Research Statement

Michael Correll

Designing effective visual interfaces to improve decision-making remains a critical challenge for both visualization and human-computer interaction research. Perceptual and cognitive biases can have harmful effects on decisions. Statistical methods can be useful, but can be difficult to explain to a general audience. In my work in **information visualization**, I focus on uniting statistical and visual approaches to interpreting data in service of better, data-driven decisions. I have designed novel techniques for the presentation of statistical concepts to non-statistical audiences, created new visual analysis systems for stakeholders in domains from Shakespeare scholarship to AIDS vaccine research, and run large-scale human subjects experiments to examine the perception of graphs. My research focuses on examining how people build up statistical information from visualizations, how new techniques can improve this understanding, and how these techniques can be instantiated into systems that address analytical needs in different domains.

Communicating Uncertainty

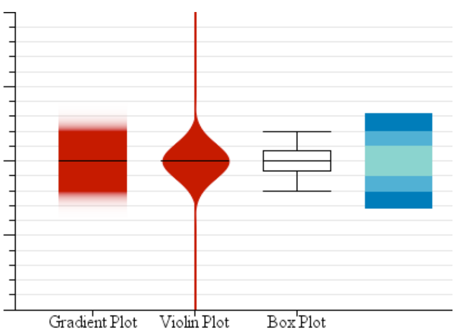

Fig. 1: Uncertainty is often communicated using bar charts with error bars. Yet, these charts can introduce bias in how uncertainty is interpreted. Gradient plots and violin plots correct for this bias [2].

Uncertainty is an inescapable part of data, but is often ignored when the data are presented, for reasons of perceived complexity for the audience, or decision biases that emerge under conditions of uncertainty. These biases can be perceptual (the way we see the data promotes poor judgments) or cognitive (the way we interpret the data promotes poor judgments). By identifying and correcting for these biases, we can improve decision-making. As an example, people tend to conflate the visual *area* of items (such as the *length* of all purple words in a word cloud) with the *numerosity* of these items (the *number* of purple words). By changing the spacing between characters, we can remove this confound [1].

A common choice for displaying uncertainty is to show a bar chart with error bars. In a crowd-sourced study, I discovered two biases with this encoding [2]. Firstly, the bar itself divides the chart into regions “inside” and “outside” the visual container of the bar. Viewers perceive outcomes that occur within the visual area of the central bar as more likely; this causes an asymmetric interpretation of the uncertainty of values. Secondly, although error bars are usually generated from procedures that rely on a *continuous* distribution, error bars create a *binary* impression that an outcome is either inside or outside the error bar (which may have very little to do with statistical significance or likelihood). I altered and tested two visualization techniques for showing mean and error, violin plots and gradient plots, which encode values in a continuous and visually symmetric way. I was able to confirm in experiments that these alternate encodings are still as easy to interpret as bar charts with error bars, but are free of the associated biases.

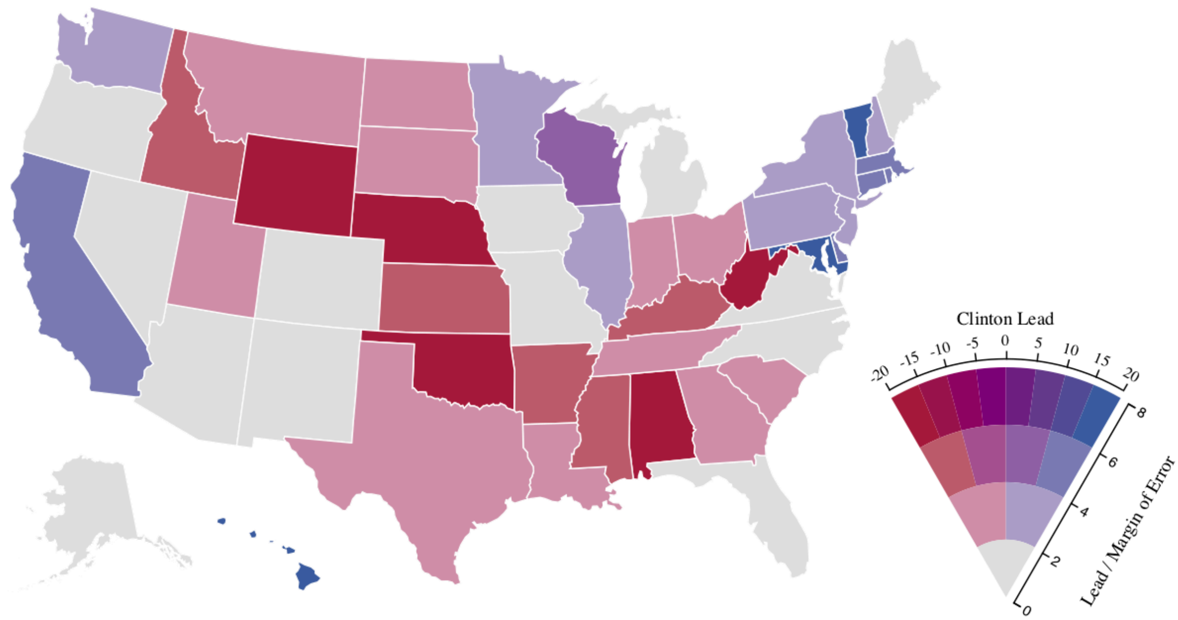

Fig. 2: Election polling prior to the 2016 U.S. Presidential Election encoded with a Value-Suppressing Uncertainty Palette. As the poll's margin of error becomes larger and larger with respect to the candidate's lead, the number of available colors shrinks. Once the uncertainty is high enough, the state is colored a single gray uncertain color. This dissuades people from drawing potentially unfounded conclusions based on ultimately highly variable predictions.

Another issue with uncertainty information is that it can be very tempting to ignore the uncertainty in our data when making decisions: for instance, we may interpret a candidate who is ahead in the polls as a sure bet in the actual election, regardless of the volatility in the polling data. My work on Value-Suppressing Uncertainty Palettes [3] is an attempt to make people more mindful of uncertainty, especially in decision-making. VSUPs are a form of bivariate map that reduce the number of categories as uncertainty increases, acting somewhat analogously to statistical effects tests in that they prevent people from speculating about minute differences when variability is too high for such differences to be reliable. In a controlled study, we found that VSUPs promote caution in judgments under uncertainty without negatively impacting decision quality compared to other types of bivariate maps.

In general, work in this area is focused on communicating statistical uncertainty (that can be esoteric or complex) to general audiences. This often involves new visual encodings. And since human decision-making under uncertainty is similarly complex, this also involves evaluating the performance of these new encodings for real tasks of inference, prediction, or estimation.

Future Work

People are increasingly exposed to decisions and recommendations made by algorithmic procedures like machine learning models. It can be tempting to view such black box systems as flawless or unbiased interpretations of the data. This elides the uncertainty in these models and their underlying data. It is an unmet challenge to communicate model outputs and their resulting uncertainty in a transparent way to the general audience. My work in exploring the visualization of distributions and outliers provides an initial set of techniques that I intend to expand to cover more complex cases like classification data and anomaly detection.

Extracting Statistical Information

Statistics is powerful, affording summarization and inference from massive amounts of data. The visual system is similarly powerful; it is capable of comparing, aggregating, and contrasting visual features quickly and reliably. Through careful design of the presentation of information, we can harness the strengths of visual perceptual system to allow people to act as *natural statisticians*, capable of making sound judgments about information in the aggregate.

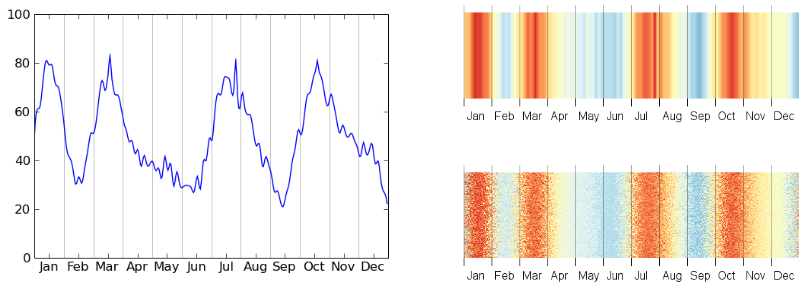

Fig. 1: Line graphs (left) are a standard design for time series data. Yet, heatmaps (right) better support estimates of values like mean or variance [7] [4]. Shuffling pixels in a process called *weaving* (bottom right) further improves these estimates.

Information on the perception of summary statistics like mean, variance, and trend is crucial to assessing our capabilities as natural statisticians. Through empirical methods developed in collaboration with perceptual psychologists, I have investigated the abilities of people, including those without statistical training, to estimate not just low-level visual features (such as the height of a particular bar in a bar chart), but also aggregate statistics (such as the average height of a group of bars). I have found that, in many cases, people are excellent natural statisticians, capable of accurately estimating averages in time series data [4], and comparing means and lines of best fit in scatterplots [5] [6]. However, in many cases, the best design for estimating aggregate statistics is not always the best for estimating point values, and vice versa [7]. “Color weaving” [4] is an example of techniques we have developed to support aggregate, rather than point, tasks. Color weaving relies on the local permutation of pixels in a heatmap to highlight group, rather than point, statistics. For instance, the noise of a local region of the heatmap corresponds to variance, and the perceived hue of a region corresponds to average value.

Future Work

A recent trend in my work has been evaluating deceptive or “Black Hat” visualizations: charts that appear to present the data truthfully but hide important features of the data [8]. I intend to continue work in this area, and identify places where human visual estimates are not sufficient to identify the statistical values of interest in the data (or, worse yet, promote the *misidentification* of such values). I also am continuing my work at attempting to create designs that “de-bias” viewers away from potentially misleading interpretations of the data.

Tools for Visual Data Analytics

As part of my interest in communicating statistical information to non-statistical audiences, I have developed a number of visualization tools for domain experts in differing fields.

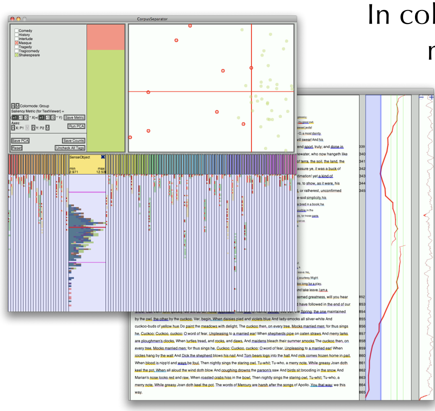

Fig. 4: CorpusSeparator and TextViewer allow literary scholars to find statistical patterns in a corpus, and see how it is instantiated in passages of text [12].

In collaboration with the Visualizing English Print (VEP) project, a multi-disciplinary effort to understand the history of early modern print culture, I developed a suite of visualization tools designed for scholars in the digital humanities. These tools assess similarity in word usage in order to detect affording stylistic differences in and between texts. Digital humanists (including the students of a graduate level English course) were able to encounter statistical patterns, and then translate them into passages of text that afforded close analysis and other traditional forms of literary analysis. The more recent SketchQuery [9] tool allows humanities scholars to analyze trends in word usage over time, without needing to learn time series analysis techniques, or special-purpose query languages.

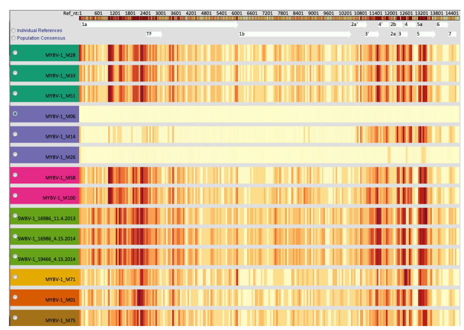

Fig. 6: LayerCake allows the exploration of patterns of mutation, finding interesting regions in dozens of viral genomes at once [10].

Working in concert with virologists interested in how to develop vaccines for viruses like HIV, I developed the LayerCake system for the analysis of variation viral genomes [10]. LayerCake allows users to simultaneously compare mutation across the entire genomes of dozens of sample populations. Dynamic thresholds of uncertainty, and varying measures of interest, allow analysts to filter out irrelevant and low-quality results to narrow their search down to only a few small areas of concern. Virologists have used LayerCake to analyze mutations in HIV, hemorrhagic fever, and avian flu, among other viruses.

A common trend in my systems is the translation of complex or obscure statistical patterns into a form that is understandable and usable by experts in different domains. This requires not just adaptation to particular datasets, but also to the *rhetorics* of discovery and proof in use in different fields. To support arguments in digital humanities, I translated from high dimensional data to passages of texts that can be subject to close reading. To support time series analysis, I created a system where anybody that can draw can immediately search for complex patterns. To support arguments in virology, I supported interactive filtering to locate important hotspots in genomes that are ordinarily too long to view in their entirety.

Future Work

I plan on collaborating with experts in different domains in order to assist them in analyzing and communicating their data. A benefit to my research perspective is that I focus on the presentation of statistical data to non-statisticians. As an increasing number of domains begin making use of quantitative data, this focus on translation to fit the rhetorical and analytical needs of an audience will be crucial for research. I have had previous success collaborating with researchers in many fields, in both the sciences and the humanities. I plan to continue direct interdisciplinary work as a central component of applying and evaluating my research contributions.

References

[1] Michael Correll, Eric Alexander, and Michael Gleicher, "Quantity Estimation in Visualizations of Tagged Text," in *Proceedings of the 2013 ACM Annual Conference on Human Factors in Computing*, 2013, pp. 2697-2706.

[2] Michael Correll and Michael Gleicher, "Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error," *IEEE Transactions On Visualization And Computer Graphics*, vol. 20, no. 12, pp. 2142-2151, 2014.

[3] Michael Correll, Dominik Moritz, and Jeffrey Heer, "Value-Suppressing Uncertainty Palettes," in *Proceedings of the 2018 ACM Annual Conference on Human Factors in Computing*, 2018, p. 642.

[4] Michael Correll, Danielle Albers, Steve Franconeri, and Michael Gleicher, "Comparing Averages in Time Series Data," in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing*, 2012, pp. 1095-1104.

[5] Michael Correll and Jeffrey Heer, "Regression By Eye: Estimating Trends in Bivariate Visualizations," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1387-1396.

[6] Michael Gleicher, Michael Correll, Christine Nothelfer, and Steve Franconeri, "Perception of Average Value in Multiclass Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 19, pp. 2316-2325, 2013.

[7] Danielle Albers, Michael Correll, and Michael Gleicher, "Task-Driven Evaluation of Aggregation in Time Series Visualization," in *Proceedings of the 2014 ACM Annual Conference on Human Factors in Computing*, 2014, pp. 551-560.

[8] Michael Correll, Mingwei Li, Gordon Kindlmann, and Carlos Scheidegger, "Looks Good To Me: Visualizations as Sanity Checks," *IEEE Transactions On Visualization and Computer Graphics*, Oct 2018, To appear.

[9] Michael Correll and Michael Gleicher, "The Semantics of Sketch: A Visual Query System for Time Series Data," in *Proceedings of the 2016 IEEE Conference on Visual Analytics Science and Technology*, 2016.

[10] Michael Correll, Adam J Bailey, Alper Sarikaya, David H O'Connor, and Michael Gleicher, "LayerCake: a Tool for the Visual Comparison of Viral Deep Sequencing Data," *Bioinformatics*, 2015.

[11] Michael Correll and Jeffrey Heer, "Surprise! Bayesian Weighting for De-Biasing Thematic Maps," *IEEE Transactions On Visualization And Computer Graphics*, 2016.

[12] Michael Correll, Michael Witmore, and Michael Gleicher, "Exploring Collections of Tagged Text for Literary Scholarship," *Computer Graphics Forum*, vol. 30, no. 3, pp. 731-740, 2011.

[13] Michael Correll and Michael Gleicher, "Implicit Uncertainty Visualization: Aligning Perception and Statistics," in *Proceedings of the 2015 Workshop on Visualization for Decision Making Under Uncertainty*, 2015.