Int. J. Human–Computer Studies 192 (2024) 103359

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Impact of interaction technique in interactive data visualisations: A study on lookup, comparison, and relation-seeking tasks

Niels van Berkel [a,*], Benjamin Tag [b], Rune Møberg Jacobsen [a], Daniel Russo [a], Helen C. Purchase [b], Daniel Buschek [c]

[a] Aalborg University, Denmark
[b] Monash University, Australia
[c] University of Bayreuth, Germany

ARTICLE INFO

Dataset link: https://osf.io/ce46r/?view_only=6ef59ba87bbb4861b54bc8a2bf3c1309

Keywords:
Data visualisation
Interaction technique
Data exploration
Accuracy
Confidence
Cognitive load

ABSTRACT

This paper presents an analysis of different interaction techniques used in interactive data visualisations to support end-users in visual analytics tasks. Our selection of interaction techniques is based on prior work and consists of the interaction techniques Select, Explore, Reconfigure, Encode, Filter, Abstract/Elaborate, and Connect. Through a within-subject study, we assessed participants' abilities to utilise these techniques when faced with three distinct types of data-driven tasks; lookup, comparison, and Relation-seeking. Our research investigates the impact of these interaction techniques on the correctness, confidence, perceived difficulty, and cognitive load of N = 80 self-identified data scientists and N = 80 non-experts. We find that interaction technique significantly impacts answer correctness and participant confidence. Participants performed best across those interaction techniques that allow for information that is deemed least relevant to be concealed, which is reflected in lower intrinsic and extraneous cognitive load. Interestingly, participants' expertise affected their confidence but not their accuracy. Our results provide insights useful for a more targeted and informed design and usage of interactive data visualisations.

1. Introduction

Data plays a substantial role in everyday life. From making sense of epidemic disease risk factors (Yang et al., 2022), trying to grasp the reasoning of decision recommendation systems (Hoque and Mueller, 2022), to everyday interactions with self-tracking technologies (Kuosmanen et al., 2022), we rely on data to understand the world surrounding us and inform our actions. While the growing reliance on data in decision-making has led to increasingly advanced tools for domain experts (see e.g., Bird et al., 2020; Bäuerle et al., 2022; Saleiro et al., 2018), the complexity of such tools often renders them unsuitable for use by anyone but trained experts. Despite this increase in expert tooling, the fundamental question of how to support the general public in engaging with the data and logic underlying much of today's decision-making systems remains unanswered. Prior work has highlighted the growing concerns among citizens regarding algorithm-driven decision-making while outlining how a lack of understanding can quickly result in a decrease in trust (Woodruff et al., 2018). Given this desire of, and need for, the general public to engage with data, we are particularly interested in how interactive visualisations can foster understanding, pattern discovery, and the ability to answer specific questions—thereby supporting sense-making (Heer and Shneiderman, 2012).

In this paper, we set out to assess different interaction techniques, as encountered in interactive data visualisations, and part of the grand challenges of immersive analytics (Ens et al., 2021). We selected seven established interaction techniques based on prior work (Lu et al., 2017) and assessed their ability to support end-users in visual analytics tasks. These interaction techniques called Select, Explore, Reconfigure, Encode, Filter, Abstract/Elaborate, and Connect, each allow for specific ways to engage with appropriate data visualisation. Through a within-subject study, we assessed participants' abilities to utilise these techniques when faced with an unknown dataset and different data-driven tasks. We compared the perceived task difficulty, confidence in their answers, self-assessed cognitive load, and participants' reflections and preferences, of an expert sample (data scientists) and a non-expert sample, with the particular intention of eliciting differences between them.

Our findings show that interaction technique significantly affects participants' correctness. We find that participants performed best across those interaction techniques that allow for the concealing of information deemed least relevant, as reflected in a lower perceived

* Corresponding author.
E-mail address: nielsvanberkel@cs.aau.dk (N. van Berkel).

https://doi.org/10.1016/j.ijhcs.2024.103359
Received 7 July 2023; Received in revised form 3 June 2024; Accepted 16 August 2024
Available online 31 August 2024

intrinsic and extraneous cognitive load. While the samples' error rates were near-identical, our non-expert participants experienced lower levels of confidence and higher levels of perceived extraneous cognitive load compared to the recruited data scientists. Overall, our results clearly identify a need for interactive tools designed to meet the general public's needs. While such tools will likely be application- and context-dependent, our results highlight that particular and well-designed interaction techniques are more suitable for supporting this target group. This provides clear direction for designers to integrate interactive guidance into their interfaces, especially in data-heavy contexts. Our work addresses the growing need to consider novice end-users when developing and inspecting data-driven systems and has strong implications for a variety of domains, e.g., informing interactive data visualisations for explainable AI applications or health-related information dashboards.

2. Related work

When facing any set of data, obtaining a sufficient level of understanding of that data is necessary to justify and inform decisions, enable user control, and allow for meaningful discovery. While explanations of data can be automated or mathematically presented, it is crucial to develop human-friendly presentations and explanations. Visualisation is often named the most human-centred explanation technique (Adadi and Berrada, 2018). However, it requires that they are not only visually attractive but also understandable by the user (Adadi and Berrada, 2018). Based on research in social science, we know that an explanation requires interaction between an explainer and the explainee, or as Miller (Miller, 2019) states: "*Explanations are social — they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs.*" This notion supports the extensive work of the visualisation community towards interactive visual analytics and is a key driver for this work. In the following, we give an overview of recent works in Visual Analytics and explain why we have to consider non-expert end-users when developing data analysis tools.

2.1. Interactive visual analytics

Visual analytics is defined as "*[combining] automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.*" (Keim et al., 2008). Here, Keim et al. stress the integral role that visual analytics plays in decision-making. Prior works have explored a wide variety of interactive approaches to support end-users in making sense of data. For example, Perin et al. present a soccer table application that provides two interaction modes which prevent interruption of users' temporal explorations of the presented data (Perin et al., 2014). The researchers found that these two interaction techniques allow users to maintain their focus on the exploration task. Hoque et al. explore the use of a natural language interface for visual analytics and found that users appreciated the ability to focus on obtaining an answer rather than interacting with an interface (Hoque et al., 2018). Schwab et al. explored pan and zoom interactions, finding that user performance can be increased by for example manipulating the visual presentation and positioning of interface elements (Schwab et al., 2019). Through an online crowdsourcing study, Mosca et al. find that interaction does not always improve accuracy in data interpretation tasks — suggesting that a well-designed static visualisation can be more effective than an interactive visualisation (Mosca et al., 2021).

Recent works have focused on interactive data visualisations in the AI domain Yu et al. (2020) and van Berkel et al. (2021). Yu et al. present an interactive visualisation approach that aims to support designers in exploring AI algorithms' trade-offs (Yu et al., 2020). The trade-offs are visualised and categorised under a model family. The results of the study show that this approach helps users better understand but also operate and manage the trade-offs between task goals and AI algorithms. In a different study, Van Berkel et al. investigate how two visualisation techniques impact the perception of fairness of AI algorithms in a general public sample (van Berkel et al., 2021). Their work shows that different visualisation techniques significantly influence the level of perceived fairness among non-expert users. While visualisation is one of the main tools for exploring complex data and its relationships, interactive visualisation additionally enables interactive explorations of such data (Tominski, 2015).

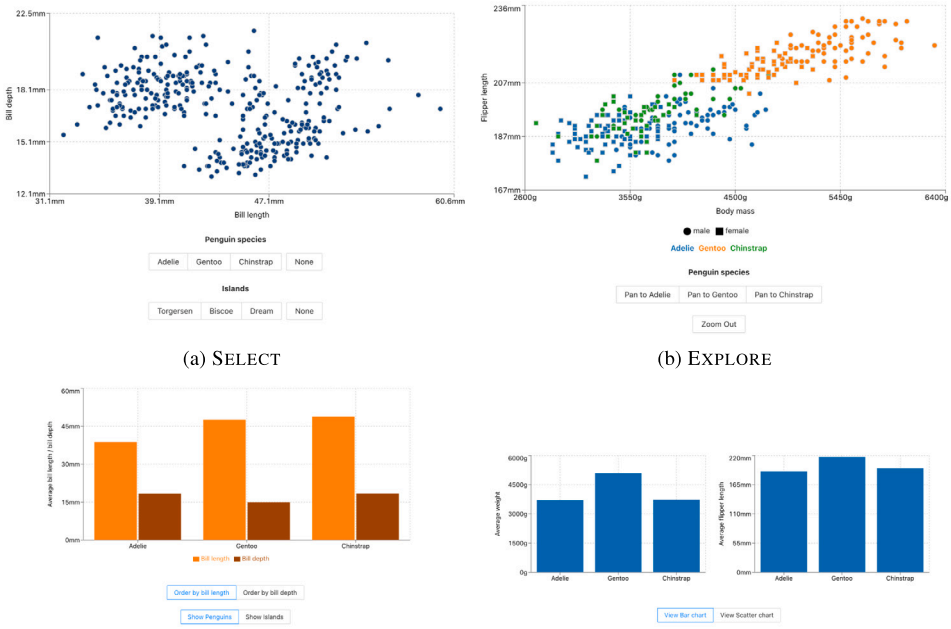
Our work contributes to the literature on interactive visual analytics through a comprehensive study of how people make sense of data and perceive the interaction with it, using seven different interaction techniques and their impact on the users' cognitive load. We chose these interaction techniques, Select, Explore, Reconfigure, Encode, Filter, Abstract/Elaborate, and Connect, as proposed by Yi et al. (2007).

2.2. Contrasting experts and the general public

In our daily interactions with electronic devices, we produce vast amounts of (personal) data. This data is increasingly processed by intelligent systems to analyse our usage patterns, generate recommendations, and create user profiles (Kugler, 2018). Interactive visual analytics has been employed to make these complex data interactions more accessible (Blackwell et al., 2018; Wanner et al., 2021). For example, Blackwell et al. conducted a series of four case studies examining how interface design can support domain experts in feeling more under control by increasing their perceived agency and confidence (Blackwell et al., 2018). Their approach aims at making the automated system appear less intelligent. They found that even domain experts often struggle with fully understanding automated decisions, further stressing the need for a more user-centred approach. To investigate how different levels of transparency influence confidence in automated decision-making, Wanner et al. propose a study comparing the impact of black-box, grey-box, and white-box systems on domain experts' levels of confidence (Wanner et al., 2021). While there has been a lot of focus on making intelligent systems more accessible and user-centred, research is not yet at the point to fully satisfy the requirements.

One contemporary use of visual analytics is in the application of AI, with the increasingly bigger role of this technology in society driving the need for transparency and understanding (Adadi and Berrada, 2018). This is especially important for non-expert users, as these users cannot only rely on AI experts to assess the quality and reasoning of AI-driven decisions in their everyday lives (Liao et al., 2020). However, as the ones primarily responsible for developing AI systems, AI engineers have been the main focus of these efforts, with non-expert users consistently overlooked (Chatzimparmpas et al., 2020). Consequently, recent efforts have stressed the importance of involving members of the general public in the development and assessment of AI systems (Liao et al., 2020; Yu et al., 2020; van Berkel et al., 2021).

Cheng et al. studied design principles that aim at supporting explainability to better provide users and other stakeholders with explanations (Cheng et al., 2019). In an online experiment they looked at the objective and subjective understanding of university admission decisions made by an algorithm, among the participants. The authors found that interactive as well as white-box explanations can significantly improve understanding of the decision-making. In an interview study with UX and design practitioners, Liao et al. identify gaps in current explainable AI (XAI) practices (Liao et al., 2020). The authors note that developers of XAI applications must consider that different user skill levels present different needs for data presentation. Therefore, the recent call for action to involve the (non-expert) end-user in XAI research seems to be a logical consequence. Jin et al. conducted a literature review of 59 papers, and developed an *end-user-friendly XAI taxonomy* based on features attributes, instance, and decision rules (Jin



(a) SELECT (b) EXPLORE

(c) RECONFIGURE (d) ENCODE

Fig. 1. Overview of four of the seven interaction techniques compared in our study. Continued in Fig. 2.

et al., 2019). Based on this, they propose a user-friendly vocabulary to help generate explanations for different user types.

A promising approach to making the understanding of AI systems more accessible was recently published by Wexler et al. (2020). Their team developed an open-source tool that enables users to better understand AI systems and the data used to build them. Through performance tests, visualisation, and analysis of different data features and their impact on the model performance, users can interact and develop a better understanding of an AI system. This is where our work is located. By investigating how data scientists (expert users) and members of the general public interact with data, we contribute new knowledge to the development of user-centred interactive visual analytics.

3. Implementation of interaction techniques

Here we outline the various interaction techniques evaluated in this study, following the categorisation of interaction techniques by Yi et al. (2007). In their hallmark paper, Yi et al. introduce seven distinct techniques encountered in interactive information visualisations (Yi et al., 2007). Within the scope of this paper and in line with Yi et al. (2007), interaction techniques represent *user intents* as opposed to a particular input modality (*e.g.*, touch, tactile) or technology (*e.g.*, AR).

For all interaction techniques implemented, we aim to follow established practice in data visualisation, including; consistent use of 2D rather than 3D visualisations, avoiding dual (y-)axis charts, and labelled axes—see, for example, Tufte (Tufte, 2001). All user actions are reversible to allow for free exploration. Finally, in designing these interactions we explicitly aimed to establish visual consistency between the interaction techniques. The seven interaction techniques are shown in these Figs. 1 and 2.

The interaction techniques were implemented using React and built on the Recharts charting library.[1] We release our implementation of the various interaction techniques as open source to support their further assessment and expansion in the literature.[2]

The SELECT interaction technique allows users to tag data points of interest — thereby making it easier to visually distinguish them from the other data points (Yi et al., 2007). This is especially valuable when many data points are shown in one view and the user objective is to distinguish data points between different categories. Our implementation, as shown in Fig. 1(a), consists of a scatter plot, commonly used in algorithmic inspection applications (van Berkel et al., 2021; Wexler et al., 2020). Upon selection of a categorical variable by the participant, the relevant scatter points change colour and outline to highlight the selection — similar to the marking of data items as presented in Yi et al. (2007). We allow for the combined selection of up to two categorical variables using the 'AND' logical operator.

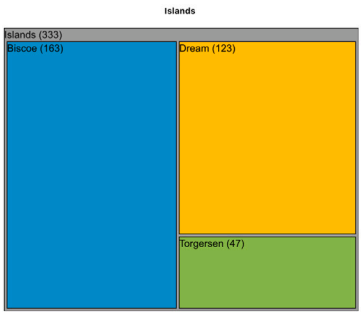
EXPLORE interactions enable users to focus on a subset of items included in a visualisation — typically by removing other data points from view. This is motivated by factors that limit users' ability to inspect the data points relevant to them, for example, limited screen size or the sheer size of the dataset (Lu et al., 2017). Our implementation, displayed in Fig. 1(b), enables users to zoom in and pan to different subsets of the dataset following the selection of a categorical variable. Yi et al.'s survey reports 'panning' to be the most common implementation of the EXPLORE interaction.

The RECONFIGURE interaction supports the user in changing the spatial arrangement of the data shown. This rearrangement may help the user in obtaining new insights (Yi et al., 2007). In our implementation, as presented in Fig. 1(c), the user can reconfigure the data in two distinct ways. First, the user can reorder the data in ascending and descending order based on two displayed numerical variables — referred to by Yi et al. as sorting and rearranging (Yi et al., 2007). Second, the user can change the grouping variable (x-axis) between two categorical variables — labelled by Yi et al. as changing the attributes presented (Yi et al., 2007). The underlying data remains identical in all the reconfigurations, but the presentation is reconfigured to enable the user to obtain different insights.

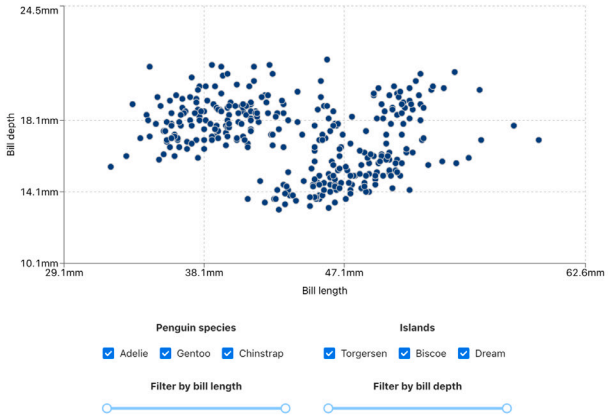
ENCODE interactions allow users to "*fundamentally alter the visual representation of the data*" (Lu et al., 2017). In changing the visualisation shown, the user can make use of the unique strengths of different visualisations in building an understanding of the data. In our implementation of ENCODE, we present participants with two different visualisations: a bar chart and a scatterplot. The underlying data is

[1] https://recharts.org/

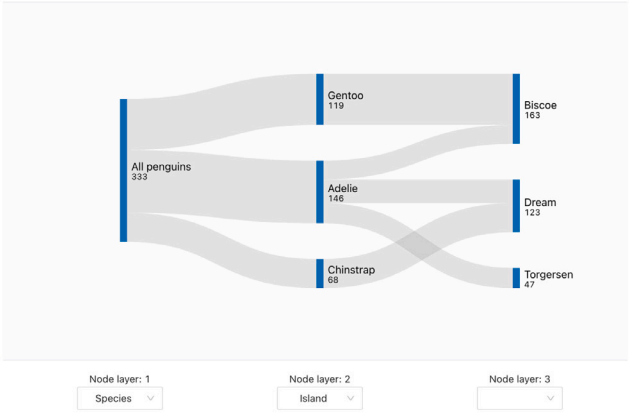
[2] https://osf.io/ce46r/?view_only=6ef59ba87bbb4861b54bc8a2bf3c1309



(a) ABSTRACT/ELABORATE



(b) FILTER



(c) CONNECT

Fig. 2. Overview of three of the seven interaction techniques compared in our study. Continuation from Fig. 1.

kept identical between the two visualisations, with the bar displaying average values and the scatterplot displaying all individual data points.

The ABSTRACT/ELABORATE interaction enables a user to obtain an overview of the data at different levels of granularity. This provides a way to understand both the distribution of data at a high level (ABSTRACT) as well as to unfold the dataset to obtain details on specific elements within the data (ELABORATE) (Lu et al., 2017). Our implementation consists of an interactive treemap in which the size of each node is proportional to the size of the category it represents (Shneiderman, 1992), as shown in Fig. 2(a). Yi et al. highlight the treemap as an example of a technique which allows for details-on-demand operations and therefore fits well to the ABSTRACT/ELABORATE interaction (Yi et al., 2007). Participants can navigate through the treemap by clicking on any of the nodes displayed, which brings them one step 'deeper' into the tree. Further, a breadcrumb navigation is provided to directly navigate up the tree.

FILTER interactions enable the user to conditionally hide or show data items. While the EXPLORE and FILTER interaction techniques both exclude data from the view, in FILTER interactions, the hiding of data is based on a (user-configured) condition, whereas the EXPLORE interaction hides items based on a limit introduced in the display of information (Lu et al., 2017). Our FILTER interaction technique, shown in Fig. 2(b), consists of a scatter chart with multiple filter options; two sets of multiple-choice checkboxes to show or hide categorical variables and two sliders to limit the numerical values included in the data display. Through this combination of filters, participants can precisely alter the conditions under which data is shown.

The CONNECT interaction displays the relationships between entities within a dataset (Yi et al., 2007). Our implementation, as displayed in Fig. 2(c), consists of an interactive alluvial plot. Alluvial plots show the distribution of data between and across different categorical variables, with the height of the vertical bars representing the volume of the data. Participants can select between one to three categorical variables to be shown as part of the alluvial plot. Each categorical variable can be selected only once, with our system automatically updating both the visualisation and the remaining available selection options.

As can be seen in Figs. 1 and 2, our seven interaction techniques follow a variety of visual representations, including barcharts, scatterplots, and other representations. We argue that there is no possible or practical way to unify these interaction techniques in one visual representation. Therefore, we sought to present prototypical implementations of each interaction technique, as inspired by existing literature and practice.

4. Method

As people encounter data in everyday situations with increasing frequency, understanding how to best support them in interpreting this information is crucial. To explore how different interaction techniques can aid laypeople in comprehending complex data, we designed and conducted a randomised within-subject study. This study assessed participants' performance in visual analytics tasks, focusing on how various interaction techniques enhance their ability to make sense of everyday data.

4.1. Materials

To capture the effect of the different interaction techniques on users' behaviours and perceptions, we collected data on both the task- and interaction technique levels. Prior to the actual study, all participants were asked to provide **demographic information**, such as age, sex, educational attainment, and experience with data visualisation.

To assess the understanding of the different tasks for each interaction technique, we analysed a set of performance measures. First, we looked at participants' **error rate** by comparing their answers to the given task with the ground truth (as established separately by the paper's authors). Second, we assessed participants' **efficiency** by recording the completion time on each individual question (*i.e.*, time between task loading time and task submission). Third, to better understand the **ease of interaction** while keeping participant burden at a minimum, we integrated two 7-point Likert scale items that were triggered after the completion of each of the three tasks for each interaction technique. Participants were asked to rate (1) their **confidence** ('Overall, how confident are you that you completed the task successfully?' [Not at all confident–Extremely confident]) and the **overall difficulty** on the (2) Single Ease Question (SEQ) (Locascio et al., 2016) ('Overall, this task was?' [Very easy–Very difficult]). The SEQ, while simple and low in demand on the participant, performs equal if not better than more complex measures such as the Subject Mental Effort Questionnaire (SMEQ) or the Usability Magnitude Estimation (UME) (Sauro and Dumas, 2009). Fourth, to better understand the impact that each interaction technique has on users' perception, we used a questionnaire to measure intrinsic, extraneous, and germane cognitive load (Klepsch et al., 2017) after the completion of each interaction technique (i.e., after each set of three tasks). While intrinsic (related to the information presented) and extraneous cognitive load (related to the given instructions) refer to the characteristics of the presented material and interaction, germane cognitive load (the effort to build functioning mental models) provides insights into user characteristics (Sweller et al., 1998; Sweller, 2010; Paas and van Gog, 2006). The user has no active control over their germane cognitive load, as it is "*purely a function of the working memory resources devoted to the interacting elements that determine intrinsic cognitive load.*" (Sweller et al., 1998). Consequently, a higher germane cognitive load indicates that the user devotes a larger share of their working memory to learning. However, when instructions are not presented in an organised and comprehensible way, users have to devote a larger share of working memory to the extraneous demand, limiting the resources available for learning (Sweller, 2010). Finally, we concluded the data collection by asking each participant to reflect on their use of the presented interaction techniques and fill in a free text field.

4.1.1. Tasks

To evaluate the seven interaction techniques, we created a set of visual inspection tasks. Each task consisted of a single question to which users had to find the answer by using one of the interaction techniques. A central goal in drafting the tasks was to ensure as much similarity as possible in the questions' difficulty. This was done to ensure that we do not measure the challenge of the task itself, but rather capture the participants' ability and experience of utilising the interaction techniques.

To create the tasks in a structured and consistent way, we follow the Andrienko Task Framework (ATF) (Andrienko and Andrienko, 2006). The ATF provides a typology of data analysis tasks focusing on the data's *characteristic* and *referential* components: Characteristic components are observations or measurements (*e.g.*, time in seconds), whereas referential components specify the context of observation or measurement (*e.g.*, a place) (Andrienko and Andrienko, 2006). ATF specifies three task types (Andrienko and Andrienko, 2006; Kerracher et al., 2015);

- *Lookup*: identify a characteristic given a reference (known as direct lookup) or identify a reference given a characteristic (known as inverse lookup).
- *Comparison*: contrast the relation between two or more components; either between characteristics (known as a direct comparison) or between references (known as an inverse comparison).
- *Relation-seeking*: determine the components associated with a defined relation. For example, such a relation could be "*a 20% increase between subsequent days*".

Finally, ATF distinguishes between elementary tasks, which "*refer to individual elements of the reference set*" (Andrienko and Andrienko, 2006), and synoptic tasks, which "*involve the whole reference set or its subsets*" (Andrienko and Andrienko, 2006). We solely considered elementary tasks. For each of the seven interaction techniques, we included one task of each of the three task types — totalling 21 questions. All of the questions were single-choice, i.e., only one answer among a selection of three choices was accepted. Table 1 shows the three questions for the *Select* interaction. We provide a complete overview of all questions and correct answers in Table 5.

The ATF specifies the task types formally and includes multiple variants per type. To ensure high consistency across our tasks we thus decided on the following additional 'rules' for our realisation of the three ATF types: For *Lookup* tasks, our questions ask to look up a referent (*e.g.*, 'which species' in Table 1, first row) given a characteristic (*e.g.*, 'bill length') and a specific absolute value (*e.g.*, '61 mm'). For *Comparison* tasks, our questions ask for a referent from a set (*e.g.*, 'Which of the three islands' in Table 1, second row) given a characteristic (*e.g.*, number of Gentoo penguins) that needs to be compared between the specific referents in the set (*e.g.*, here: set of islands). For *Relation-seeking* tasks, our questions ask for a referent from a set (*e.g.*, 'which species' in Table 1, third row) given a relation defined via relative statements about (multiple) characteristics (*e.g.*, *deepest* bill).

4.1.2. Dataset

Following our intention to assess interaction techniques, we purposefully chose a dataset unlikely to rouse any emotion in the participants. We selected the 'Palmer penguins' dataset (Horst et al., 2020) given the variety and distribution of its variables, allowing for a wide range of questions while maintaining the same context throughout the experiment. The dataset contains a variety of characteristics (*e.g.*, species type, weight) as collected from penguins on three islands in the Palmer Archipelago, Antarctica. Following the removal of penguins with missing data, our dataset consists of 333 penguins.

4.2. Design

The study follows a within-subject design, with each of the seven interaction techniques presented to each participant in randomised order. For each technique, participants were asked to complete three distinct task types—as described in Section 4.1.1. These two variables, *interaction technique* and *task type*, are the independent variables in this study.

The primary dependent variables are *participants' correctness* on the task (correct or incorrect), *participants' self-reported confidence* in their answer, and their *perceived difficulty* of the task. We furthermore record participants' *task completion time* and their self-reported *cognitive load*. The measures used to collect these variables are detailed in Section 4.1.

4.3. Procedure

Participants first completed the three tasks for each of the seven interaction techniques before proceeding to the next technique. The interaction techniques and the three tasks included with each technique were presented in randomised order. Following the completion of all 21 tasks, participants were asked to complete the Short Graph Literacy

Table 1
Task type and respective questions as presented to participants in the Select interaction.

Task type	Question
Lookup	To which species does the penguin with a bill length closest to 61mm belong?
Comparison	Which of the three islands contains most Gentoo penguins?
Relation-seeking	To which species does the penguin with the deepest bill on the Dream island belong?

Table 2
Demographic overview of the two study samples.

	General public	Data scientists
Women	40	40
Men	40	40
Age range	19–51	19–54
Mean age (SD)	26.40 (6.9)	29.74 (7.8)

(SGL) scale (Okan et al., 2019). The SGL is a validated four-item questionnaire used to assess an individual's ability to understand and interpret information presented in graphical form. As a final task, we asked two open-ended questions to gather qualitative feedback on (1) what helps our participants make sense of data (*e.g.* charts, visualisations, and styling) and (2) potential specific tools or interactions to better make sense of data.

4.4. Participants

We recruited our participants through Prolific Academic. We restricted participation to crowd-workers with an acceptance rate of 95% or above. We excluded mobile or tablet users from participating to ensure sufficient screen real estate. Following acceptance of the task, we routed participants to our study website. We compensated each participant with a fixed compensation of $3.50. With an expected task completion time of 22 min, as based on pilot tests, this comes down to an hourly compensation of $9.50.

We computed an *a prior* power analysis to minimise type II errors and determine the sample size as follows. We follow a medium effect size ($f^2 = 0.15$), an alpha level of 0.05, and a statistical power of 0.95. With four predictors (interaction technique, task type, participant role, and SGL score), G*Power's power calculation prescribes a minimum sample size of 129 (Faul et al., 2009). We recruited a total of 160 participants: 80 participants were recruited without additional requirements, and the remaining 80 were recruited among Prolific participants with 'Data scientist' listed as their current business role. All participants were limited to taking part only once in this study.

5. Results

Our study sample was balanced evenly between male and female participants, both across members of the general public and data scientists (as per Prolific's 'balanced study distribution'). The average age of our participants is 28.02 years (SD = 7.6). We provide an overview of these demographics in Table 2, as split between members of the general public and data scientists. Participants were recruited without any restrictions on their country of residence or nationality, resulting in an international sample. The most represented countries include South Africa, Poland, Portugal, and the United Kingdom.

The mean completion time for an individual task was 48.36 s (SD = 66.62). The mean completion time was similar between participants from the general public and data scientists at respectively 48.67 and 48.05 s. We report the mean completion time per interaction technique in Table 3. A Kruskal–Wallis test finds a significant difference in completion time between the different interaction techniques ($\chi^2(6) = 207.86$, $p < 0.001$). Follow-up with multiple pairwise comparisons using Bonferroni correction indicates that task completion time for Filter was significantly longer than all other techniques. Completion time on Connect was significantly longer as compared to Explore, Encode, and Abstract/Elaborate. Task completion for Reconfigure tasks took significantly longer than Explore, Encode, and Abstract/Elaborate. Both Select and Explore were significantly slower than Encode and Abstract/Elaborate. Finally, task completion for Encode was significantly slower as compared to Abstract/Elaborate.

Table 3
Average completion time and task error rate between the different interaction techniques.

Interaction technique	Avg. completion time	Correctness
Select	49.9 s	86.7%
Explore	45.6 s	79.2%
Reconfigure	52.2 s	76.5%
Encode	40.0 s	92.9%
Abstract/Elaborate	37.2 s	74.2%
Filter	59.4 s	92.3%
Connect	54.3 s	94.0%

Next, a Kruskal–Wallis test indicates a significant difference in completion time between the three task types ($\chi^2(2) = 143.39$, $p < 0.001$). Follow-up with multiple pairwise-comparison using Bonferroni correction indicates that task completion time for Relation-seeking tasks (mean 54.9 s) was significantly longer than the completion time for both the Comparison (49.4 s) and Lookup (40.8 s) tasks, with $p = 0.041$ and $p < 0.001$ respectively. We similarly find a significant difference between the completion time for the Comparison (49.4 s) and Lookup (40.8 s) tasks, $p < 0.001$.

The average Short Graph Literacy scale (SGL) score across our participant sample was 2.44 (SD = 1.10). This is in line with prior work (van Berkel et al., 2021) and slightly above an earlier assessment of USA (mean score of 2.2, $N = 492$) and German (mean = 2.0, $N = 495$) participants (Okan et al., 2019). A Mann–Whitney–Wilcoxon test found no significant difference between the SGL score among our participants from the general public (mean = 2.3) or data scientists (mean = 2.5), W = 3601, $p = 0.156$.

5.1. Correctness

We next report on participants' correctness across the given tasks. Across all tasks, a total of 85.1% of tasks were completed correctly. Surprisingly, the share of correct answers was nearly identical between participants from the general public (85.06%) and data scientists (85.12%). We report the share of correct answers per interaction technique in Table 3.

A two-way repeated measures ANOVA, using interaction technique and task type as independent variables and task correctness as dependent variable, shows a significant main effect of interaction technique ($F(6, 3339) = 29.42$, $p < 0.001$) and a significant main effect of task type ($F(2, 3339) = 4.34$, $p = 0.013$). We furthermore find an interaction effect between interaction technique and task type ($F(12, 3339) = 15.09$, $p < 0.001$).

We visualise the main effects of interaction technique in Fig. 3-A. Following this, we ran Tukey's post hoc test for multiple comparisons for interaction technique. We find that Connect resulted in significantly higher accuracy than Select, Explore, Reconfigure, and Abstract/Elaborate. Each, Filter, Select, and Encode led to significantly higher accuracy than Explore, Reconfigure, and Abstract/Elaborate. See Table 6 for all details.

The distribution of correct answers per task type is as follows: 84.2% for Comparison, 87.5% for Lookup, and 83.6% for Relation-seeking

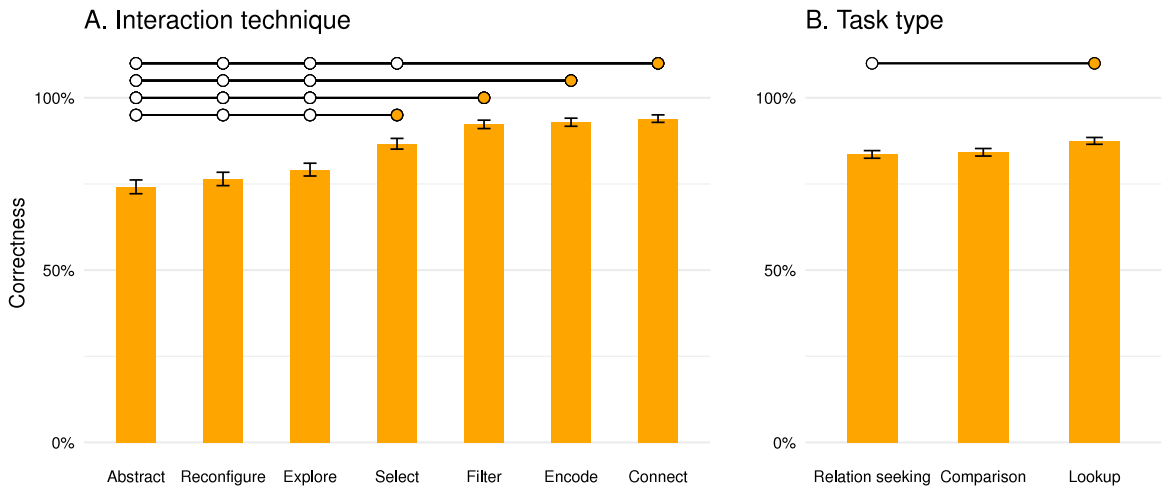


Fig. 3. Mean correctness and standard error across interaction techniques (left) and task type (right). Filled circles highlight a significant difference from the empty circles.

tasks, see Fig. 3-B. Tukey's post hoc test for multiple comparisons for task type showed that the Relation-seeking task resulted in a significantly lower accuracy than the Lookup task (see Table 7 for all comparisons).

5.2. Confidence & perceived difficulty

Next, we assess the correlation between self-reported confidence and perceived difficulty across tasks. Given the non-independence of our observations (participants completed multiple tasks), we assess the repeated measures correlation using the R package *rmcorr* (Bakdash and Marusich, 2017). We find a strong negative correlation of 0.7 between self-reported confidence and perceived difficulty ($r(3199) = -0.70$, $p < 0.01$). This follows the expectation that participants had lower confidence in completing tasks they perceived as difficult.

We conducted a two-way repeated measures ANOVA to evaluate the effects of interaction technique and task type on participants' self-reported confidence. The results indicated a significant main effect for interaction technique ($F(6, 3339) = 10$, $p < 0.001$) and task type ($F(2, 3339) = 27.91$, $p < 0.001$). We furthermore find an interaction effect between interaction technique and task type ($F(12, 3339) = 4.42$, $p < 0.001$). We visualise the effects of interaction technique in Fig. 4-A. Following this, we ran Tukey's post hoc test for multiple comparisons for interaction technique. We find that participants were significantly more confident when using Encode as compared to Select, Explore, Reconfigure, and Abstract/Elaborate. Furthermore, we find that participants were significantly less confident when using Abstract/Elaborate in comparison with Explore, Filter, and Connect. See Table 8 for all details. Tukey's post hoc test for multiple comparisons for task type showed that participants reported significantly higher confidence in the Lookup task as compared to both the Comparison and Relation-seeking tasks (see Table 9 for all comparisons).

Subsequently, we assessed the impact of interaction technique and task type on the perceived difficulty of the tasks. A two-way repeated measures ANOVA again highlight a significant main effect for interaction technique ($F(6, 3339) = 14.65$, $p < 0.001$) and task type ($F(2, 3339) = 40.16$, $p < 0.001$). We further find an interaction effect between interaction technique and task type ($F(12, 3339) = 6.17$, $p < 0.001$). Fig. 5-A shows the perceived difficulty across the seven evaluated interaction techniques. Tukey's post hoc test for multiple comparisons for interaction technique highlights that Encode is perceived as significantly less difficult than all other interaction techniques. Further, Abstract/Elaborate was perceived as significantly more difficult than Explore, Reconfigure, Filter, Connect, and—as mentioned—Encode. Posthoc results for task type showed that participants perceived both the Comparison task and Relation-seeking task as significantly more difficult as compared to the Lookup task. Tables 10 and 11 outline the exact details of these posthoc tests.

Finally, we assess differences between our self-identified data scientists and non-expert participants. A Wilcoxon rank sum test showed that the mean confidence score was significantly higher for the self-identified data scientists (M = 5.86) than the non-expert participants (M = 5.55), $W = 1577588$, $p < 0.001$. Similarly, we find that our self-identified data scientists reported a significantly lower perceived difference (M = 2.63 vs M = 2.47), $W = 1335000$, $p = 0.004$.

5.3. Cognitive load

We present the distribution of the cognitive load scores in Fig. 6. Mean cognitive load scores across all tasks were respectively 3.88 (SD = 1.50) for the intrinsic cognitive load (ICL), 3.20 (SD = 1.63) for the extraneous cognitive load (ECL), and 5.49 (SD = 1.11) for the germane cognitive load (GCL). A Friedman rank sum test indicated a significant difference between the types of cognitive load reported ($\chi^2(2) = 230.44$, $p < 0.01$). Post-hoc pairwise comparisons using Conover's test with Bonferroni correction indicate that GCL is significantly higher than both ICL and ECL. Further, ICL is significantly higher than ECL, all $p < 0.01$.

We next contrast the cognitive load scores between the general public and data scientists for each of the three cognitive load types using Mann–Whitney U tests. For both ICL and GCL, no significant difference in cognitive load scores was found ($W = 150282$, $p = 0.226$ and $W = 155262$, $p = 0.775$ respectively). We find a significantly higher ECL score for the general public (mean = 3.31) as compared to the data scientists (mean = 3.10), $W = 144424$, $p = 0.022$. Finally, we evaluate the effect of interaction technique on the three types of cognitive load. Using a Friedman rank sum test, we find a significant effect for interaction technique on ICL ($\chi^2(6) = 63.35$, $p < 0.01$) and ECL ($\chi^2(6) = 56.71$, $p < 0.01$), but not for GCL ($\chi^2(2) = 230.44$, $p = 0.05$). We subsequently conduct a post-hoc pairwise comparison between the interaction techniques for ICL and ECL. Using the R package *multcompView*, we report a compact letter display of these comparisons in Table 4. Compact letter displays group treatments into groups that are not significantly different by pairwise comparisons. As an example, the Select interaction technique has a significantly different ICL score (mean 3.84, ICL group 'a' and 'b') than the Encode, Abstract/Elaborate, and Filter interaction techniques (none of which are part of ICL group 'a' or 'b').

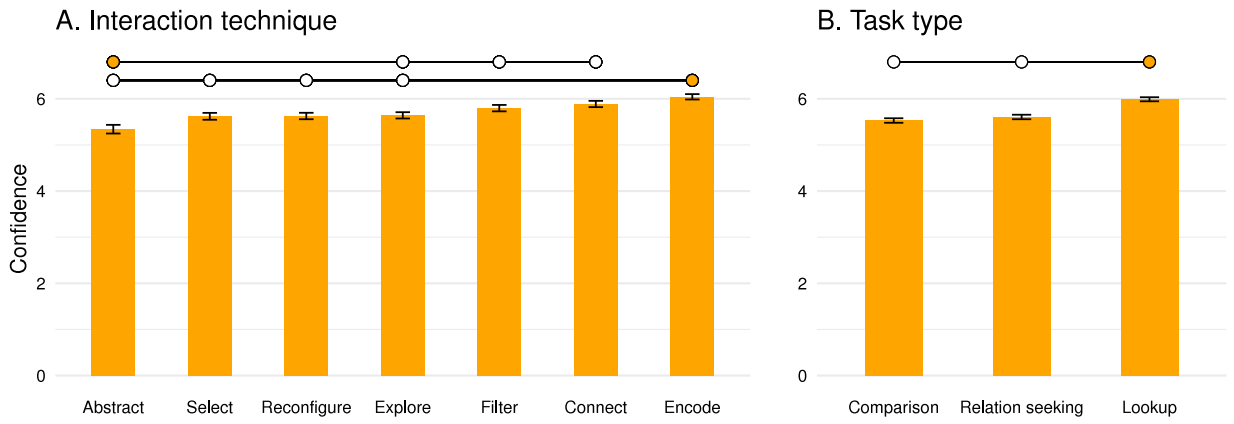


Fig. 4. Mean answer confidence and standard error across interaction techniques (left) and task type (right). Filled circles highlight a significant difference from the empty circles.

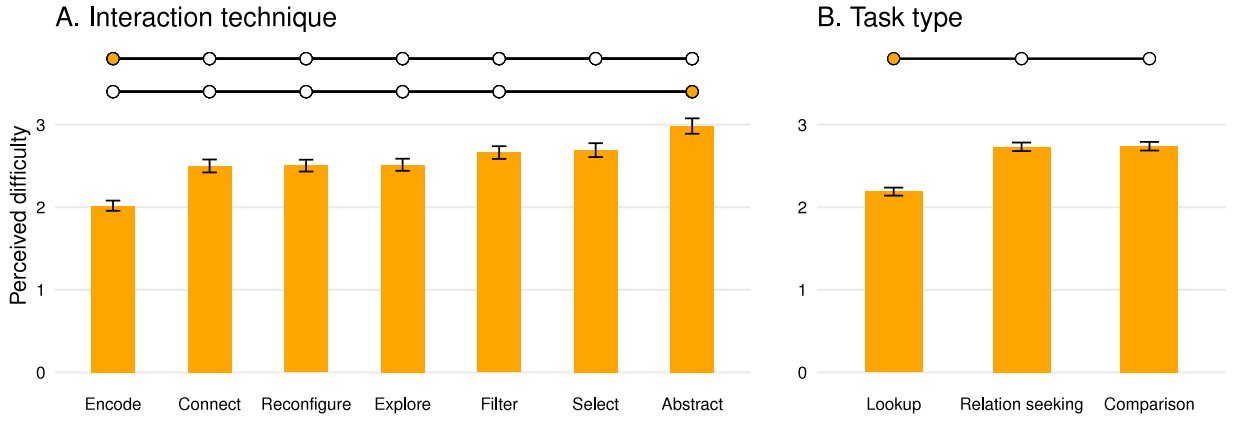


Fig. 5. Mean perceived difficulty and standard error across interaction techniques (left) and task type (right). Filled circles highlight a significant difference from the empty circles.

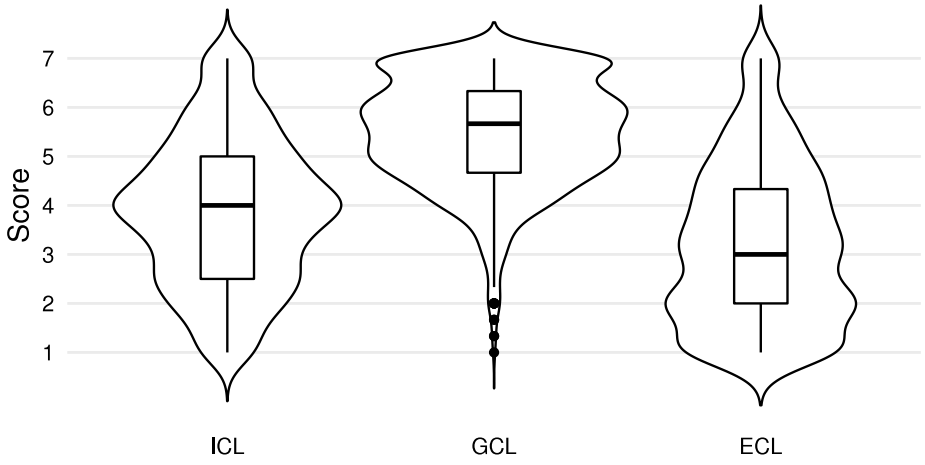


Fig. 6. Cognitive load distributions as split by intrinsic cognitive load, extraneous cognitive load, and germane cognitive load.

Table 4
Mean cognitive load scores per technique across ICL, ECL, and GCL. We use a compact letter display to indicate significant differences between techniques, with unique letters for techniques that are significantly different from one another.

Technique	Mean ICL	ICL sig. dif.	Mean ECL	ECL sig. dif.	Mean GCL	GCL sig. dif.
Select	3.84	a b	3.29	a	5.43	–
Explore	3.73	a c	3.19	a b	5.48	–
Reconfigure	3.80	a b c	3.14	a b	5.54	–
Encode	3.39	c	2.75	b	5.52	–
Abstract/Elaborate	4.41	d	3.92	c	5.33	–
Filter	4.08	b d	2.98	a b	5.56	–
Connect	3.88	a b d	3.16	a b	5.53	–

5.4. Participant reflections

From the open-ended questions, we collected insights on what could improve sense-making in data-solving tasks. A primary observation in the answers is how people sought to limit the possible answers to the questions. Either by filtering (removing options) or making clear distinctions within the data. "*Filtering is helpful, and in a scatter plot strong colour contrasts are better than similar colours. The ability to order the distinctions is particularly helpful and to keep all data on display whilst highlighting different aspects of it*" ([P57], general public). This point is also shared by the data scientists; "*The best data visualisations in my opinion where those where the information could be filtered out or ordered by a given variable. Specially when several variables are studied*" ([P124], data scientist).

With the interaction techniques intentionally evaluated separately, several participants considered the benefit of having different graphs and techniques available simultaneously. For instance, by having pane constellations for answering questions where different techniques may be used for different parts of the inquiry. "*Maybe an option to display multiple of these graphs at the same time. Like I could zoom in or have one constellation of information on the left side pane and start over on the right side by making a different set-up*" ([P77], general public).

It was clear from the responses that the experience with different tools and techniques was higher for the data scientists. We encountered many references to existing applications such as R, PowerBI, Tableau, Looker, Google Data Studio, and SAP. For example: "*I would say an R shiny application or D3 JavaScript application would allow for an easy GUI to transform and analyse data*" ([P118], data scientist). Tools mentioned by the general public were less specialised: "*I've only used Excel, but I would like to know other tools*" ([P114], general public).

6. Discussion

Our evaluation of seven interactive visualisation techniques provides an empirical understanding of their use by the general public and more specialised data scientists. We find that the interaction technique that is used significantly impacts participants' correctness, with Encode, Filter, and Connect providing the best results. A common characteristic of these techniques is the relatively high degree of control they offer end-users over hiding information deemed irrelevant compared to the other four techniques. In Encode, users have the freedom to choose *how* they wish to visualise the information, whereas Filter and Connect allow the user to control *what* information is shown. This corroborates the self-reported cognitive load scores (see Table 4), highlighting a relatively low ICL and ECL score for these interaction techniques. In other words, the inherent difficulty of processing the data and the interaction (ICL) is low and the presentation of that information to the user (ECL) supports the cognitive processing (i.e., easy to grasp) (Sweller et al., 2011). Participants' responses further indicated their preferences for interaction techniques that allow them to narrow down the data shown to limit the possible answers.

We hypothesise that the ability to hide irrelevant information, as most prominent in the Filter interaction, or the ability to choose what information is shown, as part of the Filter and Connect interactions, substantially contribute to lowering the experienced cognitive load, and subsequently foster a higher performance. In their work, Yu et al. investigated non-experts' understanding of an algorithm, comparing a *baseline without explanation* against a *confusion matrix* and a *textual summary* (Yu et al., 2020). Their results show that only the textual summary significantly increased participants' understanding. Based on our results, we argue that this could (partly) be explained by a higher cognitive load experienced when participants were faced with the confusion matrix.

An interesting observation in our data is the similar error rate between the general public and the data scientists' sample. While these results surprised us, they corroborate the perception of interactive visualisation as a tool to make data more accessible to non-experts (Correll, 2019; Harold et al., 2016). Given this surprising outcome, we suggest future work to reassess these findings with a participant sample of which their current professional role or expertise can be more thoroughly validated. Even though our sample of non-experts performed equally well as our data scientists, this does not necessarily mean that they approached the tasks or experienced the interaction techniques equally as well. This is supported both by the higher ECL as well as the reported lower confidence of the non-experts, as compared to the data scientists. As we provided no instructions to either of the participant groups on how to operate the interaction techniques, participants of the non-expert group potentially had to deal with a bigger handicap. The difference in ECL can be explained by the fact that our sample of data scientists was simply more familiar with the visualisations of data, and interaction techniques in general, as also observed in their reference to existing tools in the open-ended questions (see Section 5.4). Prior work found no effect of visualisation style on end-user confidence in interpreting the presented information (Walsh et al., 2021). To the best of our knowledge, no prior work has directly compared the general public's confidence against that of data scientists in data visualisation tasks. We further find that graph literacy scores are a significant predictor of task correctness, but not for participant confidence or experienced difficulty. Prior work reported a significant correlation between graph literacy and users' perceived fairness in algorithm-driven decision-making (van Berkel et al., 2021). Our results indicate that this difference might be partially explained by a difference in the correct understanding of the presented data.

We next reflect on our findings in light of existing and future work in interactive visual analytics.

6.1. Interactive visual analytics and the general public

The research community has repeatedly called for broader public involvement in algorithmic decision-making applications, in which interactive visual analytics play a major role (Liao et al., 2020; Chatzimparmpas et al., 2020; Correll, 2019). However, the needs of the general public have thus far been largely neglected in this scope. With a lack of understanding of potential end-users' needs, data analytics applications will continue to primarily serve expert users. This is problematic, since the general public is the largest user group of intelligent systems, knowingly and unknowingly. Therefore, we next outline the implications of our work for promoting the explainability of algorithmic systems, focusing on the starting point of most 'explainability' work: understanding and assessing the training data used to build these systems.

First, we highlight the importance of minimising the intrinsic and extraneous load demand in end-users of interactive visual analytics applications. Our results indicate that allowing the user to narrow down the data to aspects that they consider as most relevant helps to reduce both ICL and ECL. Narrowing down the data on display can help to manage visualisation complexity (Correll, 2019), thereby 'freeing up cognitive resources' (creating 'mental space') for comprehending other critical or more relevant aspects of the data. As prior research has shown, lower cognitive load designs result in significantly better learning outcomes (Paas, 1992). As neither the end user, nor the developer can directly influence the GCL (Sweller et al., 1998), designers have to aim for a concise, transparent, and accessible data presentation, especially when developing tools for the public. This can be achieved by keeping instructions organised and comprehensive so that users do not have to devote a bigger share of their working memory to this extraneous demand, as this would unnecessarily limit the resources required for learning (Sweller, 2010).

Second, our results show that participants were more correct, experienced lower task difficulty, and had higher confidence for 'lookup' tasks as compared to 'comparison' and 'Relation-seeking' tasks. This

Table 5
Questions as presented to participants in our study. The correct answers are underlined.

Technique	Task type	Question	Answer options
Select	Lookup	To which species does the penguin with a bill length closest to 61mm belong?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three islands contains most Gentoo penguins?	[Biscoe, Dream, Torgersen]
	Relation seeking	To which species does the penguin with the deepest bill on the Dream island belong?	[Adelie, Gentoo, Chinstrap]
Explore	Lookup	To which species does the penguin with a body weight closest to 2600 g belong?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three species has the lowest average flipper length?	[Adelie, Gentoo, Chinstrap]
	Relation seeking	To which species does the male penguin with the lowest body mass belong?	[Adelie, Gentoo, Chinstrap]
Reconfigure	Lookup	Which of the islands has an average bill length closest to 45mm?	[Biscoe, Dream, Torgersen]
	Comparison	Which of the three islands has the highest average bill depth?	[Biscoe, Dream, Torgersen]
	Relation seeking	Which of the three species with an average bill depth above 15mm has the longest average bill?	[Adelie, Gentoo, Chinstrap]
Encode	Lookup	To which species does the penguin with a flipper length closest to 240mm belong?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three species has the most penguins with a flipper length below 187mm?	[Adelie, Gentoo, Chinstrap]
	Relation seeking	Which of the three species has both the heaviest individual penguin and the highest average weight?	[Adelie, Gentoo, Chinstrap]
Abstract/Elaborate	Lookup	Which species has more than 100 penguins on the Biscoe island?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three islands has the most female Adelie penguins?	[Biscoe, Dream, Torgersen]
	Relation seeking	Which of the three species is only represented on the Dream island?	[Adelie, Gentoo, Chinstrap]
Filter	Lookup	To which species on the Biscoe island do the penguins with a bill length below 38mm belong?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three islands has the most penguins with a bill depth below 15mm?	[Biscoe, Dream, Torgersen]
	Relation seeking	To which species do the penguins belong with a bill length above 50mm and a bill depth lower than 15mm?	[Adelie, Gentoo, Chinstrap]
Connect	Lookup	Which of the three species has more than 140 penguins?	[Adelie, Gentoo, Chinstrap]
	Comparison	Which of the three islands has the most female penguins?	[Biscoe, Dream, Torgersen]
	Relation seeking	Which island hosts both Adelie and Chinstrap penguins?	[Biscoe, Dream, Torgersen]

Table 6
Pairwise comparisons with TukeyHSD tests for correctness.

Comparison	Difference	95% CI	*p-value*	
Explore-Select	−0.075	[−0.140, −0.010]	0.011	*
Reconfigure-Select	−0.102	[−0.167, −0.037]	<0.001	***
Encode-Select	0.063	[−0.002, 0.127]	0.065	
Abstract-Select	−0.125	[−0.190, −0.060]	<0.001	***
Filter-Select	0.056	[−0.008, 0.121]	0.136	
Connect-Select	0.073	[0.008, 0.138]	0.015	*
Reconfigure-Explore	−0.027	[−0.092, 0.038]	0.880	
Encode-Explore	0.138	[0.073, 0.202]	<0.001	***
Abstract-Explore	−0.050	[−0.115, 0.015]	0.252	
Filter-Explore	0.131	[0.067, 0.196]	<0.001	***
Connect-Explore	0.148	[0.083, 0.213]	<0.001	***
Encode-Reconfigure	0.165	[0.100, 0.229]	<0.001	***
Abstract-Reconfigure	−0.023	[−0.088, 0.042]	0.943	
Filter-Reconfigure	0.158	[0.094, 0.223]	<0.001	***
Connect-Reconfigure	0.175	[0.110, 0.240]	<0.001	***
Abstract-Encode	−0.188	[−0.252, −0.123]	<0.001	***
Filter-Encode	−0.006	[−0.071, 0.058]	1.000	
Connect-Encode	0.010	[−0.054, 0.075]	0.999	
Filter-Abstract	0.181	[0.117, 0.246]	<0.001	***
Connect-Abstract	0.198	[0.133, 0.263]	<0.001	***
Connect-Filter	0.017	[−0.048, 0.081]	0.988	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 7
Pairwise comparisons with TukeyHSD tests for correctness.

Comparison	Difference	95% CI	*p-value*	
Lookup-Comparison	0.033	[−0.001, 0.067]	0.056	
Relation seeking-Comparison	−0.006	[−0.040, 0.027]	0.901	
Relation seeking-Lookup	−0.039	[−0.073, −0.006]	0.017	*

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 8
Pairwise comparisons with TukeyHSD tests for confidence.

Comparison	Difference	95% CI	*p-value*	
Explore-Select	0.021	[−0.277, 0.319]	0.999	
Reconfigure-Select	0.006	[−0.292, 0.304]	1.000	
Encode-Select	0.423	[0.125, 0.721]	<0.001	***
Abstract-Select	−0.279	[−0.577, 0.019]	0.083	
Filter-Select	0.177	[−0.121, 0.475]	0.579	
Connect-Select	0.267	[−0.031, 0.564]	0.114	
Reconfigure-Explore	−0.015	[−0.312, 0.283]	1.000	
Encode-Explore	0.402	[0.104, 0.700]	0.001	***
Abstract-Explore	−0.300	[−0.598, −0.002]	0.047	*
Filter-Explore	0.156	[−0.142, 0.454]	0.716	
Connect-Explore	0.246	[−0.052, 0.544]	0.184	
Encode-Reconfigure	0.417	[0.119, 0.714]	<0.001	***
Abstract-Reconfigure	−0.285	[−0.583, 0.012]	0.070	
Filter-Reconfigure	0.171	[−0.127, 0.469]	0.621	
Connect-Reconfigure	0.260	[−0.037, 0.558]	0.132	
Abstract-Encode	−0.702	[−1.000, −0.404]	<0.001	***
Filter-Encode	−0.246	[−0.543, 0.052]	0.184	
Connect-Encode	−0.156	[−0.454, 0.142]	0.716	
Filter-Abstract	0.456	[0.158, 0.754]	<0.001	***
Connect-Abstract	0.546	[0.248, 0.844]	<0.001	***
Connect-Filter	0.090	[−0.208, 0.387]	0.975	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 9
Pairwise comparisons with TukeyHSD tests for confidence.

Comparison	Difference	95% CI	*p-value*	
Comparison-Lookup	−0.461	[−0.616, −0.306]	<0.001	***
Relation seeking-Lookup	−0.384	[−0.539, −0.229]	<0.001	***
Relation seeking-Comparison	0.077	[−0.078, 0.232]	0.476	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

perhaps indicates that the evaluated interaction techniques do not sufficiently support users in comparing two or more data points. However, enabling users to compare two or more data points is a crucial element of data exploration. Fairness criteria such as 'group fairness' and 'demographic parity' extensively rely on comparing groups of data points. Prior work has indicated the challenges in contrasting data points in ranking tables (Perin et al., 2014), pointing to the possibility of bringing selected or highlighted items into the same view to support comparison. Such an approach could be evaluated by expanding and combining the Filter and Encode techniques. Wang et al. have presented

Table 10
Pairwise comparisons with TukeyHSD tests for difficulty.

Comparison	Difference	95% CI	*p-value*	
Explore-Select	−0.177	[−0.494, 0.140]	0.652	
Reconfigure-Select	−0.188	[−0.505, 0.130]	0.587	
Encode-Select	−0.673	[−0.990, −0.356]	<0.001	***
Abstract-Select	0.292	[−0.026, 0.609]	0.096	
Filter-Select	−0.029	[−0.346, 0.288]	1.000	
Connect-Select	−0.192	[−0.509, 0.126]	0.560	
Reconfigure-Explore	−0.010	[−0.328, 0.307]	1.000	
Encode-Explore	−0.496	[−0.813, −0.179]	<0.001	***
Abstract-Explore	0.469	[0.151, 0.786]	<0.001	***
Filter-Explore	0.148	[−0.169, 0.465]	0.815	
Connect-Explore	−0.015	[−0.332, 0.303]	1.000	
Encode-Reconfigure	−0.485	[−0.803, −0.168]	<0.001	***
Abstract-Reconfigure	0.479	[0.162, 0.796]	<0.001	***
Filter-Reconfigure	0.158	[−0.159, 0.476]	0.762	
Connect-Reconfigure	−0.004	[−0.321, 0.313]	1.000	
Abstract-Encode	0.965	[0.647, 1.282]	<0.001	***
Filter-Encode	0.644	[0.326, 0.961]	<0.001	***
Connect-Encode	0.481	[0.164, 0.799]	<0.001	***
Filter-Abstract	−0.321	[−0.638, −0.004]	0.045	*
Connect-Abstract	−0.483	[−0.801, −0.166]	<0.001	***
Connect-Filter	−0.163	[−0.480, 0.155]	0.738	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 11
Pairwise comparisons with TukeyHSD tests for difficulty.

Comparison	Difference	95% CI	*p-value*	
Comparison-Lookup	−0.550	[0.385, 0.715]	<0.001	***
Relation seeking-Lookup	0.543	[0.378, 0.708]	<0.001	***
Relation seeking-Comparison	−0.007	[−0.172, 0.158]	0.994	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

positive results of supporting expert users in comparing differences between datasets (Wang et al., 2022), providing further inspiration for future work aiming to involve the general public.

Third, we highlight the importance of the datasets used in studying interactive data visualisations. We specifically chose our dataset on Palmer penguins (Horst et al., 2020) due to its lack of contentious characteristics. This stands in contrast with prior work, which includes, for example, the widely studied recidivism dataset (Shen et al., 2020; van Berkel et al., 2021). While using such a topical and timely dataset can provide valuable insights into the general public's fairness perceptions, it introduces a confounding factor when studying interaction techniques. To overcome this, Van Berkel et al. suggested researchers to "*validate results across scenarios/datasets*" (van Berkel et al., 2021). Using a neutral dataset may offset this requirement when the primary relevance is the evaluation of data interaction techniques. We further stress the value of using a publicly available dataset, enabling other researchers to compare the impact of their implementations on *e.g.* data comprehension.

6.2. Limitations and future work

We recognise several limitations concerning the study's internal and external validity. First, we stress that our evaluation of the seven interaction techniques is limited to a specific representation of each technique. While we took the utmost care to construct prototypical implementations of each interaction technique, taking into consideration both prior work and existing systems (see Section 3), we recognise that our specific representations may have affected participants' perceptions. By releasing the source code of our study materials, we hope to support the further evaluation and expansion of our representation of these interaction techniques.

Second, our study was limited to the individual assessment of seven interaction techniques. Real-world data explorations may require end-users to combine these interaction techniques, thereby introducing additional challenges to understanding the data. However, evaluating the combination of all interaction techniques would have exorbitantly inflated the number of study conditions and inhibited the evaluation of individual interaction techniques.

Third, the recruitment of our participant sample introduced some limitations. While we purposefully chose not to limit the geographical location of participants, our participant sample is still biased towards those from the Global North and of relatively young age. Prior work highlights that demographic factors such as age may impact the preference and understanding of data visualisations (Shamim et al., 2016; van Weert et al., 2021). Due to the recruitment method, our information on participants is relatively sparse. While our 'data scientist' participants have reported this as their current business role, we have no data on, for example, their experience in this role. Future work may, therefore, seek to compare our results with a sample of participants currently employed as (senior) data scientists. On the other hand, our 'general public' participants might have read the study description and had a pre-existing interest in the domain. Lastly, our participants completed these tasks alone rather than collaborating with others. Real-world visual analytics is likely to involve collaboration, and earlier work has highlighted that the diversity of collaborative groups can indeed impact such data assessments (van Berkel et al., 2019). Future work may explore how collaboration across various roles may impact the use and preference for interaction techniques.

7. Conclusion

In this paper, we investigated how different interaction techniques impact end-user performance in the visual inspection of data. While the ability to understand data is becoming increasingly important in everyday life, our results show that the interaction technique impacts users' performance, perceived difficulty, intrinsic and extraneous cognitive load, and—in case of non-experts—their confidence. Based on our analysis, we highlight the value of interaction techniques that enable users to narrow down the information shown to them. By enabling members of the general public to better comprehend data, we seek to support the development of future visual analytics tools aimed at a broader target audience than technical experts.

CRediT authorship contribution statement

Niels van Berkel: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Benjamin Tag:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rune Møberg Jacobsen:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Investigation, Data curation. **Daniel Russo:** Methodology, Conceptualization. **Helen C. Purchase:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Daniel Buschek:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Niels van Berkel reports financial support was provided by Carlsberg Foundation.

Data availability

The application's code is available at https://osf.io/ce46r/?view_only=6ef59ba87bbb4861b54bc8a2bf3c1309.

Acknowledgement

This work is supported by the Carlsberg Foundation, Denmark, grant CF21-0159.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Andrienko, N., Andrienko, G., 2006. Exploratory analysis of spatial and temporal data: a systematic approach. Springer Science & Business Media.

Bakdash, J.Z., Marusich, L.R., 2017. Repeated measures correlation. Front. Psychol. 8, http://dx.doi.org/10.3389/fpsyg.2017.00456.

Bäuerle, A., Cabrera, Á.A., Hohman, F., Maher, M., Koski, D., Suau, X., Barik, T., Moritz, D., 2022. Symphony: Composing interactive interfaces for machine learning. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3491102.3502102.

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R.M., Kostakos, V., 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), http://dx.doi.org/10.1145/3359130.

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., Skov, M.B., 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445365.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report, MSR-TR-2020-32, Microsoft.

Blackwell, A., Church, L., Hales, I., Jones, M., Jones, R., Mahmoudi, M., Marasoiu, M., Meakins, S., Spott, M., 2018. Computer says ' don't know' - interacting visually with incomplete AI models. In: Tanimoto, S., Fan, S., Ko, A., Locksa, D. (Eds.), Proceedings of the Workshop on Designing Technologies To Support Human Problem Solving. University of Washington, pp. 5–14.

Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F., Kerren, A., 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. Comput. Graph. Forum http://dx.doi.org/10.1111/cgf.14034.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H., 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. Conf. Hum. Factors Comput. Syst. - Proc. 1–12. http://dx.doi.org/10.1145/3290605.3300789.

Correll, M., 2019. Ethical dimensions of visualization research. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3290605.3300418.

Ens, B., Bach, B., Cordeil, M., Engelke, U., 2021. Grand challenges in immersive analytics. Conf. Hum. Factors Comput. Syst. - Proc. http://dx.doi.org/10.1145/3411764.3446866.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. Behav. Res. Methods 41 (4), 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149.

Harold, J., Lorenzoni, I., Shipley, T.F., Coventry, K.R., 2016. Cognitive and psychological science insights to improve climate change data visualization. Nature Clim. Change 6 (12), 1080–1089. http://dx.doi.org/10.1038/nclimate3162.

Heer, J., Shneiderman, B., 2012. Interactive dynamics for visual analysis. Queue 10 (2), 30–55. http://dx.doi.org/10.1145/2133416.2146416.

Hoque, M., Mueller, K., 2022. Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making. IEEE Trans. Vis. Comput. Graphics 28 (12), 4728–4740. http://dx.doi.org/10.1109/TVCG.2021.3102051.

Hoque, E., Setlur, V., Tory, M., Dykeman, I., 2018. Applying pragmatics principles for interaction with visual analytics. IEEE Trans. Vis. Comput. Graphics 24 (1), 309–318. http://dx.doi.org/10.1109/TVCG.2017.2744684.

Horst, A.M., Hill, A.P., Gorman, K.B., 2020. Palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0.

Jin, W., Carpendale, S., Hamarneh, G., Gromala, D., 2019. Bridging AI developers and end users: an end-user-centred explainable AI taxonomy and visual vocabularies. In: IEEE Visualization Conference. VIS, IEEE.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G., 2008. Visual analytics: Definition, process, and challenges. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (Eds.), Information Visualization: Human-Centered Issues and Perspectives. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 154–175. http://dx.doi.org/10.1007/978-3-540-70956-5_7.

Kerracher, N., Kennedy, J., Chalmers, K., 2015. A task taxonomy for temporal graph visualisation. IEEE Trans. Vis. Comput. Graphics 21 (10), 1160–1172. http://dx.doi.org/10.1109/TVCG.2015.2424889.

Klepsch, M., Schmitz, F., Seufert, T., 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front. Psychol. 8 (NOV), 1–18. http://dx.doi.org/10.3389/fpsyg.2017.01997.

Kugler, L., 2018. The war over the value of personal data. Commun. ACM 61 (2), 17–19. http://dx.doi.org/10.1145/3171580.

Kuosmanen, E., Visuri, A., Kheirinejad, S., van Berkel, N., Koskimäki, H., Ferreira, D., Hosio, S., 2022. How does sleep tracking influence your life? Experiences from a longitudinal field study with a wearable ring. Proc. ACM Human-Computer Interact. 6 (MHCI), 1–19. http://dx.doi.org/10.1145/3546720.

Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, pp. 1–15. http://dx.doi.org/10.1145/3313831.3376590.

Locascio, J., Khurana, R., He, Y., Kaye, J., 2016. Utilizing employees as usability participants. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 4533–4537. http://dx.doi.org/10.1145/2858036.2858047.

Lu, Y., Garcia, R., Hansen, B., Gleicher, M., Maciejewski, R., 2017. The state-of-the-art in predictive visual analytics. Comput. Graph. Forum 36 (3), 539–562. http://dx.doi.org/10.1111/cgf.13210.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

Mosca, A., Ottley, A., Chang, R., 2021. Does interaction improve Bayesian reasoning with visualization? In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445176.

Okan, Y., Janssen, E., Galesic, M., Waters, E.A., 2019. Using the short graph literacy scale to predict precursors of health behavior change. Med. Decis. Making 39 (3), 183–195. http://dx.doi.org/10.1177/0272989X19829728, PMID: 30845893.

Paas, F.G.W.C., 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. J. Educ. Psychol. 84 (4), 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429.

Paas, F., van Gog, T., 2006. Optimising worked example instruction: Different ways to increase germane cognitive load. Learn. Instr. 16, 87–91. http://dx.doi.org/10.1016/j.learninstruc.2006.02.004.

Perin, C., Vuillemot, R., Fekete, J.-d., 2014. A' table! improving temporal navigation in soccer ranking tables. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 887–896. http://dx.doi.org/10.1145/2556288.2557379.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R., 2018. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.

Sauro, J., Dumas, J.S., 2009. Comparison of three one-question, post-task usability questionnaires. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09, Association for Computing Machinery, New York, NY, USA, pp. 1599–1608. http://dx.doi.org/10.1145/1518701.1518946.

Schwab, M., Hao, S., Vitek, O., Tompkin, J., Huang, J., Borkin, M.A., 2019. Evaluating pan and zoom timelines and sliders. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–12. http://dx.doi.org/10.1145/3290605.3300786.

Shamim, A., Balakrishnan, V., Tahir, M., Qureshi, M.A., 2016. Age and domain specific usability analysis of opinion visualisation techniques. Behav. Inform. Technol. 35 (8), 680–689. http://dx.doi.org/10.1080/0144929X.2016.1141235.

Shen, H., Jin, H., Cabrera, Á.A., Perer, A., Zhu, H., Hong, J.I., 2020. Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. Proc. ACM Hum.-Comput. Interact. 4 (CSCW2), http://dx.doi.org/10.1145/3415224.

Shneiderman, B., 1992. Tree visualization with tree-maps: 2-d space-filling approach. ACM Trans. Graph. 11 (1), 92–99. http://dx.doi.org/10.1145/102377.115768.

Sweller, J., 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ. Psychol. Rev. 22 (2), 123–138. http://dx.doi.org/10.1007/s10648-010-9128-5.

Sweller, J., Ayres, P., Kalyuga, S., 2011. Cognitive load theory. Cognitive Load Theory. Springer New York, New York, NY, pp. 17–25. http://dx.doi.org/10.1007/978-1-4419-8126-4.

Sweller, J., van Merrienboer, J.J.G., Paas, F.G.W.C., 1998. Cognitive architecture and instructional design. Educ. Psychol. Rev. 10 (3), 251–296. http://dx.doi.org/10.1023/A:1022193728205.

Tominski, C., 2015. Interaction for visualization. Springer International Publishing, Cham, pp. 9–25. http://dx.doi.org/10.1007/978-3-031-02600-3.

Tufte, E.R., 2001. The Visual Display of Quantitative Information, second ed. Graphics press Cheshire, CT.

van Weert, J.C., Alblas, M.C., van Dijk, L., Jansen, J., 2021. Preference for and understanding of graphs presenting health risk information. The role of age, health literacy, numeracy and graph literacy. Patient Educ. Couns. 104 (1), 109–117. http://dx.doi.org/10.1016/j.pec.2020.06.031.

Walsh, E.I., Sargent, G.M., Grant, W.J., 2021. Not just a pretty picture: Scientific fact visualisation styles, preferences, confidence and recall. Inf. Vis. 20 (2–3), 138–150. http://dx.doi.org/10.1177/14738716211027587.

Wang, A.Y., Epperson, W., DeLine, R.A., Drucker, S.M., 2022. Diff in the loop: Supporting data comparison in exploratory data analysis. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3491102.3502123.

Wanner, J., Herm, L.V., Heinrich, K., Janiesch, C., Zschech, P., 2021. White, grey, black: Effects of XAI augmentation on the confidence in AI-based decision support systems. International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global (ML), 1–9.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J., 2020. The what-if tool: Interactive probing of machine learning models. IEEE Trans. Vis. Comput. Graphics 26 (1), 56–65. http://dx.doi.org/10.1109/TVCG.2019.2934619.

Woodruff, A., Fox, S.E., Rousso-Schindler, S., Warshaw, J., 2018. A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, pp. 1–14. http://dx.doi.org/10.1145/3173574.3174230.

Yang, C., Zhang, Z., Fan, Z., Jiang, R., Chen, Q., Song, X., Shibasaki, R., 2022. EpiMob: Interactive visual analytics of citywide human mobility restrictions for epidemic control. IEEE Trans. Vis. Comput. Graphics PP, http://dx.doi.org/10.1109/tvcg.2022.3165385.

Yi, J.S., ah Kang, Y., Stasko, J., Jacko, J.A., 2007. Toward a deeper understanding of the role of interaction in information visualization. IEEE Trans. Vis. Comput. Graphics 13 (6), 1224–1231. http://dx.doi.org/10.1109/TVCG.2007.70515.

Yu, B., Yuan, Y., Terveen, L., Wu, Z.S., Forlizzi, J., Zhu, H., 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference. Association for Computing Machinery, New York, NY, USA, pp. 1245–1257. http://dx.doi.org/10.1145/3357236.3395528.