J475: Concepts and tools for data analysis and visualization
Fall 2014, 4 credits
Tuesdays and Thursdays, 2116 Vilas Hall
Dr. Chris Wells (cfwells@wisc.edu)
Office: 5004 Vilas Hall
Office hours: Wednesday, 1:30-3:30, and by appointment
http://dataviz.journalism.wisc.edu/

Introduction

Ours has become a data society. Like at no other time, our world--the natural world, from storm systems to diseases; governments and companies; and our conversations with friends and relatives, even our movements--is recorded in digital format. A few years ago, Google CEO Eric Schmidt famously stated that "There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing." (An exabyte is 1 trillion megabytes.)

The result is that researchers, companies and organizations now literally have more data than they know what to do with. There is tremendous need across our society for people who are able to use data to investigate important questions, draw useful insights from those data, and communicate those insights to others--and also to be realistic and honest about what data can and *cannot* do.

That is what this class is for: it is an introduction to the world of data and how data can be used to answer questions. More specifically, we will offer a combination of conceptual training, instruction in specific tools for data analysis and visualization, and the opportunity to put new skills to use in a final project.

Technical abilities

All this talk about data may be scary for the math-apprehensive among us. Don't worry. This course is specifically designed for students in the social sciences (not engineering, not computer science). So we do not expect **any** familiarity with programming or higher math, for example. The most important thing that you bring is curiosity: willingness to think about interesting questions and how they might be answered, and the patience to endure some trial and error in answering them.

There will be a little bit of math. We will go over some key concepts in statistics, and topics in what we have come to call simply **numeracy**. Numeracy does not mean calculus or really anything fancy like that. It means having some basic skills to treat data appropriately and spot misuses of data--and it is one of the most sought-after things, according to our conversations with people at the top of the data field. What is more, our approach to numeracy is fundamentally hands-on: it is not about abstract equations but about why uses of numbers and data are appropriate in certain instances and not in others. We think this gives that aspect of the course a much more intuitive and useful foundation.

You will be learning some new software in the course, as you have many times before. The goal here is twofold: one is simply that there are certain tools that are very useful and it is good to have in your personal toolkit. At the same time, tools do change quite rapidly--so the second goal is for you to develop the skill of picking up new software skills fairly rapidly.

Structure of the course

You will be working alongside a diverse set of students: students from journalism, strategic communication, library and information studies, graduate students focused on research, undergraduates thinking about their first job, and so forth. While elements of the course are somewhat tailored to one or another of these specialties--and it is useful for you to know that--what we have found in our conversations with researchers, journalists, marketers, advertisers and others is that their concerns about working with digital data are very similar: they are concerned with how they can take the data they have, draw useful insights from it using evidence-based reasoning and visualizations, and use those insights to say something meaningful or create a strategic course of action.

Thus, what we cover in the course will have broad applicability for students. That said, as graduate students and upper-level undergraduates, we expect you are beginning to develop your own particular interests and expertise, and we hope you use this class to build on them. As a result, several of the assignments for the class--and the final project especially--are quite open-ended as to topic and presentation. You may choose to produce a traditional newspaper story on water quality issues in Wisconsin, illustrated with charts and county-level maps; or you might create a marketing report for a company building on your analysis of that brand's reputation on Twitter; or it might be an analysis of public opinion trends on transportation issues in the Midwest over the past decade; or it could be a sports story about the NFL careers of Wisconsin Badgers. As you will see, the specific project you choose to pursue will be less important than that you demonstrate your ability to: (1) ask an interesting question; (2) find data appropriate for asking that question; (3) analyze your data appropriately; (4) visualize the data--both for analysis and presentation; and (5) draw meaningful insights from your analysis.

Flexibility: I am open to some flexibility in how students pursue the class. Graduate students especially may have specific interests or goals to pursue, and I am open to modifying the class requirements in such situations: for example, a student may wish to focus on learning python code and R in place of Tableau. Please talk to me well in advance of assignments and deadlines so that we can come to a clear, mutually agreed-upon plan.

Weekly plan: In a week, we will typically use Tuesday as a lecture and discussion, and Thursday as time in the lab working on exercises, assignments or the final project.

Class attendance: Come to class. Ten percent of your grade will be based on attendance and participation. Contribute ideas and critiques to discussions. Help your peers and learn from them during lab time. Moreover, this is a small class, and absences disrupt class flow.

Software: We will focus mainly on three software packages: Excel, Google Charts and Fusion Tables, and Tableau. Google products are available for free to everyone with an account. Computers in the lab will have access to recent Excel and Tableau licenses. On your own machine, you will want a fairly recent version of Excel (2010 or later) in order to make good use of Pivot Tables; Tableau is also generously giving us licenses for students to use on personal machines.

Books

- Alberto Cairo, The Functional Art
- Stephen Johnson, The Ghost Map

Optional/recommended books and resources

- Edward R. Tufte, The Visual Display of Quantitative Information (This is considered the bible of data visualization. It has lots of great examples, though sometimes enigmatic and snarky commentary.) Worth purchasing in print, though it is also available here: https://docs.google.com/file/d/0B2qjsD0S2vNUZGMwOGJlOGItZDFkNi00ODJiLTliMGYtNTcwZDBjZGYzNTYz/edit
- James Gleick, The Information
- *Advertising Age*'s section on data: http://adage.com/channel/data/42
- Visualizing Data, 10 significant visualisation developments: January to June 2014: http://visualisingdata.com/index.php/2014/08/10-significant-visualisation-developments-january-to-june-2014/
- Jonathan Stray, A computational journalism reading list: http://jonathanstray.com/a-computational-journalism-reading-list

Assignments and grading

Assignment 1: How to visualize	Sept. 18	5%
Assignment 2: Pivot tables	Sept. 25	10%
Assignment 3: Data cleaning	Oct. 2	5%
Assignment 4: Statistics	Oct. 9	5%
Assignment 5: Asking questions	Oct. 14 (Tuesday)	5%
Assignment 6: Fusion tables	Oct. 23	10%
Assignment 7: Good & Bad	Oct. 30	5%
Assignment 8: Tableau	Nov. 13	10%
Assignment 9: Project proposal	Nov. 20	10%
Assignment 10: Final project	Dec. 16	25%
Participation and preparation		10%

Schedule

Week 1: Sept 2-4: The data society
READ:

- Paskin, J. Serious fun with numbers http://www.cjr.org/reports/serious_fun_with_numbers.php?page=all
- Nate Silver, What the fox knows http://fivethirtyeight.com/features/what-the-fox-knows/
- Joshua Yaffa, The information sage: http://www.washingtonmonthly.com/magazine/mayjune_2011/features/the_information_sage029137.php?page=all

LAB:

- Excel refresher

Prep: Lynda, Excel 2013: Managing and Analyzing Data (sorting, filtering and formulas)

WATCH:

- Journalism in the Age of Data http://datajournalism.stanford.edu/

Week 2: Sept 9-11: Data and databases
READ:

- The Functional Art, Ch. 1-2
- Lazer, Life in the Network (PDF)
- Leaked NYTimes "Innovation Report": http://www.niemanlab.org/2014/05/the-leaked-new-york-times-innovation-report-is-one-of-the-key-documents-of-this-media-age/ (Read Nieman Lab's story and the report itself.)
- Simon Dumenco, The brutal truth about 'big data': http://adage.com/article/dataworks/brutal-truth-big-data/240364/
- Steve Lohr: Google Flu Trends: The limits of big data: http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/
- Alexis C. Madrigal, In defense of Google Flu Trends: http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/
- Check out Google Flu Trends yourself: http://www.google.org/flutrends/us/#US

LAB:

- Excel pivot tables

Prep: Lynda, Excel 2013: Pivot Tables in Depth with Curtis Frye (Watch Part 1; skim other sections)

Week 3: Sept 16-18: Communicating with visual information
READ:

- The Caging of America: http://www.newyorker.com/magazine/2012/01/30/the-caging-of-america
- Edward Segel & Jeffrey Heer, Narrative visualization: Telling stories with data (PDF)
- The Functional Art, Ch 3-4, Profile 1&6
- Brian Melmed, Six ways to make your data more human: http://adage.com/article/digitalnext/ways-make-data-human/293458/

LAB: Pivot tables

ASSIGNMENT 1 due Thursday: How to visualize data from a story?

Week 4: Sept 23-25: Data acquisition and cleaning
READ:

- Steve Lohr, For big-data scientists, 'janitor work' is key hurdle to insights: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
- Abromowitz, D. (2011, June 28). Unhelpful But Accurate Graphs. *CollegeHumor*. http://www.collegehumor.com/article/6550869/unhelpful-but-accurate-graphs

LAB: Cleaning data

ASSIGNMENT 2 due Thursday: Pivot tables

Week 5: Sept 30-Oct 2: Research methods and statistics
READ:

- Nicholas Diakopoulos, The rhetoric of data: http://towcenter.org/blog/the-rhetoric-of-data/
- Chris Anderson, The end of theory?: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory
- xkcd, Significant: http://xkcd.com/882/ (What's the problem here?)
- The Functional Art, Profile 7
- Robert Niles, Statistics every writer should know: http://www.robertniles.com/stats/
- Tufte, Visual Explanations, Ch. 2: Visual and Statistical Thinking (PDF)

WATCH:

- Rosling, Hans. "The best stats you've ever seen," June 2006. http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

LAB: Fusion tables

ASSIGNMENT 3 due Thursday: Data cleaning

Week 6: Oct 7-9: Text mining and Twitter
READ:

- Frederick W. Gibbs & Daniel J. Cohen, A conversation with data: Prospecting Victorian words and ideas (PDF)
- Aiden & Michel, Uncharted, Chapters 1 and 5 (PDF)
- Nate Silver, Democrats are way more obsessed with impeachment than republicans: http://fivethirtyeight.com/datalab/obama-impeachment-msnbc-fox-news/
- The Functional Art, Profile 10
- David Leonhardt, Inequality and web search trends: http://www.nytimes.com/2014/08/19/upshot/inequality-and-web-search-trends.html
- Google Trends! Check it out: http://www.google.com/trends/
- Google N-Grams: https://books.google.com/ngrams

LAB: Fusion tables

ASSIGNMENT 4 due Thursday: Statistics

Week 7: Oct 14-16: Asking questions and storytelling
READ:

- Stephen Johnson, The Ghost Map
- The Functional Art, Ch. 8-9
- Where People in Each State were Born: http://www.nytimes.com/interactive/2014/08/13/upshot/where-people-in-each-state-were-born.html?hp&action=click&pgtype=Homepage&version=LargeMediaHeadlineSum&module=photo-spot-region&region=photo-spot&WT.nav=photo-spot

LAB: Fusion tables

ASSIGNMENT 5 due TUESDAY: Asking useful questions

Week 8: Oct 21-23: Opinion/survey data
READ:

- Nate Cohn, Explaining online panels and the 2014 midterms: http://www.nytimes.com/2014/07/28/upshot/explaining-online-panels-and-the-2014-midterms.html?abt=0002&abg=1
- Pew Research Center: Q/A: What the New York Times' polling decision means: http://www.pewresearch.org/fact-tank/2014/07/28/qa-what-the-new-york-times-polling-decision-means/

- David Rothschild & Andrew Gelman, Modern polling requires both sampling and adjustment: http://www.huffingtonpost.com/david-rothschild/modern-polling-requires-b_b_5646174.html?utm_source=twitterfeed&utm_medium=twitter
- FiveThirtyEight, Methodology: http://fivethirtyeight.blogs.nytimes.com/methodology/
- (Optional) Mark Blumenthal, Rating Pollster Accuracy: How Useful? http://www.pollster.com/blogs/rating_pollster_accuracy_predi.php?nr=1
- Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment (pp. 178–185). Presented at the Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Retrieved from http://scholar.google.de/scholar.bib?q=info:mc319eHjea8J:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=28
- Jungherr, A., Jürgens, P., & Schoen, H. (2011). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment." *Social Science Computer Review*, 0894439311404119. doi:10.1177/0894439311404119

LAB: Tableau

ASSIGNMENT 6 due Thursday: Fusion tables

Week 9: Oct 28-30: Human vision, perception, cognition (and design!)
READ:

- The Functional Art, Ch. 5-7
- Ian Spence, William Playfair and the psychology of graphs: http://www.psych.utoronto.ca/users/spence/Spence%20(2006).pdf
- Tufte, Visual Display of Quantitative Information, Ch 5: Chartjunk
- Reingold, Tufte's Rules: http://www.sealthreinhold.com/tuftes-rules/rule_one.php
- Stephen Few, Should data visualizations be beautiful?: http://www.perceptualedge.com/blog/?p=1169
- Stephen Few, the Chartjunk debate (PDF)

LAB: Tableau

ASSIGNMENT 7 due Thursday: Good and bad visualizations.

Week 10: Nov 4-6: Geographic data
READ:

- Tom Giratikanon et al., Up close on baseball's borders: http://www.nytimes.com/interactive/2014/04/23/upshot/24-upshot-

baseball.html?action=click&region=Footer&module=Promotron&pgtype=article&abt=0002&abg=1

- Cartographia, Charles Joseph Minard posts: http://cartographia.wordpress.com/category/charles-joseph-minard/
- Zack Beauchamp, Timothy B. Lee & Matthew Iglesias, 40 maps that explain World War I: http://www.vox.com/a/world-war-i-maps
- Craig Gilbert, Dividing lines: http://www.jsonline.com/news/statepolitics/democratic-republican-voters-worlds-apart-in-divided-wisconsin-b99249564z1-255883361.html
- The Functional Art, Profile 5

LAB: Tableau

Week 11: Nov 11-13: Ethics
(Katy Culver to lead class on Tuesday.)
READ:

- Levine, When society becomes fully transparent to the state: http://peterlevine.ws/?p=14109
- Fiske, S. T., & Hauser, R. M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1414626111

WATCH:

- CNN on NY surveillance: http://www.cnn.com/video/data/2.0/video/tech/2014/05/25/cot-nyc-surveillance.cnn.html

LAB: Tableau, Project proposal preparation

ASSIGNMENT 8: Tableau assignment due

Week 12: Nov 18-20: Network science
READ:

- Reading TBA
- The Functional Art, Profile 9

WATCH:

- BBC, Gay marriage: Can online activism make a difference?: http://www.bbc.com/news/magazine-21998353

ASSIGNMENT 9: Project proposals due

Week 13: Nov 25 & Thanksgiving: Final project work

No READING this week

LAB (Tuesday): Project work

Week 14: Dec 2-4: Advanced techniques in data analysis
READ:

- Reading TBA
- The Functional Art, Profile 2

LAB: Final project work

Week 15: Dec 9-11 (Last week of classes): Presentations
No READING this week

Week 16: Dec 16-18 (Finals week): Final projects due Dec 16 at 5:00pm