Concept-Driven Visual Analytics: an Exploratory Study of Model- and Hypothesis-Based Reasoning with Visualizations

In Kwon Choi
Indiana University-Purdue University Indianapolis
inkwchoi@iu.edu

Taylor Childers
Indiana University-Purdue University Indianapolis
tayrchil@iu.edu

Nirmal Kumar Raveendranath
Indiana University-Purdue University Indianapolis
niraveen@iu.edu

Swati Mishra
Cornell University
sm2728@cornell.edu

Kyle Harris
Indiana University-Purdue University Indianapolis
kylwharr@iu.edu

Khairi Reda
Indiana University-Purdue University Indianapolis
redak@iu.edu

ABSTRACT

Visualization tools facilitate exploratory data analysis, but fall short at supporting hypothesis-based reasoning. We conducted an exploratory study to investigate how visualizations might support a *concept-driven* analysis style, where users can optionally share their hypotheses and conceptual models in natural language, and receive customized plots depicting the fit of their models to the data. We report on how participants leveraged these unique affordances for visual analysis. We found that a majority of participants articulated meaningful models and predictions, utilizing them as entry points to sensemaking. We contribute an abstract typology representing the types of models participants held and externalized as data expectations. Our findings suggest ways for rearchitecting visual analytics tools to better support hypothesis- and model-based reasoning, in addition to their traditional role in exploratory analysis. We discuss the design implications and reflect on the potential benefits and challenges involved.

CCS CONCEPTS

• **Human-centered computing** → **Visual analytics**; *Information visualization*;

KEYWORDS

Visual analytics; sensemaking; hypothesis- and model-based reasoning; mental models

1 INTRODUCTION

Visualization transforms raw data into dynamic visual representations that, through human interpretation, provide new insights [6]. 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 and hypothesis-driven activities when making sense of data [25]. Statisticians also recognize the need for *both* exploratory and confirmatory analyses [54], and have developed a range of statistical methods to support both analysis styles.

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 primarily on supporting data-driven tasks (e.g., overviewing the data, browsing clusters) [2, 5, 46], but often neglect features that would aid users in testing their predictions and hypotheses, making it less likely for users to engage in these

activities. It has been suggested that expectation-guided reasoning is vital to conceptual change [11, 26]. The lack of hypothesis-driven workflows in visual analytics could thus be a stumbling block to discovery. An opportunity exists to re-architect visualization tools to capture some of the cognitive advantages of hypothesis-centric science.

One way to counteract the imbalance in tools is to introduce visual model-testing affordances, thereby enabling users to externalize their conceptual models, and use them as entry points to sensemaking. In such *concept-driven* approach, the user shares his/her hypotheses with the interface, for instance, by describing expected data relationships in natural language. The system analyses these specifications, selects the pertinent data features, and generates conceptually relevant data plots. In addition to showing the data, concept-driven visualizations would incorporate a representation of the user's model and depict its fit. The visualization could also provide the user with targeted interactions for digging into model-data discrepancies.

Researchers and practitioners have started exploring designs that explicitly engage a viewer's conceptual model. For example, the New York Times featured a visualization that initially displayed a blank line chart, inviting the viewer to first predict and sketch the likelihood of a person attending college based on their parent's income level [1]. The empirical relationship is then visualized alongside the sketch, allowing the viewer to compare their prediction against real data. Kim et al. experimentally studied the effects of this kind of interaction, and found that it improved participants' recall of the data, even when they had little prior knowledge about the topic [23]. However, it is still unclear what effects such model-testing affordances might have when introduced broadly in general-purpose visualization tools. Given the opportunity, might users take the time to share their models and data expectations with the system? Or would they continue to adopt a purely exploratory approach to analysis, as is traditionally the case? What kind of conceptual models would users express? How might the availability of model-testing affordances impact users' analytic behaviors?

To investigate the above questions, we conducted a mediated experiment to explore how people might interact with a concept-driven visualization interface in a sensemaking context. Participants were asked to iteratively pose data queries and, optionally, provide their data expectations in natural language. They were then presented with manually crafted visualizations that incorporated a representation of their expectation alongside the data. We present a qualitative analysis of participants' interactions with this model interface and describe the emerging analytic behaviors. We then present an abstract typology of data models encompassing the variety of expectations we observed in the study. Our findings suggest design implications for incorporating model-testing affordances into visualizations tools. We assess the potential benefits and difficulties involved in realizing such designs and outline future research directions.

2 BACKGROUND & RELATED WORK

Concept-driven analytics is inspired by research on sensemaking and scientific reasoning. We examine relevant works from cognitive science and visual analytics. We then discuss attempts at making adaptive visualizations that respond to user tasks or models, and to natural language queries.

Sensemaking and Scientific Reasoning

Sensemaking refers to a "class of activities and tasks in which there is a native seeking and processing of information to achieve understanding about some state of affairs" [29]. Several models have been proposed over the years to capture the main activities of sensemaking. Pirolli and Card observe the following sequence in sensemaking: analysts approach the source data by filtering for relevant information [37], extracting nuggets of evidence, and re-expressing the evidence in a Schema. In this model, data is progressively funneled into increasingly sparser and more structured representations, culminating in the generation of hypotheses or decisions. While the model allows for feedback, it is often conceived as bottom-up and data-driven sensemaking.

An alternative model is Klein et al.'s data-frame theory, which posits that, when people attempt to make sense of information, "they often begin with a perspective, viewpoint, or framework—however minimal" [26, 27]. This initial "frame" can take the form of a story, map, timeline, hypothesis, and so on. The frame is, in essence, a meaningful conceptual model that encompasses the main relationships one expects to see. Here, sensemaking is primarily an expectation-guided activity: the analyst iteratively questions his/her frame by testing its fit against the data. Poor fit can lead one to elaborate the frame by adding new "slots" to account for specific observations, or, alternatively, cause the analyst to adopt an entirely new frame. Klein et al. argue that a purely bottom-up approach to sensemaking is counter-productive, because it blocks analysts from deliberately testing their expectations, which is essential to refining one's knowledge [26].

Dunbar echoes the above perspective in his empirical findings on the cognitive basis of scientific discovery. In this work, subjects were placed in a scientific investigation modeled after a prominent discovery of a novel gene-regulation mechanism [11]. Using simulated experimental tools similar to what scientists had used at the time, subjects were tasked to identify genes responsible for activating enzyme production in E. coli. Only a small fraction of participants discovered the novel inhibitory mechanism. The researcher observed that participants who made the discovery generally

followed a strategy of consistently comparing the data they gathered to their expectations, and setting goals for themselves to explain discrepancies. Successful discovery was, in essence, guided by frames and deliberate goals, rather than spontaneously arising from the data. Conversely, subjects who neglected setting expectations and opted for bottom-up experimentation experienced a mental "block"—they often failed to make the discovery despite having seen the necessary evidence [11]. In subsequent field research, Dunbar noted that successful scientists pay special attention to results that are inconsistent with their hypotheses, and, in what can be construed as a form of top-down reasoning, conduct new experiments in order to explain inconsistencies [13].

The above views of sensemaking emphasize model-based analysis and hypothesis-testing as critical components of discovery, and show that this type of reasoning is quite pervasive in the scientific enterprise [12]. Frames and expectations aid the discovery process by providing "powerful cognitive constraints" [11]. Even when having large amounts of information at their disposal, analysts often rely on existing frames, testing data against provisional hypotheses, more readily than they can spontaneously create new frames from data [28]. Unfortunately, though, current visualization tools do not provide interactions to scaffold this kind of expectation-guided analysis.

Visualization Tools for Sensemaking

Visualization plays an important role in data-driven science [22], helping people quickly observe patterns [6], which can raise new questions and lead to insights. Visualization researchers have developed design patterns to enable users to interactively navigate the information space and transition among data plots [20]. For instance, Shneiderman proposed "overview first, zoom and filter, then details on demand" [46]. The overview display depicts the major data trends and clusters to orient the user, with subsequent interrogations occurring through a combination of zooming and filtering actions. Although widely adopted, the layout, available interactions, and information content of "overview first" visualizations are often predetermined by the interface designer; they are thus less responsive to users' mental models, and do not explicitly support visual hypothesis or model testing.

A few visualization tools provide built-in features for users to externalize and record their hypotheses and insights from within the interface [18, 47, 49]. On the surface, this appears to provide an outlet for users to articulate their conceptual models and adopt concept-driven reasoning. However, a key limitation in these tools is that they treat hypotheses and knowledge artifacts as 'products'. Externalizing hypotheses is thus merely intended to help people recall their discoveries [17, 39], but not as potential entry points into the sensemaking process. Interestingly, most of these tools

were inspired by Pirolli and Card's bottom-up sensemaking model [37]. To our knowledge, there are no visualization tools that support top-down, expectation-guided analysis, as espoused by the data-frame theory [28]. Our work explores how people might utilize such a tool.

Some effort has been made to design predictive visualizations that adapt to user goals and models. For instance, Steichen et al. detect the viewer's current task from their eye gaze behavior [50]. Endert et al. introduced 'semantic interaction', which deduces conceptual similarity between data items based on how users manipulate the visualization, and accordingly adjust an underlying data model [14]. While these techniques provide a form of adaptation to user goals and models in real-time, they are limited to inferring low-level features of people's mental models. Furthermore, these techniques do not provide explicit hypothesis and model validation affordances, which we seek to establish.

Natural Language Interfaces for Visualization

Natural language has emerged as an intuitive method for interacting with visualizations [48]. There now exist tools that allow users to speak or type their queries in natural language, and automatically receive appropriate data plots [52]. Generally, these tools work by identify data features referenced in the query [10, 45], resolving ambiguities [15], and then generating appropriate visualizations according to perceptual principles and established designs [33]. Natural language processing (NLP) is primarily couched as a way to lower the barrier for creating visualizations, especially among novices [3, 19]. However, NLP also presents a compelling approach to tap into users' internal models and predictions about data; users could verbalize their hypotheses to the interface and, in return, receive custom visualizations that either validate or question those hypotheses. That said, current NLP-based tools are merely capable of responding to targeted questions about the data, and do not make an effort to infer users' implied models or data expectations.

While there exist intelligent tutoring systems that are capable of inferring mental models via NLP [8, 36], these systems are mainly intended for assessing a learner's mastery of fundamental scientific facts [30] (e.g., whether a student grasps the human body's mechanism for regulating blood pressure [16]). It is unclear if such techniques can be used to infer users' mental models in visual analytics. Furthermore, there is limited empirical knowledge about the nature of conceptual models people express when interacting with visualizations. With few exceptions (e.g., [32, 42, 55]), mental models are often treated as products of the visual analytic process — domain-embedded 'insights' gained after interacting with the visualization [21, 35, 44]. But do insights reflect conceptual models that people retain for the long-term, and

subsequently activate as frames when making sense of unseen data?

Our work directly addresses the above gaps; we collect example models from participants specified in natural language, and in the context of visual analysis sessions involving realistic datasets. Rather than limiting our analysis to 'insights' [7], we give participants the opportunity to engage in hypothesis- and model-based reasoning, by inviting them to articulate expectations to test against the data (before seeing the latter). As such, the sample specifications we collect reflect deeper conceptual models likely encoded and recalled from participants' long-term memories. We inductively construct a typology from these examples to understand the types of data models people come to express in visual analytics.

3 METHODS

We conducted an exploratory study to investigate how people might interact with a visualization interface supporting a concept-driven analysis style. Such an interface would give users the option of sharing their conceptual models and receiving relevant data plots in response. Because, to our knowledge, no such interface exists, we opted for a mediated study: subjects entered their queries and models in natural language through an interface, but the visualizations were generated manually by an experimenter (hereafter, the mediator). This design was inspired by earlier InfoVis studies [19, 57] and is particularly suited for this formative stage. The study addresses two research goals (RG 1, RG 2):

- **RG 1**—We sought to determine whether participants would opt to share their models and expectations during visual analysis, as opposed to following an exploratory approach. How often do people engage in expectation-guided analysis? Where do their expectations come from? How does the availability of model-testing affordances impact participants' sensemaking activities?
- **RG 2**—Given the opportunity to engage in expectation-guided analysis, what kind of models do people express? Can we classify these models according to the fundamental data relationships implied, while abstracting natural language intricacies?

Participants

We recruited 14 participants between the ages of 22 and 50 from a large, public university campus. Participants were required to have had at least 1-year experience in using a data analysis tool (e.g., R, SPSS, Excel, Tableau). Selected participants represented a range of disciplines: eight majored in informatics, two in engineering, two in natural sciences, one in social science, and one in music technology. Participants were compensated with a $20 gift card.

Setup

The study comprised two 40-minute visual analysis sessions. In each session, participants were asked to analyze a given dataset and verbally report their insights. We sought to place participants in an open-ended sensemaking context to elicit realistic behaviors. We thus did not provide them with specific tasks or questions. Rather, they were instructed to analyze the provided datasets by developing their own questions. This setup is thus similar to earlier insight-based studies [31, 40, 41, 43] but with one crucial difference: our interface is initially blank and contained only two empty text boxes: 'query' and 'expectation'. Participants typed their query and optionally provided an expectation in natural language. The mediator interpreted the query-expectation pair in real-time, and generated a response visualization using a combination of Tableau and R. If an expectation was provided, the mediator manually annotated the visualization to superimpose the expected relationship onto the plot, while highlighting any discrepancy between the data and the expectation. The result was then shown to the participant as a static visualization (see Figure 1). To help maintain participants' chain of thought, the interface allowed them to see a a history of the last five visualizations.
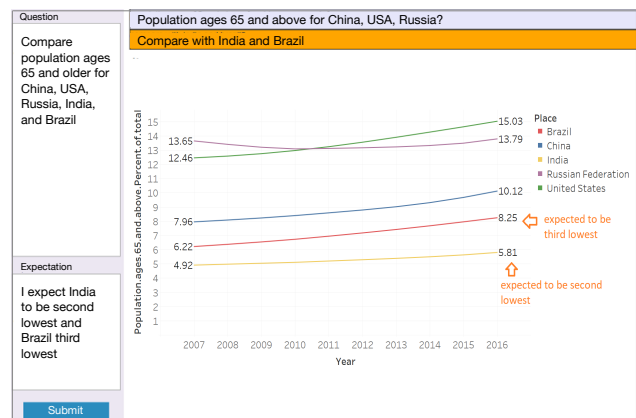
Figure 1: The experimental interface. Participants entered a query and, optionally, an expectation (left). They received a data plot with their expectation annotated to highlight the difference between and their model and the data.

Procedures

Participants analyzed two datasets over two separate sessions. The datasets comprised socio-economic and health indicators (e.g., labor force participation, poverty rate, disease incidence), depicting the variation of these attributes over geography and time. The first dataset comprised 35 health indicators for 20 major cities in the US in 2010-2015 [9]. The second dataset is a subset of the World Development Index with 28 socio-economic indicators for 228 countries in

2007-2016 [4]. We provided participants with a paper sheet containing a summary of each dataset and a list of attributes contained within. The list also included a brief description of each attribute along with its unit of measurement. Additionally, we provided printed world and US maps with major cities indicated to facilitate geo-referencing.

At the beginning of the study, participants were given a few minutes to simply read the attribute list and scan the maps. In pilots, we found that this step is conducive to the analysis; it seemed participants were using this time to formulate initial lines of inquiry and activate prior knowledge, form which they seem to derive concrete expectations. Next, the experimenter demonstrated the interface mechanics by typing two example query-expectation pairs and asking participants to try examples of their own, emphasizing that input is in free-form, natural language. Participants sat at a desk and viewed the interface through a wall-mounted 50-inch monitor. They interacted using keyboard and mouse. In addition to the mediator who sat behind and away from the participant, another experimenter was present in the room to proctor the study and answer questions. The study was video and audio recorded. Participants were also instructed to think aloud and verbalize their thoughts. Additionally, screen captures were recorded periodically.

Analysis and Segmentation

In addition to textual entries that were manually entered through the interface, we transcribed participants' verbal utterances from the audio recordings. The combined verbal and typed statements were then segmented into clauses following verbal protocol analysis techniques [53]. This resulted in a total of 835 segments, after discarding technical questions about the interface and utterances not germane to the study. We then classified segments into one of three categories: *question*, *expectation*, and *reaction*. Questions are concrete inquiries about data attributes (e.g., P3: "Can I get the adult seasonal flu vaccine [rates] from 2010-2015 for all cities?"). Expectations comprised statements that are indicative of predicted data relationships participants expected to observe, before actually seeing the visualization (more on this in Section 5). Reactions include comments participants made after seeing the resulting visualization (e.g., P12: "Ok so it's much higher than expected.").

Because a single question was often followed by multiple verbalized expectations, our method resulted in more expectations than questions. However, we maintained a link between each expectation and its origin question to adequately capture the context. Hereafter, we refer to the cycle of formulating a question, providing one or more data expectation (optional), and receiving a visualization as a *'query'*. In total, we collected 277 distinct queries and 550 expectations from 14 participants (distribution illustrated in Figure 2).
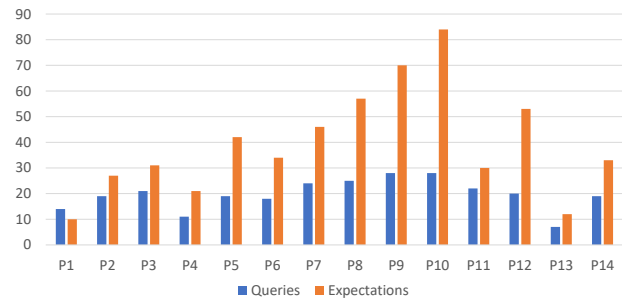

Figure 2: Distribution of queries and expectations across participants.

Selecting Visualization Templates

In response to queries, the mediator sought to generate familiar visualizations, such as line charts, bar charts, and scatterplots. The appropriate visualization was determined after considering the query provided by the participant and the types of attributes involved. For example, when the participant used words like 'correlation' or 'relationship' between two quantitative attributes (e.g., P3: "Is there a relationship between country's population density and total unemployment rate in 2016?"), the mediator used a scatterplot. On the other hand, when the participant referenced temporal trends (e.g., "across the years" or by providing a specific time range), the mediator employed a line chart and represented the referenced attributes with one or more trend lines (e.g., P14: "School enrollment of females vs males in Pakistan vs Afghanistan over time").

Visualizing Participants' Expectations

The mediator employed a variety of strategies to superimpose participants' expectations onto the visualization, aiming to visually highlight any mismatch between data and expectations. We outline some of these strategies but note that the resulting designs are not necessarily optimal. Rather, the resulting visualizations reflected the mediator's best attempt to respond in real-time and with minimal delay so as not to frustrate participants or disrupt their thought process. These constraints made it difficult to optimize the design, so in most cases the mediator opted for simplicity over optimality.

It was relatively straightforward for the mediator to superimpose participants' expectation onto the visualization, when the expectation can be directly mapped to specific marks in the visualization. For instance, in a query about female unemployment in Iran vs. the US, P14 expected that the "percentage in Iran is higher". The corresponding visualization comprised two trend lines depicting how unemployment changed in the two countries over time. Here, the mediator added an arrow pointing to Iran's along with a text
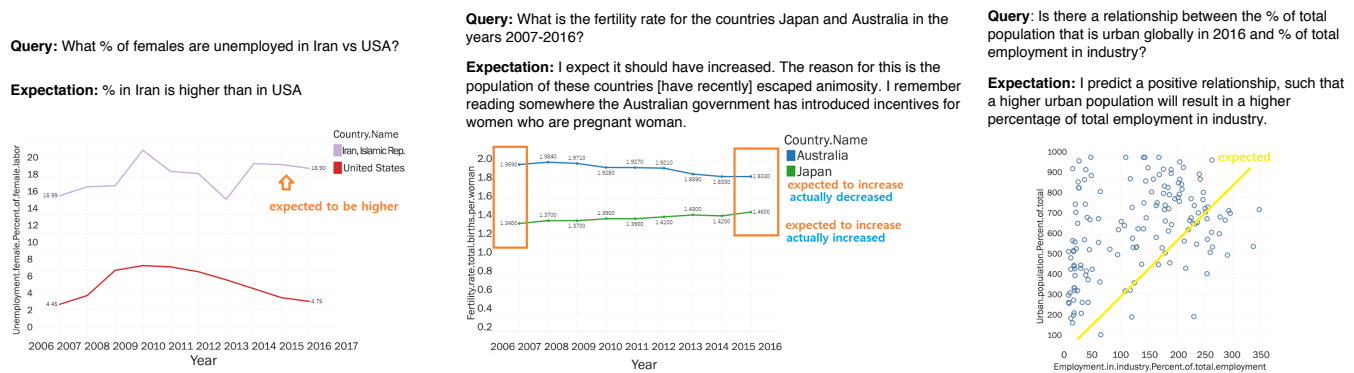
Figure 3: Three example query-expectation pairs from the study and the associated response visualizations as created by the mediator in real-time.

annotation reading "expected to be higher" (see Figure 3-left). During pilots, we found that adding textual annotations improved the readability and prompted more reflection. Generally, whenever the participant's model implied inequality (e.g., between two or more geographic locations), an arrow was added to the mark(s) referenced along with a text annotation indicating the predicted ordering. Similarly, when the participant predicted a specific quantity, a text annotation restating the expectation (e.g., "60 expected") was added near the corresponding mark. Such annotations, while simplistic, served to focus participants' attention onto parts of the plot most relevant to their expectation, enabling them to efficiently compare their model and the data.

Expectations involving correlations and trends required some interpretation on the part of the mediator, as those often lacked information about the predicted strength of the correlation or the slope of the trend. The mediator attempted to infer the implied strength from context, and visualized such expectations as line marks overlaid onto the plot. For instance, Figure 3-right shows a regression line (yellow) for an expected "positive relationship" between urban population and total employment in industry. When the strength of predicted trends could not be readily determined, the mediator employed text annotations and indicated whether, subjectively, the data had met the expectation (Figure 3-center).

4   FINDINGS — USAGE PATTERNS

We first discuss our observations on how participants approached the sensemaking task, and how they interacted with the model concept-driven interface (RG 1). In Section 5, we systematically examine the types of expectations externalized by participants (RG 2).

Analytic Behaviors

We observed two distinct analytic behaviors: model validation and goal-driven querying.

Model Validation. The most frequent analytic behavior, occurring in 218 queries (78.7% of total), can be classified as _model validation_; the participant sought to explicitly test his/her expectation against the dataset. When seeking to validate models, participants typically provided a query that directly probed their model along with a clear expectation to be tested. For instance, P4 asked about the relationship between alcohol abuse and chronic disease, expecting to "see a clear positive correlation." During the study, participants were asked to indicate whether their expectations had been met after receiving the visualization: 107 model-validation queries (53.5%) resulted in data that did not fit participants' expectations, 18 (9%) resulted in partial fit, and 75 (37.5%) were described as having a 'good fit'.

Reaction to the visualization varied depending on whether the participant's expectation had been met. When the visualization showed data in agreement with the model, the participant typically verbally indicating that their expectation had been met, often repeating their prediction so as to seemingly emphasize the validated model. However, when data is presented that contradicted the expectation, we observed several types of reaction.

In 22 instances (17.3% of queries resulting in unmet expectations), participants formulated new hypotheses in an attempt to explain the mismatch. For example, P2 asked a question about the percentage of youth population in five U.S. cities, expecting to see a higher percentage in Washington D.C. compared to Denver. This expectation was not borne out, and upon examining the graph, the participant responded with a possible explanation: "maybe because D.C. has more government offices, or people there work in governmental offices mostly, so because of that there are more old people than youth, comparatively." Occasionally, participants drew on factors external to the dataset to justify the mismatch.

In a second type of reaction, unmet expectations led a few participants to generate follow-up queries in order to explain the mismatch. For example, P7 predicted increasing unemployment in Italian industry. When his expectation did not match the visualization, a subsequent line of exploration was developed to investigate other sectors "so we are seeing that unemployment has increased in 2014 and then it went down twice, so can we see the employment in the services area?" However, this type of analysis directed at model-data discrepancy was rare, occurring in 3 queries only (2.4%).

On the other hand, some participants responded to unmet expectations by questioning the veracity of the data. This behavior was observed 14 times (in 11% of queries resulting in model-data mismatch). Here, both subjective impressions and objective anecdotes were invoked to justify the rejection of the result. For example, P12 asked about the birth rate per woman in his home country of Bangladesh. When the response came back as approximately half his expectation (2.13 instead of 4 predicted births), the participant responded that "[he's] unsure how they actually got this data... it really doesn't reflect the conditions." This type of reaction was generally pronounced when the participant had personal knowledge about the subject. As another example, P1 expressed an expectation of a correlation between rural population and agricultural employment but when the data contradicted his views he stated "this displays the opposite of what I expected... is this the real data?" Lastly, in the majority of cases (88 queries, or 69.3%), participants simply moved on to another query, without explicitly addressing the model-data discrepancy.

Goal-Driven Querying. We classified 52 queries (18.8%) in which participants sought to answer well-defined questions as *goal-driven.* In contrast to model validations, goal-driven queries lacked expectations despite stemming from concrete epistemological objectives. For instance, when querying lung cancer mortality in Minneapolis, P7 simply noted: "I don't know what to expect".

The remaining 7 queries (2.5%) comprised questions about the structure of the study, the nature of "questions that can be asked", or about the meta data (e.g., P5: "How complete is the data available on access to electricity in 2007?")

Sources of Expectations

To understand where participants' data expectations come from, we analyzed verbal utterances looking for references to the origin. In queries where the source could be identified, participants drew upon two sources of information when conceptualizing expectations: their prior knowledge and information gleaned from earlier visualizations seen in the study. In 19 of 26 expectations (73.1%), participants relied on their memory when formulating predictions. This included

information remembered from the media as well as personal perceptions and anecdotes. For example, when expecting a higher employment rate in Detroit compared to other major cities, P2 noted that she had "heard that there are lots of jobs in Detroit" because "lots of friends after graduating from here moved to either Texas or Detroit." On the other hand, 7 expectations (26.9%) were formed or derived based on information uncovered through previous queries. An example of a study-informed expectation is when a participant, after asking about the time needed to start a business in Bangladesh, followed up with a query requesting the same data for the US. In specifying his expectation, the participant predicted that "if it is 19.5 days [in Bangladesh] then it should be around 3 [in the US]". Most participants used a combination of the two methods, conceptualization hypotheses from their memory and drawing upon insights they uncovered during the study, although the former was more common, on average.

5 FINDINGS − MODELS ABOUT DATA

To understand the mental models participants held and externalized during the study, we analyzed and coded their expectations. We then synthesized a typology comprising the most frequently observed models. We formally account for each model type with a set of abstract, language-independent templates. We begin by describing the coding methodology.

Coding Methodology

Two coders inductively coded participants' expectations (n=550) using a grounded theory approach [51]. Because expectations were often incomplete on their own, the coders also considered the associated queries to resolve implicitly referenced attributes. Additionally, the coders considered the resulting visualizations, adopting the mediator's interpretation to resolve any remaining ambiguities. Long sentences that had no interdependencies between the constituting clauses were divided into multiple expectations during the segmentation phase (see 'Analysis and Segmentation').

Throughout the coding process, the emerging themes were regularly discussed with the rest of the research team and revised iteratively, leading up to a final coding scheme. The entire expectation dataset was then recoded using the finalized scheme. Coding reliability was measured by having the two coders redundantly code 40 expectations (approximately 8% of the dataset). Intercoder agreement was 90%, with a Cohen's kappa of 0.874, indicating strong agreement [34]. The scheme was organized as a two-level hierarchy; codes that represent conceptually similar expectations were grouped under 'types' and 'subtypes'.

Building on the final coding scheme, we synthesized abstract templates representing each type and subtype of expectation observed. The templates capture fundamental data relationships entailed, but omit the exact attributes and trends.

Model	Subtype	Goodness of fit	ID	Template
Relationships	Cause and effect	114 / 114	R1.	[X] causes [Y] (to increase | decrease) [in locations C+] [between T1-T2]
	Mediation	10 / 10	R2.	[X] [pos | neg] affects [Y] , which in turn [pos | neg] affects [Z] [in locations C+]
	Causal sequence	3 / 3	R3.	A series of past events (E1, E2, E3, ...) caused [X] (to increase | decrease) [in locations C+] [between T1-T2]
	Correlation	73 / 77	R4.	[X] [pos | neg] correlates with [Y] [in locations C+] [between T1-T2]
Partial Ordering	Direct comparison	119 / 120	C1.	Locations C1+ are (higher | lower | similar | dissimilar) compared to locations C2+ in [X] [between T1-T2]
		5 / 6	C2.	[X] is (higher than | lower than | similar to | different from) [Y] [in locations C+] [between T1-T2]
	Ranking	2 / 2	K1.	Locations (C1, C2, ...) are in (ascending | descending) order with respect to [X]
		13 / 14	K2.	Location A is ranked (Nth | lowest | highest) from [a set of locations C+ | all available locations] in [X]
Trends & Values	Values	102 / 102	V1.	[Average of] [X] (equals | greater than [equal] | less than [equal]) [V] [in locations C+]
		5 / 5	V2.	[X] is in the range of V1-V2 [in locations C+]
		18 / 18	V3.	[X] is [not] (high | low) [in locations C+]
	Trends	31 / 33	T1.	[X] (increases | decreases) between T1-T2 [in locations C+]
		8 / 8	T2.	[X] exhibits a (geometric shape) between T1-T2 [in locations C+]

Figure 4: A typology of data models held and externalized by participants. The three major model types (Relationships, Partial Ordering, and Trends and Values) are further divided into subtypes. We formally describe each model subtype with language-independent templates. Color-coded slots in the templates represent data attributes (purple), geographies (cyan), time periods (green), event sequences (yellow), quantitative data values and trends (blue), and other qualifiers (grey). Parameters not always specified in participants' expectations are considered 'optional' and enclosed in brackets.

Instead, the templates define 'slots' that serve as placeholders for attributes, trends, values, and other qualifiers. Taken together, these templates constitute a typology of models participants externalized and tested against the data. Expressing an expectation can accordingly be thought of as selecting an appropriate template from the typology and 'filling' its slots to complete the specification.

Of the 550 expectations we recorded, 503 (91.4%) could be classified under the typology. The remainder could not be fully classified either because they were observed rarely (no more than once) and thus did not warrant inclusion as unique templates, or because, while they could be matched with a suitable template, they included additional information that could not be fully expressed by the matching template. The resulting typology is outlined in Figure 4. We observed three major model types: Relationships, Partial Ordering, and predicted Trends and Values. In the following, we discuss each model type in detail and outline its variants (hereafter referred to as subtypes). Figure 5 illustrates the number of times participants invoked the different models.

Relationships

The most frequently observed model (204 expectations, or 39.8%) entailed an expected relationship or interaction between two or more attributes. We classified this model into four subtypes, based on the nature of the relationship:

Cause and Effect (R1). Participants frequently expected an attribute to unidirectionally affect a second attribute. For example, P14 expected that "Because there are a lot of rules
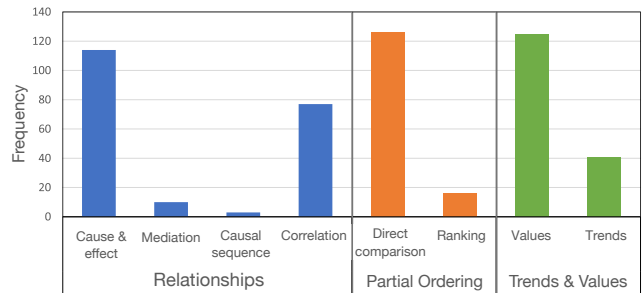

Figure 5: Frequency of model occurrence in expectations.

and regulations [that are] kind of not in favor of women's rights... I assume Iran should be lesser than USA" in regards to the rate of female unemployment. This example can be rephrased to fit template *R1* (see Figure 4): Rules and regulations that limit women's personal freedom [X] causes female employment [Y] to decrease in Iran [C]. The second clause from the above example was separately classified as a Direct Comparison between Iran and the US (more on this later).

Notably, participants frequently referenced causal factors that were outside the dataset. For instance, in the above example, there were no attributes about 'rules limiting personal freedoms' in our dataset. Additionally, participants occasionally qualified the expected cause-and-effect relationship to specific locations (e.g., 'Iran') or to a certain time window.

We determined a goodness of fit for each template (reported in Figure 4) by counting instances where expectations

could be fully expressed in the template by filling slots in the latter. All 114 expectations implying simple cause-and-effect relationships could be expressed with this template.

Mediation (R2). In addition to simple cause-and-effect relationships, participants also expected mediations; an attribute affecting another attribute indirectly through a mediator. For example, P5 stated: "I'm assuming the more [people] go to school the more educated they are, and the more life expectancy will increase for all of them." This example can be expressed under template *R2* with educational attainment `[Y]` acting as a mediator between the rate of secondary school enrollment `[X]` and life expectancy `[Z]`. Consider a second example from P9: "I'm guessing people generally don't really think about settling down in New York City unless you have some high income source so." Here, cost-of-living `[X]` negatively affects the percentage of population without stable income `[Y]`, which in turn negatively affects unemployment rate `[Z]` in New York City `[C]`.

Causal Event Sequence (R3). A chain of past events is hypothesized to affect an attribute. Consider an example from P6 who expected that "after the war on Iraq, for whatever reasons, and the [capture] of Saddam and installing [of a] new government... I think, around this time will come for [American forces] to leave the country and [return] autonomy to local Iraqi government... I wanted to see how education has progressed between 2008 and 2016." In this model, which can be expressed with template *R3*, the first event is the invasion of Iraq, the second is the capture of Saddam Hussein, followed by returning control to an Iraqi government. This chain of events was expected to affect secondary school enrollment `[X]`, an attribute in the dataset. Consider a second example where P6 hypothesized that the fertility rate in Australia would have increased because of a government policy: "I remember reading somewhere the Australian government has introduced incentives for women who are pregnant woman... So I just wanted to know how the thing has fared." Causal event sequences were rare, occurring only in 3 expectations, all of which fit the template.

Correlation (R4). A decidedly weaker relationship than cause-and-effect. Generally, participants indicated whether they expected a "positive" or "negative" correlation, without implying causality. For example, P3 "[predicted] a positive relationship" between opioid-related mortality and poverty. In a few cases, participants did not concretely specify the direction of the correlation, simply stating that there would be one (e.g., P5: "I'm assuming there is a correlation"). However, based on their reaction to the visualization, the expected correlation was almost always conceived as positive. Similarly, participants rarely specified the strength of an expected relationship; only a handful of participants used the words

"strong", "significant", or "linear" to describe a correlation. In a few instances, the expected correlation was qualified to a specific location or a time window (e.g., P3: "I predict a positive correlation [between country population density and total unemployment rate] in 2016"). This template expresses 73 of 77 expected correlations, with the exceptions of utterances implying loose relationships that could not clearly identified as causal or correlative.

Partial Ordering

This model defines a partial order or ranking for a set of data items under a specific criterion. Often, the referenced data items denote cities, countries, or clusters of geographical locations with similar characteristics (e.g., "big cities"). The model, observed in 142 expectations (27.7%), can be divided into two subtypes: Direct Comparison and Ranking.

Direct Comparison (C1, C2). Participants frequently drew direct comparison between two sets of items, expecting one set to exhibit higher or lower values with respect to a specific attribute. For instance, P9 expected "[preschool enrollment] to be higher in Boston as opposed to in Cleveland." The magnitude of the difference was often unspecified.

Most direct comparisons were limited to two specific data items (e.g., 'Boston' and 'Cleveland'), but in a few cases multiple items were collectively compared to a second set of item (e.g., P14: "I think Spanish speaking population is more on the west coast compared to east coast"). As this example suggests, it was often unclear whether the inequality was expected to held over the averages of the two sets ('east' and 'west' coast cities), or whether a direct pairwise comparison between items in the two sets was implied. Barring the ambiguity, this type of expectation can be expressed with template *C2*.

Ranking (K1, K2). In addition to making direct comparisons, participants also ranked a set of items (usually geographic locations) with respect to a specific attribute. For instance, P1 expected "LA to be the highest [in excessive housing cost burden] followed by Boston, Denver, Portland and Houston." Such models were observed only a handful of times in our study.

In a more common subtype of this model (template *K2*), participants predicted the rank of a specific location relative to the rest of the dataset. For instance, P8 expected "India will be second lowest [with respect to population aged 65 and above]." Participants also predicted the locations ranking highest or lowest in regards to an attribute (e.g., P8: "Maybe, [for unemployment rate of females], China [is] lowest, and Saudi Arabia [is] highest"). In rare cases, participants invoked a multi-attribute ranking criterion. For instance, when

asking "Which city has the lowest combination of teen smoking, unemployment, and children's blood lead levels", the participant expected "San Jose or Washington DC". However, this happened rarely, and thus not accounted for in our typology.

Trends and Values

This model, observed 166 times (32.4%), entails expectation of encountering specific values or trends within the data. Participants typically invoked a single attribute and specified explicit quantitative expectations on how that attribute manifests at specific geographies, or in certain time periods. We observed two subtypes: Values and Trends.

Values (V1–V3). Expectations containing specific quantitative values for an attribute were very common. For instance, P10 expected that "in developed countries, college enrollment rate will be a little higher than 25 percent", which can be expressed with template *V1*. In some instances, participants explicitly referred to the 'average' expected value in different locations, or over the entire dataset. We also observed participants expecting a range of values, rather than a single concrete quantity (e.g., P9: "I expect it to be around 85-90%"). Lastly, in 18 value-based expectations, participants predicted an attribute to be simply 'high' or 'low', without elaborating (e.g., P12: "I think the number is very low.")

Trends (T1, T2). Hypothesized trends were typically conceptualized as increasing or decreasing values over time for a single attribute. For instance, P4 predicted "the number [of unemployed females] in China will be decreasing". In some cases, participants provided a more concrete geometric description of the predicted trend, a model that can be loosely captured with template *T2*. For example, P2 predicted that "[in] Miami, [one] might see some peaks [in opioid-linked mortality rate] during especially the seasonal times"). With the exception of two instances, all 166 expectations involving concrete values and trends fit one of the above templates.

6 DISCUSSION & DESIGN IMPLICATIONS

Our study sheds light on how people might adopt hypothesis- and model-guided reasoning in visual analytics. We reflect on the main findings and, where appropriate, consider the design implications (highlighted with a ★).

Sensemaking with a Concept-Driven Interface

Given the opportunity, we found that a majority of participants engaged in model validation, by explicitly outlining their expectations when submitting a query. Overall, 78.7% of queries were accompanied with concrete predictions about the data (recall that participants had the option of not providing expectations and issue questions only). However, it

was difficult to gauge whether an externalized expectation was primarily intended to validate a model, or whether it simply reflected a participant's best guess in what could be an information gathering task.

Interestingly, a majority of expectations (73.1%) whose source is indicated were conceptualized or recalled from memory; a smaller portion could be directly attributed to information seen in an earlier visualization. This latter point suggests that participants' strategy, to a large extent, comprised exploration of a 'hypothesis space' [25]. In other words, participants did not seem to adopt a purely data-driven approach, in which the analyst 'explores' the information space and develops insights from data. Rather, it seems that the interface helped scaffold hypothesis-based inquiry, driven primarily by conceptual models participants held, developed, and tested throughout the experiment.

★ *Design Implication:* Our findings suggest that letting users *optionally* externalize their models in visualization tools, and use those models to drive the generation of data plots would provide a viable visual analysis workflow. Furthermore, findings suggest that, in the majority of queries, users would leverage such affordances by activating their mental models as entry points to the analysis. That said, it is unclear whether a concept-driven analysis style would be beneficial to discovery. Being exploratory in nature, our study did not assess participants' ability to discover new knowledge. This issue could be investigated in a future controlled study comparing the different analysis styles (exploratory vs. concept-driven, or a mixture thereof).

Interacting with Visualized Expectations

Eliciting and visualizing one's data expectations is shown to statistically improve data recall [23, 24]. However, it is unclear how individuals respond to such visualizations. Our study contributes qualitative insights on how people react to seeing their expectations alongside the data.

When a participant's expectation fit the data, the reaction was brief and typically included a verbal response emphasizing that the expectation was borne out. Unmet expectations, however, evoked a variety of responses. Some participants attempted to justify the mismatch by invoking explanations or factors external to the dataset. On the other hand, a tiny minority of participants reacted to unmet expectations by developing additional lines of inquiry in an attempt to explain the mismatch. This response is reminiscent of behaviors in Dunbar's experiment, in which motivated subjects were able to make a conceptual leap in their understanding by deliberately digging into unexpected results [11]. Similarly, very few participants in our study persisted, attempting to investigate reasons behind the discrepancy between their models and the data. It is unclear, though, whether this behavior

resulted from intrinsic goal-setting on behalf of the participant, or whether it was partially prompted by reflecting on the visualized expectation.

Another type of response to unmet expectation is characterized by distrust; questioning the veracity of the data or the visualization. Participants exhibited this behavior when they had personal knowledge about the topic (e.g., when looking at economic data about their country of origin). On one hand, such reaction could reflect a healthy skepticism when encountering what might seen as implausible data. On the other hand, it may reflect a difficulty in changing one's model, when the latter is derived from deeply-held beliefs or personal impressions. Lastly, in the majority of unmet queries (69.3%), participants simply moved on without attempting to reconcile or explain the mismatch.

★ *Design Implication:* It seems plausible to suggest that simply visualizing the gap between one's mental model and the data is enough to prompt reflection and, by extension, improve understanding [23]. Nevertheless, the low rate of follow-up onto unmet expectations (2.4%) indicates a need for carefully thought out designs and interactions that motivate users to dig into model-data discrepancy, when the latter arises. Similarly, our findings suggest that users may still ignore flawed models and continue to have 'persistent impressions' in the face of discrediting evidence [56]. This in turn suggests a need for persistent visual highlighting of model-data inconsistencies that people cannot easily ignore.

Thinking about Data

Our model typology suggests that participants thought about the data with varying levels of generality. At one end of the spectrum, Relationship models suggest generalizable processes thought to cause an interaction between two or more attributes. Such models were validated by finding an exemplifying data item (often one that registers extremely), and testing whether its attributes are consistent the projection. For instance, the participant who hypothesized a negatively causal relationship between cost-of-living and unemployment opted to select New York City as a test bed (and where living costs are unusually high).

Similarly, expectations involving trends hinted at conceptualized phenomena that unfold over time, which in turn was thought to impact an observable attribute. However, trend-inducing processes were usually not stated concretely as part of the expectation. For instance, comments from the participant who predicted a trend of "decreasing female unemployment" in China implied a process where by more females are entering the workforce due to a combination of government policies and a shifting socio-economic landscape. However, unlike Relationships, an explanatory model was often not specified.

At the other end of the spectrum, Partial Orderings and Value-based models were less systematic in nature; these models often reflected personal observations and anecdotal facts remembered by participants. For instance, Direct Comparisons often invoked expectations about geographies with which a participant had some familiarity.

From Conceptual Models to Data Features

An important part in validating conceptual models is connecting them to concrete data measures. While participants were generally able to articulate models and derive meaningful expectations, they often did not establish clear correspondence between their models and the relevant data features in the dataset at hand. For instance, one participant expected a "correlation between children's blood lead levels and all types of cancer mortality". Although the first factor could be immediately resolved to a unique attribute, the dataset had several attributes related to cancer incidence. Participants often did not identify which attributes they intended to refer to, or whether the expected relationship is projected to hold over an aggregate (e.g., average incidence across all cancer types). Consequently, many of the models had ambiguities that were left for interpretation by the mediator.

We also observed participants referring loosely to data features, even when the corresponding attribute could be resolved unmistakably. This often presented as ambiguity in the level of granularity at which attributes are being 'accessed'. Consider the following expectation: "College enrollment rate will be a little higher than 25 percent." Bearing in mind that this attribute is time-varying, is the participant referring to the average value across the time series? Or is the model expected to hold individually in each year? Consider another example where a participant wanted to look at the prevalence of adult smoking in Philadelphia and Indianapolis, expecting the former to have a higher rate. Although the prevalence of smoking was available over several years (2010-2015), the participant did not specify whether the predicted relationship was expected to hold consistently over time, over an average of all years, or in the most recent year. Naturally, the mediator responded with a line chart containing two trend lines for each city. Nevertheless, the ambiguity made it difficult for the mediator to properly annotate the expectation onto the visualization; should she call out every year where the inequality was violated? Or would the participant tolerate a few inversions, as long as the slope of trend lines is consistent with the expectation? This sort of ambiguity might suggest a cognitive difficulty in connecting one's conceptual model to concrete data features.

★ *Design Implication:* It is possible to resolve some model ambiguities at the interface (e.g., via a selection menu showing attribute suggestions for ambiguous concepts [15]). More generally, it seems desirable to allow users to partially and

iteratively specify their model with feedback from the interface. However, if iterative model specification is enabled, the issue of multiple comparisons should be considered carefully; as multiple similar models are tested iteratively, the analyst runs a higher risk of making false discoveries [38, 58]. Tools should therefore track the number of model revisions or the number of data 'peeks', thereby allowing analysts to make reasonable model adjustments while calling out the potential for false discovery and overfitting.

External References and Mental Model Gaps

Participants routinely invoked factors that were external to the data in their models. For instance, one participant expected women's workforce participation to be negatively affected by laws that limit personal freedom. The latter measure was not present in the provided dataset, and thus could not be visualized. References to external factors were particularly common in expectations involving causality. This is not surprising, as causal reasoning often entails considerations that are not directly accessible to the analyst. However, the need to invoke external factors may present a design bottleneck for enabling a truly concept-driven sensemaking experience. If the user is not allowed to 'see' model-related attributes that are outside the immediate data at hand, many models could potentially go unvalidated or under-explored. That, in turn, can be frustrating to users or, at the very least, disruptive to their thought process; some of our participants quickly abandoned what could have been fruitful lines of inquiry upon encountering unavailable data.

★ *Design Implication:* A potential mitigating solution is to incorporate a 'data discovery' mechanism in sensemaking tools, allowing automatic search and importation of external data as needed. Such feature could help in validating models that invoke concepts external to the immediate data at hand.

In addition to referencing external factors, participants occasionally invoked references to non-specific quantities in their expectations. As template *V3* suggests (see Figure 4), participants expected that certain attributes will be simply "high" or "low". This can be taken to mean one of two things: the participant is expecting the attribute (often at certain geographies) to be higher or lower than average, or, alternatively, the participant had an idiosyncratic, possibly uncertain quantity in mind to compare against, but neglected to specify it concretely. We also observed ambiguous references to geographies. For instance, some participants referred to "developing countries" and "big cities".

★ *Design Implication:* In addition to interface elements that prompt users to manually resolve ambiguities in their model, certain references (e.g., to "big cities") could be resolved algorithmically by consulting appropriate ontologies. That said, we still do not fully understand the reasons behind the ambiguities in participants' model specifications. Although some omissions could be attributed to users being parsimonious in their specification, we suspect uncertainty in people's mental models to play a role in incomplete specifications. The latter presents a more difficult issue to address. Additional research is needed to understand the gaps and ambiguities in people's data expectations, and to devise strategies to resolve them.

7 LIMITATIONS & FUTURE WORK

Our study has several limitations that should be contextualized when interpreting its findings. First, although participants had experience in using a variety of data analysis tools, they did not have specialized knowledge in the datasets posed by the study. This may have impacted their ability to articulate expectations. In contrast, a domain expert will likely poses more sophisticated mental models, which may not be reflected in the current typology. Moreover, it is possible that the models we observed would be different, had we used different datasets. Replicating our study with domain experts examining their own data could provide richer insight into the kinds of models users may seek to externalize.

During the study, we observed a relatively large number of model validations (over three quarters of total queries). This suggests that concept-driven affordances would be frequently utilized, if introduced into tools. However, our study design could have inflated this number, given that participants were encouraged to both type and verbalize their expectations. The think-aloud setup enabled us to increase the fidelity of the analysis, by including models participants did not have the bandwidth to formally enter through the interface. That said, it is possible that actual users would be more deliberate when externalizing models and, hence, utilize model-testing affordances less frequently compared to our participants.

A third limitation stems from the type of visualizations used in the study. We opted for minimalistic, static visualizations so as not to bias participants to behave or interact in a certain way. However, compared to interactive representations, static plots offer limited potential for exploration. This might have impacted participants' ability to inquire further, especially when results contradicted their models.

Lastly, in analyzing the models expressed by participants, we abstracted away the subtleties of natural language, focusing instead on characterizing the conceptual relationships entailed. We also did not attempt to infer implied relationships, beyond what is being directly uttered or typed. Therefore, our model typology represents the results of a first-order, 'low-level' analysis. Nevertheless, we believe the typology represents a significant first step towards understanding the full complexity of models people draw upon when making sense of data. The typology will also guide our future work, as we seek to develop algorithmic techniques for parsing users' expectations and encoding them in visualizations.

8 CONCLUSIONS

We conducted an exploratory study to investigate how people might interact with a concept-driven visualization interface. Participants analyzed two datasets, asking questions in natural language and, optionally, providing their expectations. In return, they received customized data plots that visually depicted the fit of their mental models to the data. We found that, in the majority of queries, participants leveraged the opportunity to test their knowledge by outlining concrete data expectations. However, they rarely engaged in follow-up analyses to explore places where models and data disagreed. Via inductive analysis of the expectations observed, we contributed a first typology of commonly held data models. The typology provides a future basis for tools that respond to users' models and data expectations. Our findings also have implications for redesigning visual analytics tools to support hypothesis- and model-based reasoning, in addition to their traditional role in exploratory analysis.

ACKNOWLEDGMENTS

REFERENCES

[1] Gregor Aisch, Amanda Cox, and Kevin Quealy. 2008. You Draw It: How Family Income Predicts Children's College Chances. http://nyti.ms/1ezbuWY. (2008). [Online; accessed 25-August-2018].

[2] Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 111–117.

[3] Jillian Aurisano, Abhinav Kumar, Alberto Gonzales, Khairi Reda, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2015. "Show me data": Observational study of a conversational interface in visual data exploration. In *IEEE VIS posters*.

[4] World Bank. 2018. World Development Index. https://data.worldbank.org/indicator. (2018). [Online; accessed 25-August-2018].

[5] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2376–2385.

[6] Stuart Card, Jock Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[7] Remco Chang, Caroline Ziemkiewicz, Tera Marie Green, and William Ribarsky. 2009. Defining insight for visual analytics. *IEEE Computer Graphics and Applications* 29, 2 (2009), 14–17.

[8] Wei Chen. 2009. Understanding mental states in natural language. In *Proceedings of the Eighth International Conference on Computational Semantics*. Association for Computational Linguistics, 61–72.

[9] Big Cities Health Coalition. 2018. Health Inventory Data. http://www.bigcitieshealth.org. (2018). [Online; accessed 25-August-2018].

[10] Kedar Dhamdhere, Kevin S McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring data with conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 493–504.

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

[12] Kevin Dunbar. 1997. How scientists think: On-line creativity and conceptual change in science. *Creative thought: An investigation of conceptual structures and processes* 4 (1997).

[13] Kevin Dunbar. 2001. What scientific thinking reveals about the nature of cognition. *Designing for science: Implications from everyday, classroom, and professional settings* (2001), 115–140.

[14] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE TVCG* 18, 12 (2012), 2879–2888.

[15] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proc. ACM UIST'15*. ACM, 489–500.

[16] Michael S Glass and Martha W Evens. 2008. Extracting information from natural language input to an intelligent tutoring system. *Far Eastern Journal of Experimental and Theoretical Artificial Intelligence* 1, 2 (2008).

[17] David Gotz and Michelle X Zhou. 2009. Characterizing users' visual analytic activity for insight provenance. *Information Visualization* 8, 1 (2009), 42–55.

[18] David Gotz, Michelle X Zhou, and Vikram Aggarwal. 2006. Interactive visual synthesis of analytic knowledge. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*. IEEE, 51–58.

[19] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 943–952.

[20] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis. *Queue* 10, 2 (2012), 30.

[21] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*. Springer, 154–175.

[22] Steve Kelling, Wesley M Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard, and Giles Hooker. 2009. Data-intensive science: a new paradigm for biodiversity studies. *BioScience* 59, 7 (2009), 613–620.

[23] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proc CHI'17*. ACM, 1375–1386.

[24] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2018. Data Through Others' Eyes: The Impact of Visualizing Others' Expectations on Visualization Interpretation. *IEEE Transactions on Visualization & Computer Graphics* 1 (2018), 1–1.

[25] David Klahr and Kevin Dunbar. 1988. Dual space search during scientific reasoning. *Cognitive science* 12, 1 (1988), 1–48.

[26] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems* 4 (2006), 70–73.

[27] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems* 21, 5 (2006), 88–92.

[28] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data-frame theory of sensemaking. In *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*. New York, NY, USA: Lawrence Erlbaum, 113–155.

[29] Christian Lebiere, Peter Pirolli, Robert Thomson, Jaehyon Paik, Matthew Rutledge-Taylor, James Staszewski, and John R Anderson. 2013. A functional model of sensemaking in a neurocognitive architecture. *Computational intelligence and neuroscience* 2013 (2013), 5.

[30] Mihai Lintean, Vasile Rus, and Roger Azevedo. 2012. Automatic detection of student mental models based on natural language student input

during metacognitive skill training. *International Journal of Artificial Intelligence in Education* 21, 3 (2012), 169–190.

[31] Zhicheng Liu and Jeffrey Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization & Computer Graphics* 1 (2014), 1–1.

[32] Zhicheng Liu and John Stasko. 2010. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization & Computer Graphics* 6 (2010), 999–1008.

[33] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. 2007. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007).

[34] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[35] Chris North. 2006. Toward measuring visualization insight. *IEEE computer graphics and applications* 26, 3 (2006), 6–9.

[36] Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review* 4, 4 (1990), 251–277.

[37] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[38] Xiaoying Pu and Matthew Kay. 2018. The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics. In *Proceedings of the 2018 BELIV Workshop: Evaluation and Beyond – Methodological Approaches for Visualization.*

[39] Eric D Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. 2016. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 31–40.

[40] Khairi Reda, Andrew E. Johnson, Michael E. Papka, and Jason Leigh. 2015. Effects of Display Size and Resolution on User Behavior and Insight Acquisition in Visual Exploration. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2759–2768. https://doi.org/10.1145/2702123.2702406

[41] Khairi Reda, Andrew E Johnson, Michael E Papka, and Jason Leigh. 2016. Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization* 15, 4 (2016), 325–339.

[42] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2014. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1604–1613.

[43] Purvi Saraiya, Chris North, and Karen Duca. 2005. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456.

[44] Purvi Saraiya, Chris North, Vy Lam, and Karen A Duca. 2006. An insight-based longitudinal study of visual analytics. *IEEE Transactions*

on *Visualization and Computer Graphics* 12, 6 (2006), 1511–1522.

[45] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 365–377.

[46] Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*. IEEE, 336–343.

[47] Yedendra Babu Shrinivasan and Jarke J van Wijk. 2008. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1237–1246.

[48] Arjun Srinivasan and John Stasko. 2017. Natural language interfaces for data analysis with visualization: Considering what has and could be asked. In *Proceedings of EuroVis*, Vol. 17. 55–59.

[49] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[50] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 317–328.

[51] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17 (1994), 273–85.

[52] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*. Springer, 184–195.

[53] Susan Trickett and Gregory Trafton. 2007. A primer on verbal protocol analysis. In *Handbook of virtual environment training*, J. Cohn D. Schmorrow and Nicholson (Eds.). Chapter 18.

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

[55] Jarke J Van Wijk. 2005. The value of visualization. In *Visualization, 2005. VIS 05. IEEE*. IEEE, 79–86.

[56] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST).*

[57] Jagoda Walny, Bongshin Lee, Paul Johns, Nathalie Henry Riche, and Sheelagh Carpendale. 2012. Understanding pen and touch interaction for data exploration on interactive whiteboards. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2779–2788.

[58] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 479, 12 pages. https://doi.org/10.1145/3173574.3174053