Visualization for Sociocultural Signature Detection

Ronald D. Fricker, Jr., Samuel E. Buttrey, and William Evans
Naval Postgraduate School

May 31, 2013

1 Introduction

When attempting to discover, distinguish and locate operationally relevant sociocultural signatures, there is no better tool than the human eye applied to appropriate displays of relevant data. Indeed, the human brain is generally far better than computer-based algorithms at finding patterns in data, particularly patterns that are *a priori* ill specified or unknown. One need look no further than the captchas[1] used on webpage login screens for confirmation of this.

Done correctly, data visualization can facilitate:

- *Insight*: Data visualization can provide insights, particularly relationships among variables and trends over time, that may not be apparent otherwise. With large and/or complicated data, appropriate visualizations can make the data easier to understand.
- *Exploration*: With interactive visualization methods, data can assessed from multiple perspectives and contexts; for sociocultural experts, interactive methods allow them to bring their expertise to bear in the visualization and ultimate interpretation of the data.
- *Impact*: "Use a picture. It's worth a thousand words."[2] Data visualization is a powerful way to communicate information, relationships, and even the stories present in data.

Data visualization can reveal relationships in the data that are simply not apparent via summary statistics. The canonical example is Anscombe's data plotted in Figure 1 (Anscombe, 1973). Visually these four sets of data clearly show very different relationships between the x and y variables, yet:

- the means and standard deviations of each of the x variables are exactly the same;
- the means and standard deviations of the y variables are also the same;
- the correlations between x and y in each of the four cases are the same; and,
- even the regression fits are the same.

In the absence of plotting the data, by just looking at some summary statistics one could be completely misled into thinking that there is little difference in the underlying phenomena.

[1] Captcha is short for Completely Automated Public Turing Test To Tell Computers and Humans Apart. See `http://www.captcha.net`.

[2] Arthur Brisbane, "Speakers Give Sound Advice," *Syracuse Post Standard* (page 18), March 28, 1911.

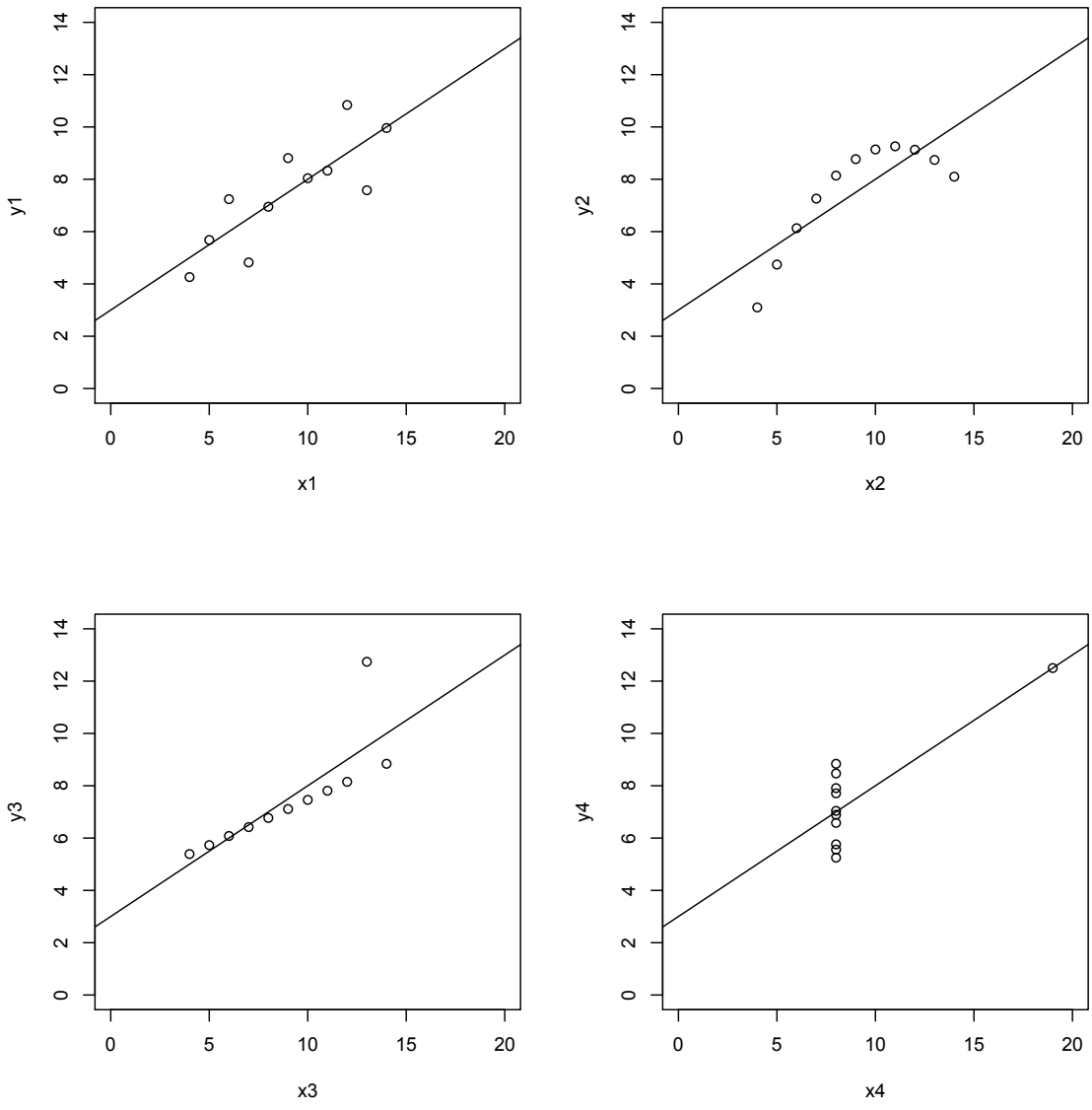

Figure 1: Plots of Anscombe's data, where all the summary statistics match, the regression fits (as shown by the lines) are the same, and yet the relationships between the x and y variables are clearly different in all four cases.

In terms of communication and collaboration, Sviokla (2009) says visualization has the following three benefits:

1. Great visualizations are efficient – they let people look at vast quantities of data quickly.
2. Visualizations can help an analyst or a group achieve more insight into the nature of a problem and discover new understanding.
3. A great visualization can help create a shared view of a situation and align folks on needed actions.

As Henry Hubbard said in Brintion's 1939 text, *Graphic Presentation*,

> There is a magic in graphs. The profile of a curve reveals in a flash a whole situation – the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces (Brinton, 1939).

In summary, there are clear benefits for using visualization, and these benefits increase in direct relation to the size of the data, though even with small data sets such as Anscombe's visualization can provide unique benefits. But unlike Anscombe's data, which is small and simple to visualize via scatterplots, sociocultural data can be large, complicated, and messy. Thus, the real question is how to appropriately visualize such data?

1.1 Visualization for Sociocultural Signature Detection

As defined in the Office of the Secretary of Defense (OSD) Human Social Culture Behavior (HSCB) modeling program, sociocultural signature detection is one of the four program capabilities: (1) Understand, (2) Detect, (3) Forecast, and (4) Mitigate. The Detect capability is defined as "Capabilities to discover, distinguish, and locate operationally relevant sociocultural signatures through the collection, processing, and analysis of sociocultural behavior data" (U.S. Government, 2013). The site goes on to say,

> Once the defining features of the sociocultural setting are understood, the next steps are to develop a persistent capability to detect sociocultural behavior signals of interest amidst complexity and noise, and to harvest data for analysis. This entails capabilities for ISR in the area of sociocultural behavior (referred to here as a "social radar"), with particular focus on the challenges associated with open source data collection. It also requires robust systems for storing and managing that data, and tools enabling timely, dynamic analysis.

Visualization, then, supports "enabling timely, dynamic analysis" both in terms of finding relevant signatures and identifying when known signatures change. The former is largely a retrospective exercise in exploring and modeling existing data to identify sociocultural signatures, while the latter is a prospective exercise in monitoring a given signature to identify if and when it changes. Both types of analysis may require a variety of cross-sectional, temporal, and spatio-temporal visualization analytical methods and visualization techniques for:

- Infrastructure, social, and other types of network data;
- Geographic information system (GIS) and similar types of spatial data;

- Surveys and related types of data;
- Social media and other types of linguistic data.[3]

Good sociocultural signature detection visualization is optimized for the particular application and user. For example, methods for best visualizing networks may be quite different than those for displaying GIS data. Similarly, GIS methods for displaying point data on maps are not appropriate for displaying areal data from surveys. On the other hand, some types of social media data can be effectively displayed as a network while other types of social data will require other visualization methods.

The general question of effective quantitative visualization design has been addressed by Tufte (1997, 2001, 2006), Cleveland (1993, 1994) and others. For example, Tufte says,

> Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should
>
> - show the data
> - induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
> - avoid distorting what the data have to say
> - present many numbers in a small space
> - make large data sets coherent
> - encourage the eye to compare different pieces of data
> - reveal the data at several levels of detail, from the broad overview to the fine structure
> - serve a reasonably clear purpose: description, exploration, tabulation, or decoration
> - be closely integrated with the statistical and verbal descriptions of the data set (Tufte, 2001, p. 13).

Of course, the devil is in the details: How to meet these criteria for a specific sociocultural analytical method in a given scenario for a particular individual analyst or researcher?

1.2 Visualization as Part of Data Exploration

Exploratory data analysis or EDA (Tukey, 1977) is an approach to analysis that focuses summarizing data so that it is easy to understand, often through the use of visual graphics and data displays, and generally without the use of formal statistical models or hypotheses. As described by the National Institute of Standards and Technology (NIST, 2012):

> Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

[3]Visualization is equally important and relevant for the Understand capability and, in fact, there is no clear dividing line between the two capabilities in terms of data visualization. Indeed, many of the visualization techniques are the same; the main distinction is in how the methods are employed. See Chapter [**Understand Visualization**] for additional discussion.

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

NIST goes on to say,

> The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:
>
> - Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots).
> - Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
> - Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

Hence, EDA is a methodology for learning from data, often using visualization. In terms of the discussion in this chapter, the emphasis is on the visual display of data but in support of information exploration for sociocultural signature detection. Some may take issue with either an over reliance on more informal exploratory methods or a failure to use confirmatory methods, however, as Tukey (1977, p. vii) said,

> Once upon a time, [researchers] only explored. Then they learned to confirm exactly–to confirm a few things exactly, each under very specific circumstances. As they emphasized exact confirmation, their techniques inevitably became less flexible. The connection of the most used techniques with past insights has weakened. Anything to which a confirmatory procedure was not explicitly attached was decried as "mere descriptive statistics", no matter how much we have learned from it. ... **Today, exploratory and confirmatory can–and should–proceed side by side.** [Emphasis in the original text.]

EDA, particularly in terms of exploring data to detect sociocultural signatures, is an inherently interactive exercise. As Yi *et al.* (2007) say,

> Interaction is an essential part of Infovis [information visualization], however. Without interaction, an Infovis technique or system becomes a static image or autonomously animated images (e.g., InfoCanvas). While static images clearly have analytic and expressive value, their usefulness becomes more limited as the data set that they represent grows larger with more variables.

1.3 Caution: Apophenia and Pareidolia

Shermer (2008) says, "...our brains are belief engines: evolved pattern-recognition machines that connect the dots and create meaning out of the patterns that we think we see in nature. Sometimes A really is connected to B; sometimes it is not." Thus, while the human brain excels at finding true patterns against noisy backgrounds, it also quite capable of over interpreting pure noise to find nonexistent "patterns." This phenomenon of finding (seemingly) meaningful patterns in random or meaningless data is called *apophenia*, and *pareidolia* is visual apophenia (Hoopes, 2011). To anyone who has spent a summer afternoon looking for animal shapes in the clouds,[4] pareidolia is the act of then believing the shapes represent something other than the coincidental arrangement of water vapor in the upper atmosphere.[5]

As Silver (2012, p. 240) says, "Finding patterns is easy in any kind of data-rich environment... The key is in determining whether the patterns represent noise or signal." In statistical terms, apophenia and pareidolia are a Type I errors: the false positive identification of "patterns" in data. This type of error is a non-trivial consideration in sociocultural analyses, particularly since an underlying assumption of much of the visualization literature seems to be that a true pattern in the data exists and the main question is how to best display that pattern.

In contrast, during exploratory analyses of data in which one is looking for patterns that may or may not be present, the potential for false positives is always there. Under these conditions, it is important to be mindful of the human ability to find fanciful camels in the clouds, and furthermore then readily construct an *a posteriori* rationale explaining the existence and meaning of the camel. But it is equally important to exploit the human facility for pattern recognition, particularly with complex sociocultural signature data. Thus, one challenge with exploratory data visualization and pattern recognition is balancing human and methodological sensitivity for recognizing true patterns against over sensitivity resulting in false positive "patterns."

2 Sociocultural Signature Detection Visualization

Visualization is the science – and the art – of discovering trends and patterns in data. That data can be quantitative or qualitative and it can be temporal, spatial, both, or neither. The appropriate visualization of the data allows the viewer to move from the specific to the general, hopefully gaining some insight into the larger realm from which the data came, perhaps along with some insight into the underlying social, cultural, or other phenomenon of interest. As such, the appropriate visualization is often specific to the particular problem or research question at hand and the available data. Of necessity, in this section we discuss visualization methods and techniques in the context of fairly general problem and data classes: networks, geographically-based data, surveys, linguistic data, and social media.

[4] As Shakespeare (1936) wrote in *Hamlet*: "Hamlet: Do you see yonder cloud that's almost in shape of a camel? Polonius: By the mass, and 'tis like a camel, indeed. Hamlet: Methinks it is like a weasel. Polonius: It is backed like a weasel. Hamlet: Or like a whale? Polonius: Very like a whale."

[5] A somewhat related phenomenon is the cherry picking of facts to suit a preconceived notion. As Sherlock Holmes warns, "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts" (Doyle, 2003, p. 189).

2.1 Classical Approaches to Data Visualization

Data visualization has historically been thought of in terms of graphing numerical data. Typically these data were either cross-sectional (i.e., collected at the same point in time) or longitudinal (i.e., collected over time) and they were either continuous or discrete, with methods developed to address each combination. For example, typical graphical methods for continuous cross-sectional data include histograms and box plots. Methods for discrete cross-sectional data include a variety of bar and pie charts, as well as simple tabular summaries. Longitudinal data are most typically plotted on some type of time series plot. Visualization methods for two variables include scatterplots, mosaic plots, side-by-side box plots and others.

The scatterplot matrix (also known as a pairs plot) is a specialized plot for simultaneously showing multiple scatterplots. It is typically displayed in a square showing n continuous variables on both the vertical and horizontal axes, where each row and each column depicts one variable's comparison with the other $n - 1$ variables. The diagonal is typically left for labels or single-variable graphs like histograms, since no meaning can be derived from a scatterplot comparing a variable to itself. The upper-triangle of the square displays the same information as the lower triangle, albeit inverted (i.e. "X versus Y" translates to "Y versus X"), so it is not uncommon to display only one side of the diagonal. All benefits of the scatterplot (e.g. visualizing correlation) are inherited, though the scatterplot matrix requires the viewer to cognitively examine and potentially explore each pair of variables individually.

For more detailed discussion of classic data visualization methods, see Tukey (1977) and Cleveland (1993, 1994). These and more recent methods, described in the context of modern statistical software, are discussed in texts such as Wilkinson *et al.* (2005), Wickham (2009), Sarkar (2008), and Murrell (2011), though these still tend to focus on statistical and quantitative graphics. Discussions of visualization from a media and/or graphical design perspective include Yau (2011, 2013), Steele & Illinsky (2010), and Wong (2010). And, of course, there are the well-known books by Tufte (1986, 1990, 1997, 2006).

Common to most of these methods and most of these discussion is that the data is quantitative, typically of low dimension (mainly one or two, and sometimes three dimensions), and the plots are mainly static. Said another way, the classical data visualization methods are generally not designed for text-based data, photo and video data, high-dimensional data, and geo-spatial data, all of which are rapidly becoming more common at an increasingly greater rate.

Text, photo, and video data are often displayed in small multiples or facets, often called lattice or trellis plots. These are particularly useful for looking at plots of data conditioned on the value of one or more categorical variables. For example, Figure 2 is a lattice plot of the age of patients presenting at a health clinic conditioned on gender (male or female) and ethnicity (unknown, Hispanic, and non-Hispanic). The plot shows there are clear differences in the age distribution of patients presenting by gender but no difference by ethnicity.

Rosling (2013) displays multiple variables via bubbles on a scatterplot. In addition to the variables represented on the two axes, the size of the bubble represents a third variable and (optionally) bubble color represents a fourth. His software, *Gapminder World*, provides over 500 variables for comparison; an example output is provided in Figure 3. The same data can also be displayed geographically, as in Figure 4, where the two axis variables from Figure 3 (life expectancy and income per person) are exchanged for country centroid

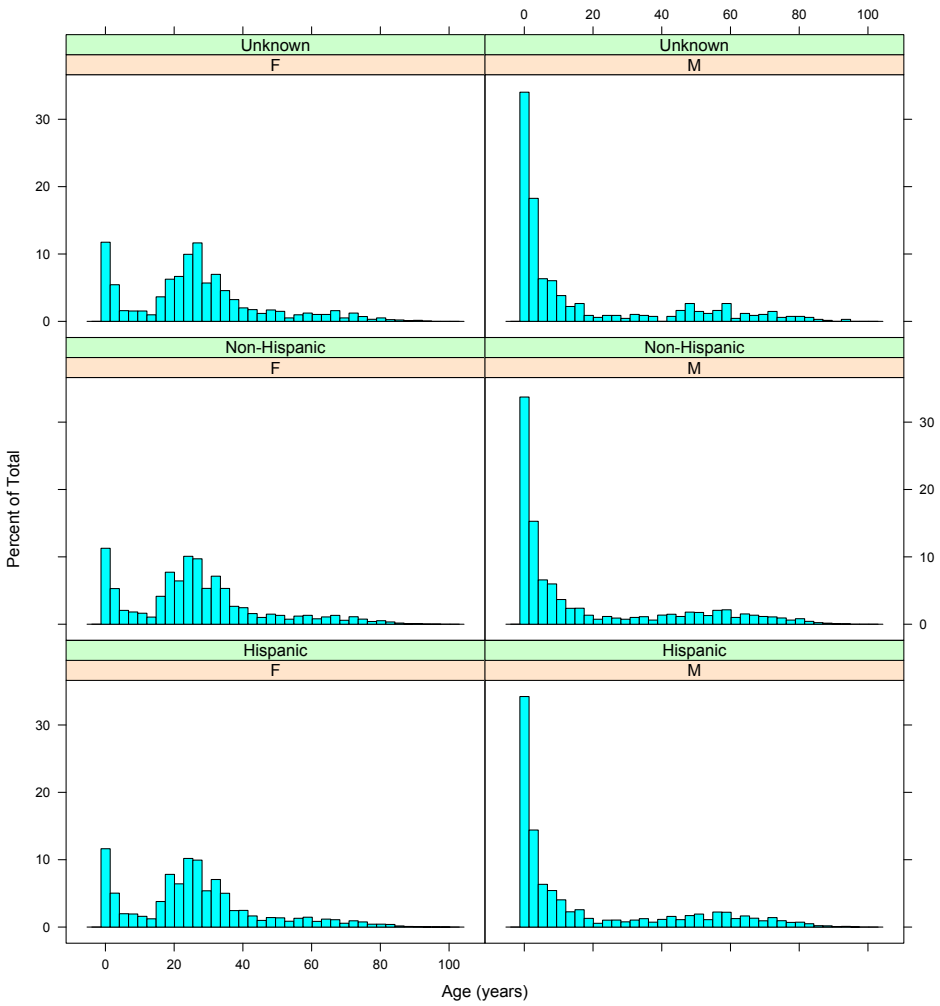

Figure 2: A trellis plot example showing clear differences in distribution of health clinic patient age by gender but no difference by ethnicity. Source: Fricker (2013).

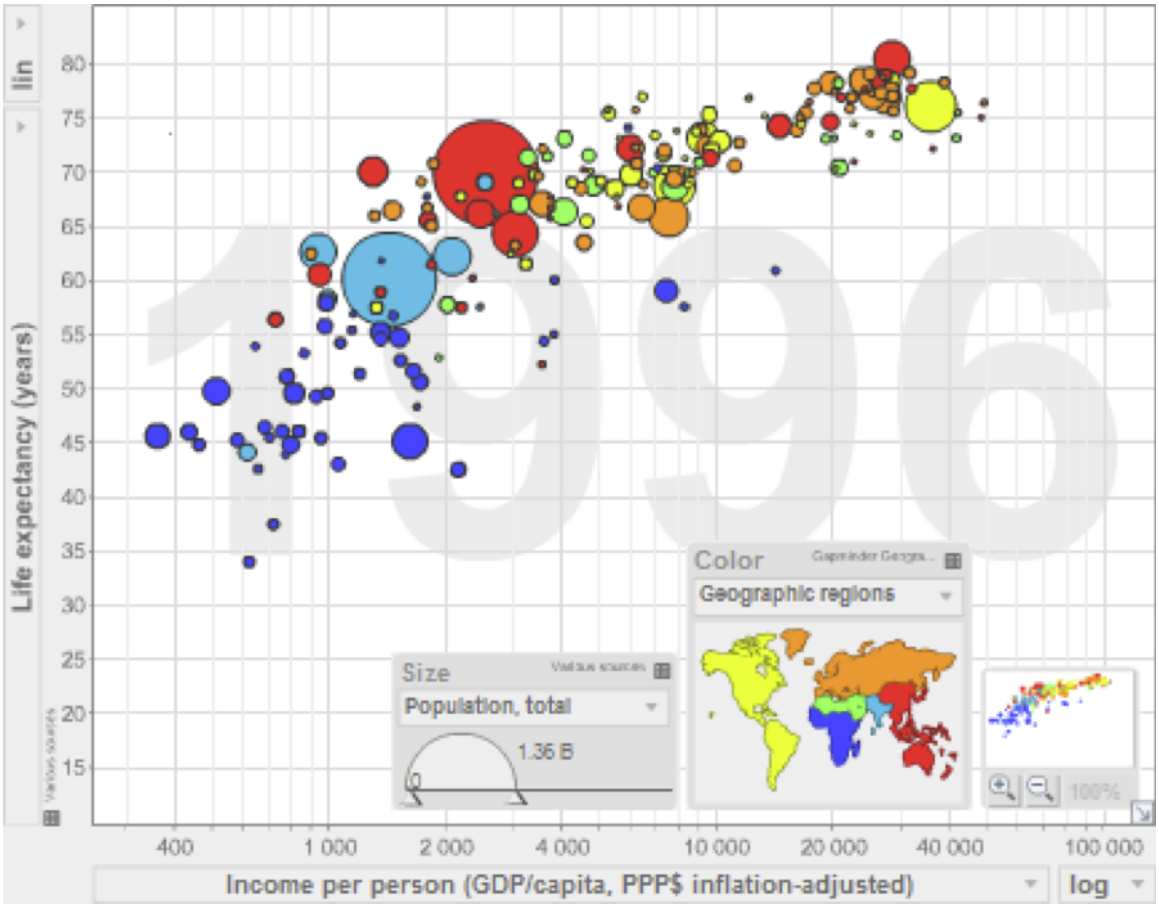

Figure 3: Rosling's *Gapminder World* software provides easily-interpretable displays of up to four variables, with a fifth presented in a time-series animation. Dimensions are provided by the two axes plus the bubble size and color. Data is current as of April 20, 2013. Adapted from *Gapminder World* software display (Rosling, 2013).

latitude and longitude, while the bubble size and color retain their original meanings.

Tufte (2001) cautions against under- or more commonly over-representing the magnitude of the data when using area to display a variable (where the mistake is not made in Rosling's work). As Tufte says, "The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented." Tufte proposes a metric called the *Lie Factor* which is the ratio of the size of the effect shown in the graphic and the size of the effect in the data. Values within 0.95 and 1.05 are preferred, while ratios outside of that range show "substantial distortion, far beyond minor inaccuracies in plotting" (Tufte, 2001, p. 57).

Enhancements to an individual scatterplot may be used to display one or more additional variables. For example, varying the dot size, color, and/or shape provides the ability to add extra variables to the graph. LOESS (locally weighted scatterplot smoothing) curves, developed by Cleveland & Devlin (1988), do not add a dimension though may provide insight to the relationship between variables.

Healey *et al.* (2008) related data display to human perception and automated search

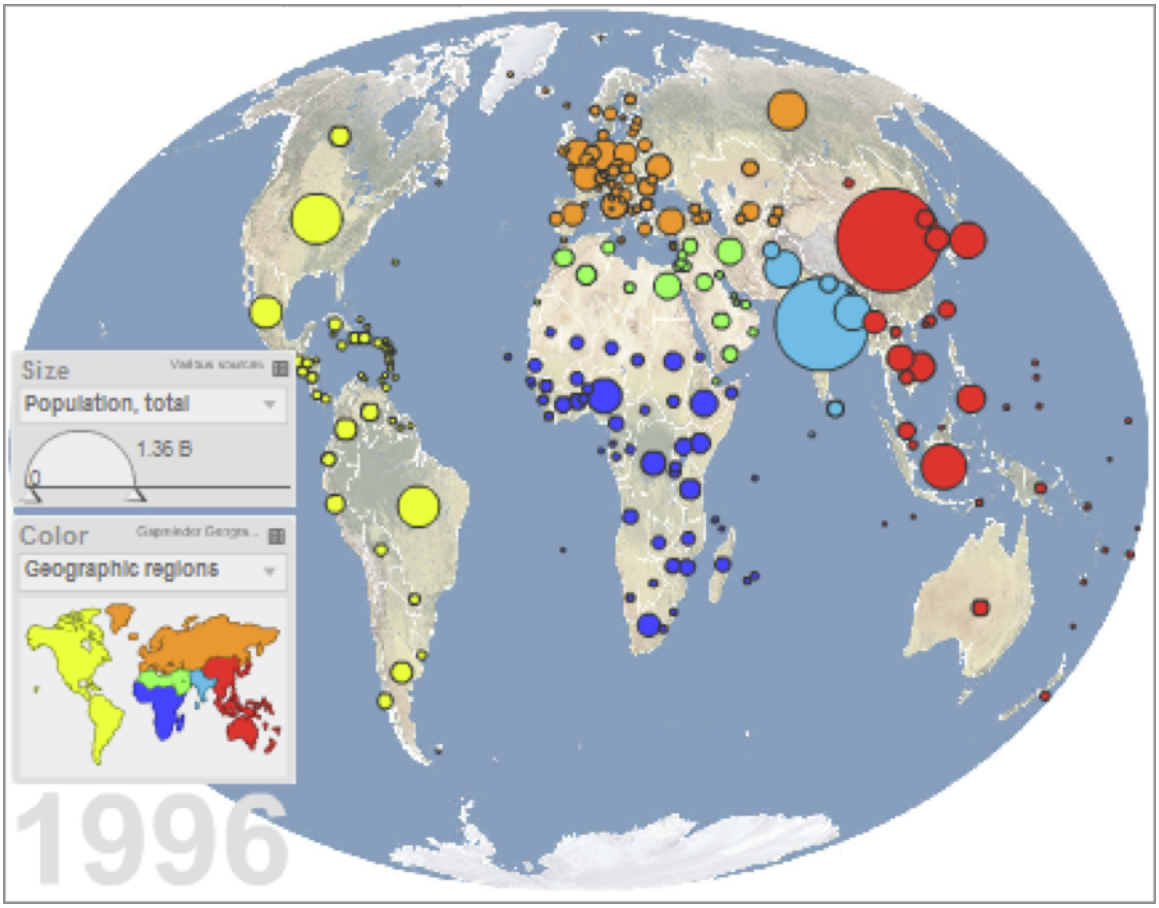

Figure 4: The two axis variables from Figure 3 are exchanged for country centroid latitude and longitude. The bubble size and color retain their relevance. Data is current as of April 20, 2013. Adapted from *Gapminder World* software display (Rosling, 2013).

Figure 5: Multivariate visualization techniques evaluated in the study. From left to right: brush strokes, data driven spots, oriented slivers, color blending, and attribute blocks. Source: Livingston & Decker (2012).

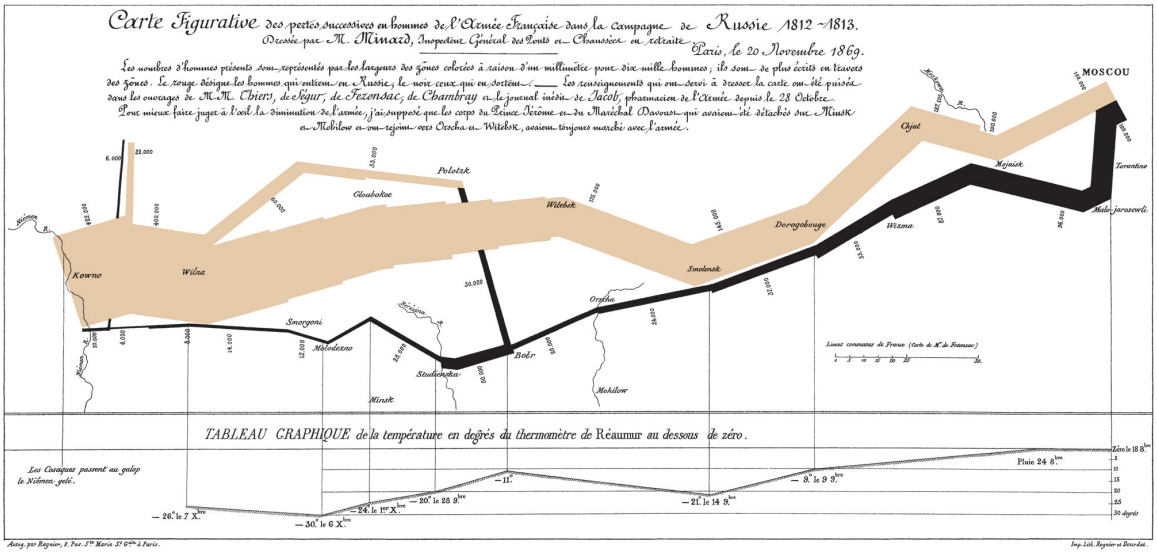

Figure 6: Classic graphic from Charles Minard (1781-1870) showing the progress and fate of Napoleon's army in its march against Moscow. This is a combination of a data map and a time-series, and displays six variables.

strategies, providing a pathway from the data to visual interpretation. They used luminance, hue, size (height), density, orientation, and regularity to a grid. Their work is continued by Livingston *et al.* (2011, 2012, 2013) and Livingston & Decker (2012, 2011), intending to display up to ten variables simultaneously on two-dimensional plots. Figure 5 shows five of their techniques.

Perhaps one of the most elegant portrayals of multi-dimensional data is Charles Minard's flow map of Napoleon's March to Moscow, Figure 6. Tufte (2001, p. 40) credits this graph as displaying "six variables: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. ... It may well be the best statistical graphic ever drawn."

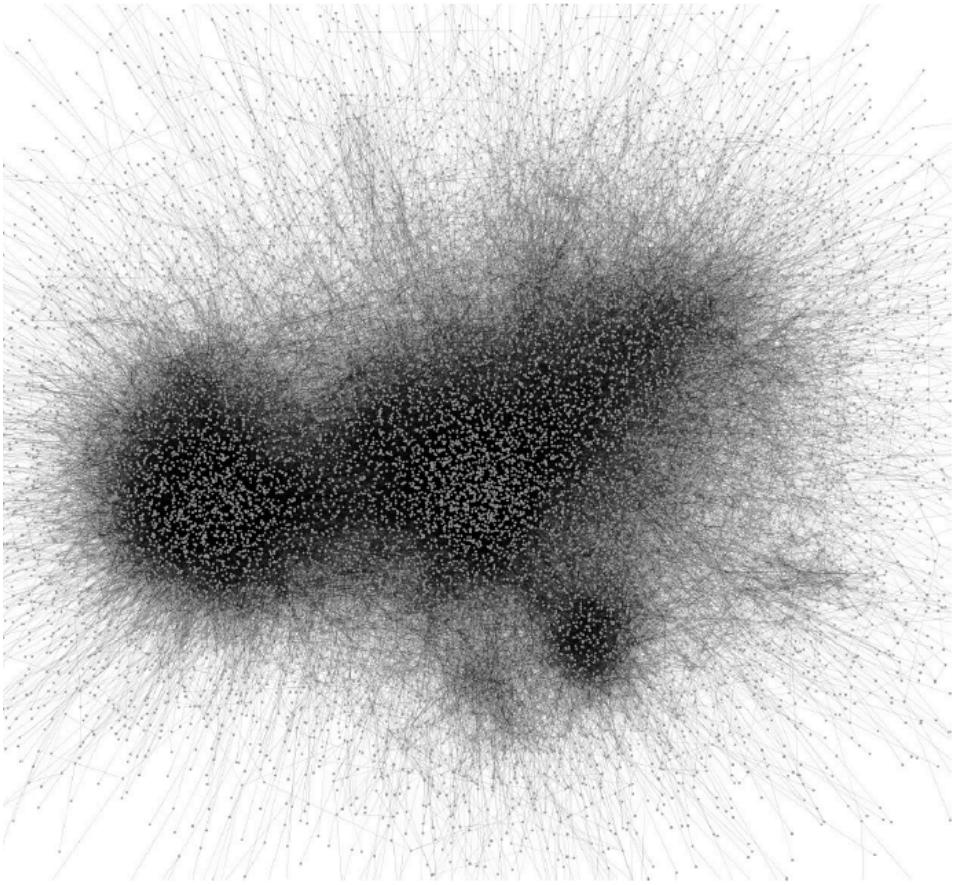

Figure 7: A network visualization of a Facebook friendship network for 15,000 users. A classic "hairball" that provides little insight into the network structure. Source: Coscia (2013).

2.2 Network Visualization

Networks occur naturally in many situations, and conceptually they are quite simple, with every entity denoted by a node in the network and linkages between entities denoted by arcs. Visualizing networks in ways useful for learning about and understanding the network, on the other hand, can be anything but simple. For example, particularly with large networks, simply displaying all the arcs and nodes results in the classic "hairball" (see, for example, Figure 7) from which little about the network can be discerned. As Krzywinski (2013) says,

> Hairballs turn complex data into visualizations that are just as complex, or even more so. Hairballs can even seduce us to believe that they carry a high information value. But, just because they look complex does not mean that they can communicate complex information. Hairballs are the *junk food of network visualization* – they have very low nutritional value, leaving the user hungry.

At issue is that "[v]isualizations are useful to leverage the powerful perceptual abilities of humans, but overlapping links and illegible labels of nodes often undermine this approach" (Perer & Shneiderman, 2006, p. 693). As a result, when visualizing networks it is often

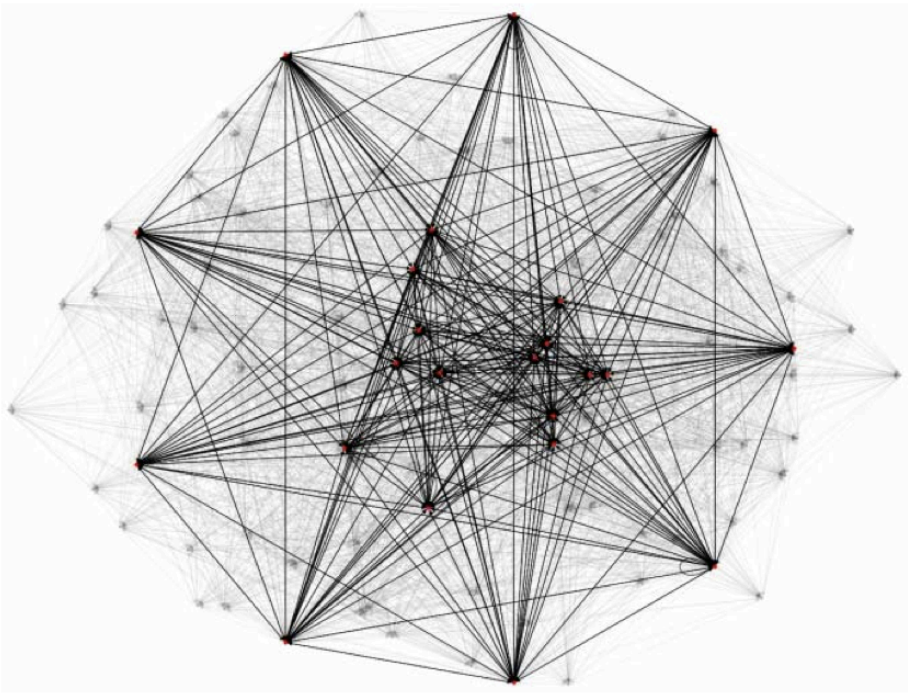

Figure 8: An example of a filtered network that highlights the most connected nodes and their connections to one another. Source: Smith *et al.* (2009, Figure 5).

important to either do the visualization in interactive software so that the user can explore the network appropriately and/or to subset or highlight the data so that structure is visible. Figure 8 is an example of the latter approach in which the most connected nodes and associated arcs are highlighted.

There are numerous ways to visualize networks. Figure 9 illustrates a few by Krzywinski (2013) using hive plot software. What should be evident is that the most relevant visualization or visualizations is tied to the specific problem at hand and associated data. Furthermore, finding the preferred visualization may require multiple attempts and will be facilitated with software that allows for interactive exploration and visualization of the network.

Interactive network visualization is very much in the spirt of Tukey's exploratory data analysis and in keeping with Perer & Shneiderman (2006, p. 40) who promote an "overview first, then details on demand" approach to visualization. Similarly, (Viégas & Donuth, 2004) say, "We posit that basic cartographic principles – such as adaptive zooming and multiple viewing modes – provide system designers with useful visual solutions to the depiction of social networks."

The variety and quantity of network visualization approaches and software are too great to permit a comprehensive review in the limited space available in this chapter. Furthermore, the practice is undergoing rapid development, so any listing is sure to become quickly dated. Instead, on-line resources such as Mobio (2013) are recommended. For example, in an interesting exercise in self-referential visualization, Figure 10 is a screen-shot of an interactive network visualization of data visualization resources.

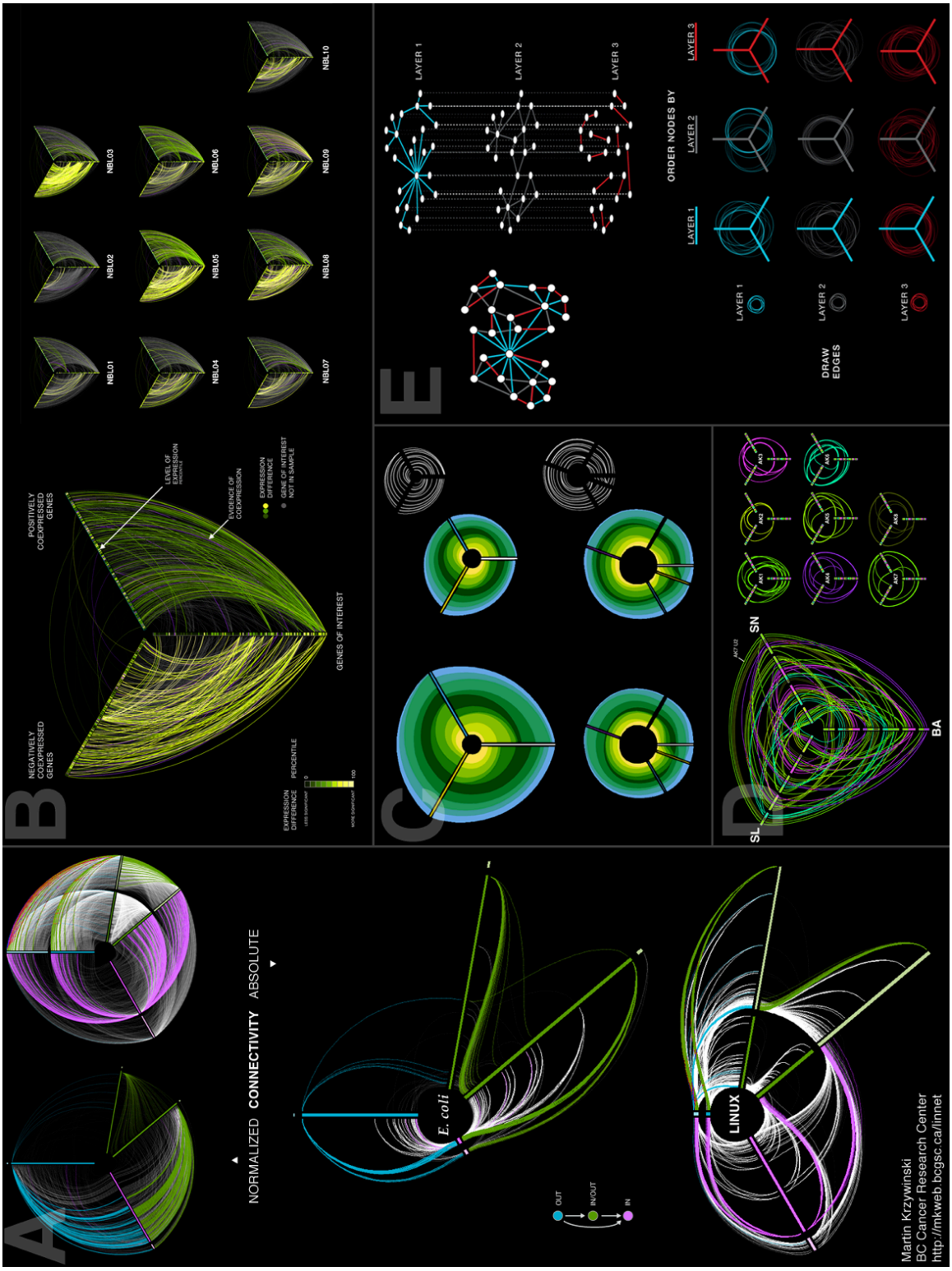

Figure 9: Alternative network visualizations: (A) normalized (top) and absolute (bottom) connectivity; (B) co-regulation networks; (C) network edges shown as ribbons creating circularly composited stacked bar plots (a periodic streamgraph); (D) syntenic network; and (E) layered network correlation matrix. Source: Krzywinski (2013).

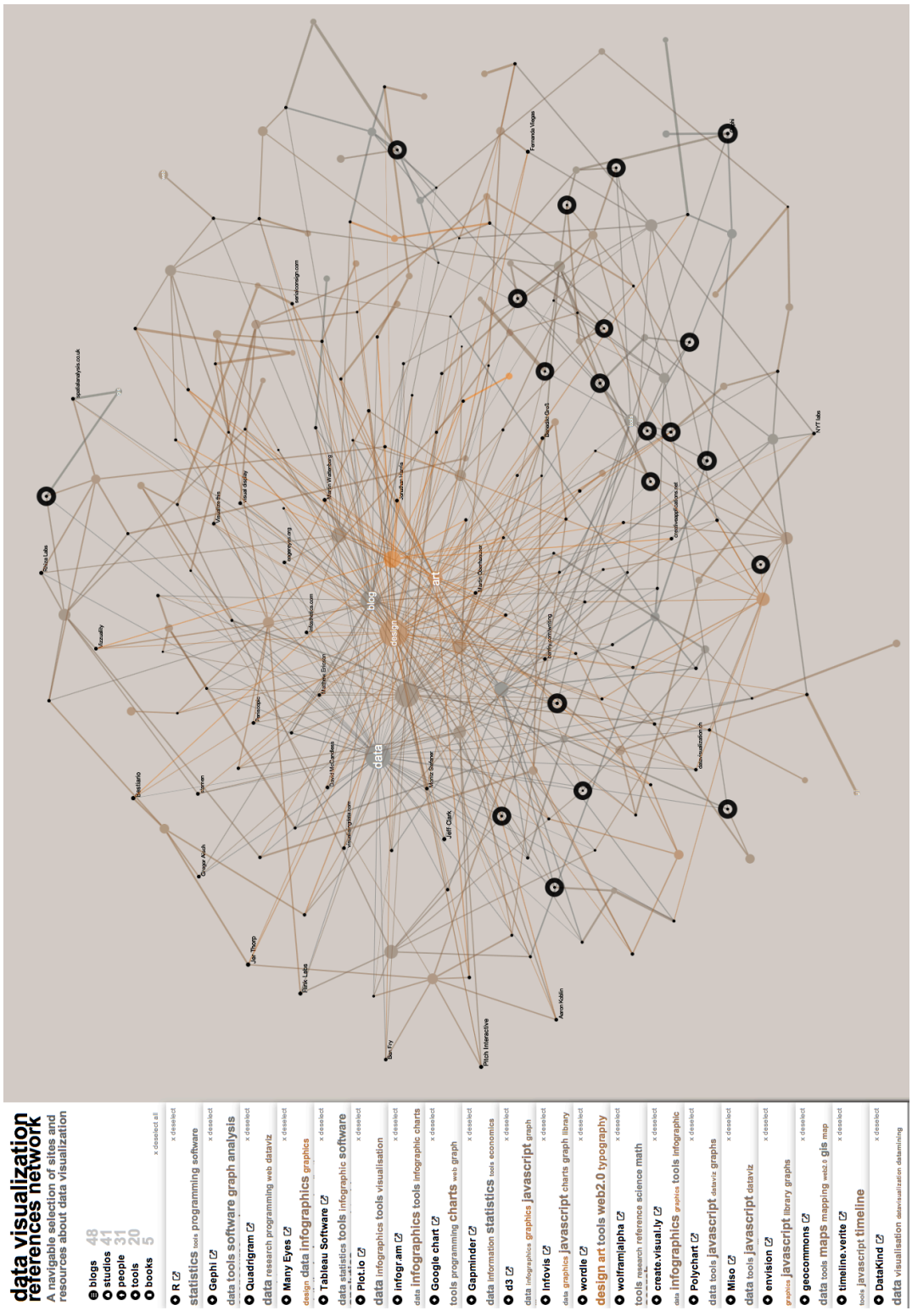

Figure 10: An interactive network visualization of data visualization resources. This particular view highlights visualization software. Source: Mobio (2013).

2.3 Geographic Information Visualization

The visualization of geographic data – the making of maps – has a long history. Developments in computing, though, have put the display of geographic data within easy reach. Furthermore, the widespread adaptation of global positioning system (GPS) devices in objects like mobile phones and freight palettes have led to great growth in the amount of georeferenced data available to consumers and organizations. Additionally it is now straightforward to perform the sort of analyses, as distinct from visualizations, that would have been much more difficult even ten years ago. By analysis, we refer here to tasks like identifying the polygons that contain particular points (say, identifying terrorist incidents with provinces), projecting points onto lines (say, identifying traffic accidents with road locations) , or computing lines of sight (locating regions that cannot be seen from a particular tower, for example). Of course, the distinction between display and anlysis is not always sharp.

2.3.1 Coordinate Systems

One aspect of geographic visualization that sets it apart from traditional approaches is that coordinate systems need to be chosen with care. Data often start off being associated with geographic coordinates – latitude and longitude, typically in degrees. Since the earth is roughly ellipsoidal, geographic coordinates need to be associated with a "datum" which characterizes the particular ellisoidal approximation in use. A latitude, longitude pair for a particular location derived under one datum might be hundreds of meters away from the same pair derived under another datum.

A number of datums[6] are available, but the WGS84 datum, used by GPS, is very much the most commonly used. A number of computer programs make it possible to convert coordinates based on one datum to another.

Over small areas geographic coordinates can be plotted directly, neglecting the curvature of the earth, but for even moderate distances (say, dozens of miles or scores of kilometers) a projection will be necessary. As one interesting example, the two towers of New York's Verrazano bridge are 1-5/8 inches farther apart at the top than at the bottom because of the curvature over the bridge's 4,260-foot length.

There are dozens of projections available, with different projections serving different roles. For example, the widely-used transverse Mercator projection preserves angles – that is, angles on the map match angles on the ground – and it therefore useful for navigation. The Albers projection shows areas accurately, which might be particularly useful in displaying political data. Figure 11 shows the lower 48 states of the USA under the Mercator (upper) and Albers (lower) projections, together with a grid of lines of constant longitude and of constant latitude. Any text on GIS (such as, for example, Bolstad, 2008) will list and compare the common projections.

Projected coordinates are often displayed as latitudes and longitudes, but other times they are converted to UTM, the Universal Transverse Mercator grid system (or perhaps its military analong, Military Grid Reference System). UTM represents the globe by a set of sixty projections, each six degrees of longitude wide by eight degrees of latitude high. Longitude bands are numbered; latitude bands are lettered. Within a zone, points are labeled with "eastings" and "northings" in meters. This makes it easy to compute

[6] In the geodetic context, the plural of "datum" is always "datums."

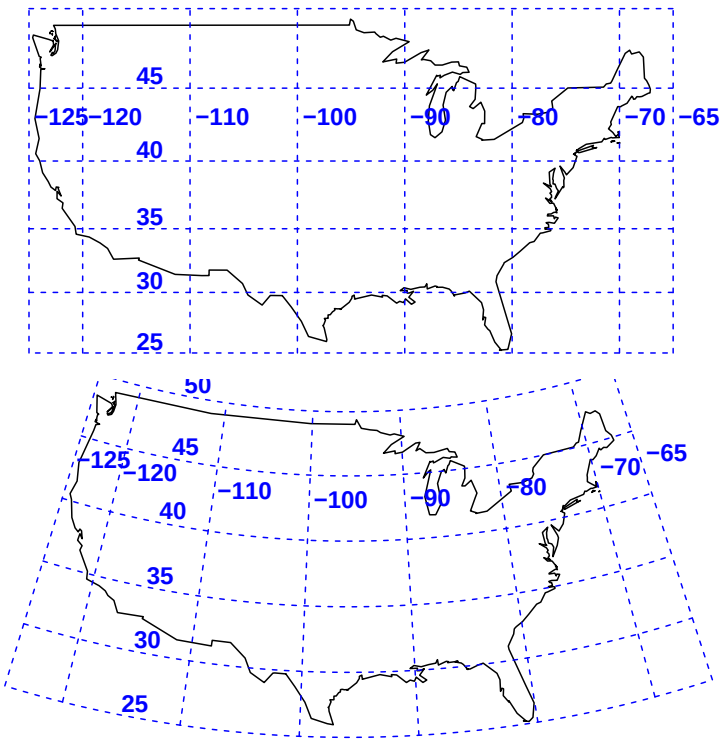

Figure 11: Two projections of the lower 48 states of the U.S.A. The upper shows the familiar Mercator projection, the lower the equal-area Albers projection.

lengths and areas directly from the coordinates, and within a zone the distortion in a length computed this way is held under 0.1%. Computing distances between points in different zones, however, requires a more difficult computation.

2.3.2 Tools and Formats for Geographic Data

The final product of a visualization will generally be a graphic image in digital form. Just as there is a vary large number of graphic formats in use, so too is there a large number of formats in which geographic data can be held, manipulated and depicted. Often the data will be saved by one piece of software in projected form, which means that it may have to be both re-projected and converted to be used in another.

There are, as well, many sets of tools for displaying geographic data. Such a tool is often called a Geographic Information System (GIS). Among the most popular of these are the ArcGIS suite from Esri Corporation, which is popular, powerful and comparatively expensive; the free, open-source GRASS GIS from the Open Source Geospatial Foundation; and the widespread Google Earth, the base version of which is currently offered free by the Google Corporation. This last offers very limited analysis capabilities, but provides a quick way to combine geographic data with pre-supplied aerial photography and road network information.

While different software uses different formats, two widespread formats deserve mention here. ArcGIS stores data in "shape files." This is a slight misnomer since a shape file is actually a collection of at least three distinct disc files. While originally intended for ArcGIS, the shape file format is now open and in fairly wide use in other GIS products. Shape files generally hold one of four types of displayable data: points (like locations of specific incidents), lines (like roads or rivers), polygons (like the boundaries of states or provinces), and raster data. The first three of these are collectively known as "vector" data. Raster data is in grid form; it might be physical, as with a photographic image of the ground, or logical, as in a map showing the availability of fresh water for each cell in a gridded region. A second format is KML, a text-based format built in XML and supported by Google Earth. KML files can contain both vector and raster data. It should also be noted that the widely used open-source statistical environment R can, with the proper libraries, read and write both shape files and KML.

The construction of a geographic image generally consists of the superposition of "layers," each layer consisting of a shape or KML file describing some feature of the area being depicted. Because these tools have been designed for map-makers, the features are almost always physical, but in principle they might be lines showing connections among individuals, numbers of web-site hits per square mile, or other non-physical data. As an example of the superimposition of layers, Figure 12 shows three layers used to build a map of Nigeria: a satellite image (top left), state boundaries (top right), terrorst events (bottom left), and then the three superimposed (bottom right). Of course, in making the map the researcher is required to select not only the layers but their attributes – color, point size, line width, transparency, and so on.

2.3.3 Coloring the Map

The researcher's task, then, is to use the correct set of layers to display, and not conceal, the essential information. A map intended to show population density by province would

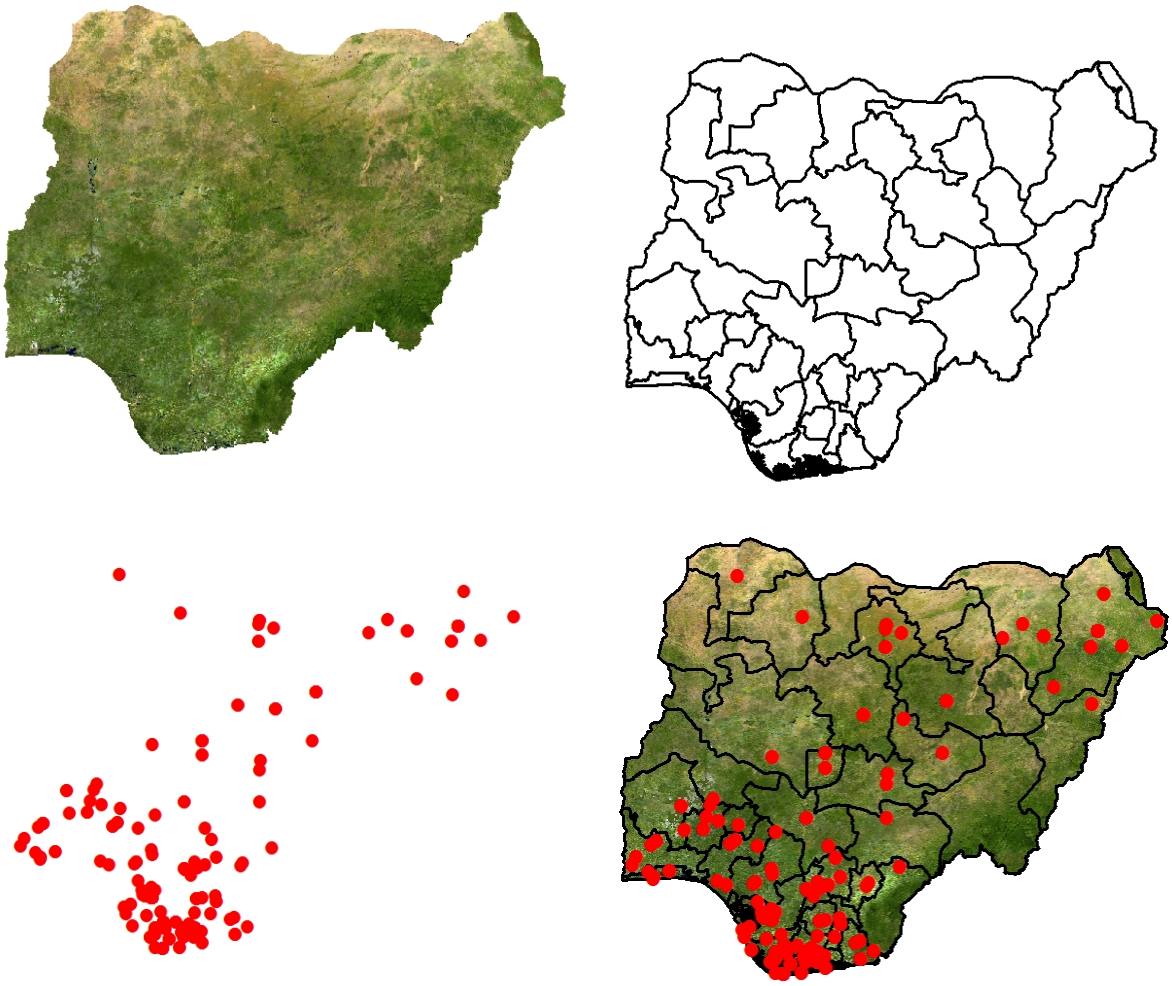

Figure 12: Three layers in a Nigeria map (upper left and right and lower left) and the resulting map in the lower right.

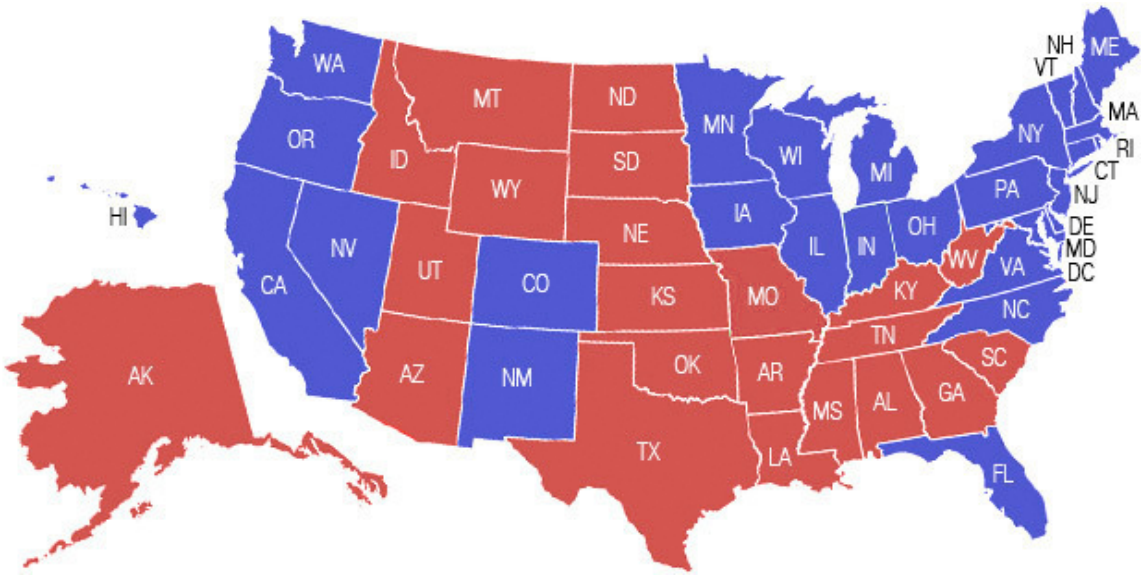

Figure 13: Map showing the states whose electoral votes were awarded to John McCain (light gray/red) and to Barack Obama (dark gray/blue), plus the District of Columbia. About 62% of the map is light gray/red, but equal areas do not denote equal densities of electoral votes; Alaska's 570,000 square miles of land are much more visible than D.C.'s 68, though the two each contribute three electoral votes. Source: Cole (2012).

not benefit from the inclusion of railroads or precipitation unless part of the story were that population tends to settle along the railroads, or in areas of higher rainfall.

In the context of human activity, though, data is often produced in the form of counts by region, say province. A natural device is to color areas according to the value being represented. Such a graph is sometimes called a "chloropleth." However the researcher should in general display not counts but count density, that is, counts per area, particularly when areas are quite different.

One interesting technique that is now available to analysts is that of the "cartogram," in which a map shows regions that are scaled to represent the quantity being displayed, while having their respective shapes retained to the extent possible. Figure 13 shows a map of the 2008 U.S. Presidential election, in which states that awarded their electoral votes to John McCain are light gray (red when the cartogram is displayed in color; see `www.npr.org/blogs/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money`), and those which awarded their votes to Barack Obama are darker gray (blue in color) (Cole, 2012). In this map, the ratio of light gray (red) area to dark gray (blue) is about 62%, yet the size of the states visually misrepresents that 68% of the electoral votes are from the states that voted for Obama. In Figure 14, the data and color-coding are the same, but each state is represented by an area proportional to its number of electoral votes (Carter, 2008). This map is about 68% darker gray (blue). Cartograms can also be drawn with curved boundaries.

Another approach to showing the density of data on maps is the so-called "heat map." A heat map represents values by colors selected from a color ramp. Interpolation is used to estimate the value at points where no data is observed; the coloring is typically performed

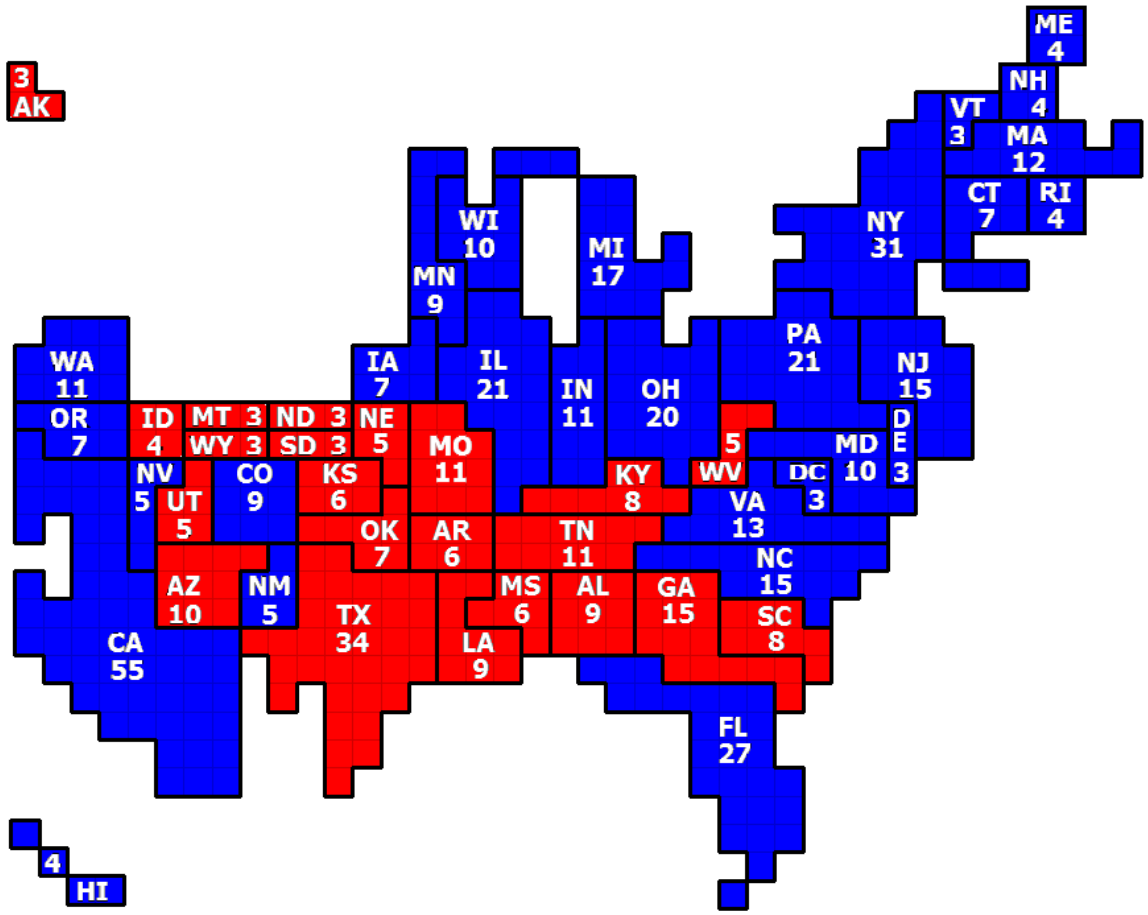

Figure 14: Map showing the electoral votes awarded to John McCain (light gray/red) and to Barack Obama (dark gray/blue) by the fifty U.S. States plus the District of Columbia. In this map, states are represented by areas proportional to the number of electoral college votes to which they are entitled. About 68% of the map is dark gray/blue, because about 68% of the electoral votes went to Obama. Source: Carter (2008).

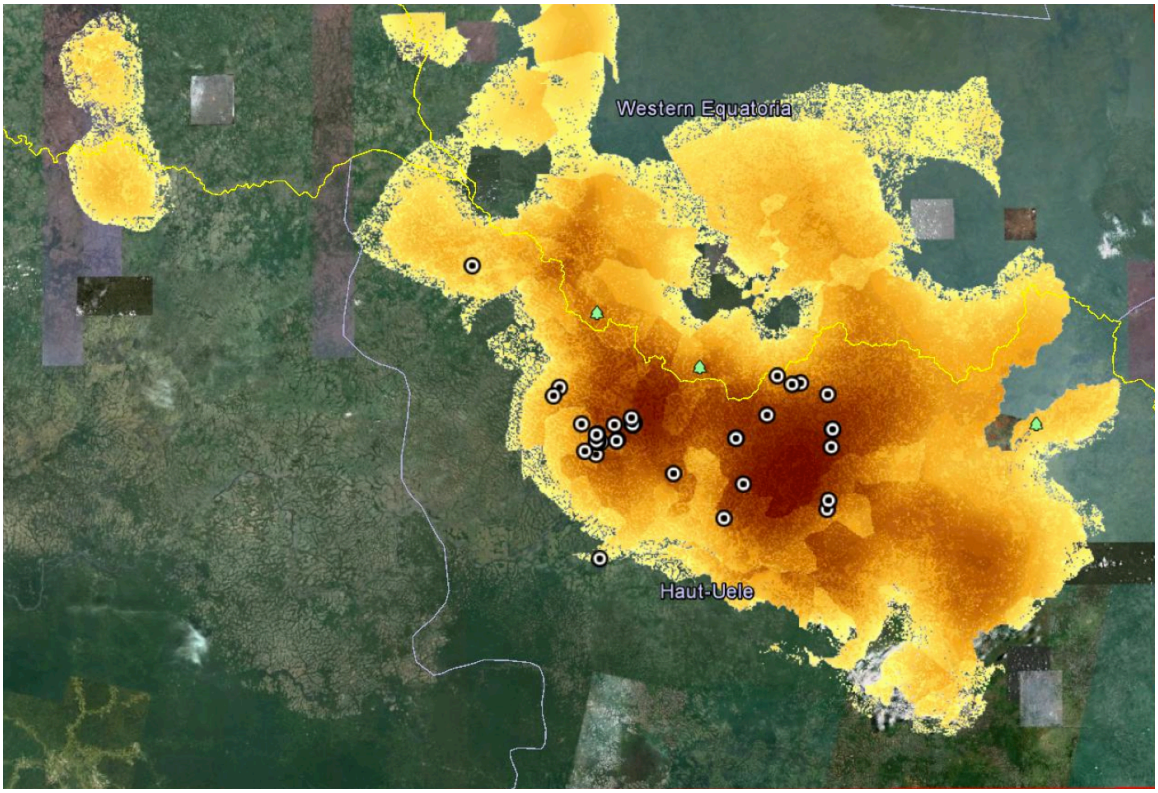

Figure 15: Heat map showing predicted Lords Resistance Army (LRA) activity in a given area based on previous attack data. Darker colors indicate regions of higher probability of future attacks and points show LRA attacks reported after the original prediction was created and disseminated. Source: McCue *et al.* (2012).

without reference to political or other interior boundaries. In Figure 15, we see a depiction of predicted Lords Resistance Army (LRA) activity in a given area based on previous attack data. Darker colors indicate regions of higher probability of future attacks and points show LRA attacks reported after the original prediction was created and disseminated (McCue *et al.*, 2012).

2.3.4 Altitude, Time and Animation

Although there is a mechanism by which shape files can hold altitude or elevation data, in fact other formats are in more common use. These include the US Geologic Survey's Digital Elevation Model; its successor, the Spatial Data Transfer Standard; and the military's Digital Terrain Elevation Data. All of these formats have been developed to support the display of actual ground elevations and do not appear to be in widespread use to represent non-geographical data like population densities (but see below).

It is also difficult to display chances associated with the passage of time on a static map. Animation is a natural tool here, and it is straightforward to create a simple animation as an animated GIF or PNG file in many GIS and other products. These formats can be viewed in any web browser, and animation is also easily understood by an audience.

However, difficulties arise with the necessity of creating, storing and displaying animations for all possible combinations of hardware and software. More importantly, even today a large number of readers still use paper to digest and store the visualization.

The simplest way to show the passage of time is by showing a number of copies of the map, each constructed using data for a different time period. Figure 16 is an example. Here we see the number of nurses per 10,000 residents for each county in Missouri at six different time periods (Courtney, 2005). Of course, the focus on counties runs the risk of the same misinterpretation as in Figure 13. For example, the independent city of St. Louis, which is very populous but not in any county, is nearly invisible, and the density scale is per 10,000 residents, not, as perhaps it should be, per square mile. Still, this sequence of maps does make it easy to track the status of specific counties from year to year. The second, smaller state map, above each larger one, shows which counties were declared to be HPSA (health professional shortage areas). The graph makes it easy to see that the set of HPSAs was mostly constant between 1991 and 1999, but that there was a big change between 1999 and 2001.

We offer two more notes on this picture. The density scale of nurses per 10,000 people is divided at values of 13, 44 and 100. The analyst is free to choose her own cutoffs, of course, but she must make sure (a) that the graph tells the story she wants to tell; (b) that a different choice of cutoffs would not tell a radically different story; and (c) that where applicable the cutoffs correspond to natural or important divisions. For example, in monthly data, 6, 12, and 24 are natural divisions, whereas in income data natural divisions would be in even thousands, or tens of thousands of dollars. Second, the color choice in this example lends itself nicely to printed reproduction. Bright colors that are very different on a color computer screen are often similar in print. Colors need to be chosen wisely, using intensity as well as hue, and perhaps with some sensitivity to the fairly common red-green form of color blindness.

As a final example, we present a map showing the concentration of arrests for prostitution in the city of San Francisco in 2009. Figure 17 (McCune, 2010) makes it easy to see that arrests in that year were concentrated in two geographic areas. The author notes that the data was "...aggregated geographically and artistically rendered. This [visualization] is meant more as an art piece than an informative visualization" (McCune, 2010). While the tools to create visualizations like these, complete with lighting and shadow effects, do not yet appear to be in widespread use, the pictures may serve as an indicator of what is now possible in the field of geographic visualization.

2.4 Survey Data Visualization

The visualization of survey data has changed little in the past few decades and the standard techniques of Section 2.1 apply whether the data is collected for sociocultural analysis or other reasons. One challenge with survey data is that they are typically discrete, often arising as responses to Likert scale questions. Standard displays of such data include pie and bar charts, though these types of graphics do not lend themselves to much more than univariate presentation. At issue is that these types of graphs are not particularly useful for displaying interactions between variables or relationships among multiple variables. Even using side-by-side and stacked bar charts, it is generally cognitively difficult to compare between subsets of the data, and thus these techniques make it challenging to identify complex sociocultural signatures.

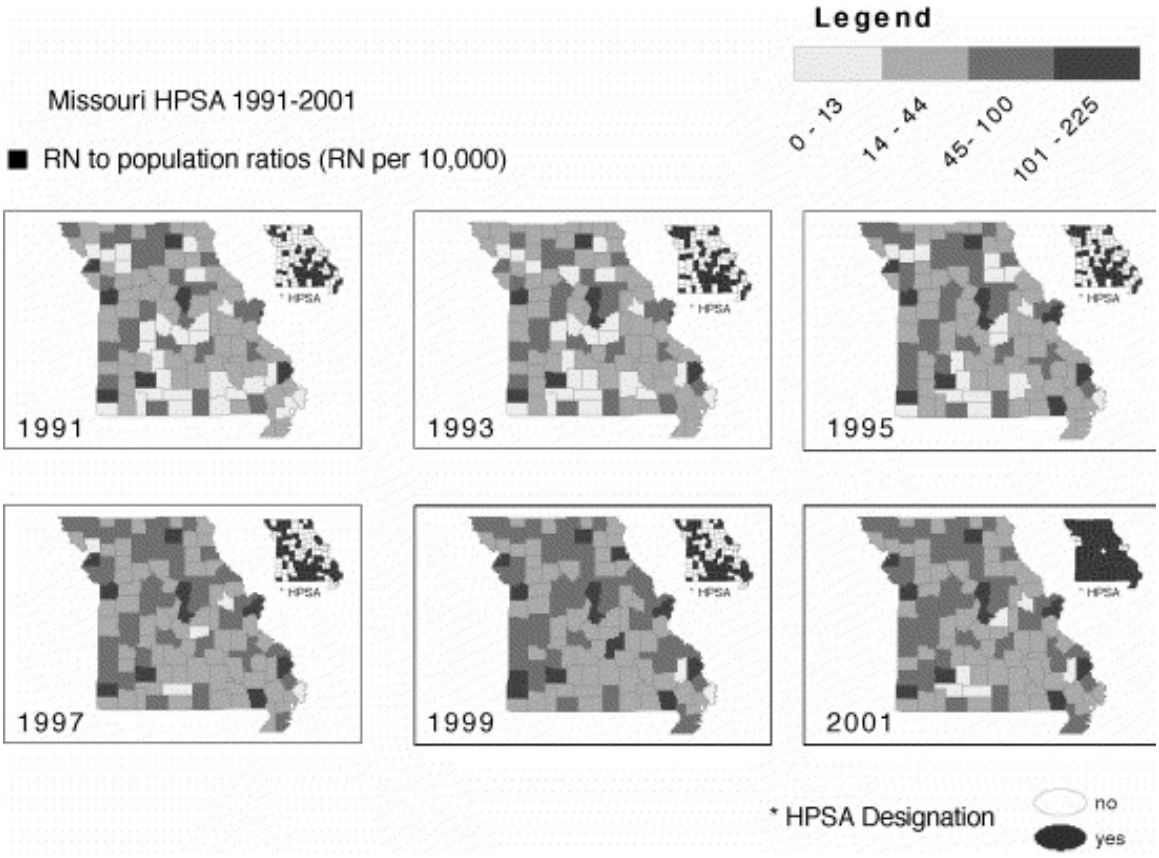

Figure 16: Maps of the density of nurses, by county, in Missouri, 1991-2001. The insets show the set of counties designated as Health Professional Shortage Areas. Source: Courtney (2005).

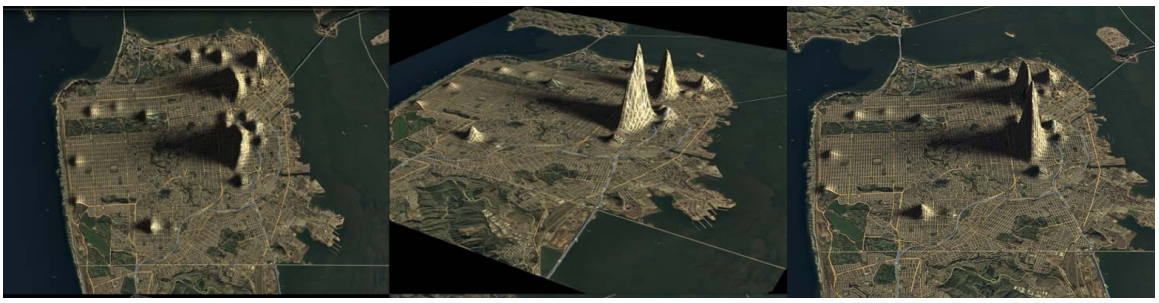

Figure 17: Visualizations of the density of prostitution arrests (vertical direction) in San Francisco in 2009. Source: McCune (2010).

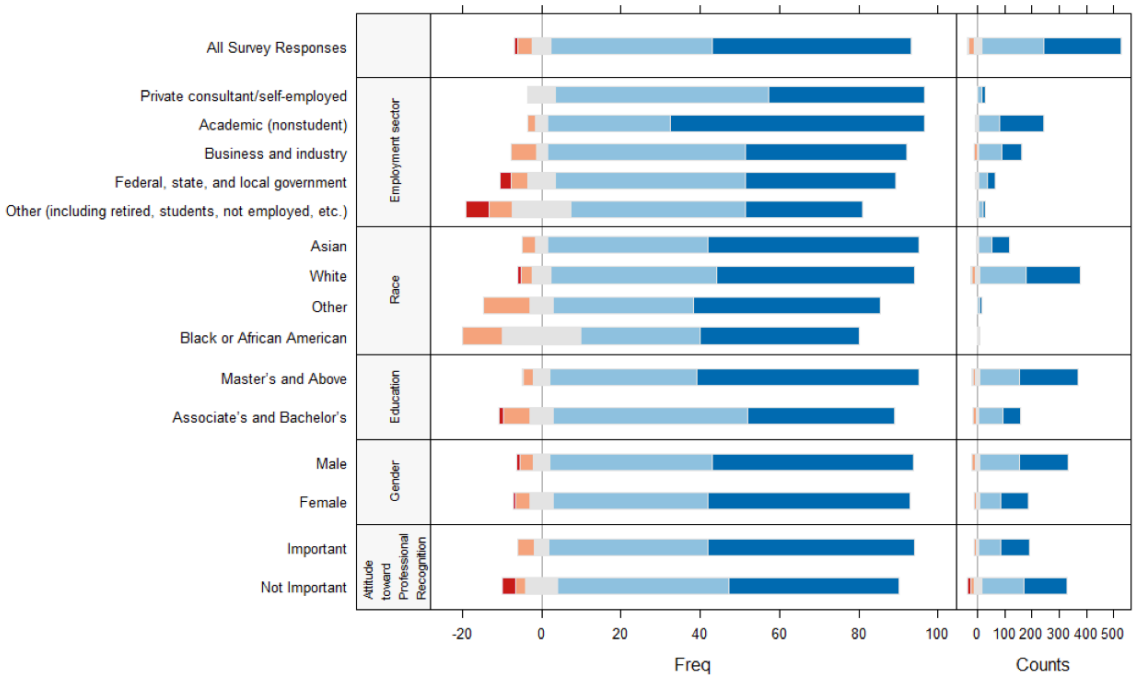

Figure 18: An example of a diverging stacked bar chart for a five point Likert scale question. This particular example shows the results for the entire sample (top bar) and then by five different demographic variables. The left set of bars shows the data in terms of percentages and the right set shows the counts. Source: Robbins & Heiberger (2011, Figure 2).

A recent advance in survey data visualization is the diverging stacked bar chart developed by Robbins & Heiberger (2011). As shown in Figure 18, the diverging stacked bar chart centers the stacked bars, typically on the neutral or central response and then each end of the bar diverges away from the neutral. These types of plots make comparing response distributions between subsets of the data relatively easy, clear, and intuitive.

A major impediment for survey data that are collected using complex sampling designs is the lack of software that would aid in exploratory data analysis, particularly software that would allow social scientists easy (yet correct and appropriate) visualization of the data. Specialized software, such as SAS, SPSS, Stata, and R, have these capabilities, but they also require specialized skills and capabilities. Rix & Fricker (2012) proposed a proof-of-concept solution, but to date there is no dedicated software solution that would facilitate visual EDA of complex survey data. At issue is that complex sampling designs require sophisticated analytical techniques in order to correctly analyze the data and these methods need to be an integral part of the software, but they also need to be relatively transparent to the user, while the software also must be designed to support effective EDA as described in Section 1.2.

An open research question with survey data that have a geographic component is how to map the data *and* simultaneously show the margin of error by geographic region. The difficulty is that this type of visualization requires four dimensions (two for the map itself, a third for the measure of interest, and the fourth for the margin of error). Most geographic

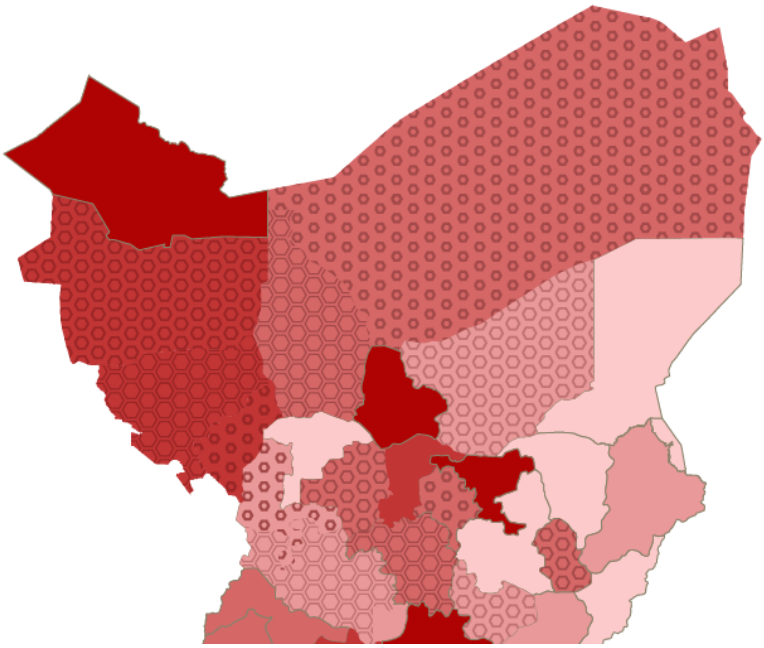

Figure 19: An illustrative map where survey results are displayed using color and uncertainty (the margin of error) is displayed using texture.

displays of survey data display only point estimates by shading or coloring regions (similar to Figures 13, 14, and 16). The need for also displaying the margin of error follows if a survey EDA system, such as described in the previous paragraph, were to be used to display survey data geographically and allow for "drill down" into successively smaller geographic regions. As one looks at successively smaller regions, the margin of error will increase as the sample size is reduced, and it is important for the user to have some indication of this increasing uncertainty in the observed estimate.

We are currently conducting research into this question and are in the process of conducting experiments to understand which designs and design elements work best at conveying the information. The goal is to produce a visualization that appropriately communicates the information without much study, and one that is intuitively understandable to the average consumer of the information. Figure 19 illustrates one likely solution, where the point estimate values are the color levels (as is the convention with showing survey data on maps) and the uncertainty is displayed via the textures. For this example, larger hexagonal shapes represent higher uncertainty in the color value and smaller shapes represent lower uncertainty (where the lack of texture indicates no uncertainty).

Finally, it is important to note that, while this discussion has focused on the visualization of survey data, accounting for how respondents visually interact with self-administered surveys is critical for good data collection. After all, there is no point in stressing good visualization of data if that data was collected poorly. See Dillman *et al.* (2009) and Couper (2008) for visual design guidelines for paper and web-based surveys, respectively.

2.5 Linguistic Analysis Visualization

Linguistic data – free-form text from literature searches, survey responses, incident reports, product descriptions, web pages, blog posts, and a multitude of other sources – is as difficult to analyze as it is widely available. Although years of work have gone into methods of analyzing text, it is fair to say that this field is still immature. Unlike structured data, linguistic information is often complex, subtle or even ambiguous, relying on advance knowledge on the part of the reader.

The two major directions of linguistic analysis have been in natural language processing and computational linquistics. A third area, keyword search and retrieval, is well-known to users as the foundation for web search. This technique is, at least for the moment, more mechanical and less analytical. As a simple example that shows both the limitation of keyword search and the sorts of difficulties facing researchers using linguistic data. Consider the two sentences "Emperor Napoleon drove his men towards Moscow in 1812" and "Four years into the Peninsular War, General Bonaparte led the French army into Russia." These express almost exactly the same idea, and every reader knows that "Emperor Napoleon" and "General Bonaparte" refer to the same person, but because the two sentences use no words in common, retrieval techniques will not detect their similarity.

Natural language processing (NLP) is the study of extracting meaning from text, perhaps by identifying parts of speech, named persons or places, or identifying key phrases. Computational linguistics examines the statistical properties of a large group of documents. These might include, for example, frequency distributions over a set of important words. The two approaches are somewhat different in their approach, but from the end users' perspective, they both try to perform the same set of tasks on documents. These tasks include classification (of documents into groups with known content, e.g. spam detection), clustering (assignment of documents into groups of similar items), and sentiment recognition (determining whether a set of blog posts, say, shows a general approval or disapproval of a person or idea).

We should note here that our discussion, like much research in language processing, will focus on English, in which a number of pre-processing steps are generally performed. The first is the removal of "stop words," that is, words that do not necessarily carry inherent meaning, like "the," "of," and "at." (Of course, stop words are sometimes important, as when naming the band The Who or preserving phrases like "to be or not to be.") A second pre-processing step is "stemming," in which varying word endings are removed so that, for example, "weapon," "weapons," and "weaponry" are all reduced to the same word. In this step irregular forms like "bought" might also be converted to their regular forms. In languages with upper- and lower-case letters like English, words will normally all be converted to one case. These pre-processing steps are almost always performed early in the analysis process by dedicated software provided with suitable word lists. Pre-processing will, of course, need to be quite different in other languages, although some steps – for example, the removal of words that are very rare or very common – may be necessary in any language.

As a final preliminary, we observe that many of the techniques for analyzing and visualizing linguistic data require measures of similarity or "distance" between documents (Weiss *et al.*, 2005, Chapter 2). These distances require some computation and will generally be performed inside software; for our purposes, we can think of these techniques as representing each document in a two- or three-dimensional space. The analysis then proceeds from

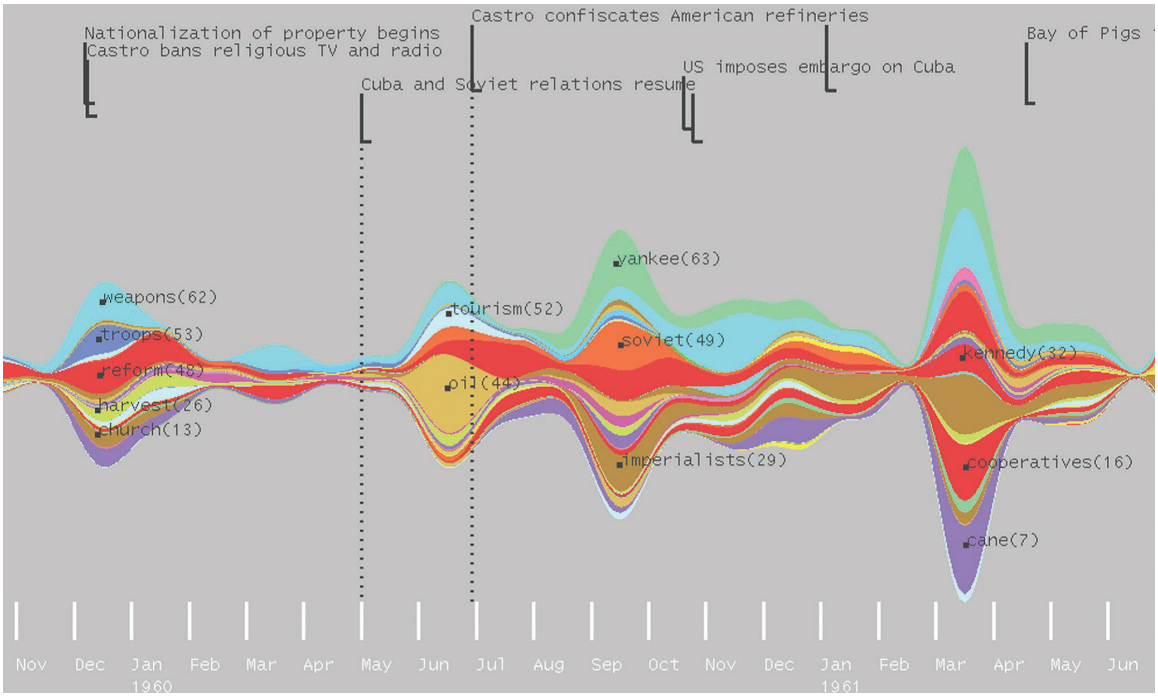

Figure 20: A "Theme River" showing the strength of a set of thirteen themes in Fidel Castro's speeches, interviews and articles from November 1959 to June 1961. Source: Havre *et al.* (2002).

there.

There is no shortage of tools for acquiring and processing text. Krallinger & Valencia (2005) lists several dozen just in the field of molecular biology, and the medical and biological fields seem to be particularly rich in research in this area, perhaps because of the large volumes of literature they generate. The development of techniques specifically for visualizing, though, seems to be in its infancy. Instead, visualizations exploit existing approaches for the tasks at hand. A simple histogram might show word frequencies, for example.

The trademarked name "ThemeRiver" forms a sequence of histograms to show the changing distribution of keywords in a series of documents over time. Figure 20 shows one taken from Havre *et al.* (2002). The horizontal axis shows time from November 1960 to June 1961, and the vertical scale shows the number of occurrences of a number of themes in the speeches, interviews and articles of Fidel Castro. (The extraction of thematic content seems to have been done manually.) The different colors represent the different themes, with each band's thickness proportional to the strength of that theme.

Perhaps the most common text mining task is clustering. In these cases, it is common to see a tree diagram (or "dendrogram") when the number of documents is small. For example, Fleuren *et al.* (2013) used the publicly available CoPub tool Fleuren *et al.* (2011) to cluster abstracts from the Medline database that related to insulin resistance. The picture in Figure 21 shows the clustering, together with the authors' additions of categories (large letters, "Blood," "Cancer," etc.) of abstracts from the Medline database that related to insulin resistance. (We note that the graph is difficult to read even in the original, and

that the numbers relate to part of the analysis that is not directly relevant here.) This technique of hierarchical clustering involves measuring the "distance" from each document to all the others, and then successively joining whatever pair of documents or clusters are closest together.

A second, widely used form of clustering is partitioning, of which the well-known k-means approach provides an example. This, too, has been used on documents with some success. One interesting example comes from Skupin (2004). Here a corpus of some 2,200 conference abstracts were divided into 25 clusters using k-means. The left panel of Figure 22 shows the positioning of the documents along two axes (it is not entirely clear what those axes are); the cluster boundaries would divide the panel up into 25 pieces, if they were shown. The colors, taken from standard geographical mapping, show the density of documents found in each region of the cluster space. So a large number of documents cluster near the labels "management" and "land" at the upper center of the picture. (The labels, it should be noted, are placed automatically, using algorithms like the ones used in geographic information systems to prevent overlapping.)

In the right panel, the author has performed what he calls "semantic zooming." A portion of the left panel is shown, on a magnified scale, but in the right panel 100 clusters are used, presumably producing a finer level of detail. In this way the analyst can interact with a graph like this to locate areas on the map – that is, sets of documents in the original corpus – of particular interest. The commercial product ThemeScape produces maps like these, which it calls "content maps."

Rather than cluster them, it is possible to represent documents directly. Imagine representing a document by a point in a high-dimensional space whose axes are, for example, the frequencies of certain keyworks. As a concrete example, suppose we score each document based on the number of occurrences of the words "cancer," "lung," "emphysema," "bone," and "fracture." Each document could be drawn as a point in a space with five axes, where each axis might measure the number of times the corresponding word appears. (In fact a more common approach is to give words in a document weights, having to do with how frequently they appear in that document and in the whole set of documents.) The statistical techniques of multidimensional scaling provide a way to reduce the dimensionality of a set of points onto a two-dimensional plane (or into three dimensions), while preserving the distances between them to the greatest extent possible. Examination of the two-dimensional projection can then give insight into the structure of the original five dimensions.

Of course, the dimensionality of the original data might be very much higher than the five in our example. Figure 23 shows an example from Lopes *et al.* (2007). Here the authors have represented a set of 574 articles on three subjects in a high-dimensional space and then projected those points into two dimensions. (The labels in the rectangles can be ignored here.) The authors colored the points manually, according to the subject of the document. The number of mismatches (blue points in mostly red areas, for example, which may not be visible in the black and white figure reproduction) is small, suggesting that this technique could lead to useful rules for detecting content.

One visualization specific to linguistic data is the word cloud (or "wordle"). This is used to "to present a visual overview of a collection of text" (Viégas & Wattenberg, 2008, p. 49). In particular, words that appear more frequently in the text appear more prominently in the cloud; the cloud-building software arranges placement and orientation. Figure 24 illustrates a word cloud that was created from the NPS Operations Research Department's application for the INFORMS Smith Prize. The prize is awarded to the "academic department or

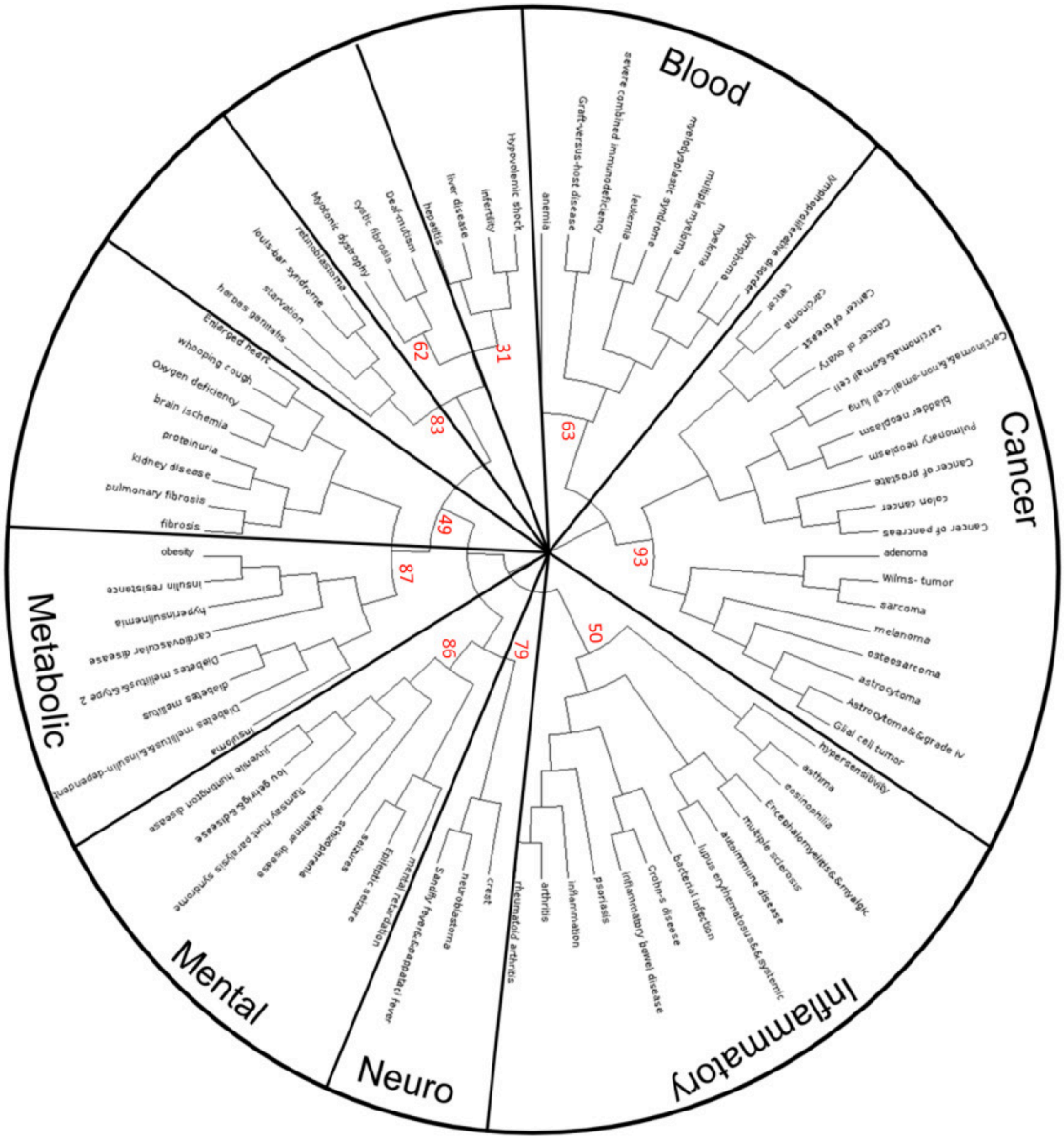

Figure 21: Hierarchical clustering from Medline abstracts. Each document is provided with a "distance" to all others; then the clustering joins documents and clusters according to distance. Source: Fleuren *et al.* (2011).

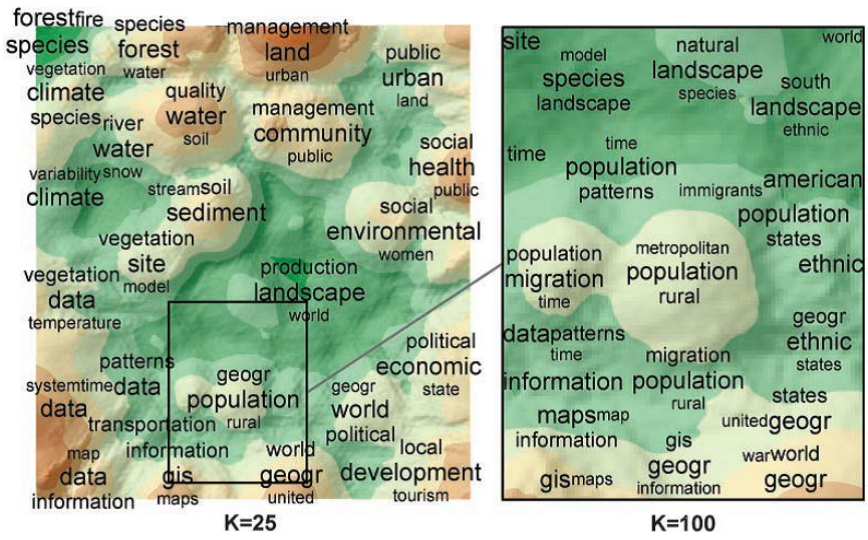

Figure 22: K-means cluster results of 2,200 abstracts. The left panel shows the division of space into 25 clusters (boundaries not shown), on top of which document densities and keywords are shown. The right panel shows a sub-portion of the same region divided into 100 clusters. Source: Skupin (2004).

program for effective and innovative preparation of students to be good practitioners of operations research, management science, or analytics" (INFORMS, 2013). The word cloud in Figure 24 clearly shows the story the department told in its prize application: students and faculty collaborating on research and practical applications. This is the point of the prize and the department subsequently won it using this application. When rendered on some sort of electronic media, word clouds can be made three dimensional, rotating, and/or interactive. Of course, there are no formal analyses associated with a cloud: the cloud is the end result. See Russell (2011, Chapter 9) for additional discussion.

HSBC data is often "streaming," that is, being produced continuously over time. For example, social networks, news aggregators and blog feeds can produce large amounts of data whose character can evolve over time or change suddenly. While some work has been devoted to the mining specifically of streaming data (for example, (Aggarwal & Zhai, 2012, Chapter 6)), this too, and associated visualization remains very much an open field.

2.6 Twitter and Other Social Media Visualization

Social media data can take many forms, and thus there are many possible visualizations, including all of those already discussed. Typically the data is text-based, though the text can arise in a large number of contexts: tweets, Search engine queries, blogs, e-mail, web sites, etc. The data may be cross-sectional, in the sense that it is essentially fixed and there is little change over time, or it may be temporal and perhaps highly dynamic. As social media data, it frequently is dyadic, meaning that it arises as an interaction between two individuals, organizations, or entities.

If the social media data does contain information about interactions, then it can often be analyzed and visualized as a network. For example, Schroeder *et al.* (2012) used network visualizations to explore what was occurring on Twitter during the Egyptian protests of

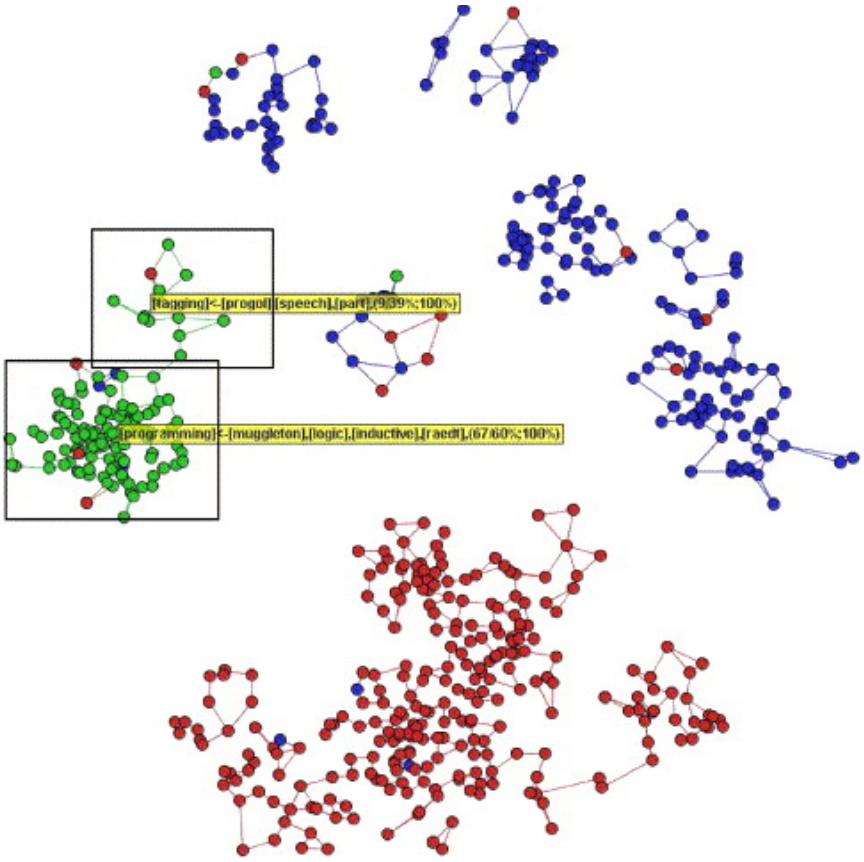

Figure 23: Example of multidimensional scaling. Points from a high-dimensional space built on keyword frequencies are projected into two dimensions in a specific way, producing obvious clusters in content. Source: Lopes *et al.* (2007).

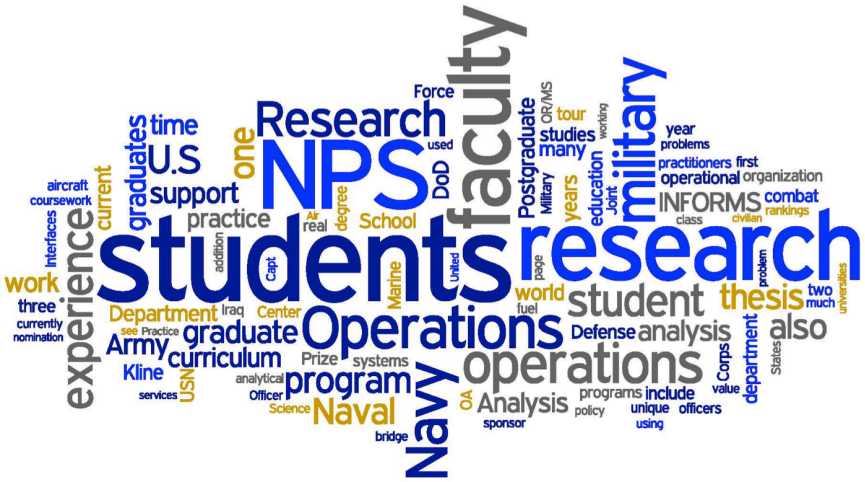

Figure 24: Illustrative word cloud.

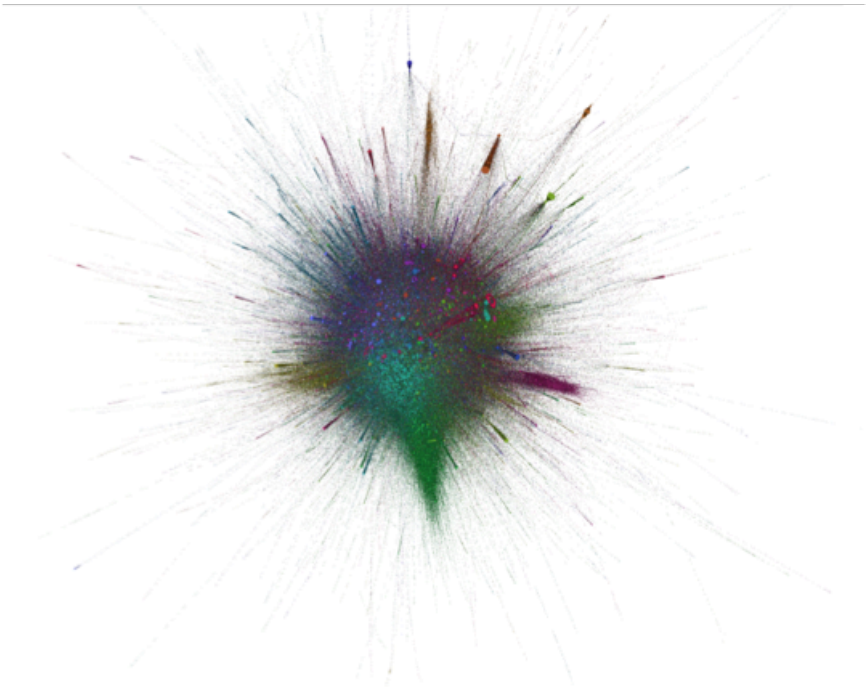

Figure 25: Network visualization of tweets during the Egyptian protests for Twitter data for January 28 and February 4, 2011. Source: Schroeder *et al.* (2012).

2011. In order to understand which users were significant conduits of Twitter information, they analyzed over one million tweets about Egypt from just two days: January 28 and February 4, 2011. They say,

> Using these data, we generated a user-by-user network [Figure 25] where a direct tie was drawn between two users if one of the users sent a message to the other, or a user retweeted the message of another. In the case of the latter, we drew a tie from the author of the original message to the user who "retweeted" the message. In the end, our user-by-user network included 196,670 users with 526,976 ties between them (Schroeder *et al.*, 2012).

However, as discussed in Section 2.2 and as shown in Figure 25, simply plotting an entire network is frequently not particularly informative. Indeed, Figure 25 is really just another "hairball" network visualization in the spirit of Figure 7. Thus, the authors applied an algorithm developed by Blondel *et al.* (2008) to identify distinct clusters (or communities) within the network by based on a partition of users that yields the highest modularity score. As shown in Figure 26, the algorithm identified a number of distinct clusters in the data, including news organization such as Al Arabiya and Al-Jazeera Arabic. In addition to the news organizations, the authors found that a Hosni Mubarak parody account was also quite central and they speculated that its tweets may have been influential in the framing of Egyptian grievances during the revolution (Schroeder *et al.*, 2012).

When analyzing a corpus of information, which may be sets of documents or other types of social media information, one is frequently interested in identifying items that are similar. To this end, Figure 27 illustrates a visualization of some Google+ activities

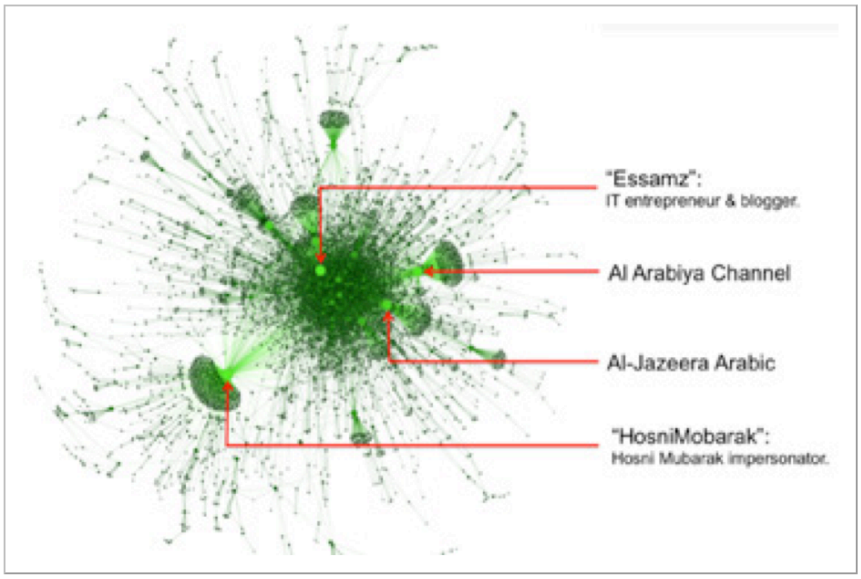

Figure 26: An alternate visualization of the tweet network after clustering users via an algorithm developed by Blondel *et al.* (2008). Source: Schroeder *et al.* (2012).

using the Protovis software[7] (Stanford Visualization Group, 2010) showing the similarity between pairs of activities (Russell, 2011, Chapter 7). In the figure the arcs indicate a linkage (similarity) between nodes, which are the Google+ activities, and where the nodes are scaled according to their degree. The software sorts the nodes to minimize visual clutter and the diagram clearly shows which nodes have the largest number of connections. Note that the diagram is designed to be interactive, where on-line the titles can be omitted because they are displayed when the associated node is moused over. In addition, clicking on a node opens a new browser window that that contains the activity represented by that node.

As previously discussed, signature detection may involve determining, deriving, or otherwise looking for sociocultural signatures or it may involve observing particular signatures to detect if and whether they change over time. Servi & Elson (2013) have worked on the question of trying to detect changes in the mood of social media users. Their method combines Linguistic Inquiry and Word Count (LIWC, 2013) with a mathematical algorithm to follow trends in past and present moods and detect breakpoints where those trends changed abruptly. The results are then plotted, as in Figure 28, to show mood changes over time. This type of temporal change point detection can be further enhanced, assuming the underlying statistic can be characterized probabilistically, by employing statistical process control methods from industrial quality control. For example, see Figures 14-16 in Chapter [**Detection_DataProc_v6.docx**].

Figure 29 illustrates another type of temporal display that, when viewed on-line, is interactive. This particular example is a visualization of important events in Christian and Jewish history from mid-1970s to 1990 (Huynh, 2013). While not a plot of social media *per se* (the data was taken from Wikipedia), the generalization of this type of plot to social

[7]Protovis has been supplanted by D3.js (Bostock, 2012), which provides improved support for animation and interaction, but the software is still available. D3.js builds on many of the concepts in Protovis.

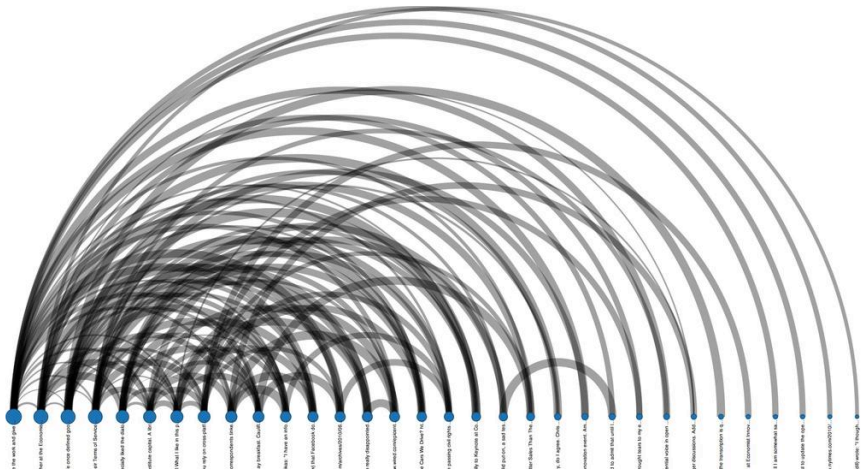

Figure 27: A visualization of the similarity of some Google+ activities. Source: Russell (2011, Chapter 7).

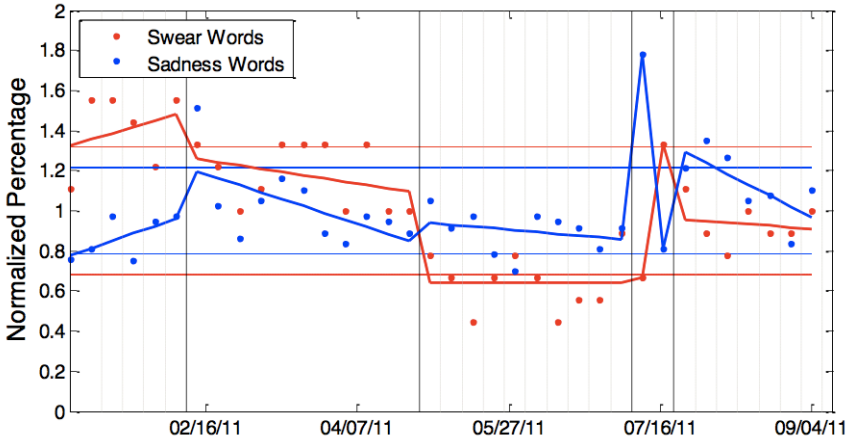

Figure 28: Twitter mood change point visualization, where changes are evident at (at least) four points during the displayed time period. Source: Servi & Elson (2013, Figure 5).

media data is obvious. See, for example Russell (2011, Chapter 3) where the timeline is applied to e-mail data.

Analysis of social media can also focus on social media meta-data. For example, Figure 30 is a word cloud (similar in spirit to Figure 24 but different in layout) of Flikr's most popular photo tags. As discussed in Viégas & Wattenberg (2008, p. 50) the Flikr "tag cloud" provides "an instant overview of the site's pictures."

Donath (2002, p. 49) makes the argument that "all visualizations will have some evocative quality. We do not think in pure abstractions; rather, our thinking is metaphoric and grounded in the spatial world." Thus, good visualizations fit with our naturally developed intuitions. For example, in Figure 31 Donath (2002) uses a garden metaphor to visualize participation on a message board. Based on the work of Xiong & Donuth (1999), in this "PeopleGarden" individual participants are each represented by a flower. The longer they have been involved in the message board, the longer the flower stem, while the more they have posted to the message board, the more petals on their flower. Initial postings are depicted in a different color from replies.

3 Discussion & Conclusions

This chapter has focused on methods for visualizing many different types of data. Each of the visualization methods is appropriate for a specific type of data, sometimes in a particular situation, and may be more or less useful for signature detection deepening on how and when employed. Broadly speaking, signature detection involves identifying either differences between subsets of data, say geographically or demographically, or changes that occur over time. Our emphasis in this chapter on exploratory data analysis is purposeful because it is impossible (at least within the constraints of a book chapter) to give a comprehensive treatment of all possible detection visualization strategies. Furthermore, while good detection strategies may be obvious in retrospect, unless the signature one is looking for is well understood and well defined, successful prospective detection is likely to require the ability to explore and search through data in multiple ways.

As we have shown in this chapter, there are a large number of visualizations relevant to sociocultural signature detection. Indeed, given the limitations of the chapter, in many ways we have only scratched the surface, particularly in terms of the possible variants of the visualizations shown herein. Furthermore, with the the ubiquitous availability of significant computing power and sophisticated software, there is a lot of innovation currently going on in visualization. Some of it results in eye candy that is not particularly well suited for good information and data communication, while others are resulting in very effective communication and research methods. Separating the former from the latter will become increasingly important. As Steve Jobs said,

> Most people make the mistake of thinking design is what it looks like... People think it's this veneer – that the designers are handed this box and told, 'Make it look good!' That's not what we think design is. It's not just what it looks like and feels like. Design is how it works (Walker, 2003).

Jobs' point is equally applicable to the design of an Apple product as it is to the design of visualization methods. It's not sufficient that a visualization look good, it must also work well for communicating information. For additional discussions about information visualization, including open challenges, see Chen (2005) and Fayyad *et al.* (2001).

Figure 29: SIMILE timeline visualization of important events in Christian and Jewish history from mid-1970s to 1990. Source: Huynh (2013).

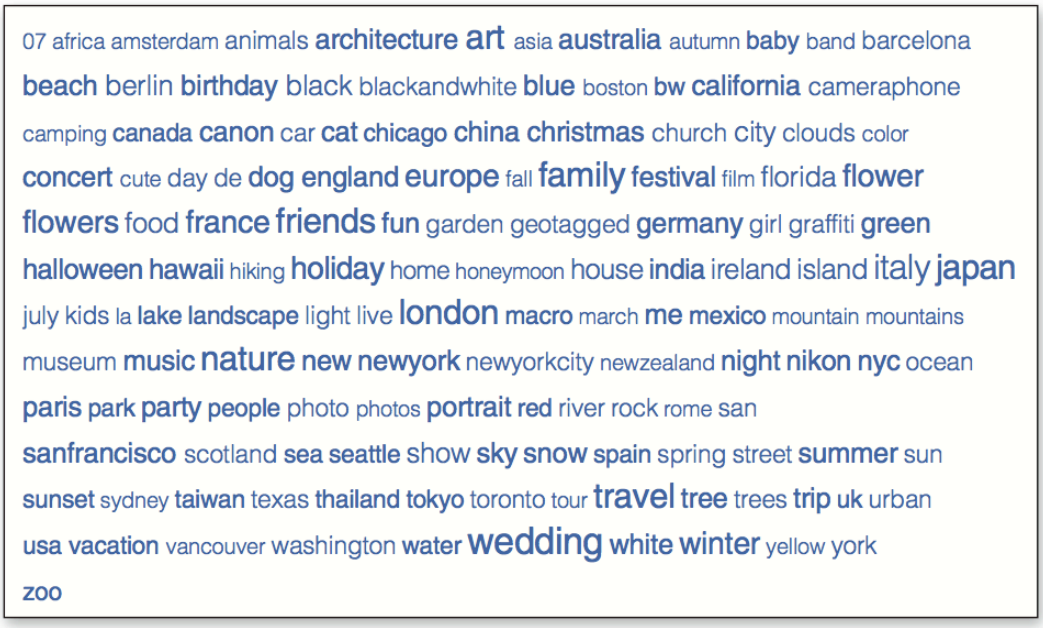

Figure 30: Example of a tag cloud visualization: Flikr's most popular tags. Source: Viégas & Wattenberg (2008, Figure 3).

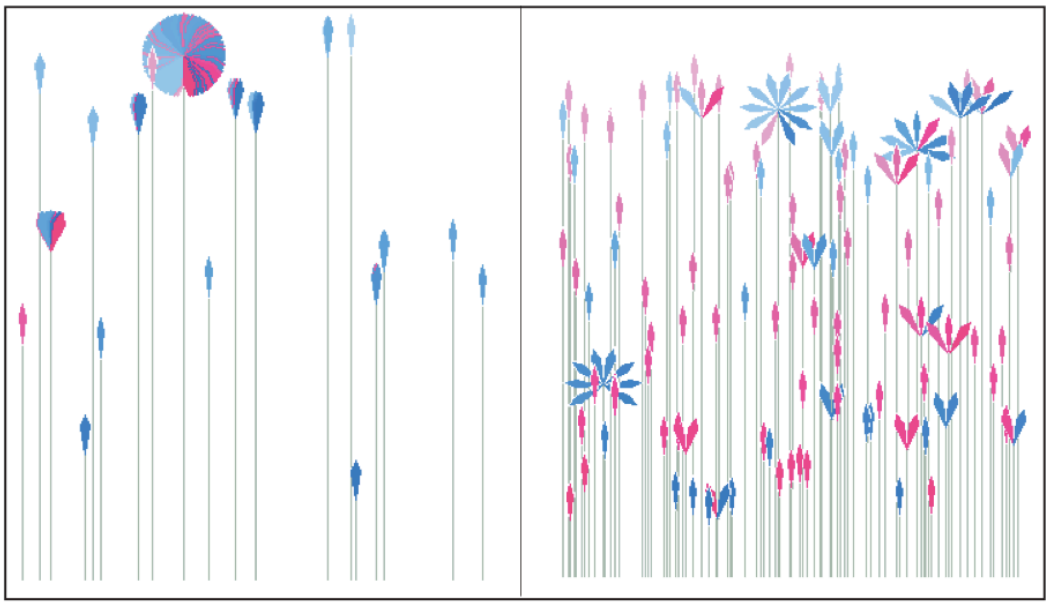

Figure 31: Example of a PeopleGarden visualization. Each flower represents a participant on a message board. Length of stem corresponds to longer involvement in the message board and the number of flower petals corresponds to the number of postings. Source: Donath (2002, Figure 2).

While there are quite a number of visualization methods that are commonly accepted and understood, this may not be the case with the new types of data visualization that are just emerging. And even with existing methods, if the goal is efficient and effective display of information, it is important to understand the strengths and limitations of of each type of visualization. For example, in spite of its ubiquity, the pie chart is much poorer method for visually comparing between groups than the bar chart because human beings are better at accurately visually comparing lengths than areas and angles (Gemignani, 2006).

A key point, as Ben Shneiderman has said, is that "the purpose of visualization is insight, not pictures" (Card *et al.*, 1999, p. 6). That is, the goal of any good visualization method should be accurate perception and comprehension that follows when the viewer correctly understands and interprets the information encoded in a visualization. Just because a visualization looks fancy or high tech or "cool" does not mean that it is accurately or efficiently communicating information. Hence, particularly when looking to design new methods, it is essential that those methods are carefully evaluated to ensure they correctly convey to the viewer what is intended as well as to identify improvements or better visualization methods.

3.1 The Leading Edge: Signature Detection Visualization Systems

With the ongoing explosion of data, particularly social media data, and the proliferation of mobile computing devices, many types of data collection, analysis, and display systems are under development. Given the limited space in this chapter, we will illustrate these with one example: the Marine Civil Information Management System (MARCIMS).

MARCIMS leverages mobile computing technology and geospatially-aware semantic knowledge management tools to support of United States Marine Corps (USMC) civil affairs (CA) personnel. As illustrated in Figure 32, MARCIMS facilitates sharing, organizing, analyzing and visualizing field-collected data in the context of sociocultural and geospatial knowledge. It is being developed as an unclassified system capable of being accessed by coalition, non-governmental organizations (NGOs), and intergovernmental organizations (IGOs). The goal is to provide the military commander and other decision makers with the ability to visualize data in both geospatial and other graphical representations (USMC, 2013b).

Figure 33 is an illustration of a tablet input device and some of the data visualizations currently part of MARCIMS. The system is designed for individuals in the field to input and upload data via mobile devices such as smartphones and tablets. Then, via these same devices, they can query and display the data in a variety of formats, including simple statistical displays such as bar charts and pie charts as well as various geospatial displays of the data.

The near-term objective of MARCIMS is to modernize current Marine Corps civil affairs data collection, storage, analysis, and dissemination. In the longer-term, it may revolutionize how civil affairs teams work. In terms of visualizing sociocultural signatures, MARCIMS presents sociocultural data in ways that a U.S. Marine can can use the result to gain insight into subgroups of people in a region/society/culture. While perhaps less than what a social scientist might desire for research, the system provides an excellent baseline capability to which additional algorithmic or visualization methods could be added. And, regardless of whether socio-cultural signature detection is of interest to the Marine Corps, MARCIMS is a nice illustration of what can be done in this realm with current technology.

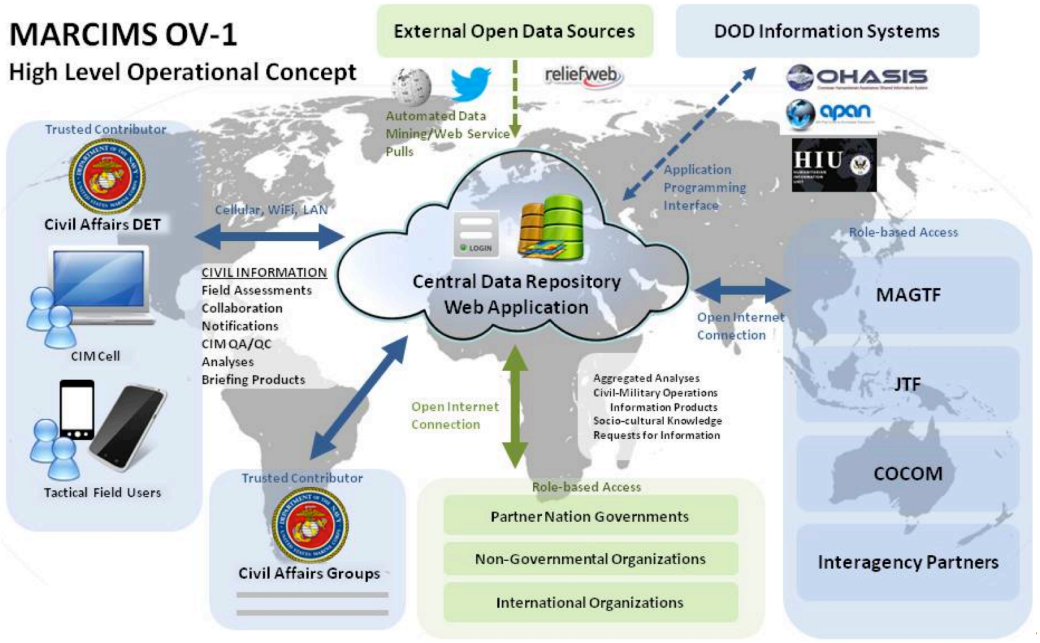

Figure 32: MARCIMS' "high level operational concept." Source: USMC (2013a).

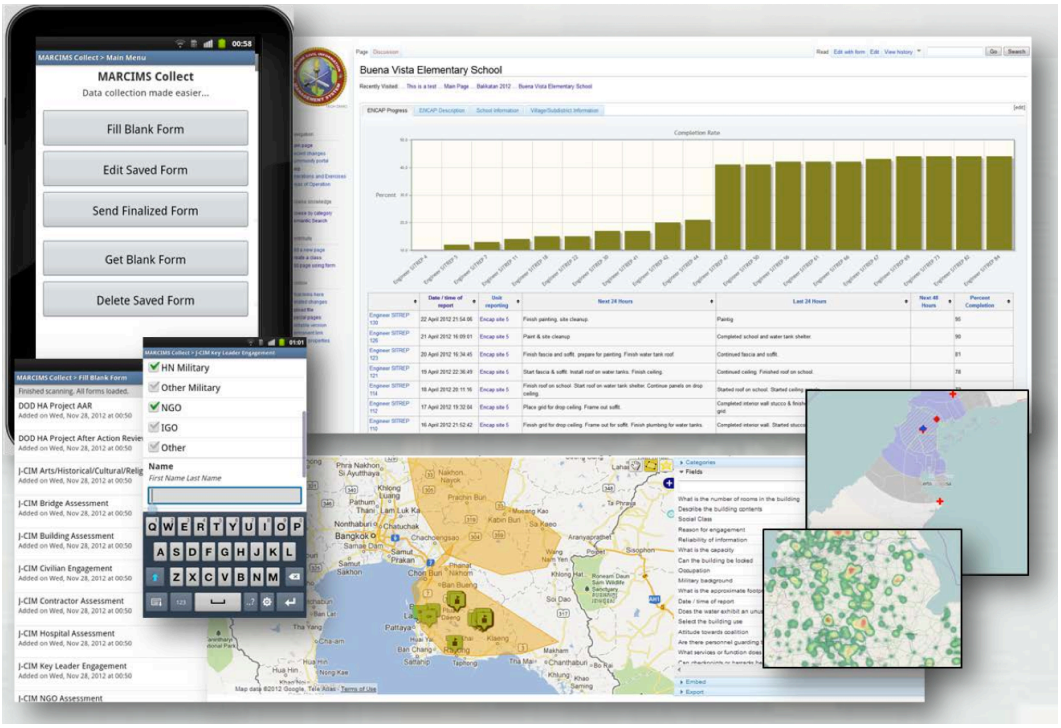

Figure 33: An illustration of a tablet input device and some of the data visualizations currently part of MARCIMS. Source: Clark (2013).

3.2 Signature Detection Visualization Research Challenges

There are a number of research challenges in the field of visualization, particularly as applied to socio-cultural signature detection. Perhaps the most important challenge is simply identifying which types of signatures and changes to signatures are important to detect. Without such information, signature detection becomes an unstructured and unbounded data exploration exercise in which it is highly likely that something "unusual" will be "detected." This brings us full-circle back to the cautions of Section 1.3.

Quantitatively-trained social scientists often address this type of problem by, for example, distinguishing between exploratory and confirmatory hypothesis testing. In particular, they employ confirmatory methods as a guard against multiple testing resulting in an increased likelihood of Type I errors. Typically this approach requires the *a priori* specification of one or more hypotheses which are tested against data unobserved at the time of hypothesis specification. Signature detection is similar in the sense that the *a priori* definition of the signatures to be detected will help guard against Type I errors.

Of course, the specification of the signatures is likely to be non-trivial in many cases and highly context dependent. However, to the extent they can be specified, visualization methods can be tailored that are best designed to facilitate detection. An example from the physical sciences is the discovery of Pluto in 1930 by Clyde Tombaugh. Prior to its discovery, astronomers predicted that there was a planet out beyond Neptune. Given its expected location:

> Tombaugh used the observatory's 13-inch astrograph to take photographs of the same section of sky several nights apart. He then used a blink comparator to compare the different images. When he shifted between the two images, a moving object, such as a planet, would appear to jump from one position to another, while the more distant objects such as stars would appear stationary. Tombaugh noticed such a moving object in his search, near the place predicted by Lowell, and subsequent observations showed it to have an orbit beyond that of Neptune (Wikipedia, 2013).

In this example, the visualization detection methodology was driven by the type of available data (photographs) and the expected signature (motion against a stationary background).[8]

Other research challenges may include legal and regulatory challenges, for example, in order to gain access to data in order to do visualization; technological challenges related to designing and implementing computer hardware and software necessary for displaying the data in a visualization; perhaps ethical issues related to managing and safeguarding the data, particularly if it is sensitive; analytical and algorithmic challenges necessary for developing and displaying the visualizations; and the managerial challenges of effectively assembling and operating an entire system, particularly if the visualizations are depending on massive and/or disparate data. Many of these challenges are described in the other chapters of this book.

Finally, simply continuing to improve upon existing visualization methods is an important research challenge. Better visualizations require innovation in the way data is visually presented *and* subsequent careful evaluation of the those methods to ensure that they work. Whether a visualization method "works" encompasses a number of dimensions, including

[8] Our thanks to an anonymous reviewer who suggested this example.

whether the visualization accurately presents the information, whether the viewer can subsequently accurately retrieve the information from the visualization, whether the visualization is intuitive and easy to interpret, etc.

Many of the classical visualization methods of Section 2.1 have been evaluated over the years using formal experiments and/or Darwinian selection. For discussions about how to experimentally evaluate visualization methods and techniques, see for example Carpendale (2008) and Cleveland & McGill (1984). With more complicated visualizations, and particularly for evaluating software used for EDA and visualization, some propose a case study approach rather than controlled experiments in a laboratory (Perer & Shneiderman, 2006).

Often, for example, effective signature detection requires more than a simple static visualization. It may require a software system with an interface design that allows the user to easily and appropriately explore the data. Important features include the ability to "drill down" into the data for details in order to, for example, facilitate easy identification and analysis of subgroups and dynamic and interactive graphics, and the ability to "tour" through the data, particularly higher dimensional data (Fricker, 2013, p. 109). With such systems Perer & Shneiderman (2009, p. 42) say,

> ...laboratory-based controlled experiments are less compelling for information visualization and visual-analytics (VA) research. VA systems are often designed for domain experts who work for days and weeks to carry out exploratory data analysis on substantial problems. Tthese types of tasks are nearly impossible to reconstruct in a controlled experiment for a variety of reasons.

Ultimately the point is that, regardless of how, new visualization methods and software should be evaluated. As most, if not all, of this work is now done via some type of computer-based system, these evaluations may also extend into the realm of human-computer interaction evaluations as well (see, for example, de Graaff, 2007, and Hunt, 2013).

3.3 Concluding Thoughts

For social scientists working with sociocultural data, this is an revolutionary time to be conducting research. As Michael Macy says, "Human beings around the globe are now communicating with each other using devices that record those interactions and have open access. I think this is an extraordinarily exciting moment in the behavioral and social sciences" (Miller, 2011, p. 1814). Furthermore, the era when research data sets were small enough that one could actually look though the raw data and learn something from it is long past. Today data, in a wide variety of formats and types, is being captured and stored at a rate that makes effective visualization perhaps the only way to effectively look at data. Welcome to the era of data visualization!

References

Aggarwal, C.C., & Zhai, C.X. (eds). 2012. *Mining Text Data*. New York: Springer.

Anscombe, Francis. 1973. Graphs in Statistical Analysis. *American Statistician*, **27**(1), 17–21.

Blondel, Vincent D., Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics*. arXiv:0803.0476v2 [physics.soc-ph]. Accessed at `http://arxiv.org/abs/0803.0476` on April 12, 2013.

Bolstad, Paul. 2008. *GIS Fundamentals*. 3rd edn. White Bear Lake, Minnesota: Elder Press.

Bostock, Michael. 2012. *D3 website*. Accessed on-line at `http://d3js.org/` on April 13, 2013.

Brinton, Willard Cope. 1939. *Graphic Presentation*. Brinton Associates. Accessed on-line at `http://www.archive.org/stream/graphicpresentat00brinrich#page/2/mode/2up` on April 13, 2013.

Card, Stuart K., Mackinlay, Jock D., & Shneiderman, Ben (eds). 1999. *Readings in Information Visualization: Using Vision to Think*. Academic Press.

Carpendale, Sheelagh. 2008. Evaluating Information Visualizations. *Pages 19–45 of: Lecture Notes in Computer Science Volume 4950*. Springer.

Carter, C.J. 2008 (November). *Track Election Night 2008 With This Electoral Cartogram*. http://tib.cjcs.com/1374/track-election-night-2008-with-this-electoral-cartogram/.

Chen, Chaomei. 2005. Top 10 Unsolved Information Visualization Problems. *Computer Graphics and Applications, IEEE*, **25**(4), 12–16.

Clark, Tim. 2013. *MARCIMS: Managing USMC Civil Information briefing*. Provided by Mr. Joseph M. Watts on April 9, 2013.

Cleveland, William S. 1993. *Visualizing Data*. Hobart Press.

Cleveland, William S. 1994. *The Elements of Graphing Data*. 2nd edn. Hobart Press.

Cleveland, William S., & Devlin, Susan J. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

Cleveland, William S., & McGill, Robert. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, **79**(387), 531–554.

Cole, Adam. 2012 (November). *It's All Politics website: A Campaign Map, Morphed By Money*. Accessed at `www.npr.org/blogs/itsallpolitics/2012/11/01/163632378/a-campaign-map-morphed-by-money` on April 16, 2013.

Coscia, Michele. 2013. *Michele Coscia – Connecting Humanities website*. Accessed on-line at `http://www.michelecoscia.com/` on April 13, 2013.

Couper, Mick P. 2008. *Designing Effective Web Surveys*. Cambridge University Press.

Courtney, Karen L. 2005. Visualizing Nursing Workforce Distribution: Policy Evaluation using Geographic Information Systems. *International Journal of Medical Informatics*, **74**, 980–988.

de Graaff, Hans. 2007. *HCI index website: Tools*. Accessed on-line at `http://degraaff.org/hci/tools.html` on April 17, 2013.

Dillman, Don A., Smyth, Jolene D., & Christian, Leah Melani. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. 3rd edn. Wiley.

Donath, Judith. 2002. A Semantic Approach to Visualizing Online Conversations. *Communications of the ACM*, **45**(4), 45–49. Accessed on-line at `http://dl.acm.org/citation.cfm?id=505271` on April 10, 2013.

Doyle, Arthur Conan. 2003. A Scandal in Bohemia. *In: The Complete Sherlock Holmes, Volume I*. Barnes & Noble Classics.

Fayyad, Usama, Grinstein, Georges G., & Wierse, Andreas (eds). 2001. *Information Visualization in Data Mining and Knowledge Discovery*. 1st edn. Morgan Kaufmann.

Fleuren, Wilco W.M., Verhoeven, Stefan, Frijters, Raoul, Heupers, B., Polman, J., van Schaik, René, de Vlieg, Jacob, & Alkema, Wynand. 2011. CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids*, **39 (suppl 2)**, W450–W454.

Fleuren, Wilco W.M., Toonen, Erik J.M., Verhoeven, Stefan, Frijters, Raoul, Hulsen, Tim, Rullmann, Ton, van Schaik, René, de Vlieg, Jacob, & Alkema, Wynand. 2013. Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining. *BioData Mining*, **6**. Accessed at `www.biodatamining.org/content/6/1/2` on April 18, 2013.

Fricker, Jr., Ronald D. 2013. *Introduction to Statistical Methods for Biosurveillance*. 1st edn. Cambridge University Press.

Gemignani, Zach. 2006. *JuiceAnalytics website: The Problem with Pie Charts*. Accessed on-line at `www.juiceanalytics.com/writing/the-problem-with-pie-charts/` on April 13, 2013.

Havre, Susan, Hetzler, E., Whitney, P., & Nowell, L. 2002. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, **8**, 9–20. doi:10.1109/2945.981848.

Healey, C., Kocherlakota, S., Rao, V., Mehta, R., & St. Amant, R. 2008. Visual perception and mixed- initiative interaction for assisted visualization design. *IEEE Transactions on Visualization and Computer Graphics*, **14**(Mar-Apr), 396–411.

Hoopes, John W. 2011. *11-11-11, Apophenia, and the Meaning of Life*. Accessed on-line at `www.psychologytoday.com/blog/reality-check/201111/11-11-11-apophenia-and-the-meaning-life` on April 6, 2013.

Hunt, William. 2013. *HCI Tools Online website*. Accessed on-line at `http://degraaff.org/hci/tools.html` on April 17, 2013.

Huynh, David F. 2013. *Timeline: Web Widget for Visualizing Temporal Data website*. Accessed on-line at `www.simile-widgets.org/timeline/` on April 12, 2013.

INFORMS. 2013. *INFORMS website: UPS George D. Smith Prize*. Accessed on-line at `www.informs.org/Recognize-Excellence/INFORMS-Prizes-Awards/UPS-George-D.-Smith-Prize` on April 12, 2013.

Krallinger, Martin, & Valencia, Alfonso. 2005. Text-mining and information-retrieval services for molecular biology. *Genome Biology*, **6**. doi:10.1186/gb-2005-6-7-224.

Krzywinski, Martin. 2013. *Hive Plots website*. Accessed on-line at `www.hiveplot.com` on April 12, 2013.

Livingston, Mark A., & Decker, Jonathan W. 2011. Evaluation of Trend Localization with Multi-Variate Visualization. *IEEE Transactions on Visualization and Computer Graphics*, **17**(12), 2053–2062.

Livingston, Mark A., & Decker, Jonathan W. 2012. Evaluation of Multi-variate Visualizations: A Case Study of Refinements and User Experience. *SPIE Visualization and Data Analysis*, 23-25 Jan. Burlingame, CA.

Livingston, Mark A., Decker, Jonathan, & Ai, Zhuming. 2011. An Evaluation of Methods for Encoding Multiple, 2D Spatial Data. *SPIE Visualization and Data Analysis*, 24-25 Jan. Burlingame, CA.

Livingston, Mark A., Decker, Jonathan W., & Ai, Zhuming. 2012. Evaluation of Multivariate Visualization on a Multivariate Task. *IEEE Transactions on Visualization and Computer Graphics*, **18**(12), 2114–2121.

Livingston, Mark A., Decker, Jonathan W., & Ai, Zhuming. 2013. Evaluating Multivariate Visualizations on Time-Varying Data. *Proceedings of SPIE Visualization and Data Analysis*, **8654**(03-07 Feb). Burlingame, CA.

LIWC. 2013. *Linguistic Inquiry and Word Count website.* Accessed on-line at `http://www.liwc.net` on April 12, 2013.

Lopes, A.A., Pinho, R., Paulovich, F.V., & Minghim, R. 2007. Visual text mining using association rules. *Computers and Graphics*, **31**, 316–316. Accessed on-line at `www.sciencedirect.com/science/article/pii/S0097849307000544#bbib12` on April 19, 2013.

McCue, C., Hildebrandt, W., & Campbell, J.K. 2012. Pattern analysis of the Lords Resistance Army and Internally Displaced Persons. *Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter*, **12**(9).

McCune, Doug. 2010 (June). *If San Francisco Crime Were Elevation website.* Accessed at `http://dougmccune.com/blog/2010/06/05/if-san-francisco-crime-was-elevation/` on April 17, 2013.

Miller, Greg. 2011. Social Scientists Wade into the Tweet Stream. *Science*, **333**, 1814–1815.

Mobio. 2013. *Visualizing.org website.* Accessed at `www.visualizing.org/visualizations/data-visualization-resources-network` on April 13, 2013.

Murrell, Paul. 2011. *R Graphics, Second Edition.* 2nd edn. CRC Press.

NIST. 2012. *NIST website: What is EDA?* Accessed on-line at `www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm` on June 13, 2012.

Perer, Adam, & Shneiderman, Ben. 2006. Balancing Systematic and Flexible Exploration of Social Networks. *IEEE Trans. Visualization and Computer Graphics*, **12**(5), 693–700. Accessed on-line at `http://hcil2.cs.umd.edu/trs/2006-25/2006-25.pdf` on April 10, 2013.

Perer, Adam, & Shneiderman, Ben. 2009. Integrating Statistics and Visualization for Exploratory Power: From Long-Term Case Studies to Design Guidelines. *IEEE Computer Graphics and Applications*, **29**(3), 39–51. Accessed on-line at `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4909117&tag=1` on April 10, 2013.

Rix, Joseph D., & Fricker, Jr., Ronald D. 2012. Displaying Survey-based HSCB Data for Decision-makers. Presented at the 2nd International Conference on Cross-cultural Decision Making. Accessed on-line at `http://faculty.nps.edu/rdfricke/presentations/Rix-Fricker.pdf` on April 8, 2013.

Robbins, Naomi B., & Heiberger, Richard M. 2011. Plotting Likert and Other Rating Scales. *Pages 1058–1066 of: Proceedings of the 2011 Joint Statistical Meetings*. American Statistical Association.

Rosling, H. 2013. *GapMinder World*. Accessed on-line at `www.gapminder.org` on April 20, 2013.

Russell, Matthew A. 2011. *Mining the Social Web*. O'Rilley.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. Springer.

Schroeder, Rob, Everton, Sean, & Shepherd, Russell. 2012. Mining Twitter Data from the Arab Spring. *CTX*, **2**(4). Accessed on-line at `https://globalecco.org/mining-twitter-data-from-the-arab-spring#All` on April 12, 2013.

Servi, Les, & Elson, Sara Beth. 2013. A Mathematical Approach to Identifying and Forecasting Shifts in the Mood of Social Media Users. *American Behavior Scientist Journal*. In submission.

Shakespeare, William. 1936. The Tragedy of Hamlet, Prince of Denmark. *In:* Morley, Christopher (ed), *The Complete Works of William Shakespeare*. New York: Doubleday & Company, Inc. Act 3, scene 2.

Shermer, Michael. 2008. Patternicity. *Scientific American*, November, 48. Accessed on-line at `www.michaelshermer.com/2008/12/patternicity/` on April 5, 2013.

Silver, Nate. 2012. *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't.* New York: The Penguin Press.

Skupin, Andrè. 2004. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, **101**(April).

Smith, Marc A., Shneiderman, Ben, Milic-Frayling, Natasa, Rodrigues, Eduarda Mendes, Barash, Vladimir, Dunne, Cody, Capone, Tony, Perer, Adam, & Gleave, Eric. 2009. Analyzing (Social Media) Networks with NodeXL. *Pages 255–263 of: Proceedings of the 4th International Conference on Communities and Technologies*. Accessed on-line at `http://dl.acm.org/citation.cfm?id=1556497` on April 10, 2013.

Stanford Visualization Group. 2010. *Protovis github website*. Accessed on-line at `http://mbostock.github.io/protovis/` on April 13, 2013.

Steele, Julie, & Illinsky, Noah. 2010. *Beautiful Visualization: Looking at Data through the Eyes of Experts*. O'Reilly Media.

Sviokla, John. 2009. *Swimming in Data? Three Benefits of Visualization*. Accessed on-line at `http://blogs.hbr.org/sviokla/2009/12/swimming_in_data_three_benefit.html` on April 6, 2013.

Tufte, Edward R. 1986. *The Visual Display of Quantitative Information*. 1st edn. Graphics Press.

Tufte, Edward R. 1990. *Envisioning Information*. Graphics Press.

Tufte, Edward R. 1997. *Visual Explanations*. Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd edn. Graphics Press.

Tufte, Edward R. 2006. *Beautiful Evidence*. Graphics Press.

Tukey, John W. 1977. *Exploratory Data Analysis.* Addison-Wesley.

U.S. Government. 2013. *HSCB website: OSD Human Social Culture Behavior Modeling Program.* Accessed on-line at `www.dtic.mil/biosys/capabilities.html` on April 7, 2013.

USMC. 2013a. *Marine Civil Information Management System (MARCIMS) System Information Brief, rev. 4.0.* Provided by Mr. Joseph M. Watts on April 9, 2013.

USMC. 2013b. *Marine Civil Information Management System (MARCIMS) whitepaper.* Provided by Mr. Joseph M. Watts on April 8, 2013.

Viégas, Fernanda B., & Donuth, Judith. 2004. Social Network Visualization: Can We Go Beyond the Graph? *Pages 6–10 of: Workshop on Social Networks for Design and Analysis: Using Network Information in CSCW*, vol. 4. Accessed on-line at `http://alumni.media.mit.edu/~fviegas/papers/viegas-cscw04.pdf` on April 10, 2013.

Viégas, Fernanda B., & Wattenberg, Martin. 2008. Tag Clouds and the Case for Vernacular Visualization. *ACM Interactions*, **15**(4), 49–52. Accessed on-line at `dl.acm.org/citation.cfm?id=1374501` on April 10, 2013.

Walker, Rob. 2003. The Guts of a New Machine. *In: New York Times Magazine.* Accessed on-line at `http://www.nytimes.com/2003/11/30/magazine/the-guts-of-a-new-machine.html` on April 13, 2013.

Weiss, Shalom M., Indurkhya, Nitin, Zhang, Tong, & Damerau, Fred J. (eds). 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis.* Springer.

Wikipedia. 2013. *Apophenia.* Accessed at `http://en.wikipedia.org/wiki/Apophenia` on April 5, 2013.

Wilkinson, Leland, Wills, D., Rope, D., Norton, A., & Dubbs, R. 2005. *The Grammar of Graphics.* 2nd edn. Springer.

Wong, Donna M. 2010. *The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures.* W.W. Norton & Company.

Xiong, Rebecca, & Donuth, Judith. 1999. PeopleGarden: Creating Data Portraits for Users. *Pages 37–44 of: Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology.* Accessed on-line at `http://smg.media.mit.edu/papers/Xiong/pgarden_uist99.pdf` on April 12, 2013.

Yau, Nathan. 2011. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics.* Wiley.

Yau, Nathan. 2013. *Data Points: Visualization That Means Something.* Wiley.

Yi, Ji Soo Yi, ah Kang, Youn, Stasko, John, & Jacko, Julie. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Trans. Visualization and Computer Graphics*, **13**(6), 1224–1231. Accessed on-line at `www.cc.gatech.edu/~john.stasko/papers/infovis07-interaction.pdf` on April 10, 2013.