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.

Techniques for De-Biasing Visualizations

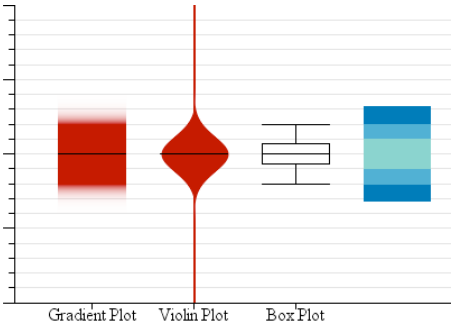

Fig. 1: Uncertainty is often communicated using bar charts with error bars. Yet, these charts can introduce bias in how uncertainty is interpreted. Gradient plots and violin plots correct for this bias [5].

Beyond the mere presentation of data, designers of visualizations have an obligation to ensure that data are responsibly and accurately used. Standing in the way of this goal are a number of potential biases in how humans understand data. 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.

Another common visualization is a thematic map, a category including heatmaps and choropleth maps, where a single variable of interest is encoded using color across a geographic region. These

maps can be overly simplistic in a way that makes signals of interest difficult to see, or creates false signals. For instance, confounding variables can hide patterns of interest, and small sample sizes can create false patterns. These biases can create maps that either hide "true" signals of interest within a sea of noise, or show "false" patterns that are based on sampling error and variability, rather than an actual spatial relationship. Employing Bayesian statistical methods, I created "Surprise Maps" as a technique for addressing these biases in thematic maps [3]. These maps rely on an initial set of models of expected density, each with an initial belief in the model's plausibility. These priors are updated with respect to the observed data, creating a set of posterior beliefs. "Surprising" regions of the map are those that cause large shifts in belief. The advantage of this approach is that, rather than focusing on regions with extreme values, these maps highlight *informative* regions of the map. Salient regions of the map are those with strong confirmatory or dis-confirmatory evidence for particular models. The weight of complex but irrelevant spatial patterns, or extreme but implausible values, is therefore reduced.

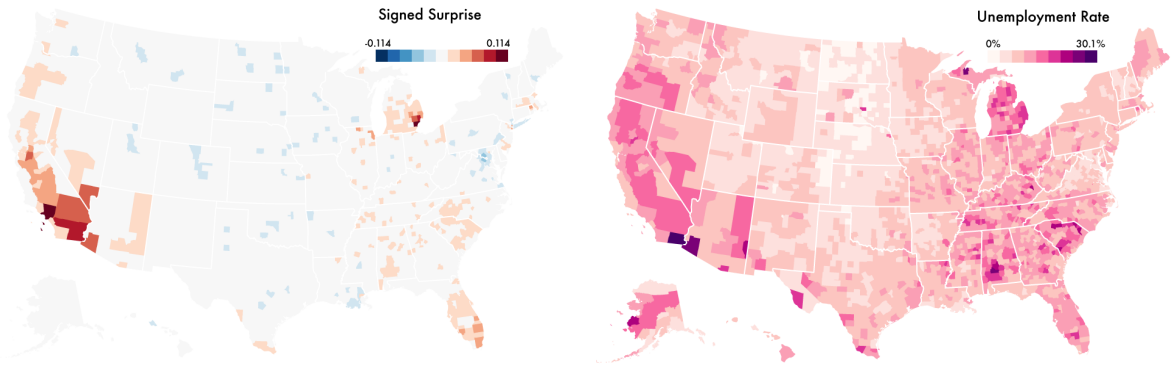

Fig. 2: Surprise maps (left) correct for confounding variables and false positives that may be difficult to see in standard choropleth maps (right), highlighting unexpected data [6].

In general, work in this area is focused on identifying where humans excel at being good intuitive statisticians, and where they need assistance. Identifying these *visual statistical affordances* is important for presenting information in a way that is usable. Identifying cognitive and perceptual *biases* in these affordances tells designers where they need to intervene to ensure that viewers are properly interpreting what they see.

Future Work

A specific research direction I am investigating is how the explicit annotation of uncertainty information (such as trend lines in scatterplots, error bars, and residual plots) can cause people to modify their assumptions about data. It is possible that human judgments about distributions and outliers might be more robust than many standard models for estimating uncertainty, in which case the explicit encoding of these simple models may cause more harm than good [4]. Of particular interest to me is how designers of visualizations can influence not just *judgments* (estimations and comparisons of particular values), but also *decisions*: can the proper design of visualizations influence people to make better decisions? Are there ways of conveying the uncertainty in data such that people refrain from making decisions based on weak evidence? In general, I believe that there are many aspects of the communication of uncertainty that would be well suited to my approach, which specifically investigates where statistical and visual judgments align or conflict.

Empirical Research on Visual Estimation of Statistical Values

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.

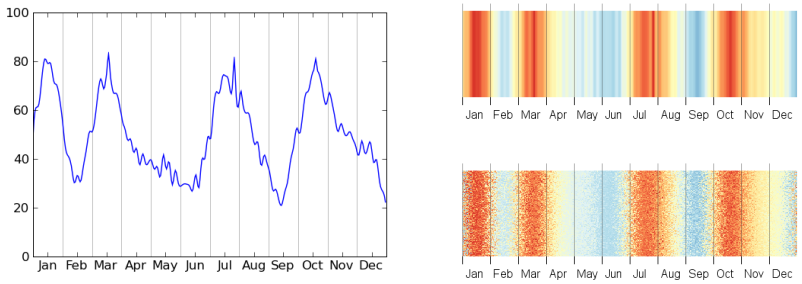

Fig. 1: Line graphs (left) are a standard design for time series data. Yet, heatmaps (right) better support estimates of values like mean or variance [4] [1]. Shuffling pixels in a process called *weaving* (bottom right) further improves these estimates.

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 the 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 [1], comparing means in scatterplots [2], and estimating proportions in word clouds [3]. However, in many cases, the best design for estimating aggregate statistics is not always the best for estimating point values, and vice versa [4]. "Color weaving" [1] 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 afford easy visual correlates between visual features and aggregate statistical properties: 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

I am currently investigating how well people can estimate trends and impute missing values from visualizations alone ("regression by eye"). I also intend to investigate how to communicate not just simple summary statistics, but more complicated statistical *models* (such as classifiers generated through machine learning, or projections from high-dimensional data). Research in this area could be beneficial not just to the general public (which often view such models as impenetrable black boxes), but also to researchers in machine learning and artificial intelligence, who must quickly evaluate multiple competing models.

Tools for Visual Data Analytics

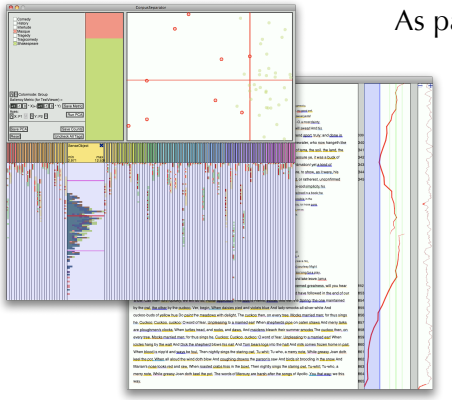

Fig. 4: CorpusSeparator and TextViewer allow literary scholars to find statistical patterns in a corpus, and see how it is instantiated in passages of text [10].

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.

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, affording the visualization of stylistic differences in texts over time, and across authors and genres. The companion tool TextViewer uses these spatial patterns to mark regions of interest within a single text. 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 [7] 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.

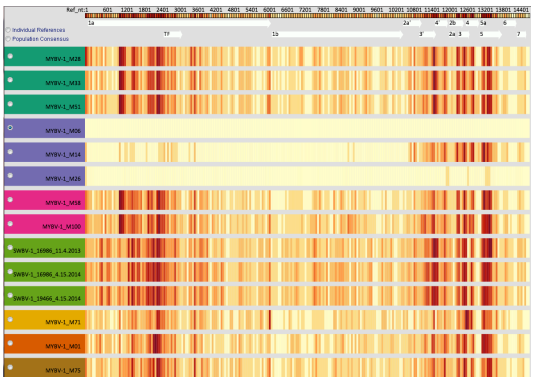

Fig. 6: LayerCake allows the exploration of patterns of mutation, finding interesting regions in dozens of viral genomes at once [9].

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 [8]. 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.

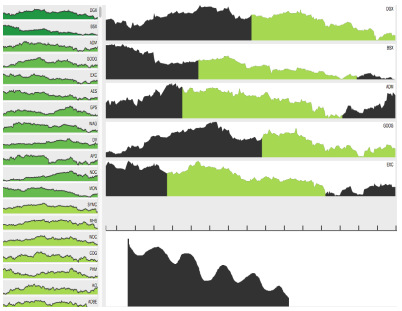

Fig. 5: SketchQuery allows users to draw patterns of interest, and look for similar patterns amongst thousands of time series [7].

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 of irrelevant or noisy data 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, 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.