Statistics 98T

Data Visualization

Location: Humanities A32
Time: 11:00 am - 12:15 pm Monday and Wednesday

Instructor: Amelia McNamara
Email: amelia.mcnamara@stat.ucla.edu
Website: www.stat.ucla.edu/∼amelia.mcnamara
Office: MS 8208
Office Hours: 1:00 - 2:00 pm Monday and Tuesday or by appointment

More and more, the products of statistical analysis that we encounter in our everyday lives come in the form of data visualizations. Visualizations have the advantage of being easier to interpret for many people, but they also give the impression of being a form of absolute truth. As with any presentation method, though, visualizations can be manipulated by their creators to show the story they are trying to tell. In this class, we will explore the many methods of information visualization, and develop intuition for when data graphics are not telling the truth. Throughout the class, you will be expected to respond to visualizations in the form of written response, but we will discuss the methods for analyzing graphics and the vocabulary for responding to them.

This course will draw upon methods from statistics, graphic design, geographic information systems, and computer science. However, we will approach these ideas from the ground up, and they will all be framed in the context of visualization. As a result, you can expect to come away from this course with a basic understanding of concepts from all of the aforementioned disciplines, as well as one place where they fit together.

Course Objectives

By the end of this course, you should be able to:

- Explain, in your own words, what a data visualization is representing.
- Provide an explanation of why the message portrayed in the visualization might be correct, but also:
- Identify possible sources of bias, problems with the analysis or assumptions behind the visualization.
- Identify and critique data visualizations "in the wild."

Class policies

Learning philosophy

This is not a typical lecture-based class. You must be an active participant in class discussion, which means coming to class prepared (having read the readings and completed the reading forum response). At the beginning of the quarter, we will agree on a set of classroom norms, to which we will all adhere. For example, I like the philosophy of "step up/step down." This norm means that if you are someone who tends to be very active in class discussions and always have something to contribute, "step down" a little to allow others the space to participate. Conversely, if you are someone who does not typically contribute to group discussions, try to "step up" a little to stretch yourself as a participant.

Academic Integrity

I do not tolerate academic dishonesty. When you provide written work to me, I expect it to be entirely your own original product, with the exception of quoted and properly cited sources. If I find work that I believe to be plagiarized, I will take the issue to the Dean of Students. As such, it's in your best interest to understand the definition and consequences of plagiarism. The Office of the Dean has a great handout on "before you begin that paper" which we will discuss at the start of class.

Communication

If you have questions for me, the best way to contact me is by emailing me at amelia.mcnamara@stat.ucla.edu or coming to my weekly office hours. If your question is something that might be of interest to other students, please feel free to post it on the CCLE discussion forum in the "questions" thread. I will monitor and respond to questions on the forum, but if you see a question that you know the answer to, please respond directly to your fellow students.

Late work

Late work will not be accepted unless you have discussed it with me beforehand, or you provide me with a doctor's note. If you know you will have trouble completing an assignment on the required schedule, please contact me as soon as possible.

Disability accommodations

I am committed to providing assistance to help you be successful in this course. Reasonable accommodations are available for students with disabilities. However, it is important that you register with the Office for Students with Disabilities as soon as possible, to arrange accommodations through their office. The OSD phone number is (310) 825-1501 (or (310) 206-6083 TTY) and they are located in the basement of Murphy Hall (A255 Murphy).

Required Materials

1. Course Reader (featuring excepts from the books listed in the Reading List, below).

Assignments

Reading Forum Responses (15% of your final grade)

Each week, we will be reading and discussing texts about visualization, as well as examining exemplary visualizations (both good and bad) chosen from a variety of sources. To kick off the in-class discussion, you will submit a response to the reading or visualizations before class in a forum on the class CCLE website. You are allowed to miss two reading forum responses without penalty. Guiding questions will be posted on the forum, but will include questions like "what assumptions were made in the creation of this graphic?" "are there any obvious biases visible?" "is there anything in the visualization or reading that you don't agree with?" (this must be supported by data) and "do you have any questions about this reading or graphic?"

Data journal entries (10% of your final grade)

In addition to the reading forum responses, you will also be responsible for keeping a data journal. We decided that this would take the form of a Facebook group for posting interesting visualizations. So, on the Facebook page, you should record interesting or notable data visualizations you encounter in your everyday life. When you find a visualization, you should post it and then write a few sentences about what drew you to this particular example and how it relates to what we've been discussing in class. I would like everyone to be posting at least once a week, but the main goal is to make you more aware of data products around you, and to spark class discussion.

Final Paper (35%)

The main product of this class will be a paper, which you will submit during finals week. This will be an exploration of a topic that has been getting statistical treatment in the media during the time of the class. For example, when I was creating this syllabus the Sochi Winter Olympics were motivating a lot of data visualizations, as was Obamacare enrollment. I will pick more recent and relevant examples during the class, and give you a list to choose from. You are also welcome to propose your own idea.

In a 15-page paper (double spaced) you will focus on two or three graphics on the same subject, each telling a different side or facet to the story. Using the readings we've discussed all quarter, you will provide a source-backed critique of each of the graphics. Topics to focus on include data processing, color theory, scaling, axes and biases implicit in any of the above.

Of course, this paper will depend on finding appropriate data visualizations from a variety of sources on the same topic, and your text will be analyzing and critiquing those graphics. However, the 15-page length is without graphics. Please include an appendix with reproductions of all the visualizations you comment on in your paper, labeled with figure numbers and appropriately cited with source information. In the text, you will refer to the figure number, but you will not have graphics in-line with the paper.

Paper Proposal (5%)

In the third week of class, you will turn in a 1-page topic proposal. The proposal consists of

- The subject matter you will be exploring
- At least one example of a graphic you will examine (preferably, two)
- One area you already find interesting, given the topics we've discussed in class so far

During the fourth week of the quarter, I will post a Doodle scheduler with my availability, and you will choose a time to come discuss your topic with me.

Some examples of topics:

- Police shootings as covered by Vox, http://www.vox.com/2014/12/17/7408455/police-shootings-map and 538, http://fivethirtyeight.com/features/how-many-americans-the-police-kill-each-year/ or this old article on the NY Times, http://www.nytimes.com/2008/05/08/nyregion/08nypd.html
- Similar, the LA times has a homicide watch: http://homicide.latimes.com/ which you could compare with something more entertainment-focused like this, http://la.curbed.com/archives/2013/10/mapping_21_of_los_angeless_most_notorious_murder_sites_1.php
- Compare the data and representations chosen by Giorgia Lupi and Stefanie Posavec in one week of their Dear Data project http://www.dear-data.com/
- As suggested in the reading forum, compare Facebook's map of NFL fandom, https://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859 with the NY Times fan maps like http://www.nytimes.com/interactive/2014/04/23/upshot/24-upshot-baseball.html

Of course, these are based on the things that have been catching my eye lately! I would recommend picking some graphics on a topic that you find interesting– if you can only find one, let me know, and I'll try to help you find another to do the compare/contrast.

Literature Review (20%)

In the sixth week, you will submit a literature review for your final paper. The literature review will be in the form of an annotated bibliography of sources that you expect to cite in your final paper. For each source, provide a short description of the information it contains and the major takeaways. Then, write a few sentences on how it ties into your final paper topic.

It's acceptable for your literature review to be mainly sources from the course reader, but you will need to include at least three additional sources. For the material we discussed in class, it is acceptable to refer to your class notes for the summary.

Your literature review should be about 5 pages long, and much of the material you produce will become part of your final paper.

Peer Review Activity (5%)

In the eighth week, we will do a peer review activity. Students will be put into pairs, and will exchange first drafts of their final papers. This first draft may not be the full 15 pages, but it should be at least 8 pages for this exercise. We will read and critique each others papers, and provide written feedback. This is similar to the peer review process that academic papers go through before they are published.

Final Presentation (10%)

During finals week, you will do a short (5 minute) presentation on your final paper topic. This is a great opportunity to point out visual components to your analysis that are hard to describe in your paper, as well as gather your classmates thoughts to help you finish your paper. I expect that most people will choose to

do a presentation with slides, but if you prefer to use a different format, just let me know ahead of time. To keep presentations moving along quickly (in order to accommodate everyone's presentations) I will ask you to send me your completed presentation the night before you present. I will also be keeping time.

Course Schedule

Week 1	**What is data visualization?** This week we'll be setting class norms and getting into the basics of data visualization. For Wednesday:	 No readings are due for the first day of class. The Functional Art, Alberto Cairo [4]. Introduction. **Reader**, pages 1-4 Visualize This, Nathan Yau [32]. Chapter 1, Telling Stories with Data. **Reader**, pages 5-14 How to Lie with Statistics, Darrell Huff [16]. Chapter 5, The Gee-Wizz Graph and Chapter 10, How to talk back to a statistic. **Reader**, pages 15-29
Week 2	**How to read data graphics** Monday: Wednesday:	 Numbers in the Newsroom, Sarah Cohen [6]. Chapter 3, Working with Graphics. **Reader**, pages 31-36 Data Points, Nathan Yau [33]. Chapter 3, Representing Data **Reader**, pages 79-90 Computational Information Design, Benjamin Fry [12]. Chapter 1, Introduction and Chapter 2, Basic Example. **Link on CCLE** The Elements of Graphing Data, William Cleveland [5]. Chapter 2, Principles of Graph Construction. **Reader**, pages 37-78
Week 3	**How to read data graphics, cont.** Monday:	 40 Years of Boxplots, Hadley Wickham [29]. **Link on CCLE** A Brief History of Mosaic Displays, Michael Friendly [11]. **Link on CCLE** Simpson's Paradox, Visualizing Urban Data ideaLab [28]. **Link on CCLE**

Week 3	**How to read data graphics, cont.**	
	Wednesday:	How to Read Histograms and Use them in R, Nathan Yau [34]. **Link on CCLE**
	Paper proposal due	Widening Inequality In Wages, NY Times. **Link on CCLE**
		Can every group be worse than average? Yes. NY Times. **Link on CCLE**
Week 4	**Maps: projections, totals versus rates**	
	Monday:	How to Lie with Maps, Mark Monmonier [18]. Chapters 2, 3, and 10. **Reader**, pages 93-118
	Meet with me to discuss final paper topics	Data Points, Nathan Yau [33]. Chapter 4, Exploring Data Visually **Reader**, pages 91-92.
	Wednesday	Avoiding Data Pitfalls, Part 2., Ben Jones [17] **Link on CCLE**
		Choropleth Maps, Penn State GEOG 486 [19] **Link on CCLE**
		When is a heat map not a heat map, Kenneth Field [10]. **Link on CCLE**
Week 5	**Perception: How color, form, and size can distort our understanding**	
	Monday:	The Elements of Graphing Data, William Cleveland [5]. Chapter 4, Graphical Perception. **Reader**, pages 139-152
	Wednesday:	The Grammar of Graphics, Leland Wilkinson [31]. Chapter 10, Aesthetics. **Reader**, pages 119-138
		Crowdsourcing Graphical Perception, Jeffery Heer and Mike Bostock [15]. **Link on CCLE**
		Ranking Visualization of Correlation Using Weber's Law, Lane Harrison, Fumeng Yang, Steven Franconeri, Remco Chang [14]. **Link on CCLE**

Week 6	**Big things: multidimensional data versus "big data"**	
	Monday:	Critical Questions for Big Data, danah boyd and Kate Crawford [3]. **Link on CCLE**
		Graphical Inference for Infovis, Hadley Wickham, et. al [30]. **Link on CCLE**
	Wednesday:	Grand Tour, Di Cook et al [7]. **Link on CCLE OR** Prim-9, John Tukey [26] **Link on CCLE**
	Literature review due	Computational Information Design, Benjamin Fry [12]. Chapter 4, Advanced Example. **Link on CCLE**
		DIY Bertin Matrix, Charles Perin et al [20] **Link on CCLE**
Week 7	**Small things: small multiples, sparklines, mini-graphs**	
	Monday:	Data Points, Nathan Yau, [33]. Chapter 4, Exploring Data Visually. **Reader**, pages 161-168
		The Generalized Pairs Plot, John Emerson, et al [8]. **Link on CCLE**
		Wee Things, Lena Groeger [13] **Link on CCLE**
	Wednesday:	Sparkline Theory and Practice, Edward Tufte [25]. **Link on CCLE**
		Envisioning Information, Edward Tufte [23]. Chapter 4, Small Multiples. **Reader**, pages 169-175

Week 8	Monday: **Chark junk: when a visualization doesn't mean anything**	WTF Visualizations [1]. **Link on CCLE** The Visual Display of Quantitative Information, Edward Tufte [24]. Chapter 5: Chartjunk: Vibrations, Grids, and Ducks. **Reader**, pages 153-160 Spurious Correlations, Tyler Vigen [27]. **Link on CCLE**
	Wednesday: **Interactive Visualizations** Peer review of first drafts	Coordinated Highlighting In Context, Stephen Few [9]. **Link on CCLE** Scatterplot Matrix Brushing, Mike Bostock [2]. **Link on CCLE** Movie Explorer, RStudio [21]. **Link on CCLE** The Tracks of Arrears, The Economist Data Team [22]. **Link on CCLE**
Week 8	**Getting technical: How do people actually visualize data?** Peer review discussion	Guest Speakers: Terri Johnson and Josh Gordon
Week 10	Final oral presentations	
Finals Week	Final paper due	

Reading List

[1] Wtf visualizations. http://wtfviz.net/, 2014.

[2] Mike Bostock. Scatterplot matrix brushing. http://bl.ocks.org/mbostock/4063663, November 2012.

[3] danah boyd and Kate Crawford. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, 2012.

[4] Alberto Cairo. *The Functional Art: An introduction to information graphics and visualization.* New Riders, 2013.

[5] William S Cleveland. *The elements of graphing data.* Wadsworth Advanced Books and Software, 1985.

[6] Sarah Cohen. *Numbers in the Newsroom: Using math and statistics in the news.* Investigative Reporters and Editors, Inc, 2001.

[7] Dianne Cook, Andreas Buja, Javier Cabrera, and Catherine Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 4(3):155–172, 1995.

[8] John W Emerson, Walton A Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, and Hadley Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 2013.

[9] Stephen Few. Coordinated highlighting in context: Bringing multidimensional connections to light. Technical report, perceptual edge, 2010.

[10] Kenneth Field. When is a heat map not a heat map. http://cartonerd.blogspot.com/2015/02/when-is-heat-map-not-heat-map.html, February 2015.

[11] Michael Friendly. A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 2001.

[12] Benjamin Fry. *Computational Information Design.* PhD thesis, Massachusetts Institute of Technology, April 2004.

[13] Lena Groeger. Wee things. http://lenagroeger.s3.amazonaws.com/talks/nicar/weethings.html, 2014.

[14] Lane Harrison, Fuming Yang, Steven Franconeri, and Remco Chang. Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943 – 1952, 2014.

[15] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *ACM Human Factors in Computing Systems (CHI)*, pages 2013–212, 2010.

[16] Darrell Huff. *How to lie with statistics.* W W Norton & Company, 1954.

[17] Ben Jones. Avoiding data pitfalls, part 2: Fooled by small samples. http://dataremixed.com/2015/01/avoiding-data-pitfalls-part-2/, January 2015.

[18] Mark Monmonier. *How to Lie with Maps*. University of Chicago Press, 1996.

[19] Penn State GEOG 486. Choropleth maps. https://www.e-education.psu.edu/geog486/node/1864, 2014.

[20] Charles Perin, Mathieu Le Goc, Romain Di Vozzo, Jean-Daniel Fekete, and Pierre Dragicevic. DIY Bertin matrix. *CHI'15*, 2015.

[21] RStudio Team. Movie explorer. http://shiny.rstudio.com/gallery/movie-explorer.html, 2014.

[22] The Data Team. The tracks of arrears. *The Economist*, May 2015.

[23] Edward Tufte. *Envisioning Information*. Graphics Pr, 1990.

[24] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Pr, 2001.

[25] Edward Tufte. Sparkline theory and practice. http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR, 2014.

[26] John Tukey. Prim-9. http://www.youtube.com/watch?v=B7XoW2qiFUA.

[27] Tyler Vigen. Spurious correlations. http://www.tylervigen.com/spurious-correlations, 2013.

[28] Visualizing Urban Data ideaLab, Lewis Lehe, and Victor Powell. Simpson's paradox: Girls gone average. averages gone wild. http://vudlab.com/simpsons/.

[29] Hadley Wickham and Lisa Stryjewski. 40 years of boxplots. http://vita.had.co.nz/papers/boxplots.html, 2011.

[30] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 2010.

[31] Leland Wilkinson. *The Grammar of Graphics*. Statistics and computing. Springer Science + Business Media, 2005.

[32] Nathan Yau. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Wiley, 2011.

[33] Nathan Yau. *Data Points: Visualization that Means Something*. Wiley, 2013.

[34] Nathan Yau. How to read histograms and use them in R. http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/, 2014.