Socrates: Data Story Generation via Adaptive Machine-Guided Elicitation of User Feedback

Guande Wu 🔗, Shunan Guo 🔗, Jane Hoffswell 🔗, Gromit Yeuk-Yin Chan 🔗, Ryan A. Rossi 🔗, and Eunyee Koh 🔗

Abstract—Visual data stories can effectively convey insights from data, yet their creation often necessitates intricate data exploration, insight discovery, narrative organization, and customization to meet the communication objectives of the storyteller. Existing automated data storytelling techniques, however, tend to overlook the importance of user customization during the data story authoring process, limiting the system's ability to create tailored narratives that reflect the user's intentions. We present a novel data story generation workflow that leverages adaptive machine-guided elicitation of user feedback to customize the story. Our approach employs an adaptive plug-in module for existing story generation systems, which incorporates user feedback through interactive questioning based on the conversation history and dataset. This adaptability refines the system's understanding of the user's intentions, ensuring the final narrative aligns with their goals. We demonstrate the feasibility of our approach through the implementation of an interactive prototype: Socrates. Through a quantitative user study with 18 participants that compares our method to a state-of-the-art data story generation algorithm, we show that Socrates produces more relevant stories with a larger overlap of insights compared to human-generated stories. We also demonstrate the usability of Socrates via interviews with three data analysts and highlight areas of future work.

Index Terms—Narrative visualization, visual storytelling, conversational agent

---

◆

---

1 INTRODUCTION

Visual data stories present data facts or insights woven together with narrative visualizations to support communication [12,34,36,57,77,78] and decision-making [13,24,53,70]. To create a compelling story, the author often needs to go through a cumbersome workflow of exploring and analyzing the data to find relevant insights, arranging the insights in a meaningful order to make a story, and building a shareable artifact to present the story to the target audience [35,39].

To make the data story authoring process easier, researchers have explored a variety of automated technologies, ranging from data insight finding [20,66] to story generation [57,58,79]. Existing story generation tools often start with the user choosing attributes from the data and configuring parameters on a story generation model [58,61], then a system will show the results to the user and provide simple interactions to refine the story. We see these current workflows as trial-and-error optimization [44] – the system takes an abstract description of the user's goals (e.g., story logicality, number of facts, etc.), and produces an optimized design, without additional clarifying input from the user. Instead, the user must iteratively modify their goals using a complex and constrained UI based on the output of the system. The user must therefore understand the intricacies of the underlying data and story generation algorithm in order to use the system effectively.

We are motivated to alleviate such requirements by simply asking the users to provide iterative feedback on the data stories [22], without configuring abstract parameters. Instead of asking designers to define specifications for charts or to fine-tune weights for the story generation algorithm, we formulate the goals by learning the design trade-offs through simple questions posed to the user. Notably, these questions don't demand extensive algorithm or data knowledge.

In this work, we propose a mixed-initiative data story generation workflow that can incorporate the storyteller's preferences into an automated story generation workflow through conversational recommendation [75]. To justify the design of the workflow, we conduct a formative study that encompasses a thorough literature review and an expert interview, which validates the benefit of conversational recommendation on adaptively collecting user feedback and applying the feedback in guiding story generation. We utilize the insights from the formative study and introduce a novel algorithm utilizing Pareto Frontier Optimization to adaptively incorporate user feedback into the data story generation process. We demonstrate the feasibility of our proposed algorithm and workflow through the implementation of a prototype system, Socrates, with support for story creators to explore data facts and story candidates and also refine the generated story.

We conducted a user study with 18 participants, in which we compare the story generated with our approach against the state-of-the-art method [58] in terms of their similarity with the story manually created by the participants. We quantify the similarity between user-created story and machine-generated story by calculating the overlap of data facts and fact transitions. The study results show that data stories generated by our workflow exhibit a higher overlap with the manually created stories and yield higher relevance scores in the subjective ratings. Through the post-study interview and a use case demonstration with three data analysts using Socrates, we also report findings from their feedback on the proposed workflow and desired future system. The key contributions of this paper include:

1. **Formative study**: A summarization of user feedback space and design guidelines for mixed-initiative data story generation workflow compiled from pilot user interviews and the literature survey;
2. **Method**: A novel data story generation method with feedback elicitation based on Pareto-Frontier optimization, which enables active incorporation of user feedback in the generated data story via iterative user feedback request.
3. **Prototype**: A prototype Socrates, that employs the proposed story generation workflow, with additional support for users to explore data facts and story candidates and also refine the story.
4. **Evaluation**: A user study with 18 participants that confirms the effectiveness of the proposed workflow in terms of reflecting user feedback in the generated story, and a use case demonstration with three data analysts that validates the usefulness of Socrates.

2 RELATED WORK

2.1 Visual Data Storytelling

Visual storytelling and narrative visualization combine visuals and storytelling to effectively communicate insights through data visualiza-

- *Work was done during the first author's internship at Adobe Research.*
- *Guande Wu is with New York University. E-mail: guandewu@nyu.edu*
- *Shunan Guo, Jane Hoffswell, Gromit Yeuk-Yin Chan, Ryan A. Rossi and Eunyee Koh are with Adobe Research. E-mail: {sguo, jhoffs, ychan, rrossi, eunyee}@adobe.com.*

tions. [68]. Previous research shows that it enhances the effectiveness, memorability, and comprehensibility of data visualizations. [3, 5, 25], resulting in increasing research interest in creating theories and guidelines [4, 30, 52, 55].

Segel and Heer [55] summarized the design components of narrative visualizations by reviewing data stories in journalism and creating a taxonomy of genre and narrative structure. Cohn [15] introduced five core visual narrative categories – Establisher, Initial, Prolongation, Peak, and Release, which was also applied in Amini et al.'s analysis of narrative structures in data videos [1]. Dykes proposed a four-stage model based on Freytag's "pyramid-based" structure in traditionalr storytelling [19] – Setting, Rising, Aha Moment, Solution, and Next Steps – for effective data storytelling [21]. Yang et al. [73] proposed a design space for applying Freytag's pyramid to data story creation. We incorporated their proposed narrative structures into our questionnaire process to optimize the data story generation and effectively collect user feedback.

Prior research has investigated the composition and authoring workflow of data stories. Lee et al. [39] defined visual data stories as a series of story pieces presented in a meaningful order to achieve communication goals. The definition is used by other recent work [37,57,58,66,77]. Based on Kosara and Mackinalay's [35] model, Lee et al. introduced a three-step authoring process for data stories, while Chevalier et al. [14] extended it with additional roles. Showkat and Baumer [59] emphasized the story-ideation process. These studies highlight the need for customization in data story authoring, including adapting stories to different audiences. Our work aims to incorporate storyteller preferences into the automated data story generation workflow.

2.2 Authoring Tools for Visual Data Stories

Researchers have created various tools to help with the time-consuming process of data exploration and data story creation. These tools allow authors to create data stories in different narrative genres, such as annotated chart [51], comic strip [33], slide show [16], and data video [2]. The goal of these tools is to provide an easy-to-use interface for transforming insights or visualizations into a presentable data story format. For example, Tableau Story arranges visualization charts into a slide show and allows users to add titles, captions, and annotations to the charts. DataClips [2] provides a library of data clip templates for users to quickly construct a sequence of clips for a new data video. While these tools make it easier to create a presentable narrative visualization, users still need to manually explore, find and connect insights.

New intelligent and automatic techniques for storytelling have emerged recently to remove obstacles in the data story creation process. For instance, ChartStory [77] arranges visualization charts into comic-style layouts and narratives. AutoClips [57] was developed to generate videos from a series of data fact visualizations.

Calliope [58] is the most closely related approach that introduces a fully-automatic data story generation method, which generates an insight sequence that achieves the best score in terms of three quality metrics – logicality, integrity, and diversity using Monte Carlo Tree Search (MCTS). Despite that Calliope allows users to assign weights to the three metrics, it does not provide the option for users to express more concrete preferences regarding the story content (e.g., the importance of certain data attributes or subspace) or story structure (e.g., narrative patterns and transitions), which are often tied closely to story creator's own understanding on the data. In contrast, Sun et al. introduced a cooperative data story editing workflow, Erato [61], which allows users to specify key story frames before generating the data story. Erato focuses on creating smooth transitions between user-specified frames and the interpolated data facts, while still requiring users to input insightful data facts as a starting point. In our work, we fully leverage the machine's effectiveness in insight discovery and orchestration. Simultaneously, we elicit and incorporate user feedback through a machine-guided question-answering workflow.

We build our work upon Calliope, utilizing its insight discovery and story generation engine to prepare a set of story candidates with satisfying logicality, integrity and diversity. In addition, we introduce mixed-initiative workflow with machine actively collects user feedback through concrete questions regarding their preference for story content and structure. The machine then utilizes user response to guide story generation process, producing increasingly better stories that are more relevant to user's goals and preferences.

2.3 Mixed-Initiative and Conversational Recommendation

The concept of mixed-initiative approaches is widely applied to applications of human-computer interaction [17, 65]. Contrasting with traditional "human-in-the-loop" guidelines, Endert et al.. [23] proposed a "human-is-the-loop" visual analytics paradigm, which suggested that user interactions and tasks of users are more preferred by users than explicit input for model steering [22]. For example, ForceSPIRE [22] infers the importance of features (i.e., keywords) for document clustering through analyzing user interactions The contrast is to let users specify the features or parameters for model steering. Similarly, visualization recommendation [8, 9, 26] aims to identify users' analytical tasks and preferences based on users' interaction with charts of interest.

Conversational recommendation serves as a potential component to understand user intention in mixed-initiative systems. With roots in natural language processing, dialogue systems, and interactive machine learning, it aims to provide personalized and adaptable recommendations through a back-and-forth dialogue [75]. These systems employ various techniques such as natural language understanding, context-aware reasoning, and reinforcement learning to refine recommendations and create engaging experiences [31]. Conversational recommendation systems have been applied to diverse domains, including media and entertainment, travel and tourism, health and fitness, education and career, and creative writing and story generation [27, 46, 52, 62, 72]. The versatility and potential of these systems make them valuable tools for enhancing personalization and user experiences in a wide range of industries and applications. However, their application on data visualization, particularly narrative visualization, remains relatively unexplored. Existing systems such as Calliope [58] and ChartAccent [51] may not be accessible to non-expert users who wish to personalize their data stories. Therefore, we propose a plug-in module that uses conversational recommendations to enhance data story generation in existing authoring systems, thereby optimizing user experience.

3 CONVERSATIONAL AGENT FOR DATA STORY CREATION

As discussed in Sec. 2.2, current tools for visual data storytelling face challenges balancing between efficiency and customizability. Highly customizable authoring tools [38] often require users to manually engage in the insight discovery and story creation processes. In contrast, fully-automated pipelines [57, 58] lack adequate support for incorporating users' intentions into the generated stories. Story creators, such as data journalists and business analysts, often construct data stories with specific communication objectives in mind, which will significantly shape the content and structure of the stories. For example, data journalists might employ dramatic narrative structure to enhance the captivating nature of the story, utilizing easily digestible data attributes to cater to layman audience. Conversely, business analysts may concentrate on a distinct group of key performance indicators while constructing business reports, opting for straightforward narratives that precisely convey the statistical information. Therefore, the focus of this work is on a user group that possesses a distinct preference for tailoring data stories, and that has particular preference in customizing the data stories, and should have a significant level of expertise to specify the desired story content or structure. To elaborate, these users are anticipated to possess a solid comprehension of data attributes and their importance within an analytical context, so that they can make decisions on what attributes the story should focus on. Furthermore, a fundamental grasp of various narrative patterns (e.g., show dramatic contrast, highlight decisive moment, etc. [73]) a story can adopt is expected, enabling them to articulate their requirements concerning the organization of the story pieces more effectively.

One potential solution involves fostering collaboration between users and machine agents. Successful stories should not only meet users' needs but also adhere to established design guidelines [55]. While general design principles can be integrated into machine agents using

quantifiable metrics [56, 58], it is critical to gather user feedback during the story generation process to reflect their customization needs and communication goals.

While existing data story authoring tools provide some ways to allow users to input their goals, either through pre-defined questionnaires [66], or configurations on the data or insight [2, 58], they require users to manually input the preset parameters as low-level details. However, setting the configurations can be time-consuming and confusing for story creators that are not familiar with the data structures or mechanisms behind insight discovery and story generation. Research on conversational recommendation systems (CRSs) [31] opens a way to acquire user needs more efficiently. Akin to a human designer, a conversational agent (CA) can adaptively propose questions to users through dialogues based on the recommendation goal and real-time user response to iteratively collect feedback; the CA can continually refine its understanding of the user's taste to make more accurate suggestions [47].

CRSs have shown advantages for collecting user feedback across various domains [42, 43], with the benefit of also improving human trust in the machine, which was acknowledged as a critical issue in fully automated story generation [58]. However, there is no existing work that utilizes CRSs in visual data story creation. To design such a conversational agent for data story creation, we will need to scope the type of questions that machine should propose as well as the feedback that story creators can provide. To this end, we first conducted a literature review on recent insight discovery and data storytelling to extract types of feedback user could incorporate in the created stories. To verify these results, we conduct semi-structured interviews with three experts experienced in data story creation. We summarize our findings into three requirements for supporting a data story generation workflow with machine-guided user feedback elicitation.

3.1 Literature Survey

To understand the potential feedback a machine can collect from user when creating data stories, we review and summarize the design components considered in existing visual data storytelling studies. Since there is little work on human-AI interaction in data story creation, we extend the survey scope to include automated and semi-automated visualization research. We begin the paper collection with two of the most recent papers [12, 61] and trace earlier studies based on their references. Two main categories of design needs were derived:

Content feedback: Information collection and fact analysis is a core step of creating a data story [35]. For tabular data, facts can be categorized into entity-focused [12] and pattern-focused [29]. Entity-focused facts analysis is based on users' interest in particular columns, rows, or range of values in the table [74], while pattern-focused fact analysis target at finding specific patterns, such as trends and distributions. To accommodate different content needs of story creators, the conversational agent should be optimized to address the particular insight discovery needs of the story creator [54, 69, 70].

Structure feedback: The structure of how the story pieces are organized can largely impact audience's perception on the story [30], which can be represented by the transitions and distributions of data facts (e.g., diversity, logicality or integrity [58]) or the narrative pattern of the overall story [73]. Depending on the communication goals, story creator can prioritize the relation factors or manipulate the narrative structure differently [1]. Therefore, it is necessary to gather users' preferences on how they wish to organize the story pieces to create data stories that align better with their communication purposes [33, 55, 79].

3.2 Pilot User Interview

To verify our conclusions from the literature review and gather initial impressions on data story creation using machine-guided elicitation of user feedback, we conducted interviews with three pilot users (PU1-PU3) who frequently author data stories. Specifically, PU1 is a product manager who writes data reports for product market analysis, PU2 is a data visualization researcher with published data story papers, and PU3 is a journalist who often writes data-centric news

articles. We selected these personas since they have been widely recognized in existing work as important stakeholders for data-driven storytelling [12, 55]. The interview process consists of two phases. First, we seek confirmation for the needs derived from our literature review by asking participants about their workflow and authoring needs when creating a data story. Second, we conducted a wizard-of-oz experiment, where we simulate a machine-guided story creation workflow and ask for participants' feedback. Specifically, we act as a conversational agent that proposes questions to gather content and structure needs from participants. We use their response to filter the data and manually organize the facts into a data story.

Workflow and story needs. When asked about their workflow when creating data stories, all three experts described a similar pipeline of "exploring the dataset" to identify key data facts, "creating a story outline" to establish the narrative structure, and "selecting the visualizations". Diving into more details of each step, PU1 primarily focuses on obtaining data facts for specific metrics, such as `hit rate` of the products, `revenue`, etc. PU3, on the other hand, focuses more on the narrative flow. Specifically, PU3 described a sample story about the electrical automotive industry distribution in China, which was created by first identifying a series of analytical directions, e.g., the temporal change of the distribution and the causal relationship with the supply chain, then filling in the content based on the chosen narrative pattern. PU2 agreed that both content and structure are important factors when creating stories, but the priorities may be different depending on the purpose and audience. Moreover, PU1 and PU3 mentioned that since they lack sufficient data analysis skills, they usually collaborate with data analysts. Both PU1 and PU3 agreed that an automated pipeline would be beneficial, but also noted that existing automated solutions do not support the level of customization they require (e.g., PU1's emphasis on particular attributes or PU3's control of the narrative flow).

Implications on wizard-of-oz experiment. All participants acknowledged that assistance from a conversational agent can be beneficial. In particular, this process reminded PU3 about her experience working with domain experts when creating a data story, where she needs to iteratively adjust the insights and narrative according to experts' feedback. However, PU3 complained about the inefficiency of back-and-forth communication and the need to confirm many details from domain experts on a complex dataset. Similarly, our experiments revealed that, after 10 questions, participants became less focused and started providing more neutral answers. It emphasizes the necessity of limiting the number of questions from machine-guided story generation.

We also noticed a gap between the analytical tasks and actions. Considering the customized insight and chart configurations that the state-of-the-art tools [58, 61] support, we included some questions related to low-level analytical actions for specifying the data facts, such as "what attributes to use in the insight" or "what chart type to display the insights." However, PU1 and PU3 expressed confusion on such questions even though they had no difficulty making decisions for high-level analytical tasks (e.g., analyzing a distribution, making comparisons, etc.). PU3 explained: "I'm interested in seeing distribution insights but not sure what attributes to use until actually looking into data." Previous study from Brehmer and Munzner [7] also pointed out this knowledge gap between high-level and low-level analytical tasks, inspiring us to prioritize questions that seek decisions on high-level tasks from the user instead of low-level analytical actions.

3.3 Design Guidelines

Based on the findings from literature review and the interview study, we have compiled a set of three design guidelines that should be considered when designing the CA for visual data story creation.

DG1: Facilitate user control over the story content and structure. All interviewed experts highlighted the importance of incorporating their needs content and story structure in the created data stories. The conversational agent should effectively collect users' feedback on the insights they want to focus on and the narrative patterns they prefer. In return, the generated data stories should better align with users' expectations by incorporating their needs.
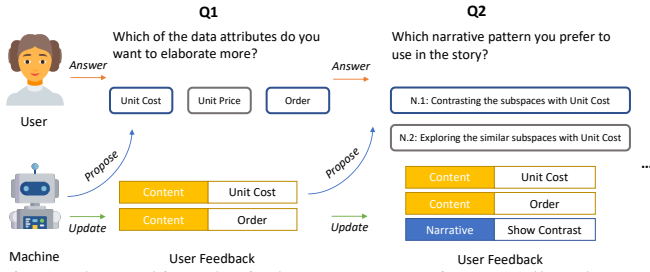
Fig. 1: The machine adaptively proposes questions to collect the user's feedback, which is incorporated into story generation.

DG2: Minimize the number of question turns. Both our findings in wizard-of-oz experiment and PU3's feedback in practice indicates a loss of patience when asked too many questions during the process of story creation, which can introduce noise in the user responses. To avoid this, the conversational agent should actively prioritize the question that is most informative to the process of searching candidate story to ensure that the most important aspects of the users' needs are covered within a limited number of question turns.

DG3: Prioritize questions that gather high-level decisions. Despite being able to specify configurations on individual chart and insight [58, 61] offers the finest level of control on the story content, it can be hard for story creators to decide without sufficient knowledge on the data and expertise on data analysis. Our finding indicates the potential benefit of prioritizing user feedback on high-level decisions, such as insight types or narrative patterns to use, instead of concrete insight specifications (e.g., chart type or data attributes for particular insight).

4 STORY GENERATION WITH USER FEEDBACK

This section introduces our proposed method to support data story generation with user feedback incorporated. In particular, we aim to search through a set of story candidates via a machine-guided feedback elicitation process to identify a subset of stories that align better with user feedback. The algorithm performs two key tasks: 1) evaluating story candidates within the context of user-provided feedback, 2) adaptively proposing questions to gather user feedback according to the searching need. Figure 1 represents an overview of an example user workflow. In the following, we first give the overall pipeline of the algorithm and then expand more details on how each task is performed.

4.1 Algorithm Overview

Following the common problem formulation in CRSs, we also formulate data story generation into a recommendation process, which starts preparing a set of story candidates S. We generate story candidates with a re-implementation of the sequential data story generation algorithm proposed by Shi et al. [58] to ensure the candidates have satisfying logicality, integrity, and insight diversity. Specifically, logicality measures the transition coherence (i.e., commonness in data subspace, measure, breakdown, and insight type) between adjacent data facts. Integrity measures the data coverage of a particular data story to prevent drill-down fallacy [40], while diversity ensures the variance of data facts in the story. To accelerate story candidate generation, we adapt the original MCTS with a beam search algorithm, as described in Appendix E.

The next step is to search through the story candidates and recommend an optimal data story based on user feedback. To incorporate diverse user feedback (discussed in Section 3.1) in the story evaluation process, we propose a faceted schema to concretely capture user feedback across various aspects (**DG1**). The facet concept is inspired by the faceted approach used in recommendation systems [32, 48, 67]. Specifically, given an input dataset, we identify a set of potential feedback facets, denoted as $L = \{l_1, l_2, l_3, ..., l_j, ..., l_J\}$. Each facet may be weighted differently for each story candidate. During the story generation process, we gather feedback from the user that indicates their preferences on the facets L denoted as $m = \{\beta_1, \beta_2, ..., \beta_j, ..., \beta_J\}$, where a $\beta_j \in \{-1, 0, 1\}$ represents the user's opinion towards the

feedback facet l_j. This allows users to express their preferences for including (e.g., $\beta = 1$) or excluding (e.g., $\beta = -1$) specific information from the data story. Therefore, the goal is to estimate m by iteratively collecting user responses and recommending a subset of data stories S^* that aligns best with m.

To evaluate the alignment of facets shown in a story candidate s with m, we define an alignment reward $r_j^s \to [0, 1]$ for each feedback facet l_j. We will expand on the facet space and reward functions in Section 4.2. Finding S^* given m can be considered a multi-objective optimization problem with the goal of maximizing the reward for each feedback facet. We employ Pareto optimization to extract the Pareto frontier [28, 45] that can best satisfy all the objectives specified in m. In the context of multi-objective optimization, the Pareto frontier is defined as the set of solutions that are not dominated by any other solution in the search space [64]. A solution dominates if it is better in at least one objective while being equal or better in all other objectives than all possible solutions. For our problem, a story candidate s dominates if it has better rewards on at least one feedback facet $l \in m$. Thus, the optimal story set can be expressed as the Pareto frontier given m:

$$S_m^* = \{s | s, s' \in S, s \neq s', \exists l \in m, r_l^s > r_l^{s'}\} \tag{1}$$

The size of the Pareto frontier can be negatively correlated with the size of the objectives (i.e., feedback facets) [64]. Initially, when no user feedback is available, all data stories are in the frontier. As we collect iterative feedback, the frontier size gradually reduces until only a limit number of story candidates remain. In particular, the user feedback is collected through multi-turn question-answering, in which we hope to minimize the number of questions to reduce user workload (**DG2**). To this end, we need to select questions that can best accelerate the search process. Specifically, in each turn t of the question-answer, the algorithm selects a question $q \in Q$ that can reduce the size of the Pareto frontier most based on the current optimization status m_t and asks the user to provide an answer a. We denote this process as:

$$m_{t+1} = \arg\min_q |S_{m_t}^*| \tag{2}$$

where m_{t+1} is the updated estimation of user preference on the feedback facets. More details on how the question q in each turn is selected and how the m_{t+1} is updated will be introduced in Section 4.4.

4.2 Feedback Facets and Alignment Reward

User feedback is captured under a set of feedback facets L to incorporate user control on story content and structure (**DG1**). Rewards are calculated for each facet to evaluate the alignment between user feedback and the story candidates. In this section, we elaborate more details on the feedback facets and the rewards.

4.2.1 Feedback Facets for Story Content

In Sec. 3.1, we discuss user feedback focusing on specific data entities and patterns. To compare desired entities and patterns with those in story candidate s, we compute an alignment reward r^s by aggregating the preference m that includes the desired data attributes and vice versa:

$$r^s = |\{l_j \in m\} \bigcap \{l_j \in s\}| \tag{3}$$

where r^s corresponds to the intersect of the facets in the user's facet set m and the story's involved facets $\{l_j \in s\}$.

4.2.2 Feedback Facets on Story Structure

As discussed in Sec. 3.1, we aim to capture feedback on story structure by assessing 1) diversity, logicality, and integrity of the data facts in a story; and 2) the narrative structure of the overall story flow. To quantify the story structure in the first aspect, we followed the metrics defined by Shi et al. [58] to estimate the variance of fact types in the story (i.e., diversity), coherence of adjacent data facts (i.e., logicality), and data coverage of a story (i.e., integrity). For the second aspect, we include questions to gather user preference on the following narrative patterns

following the topology proposed by Yang et al. [73] and identify the presence of a narrative pattern with the heuristics described below:

C1 Showing contrast emphasizes differences between data facts to create "highlight" effect. Following the methodology of prior research [30, 58], we determine the presence of this narrative pattern by detecting facts with contradicting trends or associations, such as the increasing average price of the Hat and the decreasing sum of orders.

C2 Showing accumulative significance aims to heighten the audience's excitement by repeating facts with similar meanings progressively. We incorporate this narrative pattern in the story candidates by expanding on data facts related to the same data subspace or measures. For instance, when presenting data facts about Japan, an initial data fact "The average price of the Hat is increasing in Japan" followed by an additional data fact related to "Japan" subspace and "Unit Price" measure will accumulate significance of Japan-specific data facts.

C3 Showing the decisive moment attempts to captivate the audience by highlighting crucial temporal points in the data facts. In particular, the decisive moment represent significant changes in quantitative trends or semantically important events [73]. We incorporate this narrative pattern by highlighting the trend facts and expanding the story with the facts related to the specific moment.

C4 Showing ranking presents data in ranked order, engaging the audience through a series of related facts about the ranked items. Specifically, we incorporate this narrative pattern by first introducing a ranking type data fact (e.g., California has the highest homeless population than other states in the nation.) and expanding related data facts sharing the same data subspace and measure.

To represent the user's feedback on the narrative patterns, we add the feedback facet on the narrative pattern to the existing facets L as $\{pattern, c\}$, where each $c \in \{C1, C2, C3, C4\}$ represents an individual narrative pattern. We adapted Equation 3 to calculate the reward on narrative pattern in a similar manner.

4.3 Question Space

In order to efficiently and effectively collect user feedback, we have designed a set of questions that cover various categories of feedback. Our approach incorporates three types of choice questions that have been proven user-friendly and efficient in conversational recommendation systems, namely rating questions [41], comparison questions [71], and multi-choice questions [76]. The algorithm adaptively selects the most suitable question type-based optimization needs, taking into account the number of facets that require user input. In the following, we describe each question type in detail. Examples of each question type can also be found in Appendix D.

Rating Question. The algorithm proposes a rating question to the user when it needs feedback primarily on one feedback facet. As shown in Fig. 2 (question 3–5). the user can select either important, not important, or neutral. If the user responds with "neutral", the system preserves the current importance score of the corresponding facet. If the user answers "important", the boolean indicator β of the feedback facet will be positive. Or, if the user chooses "Not Important", β will be negative.

Comparison Question. The comparison question is presented when the optimization direction depends on the choice of two facets, which asks the user to compare two facets and select the one that is relatively more important (as shown in question 2 of Fig. 2). After the user gives a response, the compared two feedback facets l_i and l_j are merged as $l^* = \{l_i, l_j\}$. And we compute the new alignment reward of l^* as $\alpha \in (0.5, 1]$ as $r^* = \alpha r_i + (1 - \alpha)r_j$, where $\alpha \in (0.5, 1]$ is a constant value. In our implementation, we set $\alpha = 0.8$, which gives more weight to the user-selected option.

Multi-Choice Question. The multi-choice question allows the user to give preference on multiple facets simultaneously (as shown in Fig. 2, question 1), providing an effective alternative when the optimization algorithm needs feedback on several equally important facets. Each option will be associated with one facet l_k, when the chosen option' indicator β_k will be set as positive when others are set as negative.

Algorithm 1 Question Selection

Input: Candidate story set $S = \{s_1, s_2, ..., s_N\}$
Feedback facet set $L = \{l_1, l_2, ..., f_J\}$
Candidate question set $Q = \{q_1, q_2, ..., q_K\}$
Current state of the questionnaire S^*

Output: The next best question q^* to propose

if $|S^*| < \varepsilon$ **then**
    return None and stop the question generation iteration
else
    for question $q_k \in Q$ **do**
      compute $|S_k^*|$ based on $E[S^*|q_k, m]$
    end for
    Set $q^* = \text{argmin}(\{|S_1^*|, |S_2^*|, ..., |S_K^*|\})$ to be the optimal question
end if

We decided to only include choice questions with the goal of minimizing user effort in giving feedback. Alternatively, we also considered ranking questions but observed longer response times in our wizard-of-oz study. Additionally, the open-ended questions may offer flexibility in user input and the potential to provide more information to the algorithm, but they can be laborious for users to respond to and may require a large language model to accurately decode user preference.

4.4 Question Selection and Story Generation

We implement our questionnaire module to select the most beneficial question based on the current user feedback m. At each turn of the conversation, the system selects a question q and asks the user to provide the answer a. We represent the question turn as $\pi = (q, a)$. The effect of each turn is to update the state of the user's feedback m shown in Equation 2, To select the optimal question to reduce the size of the Pareto frontier S^*, the questionnaire module exploits a Monte-Carlo strategy to compute the expectations of the question candidates.

$$E[|S^*||q, m] = \sum p(a|q, m) * |S^*(qa(m))|, \quad (4)$$

where $E[|S^*||q, m]$ indicates $|S^*|$'s conditional expectation based on the question candidate m. For simplicity, we use the uniform distribution $p(a|q, m_b) = \frac{1}{N_a}$, where N_a is the number of the possible user's answers. The detailed algorithm can be found in Algorithm 1.

Algorithm 1 operates iteratively to minimize the size of the Pareto frontier. Once the size of the frontier reaches a predefined threshold, $|S^*| < \varepsilon$, the iteration terminates, and the algorithm outputs story candidates on the frontier and recommends the optimal data story to the user. The selection process involves identifying the data story that best satisfies the majority of the feedback facets present in m, ensuring that the generated story is in alignment with user preferences and intentions. The proposed question selection algorithm has an adequate scalability to large question set with a time complexity of $O(KN)$, where K is the size of the question set and N is the size of the candidate set. The space complexity is $O(N + J + K)$. When dealing with excessively large datasets, the story set can be trimmed to accommodate efficiency requirements. Additionally, the scalability of the algorithm may be impacted by the facet size, which is proportional to the number of data columns in the dataset. In Sec. 5.6, we present a runtime analysis of the question selection process regarding the dataset complexity.

The convergence of the algorithm in our approach relies on the story candidate set and the user's answers. In our approach, we establish a lower bound on the convergence rate, denoted as σ, in a user session using Algorithm 1. This lower bound can be calculated as $\sigma = \min_{i=1, i \leq T} \frac{E[S_i^*]}{S_i}$, where T represents the number of turns in the session. The iteration terminates when the size of the story candidate set, $|S_i|$, falls below a predefined threshold ε. By analyzing the convergence condition, we can derive an upper bound on the number of turns given by $T \geq \frac{\ln(S_0)}{\ln(\varepsilon)}$, where S_0 is the initial story candidate set. Consequently, our algorithm exhibits a logarithmic convergence rate, ensuring scalability even for large datasets. In our user study imple-

mentation, we observed that the algorithm typically converged within six turns of questions, as reported in Section 5.5.3.

4.5 Adaptive Workflow

Our proposed method utilizes an adaptive workflow between the story creator (i.e., user) and the optimization algorithm (i.e., system) to incorporate user feedback in the automated story generation. We design the workflow following the three-stage taxonomy proposed by Sperrle et al. [60] for the adaptive system: initialize, refine, and automate.

In the initialization stage, the system generates a list of story candidates under quality control (as introduced in Sec. 4.1) without emphasizing particular feedback facets L. The story candidates are used to create an initial Pareto frontier. Next, the system enters an iterative loop of refinement and automation.

In the refinement stage, the system utilizes Algorithm 1 to select the most informative question for determining the optimization direction. The user is then presented with the question, for example, "Q1: Which of the following attributes do you want to elaborate more about?" (as shown in the Fig. 1 example). The user provides feedback by selecting "unit cost" and "order", and the system updates user feedback m with indicators $\beta = 1$ on the corresponding facets to imply user interests and indicator for "unit price" $\beta = -1$.

In the automation stage, the system recomputes the alignment reward for each story candidate with the updated user feedback m and generates a new Pareto frontier for a new round of "refinement-automation" (i.e., question-feedback) loop. The system keeps filtering the story candidates with the user feedback until either the size of the Pareto frontier falls below a predefined threshold, or the number of question-feedback loops reaches the upper limit.

In our implementation, we set the threshold for the size of the Pareto frontier to 5, and the system will output 5 story recommendations for users to explore. In addition, considering the insights from our pilot user interviews (Sec. 3.2), we set the upper limit of question-feedback turns to 10 to avoid overwhelming the user with excessive questions.

5 USER STUDIES

To validate the effectiveness of the proposed workflow, we developed Socrates: a prototype system that incorporates our data story generation model (Section 4) along with an interactive interface (described below). We then evaluate Socrates through a quantitative user study with 18 participants as well as an objective comparison between machine- and user-generated data stories. We have two primary hypotheses:

H1 Socrates generates data stories with better quantitative ratings compared to prior weight-fine-tuning approaches;

H2 Socrates better reflects the user's feedback and generates a data story with greater similarity to manually, user-created stories.

5.1 Socrates: Prototype Interface

Socrates combines a front-end interface, developed using React.js and D3.js [6], with a back-end server based on MongoDB and Flask. The back-end server processes user-uploaded datasets and generate an initial question; follow-up questions are generated based on the user's response in the interface (Fig. 2A). Once enough feedback has been collected, the back-end server generates the final story, which can be refined by the user through the interface.

The design of Socrates follows a co-adaptive guidance process [10, 60], enabling bidirectional guidance between the user and machine. Specifically, users provide prescribing guidance to the system by offering preference feedback (Fig. 2A), while the system provides orienting guidance for users to give response through the *interesting facts view* (Fig. 2C). In particular, this view helps bridge the knowledge gap for users who may not be very familiar with the dataset, facilitating a better understanding of the data. Once all feedback is collected, the system presents the most recommended story in the *story preview* (Fig. 2D), offering prescriptive-level guidance for story generation. However, recognizing that the most recommended story may not always be satisfactory, Socrates allows users to make high-level modifications to the story structure through *data story flow view* (Fig. 2B), which visually

depicts the relationships between selected and alternative facts based on different user responses. The longest flow corresponds to the default narrative derived from user feedback, while shorter flows represent potential narrative developments based on alternative responses. This view not only provides additional directing-level guidance for story generation but also helps mitigate the potential issues of the optimization algorithm, such as overfitting or local minima by allowing users to explore various options. In addition, the *story preview* enables users to flexibly adjust the story by deleting facts or adding interesting data facts from the *interesting facts view*. Users can also edit the accompanying text to ensure cohesive and accurate narrative visualization.

5.2 Participants and Dataset

We recruited 18 participants (9 female, 9 male) through social media and email groups. The participants' backgrounds include visualization (P2, P6, P9-10, P12, P15, P18), data science (P5, P13-14), machine learning (P1, P3-4, P8, P11, P17), biology (P7, P16), and education (P11). When recruiting the participants, we make sure they all have sufficient experience in data analysis and visualization to meet our requirement for the target audience (discussed in Sec. 3.1). For data story creation specifically, four participants have more than two years of data story creation experience, ten participants have less than two years of experience, and the remaining four participants have no data story creation experience. The participants were not compensated.

The user study was conducted through online meeting with an open dataset: US Regional Sales Data[1]. The dataset contains the orders of a US-wide sales company from 2018 to 2020. The dataset has five categorical attributes, i.e., sales channel, state, customer name, sales team, and product name, and four numeric attributes, i.e., unit cost, unit price, order quantity, and discount applied. Detailed information can be found in Appendix B.

5.3 Experimental Procedure

Before the story, we gave the users a 10-minute introduction covering the system interface, dataset context, and study procedure. We also introduce the important terms, i.e., data entity, data measure and subspace with examples from another dataset. Each participant is first asked to explore a dataset and the automatically-generated data facts. The generated facts are presented in a simple web-based system that consists of only the *interesting facts* (Fig. 2C) and *story preview* (Fig. 2D) panes of Socrates. After analyzing the dataset, the participant is required to create one primary data story and two auxiliary data stories. The two auxiliary data stories should be authored with the same design considerations in mind but can have different facts and orders. To ensure the length of the stories created, we instructed participants to include at least 5 data facts in the *interesting fact view* they found insightful. We also informed participants that they would be asked to present the story during the post-study interview session and encourage them to create meaningful narratives on the selected data facts. The user-created data stories are compared with the machine-generated stories in the later quantitative evaluation.

After finishing this initial story creation, the participant is presented with two user feedback acquisition approaches: weight fine-tuning and Socrates. The sequence of the two approaches is counterbalanced. The weight fine-tuning approach is based on Calliope [58], which allows the user to adjust the weights for the different story evaluation measures: diversity, logicality, and integrity. The user is asked to assign a weight from 0 to 1 on the three metrics according to their previous design considerations during the initial data story creation step. The second approach is our prototype system Socrates, which presents a series of questions for the user to answer in order to provide feedback on the story generation. After using both approaches, the resulting stories are presented side-by-side to the user, who is asked to rate them on a scale from 1 (worst) to 5 (best) based on the story's engagement, insightfulness, logicality, relevance, and understandability. Finally, we conducted a semi-structured interview to better understand participants'

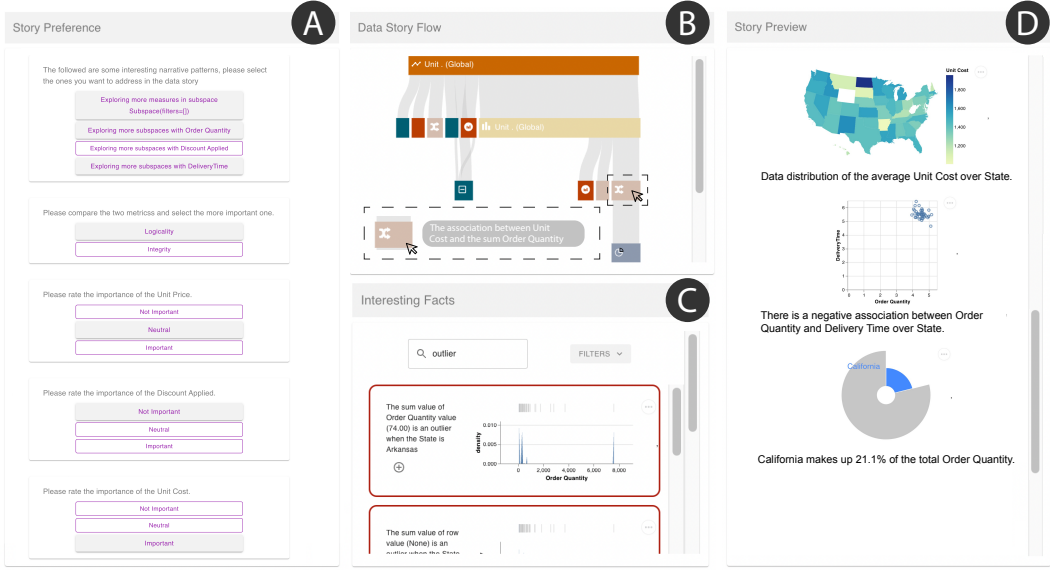[1]US Regional Sales https://data.world/dataman-udit/us-regional-sales-data

Fig. 2: The prototype interface for Socrates consists of four views. The **Story Preference View (A)** shows the system-generated questions and user-provided feedback. The **B. Data Story Flow View** presents a flowchart representing the structure of the data story, as well as alternative facts that could be added to the data story. Each node represents a fact, and the width encodes the score of the narrative transition between the two connected facts. Some facts; descriptions are hidden due to the limited space. The user can hover over a fact to view the detailed description and choose whether to add it to the story. The **C. Interesting Facts View** presents all the potential facts from the dataset. This view provides a filter panel and search bar to help the user explore the facts. The **D. Story Preview** allows the user to scroll through the generated data story.
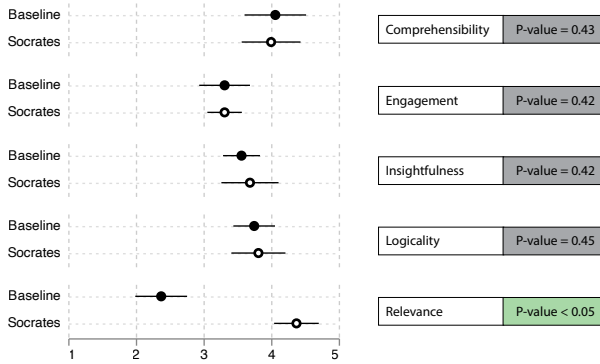

Fig. 3: Average user study ratings with one standard deviation. Our approach achieves a significantly superior rating in terms of *relevance*.

experience with Socrates and the baseline approach. The three user-created stories are compared with the two machine-generated stories using objective evaluation metrics, including *fact overlap*, *transition overlap*, and *fact component overlap*. This objective evaluation aims to assess how well the two approaches incorporate and reflect the user's feedback in order to create a satisfying data story.

5.4 Subjective User Ratings

We hypothesize that Socrates can generate a more satisfying data story in terms of the user's subjective evaluation compared to prior weight fine-tuning approaches (H1). To verify this hypothesis, we asked participants to compare the two machine-generated stories on a scale from 1 (worst) to 5 (best) for five subjective metrics: relevance, comprehensibility, engagement, insightfulness, and logicality (Fig. 3).

We run a Mann-Whitney U-test on the user's subjective ratings between the baseline [58] and our approach and report the p-value. Results show that the *relevance* score of our approach significantly surpassed the baseline with a p-value of 1.5e-6. However, for the other metrics, the result does not show a significant difference between our approach and the baseline. It shows that our plug-in module does not trade off the story quality for better customizability.

5.5 Objective Benchmark for the Relevance

We hypothesize that our approach can generate a data story that better reflects the user's feedback and is more similar to the user-created data stories (H2). To mitigate human bias in the evaluation, we propose a novel objective benchmark approach according to the overlap between the user-created and system-created data stories. The evaluation protocol is inspired by BLEU score [50], which compares the reference text and machine-generated sentences for translation tasks. In the user study, the participants manually create three data stories, which are compared with each of the two machine-generated data stories: the baseline and Socrates. Specifically, we compare the following three scores: *fact overlap*, *transition overlap*, and *fact component overlap*.

5.5.1 Fact and Transition Overlap

The *fact overlap* evaluates the overlap of the stories at the fact level. This measure is useful because the delivery of the data patterns heavily depends on individual facts. A larger fact overlap can indicate that the two stories convey similar information. We first compute the intersection between the user-created and machine-generated stories as $O(S, S^*) = |S.\text{facts} \cap S^*.\text{facts}|$, where $S^* \in \Delta$ is a user-created data story, and Δ is the set of user-created stories. Please note that, for each participant, there are three reference data stories shown in Sec. 5.3. Then, we can compute the recall and precision as follows:

$$\text{precision}(S, S^*) = \frac{O(S, S^*)}{|S.\text{facts}|}, \text{recall}(S, S^*) = \frac{O(S, S^*)}{|S^*.\text{facts}|} \quad (5)$$

We can compute the standard F-1 measure, which is denoted as $F_1(S, S^*)$ and derive both the maximum and average F-1 measures as,

$$F_1^{\max}(S) = \text{MAX}\{F_1(S, S^*)|S^* \in \Delta\}, \quad (6)$$

$$F_1^{\text{avg}}(S) = \text{AVG}\{F_1(S, S^*)|S^* \in \Delta\} \quad (7)$$

The reason for using both the maximum and average simultaneously is that multiple data stories can satisfy the same user's feedback. Evaluating the maximum F-1 measure can determine whether the machine-generated story can meet the user's expectations.

To evaluate the alignment of the narrative transitions between the data stories, we also incorporated transition-level overlap. we first define the transition as the connection between two adjacent facts in a data
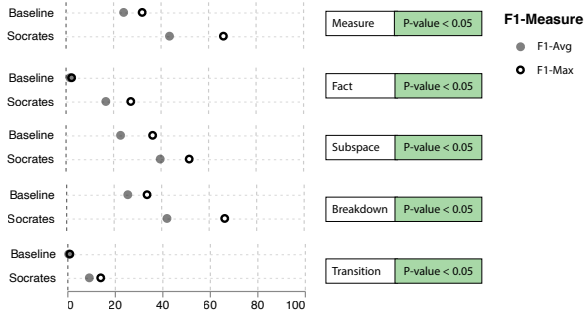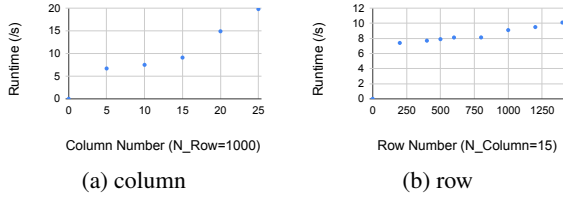
Fig. 4: The objective evaluation result

(a) column   (b) row

Fig. 5: Runtime analysis with different row and column numbers

story, transition$_j = \{x_j, x_{j+1}\}$, where x_j is the jth fact in a data story. Then we refer to the transitions in the data story S by $S.transitions$. We can derive the *transition overlap* as $O^t = |S.transitions \cap S^*.transitions|$. Following the same protocol as the *fact overlap*, we can compute the maximum and average F-1 measures for the *transition overlap*.

5.5.2 Fact Component Overlap

While the *fact overlap* and *transition overlap* can evaluate the delivery of the information in a data story, they are sensitive to the individual facts and not discriminative when the overlap is minor (e.g., no fact is shared by the two data stories). In particular, the data story can have a large search space for possible facts, which may convey similar information. For example, the fact "USA occupies the largest proportion (30.1%) of the hat orders" has a similar meaning to the fact "USA has the largest order number of the hat" but this similarity would not be captured by the *fact overlap* or *transition overlap*.

To address these issues, we propose to evaluate the overlap on the fact's component level. The fact component refers to the elements of the fact's 5-tuple. Specifically, we evaluate the overlap of the *subspace*, *measure*, and *breakdown*. We exclude the fact *type* because the fact types are quite limited, and even a random approach can achieve a higher score. To evaluate the overlap, we extract the subspace, measure, and breakdown from the story as $S.subspaces$, $S.measures$, and $S.breakdowns$. Then, we exploit the same approach with *fact overlap* and *transition overlap* to compute the max and average F-1 measures.

5.5.3 Results

The evaluated results for each type of overlap can be found in Fig. 4. These results show that our approach achieves significantly better performance compared with the baseline. The figure shows that Socrates achieves a significantly better *fact overlap* than the baseline. Since the *fact overlap* reflects the similarity of the information contained in the story, this result shows that our approach can better extract the user's intended information from the dataset. Socrates also performs better on the *transition overlap*, which reflects the narrative relations between facts used in the data story. This result suggests that Socrates's questions about the narrative patterns can be effective for acquiring the user's narrative preferences. Furthermore, Socrates is also superior to the *fact component overlap*. This observation is reasonable because our proposed questions directly cover the subspace, measure, and breakdown. To sum up, the quantitative evaluation on *fact overlap*, *transition overlap*, and *fact component overlap* verify that Socrates can effectively capture the user's feedback to generate desirable data stories.

In addition, we collected data on the number of question-answer exchanges required for the final output, which ranges from 4 to 7 (mean

= 5, SD = 0.8). Despite our algorithm allowing up to 10 question-answer turns, the median number of turns for participants to obtain the generated story was 5. Furthermore, out of 18 generated stories, only one required the participant to give more than 6 responses. These findings highlight the efficient and streamlined nature of our approach.

5.6 Runtime Analysis

We performed a runtime analysis on the question selection algorithm (introduced in Sec. 4.4) for various data complexities, including different column and row numbers.d Specifically, we generate three questions with randomly selected answers for each data size and record the average time taken for the question selection. The results (shown in Figure 5) reveal that the runtime is more sensitive to the changes in the number of columns than rows, and our approach maintains reasonable runtime (less than 10 seconds) when the number of columns is less than 10. This sensitivity is because the number of columns directly affects the space of facets, consequently influencing the search space for the questions. On the other hand, an increase in the number of rows may result in a larger set of story candidates for the algorithm to consider initially. To handle large datasets efficiently, users can choose to restrict the number of story candidates, expediting the interaction process.

6 QUALITATIVE USER FEEDBACK

To better understand users' impressions of Socrates, we collected qualitative feedback during both the user study (Section 5) and conducted additional expert interviews on Socrates with practitioners, including two business analysts (E1, E2) and one data engineer (E3). All three experts have more than 10 years of experience in their current roles and regularly create data stories to communicate insights to stakeholders. In each session, we demonstrate the usage of Socrates through a usage scenario on the user study dataset (see the supplemental material for the video demo) with a detailed introduction to the dataset and Socrates's design. We also asked experts to have hands-on experience with the system and provide feedback. All interviews were conducted via online meetings lasting between 45 minutes to one hour, and participants were given access to Socrates through a remotely controlling shared screen. In the following, we discuss participants' feedback on the story generation workflow, the general usability of Socrates, and challenges and opportunities for data story generation, referring user study participants using P# and the expert practitioners using E#.

6.1 Data Story Generation Workflow

According to our user study participants, the most common workflow consists of three primary steps (P2-P5, P10, P12): (1) selecting a story topic, (2) selecting supporting facts, and (3) refining the output story. Eleven of our eighteen user study participants (P1, P3, P6, P7, P10, P12-P15, P17, P18) noted that the data story topic was the most important consideration when creating the story. For example, P15 explained that *"I first pick an interesting topic before determining supporting facts."*

However, methods for topic selection varied; P14 relied only on the meta-data: *"Column names guide me to interesting topics,"* which risks overlooking important insights, while P5 focused on exploring facts regarding different patterns, such as increasing trends. P15 and P16 performed a trial-and-error process by repeating Steps 1 and 2 multiple times to find the most satisfying topic and set of facts. Others, like P3 and P17, used the raw data and facts; for example, P17 reversed Steps 1 and 2, first selecting the facts he was most interested in before settling on the topic, noting that *"It was only after I had seen enough insights from the data that I could decide on the story topic."* The experts similarly emphasized the importance of understanding the data first, and all of our experts noted that they begin with a dashboard for data analysis and then manually transform the visualizations into a story format (e.g., PowerPoint slides, word document, email, etc.).

Once a topic has been selected, there are still many ways that it can be organized into a story. For example, P1 and P6 used the initial data fact as the topic sentence, while P11 introduced the topic with an opening question to provide context. P11 explained that *"The audience will be attracted to the story instantly by the opening question."* The introductory sentence can act as a hook to draw the reader into the story

An important part of how the story is organized is the overall flow. Seven of the eighteen user study participants (P2, P4, P6, P9, P11, P12, P15) mentioned that the narrative transitions between the facts should be coherent and reasonable. P12 further emphasized considering both local transitions between facts and the global story arc.

6.2 Usability and Usefulness of Socrates

During the usage scenario interviews, all experts spent a large amount of time exploring the *interesting facts*, and expressed their desire to have this quick insight view in their existing analytical environments. For example, when considering the system overall, E1 noted that *"This is very helpful. We have tens of KPI metrics in our database, but we usually just pick 6 to 7 metrics for the report."* E2 had similar comments about exploring the relevant facts, noting that *"Our focus of data when creating the report can constantly change depending on the business focus at the time. Having the system know what the focus is is really useful."* When further reflecting on the design of the *story preference* view, E2 explained that *"the system asking one question at a time makes it not too overwhelming [for users]."* E3 similarly liked that the *story preference* view has the history of question answers for users to trace back: *"I can compare the generated story with the questions and answers to see if the story can actually reflect my choices."*

In order to really understand what the story generation model was doing, our experts often leveraged the *data story flow* view. E1 noted that it *"provides easy access to explore other related insights,"* while E2 explained that it *"shows multiple options in a very neat way."* When it comes to the *story preview* itself, E1 appreciated that *"the visualization and the text are arranged in the form of a document, which is similar to what we usually do when creating the data story"* and E3 expressed a similar sentiment, noting that *"I wish I could just export this [content in the story preview] to email and make it recurrent reporting."*

6.3 Challenges & Opportunities for Socrates

Though our experts agreed on the usefulness of Socrates, they expressed some concerns in terms of the learning curve for getting familiar with the visualization in the *data story flow* view (E1) and the complexity of the overall user experience (E2). E1 noted that *"It feels pretty easy to understand [the data story flow view] once you explained everything, but it may be different to remember."* E2 felt that Socrates should *"display one or two views at a time"* to simplify user tasks.

Our user study participants faced three main challenges in manual story creation: burdensome topic selection, numerous data facts, and difficulty finding relevant facts. These concerns echo aspects of the workflow described in Section 6.1. In the beginning of our study (Section 5), we provided users with a basic system to browse charts and captions in order to manually create a data story; however, participants generally felt overwhelmed with the number of data facts (P3, P7, P9, P11, P16, P18), and suggested presenting alternative representations like a 'map' for guidance on how to explore the data (P18). This "cold-start" problem can occur when participants are unfamiliar with the dataset, and struggle to decide on story development or data exploration directions. To address the "cold-start" problem, participants suggested incorporating a more intelligent fact recommendation approach. For example, E2 expressed some uncertainty about *"whether [what the] insights model considers as interesting (statistically significant) equal to what the user perceives as interesting."* E2 further wondered *"what if some attributes that I care about never show up in the questions or options?"* While Socrates allows users to add particular data facts of interest to adjust the generated story, which mitigates this problem to some degree, it is worth exploring new user-initiated interactions for communicating preferences to the story generation model. For example, Socrates supports recommending related facts when a central fact or fact component is specified.

The second type of assistance requested by our user study participants is an automatic evaluation of the story generated by Socrates (P10 and P12). Though the participants have basic knowledge about narrative visualization and data story creation, they mentioned that an evaluation system to remind the user of potential errors could be useful.

To this end, we believe that automatic linting or completing approaches similar to the work from Chen et al. [11] can be beneficial for the user.

7 DISCUSSION

In this section, we discuss the design implications that arose through the process of developing and evaluating Socrates. We also point out the current limitation of our method with potential future improvements.

Generalizability. In this work, we demonstrate the use of machine-guided feedback elicitation on data story generation. Yet, our proposed method of actively incorporating user feedback into the generation process is not limited to generating data stories. Recent studies have explored the automatic generation of various other storytelling artifacts, such as dashboard [49, 70], infographics [18] and storyline [63]. In fact, our proposed method can be adapted to most of the generation procedures with the feedback space and question space identified for the particular application to enable active incorporation of user feedback.

System design and evaluation. We develop Socrates prototype with the goal of testing our method's effectiveness in integrating user feedback during story generation. We acknowledge that there is room for more advanced features, such as supporting freeform input and employing a language model for feedback extraction. However, our primary focus leads us to collect user feedback more directly. Our goal is to align the data stories more closely with user preferences through feedback incorporation. Thus, our user study is designed from the story creators' viewpoint, evaluating whether the output aligns with their intent. While our results suggest improved alignment between user preferences and the generated story, whether this improvement is perceptible by the story's audience remains an area for future exploration.

Limitations and future directions. The primary limitation in our method is the time complexity on the story preparation. Our method requires the preparation of story candidates via MCTS [58], which has a relatively high time complexity. Despite that we mitigate this issue by developing a beam search algorithm, which is still hard to complete in real-time with large datasets. In our current implementation, we pre-calculate and store the story candidates, then conduct the story search in real time. Two potential future improvements for our system include implementing a more efficient cache mechanism and incorporating parallel computing. This would optimize the question enumeration Q in Algorithm 1, reducing computing time and enhancing the user experience. Another limitation is the potential overfitting of the questionnaire model used in our study. While the model was specifically crafted to suit our research objectives, there is a risk that it may become too tailored to the specific dataset or user context, thus compromising its generalizability. To address this limitation, future work could focus on calibrating the model with a more diverse and representative annotated dataset, preventing overfitting specific examples. Another potential improvement is to enhance the system's support for complex data patterns, such as sequential or graph data patterns. This can be achieved by integrating a more flexible fact-generation module that leverages a large language model to generate customized data facts specifically tailored to complex patterns. The complex data patterns can pose new challenges for the users to understand the options in the questionnaire. An improved interface with visual illustrations can alleviate the issue by showcasing the relevant facts.

8 CONCLUSION

We present a mixed-initiative approach for generating data stories with adaptive, machine-guided user feedback elicitation. To demonstrate the feasibility of our approach, we implemented Socrates: a prototype system with additional support for data fact exploration, navigation of the narrative flow, and flexible story editing functionalities. Evaluation results showed that Socrates generated data stories with a higher overlap of insights and greater story relevance compared to manual creation. Interviews with practitioners validated the usability and usefulness of our prototype, suggesting future directions such as integrating a story evaluation mechanism and enabling user-initiated communication.

REFERENCES

[1] F. Amini, N. Henry Riche, B. Lee, C. Hurter, and P. Irani. Understanding data videos: Looking at narrative visualization through the cinematography lens. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1459–1468, 2015. doi: 10.1145/2702123. 2702431 2, 3

[2] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani. Authoring data-driven videos with dataclips. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):501–510, 2016. doi: 10.1109/TVCG.2016. 2598647 2, 3

[3] B. Bach, N. Kerracher, K. W. Hall, S. Carpendale, J. Kennedy, and N. Henry Riche. Telling stories about dynamic networks with graph comics. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3670–3682. ACM, New York, 2016. doi: 10. 1145/2858036.2858387 2

[4] B. Bach, M. Stefaner, J. Boy, S. Drucker, L. Bartram, J. Wood, P. Ciuccarelli, Y. Engelhardt, U. Koeppen, and B. Tversky. Narrative design patterns for data-driven storytelling. In *Data-driven storytelling*, pp. 107–133. AK Peters/CRC Press, 2018. 2

[5] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, 2015. doi: 10.1109/TVCG.2015.2467732 2

[6] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185 6

[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. doi: 10.1109/TVCG.2013.124 3

[8] Y.-R. Cao, X.-H. Li, J.-Y. Pan, and W.-C. Lin. Visguide: User-oriented recommendations for data event extraction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, New York, 2022. doi: 10.1145/3491102.3517648 2

[9] Y.-R. Cao, J.-Y. Pan, and W.-C. Lin. User-oriented generation of contextual visualization sequences. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020. 2

[10] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer graphics*, 23(1):111–120, 2016. doi: 10.1109/TVCG.2016.2598468 6

[11] Q. Chen, F. Sun, X. Xu, Z. Chen, J. Wang, and N. Cao. Vizlinter: a linter and fixer framework for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):206–216, 2022. doi: 10. 1109/TVCG.2021.3114804 9

[12] Z. Chen and H. Xia. Crossdata: Leveraging text-data connections for authoring data documents. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, and K. Yatani, eds., *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2022. doi: 10.1145/3491102.3517485 1, 3

[13] Z. Chen, S. Ye, X. Chu, H. Xia, H. Zhang, H. Qu, and Y. Wu. Augmenting sports videos with viscommentator. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):824–834, 2022. doi: 10.1109/TVCG.2021. 3114806 1

[14] F. Chevalier, M. Tory, B. Lee, J. van Wijk, G. Santucci, M. Dörk, and J. Hullman. From analysis to communication: Supporting the lifecycle of a story. In *Data-Driven Storytelling*, pp. 151–183. AK Peters/CRC Press, 2018. 2

[15] N. Cohn. Visual narrative structure. *Cognitive science*, 37(3):413–452, 2013. doi: 10.1111/cogs.12016 2

[16] M. Conlen and J. Heer. Idyll: A markup language for authoring and publishing interactive articles on the web. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 977–989. ACM, New York, 2018. doi: 10.1145/3242587.3242600 2

[17] K. Cook, N. Cramer, D. Israel, M. Wolverton, J. Bruce, R. Burtner, and A. Endert. Mixed-initiative visual analytics using task-driven recommendations. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 9–16. IEEE, 2015. doi: 10.1109/VAST.2015. 7347625 2

[18] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J.-G. Lou, and D. Zhang. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):906–916, 2019. doi: 10.

[19] J. E. Cutting. Narrative theory and the dynamics of popular movies. *Psychonomic bulletin & review*, 23(6):1713–1743, 2016. doi: 10.3758/ s13423-016-1051-4 2

[20] R. Ding, S. Han, Y. Xu, H. Zhang, and D. Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 317–332. ACM, New York, 2019. doi: 10.1145/3299869.3314037 1

[21] B. Dykes. *Effective data storytelling: how to drive change with data, narrative and visuals*. John Wiley & Sons, 2019. 2

[22] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012. doi: 10. 1109/TVCG.2012.260 1, 2

[23] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435, 2014. doi: 10. 1007/s10844-014-0304-9 2

[24] A. Franklin, S. Gantela, S. Shifarraw, T. R. Johnson, D. J. Robinson, B. King, A. M. Mehta, C. L. Maddow, N. R. Hoot, V. Nguyen, A. Rubio, J. Zhang, and N. G. Okafor. Dashboard visualizations: Supporting real-time throughput decision-making. *J. Biomed. Informatics*, 71:211–221, 2017. doi: 10.1016/j.jbi.2017.05.024 1

[25] N. Gershon and W. Page. What storytelling can do for information visualization. *Communications of the ACM*, 44(8):31–37, 2001. doi: 10. 1145/3299869.3314037 2

[26] D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pp. 315–324. ACM, New York, 2009. doi: 10.1145/1502650. 1502695 2

[27] N. Hariri, B. Mobasher, and R. Burke. Context-aware music recommendation based on latenttopic sequential patterns. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, pp. 131–138. ACM, New York, 2012. doi: 10.1145/2365952.2365979 2

[28] M. Hinne, M. van der Heijden, S. Verberne, and W. Kraaij. A multidimensional model for search intent. In *Proceedings of the Dutch-Belgium Information Retrieval workshop (DIR 2011)*, pp. 20–24, 2011. 4

[29] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2707–2716. ACM, New York, 2013. doi: 10.1145/2470654.2481374 3

[30] J. Hullman, S. Drucker, N. H. Riche, B. Lee, D. Fisher, and E. Adar. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics*, 19(12):2406–2415, 2013. 2, 3, 5

[31] D. Jannach, A. Manzoor, W. Cai, and L. Chen. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021. doi: 10.1145/3453154 2, 3

[32] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, vol. 37, pp. 18–28. ACM New York, NY, USA, 2003. 4

[33] N. W. Kim, N. Henry Riche, B. Bach, G. Xu, M. Brehmer, K. Hinckley, M. Pahud, H. Xia, M. J. McGuffin, and H. Pfister. Datatoon: Drawing dynamic network comics with pen+ touch interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019. 2, 3

[34] C. N. Knaflic. *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons, 2015. 1

[35] R. Kosara and J. Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013. doi: 10.1109/MC.2013.36 1, 2, 3

[36] S. Latif, S. Chen, and F. Beck. A deeper understanding of visualization-text interplay in geographic data-driven stories. In *Computer Graphics Forum*, vol. 40, pp. 311–322. Wiley Online Library, 2021. 1

[37] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2021. 2

[38] B. Lee, N. H. Riche, P. Isenberg, , and S. Carpendale. More than telling a story: A closer look at the process of transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 2015. 2

[39] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5):84–90, 2015. 1, 2

[40] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran. Avoid-

ing drill-down fallacies with vispilot: Assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 186–196, 2019. 4

[41] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 304–312, 2020. 5

[42] W. Lei, G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T.-S. Chua. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2073–2083, 2020. 3

[43] Y. Lu, J. Bao, Y. Song, Z. Ma, S. Cui, Y. Wu, and X. He. Revcore: Review-augmented conversational recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1161–1173. ACL, 2021. doi: 10.18653/v1/2021.findings-acl.99 3

[44] D. Meignan, S. Knust, J.-M. Frayret, G. Pesant, and N. Gaud. A review and taxonomy of interactive optimization methods in operations research. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(3):1–43, 2015. doi: 10.1145/2808234 1

[45] K. Miettinen. *Nonlinear multiobjective optimization*, vol. 12. Springer Science & Business Media, 1999. 4

[46] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 19–34, 2018. 2

[47] F. Narducci, P. Basile, M. de Gemmis, P. Lops, and G. Semeraro. An investigation on the user interaction modes of conversational recommender systems for the music domain. *User Modeling and User-Adapted Interaction*, 30:251–284, 2020. 3

[48] B. V. Nguyen and M.-Y. Kan. Functional faceted web query analysis. In *WWW2007: 16th International World Wide Web Conference*, 2007. 4

[49] A. Pandey, A. Srinivasan, and V. Setlur. Medley: Intent-based recommendations to support dashboard composition. *IEEE Transactions on Visualization and Computer Graphics*, 1912. 9

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. ACL, Philadelphia, 2002. doi: 10.3115/1073083.1073135 7

[51] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe. Chartaccent: Annotation for data-driven storytelling. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 230–239. Ieee, 2017. 2

[52] N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-driven storytelling*. CRC Press, 2018. 2

[53] L. Rundo, R. Pirrone, S. Vitabile, E. Sala, and O. Gambino. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *J. Biomed. Informatics*, 108:103479, 2020. doi: 10.1016/j.jbi.2020.103479 1

[54] A. Satyanarayan and J. Heer. Authoring narrative visualizations with ellipsis. In *Computer Graphics Forum*, vol. 33, pp. 361–370, 2014. 3

[55] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179 2, 3

[56] D. Shi, Y. Guo, M. Guo, Y. Wu, Q. Chen, and N. Cao. Talk2data: High-level question decomposition for data-oriented question and answering. *arXiv:2107.14420*, 2021. 3

[57] D. Shi, F. Sun, X. Xu, X. Lan, D. Gotz, and N. Cao. Autoclips: An automatic approach to video generation from data facts. In *Computer Graphics Forum*, vol. 40, pp. 495–505, 2021. doi: 10.1111/cgf.14324 1, 2

[58] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):453–463, 2020. doi: 10.1109/TVCG.2020. 3030403 1, 2, 3, 4, 5, 6, 7, 9

[59] D. Showkat and E. P. Baumer. Where do stories come from? examining the exploration process in investigative data journalism. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–31, 2021. doi: 10.1145/3479534 2

[60] F. Sperrle, A. Jeitler, J. Bernard, D. Keim, and M. El-Assady. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics*, 100:93–105, 2021. doi: /10.1016/j.cag.2021.06.016 6

[61] M. Sun, L. Cai, W. Cui, Y. Wu, Y. Shi, and N. Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):983–993, 2022. doi: 10.1109/TVCG.2022. 3209428 1, 2, 3, 4

[62] R. Swanson and A. S. Gordon. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–35, 2012. 1

[63] T. Tang, R. Li, X. Wu, S. Liu, J. Knittel, S. Koch, T. Ertl, L. Yu, P. Ren, and Y. Wu. Plotthread: Creating expressive storyline visualizations using reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):294–303, 2020. 9

[64] E. Van Den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *Siam journal on scientific computing*, 31(2):890–912, 2009. 4

[65] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert. Podium: Ranking data using mixed-initiative visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):288–297, 2017. 2

[66] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905, 2019. 1, 2, 3

[67] S. Whiting, K. Zhou, J. Jose, and M. Lalmas. Temporal variance of intents in multi-faceted event-driven information needs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 989–992, 2013. 4

[68] W. Wojtkowski and W. G. Wojtkowski. Storytelling: its role in information visualization. In *European Systems Science Congress*, vol. 5, pp. 1–5, 2002. 2

[69] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2648–2659. ACM, New York, 2017. doi: 10.1145/3025453.3025768 3

[70] A. Wu, Y. Wang, M. Zhou, X. He, H. Zhang, H. Qu, and D. Zhang. Multivision: Designing analytical dashboards with deep learning based recommendation. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):162–172, 2022. doi: 10.1109/TVCG.2021.3114826 1, 3, 9

[71] Z. Xie, T. Yu, C. Zhao, and S. Li. Comparison-based conversational recommender system with relative bandit feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1400–1409, 2021. 5

[72] B. Xu, J. Bu, C. Chen, C. Wang, D. Cai, and X. He. Emr: A scalable graph-based ranking model for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):102–114, 2013. 2

[73] L. Yang, X. Xu, X. Lan, Z. Liu, S. Guo, H. Shi, H. Qu, and N. Cao. A design space for applying the freytag's pyramid structure to data stories. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):922–932, 2021. 2, 3, 5

[74] S. Zhang and K. Balog. Entitables: Smart assistance for entity-focused tables. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 255–264, 2017. 3

[75] Y. Zhang, X. Chen, Q. Ai, L. Yang, and W. B. Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 177–186. ACM, New York, 2018. 1, 2

[76] Y. Zhang, L. Wu, Q. Shen, Y. Pang, Z. Wei, F. Xu, B. Long, and J. Pei. Multiple choice questions based multi-interest policy learning for conversational recommendation. In *Proceedings of the ACM Web Conference 2022*, pp. 2153–2162, 2022. 5

[77] J. Zhao, S. Xu, S. K. Chandrasegaran, C. Bryan, F. Du, A. Mishra, X. Qian, Y. Li, and K. Ma. Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1384–1399, 2023. doi: 10.1109/TVCG.2021.3114211 1, 2

[78] C. Zheng and X. Ma. Evaluating the effect of enhanced text-visualization integration on combating misinformation in data story. In *15th IEEE Pacific Visualization Symposium, PacificVis 2022, Tsukuba, Japan, April 11-14, 2022*, pp. 141–150. IEEE, 2022. doi: 10.1109/PacificVis53943.2022.00023 1

[79] C. Zheng, D. Wang, A. Y. Wang, and X. Ma. Telling stories from computational notebooks: Ai-assisted presentation slides creation for presenting data science work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2022. doi: 10.1145/3491102.3517615 1, 3