Escaping Flatland: A Discussion of Data Visualization in Libraries

Andrew Weiss, Digital Services Librarian, The California State University, Northridge

Abstract

After an opening exercise in which participants were asked to collaborate on how best to organize, draw, and present a set of synonyms, the session introduced Edward Tufte's data visualization work and then presented working examples of how digital visualization is being used by various organizations. The theme was meant to be provocative and to challenge assumptions librarians may have about their users, including their web-use and information seeking behaviors. Discussion topics revolved around several essential questions, including "What obstacles prevent libraries from better utilizing data visualization?", "What makes effective data visualization?", "How can libraries improve upon existing tools?", and "What have others done to present information better?" Session participants were also asked to provide critiques of both good and weak data visualizations. Additionally, the presenter shared two data visualization projects that he had created from MIT Simile widgets to interact with collection housed in the digital management platform CONTENTdm.

Discussion

Introduction

Data visualization has been an important part of sharing information since the groundbreaking work of mathematicians such as Galileo, Leonardo, and Newton. Edward Tufte's studies into the impact and aesthetics of data visualization provide a solid framework for the discussion of improving the presentation of information in libraries and library data systems. The impact of data visualization can be best summarized by Anscombe's Quartet, which demonstrates that number sets with identical statistical properties can sometimes plot with widely varying results. Providing visualization of the number sets helps viewers to more easily grasp outliers and other anomalous data points that otherwise might go unnoticed (Tufte, 1983, p. 14).

Improving data visualization for libraries becomes more important in the digital age as more data becomes aggregated from ever-expanding sources, as is evidenced in massive online digital aggregated libraries such as Google Books, Hathi Trust, and the like. MIT has been a pioneer in the creation of open source and open access data visualization with their SIMILE widgets project (MIT, 2012). Their open source programs allow libraries to harness metadata and present it in more user-friendly ways. One example is their timeline function which can plot data points along a time-frame of the designer's choosing. The presenter used this widget to great effect with a digitization project completed with the Kansas Cosmosphere and Space Center (Weiss, 2011a, 2011b).

Although data visualization can have a powerful impact upon presentation and can help users draw better connections and stimulate ideas in users, weak and manipulative examples of data visualization are prevalent. As a result, a few rules of thumb will help users identify good data visualization: 1) there should be a combination of substance, statistics and design; 2) the viewer should be able to get the greatest number of ideas in the shortest time within the smallest space possible; and 3) there should be presentation of truth and fact (Tufte, 1983, p. 51). Ultimately good data visualization will include user-friendly characteristics such as providing scale and time, showing progressions and changes, presenting clear causal relationships and evincing new connections. Weak or manipulative examples of data visualization include things such as "chart junk," inaccuracies, exaggerations, lack of context, and an unclear purpose (Tufte, 1983, p. 107).

Description

Activities and structure were designed to encourage the participation of all audience members. Using a Prezi presentation (link: http://prezi.com/w6_sp72jymiy/escaping-flatland-a-discussion-of-data-visualization-in-libraries/) as the narrative base and prompt, the topic of data visualization was tackled by the presenter with feedback from the audience (Weiss, 2012). After a brief introduction to the topic of data visualization, the presenter provided audience members with a large sheet of paper and colored pens. Audience members were then provided with a short list of terms, which included synonyms of the word "fair" (See appendix A, side 1, for details of Activity I). Divided by up to eight per table, audience members were asked to work together to determine how they might best present this set of terms. Audience members spent about 15-20 minutes discussing and collaborating on their visualization of this list of words. Each group was then asked to share the results of their collaboration and to explain the reasons for organizing the information in this manner.

Following Activity I, the presenter continued the Prezi slide narrative, demonstrating how commonplace assumptions about library users might be challenged by current studies, including studies in web-usability, reading habits, storytelling, and mathematics (Weiss, 2012).

After a brief description of the presenter's background working with digitization projects at Fort Hays State University and the California State University, Northridge, the presenter continued by discussing Anscombe's Quartet, which effectively demonstrates the power of data visualization. The quartet provides a clear example of how sets of data that are otherwise statistically identical visualize completely differently. At this point, the presenter raised the first audience-wide discussion question: "Given the powerful impact that it can provide users as shown with Anscombe's Quartet, the question posited, why are libraries not doing more with data visualization? In other words, what obstacles prevent libraries from better implementing it?" Discussion of these points ensued with significant audience participation.

Following the first discussion, the presentation narrative continued by examining the tenets of data visualization. Subsequent slides also discussed what constitutes good data visualization, using as a model two projects that the presenter had completed at Fort Hays State University in 2011. The presenter's collaboration with the Kansas Cosmosphere and Space Center were discussed, including the background of the Jack Swigert Apollo documents collection (FHSU, 2010), (KCSC, 2012). The presenter displayed the way in which the MIT

Simile Timeline was used for the project. He also provided information on how Simile Timeline meets the criteria for good data visualization.

The last part of the session included a discussion of poor data visualization and used the often-used word cloud as the prime example. Audience members were solicited for their opinions on how the word cloud could be improved upon for library uses. Finally, the presenter provided background and two examples of "chart junk," the types of data visualization that distract viewers and take away from effective and truthful representation. Audience members were asked to provide feedback and impressions about how or why the examples were poorly constructed.

Key Points

There were a significant number of key points to take away from the discussion session originating from the source material, the presenter and the audience.

Source material

Edward Tufte

Edward Tufte's work was the starting point for the session's exploration of data visualization, and provided the analytical and topical framework for the ensuing audience discussion and feedback. One of the defining issues from Tufte's work is how to judge the quality of data visualizations. According to his work, which was outlined in several of his major texts, including *The Visual Display of Quantitative information*, *Visual Explanations*, and *Envisioning Information*, the best visualizations contain graphical excellence, which is "the well-designed presentation of interesting data—a matter of substance, of statistics, and of design….[is] complex ideas communicated with clarify, precision, and efficiency….[and] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" (1983, p. 51) (1990) (1997).

Being able to judge the quality of these visualizations will help viewers not only to understand complex ideas, it will also help them to avoid manipulations that can occur with the misrepresentation of data. Along with providing clear criteria for the evaluation of data visualization, Tufte's scholarship also helps to define another troubling aspect in data representation: "chart junk" (1983, p. 107). Tufte defines "chart junk" as decoration generating "a lot of ink that does not tell the viewer anything new" (1983, p. 107). Tufte includes three main categories of this, including unintentional optical art (distracting designs that add vibration and movement to a figure), grids, and decorative forms that overtake the design (1983, p. 117).

MIT's Simile Widgets

MIT has been on the cutting edge of open source software since the 1990s. MIT is most well-known in library circles for its development of the institutional repository platform DSpace (DSpace, 2012). MIT has also developed a series of open source data visualization widgets to help institutions better present data sets. The two most relevant to library data would be Simile

Timeline, which presents data points across a pre-set time line that users can scroll forward and backward in time. Their demonstration of the assassination of John F. Kennedy is a prime example of the power of these visualizations (MIT, 2009a). Another widget designed by MIT is Simile Exhibit, which allows for search filtering and the possibility of displaying set results either within a timeline or within a tiled results list (MIT, 2009b). These widgets, written in JavaScript, can be downloaded and are easily customized by users who know some coding languages such as HTML and XML. Data points are populated with either XML files or with JavaScript Object Notation (JSON) files.

The Presenter

The presenter provided the audience with two distinct points. The first was providing challenges to certain assumptions that librarians may have about users. These challenges included an examination of the assumption of the "linearity" of a user's experience. Often, librarians assume users will approach reading, using a website or storytelling as a logical and linear process, moving from point A to point B. However, many of the examples provided in the slide presentation included data visualizations of how people actually read, how people search web-pages and how people process information (Nielsen and Pernice, 2009). Many of the examples show non-linear, disjointed patterns, suggesting that many assumptions about the user experience may be mistaken. A final point was shown about how so-called primitive cultures have also created highly-developed architecture based on fractal geometry, further calling into question the methods in which most library online catalogs and data management systems present information (Eglash, 2007).

The second point was to demonstrate actual data visualization projects the presenter had completed. The first project was a digitization collaboration between the Fort Hays State University and the Kansas Cosmosphere and Space Center. Using the MIT Simile Timeline widget, the presenter showed Apollo-era documents within the temporal context of the space race from Sputnik in1957 through to the Apollo-Soyuz missions of 1975 (Weiss, 2011a). The second project was the digitization of the FHSU Master's Thesis collection. Using the MIT Simile Exhibit widget, the presenter showed how searching a thesis collection might be more intuitive by allowing a timeline search, keyword and date filtering, and tiled set results that included author names, degree program, abstracts and dates (Weiss, 2011b).

The Audience's Insights

The audience members were very active in their participation and provided valuable insights regarding several of the issues raised during the presentation.

Activity I

During the opening activity, each table was asked to discuss and propose a method for organizing a set of synonyms related to the word "fair". The challenge posed was for the table groups to come to a consensus on how to best display not only the 15 terms related to fair, but how to demonstrate relationships between terms. Several of the tables made clear diagrams that mapped concepts together. A few other tables provided imaginative drawings that explained the meaning of several of the words. For example, one group drew a circus tent to exemplify the funfair meaning of fair. Other groups tried pictures to demonstrate one or more concepts. Another group suggested that spreading the synonymous terms along a time line would be a possible solution. However, the ability to carry such a plan was considered by them to be beyond their drawing capabilities. Overall, the audience provided imaginative solutions to the problem.

Discussions and critiques of good and bad data visualization

The audience was posed with several questions regarding data visualization during the session. One question posed was why libraries were not better at implementing data visualization projects. Many of the audience felt that several barriers existed. First of all, many felt that the technology hadn't existed until recently to handle the amount of data used by libraries. The lagging adoption of Web 2.0 technology, the lack of sufficient staffing and time to devote to these projects and the lack of skills or training for librarians were cited as major institutional obstacles to the implementation of data visualization in many libraries.

Regarding the critiques of some data visualization tools, fruitful results came out of the analysis of the *Wordle* word cloud (Feinberg, 2011). The presenter had used the word cloud as an example of weak data visualization. However, he asked the audience if there were ways that it might be improved. One suggested that the words be made larger; another suggested that the colors used be meaningful or representative of something. Another person suggested that depth to distinguish between words could improve usability. Finally, another suggested that having the sizes of the words correspond to meaningful statistics would improve usability. Essentially, a lack of clarity of purpose was the main criticism of the word cloud.

Overall, audience participation added significant insights into the discussion of data visualization. One final audience activity was scheduled for the session, but time was limited and therefore could not be carried out.

References

DSpace. (2012). *DSpace Home.* Retrieved April 7, 2012 from http://www.dspace.org/

Eglash, Ron. (2007). *Ron Eglash on african fractals*. Retrieved March 30, 2012 from http://www.ted.com/talks/ron_eglash_on_african_fractals.html

Feinberg, J. (2011). *Wordle Home*. Retrieved April 7, 2012 from http://www.wordle.net/

Fort Hays State University (FHSU). (2010). *Digital repository of space exploration*. Retrieved April 7, 2012 from http://contentcat.fhsu.edu/cdm/landingpage/collection/cosmosphere

Kansas Cosmosphere and Space Center (KCSC). (2012). *Kansas cosmosphere and space center home*. Retrieved April 7, 2012 from http://www.cosmo.org/

Massachusetts Institute of Technology (MIT). (2009). *Simile widgets exhibit*. Retrieved April 7, 2012 from http://www.simile-widgets.org/exhibit/

Massachusetts Institute of Technology (MIT). (2009*). Simile widgets timeline*. Retrieved April 7, 2012 from http://www.simile-widgets.org/timeline/

Massachusetts Institute of Technology (MIT). (2012). *Simile data visualization widgets*. Retrieved April 7, 2012 from http://www.simile-widgets.org/

Nielsen, J. and Pernice, K. (2009). *Eyetracking web usability*. Berkeley, CA: New Riders Press.

Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tufte, E. (1997). *Visual explanations.* Cheshire, CT: Graphics Press.

Weiss, A. (2011). *Digital repository of space exploration – Timeline search 1957-1975*. Retrieved April 6, 2012 from http://scatcat.fhsu.edu/~apweiss/timeline-space-2.html

Weiss, A. (2011). *FHSU master's thesis collection*. Retrieved April 6, 2012 from http://scatcat.fhsu.edu/~apweiss/thesis-contentdm-final.html

Weiss, A. (2012). *Escaping flatland: A discussion of data visualization in libraries*. Retrieved April 7, 2012 from http://prezi.com/w6_sp72jymiy/escaping-flatland-a-discussion-of-data-visualization-in-libraries/

Appendix I:

Escaping Flatland: A Discussion of Data Visualization in Libraries | Andrew Weiss, CSUN | 4/6/12

Activity I:

Using the provided markers & paper, organize these synonyms of ***FAIR*** *however you see fit:*

FAIR =

Average
Bazaar
Bonny
Bonnie
Clean
Comely
Carnival
Fairish
Funfair
Honest
Just
Mediocre
Middling
Reasonable
Sightly

Activity II:

A. Discuss & brainstorm how to improve the library ILS/OPAC using Data visualizations.

{OR}

B. Discuss & brainstorm ways to improve the Word Cloud for library use.

{OR}
C. Discuss & brainstorm other areas where libraries might use data visualization.

Then share results with your colleagues.

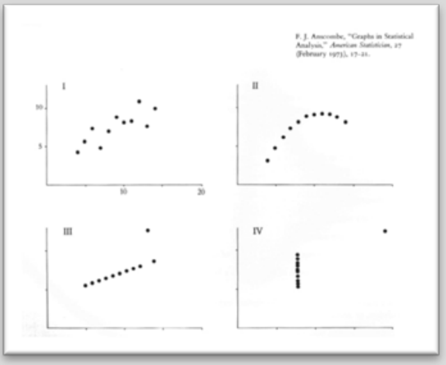
THE BASIC TENETS OF DATA VISUALIZATION: (from Edward Tufte)

- Combination of substance, statistics and design
- Complex ideas communicated with clarity, precision & efficiency
- Viewer gets greatest number of ideas in the shortest time, with the least amount of 'ink' in the smallest space possible
- It gets at "truth", not "truthiness"

GOOD EXAMPLES:

Will: Often show scale and time
Often show progressions & changes
Show clear causal relationships
Evince new connections

Anscombe's Quartet:

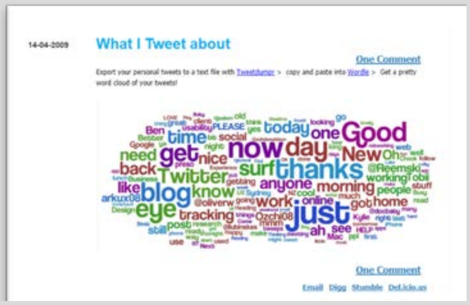


F. J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (February 1973), 17-21.

BAD EXAMPLES:

May have: "Chart junk"
Inaccuracies
Exaggerations
Little or no context
An unclear purpose

Wordle:



14-04-2009

What I Tweet about

One Comment

Export your personal tweets to a text file with Tweetdumpr > copy and paste into Wordle > Get a pretty word cloud of your tweets!

One Comment

Email Digg Stumble Del.icio.us

Data Visualization Further Reading & Study

- Bertin, Jacques. *Semiology of Graphics*. William J. Berg (Trans.). University of Wisconsin Press, 1983.
- Dodge, Martin and Kitchin, Rob. *Atlas of Cyberspace*. Pearson Education Ltd. London, 2001.
- MIT. *Simile Data Visualization Widgets*. http://www.simile-widgets.org/, 2012.
- Tufte, Edward. *Envisioning Information*. Graphics Press. Cheshire, Connecticut, 1990.
- Tufte, Edward. *The Visual Display of Quantitative information*. Graphics Press. Cheshire, Connecticut, 1983.
- Tufte, Edward. *Visual Explanations*. Graphics Press. Cheshire, Connecticut, 1997.
- Weiss, Andrew. *Digital Repository of Space Exploration – Timeline.* Forsyth Library Digital Collections. http://scatcat.fhsu.edu/~apweiss/timeline-space-2.html, 2011.