UGRiD Spring Plans and Data Viz Talk

Undergraduate Researchers Interested in Data

2/11/18

Welcome back!

- Thanks everyone for coming today!
- We have many exciting things planned for the semester, including...
  - Revisiting and expanding on our partnership with the Graduate Researchers Interested in Data club.
  - Collaborating with the Center for Data Science on campus.
  - Five College DataFest 2018
- We'll finish the meeting with a discussion about data visualization and a look at some cool and accessible options for creating interactive visualizations of your own, so be sure to stick around!

UGRID -

Plans for Spring 2018

Collaboration with GRiD

- Academic talks
  - We will be presenting two more talks regarding the “Recurrent Systems for Agent Decision Making and EMG-based Motor Control” project from graduate researchers that we began last semester.
- Graduate student panel
  - The president of GRiD has expressed interest in hosting a panel of students at one of our meetings to answer questions about graduate school, research, math, computer science, and how any of the mentioned topics relate to data science.
- Hackathon event
  - GRiD welcomes advanced undergraduates to participate in a data science hackathon they will be hosting during the semester.

Center for Data Science

- Collaboration with the Town of Hadley Public Safety Committee
  - Along with GRiD and the CDS, UGRiD is hoping to become involved in a real-world data science project that would help the local community.
  - We are still at the very early stages of communication, but stay tuned!
  - This project highlights the importance (and difficulty) of the entire data science pipeline - from cross-discipline introductory talks to vet the problem as being solvable by “data science”, to the logistics of data gathering and distribution, to modeling and visualization, and everything in between.
  - One potential question: can we predict how many calls the police department will get on a given day, in order to help inform staffing decisions?

Five College DataFest

- DataFest is a nationally coordinated undergraduate competition where groups of students work over the weekend to extract insight from a rich and complex dataset, ultimately presenting their findings to a panel of judges at the end of the weekend.
- Datasets often highlight business problems, and feature an extremely wide array of approaches - past clients include Ticketmaster and Expedia.
- This semester, we will be focusing on preparing for (and eventually participating in) Five College DataFest!

DataFest and UGRiD

- We can help...
  - Find group members to form a team.
  - Give tips on what to expect or how to be most successful.
  - Offer feedback on analysis you've completed in the past.
  - Give you a chance to present your work in front of others.
  - Offer a collaborative environment similar to DataFest to perform analysis in.
- Please talk to us if you have any more questions. We really hope you consider participating!

Meeting Plans - Spring 2018

- 2/11/18 (Today!) - Back to Semester Meeting/Data Viz
- 2/25/18 - TBA (DataFest prep/topic of interest)
- 3/4/18 - EMG Research Talk #1
- 3/11/18 - Spring break (no meeting, unless there's interest!)
- 3/25/18 - DataFest prep
- 4/8/18 - EMG Research Talk #2
- 4/22/18 - Grad Student Panel (tentative) / Semester wrap-up

Other Ideas

- Please let us know if there is something you want to hear more about, and we will do our best to fit it in over the semester!
- Some potential topics we could go over in a meeting include...
  - Specific Python/R packages (e.g. dplyr, ggplot2, scikit-learn, pandas)
  - Other languages/technologies (e.g. SQL, RShiny, personal portfolio websites through Github)
  - Mathematical foundations of data science (e.g. specific algorithm, or just broad techniques.)
  - Specific domain area (e.g. health, sports, video games, NLP)

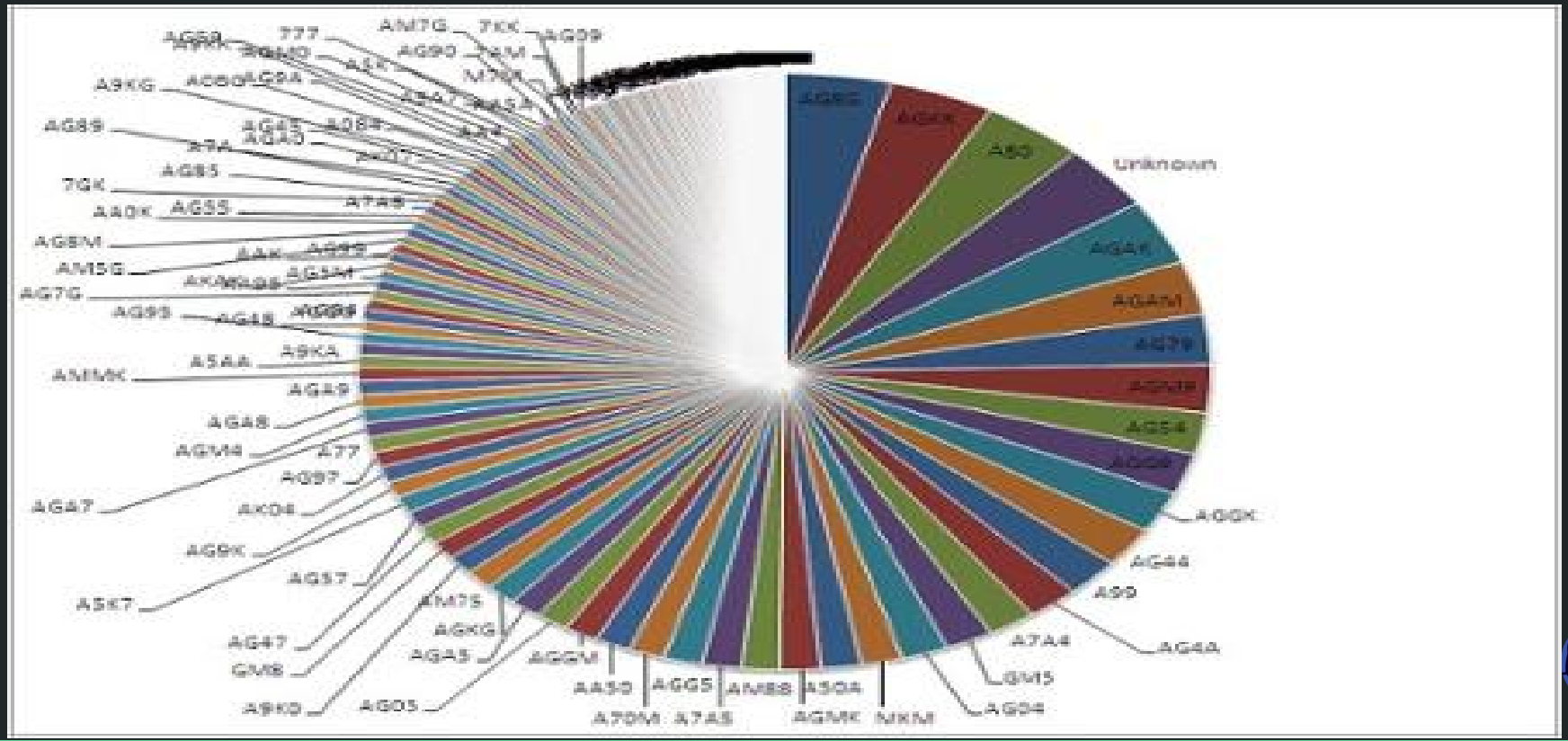
Today's thought -

What makes an effective data visualization?

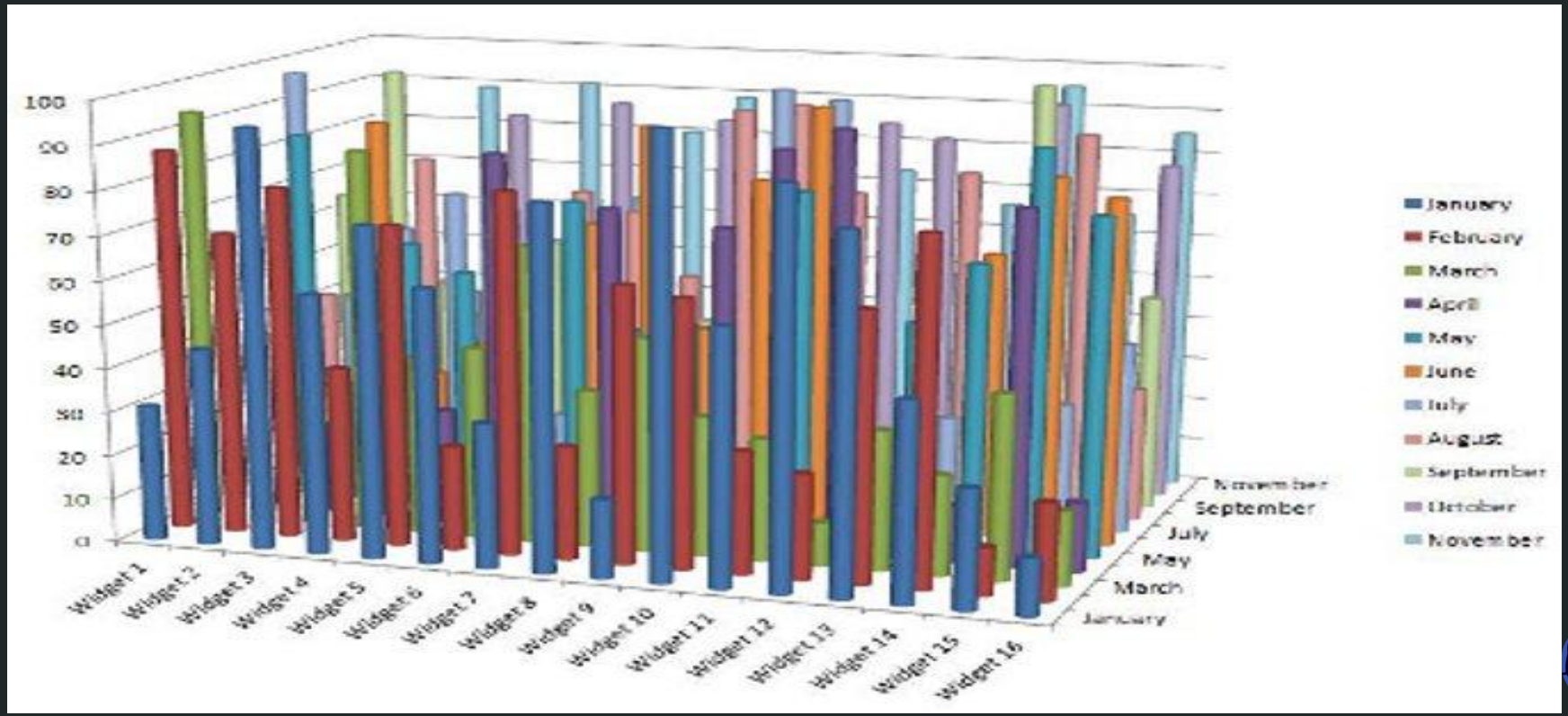
“Good” Visualization

- It may be difficult to present an objectively “good” graphic, but here’s some we liked...
  - Les Mis Co-occurence
  - Radiation Dose Chart
  - FluSight Network
  - Many more examples...(D3 gallery)
- Do you have any more suggestions?

“Bad” Visualization

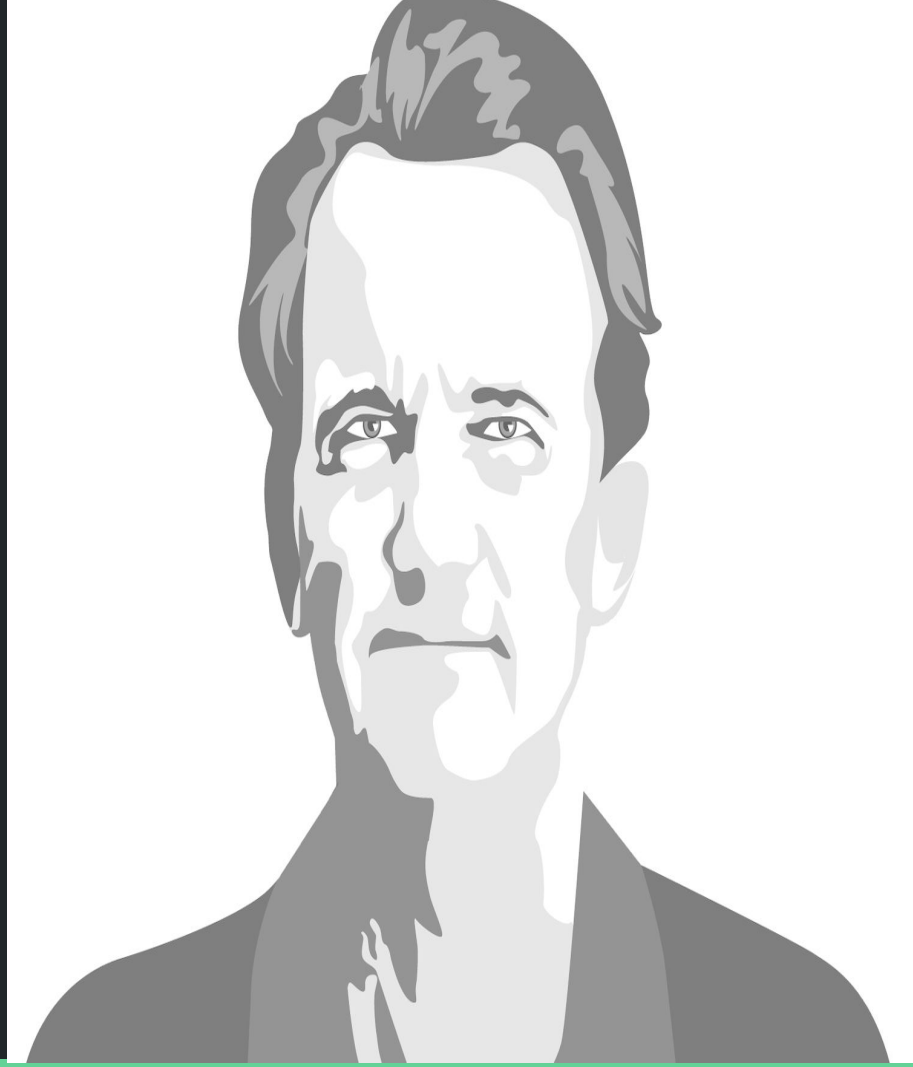


“Bad” Visualization



Enter Tufte

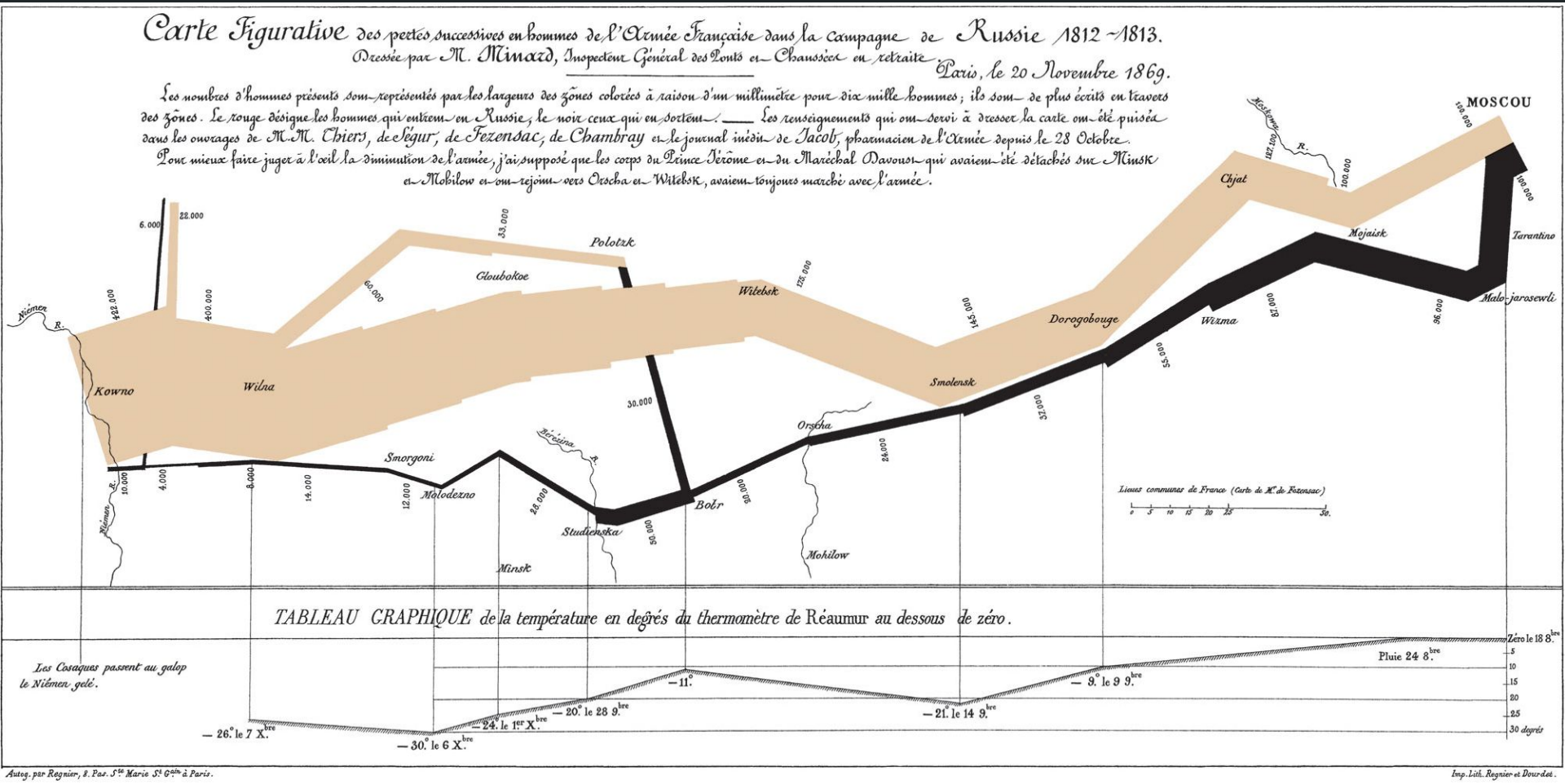
- While it may be impossible to establish a set of universal rules that dictate what makes an effective visualization, one man has set out to do just that.
- **Edward Tufte** is a statistician and information designer famous for being a pioneer of data visualization.
- His books *“The Visual Display of Quantitative Information”* and *“Envisioning Information”* are still referenced to this day for their broad, principled guidelines to information design.
- Ironically, he hates Powerpoint.



Broad Summary of Tufte's Guidelines

- There are a few main adages of effective data visualization and display of quantitative information:
  - "Show" the data
  - Maximize the data-to-ink ratio
    - Present many numbers in a small space
  - Encourage the eye to compare data
    - Highlight several different levels of detail at once
  - Provoke thought about the subject at hand.
- Graphical excellence:
  - is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
  - is nearly always multivariate.
  - requires telling the truth about the data.

Minard's Map of Napolean's Campaign into Russia



Japanese Train Map

Escaping Flatland

Even though we navigate daily through a perceptual world of three spatial dimensions and reason occasionally about higher dimensional arenas with mathematical ease, the worlds portrayed on our displays of information are caught up in the two-dimensionality of the endless flatlands of paper and video screen.[1] All communication between the readers of an image and the makers of an image must now take place on a two-dimensional surface. *Escaping this flatland is the essential task of envisioning information—for all the interesting worlds (physical, biological, imaginary, human) that we seek to understand are inevitably and happily multivariate in nature. Not flatlands.*

[1] "Flatland" is described in the classic by A. Square [Edwin A. Abbott]. *Flatland: A Romance of Many Dimensions* (London, 1884). A statement from a modern artist's viewpoint—how can abstract painting escape flatland?—is found in Frank Stella, *Working Space* (Cambridge, 1986).

This chapter outlines a variety of design strategies that sharpen the information resolution, the resolving power, of paper and computer screen. In particular, these methods work to increase (1) the number of dimensions that can be represented on plane surfaces and (2) the data density (amount of information per unit area).

In this Japanese travel guide, an engaging hybrid of design technique, the abrupt shift from friendly perspective to hard flatland shows the loss suffered by giving in to the arbitrary data-compression of paper surfaces. A bird's-eye view with detailed perspective describes local areas near the architecturally renowned Ise Shrine; then, at the right margin, a very flat map delineates the national railroad system linking the shrine to major cities, compensating for loss of a visual dimension with a broad overview. A change in design accommodates a change in the scale of the map. All in all, local detail is seen within national context, a mixed landscape of refuge and prospect. Horizontal layout harmoniously combines with the vertical orientation of the language, so that the stand-up labels point precisely to each location.

Guide for Visitors to Ise Shrine (Ise, Japan; no date; published between October 1948 and April 1952, according to The Library, Ise Shrine, Mie Prefecture).

More on Tufte’s Design Principles

- These slides, courtesy of Oregon State University, go into further detail about Tufte’s principles better than we could attempt to.

Modern Tool for Data Visualization

- R
  - ggplot2
  - Shiny (for interactive web applications)
- Python
  - Matplotlib
  - Seaborn
  - Bokeh/Dash (for interactive web applications)
- JavaScript/General
  - Plotly
  - D3
  - Tableau
  - Leaflet (for interactive map plots)
- ...and many more!

Conclusion

- Making a universally well-received data visualization is difficult, but there are guidelines one can follow to make them generally more effective at conveying your intended meaning.
- Modern tools for allow for a high degree of interactivity in visualizations, mostly due to the advent of HTML/JavaScript based design libraries and their integration into common statistical software platforms like R and Python.
- Don’t underestimate the importance of effective visualization in communicating results, especially to non-technical audiences.
- Try out some of these tools on your own or at one of our meetings, and have fun!
    - Many online resources exist for learning them, and we are always available to help answer questions!