The Interactive Visualization Gap in Initial Exploratory Data Analysis

Andrea Batch and Niklas Elmqvist, *Senior Member, IEEE*

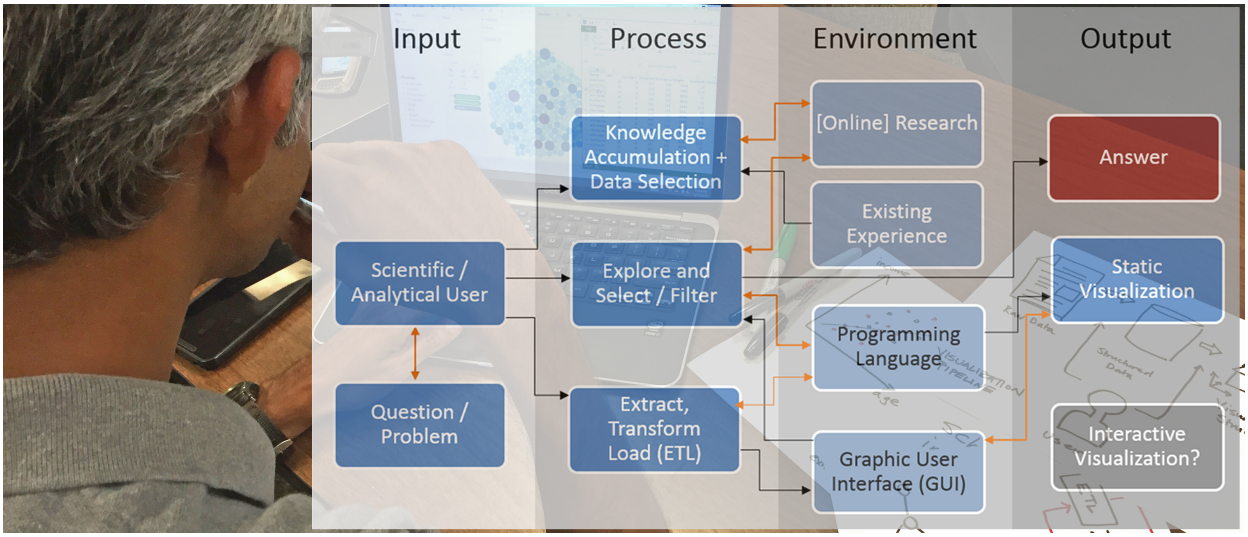

Figure 1. Elements of the exploratory data science workflow identified in this paper (background: analyst exploring a dataset).

Abstract—Data scientists and other analytic professionals often use interactive visualization in the dissemination phase at the end of a workflow during which findings are communicated to a wider audience. Visualization scientists, however, hold that interactive representation of data can also be used during exploratory analysis itself. Since the use of interactive visualization is optional rather than mandatory, this leaves a "visualization gap" during initial exploratory analysis that is the onus of visualization researchers to fill. In this paper, we explore areas where visualization would be beneficial in applied research by conducting a design study using a novel variation on contextual inquiry conducted with professional data analysts. Based on these interviews and experiments, we propose a set of interactive initial exploratory visualization guidelines which we believe will promote adoption by this type of user.

Index Terms—Data science, visualization, visual analytics, contextual inquiry, semi-structured interviews.

1 Introduction

The visualization field is at something of an impasse. On the one hand, rapid advances in the power, simplicity, and familiarity of visualization combined with an increasing awareness of the potential of visual communication have pushed the field to the cusp of mainstream breakthrough in society [7, 34]. Interactive treemaps, Sankey diagrams, and other complex visual representations nowadays routinely appear on the first page of the New York Times and the Washington Post as well as other newspapers [54], visualization-based companies such as Tableau, Spotfire, and Qlik are seeing wide market success, and even complex visualizations such as networks, trees, and streamgraphs are shared on social media platforms every day [8, 58]. On the other hand, all of this success is easily dwarfed by the tremendous victories of data science in the last decade, where big data has been likened to the "next frontier" of computing [35], where statisticians and data scientists are touted as the "sexiest careers" of the 21st Century [14], and where data-driven business intelligence spawned an industry overnight [36]. Because of the ubiquity of data in all walks of life, statisticians and data scientists have become hot commodities [43], essentially having their pick of employment in a wide range of organizations such as media conglomerates [48], political campaigns [61], and tech giants as well as startup companies.[1] Tapping into this "data science boom" would be a surefire way to bootstrap the visualization field to the next level, and the leap does not appear to be that large given that both disciplines deal with deriving insights from data.

Still, very few data science tools and workflows actually employ interactive visualization as anything more than a mere communication tool used to disseminate results at the end of an investigation to stakeholders and the general public [18] (Figure 1). As a result, data storytelling has become a major point of interest in big data [15, 17]. In fact, data science disciplines such as statistics, machine learning, and knowledge discovery have a vested interest in striving towards fully or semi-automated workflows [10, 26], since this leads to faster and thus cheaper as well as less ambiguous (albeit possibly flawed) results. In contrast, the premise of visualization and visual analytics is that the existence of a human in the sensemaking loop can significantly improve the outcome of an analysis. For this reason, visualization researchers and practitioners are convinced that typical data science workflows could be significantly improved—both in terms of speed as well as quality of the results—if visualization was adopted as part of the analytical process. The challenge for the visualization field is that the data science field is generally satisfied with the *status quo* and has little incentive to change. This represents a gap between data science

• *Both authors are with the University of Maryland, College Park, MD, USA. E-mail: {ajulca, elm}@umd.edu.*

Submitted to IEEE VAST 2017. Do not redistribute.

[1] http://www.kdnuggets.com/2016/04/best-data-science-blogs-companies-startups.html

and visualization that is unlikely to improve on its own. In other words, a philosophy of *"if you build it, they will come"* is not effective, yet is precisely what many visualization researchers are doing. Rather, the onus is on the field of visualization to build the bridges, engage in the conversation, and promote visualization to data scientists.

However, to begin this process, we need a better, actionable, and more nuanced understanding of how this "visualization gap" can be bridged from the direction of visualization research. In this paper, we try establish such an understanding through a variation of contextual inquiry [23, 60] involving a number of professional data scientists and analysts working at various U.S. federal agencies in Washington, D.C. Our inquiries were conducted in the actual workplace of each participant, involved data relevant to their everyday work, and focused on their current practice, workflow, and tools. Our multi-stage protocol first asked participants to describe their process in an effort to elicit a common data science workflow, and then engaged each participant in an hour-long extended data analysis session with a given problem set. The study concluded with a semi-structured interview where participants were asked to reflect on their analysis activity, discuss their use of visualization, and brainstorm on whether and how visualization could be made into a permanent fixture in their daily work. Each inquiry session lasted two hours per participant, which is uncommon in similar work in the literature given the highly protected and unavailable nature of professional analysts such as our participants.

We analyze the results from our contextual inquiry using qualitative methods and derive both expected and surprising findings. More specifically, our results confirm that visualization is primarily seen as a communication tool among professional analysts and that few of our participants ever use visual representations of their data in the middle of an analysis process. The reason stated for this is that visualization tools are generally seen as endpoints in the process in that they (a) are separate from the computational tools that data scientists typically use (R, Matlab, SPSS, JMP, among others), (b) require extensive data wrangling [28] to use, and (c) provide poor functionality for exporting insights, operations, and filters used in the visualization. Nevertheless, we note that our participants have a quite pragmatic view of the use of visual representations; visualization is just yet another tool, and they claim no intrinsic bias against its use if it provides clear utility. This represents a promising opportunity for the visualization field provided that our tools can be better integrated in data science workflows.

The rest of this paper is structured as follows: We begin by surveying the literature on data science, visualization, and the use of contextual inquiry in understanding current practice. We then describe our method for the inquiry involving professional data scientists. The following section describes our results in detail. We next discuss these results at length and identify actionable outcomes. We close the paper with our conclusion and plans for future work.

2 Background

Our work is a contextual inquiry into current practice of data scientists and analytic professionals with the purpose of eliciting needs, requirements, and plans for the use of visualization. Here we review the main topics of our work: the rise of big data, the current use of visualization in data science, and contextual inquiry for informing this usage.

2.1 The Rise of Big Data

As virtually every aspect of our world has become instrumented in the last few decades, there is an increasing prevalence of data being recorded about life, society, and the human condition [45]. This idea of "big data" has quickly become a major factor in the new millennium, to the point where it has been dubbed the "next frontier" (if not the last) [35] of computing, where its use has been suggested as a replacement for theory [2], and where it prompted the White House to launch a big data initiative in 2012 [41]. As a result, careers involving data analysis have skyrocketed in popularity [14], with companies in all industries hiring statisticians and data scientists to inform decisions.

However, such tremendous opportunities afforded by massive data are matched by equally magnificent challenges in terms of managing, transforming, and extracting insight from this data. For one thing, the models and algorithms that are used are often oversimplified and lack the necessary discriminating power to model the phenomena they claim to capture [30]. For another, even with large data volumes, there is a risk of overfitting models or basing them on uncorrelated phenomena; for example, with the since-debunked claims by Google in 2009 to provide early influenza warning based solely on search terms [20], the model was confounded by other seasonal patterns [31]. These, and other examples, serve as cautionary tales against putting blind trust into data as the primary and only grounds for decision making. Nevertheless, the data-driven society [45] is here to stay, data science is the engine behind it, and data scientists are its priesthood.

2.2 Data Science Workflows

Digital tools are critical to data science and analytics workflows, and current practice spans data analysis tools such as R[2], Pandas [37][3], and SAS[4]; data warehousing services such as MySQL[5], MongoDB[6], or Amazon Redshift[7]; and machine learning libraries such as scikit-learn [44][8] and Apache MLlib [39][9]. While there is no formally standardized workflow or process that fits every data scientist, and every professional tends to establish their own, a common process typically consists of the following general stages [3, 29]:

1. *Discovery:* Formulating an interesting question and determining the data necessary to answer it;
2. *Acquisition:* Locating, organizing, and preparing data so that it is accessible to the chosen analysis environment;[10]
3. *Exploration:* Investigating and analyzing the dataset in order to collect insights and understand the data;
4. *Modeling:* Building, fitting, and validating a model that can explain the dataset and the observed phenomena; and
5. *Communication:* Disseminating the results to stakeholders in reports, presentations, and charts.

2.3 Visualization in Data Science

Static visualization is commonly used in the communication phase of data science workflows, and data scientists sometimes use them as part of the analysis as well [21, 29]. For example, John Tukey's notion of exploratory data analysis [57] is firmly entwined with visual methods. However, interactive visualization is generally not standardized components of this workflow, and visualization beyond static line and bar charts is often relegated to the final communication phase of the workflow [18]. Nevertheless, tools such as Tableau [56], Spotfire [1], and ggplot2 [59] provide a wide variety of static visualization techniques in a format easily accessible and usable by data scientists.

Satyanarayan et al. [50] begin to address this by introducing a high-level grammar of graphics, "Vega-Lite," which presents a set of standardized linguistic rules for producing interactive information visualizations using a concise JSON format for data to be represented by the grammar. The creators of Vega-Lite do not themselves discuss implementing Vega-Lite directly in analytical environments, but their grammar of interactive graphics has been implemented in R via the "ggvis" package using the same—albeit slightly lower-level—tool that Vega-Lite is built around (namely, Vega [51]).

While examples of visualization researchers developing techniques using environments popular with data scientists do exist [46], they are not commonplace. It is generally considered a fundamental principle

[2] http://r-project.org/
[3] http://pandas.pydata.org/
[4] http://www.sas.com/
[5] http://www.mysql.com/
[6] http://www.mongodb.com/
[7] https://aws.amazon.com/redshift/
[8] http://scikit-learn.org/
[9] https://spark.apache.org/mllib/
[10] Also often called "ETL," meaning "extract, transform, and load."

of effective interface design to directly build implementations within the environments most often used for analysis by researchers [40, 55]. Furthermore, such implementations should fit well within that environment [40, 55]; we argue that this is as true of the more technically-demanding programming and scripting environments as it is in higher level, "point-and-click" interfaces.

While there are few studies which explicitly investigate visualization amongst data scientists, there are several related studies which, while they do not necessarily refer to the field as "data science," are closely related enough to be considered within the same domain as this study. An investigation into existing case studies of visual parameter space analysis—the use of visual analysis in structured data input-output workflows—by Sedlmair, et al. [52] identified three primary visualization research gaps: data acquisition, data analysis, and cognition (facilitation of understanding). To explore the visual analytics needs of constraint programmers (i.e., developers who use solutions and models which satisfy many constraints simultaneously), Goodwin et al. [21] discuss their findings from group workshop sessions and online surveys regarding the requirements for visualization. Their results, much like our our own, indicate that there is indeed a strong need for more visualization tools in the modeling, discovery, and decision-making processes of their participants' workflows.

2.4 Contextual Design

In their seminal paper, Wixon et al. introduce "contextual design" as a systems development method in which the researcher partners with the user at the user's place of work to "develop a shared understanding" of the user's activities, and they define *contextual inquiry* as the first part of the broader process [60]. Specifically, contextual inquiry is the data collection step of the field research element of the contextual design method, and it emphasizes four essential principles: (1) the *context* of the activity being performed by the user, (2) the *partnership* between the researcher and the participant, (3) the spoken verification that the investigator's *interpretation* of the activity matches the user's, and (4) the *focus* of the study as central to the approach taken by the interviewer [5, 23]. The most typical application of contextual inquiry is in the form of a *contextual interview*, which begins in the user's actual work environment as a traditional interview regarding the user's recollections of their work activities, and, within fifteen minutes, is transitioned o an activity in which the participant conducts their work while the researcher watches and takes a participatory role by sharing and summarizing their understanding of the user's work [5, 23].

The remaining stages of the contextual design method include the *interpretation* of field research data, the *consolidation* of the data based on a *model* to build *user personas* and an affinity diagram grouping notes and imagery gathered during the interviews into categories, *ideation* workshops in which teams determine which elements of the affinity diagram should be synthesized into a vision for a product, and finally, detailed design [23]. Cooper originates the concept of the "persona," as it is used in contextual design, as a generalized representation of the user—an archetype—i.e., based on users' behavioral data [12, 24].

2.5 Contextual Inquiry and Data Scientists

While alternatives to the qualitative methods for developing personas may be applicable in certain cases (e.g., where mouse events are the most notable method for human-computer interaction) [62], this approach is more difficult to apply in design for the sciences beyond simply categorizing event sequence structures [32]. Field survey methods are still a popular approach for determining the direction of design targeting scientific users [33, 47], including data scientists [4, 29].

Kandel et al. [29] conducted what might be considered a contextual interview study similar to our own in that they analyze data scientists' self-reported work processes, and attempted to interview participants at their place of work in as many cases as possible. They propose three main archetypes that data scientists may be classed into: *Hackers*, who build processes chaining together multiple programming languages of different types (analytical, scripting, and database languages, for example) and who use visualization in a variety of environments; *Scripters*, who perform most of their analysis in an analytical environment (e.g., R) and perform the most complex statistical modeling of the types but who do not perform their own ETL; and *Application Users* who performed most or all of their work in an application such as Excel or SPSS and, like Scripters, relied on others (namely, their organizations' IT departments) for ETL. The appropriateness of contextual inquiry for analytical professions in more contemporary research is further evidenced by the recent, complete contextual design study of data scientists [42] conducted by IBM, a notable employer of data scientists.

3 Study Design Rationale

We adopt a variant of contextual inquiry (described in greater detail at the beginning of Section 4) to observe analysts in their work environment. For this study, this method required access to U.S. federal government facilities. This approach comes with two major constraints: our participants are professionals skilled in a highly-demanded (and demanding) set of talents who would need to be willing to sacrifice at least two hours of their time, and their organizations have protocols regarding facility entry and the information that employees can and cannot share with the public. We believe that this latter point would be true of many, if not most, organizations which employ data scientists and similar analysts. In this sense, our participants are members of a demographic that is often "protected" from engaging in this sort of research except in ways that are internally managed and motivated primarily by the mission of their employers.

We were able to gain unfettered access to participants' facilities, but still needed to address the issue of making the most of the time we have with a group of people who often only have time to do their job (and not engage in additional activities that do not directly contribute to their job). To achieve this, we constructed an activity to distill their workflows, as captured by artifacts of their routine procedures, into problem sets that would mimic the early stages of their analytical process but, could be captured within the small blocks of time that we had with them. We created problems that required them to use real data of their choosing to answer a question from a list populated based on the kinds of issues they typically addressed on the job, and gave them a comparatively large number of options so that our sessions, much like their work, would be semi-structured but fairly open-ended.

4 Method

We conducted our study as a contextual inquiry [23], where we first interviewed participants to establish their everyday work practice. However, our study deviated slightly from standard contextual inquiry protocols in that we then asked participants to solve specific problems that we provided (instead of using their own datasets). These problems were based on (1) *artifacts* used throughout the participants' work process, including code, databases, spreadsheets, methods documentation, and checklists; (2) on our prior knowledge of data science workflows; and (3) on user feedback gathered during beta testing of an R library developed to aid in the extract, transform, and load (ETL) processing of data from a major producer of economic statistical indicators.

Our motivation for the modification was that we already have a reasonable understanding of current data science practice (e.g., as described by Anderson [3] and Kandel et al. [29]), the practices of our participants based on their organizational artifacts and their feedback, and we were more interested in directing participants towards specific tasks to elicit a better understanding of the initial exploratory stages of the data analysis process. We believe that inferences about these stages would be difficult to make if participants were instead asked only to walk through routine data product maintenance procedures or to give a verbal explanation of already completed projects. By controlling the tasks and problems to work on, we hoped to eliminate some of the wide variation in tools and approaches that individual analysts may exhibit.

4.1 Participants

We recruited eight data scientists and economists from several federal agencies in Washington, D.C., USA to participate in our experiment. Five of the participants were male and three were female, their ages ranged from 26 to 50 (mean age: 35.5), and they all had normal or corrected-to-normal vision (self-reported). Six participants had

earned masters degrees in quantitative fields, one had started—but not finished—a Ph.D. program in economics, and the remaining participant was in the process of earning a masters degree in economics. The participants' experience in their fields ranged from 4 years to 20 years (self-reported). Participants were screened to be experts in data analysis; all participants reported routinely using data management and analysis operations in their daily work and had several years of experience working on this type of duties. Four of these participants had developed or contributed to the development of interactive data visualization projects.

Once screened, participants self-selected in response to emailed requests for their involvement in our study. The self-selection and small sample size must be acknowledged as a limitation to how representative this study may be, but is not uncommon in field studies involving the entry of researchers into the personal or professional environments of the participants [11, 25, 38]. Similarly, the sample was selected based on their employment with, and roles within, federal agencies, which must also taken into consideration with respect generalizing based on our results.

4.2 Apparatus and Locale

All inquiry sessions were conducted in the workplace of the participant and using their everyday computing environment to ensure their familiarity and comfort during the study. The exact computing platform, hardware setup, and data analysis software thus varied significantly between participants. Because of this difference, screen recording tools varied across two organizations; one organization had a preexisting screen recording utility and security settings prevented the use of external screen recording software, and the other participants used a free screen recording application. All participants used pencils and paper provided by the researchers for the sketching activity.

4.3 Procedure

A single inquiry session consisted of the study administrator arriving at the participant's workplace, collecting informed consent, and then giving a brief background of the study. Significantly, at **no time**—either in recruitment or during the introduction of the session—did the administrator mention the visualization theme of our study. The reason for this omission was to avoid priming and potentially biasing participants with regards to their use of visualization. The rest of the study then consisted of four primary steps:

1. A preliminary interview regarding the participant's work processes and tools used in their work (10 to 15 minutes);
2. A data analysis activity designed to mimic a standard data science workflow [3] (approximately 1 hour);
3. A formative design activity during which the participants were asked to sketch visualizations appropriate to tasks in the preceding analysis activity (20 to 30 minutes); and
4. A final semi-structured interview on visualization in the context of the participant's workflow (10 to 15 minutes).

Each session lasted approximately two hours. After finishing a session, the administrator summarized the participant's findings, asked for clarifications or corrections, and answered any remaining questions.

4.4 Problem Set

Each participant was asked to pick one of the four questions below to answer using real, public data by the end of the Stage 2 within one hour of making their selection (see the Appendix for more details):

1. "How has the rate of a specific type of crime changed over the last few years?"
    - **Optional:** "What might be causing this change?"
2. "Tell me something interesting about the careers or personal finances (e.g., income, spending habits, or employment) of a particular group of people compared to (an)other group(s)."
    - **Optional:** "Suggest an explanation for your observations."
3. "When and where has a number of major catastrophic events occurred? Do they share anything in common with events you didn't expect to exhibit similar characteristics"
    - **Optional 1:** "How frequently and how long after the fact did people talk about/reported on these events?"
    - **Optional 2:** "What was the weather like in the area of the event before and afterward?"
4. "What's been going on with gasoline for the past few decades? Tell me as many things about it as you can."

As noted at the beginning of the methods section, these questions were based mainly on artifacts used throughout the participants' work process (code commentary, spreadsheet notes, process documentation, and so on). Questions were made fairly open-ended so that analysts could use their experience to not only determine how they would answer it, but also to decide what constitutes a satisfactory solution.

4.5 Data Collection and Analysis

Participant voices and on-screen activities were recorded during each session, and some participants drew sketches which were retained by the researchers. Furthermore, the test administrator took extensive notes of observations as well as discussions with the participants during the session. These transcripts and notes form the primary data collected from the study.

We followed a basic qualitative interview analysis method when extracting insights from these transcripts. We first listened through the audio recordings in their entirety to form a general understanding of the themes and topics of the discussion. We then used the interviews to start coding these themes and topics. While we did not use a formal Grounded Theory approach, we did apply an open-coding scheme and regularly stopped to calibrate and merge codes as needed.

5 Results

We report our results for each of the four different stages of the evaluation: (1) preliminary interview, (2) data analysis using a problem set, (3) formative sketching, and (4) final post-experiment interview.

5.1 Stage 1: Pre-experiment Interview

With one exception, all participants described their work procedures to largely occur within the context of existing information systems and data structures.

5.1.1 Self-Reported Workflows

The work processes reported by all participants began at the point understanding the problem or issue they were addressing in their analyses. Participants all moved on to describing the sources of their data, and all participants described a central component to their work being to join or infer relationships between series across different data stores. Three participants noted that the most frustrating part of their work process is often these first two stages when it required communication with data providers. In describing the methods used, all analysts described a need to extract data from an external source and transform it for use with statistical programming languages (R, FAME, and Python).

Participants described using models of varying complexity in their typical work process; most notably, they mentioned statistical language processing and other information matching and retrieval methods, as well as hierarchical and relational structures. Three participants reported the end of their workflow as generally being the communication of their findings, with the remainder reporting archival as the final stage. Five participants reported recent work projects ending in the completion and deployment of tools for data manipulation or analysis; the remaining three conducted their analyses using existing tools.

5.1.2 Work Focus

All participants had recently (within the last year) conducted independent analytical or development projects for which they were the lead or sole contributor. One participant described his work as consisting of running projects that primarily start from scratch. This participant recently developed a search method for large, unstructured, and highly technical text data that had been accruing for roughly forty years.

The four remaining participants reported that the primary focus of their work was in the context of an existing information system. Three of these had made lasting and substantive methods contributions to the body of data science or analytical systems within their current agencies: one had built a user interface for querying agency databases; another had restructured a complex, hierarchical data structure; the third had constructed a revision analysis tool referencing a node aggregation structure.

5.1.3 Self-Reported Tool Use: Revisiting Kandel's Archetypes

In some ways, the results from the study by Kandel et al. [29] are similar to ours (e.g., finding appropriate data, ETL, and integrating datasets from several sources took up a large share of many of the analysts' time). However, in contrast to the findings that lead them to propose their three archetypes, interview question responses from the participants in our study indicate that they invariably straddled the "Hacker" and "Scripter" role; not one of them relied on others within their organization for data ETL (although some reported receiving data from external providers under contract as part of a wider process that involved conducting their own ETL). Perhaps even more importantly, all of our respondents reported performing the bulk of their analyses in a scripting or analytical language and had used multiple languages on the job. This difference may, admittedly, be a result of our small sample size, but it may also be an indicator that their third archetype, the "Application User," has become passé in analytical professions. Alternatively, it may mean that we have not yet reached a tool maturity where this archetype can become dominant.

In our study, one participant reported mainly using Python, and noted that the SciPy, NumPy, multiprocessing, and glob libraries were essential for recent work, but that a number of additional libraries made their work easier, with the "ujson" library being among their most favored. This participant also made a note of recent work made use of the Python interface for the Stanford Network Analysis Project (SNAP). Four participants reported using R, but only two of these reported using it regularly on the job. Four participants reported developing interactive visualizations using Plot.ly, Leaflet, and D^3 [6], among other tools, at least once in the past. Three of these also reported using JavaScript/HTML/CSS infrequently on the job to communicate output from statistical models to colleagues. These same three participants further reported having used Python, but this was mainly used for personal projects (e.g., combining the use of an API of a financial newspaper, a string pattern recognition algorithm, and a text-to-speech function in order to find and produce audio summaries of news related to their interests which they could no longer find the time to read through manually). Five participants reported used Excel and the time-series database and programming environment FAME ("Forecasting Analysis and Modeling Environment") as the primary environment for analysis on the job.[11] For all of these participants, FAME was described as the environment used most heavily for analysis, whereas Excel was described as being used mainly for the purpose of viewing data and communicating analysis results to others.

5.2 Stage 2: Problem Set

Of the eight participants, two partly answered the question asked in the problem set to their own satisfaction, and the remaining six participants fully answered the question. In all cases, the main stage that participants found impediments to their progress was in the "Discovery" stage. Interactive visualization was not implemented at any stage of the problem set activity, but static visualization was used by a majority of participants (Table 5.2).

[11] FAME is a time-series database with many easily accessible APIs and a domain-specific programming language.

Table 1. Participant time use and static visualization rate by task types. Participants spent by far the most time in discovering the appropriate dataset to use in answering their selected question. "Static Visualization Rate" in this context refers to the percentage of participations who created static visualizations during their activity.

Task	Average Time	Static Visualization Rate
Discovery	37 minutes	50.0%
Data ETL	9 minutes	0.0%
Exploration	14 minutes	62.5%

Several participants used interactive visualizations built by others regarding the data they were considering using to answer the problem. We also observed that all participants using programming environments either received syntax error messages or had minor difficulties reshaping the data which required minutes to resolve.

5.2.1 Summary of Tools and Visualizations Used

During the activity, one participant used Python without an IDE, three participants used R in RStudio, and five used Excel. For direct manipulation and analysis of the data, three participants *only* used Excel, and two participants *only* used R in RStudio. Of the participants who stated during the interview section that their primary analytical environment was FAME, if any visualization was produced during their session, both the visualization and the analysis itself were done using Excel. None of the participants in this study used any visualization tools outside of those built into their analytical environments. All participants used the "look at the data" (or "show me the numbers" [16]) approach as primary means of verifying the relevance and completeness of the data prior to communication stage (i.e., looking at the data in whatever format it was stored). The two most experienced users in this study did not use visualization at any stage of the problem set.

5.2.2 Discovery

The discovery stage was by far the most time-intensive activity for all participants during the approximately 1-hour-long problem set activity, taking participants on average **37 minutes** to complete. Of this time spent in discovery,

- An average of approximately 22 minutes was spent *reading reference material* (excluding metadata) to find potential causal factors, and to explore statistical methods including syntactical options within analytical environments. The participants referred to a combination of news, academic, and data science blog articles to assist with this stage of their process. Three participants mainly referenced articles, two of whom read online tutorials (e.g., R cookbook), StackOverflow, and R help documentation; of these, one also referred to API documentation and metadata, and the other participant mainly referenced financial news, academic articles, and statistical reports from government agencies. The third of these participants mainly referenced popular press articles and data science blog posts. Two participants made a point of referring to visualizations produced by others in their readings.

- An average of approximately 15.25 minutes was spent *referencing site or API metadata and conducting searches* as a means to find the location of the correct data. One participant spent the large majority of the discovery stage searching and exploring site metadata, and virtually no time reviewing other reference material. No visual representation of the reference metadata was referenced or created by any of the participants.

All participants exclusively selected government data; one used local government data for crime statistics, while all others used federal government data.

5.2.3 Acquisition and Transformation

None of the participants used visualization during this stage. The average amount of time spent on data acquisition (ETL) was approximately **9 minutes**.

- *Data extraction and loading* took, on average, approximately 2.25 minutes, which was skewed upward by a participant who needed to extract several large datasets from a site, and skewed downward by a participant who extracted the data using an API request that took only the amount of time required to write the request function (approximately 10 seconds). One participant used a REST API, and the remaining three exclusively used site download tools.
- Once it was loaded into the analytical environment, *transforming* the data to prepare it for modeling took slightly longer for participants across all environments, taking an average of approximately 7.75 minutes. This process was lengthier in cases where the structure of the source data being used in the model was more complex, and in cases where the data was being manipulated using a programming language, and was skewed downward where Excel was used with minimal transformation.

5.2.4 Exploration, Modeling, and Communication

This process took, on average, approximately **14 minutes**. The most complex model attempted was a basic linear regression model. One participant attempted a categorical parent/child aggregation hierarchy, but was unable to finish the analysis. The participant using this hierarchy did not use visualization at any stage of the problem set activity. Of the remaining participants, one examined a cross-section of ratios across geographic categories; this participant produced a column chart comparing public sector employment rates against private sector employment rates by state using ggplot2 (Figure 2). This participant also expressed a desire to create a grid of faceted bar charts (also using ggplot2), but decided against it because it would take too long.

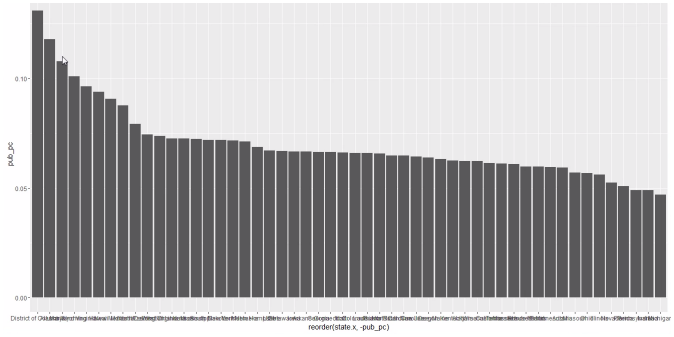

Figure 2. One user produced a column chart with U.S. Census Bureau data in RStudio using the ggplot2 library

One participant examined the rate of change of two potentially related time series with different units of measurement, and produced a line chart comparing the series scaled to different axes to explore the potentially causal relationship. This participant was the only one who used a chart to inform the later stages of analysis, first charting one series and using that information to search the time period of interest, and was the only participant to perform comparative data analysis. The remaining participant examined the rate of change in a single series and produced a line chart representation of the series. All charts used or produced during this activity were static. All participants who used visualization for exploration used the same charts as part of the communication of their findings.

5.3 Stage 3: Sketching

As in other studies [9], we opted to a sketching activity to allow for the creation of visualization in instances which may otherwise have been constrained by either technological barriers or the time limitations of our interview sessions. The most common theme in participant sketches of potentially helpful visualizations during this stage was that most participants viewed a table as the *most* beneficial visual aid. Only four of them drew a chart, and in one of these cases, it was mainly as an afterthought. All participants focused on the work involved in data discovery as the most difficult element of the activity, including participants who were already familiar with the source of the data they selected. All participants were most strongly interested in methods for multistage search-and-filter interface design; all participants included either drop-down menus or search bars (or both) in their sketches. Three participants also included tables in their sketches; two of these sketches contained lists of potential data sources, the third contained the data itself (Figure 3). One participant expressed interest in a related-data search and discovery tool inside the RStudio IDE.

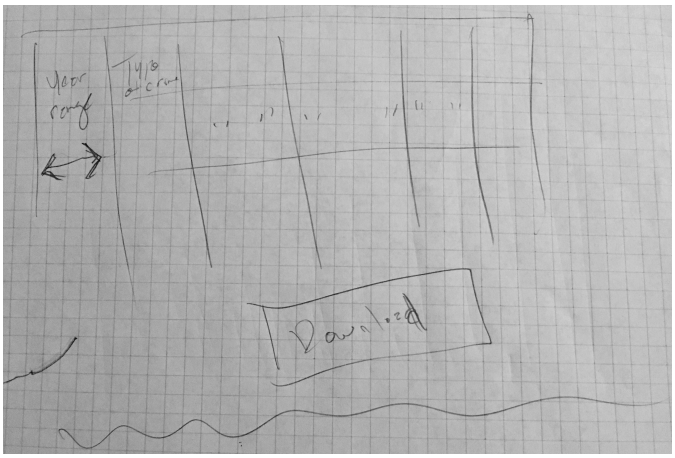

Figure 3. The fifth participant simply sketched a data table and, not without sarcasm, added a “download” button.

Of the participants whose sketches extended beyond search-and-filter methods for data discovery, one drew a bar chart representation of a hierarchical time series and expressed an interest in better illustrating the hierarchy. Another participant expressed a desire to represent autoregression models of the series used during the problem set activity, and noted that it would have been easier for them to do using Stata. A third participant, who we consider to have the most experience in developing interactive visualizations within the study cohort, incorporated interactive elements within his sketch as a small window which appears on mouse-over (i.e., a tooltip) with details regarding data linked to a visual object within the view (Figure 4).

5.4 Stage 4: Post-experiment Interview

When asked about reasons for not using visualization, three recurring themes arose during the post-experiment interviews: (1) Visualization was too time-consuming to be worth their effort, (2) numeric data provided more detail in many instances than visualization could, and (3) visualization was just not needed.

5.4.1 Not Enough Time

Five of the eight participants stated that visualization is important, but that they did not have time to do it often. One participant said that only one of their projects, not a routine part of their work process, involved visualization in order to check the accuracy of predictive models. This participant said that building visualization into their typical workflow was difficult due to time constraints. When asked about the tools they typically use for visualization, they responded that use of Excel was most common, but that they have used Stata, Eviews, and R for visualization as well in their free time or as a student. Regarding ggplot2, one participant remarked: ”The syntax just doesn’t feel right[...] to come up with one beautiful graph, if I put it in a nice block format, it would be like fifteen additional lines. To me, that

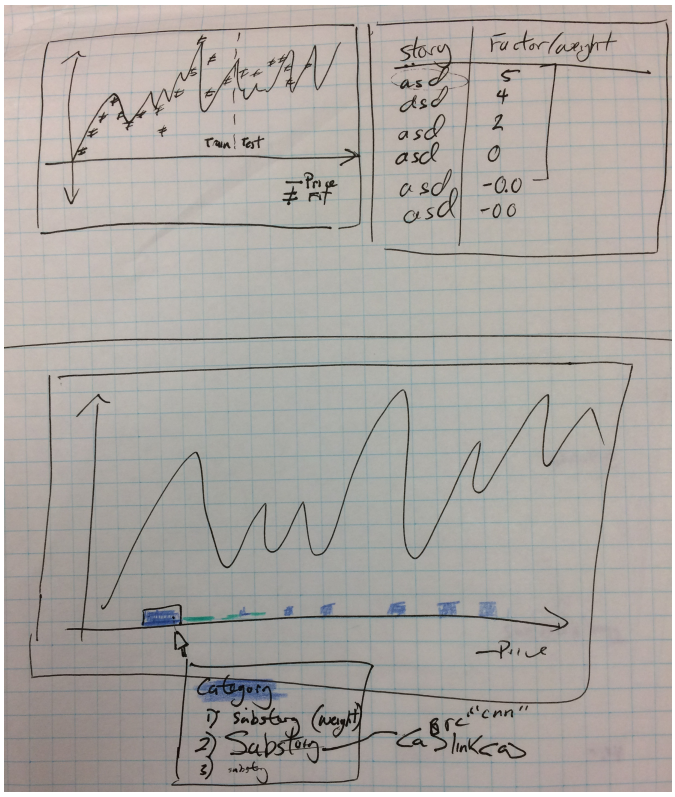

Figure 4. Interactivity appeared only once in our study, in a sketch; this indicates that the desire to build interactive views is present within the data science community, but the costs of using the tools outweigh the need during initial exploration.

seems superfluous. I also don't like this syntax—using 'plus' signs between each line. R's syntax is more functional—traditional functions have commas, all within the same parens; I understand that maybe the philosophy is that you have to be explicit about [features...] but that seems like overkill." We found this emblematic of the guidelines we propose: It is not enough to build tools for interactive visualization, or even to port them to the researcher's environment—we must also make it syntactically familiar, concise, and convenient to use within that environment.

5.4.2 Show me the Numbers!

One participant said that they occasionally use a line graph to track rates of change, but that they typically just look at the numeric representation of a time series when checking for volatility or revisions, as they find it clearer and more accurate than the line chart. This participant noted, however, that representing thresholds or other important characteristics by changing the color of the number or background was helpful.

5.4.3 Visualization is Unnecessary

Five participants noted that the data was straightforward enough that there was not a strong need to visualize it, and one of these, along with one other participant, noted that familiarity with the conceptual context of the data coupled with a quick examination of the numeric data was sufficient for their purposes. One data scientist stated that they virtually never used visualization except to communicate their findings with others, and during the post-activity interview, noted that the exception to this was in cases where data was either structurally complex (e.g., representing networks), or when it was intrinsically spatial.

6 Implications for Design

To our knowledge, this is one of the first studies using contextual inquiry specifically into the ways professional data scientists and analysts use interactive visualization and making suggestions on strategies for closing the gap between visualization and data science. In the discussion below, we first attempt to explain our results. We then discuss how they generalize to different populations, problems, and technology. Finally, we build on our findings to derive a set of action items for visualization researchers and practitioners to focus on for the purpose of bridging the gap between our field and general data science.

6.1 Explaining the Results

Experience played a role in our findings: more experienced participants were less likely to use visualization, and more likely to work with complex models. This, and respondents' remarks during the interview section, point to the primary impetus that drives data scientists to use visualization during their exploration and analysis stages is for sensemaking—the process by which the analyst applies their knowledge and understanding to interpret meaning from data [19, 22, 27, 49]. Confidence in their sensemaking ability based on familiarity with standard methods, models, or structures may present a disincentive for analytical professionals to set aside time to build visual representations of data prior to sharing their informed perspective with others.

Furthermore, our methods placed artificial time constraints on the participants, and visualization is generally sufficiently time-consuming that it may discourage participants from doing it. This is particularly true under the assumption that an experienced analytical professional may not consider visualization to be worth the trouble under the most leisurely conditions, as several of the participants noted. Since data scientists generally work under high time pressure in virtually all their projects, it is not surprising that components that are seen as non-essential receive little attention.

This may also point to an underlying explanation: Human nature. Given the complexity of the tasks involved in data science and exploratory data analysis, data scientists and analysts will always be looking for automated solutions to the challenges they face on a daily basis. Interactive visualizations are, by definition, not automatic, and their use does not generally lead to reusable solutions that can be easily automated. This may help explain why more experienced analysts tend to steer clear of such tools, or use them exclusively for rare deep dives into new and unknown datasets or problems.

6.2 Generalizations and Limitations

Our study is not intended to be representative of all data scientists, so we must be careful about how our findings can be generalized. While our participants were all professionals who engaged in data-driven analysis on a daily basis, we were only able to recruit eight individuals to our study. However, data scientists are generally a protected population that are difficult to engage in studies such as ours. In other words, to our knowledge, our work is the first extended contextual inquiry to study this particular population of information professionals for the express purpose of understanding when and how they use interactive visualization, particularly in their initial exploratory analysis. For this reason, our findings provide at least an initial understanding of this type of visualization for data science from a human-computer interaction and visual analytics perspective.

Furthermore, all of our participants were employees of the U.S. federal government, which may also bias the type and scale of analysis projects they perform. It is possible that data analysts from industry, or even from outside the United States, may have a different outlook, process, or dataset scale and type. This may have an impact on how widely our results can be applied. However, from our informal discussion with our participants, we are under the impression that the data science process—while far from standardized—looks similar across both government and industry. Our participants were all well-versed in the tools and software that data scientists use, and did not appear to be artificially constrained—in terms of budget, philosophy, or expertise—by the government agencies they worked for. As for scale, U.S. federal agencies remain one of the top clients for big data [41].

We did not anticipate that the majority of the second activity would, for all participants, be spent searching and reading—either about the data or about methods. In other words, participants spent a large portion

of the sensemaking process in the early discovery phase even before the data was extracted, transformed, and loaded, and from participant responses, it is likely that they formed much of their intuitions about the data during this early stage. That much of data analysis is spent diagnosing, cleaning, and transforming a dataset prior to starting the actual analysis process has already been recognized as a major challenge [28]. In fact, Dasu and Johnson [13] estimate that up to 80% of the development time is spent on data cleaning. However, an interesting secondary finding from our results is that some of the sensemaking may already be happening during this discovery stage.

Another corollary from our study is that data scientists' actual work processes have left them, as users, to sit at a desk using a keyboard and mouse to navigate largely GUI-free lines of characters, both in discovering external data and for the purposes of syntactical error management. While such command-line interfaces are often powerful and effective for expert users, they make integration with interactive visual representations challenging. For example, tools such as R and Matlab do provide dedicated rendering systems to produce visualization windows, but these are more or less static and do not let the user interact with them in a meaningful way. RStudio extends R with, among other features, a viewer which interprets HTML/CSS/Javascript, and a tabular view panel for data.frame class objects.

6.3 Closing the Interactive Visualization Gap

As visualization researchers ourselves, we are interested in finding actionable and direct measures that we can take to close this gap between visualization and data science. Based on our conversations with the data scientists involved in our contextual inquiry, we can now outline a few such measures that the visualization field should focus on:

- For visualization scientists collaborating with data scientists, **use the same programming environments and syntax that they do and build visualization elements into "data discovery" libraries**, creating or tying together data ETL tools that can be used in a non-interruptive step within the analytical environment to facilitate sensemaking. Sensemaking is often described as a cognitive skill requiring human intervention [19, 22, 27, 49], and libraries within statistical environments are nothing if not artifacts of data scientists' efforts to simplify that process for their peers.

- **Conduct user experience (UX) design sessions with data scientists** to investigate ways to soothe the frustration of errors and data foraging. All of our participating data scientists noted that the user experience of their most commonly used tools left much to be desired. Unfortunately, given their small population size and because of the haphazard and highly personal data science process, not enough attention has been spent on this topic.

- **The verdict on data tables: Not bad**. Participants of this study gravitated toward the data table format as their visual representation of choice, and every single participant viewed the data in a tabular format. Those using Excel, which links chart creation with table views, were able to more quickly and successfully visualize their data; however, many of these users expressed a degree of embarrassment at resorting to Excel. Those using R or Python either did not attempt to visualize, or found the syntax to be inconvenient. Bridging visualization and data science may require visualization researchers spending more time on augmenting basic representations such as tables with additional functionality rather than designing entirely novel visual representations.

- **Design self-contained, visualization components** that can integrate into the command-line interfaces that data scientists routinely use while still allowing for full-fledged interaction (zooming and panning, filtering, details-on-demand, etc) [53]. The syntax of calling the components must match that of the target environment; for instance, calling visualizations using single-line functions with parenthetical variables and specifications was a feature more than one of our respondents mentioned finding desirable. Furthermore, these visualization components should be first-class members of the analytical process so that actions and transformations interactively performed in the component can be exported and passed on to the next component in the sequence.

- **Education, not evangelization** is what is primarily needed to improve visualization adoption within data science, including providing easily accessible galleries of useful visualization techniques based on data type and tasks, giving examples of best practices, and finding allies within the data science community who can evangelize on our behalf.

7 Conclusion and Future Work

We have presented results from a contextual inquiry of current visualization practice for a collection of data scientists and analytics experts from several U.S. federal agencies. Our results highlight the quandary of visualization in professional data science: visualizations—including static visualizations—are rarely seen as obligatory or even useful components of the initial analytical process, and are instead relegated to the final checking and dissemination stages of the process. In other words, a dynamic visual representation is considered a good tool for communicating results with a lay audience, but is not considered vital when trying to understand which results to communicate in the first place. This means that visualization still has a long way to go in order to fully capitalize on the data deluge that our society has come under.

As visualization researchers ourselves, we see this work as a call to action for closing the gap between visualization and data science. Our action items suggest many possible venues for future work: better integration of dynamic visualization functionality into the very tools that data scientists are already using, improving the provenance and output filters of visualization so that they can become components in the overall tool ecosystem, and creating educational material to help data scientists select the right visualizations depending on their data and problem.

Acknowledgments

This work was partially supported by the National Socio-Environmental Synthesis Center (SESYNC) through a grant from the U.S. National Science Foundation award #1052875. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

Appendix: Semi-Structured Interview Questions

The questions below are examples of prompts which we used to guide the interview in order to determine more detailed areas of discussion.

More targeted questions were prepared based on participants' areas of specialization. The direction of the conversation and participant task performance may lead to additional questions. However, all questions were specifically related to participants' work procedures, tasks performed during the experiment, and feedback in the scope of their expertise.

Pre-Experiment Questions

These questions were asked at the beginning of the session. To avoid skewing the respondents' activities, we did not, at any point, reference visualization (interactive or static) at this stage unless the participant brought it up.

1. "Tell me about two or three interesting data analyses you've done recently."
    - *Motivation:* To engage the participant and get a sense of the inputs, outputs, and methods of their workflows.
    - *Follow-up:* "What data did you use? How did you get it? How did you transform it before doing analysis? What did your final analysis look like?"
2. "Could you briefly describe the work practices involved in your general routine and/or recent projects?"

- *Motivation:* To identify existing practices.
- *Follow-up:* "Is there a particular direction you foresee/want your work to take in the future?"

3. "Can you describe points in your work processes where gaining a complete understanding of the data is the most difficult?"

   - *Motivation:* To improve our identification of the areas where visualization may be most beneficial in data science workflows.
   - *Follow-up:* "What makes insight so difficult at these points in your process?"

4. "What sorts of tools do you and your colleagues most commonly use in your work?"

   - *Motivation:* To identify the characteristics, skills, and preferences of study participants.
   - *Follow-up:* "Please describe how and why those tools are be used. Is there a reason alternatives are not used?"

Post-Experiment Questions

The questions below were asked at the end of the session to assess the participants' perspective on the activities and to inquire about the use of visualization.

1. "Did you feel that tasks you perform in the early stages of your analytical process in your field were, at a very general level, reflected in the problem set in the first part of the experiment?"

   - *Motivation:* To verify the relevance of activity to the participants' real workflows.
   - *Follow-up:* "Would you say that the same is generally true for your colleagues? If not, what is different about their work practices?"

2. "What parts of your work process, if any, usually involve data visualization?"

   - *Motivation:* To improve our identification of existing practices around visualization in data science workflows.
   - *Follow-up:* "What makes visualization more important in those parts of the process?"

3. "In the second section of the activities (with the sketches), were there types of diagrams which you would like to use in your own workflow, but have not found an appropriate tool for the task? If so, please describe the stages of your work process, and the diagrams you already use, if any."

   - *Motivation:* To identify technical and design barriers to visualization.
   - *Follow-up:* "Which part of constructing your own charts do you find most difficult? Least difficult?"

References

[1] C. Ahlberg. Spotfire: an information exploration environment. *SIGMOD Record*, 25(4):25–29, Dec. 1996.

[2] C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, June 2008.

[3] P. Anderson, J. Bowring, R. McCauley, G. Pothering, and C. Starr. An undergraduate degree in data science: Curriculum and a decade of implementation experience. In *Proceedings of the ACM Symposium on Computer Science Education*, pp. 145–150, 2014.

[4] A. Begel and T. Zimmermann. Analyze this! 145 questions for data scientists in software engineering. In *Proceedings of the ACM International Conference on Software Engineering*, pp. 12–23, 2014.

[5] H. Beyer and K. Holtzblatt. *Contextual Design: Defining Customer-Centered Systems*. Interactive Technologies. Elsevier Science, 1997.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D^3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

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

[8] L. Byron and M. Wattenberg. Stacked graphs - geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[9] M. Cherubini, G. Venolia, R. DeLine, and A. J. Ko. Let's go to the whiteboard: How and why software developers use drawings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 557–566. ACM, New York, NY, USA, 2007. doi: 10.1145/1240624.1240714

[10] W. Cho, Y. Lim, H. Lee, M. K. Varma, M. Lee, and E. Choi. Big data analysis with interactive visualization using r packages. In *Proceedings of the ACM Conference on Big Data Science and Computing*, pp. 18:1–18:6, 2014.

[11] T. Clegg, E. Bonsignore, J. Yip, H. Gelderblom, A. Kuhn, T. Valenstein, B. Lewittes, and A. Druin. Technology for promoting scientific practice and personal meaning in life-relevant learning. In *Proceedings of the 11th International Conference on Interaction Design and Children*, pp. 152–161. ACM, 2012.

[12] A. Cooper. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999.

[13] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.

[14] T. H. J. Davenport and D. J. Patil. Data scientist: The sexiest job of the 21st Century. *Harvard Business Review*, pp. 70–76, October 2012.

[15] B. Dykes. Data storytelling: The essential data science skill everyone needs. *Forbes*, March 2014.

[16] S. Few. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press, 2004.

[17] S. Few. Statistical narrative. *Visual Business Intelligence*, July/August 2009.

[18] S. Few. Why do we visualize quantitative data? *Visual Business Intelligence*, May 2014.

[19] S. Few. Data sensemaking requires time and attention. *Visual Business Intelligence*, June 2015.

[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

[21] S. Goodwin, C. Mears, T. Dwyer, M. G. de la Banda, G. Tack, and M. Wallace. What do constraint programming users want to see? exploring the role of visualisation in profiling of models and search. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):281–290, Jan 2017. doi: 10.1109/TVCG.2016.2598545

[22] G. Grolemund and H. Wickham. A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204, 2014. doi: 10.1111/insr.12028

[23] K. Holtzblatt and H. Beyer. *Contextual Design: Evolved*. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool Publishers, 2014.

[24] K. Holtzblatt, J. B. Wendell, and S. Wood. Rapid contextual design: A how-to guide to key techniques for user-centered design. *Ubiquity*, 2005:3–3, Mar. 2005. doi: 10.1145/1066322.1066325

[25] H. Hutchinson, W. Mackay, B. Westerlund, B. B. Bederson, A. Druin, C. Plaisant, M. Beaudouin-Lafon, S. Conversy, H. Evans, H. Hansen, et al. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 17–24. ACM, 2003.

[26] B. Johnson, Y. Song, E. Murphy-Hill, and R. Bowdidge. Why don't software developers use static analysis tools to find bugs? In *Proceedngs of the International Conference on Software Engineering*, pp. 672–681, May 2013. doi: 10.1109/ICSE.2013.6606613

[27] S. Jolaoso, R. Burtner, and A. Endert. *Toward a Deeper Understanding of Data Analysis, Sensemaking, and Signature Discovery*, pp. 463–478. 2015. doi: 10.1007/978-3-319-22668-2_36

[28] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling. *Information Visualization*, 10(4):271–288, 2011.

[29] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, Dec 2012. doi: 10.1109/TVCG.2012.219

[30] J. Klein. When big data goes bad. *Fortune*, November 2013.

[31] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[32] J. Liu, A. Wilson, and D. Gunning. Workflow-based human-in-the-loop data analytics. In *Proceedings of the ACM Workshop on Human Centered Big Data Research*, pp. 4949–4952. ACM, 2014.

[33] K. Madhavan, N. Elmqvist, M. Vorvoreanu, X. Chen, Y. Wong, H. Xian, Z. Dong, and A. Johri. DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1823–1832, Dec 2014. doi: 10.1109/TVCG.2014.2346747

[34] L. Manovich. Visualization methods for media studies. *Oxford Handbook of Sound and Image in Digital Media*, 2014.

[35] J. Manyika and M. Chui. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute, May 2011.

[36] A. McAfee and E. Brynjolfsson. Big data: The management revolution. *Harvard Business Review*, 90(10):60–66, October 2012.

[37] W. McKinney. Data structures for statistical computing in Python. In *Proceedings of the Python in Science Conference*, pp. 51–56, 2010.

[38] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: Interactive visual exploration of system management time-series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1483–1492. ACM, New York, NY, USA, 2008. doi: 10.1145/1357054.1357286

[39] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. MLlib: Machine learning in Apache Spark. *Journal of Machine Learning Research*, 17:1–7, 2016.

[40] T. Mühlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit. Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1643–1652, Dec 2014. doi: 10.1109/TVCG.2014.2346578

[41] W. H. O. of Science and T. Policy. National big data research and development initiative, March 2012.

[42] Z. Padgett and E. Davidovits. *Finding the user in data science*. IBM Data Science Experience Blog, http://datascience.ibm.com/blog/finding-the-user-in-data-science/, June 2016.

[43] C. Palmer, C. Thompson, K. Baker, and M. Senseney. Meeting data workforce needs: Indicators based on recent data curation placements. In *iConference Proceedings*, pp. 522–537, 2014. doi: 10.9776/14133

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] A. S. Pentland. The data-driven society. *Scientific American*, 309:78–83, 2013.

[46] V. G. Pinto, L. Stanisic, A. Legrand, L. M. Schnorr, S. Thibault, and V. Danjean. Analyzing dynamic task-based applications on hybrid platforms: An agile scripting approach. In *Proceedings of the Workshop on Visual Performance Analysis*, pp. 17–24, Nov 2016. doi: 10.1109/VPA.2016.008

[47] D. Pugmire, J. Kress, J. Choi, S. Klasky, T. Kurc, R. M. Churchill, M. Wolf, G. Eisenhower, H. Childs, K. Wu, A. Sim, J. Gu, and J. Low. Visualization and analysis for near-real-time decision making in distributed workflows. In *Proceedings of the IEEE Parallel and Distributed Processing Symposium Workshops*, pp. 1007–1013, May 2016.

[48] M. T. Rodríguez, S. Nunes, and T. Devezas. Telling stories with data visualization. In *Proceedings of the Workshop on Narrative Hypertext*, pp. 7–11, 2015. doi: 10.1145/2804565.2804567

[49] D. Russell, M. Stefik, P. Pirolli, and S. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT and ACM Conferences on Human factors in computing systems*, pp. 269–276, 1993.

[50] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017.

[51] A. Satyanarayan, K. Wongsuphasawat, and J. Heer. Declarative interaction design for data visualization. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 669–678, 2014.

[52] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, 2014.

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

[54] B. Shneiderman, C. Dunne, P. Sharma, and P. Wang. Innovation trajectories for information visualizations: Comparing treemaps, cone trees, and hyperbolic trees. *Information Visualization*, 11(2):87–105, 2012.

[55] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publishing Company, USA, 5th ed., 2009.

[56] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis and visualization of multi-dimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[57] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

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

[59] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2009.

[60] D. Wixon, K. Holtzblatt, and S. Knox. Contextual design: An emergent view of system design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 329–336, 1990.

[61] A. Woodie. Big data analytics give electoral edge. *Datanami*, June 2013.

[62] X. Zhang, H.-F. Brown, and A. Shankar. Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 5350–5359, 2016.