Using visualization to understand big data

By T. Alan Keahey, Ph.D., IBM Visualization Science and Systems Expert

IBM

Contents

Introduction

Studies have shown that the human short-term memory is capable of holding 3 - 7 items in place simultaneously, which means that people can only juggle a few items in their heads before they start to lose track of them. Visualization creates encodings of data into visual channels that people can view and understand. This process *externalizes* the data and enables people to think about and manipulate the data at a higher level. This externalization enables humans to think more complex thoughts about larger amounts of information than would otherwise be possible.[1]

Visualization exploits the human visual system to provide an intuitive, immediate and language-independent way to view and show your data. It is an essential tool for understanding information. The human visual system is by far the richest, most immediate, highest bandwidth pipeline into the human mind. The amount of brain capacity that is devoted to processing visual input far exceeds that of the other human senses. Some scientific estimates suggest that the human visual system is capable of processing about 9 megabits of information per second, which corresponds to close to 1 million letters of text per second.

Visualization research over the past decades has discovered a wide range of effective visualization techniques that go far beyond the basic pie, bar and line charts used so pervasively in spreadsheets and dashboards. These techniques are especially useful now that most organizations are being confronted with big data. The majority of organizations are struggling to make sense of output from data sources that include RFID communications, social media text, customer surveys, streaming video and more, along with data captured over very long periods of time. For the IBM Institute for Business Value report on big data, IBM surveyed more than 1100 business and IT professionals and found that less than 26 percent of respondents who had active big data efforts could analyze extremely unstructured data such as voice and video and just 35 percent could analyze streaming data.[2] Visualization plays a key role in enabling the understanding of these complex data analytics, and it can convey the key analytical nuggets of information to other people in the organization who have less expertise in analytics.

When companies can analyze big data, they benefit. In that same IBM survey, 63 percent of respondents reported that they believe that understanding and exploiting big data effectively can create a competitive advantage for their organizations.[3] Big data analysis can help them improve decision making, create a 360-degree view of their customers, improve security and surveillance, analyze operations and augment data warehousing. Visualization can play a vital role in using big data to get a complete view of your customer. This paper covers how.

Using the visualization of big data for a complete picture

Businesses in the modern economy require a fuller picture than ever of their customers in order to compete. Such a picture requires a complete understanding of not only how each customer is transacting with your company, but how each customer is finding out about offerings, comparing alternatives, discussing products and services in their social networks and interacting with related products and services. Each of these aspects represents a separate analytics task that can be difficult for business users without an analytics background to master; when combined, the challenges become even greater for obtaining this 360-degree view of your customers.

Visualization can play a key role by making the individual analytic components understandable and by tying them together into a comprehensible “big picture.” In addition, visualization can be used in several distinct ways to help tame the scale and complexity of the data so that it can be interpreted more easily. To understand how, you can start with a simple customer data set and add more views of the customer, including those from big data.

Simple customer data

Many dashboards and reporting tools show data simply as a set of one or more basic charts, such as the bar, line or pie chart. These work fine for conveying basic information such as historical key performance indicators (KPIs); however, their effectiveness becomes more limited when you want to understand multiple KPIs or other measures in a bigger picture. Combining many simple charts into a single page can quickly lead to overwhelming clutter. If they are placed on different pages so a user must navigate them, he or she can have problems with internally integrating them or relating the different measures on different pages to one another. The bar chart and line chart can be extended at this point (Figure 1) to show a single measure for multiple categories in a single chart by means of a variety of common techniques, including stacked or clustered bar charts and multiple series line and area charts.

These techniques can typically be used to show 4 - 8 categories in a single chart; however, for many big data scenarios, the number of measures (columns in a spreadsheet) can run into the thousands. In these cases, no single visualization technique is adequate for conveying the raw data. Some sort of analytical or dimensional reduction technique should be applied to the data first before attempting to apply visualization. A wide variety of such data reduction techniques are available that can be applied, including segmentation, clustering, linear regression and more. The idea behind these is to come up with a mathematical model that reduces the complexity of the data, either the number of dimensions or the number of data points, while still capturing the essential characteristics of the data.

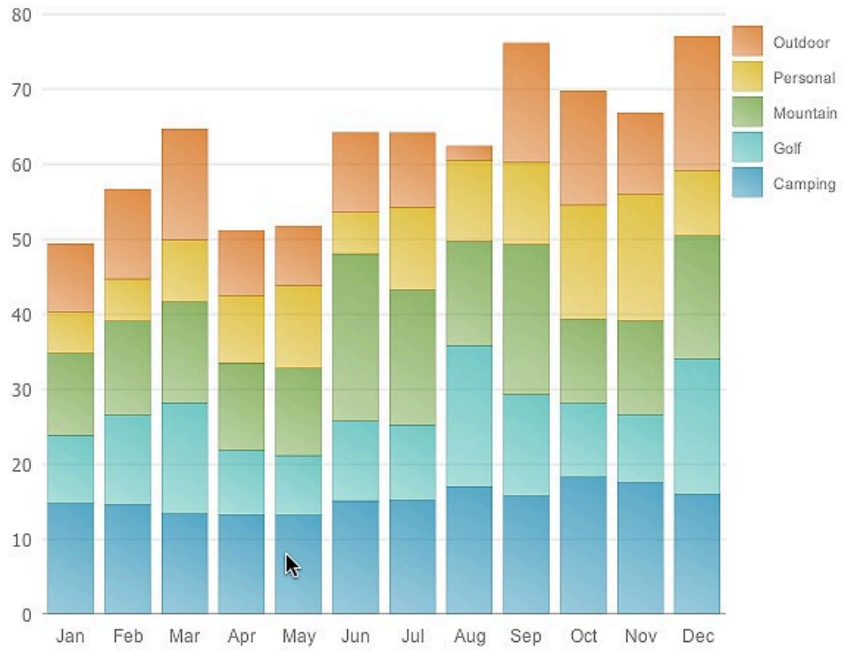

Figure 1: A bar chart to show a single measure for multiple sales categories.

Adding time to the customer equation

Customer patterns are often tied to time cycles, such as the 24-hour diurnal cycle or monthly payroll cycles. Although line charts can be used to convey these patterns over time, other more advanced visual metaphors can more realistically convey these temporal (that is, time-based) patterns.

The radar chart (Figure 2) is often a good choice for showing regularly cyclical data such as daily data over a weekly timeframe, or monthly data over a yearly time frame. Care should be taken with hourly data, however, because it might confuse viewers who are used to the 12- hour clock cycle.

Figure 2: A radar chart that shows weekly cycles of sales data that are arranged in a circular fashion.

The calendar also provides a powerful and well-known metaphor for showing time. One effective visualization uses a "heat map" scale and color to encode calendar days with a value (Figure 3). The result is a very compact, intuitive visual representation that conveys weekly and monthly patterns effectively, whether for just a few months of data or for many years.

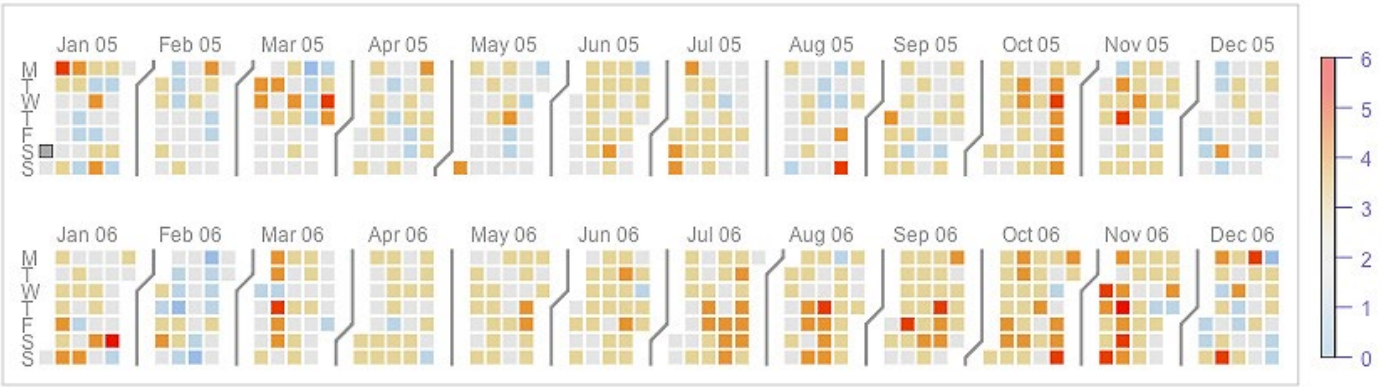

Figure 3: Calendar heat map example that shows two years of changes (in percentages) in customer web orders by year (row), month (column), day of week (sub row), week (sub column) and day.

Compare the calendar heat map with a line chart (Figure 4) that has the same information and you can see how effective the heat map can be.

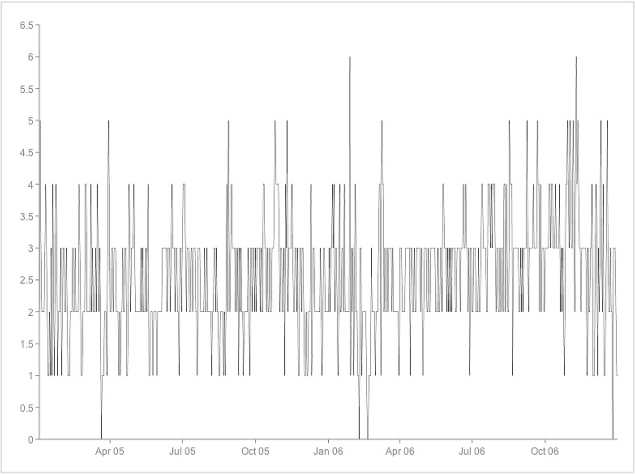

Figure 4: Chart view of the same data shown in the Calendar Heat Map in Figure 3.

Understanding customer sentiment

Getting a complete view of your customers requires more than just understanding the key transaction metrics. You must track what your customers (potential and actual) are saying about your company, products and services, along with what they are saying about your competition. This form of big data can be collected from a number of sources, including call center logs, social media and customer surveys. Sentiment analysis and other techniques can be used to process this big data to extract patterns, and visualization is an essential tool for conveying many of those patterns to the business user.

One prominent technique extracts the key words and phrases from a set of customer communications and then tracks how the use of those words and phrases changes over time. The "theme river" visualization is well suited for showing this type of information (Figure 5). The thickness of the band at a point in time corresponds to the frequency count for the associated word or phrase. With this visualization, you can get a feel for the ebbs and flows of your customer's sentiments.

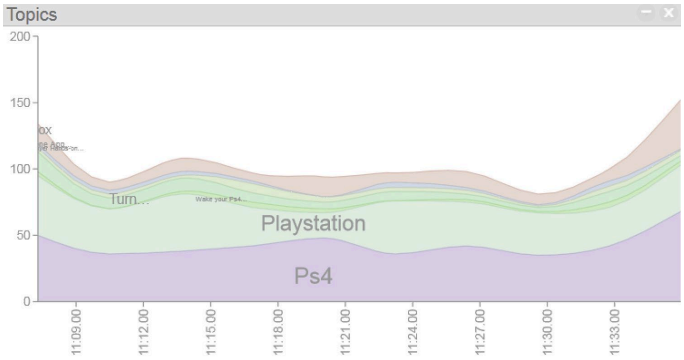

Figure 5: Theme river visualization showing phrase popularity related to gaming platforms over time.

Uncovering customer relationships

Relationships are a critically important aspect of many big data scenarios. Social networks are perhaps the most prominent example in this regard, and mastering them means that you can understand and influence not just individual customers but also their associated networks of friends and family members. These types of relationships are very difficult to understand in text or tabular format; however, applying visualization (Figure 6) can make emerging network trends and patterns apparent.

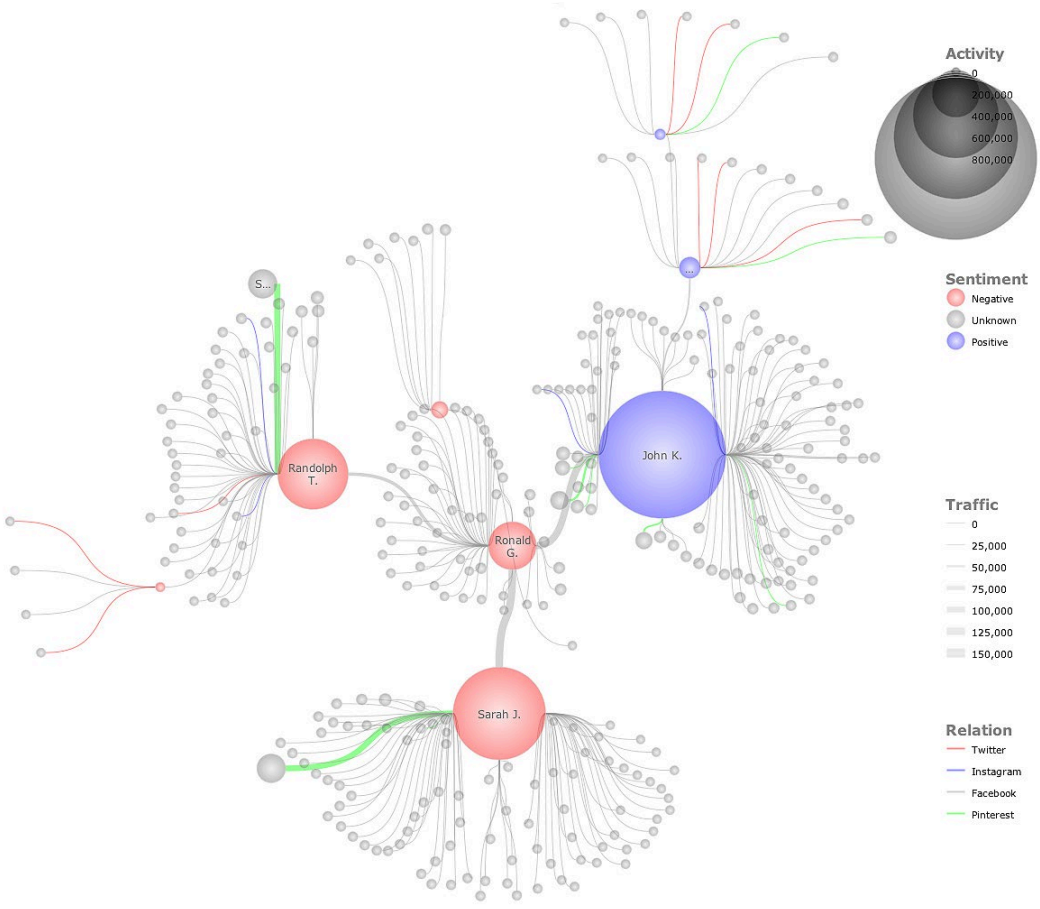

Figure 6: A social network visualization can show patterns of customer sentiment, key influencers and their reach.

Understanding customers at different levels of detail

Hierarchies are powerful data abstractions for aggregating information into broader categories so you can make sense of it at a higher level. One way in which these hierarchies are used is to aggregate the time dimension: customer activity measured in days, then months and then into years creates a three-level hierarchy. Another common type of hierarchy used to aggregate larger data sets into more understandable abstractions is based on geospatial properties: for example, customer sales in individual cities can be rolled into state level sales and then into national sales. A third example involves organizing a product catalog into broad categories (outdoor, recreation, sports) and then into subcategories (baseball, tennis) and finally product. Many other hierarchies are possible.

Hierarchies are very popular in data analytics; however, they should be used with care, especially with big data, because the chosen roll-up mechanism can sometimes obscure important details at lower levels. A traditional way of enabling the understanding of information at multiple levels of hierarchical detail is to present the individual levels in a series of tabbed reports, each report showing a single section of the hierarchy (for example, all tennis products). Using visualization in hierarchies provides a fuller understanding of the information because it shows multiple levels of the hierarchy simultaneously. A wide range of visualization techniques is available for viewing hierarchies; the example here (Figure 7) shows a fairly direct approach where each entity in the hierarchy is represented by a node in the chart. Size and color are used to show various properties of the nodes, and lines between the nodes show the hierarchical relationships.

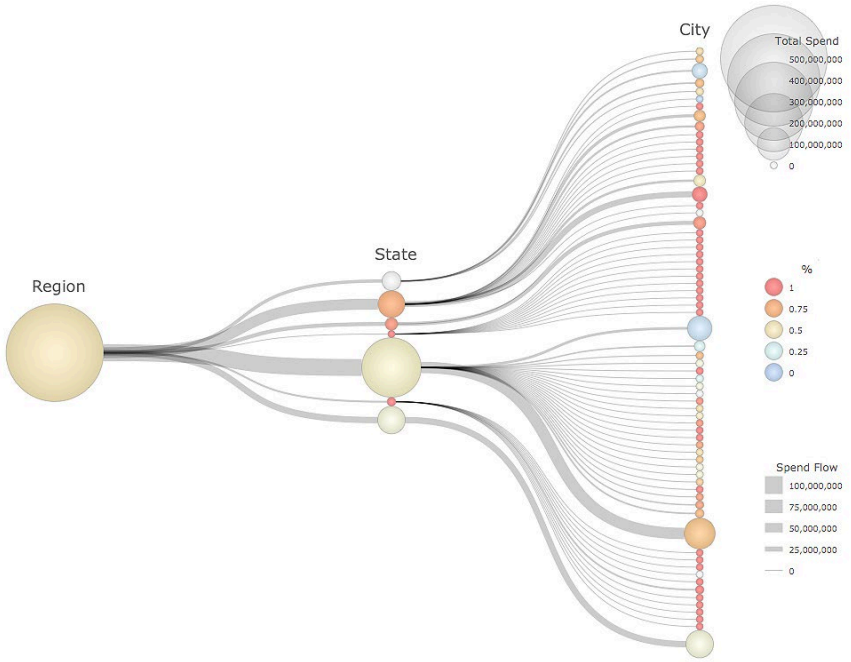

Figure 7: Hierarchy visualization of data that shows the number of targeted campaign responses on the regional, state and city levels. Each level is represented by a bubble or bubbles.

In Figure 7, bubble size indicates the number of campaign responses, and color indicates another measure such as change from prior year. Red is low, white is neutral and blue is high.

Because of their branching structures, hierarchies are often referred to as "trees" in the visualization research community. Another very powerful and effective method for visualizing hierarchies is the tree map, in which the outer rectangle represents the sum total for the whole hierarchy and is recursively subdivided according to the divisions of the hierarchy. The size of each sub-rectangle can represent one measure, while color is often used to represent another measure of the data. Figure 8 shows a tree map of a collection of choices for streaming music and video tracks by a social network community that a media service could find useful when designing personalized offers of music and videos for download. Color represents the genres of the selected tracks, with each genre subdivided into rectangles for each artist. Size of rectangle for both genre and artist represents the number of track plays in that category.

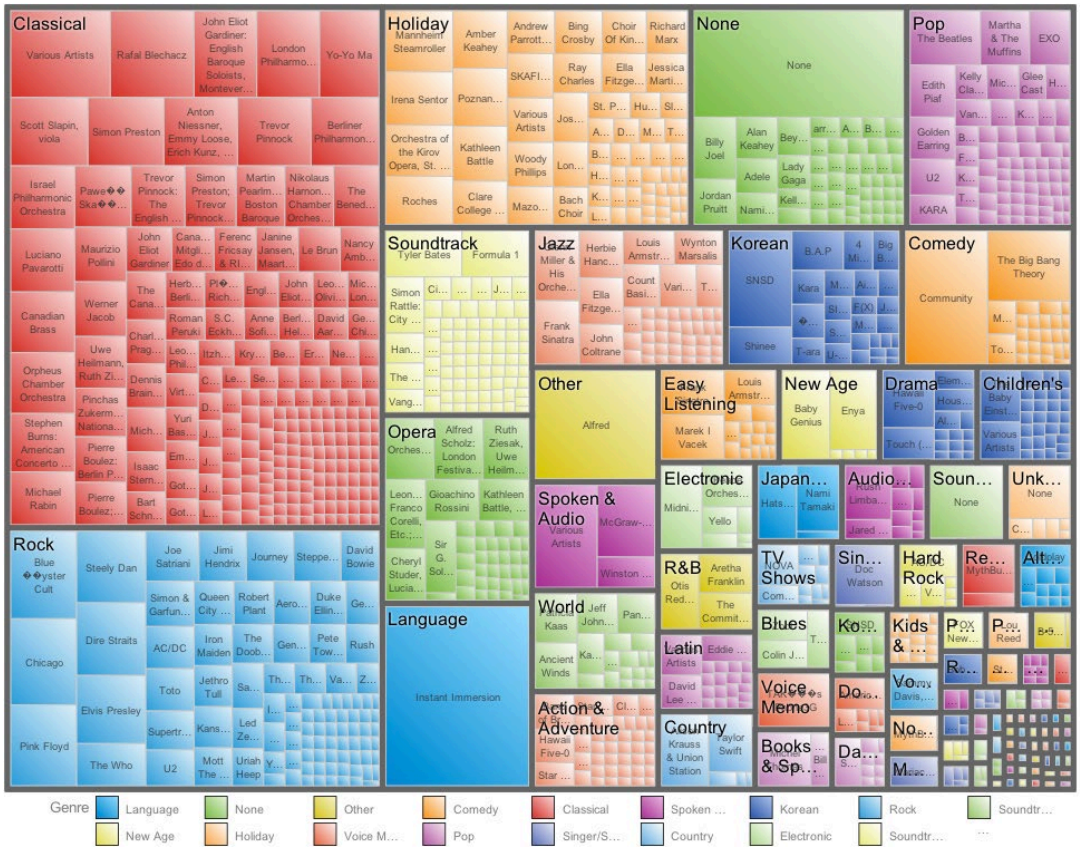

Figure 8: Tree map view of a social network's track selections from a streaming media service.

Hundreds of different tree visualization methods have been explored in the research community, many of them finely tuned for specific types of tree data such as genome sequencing, large social graphs and tournament matches.

Some of these tree visualization methods are capable of showing hundreds or thousands or even millions of entities arranged in a hierarchical structure.

How much big data can we visualize directly?

A frequently asked question is how much big data can people view and understand directly with visualization techniques. The answer depends greatly on what type of data is being viewed, and what sorts of questions and answers the viewer wants to develop. However, for most cases, direct visualization of big data sources is not possible or effective. Visualization for large data works best with analytics techniques, which has given rise to an entire field of research known as visual analytics.

How much data can be effectively visualized directly is still worth considering, however. The answer to that depends on a number of factors: the scale and structure of the data, the size of the display device, computational scalability, collaborative and sharing needs, and the scalability of the visual layout.

Some general rules of thumb for the amount of data items that can be effectively shown with some of the common visualization techniques are:

- Pie chart: 3-10
- Bar chart: fewer than 50
- Line chart: fewer than 500
- Bubble plot: fewer than 500
- Scatter plot: fewer than 10,000

More advanced visualization techniques can show a greater number of items. These advanced styles can provide somewhere between 3 - 6 orders of magnitude (1,000 - 1,000,000 items) of direct data visualization, perhaps 9 orders of magnitude (1,000,000,000) for extremely special cases. Although the visualization can provide a significant reduction in scale, it clearly can only be part of the solution if the goal is to process a terabyte (12 orders of magnitude) or petabyte (15 orders) of big data. Analytics plays a key role by helping to reduce the size and complexity of big data to a point where it can be effectively visualized and understood. In the best scenario, the visualization and analytics are integrated so that they work seamlessly with each other.

Taming the complexity of big data with IBM

Visualization is an essential tool for making sense of big data. It provides a far richer view of big data than can be obtained from tables and statistics alone. However, the key to effective analysis of big data is the integration of visualization into analytics tools so that all kinds of users can interpret big data from a wide range of sources—clickstreams, social media, log files, videos and more. IBM has embedded visualization capabilities in a number of solutions and also offers extensible visualizations that can be downloaded for use in business analytics solutions. All the visualizations in this paper were created with IBM solutions and standards.

Visualization and big data solutions

Because IBM understands that big data analytics contributes significantly to competitive advantage and that visualization is a key ingredient in such analytics, IBM has embedded visualization capabilities into business analytics solutions. What makes this possible is the IBM Rapidly Adaptive Visualization Engine (RAVE).

RAVE is increasingly used as the standard for IBM visualization capabilities because it enables the rapid development of common and new visualization types. Because interpreting big data is still an emerging concept and ways to understand it are still developing, the ability of RAVE to create new kinds of charts that are as yet unknown is especially compelling.

IBM products, such as IBM® InfoSphere® BigInsights™ and IBM SPSS® Analytic Catalyst, use visualization libraries and RAVE to enable interactive visualizations that can help you gain greater insight from your big data. InfoSphere BigInsights is software that helps firms discover and analyze business insights hidden in big data, and the solution includes visualizations to simplify analysis of the data. SPSS Analytic Catalyst automates big data preparation, chooses the proper analytics procedures and can display the results as interactive visualizations.

Extensible visualization

With the future of big data still developing, having the capability to respond with new visualization types as you need them helps to meet the challenge of dealing with big data. An increasing number of IBM business analytics solutions, including IBM Cognos® Business Intelligence, are using new extensible visualization capabilities. Extensible visualization enables users to download new visualizations from an ever-increasing library on IBM Analytics Zone as needed. Access to this ever-changing set of visualizations frees business users and report authors from the constraints of a prescribed library of in-product visualizations and also offers opportunities to use newly developed visualizations with big data.

The chord diagram visualization (Figure 9) is an example of what can be produced with extensible visualization. It is an elegant and compact way to show networks of relations between items such as products, individuals or groups. The width of each chord shows the amount of traffic between the groups that are located around the circumference. Color on the chords and groups can also be used to convey additional information. This particular example relates customer support request types, which are shown on the right side of the circle, to the company support group that is handling the request, which is shown on the left side of the circle. This single visualization can represent a huge amount of customer interactions over the period of a year or more.

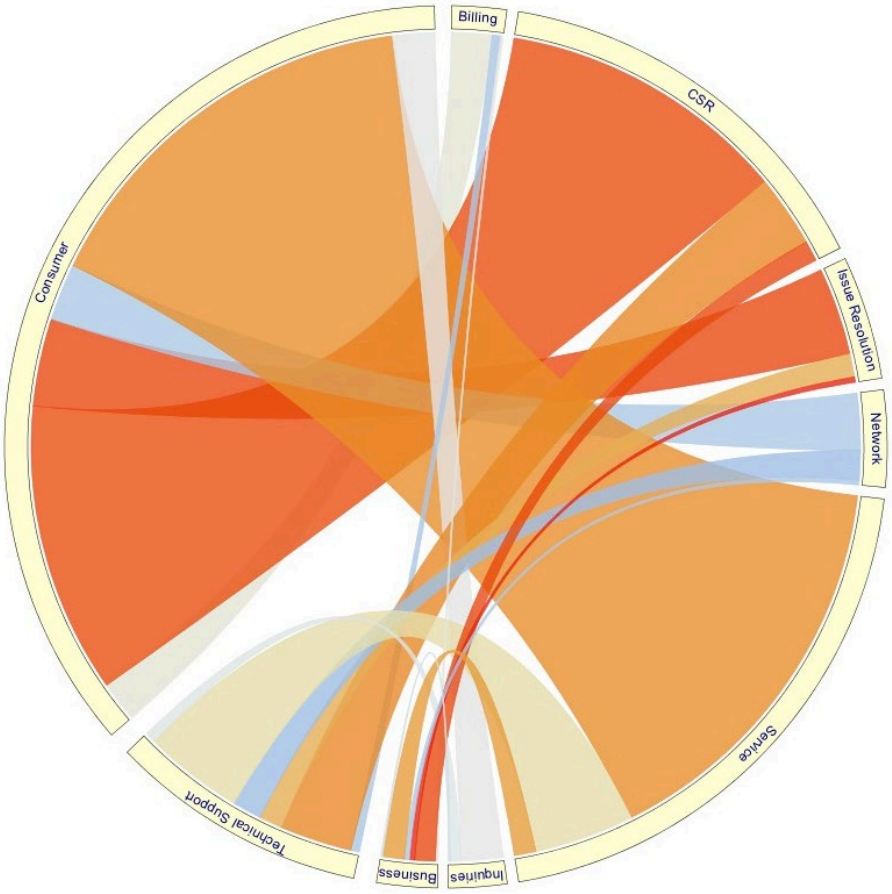

Figure 9: A chord diagram visualization created with extensible visualization technology.

Conclusion

Visualization is an essential tool for understanding information and uncovering insights hidden in your data, including your big data. With an understanding of big data, you can accomplish a number of things that can help your business, including creating a complete view of your customers. New visualization methods are available that are well suited to the particular needs of big data in many areas, such as social media analysis, geospatial analysis and sentiment or text analysis. These new visualization methods go far beyond the traditional tables and bar or line charts. They include radar charts, chord charts, calendar heat maps and more. IBM technology, such as RAVE and extensible visualization capabilities, can help you create and use effective visualizations that provide you with a better understanding of your big data.

For more information

To learn more about IBM and advanced visualization, visit the IBM advanced visualization web page: **ibm.com**/software/analytics/many-eyes/

To learn more about extensible visualization and to see extensible visualizations that are currently available from IBM, visit the Extensible Visualization Community in the Analytics Zone: analyticszone.com/visualization

About the Author

Dr. T. Alan Keahey has played a leading role in the research and development of highly innovative information visualization systems for close to 20 years. His experience spans a wide range of environments, including national labs research scientist, research director at a Lucent Bell Labs spin off and founder of his own visualization research and development company. He thrives on anchoring connections between the capabilities created in research environments and the real-world needs of business customers. Alan is currently a Visualization Science and Systems Expert at the IBM Business Analytics Office of the CTO.

Blog: http://www.HolisticSofa.com
LinkedIn: http://www.linkedin.com/in/truviz/

About IBM Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management, and risk management.

Business Analytics solutions enable companies to identify and visualize trends and patterns in areas, such as customer analytics, that can have a profound effect on business performance. They can compare scenarios, anticipate potential threats and opportunities, better plan, budget and forecast resources, balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision-making to achieve business goals. For further information please visit **ibm.com**/business-analytics

Request a call

To request a call or to ask a question, go to **ibm.com**/business-analytics/contactus. An IBM representative will respond to your inquiry within two business days.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2013

1 Donald A. Norman, The Design of Everyday Things. New York: 2002.

2 "Analytics: The real-world use of big data." IBM Institute for Business Value, in collaboration with Said Business School at the University of Oxford. 2012. http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03519usen/GBE03519USEN.PDF

3 "Analytics: The real-world use of big data."

Please Recycle

YTW03349-USEN-00