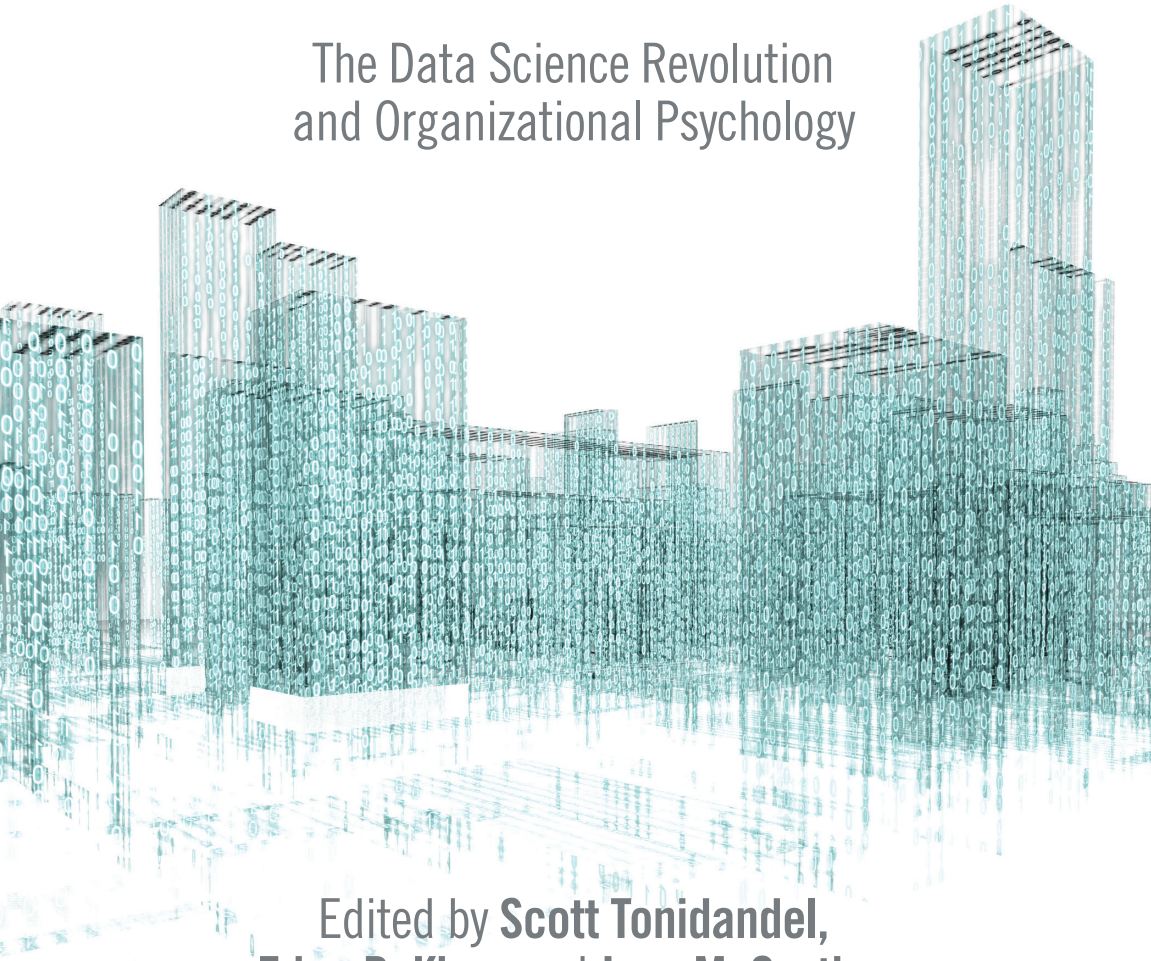
SIOP SOCIETY for INDUSTRIAL and ORGANIZATIONAL PSYCHOLOGY
ORGANIZATIONAL FRONTIERS SERIES

ROUTLEDGE

Big Data at Work

The Data Science Revolution and Organizational Psychology



Edited by **Scott Tonidandel**, **Eden B. King**, and **Jose M. Cortina**

First published 2016
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Big data at work : the data science revolution and organizational psychology / edited by
Scott Tonidandel, Eden B. King, & Jose M. Cortina.
pages cm. — (The organizational frontiers series)
Includes bibliographical references and index.
1. Organizational behavior. 2. Big data. 3. Organizational sociology. 4. Psychology, Industrial. I. Tonidandel, Scott. II. King, Eden. III. Cortina, Jose M.
HD58.7.B534 2016
302.3'5—dc23
2015010678

ISBN: 978-1-84872-581-2 (hbk)
ISBN: 978-1-84872-582-9 (pbk)
ISBN: 978-1-31578-050-4 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

5

Data Visualization

Evan F. Sinar

In the era of big data, improvements in technology have made it easy for organizations to collect huge volumes of information of a vast variety of types and characteristics. From consumer shopping habits to real-time electricity usage, from internet connectivity to weather patterns, from social media data to employee accident and error rates, terabytes of data are constantly and meticulously collected, filed, sorted, and stored by automated and hand-entered systems. Big data provide the raw information needed to perform complex analysis and discern key patterns and trends to a degree that would have been impossible 20, or even five, years ago.

Despite the advancement in data tracking and accumulation, however, the human brain has not advanced at the same rate. It is simply not possible to readily make use of such a massive scope and scale of data in its raw form. This limitation is a key barrier to individuals and organizations seeking to leverage the power harnessed within newly accessible large-scale datasets. Even the most comprehensive and expansive databases are worthless without a way to understand and process their qualities and to translate these qualities into actionable insight. Although the potential of big data is substantial, methods for identifying and comprehending new aspects of knowledge hidden within these data are essential. Without approaches that enable this and that do so in a manner that increases accessibility of information to the broad array of those charged with extracting value from big data, this value will be underutilized at best, or ignored at worst. Data visualization brings accessibility and interpretability to big data.

In this chapter, I will review the topic of data visualization and its applications to big data in five major sections. First, I define data visualization and overview its emergence, function, and advantages in general, business, and big data contexts. Second, I briefly discuss the perceptual foundations for visualization. Third, I review several examples of specific data visualization types, applications for I-O psychologists, and publicly available tools

to create them. Fourth, I expand on a discussion of research questions potentially well suited to visualization approaches and key design considerations in creating them. Finally, I discuss key issues and risk areas associated with data visualization and future opportunities for I-O psychologists to both advance the knowledge base and harness the advantages of data visualization within their own practice areas.

DATA VISUALIZATION: DEFINITION AND GOALS

In its simplest form, data visualization is a set of methods for graphically displaying information in a way that is understandable and straightforward, ideally while also incorporating aesthetic considerations to drive engagement and interest to in turn capture the attention of the intended audience. Data visualizations use distinctive techniques and design choices to guide users to easily absorb, understand, and make decisions based on information. How well this goal is accomplished is dependent not only on the qualities of the data themselves, but also on the skills of the researcher and visualization creator in choosing the right presentation method, and in guiding the user to observe key features in the data—while simultaneously considering the appropriateness of the format and guidance provided.

From an analytic perspective, visualization serves two primary functions—to explore data and to explain it (Iliinsky & Steele, 2011). The exploratory purpose of data visualization is to discover patterns, relationships, hierarchies, and differences that would be difficult or impossible to detect based solely on statistical procedures or by reviewing textual or tabular forms of data presentation. It is important to note that data visualization typically plays an inductive role in the analytic process, detecting observations and findings that themselves have a distinctive value, yet can also serve as centering points for further hypotheses and investigation using more traditional analytic techniques. That is, an exploration-focused use of data visualization can be an outcome in itself, or it can be a precursor to further analyses of high-level patterns detected.

A second function of visualization is to explain patterns, trends, or relationships involving variables of interest. Visualizations that originate not in a raw dataset, but rather in a research question or business objective, can graphically display alternative hypotheses, allowing the user to gauge which is most likely. The explanation function extends to communication of the findings themselves to reach a broader audience, to efficiently orient a user to a topic area, and to incite interpretation, inferences, and decisions

to a degree that would not be possible using text- or numerically-based communication formats.

EMERGENCE OF VISUALIZATION FOR INFORMATION COMMUNICATION

Data visualization itself is not a new idea, nor is its ability to drive action novel—some of the most influential data visualizations emerged well over a century ago, such as John Snow's 1854 London cholera map, Florence Nightingale's 1858 war mortality graph, and Charles Minard's 1869 march on Moscow chart. Visualization has long played a role in communication of quantitative information through the foundational work of Tufte (1983, 1990, 1997), Cleveland (1993), and others decades before the term "big data" came into use. Making sense of complex information has always been necessary, and past a certain point, additional data beyond that available decades ago do not become meaningfully "bigger."

However, though visualization as a communication technique is not new, what has changed is the range of individuals charged with processing and making decisions based on data. Research by Manyika et al. (2011) projected a 2018 deficiency of 140,000 to 190,000 positions for data analytics experts, and more broadly a shortage of 1.5 million managers and analysts who—as a component of their job rather than their full-time employment—must make sense of and decisions based on large-scale datasets. Visualization is a critical component in this equation that enables broader information exchange and processing efficiency due to the advantages visualizations can provide.

The surge in public usage of graphical information presentation formats is also relatively recent. As one indicator of the growth of their prevalence as a data communication mechanism, interest in the term "infographics," as indexed by Google Trends, has increased five-fold in only three years, from 2011 to 2014. While infographics and data visualizations are considered somewhat distinct—infographics are usually designed for stylistic rather than analytic purposes and are less amenable to big data applications—the proliferation of infographics nonetheless has established a foundation for visualizations of all types. The acknowledgement by media companies of the value of graphical information formats for communicating complex concepts to a broad audience is also clearly evident from their rapid adoption of such approaches. Indeed, many of the leading practitioners of advanced data visualizations are based in large and influential media

outlets such as the *New York Times* and the *Wall Street Journal*. Websites such as www.dadaviz.com have also emerged to compile visualizations of broad interest.

Visual analysis and production skills are integral to many projections of future work skills. For example, the Institute for the Future (Davies, Fidler, & Gorbis, 2011) defines a future need in response to the new media ecology—new communication tools requiring new media literacies beyond text for new media literacy. In their view, the next generation of workers, "will also need to be comfortable creating and presenting their own visual information. [. . .] As immersive and visually stimulating presentation of information becomes the norm, workers will need more sophisticated skills to use these tools to engage and persuade their audiences" (p. 13).

As more professionals and managers find themselves in the role of data analysts and presenters, it is likely that many forms of analysis will take a visual rather than purely quantitative form in order to reduce the gap between the relatively small number of quantitative specialists and the much larger employee base of those who can readily interpret, critically evaluate, and act upon information presented graphically.

ADVANTAGES OF VISUALIZED DATA

Increasingly, information presented in daily life and in business settings is presented visually. Visualization makes data approachable to a broad audience. It democratizes data access, interpretation, and analysis by drawing upon our substantial visual skills and by leveraging common visual referents. Through use of these cues, accessibility increases and training time to interpret the visuals is reduced to the degree that these cues are already inculcated in the audience. Visualization, regardless of the size of data to which it is applied, is also advantageous in comparison to textual or tabular forms of data presentation. They enable detection of relationships that would otherwise remain hidden, and do so efficiently. Visualizations facilitate integration of multiple data sources through the use of common visual referents that place different types and scales of data into a singular view. Through their influence on human cognition, visualizations produce benefits for decision-making, learning, and analytical reasoning (Parsons & Sedig, 2013). Although the number of studies into the persuasive impact of data visualizations is very limited at this time, early investigations have been promising: as one of the most recent examples, Pandey,

Manivannan, Nov, Satterthwaite, and Bertini (2014) found that information presented using graphs was more likely to persuade an audience, and to have a greater effect on their attitude change, than the same information presented using tables.

Visualization is increasingly being recognized across scientific domains as a valuable and effective approach for distributing information to populations with varying levels of information processing experience and literacy in the content domain. As one notable example from the health sciences, researchers developed a series of visualizations to display health status indicators (Arcia et al., 2013). This approach, to address communication gaps limiting the transfer of key quantitatively derived knowledge to target populations, is indicative of the broad value that data visualizations can generate.

VISUALIZATION AS A BIG DATA IMPERATIVE

Big data describe a class of information that is too large and complex to process using traditional database management techniques or conventional data refinement software—these challenges extend across functions including searching, capturing, sharing, storing, curating, transferring, and exploring, as well as visualizing (Russom, 2011). Visualization provides a method for introducing structure and a graphical representation to the data, and can include both static and interactive visualization approaches (Thomas & Cook, 2006). Visualization also provides access to data elements—and resulting discoveries—that would be difficult or impossible to connect using traditional methods (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). In response to the rapidly growing level of data and aided by improvements in technology, visualizations counteract a state of data overload by enabling data processing at a more rapid rate, guiding decisions accordingly (Viégas & Wattenberg, 2007). Visually-based analytic methods combine the strengths of human cognition with new data tracking and storage technologies and increased availability of visualization toolsets to collectively produce valuable outcomes (Connolly & Wooledge, 2012).

Visualization provides specific advantages for each facet of big data: volume, velocity, and variety. In response to the volume parameter, big data generate sheer amounts of information that are too large to process without visual representations and interpretations. In response to velocity, it becomes inefficient and infeasible to have manual steps in the analysis process when data are arriving so rapidly. Visualizations provide a data

structure that can be quickly updated and revised as more data become available, improving response time between data retrieval and decision guidance and reducing the need for users to re-acclimate themselves each time. Without the speed and synthesis advantages that visualizations provide, organizations will struggle to harness big data's potential for accelerating their growth and bolstering their competitive advantage. The velocity of big data also matches well with visualization's ability to show trending and time-series data. Visualization addresses the variety aspect of big data by visually aligning and integrating different data sources, providing a common visual structure for interpreting various forms of data. Visualization techniques can also incorporate unstructured as well as structured varieties of data, important as the former becomes increasingly prevalent and often can be more easily processed, categorized, and acted upon when visualization is a step in the analysis process. Some definitions of big data introduce a fourth "V," veracity, dealing with the quality of the data. Visualizations can serve a useful function for this component also, as they are ideal for identifying outlier and low-base rate occurrences within big data information sources to drive further investigation of the nature of these cases and, if appropriate, removal or reconciliation to produce a cleaner dataset for further analysis.

BUSINESS APPLICATIONS OF DATA VISUALIZATION

Alongside immense business pressures to generate and utilize big data, visualization is viewed as a key mechanism for unlocking the value of these data. Organizations accumulate an abundance of raw information, and their ability to produce smart enterprise decisions is in part based on their acumen and speed in examining and interpreting these data (Bonneau, Ertl, & Nielson, 2006). If critical data insights are present but not uncovered, it becomes increasingly likely that organizations will make weak business decisions as a result and will fall behind their competitors with higher proficiency in big data management. As organizations are charged with data-driven decision making, this also requires that a broader range of individuals becomes comfortable with, and sophisticated in, using data to guide decisions—and in this context, visualization's role to enable decisions becomes paramount. As big data are seen as an underleveraged source of business-critical insights, decision makers at all levels need to become much more adept at rapidly capturing, analyzing, and extracting value from the data they are accumulating.

As noted above, recent studies of the workforce impact of big data and analytics trends have noted the immense need for more individuals with analytical skills. Visualization is seen as one mechanism for reducing this gap—for allowing data-driven decision making without introducing a heavy reliance on quantitative skills. Within a business environment, visualization also drives collaboration among groups, providing a shared view for interpretation (Eppler & Bresciani, 2013). Visualization can also aid organizations in ensuring that their data are sufficiently clean and relevant to the business questions of interest, as it allows individuals who lack traditional quantitative analysis skills to nonetheless get close to the data and to identify issues for resolution. Further, visualizations serve an important function to combat data fatigue and skepticism: they provide a direct response to the "have to see it to believe it" viewpoint. Decision makers charged with making decisions at a higher pace and supported by a solid evidence foundation can more easily synthesize and communicate information when presented visually (Al-Kassab, Ouertani, Schiuma, & Neely, 2014; Parsons & Sedig, 2014). It is difficult to overstate the potential value of visualization for framing research findings in a manner that engages a non-technical audience, helping them interpret and internalize and simply making them want to share the information with others.

VISUAL PERCEPTION AS A FOUNDATION FOR DATA VISUALIZATION

The impact and advantages of data visualizations are based in their ability to leverage our extremely well-developed systems of perception, attention, and memory, enabling us to process visual information at a similar rate to that of an Ethernet connection (Koch, Mclean, Segev, Freed, Berry, & Balasubramanian, 2006). In particular, the speed advantages of visualizations begin with the initial stages of perception and the use of certain visual features, termed pre-attentive, that can be detected during a single glance lasting approximately 200 to 250 milliseconds (Healey & Enns, 2012). These features—such as position, length, density, and color—serve as the core elements of a visualization, drawing on our most deeply rooted perceptual skills. Though these pre-attentive features all provide an expedited path to perception, they can also be ordered in terms of the accuracy with which they can be interpreted (Cleveland & McGill, 1985). MacKinlay's (1986) research produced such an ordering of visual features—in this ranking, those higher on the list are generally better-suited for visual design features

to enable interpretative accuracy due to their higher precision, and those lower on the list less-suited.

TABLE 5.1

	Categorical (Nominal)	**Categorical (Ordinal)**	**Quantitative (All)**
Most accurate	Position	Position	Position
	Color hue	Density	Length
	Texture	Color saturation	Angle
	Connection	Color hue	Slope
	Containment	Texture	Area
	Density	Connection	Volume
	Color saturation	Containment	Density
	Shape	Length	Color saturation
	Length	Angle	Color hue
	Angle	Slope	Texture
	Slope	Area	Connection
	Area	Volume	Containment
Least accurate	Volume	Shape	Shape

In addition to the processing speed benefits of visualizations that take advantage of pre-attentive perception, our memory for visuals in comparison to text-only information is also enhanced (Schnotz, 2002). This advantage is due to the use of two different cognitive subsystems applied to the processing of visual information—a concept referred to as dual coding theory (Clark & Paivio, 1991). In this theory, textual information is processed only within the verbal cognitive subsystem, yet visual information is encoded in the imagery subsystem as well.

Despite the extensive research base on perceptual and cognitive aspects of information processing, many unanswered questions remain for most facets of data visualization, particularly for newer types of visualizations (Johnson, 2004; Chen, 2005; Chen, 2010). Though early indications are positive about the ability of data visualizations to influence and create shared knowledge for a managerial audience (Al-Kassab, Ouertani, Schiuma, & Neely, 2014) and to exert more persuasive influence than purely numerical forms of information presentation (Pandey et al., 2014), much research remains to be conducted to validate these early signs of promise within applied settings.

DATA VISUALIZATION TYPES AND TOOLS

The field of available visualization methods is expanding rapidly—regardless of the type of data, there are an increasing number of approaches for communicating a message, and a growing number of tools available to facilitate visualization creation. Deciding on an appropriate type of visualization can involve a wide array of choices, but begins with two straightforward questions: what is the intended message and who is the intended audience. Only after making those considerations can the right visualization technique be chosen (Zhu, 2007). Choice of visualization method may also be enhanced by taking into consideration the input of end users (Fox & Hendler, 2011) or past examples of visualizations that have been more or less effective to increase the probability of a successful outcome. Due to their nature, visualizations are also extremely well-suited to "window shopping" examples drawn from other contexts, curated and compiled by visualization experts such as Stephen Few (2009), Alberto Cairo (2013), and Andy Kirk (2012); even a simple Google image search for any of the visualization types described in this chapter will typically produce thousands of examples.

Numerous taxonomies exist for classifying data visualization methods. I focus on the five-category structure proposed by Kirk (2012): comparing categories, assessing hierarchies and part-to-whole relationships, showing changes over time, plotting connections and relationships, and mapping geospatial data. I will highlight and briefly overview each of these categories, including a representative example (in most cases smaller-scale to allow legibility; in practice many visualizations would be displayed in a larger printed size or on a computer screen) of methods that are relatively common and that are well-suited to big data applications, corresponding to the focus of this volume. Of course, few, if any, of the techniques presented are exclusively associated with big data, but several derive greater value as a communication mechanism as the underlying datasets they represent grow larger. In addition, any of the visualization types shown can also be represented in interactive form, enhancing their utility for large-scale datasets. For more expansive lists and classification systems of graphical and visualization methods beyond those listed in the following sections, including those that are either less unique to big data visualizations or less frequently used, refer to Abela (2008), Few (2009), Heer, Bostok, and Ogievetsky (2010), and Kirk (2012).

COMPARING CATEGORIES

This data visualization category includes visualizations designed to compare groups based on their corresponding values. For I-O psychologists, this set of visualizations can be beneficial for exploring research questions comparing years, employee or organizational subgroups, and individual or organizational characteristics, for example. The sample visualizations below provide further illustrations of potential applications.

Slopegraph

A slopegraph compares various categories by visually connecting their values using a line. Common category examples are year (e.g., 2013 versus 2014) and level (e.g., front-line leader versus senior executive). By connecting the same category across representations, a slopegraph visually indicates rate of change between the representations, as well as each category's relative order compared to others to show how rank ordering

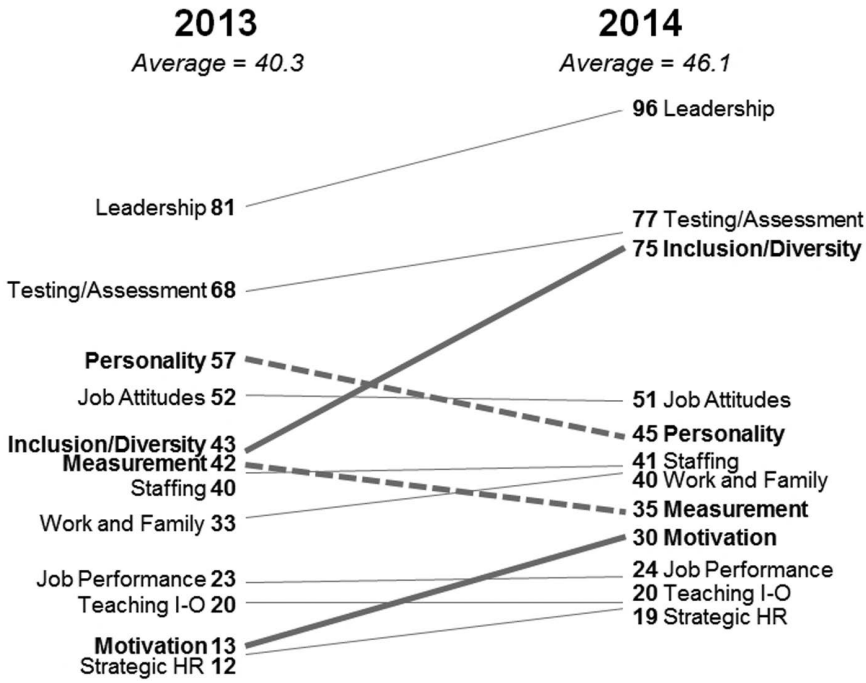


FIGURE 5.1
Slopegraph chart

varies. Slopegraphs can also benefit from the use of additional visual cues for the lines connecting categories—for example, coloring lines of upward slope green, downward slope red, and nominal slope gray or using bold or dashed lines to show noteworthy upward or downward trends. An example of a basic slopegraph is shown in Figure 5.1—this graph shows a subset of topic areas within the Society for Industrial and Organizational Psychology (SIOP) annual conference and compares the number of conference sessions on that topic between 2013 and 2014. Content areas with a high percentage increase are shown with a bold line, those with a high percentage decrease are shown with a dashed line.

Radial Chart

A radial chart uses a circular display format, with data categories ordered around the circle to show the standing of each category on a common numerical or ordinal scale ranging from the inside to the outside of the circle. In addition to distance from the center, radial charts can also incorporate color and shading variations to indicate other distinctions among categories, and their flexibility provides numerous advantages for large-scale data representation. Florence Nightingale's seminal 1858 chart of war mortality was a form of radial chart; a more modern radial chart using seven of the same SIOP Annual Conference content areas from the slopegraph example is shown in Figure 5.2.

Sankey (Alluvial) Diagram

A Sankey, also called alluvial, diagram shows how various categories flow together or apart across stages (indicative of their intercorrelations), with stages often representing, but not limited to, multiple time periods. Sankey diagrams can be thought of using a water analogy, where tributaries join to form larger streams or rivers split to form various branches. The width of the water flowing shows how categories change in size. Charles Minard's 1869 influential march on Moscow graphic, while it incorporates other visualization techniques as well, is a form of Sankey diagram, showing the joining and dividing of Napoleon's army during its attempted incursion into Russia. The example Sankey diagram (Figure 5.3) shows how a set of individuals experiencing job transitions moved between levels, displaying the number and proportions following each flow. This visualization type is particularly well suited to "to" and "from" research questions within I-O psychology, for example to illustrate employee flows between levels, countries, and companies.

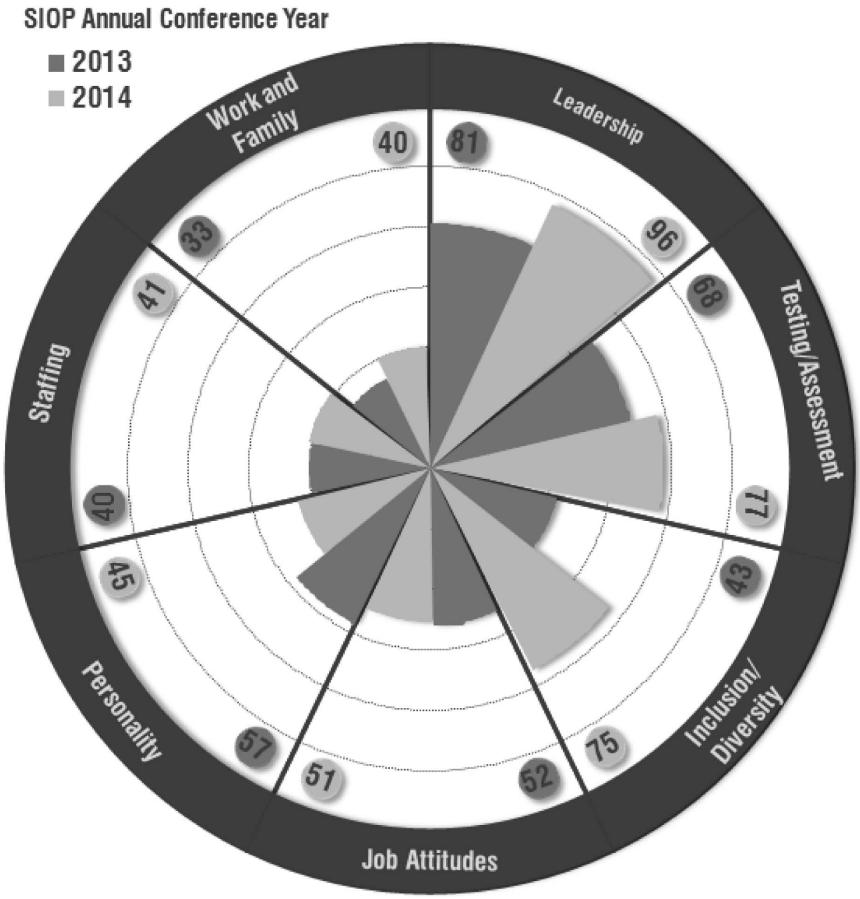


FIGURE 5.2
Radial chart

Small Multiples (Trellis) Chart

Small multiples charts, also referred to as trellis charts or sparklines, take advantage of humans' visual skill at detecting patterns to simultaneously depict—using various panels—several variables displayed individually but with common horizontal and vertical scales. Common examples of scales to enable accurate comparisons in a small multiples format are time along the horizontal axis (i.e., using a common start and end date), and percentage change for each variable along the vertical axis. Though the component graphs are typically line or area graphs, it is nonetheless useful to think of small multiples as a distinct visualization type due to the incremental visual impact and diagnostic potential of an integrated view

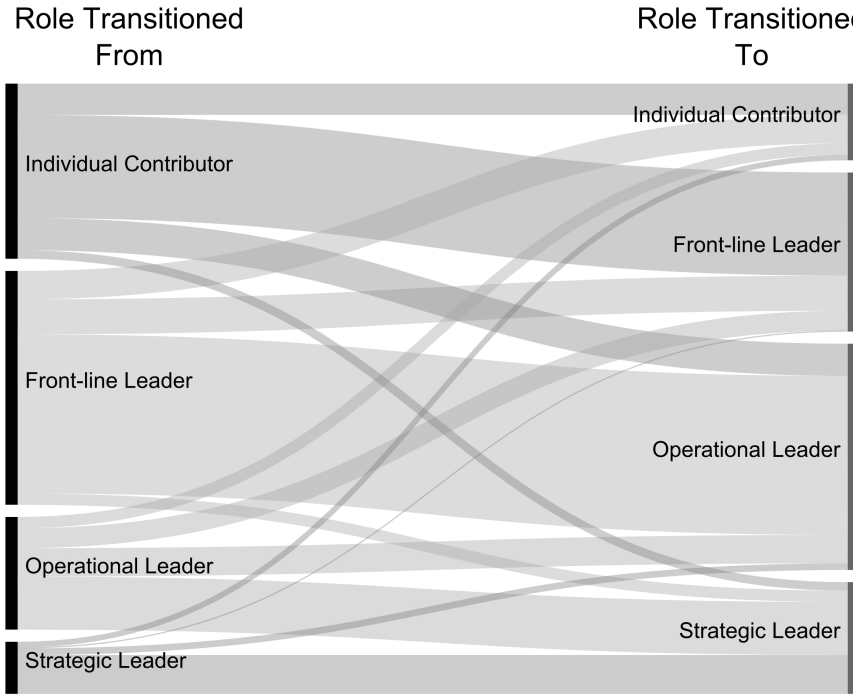


FIGURE 5.3
Sankey (alluvial) diagram

of the individual graphs. A small multiples example is shown Figure 5.4, again referencing a subset of content areas within the SIOP annual conference, but now extending that view across a wider range of years. For I-O psychologists, this visualization type can be useful for exploring virtually any type of longitudinal dataset. Alternatively, the X-axis can be used to represent ordered categories such as job level or continuous variables such as tenure.

Word (Tag) Cloud

A word cloud is a particularly distinct visualization type, as it is expressly designed for use with unstructured, textual data as an initial input and to display individual words and phrases based on their frequency and other characteristics that can be qualitatively extracted, such as word tone. The shape of the word cloud itself can also be modified to further

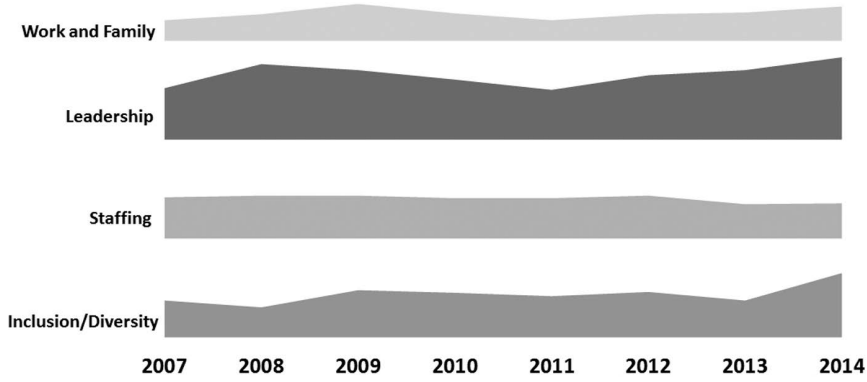


FIGURE 5.4
Small multiples (trellis) chart

emphasize the characteristics or context of the text represented (e.g., the responses from a country represented in the shape of that country). For this type of visualization, it is particularly critical to evaluate and preprocess the underlying data before creating the visualization, as misspellings, hyphenations, and overrepresentation of less valuable words such as articles are unfortunately very prevalent in any large-scale text database. More so than most visualization methods, the visual appeal of word clouds has driven very widespread use—a colleague of the author even reported that her 8-year old daughter created word clouds as a third grade class project. Thus, the potential for oversaturation of this approach should be recognized accordingly. However, an advantage of this method's uniqueness is that it has also driven more focused research and discussion (see Feinberg, 2009; Viégas, Wattenberg, & Feinberg, 2009). An example word cloud is shown in Figure 5.5, showing the most common responses among 13,000 leaders indicating the one word that best described their organization's leadership development program, with positive terms shaded in lighter gray and neutral or negative terms shaded in darker gray. Though this use, to display words or phrases sized based on frequency, is more common, I-O psychologists should also consider less traditional uses of word clouds. For instance, multiple word clouds can be generated to represent relative frequencies of competencies, countries, personality characteristics, or any other category with a large membership, with each limited to a particular subgroup (e.g., based on demographic characteristics, tenure, or turnover risk category) in order to facilitate visual exploration and comparison of their responses.

FIGURE 5.5
Word cloud

ASSESSING HIERARCHIES AND PART-TO-WHOLE RELATIONSHIPS

This data visualization category includes visualizations designed to depict how a particular category relates to a broader set as a component of this total or how a hierarchical structure extends through a dataset. This hierarchical structure, and the many layers of detail that can be present, also makes this category of data visualization ideal for interactive approaches such as those discussed in a later section. Visualization types within this section are useful for I-O psychology applications involving hierarchical data structures such as regions and countries, job levels, and business units, for example.

Circle Packing Diagram

Circle packing diagrams group circles into larger circles, with smaller circles representing subsets of the data, and larger circles representing supersets. Circles within a grouping are sized based on their amount or other quantitative characteristics. Circle packing methods are considered space-efficient for big data applications, yet are often unsuitable for trend and time investigations and cannot accommodate negative values used for sizing elements, making them less suited for certain variety and velocity

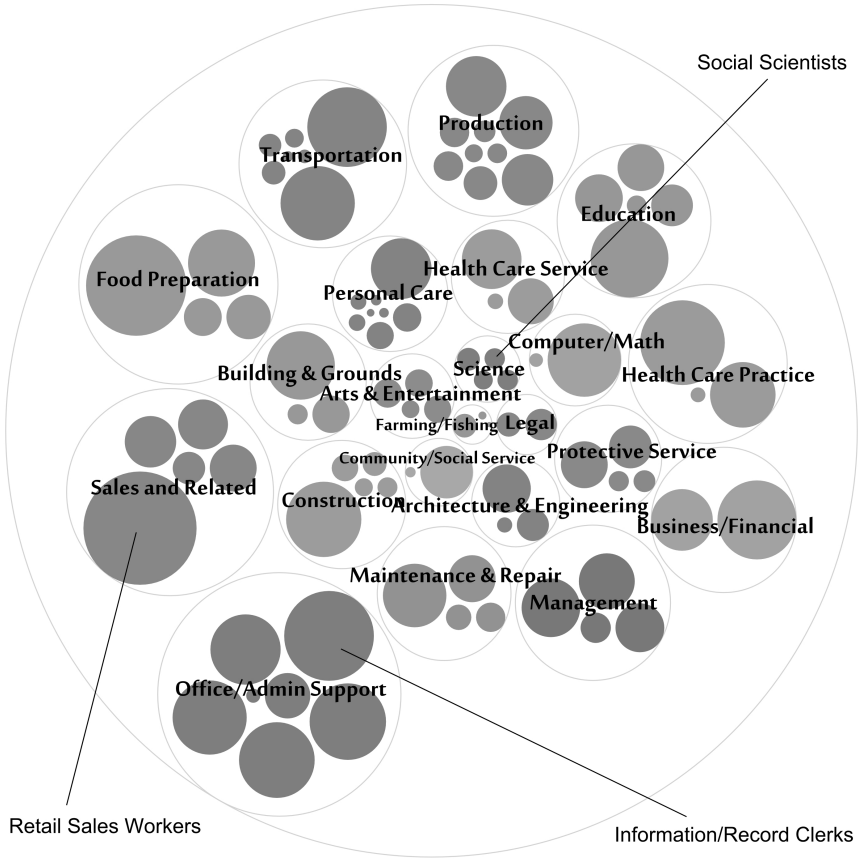


FIGURE 5.6
Circle packing diagram

aspects of big data (Gorodov & Gubarev, 2013). An example circle packing diagram is shown above (Figure 5.6), displaying two levels of structure and sizing of job categories within the United States based on 2013 employment levels, both sized to their relative proportions as well as highlighting three more specific job classifications.

Tree Map

Tree maps use a rectangular format to partition a dataset's components based on their size or other relative value (e.g., a larger component is displayed as a larger proportion of the overall rectangle). In addition to size, color can also be used to differentiate components—using more than one

FIGURE 5.7
Tree map

distinguishing characteristic often becomes critical when multiple hierarchies are displayed within the same visualization. Regarding applicability to big data applications, tree maps are considered less space-efficient than circle packing and are subject to a similar set of disadvantages and some degree of insufficiency in response to big data variety and velocity (Gorodov & Gubarev, 2013). Tree maps are effective at showing hierarchical groupings and data outliers in the forms of particularly large and small groups relative to others. A well-known example of an interactive tree map is the website Newsmap (http://newsmap.jp/), which displays news grouped into categories such as sports, business, entertainment, and technology, with the size of a particular news story a function of its popularity. An example tree map is shown above (Figure 5.7) using the same data as above and with three more specific job classifications indicated.

Sunburst Diagram

Sunburst diagrams provide a concentric layout for a hierarchical structure, with subsets of data extending out from their supersets. Each new ring outward represents another layer in the hierarchy. This property, to efficiently display multiple levels of the structure, allows sunbursts to show data proportions as well as category depth since some categories

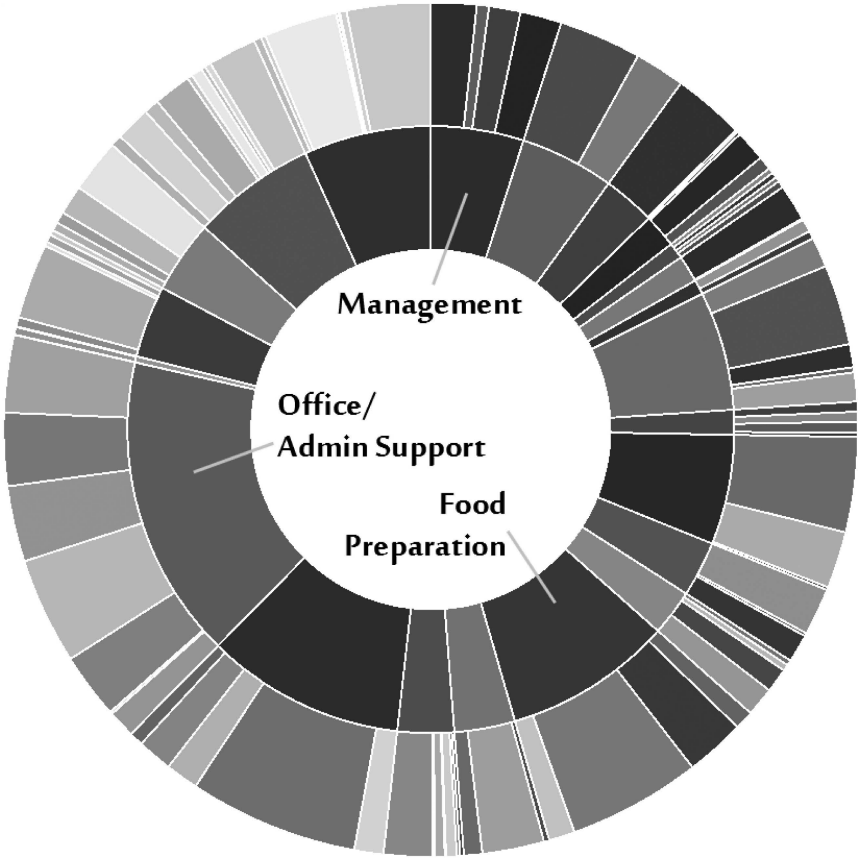


FIGURE 5.8
Sunburst diagram

will extend farther from the center than others. When viewed in the context of big data applicability, sunbursts share the same disadvantages as tree maps and circle packing for high-variety scenarios, yet do provide advantages for velocity due to the possibility of representing data through animation (Gorodov & Gubarev, 2013). An example sunburst diagram is shown above (Figure 5.8), displaying the same United States employment data as referenced earlier down to two levels of job categorization. Due to space limitations, only representative top-level categories are indicated. Sunburst diagrams have similar applications to circle packing and tree maps for displaying and guiding research exploration of hierarchical data structures within I-O psychology.

Cluster Dendogram

A cluster dendogram is a form of node-link diagram that presents a hierarchy in a circular format, with the deepest levels of the hierarchy placed uniformly along the outer ring and higher-order levels, or clusters, placed inward toward the center. These diagrams do not show the relative sizing of categories as do the prior types in this section but can accommodate a larger amount of text descriptors and show numbers of subcategories more efficiently than do the types above. An example cluster dendogram created based on three levels of structure within the United States employment dataset referenced above is presented in Figure 5.9. Though the text

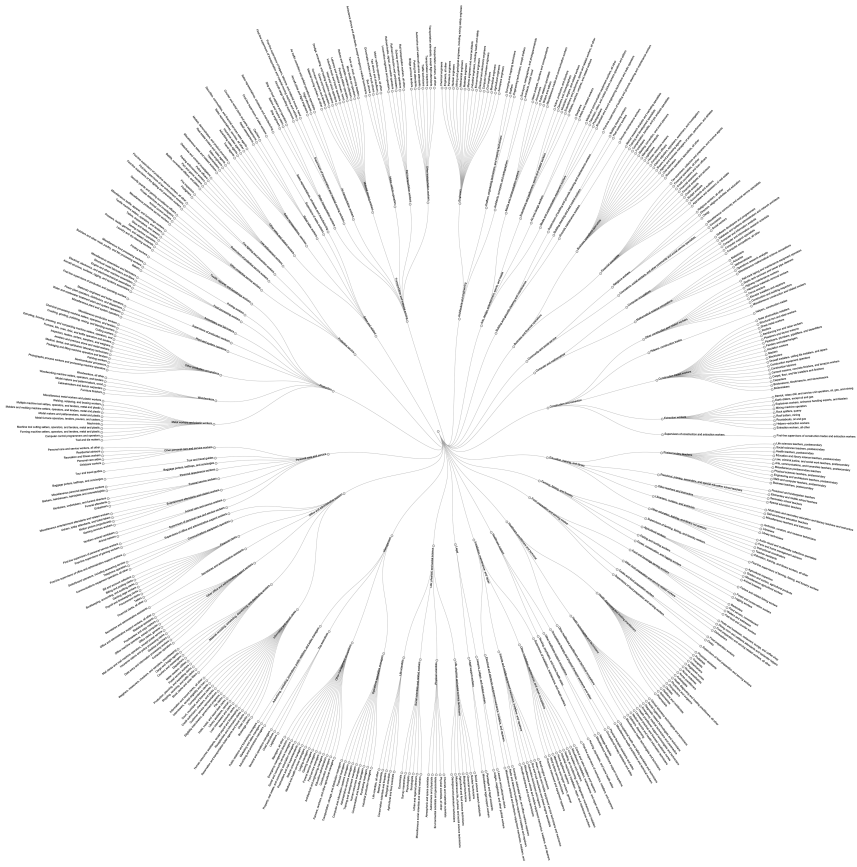


FIGURE 5.9
Circle dendogram

necessary to represent this complex structure is too small to be legible within the available page size, this example nonetheless shows the nature of this visualization type.

SHOWING CHANGES OVER TIME

This data visualization category includes visualizations designed to show variation across a time span. For I-O psychologists, these visualization types are useful for any form of longitudinal data where the focus is displaying change or absolute level of a set of values over time. These visualization types can also be adapted such that their X-axes are any ordered category (e.g., to show how various skills grow or fall in importance from one job level to the next).

Horizon Chart

A horizon chart is designed to efficiently represent time-series data that include both negative and positive values by using coloring and shading to represent negative values that have been transposed above the baseline, or "horizon." This property allows these charts to provide a higher rate of data density and rapid interpretation of over-time patterns. An example horizon chart is shown in Figure 5.10. This chart depicts the year-to-year

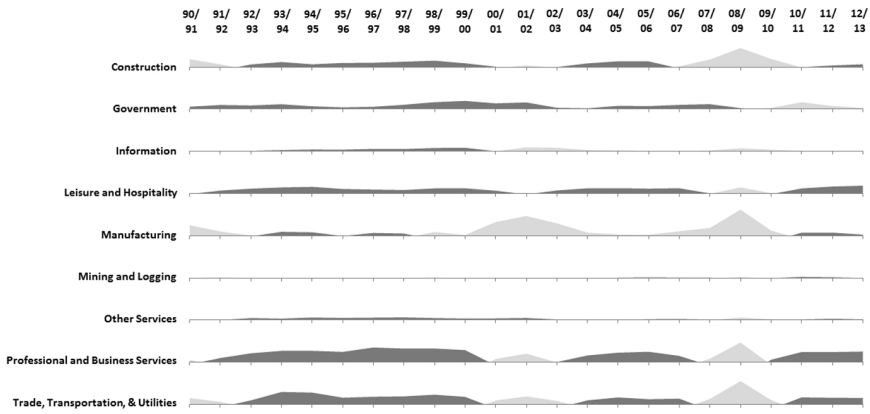


FIGURE 5.10
Horizon chart

United States employment change by industry from 1990 until 2013—darker-gray indicates a positive change from the prior year, lighter gray indicates a negative change, and in both cases height indicates the absolute number of jobs gained or loss within that industry. All industries use the same vertical scale to allow direct comparisons.

Stream Graph

A stream graph is a form of stacked area graph that displays how the relative proportions of data vary over time. While stream graphs do not support negative numbers and can obscure precise differences in some cases, they can facilitate identification of flow patterns, particularly when data are available over long periods of time and where a relatively small number of categories are present. Stream graphs are also well suited to interactivity, allowing exploration of individual segments of the dataset. An example stream graph is shown below displaying the same year-to-year United States employment change by industry from 1990 until 2013 data as above. In this case, however, the graph focuses on the absolute value of each industry's employment rather than the year-to-year change.

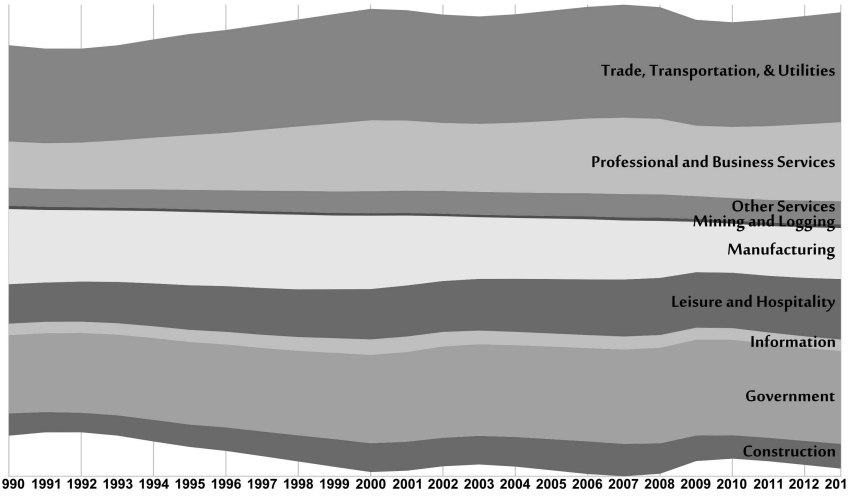


FIGURE 5.11
Stream graph

PLOTTING CONNECTIONS AND RELATIONSHIPS

This data visualization category includes visualizations designed to show how two or more variables relate to one another. As such, these visualizations can be useful for a wide range of research questions explored by I-O psychologists.

Bubble Plot

A bubble plot is a scatter plot—one dimension on the X-axis, one on the Y-axis, and data elements plotted in terms of their relative positions on these scales—but adds a third additional dimension of bubble size to indicate magnitude or another quantitative property. In some cases, a fourth dimension may be denoted by color. An example bubble plot is shown below, displaying several countries (intentionally left unlabeled for sample purposes, although the data are real) plotted in terms of changes in current leadership quality on the X-axis and leader "bench strength" (i.e., potential future leaders' projected abilities to fill key roles over the next three years) on the Y-axis between 2011 and 2014. Bubble size indicates the GDP growth rate for each country.

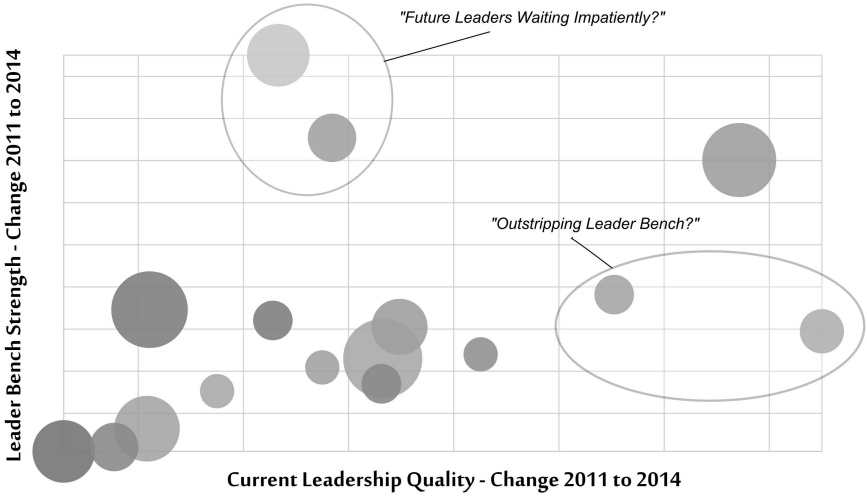


FIGURE 5.12
Bubble plot

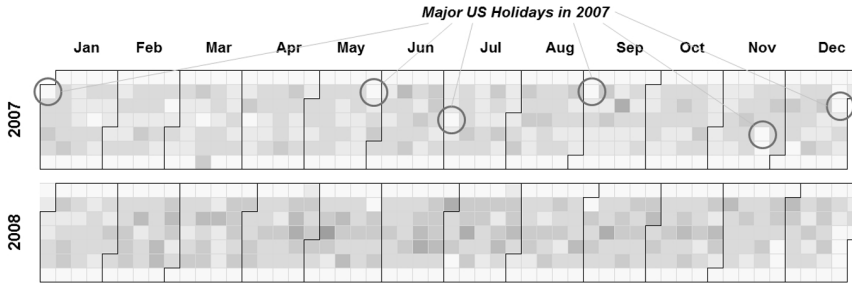


FIGURE 5.13
Heat map

Heat Map

A heat map displays the strength of relationship between any two variables in a matrix format using colors or shading to denote stronger connections. Because the level of detail represented within a heat map is relatively low (which also limits its precision for visual analytic purposes), this approach can make use of individuals' perceptual skills at detecting color and hue variation to display an extensive set of variables simultaneously, in some cases including hundreds of individual relationships. An example heat map—using a calendar view—is shown above (Figure 5.13). This example shows the candidate volumes for an organization's employment testing program over the course of two full years, 2007 and 2008 (the first row of each year represents Sundays, with further days of the week completing the other six rows).

Parallel Coordinates

Parallel coordinates are designed to visualize multivariate data by arraying a series of variables along an X-axis, showing the position of each data element on each axis using a point along the Y-axis, and then connecting these points with lines. The visual connections provided by these lines allow the user to rapidly view the profile of a particular case of data compared to others and when paired with interactive functionality, to identify a target profile and detect other cases with a similar profile. Parallel coordinates are viewed as a particularly versatile and powerful visualization method. In Gorodov and Gubarev's (2013) analysis of six common data visualization methods, parallel coordinates were the only method designated as well-suited for big data's velocity and variety characteristics, as well as its

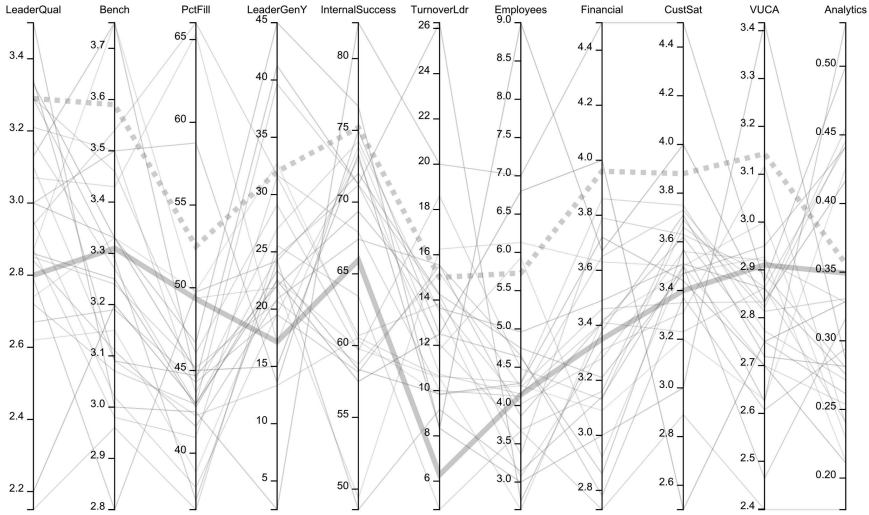


FIGURE 5.14
Parallel coordinates

volume. An example interactive parallel coordinates visualization displays data extracted from the USDA Nutrient Database: http://exposedata.com/parallel/. The example parallel coordinates visualization above (Figure 5.14) shows how 28 countries (intentionally unlabeled for this example) vary across average scores on 11 variables of organizations within each country.

Chord Diagram

Chord diagrams use a circular structure to display relationships between category members as lines connecting every pairing. Chord diagrams can incorporate dozens of individual category members and can also incorporate line color, shading, and thickness to show the strength and nature of interrelationships, as well as a category member's length along the circumference as a further indicator of its collective degree of interconnectivity. An example chord diagram is shown below. This chord diagram example, based on the co-occurrence of 10 leadership skills as a development focus within a large-scale organization sample, is best viewed in three parts: First, the outer ring—in this ring, skills extending across a greater portion of the circumference are more correlated, on average, with other skills (i.e., the skill most intercorrelated with the others is integrating culturally). Second, the lines extending from the right half of each skill's portion of the ring—the width of

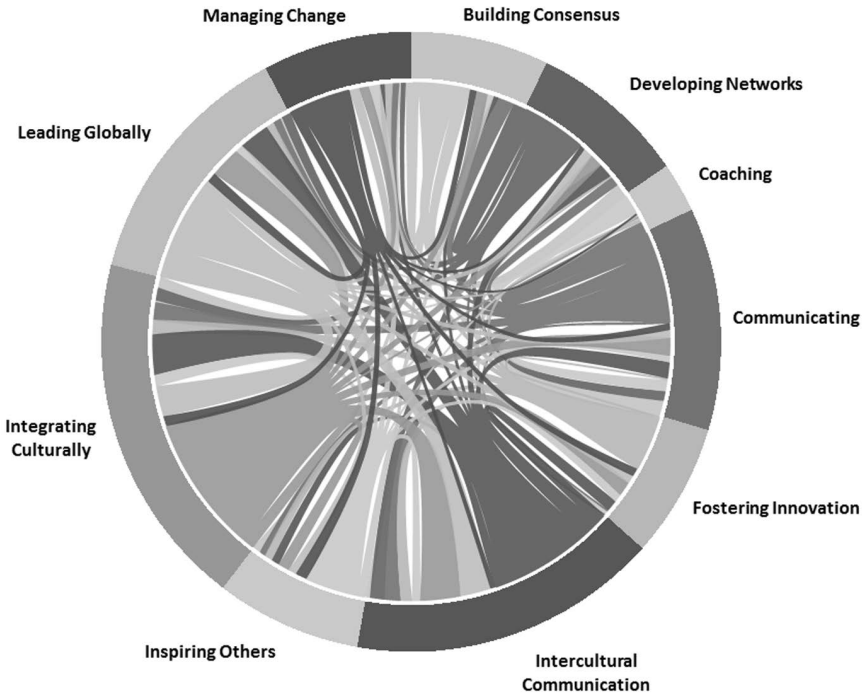


FIGURE 5.15
Chord diagram

these lines indicates the intercorrelation or co-occurrence of those skills (i.e., intercultural communication is most intercorrelated with integrating culturally). Third, the lines extending from the left half of each skill's portion of the ring—the width of these lines indicates the corresponding intercorrelation of each other skill (i.e., intercultural communication has a low co-occurrence with managing change).

Table Lens

A table lens is designed to enable rapid detection of intercorrelated variables by showing numerous variables simultaneously. After choosing and sorting a single column's values highest to lowest, all other columns inherit that sort order accordingly, allowing the user to see which other columns are positively (with a similar pattern of long to short horizontal bars) or negatively (with the opposite pattern of short to long bars) correlated. Due

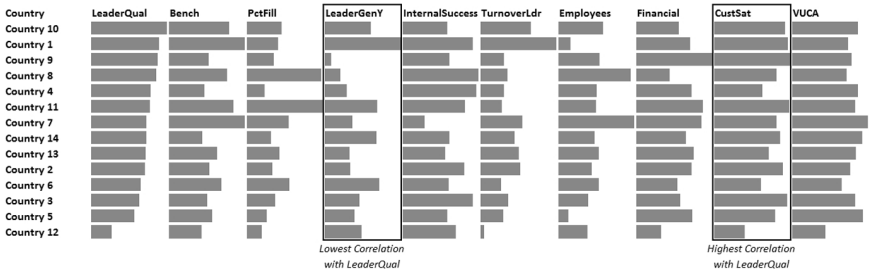


FIGURE 5.16
Table lens

to its efficiency in allowing the user to explore numerous potential intercorrelations in a single view, a table lens can serve as an initial exploration step for a new dataset. An example table lens is shown above (Figure 5.16), displaying data from 12 countries (intentionally unlabeled for this example) on 11 variables sorted by the first variable (average leader quality). Of the other variables, those that share a similar high to low pattern have a high positive correlation with leader quality. If any variables had a high negative correlation, they would show a low to high pattern. Those that show little correspondence have a nominal correlation.

MAPPING GEOSPATIAL DATA

Choropleth and Dot Plot Maps

A choropleth map uses a similar concept as a heat map but applies the color and shade variations to a map rather than a matrix. This can be done at the lowest-level visual, where mutually-exclusive borders can be established and displayed—depending on the size of the map, this could be counties, states, countries, continents, or other existing groupings. A minor variation is a dot plot map, which represents distinct geographic entities as dots rather than by their actual shapes (an example dot plot map showing the spread of the Code Red computer virus over a single day in 2001 can be viewed at http://www.caida.org/research/security/code-red/newframes-small-log.gif). An example choropleth map is shown below—this chart displays the proportion of SIOP annual conference attendees from each

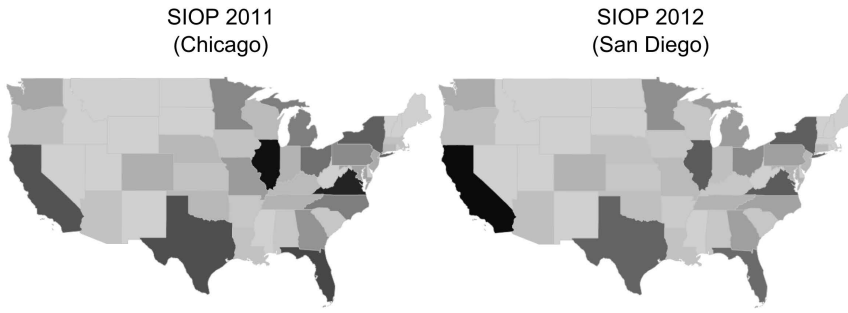


FIGURE 5.17
Choropleth

state in the continental US for 2011, when the conference was held in Chicago, Illinois, and for 2012, when the conference was held in San Diego, California.

ANIMATED DATA VISUALIZATIONS

Animated visualizations are a distinct form of data presentation. The most notable example of an animated visualization—noteworthy both because of the heavy influence it exerted across a broad audience and because of its compelling topical focus—is that designed by Hans Rosling and made available in various forms, including a TED talk drawing over eight million views (http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen) and an extensive website, Gapminder, drawing on an expanded set of data and accompanying videos displaying various visualizations in animated form (www.gapminder.org). The popularity of this particular graphics-focused presentation speaks positively to the potential benefits of animation to enhance visualization communication.

While the available research on static visualizations is limited, studies on animation are even more so (Shah & Hoeffner, 2002). Some studies have shown that applying animation to visual displays can facilitate detection of three-dimensional aspects of the data such as clustering and interrelationships among three variables concurrently (Becker, Cleveland, & Wilks, 1988; Marchark & Marchark, 1991). However, other studies have raised questions about an audience's ability to appropriately

interpret relationships and trends when animated (e.g., Huber, 1987; Stuetzle, 1987). As a further indicator of the complexity and relatively early state of research into animated forms of visualization, Kriz and Hegarty (2007) detected an interaction between animation exposure and learner knowledge such that high-knowledge learners were more likely to revise their mental models after multiple exposures to the animated visualization compared to low-knowledge learners.

INTERACTIVE DATA VISUALIZATIONS

Though static forms of visualization are sufficient for many big data applications, interactive data visualizations are often critical for deep exploration of large-scale datasets. Interactive visualizations move beyond the representation of data to allow users to dynamically change and focus their view of the information. Importantly, many forms of visual analytics, such as the parallel coordinates example above, are limited or impossible with static visualizations alone—interactivity enables a form of "self-service" analytics that can greatly expand the impact and utility of visualization approaches.

Whereas static visualizations are created for a predefined purpose and application, interactive visualizations allow—often through controlled access to the full underlying dataset—the audience to define the message and insight of greatest interest to a particular context and business question. In many cases, the data views derived from an interactive approach can be saved for repeated use. Because the users of interactive visualization methods bring new content knowledge and expertise, they may detect findings that the original researcher had not. For this reason, interactive visualizations are particularly valuable when information is being shared with an audience with deeper or differing perspectives. Well-designed interactive approaches also provide a feedback loop to the original researcher to help progressively define new research questions.

Interactivity in the context of big data visualization can be classified on a number of dimensions—Soo Yi, Kang, Stasko, and Jacko (2007) proposed a seven-category taxonomy for interaction techniques based on their review and synthesis of existing frameworks, emphasizing the interplay between the user's goal in seeking a particular form of interaction and the specific mechanisms used to accomplish these objectives. This taxonomy also covers the majority of interaction techniques available in commercial data visualization products.

Soo Yi et al.'s (2007) first interaction technique is to select, or to assign a marker to, a particular data element for revising and further investigation. Importantly, once a selection is made, it is retained through further manipulations of the data view to facilitate tracking of, and the ability to easily return to, a particular case or set of cases throughout. In many software packages, selecting is accomplished simply by clicking on the case or set of interest.

A second interaction technique from this taxonomy is to explore a different portion of the dataset, typically a parallel partition (e.g., shifting between countries). In visualization software packages, exploring often involves panning the view to the new target or clicking on a segment, which in turn reorients the visual display to place the new target at the center. Exploration functionality is often essential for big data visualizations since the scope of data precludes full visibility in any one on-screen viewpoint.

A third interaction technique is to reconfigure the user's perspective on the data. Common reconfigurations include sorting, changing axis ranges, establishing a new baseline, rotating a 3D view of the data to improve the visibility of certain cases, and reordering data elements such as columns to better align with a more intuitive ordering. A particularly unique form of reconfiguring deployed in visualization software—useful when many data points overlap, thus making their density invisible—is a "jitter" technique, which shifts the position of each case very slightly and randomly.

A fourth interaction technique from Soo Yi et al.'s (2007) framework is to visually encode properties of data elements such as their size, color, and shape. While such encoding choices can seem relatively minor, they can have a substantial influence on the interpretability of a data visualization. For example, poorly chosen colors or shapes can slow the pre-attentive velocity that provides key advantages for visualizations above other data presentation methods, as discussed above. Whereas software packages will produce a default encoding, the researcher may alter this based on the objectives of the visualization.

A fifth interaction technique—and one particularly relevant to big data applications—is to abstract/elaborate, to orient the view to a lower level of the data structure (e.g., individual business units versus the organization as a whole) or back to a higher-order structure. In some cases, this may allow examination at the level of individual cases. Certain visualization types, such tree maps and sunbursts, utilize abstract and elaboration approaches extensively. For other visualization types, such as scatterplots,

elaboration through a zooming function is essential to spread out and detect the details of individual positions within a tightly clustered set of data points.

A sixth interaction technique within this taxonomy is to filter, to reduce the number of data elements displayed in the visualization based on predefined parameters. When a filter is applied, data elements not meeting the specified condition are removed from or de-emphasized within the view. Filtering is a major feature set within visualization software packages, in many cases driven dynamically using slider bars placing parameters on a high to low scale or by using checkboxes or radio buttons to set categories. In many business applications of big data, compiled datasets serve a wide variety of functions, and for any one visualization objective, a sizeable portion of the dataset may be irrelevant. Filtering techniques ensure that extraneous data can be excluded and do not cloud messaging and interpretation.

Soo Yi et al.'s (2007) final interaction technique is connect, to show related items across or within visualizations. If multiple visualizations are presented for the same dataset, connect interactivity allows the researcher to simultaneously view the appearance of a single data point in each view. Parallel coordinates also rely heavily on this interaction technique, as one of that visualization type's primary advantages is its ability to guide identification of other data elements possessing a similar profile across the variables displayed.

Few big data visualizations reach their full illustrative and diagnostic potential without interactivity. While interactive visualizations undoubtedly provide major advantages over static forms, these advantages are accompanied by three important consequences for when and how they are deployed by psychologists. First, some—though not all—of the interactive techniques require the availability of and the researcher's proficiency in commercial visualization software. This dependence, paired with a situation where only one such toolset may be available, makes interactive capabilities a key decision criterion when evaluating visualization software for purchase and adoption. Second, it creates a potential disconnect between the full potential of data visualizations (i.e., incorporating their interactive components and presentation/publication avenues for the work). While less an issue when interactive visualizations are being shared internally to an organization, it does mean that visualizations published in static formats such as print will be relatively limited in their ability to inform and influence new audiences. Third, interactive visualizations rely on deep data access and availability that may not always be

feasible and appropriate, particularly when sharing these interactive views outside of the core research team.

ONLINE VISUALIZATION TOOLS

Due to the vast range of data visualization software packages available, this chapter will not be reviewing specific options due to their number and commercial nature. However, numerous online—and free in basic form—tools now exist to enable a wide range of visualizations, encompassing many of those described above and others. Four specific such tools are briefly discussed here: Wordle, Infogr.am, Raw, and D3.js. Despite its commercial nature, due to its widespread availability, I also include a brief description of visualization capabilities available using Microsoft Excel and its extensions.

Wordle (http://www.wordle.net) is an online tool for creating word and tag clouds, using raw data, or (in its advanced form) word frequencies as input. Wordle (used to generate the word cloud earlier in this chapter) provides a range of text generation options including the ability to color words distinctly based on their properties as entered into the site.

Infogr.am (http://infogr.am), which offers a free as well as a paid version, is a toolset for producing static and interactive infographics. As part of its functionality, Infogr.am allows users to generate a range of visualization types including radial charts, bubble charts, word and tag clouds, and tree maps.

Raw (http://app.raw.densitydesign.org/) allows users to enter their own data and to generate Sankey (alluvial) diagrams, circle packing, cluster dendograms, parallel coordinates, tree maps, and stream maps, among other data visualization types. When paired with a vector graphics editor such as Inkscape (http://www.inkscape.org/en/), Raw can rapidly produce free, high-quality visualizations.

D3.js (http://d3js.org/) involves a heavier requirement for programming knowledge, yet is perhaps the most expansive of the tools mentioned, providing JavaScript functionality to generate chord diagrams, circle packing, stream graphs, sunbursts, parallel coordinates, scatterplot matrices, tree maps, choropleths, Sankey (alluvial) diagrams, word clouds, and others (it is also the toolset underlying Raw).

Finally, Microsoft Excel, though relatively limited in its native visualization capabilities, can be configured to produce many of the visualization types listed above; a large and often-updated compilation of such

templates is available at https://sites.google.com/site/e90e50charts/. The 2013 version of Excel also enables numerous visualization types through add-on apps designed for this purpose (Knies, 2013). Specialized software is not needed to generate data visualizations; all examples presented above were generated with real data using either Raw or Excel 2010.

VISUALIZATION APPLICATIONS AND RESEARCH QUESTIONS

Though the versatility of visualizations is substantial, certain research questions are better matched to visualization techniques than others due to characteristics of the available data—not only volume, variety, and velocity considerations, which extend across all forms of big data, but also other properties as well. Paired with the visualization types described above and their associated examples, I have attempted to overview various representative examples. More generally, longitudinal and time-series data are often very well-suited to visualized display. Many visualization types, such as those described above, are specifically designed to illustrate time-based trends and patterns. Data with a complex hierarchical structure, too, can draw on an entire class of visualizations to represent this structure in static and interactive formats. Geographical, map-based data are common source material for visualizations produced by the media, and as such, familiarity with associated visualization types has been instilled in many potential users. More generally, multivariate datasets in which categories can be distinguished on many dimensions, especially when each dimension uses a different scale, have visualization types specifically designed for exploring these data efficiently and insightfully. Unstructured text data are also associated with visualization types to guide initial review and interpretation of this information.

In the context of data commonly available to I-O psychologists, research and content domains that may be fruitful for further application of visualization techniques include cultural differences (which can draw on visual techniques for linking category members as well as displaying geographic data), teams (where intact groups can be readily compared on key properties), customer service (due to the vast amounts of data being tracked), standardized testing (for educational and large-scale employment purposes), technology-captured data (e.g., social networking; wearable sensors), diversity-related topics (which involve comparison of different groups, and often produce compelling interpretations that can be represented

accordingly through visualizations), historical/over-time datasets for score trends, and more generally, interventions targeting large segments of the employee population, such as hourly and entry-level employees.

KEY CONSIDERATIONS WHEN CREATING BIG DATA VISUALIZATIONS

Objectives

Any visualization design process must begin not with the data, but rather with a careful inventory of the objectives of the resulting output. This step is a critical precursor to choosing among the various visualization types and techniques, and to mitigate risk of a visualization with aesthetic properties that far exceed its desired impact and influence on the intended audience. Common objectives of data visualizations include decision initiation or modification (i.e., what will the user do for the first time, or do differently than they do currently, as a result of the visualization), enhancing understanding (i.e., what will the user know after viewing and interpreting the visualization that they did not previously), and communication expansion (i.e., will the visualization allow you to reach a new audience with an existing message).

Source Data

Stephen Few (2009) proposes several "traits of meaningful data," many of which intersect with big data applications and are notable when considering how to identify, structure, and clean data prior to initiating the visualization design process. First, meaningful data well suited for visualization is high volume—visualizations will be less advantageous over numerical or tabular forms of information presentation when the volume of data is relatively low. Second, historical—as noted above, many advanced visualization methods are specifically designed to illustrate trends in over-time data. A third set of traits is data consistency, clarity, and cleanliness—data veracity of these forms is a common challenge for large-scale datasets, and so a key focus for data qualification and cleaning efforts occurring before visualization must be generated. Fourth, multivariate—similar to volume, visualizations become exponentially more useful as an exploration technique when more variables are present. Fifth, richly segmented—if visualizations can draw on data that have been presegmented into meaningful

groups, this pre-established logic for categorization helps to ensure that resulting interpretations of the data have an inherent and interpretable meaning. By selecting data that innately fulfill most or all of these criteria, or that can be made so prior to beginning the visualization creation process, resulting outputs will be more likely to produce interpretable, prescriptive guidance for the user.

Information Transfer to the Audience

When designing visualizations, it is critical to consider the process by which the information gets transferred to the viewer. A data visualization applies an interpretive framework to the data toward the goal of communicating new information to the user. In this model, the designer of the visualization encodes information to in turn drive an intended form of decoding and to produce insight and understanding for the user's benefit (Cairo, 2013). Ensuring that this transfer occurs requires careful consideration of the audience for the visualization—the researcher may make very different design choices for an audience comprising senior executives compared to one of fellow researchers or students, as these audiences have different levels of experience with visualized information formats and different orientations to the depth of attention and time that they are willing to give to the interpretive process.

Though visual forms of information have become increasingly common, variation in sophistication levels for processing visualizations remain. Shah and Hoeffner (2002) discuss mechanisms for teaching students this form of "graphical literacy" and these recommendations—including using multiple representations of data when possible, focusing on the meaning associated with different visual features, and guiding users to consider visualization a critical evaluation and interpretation opportunity—extend to non-student audiences as well. As I-O psychologists have opportunities to foster these skills in our roles interacting with students and professionals, our orientation toward these factors, particularly the "meta-cognition" (Shah & Hoeffner, 2002) associated with the act of visualization reading itself, will produce benefits for future as well as current instances of visualized information transfer.

Design

Design plays an essential role in the success of data visualizations in achieving their goals, with practical benefits that far outweigh the increased demands placed on the visualization's creator (Vande Moere & Purchase,

2011). An initial set of design considerations relates to the use of visual features that combine the benefits of pre-attentive processing with accuracy of interpretation for the types of data represented within the visualization as summarized above. A second early step in the design process is to choose a particular class of visualization—for example, to show connections and relationships or to display trends in data over time—to match the research question and intended use. Within a visualization class, a designer may choose to select and test multiple options for specific types, taking into account the visual metaphor that best blends the message, the need to prioritize depiction of either patterns or details, and the aesthetic qualities of the output (Kirk, 2012). Once a narrow range of options has been selected, pretesting with individuals representative of the target audience can help in confirming a final option.

Kelleher and Wagener (2011) propose a set of useful guidelines for effectively visualizing data in scientific publications. These guidelines are applicable across levels of sophistication from the basic to the advanced, and a subset of recommendations from this discussion that relate particularly well to big data visualization applications is summarized here. A first consideration is simplicity—to create the simplest graph that still depicts the intended message (Tufte, 1983). While many of the examples displayed above are far from simple, they also are designed for use with extremely complex and multifaceted datasets, so they may still be the most efficient way to present information of big data's typical scope and scale. A second consideration highly relevant to such datasets is how to address density of data points in a way that can render them still visible. Options for dealing with this condition include making data points transparent or as unfilled circles or reducing the size of the points to allow more to be visible simultaneously. A final design consideration is color choice—to select an appropriate color scheme based on the qualities of the data (which can also include key distinctions such as significance or non-significance, with the latter indicated by a distinct coloring) and the message the visualization should convey. Color can either detract from the brain's ability to process a graphic or enable it, and the temptation to overuse color (an unfortunate default characteristic of many visualization tools) should be avoided—Cairo's (2013) guidance accordingly is that:

> The best way to disorient your readers is to fill your graphic with objects colored in pure accent tones. Pure colors are uncommon in nature, so limit them to highlight whatever is important in your graphics, and use subdued hues—grays, light blues, and greens—for everything else. (p. 105)

Two related sets of color considerations are for printing and color-blindness. Various websites provide self-evaluation tools to gauge and avoid potential issues in these areas.

GRAPHICAL OVERLAYS AND ANNOTATIONS

A special class of design considerations is graphical overlays and annotations—for certain forms of big data visualizations, these are very useful design features to illustrate trends not directly observable from the graphic itself. These features can serve a valuable function to guide the reader to a level of clarity and form of interpretation not otherwise assured. Although graphical overlays and annotations play a critical role in transferring the intended visual message to the user, they are not automatically generated by data visualization tools in most cases; rather, they must be added by the researcher and visualization creator. Although adding graphical overlays and annotations requires additional design steps, their potential value should not go unrecognized, as they have been shown to improve memorability of visually-presented information (Borgo et al., 2012) and reduce working memory demands placed on a user, a common risk area for large-scale visualization (Shah & Hoeffner, 2002). Borgo et al. (2012) also found that the use of graphical overlays can negatively impact visual search, so these advantages can come at a cost.

Kong and Agrawala (2012) define five types of graphical overlays applicable to data visualizations. While these are also relevant to more traditional visualizations such as bar, line, and pie charts, they are worth noting for big data visualizations due to their ability to reduce working memory demands (Shah & Hoeffner, 2002).

Reference Structures

Overlays can serve as reference structures to clarify the linkage between the underlying data and its visual representation. Reference structures can include gridlines at standard intervals, such as might be generated by default from a software package. The placement of these lines can also be placed in accordance with the desired interpretation, for example to denote the high and low value range for a particular group. An example reference structure is shown above in Figure 5.12 (the gridlines within the bubble plot).

Highlights

Highlights use color and shading to emphasize particular key components of the data or conversely, to de-emphasize others. This form of annotation can be useful to show the scores for one particular country's growth over time among all other countries displayed, for example. Highlights can also be added and removed sequentially to progress through an interpretive view of the data. Certain big data visualization methods, such as parallel coordinates, make frequent use of highlights to add prominence to selected elements within the visual field, without which trend and pattern detection would be difficult or impossible. Examples of highlighting are shown above in Figure 5.7 (highlighting certain specific job classifications in a distinct color) and in Figure 5.14 (thicker and dotted lines to emphasize particular countries' patterns).

Redundant Encodings

Redundant encodings can be simply data labels to indicate values within a visualization. These can be applied either across the full set of data or selectively to indicate key points of note. Redundant encodings can also include supplementary indications of trends within the data that do not appear within the standard visualization (e.g., an additional line connecting two specific data points). Redundant encoding is also shown above in Figures 5.6 and 5.7 (the labeling of certain specific job classifications).

Summary Statistics

Summary statistics can be very useful as annotations to display a common reference point within the visualization based on an average, maximum, or minimum value. This supplementary visual information can aid the user in putting individual data points in context. Slopegraphs make frequent use of summary statistics to display an average set of values. An example of summary statistics is shown above in Figure 5.1 (average values for 2013 and 2014).

Annotations

Annotations are textual notes or comments added directly to the visualization as a mechanism for communicating directly with the user about aspects of or information within the data that he or she may not otherwise detect or that warrant special emphasis. The use of annotations is more appropriate for data visualizations with a focus on explanation, and less

so for those where the user should be allowed to explore the dataset independently. However, when used responsibly, annotations can substantially reduce the risk of mis- or underinterpretation of the visualization. Many visualizations presented in news media use annotations—as well as other graphical overlays—to bridge the gap between the expertise level of the researcher and the user in order to allow the latter to benefit from the context and experience of the former. Examples of annotations are shown above in Figure 5.12 (labels applied to certain clusters of countries), Figure 5.13 (indicating US holidays), and Figure 5.16 (indicating the variables most and least correlated with leader quality).

KEY ISSUES AND RISKS

As the expectations for data visualization are growing, so too does the risk of unfulfilled hopes for their ultimate value. These risks are particularly salient for poorly designed and presented visualizations. In this section, we focus on several risks and disadvantages that are unique to or exacerbated by big data applications. For an expanded list, Bresciani and Eppler (2009) provide a thorough representation of key risks associated with data visualizations of all types.

Imprecision and Inaccuracy

A primary disadvantage of visualizations is that they display information at a lower level of precision and accuracy than numerical or tabular formats. Though the human eye can, based on a well-designed visualization, easily spot patterns and variations in the pre-attentive variables described above, we are much less proficient at detecting minute differences between individual data points unless guided to do so by a graphical annotation or other indicator. Certain relationships are also frequently overstated in visual as compared to numerical formats—for example, correlation magnitude is often over-estimated with high-density scatterplots (Cleveland, Diaconis, & McGill, 1982; Lauer & Poster, 1989).

"Optical Significance"

A critical limitation of data visualizations is their treatment of and, in some cases, inability to incorporate statistical and practical significance concepts. If a user of a specific visualization detects a pattern he or she

feels is meaningful and that warrants corresponding action, this reaction is often driven by an idiosyncratic interpretation of the finding as it is visually depicted. Beyond practical and statistical significance indices, visualizations are susceptible to a third form of significance, which we are terming "optical significance," such that the viewer will interpret a difference or pattern as meaningful based on his or her perception, often without corresponding quantitative evidence to support this interpretation. This issue is further clouded by design choices made when constructing the visualization. Most data visualization software packages provide no easy way to denote significance—however, the designer can add annotation or shading variation to a visualization to help reduce the risk of the audience either over-interpreting an effect that appears large but does not meet traditional thresholds for either practical or statistical significance, or under-interpreting an effect that is noteworthy despite its trivial appearance.

When significance concepts are incorporated alongside data visualizations, this occurs in one of three ways: First, the visualization designer may have only included patterns within the visualization that met a minimum threshold for statistical or practical significance. Second, as noted above, significance can be indicated by distinctive coloring and shading or graphical overlays such as annotations to denote findings that meet these criteria. Third, visualizations can serve a screening function to detect preliminary findings that can subsequently be subjected to further significance testing.

Visualization Oversaturation

Overuse and excessive prominence of data visualization is the downside of the increasing popularity of these approaches, impacting those who wish to make use of these methods in two ways. First, while the proliferation of data visualizations has led to a vastly increased number of high-quality examples, it has also led to a dramatic increase—possibly even a steeper one—in deficient and flawed visualizations. Blogs such as JunkCharts (http://junkcharts.typepad.com/) are dedicated to visualizations appearing in widely read sources yet lacking, at least in the view of the website's author, key elements of quality. The lack of high-quality, peer-reviewed research on many aspects of visualization also means that less evidence is available to provide solid guidance to creators.

Second, the state of overuse itself can engender skepticism and cynicism about such methods—as with big data as an overarching topic (see, for example, the diagnosis by LeHong, Fenn, and Leeb-du Toit for research

firm Gartner in 2014 of big data passing the peak of inflated expectations and sliding into the trough of disillusionment). This risk area can be mitigated by setting careful expectations for one's own as well as others' visualizations, and as a researcher, by solid design choices when creating these graphics as outlined above.

FUTURE DIRECTIONS FOR I-O PSYCHOLOGISTS AND DATA VISUALIZATION

Advanced forms of data visualization are proliferating within the commercial software market and in popular media, yet remain understudied not just as applied to big data and I-O psychology, but more generally as well. As new research studies are conducted, I-O psychologists must extend beyond their natural research domains and information sources to remain current about new developments. We must also push for greater use of visualizations, for appropriate purposes, in our practice and publication efforts—to do otherwise risks missing an opportunity to convey our ideas and extend our influence to audiences who would otherwise fail to connect with our messaging. The use of visualizations within scientific publications is itself a challenging issue, as many forms of visualization are poorly-suited for traditional paper forms of information presentation such as professional journals—as more journals offer online access to their articles, and in interactive formats to draw on the full scope of visualization options outlined above, this will extend the possibilities of these methods to gain a stronger hold within scientific discourse. Researchers who can make their datasets accessible to others in raw format can also benefit from doing so if it allows others to explore the dataset to find otherwise unseen patterns, and conversely, those who are themselves proficient in data visualizations can leverage existing datasets to detect and display new findings not visible through traditional methods. Finally, data visualization can be seen as a form of narrative storytelling (Segel & Heer, 2010) to enhance our ability to convert the inherent but largely unleashed potential of big data into influence and impact. I-O psychologists are extremely well-positioned to be the conduit between data and insight for our constituents within the scientific, practice, and general public communities—our awareness and mastery of data visualization techniques and applications will be an increasingly critical enabler of our success in this role.

REFERENCES

Abela, A. (2008). *Advanced presentations by design: Creating communication that drives action*. San Francisco, CA: Pfeiffer.

Al-Kassab, J., Ouertani, Z. M., Schiuma, G., & Neely, A. (2014). Information visualization to support management decisions. *International Journal of Information Technology & Decision Making*, *13*, 407.

Arcia, A., Bales, M. E., Brown, W., Co., M. C., Gilmore, M., Lee, Y. J., . . . & Bakken, S. (2013). Method for the development of data visualizations for community members with varying levels of health literacy. *Proceedings from AMIA Annual Symposium 2013*, 51–60.

Becker, R. A., Cleveland, W. S., & Wilks, A. R. (1988). Dynamic graphics for data analysis. In W. S. Cleveland & M. E. McGill (Eds.), *Dynamic graphics for statistics* (pp. 1–50). Pacific Grove, CA: Brooks/Cole.

Bonneau, G. P., Ertl, T., & Nielson, G. M. (2006). *Scientific visualization: The visual extraction of knowledge from data*. New York, NY: Springer.

Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P. W., Reppa, I., Floridi, L., & Chen, M. (2012). An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2759–2768.

Bresciani, S., & Eppler, M. J. (2009). The risks of visualization: A classification of disadvantages associated with graphic representations of information. In P. J. Schulz, U. Hartung & S. Keller (Eds.), *Identität und Vielfalt der Kommunikationswissenschaft* (pp. 165–178). Konstanz, Germany: UVK Verlagsgesellschaft mbH.

Cairo, A. (2013). The functional art: An introduction to information graphics and visualization. Berkeley, CA: New Riders.

Chen, C. (2005). Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, *25*(4), 12–16.

Chen, C. (2010). Information visualization. *Computational Statistics*, *2*(4), 387–403.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*, 149–210.

Cleveland, W. S. (1993). *Visualizing data*. Murray Hill, NJ: AT&T Bell Laboratories.

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216, 1138–1141.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833.

Connolly, S., & Wooledge, S. (2012). *Harnessing the value of big data analytics*. Retrieved from https://site.teradata.com/Microsite/wc-0217-harnessing-value-bigdata/landing/.ashx

Davies, A., Fidler, D., & Gorbis, M. (2011). Future work skills 2020. University of Phoenix. Retrieved from http://asmarterplanet.com/studentsfor/files/2012/10/future_work_skills_2020_full_research_report_final_1.pdf

Eppler, M. J., & Bresciani, S. (2013). Visualization in management: From communication to collaboration. *Journal of Visual Languages & Computing*, *24*(2), 146–149.

Feinberg, J. (2009). Wordle. In J. Steele & J. Illinsky (Eds.), *Beautiful visualizations* (pp. 37–58). Sebastopol, CA: O'Reilly Media, Inc.

Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Oakland, CA: Analytics Press.

Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, *331*, 705–708.

Gorodov, E. Y., & Gubarev, V. V. (2013). Analytical review of data visualization methods in application to big data. *Journal of Electrical and Computer Engineering*, *2013*, 1–7.

Healey, C., & Enns, J. (2012). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 1170.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*, *53*(6), 59–67.

Huber, P. J. (1987). Experiences with three-dimensional scatterplots. *Journal of the American Statistical Association*, *82*, 448–453.

Iliinsky, N., & Steele, J. (2011). *Designing data visualizations*. Sebastopol, CA: O'Reilly.

Johnson, C. (2004). Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, *24*(4), 13–17.

Kelleher, C., & Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, *26*(6), 822–827.

Kirk, A. (2012). Data visualization: A successful design process. Birmingham, UK: Packt Publishers.

Knies, R. (2013, May 23). New ways to visualize your data. Retrieved from http://blogs.technet.com/b/inside_microsoft_research/archive/2013/05/23/new-ways-to-visualize-your-data.aspx

Koch, K., Mclean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., . . .& Sterling, P. (2006). How much the eye tells the brain. *Current Biology*, *16*(14), 1428–1434.

Kong, N., & Agrawala, M. (2012). Graphical overlays: Using layered elements to aid chart reading. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 2631–2638.

Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, *65*(11), 911–930.

Lauer, T. W., & Post, G. V. (1989). Density in scatterplots and the estimation of correlation. *Behaviour and Information Technology*, *8*, 135–244.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*, 21–32.

LeHong, H., Fenn. J., & Leeb-du Toit, R. (2014). *Hype cycle for emerging technologies, 2014*. Retrieved from https://www.gartner.com/doc/2809728

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, *5*(2), 110–141.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Marchak, F. M., & Marchak, L. C. (1991). Interactive versus passive dynamics and the exploratory analysis of multivariate data. *Behavioral Research Methods, Instruments, & Computers*, *23*, 296–300.

Pandey, A. V., Manivannan, A., Nov, O., Satterthwaite, M. L., & Bertini, E. (2014). The persuasive power of data visualization. *New York University Public Law and Legal Theory Working Papers, Paper 474*.

Parsons, P., & Sedig, K. (2013). Common visualizations: Their cognitive utility. In W. Huang (Ed.), *Handbook of human centric visualization* (pp. 671–691). New York, NY: Springer.

Parsons, P. & Sedig, K. (2014). Adjustable properties of visual representations: Improving the quality of human-information interaction. *Journal of the American Society for Information Science & Technology*, *65*(3), 455–482.

Russom, P. (2011). *Big data analytics*. Retrieved from http://tdwi.org/research/2011/09/~/media/TDWI/TDWI/Research/BPR/2011/TDWI_BPReport_Q411_Big_Data_Analytics_Web/TDWI_BPReport_Q411_Big%20Data_ExecSummary.ashx

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, *14*, 101–119.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1139–1148.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*(1), 47–69.

Soo Yi, J., Kang, Y., Stasko, J. Y., & Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1224–1231.

Stuetzle, W. (1987). Plot windows. *Journal of the American Statistical Association*, *82*, 466–475.

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications*, *26,* 10–13.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Vande Moere, A., & Purchase, H. (2011). On the role of design in information visualization. *Information Visualization*, *10*(4), 356–371.

Viégas, F. B., & Wattenberg, M. (2007). Artistic data visualization: Beyond visual analytics. *Online Communities and Social Computing*, *4564*, 182–191.

Viégas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1137–1144.

Zhu, Y. (2007). Measuring effective data visualization. *Advances in Visual Computing*, *4842*, 652–661.