CSSS 569 · Visualizing Data and Models

Winter Quarter 2021

University of Washington

Christopher Adolph · Associate Professor · Political Science and CSSS

Class Meets
TTh 4:30–5:50 PM
Taught via Zoom

Section Meets
F 3:30–5:20 PM
Taught via Zoom

Office
All meetings by Zoom
cadolph@uw.edu

Teaching Assistants
Brian Leung
kpleung@uw.edu
Kenya Amano
kamano@uw.edu

Overview. Visual displays are an integral part of most social science presentations and can make or break a paper. Good visuals help researchers uncover patterns and relationships they would otherwise miss. Ever more sophisticated statistical models cry out for clear, easy-to-understand visual representations of model findings. Yet social scientists seldom put as much care into designing visual displays as they devote to crafting effective prose. This course takes the design of graphics and tables seriously and explores a variety of visual techniques for investigating patterns in data, summarizing statistical results, and efficiently representing the robustness of such results to alternative modeling assumptions. Emphasis is placed on the principles of effective visualization, examples from the social sciences, novel visual displays, and the implementation of recommended techniques using the R statistical environment and the R packages `tile`, `simcf`, and `ggplot2`.

Prerequisites. No specific courses are required, but some graduate level quantitative methods coursework is prerequisite, as many of the applications we consider will assume familiarity with the basics of research design and quantitative inference (linear regression & elementary maximum likelihood).

Office Hours. Chris Adolph: Thursdays, 3:00 – 4:15 PM and by appointment via Zoom. Brian Leung: Tuesdays, 3:00 – 4:00 pm PM and by appointment via Zoom. Kenya Amano: Thursdays, 1:00 – 2:00 pm PM and by appointment via Zoom.

Course Website. Consult `http://faculty.washington.edu/cadolph/vis` for problem sets, notes, and announcements.

Notice Required by State Law. *Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW's policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy* (`https://registrar.washington.edu/staffandfaculty/religious-accommodations-policy`). *Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request form* (`https://registrar.washington.edu/students/religious-accommodations-request`).

Other relevant university policies. See this website: `https://registrar.washington.edu/staffandfaculty/syllabi-guidelines`

Course Requirements

Homework (30%) I will assign three homeworks covering topics to include exploring datasets, visualizing the results of statistical inference, and designing and programming new visualizations. For some assignments, it will be possible to use a variety of graphics packages to complete the assignment, but for most problems, R will be required or strongly recommended. Help will be available for R and any other package specifically recommended for the assignment, but not for other packages.

Breakout Groups (30%) Starting *next week*, students will self-select into a small discussion group investigating the application of visual displays to a specific scientific problem or area. This problem might consist of a difficult kind of model or dataset to visualize. Alternatively, it might be a problematic or promising visual display method used fre-

quently in the student's field which the student hopes to replace, improve, or perfect. In past years, students investigated interactive graphics, animations, and visualizations for text data, network data, hierarchical and multilevel data, spatial data, and time series, respectively, among other topics. Students may choose among these topics or propose their own. I reserve the right to decide which groups are large enough to be viable and to combine groups if needed.

Before our joint Zoom meeting, each member of the breakout group will write and circulate to the group and to me a 2–5 page memo, complete with (original or borrowed) graphics, illustrating a relevant data visualization problem they wish to tackle and briefly sketching possible strategies for solving it. This memo need not *solve* the data visualization problem and may not necessarily even present an actual data analysis; the goal is to start a conversation about how we might approach a student-selected visualization challenge. Each group will meet at least once for discussion of their problem area and memos; this meeting will occur no earlier than the start of Week 3 (Tuesday, 19 January) and no later than the end of Week 6 (Friday, 12 February).

By 9 AM Monday, 22 February, each group will email to the class a 5–8+ page essay sharing lessons learned, recommendations for best practices, and outstanding problems in the area studied by the group. During the week of 22 February, I will facilitate a (written) online discussion in which members in the class may ask any other group questions about their topic and conclusions. Each member of the class should ask (at least) one original question of another group, and each member should help answer at least one question directed at their own group.

Credit for this portion of the course will be based on the individual memo, participation in breakout discussions, the final essay, and participation in the online class discussion.

Final presentation (40%) Over the final two weeks of the course, each student will present a poster[1] applying the tools learned in class to their own research. Alternatively, students can take a published article in their field and show how better visuals would either more clearly convey the findings or cast doubt on them, or present an innovation in statistical graphics, preferably one which comes with software to help

1 Posters are used as an alternative to slide presentations in many fields. Guidance on poster construction will be provided later in the quarter for students who have never made a scientific poster. Students presenting interactive graphics as part of their final presentation should bring a laptop displaying the interactive graphic, perhaps with a supporting poster explaining the project if needed.

implement the innovation. The final presentation may address problems related to the topics pursued in the breakout group, but should represent primarily the work of the presenting student, not the group: this is a separate assignment, and it is usually more fruitful to tackle a second problem for the final presentation. Likewise, it's useful for the final poster to be substantially different from the homeworks, though it may represent an evolution of a project explored in the homework assignments. Final presentations must be emailed to your instructor in PDF format for credit to be given.

Group projects permitted, but each member must have primary responsibility for at least one figure, and this should be indicated in the email sending the poster to me (but not in the poster itself).

NB: We will use Google Sheets to coordinate formation of breakout topics and groups, scheduling of breakout meetings, and scheduling of final posters. Google Sheets requiring your attention will be announced on the course mailing list. ***Prompt attention to Google Sheets requests is essential to keeping the course on schedule.***

Course texts

Visual display books are expensive; students should order based on their interests. Descriptions at right may help select the most useful texts for permanent purchase. The starred texts are the most essential for purchase.

*Kieran Healy. 2018. *Data Visualization*. Princeton University Press. (Amazon: $40.00)	*Up-to-date guide to data visualization implementing many of this courses' recommendations in R's* `ggplot2`.
*Edward R. Tufte. 2001. *The Visual Display of Quantitative Information*. Graphics Press. 2nd ed. (Amazon: $29.49)	*The most famous and possibly the best book on data visualization ever written. Fun to read and essential.*
William S. Cleveland. 1993. *Visualizing Data*. Hobart Press. (Amazon: $49.00)	*Classic on the design of data visuals from a statistical perspective, especially for exploratory data analysis with many conditioning variables.*

Paul Murrell. 2018. *R Graphics*. Chapman & Hall. 3rd ed. (Amazon: $86.48)	*The authority on R's various graphics engines; excellent technical reference for both beginners and programmers.*
Colin Ware. 2012. *Information Visualization*. Morgan Kaufman. 3rd ed. (Amazon: $55.21). *NB: 4th edition published March 2020.*	*Collects a wealth of cognitive science research on how people see and process data visuals. Helpful background; less emphasis on application.*
Nathan Yau. 2011. *Visualize This*. Indianapolis: Wiley. (Amazon: $35.99)	*Gentle introduction to use of R and other packages to perform exploratory data analysis and make beautiful visual displays.*

Recommended for further reading

Chris Beeley. 2013. *Web Application Development with R Using Shiny*. Packt Publishing.

Jacques Bertin. 1967. [2010.] *Semiologie graphique*. [*Semiology of Graphics*.] trans. William J. Berg. ESRI Press.

R. Dennis Cook. 1998. *Regression Graphics*. Wiley Interscience.

Dianne Cook & Deborah F. Swayne. 2007. *Interactive and Dynamic Graphics for Data Analysis*. Springer-Verlag.

Michael Friendly. 2000. *Visualizing Categorical Data*. SAS Publishing.

Ben Fry. 2007. *Visualizing Data*. O'Reilly.

Kosuke Imai. 2017. *Quantitative Social Science: An Introduction*. Princeton Univ. Press.

Julie Steele and Noah Iliinsky, eds. 2010. *Beautiful Visualization*. O'Reilly Media, Inc.

David McCandless. 2009. *The Visual Miscellaneum*. Harper Design.

Isabel Meirelles. 2013. *Design for Information*. Rockport Publishers.

Oscar Perpinan Lamigueiro. 2014. *Displaying Time Series, Spatial, and Space-Time Data with R*. Chapman & Hall/CRC.

Deepayan Sarkar. 2008. *Lattice: Multvariate Data Visualization with R*. Springer-Verlag.

Edward Tufte. 1990. *Envisioning Information*. Graphics Press.

Edward Tufte. 1997. *Visual Explanations*. Graphics Press.

Edward Tufte. 2006. *Beautiful Evidence*. Graphics Press.

Howard Wainer. 2005. *Graphic Discovery*. Princeton University Press.

Hadley Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
Leland Wilkinson. 1999. *The Grammar of Graphics*. Springer-Verlag.
Graham Wills. 2012. *Visualizing Time: Designing Graphical Representations for Statistical Data*. Springer.
Yihui Xie. 2013. *Dynamic Documents with R and knitr.* Chapman & Hall/CRC.

Tools

It's easier than ever to create beautiful and effective scientific graphics, but not all graphical software is created equal. Many commonly used packages – particularly Microsoft Excel and its clones – combine inflexibility with poor default settings.

For the most part, students are not required to use a specific package, but are encouraged to use software that allows: (1) flexible generation of virtually any diagram, (2) command line or code interface, perhaps in addition to a graphical interface, and (3) widely usable output, such as postscript or PDF.

Recommended Software for Visual Display

R. In-class code examples will use the R statistical language, which has all these virtues in addition to being free, open source, and widely used. You can obtain R at `http://www.r-project.org`. Throughout the course, I will provide example code in R and can only promise detailed homework help for the R package. At least one homework will require students to use R, so it's worth downloading now. In particular, the course will provide readings and examples drawing on the popular `ggplot2` graphics package and the instructor's own `tile` graphics package, both available for R.

Illustrator. Adobe Illustrator is the industry standard for retouching postscript and PDF graphics. Unfortunately, it is also (a) very expensive, even with an academic license and (b) now only available as part of a subscription to a package of Adobe software (see the Tech Center page at the University Bookstore's website for details). Illustrator is *not required for the course* but is worth considering as students develop their visualization skills, especially for touching up final illustrations.

Other free tools. Yau's *Visualize This* discusses other tools for getting data off the web (like the Python programming language), constructing interactive graphics (like the processing language), and for working with maps (using SVG). Although we will not

cover these tools in class, they may be of use for student projects. A wealth of tools have emerged to work in conjunction with R to create interactive graphics, animations, and slides for the web (especially `Shiny`, but also `rCharts`, `Slidify`, `gridSVG`, and others).

Course outline

The readings for this course are complementary to the lectures and often cover topics or directions we don't have time to get to in lecture. It is thus more important than usual for a statistics class that students should come to class having read the material assigned for that day. The reading load for this class is considerably longer (in pages, if not minutes) than the typical statistics class but is fun, quick, and essential: the best way to learn effective visualization is to see how other scholars do it. Some of the readings, particularly from the *Journal of Computational and Graphical Statistics* (JCGS), have technical portions, but in most cases these details can be skimmed unless/until you want to code up these methods for yourself. Readings marked *Optional* are intended to be read *now* if you are interested in or working on the graphical problem described therein.

Note that if you are *not* familiar with R, you should begin reading the "optional" selections from Zuur immediately.

On some days, we will open class with a "Gallery" in which I will present for discussion several innovative or problematic visualizations (see the course site for a list). This will give everyone a chance to see the principles of the course in action, and learn from both the successes and mistakes of other scientists, including your instructor.

Part I: Theory of Visualization

Tuesday, 5 January 2019 · Introduction

Optional: Tufte, *Visual And Statistical Thinking*, pp. 5–15

Thursday, 7 January – Tuesday, 19 January · Principles of Information Visualization

Required: Tufte, *VDQI*, all

Richard A. Feinberg and Howard Wainer. 2011. "Extracting sunbeams from cucumbers." *JCGS* 20:4.

Optional: Alain F. Zuur, Elena N. Ieno, and Erik H.W.G. Meesters. 2009. *A beginner's guide to R*. Springer. Ch. 1–4 (for new R users)

Thursday, 21 January – Tuesday, 26 January · Cognitive Issues in Visualization

Required: Healy, Chapter 1

Jeffrey Heer and Michael Bostock. 2010. "Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design." *ACM Human Factors in Computing Systems (CHI)*. 203–212.

Ware, Ch. 1, 4, 5

Optional: Alain F. Zuur, Elena N. Ieno, and Erik H.W.G. Meesters. 2009. *A beginner's guide to R.* Springer. Ch. 5–6 (for new R users)

Ware, Ch. 6

Rick Wicklin. 2011. "Visualizing airline delays and cancelations." *JCGS* 20.2 (heatmap example)

Yau, Ch. 3–4

PROBLEM SET 1 DUE THURSDAY, 21 JANUARY BY EMAIL

Thursday, 28 January – Tuesday, 2 February · Programming Visual Displays

Required: Healy Ch. 2–5, 8

Suggested: Murrell, Ch. 1–3, 6–7, 9–10

Optional: Murrell, Ch. 4–5, 8, 11–17 (on `lattice`, `ggplot2`, advanced `grid`, categorical data, maps, networks, 3D, dynamic and interactive graphics)

Hadley Wickham. 2010. "A Layered Grammar of Graphics." *JCGS* 19:1. (on `ggplot2`)

Yau, Ch. 5–6

Part II: Visualization for Statistical Applications

Thursday, 4 February – Tuesday, 9 February · Exploratory Data Analysis

Required: Cleveland, *Visualizing Data*, selections.

W. N. Venables and B. D. Ripley. 2010. *Modern applied statistics with S*. 4th ed. Springer. Ch. 5 & 11.

Ben Fry. 2007. *Visualizing Data*. O'Reilly. Ch. 1, 2, 4.

William G. Jacoby. 1998. "Statistical Graphics for Visualizing Multivariate Data." *Sage Papers on Quantitative Applications in the Social Sciences*, selections.

Catherine B. Hurley. 2004. "Clustering visualizations of multidimensional data." *JCGS* 13:4.

Rida E. Moustafa, Ali S. Hadi, and Jürgen Syzmanik. 2011. "Multi-class data exploration using space transformed visualization plots." *JCGS* 20:2. (read for essential points and graphics)

Healy, Ch. 7 (on maps)

Optional: Yau, Ch. 7–8 (on maps)

Danny Holten. 2006. "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data." *IEEE Transactions on Visualization and Computer Graphics*. 12:5 (on network data).

Christopher G. Healey. 2001. "Combining perception and impressionistic techniques for nonphotorealistic visualization of multidimensional data." SIGGRAPH Paper.

Christopher Adolph. 2003. "Visual interpretation and presentation of Monte Carlo results." *The Political Methodologist*

Thursday, 11 February – Thursday, 18 February · Visualizing Model Inference

Required: Gary King, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Interpretation and presentation." *American Journal of Political Science* 44:2

Healy, Ch. 6.

Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's practice what we preach: Turning tables into graphs." *The American Statistician* 56:2.

Andrew Gelman. 2011. "Why tables are really much better than graphs (with responses and rejoinder)." *JCGS* 20:1.

Optional: Rob J. Hyndman and Han Lin Shang. 2010. "Rainbow plots, bagplots, and boxplots for functional data." *JCGS* 19:1.

Ying Sun and Marc G. Genton. 2011. "Functional boxplots." *JCGS* 20:2 (note final figure for 3D confidence intervals).

PROBLEM SET 2 DUE TUESDAY, 16 FEBRUARY BY EMAIL

Tuesday, 23 February · Visualizing Model Robustness and Interactions

Required: Andrew Gelman. 2004. "Exploratory data analysis for complex models (with response and rejoinder)." *JCGS* 13:4

Optional: Achim Zeileis, David Meyer, and Kurt Hornik. 2007. "Residual-based shadings for visualizing (conditional) independence." *JCGS* 16:3.

Christopher Adolph. 2013. *Bankers, Bureaucrats, and Central Bank Politics: The Myth of Neutrality.* Cambridge University Press. Selected chapters on the display of interactive specifications.

Thursday, 25 February · Interactive Visual Displays

Required: "Tutorial: Building 'Shiny' Applications with R."
`shiny.rstudio.com/tutorial`

Tuesday, 2 March · Advanced LaTeX for Scientific Typesetting (*lecture if time permitting*)

Recommended: Tobi Oetiker. 2018. *The Not-So-Short Introduction to LaTeX*. Version 6.3. Ch. 1–3 and possibly 6.

Optional: Will Robertson. "The `fontspec` package." 2018. Version 2.46h. (full modern type support for advanced LaTeX users).

Part III: Student Presentations

Thursday, 4 March – Thursday, 11 March · Final Poster Presentations

Students will have a chance to express preferred presentation dates, which we will accommodate as far as is feasible given the constraint of keeping the number of presentations roughly equal across dates.

PROBLEM SET 3 DUE THURSDAY, 11 MARCH BY EMAIL