Data Visualization Literacy of Youth and Adult Science Museum Visitors

Katy Börner, Adam V. Maltese, Russell N. Balliet INDIANA UNIVERSITY, BLOOMINGTON, IN
Joe E. Heimlich CENTER OF SCIENCE AND INDUSTRY, COLUMBUS, OH
Stephen Uzzo NEW YORK HALL OF SCIENCE, CORONA, NY

OBJECTIVES

The *Sense Making of Big Data project* was designed to study how audiences in public spaces relate to and make sense of representations of large data sets. Ultimately, the project will inform the development of a traveling, hands-on exhibition that will enable visitors to create and utilize representations of data such as graphs and maps. In addition, the project hopes to create a foundation for the design of informal learning experiences that encourage participants to explore, engage, and make better sense of small and big data.

This research is timely as the amount of data in our world is exploding, and the capability to analyze large data sets *(big data)* will become a key basis for all citizens to be data literate decision-makers. For the sake of this project, we operationally define *big data* as datasets that are nationally representative (e.g., energy usage by demographic), those that are nearly exhaustive for a category (e.g., calorie counts for common foods), or that allow for extended longitudinal analysis (e.g., climatology records).

The big data phenomenon is fueled by cheap sensors and high-throughput simulation models, the increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet. It exists from social media to cell biology offering unparalleled opportunities to document the inner workings of many complex systems (Barabási, 2011). Research by MGI and McKinsey's Business Technology Office argues that there will be a shortage of talent necessary for organizations to take advantage of big data. "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (Manyika et al., 2011).

THEORETICAL FRAMEWORK

Building on major works in statistics, information visualization, and graphic design detailed in (Börner & Polley. 2014; Börner, 2015), we we argue that each visualization can be decomposed into three parts:

1. **Reference system**, also called visualization type, e.g., two-axis coordinate system or a geospatial map,
2. **Data overlay** using graphic variable types such as geometric, linguistic, or pictorial symbols, and
3. **Visual encoding** of additional data variables by graphic variable types, e.g., spatial (e.g., x-y position) or retinal (e.g., size, shape, color coding).

Specifically, the visualization framework distinguishes four general visualization types:

- **Charts:** No reference system, e.g., Wordle.com, pie charts
- **Graphs:** Quantitative or qualitative (categorical) axes. Timelines, bar graphs, scatter plots.
- **Geospatial maps:** Use latitude and longitude reference system. World or city maps.
- **Network graphs:** Node positions might depend on node attributes or node similarity: ***Tree graphs:*** hierarchies, taxonomies, genealogies ***Networks:*** social networks, migration flows.

Note that the different visualization types are preferentially used to answer different types of questions. Timeline graphs are frequently used to answer "When" questions. Geospatial maps are common for answering "Where" questions. Network graphs are good for depicting relationships, i.e., answering "With Whom" questions.

Furthermore, the framework identifies key graphic variable types that are commonly used to encode additional data variables as part of the data overlay:

- **Position:** x, y; possibly z
- **Form:** Size, shape, orientation/rotation
- **Color:** Value (lightness), hue (tint), saturation (intensity)
- **Texture:** Pattern, rotation, coarseness, size, density gradient

Note that the position of a data point in a coordinate system is determined by the axis values. However, the positioning of an object on a geospatial map requires a lookup table to convert address data into latitude and longitude information. In addition, a good visualization features a title, labels, legend, explanatory text and author information that are important for reading and interpreting the visualization. We use this framework to design stimuli that cover major visualization types and data overlays and to study if and how museum visitors decompose visualizations into the different components when reading them.

METHODS FOR DATA COLLECTION AND ANALYSIS

Data collection occurred at the New York Hall of Science in Queens, New York, the Center of Science and Industry (COSI) in Columbus, Ohio, and WonderLab Museum in Bloomington, Indiana. The 20 visualizations were compiled using the above mentioned visualization framework, had labels and legends removed, and were printed in color and laminated so that they had a finished, bright look and would last through the study. The purpose in removing labels and legends was so that individuals would respond to the data representation and would then attempt to explain how it would be interpreted without the visual cues provided by the actual legends, titles, and keys. Each of the three participating museums was provided a set of laminated data representation cards. The cards were coded on the back by letter for identification and organized into four discrete sets with each set having five visuals covering different representations. These sets were used in constant rotation. After museum visitors gave verbal consent to participate, the data collector asked the subject the following four questions for each visual within a set:

- Does this type of data presentation look at all familiar? (probe: Where have you seen images like this?)
- How do you think you read this type of data presentation?
- What would you call this type of data presentation?
- What types of data do you think would makes sense for this type of presentation?

Each individual/group did this for five visualizations (see Figure 1) while the data collector captured comments on the data collection sheet coded for each of the five figures. Perceived sex was also noted to minimize gender bias in who was asked to participate. The results discussed here come from analysis of the first three of these questions.

Each data collection sheet was transcribed into Microsoft Excel, before being imported into *NVivo 10* software for analyses. A qualitative, iterative, open-ended approach was used in the identification and definition of emergent themes in the data (Glaser, 1965). As expected, each prompt elicited distinct responses, requiring each prompt to be analyzed independently and leading to the formation of separate coding schemes (Tables 1 & 2).

We coded at the phrase level, where appropriate, to include any relevant context of a response. Respondents typically mentioned multiple ideas and thus there are many more instances than participants. In coding the *Where* responses, one of the authors initiated coding of about 20% of the 1000+ responses. A second author coded the same responses using the established codes but open to adding/editing codes if necessary. After the second pass, the first coder went back and reviewed his coding using the updated codes. At this point the average inter-rater reliability (measured through Cohen's Kappa) was 0.822. Discrepancies were discussed until agreement was reached and codes/definitions were updated. Once this was done, a single coder coded the remaining instances. A similar approach was used with the *How would you read this?* response data. For the *What would you call this?* responses, two of the authors independently came up with an 'answer key' for the visual types and then iteratively worked to create a list of equivalent terms. Examples of these are display in Figure 2.

Two audiences were included in the sample population: youth (and parents/guardians) ages 8-12 and adults (over 18). The youth group (n=127) was 45% female and the adult group (n=146) was 50% female. Data were collected within the gallery setting of each institution on a weekend day.

RESULTS

In sum, we analyzed nearly 1400 responses for the visuals presented to participants. Participants recognized the visuals presented to them 61% of the time, but age was an important factor as 45% of youths responded that the visuals were unfamiliar to them versus only 31% of adults. When asked where they had previously encountered similar visuals, youths reported school at a very high rate (51%), with books next at approximately 9% of responses (Table 1). Even then, many of these book references were to textbooks; this is not surprising as the majority of a youths' time is spent in school or engaging in school-related activities. Work was commonly mentioned by adults, but interestingly adults reported school as the most common place they saw these visuals (20%) suggesting they're still in school, referring to their children's schooling, or drawing on prior experiences. Results also show that in general adults encounter visuals in a wider variety of places more frequently than children and those places tend to be a bigger part of their non-work, non-school activities (news, entertainment, finances, etc.).

When asked how they would read the visual presented to them responses fell into two different categories (Table 2): They described components of the visual that they found useful to reading the visual (Key Visual Features; 81% of responses) and/or they described what style of information was being conveyed by the visual (What the Visual Communicates; 19%). Adults tended to describe the trends, relationships, and general information in the visual more often than children (22% and 16% respectively). Both youths and adults most frequently identified color as a key aspect to understanding the visual (16%), as well as size and quantity (13%), and the presence of objects (e.g. circles, pictures, icons, etc.; 12%). A small number of responses (n=23; 78% by adults) noted that certain aspects were missing from the visual and indicated that these pieces would help them better understand the visual. Similarly, many stated that they would use the title, legend, or key to help them read the visual (Table 2) even though those were not available for them to use. When discussing what information the visual provided, participants talked about comparisons or relationships (e.g. "Colors of lines equals different data."; 15% of instances), with 62% of these responses provided by adults. Very few responses (< 1%; not in table) indicated that reading a visual was a sequential process and most listed the key pieces with no intended order; most that mentioned this were adults (62%). A small percentage of responses (<1%; not in table) provided a specific descriptions of the visual (e.g. " [It's] letting us know where energy comes from"), responded that they didn't know how to read the visual or had a negative response (e.g. "It means nothing").

Figure 2 highlights four of the visual stimuli used in this study and the terms participants used to label them. These examples illustrate that participants had a wide range of labels for the visuals presented to them, and very few would be considered accurate or equivalent terms. This suggests that while participants commonly encounter these visuals in their daily lives, they are unfamiliar with their technical names. Furthermore, the preponderance of different labels the participants assigned to the visuals, even basic ones, suggests that more raining and better communication is required before the general public uses a common language to refer to different visualization types.

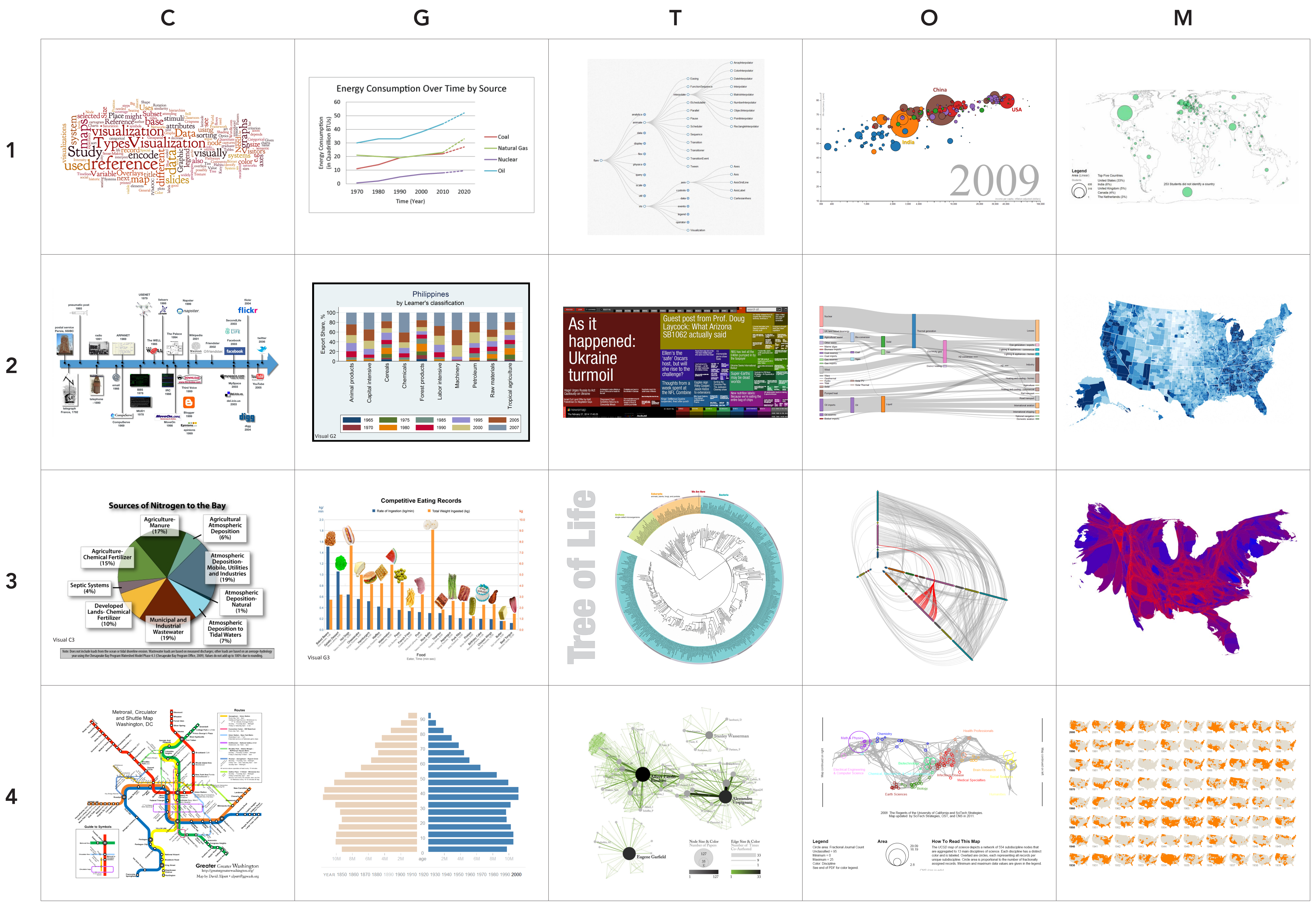


Figure 1: Thumbnail versions of all 20 visualizations used in the study. Visualizations were selected from textbooks, news, widely used online visualization libraries like http://d3js.org, or designed using the Sci2 Tool. Each row represents one set, e.g., subject A may see C1, G1, T1, O1 M1, subject B may see C2, G2, T2, O2 M2, etc. The visualizations are of type chart (C1, C3), graph (G1, O1, C2, G2, G3, G4), map (M1, M2, M3, C4, M4), and network layout (T1, T2, O3, T3, O3, T4, O4).

SIGNIFICANCE OF THE STUDY

Collectively, the findings show a very low data visualization literacy for youth and adults. This result runs contrary to the belief that most U.S. citizens can read visualizations shown in newspapers, textbooks, or encyclopedias and has a number of serious implications.

First, in the information age, the ability to read and write data visualizations might be as important as being able to read and write. Hence, it seems highly desirable to teach data visualization reading and writing skills in formal and informal education settings.

Second, visualization designers in journalism, government, industry, and academia interested in using visualizations for more than "eye candy" should provide information on how to read visualizations. A simple note below a visualization telling: "If you do not know how to read and interpret this visualization please go to p. xx in appendix." and a more detailed description and exemplification of this particular visualization in the appendix should suffice.

Third, we are interested to replicate the study in different environments and countries. Therefore, all study materials have been made available online at http://cns.iu.edu/2015-VisLit.html. Please feel free to contact the authors for more information. We are interested to collaborate on meta-analyses that compare data from different settings/countries to understand the impact of different education systems on data visualization literacy.

REFERENCES

Börner, Katy, and Ted Polley. 2014. *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

Börner, Katy. 2015. *Atlas of Knowledge: Anyone Can Map*. Cambridge, MA: The MIT Press.

Barabási, A. (2011). The network takeover. *Nature Physics*. Vol. 8, No. 1. London: Nature Publishing Group. 14-16.

Glaser, B. G. (1965). The Constant Comparative Method for Qualitative Analysis. *Social Problems*. Vol. 12, No. 4. Berkeley, CA: University of California Press. 436-445.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Report for the McKinsey Global Institute. New York: McKinsey & Company.

TABLES & FIGURES

Table 1: Percentage of respondents indicating Where they have previously seen visuals, by age. These results exclude missing responses and answers for "I don't know."

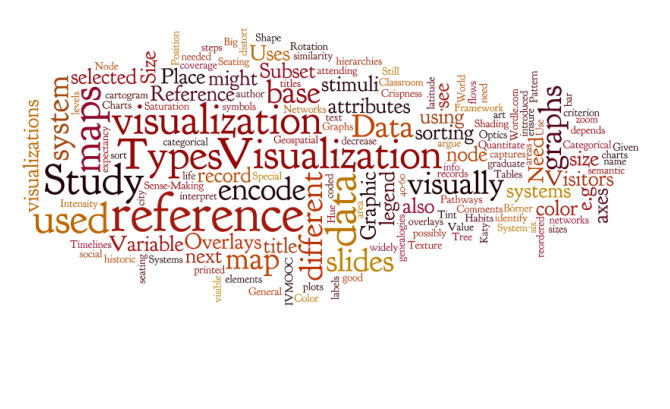
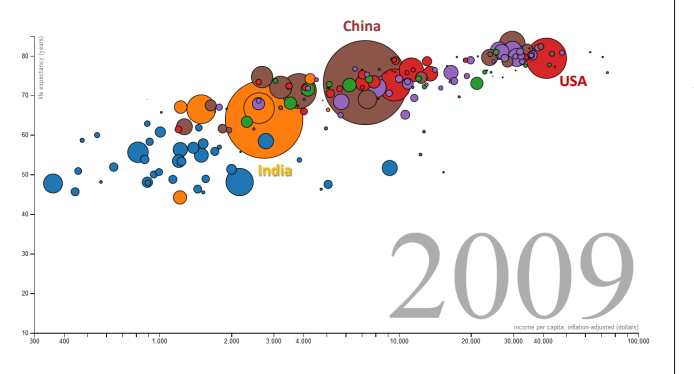
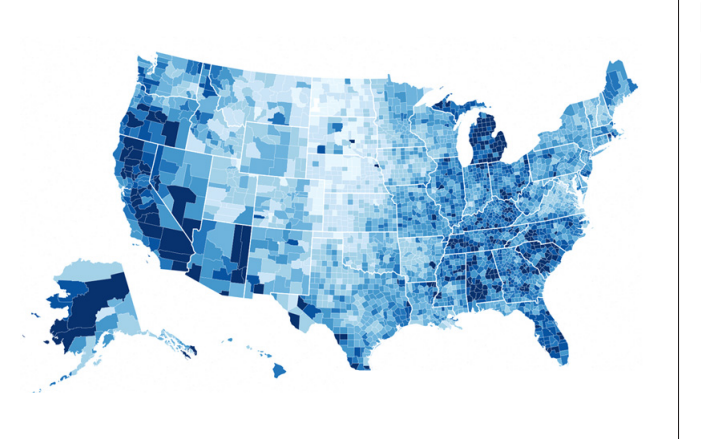
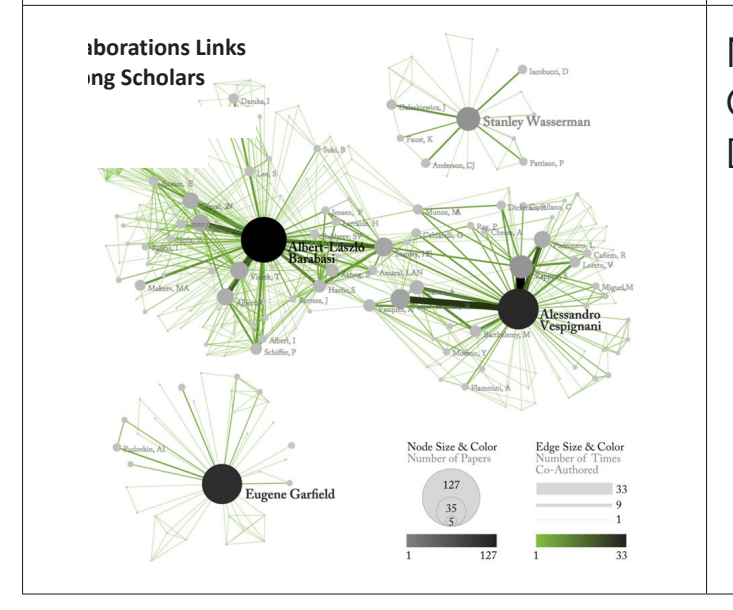
Location Code	Example Quotes	Adult	Youth
Advertisement	Seen it in ads	2.8	1.0
Art	Artwork	1.1	0.7
Books	History Book	9.5	8.7
Entertainment	American Idol	3.6	4.2
Everyday Life, Business, Politics	Weather, Stock Market, Voting	11.6	3.9
Internet	Websites	8.1	6.5
Magazines and Brochures	Magazines	5.6	1.5
Maps	Anywhere a map would be	4.2	7.0
Medical	Doctor's office, Psychology tests	1.7	1.2
Museum	MOMA in NYC	0.7	1.5
News	Newspapers	9.8	2.5
Posters and Presentations	Business presentation	4.2	0.7
Public Spaces	I think I've seen this at the mall	4.1	4.5
School	Classes- history	19.9	51.0
Science and Research	Scientific Journals	3.1	0.5
Technology Devices	iPod	0.8	2.0
Work	At my job in marketing	7.7	0.2
Other		1.5	2.2
Total Coded Elements		**715**	**402**

Table 2: Percentage of respondents indicating How they would approach reading the presented visual.

Response Rate by Age (%)

Response Code	Adult	Youth	Total Response (n)
Key Visual Features	78.4	83.6	1956
Axes	3.9	3.8	94
Color	14.7	17.5	386
Generic "Visual"	3.7	2.9	81
Legend or Scales	8.8	8.2	208
Lines	4.6	4.8	114
Location or Orientation of Features	3.2	3.5	80
Objects	10.9	12.8	283
Size and Quantity of Items	13.4	12.1	311
Time	4.1	5.4	112
Titles and Labels	3.8	4.5	99
Visual is Missing Pieces	1.3	0.5	23
Words and Numbers	6.2	7.7	165
What the Visual Communicates	21.6	16.4	469
General "information"	1.1	0.9	24
Trends	2.7	1.2	51
Groupings	1.7	1.1	35
Comparisons or Relationships	16.0	13.2	359
Total Coded Elements	**1382**	**1043**	**2425**

Figure 2: Matrix of visual representations, labels and participant responses. Frequency counts for the instances are included in parentheses. Please see http://cns.iu.edu/2015-VisLit.html for complete set of 20 visualizations.

Visual	Visualization Type: Technical Label	Equivalent Phrases	Examples of Related Phrases	Examples of Unrelated Phrases
	Chart::Word cloud/ tag cloud	Wordle (6) Word Chart (2) Word Cloud (1) Tag Cloud (1)	Info Graphic (1) Visual Thesaurus (1) Word Cluster (1) Word Art (1) Word Scramble (1) Word Painting (1) Word Blob (2) Word Map (3) Word Pattern (1)	Messaging (2) Mess (1) Randomness (1) Mixed Up Graph (1) Advertising (1) Art (1) Collage (1) Educational Chart to Find Words (1) Informatics (1)
	Graph::Scatter graph with proportional symbol coding	Scatterplot (3) Bubble Graph (5) Rainbow Scatter Graph (2) Bubble Chart (4) Weighted Scatter Plot Graph (1)	Population Graph (1) Graph (9) Graph to Show Income and Life (1) Dot Chart (2)	Chart (2) Paint Dots (1) Countries of the World (1) Open Map (1) Circley Graph (1) Crumb Graph (2)
	Map::Choropleth Map	Map of U.S. (4) Density Distribution Map (1) Density Map (2) Map with Concentrations (1)	Graph (4) Map (12) Population Chart (2) Population Map (1) Percentage Map (2) Saturation Concentration (1) Map Density (1) Hue Map (1)	
	Network Graph::Force-Directed Layout	Network Graph (1) Relationships Graph (1)	Connection Chart (2) Networking Map (1) Relational Graph (1)	Scribbles (1) Spider Web Diagram (2) Spider Web Population Density (2) Star Graph (1) Web Charts (1) Web Graph (1) Iris Chart (2) Graph (2) Line Graph (1)