DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

A Grounded Theory Study on the Language of Data Visualization Principles and Guidelines

Eser Kandogan, IBM Research, San Jose, CA USA; Hanseung Lee, Google, Mountain View, CA USA

Abstract

Current automated visualization techniques use only a small set of basic guidelines, mostly based on perceptual principles. Yet, in literature a broad set factors are discussed for effective visualization design. In this paper, we analyzed principles and guidelines extracted from academic papers, books, and blogs and examined factors that influence visualization design to aid the design of automated visualization systems. We used grounded theory to code and examine concepts and relationships extracted. We found a variety factors, including domain, semantics, user tasks, insight, and display characteristics and found a variety of interactions between these factors including causal, contextual, and logical relationships. Our findings suggest that automatic approaches need to handle not only broader set of factors but also complex interactions among them. Furthermore, several factors such as domain and semantics are likely to gain even more importance, necessitating flexibility in terms the specification of design know-how as input to automatic visual analytics systems.

Introduction

In science and in business, there are increased demands for open analytics platforms to support collaborative exploration and analysis of data to accelerate discovery. A major challenge is to design tools that effectively support data visualization needs for a broad range of users, tasks, and devices. Automatic or semi-automatic visualization are promising but current techniques are limited in that they only a small set of basic guidelines, mostly based on perceptual principles, and match the data to known set of visualizations, rather than generating visualizations that takes into account diverse set of factors, typical on open analytics platforms.

To inform the design of automatic data visualization techniques for open analytics platforms we conducted a grounded theory study to examine the language and content of data visualization principles and guidelines. We extracted guidelines from books, papers, and blogs on data visualization and examined factors and relationships among them that influence effective visualization design. We found a variety factors, including data domain and attribute semantics, user perception, user tasks, insight type, and display characteristics, among many other factors, and found a variety of interactions between these factors including causal, contextual, and logical relationships.

In this paper, first we describe our study method and resources. Next, we present our findings, particularly patterns of form and content in data visualization guidelines. Then, we discuss our findings, particularly as they relate to the design of automatic visualization techniques. We also discuss gaps we found in literature, particularly on interaction and collaboration; data domain and semantics; and scalability and complexity.

Based on findings from our studies, we argue (1) for an *optimization-based approach* to automatic visualization, (2) driven by a *visual design language*, capable of expressing a large number of factors and complex relationships, where by (3) *optimization techniques are decoupled from design guidelines* via (4) *design guideline repositories,* for specific application domains, which can grow as the know-how on effective visualization design expands.

Related Work

Many visualization principles and models have been proposed to guide the generation of useful visualizations. One of the most frequently used principles is Shneiderman's Visual Information-Seeking Mantra, "Overview first, zoom and filter, then details-on-demand", describing a recipe for efficient design steps for users to transform the data to useful visualizations [1][2].

There are several other principles and guidelines, which focus on different factors, such as data, interaction, users, tasks, and domain, and have different strengths and weaknesses [3][4][5][6][7][8][9][10][11]. Carr, for example, analyzed different areas of information visualization and suggested seven general guidelines for designing information seeking applications [3]. Conati and Maclare found individual differences, such as user's cognitive abilities, to be an important factor in visualization effectiveness [4]. Munzner's nested model emphasizes the problem domain and suggests designers to characterize the task and data in the vocabulary of the problem domain and design visual encoding and interaction techniques, and algorithms accordingly.

However, very few studies connect these guidelines with the context of use, especially when applied to automatic visualization systems. Dias et al. analyzed visualization rules using grounded theory techniques over five different parameters: data type, task type, scalability, dimensionality, and positioning of attributes and characterized several visualizations based on these parameters [12]. While similar in method in spirit our present study goes much further, covering not only broader set of concepts but also deeper in regards to details of concepts and their interactions.

Study

Method

We employed grounded theory [13] as the primary method of analyzing our data. Grounded theory is commonly used in social sciences to develop a theory based solely on data. It is based on iterative multi-level coding, where concepts emerge from analysis of the data. In our case, our purpose was not to develop a theory, but the iterative coding and the data-driven concept development process of grounded theory was what we needed.

Resources

We used 5 books and 18 papers as resources for our analysis. Books chosen were specifically intended for effective visualization design and included Beautiful Visualization by Steele and Illinsky [14], Data Points by Yau [15], and Wall Street Journal Guide to Information Graphics by Wong [16], and [17][18]. Papers included work on automatic data visualization, visualization in knowledge discovery, visualization recommendations for science, and visual

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

analytics, such as [19][20][21][22]]. In addition, we also examined a visualization blog (School of Data, schoolofdata.org) and a documentation of a visualization product. Overall, we extracted about 550 guidelines, ranging in length from 5 to 140 words.

Analysis

In grounded theory there are three stages of coding: Open coding is the first level of coding, in which concepts, properties, and dimensions are identified at the desired level of granularity. In the second level, axial coding, the goal is to relate concepts, and identify context of these relationships. In the last level, selective coding, the goal is to integrate all these to form a larger theoretical scheme. Since our goal in this study was not to develop a theory we only performed open coding and axial coding of our data.

We used iterative coding in which we selected a small subset of the guidelines and two researchers coded it independently. We then gathered together and resolved issues in our coding scheme, and coded a different set of guidelines. Codes used in our iterations emerged from the data, as it should in grounded theory analysis. After a couple of iterations each researcher felt comfortable with the scheme and coded the rest of the guidelines independently.

Our coding scheme was hierarchical. We identified 5 1st level codes, relating to Data, Visualization, and User. Each level described more detailed sub-concepts. For example, Data/Attribute was a 2nd-level concept, Data/Attribute/Measurement was identified as a 3rd-level concept. We coded varying instances of a concept in parenthesis, for example, nominal measurement as Data/Attribute/Measurement (nominal).

In total, we derived 515 concepts at several levels, including all the varying cases. On average, each guideline was coded with 6 (hierarchical) concepts, and the longest guideline contained 17 concepts. In guidelines we examined 3% contained 1st-level, 56% 2nd-level, 30% 3rd-level, and 10% higher level concepts. See Table A and B for a detailed description of the coding scheme.

Limitations

While we did our best to reach a high-level of validity of our findings, we acknowledge several limitations. First, we could only analyze a limited set of resources, given the level of detailed coding that we needed to develop a deep understanding of the guideline content and structure with all their complexities. We tried to cover as many varieties of resources as possible from general public books, to highly academic papers, to visualization blogs so that we get a balanced representation. With more data, distribution of the concepts would change to some degree but even with limited resources we identified complexity beyond what is used in automated techniques. Our goal in this paper was focused on the structure of guidelines, i.e. examination of the various patterns and concepts and how they related to each other. Secondly, given the scale of data and depth of the analysis not all guidelines were double-coded but we did random sampling to verify consistency. After several iterations, during the development of coding, we reached a consensus on the coding scheme, each coder worked on a different set of guidelines.

Findings

Below, we present our findings from analyzing the 550 guidelines we identified from our data. First, we will describe the high-level concepts and relationships that exists among these concepts. Then, we will examine the key patterns we observed, particularly in conditional, ranking, and explanation type guidelines.

Concepts

During our iterative coding, five high-level concepts emerged from our data: Data, Visualization, User, Insight, and Device. We also identified a refining/contextualizing concept called Qualifier. Below we explain each of these concepts in detail:

Data (D): Data covers all aspects related to information, including its attributes, schema (e.g. hierarchical), operations (e.g. aggregation), domain (e.g. business) and issues related to domain such as trust, privacy, validity, etc. Most of these aspects constitute second level concepts, with further sub-concepts. For example, Data/Attribute concept was further refined regarding data size, semantics (e.g. time), role (e.g. measure), measurement (e.g. nominal, ordinal), distribution (e.g. sparse, normal), etc.

Visualization (V): Visualization captures all aspects related to representation and interaction of information, such as elements (e.g. bar), attributes (e.g. length), components (e.g. legend, axis), class (e.g. scatter plot), operations (e.g. group), and interaction (e.g. zoom). As in Data, each of these have further sub-concepts. For example, Visualization/Attribute further contains various visual attributes such as size, position, color, etc.

U*ser (U):* User entails all aspects related to human viewing and interacting with visual representation of data, including tasks (e.g. compare, communicate), (dis)abilities (e.g. read, recognize, recall), and social aspects (e.g. conventions).

Insight (I): Insight represents all intuition, understanding, and knowledge to be gained from data visualization. Insight is related to user task. Insight is the outcome, task is the process. Insight includes sub-concepts such as trends, outliers, relationships, variance, extrema, ranks, etc.

Device (M): Device covers the physical aspects of the visualization medium, particularly related to display and interaction capabilities.

Qualifier (Q): Qualifiers serve to refine and contextualize quality and quantity of concepts. For example, it may refer to a specific quantity, such as "two nominal attributes", or refer to a less well-defined quality such as "large data" or "dirty data". We identified over 35 such qualities in our data. Qualifiers may also refer to computed aspects of concepts such as "length of the labels".

Table A lists descriptions of the concepts identified. Figure 1 shows frequency of 1st level codes in our data, aggregating any sub-concepts. Figure 2 shows the top 20 ranked codes that occur most frequently in our data. Our analysis of frequency conforms to Zipf's law, as it should [23], where frequency is inversely proportional to rank.

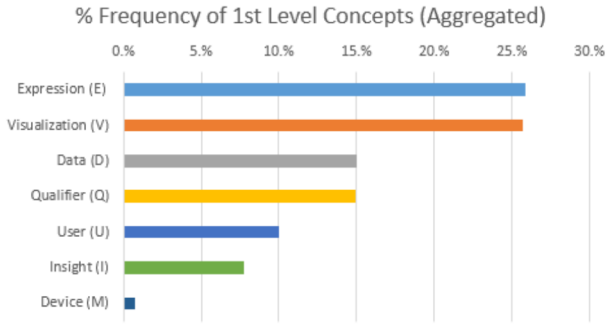

Figure 1. Frequency of 1st-level concepts, aggregated

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

Relationships

In order to express the relationships among (groups of) concepts, we created an Expression concept, which entails logical (e.g. and, or), (in)equality (e.g. more than), similarity, existential (e.g. contains at least 2, many), rank (e.g. top among), and prepositional (e.g. in, with) relationships. In addition, groups of concepts might also be related in the way guideline is formed structurally representing conditional (e.g. if then), copulative (e.g. and, furthermore), causal (e.g. due to), adversative (e.g. but), and supportive relationships. Detailed description of the Expression relationship is in Table B.

We also analyzed pair of codes that have a strong association through co-occurency (Figure 3). The strength of association is calculated using the G-test [24], which is equivalent to the chi-squared test for goodness of fit, but it is more accurate for small sample sizes. Figure 3 shows the top 20 pair of codes that have low p-values. Some associations identified were expected codes are clearly related such as: V/Component/Axis(h) and V/Component/Axis(v), but others point to interesting patterns such as D/Attribute and E/Conditional suggesting attributes where commonly used with conditionals.

Patterns

In our data we found three forms of guidelines: (1) Imperatives, (2) Declaratives, and (3) Conditionals.

Imperatives are basically guidelines that provide an essential direction to follow, typically the do's and don'ts in visualization design, for example, "Don't create shadows behind bars", "Always extend bar charts to zero baseline", or "Reveal data at several levels of detail". Overall we found that 11% of the guidelines were imperatives. Imperatives tend to be shorter in description and frequently have a negative form (e.g. "Do not...", "Never..."). Imperatives were about 52.5% shorter than declarative forms and about 41.0% were expressed in negative form.

Declaratives are statements regarding a design rule, principle, or opinion often with an explanation. They were by far the most common form of guidelines, constituting about 73% of the guidelines we examined.

Conditionals are guidelines that state a condition and a consequence, which holds true only if the condition is satisfied. We observed several patterns both in the condition and consequence parts. Conditionals made up about 16% of the guidelines.

Whether specified in conditional, imperative, or declarative forms, guidelines provides 1) instructions, describing what should be done and how, 2) rankings (or comparisons), stating preferences of one concept over others, and 3) explanations, providing a reasoning beyond the suggested action or statement. Below, we provide our analysis of the patterns identified in the form and content of the guidelines.

Instructions

Instructions make up the core part of the guideline, they express what actions to take and how. Instructions constitute the bulk of the content in imperative and declarative forms. In conditionals, instructions only exist in the consequence part (i.e. not in the condition part).

Instructions typically involve a data segment, which refers to a sequence of data attribute, schema, size, etc., optionally with data operations and qualifiers, and/or a visual segment, which refers to a sequence of visualization attribute, element, component, or type, optionally with one or more visual operations and qualifiers. Simple instructions include either a data or a visualization segment, e.g. "group data into bins", coded as D/Operation/Aggregate(bin), or "use a bar chart", coded as V/Class (bar), respectively.

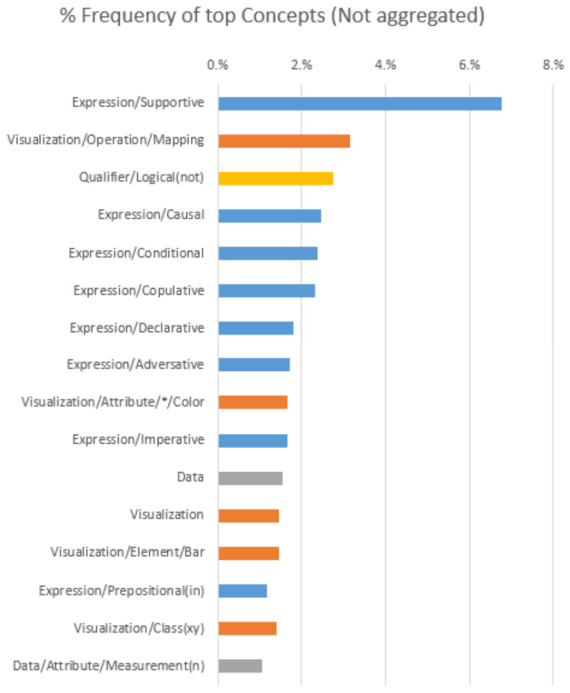

Figure 2. Frequency of top concepts, not aggregated.

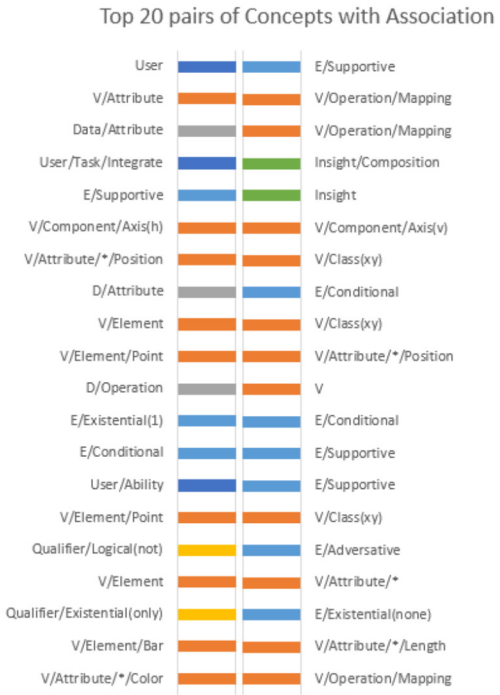

Figure 3. Top 20 pairs of co-occurring concepts with significance ($p < 0.002$)

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

More complex instructions involve several segments, often with a logical expression (e.g. “two measures and a lot of data values”, coded as Q/Existential(2), D/Attribute/ Role(measure), Expression/Logical (and), Q/Qualitative (large), D/Size). Other expressions such as (in)equalities (e.g. “median is a more robust measure than average”, coded as D/Operation/ Aggregation (median), Expression/ Inequality(more), Q/Qualitative (robust), D/Operation/ Aggregation (average)), similarities (e.g. “keep axis ranges similar”, coded as V/Component/Axis/Range, Expression/ Similar, V/Component/ Axis/Range), prepositional or possessive expressions such as “in”, “with”, and “of” (e.g. “labels in graphs”, coded as V/Element/Label, Expression/Preposition(in), V/Class (graph)), and other expressions that requiring a computation (e.g. “legend separated from the line”, coded as V/ Component/Legend, Expression/Spatial/Distance(far), V/Element/ Line).

In our data we found 70.4% of the guidelines to be in simple form. Even so, 42.8% of these simple forms included one or more visual or data operations. On the other hand, 22.3% of the guidelines contained one, and 7.3% two or more expressions. Logical expressions are used in 9.9% of the guidelines, while in(equalities) and similarities were included in 4.2% and 1.1% of the guidelines, respectively. Prepositional expressions were used in 9.0%. Other expressions (e.g. Spatial) were used rarely, i.e. 0.2%. Qualifiers existed in about 47.2% of the guidelines, of these 68.2% only one, 20.5% two, and 11.2% three or more.

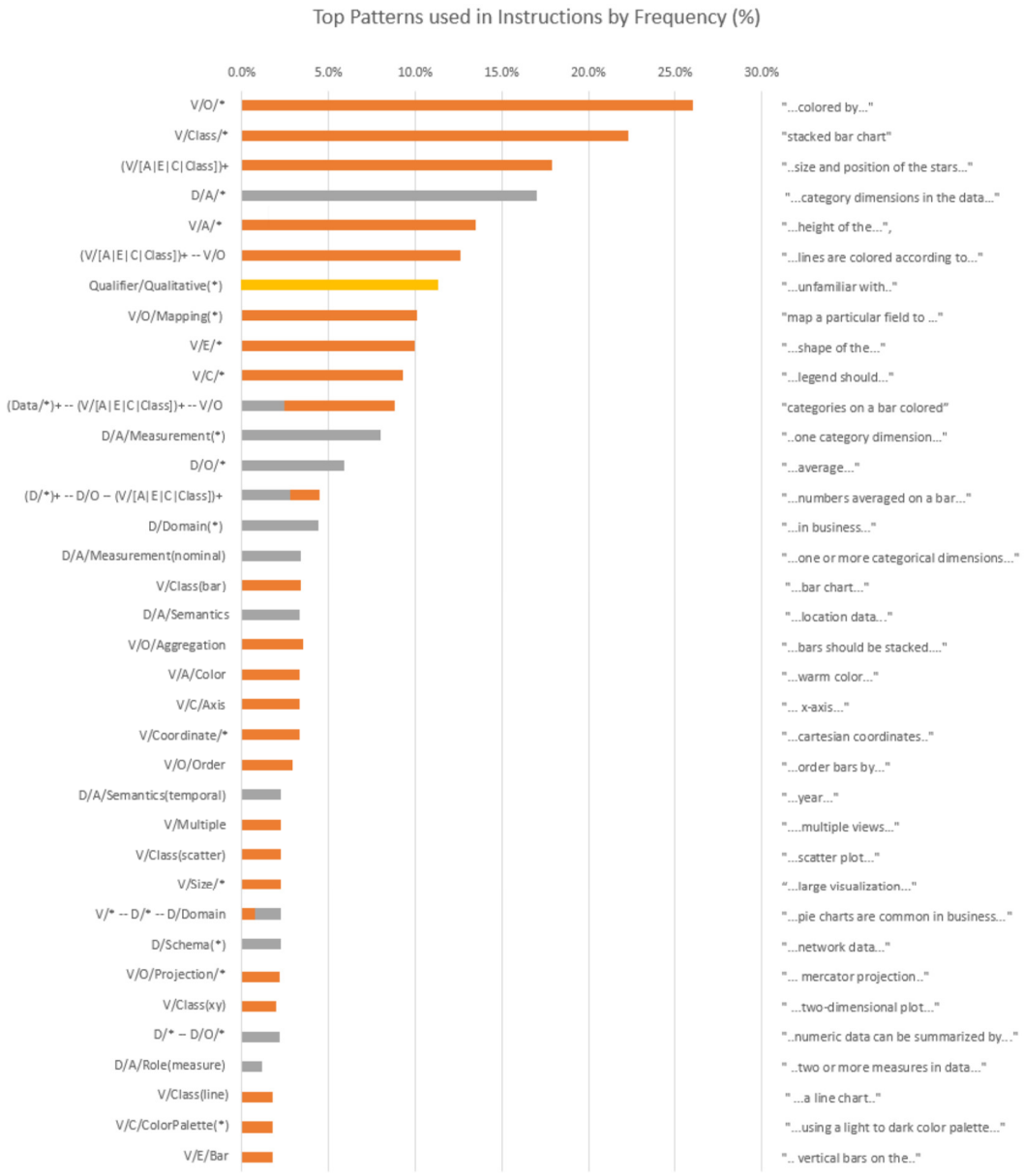

Figure 4. Top patterns used in segments of Instructions by percent frequency.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

In Figure 4, we list the common patterns in data and visual parts of the instructions. Note that the patterns here can be combined with others to make up the guideline through.

Conditions

In our data, we found that data and visualization concepts are heavily utilized in the condition part of the guidelines with 69% and 25% coverage respectively, while on the other hand user, insight, and device concepts are rather underutilized, with 9%, 6%, and 3% coverage respectively.

Overall patterns in the condition part are of the form: (1) Concept/*, which states that a certain concept is valid, for example, “hierarchical schema”, coded as D/Schema(hierarchical), (2) Concept/* (--Q/*), which states a specific concept, with a certain qualifier, for example, “large number of bars”, coded as V/Element/Bar--Q/Qualitative(many), (3) Concept/* -- Expression/* -- Concept/*, which states an expression or operation applied on concepts, for example, “data contains time”, coded as Data -- Expression/Existential -- D/Attribute/Semantics (temporal), and (4) Operand -- Expression/Logical(*) -- Operand, a logical expression composed of above primitive forms. We found that in our data only 12.6% is in the first primitive form, while 19.4% in the second form, and 37.9% and 27.6% in third and fourth form, indicating the complexity of the condition part of the guideline.

In Figure 5, we list frequently observed patterns (ordered from most frequent to least) in the condition part, along with examples and coverage, indicating % of conditions in which pattern occurred. At the top of the figure are mostly data concept related conditions, e.g. whether a data attribute of certain measurement, semantics, range, distribution quality is satisfied. On the other hand, interaction (Visualization/Interaction/*) and data domain (Data/Domain/*) concepts (not listed) came at the bottom, with 1.2% each.

Rankings

Rankings are statements that indicate a preference between two or more concepts in a guideline. Rankings can be relative in which two or more concepts are compared, and given a relative ordering. They can be absolute, suggesting one or more concepts are given a fixed rank independently.

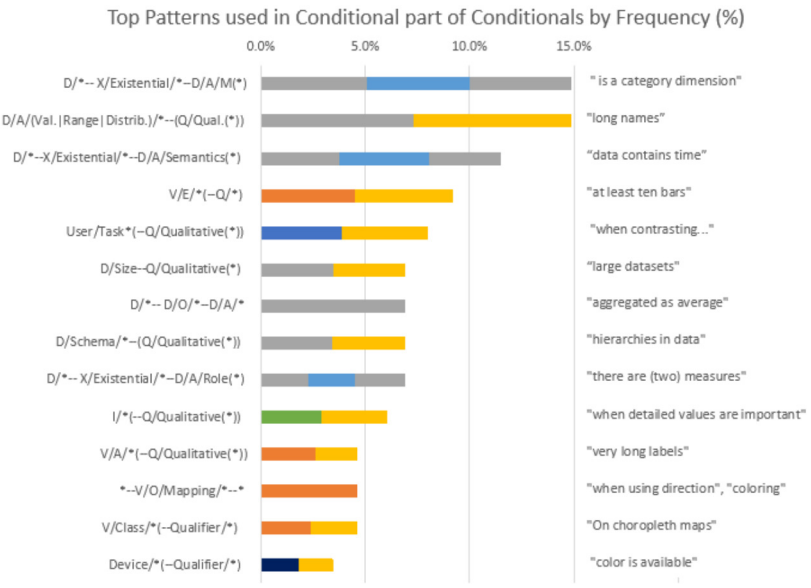

Figure 5. Patterns in the condition part of a conditional guideline

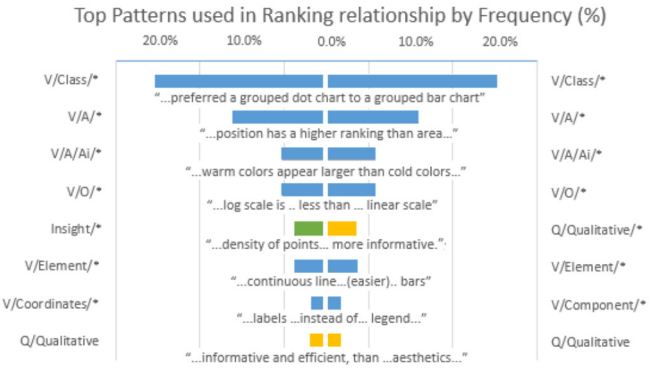

Figure 6. Patterns used in ranking relationship (Concept vs. Concept)

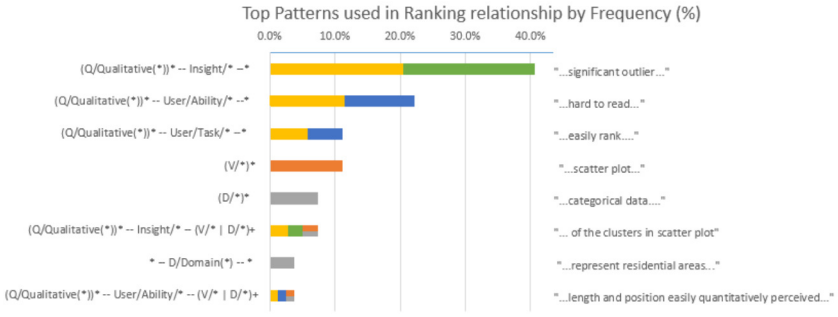

Figure 7. Patterns used in explanations

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

Guidelines can also talk about changes in ranking in response to a specific choice in the guideline. In our data about 5% of the guidelines contained a ranking. (Figure 6).

We also analyzed rankings to see whether a cause or supportive argument is made or not and examined the arguments. About 74% of the rankings contained a cause or supportive argument. Insight and user ability/task related concepts (> 80%) are identified as top concepts as part of the arguments. It is also interesting to note that 16.7% of the supportive argument is made in negative form.

Explanations

Explanations are basically parts of the statements that provide reasoning in support of the guidance. In our data, about 45.3% of the guidelines contained either a cause (E/Cause) or supportive (E/Supportive) argument.

Figure 7 shows a breakdown of the content of the explanations. At the top are insight, user ability, and user tasks, optionally qualified by qualitative attributes. Next come explanations that indicate support for a particular visualization aspect, followed by data aspects, in general. Then, we saw insight, user ability, and user's tasks with additional context provided in reference to a specific visualization or data aspect. We saw little use of visualization interaction as a cause. Device characteristics, data size, privacy, etc. were very rarely specified as argument for the explanation.

Discussion

Visualization design is not a simple task. Our grounded theory analysis reveals that there are many including aspects of data, visualization, users, insight, and devices, and relationships among these. Of these, not surprisingly, visual aspect came at the top, followed by data. Aspects related to user and insight combined were also nearly as significant as that of data aspects. Interestingly, device related aspects came last by a significant margin.

Perhaps more strikingly, relationships among these basic concepts were as frequent as the top factor, visual aspects. These were the expressions that connected basic and more complex concepts through logical, causal, supportive, and adversative relationships. Similarly, data and visual operations also connected basic concepts. Of these relationships, supportive expressions came at the top. Along with causal expressions, which were also significant, indicate that many of the guidelines came with explanations. Explanations not only serve to help the designer understand but also allow them to make the proper trade-offs when considering a number of design options. We also observed significant presence of conditional, copulative, and adversative expressions suggesting that guidelines were expressed in detail with many statements combined to contextualize and support a design guideline. Prepositional expressions were lower in frequency but they existed in many guidelines and helped properly describe the intended relationships of concepts in the guideline.

Significant presence of expressions suggests that there is a lot more complexity that goes into the visualization design than considering basic guidelines alone. There are significant interactions among factors and design without considering such interactions is likely to result in failure. Furthermore, qualifiers that bring about further attributes of the concepts surfaced significantly in the guidelines, in fact more than user and insight aspects combined. These included qualitative attributes, such as efficiency and importance, as well as computed attributes such as size, length, distance, etc. The significant emergence of qualifiers not only further adds to the complexity argument but also brings up another argument, the qualitative and judgmental aspect of design.

Our hierarchical coding of concepts allowed us to examine factors and interactions among them at several levels. Our analysis suggests strong relationships among integration task and composition insight. Similarly, other interactions among several visual attributes, components, and types of visualizations, such as between position attribute and 2d plots, length attribute and bar element, size attribute and shape elements.

Examining the content and form of the core part of guidelines (i.e. instructions) reveals further insight into visualization design. From Figure 4, we again see significant presence of visualization related concepts. Visual operations came at the top, followed by classes of visualizations, then by combinations of several visualization concepts, particularly of visual attributes, elements, and components, in order of frequency. Color emerged as the most discussed visual attribute, in fact color palette was also significant.

After visual aspects came data related aspects. Specifically data attribute was amount the top concepts, particularly its measurement type. Nominal was the most commonly referred measurement type. Attribute semantics also played a key role in design guidelines, particularly temporal semantics. Attribute role was also important but not as important as attribute measurement or semantics.

We saw several patterns that combined visual and data aspects along with data and visual operations. Visual operations topped the list, mapping operation was the most significant, followed by aggregation and order. Among data operations group and aggregation operations were common though we observed others such as data reorganization (e.g. slice) and transformation operations.

Examining the conditional guidelines, we again see significant use of the visualization and data concepts, perhaps not surprisingly, in the condition part. On the other hand, user, insight, and device concepts were rather low. Our data suggests that close to two-thirds of the conditions were complex, involving one or more expressions. For example, existential expression, requirement related to existence of a particular type of attribute, was the top expression used, particular as it related to measurement types, followed by semantics, and role. Qualitative qualifiers were also significantly used in the conditions. Consequence part of the conditionals exhibited similar patterns frequencies, with combinations of visual aspects coming at the top, followed by data related aspect. Again, data schema and domain aspects were not to be ignored.

In rankings, we see that visual aspects were compared mostly, particularly, visualization class, attribute, operation, and element in order of frequency. We also noted comparisons between intended insight and qualitative aspect of visualizations, and among qualitative aspects, suggesting descriptions of trade-offs between alternatives.

In summary, we believe three themes emerge from our analysis: (1) complexity of visualization guidelines in form and content, particularly important as it relates to big data and broader use of analytics, (2) importance of qualitative aspects of visualization design, particularly relevant as visualizations becoming more a commodity for data analysis in several domains, and (3) relatively low utilization of some concepts, such as interaction and collaboration.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

Table A. Concept-related codes and their sub-concepts (in capital-case), along with instances (in lower-case)

Concept	Sub-concept(s) and instances
Data/Schema(*)	2d, 3d, multid, multivariate, mixed, sequential, hierarchical, graph
Data/Attribute/*	Value (low, high, avg), Size (large, small), Semantics (temporal, spatial, geographical, temperature, percentage, count), Unit, Role (measure, control), Measurement (nominal, ordinal, interval, ratio, numeric, continuous, discrete, cyclical), Distribution (sparse, dense, uniform), range (wide, narrow)
Data/Operation/*	Aggregation (spatial, average-mean, average-median, sum, min, max), Outlier, Sort (increase, decrease, alphabetically), Union, Annotate, Explain, Slice, Filter, Simplify, Group (small, range, type, region, time), Transform (text-numeric, text-abbreviation), Fit (line, curve), Ratio, Round, Reduce
Data/Domain(*)	geography, business, politics, it, survey, science, technology, media, sports, meteorology
Data/*	Size (large, small), missing, incorrect, new, privacy, trust, precision, related (time), context, source, selection, organization
Visualization/Element	point, line, shape (circle, rectangle, cube, sphere, 2d, 3d, fill), bar, pie, text, node, edge, glyph, error bar
Visualization/Attribute/*	size, radius, position, color (hue-bright/dark/red/gray, saturation, brightness-light/dark), length, area, angle, direction, orientation, pattern, shape, label, weight, texture, closure, connection, volume, transparency, style, typography
Visualization/Operation/*	label, mapping, projection (3d, row, column), aggregation (stack, cluster, overlay), nesting, order (top-bottom, clockwise, time-incr, cyclical), rearrange, distortion, highlight (color, multi), group (bottom-N, 3, 5), scale, position (x, y, start, end, middle, below, aligned, left-aligned, right-aligned, center-aligned, baseline-aligned, point-aligned)
Visualization/Component/*	Background, Title, Axis (negative, h, v, tick, range), Grid, Label, Legend, Color palette (redgreen, lightdark, warm, cold, alternating), Shape palette, Description, Source, Thumbnail, Table (row, column, cell)
Visualization/Coordinate	cartesian-scale (double, range-small/baseline, numeric-linear/log, categorical), polar, geo (projection-mercator/albers)
Visualization/Class	XY, scatter, bar, list, table, graph, 2d, 3d, map, pie, donut, radial, calendar, histogram, hierarchy, abstract, boxwhisker, symbols, line, area, cycle, treemap, mosaic, star, contour, choropleth, cartogram, heatmap, aparallelcoords, dot, density, surface, pictogram, graphical
Visualization/Interaction(*)	change metaphor, filter, pan, rotation, scroll, select, zoom (multiple), link to source, annotate, explain, overview, detail, sort, split, animation, highlight, style (direct manipulation)
Visualization/*	Standards, Organization, Size (small, large), Aspect, Layer (map), Multiple, Aesthetics
User/Task	process, context, goal, reflect, emotion, perspective, comparison, integrate, search, describe, communicate, find, browse, explore, analyze, present, explain, monitor, decision making, classify
User/Ability	Attention, perceive (differentiate, order, measure, change), understand, recognize, read, learn, retain, recall, locate
User/	Disability (colorblind), social (conventions), action (click, move)
Qualifier/Qualitative	good, bad, relevant, easy, accurate, consistent, important, powerful, useful, precise, noisy, simple, fancy, informative, novel, efficient, beautiful, successful, appropriate, quick, familiar, redundant, usable, distorted, dense, cluttered, lossy, clear, explicit, implicit, flexible, compact, narrow, engaging, different, continuously, discretely
Qualifier/	Numeric (1, 2, 3, 4, 5, … , N, many, few), Cardinality (N, many, few), rank (N, top, bottom), inequality (more, less, N), Existential (many, some, none, only, all, except, emphasis), Logical (not), Temporal (recent), Spatial (width)
Insight/Trend	change, past, steady, strength, ragged, change, absolute, time, linear, cyclical
Insight/	relationship (multi, part-to-whole), correlation (pairwise, multi), composition, variance, extrema, value, comparison, distribution, progress, rank, quality, summarization, structure (hierarchy), cluster, causality, message, guide, meaning, outlier, gaps, percentage, similarity, pattern, detail, overview, focus
Device/	Display (size, resolution, aspect, density), Input (mouse, keyboard)

Complexity, Scalability, Automation

As big data becomes a major topic in enterprise analytics and consequently as the user base of analytics broadens, automating the design of effective visualizations (along with the underlying data and analytics pipeline) is now more critical than ever. This has several implications.

First, as discussed earlier, visualization design is a complex task, requiring consideration of several factors at once. Simple matching between data attribute characteristics and a set of visualizations may not do justice to represent the complexity of visual analytics. For big data, the situation is worse, as data must be carefully transformed and organized before even visualization is considered. Interactive visualization of big data is even harder. In our analysis, we found relatively lower use of data operations, particularly as they related to visualization and interaction with data. We need to systematically identify patterns of visual interaction with big data that exhibit high utility. Secondly, we argue that visualization and analytics should be considered, and optimized, in an integrated manner, to increase the overall effectiveness of the whole process. This requires a systematic approach in which data and visual operators are represented on equal terms, potentially requiring visual analytics algebra and language to define data, visualization, and interaction altogether, that can be optimized (as in relational algebra), as such making automatic visualization not a matching problem but rather an optimization problem

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

Domain, Semantics

Another implication of the broadening user base is that visualization is becoming more a commodity and used in several domains. In our data, we see relatively low use of data domain and semantics (beyond time) in guiding effective data visualization. However, we argue that domain and semantics will be more important as visualizations will be used in different domains, as part of everyday tools. There are several implications of this.

First, we need to develop more guidelines that leverage domain and semantics. For example, temperature in physics and medicine are very different concepts. Though some general principles apply, based on the measurement type, much is left to design, particularly as it relates to qualitative issues. In different domains, possible value ranges are different, more importantly the meaning associated with values are different. For example, in medicine, the normal body temperature ranges should be considered, while in physics, for example, it could be the melting point. These impact visual design, choices of colors, axis ranges, emphasis on critical points and ranges, etc.

Secondly, in our data we observed only very little consideration of multiple views and datasets. In almost any domain, decision making involves consideration of multiple data and how they relate to each other. This is clearly another area where guidelines need to be developed for effective visualizations.

Lastly, given the wide range of possibilities to incorporate domain and semantics into the equation of visualization design, again we may need to consider a systematic approach to design, one that involves a language that facilitates specification of different considerations. The language should be flexible to express different semantics and design know-how. A language based approach to automation would support customization quite effectively by building different repositories of design know-how for different domains and incorporating desired repository into the system based on domain of the data analytics problem.

Interaction, Collaboration, Presentation

Driven by broader use of visualization in decision-making in business and public, we need to support different phases of the process, including the analysis phase, which includes a lot of interaction with data but also the collaborative aspect of it. A critical and often overlooked phase is that of presentation of analytics work directly in support of decision makers. Unfortunately, we saw few guidelines that incorporated aspects of interaction, and almost none that considered collaboration and presentation.

Conclusion

Our grounded theory analysis of a large set of data visualization guidelines suggests that a variety factors are considered for effective visualization design, including aspects of data, visualization, user, and insight and a variety of relationships between these factors including causal, contextual, and logical relationships. We also reported gaps and potential directions we found in the literature requiring further work on principles and guidelines, for example, regarding effective support of collaboration and presentation. Our study suggests that, as visualization becomes a commodity used in several different fields by a variety of users of different backgrounds, we need a systematic approach to visualization design. Such an approach should preferably be based on a sound algebra and language for visual analytics, (1) able to express complex set of guidelines, (2) be flexible enough to customize for different domains, and (3) allow optimization of the complete pipeline, from data processing to visual representation and interaction.

References

[1] Shneiderman, B., The eyes have it: A task by data type taxonomy of information visualizations, Proc. IEEE Visual Languages '96, pp. 336-343, 1996.

[2] Craft, B. and Cairns, P., Beyond Guidelines: What Can We Learn from the Visual Information Seeking Mantra? In Proceedings of the Ninth International Conference on Information Visualization (IV '05). IEEE Computer Society, Washington, DC, USA, 110-118, 2005.

[3] Carr, D., Guidelines for designing information visualization applications, Proceedings of the 1999 Ericsson Conference on Usability Engineering, Stockholm, 1999.

[4] Conati, C. and Maclare, H., Exploring the role of individual differences in information visualization. In Proceedings of the working conference on Advanced visual interfaces (AVI '08). ACM, 199-206, 2008.

[5] Munzner, T., A Nested Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics 15, 6, 921-928, 2009.

[6] Brehmer, M. Munzner, T., A Multi-Level Typology of Abstract Visualization Tasks," IEEE Transactions on Visualization and Computer Graphics, vol.19, no.12, pp.2376-2385, Dec. 2013.

[7] Tory, M., Moller, T., Rethinking Visualization: A High-Level Taxonomy, IEEE Symposium on Information Visualization, 2004. INFOVIS 2004, vol., no., pp.151-158, 2004.

Table B. Relationship-related codes (e.g. Expression) and their usage

2nd Level codes	Usage	2nd Level codes	Usage
Expression/Logical	and, or	Expression/Alternative	X ... alternatively... Y
Expression/Inequality	more, less	Expression/Exemplar	such as …, for example, ...
Expression/Existential	1,2, many, some, none, only, all, at least,…	Expression/Causal	So that ...,because of X, … Y, X leads to Y
Expression/Rank	If X is ranked * than Y	Expression/Possessive	X of Y, X with Y
Expression/Conditional	If X then Y	Expression/Prepositional	X in Y, X below Y, X with Y
Expression/Copulative	Connecting X and Y	Expression/Declarative	X is Y
Expression/Optional	Optionally, ... Y	Expression/Imperative	do X
Expression/Supportive	X supports Y, X allows Y	Expression/*	Equal, Similar, Spatial

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-132

[8] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S., Empirical Studies in Information Visualization: Seven Scenarios, IEEE Transactions on Visualization and Computer Graphics, vol.18, no.9, pp.1520-1536, Sept. 2012.

[9] Card, S. K. and Mackinlay, J., The structure of the information visualization design space. In Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97), pages 92–99, 1997.

[10] Keller, P. R., and Keller, M. M., Visual Cues: Practical Data Visualization, Alamitos, CA: IEEE Computer Society Press, 1994.

[11] Gotz, D., Zhou, M.X., Characterizing users' visual analytic activity for insight provenance, IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST '08, pp.123,130, 2008.

[12] Dias, M. M., Yamaguchi, J. K., Rabelo, E., and Franco, C., Visualization Techniques: Which is the Most Appropriate in the Process of Knowledge Discovery in Data Base? In Advances in Data Mining Knowledge Discovery and Applications, Adem Karahoca (Ed.).

[13] Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge, United Kingdom: Cambridge University Press.

[14] Steele, J., and Iliinsky, N., *Beautiful Visualization: Looking at Data Through the Eyes of Experts* (1st ed.). O'Reilly Media, Inc, 2010.

[15] Yau, N., *Data Points: Visualization that Means Something* (1st ed.). Wiley Publishing, 2013.

[16] Wong, D., The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures. W. W. Norton & Company, 2013.

[17] Ward, M., Grinstein, G., and Keim, D., *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA.

[18] Battiti, R., Mauro B., Reactive Business Intelligence. From Data to Models to Insight, Trento, Italy: Reactive Search Srl, 2011.

[19] Mackinlay, J., Automating the design of graphical presentations of relational information. ACM Trans. On Graphics, 5(2):110-141. Apr. 1986.

[20] Polowinski, J., and Voigt, M., VISO: a shared, formal knowledge base as a foundation for semi-automatic infovis systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13). ACM, 1791-1796, 2013.

[21] McGill, C. Graphical Perception: Theory Experimentation and Application to the Development of Graphical Models, Journal of the American Statistical Association, Vol. 79, No. 387., pp. 531-554, 1984.

[22] Kelleher, C., and Wagener, T., Short communication: Ten guidelines for effective data visualization in scientific publications. *Environ. Model. Softw.* 26, 6, 822-827, 2011.

[23] Wentian Li (1992). Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory 38 (6): 1842–1845.

[24] Hoey, J., The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test, 2012. eprint arXiv:1206.4881.

Author Biography

Eser Kandogan received his Ph.D. in computer science from the University of Maryland, College Park (1998). Currently, he is a research staff member in the Accelerated Discovery Lab. at IBM Research – Almaden. His research interests include collaborative visual analytics, human-computer interfaces to data, and workplace studies on data analysis.

Hanseung Lee obtained his M.Sc. in computer science from the University of Maryland, College Park (2014). Currently, he is a software engineer at Google. His research interests include information visualization and human computer interaction.