental methodology might better assess the impact on decision making, for example an experiment where stakes are higher might more clearly show differences between encodings. While our experiments show that encodings that follow our design guidelines provide advantages over bar charts with error bars, we have not fully explored the space of designs of mean and error encodings. We believe other designs that fit our guidelines should also have these advantages. Our experiments suggest that some encoding other than bar charts with error bars should be used, but are less specific in recommending the best replacement.

This is not to say that bar charts do not have utility. There are tasks where asymmetric encodings outperform symmetric encodings; for instance, comparing ratios can be done quickly and more accurately with bar charts as compared to dot plots or other encodings where area under the bar is more difficult to estimate [Cleveland and McGill 1984]. There are also cultural costs involved in adopting non-standard encodings — viewers might prefer to see familiar but known suboptimal encodings.

Limitations and Future Work

One area not well-covered by our experimental tasks was decision-making: does the presentation of different sorts of statistical graphs result in different actions (beyond mere predictions)? Assessing this facet of inferential behavior would require a more involved series of experiments, with real-world stakes. Likewise, our experiments did not collect a great deal of qualitative data such as viewer preferences for the different chart types: the aesthetics of information visualizations can be an important consideration for how data are perceived and used [van der Geest and van Dongelen 2009], especially for issues of trust and uncertainty. In the future we hope to modify or extend our set of proposed encodings to cover a wider range of inferential scenarios, including the perception of outlier values, regression, and multi-way comparison, and to deal with additional known biases in human reasoning.

Our data and experimental design also did not reveal many significant differences between our two proposed encodings. Our data do not support the use of one over the other for decisions tasks, however paper authors, reviewers, and colleagues have

stated differing preferences between the two on aesthetic and theoretical grounds. We present both in this chapter to promote critique, but further work remains to assess both encodings in a principled way.

We also did not investigate how performance might differ with different design decisions. For instance, we colored the gradient chart to make the region within the margin of error fully opaque, but we could have encoded the pdf of the t-distribution directly. We chose a single set of color ramps for our encodings, but it is possible that other choices might bias viewer judgments (for instance, viewers might overestimate the likelihood of outcomes in red violin plots [Cleveland and McGill 1983]).

Conclusion

In this chapter we illustrate that the most common encoding for displaying sample mean and error — bar charts with error bars — has a number of design flaws which lead to inferences which are not very well correlated with statistical expectation. We show that simple redesigns of these encodings which take into account the semiotics of the visual display of uncertain data can improve viewer performance for a wide range of inferential tasks, even if the viewer has no prior background in statistics. We show that the general audience can achieve good performance on measurable decision tasks with encodings which are less well-known than the standard bar chart. These results provide evidence that the choice of visual statistic important for measures beyond task accuracy (the main measure in Ch. 3), but also higher level measures such as confidence. It also provides evidence that, although there existence perceptual and cognitive biases in how visual statistics are interpreted, designers can partially or totally "de-bias" through careful design.

In the following chapters we present systems that are mindful of differences in argumentation and persuasion, and that are mindful of the affordances of visual aggregation (as presented in Referencesch:visagg and explored in these last three chapters). In both, cases the analysts may or may not have strong statistical backgrounds, but through careful design the visualization tools can support data- and statistically-driven analysis and decision-making.

5 TEXTVIEWER AND CORPUSSEPARATOR: A TOOL SUITE FOR THE DIGITAL HUMANITIES

5.1 Introduction

We have previously discussed a continuum of trust we as designers must consider: e.g. can we trust the viewer to correctly estimate the mean of a particular cluster of points, or gauge the certainty of a particular inference? When dealing with models we have additional trust concerns: can we communicate highly complex and rather abstract statistical gestalts to our audience? If so, can we do so in such a way that the viewer both trusts that the model correctly captures an important property of the data? If so, can we convey this property in an explicable way?

In this chapter we describe the creation of a suite of visualization tools designed to meet the needs of literary scholars as they incorporate computational tools that operate over large text corpora, specifically those using tagging schemata to perform "algorithmic criticism." This algorithmic criticism requires the communication of a statistical model to a potentially non-statistical audience, and furthermore requires situating the statistical information gained through analysis in a way that is rhetorically relevant for literary scholars. Through a careful design process, we have developed a solution (the CorpusSeparator+TextViewer tool suite) tailored to the needs of these scholars. This solution allows for the discovery and examination of patterns of tagged text at the corpus-wide level, but also supports the process of passage analysis for grounding digitally inspired arguments in specific human artifacts.

Advances in digital storage, curation, and collaboration are beginning to produce seas of data humanities scholars are currently ill-equipped to handle [Ellis 2005]. Typically when faced with a situation such as this, visualization techniques are applied to aid scholars in making sense of large-scale data. Unfortunately, humanities scholars are not well served by most existing information visualization techniques. Humanities argumentation is traditionally done by reference to specific exemplars. For example, while a chart showing statistical properties in a large dataset might be sufficient proof for an argument in the sciences, literary arguments prefer to

reference and analyze specific text passages.

Our domain collaborators needed to analyze large corpora of texts that had been tagged based on rhetorical content of specifics words and phrases. To assist them in this effort we created two tools: the CorpusSeparator for visualizing corpus-wide rhetorical patterns and trend-specific selections, and the TextViewer for visualizing and selecting salient passages of specific texts. In concert, these tools allow literary scholars to very quickly formulate a hypothesis about a corpus and then provide evidence for this hypothesis by drilling down to the specific passages of text.

These tools defend my thesis in that the data set (which is based on applying a potentially complex and artificial statistical process to labeled data) was presented to scholars with no explicit statistical background. However, by the thoughtful application of visual statistics, our collaborators were not only able to successful use the tool to generate new insights in their domain, but communicate these insights to others in their field. This work was originally published as Correll et al. [2011c].

Problem Overview

Rhetorical analysis is a form of literary scholarship that seeks to understand the ways that language is used, rather than the message that is being conveyed. Often, scholars aim for analysis that is *distanced* from the meaning, so that their findings are generalizable beyond the particular content of a text. At the same time, scholars prefer arguments with specific examples of passages that support their analysis. Such arguments are made by *close* reading: careful analysis of specific passages of text, with an eye toward exemplifying some more global pattern.

Performing rhetorical analysis at a large scale, e.g. on significant corpora of texts, offers the possibility of finding trends and patterns in language usage, for example, to resolve questions of authorship, to see "signatures" of structure in different genres, or to observe the historical development of language. Unfortunately, the traditional approach to analysis (close reading of specific passages) does not scale well due to the sheer amount of text that must be closely read, especially if the scholar attempts to maintain distance. As scanning, curation, and sharing efforts provide scholars access to larger sets of text to analyze, new analysis methods must also be developed that scale appropriately.

Algorithmic criticism is an emerging method for applying rhetorical analysis at a large scale [Ramsay 2003]. The approach exploits the disconnect between deep semantic interpretation of text and the surface features of natural language: by focusing on simple, low-level properties that can be discerned algorithmically, these "prosthetic readings" of texts are necessarily distanced. Statistical analysis of low-level properties are viewed to identify patterns and trends. This mode of analysis is sometimes referred to as "iterative criticism" to indicate the need to reincorporate feedback into the process, which emphasizes the need for a fluent workflow that enables iteration.

One form of algorithmic criticism focuses on the roles of words. Thus, each word is replaced by a "tag" corresponding to a rhetorical or semantic category. Automated or semi-automated tagging systems use algorithmically simple means, such as regular expressions, to determine the tags for each word. A text (or a portion of one) can be represented as a vector in n-dimensional space, where there are n different tag types. Each element of the vector is the count of the corresponding tag. Statistical tools such as Principal Component Analysis can then be used to see where different categories of a corpus fall within the n-dimensional space.

While standard text taggers are available and even preferred to ensure consistency in the tag schema, there are few tools to help analyze the tagged data. No existing tools (see below) had been created to support the unique scholarly process that must ultimately connect the distanced reading with the close passage analysis. Scholars have been using standard statistical packages to view corpus-wide data, but this is not a panacea. Even if an interesting pattern does emerge (for instance, one author's rhetorical style results in a very different distribution of tag counts as compared to the corpus at large), providing passages that present literary evidence of this pattern requires manually tagging, and sampling many texts, then looking through interesting passages by hand until a rhetorical pattern is pinpointed. This approach does not scale to large corpora, nor is it particularly efficient.

Solution Overview

The primary insights from our work are an understanding of how the methods of humanities scholars lead to unique needs for their tools. While our study has

been specific to algorithmic criticism, we believe these needs, particularly the ability to connect large scale statistics to specific exemplars, exist across many forms of humanities scholarship and beyond. The primary novelty in our system is its overall design that assembles components to support scholars' workflow. However, in realizing this approach, we have developed a number of novel components:

- On-the-fly thresholding and filtering for visualization and re-computation of statistical analysis on text corpora.
- At a glance information about distribution of tags across a corpus, with details on demand.
- Visual links between loadings on principal components and individual passages of tagged text.
- Focus+context techniques for selection of passages.

Contributions: Our work contributes an example of a case study, showing how a process of following task analysis with iterative development can lead to interesting results. We provide specific insights on the work of literary scholarship, and how scholars' needs create demands on tools. The design we propose combines standard visualization techniques with several novel ones (see above) to address these needs Finally, the actual systems themselves, CorpusSeparator and TextViewer, are a contribution as they have shown immediate utility for our collaborators.

5.2 Related Work

Many existing visualization tools for use with large text corpora are intended mainly for topic clustering and curation. While these tools can generate compelling results for these tasks, they are often abstracted from the actual content of the text, relying on "bag of words" vector representations of texts to generate their results, as with the common terrain or starfield corpus visualization techniques [Wise et al. 1995]. Thus while important properties of the text are visualized, there is no clear mapping from these properties to specific text, even in tools such as Docuburst that are otherwise useful for visualizing how different texts differ [Collins et al. 2009a].

Some visualizations, such as implicit shapes [Rohrer et al. 1998], do not clearly connect visual features to either text or concept, preventing even surface level analysis of properties. This disconnection is inappropriate for our domain. In general, corpus-level overview tools that rely on important words to create groups are not appropriate for specifically literary tasks, even tools such as ThemeRiver [Havre et al. 2000] that otherwise do a good job of displaying temporal changes in a corpus. Visualizations combining tag clouds with other visualization paradigms provide some connection to the text, but this is still unsuited to passage analysis [Collins et al. 2009b][Cao et al. 2010].

Some tools combine both high level corpus overviews with text views, such as Jigsaw [Stasko et al. 2008]. The generality of these systems creates drawbacks for scholarly applications, including high learning curves and demands for user creation and labeling of salient features. They also do not help with identifying paradigmatic passages. One might use these tools to be able to see how many times a certain word or pattern appears in a text, but searching by hand through all of these "hits" without the ability to set thresholds of relevance does not scale well.

One method of computer-aided text analysis relies on multiple tools, some of which incorporate views of the text per se. Using a suite of different visualization tools it is possible to conduct intricate text analysis [Clement et al. 2009]. These tools still rely on specific words, although stemming is often used to remove the effects of changes in tense or number. Rhetorical patterns (and other literary methods of analysis) are concerned less with word choice but more with word meaning, making existing tool suites difficult to adapt to problems of rhetorical analysis.

Wanting to connect visualization of a large corpus with small sections of text is a problem encountered in the wider software visualization community [Ball and Eick 1996]. The analogy is not perfect, however, as softVis typically deals with issues such as version control and collaboration that do not have clear parallels in literary analysis, which usually focuses on the final product rather than the production process.

5.3 Design Setting

Our domain collaborators are literary scholars operating in the DigHum (Digital Humanities) working group at the University of Wisconsin-Madison. We observed their current tagged text analysis process in person, as well as as described in their published works. They had access to a number of corpora of English literary works. Their avenues of research included pinpointing the development of specific genres, differentiating rhetorical style between genres of Renaissance dramas, and using rhetorical "signatures" to back up grounded claims of authorship [Witmore 2010]. This methodology is already successful in identifying how "micro" linguistic level activities, such as word choice, noun forms, syntax, or article and pronoun use, are connected to higher level phenomena, like authorship, genre, and meaning.

Their efforts make use of the Docuscope tagging schema, which categorizes millions of regular expressions (usually on the scale of one or two words) into roughly 100 different tags based on different rhetorical categories (e.g "FirstPerson" or "DirectAddress" language) [Collins and Kaufer 2001]. Since Docuscope was meant to tag modern English speech, modernization techniques were used on the largely antiquated English corpora to increase the percentage of raw text being tagged.

The existing workflow relied on Principal Component Analysis to analyze clusters in a particular corpus. Texts in the corpus were divided into multiple subsections based on an empirically justified word length window. The Principal Components that were sufficient to isolate a sub-corpus of interest were saved, and the subsections that scored particularly well on these axes were examined by hand using traditional literary methods of passage analysis, combined with a subjective appraisal of weights on tags in each Principal Component. This workflow generated results that have been published in mainstream Humanities journals [Hope and Witmore 2010].

Of note in this workflow is that it differs from traditional (and well-studied) text clustering problems in that it is not about finding clusters in the text (since these clusters are already known, e.g. the genre of texts are known in advance), but rather showing how known a priori groupings are ultimately reflected in low-level properties in specific passages of text. The existing workflow was not optimized

for this purpose, and required switching between statistical packages like JMP[1], a proprietary Docuscope reader program, and raw analysis of large matrices of data. This approach was already difficult at the scale of the less than 100 plays of Shakespeare, and could not even in principle scale to larger corpora (such as Google Books' millions of texts).

Another problem not supported by manual methods was that it was often possible to find interesting passages (say by using the PCA methods in JMP), but generating "slices" of the texts for this purpose was somewhat arbitrary, generating high frequency, noisy results in PCA space. To minimize this issue our collaborators would use overlapping sliding windows of text a few hundred words long. Changing this window size had the effect of reducing the impact of outlier sections of text, but it was difficult to perform this sort of analysis on the fly without having to recreate the text slices.

Requirements Analysis

In order to improve (and hopefully supplant) the existing workflow, we compiled an initial list of requirements:

- The tools must scale to corpora with sizes from just a few items to potentially thousands of entries, where each item could be as small as a single passage or as large as an entire novel.
- In order for the results generated to be of use from an argumentation standpoint, there must be a direct link between a visualized item and specific text.

After constructing an initial prototype (see below) we realized that our collaborators had already "learned" the various idiosyncrasies of the Docuscope software, and were reluctant to modify existing knowledge about color mappings or file formats. In addition, other collaborators were reluctant to use new tools at all if they only relied on the Docuscope tagging scheme. Thus we modified our list to include two additional, somewhat contradictory requirements:

[1] A proprietary statistical analysis software suite, www.jmp.com

- Since the scholars were used to operating with the original Docuscope software, our software ought to mesh well with the esthetics and structure of Docuscope.
- Since not all of the scholars thought the Docuscope scheme was equally useful, our software should generalize to different tagging schemata.

In addition, further analysis of the workflow clarified the existing taks, which was more than analyzing tag counts per se. In particular, most of the existing insight-generating work seemed to follow this model:

1. Identify a useful weighting of each tag that distinguishes certain groups of texts from others.
2. Using this weighting, find important texts that typify a group, or are outliers from a group.
3. Within these texts, find important passages that score in important ranges using the current weighting scheme (e.g., passages that "explain" a text's location).

Existing tools were insufficient for this model, and there was no automated way to accomplish the task at different levels of scale. It was not enough to augment this workflow with better tools, since for the most part these tools did not exist. Thus we added a final item to our list of requirements:

- Our tools must either supplant or provide a superset of abilities found in the current workflow: i.e., they must provide the ability to find a salient weighting of tags, find salient texts based on this weighting, and lastly find salient passages within these texts.

5.4 Design Rationale

For the design of our tools, we followed an iterative model where we would conduct ethnographic observation of workflow, prototype initial tools, and then refine existing tools or create new prototypes based on feedback. This cycle was repeated several times as our understanding of the needs of our collaborators was deepened.

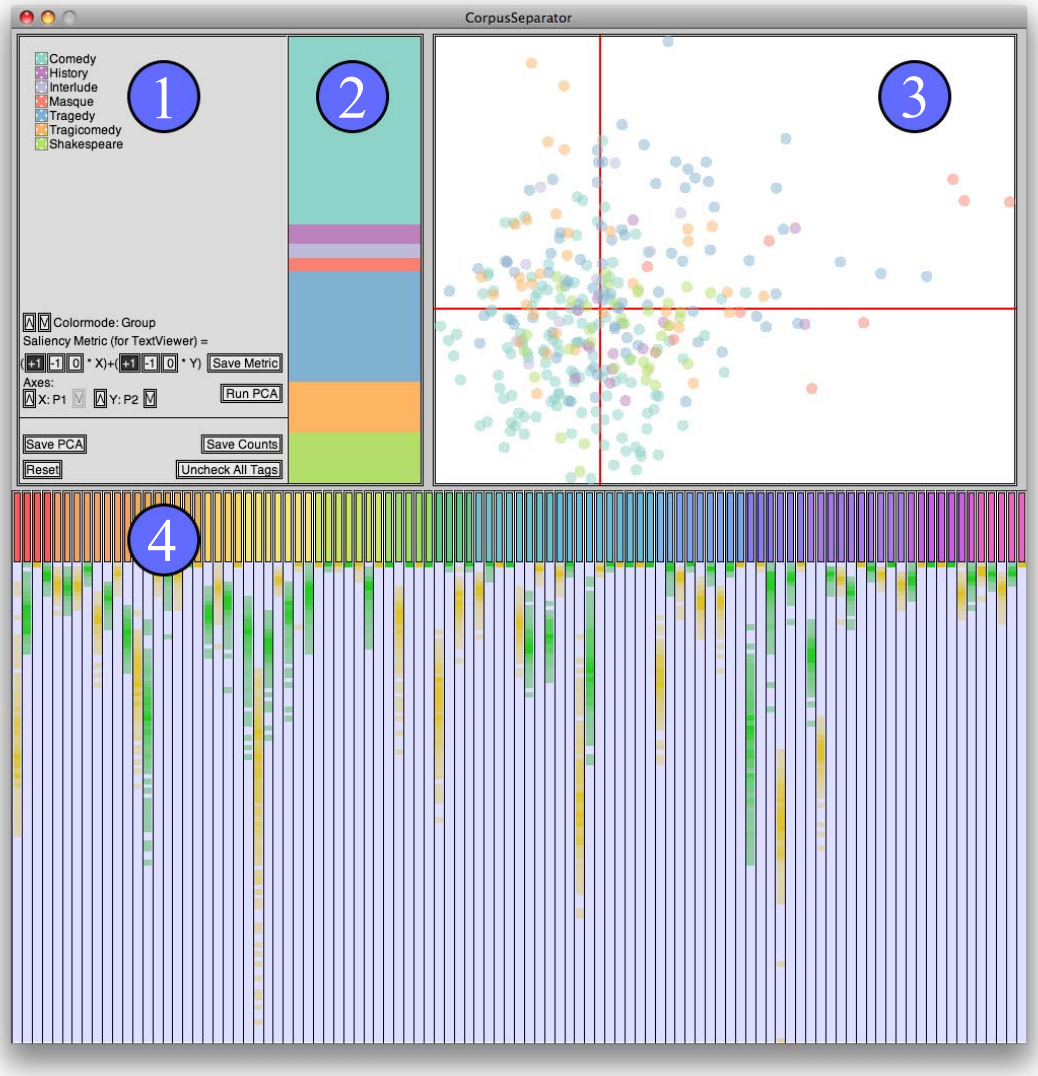


Figure 5.1: A view of the CorpusSeparator tool after importing a corpus of 300 16th and 17th century English plays.

1. *The upper left panel displays filtering information and principal component options, as well as options to save a weighted sum of tags as a combination of one or two principal components for later use by the TextViewer. One can scalar multiply each axis' principal component by 0, -1, or +1. Both multiplied axes are then added together to generate an ad-hoc saliency metric.*

2. *The composition of the corpus (in terms of blocks of different groups). Users can color based on a priori groups like genre, but also metadata such as composition year or author.*

3. *The upper right panel shows the projection of each text as an n-dimensional vector of tag counts into a two dimensional subspace based on Principal Component Analysis. Conventional axes, in red, orient the user.*

4. *The lower panel is an accordion view of all tags (detailed in Figure 5.3), with the distribution of normalized tag counts per text represented as bands of color.*

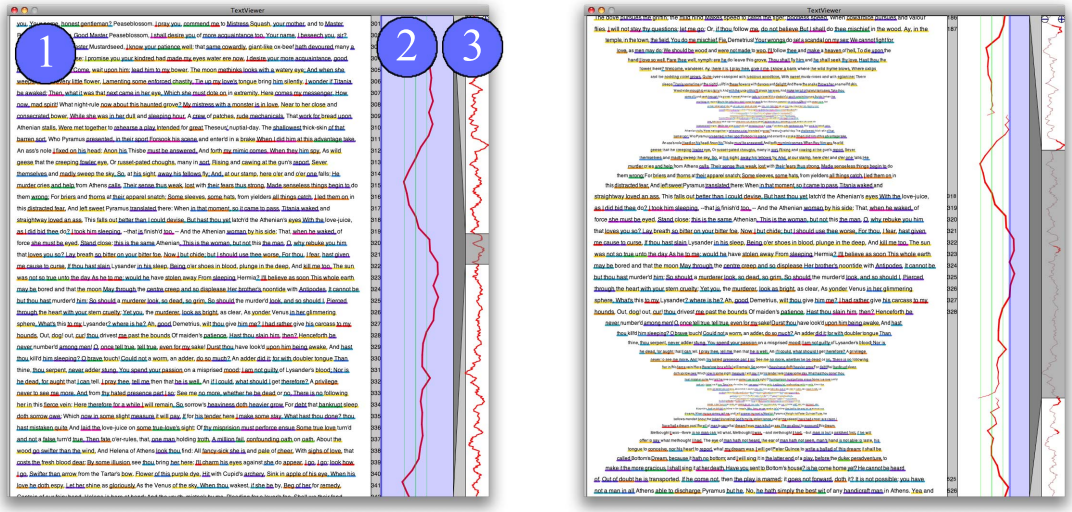


Figure 5.2: Two views of the TextViewer tool on Shakespeare's "A Midsummer Night's Dream." The left view shows an initial view of the text, whereas the right view shows the same location in the text after a threshold has been set. Lines with scores at or above the threshold are in focus. This creates bubbles of salient passages that are in focus, while the rest of the text recedes.

***1**. The left panel of each image is the raw text, with colored underlines for text that has been tagged.*

***2**. The larger vertical line graph (the local graph) shows the score of a window of text centered on each visible line based on a weight on tags generated in the CorpusSeparator or by hand.*

***3**. The smaller, rightmost graph is the global graph of the scores for the entire play. The gray rectangle of the global graph represents the text currently in view. Note that much more of the text is within the window once a threshold has been set.*

Initial Prototype

Our initial efforts focused on a specific facet of the existing workflow: our collaborators did not have a good way of visualizing differences in tag counts across the entire corpus other than the manual analysis of a large data matrix. The initial CorpusViewer tool was meant to be a simple prototype that allowed users to visualize outliers between groups.

A stacked bar chart with multiple levels of detail, aggregation, and mouseover annotation was prototyped and released to the domain collaborators. Since Docuscope has more tags than one could conceivably use for a mutually-distinguishable

color palette, a palette of low-saturation pastels with alternating bands of small hue difference was used. This proved to be difficult for collaborators used to the rainbow spectrum scheme of the Docuscope program, and so we gave users the option of using our banded saturation palette or using the original rainbow Docuscope palette. The ability to keep the old palette eased the transitions, despite the known deficiencies of the rainbow color scheme [Borland and Taylor 2007].

While the CorpusViewer was used for gross analysis of tag patterns, we had hoped to supplant what we saw as an inefficient workflow, as well as allow previously impossible capabilities. In particular our domain collaborators had no ready way of visualizing what different tag patterns meant at the textual level. We began prototyping tools that could interface with both a binary list of "interesting" tags generated from the CorpusViewer, as well as the axes in tag space generated by PCA or other embeddings (such as other MDS methods, or a distance from a hyperplane generated by a Support Vector Machine).

CorpusSeparator

The CorpusSeparator [Fig. 5.1] is meant to provide an overview of the entire corpus, separating out clusters or patterns of interest at the level of specific texts. Each item in a corpus is represented as a vector of tag counts. After normalization of each item to control for differences in item length and tag density, Principal Component Analysis is used to collapse this potentially high dimensional space into a linear subspace. Users observe the projection of the corpus into a two dimensional subspace of Principal Component space, find clusters of texts that are separated from other groups, and use the relevant Principal Component axes to generate an ad-hoc saliency metric that assigns an importance rating to each tag. Individual texts that are particularly good exemplars of their group, or are particularly troubling outliers, can then be examined using the TextViewer (see below) with the aide of the chosen saliency metric. Problems of occlusion, distance gauging, and general complexity limit the user to viewing only two dimensions of the space at a time, but users can select any arbitrary two dimensional subspace.

One capability that was currently lacking from the existing workflow was to be able to filter out outliers, both in terms of tags and texts. Certain tags were rare

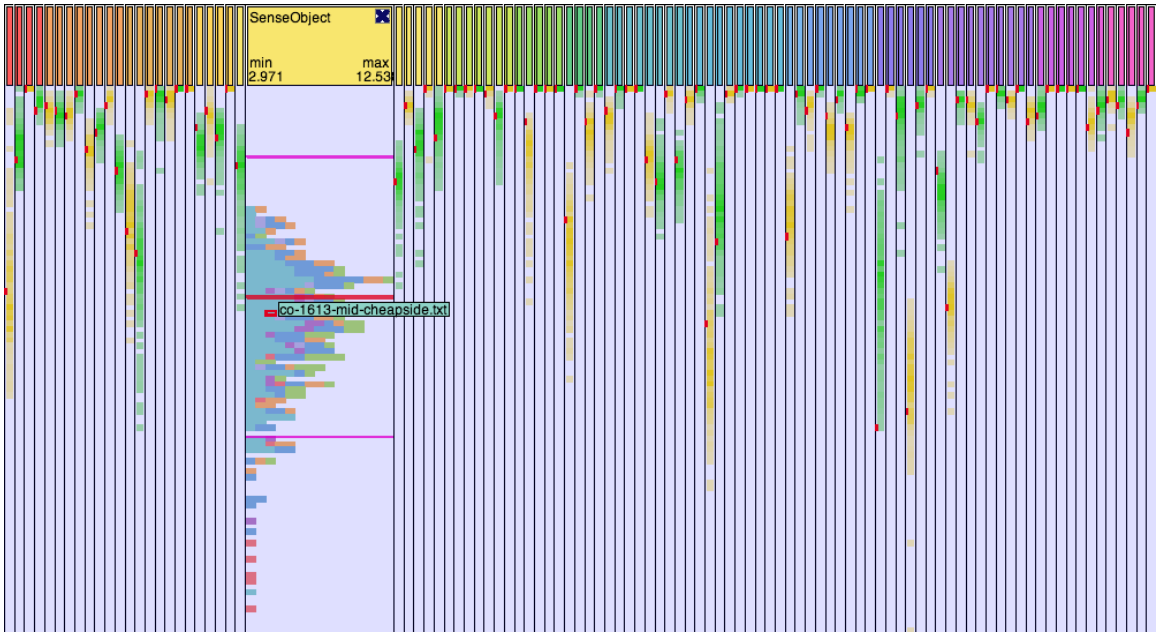


Figure 5.3: An example of the accordion view of per tag "cards" in CorpusSeparator. When a mouse is over a card, it expands into a cumulative distribution histogram of a particular tag. The count of a particular tag's appearance in each text is represented by a block. Lines for the mean and the standard deviation in both directions are represented by red lines. To exclude outliers users can drag thresholds up or down, removing texts outside of the threshold from consideration in any future Principal Component Analysis. Collapsed cards represent distribution as saturation values of blocks of color, allowing tags with outliers to be seen at a glance without having to expand every card.

enough that they did not noticeably contribute to understanding the corpus, and certain texts were poorly tagged for one reason or another (poor modernization for instance) and so not representative of a group in the corpus. Other texts or tags that would appear to be outliers were actually useful for analysis; filtering decisions had to be made a posteriori. To facilitate this filtering we included "cards" for each tag showing the distribution of normalized tag counts across the entire corpus. Originally users would scroll across each card, setting thresholds for inclusion or exclusion of texts, or binary choices for inclusion or exclusion of the entire tag in the PCA. This approach was useful and allowed a lot of fine control, but did not scale to Docuscope's over 100 tags (and thus over 100 cards). An accordion view was chosen, allowing quick navigation of tag distributions [Fig. 5.3]. While this

accordion view does not scale fully to the level of hundreds of tags, it works well for the number of tags in Docuscope. Since each tag type is meant to encode a different semantic category, the workflow used by our collaborators would not support a tag encoding scheme much larger than Docuscope's even if one did exist.

Using the CorpusViewer, our collaborators were able to develop richer intuitions about the dimensionality reduction process, notice outliers, and see the results of moving to different embedding subspaces on the fly. While other PCA visualization tools exist, many relying on the same scatterplot presentation, our tool unites the ability to quickly filter outliers, gain statistical information about the distributions of values in the original dimensions of the space, and switch to new PCA embeddings.

TextViewer

The TextViewer [Fig. 5.2] is meant to tie large scale tag patterns to specific passages of text. It takes in two arguments: a weighted list of tags, and a specific tagged text. This weighted list can be generated using PCA (e.g. using the CorpusSeparator), or can simply be a binary list of inclusion or exclusion of tags. Text is then rendered with colored underlines, with the saturation value of the underline corresponding to the absolute value of the tag weight. These underlines (especially with tensely tagged documents) are distracting to an untrained reader, but are representative of the output of the original Docuscope program and thus our collaborators were acclimatized to reading documents in this form. The goal of a TextViewer session is to visualize how a tag pattern is realized in a particular text, and more specifically pick out and perform close analysis on passages that are exemplars of these patterns. A text might be quite large, but the number of salient passages quite small, and the requirements of saliency fluid. A user would need an overview of the entire text, but also the ability to quickly move between important passages. A focus+context view was then the natural choice [Sarkar et al. 1993]. The raw count (or in the PCA case, the weighted sum of raw counts) of tags in a particular window of text is a one to one mapping of a text to a signal. By setting thresholds of importance, the user creates foci on sections of text that lie within the desired threshold. Virtual lenses placed on these foci causes these passages to come forward, while less important passages recede.

The process of direct analysis in the existing workflow differed in several key areas from the "natural" reading environment, and so we had to adapt TextViewer to more closely resemble this artificial method of reading. Our domain collaborators would strip out information such as line breaks and stage directions from texts, since those vary from edition to edition (and as such do not have explicit, invariant rhetorical content). The lines in the TextViewer are thus not natural units of a text (although it is possible to set canonical line breaks). This conflicted with our desire for there to be a one to one relationship between a line of text and a point on the local graph. We also did not want to have to normalize to account for lines with shorter words having more tags (and thus higher absolute values of scores). We decided therefore to treat the score associated with a line as being the score of a window of words centered at that line (e.g. if the window size is 256 words and the line is 56 words, the previous 100 and next 100 words are scored and added to the total). These windows naturally overlap with the windows of other nearby lines. Thus larger window sizes perform a convolution of the high-frequency saliency signal into a smoother, more manageable one. For some texts larger or smaller window sizes are the natural unit of division. Since the "best" window size for a particular text is not known a priori, we gave users the option to adjust this window size on the fly.

Focus+context interfaces in other domains benefit in that the minimum size of an item not in focus may be quite small: one might still glean information from e.g. a tree if sections of it are compressed to one or two pixels [Munzner et al. 2003]. Unfortunately, in the text domain words must be of a minimum size in order to be legible. Originally, text not within a focus was greeked to provide context information about a passage's location with regards to the rest of the text. Since this greeked text was not useful for the task at hand, we removed this functionality and instead had text recede until it was no longer legible, and then removed the out of focus text entirely, creating "bubbles" of text with discontinuities between them. The user thus still has local context within the passage without having to consider large sections of greeked, meaningless text. Global context is also provided by a small "global graph" overview to the right of the text, with a colored region indicating how much of the document is currently in focus.

Another issue with a focus+context interface in the realm of text is that sentences

may begin or conclude in an area currently out of focus, preventing the full context of a passage from being known. To prevent this issue we used a bridge lens as our method of focus: focal lines are in focus, as are lines ϵ lines above and below. After that point, there is a linear decrease in size for δ additional lines until finally the text is completely out of focus. These parameters are controlled by the user to account for line length and total length of texts. For e.g. a poem important phrases are likely to begin and end on one line and so smaller ϵ and δ values can be selected.

TextViewer's focus+context interface is a unique way of scanning and visualizing saliency in potentially large text documents, and the ability to load in arbitrary saliency metrics works in concert with this interface to allow fast textual analysis that combines the benefits of distanced readings with close passage analysis.

Implementation Details

From our requirements analysis we decided to develop our tools as portable Java clients that manipulated the human-readable labeled comma-delimited files already in use by our collaborators. The Docuscope tagging software generated such .csv at the corpus level, and the JMP statistical software output .csv files as well. The Processing library was chosen for its ease of implementation and inherent design as a cross-platform graphical prototyping tool.

5.5 Case Study

Our collaborators adopted these tools immediately, sharing them with fellow researchers in and with students (graduate and undergraduate) who are now being trained to use them. Strikingly, they found uses for the tools we had not anticipated, allowing us to see a number of use cases that differ from the "typical" ones described above. These adaptations are worth discussing in detail.

The tool was immediately put to work on an expanded corpus of works. Our collaborators had already been working with a collection of 36 Shakespeare plays whose texts have been edited and modernized through centuries of scholarship. (Lacking authoritative autograph manuscripts for Shakespeare's, literary critics must reconsruct them from multiple or variant sources.) Soon, however, they began

to work with larger collections: 318 plays written between 1509-1669, including the original 36 by Shakespeare. Analysis of these data did not scale well with existing tools, offering a good use case for the tools and concepts we developed.

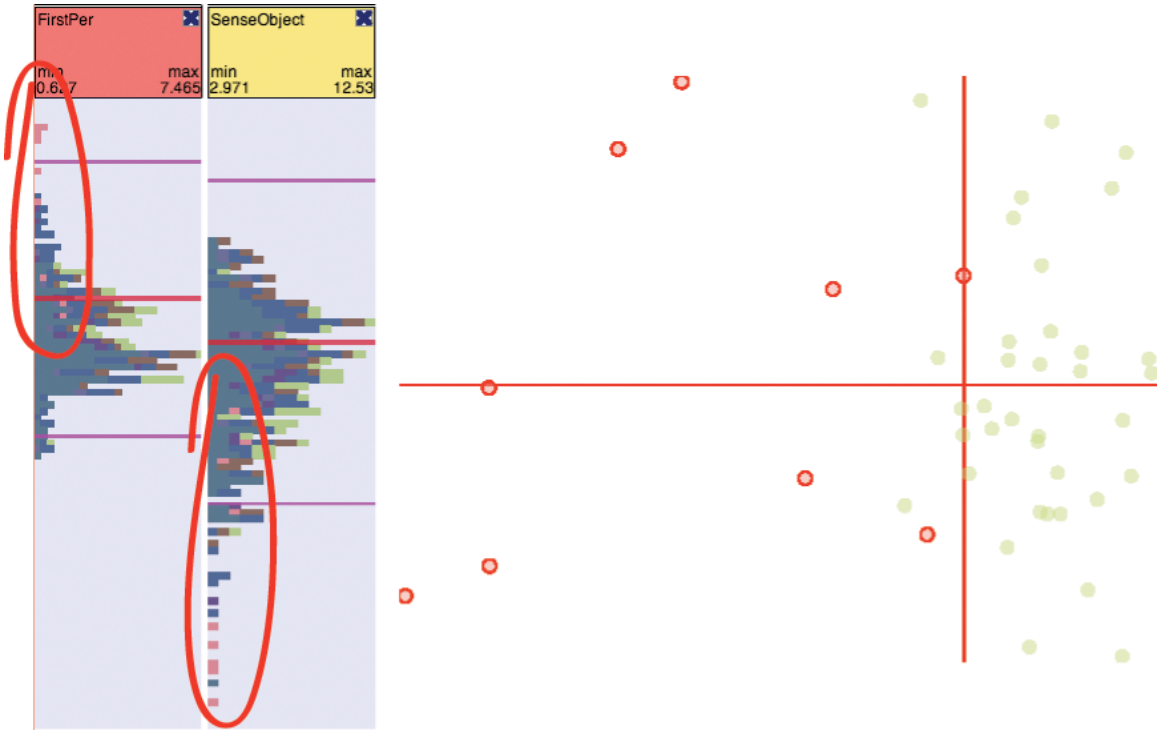


Figure 5.4: Court Masques (highlighted in red) vs. Shakespeare, in views from CorpusSeparator. The masques are "pushed" leftward as a result of the current choice of principal components. We can see that certain tags (such as the FirstPer tag) do a good job of distinguishing the two groups, with all of one type clustered at one end or another of the distribution. TextViewer allows users to zoom in and see "masque-like" vs. "Shakespeare-like" passages.

Our collaborators began by looking for distinct types of writing – genres – based on existing scholarly classifications of these texts (corpus-specific metadata). Using the CorpusSeparator's PCA capabilities, they isolated a subgenre known as the court masque, a forerunner of modern opera which includes spoken poetry, music and dance. CorpusSeparator allowed them to identify the most exemplary text from this genre, a masque by Ben Jonson. Accordion view then showed which tags were associated with these plays. In the case of the masque, it turned out to be words or strings of words referring to things in the world that are present in

large numbers, whereas these plays are conspicuously lacking in questions. At first it was not obvious why these features would characterize the masque as a genre. Looking at an exemplary passage as indicated in the TextViewer, however, they were able to make sense of the pattern and connect it to observations that have already been made about the masque, but not in linguistic terms: because masques are a courtly genre that requires long, set speeches and little dynamic interaction among characters (they are not particularly dramatic), there are few questions and the dialogue that does exist tends to be descriptive – evoking real or imaginary scenes by describing them slowly and in detail (a practice known as *ekphrasis* in literary criticism). They were also able to take this set of weightings from PCA and apply it to Shakespeare, in effect finding the passages in Shakespeare's works that most resemble a court masque. Since there has been much speculation in the secondary literature on this issue, their ability to make this identification is significant.

Simple PCA indicated that the plays written by Shakespeare were noticeably different from the rest penned by dozens of authors. But this apparent distinctiveness turned out to be a problem: because the plays by Shakespeare had been hand-corrected and modernized in the nineteenth century, whereas the plays by the rest of the authors had been semi-algorithmically "modernized" by one of their collaborators, this difference may have been editorial. The only way to know this would be to find the passages in Shakespeare that were the most dramatically different – or better, representative of Shakespeare's difference – from other authored texts. Using CorpusSeparator, Witmore and Hope identified two principle components that effectively distinguished Shakespeare's plays from other works. This they were able to do by varying displayed components in the CorpusSeparator and looking for significant clusterings. They then saved these components and used them in TextViewer to find exemplary, Shakespearean passages.

In TextViewer, they discovered that curation and modernization were at least partially responsible for the distinctness in Shakespeare's texts: the exemplary passages they viewed in TextViewer revealed certain tokens were being systematically modernized in the Shakespeare corpus but were more variable in the semi-algorithmically modernized corpus. (Renaissance orthography is significantly more variable than contemporary orthography, in part because of the introduction of dictionaries in the eighteenth century.) Knowing this allowed them to correct the

modernization algorithm and obtain higher quality texts for analysis. A significant area in their workflow was now better understood and timely, practical interventions became possible.

Having eliminated result-skewing tokens that were an artifact of editorial procedure, they could repeat their survey of corpus-wide variation at different levels of abstraction - finding, for example, patterns that are associated with time of composition (which show up as corpus size increases), but also patterns associated with authorship. With respect to Shakespeare, for example, they were able to see how this writer's "authorial signature" is connected to the ways in which his characters use words that describe properties of people (professions, social status, age, and the like). At certain points in the analysis, they became aware of the need to limit outliers – a difficult issue in literary studies, since there are few "natural" sources of constraint in the production of words that might ensure a Gaussian distribution (as there might be in a biological system). They were able to investigate this phenomenon – outlier status and its significance to literary analysis – because they had a tool that could demonstrate immediately the consequences of excluding outliers. Such exclusions could, for example, bring a particular writers' work into clearer focus. The iterative design and use process, then, produced improvements in the tools, improvements in the data, and a new question for research: what is the nature of variation in a non-physically bounded system like a literary text?

Our collaborators offered a presentation at a literature forum at the St. Louis University in which they presented a large dendrogram representing Shakespeare's plays in the context of the other early modern dramatic texts in their corpus. The dendrogram itself is ungainly visually: when printed it is over 5 feet long. However, individual clusters show interesting patterns: they are produced by a single author, for example, or they are texts that are all written by members of a particular court circle. But in order to understand why the cluster occurs, critics need to be able to see an exemplary passage from the group and so apply their domain knowledge to the results. The TextViewer was used to isolate one of the clusters composed of plays written primarily by Shakespeare. When this image was put before the group of over thirty experts in Renaissance drama, they immediately began producing hypotheses about its Shakespearean character. Thus literary analysis, which is often done by individuals in their solitary reading, can now be done collectively, multiplying the

power of their domain knowledge.

These tools were later distributed to a class of English Literature graduate students enrolled in the Digital Studies of Renaissance Genres (ENG764) at the University of Wisconsin-Madison, after an initial tutorial. Many students presented final results garnered from the use of this tool, including an analysis of the language of fairies in the works of Shakespeare, as well as the treatment of femininity and "tomboy" behaviors in late Victorian drama. This wide adoption by a group without explicit training in statistics, and to a large variety of datasets, speaks to the wider utility of the tool suite.

5.6 Conclusion

By careful consideration of design decisions as well as close collaboration with domain experts, we created a suite of tools well suited for use by literary scholars. These tools allow for previously impossible connections between statistical phenomena in large text corpora, and low-level patterns in text. This capability is not desired exclusively by literary scholars. The tools we created are flexible enough to adapt to arbitrary text-tagging schemata; sociologists could examine XML coded ethnographic documents to observe and qualitatively analyze social patterns, web designers could compare the HTML of a large number of websites to determine what "works" and what does not in website design. Even the relatively novice user could compare sections of his or her own written work via tags to pick out error-prone sections of code in software, or overly dense sections of prose in documents.

Outside of the domain of texts, future users could use CorpusSeparator to perform quick analysis of multidimensional scaling results from arbitrary high dimensional vectors. The accordion viewer of distribution provides quick overview of a high dimensional dataset in individual dimensions for outlier selection and exclusion. The visual principles in the TextViewer are also naturally extensible outside of the domain of tagged text. Numerous domains (including SoftVis, Natural Language Processing, and more generalized rhetorical analysis) are capable of creating complicated saliency metrics. The ability to zoom quickly to sections of text using focus+context techniques is of immediate use whenever there is a one to one mapping from text to saliency.

This work represents a first attempt to provide tools that support literary scholarship. At a basic level, it is limited to a specific type of scholarship (tag-based algorithmic criticism), and is specific to a particular set of tools (PCA). However, the basic ideas emerging from the work can generalize to a broader category of tools (for example, other statistical analysis like clustering or interpolative decompositions), and even to other domains of scholarship. For the specific tasks we considered, the current implementations have a number of limitations. For example, displaying two dimensional projections in PCA space quickly becomes cluttered as corpus size increases. Performing dynamic subsampling of plays into windows is fast at the level of individual text but slow at the corpus level. Our tools are currently optimized for performing PCA on a corpus of hundreds of texts, with up to one hundred tags. Future work will need to scale to larger sizes of corpora, larger sets of domains, and a larger set of statistical analyses. Still, the tools as they currently stand are easy to use and adequate for many immediate lines of inquiry in the current domain.

Natural extensions to the CorpusSeparator include the ability to use a wider range of dimensionality reduction schemes (including more generalized MDS), and using machine learning techniques (such as SVMs) to simplify the process of separating different groups in the corpus in a way that is readable by the TextViewer. The TextViewer tool, for its part, would need to have its colored underline visual encodings modified to deal with tagging schemes where the mapping from tag to text is not one to one.

In addition to the possibility for future exploration with our flexible and easily-deployable tools, learning to work closely with the needs and cultural milieu of humanities researchers has laid the groundwork for future collaboration and results. This collaboration is useful for humanities scholars, some of whom are skeptical of the place of computational techniques in their discipline, as well as visualization scientists, who can overlook the increasing need for compelling visualization tools in humanistic domains.

6 LAYERCAKE: VISUALIZATION OF VIRAL POPULATION DYNAMICS

New data acquisition techniques that provide large amounts of data provide both opportunities and challenges. While such datasets can enable new kinds of questions to be explored, they also mean that larger-scale analysis and pattern-finding must be sought. Larger data sizes also often bring issues with uncertainty and uneven data quality: large data sets are often merged from multiple sources, acquired over long time periods, and/or use acquisition modalities where quality vs. quantity tradeoffs must be made. To make use of this newly available data, scientists will need tools that explicitly address three challenges: novel questions, large-scale data, and variable data quality. In this chapter, we consider a specific application that provides a case study where these three challenges must be addressed.

We consider tools for comparing the genetic variability in populations of viruses. Such study demands large amounts of sequence data, sufficient to capture the variability in the populations of viruses collected from multiple samples. Collecting such data has been made practical with the advent of "next generation" high-throughput sequencing technology. While this technology makes large scale data collection practical, it also introduces several sources of uncertainty and variability, meaning that data quality must be considered as part of any analysis. A visualization tool must enable a scientist to find interesting patterns of variations in the genetic populations, across multiple populations to find evidence of phenomena such as a viral populations changing over the course of an infection in response to immune system features and responses. The challenge comes not only from the fact that such patterns comprise small features across large datasets, but also that the data has variable quality such that the confidence in the significance of any finding must be considered.

Our result is a novel tool, LayerCake, specifically designed to assist scientists with using next-generation sequence data to study genetic variability in viral populations. We use a color field design (like a heat map) with salience enhancing aggregration and context-preserving zooming that allows for patterns of interest to be rapidly identified and studied. We provide a similar view of the various data confidence

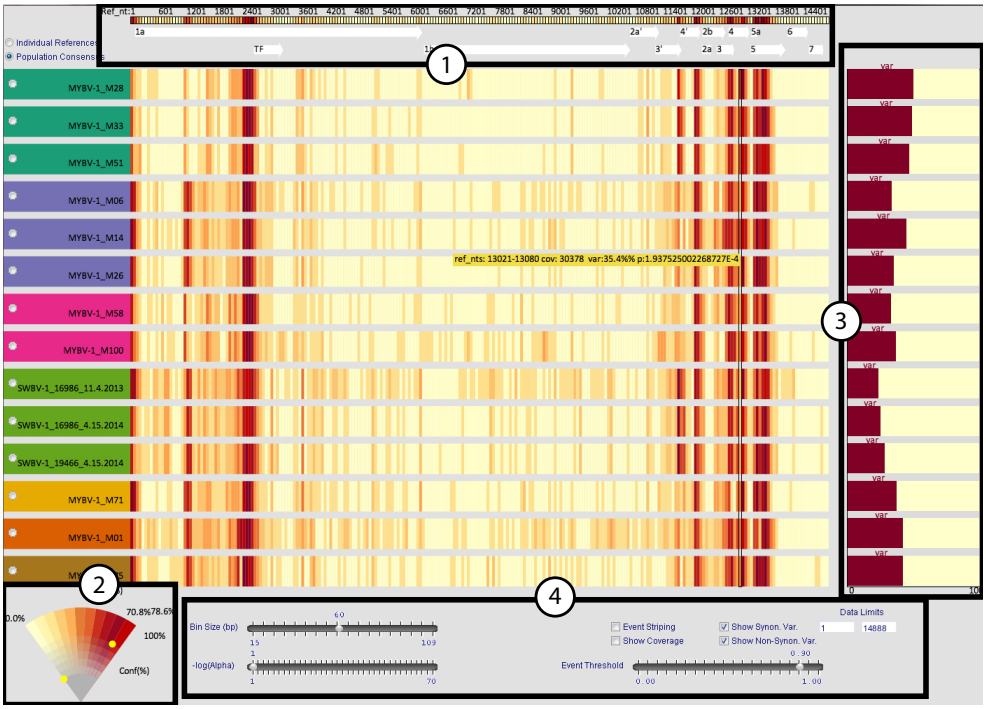


Figure 6.1: An overview of LayerCake, on the SAV dataset (see §6.3). Central to the display are a series of *layers*, each representing a sample of a viral population. Red sections of the layers correspond to areas with high deviation from a reference. The radio buttons on the left allow the viewer to choose between different conceptions of a reference (see §6.2).

1: An overview of variation across all samples, as a colored *histogram*. Red regions correspond to sections of the genome with high variation. Open reading frames are depicted as directional arrows.

2: The *color wedge*, which is both legend and interactive filtering tool. Viewers can move the yellow dots to define their owns standards of important amounts of variation and acceptable levels of uncertainty (see §6.2).

3: If the viewer mouses over a particular region of interest, the *detail view* shows histograms of variation for each population. If the viewer is zoomed into a particular bin, this will show variation information at the level of individual nucleobases.

4: *Interaction tools* for manipulating the range of data, the size of bins, and minimum standards of uncertainty, among other options.

factors, allowing the viewer to create a "fog" over the data color field, reducing the salience of less certain data. "Event striping," a visual technique for highlighting outliers in time series data (see chapter 3 for more discussion) makes specific locations with high variation immediately visible regardless of the aggregation schema. This combined scheme allows interesting data events to be made salient, but only when viewer-defined confidence criteria are met. Our design uses a number of interactive tools to explore the data set, expose relationships between data value and quality, and highlight features of potential interest. The design allows for exploring details on demand: once overall patterns are identified specific data, can be examined.

LayerCake represents another example of applying our experimental work towards improving visual statistics to a concrete domain problem. The colorfield design is meant to afford aggregate tasks and at-a-glance overview judgments (*q.v.* chapter 3). The event striping technique makes outliers, which would otherwise be hidden, readily apparent. Work from this chapter is taken from Correll et al. [2015] (in press).

6.1 Background

Comparative sequence analysis can reveal evolutionary relationships that could otherwise not be discerned. Sequence comparisons can also identify signatures of natural selection and, when analyzed in conjunction with appropriate phenotypic data, can be used to infer the "pressures" driving selection processes.

Prior to the arrival of next generation sequencing (NGS), comparative sequence analysis was largely restricted to the comparison of consensus sequences; that is, sequences represented by the most abundant nucleotide at a given position in a particular sample. The limitations of consensus-level sequence analyses are particularly apparent when examining RNA viruses, as samples often contain a highly heterogeneous "swarm of mutants" — the diversity of which cannot be represented by a consensus sequence. NGS yields thousands of short "reads" that together represent the full diversity of virus sequences in a sample. The assembly of these sequencing reads using either a pre-determined reference, or a reference assembled *de novo* from the reads themselves, allows for the reconstruction of coding-complete viral genomes with the detection of nucleotide variants that exist in as little as 1% of

a viral population. With this paradigm, it is now possible to overlay useful information such as nucleotide polymorphisms, polymorphism frequencies, and sequencing coverage depth onto every position of a whole-genome consensus sequence. Conveying the read depth at each position in conjunction with the above information creates a large multi-dimensional matrix, which can be difficult to display visually in a manner that facilitates the discovery of motifs by the investigator. This problem is further compounded when NGS data from multiple samples ("isolates") is compared, especially when the virus in question has a high degree of intra-sample sequence variability. However, it is in precisely these contexts that a visualization tool can be most useful for evaluating variation across genomes.

Relevant to the discussion of genomic sequence variability is the notion of a sample. Rather than a *sequence* of nucleotides, an individual sample contains the *population* of nucleotides observed at different locations along a genome, derived from NGS data. These populations can be compared to a reference sequence (or "reference population," see §6.2) to define a certain proportion of variability at each location. By visualizing different samples simultaneously we can observe change in variability over time (if we take multiple samples from the same infected organisms, but at different time points), or observe subgroups within a particular virus (if we take samples from multiple organisms and compare them). In both cases the analyst compares multiple samples at once.

We have therefore developed the LayerCake visualization tool to address the problem of visualizing sequence variability in viral populations. In LayerCake, samples are visualized as a colored row or layer in a single view, with variability and confidence information encoded as color. LayerCake automatically aggregates regions of the genome into discrete bins, the size of which can be controlled by the user. This design allows viewers to immediately receive an overview of the entire dataset and quickly locate regions of interest within or among samples. Zooming and side displays allow the user to retrieve detailed, nucleotide-level statistics with a single click. Interaction allows the user to adjust the aggregation, update the metrics used to define variation, or update metrics related to data quality or importance.

Related Works

There are a number of general purpose genome browsers which employ principles from visualization (see [Nielsen et al. 2010] for a survey and discussion of the difficulties in building such systems), including some which are track-based (in which different samples or data types are placed in their own distinct rows and visualized simultaneously). Many of these systems are visually similar to LayerCake in design, relying on comparison across rows or tracks and the heavy use of color to encode value (e.g. see [Robinson et al. 2011, Zhou et al. 2011, Zhu et al. 2009]). The LayerCake system differs in two key ways from these systems: firstly, it supports flexible aggregation and zooming, allowing the analyst to compare across an entire genome and examine small regions of interest simultaneously. Secondly, LayerCake is tailored for NGS data models and can adapt to the specificities of examining this sort of sequencing data (as opposed to treating each of the variables involved in NGS sequencing and alignment as orthogonal tracks).

Tools for the visualization of NGS data specifically must display the heterogeneity of reads at particular locations. Most of these NGS tools have relied on the "scaffold view" in which sequencing reads are assembled against a reference sequence and stacked atop one another. Nucleotides that vary from the reference are highlighted within their respective read, and the frequency of these variants is represented by proportional sequence logos at the bottom of the stack (see [Carver et al. 2012, Hou et al. 2010, Milne et al. 2010, Schatz et al. 2007] for a partial list of NGS visualization tools employing the scaffold view). These sequence logos are notoriously difficult to interpret (see [Maguire et al. 2014, Ray et al. 2014]), making it difficult for analysts to compare variation at individual locations, let alone large regions of a genome. Even if other aggregation strategies are used, the scaffold view is most useful when examining a single sequence of reads, since each scaffold is large and visually complex (requiring the display of potentially thousands of reads, hundreds of base pairs long). Even tools which do not use the scaffold metaphor are still limited to the exploration of variants within a single NGS sample (such as [Ferstay et al. 2013]). A survey of tools for NGS variant analysis ([Pabinger et al. 2014]) confirmed that most tools for this task afford the viewing of only a few separate tracks of reads at a time (one or two per window), although some tools allow the analyst to dynamically combine samples ([Bigelow et al. 2012]).

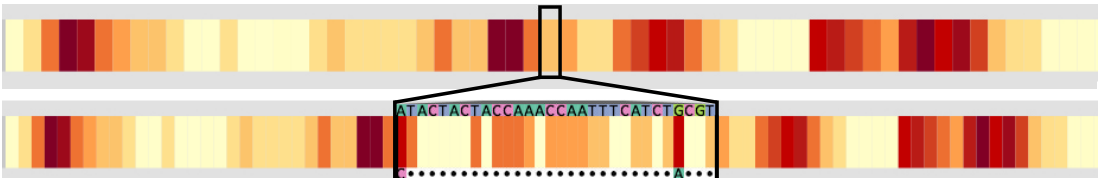


Figure 6.2: A LayerCake layer. Variation at multiple sequential locations on the genome is averaged together into bins, presenting an overview of the entire genome at once (above). By right clicking on a bin (below), the viewer can recover specific information about a section of the genome while keeping the overview in context.

One exception to the tools which present only one (or a few) samples at a time is the Sequence Surveyor tool ([Albers et al. 2011]). Originally designed for the analysis of linkage and conservation across large numbers of genomes, Sequence Surveyor encodes each genome as a row in a large display and aggregates sections of the genome into discrete colored blocks, allowing hundreds of sequences of millions of base pairs in length to be summarized on a single screen. [Swihart et al. 2010] recommend a similar layered design for observing trends in longitudinal data. The initial design of LayerCake adapts Sequence Surveyor techniques to the variant analysis task while maintaining a scalable design based on the arrangement of colored rows of blocks. A key difference between the two methods is that Sequence Surveyor is for viewing static sequences of genes or nucleobases. LayerCake deals with the simultaneous comparison of multiple sample *populations* of sequentially organized reads. Instead of one bit of information per location (for instance, "what is the nucleobase at this location?"), LayerCake must contend with at least four (how many of each type of nucleobase are at this location?). This problem becomes even more challenging when we compare populations to each other. §6.2 expands on this formalistic difference.

6.2 System and Methods

LayerCake, as a tool for the quick, visual comparison of large amounts of genomic variability data, has three primary design components:

1. Techniques for visually **aggregating** large amounts of genomic variability data from multiple samples and populations.

2. Techniques for calculating and displaying various conceptions of **variation and reference**

3. Techniques for calculating and displaying various conceptions of data quality and **confidence**.

Central to LayerCake is the notion of a layer — each separate sample of viral sequence data is visually represented as a row of colored glyphs. Figure 6.2 shows an example of a LayerCake layer; red regions of the layer correspond to locations along the genome for which this particular population has high variance compared to the current reference. Figure 6.1 shows the entire LayerCake system: dozens of discrete layers organized and displayed simultaneously, with annotations and tools for viewer interaction.

Aggregation

While viral genomes are smaller in length than mammalian genomes (tens of thousands of nucleobases rather than billions), it is still not feasible to visually present all the information from dozens of samples simultaneously. In order to present a meaningful overview in limited space, LayerCake compresses the sequence and chooses a visual representation of each sample that is compact enough to afford the simultaneous presentation of many samples in a single screen. Sequence compression must be considered not just in pixels, but also in visual complexity. By definition, this compression inherently aggregates some information, but LayerCake gives the viewer the ability to recover these details on demand.

The primary form of aggregation LayerCake supports is binning: contiguous locations in the genome are aggregated together into discrete blocks. The resulting color of the block represents the average variation of all sites within the block. We used color to encode data rather than, for instance, vertical position (as in a line graph or scatterplot) as prior work has shown that viewers are better at estimating and comparing average color values from sequences as opposed to average positional values ([Albers et al. 2014, Correll et al. 2012b]). A typical viral genome consisting of tens of thousands of nucleobases can then be reduced to a few hundred blocks, which can easily fit within the dimensions of a standard computer monitor. The viewer can interactively choose how many base pairs are contained within a single

bin, which alters the aspect ratio of each block as the entire layer is stretched to fit the available space. To guarantee the visibility of each block, the number of nucleobases within a block cannot be reduced to a number so low that a block would be less than a pixel wide. Conversely, the number of nucleobases within a block cannot be a number so high that the display of a bin's contents will not fit in the available space. In practical use cases, viewers tend to make bins dozens of base pairs large, to reduce the visual complexity of the display while still permitting the investigation of small-scale features in the data.

Recovering Detail

LayerCake averages together multiple locations into a single bin; this aggregation can create ambiguity (is this location somewhat red because many of the locations within it are somewhat variant, or is it because there is one highly variant location surrounded by locations with little or no variation?) and erase details (since a region of interest in the overview could ambiguously refer to any location within a region dozens or hundreds of nucleobases long). We therefore include two techniques to recover detail: focus+context lenses and "event striping."

When the viewer right-clicks on a particular bin, LayerCake expands the contents of the bin to a detail view and shrinks the rest of the layer to maintain total length. Since this zooming occurs discontinuously, this is a "table" or "Manhattan" lens (see [Carpendale and Montagnese 2001] for an overview of this and other lens types for information displays). This detail view explicitly shows the variation at each location within a bin. Figure 6.2 shows an example.

The overview merely shows the average value of each bin. A single point of high variation can be lost in this averaging process. If the viewer wishes to see small scale (but important) features, we support a technique called "event striping" (see Fig. 6.3). When enabled, the viewer selects a threshold of interest, and then LayerCake will draw thin red stripes on bins which contain locations where variation meets or exceeds this threshold. For instance, a viewer might use event striping to highlight locations on the genome where more than 50% of reads are variant. An individual bin in the main display might, on average, have significantly less than 50% variation, but still have a number of visible red stripes which suggest that the viewer might wish to investigate this bin with zooming. Event striping increases

(a) The original layer.

(b) The same layer, with "event striping" enabled.

Figure 6.3: An example of event striping — since each bin contains the average of information from many locations, it is possible that specific locations with high variation will be drowned out by their low variation neighbors (as in Fig. 6.3a, where there appears to be very little variability in the last few bins). Event striping draws a red bar on high variation outliers, adding to the visual complexity of the display but showing outliers that could be missed when data are aggregated (as in Fig. 6.3b, where three specific points of high variation are now visible).

the visual complexity of the display (since the number of events is only limited by the number of locations in the data set), but allows viewers to find locations that would otherwise be lost in the averaging. Prior work has shown the utility of event striping for identifying outliers in sequence data ([Albers et al. 2014]).

Defining Variation and Reference

Variation presupposes a non-variant sequence or population from which deviation can be measured — a reference. Typically this is a *reference sequence*, however in LayerCake we expand on the definition of reference to include more complex situations — for instance we may be concerned in how a viral population has changed compared to a particular time point, as opposed to some initial pre-infection reference. Different data sets will have different *references* (sequences, pseudo-sequences, or populations against which we define variation), but they also might have different *definitions* of what constitutes a valid reference. These definitions might even change dynamically over the course of a session.

Variation from a Static Reference Sequence

Let $\overrightarrow{Reads}_n$ be a four dimensional convex vector whose components sum to 1.0, denoting the population of all reads at a location n. $\overrightarrow{Reads}_{n,A}$ would then be the

proportion of reads at location n that were identified as adenine. Let Ref_n denote the reference at n. If Ref_n is a static, single base pair, then the *variation* from the reference at n is straightforward to compute. Namely, it is the percentage of reads which do not match the reference base pair:

$$1.0 - \overrightarrow{Reads}_{n,Ref_n} \tag{6.1}$$

Variation from a Reference Population

In real tasks the assumption of a static reference is frequently violated. For instance, we might want to compare against a population at a particular timepoint, or an individual might have been infected by a diverse population of viruses rather than a single homogeneous population. In this case we would represent not just the sample, but also the *reference* as another four dimensional vector $\overrightarrow{Ref}_n$. Variation should then be represented as some sort of *distance* from one vector to another. Many possible distance metrics exist; however, for this task the distance metric ought to be easily comparable to equation 6.1 above: it should preserve the semantic meaning of "more" or "less" variation, and have a range in the interval [0,1].

We chose a distance metric based on the central metaphor of swapping. I.e., to make two locations identical one would change individual reads until the distributions matched. For instance, if the population was entirely adenine at a location, but the reference was one entirely cytosine, one would "swap" out 100% of the adenine and replace it with 100% cytosine: 100% of the reads would be swapped, so the total variation would be 100%. Likewise if the reference was 50% A and 50% C, only half as many swaps would need to be performed, so variation would be 50%. This behavior of examining distance at each dimension (or nucleotide) individually and then summing up, is captured by the ℓ^1-norm, or Manhattan distance. To avoid double-counting swaps (adding more adenine by necessity means subtracting quantities for another nucleotide), we divide the ℓ^1-norm by 2.0 to derive the final metric for variation between two populations:

$$\frac{\left\| \overrightarrow{Reads}_n - \overrightarrow{Ref}_n \right\|_1}{2.0} \tag{6.2}$$

Synonymous and non-synonymous variation

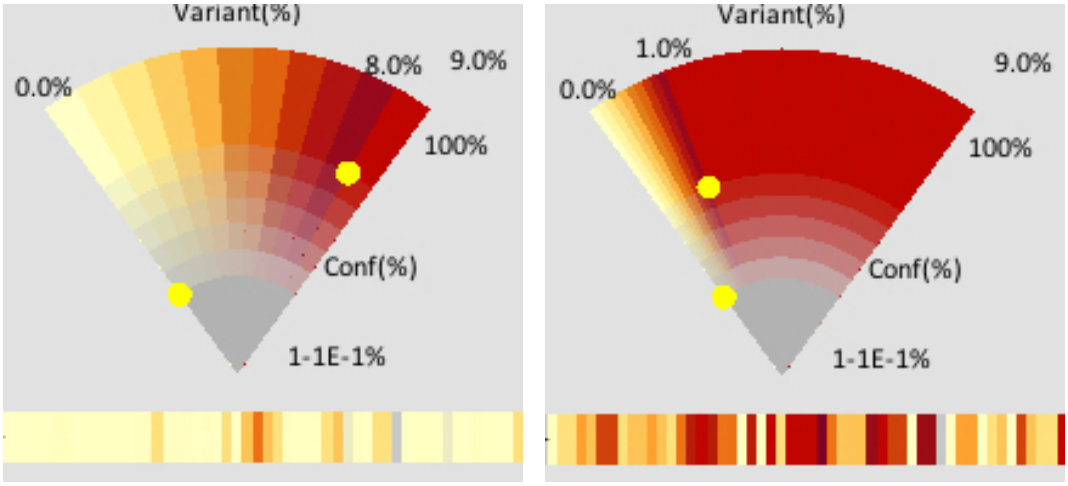
The analysis of viral NGS data within a sample or from multiple samples can be used to identify signatures of natural selection — an exercise that can can yield powerful biological insights, especially when supported by phenotypic data. At the core of this analysis is the identification of "non-synonymous" mutations: those which change the amino acid sequence of the encoded protein. Since mutations are generated randomly, a high density of non-synonymous mutations in a particular region is indicative of natural selection favoring diversification of the respective protein sequence: a phenomenon referred as "positive selection." The opposite is also true: a paucity of non-synonymous mutations indicates selection against protein sequence changes, (ie. "purifying selection"). To enable the visualization of non-synonymous variation across the genome, LayerCake can display either non-synonymous mutations, synonymous mutations, or both when open reading frame (ORF) annotations are included in the input reference sequence. A mutation is considered non-synonymous if it would result in a changed amino acid for even one of the relevant ORFs. The metrics presented above extend to this case by filtering out the relevant types of variation before calculating total variation.

Defining References in LayerCake

LayerCake considers three different reference scenarios:

1. **Individual References**: In this scenario each discrete population considers variation *separately* — for each population the user either provides a reference sequence (for instance from a FASTA file), or LayerCake will generate a consensus sequence for each sequence. This scenario highlights regions which have systematically high variation *within* a sample.

2. **Population Consensus**: In this scenario variation is defined with reference to a single reference sequence. This sequence is either provided from a source file (for instance a GFF file), or LayerCake will generate a single consensus sequence by voting. That is, if there are 10 populations in the data set, and 6 of them have an adenine at a given position, then the population consensus will also be adenine, regardless of the read depth of any individual sequence. This



(a) High standard of interest. (b) Low standard of interest.

Figure 6.4: The LayerCake color wedge, showing the mapping from uncertainty and variation to color. Highly uncertain data are all mapped to the same grey color, giving the visual impression of data receding into a "fog" of unimportance. The two yellow dots can be moved by the viewer to redefine standards of interest and importance. On the right the viewer has moved the topmost yellow dot counterclockwise, making all locations with more than 1% variation bright red, which is interactively reflected in the layers.

scenario affords the quick apprehension of particular regions of particular samples that have high variations.

3. **Per Sample Comparison**: Individual samples, through the method described in equation 6.2, can be used as a pseudo-reference for the rest of the data set. This scenario readily shows variation *between* samples, and also the identification of sub-groups of samples. See Fig. 6.7 for an example.

Users may dynamically choose between different reference scenarios, even in the course of a single session. For instance, if one is interested in general regions where variation occurs, they might begin with individual references. Once those locations are identified, they might choose a particular population as a reference, to see if there are groups of populations that have different sorts of variation in these hotspots.

Confidence Visualization

Uncertainty about variation at a particular location on the genome can occur for a number of reasons. There can be error in assembling reads, aligning reads, identifying base pairs, and sampling error that could arise from insufficient read coverage at a location.

Uncertainty data, no matter the source, must be visualized along with the variation information, especially for tasks where the viewer must decide which locations of the genome require more detailed analysis — highly variant but uncertain information might warrant less attention than a location with less variation but little uncertainty.

In LayerCake, color is used in each layer to display information. Color has been shown in [Albers et al. 2014] to be a useful visual variable for helping analysts to quickly find outliers as well as estimate average value in regions. Since we have two types of information to display (frequency of variation, and average uncertainty), this means that we must use a bivariate color map to represent the data. In order to avoid many theoretical obstacles to creating these color scales (see [Trumbo 1981]), we presume that highly uncertain values are unimportant, regardless of the variation at this location. Thus, rather than our color map resembling a square (two equal orthogonal axes), our color map resembles a wedge (with the uncertainty axis converging to a point). This makes the choice of colors significantly easier, while maintaining the desired visual behavior (important regions are highly visible, unimportant regions recede into the background). While we interpolate in multiple color attributes (both hue and saturation) to make discriminability easier, as confidence decreases it is intentionally more difficult to distinguish colors; in effect we have fewer distinct color values as we descend the wedge, replicating the intended effect of making value less important as confidence decreases. Figure 6.4 shows the color wedge in detail. While a bivariate encoding (such as color and size, or color and orientation) would allow us to faithfully present value and confidence simultaneously, we wished to make it easier for analysts to filter out uncertain (and likely irrelevant) portions of the dataset without having to integrate multiple channels of visual information.

6.3 Discussion

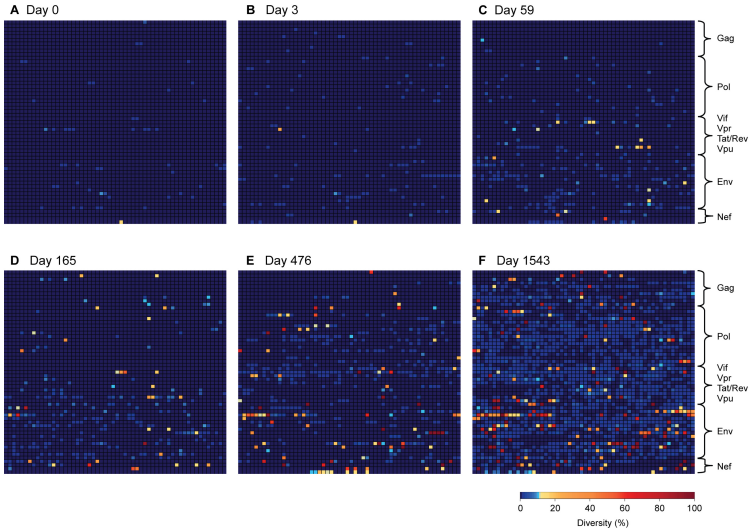
The LayerCake system has been widely deployed across a number of viral datasets. In this section we consider three datasets under analysis by our collaborators that highlight the benefits of the LayerCake system: the presentation of a genome-scale overview of data, the ability to interactively alter notions of reference and variation, and the alignment and aggregation of many samples in a single, all-encompassing display.

In addition to finding regions of high variation (as described in Correll et al. [2011b]), LayerCake affords longitudinal comparison of variation (as in Fig. 6.5), and allows for the identification of subgroups with similar variation signatures (as in §6.3).

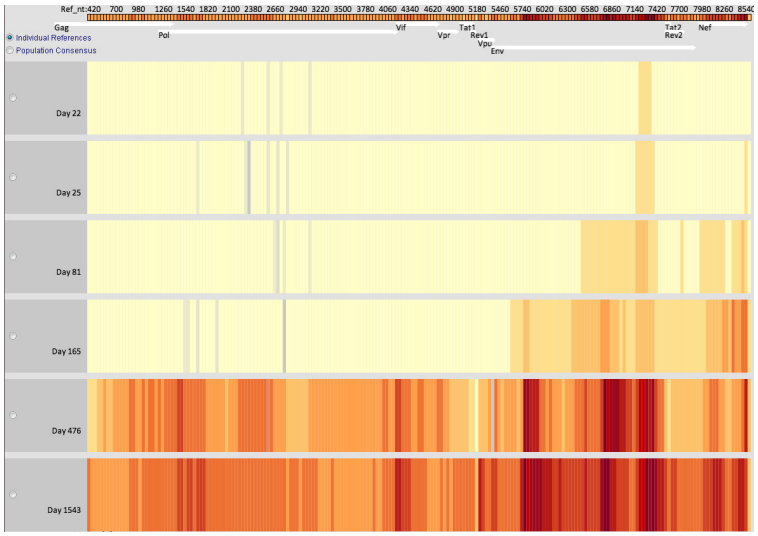
Simian immunodeficiency virus (SIV)

LayerCake allows for the quick appraisal of regions of high variation. In [O'Connor et al. 2012], 18 Mauritian-origin cynomolgus macaques (MCMs) were infected with clonal simian immunodeficiency virus (SIVmac239). At 4 weeks post-infection whole-genome sequences of SIV from each individual were obtained and examined for patterns of non-synonymous variation in LayerCake. Although SIV from each animal had unique variations, consistent patterns of variation among animals were also apparent, with the three most prominent corresponding to known CD8+ T cell epitopes — i.e. regions of the corresponding viral protein targeted by adaptive cellular immune responses. Functional characterization of these immune responses were able to link the accumulation of non-synonymous variations within these sites to the magnitude of the epitope-specific CD8+ T cell response, providing a compelling mechanism for the accumulation of these mutations.

Visualizing systematic patterns of variation between these subpopulations was difficult with existing visualization tools. The initial deployment of LayerCake made these patterns visually clear, and allowed for in depth exploration of different viral populations. Figure 6.6 shows a comparison of LayerCake with the scatterplot visualizations used in the original research for displaying systematic variation between population groups.

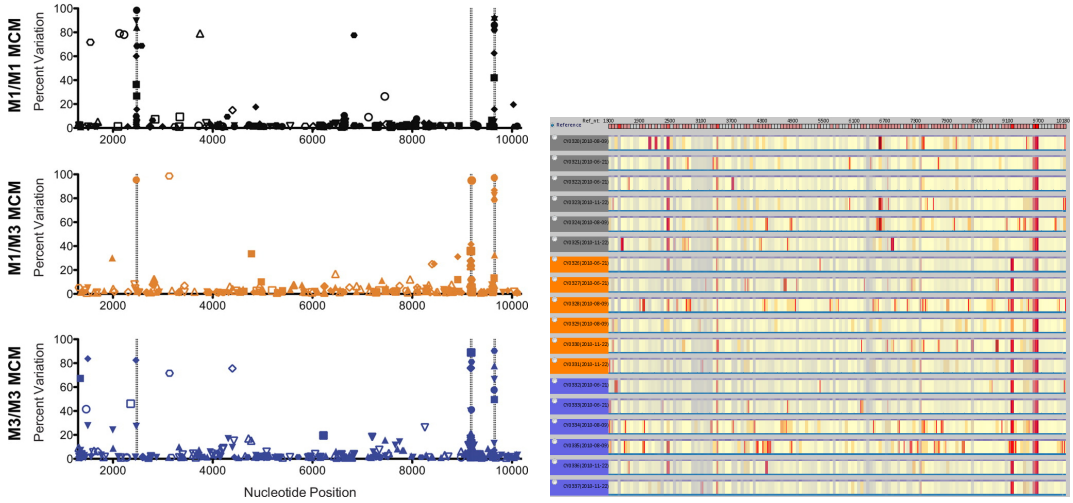


(a) A heatmap of non-synonymous diversity in HIV-1 over the course of an infection.



(b) A LayerCake view of the same data.

Figure 6.5: An example of the utility of LayerCake for viewing systematic patterns of variation, illustrated by examining the evolution of HIV-1 in an infected individual over time. While the standard heatmap display (Fig. 6.5a) makes the overall trend visible (variability increases over the course of the infection), it is difficult to compare specific locations over time. In LayerCake (Fig. 6.5b), each row represents the viral population at a different timepoint in the infection. Change over time at a particular location can be estimated by visually scanning a particular column. Annotations (across the top of the LayerCake display) also adds context to the pattern of variation accumulated over time.



(a) Scatterplots of variation in several viral populations.

(b) A LayerCake view of the same data.

Figure 6.6: An example of the utility of LayerCake for viewing systematic patterns of variation, here on a pyrosequencing dataset of a particular simian immunodeficiency virus (SIVmac239), reproduced from O'Connor et al. [2012]. Each color represents a different population subgroup. In the scatterplot, overplotting makes the display very cluttered, and it is difficult to see if a point of high variation is an outlier, or part of a systematic pattern. Dotted lines represent sections of the genome where the three groups have differing patterns of variation. For instance the black M1/M1 group has a large amount of variation at the location of the first dotted line, but the other two groups do not. In LayerCake, each sample has dedicated space for analysis, and interaction allows the viewer to choose thresholds of interest. The six grey rows correspond to the black group in the first scatterplot 6.6a, and the orange and blue rows correspond to the second and third scatterplots. The same patterns are visible in the LayerCake display, but by placing each sample into a distinct row, it is easier to distinguish between outliers and general, consistent patterns.

Figure 6.7: An example of how changing the conception of the reference in LayerCake can identify intrasequence patterns of variability, here on a dataset of simian arterivirus (SAV). Dataset from Bailey et al. [2014a]. By defining variation from a particular population rather than a reference sequence, we can easily identify subgroups. The first row is selected as the reference population. Here the first three rows are very similar to each other, but not to the other sample, indicating a meaningful subgroup.

Simian arterivirus (SAV)

LayerCake also allows for the description of nucelotide variation and deep population analysis of novel viruses for which little or no prior data on sequence evolution exists. In [Bailey et al. 2014a;b], we used LayerCake to examine nucleotide variation in novel, highly-divergent simian arteriviruses that we discovered in wild red colobus monkeys and yellow baboons living in Uganda and Tanzania, respectively. With a population-wide consensus selected (the Population Consensus option described in §6.2), LayerCake revealed several genomic regions with high levels of non-synonymous diversity. Follow-up analysis showed that the region with the most intense signal was within the ORF encoding the major envelope glycoprotein. When compared with functional data from more extensively-characterized arteriviruses, this region aligned with the primary neutralizing antibody epitope of

these viruses (ie. the region of the viral protein targeted by adaptive humoral immune responses) — again providing mechanistic insight into the selective pressures driving the accumulation of non-synonymous mutations. Selecting individual references in LayerCake (the Per Sample Comparison option described in §6.2) quickly revealed varying degrees of viral sequence homology between animals, reflecting the pattern of transmission among individual monkeys (see Figure 6.7). In the red colobus, this exercise identified one animal that was super-infected with two unique virus strains.

HIV-1

Another use case for LayerCake is the visualization of viral evolution. By comparing longitudinal samples from a persistently-infected individual LayerCake allows the investigator to examine patterns of sequence change over time. Henn et al. [2012] uses heatmaps similar to LayerCake to visualize this data, but the metaphor of mutually aligned layers, combined with the ability to interactively adjust parameters of interest, affords robust analysis of temporal variation data. Figure 6.5 compares LayerCake's view of variability over time to a standard heatmap display.

6.4 Conclusion

LayerCake is a full-featured visualization tool for exploring patterns of variability in viral genomes. We have deployed LayerCake to experts in the field and incorporated their feedback into further refinements. The tool, and more broadly the analytics and visual metaphor of the per-sample layer, has been applied to a large number of data sets, with positive scholastic results. The LayerCake tool is freely available and extensible to data sets beyond those we present.

The LayerCake system contains concrete scenarios where we could apply lessons learned in Ch. 3: to support statistical tasks such as aggregation and outlier-finding, we needed to adopt non-standard encodings like focus+context binning and event striping. In many cases the analysis required looking for patterns at multiple scales, and incorporating them with specific domain knowledge (for instance, the location

of important sites along the genome). Visual statistics, supported by interaction, supports this potentially complex analysis.

7 DISCUSSION

In this work I defend the thesis that we can improve visual statistics by making informed choices about what aggregate information to display, and by creating encodings which are mindful of how humans reason and argue about statistics and uncertainty. I explore specifically situations where visualization (which exploits the powerful human perceptual system) and statistics (which exploits the power of mathematics to summarize and structure data) can work in concert to accomplish the goal of improving human decision-making.

Human beings, through the techniques of visual aggregation, can perform sophisticated statistical tasks reliably and effectively even without explicit statistical training. As designers, we can choose to support these higher level statistical tasks, even if doing so comes at the expense of obscuring the raw data, or making lower level tasks more difficult.

7.1 Limitations

Much of the empirical work presented in this document is based on crowd-sourced experiments. Crowd-sourcing gives us access to a larger subject pool, and faster response times than in-person human subjects research. It also allows us to measure performance of populations who, on the whole, do not have explicit statistical training. However, the tools I showcase in this document are primarily deployed to experts in their respective fields, intended for long-term analytical work. Cognitive concerns such as prior knowledge, expectation, and domain-specific beliefs might fundamentally change how visualizations are interpreted, and how statistics are calculated and compared. It remains future work to conduct the longitudinal and somewhat qualitative work of analyzing patterns of interaction between statistics and visualization as analysts use tools to solve domain-specific problems.

There are also ethical considerations to visual statistics which are outside the scope of this work. It is not the case that "the data speaks for itself," especially when higher-level statistics are the driving force behind analysis and decision-making. In many cases, the encodings I evaluate *obscure* the actual data values. This places a burden on the designer to make sure that we do not obscure important data, or that

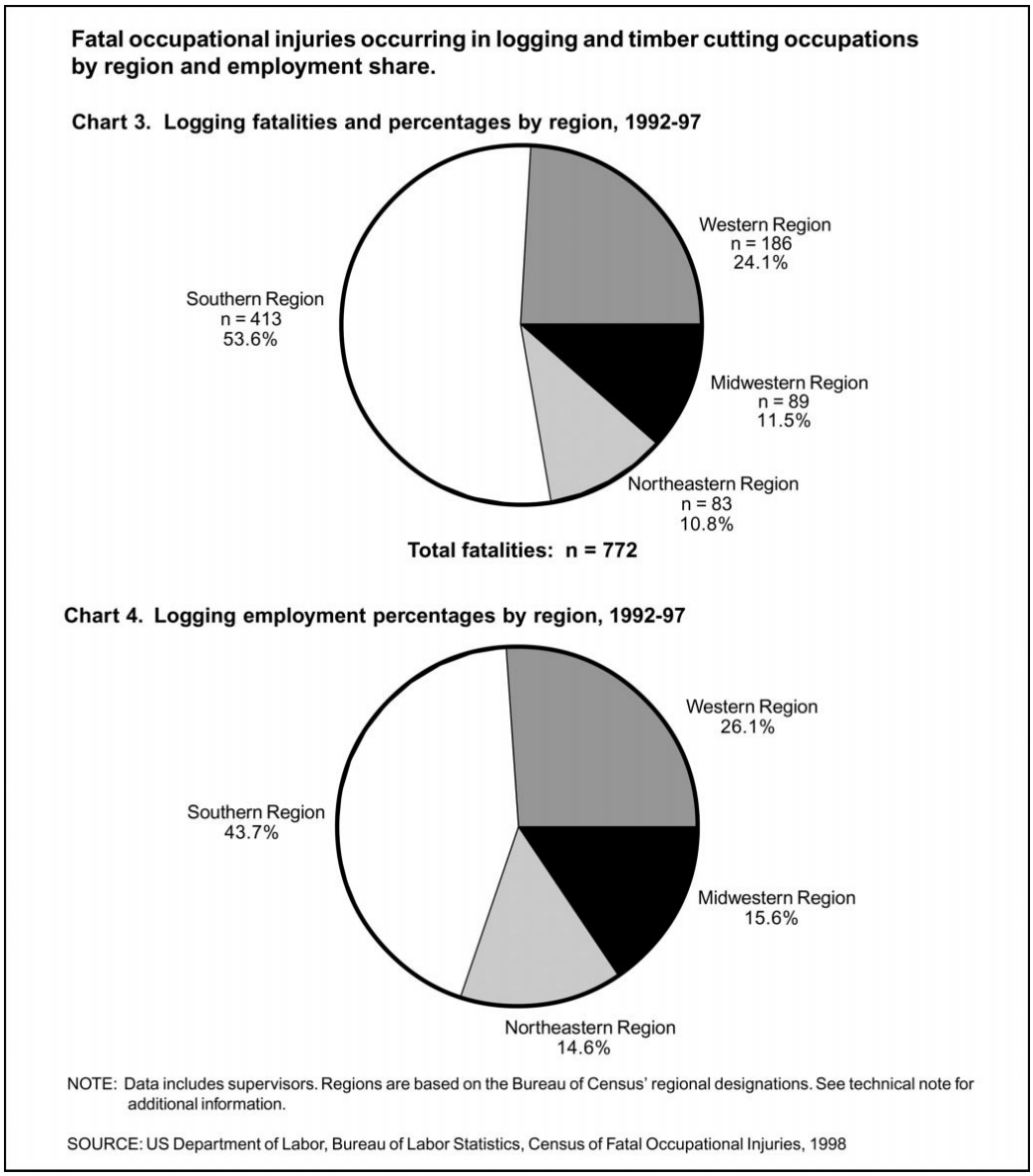


Figure 7.1: "Cruel" pies from Sygnatur [Sygnatur 1998], as used in Dragga & Voss [Dragga and Voss 2001]. While in the abstract both pie charts are encoding identical data types, one pie shows employment records while the other shows fatalities. By excluding the human element and moral connection to the data at the expense of abstract similarities in encoding, Dragga & Voss argue that this information visualization is "cruel." Adornment and embellishment beyond the minimalist pie chart could better emphasize the qualitative difference in information types.

in our haste to *guide* viewers into performing statistically-savvy analysis we do not *manipulate* them into seeing to data in only the restrictive ways we intended. For instance, the mere decision to present *any type of chart at all, even an irrelevant one* can create measurable impacts in how data are perceived [Tal and Wansink 2014]. Data domains are also not created equal. For instance, if I present two identical styles of chart, one of sales figures and another of casualty figures, by treating these two data types identically I am giving them equal weight to the viewer even though from a moral or ethical standpoint the latter ought to have vastly more weight than the former. Dragga and Ross call this equal weighting "cruel," and call for embellishment to lend additional emotional weight to datasets that are otherwise structurally identical to more "neutral" sources [Dragga and Voss 2001] (see Fig. 7.1). In general, creating an ethics of visualization is a largely unsolved problem, both in this thesis work and in the field generally.

7.2 Future Work

"Visual statistics" as I define it is intentionally broad; likewise the language of "improving" implies a constant evolution rather than a termination. To this end, I intend to conduct a great deal of additional research; some of the projects I will mention are currently in process, but others have yet to be fully fleshed out.

- **Characterizing visual aggregation under uncertainty**. Many of the experiments I have already conducted on the human capability to extract statistical information from visualizations have been discussed only briefly or not at all in this document for reasons of scope, space, and simplicity. A particular direction of future research is visual aggregation in domains where there exists risk or uncertainty. A research question I have begun exploring in this space is **can human uncertainty in visual aggregation tasks be used as a proxy for statistical models of uncertainty?**. That is, can humans perform the equivalent of "visual t-tests" — using their own pattern recognition and summarization capabilities to perform perceptual inferences that can be equated to statistical tests? Our work with error bars (chapter 4), as well as work such as Wick-

ham et al. [2010], provide initial evidence that such judgments are possible, although we have yet to fully characterize and evaluate these capabilities.

- **Measuring how visual aggregation affects decision-making**. Estimation and comparison are often just steps along the road of analysis. The intended "product" of analysis is often a concrete decision. One contention in this work is that there is benefit in trusting the viewer to perform visual aggregation rather than supplying the "answer" (explicitly encoding the relevant statistic). While we justify this claim in the document through reasons of complexity (it is impractical to display every potentially relevant statistic simultaneously), I believe that there are concrete benefits to letting viewers build up the statistics by themselves. Visual aggregation might be a "beneficial difficulty," which increases cognitive engagement with the data [Hullman et al. 2011]. Visual aggregation might also create more trust and ownership in conclusions derived from the data, as opposed to if these conclusions were simply automatically generated by the system. Lastly, the viewer might be more sophisticated than the statistical model used by a particular system, with a more robust capability to filter out outliers and fit models to data. All of these benefits ought to have measurable impacts on decision-making, and I intend to begin the process of constructing experimental methodologies which can detect these differences in the visualization context.

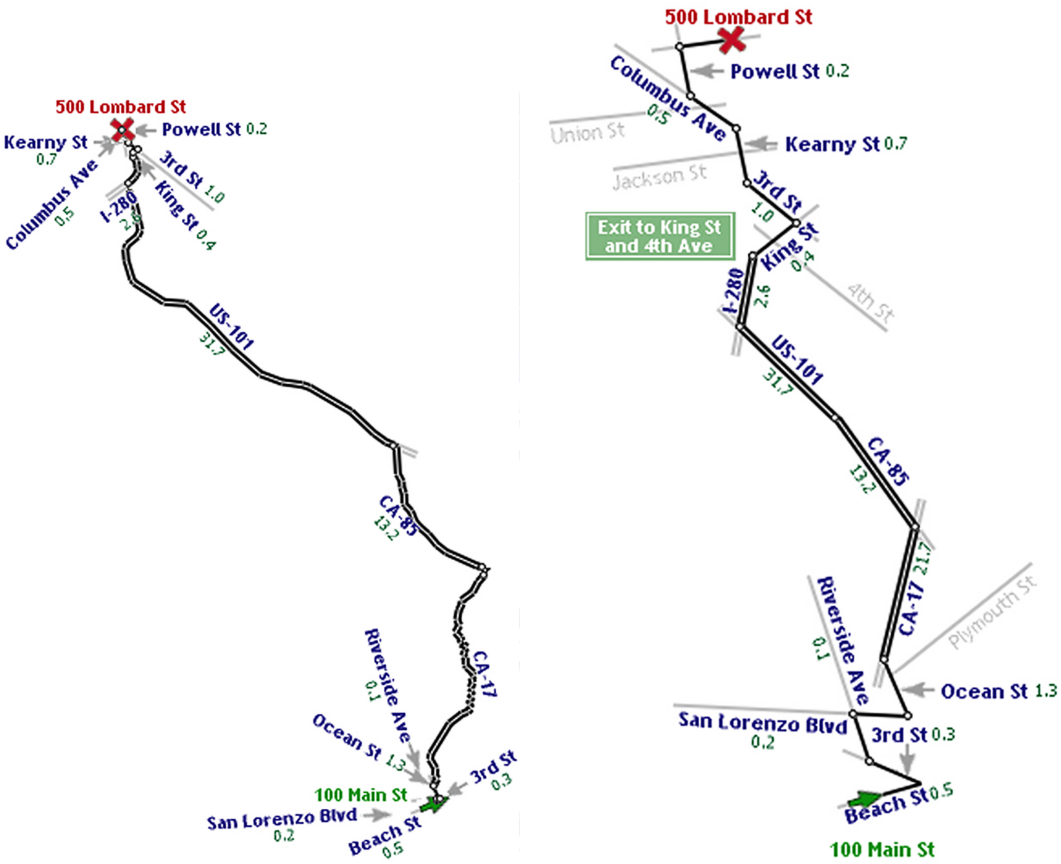
- **Creating additional applications**. Many of the experimental results presented in this document, as well as other related experiments I have performed, remain somewhat abstract. By situating them in the framework of visual statistics, I tacitly maintain that the capabilities I measure are directly applicable to real-world analysis problems. By creating more tools that make good use of visual statistics, I can provide concrete benefit to analysts in a variety of fields. These interventions are especially timely given the ubiquity (and to some extent democratization) of data collection and curation: many domains have access to data, and access to statistical techniques to interpret this data, but not necessarily visual tools which express the data and statistics in widely interpretable way. It is impractical to give everyone with access to data a graduate education in statistics, but we might be able to create simple, widely

applicable tools which supplant some of this lack of knowledge.

7.3 Towards a Rhetoric of Visual Statistics

A central component of my thesis work is that **how data are presented impacts the decisions we make.** This impact is not just in terms of how the data are perceived, but also in what we decide to do with the data we have. For instance, if I present risk factors to you textually (e.g., "1 in 5 people with small cell lung cancer survive for at least 5 years") you might perceive the risk as lower than if I were to present the same information graphically, but also *invest* less money in order to offset this risk [Lipkus and Hollands 1999]. Our job as designers is not to dispassionately present facts to viewers, but to be advocates for rational decision-making, and mindful of how we persuade with visualizations. While recent work has begun to adopt quantitative evaluation techniques to explore these factors [Pandey et al. 2014], often these considerations are somewhat opaque to traditional evaluation techniques such as measuring task performance. One solution is to borrow the language of critique from the arts and humanities, the somewhat idiosyncratic assessment of how we create meaning [Kosara 2007].

Often, thoughtful considerations of the nuances of design comes at the expense of generally accepted design guidelines. That is, if we approach the process of designing visual statistics as "merely" the presentation of data, or worse still the "efficient" presentation of data, we lose out on the nuances and higher goals of persuasion and utility. Visualization, while *data-driven*, ought to be *human-centric*. Human beings are the ultimate targets of our designs; it is not enough to simply *show* them the data, we must also make sure that human beings *comprehend and responsibly use* the data. There are many cases where designs that are "good for the data" may not be "good for people" — perceptual and cognitive biases, human methods of analysis and reasoning, and human priorities all contribute to situations where we might wish to distort, obfuscate, or scramble the data in our visualization in order to make things clearer or easier for humans. This assistance is frequently necessary — humans make habitual and severe errors when making judgments about uncertain information, and our intuitions about risk and reward are frequently flawed [Tversky and Kahneman 1974].



(a) A traditional geospatial route.

(b) Route Map from Agrawala and Stolte [Agrawala and Stolte 2001].

Figure 7.2: A beneficial distortion for the task of navigating using a map. In the non-distorted version the long (but from a direction-finding sense uninteresting) highway portion of the route dominates the display, whereas tight turns that take up very little geographic space are hard to make out. Distorting the map makes these turns easier to see.

Chapter 4 presented an example of "de-biasing" — we identified a distortion in human reasoning, and adjusted the presentation of information to correct for this. This presupposes both that there is a "correct" interpretation of the data, and that our objective as designers is to present the data as clearly as possible. In many cases these assumptions are violated: there may not be a ground truth, or, if there is, it might be more important to present something that is not strictly true, but is strictly useful.

An example of a situation where visual fidelity to the data is potentially harmful is Agrawala and Stolte's "Route Maps" [Agrawala and Stolte 2001] for depicting suggested routes for travelers to follow. Rather than naïvely present the map with the route superimposed over it, Route Maps systematically distorts the data to give greater weight to areas of the route that are more important or require more attention from the viewer. For instance, the first and last few steps of a route from one city to another might involve a large number of turns in relatively small geographic areas (maneuvering in city blocks), whereas the middle of the route might be a small number of turns in a large geographic area (the large stretch of highway between the two cities). Route Maps magnify and exaggerate these critical steps and the beginning and end, and simplify and compress the large stretch of highway driving (see Fig. 7.2). Maps more generally are a case where distortions are the cost of doing business: simply projecting the pseudo-sphere of the globe to two dimensions introduces distortions, and many techniques common in the design of maps are informed choices of what distortions to include or exclude in a design [Monmonier 1991].

Even the choice of what visual variable to use to encode a data variable can create conflicts between utility and clarity. For instance, different visual variables have different levels of accuracy for comparison tasks [Munzner 2014], but different variables also have a different *semiotic* connection with quantities of interest [MacEachren et al. 2012]. Blur and saturation are, in the mental model of many viewers, tightly coupled with uncertainty. This is despite the fact that viewers are remarkably poor at comparing blur and saturation in actual displays, perhaps to only a few qualitative levels of difference [Boukhelifa et al. 2012]. Designers are therefore faced with two seemingly opposed definitions of clarity: task clarity (ease of interpretation for specific comparison task) or semiotic clarity (ease of interpretation in terms of alignment with the mental model of the viewer). We can make an informed decision to support either type of clarity. An example is in the work of Borkin et al. [2011], where the rainbow colormap was self-reported as being "clearer" for their users to interpret (since it was the ramp they were used to using in their existing tools). It was only after concrete measurement of the costs of such a colormap in terms of errors in analysis that users were willing to switch to superior non-spectral versions.

In general, we should remember that statistics are inherently rhetorical — they are collected and curated for the purpose of argument [Rosenberg 2013]. How we present the data ought to be in service to our goals, which may or may not have much to do with the clarity of this presentation. This "clarity" itself may have a multitude of meanings: for instance journalists, technical writers, and academics all strive to clearly communicate in writing, and yet it is stylistically very easy to distinguish a news article from a technical manual from a journal article. What it means for a visualization to clearly make its case to the viewer may have little to do with the literal visual clarity of data points. This work offers initial evidence that, in some cases, the presentation of statistical values is better supported by non-standard encodings.

The empirical work presented in this thesis, as well as the systems we deployed, support the language of tradeoffs— there is no one encoding which will perfectly present all relevant statistics to all viewers. We must, as designers, make informed choices about what information to explicitly encode, make easy to extract, or distort in service of our tasks. It is my hope that this dissertation provides experimental evidence to guide this decision-making, and provides concrete scenarios where the design of visual statistics has had direct, practical applications.

REFERENCES

Agrawala, Maneesh, and Chris Stolte. 2001. Rendering effective route maps: improving usability through generalization. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques*, 241–249. ACM.

Aigner, Wolfgang, Alexander Rind, and Stephan Hoffmann. 2012. Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions 31(3pt2):995–1004.

Albers, Danielle, Michael Correll, and Michael Gleicher. 2014. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the 2014 acm annual conference on human factors in computing systems*, 551–560. ACM.

Albers, Danielle, Colin Dewey, and Michael Gleicher. 2011. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics* 17(12):2392 – 2401.

Alvarez, G.A. 2011. Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences* 15(3):122–131.

Alvarez, George A, and Aude Oliva. 2008. The representation of simple ensemble visual features outside the focus of attention. *Psychological Science* 19(4):392–398.

———. 2009. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences* 106(18):7345–7350.

Ariely, Dan. 2001. Seeing sets: Representation by statistical properties. *Psychological Science* 12(2):157–162.

Association, American Psychological. 2005. *Concise rules of apa style*. American Psychological Association.

Bailey, Adam, Michael Lauck, Andrea Weiler, Samuel Sibley, Jorge Dinis, Zachary Bergman, Chase Nelson, Michael Correll, Michael Gleicher, David Hyeroba, Alex Tumukunde, Geoffrey Weny, Colin Chapman, Jens Kuhn, Austin Hughes, Thomas

Friedrich, Tony Goldberg, and David O'Connor. 2014a. High genetic diversity and adaptive potential of two simian hemorrhagic fever viruses in a wild primate population. *PLoS ONE* 9(3).

Bailey, Adam L., Michael Lauck, Samuel D. Sibley, Jerilyn Pecotte, Karen Rice, Geoffrey Weny, Alex Tumukunde, David Hyeroba, Justin Greene, Michael Correll, et al. 2014b. Two novel simian arteriviruses in captive and wild baboons (papio spp.). *Journal of Virology* 88(22):13231–13239.

Balas, B., L. Nakano, and R. Rosenholtz. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision* 9(12).

Ball, T., and S.G. Eick. 1996. Software visualization in the large. *Computer* 29(4): 33–43.

Behrens, John T., William A. Stock, and Catherine Sedgwick. 1990. Judgment errors in elementary box-plot displays. *Commun. Stat.-Simul. C* 19(1):245–262.

Belia, Sarah, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods* 10(4):389–96.

Belopolsky, Artem V, Laura Zwaan, Jan Theeuwes, and Arthur F Kramer. 2007. The size of an attentional window modulates attentional capture by color singletons. *Psychonomic bulletin & review* 14(5):934–8.

Best, L.A., L.D. Smith, and D.A. Stubbs. 2007. Perception of linear and nonlinear trends: Using slope and curvature information to make trend discriminations 1, 2. *Perceptual and motor skills* 104(3):707–721.

Bigelow, Alex, Miriah Meyer, and Nicola J. Camp. 2012. compreheNGSive: A tool for exploring next-gen sequencing variants. *Poster Proceedings fo the 2nd IEEE Symposium on Biological Data Visualization (BioVis 2012)*.

Borkin, Michelle, Krzysztof Z. Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank J. Rybicki, Charles L. Feldman, Hanspeter Pfister, et al. 2011. Evaluation of artery visualizations for heart disease diagnosis. *Visualization and Computer Graphics, IEEE Transactions on* 17(12):2479–2488.

Borland, David, and M Russell Taylor. 2007. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications* 27(2):14–7.

Boukhelifa, Nadia, Anastasia Bezerianos, Tobias Isenberg, and J Fekete. 2012. Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 18(12):2769–2778.

Buonomano, D.V., and M.M. Merzenich. 1998. Cortical plasticity: from synapses to maps. *Annual review of neuroscience* 21:149–86.

Callaghan, T C. 1984. Dimensional interaction of hue and brightness in preattentive field segregation. *Perception & psychophysics* 36(1):25–34.

Callaghan, Tara C. 1989. Interference and dominance in texture segregation: hue, geometric form, and line orientation. *Perception & psychophysics* 46(4):299–311.

Cao, Nan, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. 2010. FacetAtlas: multifaceted visualization for rich text corpora. *IEEE transactions on visualization and computer graphics* 16(6):1172–81.

Card, S. 2002. Information visualization. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, ed. A. Sears and J.A. Jacko, 542–542.

Carpendale, M Sheelagh T, and Catherine Montagnese. 2001. A framework for unifying presentation space. In *Proceedings of the 14th annual acm symposium on user interface software and technology*, 61–70. ACM.

Carver, Tim, Simon R. Harris, Matthew Berriman, Julian Parkhill, and Jacqueline A McQuillan. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 28(4):464–469.

Chong, Sang Chul, and Anne Treisman. 2005. Statistical processing: computing the average size in perceptual groups. *Vision research* 45(7):891–900.

Chong, S.C., and A. Treisman. 2003. Representation of statistical properties. *Vision research* 43(4):393–404.

Choo, Heeyoung, Brian R Levinthal, and Steven L Franconeri. 2012. Average orientation is more accessible through object boundaries than surface features. *Journal of Experimental Psychology: Human Perception and Performance* 38(3):585.

Clark, JH. 1924. The ishihara test for color blindness. *American Journal of Physiological Optics*.

Clement, Tanya, Catherine Plaisant, and Romain Vuillemot. 2009. The Story of One: Humanity scholarship with visualization and text analysis. *Relation* 10(1.43):8485.

Cleveland, William S., and Robert McGill. 1983. A color-caused optical illusion on a statistical graph. *The American Statistician* 37(2):101–105.

———. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association* 79(387):531–554.

Cohen, Jacob. 1994. The earth is round ($p < .05$). *The American Psychologist* 49(12): 997.

Collins, Christopher, Sheelagh Carpendale, and Gerald Penn. 2009a. Docuburst: visualizing document content using language structure. In *Computer graphics forum*, vol. 28, 1039–1046. John Wiley & Sons.

Collins, Christopher, Fernanda B. Viegas, and Martin Wattenberg. 2009b. Parallel Tag Clouds to explore and analyze faceted text corpora. *2009 IEEE Symposium on Visual Analytics Science and Technology* 91–98.

Collins, Jeff, and Dave Kaufer. 2001. Description of DocuScope (November).

Correll, M., D. Albers, S. Franconeri, and M. Gleicher. 2012a. Comparing averages in time series data. In *Proceedings of the 2012 acm annual conference on human factors in computing systems*, 1095–1104. ACM.

Correll, M., S. Ghosh, D. O'Connor, and M. Gleicher. 2011a. Visualizing virus population variability from next generation sequencing data. In *Biological data visualization (biovis), 2011 ieee symposium on*, 135–142. IEEE.

Correll, Michael, Danielle Albers, Steve Franconeri, and Michael Gleicher. 2012b. Comparing averages in time series data. In *Proceedings of the 2012 acm annual conference on human factors in computing systems*, 1095–1104. ACM.

Correll, Michael, Eric Alexander, and Michael Gleicher. 2013. Quantity estimation in visualizations of tagged text. In *Proceedings of the 2013 acm annual conference on human factors in computing systems*, 2697–2706. CHI '13, ACM.

Correll, Michael, Adam L. Bailey, Alper Sarikaya, David H. O'Connor, and Michael Gleicher. 2015. Layercake: a tool for the visual comparison of viral deep sequencing data. *Bioinformatics*.

Correll, Michael, Subhadip Ghosh, David O'Connor, and Michael Gleicher. 2011b. Visualizing virus population variability from next generation sequencing data. In *2011 ieee symposium on biological data visualization (biovis)*, 135–142. IEEE.

Correll, Michael, and Michael Gleicher. 2013. Error bars considered harmful. In *IEEE visualization poster proceedings*. IEEE.

———. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *Visualization and Computer Graphics, IEEE Transactions on* 20(12): 2142–2151.

Correll, Michael, Michael Witmore, and Michael Gleicher. 2011c. Exploring collections of tagged text for literary scholarship. In *Computer graphics forum*, vol. 30, 731–740. Wiley Online Library.

Cumming, Geoff, and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *The American Psychologist* 60(2):170–80.

Doherty, Michael E, Richard B Anderson, Andrea M Angott, and Dale S Klopfer. 2007. The perception of scatterplots. *Attention, Perception, & Psychophysics* 69(7): 1261–1272.

Dragga, Sam, and Dan Voss. 2001. Cruel pies: The inhumanity of technical illustrations. *Technical communication* 48(3):265–274.

D'Zmura, Michael. 1991. Color in visual search. *Vision research* 31(6):951–966.

Eaton, C., C. Plaisant, and T. Drizd. 2005. Visualizing missing data: graph interpretation user study. *Human-Computer Interaction-INTERACT 2005* 861–872.

Egeth, Howard E, Carly J Leonard, and Andrew B Leber. 2010. Why salience is not enough: reflections on top-down selection in vision. *Acta psychologica* 135(2):130–2; discussion 133–9.

Ellis, D. 2005. The English literature researcher in the age of the Internet. *Journal of Information Science* 31(1):29–36.

Elmqvist, N., and J.D. Fekete. 2010. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *Visualization and Computer Graphics, IEEE Transactions on* 16(3):439–454.

Enns, J T, and R A Rensink. 1990. Influence of scene-based properties on visual search. *Science* 247(4943):721–3.

Ferstay, Joel A, Cydney B Nielsen, and Tamara Munzner. 2013. Variant view: Visualizing sequence variants in their gene context. *IEEE Transactions on Visualization and Computer Graphics* 19(12):2546–2555.

Fouriezos, G., S. Rubenfeld, and G. Capstick. 2008a. Visual statistical decisions. *Attention, Perception, & Psychophysics* 70(3):456–464.

Fouriezos, George, Sara Rubenfeld, and Gary Capstick. 2008b. Visual statistical decisions. *Attention, Perception, & Psychophysics* 70(3):456–464.

Franconeri, SL, DK Bemis, and GA Alvarez. 2009. Number estimation relies on a set of segmented objects. *Cognition* 113(1):1–13.

Franconeri, Steven L, George A Alvarez, and Patrick Cavanagh. 2013. Flexible cognitive resources: competitive content maps for attention and memory. *Trends in cognitive sciences*.

Freeman, J., and E.P. Simoncelli. 2011. Metamers of the ventral stream. *Nature neuroscience* 14(9):1195–1201.

Fuchs, Johannes, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. 2013. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the sigchi conference on human factors in computing systems*, 3237–3246. ACM.

van der Geest, Thea, and Raymond van Dongelen. 2009. What is beautiful is useful-visual appeal and expected information quality. In *Ieee international professional communication conference*, 1–5. IEEE.

Gershon, Nahum. 1998. Visualization of an imperfect world. *IEEE Computer Graphics and Applications* 18(4):43–45.

Gleicher, Michael, Michael Correll, Christine Nothelfer, and Steve Franconeri. 2013. Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 19(12):2316 – 2325.

Goldstein, Norm. 1994. *The associated press stylebook and libel manual. fully revised and updated.* ERIC.

Greene, M.R., and A. Oliva. 2009. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive psychology* 58(2):137.

Haberman, Jason, and David Whitney. 2007. Rapid extraction of mean emotion and gender from sets of faces. *Current Biology* 17(17):R751–R753.

Hagh-Shenas, H., S. Kim, V. Interrante, and C. Healey. 2007. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *Visualization and Computer Graphics, IEEE Transactions on* 13(6):1270–1277.

Halberda, Justin, Sean F. Sires, and Lisa Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological science* 17(7):572–6.

Haroz, Steve, and David Whitney. 2012. How Capacity Limits of Attention Influence Information Visualization Effectiveness. *IEEE Transactions on Visualization and Computer Graphics* 18(12):2402–2410.

Harrison, Lane, Fumeng Yang, Steven Franconeri, and Ronald Chang. 2014. Ranking visualizations of correlation using weber's law. *Visualization and Computer Graphics, IEEE Transactions on* 20(12):1943–1952.

Harrower, M., and C. Brewer. 2003. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40(1):27–37.

Havre, S., B. Hetzler, and L. Nowell. 2000. ThemeRiver: visualizing theme changes over time. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* 115–123.

Healey, C.G., and J.T. Enns. 1998. Building perceptual textures to visualize multidimensional datasets. In *Visualization'98. proceedings*, 111–118. IEEE.

Healey, Christopher G., Kellogg S. Booth, and James T. Enns. 1996. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction* 3(2):107–135.

Heer, J., and M. Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on human factors in computing systems*, 203–212. ACM.

Hintze, J.L., and R.D. Nelson. 1998. Violin plots: a box plot-density trace synergism. *The American Statistician*.

Hope, Jonathan, and Michael Witmore. 2010. The Hundredth Psalm to the Tune of" Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. *Shakespeare Quarterly* 61(3):357–390.

Hou, Huabin, Fangqing Zhao, LingLin Zhou, Erle Zhu, Huajing Teng, Xiaokun Li, Qiyu Bao, Jinyu Wu, and Zhongsheng Sun. 2010. Magicviewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Research* 38(suppl 2):W732–W736.

Huang, Liqiang, and Harold Pashler. 2007. A boolean map theory of visual attention. *Psychological review* 114(3):599.

Huff, Darrell. 1993. *How to lie with statistics*. WW Norton & Company.

Hughes, H.C., G. Nozawa, and F. Kitterle. 1996. Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience* 8(3):197–230.

Hullman, Jessica, Eytan Adar, and Priti Shah. 2011. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics* 17(12):2213–2222.

Ibrekk, Harald, and M Granger Morgan. 1987. Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis* 7(4):519–529.

Inbar, Ohad. 2009. Graphical representation of statistical information in situations of judgment and decision-making. Phd. thesis, Ben-Gurion University of the Negev.

Izard, Véronique, and Stanislas Dehaene. 2008. Calibrating the mental number line. *Cognition* 106(3):1221–1247.

Jackson, Christopher H. 2008. Displaying uncertainty with shading. *The American Statistician* 62(4):340–347.

Javed, W., B. McDonnel, and N. Elmqvist. 2010. Graphical perception of multiple time series. *Visualization and Computer Graphics, IEEE Transactions on* 16(6):927–934.

Johnson, Douglas H. 1999. The insignificance of statistical significance testing. *The Journal of Wildlife Management* 763–772.

Kampstra, Peter. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* 28(1):1–9.

Kittur, A., E.H. Chi, and B. Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the twenty-sixth annual sigchi conference on human factors in computing systems*, 453–456. ACM.

Kosara, R., C.G. Healey, V. Interrante, D.H. Laidlaw, and C. Ware. 2003. Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications* 23(4):20–25.

Kosara, Robert. 2007. Visualization criticism-the missing link between information visualization and art. In *Information visualization, 2007. iv'07. 11th international conference*, 631–636. IEEE.

Kruskal, Joseph B, and James M Landwehr. 1983. Icicle plots: Better displays for hierarchical clustering. *The American Statistician* 37(2):162–168.

Kumar, N., and I. Benbasat. 2004. The effect of relationship encoding, task type, and complexity on information representation: An empirical evaluation of 2d and 3d line graphs. *MIS Quarterly* 28(2):255–281.

Lam, H., T. Munzner, and R. Kincaid. 2007. Overview use in multiple visual information resolution interfaces. *Visualization and Computer Graphics, IEEE Transactions on* 13(6):1278–1285.

Levinthal, Brian R, and Steven L Franconeri. 2011. Common-fate grouping as feature selection. *Psychological science* 22(9):1132–1137.

Li, Jing, Jean-Bernard Martens, and Jarke J. van Wijk. 2010a. A model of symbol size discrimination in scatterplots. In *Proceedings of the 2010 ACM annual conference on Human Factors in Computing Systems (CHI 2010)*, 2553–2562. ACM.

Li, Jing, Jarke J. van Wijk, and Jean-Bernard Martens. 2010b. A model of symbol lightness discrimination in sparse scatterplots. *2010 IEEE Pacific Visualization Symposium (PacificVis)* 105–112.

Lipkus, I M, and J G Hollands. 1999. The visual communication of risk. *Journal of the National Cancer Institute. Monographs* 27701(25):149–63.

MacEachren, Alan M, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. 2012. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics* 18(12):2496–2505.

MacEachren, AM. 1992. Visualizing uncertain information. *Cartographic Perspective* 13(3):10–19.

Maguire, Eamonn, Philippe Rocca-Serra, Susanna-Assunta Sansone, and Min Chen. 2014. Redesigning the sequence logo with glyph-based approaches to aid interpretation. *In Proceedings of EuroVis 2014 ,Short Paper*.

Marchant, Alexander P, Daniel J Simons, and Jan W de Fockert. 2013. Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta psychologica* 142(2):245–50.

Maule, John, and Anna Franklin. 2015. Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of vision* 15(4):6–6.

Melcher, David, and Eileen Kowler. 1999. Shapes, surfaces and saccades. *Vision Research* 39(17):2929–2946.

Meserth, T.A., and JG Hollands. 1999. Comparing 2d and 3d displays for trend estimation: The effects of display augmentation. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 43, 1308–1312. SAGE Publications.

Micallef, Luana, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18(12):2536–2545.

Michael, Elizabeth, Vincent de Gardelle, and Christopher Summerfield. 2014. Priming by the variability of visual information. *Proceedings of the National Academy of Sciences* 111(21):7873–7878.

Milne, Iain, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. 2010. Tablet–next generation sequence assembly visualization. *Bioinformatics* 26(3):401–2.

Monmonier, Mark. 1991. How to lie with maps.

Munzner, Tamara. 2014. *Visual analysis and design*, chap. 5, 94–114. A K Peters/CRC Press.

Munzner, Tamara, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. 2003. TreeJuxtaposer: scalable tree comparison using Focus+ Context with guaranteed visibility. *ACM Transactions on Graphics (TOG)* 22(3):462.

Newman, George E, and Brian J Scholl. 2012. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review* 19(4):601–607.

Nickerson, R S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* 5(2):241–301.

Nielsen, Cydney B, Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. 2010. Visualizing genomes: techniques and challenges. *Nature methods* 7:S5–S15.

Nourbakhsh, M.R., and K.J. Ottenbacher. 1994. The statistical analysis of single-subject data: a comparative examination. *Physical therapy* 74(8):768–776.

O'Connor, Shelby, Ericka Becker, Jason Weinfurter, Emily Chin, Melisa Budde, Emma Gostick, Michael Correll, Michael Gleicher, Austin Hughes, David Price, Thomas Friedrich, and David O'Connor. 2012. Conditional CD8+ t cell escape during acute simian immunodeficiency virus infection. *Journal of Virology* 86(1): 605–609.

Pabinger, Stephan, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, Johannes Zschocke, and Zlatko Trajanoski. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* 15(2):256–278.

Pandey, Anshul Vikram, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The persuasive power of data visualization. *Visualization and Computer Graphics, IEEE Transactions on* 20(12):2211–2220.

Pandey, Anshul Vikram, Katharina Rall, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of the 2015 ACM annual conference on Human Factors in Computing Systems (CHI 2015)*. ACM.

Paolacci, G., J. Chandler, and P. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5(5):411–419.

Parkes, L., J. Lund, A. Angelucci, J.A. Solomon, M. Morgan, et al. 2001. Compulsory averaging of crowded orientation signals in human vision. *Nature neuroscience* 4(7): 739–744.

Perlin, K. 1985. An image synthesizer. *Comput. Graph. (SIGGRAPH'85)* 19(3):287–296.

Potter, Kristin, Joe Kniss, Richard Riesenfeld, and Chris R Johnson. 2010. Visualizing summary statistics and uncertainty. *Computer Graphics Forum* 29(3):823–832.

Price, Paul C., Nicole M. Kimura, Andrew R. Smith, and Lindsay D. Marshall. 2014. Sample size bias in judgments of perceptual averages. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Ramanarayanan, G., K. Bala, and J.A. Ferwerda. 2008. Perception of complex aggregates. In *Acm transactions on graphics (tog)*, vol. 27, 60. ACM.

Ramsay, S. 2003. Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism. *Literary and Linguistic Computing* 18(2):167–174.

Ray, William C, R Wolfgang Rumpf, Brandon Sullivan, Nicholas Callahan, Thomas Magliery, Raghu Machiraju, Bang Wong, Martin Krzywinski, and Christopher W Bartlett. 2014. Understanding the sequence requirements of protein families: insights from the biovis 2013 contests. In *Bmc proceedings*, vol. 8, S1. BioMed Central Ltd.

Rensink, Ronald A, and Gideon Baldridge. 2010. The perception of correlation in scatterplots. In *Computer graphics forum*, vol. 29, 1203–1210. Wiley Online Library.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. Integrative genomics viewer. *Nature biotechnology* 29(1):24–26.

Rogowitz, Bernice E, Lloyd A Treinish, Steve Bryson, et al. 1996. How not to lie with visualization. *Computers in Physics* 10(3):268–273.

Rohrer, R.M., D.S. Ebert, and J.L. Sibert. 1998. The shape of Shakespeare: visualizing text using implicit surfaces. *Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)* 121–129,.

Rosenberg, Daniel. 2013. Data before the fact. In *"raw data" is an oxymoron*, ed. Lisa Gitelman, 15–40. MIT Press.

Rosenholtz, R., A. Dorai, and R. Freeman. 2011. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)* 8(2):12.

Rosenholtz, R., J. Huang, and K.A. Ehinger. 2012. Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Frontiers in psychology* 3.

Rosenholtz, R., Y. Li, J. Mansfield, and Z. Jin. 2005. Feature congestion: a measure of display clutter. In *Proceedings of the sigchi conference on human factors in computing systems*, 761–770. ACM.

Ross, J., L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems*, 2863–2872. ACM.

Roth, Robert E, and Alan M MacEachren. 2015. Geovisual analytics and the science of interaction: an empirical interaction study. *Cartography and Geographic Information Science* (ahead-of-print):1–24.

Sanyal, Jibonananda, Song Zhang, Gargi Bhattacharya, Phil Amburn, and Robert Moorhead. 2009. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE TVCG* 15(6):1209–1218.

Sarkar, Manojit, Scott S. Snibbe, Oren J. Tversky, and Steven P. Reiss. 1993. *Stretching the rubber sheet*. New York, New York, USA: ACM Press.

Schatz, Michael C, Adam M Phillippy, Ben Shneiderman, and Steven L Salzberg. 2007. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology* 8(3):R34.

Schmidt, Frank L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1(2): 115–129.

Schmidt, Frank L, and JE Hunter. 2013. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In *What if there were no significance tests?*, ed. Lisa L Harlow, Stanley A Mulaik, and James H Steiger, 37–64. Psychology Press.

Serences, John T, and Geoffrey M Boynton. 2007. Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron* 55(2):301–312.

Shah, Priti, Richard E Mayer, and Mary Hegarty. 1999. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology* 91(4):690.

Singh, M., and D.D. Hoffman. 1997. Constructing and representing visual objects. *Trends in Cognitive Sciences* 1(3):98–102.

Stasko, John, Carsten Görg, and Robert Spence. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization* 7(2): 118–132.

Stock, William A, and John T Behrens. 1991. Box, line, and midgap plots: Effects of display characteristics on the accuracy and bias of estimates of whisker length. *Journal of Educational and Behavioral Statistics* 16(1):1–20.

Stone, Maureen. 2012. In color perception, size matters. *IEEE CG & A* 32(2):8–13.

Swihart, Bruce J, Brian Caffo, Bryan D James, Matthew Strand, Brian S Schwartz, and Naresh M Punjabi. 2010. Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology* 21(5):621–5.

Sygnatur, E.F. 1998. Logging is perilous work. *Compensation and Working Conditions* 3(4):3–9.

Tal, Aner, and Brian Wansink. 2014. Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science* 0963662514549688.

Toet, Alexander, Jan van Erp, and Susanne Tak. 2014. The perception of visual uncertainty representation by non-experts. *IEEE Transactions on Visualization and Computer Graphics* 20(6):935–943.

Trafton, J Gregory, Sandra P Marshall, Farilee Mintz, and Susan B Trickett. 2002. Extracting explicit and implict information from complex visualizations. In *Proceedings of the second international conference on diagrammatic representation and inference*, 206–220. Springer-Verlag.

Treisman, A. 1985. Preattentive processing in vision. *Computer vision, graphics, and image processing* 31:156–177.

Treisman, A. 2006. How the deployment of attention determines what we see. *Visual cognition* 14(4-8):411–443.

Treisman, A, and S Gormican. 1988. Feature analysis in early vision: evidence from search asymmetries. *Psychological review* 95(1):15–48.

Treisman, A M, and G Gelade. 1980. A feature-integration theory of attention. *Cognitive psychology* 12(1):97–136.

Trumbo, B.E. 1981. A theory for coloring bivariate statistical maps. *The American Statistician* 35(4):220–226.

Tversky, A, and D Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185(4157):1124–31.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. Graphical inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics* 16(6):973–9.

Wilkinson, Leland. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 54(8):594.

Wise, James A., James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. 1995. Visualizing the Non-Visual: Spatial ananlysis and interaction with information from text documents. *IEEE Symposium on Information Visualzation* 51:51–58.

Witmore, Michael. 2010. The Funniest Thing Shakespeare Wrote? 767 Pieces of the Plays.

Wolfe, J.M., and S.C. Bennett. 1997. Preattentive object files: Shapeless bundles of basic features. *Vision research* 37(1):25–43.

Wood, Jo, Petra Isenberg, Tobias Isenberg, Jason Dykes, Nadia Boukhelifa, and Aidan Slingsby. 2012. Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 18(12):2749–2758.

Zacks, Jeff, and Barbara Tversky. 1999. Bars and lines: A study of graphic communication. *Memory & Cognition* 27(6):1073–1079.

Zhou, Xin, Brett Maricque, Mingchao Xie, Daofeng Li, Vasavi Sundaram, Eric A Martin, Brian C Koebbe, Cydney Nielsen, Martin Hirst, Peggy Farnham, et al. 2011. The human epigenome browser at washington university. *Nature methods* 8(12): 989–990.

Zhu, Jingchun, J Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M Kuhn, Donna Karolchik, John Archie, Marc E Lenburg, Laura J Esserman, et al. 2009. The ucsc cancer genomics browser. *Nature methods* 6(4):239–240.