PIIM

Design and Visualization Best Practices for Big Data:
Enhancing Data Discovery through Improved Usability

Data Visualization for Big Data

Award No. FA8750-12-2-0325

February 11th, 2014

Chris Goranson, Xinle Huang, William Bevington, Jihoon Kang

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

THE NEW SCHOOL

PARSONS INSTITUTE
FOR INFORMATION MAPPING

68 5th Avenue
Room 200
New York, NY 10011

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

PIIM

TABLE OF CONTENTS

THE NEW SCHOOL

PARSONS INSTITUTE
FOR INFORMATION MAPPING

68 5th Avenue
Room 200
New York, NY 10011

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

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

INTRODUCTION

In late 2012, the Parsons Institute for Information Mapping began documenting data visualization strategies for big data. Defining visualization strategies specifically for big data is difficult, since not everyone agrees on what big data is. "Big data" can generally be thought of as describing "data sets so large and complex that they become awkward to work with using standard statistical software" (Snijders, Matzat and Reips 2012) What is considered big data to one user might not be to another, and this might be partially reflected by what kind of processing power and abilities each user has. Often the challenge for the data scientist is both one of dealing with extremely large datasets as well as attempting to combine datasets together in a way that leads to new insights. On one hand this is a clear technical challenge – the data scientist must have access to a robust system capable of storing, quickly retrieving and searching the datasets. On the other, the data scientist must hold or have access to enough subject matter expertise that they can interpret meaningful insights from the data through well-designed visualizations. Without access to the technical wherewithal, domain knowledge and a well-designed data visualization, a data scientist cannot fully leverage the value of extremely large datasets for analysis purposes.

As of 2012, three areas were identified as existing gaps in the current visualization research: explanation of visual recommendations, critiquing of designs, and rich user experiences for mixed-initiative design (S Langevin 2012). The hope is that this document provides some of the foundation necessary for communicating and quantifying the value of design in a big data environment.

From a design perspective, there are many challenges that are evident when working with extremely large, often unstructured and disparate datasets. First and foremost is an ever-present assumption that in order to visualize big data you need novel data visualization strategies. Based on the research conducted prior to the XDATA Summer Workshop in 2013 and observations from it, it can be ascertained that traditional data visualizations already familiar to an analyst (e.g. histograms, scatter-plots, maps) may still be the most effective way to display big data; however doing so requires a higher-level of interaction on behalf of the user. For example, a binned histogram may work well to show aggregate data at a certain scale, but giving the user the added ability to "zoom in" on the histogram, becoming more exposed to increasingly detailed attributes about the dataset maintains perspective on the data while providing an ever-richer experience through an interface the analyst already understands. This is not to say that there isn't a place for novel data visualizations either. Indeed there is – but one shouldn't discount the value of more traditional data visualizations if they can be intelligently engineered.

For the capable data scientist, one challenge is simply in understanding how best to visualize key findings and present them to decision-makers. While this challenge exists for visualizing any dataset, big data presents unique circumstances for design and data visualization to support better realization of information. For end-users, interface design challenges exist in adapting the tools and techniques developed by the data scientist in a manner that is complementary to the activities of an analyst. Because big data, much like a map, requires a significant amount of interaction, static products like reports, screenshots and pictures do little to convey knowledge since most knowledge is gained by

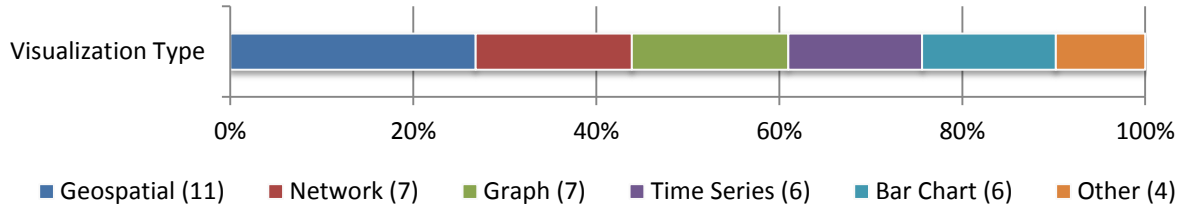
68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

interacting with the analyst. This means that the path from analyst to decision-maker will require tools and resources that support data exploration in a dynamic environment.

One theme that emerged throughout the first year XDATA was that certain data visualizations are commonly used over and over again to visualize big data sets. Furthermore, a single data visualization was commonly used to anchor the rest of the application – often times a map. This is telling, and begins to establish a sort of hierarchy of visualizations that work well with extremely large datasets. Another theme that emerged was the interactions that data scientists were using to explore the data. Finally, the stylistic renderings of the data visualizations are perhaps the most diverse as they are often the last and sometimes largely ignored elements. This document describes the most sought-after data visualizations and how they were implemented in existing solutions. The document also identifies interactions utilized by various tools, and the supporting styles that seem most beneficial to highlighting findings in datasets. This is important, as it is helpful to document and further establish the big data visualization language that is emerging through practice. As software tools continue to improve and analytic approaches develop, this language will evolve to encapsulate more data visualization approaches that complement large, often dynamic datasets.

PRE-SUMMER WORKSHOP FINDINGS

Results of a survey conducted by Draper Labs and The New School established that performers interpreted both common and less-common approaches would be used towards the visualization of big data. When asked to list up to ten UI/visualization features that might be integrated within XDATA tools, the responses tended to either address a specific type or combination of data visualization types, interface tools and workflows, or interaction methods with the user. By far the latter was the most innovative of all three. Of the 43 total responses that were considered visualization features, the results were tallied as follows:



This helped to establish that performers on the program envisioned that primary visualization features clustered around a couple key visualization types, and most of the rest tended to be well known. Many survey respondents consistently cited examples included map/geospatial components, network diagrams, bar charts/histograms, and common graphs and plots. What is interesting about the survey is the performers accurately predicted that developed components would have significant spatial components: falling specifically in geospatial and network visualizations. What they did not predict was the introduction of novel types of visualization methods. This held true through the majority of the summer workshop, with a few exceptions. Oculus introduced a Sankey diagram, Next Century

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

utilized a data clock, and Continuum created a 3D representation of a GIS/temporal data trace using sensor data. By far, however, the performers opted for well-known, recognized and established data visualizations.

Obviously, the performers were, after all, responsible for developing the resources as part of the summer workshop, so it should be no surprise that they accurately predicted at least some of the common visualization features. What is striking is how little new visualizations were introduced. Overwhelmingly the visualizations tended to be well known and well-established. Novel features were generally suggested as ways to interact with the traditional data visualization types, but the data visualization types themselves were quite common. This further suggests that it is the interaction that is important rather than the visualization itself. Suggested interaction features included visualizations that included such things as expression-based visualizations, auditory feedback, 3D stereoscopic displays, multi-touch, head tracking, and immersion technologies.

Finally, the teams focused on interaction elements of a developed GUI. Key utilities of a user interface could include ways of interacting with the data (brushed linking, interactive faceting), capturing and/or improving the workflow (scripting workflows, decision trees, sharing visualization methods, ease of access to data analytics) or features of the software (import/export files, save/reload current session, collaboration support).

THE MAP AS A MODEL

When writing this document, a couple of important visualization themes emerged. First, much of the present data visualization strategy for big data deals with the visualization of large datasets much in the same way that a web-based map engine deals with sharing contextual information back to the user. Views are "scaled" - meaning that on a map at a certain scale, certain features can be readily seen, while others are not. When viewing a map from a 1:1,000,000 scale (where one unit in any measurement represents 1,000,000 of the same unit of measurement on the map), a viewer is able to pick out major highways, geographic features like rivers and some coastline detail. At 1:500,000 major roads are now visible, some additional labels of surrounding areas and topographic features are more easily viewed. At 1:24,000 the viewer can see all roads, make out some building footprints and clearly define some patterns in land use. At such predefined scales, the caching of the imagery representing the various view extents allows the user to transition seamlessly between various view extents. The same approach is suitable for big data; while not all big data will be displayed on a geographic map, it is often applied within Cartesian coordinate space, which provides the possibility for exercising the same scalable approach. Therefore, it is possible for the end-user to "zoom in" to data visualizations representative of big data. In order for the same rules to apply as they do in a cartographic map product, this ability is contingent upon defining which dataset details are available at which extents. As the user "dives" into the data, they are exposed to more and more contextual detail. Continuum's Bokeh is one analytic toolset that leverages this interaction well.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

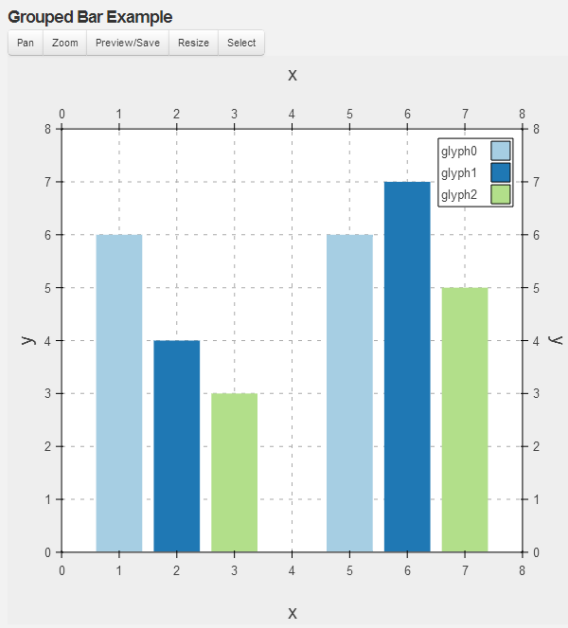


Figure: Bokeh's scalable grouped bar visualization (source: http://continuumio.github.io/bokehjs/group.html)

Further evidence of a map legacy as a driving force between all data visualization aspects of this program exists as many occur across three domains: geography, Cartesian space, and/or time. Data visualizations that function well within these domains are most often used in big data visualizations. All three provide an important context by which the data can be retrieved, parsed and queried against. Many successful visualizations make use of more than one of these domains at the same time, thereby allowing the user to scan space and time within the same interface. The Aperture JS framework acknowledges the similarities between GIS approaches to the design of an eventual interface, borrowing a number of concepts including layers, legends, mappings and map keys in explaining its visual assembly (Oculus 2013).

THE FIVE MAJOR DATA VISUALIZATION TYPES: A COMMON DATA VISUALIZATION LANGUAGE

A common language can be helpful as it defines the language by which performers, often coming from different disciplines, interact. Datasets, when abstracted into common types, can assist users in understanding both the merits and limitations of common datasets. This section of the paper outlines common data visualization types that were systematically used to analyze big data. This section also documents some of the advantages and disadvantages associated with the selected data visualizations.

As touched on earlier, just as maps provide a framework for geographical theory (Bunge 1966), a map for data visualization is the intersection of the combined representation and scalable visual assets. Practitioners in the big data space have, to some degree, self-selected frameworks that are successful at

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

visualizing big data. Big Data does however also introduce some key challenges for data visualization: namely scalability and rendering speed, and that can in turn affect the usability of the selected data visualization. Therefore, certain data visualizations will conditionally work well in some composite UI settings or against high-volume datasets, where others will not. A modifiable survey that can be useful for quickly assessing user input for a composite UI is included in the appendix. Surveys like these can be helpful for analyzing design recommendations before and after deployment, and assessing user satisfaction (Edsall and Adler 2012).

At a high level, most big data visualization components found in composite user interfaces tend to render a map (geographic data); networks (node-and-link); tables; charts; graphs or trees. Common relationships linking the various visualizations tend to be relational, geographic or temporal. The categories represented by the teams therefore can be formally defined as falling into one of these categories **(1) Maps, (2) Graphs, (3) Charts, (4) Diagrams, and (5) Tables.** The primary differentiator between data visualizations being suitable for big data lies in the interactions that are possible. Scalable (e.g. zoomable) interactions appear to be the most common method for interacting with the visualization.

A note on "graphs" and "charts": often used interchangeably, graphs and charts often represent the same data but by including slightly different features. For the purpose of this paper, a "graph" is a data visualization that includes a quantitative data measurement on at least one scale (vertical, horizontal) and often two or more. A chart, on the other hand, is generally an abstraction, including similar representations of the data as would be found in a basic graph but without a scale, grid lines, or tick marks (Harris 1999) – thereby making it an abstraction of sorts of the original graph. Maps capture geographic data, and diagrams illustrate relationships between data. Tables are the most basic representation of the data, but are important because they provide direct access to the information without the interpretation of the visualization.

Each major type in turn has many minor types (Charts includes area charts and bar charts, as an example). The infographics in the appendix identify some commonly used data visualizations against different types of data, and identifies some common visualization types used in the XDATA Summer Workshop.

MAPS

Maps are commonly found in big data visualizations because (a) they provide immediate context to the data; and (b) maps are typically scalable, allowing the user to zoom in, out or pan the map to specific areas of interest. Contextual information can be very useful as the viewer is immediately able to gain access to other data held in the map layers themselves that may contribute to more understanding of a particular dataset. Geographic Information Systems are built specifically to address datasets in context to one another and allow the user to query subsets of one based on the others. It therefore is no surprise that GIS methods and data often are present in user interfaces.

The type of background map also varied greatly amongst teams. While the "satellite" view is always a popular pick, it introduces some key challenges to the design of the application. First, the colors used in the satellite view can easily cause confusion with users since they may appear similar to colors selected to represent data. Second, they may require more time to cache, and may not provide uniform

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

value. Therefore, satellite imagery can be provided as an option but should only be used if it materially contributes to the user better understanding the data. A better choice is a simple gray topographic or area map, where the image supports the user understanding where the data resides but otherwise does not complicate understanding the data visualization.

Map projections, while still usually falling under the Web Mercator default used by Google and others, are increasingly seeing use in D3.js, Three.js and elsewhere. Map projections also provide an important tool for understanding the data, and should be selected to best represent the data. If representing maritime traffic for example, consider a map projection suitable for use in a maritime environment. Map projections can also change based on the view extent. One may be used for a global view, another for a continental view, and yet another for a state/local view. This helps support the notion that if the maps are used elsewhere or in conjunction with other data and/or applications, they will be more likely to support the user.

For big data, the temptation is to represent larger datasets simply with more classes. A better solution is to keep the number of classes lower but provide added detail when the user manipulates the view. An example would be a statewide choropleth representing political party affiliations that change to a countywide distribution once the user zooms in.

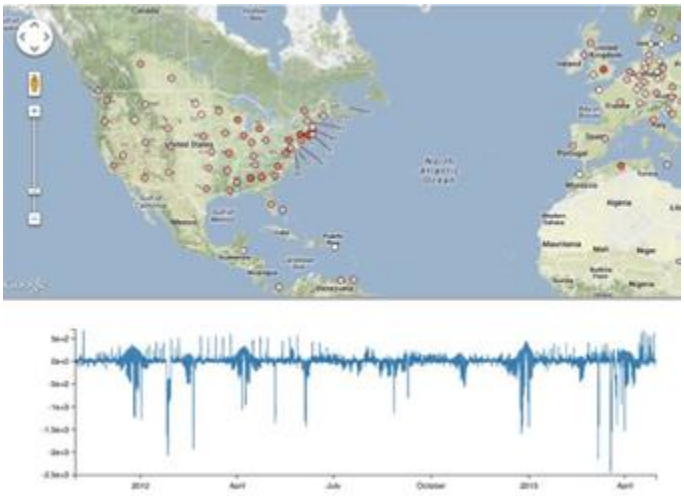


Image: Kitware's visualization of Internet browsing habits using a thematic point layer superimposed on Google's base map. Using a color thematic point symbology as opposed to different point sizes effectively maintains the denser areas (East Coast) while still indicating high-count areas. A histogram provides additional contextual information.

Common Types for Big Data: Cartogram, Choropleth, Heat Flow, Topographic

Example Map Type: Choropleth (Thematic Map)

Description: A choropleth, or thematic map is designed to illustrate a theme over a geographic area. Choropleth maps are popular because they can be quickly interpreted and provide immediate contextual information to a dataset if the user knows the geographic area well.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

Advantages: Thematic maps are most effective when representing a particular "theme" across an area in an easy-to-interpret manner. They can indicate spatial patterns or change across different time periods. Two variables can effectively be compared against one another in a thematic map.

Disadvantages: More than two thematic variables on a given map can be problematic. Known associations between two variables are best – otherwise a choropleth map can be abused to suggest causation in the data that does not exist.

Design & Style: XDATA performers often used graduated symbols to represent data on a map. One challenge was that while such a representation is an effective means by which to represent a proportional count of something across an area, they become less effective if there are pockets of extremely dense points. For obvious reasons, clustered graduated symbols will obstruct other symbols, and drawing symbol levels are ineffective since the stacking that occurs still often obstructs too many other points. In other cases both a graduated color ramp and a graduated point size were used, which can sometimes be redundant. A better solution is to keep point sizes the same when zoomed out at national or global scales, and use a color ramp to illustrate intensity of the variable. For dynamic visualizations the graduated symbols could be used once the user zooms in on the visualization.

The use of monochromatic colors can be more effective when representing variables. If representing percentages or rates, the colors should be intuitive to the viewer (higher percentages = darker colors) unless expressly explained elsewhere. Developers should avoid the temptation to represent more than two variables together on the same map, if using choropleth maps. A wide range of users can interpret the very best choropleth maps correctly without the need for a map legend. Where possible, uniform methods for class breaks should be used (e.g. quartiles, equal intervals, etc.). Special care should be taken to avoid line thickness or colors that compete with the color range selected.

When selecting colors, a good rule-of-thumb can be to follow one of the very well designed color ramps found in Cynthia Brewer's ColorBrewer (Brewer and Harrower 2009).

GRAPHS

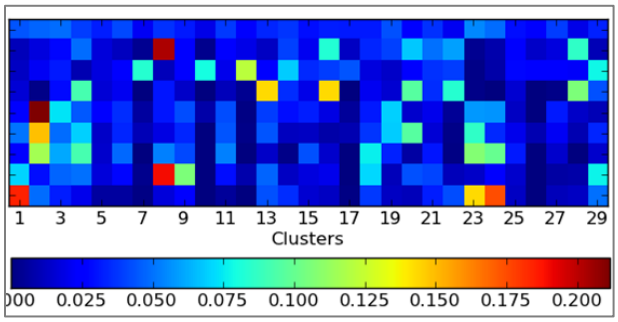


Image: A row-normalized confusion matrix graph by Carnegie Mellon University / Phronesis on a traceroute dataset.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

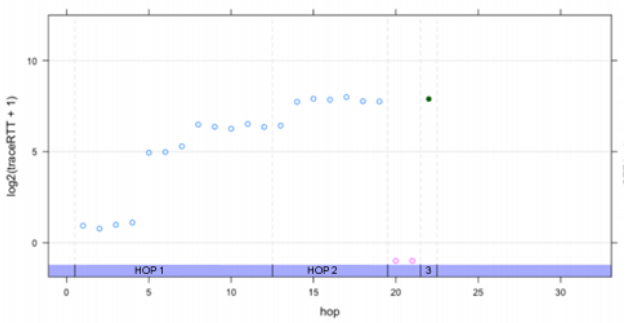


Image: A graph plot showing an edgescape of traceroute data by Pacific Northwest National Laboratory / Purdue University / Stanford University.

Traditionally more difficult to interpret than charts, graphs are usually the byproduct of direct data visualization output of large datasets. They generally require interpretation unless the user has taken the time to spell out how best to interpret the information. They can be information rich but sometimes overly so – data can first be plotted to a graph and later summarized well in a chart, for example.

Common Types for Big Data: Line, Area, Stack, Surface, Candlestick, Scatter Plot, Box Plot, Parallel Coordinate Plot, Stem Plot

Example Graph Type: Scatter Plot

Description: A scatterplot, often accompanied with a trend line, shows the distribution of values across the range of data. Certain characteristics of the data, like the concentration of values on either end of a distribution or all towards one quadrant, can be very informative depending on the source of the data. Scatterplot data can also be viewed and manipulated in 3D. A variation is the Scatter Plot Matrix and the Biplot.

Advantages: Scatter Plots are universally an excellent way to assess the distribution of a dataset. A Scatter Plot Matrix can provide a great deal of insight.

Disadvantages: Too many observations may make scatterplots difficult to interpret. Overlapping points are often hard to distinguish.

Design & Style: Use contrasting colors to differentiate the pairs of data used in the Plot. Continuum/Indiana's approach at identifying areas of over-plotting through abstract rendering is an example of an innovative exploitation of a traditional data visualization method. Allowing zoomable views of the scatterplot with dynamic axis is another effective way of improving the exploration of the data, especially when linked to substantive attribute information.

CHARTS

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

Charts are commonly used for representing aggregate data or summary information about big data sets. They have the ability to communicate information easily and effectively with an audience, but also tend to provide less information. In essence, they tend to sacrifice content for form.

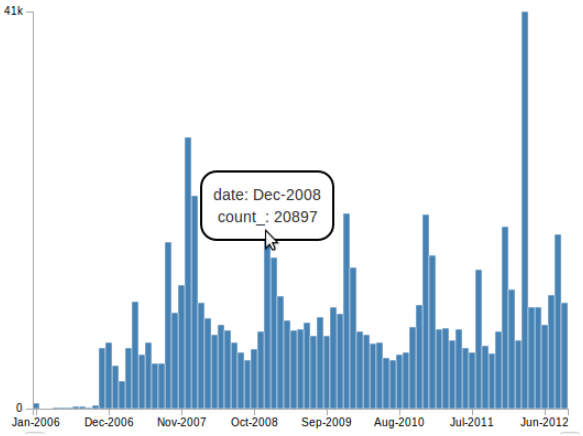


Image: In Next Century's histogram of lenders and number of loans, the introduction of a bit of space between each bar makes the entire image much more legible. Common charts were common data visualization types in big data, and effective at communicating information easily.

Example Chart Type: Bar Chart (Histogram)

Common Types for Big Data: Bar, Bubble, Sunburst, Kagi, Matrix, Pie

Description: A bar chart provides a way to compare a series of values across a range. The bars can be grouped by category, and can also be stacked to varying degrees if they represent a compilation of values. When a bar chart is plotted horizontally it is generally referred to as a column chart, or vertical bar chart. A variation is the multiple (or multiset) bar chart, or isometric bar chart.

Histograms are extremely valuable for showing the overall distribution of data in a dataset. The shape of the distribution itself in the right context can inform the viewer of certain characteristics (for example, to determine if the data normally distributed). Variations include the Ordinary Histogram the Cumulative Histogram, and the Stem and Leaf Plot.

Advantages: Discrete data is best represented in a bar chart. Bar charts are an effective way to compare fixed values over time (e.g. U.S. auto sales by year). Stacked bar charts allow the viewer to discern categories within the bar charts and compare growth within subcategories against the total. These trends are helpful in uncovering what proportion of a total bar is due to growing or shrinking subcategories. Histograms are good at showing the overall distribution of the dataset and identification of potential outliers. Histograms can also be used as a data exploratory tool in a composite user interface.

Disadvantages: As with many data visualizations, an over-saturation of values can make bar charts unintelligible. Scalable or collapsible bar charts are one way to address this constraint. Labeling the

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

axis can be problematic given a number of values represented. Stacked bar charts are best when representing a small range of categories and are easily discernible. Comparing categories between stacks can be problematic, and may best be augmented with separate bar charts where appropriate, or a label with value.

Design & Style: If data represented in a bar chart is fluid in nature, a line chart is a natural complement or alternative. A bar chart and line chart can be combined for an easy comparison across two variables. An ordered bar chart can be helpful for showing variation and comparing high and low values in the range. Stacked bar charts should avoid using similar colors unless the categories are explicitly related. Histograms are effective at multiple scales, but best when there are enough observations to discern the distribution of the data. If the user needs to see the actual values contained in the bins, consider using the Stem and Leaf Plot. Consider keeping a histogram one color. If using linked-brushing to highlight a subset of the data, simply grey out the other areas of the histogram. Rescaling should shrink the width of the columns without misrepresenting the data distribution.

DIAGRAMS

Network visualizations, or node-and-link visualizations are often popular in big data applications because they allow the user to visualize the entire universe of connecting nodes (or points) and links (or edges) existing in a dataset, thereby providing an overall picture of the data. However, there are many variations on the visualizations themselves, and some are more useful than others. Interactions are very important when working with node-and-link visualizations because connections between nodes can be difficult to follow if assistance isn't given towards helping the user trace a particular path. The organization of the nodes may or may not be important; in some cases hierarchies are established either by the size of the node or the arrangement of the nodes on the screen. In other cases the node-and-links are randomly distributed.

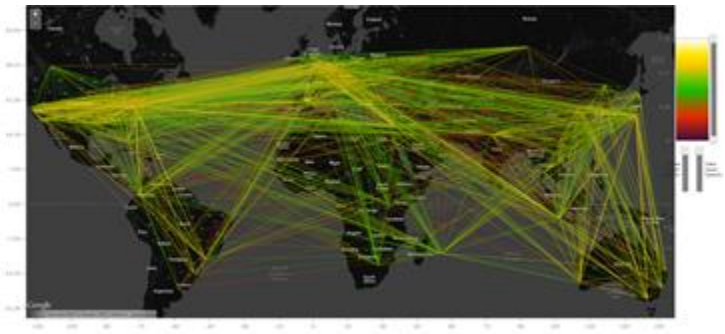


Image: Oculus utilizes a node-and-link diagram to illustrate traceroute data. By applying Colin Ware's color scheme, high traffic patterns are easier to identify. A transparency slider allows the underlying map to provide additional context for the data.

Common Types for Big Data: Node-and-Link, Arc, Tree, Sankey

Example Diagram Type: Node-and-Link Diagram

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

Who's using them: Kitware, Oculus, Continuum, Next Century

Description: Node-and-Link Diagrams are often popular representations of big data because they can appear to represent the entire universe of data in one visualization. Node-and-Link visualizations represent the relationship between two or more entities, and subsequently represent the flow of information through a network of interconnected nodes.

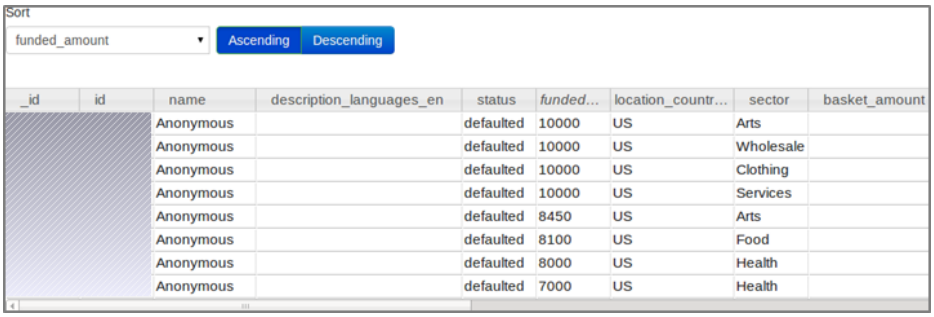
Advantages: Node-and-Link visualizations can be a convenient way to show the interconnectedness of the entire data visualization. They also exist in unconstrained space; therefore they can be easily manipulated without necessarily sacrificing knowledge about the data. Because they can be abstract in nature, they may tend to represent data that is less subject to the user's own bias about particular places or known organizations of information.

Disadvantages: Node-and-Link diagrams may not necessarily contribute meaningful information to the user if used incorrectly. They may be fun to look at, but by their nature can be unwieldy and overwhelming.

Design & Style: There are many variations of node-and-link diagrams, but they often times are selected because they are interactive and allow exploration of the data. Many times the visualizations allow for zooming, selecting and moving nodes. Sometimes they also allow for the rotation of the visualization in three dimensions. Nodes are often times colored based on an attribute, and the size may represent the relative intensity of the data they represent. In this regard they are similar to graduated symbols used in maps. Links may also be used to represent the intensity of the connection between two nodes – for example, if proportionally high rates of traffic occur between two nodes, the link may appear thicker. Occasionally, links are colored to represent different traffic.

While manipulating the location of nodes or group of nodes can be meaningful to the user, so can restricting movement of nodes. Hierarchical node-and-link visualizations can be an effective way to show the organization of data, and may be more meaningful. Colors should be used wisely; an easy way to highlight a particular path is simply to show a highlighted path in red with all others in grayscale. Too many colors can cause confusion, and a user that has to keep referring to the legend to understand what he's looking at will eventually tire of the visualization altogether.

TABLES



_id	id	name	description_languages_en	status	funded...	location_countr...	sector	basket_amount
		Anonymous		defaulted	10000	US	Arts	
		Anonymous		defaulted	10000	US	Wholesale	
		Anonymous		defaulted	10000	US	Clothing	
		Anonymous		defaulted	10000	US	Services	
		Anonymous		defaulted	8450	US	Arts	
		Anonymous		defaulted	8100	US	Food	
		Anonymous		defaulted	8000	US	Health	
		Anonymous		defaulted	7000	US	Health	

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

Image: Next Century's attribute table depicting lenders. In this example, each column has a meaningful header and the data is organized in a way that makes it easy to interpret and sort through results.

Example Table Type: Attribute

Who's using them: Kitware, Oculus, Continuum, Next Century

Description: Tables are used to show attribute data in rows and columns. Users commonly use tables to uncover attribute information about a selected item on another visualization layer as it provides context to the data that otherwise can't generally all be visualized in the graphic representation. Data from attribute tables for a single record are often represented in attribute pop-up windows, other times summary statistics on a selected portion of data may be represented in a table. In all cases, users are often very familiar with table mechanics and often expect a certain amount of interaction.

Advantages: Tabular data represents the raw attribute data behind a data visualization. They show the data in an uninterpreted manner and lead users to better conclusions about the information they are viewing. Tables are familiar items for most users, and help to build trust in a visualization if available.

Disadvantages: Tables should behave as the user expects. If data is not uniformly represented and the table design does not allow certain types of interaction, access to the table may be incredibly frustrating. For example, if a particular column representing an attribute's name is being truncated due to an insufficient column width and the user cannot expand the column width, the table may not be useful.

Design & Style: Access to the tabular data should be a feature of most data visualizations, because it assists the user in better understanding the data being represented, and how the data visualization is likely interpreting it. Columns should be aliased if possible prior to being presented to the user. This can be done by renaming the columns themselves or providing metadata on the column description to the user through a mouse-over or other interaction.

For more detailed information on recommendations for structuring elements within these and other data visualizations, please refer to the Design Guidelines document.

OFTEN-USED VISUALIZATION TYPES

Within each common visualization type, there are multiple components used throughout the program that render the relevant information within either (1) Maps, (2) Graphs, (3) Charts, (4) Diagrams, or (5) Tables. In some cases a component may be used to visualize data across multiple types (bar charts or bar graphs appear similar but can be used differently), or used to enhance another type of visualization (e.g. a bar chart superimposed on a map). The 2013 XDATA Summer Workshop was an excellent opportunity to gain insight into what eventually became the most commonly used components of the five major types.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

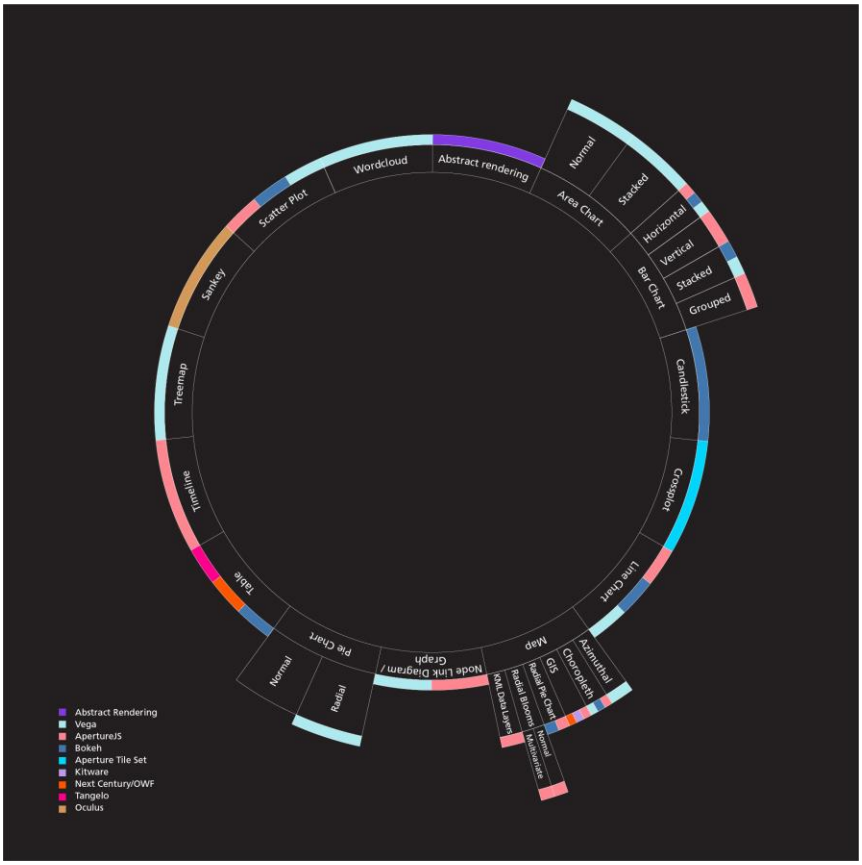


Image: An infographic depicting commonly used data visualization components in the XDATA Summer Workshop. Colors denote the provider or framework. Various kinds of maps, area and bar charts were used most often during the workshop. (Infographic: Xinle Huang)

A BEGINNER'S GUIDE FOR SURVIVING AN XDATA SUMMER WORKSHOP

A key challenge to being able to adequately address the challenges encountered during a DARPA Summer Workshop is to understand the complexities that face any performer. First and foremost, the format of a summer workshop is unlike any other a performer is likely to encounter in a government-sponsored program. Following a sort of hackathon meets summer camp model, the success of a performer will likely hinge on either (a) the ability of a performer to meet any and all requirements of a contract requirement alone; or (b) the ability to network, cooperatively build and test development environments and final projects with other performers who may be competitors outside of summer camp. Given the scope generally encountered through the various challenge problems, performers that can and still choose to fall into (a) above are in a distinct minority. Most performers will need to work together to meet challenges and greater transparency and open cooperation between the teams is essential to completing meaningful work during the workshop.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

A prohibitive factor mentioned elsewhere in this paper is the lack of subject matter experts or potential users of systems under developed prior to the workshop. This requires again that either (a) teams already have a very thorough understanding and idea of who the potential user-base might be; or (b) performers must work together to identify experts in each identified field prior to attempting to address any challenge problem. Again, most performers will fall into (b) above. Therefore it is essential that subject matter experts are identified and teams are formed early on in order to adequately solve the challenge problems

UNDERSTANDING THE SUBJECT MATTER AT-HAND

A key challenge to any developer of software solutions is understanding the contextual workflows and decisions that a user of a purported system will follow. Without a solid understanding of the user base, it is extremely difficult for someone to design a system that complements the user's existing workflow and answers the questions in a way they intend to ask them. Bloomberg's financial visualization tools are extremely successful because they are already experts working with financial services data. The FitBit device interface is successful because the visualizations are designed with a thorough understanding of the end user's goals (often to lose weight, eat better, and/or exercise more). From a design perspective, given that eventually those who are experts in their fields will apply tools, it is essential to understand how they would expect to view data in the first place. Choosing data visualizations that are complementary to the subjects will encourage adoption by the end users.

During the summer workshop, the challenges largely required subject-matter-expertize that would help establish relevance for the following types of data. Below is a list of the most common data types encountered through the challenge questions.

Social Media Data – Understanding the "pattern of life" and deviations from baseline activity were one major component of social media data. At a macro level this provided a method for tracking things like Twitter trends, or more sophisticated metrics like changing demographics of a neighborhood over time.

Network Traffic Data – In these types of challenges, it was important to understand how network traffic routing worked, changed and could be further optimized over time. Anomaly and/or change detection and predictive analysis were considered essential components of any solution.

Financial Data – Challenges that involved the ability to detect money laundering or other financial fraud were also core components of the program. As with the examples above, both the macro and micro view of the data through visualizations was considered important towards understanding the flow of information and funds.

Sensor Data – Remote and/or augmented sensing examples were indicative of attempts to understand the flow of a population where no previous information exists. In these challenges performers were often required to learn complicated data sources and collection techniques that are not typically encountered elsewhere. Through these data however, the end-goal was still to provide a way to understand deviations from normal patterns and identify types of activity based on movement or congregation.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

A near universal application of maps, timelines, and bar charts (or histograms) was found throughout the visualizations. Other common data types includes scatterplots, line graphs, and text/tabular data. It was evident that while some data visualizations tended to be more specific than others to the type of data being visualized, other data visualizations often worked well together and tended to be much more complementary. There were few cases for example, where someone would want to look at a distribution of points on a map, but not take time into account as part of the analysis. Therefore for many performers, the anchor components were built on visualizations that included:

(1) A map
(2) A timeline
(3) Summary attribute information (e.g. histogram)

USER EXPERIENCE LESSONS LEARNED FROM THE WORKSHOP

Good data visualizations should make the job of a data scientist or user easier. When we begin looking at what makes a good system "good" or a bad system "bad," users sometimes refer to systems as being "difficult to use," or "non-intuitive," or perhaps "the system is doing what I want it to." Part of this user experience is based on the user's interaction intentions (does the user want to touch or squeeze the data, or sort the data in a table), and part of this user experience is based on how they've learned to interact with other systems. However the best systems should be designed in order to lower the requirement that the user has to "learn" the system - the less work the user has to do before they are able to manipulate and understand the data, while gaining the most value and benefit from those interactions the better. Understanding the difference between overall appeal of a system vs. performance can be a delicate balancing act, as visually appealing systems do not necessarily equate to systems in which the analyst makes fewer errors - in other words, visualizations that simply look good may not necessarily be better ways to visualize the information (Purchase 2000). Lee (Lee, Butavicius and Reilly 2003) points out that still others suggest that even if interpreting a visualization technique takes longer - say a Chernoff face instead of a binary data representation, users may be willing to spend more time analyzing it (Everitt and Dunn 1991). When meaning is added to a graphical representation it tends to slow down decision-making, while maintaining shape characteristics of multivariate representations like star plot glyphs tends to lead to quicker classifications (Klippel, et al. 2009).

The visualization of data itself can ultimately lead to "insight through images," where computer-generated graphics merge with art and cognition to (hopefully) create an intuitive visual model based on a "collection of application dependent mappings" (Inselberg 2004).

The visualization strategies of various performers differed in their approach to visualizing many of the sample datasets identified for use during the program. Through the experiences gleaned from the providers and observed during various demonstrations, the following lessons emerged and may provide a foundation for measuring the impact of well-designed systems for big data:

1. **Does the visualization utilize a standard theme?** Visualizations that utilized recognized typography and color treatment rules fared better than those that did not. If an observer was forced to try and interpret what a particular color meant, or had a difficult time reading

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

descriptive text, the visualization was more likely to fail. Noted examples where this rule was broken: colors that changed meaning from one visualization to the next; illegible fonts; no common color scheme used or competing color schemes used.

2. **Is the anchor data visualization prominently featured in the composite GUI?** Screen real-estate tends to convey value, so if a particular data visualization (often a map) was the central focus of the GUI, it should therefore inhabit an area consistent with its overall importance. Examples where this rule was broken: assigning all visualizations equal areas on the screen; anchored data visualization smaller than secondary data visualizations; too much screen real estate assigned to supportive data visualizations.

3. **Can a user interpret and/or use the visualization without explanation?** In many cases the applicability of the visualization to the dataset and challenge problem was obvious enough once explained. However, very few data visualizations are initially developed with this goal in mind. In the case of the Summer Workshop this was primarily a result of simply not having enough time. Examples where this rule was broken: cryptic option menus, non-aliased column headers; no legend or scale; no introduction or help menu; the visualization doesn't answer a question posed by the challenge data sets.

4. **Does the application behave the way it's supposed to?** In this case this "supposed to" would mean, "as most users would attempt to use the application." If users are consistently trying to zoom in on the map using the scroll mouse, the map better zoom in! If users are consistently changing values and expecting to see the screen refreshed but it isn't, the visualization fails. Examples where this rule was broken: Submit and/or query buttons that are difficult to see; mouse behaviors that are unfamiliar or unexpected; constant tutoring required in order to use the visualization.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

Image: An example of an evaluation scorecard for rapid design review by a Red Team. (Image: Xinle Huang, Jihoon Kang)

CONCLUSION

Among the lessons learned was just because its big data doesn't necessarily mean that the method in which one wishes to view the data will change, but the method in which the user wants to interact with the data likely will. Just like services similar to Google Maps changed the way users expect to interact with maps, big data is changing the way users expect to interact with maps, charts, graphs,

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

diagrams and tables. Data visualizations must be flexible and intuitive in design, not sacrifice content for form and rapidly respond and change to a user's wishes. Designers, therefore, must also be flexible in their approach, marrying the correct data visualization workflow to the identified task at-hand. Understanding the workflow of not only the user is extremely important, but also understanding the multiple scales available for any given data visualization. More and more users expect to see visualizations work in concert with one another, and the best systems take advantage of a dashboard-like design to promote information awareness through multiple channels all working together with one another. In some cases users are also expecting to be able to instruct the visualization on how they should behave, look and feel. Capable systems that provide this additional flexibility will provide subject matter experts with the added advantage to fine-tune interfaces based on what they need for any given assignment.

68 5th Avenue
Room 200
New York, NY 10011

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

BIBLIOGRAPHY

Brewer, Cynthia, and Mark Harrower. *Color Brewer 2.0.* 2009. http://www.colorbrewer2.org (accessed October 15, 2013).

Bunge, William. *Theoretical Geography. Second Edition. Lund Studies in Geography Series C: General and Mathematical Geography, No. 1.* . Lund: Gleerup, 1966.

Edsall, R.L., and K.G. Adler. "The 2012 EHR use satisfaction survey: Responses from 3,088 family physicians." *Family Practice Management*, 2012: 19(6): 23-30.

Everitt, B S, and G Dunn. *Applied Multivariate Data Analysis.* London: Edward Arnold, 1991.

Harris, Robert L. *Information Graphics: A Comprehensive Illustrated Reference.* New York: Oxford University Press, 1999.

Inselberg, Alfred. "Lecture Notes: Parallel Coordinates." 2004. http://astrostatistics.psu.edu/su06/inselberg061006.pdf (accessed October 15, 2013).

Klippel, A, F Hardisty, R Li, and C Weaver. "Color Enhanced Star Plot Glyphs - Can Salient Shape Characteristics be Overcome?" *Cartographica - prefinal draft*, 2009.

Lee, M D, M A Butavicius, and R E Reilly. "Visualizations of binary data: A comparative evaluation." *International Journal of Human-Computer Studies 59 (2003)*, 2003: 569-602.

Oculus, Inc. *Layer Based Visualization Assembly.* 2013. http://aperturejs.com/tour/ (accessed October 3, 2013).

Purchase, HC. "Effective information visualization: a study of graph drawing aesthetics and algorithms. ." *Interacting with Computers*, 2000: 13 (2), 147–162.

S Langevin, B Cort. *A Comparative Analysis of Existing Approaches for Visualization Recommenders.* AVA Technical Report No. 1, Toronto: Oculus, 2012.

Snijders, C, U Matzat, and U Reips. ""Big Data": Big Gaps of Knowledge in the Field of Internet Science." *International Journal of Science*, 2012: 7 (1), 1-5.

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

APPENDIX A – SURVEY QUESTIONS

1. I can answer my research questions easily and efficiently.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

2. I can find information I need easily and efficiently.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

3. It clearly displays the information I need without unnecessary information or clutter.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

4. It allows me to complete tasks efficiently without unnecessary steps.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

5. It helps me focus on answering a question rather than on the computer.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

6. It presents information that is helpful and appropriate.

   ❑ Strongly agree
   ❑ Agree
   ❑ Neither agree nor disagree
   ❑ Disagree
   ❑ Strongly disagree

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

7. It helps me do my job.

- ❑ Strongly agree
- ❑ Agree
- ❑ Neither agree nor disagree
- ❑ Disagree
- ❑ Strongly disagree

8. Learning to use it was easy.

- ❑ Strongly agree
- ❑ Agree
- ❑ Neither agree nor disagree
- ❑ Disagree
- ❑ Strongly disagree

9. There is excellent support – I can get answers easily to my questions.

- ❑ Strongly agree
- ❑ Agree
- ❑ Neither agree nor disagree
- ❑ Disagree
- ❑ Strongly disagree

10. I enjoy using it.

- ❑ Strongly agree
- ❑ Agree
- ❑ Neither agree nor disagree
- ❑ Disagree
- ❑ Strongly disagree

11. I am highly satisfied with it.

- ❑ Strongly agree
- ❑ Agree
- ❑ Neither agree nor disagree
- ❑ Disagree
- ❑ Strongly disagree

12. Please indicate your job category.

- ❑ Analyst
- ❑ Programmer
- ❑ Systems Administrator
- ❑ Manager
- ❑ Other

Please specify: _____

13. How many years have you worked with information fusion products?

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

- ❑ Less than 1 year
- ❑ 1 – 3 years
- ❑ 4 – 6 years
- ❑ More than 6 years

14. How would you rate yourself as a computer user in general?

- ❑ I'm an expert user
- ❑ I'm an average user
- ❑ I'm new to computers or don't feel very confident about my computer skills

15. How would you rate yourself as a computer programmer in general?

- ❑ I'm an expert programmer
- ❑ I'm an average programmer
- ❑ I'm a novice programmer
- ❑ I am not a programmer

16. How would you rate your knowledge of data visualizations?

- ❑ I have an expert understanding of data visualization.
- ❑ I have an average understanding of data visualization.
- ❑ I have a basic understanding of data visualization.
- ❑ I do not have any understanding of data visualization.

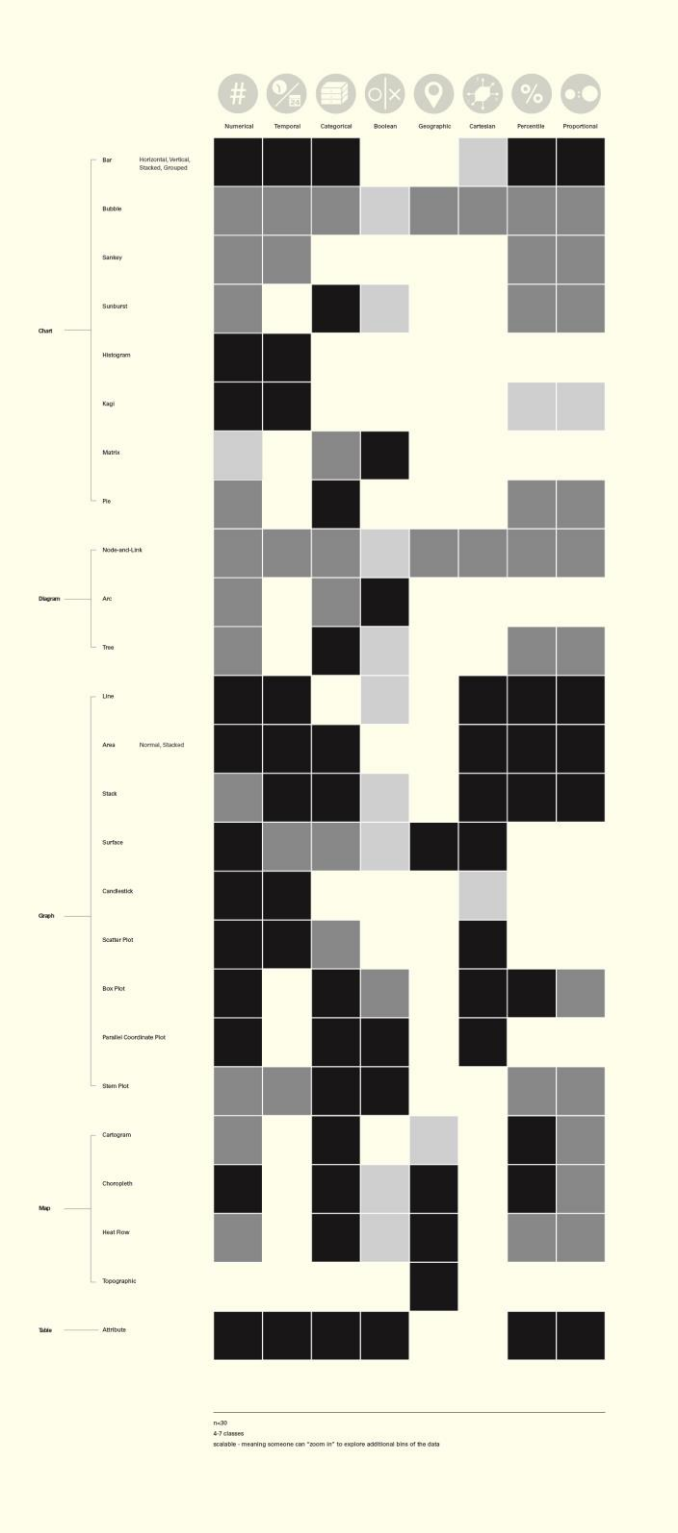
17. What is your employment status?

- ❑ Government Employee
- ❑ Contractor

68 5th Avenue
Room 200
New York, NY 10011

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

APPENDIX B – POPULAR DATA VISUALIZATION METHODS FOR BIG DATA



This infographic was developed through research conducted during before, during and after 2013 XDATA Summer Workshop, part of DARPA's XDATA program. Visualization types were identified, categorized and ranked based on their appropriateness for visualizing big data. The visualizations themselves are not exclusive in nature as many are commonly utilized together to tell a story about a dataset (a map with a histogram, for example). The graphic maps the applicability of a data-type to each visualization. A darker color represents a better fit with the associated data type (bar chart and numerical data, for example). This graphic should only be used as a guide help support design choices when working with big data – individual circumstances may dictate deviation from these recommendations. (Infographic: Xinle Huang)

68 5th Avenue
Room 200
New York, NY 10011
T: 212 229 6825
F: 212 414 4031
http://piim.newschool.edu

APPENDIX C – COMMON DATA VISUALIZATION PAIRINGS FOR BIG DATA

This infographic depicts examples of common pairings identified during the XDATA Summer Workshop. Maps or node-and-link diagrams were often used as anchor data visualizations supported by various secondary data visualizations. The anchor visualizations are depicted below with larger circles. Visualizations used by various performers can be identified by their color. (Infographic: Xinle Huang)