PIIMPAPER

DATA VISUALIZATION DESIGN GUIDELINES

JIHOON KANG, PIIM, THE NEW SCHOOL
ILLUSTRATIONS: ANN YI, JIHOON KANG

DESIGN AND VISUALIZATION BEST PRACTICES
FOR BIG DATA (FA8750-12-2-0325)

THE NEW SCHOOL

PARSONS INSTITUTE
FOR INFORMATION MAPPING

68 5th Avenue
Room 200
New York, NY 10011

T: 212 229 6825
F: 212 414 4031
piim.newschool.edu

INTRODUCTION

The effective visualization of data benefits users and consumers through an improved ability of data analysis for decision-making. Different data visualizations may be necessary depending on the user's objectives. Mathematicians, engineers, and artists may choose from a number of different data visualization methods. These data visualization methods for big data include the broader categories of maps, charts, diagrams, graphs, and tables.

The core objective of visualizing data is to enhance the communication between the presenter and the viewer. It is the presenter's goal to help the viewer process complex information effectively and accurately through well-crafted data visualizations. However, a well-crafted data visualization requires more than just a familiarization with the data; it also involves graphic design acumen and decisions affecting the use of color, graphic treatment, typography, and composition. Of course, the development of graphic design skills does not happen in one day. However, the application of some very fundamental graphic design tips with respect to data visualizations can bring obvious advantages to the rendering. The purpose of this paper is to provide some fundamental graphic design guidelines and tips to those who seek to improve their own visualizations of data. This document covers the common core types of data visualizations, and provides examples of both good and bad practices.

I serve as the creative lead at the Parsons Institute for Information Mapping, of The New School. Working with information designers, software developers, and researchers, much of my job involves providing art direction for the creation of data visualizations and Graphic User Interfaces (GUI). This paper reflects the way I share my thoughts with those who visualize data. It is not always easy to improve the graphic quality (e.g., typography, color, composition) for graphs and charts without applying commensurate graphic design skills. However, hopefully it will make your job much easier if you are familiar with some of the fundamentals of graphic design. The fundamental areas covered in this paper are where you should pay attention when creating graphs and charts. If you do, you will create data visualizations that are more visually appealing and communicate information better.

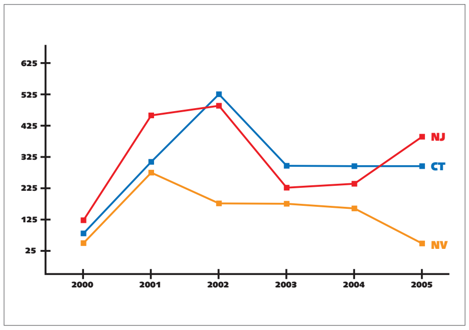


FIGURE 1: *A line graph drawn with axes and plots treated with the same line weight.*

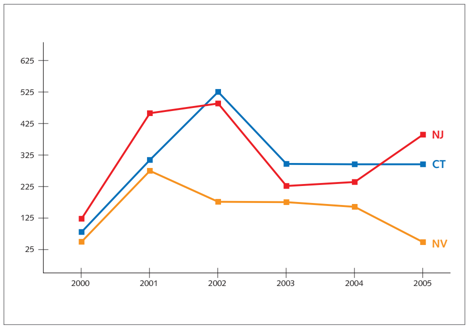


FIGURE 2: *A line graph drawn with two different line weights.*

LINE GRAPH

A line graph is one of the most popular and common data visualization methods. The use of a line graph is particularly effective for visualizing trends for allowing comparisons among multiple variables. Generally, a traditional two-dimensional (2D) line graph employs a vertical axis, horizontal axis, plots, ticks, labels, axes descriptions, legend, title, and underlying grid. It is up to the presenter to choose which of these to include. Placing a grid behind the plots can assist the viewer by providing additional reference and context. However, it may be distracting, overwhelming, or unnecessary in many cases.

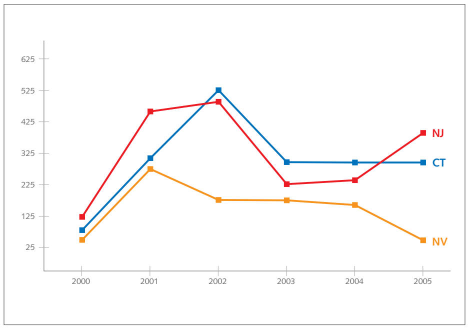


FIGURE 3: *Axes with the lower contrast in value shown here.*

AXES AND SERIES

Axes and series on a line graph share visual similarity, as both of them are rendered as lines. I have seen many line graphs with axes and series lines competing simply because of the graphic treatment applied to these lines (see FIGURE 1). It is important to apply the line style based upon the significance of each within context. What is most important between axes and series? A series represents the actual information pulled from the data, whereas the axes represent variables. It is a logical design decision to put more emphasis on the series while understating the axes. To do this we need to create visual contrast between these two; there are a few ways to do this: line weight, color (value if grayscale is used), or opacity (if applicable). FIGURE 2 is an example showing the contrast between the axes and series; you can see the series are more prominent than the axes providing references.

Now, you may have noticed that there is high contrast in value between the axes and white background. We can lower this contrast by lowering the intensity of grayscale (see FIGURE 3). If applicable, you may choose to lower the opacity instead as this creates a similar effect.

When multiple series are displayed, make sure to pick colors that the eyes can easily distinguish between. Create a set of color swatches and juxtapose them to test the range. Adjust value (brightness or luminosity), temperature (warm and cool), and saturation (color density). If they are not distinctive from each other, adjust the series. Also, make sure the series can be read comfortably on the plot area by taking the color and tonality of the background into consideration. For example, a yellow line on a white background or a violet line on a background are not very visible.

Viewers do not find it readable deciphering too many series on one graph. There is not a maximum number of series per graph; it depends on the size of the graph, nature of variables, display medium (e.g., computer monitor, tablet, mobile phone, ink on paper) and other factors. The best solution is to test samples for readability. If you find it hard to comprehend the information because of the large number of series, then come up with alternatives, such as limiting the number of series, letting the user choose the series for display, grouping the series, or making modifications and seeing what works.

DOTS ON SERIES

We often see dots on each series line. Normally, these dots represent the actual data values, and the segment between two dots helps the viewer see the trend. Dots are useful if it is your intent to show the actual data value clearly. Dots can be omitted if it is not important to display the specific data points; in fact, the graph will look cleaner without the dots if they can be omitted.

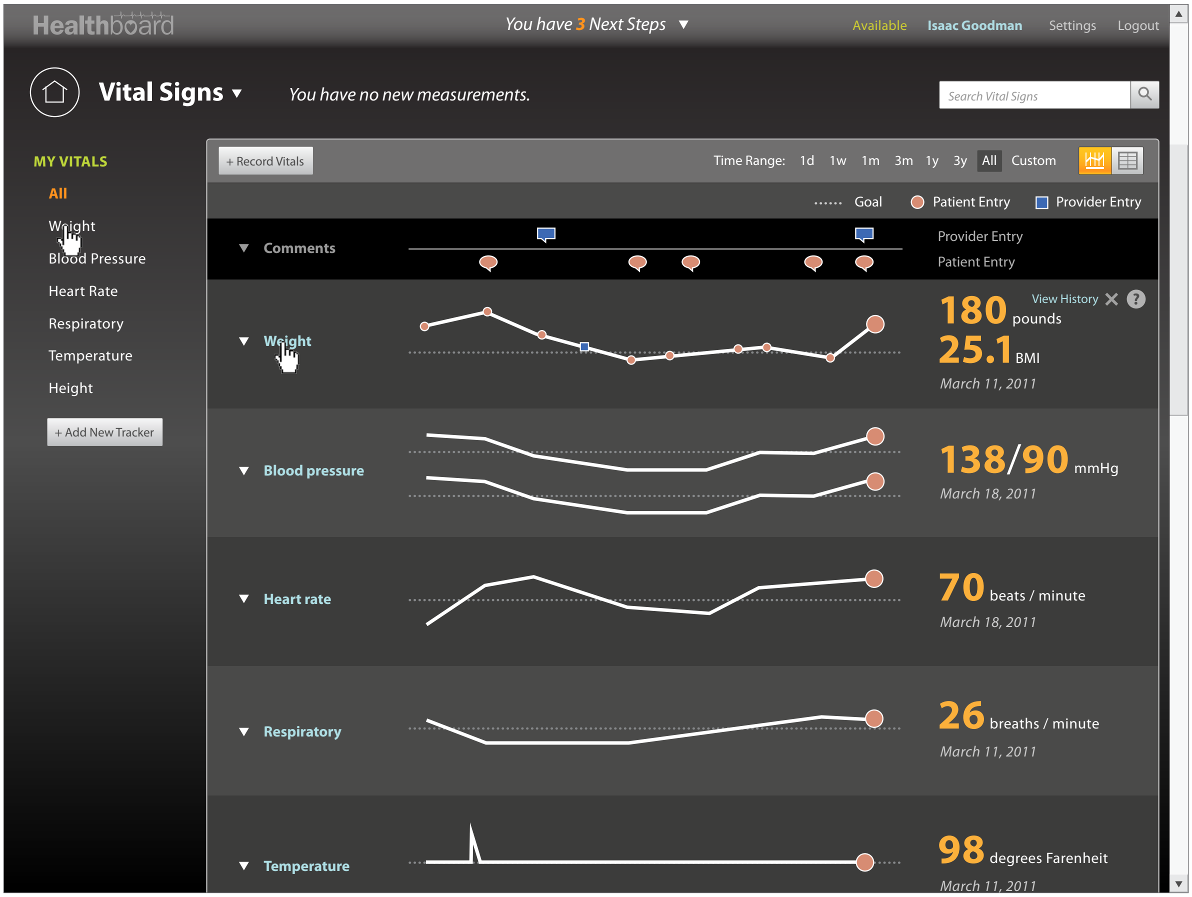


FIGURE 4: *A line graph from the Healthboard user interface that has two different symbols representing the data sources.*

Dots can also help the viewer differentiate series lines, in addition to color, when different symbols are utilized (see FIGURE 4). Make sure to use simple shapes that are distinctive from each other when you select the symbols. The symbols themselves can also carry specific information as a variable. FIGURE 4 is a screen shot of Healthboard designed by PIIM.[1] This line graph uses two symbols that are placed on a single series for patient's weight (lb). The squares represent the data entered by the healthcare professionals and the circles represent what the patient has reported. So, use dots when it is appropriate for your presentation, but don't feel obligated to add them only because you have seen them elsewhere; use the fewest elements to convey the most intelligence.

TICKS AND GRID

Ticks are short marks on each axis. They indicate the value of each variable. Make sure they do not compete with axes or series lines. Having too many ticks can overwhelm the viewer; having too few can hinder user's comprehension. As with the dots on the series line, ticks can be omitted for certain cases. Ticks are often paired with labels (see FIGURE 5). It is important to create the visual connection between the tick and its label through graphic treatment and arrangement. Some ticks can look more prominent than others when you want to emphasize certain milestones. For example, the treatment for the starting point of each year differs from other months on the time variable of FIGURE 5.

A grid can improve readability when it is used appropriately. It is particularly helpful for a large plot area where points and segments are far from the axes. You may emphasize certain grid lines as needed in a familiar manner to the way certain ticks are emphasized. A grid is a background element; hence it should not compete with the foreground elements. Avoid thick lines and intense colors that can overpower the main characters. A grid can also be replaced by references when appropriate (see FIGURE 6).

LEGEND

A legend must be present unless you are presenting only a few series that the viewer may already be generally familiar with, even so there must be clearly labeled within the plot area. It should include definitions of each graphic element (such as meaning of color and symbols). Because the legend does not present the actual information, it needs to be clearly isolated from the plot. Placing the legend outside of the plot would be the safest design decision. If you decide to place the legend within the plot, provide distinction through the use of a frame. If you have two sets of line graphs that share the same legend, find a location for the legend to let the viewer know that the legend belongs to both graphs. If you are presenting graphs on small screens where there is inadequate space to place the legend, consider adding "Show Legend" and "Hide Legend" buttons.

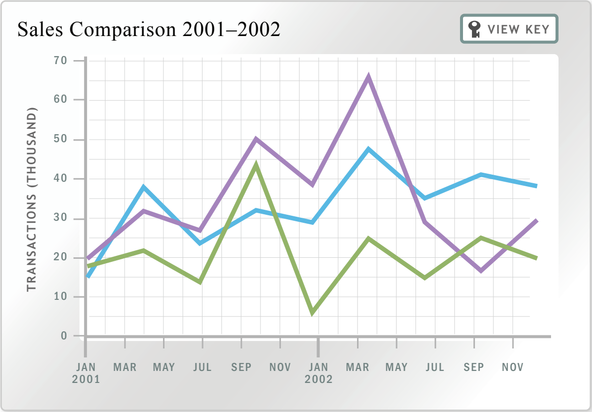


FIGURE 5: *A line graph with ticks, labels, and grid.*

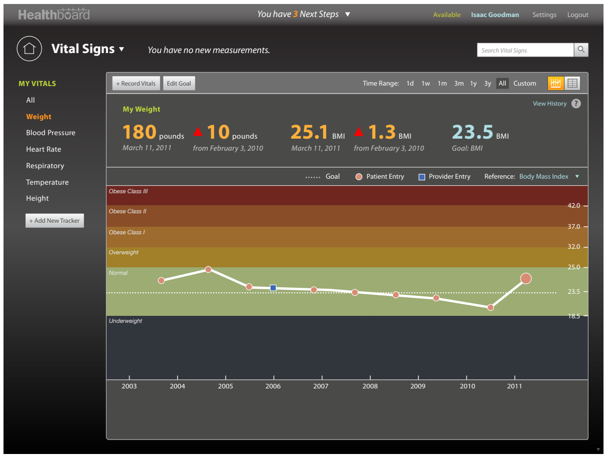


FIGURE 6: *A line graph from the Healthboard user interface that combines both series and reference*[2]

TITLE, LABELS, AND DESCRIPTIONS

A line graph, as with many other graphs and charts, include textual elements such as a title, labels, and descriptions. As each element is a taking different role, these textual elements need to be organized typographically. The organization logic is rather simple: think of the hierarchy of information. What is most important? What is less important? What is even less important? What order do you want to present the information? Let's say you have a line graph including: the title of the graph, labels for tick marks, and descriptions for both x- and y-axes. Let's also say your intent is to guide the viewer to read the title first—so they understand what the subject matter is, then move their eyes to descriptions for both axes so they know what the variables are, and then to finally view the series lines while comparing them with the ticks and labels on the axes, typography is the right tool to visualize such a hierarchy of information. Here are some typographical ingredients used to visualize the hierarchy of information: position, scale, typeface, and weight. You may mix all these characteristics or choose one or two of these attributes to differentiate the text groups. Applying typography to visualize the information hierarchy is very common and effective in the page design. When we read a magazine article, we generally read the header or images first, decks and captions next, then the body text (FIGURE 7).[3] Most of time we are not even aware of the page structure and order because these pages are carefully crafted by the designers who had thought about the hierarchy of information and applied typographic principles to enhance readability. The same kind of readability matters when presenting graphs.

HISTOGRAM

A histogram is structurally similar to a line graph. You may apply the same kind of graphical resolution regarding ticks, grid, legend (if applicable), title, labels, and descriptions for line graphs presented above. A histogram consists of adjacent rectangles and there are design standards for rendering these rectangles.

OUTLINES AND COLOR

Apply outlines onto each rectangle to for better readability. Without them the viewer may have a difficult time differentiating the intervals represented as rectangles (FIGURE 8). When you apply the outlines, you need to make sure not to over-emphasize the borders. Because the outlines are only used to divide the clustered rectangles, the lines need to be neither too thick nor too colorful. Overpowering outlines will make the graph look busy and decrease readability. Use color variations to divide the bars if you cannot draw outlines onto the rectangles for any reasons.

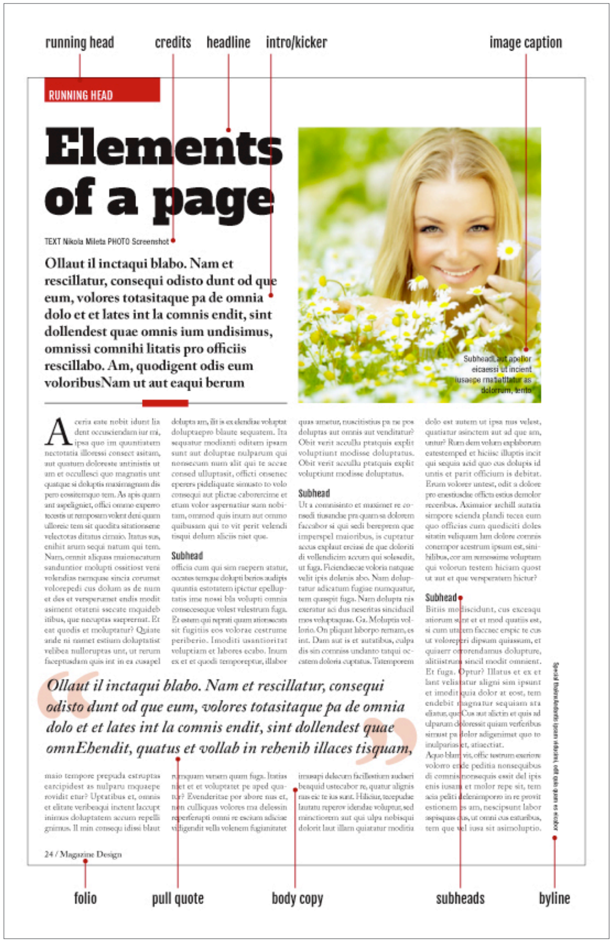


FIGURE 7: *A sample layout introducing elements of a magazine page.*

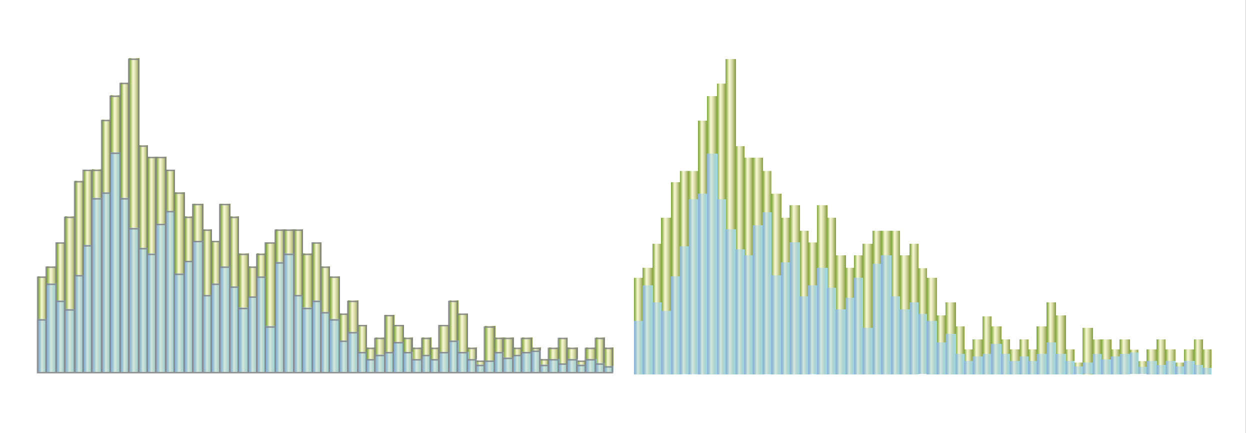


FIGURE 8: *Readability comparison for intervals with and without outlines.*

SCALE

When you present your data with a histogram, the size of rectangles presents the quantity. In order to communicate with your viewers more effectively, you need to carefully determine the height and width of the graph. The graph becomes uncomfortable to view when there are too many, or too few intervals, which determines the width of the graph. Set the width in consideration of the total number of intervals. The height of graph needs to be determined by the values of intervals. Do not leave too much negative space above the greatest interval. If the scale can be adjusted by the quantities within the data the viewer will be able to compare intervals comfortably.

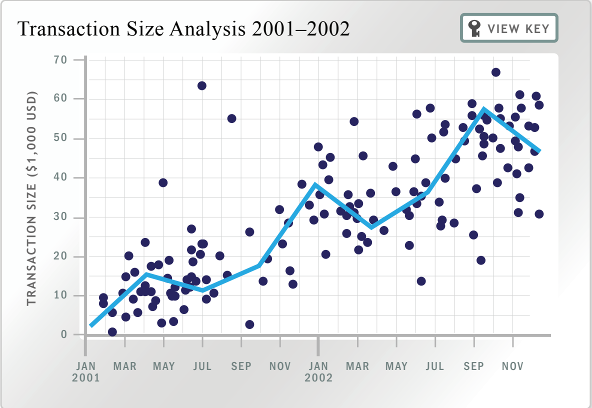


FIGURE 9: *A sample scatter plot.*

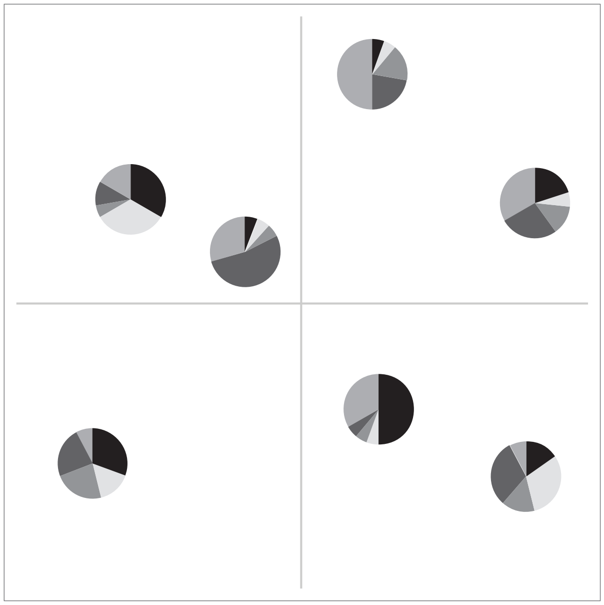


FIGURE 10: *A scatter plot with glyphs for a multivariate data display.*

SCATTER PLOT

DATA POINTS AND VARIABLES

A conventional 2D scatter plot consists of data points and two variables. As with a line graph, it is important to draw these elements based on the hierarchy of information. In most scatter plots, data points carry the most important information, then the variables. Keep the data plots prominent while understating the axes. Use color, weight, and opacity as previously explained (see FIGURE 9).

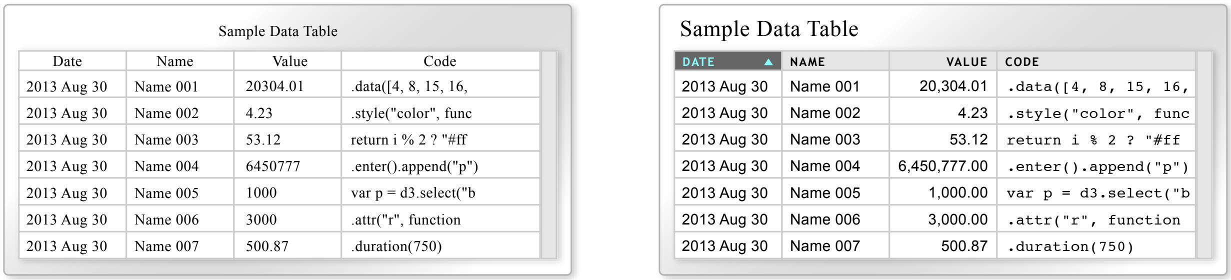
GLYPHS

Some data presenters replace data points with glyphs to display additional information. FIGURE 10 illustrates a scatter plot using pie charts as data points. Be aware that you are asking the viewer to process considerable information at once through fusing two different data display methods together. This might overwhelm viewers who are not familiar with this style. Also, make sure not to display too many glyphs on the plot, as even perceptive viewers can suffer from processing this much data. When glyphs are used, make sure to include the legend accordingly.

TABLE

Never dismiss the value of tables. Although it is not an innovative data display method, it is still one of the most practical and popular ways of displaying data. Because tables display letter and numerals, I have some specific tips on utilizing the basic typography principles to improve your tables.

One common mistake made in tables is textual misalignment. We read from left to right, hence, it is almost always better to use flush left. It will also allow you to align the contents on the left side—this keeps the table neat. For these reasons, I do not recommend the center alignment unless you have special reasons. Do not mix alignments like assigning "centered" to



Sample Data Table

Date	Name	Value	Code
2013 Aug 30	Name 001	20304.01	.data([4, 8, 15, 16,
2013 Aug 30	Name 002	4.23	.style("color", func
2013 Aug 30	Name 003	53.12	return i % 2 ? "#ff
2013 Aug 30	Name 004	6450777	.enter().append("p")
2013 Aug 30	Name 005	1000	var p = d3.select("b
2013 Aug 30	Name 006	3000	.attr("r", function
2013 Aug 30	Name 007	500.87	.duration(750)

Sample Data Table

DATE ▲	NAME	VALUE	CODE
2013 Aug 30	Name 001	20,304.01	.data([4, 8, 15, 16,
2013 Aug 30	Name 002	4.23	.style("color", func
2013 Aug 30	Name 003	53.12	return i % 2 ? "#ff
2013 Aug 30	Name 004	6,450,777.00	.enter().append("p")
2013 Aug 30	Name 005	1,000.00	var p = d3.select("b
2013 Aug 30	Name 006	3,000.00	.attr("r", function
2013 Aug 30	Name 007	500.87	.duration(750)

FIGURE 11: *Tabular data display samples. The model on the right side is rendered with typographical considerations.*

the top row and “flush left” to the rest. They should be aligned consistently as they are related. Textual alignment to conveys this relationship well.

“Flush right” is not as common as “flush left” and “centered.” Although not recommended for aligning letters, it is appropriate for numerals. Take a look at the two examples of aligning numerals (FIGURE 11). The example on the left with the numerals aligned to right is more readable than the one on right. It also gets easier to compare values on multiple rows. Consider using commas to group digits when displaying large numbers and align the decimal marks.

TYPEFACE

I am not going to guide you as to what the precisely appropriate typeface is when choosing to create a table or graphs. That is always debatable and often depends on the purpose of presentation or presenter’s personal preferences, as well as what users have become familiar with. Your choice is good as long as all text can be read comfortably, and the viewer is not “aware” of your font choice—because everything seems “normal.” If you are not confident in working with type, choose legible sans serif typefaces like Helvetica or Arial, which are also part of the system fonts and have families that allow for emphasis. Be careful in using script fonts, black letters, or other display fonts; these are generally less legible.

Working with monospaced typefaces such as Courier or Lucida Console is a bit tricky. Generally I do not recommend monospaced typefaces for displaying data because they tend to take up more space and may slow down the reading pace. Usually, a proportional typeface brings better readability. A monospaced typeface, however, does work well for displaying code. It is crucial to clearly recognize each character when working with code and makes it easier for the viewer to differentiate similar characters, such as the upper case “O” and the number “0,” as well as the lowercase “l” and uppercase “I.” If you need to display the actual code on a table, consider using a monospaced font. Use a proportional non-decorative font for the rest.

There are a few more tips that I want to add specific to interactive tables. Normally column heads are displayed on the first row followed by data cells. If you are allowing the viewer to sort content within a column, make sure you add a visual clue that these column heads are clickable. If the data is sorted after one header cell is clicked on, also add a visual clue that this is the active header cell. Use an arrowhead graphic to point upward or downward if you wish to allow the viewer to change the sorting order from the ascending to descending, or vice versa.

NODE-AND-LINK

A node-and-link diagram is commonly used to visualize networks. A node-and-link can be a simple diagram with uniform nodes and lines, or it can include millions of variable nodes and lines (see FIGURE 12). It is also highly scalable; you can start from the global level network and zoom into the personal level. Obviously, there are multiple tiers between these two levels. With a node-and-link diagram you can see the *trees*, zoom out to see the *forest*, from there to an entire *universe*. It is also flexible; nodes can be georeferenced or positioned like tree branches within a relational contest to emphasize their hierarchies (see FIGURE 13).

ADDING VARIABLES

A simple node-and-link diagram communicates only one or two variables. However, you may scale up these variables as needed by the way your render nodes and lines. You can present various meanings with the nodes: color, position, shape, size, and symbols (or glyphs) (see FIGURE 14). For example, you initiate the diagram with dots and lines to show how the data are related. Then use color to show class, size for quantity, position for geographical reference, and shapes for trend. You can create different meanings with the line quality: weight (thickness), color, style (solid vs. dotted), and shape (straight vs. curved) (see FIGURE 15). Be aware that the diagram will turn less intuitive as you add more variables where viewers need to process more and more information.

CHOROPLETH MAP

A choropleth map, also known as a thematic map, is a common method for displaying the statistical distribution over polygons within a geographic display. A choropleth map combines predefined regions (e.g., counties, states, counties, zip codes, census block groups) represented through color, shades, or patterns. This method is effective when the presenter intends to draw a comparison between regions within the same variable or a comparison between variables of the same (or different) region (e.g., population density in 2000 vs. 2010).

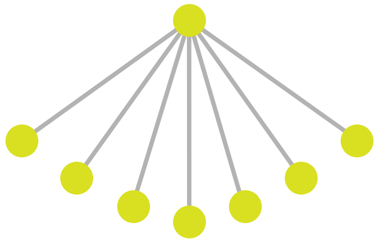


FIGURE 12: *A node-and-link diagram with uniform nodes and links.*

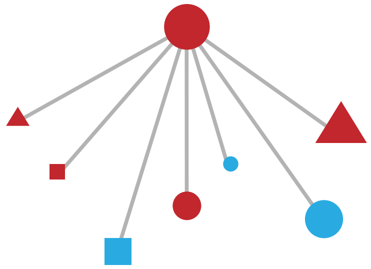


FIGURE 14: *A multivariate node-and-link diagram with variations of nodes.*

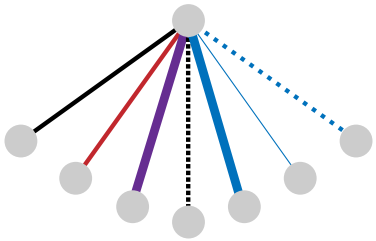


FIGURE 15: *A multivariate node-and-link diagram with variations of links.*

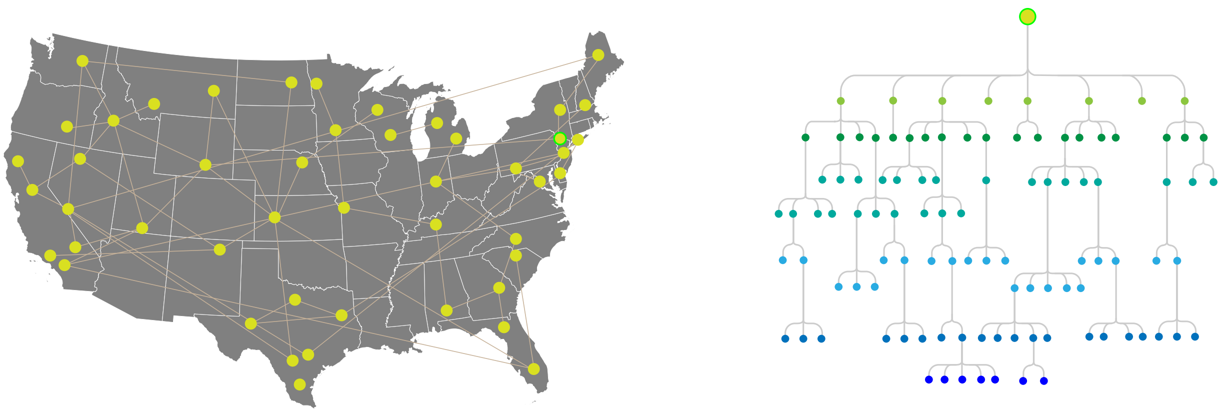


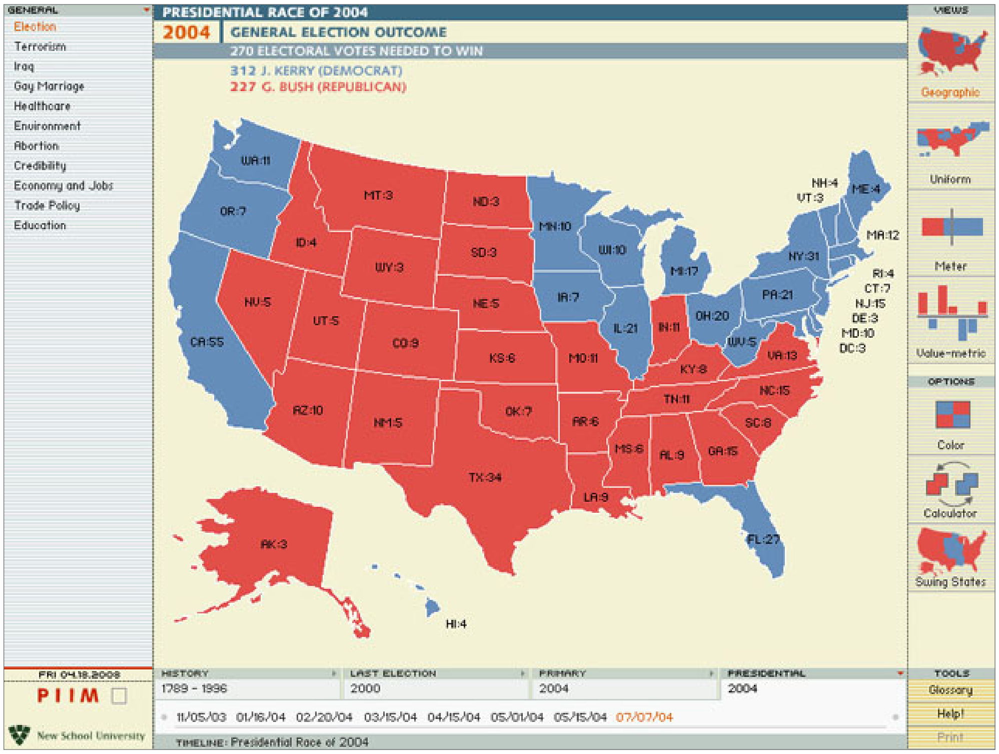
FIGURE 13: *Georeferenced model vs. hierarchical model.*

COLOR

Since the data on a choropleth map is generally represented through color, selecting the right color scheme is crucial. The viewer can easily be frustrated if the statistical data is not rendered properly. You can render the map using multiple colors which can also be combined with shades; single color with multiple shades (monochromatic scheme); or patterns which can be used alone (e.g., lines, dots, hatching, cross-hatching); or, combined with a colored theme to provide additional information. Apply a monochromatic scheme if you're going to present percentages (or intensity) and make sure to synch the direction of numeric intensity with the color intensity. Limit the number of color categories you use as our eyes cannot process extensive color variations. Because you want the viewer to be able to differentiate the categories without deeper concentration of visual focus, make sure to apply enough contrast within the range.

A bi-polar color scheme (using two contrasting hues) is appropriate to display when you want the color to illustrate two opposite or contrasting variables (e.g., rivals). A classic example would be election maps. FIGURE 16 is the Electoral College map for the 2004 US presidential election using a bi-polar color scheme to represent the Democratic Party (in blue) and the Republican Party (in red).[4] The color choice for this map is effective because it uses only two colors matching the identity of each variable. When selecting colors for a chropleth map, make sure the two colors are distinctive. Such distinction can be made through chromatic (hue) distance (e.g., blue vs. orange), value (luminosity), temperature (warm vs. cool), etc. In addition, use a color scheme that viewers with color vision deficiency (color blindness) can easily decipher. For example, pairing red and green should be avoided. In general similar values will also cause a challenge.

You may use the bi-polar method with additional color palettes to illustrate the opposite variables with intensity. FIGURE 17 is a choropleth map representing New York Yankees and Boston Red Sox fans in Connecticut.[5] Blue and bluer regions have more Yankees fans; red and redder regions have more Red



PRESIDENTIAL RACE OF 2004

2004 GENERAL ELECTION OUTCOME

270 ELECTORAL VOTES NEEDED TO WIN

312 J. KERRY (DEMOCRAT)
227 G. BUSH (REPUBLICAN)

TIMELINE: Presidential Race of 2004

FIGURE 16:
A screen of the Public Opinion Tool by the Parsons Institute for Information Mapping using a bi-polar color scheme to represent the presidential election outcomes.

Sox fans. Regions with blended fans appearing in the central part of the state are colored with purple variations. Most of us can easily differentiate blue and red because these are contrasting hues with opposite color temperatures; red is warm and blue is cool. However, it is far less easy to differentiate the purple variations blended in-between red and blue.

There are a few ways to resolve this. One is to reduce the number of categories. This map can be colored with 5 categories of blue (100% Yankees Fans), blue-purple (75% Yankees), purple (50%), red-purple (75% Red Sox), red (100% Red Sox), instead of 11 categories. If keeping all 11 categories is a necessity, then I recommend blending each color (blue and red) with white. This way the regions with 100% Yankees fans will have the same blue and the regions with 60% Yankees fans will be in light blue. 60% Yankees fans (light blue) and 60% Red Sox fans (pink) will be still distinguishable through the color temperature because you are only modifying the chroma (color saturation) while keeping the same color temperature (FIGURE 18).

OUTLINING

There are a few design tips for outlining regions on a choropleth map: color choice and thickness. Outlining regions such as states, counties, and zip codes is a useful reference that defines the boundaries of each region. However, the outlines should not compete with the colored regions which represent statistical data. As applying bright color can create such interference, I suggest using calm and neutral color for outlines. Thickness or other line quality renditions can be applied discretely; the line thickness should not overpower the colored regions that carry the main information. The thinnest outline that allows the viewer to see the boundaries comfortably is appropriate. You may then use different weights to present the hierarchical relationship between the segments. For example, when you display both states and counties on a map, apply thicker lines for states and thinner lines for counties to draw the hierarchy. If you have an interactive map for users to select regions, increase the line weight to indicate that the selection is now active. For this case it also makes sense to apply a brighter color to the outline.

LAYERING

It is not uncommon to combine a chropleth map with other layers such as aerial and satellite imagery, street maps, labels, symbols, and so forth. Most likely you will make the map less intuitive and less readable by adding more layers above or below the layer of a chropleth map. You need to carefully assess the need of each layer before adding them. For example, it may not be necessary to show street maps when you present the statistics of baseball fans. If it is debatable, consider allowing the user to turn on and off these reference layers on and off (applicable for an interactive map). When labels and symbols are combined, they are usually overlaid on top of a chropleth map. Once again, references should serve as references and they should not compete with the key data. When you notice text-based contents like labels are competing with the data layer, make them inconspicuous by changing typeface, font size, weight, color, and opacity, yet be sure to keep them legible. The same rule should be applied to the symbols; they should not compete with the key data. Pay attention to all visual elements like scale, color, opacity, and patterns. If the text and symbols are difficult to read because it is blended with what is underneath, apply an outline or drop-shadow.

LEGEND

The meaning of each class should be clearly identified through the legend. Position all color swatches and symbols used on the map and place labels accordingly. The legend should be isolated from the map. It is safer to place them on the bottom or side of the map. If you have to place the legend within the map boundaries, make sure it is not obstructing objects carrying information. You may have a collapsible or expandable legend for an interactive map as long as the user is able interact with it.

GEOGRAPHICAL SCALE AND ZOOM

When you display multiple geographical scales (e.g., states » counties » census block groups), consider breaking it down to multiple zoom levels and display regions according to the appropriate scale. For example, display the state-level distribution when the map is viewed from the national level. When the user selects a specific state, it will first zoom in to the state and display the county-level distribution. If one county is selected, it will zoom in to the census-block-level (or other appropriate types).

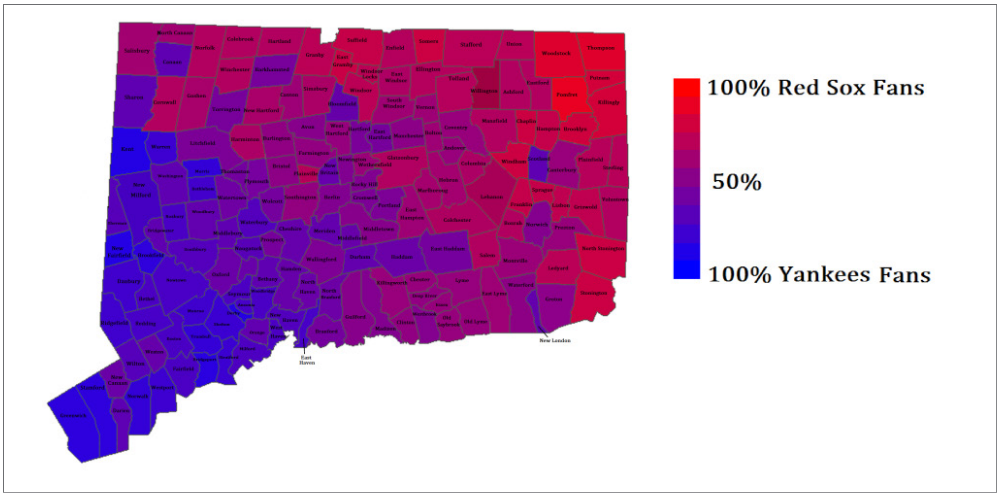
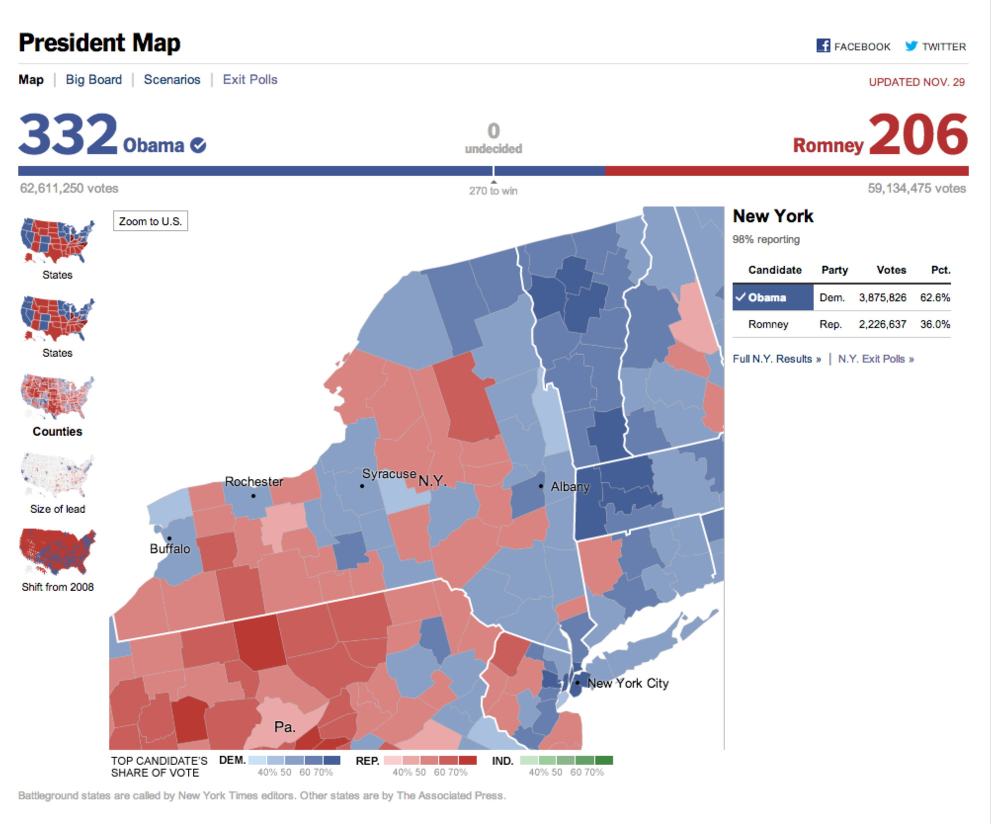


FIGURE 17: *A bi-polar choropleth map representing the distribution of the New York Yankees and Boston Red Sox fans in Connecticut.*



President Map

FACEBOOK TWITTER

Map | Big Board | Scenarios | Exit Polls

UPDATED NOV. 29

332 Obama ✓ 0 undecided Romney **206**

62,611,250 votes 270 to win 59,134,475 votes

New York

98% reporting

Candidate	Party	Votes	Pct.
✓ Obama	Dem.	3,875,826	62.6%
Romney	Rep.	2,226,637	36.0%

Full N.Y. Results » | N.Y. Exit Polls »

Battleground states are called by New York Times editors. Other states are by The Associated Press.

FIGURE 18: *A screen of President Map by The New York Times displaying multiple intensities with two visually contrasting color groups.*

NOTES

[1] The Parsons Institute for Information Mapping, The New School. “Healthboard.” Web Application. http://piim.newschool.edu/healthboard/

[2] Ibid

[3] “Elements of a Magazine.” *Magazine Designing*, March 26, 2013. http://www.magazinedesigning.com/magazine-page-elements/

[4] The Parsons Institute for Information Mapping, The New School. “Public Opinion Tool.” Web Application, 2004. http://piim.newschool.edu/tools/votingtool/

[5] Blatt, Ben. “Finding the True Border Between Yankee and Red Sox Nation Using Facebook Data.” *The Harvard Sports Analysis Collective,* August 17, 2012. http://harvardsportsanalysis.wordpress.com/2012/08/17/finding-the-true-border-between-yankee-and-red-sox-nation-using-facebook-data/

[6] The New York Times, *“President Map.”* November 29, 2012 (late update). http://elections.nytimes.com/2012/results/president

This work was supported by the Defense Advanced Research Projects Agency (DARPA) Award No. FA8750-12-2-0325. Any design guidance, opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not necessary reflect the views of DARPA.