Data Visualization syllabus

Session 1, Introduction to data visualization

In Session 1, we will discuss what data visualization is, how data differs from information and knowledge, and why we visualize data. We will consider these issues in relation to the advent of 'big data'.

Readings on DIKW

- Ted talk by David McCandless (2010), 'The Beauty of data visualizations', https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization
- Alberto Cairo (2013), 'Chapter 1: Why Visualize: From Information to Wisdom', *An introduction to information graphics and visualization* – available here: http://ptgmedia.pearsoncmg.com/images/9780321834737/samplepages/0321834739.pdf
- Jennifer Rowley (2007), 'The wisdom hierarchy: representations of the DIKW hierarchy', *Journal of Information Systems*, 33:2, pp. 163-180

Readings on big data visualizations

- Wang, Wang and Alexander (2015), 'Big data and visualization: methods, challenges and technology progress', *Digital Technologies, 1:1, pp. 33-38.*
- LaValle et al. (2011), 'Big data, analytics and the path from insights to value', *Sloan MIT Management Review*, 52: 2.
- Nikos Bikakis (2018), 'Big data visualization tools', available at: https://arxiv.org/pdf/1801.08336.pdf

Further readings on DIKW

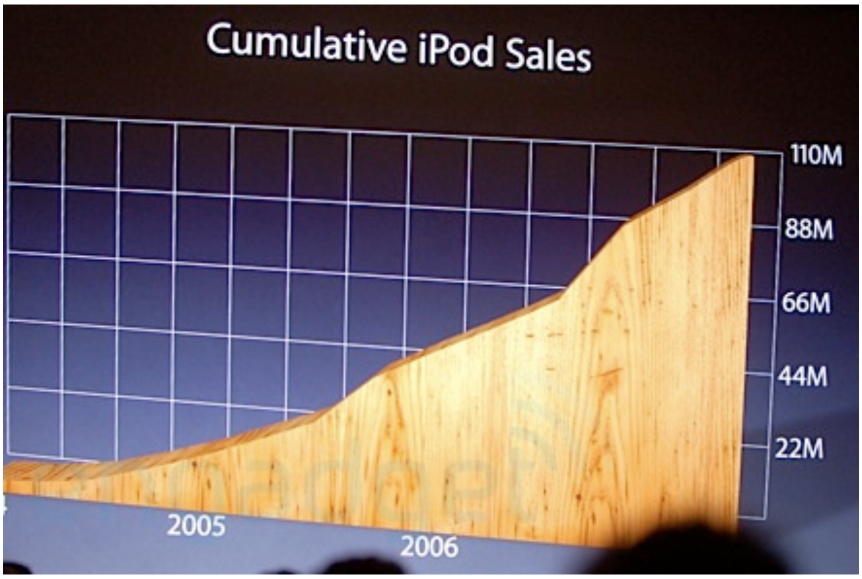
- Paul Cooper (2016), 'Data, information, knowledge and wisdom', *Anaesthesia and intensive care medicine*, 18: 1, pp. 55-56
- Cole Nussbaumer Knaflic, 'Introduction', Storytelling with data: a data visualization guide for business professionals – available on SOLO
- Tanja Keller, Sigmar-Olaf Tergan (2005), 'Visualizing Knowledge and Information: An Introduction' in Tergan and Keller (eds.), *Knowledge and information visualization: searching for synergies*, Germany: Springer – available at https://link.springer.com/content/pdf/10.1007%2Fb138081.pdf – NOTE: only read the first part, up to the description of the other papers in the collected edition.

Further readings on big data

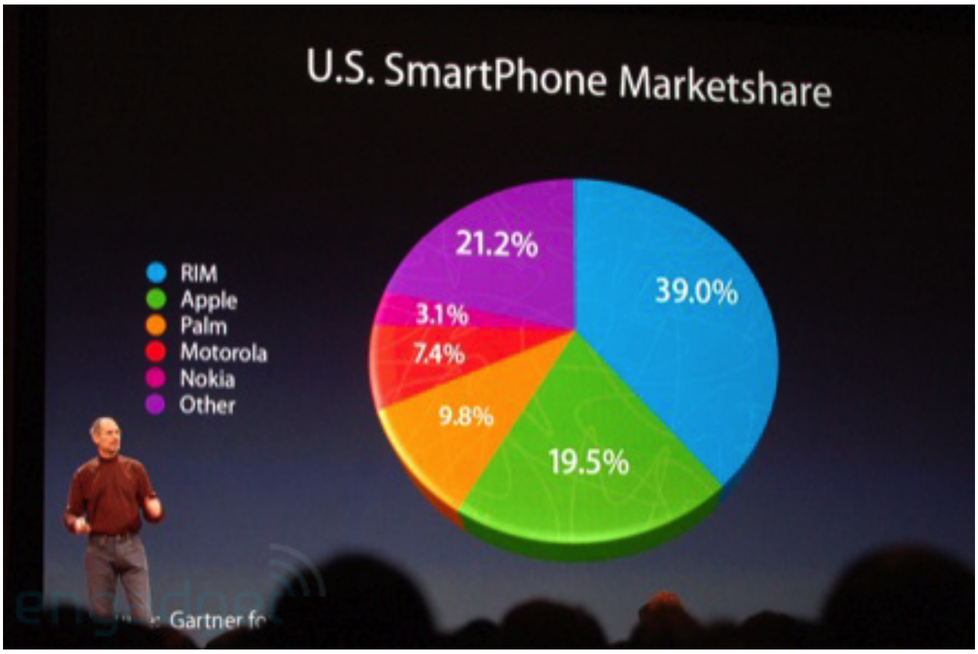
- Kitchin and McArdle, 'What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets', *Big Data & Society,* January–June 2016, pp. 1–10
- danah boyd & Kate Crawford (2012), 'Critical questions for big data', *Information, Communication & Society*, 15:5, pp. 662-679
- Roger Burrows and Mike Savage, 'After the crisis? Big Data and the methodological challenges of empirical sociology', *Big Data & Society*, April–June 2014, pp. 1–6

Practicals

1. When would you use a bar chart and when would you use a histogram?
2. When would you use a scatter plot and when a line plot?
3. What are the risks of using a cumulative diagram, compared with a frequency diagram? Think about this cumulative diagram below (from https://paragraft.wordpress.com/2008/06/03/the-chart-junk-of-steve-jobs/)



4. What is problematic about this pie chart?



Discussion points

1. What is the difference between data, information and knowledge?
    a. Does data become information once we visualize it?
2. How has the ‘advent of big data’ changed the nature of visualization?
3. How can visualizations help us to handle the ‘data deluge’?

Assignment

Prepare a short statement about the difference between data, information and knowledge (approx. 1 page), situated in relation to ‘big data’ – we will use this as a springboard for our discussions.

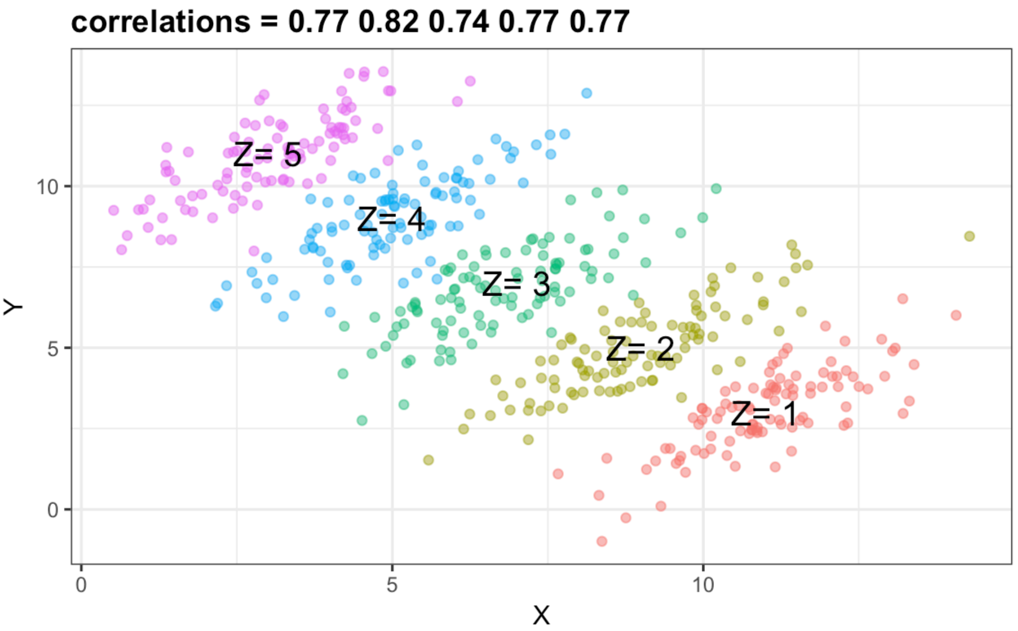
Assessment: This week will be assessed based on both the statement and our discussion.

Submission: 1 day prior to meeting (Wednesday 24th)

Session 2, The purpose of visualization: scientific inquiry

Readings

- Peter Fox and James Hendler (2011), ‘Changing the Equation on Scientific Data Visualization’, *Science*, 331, pp. 705-708.
- N. Cox and K. Jones (1981), ‘Exploratory data analysis’
    - Ravi Parikh (2014), ‘Anscombe’s Quartet, and Why Summary Statistics Don’t Tell the Whole Story’ available at: https://heapanalytics.com/blog/data-stories/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story
- Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) ‘Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm’, *PLoS Biol*, 13:4
- McInerny et al. (2014), ‘Information visualisation for science and policy: engaging users and avoiding bias’, *Trends in Ecology and Evolution*, 29: 3, pp. 148-157
- Kandel et al. (2011), ‘Research directions in data wrangling: Visualizations and transformations for usable and credible data’, *Information Visualization*, 10:4, pp. 271-288.
- Graves and Hendler (2013), ‘Visualization tools for open government data’
- Simpson’s paradox
    - Gerta Rücker and Martin Schumacher (2008), ‘Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis’, *BMC Medical Research*, 8, pp. 1-8 [NOTE: just focus on the first 3 pages! Don’t worry about their example]
    - https://www.thoughtco.com/what-is-simpsons-paradox-3126365
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5936043/
    - Example diagram of Simpson’s paradox – if we didn’t split the data into groups it would look like overall there is a downward trend.



Further readings

- Bertini and Lalaane, ‘Investigating and Reflecting on the Integration of Automatic Data Analysis and Visualization in Knowledge Discovery’
- Franco Moretti (2003), ‘Graphs, Maps and Trees’, *New Left Review*, Nov-Dec, pp. 67-93
- Keim et al. (2006), ‘Challenges in Visual Data Analysis’, *Information Visualization*, July 5-7, London, IEEE, pp. 9-16,
- Perer and Shneiderman (2008), ‘Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis’, *CHI 2008 Proceedings – Visual synthesis, April 5-10, Florence, Italy.*

Practicals

1. Explain Anscombe’s quartet and Weissgerber’s argument
2. Explain Simpson’s paradox

Discussion points

3. What is the role of data visualization in scientific inquiry?
4. How can visualizations be used in scientific reporting/communication?
5. In what cases can visualizations illuminate or cloud insight from scientific research?
6. How do visualizations differ for explorative and confirmatory analyses?
7. What is the relationship between data visualization and data modelling?

Assignment

1,000 to 2,000 word essay. You can use one of the five discussion points as a title.

Assessment: This week will be assessed solely based on the quality of your essay. A good answer will think critically about the utility of visualization, the perils of over-visualizing, and how visualizing can be used not only as a *communicative* tool but also an *exploratory* tool for gaining insight. It will also provide examples (the examples in the readings can be used)

Submission: 1 day prior to our meeting.

Session 3, What makes a good visualization?

Readings

- BCG, 'Chicken slides' (YouTube video) https://www.youtube.com/watch?v=V8Mmcce5D0k
- Cole Nussbaumer Knaflic, 'Chapter 1: The importance of context', *Storytelling with data: a data visualization guide for business professionals* – available on SOLO
- Cole Nussbaumer Knaflic, 'Chapter 2: Choosing an effective visualization', Storytelling with data: a data visualization guide for business professionals
- Cole Nussbaumer Knaflic, 'Chapter 3: clutter is your enemy!', Storytelling with data: a data visualization guide for business professionals
- Cole Nussbaumer Knaflic, 'Chapter 4: focus your audience's attention', Storytelling with data: a data visualization guide for business professionals
- Colin Ware (2004), *Information visualization: perception for design* [a chapter-length short book], Elsevier.
- Alberto Cairo (2015), 'Graphics lies, misleading visuals: reflections on the Challenges and Pitfalls of Evidence-Driven Visual Communication' in *New Challenges for Data Design*.

Further readings

- Jock MacKinlay (1987), 'Automating the design of graphical representations of relational information' – in particular, read sections 5, 6, 7 & 8
- Alberto Cairo (2013), 'Chapter 3: The beauty paradox: art and communication', *An introduction to information graphics and visualization*
- Manovich (2011), 'What is visualization?' *Visual studies*, 26: 1, 36-49
- Blogpost (2018), '9 Principles of Design: What Are They and How Can You Use Them?' – available at: https://www.idashboards.com/blog/2017/07/26/data-visualization-and-the-9-fundamental-design-principles/
- Blogpost (2018), 'Think before you pie chart' – available at: https://medium.com/geckoboard-under-the-hood/think-before-you-pie-chart-and-more-effective-ways-to-visualize-your-data-862ea3456b26

Practicals

- During the session we will look at this paper and critically discuss the visualizations - http://rsos.royalsocietypublishing.org/content/2/8/150266?elq=0f0c9e616de044df94be7db8b4a4da5c&elqCampaignId=6&elqaid=13505&elqat=1&elqTrackId=55cab36c34954760aed099f81d1d49ca
- Critically discuss the BCG Chicken video.
- Look at a plot using ggplot2 with the defaults, and the cowplot package – critically discuss the merits of each
- Critically discuss the Cairo visualization on page 10

Discussion points

- How do we produce elegant informative visualizations?
  - Practical principles of data visualization – e.g. ink to information ratio
  - Removing clutter / adequate labelling
  - Making the key features 'pop' – what is the message we want to convey?
  - Adequate labels and titles – but not overkill!
  - How can we convey our main message as effectively as possible?
- How can we know what type of visualization we should use?
  - Type of data
  - Purpose/message
  - Audience
- What principles could we agree on for making good visualizations?
- What are the limitations of assuming that everyone has similar cognition / way of viewing data? What about black and white and non-colourful visualization?

Assignment 1 (75%)

Create 3 'good' visualizations in R. Create or find from other research/online sources 3 'bad' visualizations (I want you to create the good visualizations but I'm happy for you to just find other people's bad ones if that is easier). Explain during the session what makes them good/bad.

- Think about how you can make your ggplot2 visualizations 'presentation ready' – e.g. using times new roman, commas for large values, meaningful colors, subheading.

Assessment: This week will be assessed solely based on how you critically examine the visualizations and explain what is good/bad about them (i.e. you will *not* be assessed directly on the visualizations themselves).

Submission: During the session.

Assignment 2 (25%)

Take Genis Carreras's *Philographics*. Find 3 visualizations that work and 3 that do not, and discuss critically during the session.

Assessment: The assessment will be based solely on your discussion.

Session 4, Creating data visualizations (finally!)

Readings

- Hadley Wickham, (2010) 'A layered grammar of graphics', *Journal of computational statistical graphics*, 19: 1, pp. 3-28
- Antony Unwin (2008), 'Good graphics?' in Chen, Hardle and Unwin (eds.), *Handbook of data visualization*, Berlin: Springer, pp. 57-78
- Marcin Kozak (2010), 'Basic principles of graphing data', *Scientia Agricola*, 67:4, pp. 483-494 - http://www.ibilce.unesp.br/Home/Departamentos/CiencCompEstatistica/Adriana/basic-principles-for-graphing-plots.pdf

Practicals

Using pen and paper, graph out the following from *Information is beautiful*. I'll describe them and you won't see them before writing them out – then we can compare you graphics with McCandless's.

1. "Scientific evidence for dietary supplements" (page 18/19) – approximately 200 values, showing:
   - Each instance is a dietary supplement, e.g. cod liver oil, Vitamin D, honey, calcium
   - The popularity of each supplement, continuous (based on Google searches) – McCandless bins into 3 groups
   - The level of scientific evidence – continuous (from None to Strong on a sliding scale) – could also order ordinally, binning into e.g. 4 groups of different strengths

2. "Alternative medicine" (page 204/205) – approximately 100 values, showing:
   - Each instance is a type of medicine, categorical, e.g. Chinese medicine, meditation, yoga, astrology
   - The level of scientific evidence, ordinal, e.g. slight/promising/good/strong
   - The popularity (imagine in terms of number of users etc.), continuous, e.g. 0 to 100
   - The class of medicine, categorical, e.g. body, psyche, mystical, natural

3. "Most profitable Hollywood stories" (page 226/227) – approximately 100 values, showing:
   - Each instance is a film from 2007
   - Average review score, percentage, e.g. 0 to 100 – from Rotten tomatoes
   - Profitability as a percentage of the budget, percentage, e.g. 0 to 1,000 – from IMDB
   - Size of budget, continuous, e.g. 25m to 150m – McCandless bins into 4 groups
   - Type of story, categorical, approx. 25 different categories

Discussion points

- What is hard about making a good visualization?
- How does data wrangling fit into the visualization pipeline?
- What are important tips to bear in mind when making a visualization?
- Explain the grammar of graphics – ideally, with examples in ggplot2

Assignment

1. Graph a very large volume of bivariate data (e.g. 10,000 values – the data can be real or synthetic). Find at least two ways of representing the data elegantly, in such a way that relationships in the data can be identified easily.
2. Graph three variables – find at least two ways of representing the data elegantly
3. Discuss in person the choices that you made in graphing the data.

Submission: 1 day prior to meeting

Assessment: The assessment will be based on how innovative the graphs are, and how critically you discuss them.

Answers to assignment 1

a. Heat plots
b. Contour lines
c. Heat scatter plots
d. Rug plots

Answers to assignment 2

a. Facet wrap
b. Use of different colour lines / points
c. Use of different icons for each data point (e.g. squares vs. triangles)

Session 5, Storytelling with data visualizations

Readings

- Cole Nussbaumer Knaflic (2015), ‘Chapter 7: Lessons in storytelling’, Storytelling with data: a data visualization guide for business professionals
- Segel and Heer (2010), ‘Narrative visualization: telling stories with data’
- Kosara and MacKinlay (2013), ‘Storytelling: the next step for visualization’, *IEEE Magazine*

Discussion points

- How can we represent the data in informative / attention grabbing ways without manipulating the data or being unfaithful to what it contains?
- How can we link different graphics together?
- What are some good principles of storytelling?
- How can we avoid using the same graphics over and over again when storytelling?

Assignment

Collect some data (e.g. World Bank data) and tell a story; make a 3-5 slide presentation

Submission: During the session

Assessment: The assessment will be based on how informative and well-made the visualizations are, how well they fit together, and how much of an effective narrative you construct. I will find 1-2 other people to weigh in and give their opinions on your presentation.

Session 6, The social aspect of data visualization

Non-academic readings (start here!)

- Mark Graham, 'Mapping internet users' at http://www.markgraham.space/blog/2017/7/10/mapping-internet-users
- Mark Graham, 'the hidden biases of Geodata', *The Guardian*, at https://www.theguardian.com/news/datablog/2015/apr/28/the-hidden-biases-of-geodata
- Mark Graham, 'Internet and information geographies', https://www.oii.ox.ac.uk/videos/internet-and-information-geographies-mark-graham-at-tedxbradford/

Academic papers

- Mark Graham, Stefano de Sabbata and Matthew Zook (2015), 'Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information', *Geography and Environment*

- Mark Graham (2015), 'Information Geographies and geographies of information' in Fard and Meshkani (eds.), *Geographies of Information*.

- Zook and Graham (2007), 'Mapping DigiPlace: geocoded Internet data and the representation of place', Environment and Planning B: Planning and Design 2007, volume 34, pages 466 - 482

- Anthony McCosker & Rowan Wilken (2014) 'Rethinking 'big data' as visual knowledge: the sublime and the diagrammatic in data visualisation', *Visual Studies*, 29:2, pp. 155-164

- Wanda J. Orlikowski & Susan V. Scott (2008) 10 Sociomateriality: Challenging the Separation of Technology, Work and Organization, The Academy of Management Annals, 2:1, 433-474

Practicals

1. Describe a Chloropleth
2. Describe a heat map (and how it differs from a Chloropleth)
3. Describe a Cartogram
4. Describe an Isoline

Discussion points / Essay titles

- How do visualizations reflect and/or create social reality? Can we 'trust' visualizations?
  - What is the relevance of Scott and Orlikowski's work for understanding visualizations?
- How can we draw attention to issues through visualizations? Are visualizations a political / communicative tool? Discuss the idea that "A picture is worth a thousand words…"
- Critically discuss the maps in Graham/de Sabbata/Zook (2015)

Assignment

1,000 – 2,000 word essay.

EITHER:

Pick one of the discussion points as the basis of an essay question (feel free to adjust as needed)

OR:

Discuss the political and social implications of visualizations, using maps as a case study.

Submission

1 day prior to meeting.

Assessment

The assessment will be based solely on the critical nature of the argument. The size of the bibliography will make no difference to the assessment (!).

Session 7, Interactive visualizations

Readings

- Dur (2014), 'Interactive infographics on the Internet', *Online Journal of Art & Design*, 2: 4, pp. 1-14
- Dick (2014), 'Interactive infographics and news values', *Digital Journalism*, 2:4, pp. 490-506
- Hall et al. (2016), 'Formalizing emphasis in information visualization', *EuroVis*, 25: 3, pp. 717-737.
- Weissgerber TL, Garovic VD, Savic M, Winham SJ, Milic NM (2016) From Static to Interactive: Transforming Data Visualization to Improve Transparency. PLoS Biol 14(6)
- Ellis and Meridian (2018), 'The visualizations of data in a digital context' in Costa and Condie (eds.), *Doing research in and on the digital: research methods across fields of inquiry*, Abingdon: Routledge

Examples

- WikiGalaxy - http://wiki.polyfra.me
- CitiBikes in the New Yorker - https://projects.newyorker.com/story/citi-bike.html
- Our World in Data - https://ourworldindata.org/world-population-growth
- NOTE: The Guardian and the New York Times are excellent sources of interactive (and static) visualizations

Discussion points

- What is the difference between a static and interactive visualization?
- What new components can we add through interactivity? What are the downsides?
- How technically challenging are interactive visualizations to use and to make? Is it a worthwhile endeavour?
- What new challenges does interactivity pose?

Assignment

1,000 – 2,000 word essay. Find three interactive visualizations from the Internet and critically discuss each in turn. Please include snapshots of the visualizations). Explain what works well and what is less effective. Consider this specifically in relation to the purpose of the visualizations and the intended audience. Extra marks will be awarded for critically considering whether the interactivity really adds value / enhances the visualization; you could consider counter-factually if the visualization was non-interactive. Note – this week, the emphasis is on critically applying the principles we've discussed previously. Your discussion should be practically oriented rather than abstract/theoretical.

Submission: 1 day prior to meeting

Assessment: solely based on the quality of the essay

Session 8, Interactive visualizations (2) – practical

Readings

- https://shiny.rstudio.com/articles/build.html
- http://rstudio.github.io/shiny/tutorial/

Assignment

Make a shiny app in R. Describe in person how you made it and what you learnt.

Submission: During the session.

Assessment: Assessment is based on the Shiny app and your description of it. If you can identify lessons learnt and what you would do differently next time, you will get bonus marks.