STATS 337: Annotated Bibliography
Kenneth Tay

Theme: Communication and Visualization in Data Science

EXECUTIVE SUMMARY

Data analysis involves several tasks, including scoping the question, collecting data, cleaning and analyzing it, and presenting the results. Unfortunately, most statistics programs heavily emphasize just one of these tasks: the modeling and analysis of data, along with making correct predictions and inferences. Some programs which brand themselves as teaching data science go a bit further in teaching experimental design and data measurement. However, it seems that very few programs (if at all) teach communication and visualization of results in a formal manner. This annotated bibliography seeks to plug this gap by compiling a list of best practices in communicating and visualizing the results of a data analysis, and to understand what some of the tradeoffs are in doing so.

Based on my literature review, it seems that very little has been written on the communication of data analyses. Most of the articles gave a handful of tips and best practices, without deeper reasoning and analysis for why they work. As such, I expanded my bibliography to include data visualization, which is a major component of communicating data insights.

Before a data analyst begins preparing a data analysis presentation, the right question to ask is what it means for the data analysis to be successful in the first place. Once this goal has been established, the data analysis presentation should be structured in a way that directly meets this goal. Almost always, this means that the presentation must take context into account. Who is the audience? What do they know, and what do they not know? What are their current beliefs and leanings? What incentives drive them? What do they expect from your analysis? The data analyst should know the answers to these questions, and should present data insights with this context in mind.

There seems to be relatively strong agreement that data analysis results should be presented as a narrative, much like how one might tell a story. Telling a story with data gives the data analysis coherence, and allows the audience to see the value of the insights more easily. While this presentation framework might seem obvious, it is radically different from how scientists are used to present their work, which reflects the actual sequence of the data analysis (“First I did this, then I tried this, then I made it better by doing this…”)

On the flip side, there were some who argued that we should be suspicious of narratives and stories because they make us fall into common cognitive biases such as confirmation bias. Narratives can also lead to a “one-sided” view of the data, when the full picture is frequently more complex than that. However, the solution was not to get rid of stories altogether. Rather, data presenters and audiences of data presentations need to be aware of these biases, test the robustness of their stories and ensure the reliability of their data and methods.

On visualization, much analysis has been done on static figures, such as scatterplots, histograms and pie charts. Overall, the data science community has a good sense of which figures to use in which data setting. We also have a good idea of how to modify these figures (e.g. shape, color, size) to make the takeaway messages more salient. Apart from some of the guides in this annotated bibliography, there have been several comprehensive books written on the subject.

Having said that, there are some aspects of visualization that require more analysis and research. The first is visualization of high-dimensional data. Most humans can only reason spatially in two or three dimensions, but many datasets we have today have dimensions several orders of magnitude greater than that. There have been many different techniques suggested for visualizing this type of data (e.g. principal components, tSNE mapping), but I think that the good practices in this space are less well-formed. The other area requiring more thinking is the use of animation in data visualization. The results here seem to be mixed: while animation seems to make data visualizations more engaging and exciting, it is unclear if they help the audience understand and retain the data better.

TOP 3 PAPERS

1. Peng, R. (2018, Jun 4). Trustworth Data Analysis. [Web log post]. *Simply Statistics*. Retrieved Jun 4, 2018, from https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/.

Instead of focusing on the nuts and bolts of making a good presentation, this blog post helps the data analyst take a step back and ask the more fundamental question of why the audience should trust their analysis. By setting up a clear framework to think about this question, the data analyst can structure the presentation to directly meet the goal of the audience accepting the data analysis results.

2. Rogers, T., & Bloom., R. Which chart or graph is right for you? [Web white paper]. *Tableau*. Retrieved May 15, 2018, from https://www.tableau.com/learn/whitepapers/which-chart-or-graph-is-right-for-you.

This is an excellent reference for when one should use each chart type, with tips on how to make each visualization more effective.

3. Davenport, T. (2013, Jan 31). Telling a story with data. *Deloitte Insights*. Retrieved May 7, 2018, from https://www2.deloitte.com/insights/us/en/deloitte-review/issue-12/telling-a-story-with-data.html.

This was an engaging article that outlined the structure for a typical data analysis story, as well as elements which make the story a compelling one. Practical tips for data analysis presentations were given, along with some historical examples of good and bad data storytelling.

BIBLIOGRAPHY

Communicating Data Analyses

Bladt, J., & Filbin, B. (2013, Mar 27). A Data Scientist's Real Job: Storytelling. *Harvard Business Review*. Retrieved May 8, 2018, from https://hbr.org/2013/03/a-data-scientists-real-job-sto.

> This article argues that data science is about transforming data into directives, and that without an appropriate framing of the problem, data will only confuse. To this end, the authors suggest three things data scientists can do to communicate their findings to their audience. While I agree with the main idea of the article, the suggested pointers were only briefly elaborated on.

Burnard, P., Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Analysing and presenting qualitative data. *British Dental Journal*, *204*(8), 429.

> While most of the other papers in this annotated bibliography deal with presenting quantitative data, this paper presents one method for presenting qualitative data: thematic content analysis. The basic idea is to generate categories while going through the data, then combine findings which belong to the same theme. This is opposed to coming at the data with pre-defined categories.
>
> The process is outlined at a very high level, so I did not find the article useful as a whole. It also seems to be a laborious process, so automated techniques such as natural language processing might fare better. Also, despite the title, the section on presenting data turned out to be very short and quite uninformative.

Davenport, T. (2013, Jan 31). Telling a story with data. *Deloitte Insights*. Retrieved May 7, 2018, from https://www2.deloitte.com/insights/us/en/deloitte-review/issue-12/telling-a-story-with-data.html.

> This article outlines the structure for a typical data analysis story, as well as elements which make the story a compelling one. Data analysts should think about their work in a narrative that the audience understands, not in the order which they carried the analysis out. More than that, we are also told what should not be included in a data analysis presentation (e.g. overly technical details).
>
> The article also presents some historical examples of good and bad data storytelling. These are useful references to have in mind when trying to make a point on data science communication. An excellent article overall.

Grognor (2011, Dec 16). [Transcript] Tyler Cowen on Stories. *LESSWRONG*. Retrieved May 16, 2018, from https://www.lesswrong.com/posts/4kphivjxngJmEdWsN/transcript-tyler-cowen-on-stories.

> This is a transcript of Tyler Cowen's TEDxMidAtlantic talk "Be suspicious of stories," where he outlines some of the ways that stories can cause us to fall into common cognitive biases. Knowing these mechanisms is helpful when one is looking at the data, to remind oneself not to look only for insights that match

one’s pre-existing narrative. It is also helps one to be a discerning audience to a data presentation.

Hucki, Z. (2017, Feb 10). Beyond the data: five important lessons we can learn from Hans Rosling. *The* Conversation. Retrieved May 7, 2018, from https://theconversation.com/beyond-the-data-five-important-lessons-we-can-learn-from-hans-rosling-72810.

Hans Rosling was famous for visualizing data in a way that was compelling and engaging. This short article extracts five lessons on data communication that data scientists can take away from Rosling’s work.

Krzywinski, M., & Cairo, A. (2013). Points of View: Storytelling.

This article suggests that analysts should approach data visualization just as they would storytelling. Data graphics can be presented with an introduction, question, conflict, buildup and resolution. They provide a concrete example of how to do this. I think that the example is a useful reference to have for what makes a good data graphic.

Lamar, C. (2012, Jun 8). The 22 rules of storytelling, according to Pixar. *io9*. Retrived May 14, 2018, from https://io9.gizmodo.com/5916970/the-22-rules-of-storytelling-according-to-pixar.

This article lists 22 rules of storytelling which Pixar storyboard artist Emma Coats posted on Twitter. While not strictly about data science, I found some of these rules applicable to the presentation of data analyses. My favorite rules were 2, 5, 8 and 16. I was unsure about how applicable rules 7 and 13 were in the data science context.

Peng, R. (2018, Jun 4). Trustworth Data Analysis. [Web log post]. *Simply Statistics*. Retrieved Jun 4, 2018, from https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/.

The author points out that the success of a data analysis has a lot to do with whether the audience trusts the analysis, and elaborates on what is meant by trust and how presenters of data analyses can use this angle to refine their presentations.

In particular, the author breaks down an analysis into 3 parts: “presentation”, “done but not presented”, and “not done”. He then relates trust to these 3 components. At the end, he gives valuable advice to presenters on how to use this three-part framework to craft their presentations and, in some cases, even direct the analysis.

Peng, R. (2018, Apr 17). What is a Successful Data Analysis? [Web log post]. *Simply Statistics*. Retrieved May 8, 2018, from https://simplystatistics.org/2018/04/17/what-is-a-successful-data-analysis/.

The author attempts to define success for a data analysis as acceptance by the audience to which it is presented. Acceptance need not mean that the audience agrees with the conclusion of the analysis, but that the analyst performed the analysis correctly and made reasonable decisions at points where different methods could be applicable. Hence, there is a distinction drawn between a data analysis being successful and its claims being true, which is what we usually think of as successful in science.

With this definition, the success of the data analysis is intimately tied with the audience, which means that communication of results could determine the success or failure of an analysis. This is quite different from what we are taught in class, but I think that it makes a lot of sense in industry. It also means that whether a data analysis is successful or not can be a subjective measure.

Risdal, M. (2016, Jun 13). Communicating data science: An interview with a storytelling expert | Tyler Byers. *No Free Hunch*. Retrieved May 7, 2018, from http://blog.kaggle.com/2016/06/13/communicating-data-science-an-interview-with-a-storytelling-expert-tyler-byers/.

This is an interview with Tyler Byers, data scientist and software developer at Comverge, on communicating insights in data science. His main thesis is that data scientists have a lot of insight from data, but it is for naught if we cannot influence decision makers. To do that, data scientists need to learn how to craft their message in a way that connects with their clients. He also shares his experience of blogging regularly and its impact on his clients, as well as practical tips and tools for learning how to communicate and tell data stories. An entertaining and useful read which puts a human face on the data scientist.

Rossi, B. (2015, Jul 6). The pitfalls of data storytelling: why you cant always rely on a narrative in analytics. *Information age.* Retrived May 16, 2018, from http://www.information-age.com/pitfalls-data-storytelling-why-you-cant-always-rely-narrative-analytics-123459777/.

The author argues that each data visualization only presents “one side of the story.” He argues that instead of getting data visualizations to tell one coherent narrative, businesses should challenge their stories and question the reliability of their data and methods.

University of Leicester. Presenting numerical data. *University of Leicester*. Retrieved May 16, 2018, from https://www2.le.ac.uk/offices/ld/resources/numerical-data/numerical-data.

This article talks about how best to present numbers, whether within the body of text, as a table of numbers, or as a data visualization. The material on data visualizations is fairly standard; what I found useful was their advice on how and when to present numerical data in text and tables. A useful reference to have.

Visualization

Bertini, E., Tatu, A., & Keim, D. (2011). Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2203-2212.

> This paper sets up an overarching framework for systematizing the different techniques used in visualizing high-dimensional data. They focus on visualization techniques which depend on “quality metrics”, i.e. some measurement used as a proxy for how good a visualization is. After elucidating the factors used to describe the techniques, they present a table categorizing the techniques put forth in 20 papers. The basis for their method is the quality metrics pipeline, which is a useful way of thinking about how each technique moves from the source data to the visualization.

Heer, J., & Robertson, G. G. (2007). Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1240-1247.

> There is much debate over the usefulness of animation in data visualization. This paper attempts to quantify the extent of the usefulness of animation. To do so, the authors set up a taxonomy of transition types (7 types in total) as well as a series of metrics as proxies for usefulness. The main result of the study is that animation seemed to be useful, both objectively (through good performance on object tracking and change estimation measures) and subjectively (through subject responses). The authors also note that there is a tradeoff between the different design principles for animation. These results come from only 2 experiments, so any conclusions on animation should be extrapolated far beyond the settings they tested.

Henry, K. (2017, May 26). In Defense of Pie Charts, and Why You Shouldn’t Use Them. *Medium*. Retrieved May 15, 2018, from https://medium.com/@KristinHenry/in-defense-of-pie-charts-and-why-you-shouldnt-use-them-df2e8ccb5f76.

> In this visual piece, Henry describes certain situations where pie charts can be useful, but concludes that these situations are uncommon. I found the visual arguments very compelling. A good companion piece to Olson 2016.

In, J., & Lee., S. (2017). Statistical data presentation. *Korean Journal of Anesthesiology*, *70*(3), 267-276.

> This article is similar in content to the article by University of Leicester in that they outline how and when to use text, tables or graphics to present data. Their advice differs slightly from the other piece, so it is helpful to read different points of view when deciding on what presentation format is best.

Mayorga, A., & Gleicher, M. (2013). Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, *19*(9), 1526-1538.

> This paper introduces Splatterplots, a new way to present point data which addresses overplotting in scatterplots. For each subgroup, dense and sparse regions are treated differently, with dense regions being represented as smooth contours with solid fill, and sparse regions being represented as with a mix of

density data and filtered points. These computations are repeated as the user zooms in or out to maintain the amount of information displayed in the graph. Subgroups are differentiated using hue, so that chromaticity and lightness of color can be used to represent density information. As with any new way to encode information, the analyst will need some time to orient themselves to the meaning of the different encodings for dense and sparse regions.

Olson, R. (2016, Mar 24). In defense of the pie chart. [Web log post]. *O'Reilly*. Retrieved May 15, 2018, from https://www.oreilly.com/ideas/in-defense-of-the-pie-chart.

Pie charts are generally frowned upon in the visualization community. This article presents some situations where using a pie chart may actually be a good idea, as well as some tips on how to do a pie chart well. I came away convinced that pie charts can be useful in some cases, but these cases do not happen that often.

Rensink, R. A., & Baldridge G. (2010). The Perception of Correlation in Scatterplots. *Computer Graphics Forum*, *29*(3), 1203-1210.

The authors attempt to quantify the effectiveness of the scatterplot in conveying correlation information by using techniques from vision science. One result from their experiments was that severe underestimation of correlation occurs for $0.2 < r < 0.6$. The authors also have estimates for how much the correlation of the data has to shift away from some base correlation value before the reader can notice the difference. Overall, the findings suggest to me that humans are good at understanding high correlations but not low ones. This gives some guidance on what statistics we should present as data scientists.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of Animation in Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1325-1332.

This paper presents two alternatives to animation (trace visualization and small multiples visualization), and ran experiments to compare their effectiveness in terms of exciting the audience and conveying information. Overall, no one method outdid the others. The conclusion section is worth reading and has very practical implications for the data presenter.

Rogers, T., & Bloom., R. Which chart or graph is right for you? [Web white paper]. *Tableau*. Retrieved May 15, 2018, from https://www.tableau.com/learn/whitepapers/which-chart-or-graph-is-right-for-you.

An extremely handy reference for when one should use each chart type, and tips to make each visualization more effective. I definitely recommend that every data analyst have this on hand.

Shneiderman, B. (2003). The Eyes Have It: A Task by Data Type Taxonomy or Information Visualizations. *The Craft of Information Visualization*, 364-371.

This paper attempts to set up a framework for thinking about different kinds of visualizations. It posits that, at a high level, there are 7 different tasks which users wish to perform with data visualizations. It then goes through 7 common data types and the challenges typically faced when trying to visualize them.

I found it useful to have these 7 tasks in mind when designing data visualizations. It helps me to be clear on what I am trying to achieve, and that helps my design be geared to achieving that. However, beyond delineating these 7 tasks, I did not find this paper informative.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLoS Biology*, *13*(4), e1002128.

This paper argues that bar graphs and line graphs often present too simplistic a view of the data, and that researchers should move toward a more complete representation of the data. Visualizations which help this include scatterplots, histograms and boxplots.

I was surprised at how much more bar and line graphs were used than the others in practice, especially since the other types of visualizations are well-known and easy to draw. It suggests to me that there are other motivations besides ignorance that is causing researchers to stick with bar and line graphs.