

The Need for Smart Data Visualization

By: Krishna Ramaswamy, PhD; Stephanie Nealis, MS; Shannon Brown, PMP; Adam Welsh

Just because you can do data visualization, should you?

BACKGROUND & UNDERSTANDING

Big Data and Data Visualization

One of the most popular buzzwords in the tech world today is Big Data. From the finance and banking industry to genomics and healthcare, to the marketing and communications sectors and even extending to political campaigns, everyone wants to generate and utilize Big Data. Advances in communications and information technology have fueled a tsunami of Big Data (social network relationships, health records, genome sequencing data, etc.). This data deluge has in turn paved the way for development of data visualization tools to help analyze and understand Big Data.

Data visualization is the graphical representation of vast amounts of quantitative and/or qualitative data. Data visualization tools allow organizations to have better control of data analysis, governance, and management by breaking down the traditional barriers to data analysis and providing easy-to-understand visual depictions of complex data models. Visualization of data could also highlight aspects that were not expected; it could highlight any flaws in the data itself, enable the viewing of macro and micro features in the data, and might help in forming new hypotheses.

Data Visualization - Then and Now

Data visualization is not a new concept. It has a storied history which can be traced back to the 17th century works of Galileo and Descartes who used rudimentary methods to collect and display data to the simple, elegant, and impactful mid-19th century works of John Snow and Florence Nightingale. It also includes graphs and pie charts used in present day board rooms, as well as the works of Tukey and Tufte¹. Today's interest in data visualization and interactive data visualization in particular, can be attributed to: (1) Advances in computational power, data analysis, and graphics which have enabled widespread access to data visualization products, (2) Generation and availability of large amounts of data which cannot be easily analyzed by traditional methods, and (3) A need for rapid analysis and decisions on the large amount of data that is generated within an organization.

Data visualization tools continue to evolve and can be non-interactive or static, such as graphs and charts, or interactive such as timelines and relationship maps. The traditional (non-interactive) tools require more time to analyze and understand, the data can be error prone, and often require in-depth knowledge of the application in use. Interactive tools afford a better understanding of relationships and trends in data sets in comparison to traditional tables and queries. Additionally, interactive data visualization tools allow a quick drill down of data to the smallest unit.

There are a number of big players with a focus on developing data visualization software including SAS, IBM, and Tableau. However, times are changing and it is no longer the purview of the big players with many small companies, startups, and open source groups developing various data analytics and visualization products or platforms (Visual.ly, Processing, ZingChart, and Gephi). In many cases, these products are free and/or being developed within the organization to meet that organization's specific needs.

¹ <http://www.datavis.ca/milestones/>

THE CHALLENGES

Are we asking the right questions?

Interactive data visualization tools primarily evolved as ad hoc solutions but their popularity has increased tremendously in the recent years. Consequently, companies (regardless of size) are racing to develop better, faster data visualization tools and, in turn, fueling an almost irrational expectation that data visualization is the magic-bullet for tackling Big Data. Though these expectations may be warranted, the more important questions to address include: Are the right questions being asked about the data? And what is more important – the tools or the method of utilizing the tools? Most data visualization tools just look at providing an immediate, narrowly focused answer in comparison to a big picture solution. Many of these tools are focused on the jazz and flash of data visualization rather than helping the user derive the right information, leading to the misconception that the data visualization and business intelligence products are adequate for decision support. Having Big Data and sophisticated data visualization tools does not necessarily mean having the right answers and can in fact make people look for non-existent patterns or force fit data to models². According to John Tukey, a pioneer in exploratory data analysis: “seeing may be believing or disbelieving, but above all, data analysis involves visual, as well as statistical understanding³.”

Structured vs. Unstructured Data

Big Data does not necessarily mean the right kind of data. A rapidly developing concern is effectively deriving structure from a vast amount of unstructured data such as social media data (tweets, posts), data formats such as .PDFs, documents, spreadsheets, and geo-spatial data. According to Karen Hsu⁴, the senior director of product marketing for Datameer (a Hadoop analytics vendor): “What we’re finding is people are spending 80 percent of time just in getting that data ready, because the data is so raw, because there’s so much of it, and it’s in so many formats.”

STATEMENT OF THE SOLUTION/APPROACH

Ideally, data visualization tools should not only provide information on what is expected but also help the user to decipher what is not expected, such as outliers and unusual trends. Only then will the data visualization tools help with decision support and lead to better management by exception. Edward Tufte, a pioneer in data visualization describes this idea as – “You want to see to learn something, not see to confirm something⁵.”

Human visual perception capabilities are a main factor that data visualization vendors may fail to take into consideration while developing tools. The human visual system is endowed with tremendous abilities to see patterns and make decisions (the animal kingdom uses a similar behavior in prey detection; refer to the Search Image Theory exhibit⁶). These abilities are governed by certain rules with regard to the size, shape, color and proximity of the objects. Thus it is imperative for data visualization vendors to know and understand these rules while creating data visualization tools. There is a vast amount of research conducted and data available in the field of human visual systems and cognition and recently some of this data is being incorporated into the design and development of the data visualization tools.

SEARCH IMAGE THEORY

In the animal kingdom, the predator-prey interaction is very much dependent on pattern detection. The predator looks for certain color, shape, and size cues to detect its prey. In doing so at times the predator becomes solely focused on one prey and is unable to detect other prey effectively. This kind of prey selection is based on search image formation by the predator. When the population of a particular prey decreases due to their increased detection, the predator has to form a new search image for the other prey in the wild. This phenomenon maintains the balance in the population of the prey, which is essential for the survival of the predator.

² http://lesswrong.com/lw/hxx/some_highlights_from_nate_silvers_the_signal_and/

³ Visual Data Mining: Theory, Techniques and Tools for Visual Analytics edited by Simeon Simoff, Michael H. Böhlen, Arturas Mazeika

⁴ http://www.datanami.com/datanami/2014-03-26/forget_the_algorithms_and_start_cleaning_your_data.html

⁵ http://americandigest.org/mt-archives/driveby/edward_tufte_the_art_of_d.php

⁶ http://evolution.berkeley.edu/evolibrary/article/0_0_0/happyface_11

Data visualization solution vendors also need to be cognizant about and develop methods to tackle the extensive amount of semi-structured and unstructured data that is being generated due to the popularity and ease of use of mobile devices and platforms like Twitter, Facebook, Tumblr, etc. A number of standalone products are available to analyze and consume the semi-structured and unstructured data. Going forward, such solutions need to be offered as part of the data visualization solutions suite.

Lastly, in spite of the user friendliness of various data visualization tools i.e., the so called democratization of data visualization tools, that does not require the managers/analysts to be computer scientists, it is clearly becoming important that the consumers of data visualization tools become savvy and comfortable with using the tools. The consumers need to become data scientists – in addition to analyzing the data, they have to look for patterns, hypotheses, outliers, and unusual trends to draw inferences.

CONCLUDING THOUGHTS

In an ideal world, data visualization tools should not only provide information on what is expected but also help to decipher what is not expected. The tools should be a means to identify outliers and unusual trends, account for various types of data (i.e. structured vs. unstructured), utilize the appropriate analysis methodology (statistical understanding), and incorporate human visual perception. Then and only then will the data visualization tools help with decision support and lead to better management by exception.

Our experiences with data visualization in an adverse event management system highlighted a number of the issues discussed above and prompted us to address them. In doing so, we were able to provide our client with tools to attain greater levels of efficiency through a cost-effective, low-risk solution. Stay tuned to hear about our success story in an upcoming ZapFlash.