Towards Concept-Driven Visual Analytics

In Kwon Choi[*], Swati Mishra, Kyle Harris, Nirmal Kumar Raveendranath, Taylor Childers, Khairi Reda[†]
Indiana University–Purdue University Indianapolis

Abstract

Visualizations of data provide a proven method for analysts to explore and make data-driven discoveries. However, current visualization tools provide only limited support for hypothesis-driven analyses, and often lack capabilities that would allow users to visually test the fit of their conceptual models against the data. This imbalance could bias users to overly rely on exploratory visual analysis as the principal mode of inquiry, which can be detrimental to discovery. To address this gap, we propose a new paradigm for 'concept-driven' visual analysis. In this style of analysis, analysts share their conceptual models and hypotheses with the system. The system then uses those inputs to drive the generation of visualizations, while providing plots and interactions to explore places where models and data disagree. We discuss key characteristics and design considerations for concept-driven visualizations, and report preliminary findings from a formative study.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics

1 Introduction

Visualization transforms raw data into dynamic visual representations that, through human interaction and interpretation, provide new insights. Well-designed visualizations enable people to explore and make data-driven discoveries—a bottom-up process. Yet, an equally important discovery pathway (indeed, considered the hallmark of good science) involves a top-down method of conceptualizing models and hypotheses, and testing those against the data to validate the underlying knowledge. Scientists are well-known for mixing exploratory (bottom-up) and hypothesis-driven (top-down) activities when making sense of data [5]. Statisticians also recognize the need for *both* exploratory and confirmatory analyses [6].

By contrast, current visualization tools, if inadvertently, discourage users from explicitly testing their expectations, nudging them instead to adopt exploratory analysis as the principal discovery mode. Visualization designers focus mainly on supporting low- to mid-level tasks (e.g., overviewing the data, browsing clusters), but often neglect features that aid users in testing predictions and validating their provisional hypotheses, thus making it less likely for users to engage in these activities. Cognitive research suggests that hypothesis-driven reasoning is vital to discovery and conceptual change. For example, in a study that simulated discovery from experimental data, researchers found that participants often failed at making new insights unless they explicitly tested their working hypotheses again the data, and set goals for themselves to explain observed discrepancies [1]. The lack of hypothesis-driven workflows could thus be a stumbling block to discovery in visual analytics.

We propose **concept-driven** visual analytics as a way to support a richer, bi-directional discourse between people and data. Concept-driven visualization tools will not only enable the exploration of attributes in a data-driven fashion (as current systems do), but will also encourage users to articulate their conceptual models and provisional hypotheses, and use those to drive the analysis. Their key characteristic is that they incorporate what the user believes into the visualization, highlighting the fit of one's model to the data, while providing interactions to encourage incremental model revisions. A typical concept-driven workflow starts with the user specifying his/her conceptual model and hypotheses, for example, by describing expected relationships between attributes in natural language or in a concept map. The system analyzes these specifications, selects data features and attributes that are relevant to those models, and visualizes them while highlighting places where model and data disagree. Interaction with the visualization is aimed at enabling users to visually dig into and reconcile model-data discrepancies. Unlike semantic interaction [2] which is limited to infering low-level features of users' metnal models, concept-driven visualizations invite people to proactively express hypotheses and models at a high-level, using those as inputs to generate conceptually relevant data plots. In doing so, we provide affordances for users to visually and incrementally explore a conceptual space of hypotheses [5].

2 Design Considerations

Concept-driven visualization beg two design questions: How do we enable people to intuitively communicate their hypotheses and models to the system? Second, how can we generate conceptually-relevant data plots from users' model specifications, while augmenting the resulting visualizations to highlight model-data fit?

2.1 Expressing and representing users' models

Natural language arguably provides an intuitive way for expressing conceptual models, allowing a user to specify hypotheses and predictions by simply typing or verbalizing them. For instance, an analyst looking at Chicago's crime patterns might hypothesize that "there should be correlation between drug use and weapon violations in the west side of the city because of more drug cartel activity, compared to the south side.". Although natural language queries have been used to generate visualizations [3], these techniques need to be extended to account for user predictions. Conceptual models can also be specified with concept maps, which allow for more structure compared to natural language. Here, a user can dynamically construct a node-link diagram to represent his/her model. Concepts corresponding to data attributes can be specified by dragging them to the canvas from a data ontology. Similarly, expected relationships can be specified from a list of propositions the editor is programmed to recognize. The concept map specification is then parsed and utilized as an active input to produce relevant visualizations. A third approach for model specification is to sketch expected relationships directly into a visualization template. Here, the user is first presented with an blank chart and prompted to graphically outline the expected data pattern, for instnace, by sketching a regression line in a scatterplot to predict the X-Y relationship [4].

2.2 Highlighting and exploring model-data discrepancy

To enable users to visually test the fit of their models, new representations are needed to encode the model-data discrepancies and facilitate exploration into unmet expectations. There are two possible design strategies here. First, we can employ salient encodes to attract users' attention to the mismatch between data and expectations. For instance, in a scatterplot, points with high squared-error

[*] email: inkwchoi@iu.edu
[†] email: redak@iu.edu

relative to a predicted regression line can be highlighted to distinguish them from points that follow the regression line more closely. A second strategy is to visually annotate the model specification to emphasize the conceptual cause for discrepancy. For example, concept map links that contradict the underlying data can be highlighted in salient red. Alternatively, a squiggly red underline, similar to how editing tools underline misspelled words, can be used on natural language specifications to pinpoint unmet expectations.

3 Formative study

We conducted a formative Wizard of Oz study to understand when and how people want to share their expectations during visual analysis, as opposed to following a purely exploratory approach. A second goal was to understand the kinds of models and hypotheses people choose to express when given the opportunity. We recruited 14 participants from a large university campus representing a variety of disciplines (health informatics, library sciences, engineering). All participants had prior experience with at least one data analysis tool (e.g., R and SPSS). The study comprised two 40s-minute sessions. In each session, participants were asked to analyze a given dataset and verbally report their insights. The datasets comprised socio-economic indicators and health-risk factors (e.g., GDP, poverty rate, infectious disease rates). We provided a sheet containing a summary of the dataset and a list of attributes along with a brief description of each. Participants interacted with a web interface that initially showed two text boxes: 'query' and 'expectation'. They typed their query and optionally provided an expectation. A wizard interpreted the query-expectation pair and generated a response visualization using Tableau and R. If an expectation was provided, the wizard manually annotated the visualization to superimpose the expected relationship on the data plot, highlighting any discrepancy. The result was then displayed to the participant as a static visualization. We opted for this minimalistic setup to minimize potential bias due to interface idiosyncrasy.

Findings: We observed three classes of queries in the study: expectation-driven, goal-driven, and exploratory. For expectation-driven queries, participants had a well-defined expectation about one or more attributes that they wanted to test against the data. Often, these expectations are based on recently acquired information (e.g., from news media) or from a long-held belief. For instance, one participant stated that she remembers "reading somewhere the Australian government has introduced incentives for women who are pregnant", so she looked at fertility rate for Australia by year, expecting to see an increase. The second class of query was goal-driven. Here, participants did not have an expectation. Nevertheless, they had a clearly formed goal or question that they wanted to answer. For instance, one participant stated that he "wanted to see how [life] expectancy for both males and females [to see if] there is any kind of inclination towards one gender.", but without specifying the expected gender-based effect. The third class of queries we observed can be classified as undirected exploration. Here, participants usually did not have a particular question, let alone an expectation. Rather, they wanted to start the analysis somewhere, often by randomly selecting one or two attributes from the data sheet.

The majority of queries (78.2%) were expectation-driven, followed by goal-driven (7.6%) and purely exploratory analyses (5.4%). The remainder queries were accompanied by expectations about the visual composition of the visualization (discussed below). These numbers suggest that expectation-driven reasoning is common, which validates the need for a concept-driven analysis style.

Types of user model: We coded models expressed by participants into four major categories: value-based, comparative, causal, or relationships. Value-based models typically expressed a guess or estimate of a trend, clustering pattern, or a specific data value (e.g., "I expect [poverty rate] would be around 64 or 65 percent"). Comparative models were often described in terms of similarity, disparity, or order between data items, often involving landmark locations. For instance, one participant expected "see high rates [of crime] in Chicago compared to other cities"). Causal models comprise a hypothesized link that explains an expected trend (e.g., "America has harsh weather conditions but not as harsh as Russia, when we see Russia is more towards the North Pole, so I wanted to see if that changes population growth"). Lastly, relationship models comprised specifications of interactions or correlations between attributes that are not necessarily causal: "I would think that if you have access to electricity then you would have access to education, medical, and other resources that can prevent HIV... and then I do think that there is relationship between prevalence of HIV and life expectancy especially if you don't have medical services".

Of the 262 expectations we coded, 31.2% were value-based, 27.8% were comparative, 19.8% were relationships between attributes, and 10.3% were coded as causal models. A fifth category comprised expectations about the graphical encoding (e.g., "I would have expected countries to be [in] different colors") or data characteristics (e.g., "There will be variation in prevalence of HIV"), but those only accounted for 6% and 9%, respectively.

Reaction to visualized expectation: When the graph matched participants' expectation (in 69% of queries), they often simply noted the validation. When the expectation was unmet, however, some participants articulated a hypothesis to explain the difference, and typically followed up with subsequent expectation-driven queries. In a few cases, participants questioned the veracity of data or entirely rejected the result. For instance, one participant commented that "It is impossible... It is not 19." upon seeing an unexpectedly low number for the average days needed to start a business in Bangladesh, claiming that he "[knows his] country". The majority of participants, however, only made a brief observation acknowledging that their expectation was not met and moved on to unrelated queries. This later point suggests that it is not sufficient to only depict the model fit. Rather, it seems important to provide custom, follow-up interactions that reduce the cost of exploring model-data discrepancy (e.g., by preselecting discrepant data points and automatically showing additional plots to explicate their attributes).

4 Conclusions and future work

We proposed concept-driven analytics, in which people explicitly articulate their conceptual models and visualization systems responds by selecting relevant data features and encoding the visual fit of models to data. A formative study suggests that this style of analysis is applicable in the majority of queries posed by participants. Our future work will focus on designing a functional tool that supports concept-driven analyses. We will also investigate designs for effectively blending exploratory and concept-driven workflows.

Acknowledgments

This work is supported by NSF award IIS-1755611.

References

[1] K. Dunbar. Concept discovery in a scientific domain. *Cognitive Science*, 17(3):397–434, 1993.

[2] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE TVCG*, 18(12):2879–2888, 2012.

[3] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proc. ACM UIST'15*, pp. 489–500. ACM, 2015.

[4] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proc CHI'17*, pp. 1375–1386. ACM, 2017.

[5] D. Klahr and K. Dunbar. Dual space search during scientific reasoning. *Cognitive science*, 12(1):1–48, 1988.

[6] J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.