VisLab: Crowdsourcing Visualization Experiments in the Wild

ANONYMOUS AUTHOR(S)

SUBMISSION ID: 1389*

When creating a visualization to understand and communicate data, we face different design choices. Even though past empirical research provides foundational knowledge for visualization design, practitioners still rely on their hunches to deal with intricate trade-offs in the wild. On the other hand, researchers lack time and resources to rigorously explore the growing design space through controlled experiments. In this work, we aim to address this two-fold problem by crowdsourcing visualization experiments. We developed VisLab, an online platform in which anyone can design and deploy experiments to evaluate their visualizations. To alleviate the complexity of experiment design and analysis of data visualization, the platform provides scaffold templates and analytic dashboards. To motivate broad participation, the platform enables anonymous participation and provides personalized performance feedback. We present use case scenarios that demonstrate the usability and usefulness of the platform in addressing different needs of practitioners, researchers, and educators.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: visualization, experiment, crowdsourcing, perception, cognition, citizen science

ACM Reference Format:

Anonymous Author(s). 2018. VisLab: Crowdsourcing Visualization Experiments in the Wild. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Data visualizations have recently gone mainstream, enabling people with diverse backgrounds—not only analysts and scientists but also journalists, designers, and even casual users—to understand and communicate complex data without requiring advanced statistical literacy [25, 32]. The growing adoption in the public further expanded the already combinatorial design space of visualizations. Designers not only need to consider perceptual effectiveness for accurate analysis but also other intricate design dimensions such as aesthetics, memorability, and engagement.

Although past empirical research provides fruitful knowledge for effective visualization design (e.g., perception [12, 16, 18, 19, 24] and cognition [5, 15, 17, 22, 33]), practitioners still rely on their hunches to make nuanced design decisions involving intricate data distributions, unconventional visuals, and narrational elements that are absent in typical controlled experiments [21]. For instance, journalists often employ unverified, but often engaging, representations such as Marimekko charts [1] and Tornado charts [3], while scientists also often come up with new visuals to cope with domain-specific data such as Muller plots [2] and Sequence Logo [1]. On the other hand, researchers lack motivations and resources to explore every possible design choice, leading to insufficient guidelines and principles for practitioners.

[1]https://en.wikipedia.org/wiki/Sequence_logo

Manuscript submitted to ACM

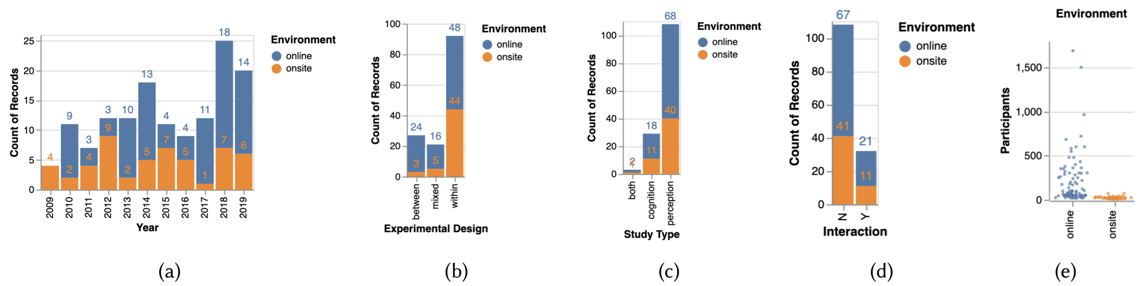

Fig. 1. The number of publications by environment per (a) year, (b) experiment design type, (c) study type, (d) interaction. (e) The participants distribution by environment.

We developed VisLab, an online voluntary experiment platform in which anyone can design and deploy experiments to evaluate their visualizations. To inform the design of the platform, we analyzed procedures, tasks, and methods in existing empirical studies. Based on the analysis, we derived experiment templates that cover perception and cognitive design dimensions. The template-based approach reduces the complexity of experiment design and is generalizable across different visualization stimuli. VisLab also supports analytic dashboards to aid the interpretation of experiment results. To enable broad participation, the experimenter can easily distribute the experiment through a shareable link, while participants receive personalized performance feedback as a learning incentive. We discuss usage scenarios for practitioners, as well as researchers and educators. VisLab builds on the similar ideas of online citizen science platforms [13, 27, 29–31] but its design is tailored to address the unique problem in the visualization field.

2 DESIGN SPACE OF EMPIRICAL STUDIES

To inform the design of the visualization experiment platform, we first explored the current landscape of visualization empirical studies below. Our analysis goal is to investigate the feasibility of online experiment platforms, the discovery of generalizable templates, and the coverage of design aspects.

2.1 Data Collection

In collecting empirical studies, the main inclusion criterion was quantitative experiments evaluating the perceptual and cognitive processes of visualizations, as they provide guidelines and principles more generalizable than other types of evaluation studies [10, 26]. We relied on VisPerception, a public repository curated by Brown [4] of which collection spans most of the seminal works in the field published from venues such as CHI, VIS, and EuroVis. It contains 223 papers published from 1926 to date with a majority of them being journal articles and conference papers (>94%). We sampled the papers from CHI and VIS from 2009 to 2019, and the final collection contains 74 papers, including 24 papers from CHI and 50 papers from IEEE VIS.

2.2 Analysis Method

We performed open coding analysis with the final collection. The goal of the analysis was to harvest representative tasks, procedures, and evaluation methods that we can incorporate into the design of our platform. Our initial codes were adapted and extended from the ones used for the analysis of crowdsourced evaluation studies by Borgo et al. [6], which include dimensions related to study design & procedure, task types, participants, and measures & metrics. Our data collection includes both in-lab and crowdsourced experiments and focuses on general perception and cognition, excluding evaluations of specific systems. To enable per-experiment analysis, we divided each paper into multiple experiments if applicable, resulting in a total of 141 experiments; an average number of experiments per paper is 1.9.

2.3 Results

2.3.1 Experiment Design. Among the 141 experiments, 88 of them were online experiments and 52 of them were in-lab experiments. Figure 1a shows that online experiments are consistently more common than in-lab experiments in recent years. In terms of experiment design, we have observed 92 within-, 25 between-, and 20 mixed-design experiments. While the study environment was evenly split into 48 online and 44 in-lab experiments in the within-design, online experiments were more common for between (24 vs. 3) and mixed (15 vs. 5) designs. Time for tasks was often constrained (47/141). Interestingly, the ratio for time constraint was higher for in-lab studies (69%) than online studies (28%).

2.3.2 Standardized Procedures. Almost all experiments followed a standard procedure including pre-survey, screening, training/practice, and post-survey, although they often do not mention whether they had each component or not. In our report, we consider the absence of presence as non-existence. Out of 141 experiments, 40 experiments (29%) had pre-surveys. Types of questions asked in the pre-surveys include gender (24), age (20), chart literacy (8), and education (7). The rate of having pre-surveys was higher for in-lab studies (46%) compared to online studies (18%).

On the other hand, 58 experiments (41%) had screening processes. A majority of them were about color blindness (28), while others include vision test (8) and the acceptance rate on Amazon's Mechanical Turk (6). In contrast to pre-surveys, online studies had more screenings (43%) than in-lab studies (38%). 60% of experiments (84) had training or practice before actual tasks. The type of practice tasks depends on task types in the experiments. 67% of in-lab studies (36/52) had training, which is higher than 55% of online studies (48/88).

At the end of the study, 32% of the experiments (45/140) had post-survey. Questions varied including demographic (9), free-form comment (7), preference (5), familiarity (3), and confidence (2). 44% of in-lab experiments (23/52) had post-surveys, while only 25% of online experiments (22/88) had post-surveys. In general, online experiments had fewer pre- and post-surveys compared to in-lab experiments.

2.3.3 Tasks, Participants, and Measures. We observed a variety of tasks in the experiments. We coded the tasks using the taxonomy by Brehmer et al. [9]. The top five tasks include *compare* (26% of 140 experiments), *identify* (21%), and *derive* (11%), *select* (5%), and *recall* (3%). The top tasks remain mostly similar across study environment types with some differences. For instance, *compare* (29) was higher than *identify* (13) in online experiments, compared to in-lab experiments (7 versus 17).

There were 144 participants on average per experiment (Median = 48, Std = 245). In online studies, the average number of participants was 216 (Median = 96, Std = 287), while the average was 24 in in-lab studies (Media = 20, Std = 14). In terms of measures, accuracy/time (96), time (54), confidence (11) were the most common. The same ranks hold true for both online and in-lab studies. In terms of reporting experimental results, diverse charts were employed. The most common forms of reporting methods were tables (24) and error bars (52), while oftentimes, simple bar charts (11) and box plots (12) were used as well.

2.3.4 Evaluation of Different Design Aspects. As expected, perception experiments (77%) were more common than cognition experiments (21%)—e.g., memorability, comprehension, and engagement. Some experiments had a combination of both (3). Only seven experiments were about animation, while only one of them was an online study. On the other

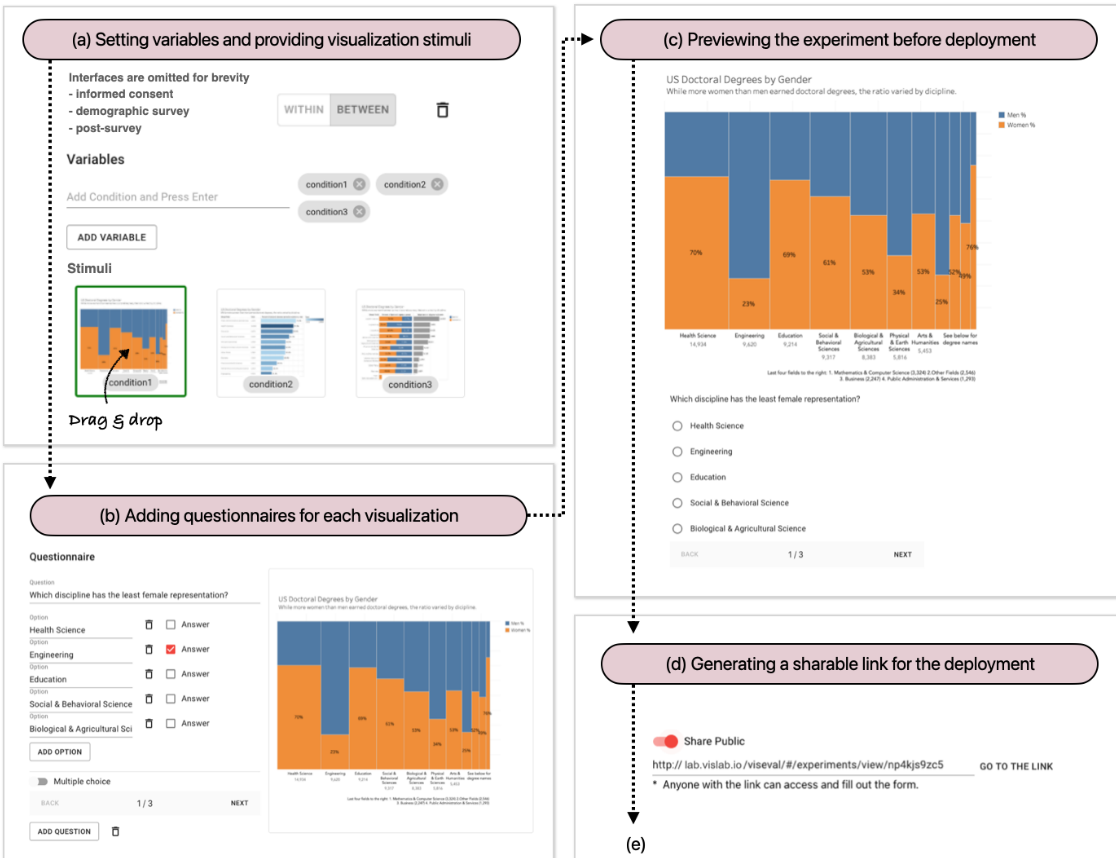

Fig. 2. A workflow for a graphical perception experiment: (a - b) experiment design interface, (c) experiment preview, (d) shareable link for recruiting participants.

hand, 32 experiments tested interaction, among which 21 of them were online. In terms of types of visualizations considered, standard charts were the most commonly tested, including scatter plots (24), bar charts (14), and line charts (7). We have observed custom charts (8) and pictograms (5) as well. The frequency ranking was more or less similar across experiment environments except all of the custom charts were tested in labs.

2.4 Takeaways

The results indicate that online experiments are now widely used and produce acceptable results. The analysis of experiment design and procedures provides insights into what components would be necessary for experiment templates, and suggests that the templates should include flexible design and dashboard interfaces to control factorial design and monitor task measures. Lastly, going beyond perception to evaluate cognition dimensions will be important for holistically evaluating visualizations in the wild.

3 VISLAB

We developed VisLab with an overarching goal of supporting anyone to evaluate their visualizations through easy-to-deploy user testing experiments (Figure 2).

3.1 Design Goals

To inform the design of VisLab, we derived the following design goals based on the current practice in the field, as well as the formative analysis and prior research in online platforms.

Enable practitioners to evaluate their visualizations. Practitioners are often interested in how their visualizations would be perceived by the intended audience. They may want to test different design variants and how potential audiences may react to the designs. An experiment currently running in the LabintheWild platform [31] demonstrates this use case in which health professionals are testing three different graphs for showing harmful chemicals [2]. We also frequently observe practitioners discussing the trade-off of a controversial real-world visualization such as the Tornado plot [3]. We aim to provide easy-to-use and off-the-shelf user interfaces to design and deploy experiments to evaluate in-the-wild visualizations.

Reduce the complexity of experiment design and analysis. Running rigorous empirical studies requires expertise in experiment design and statistical analysis, which is not typically available for practitioners. In addition, it requires complex programming to deploy visualization experiments including the LabintheWild example mentioned above. Our aim is to provide scaffolding to ease the design, analysis, and deployment of experiments in a way that practitioners only supply their visualizations with simple configuration parameters.

Enable easy and scalable participation from a broader audience. Having a larger sample size and participant diversity would provide better reliability and validity of evaluation results. To reach a wider audience quickly, we aim to support the easy-sharing of experiments in online media. In addition, while providing extrinsic monetary motivation may be useful for attracting potential participants, practitioners may not have resources to afford the cost. Thus, we seek to leverage intrinsic motivation by providing personalized feedback that would be useful for improving self-awareness of the participant's perceptual and cognitive skills [31].

3.2 Designing Experiments Using Scaffold Templates

VisLab provides three templates supporting experiments for graphical perception [11, 20], memorability [7, 8], and attention tracking [23], which cover both major perceptual and cognitive dimensions. We integrated a standard experiment procedure into the templates including informed consent, demographic survey, tasks, and post-survey. Each component in the procedure is tailored or turned-off as necessary, such as excluding a post-survey and customizing informed consent. The main difference across the templates lies in the task design interface that supports different configuration parameters. In the graphical perception template (Figure 2), the user can add analytical questions testing perceptual abilities with visualizations, while memorability and attention tracking uses the same tasks in the original studies along with customizable options (e.g., image display time and blur amount) (Figure 4a) [8, 23]. The users can simply drag and drop the visualization stimuli into the thumbnail containers in the task design interface. VisLab supports an image, as well as Vega-lite spec [34] that can support interactions.

[2] http://graphs.labinthewild.org/
[3] https://twitter.com/emmawage/status/1255172980788785152

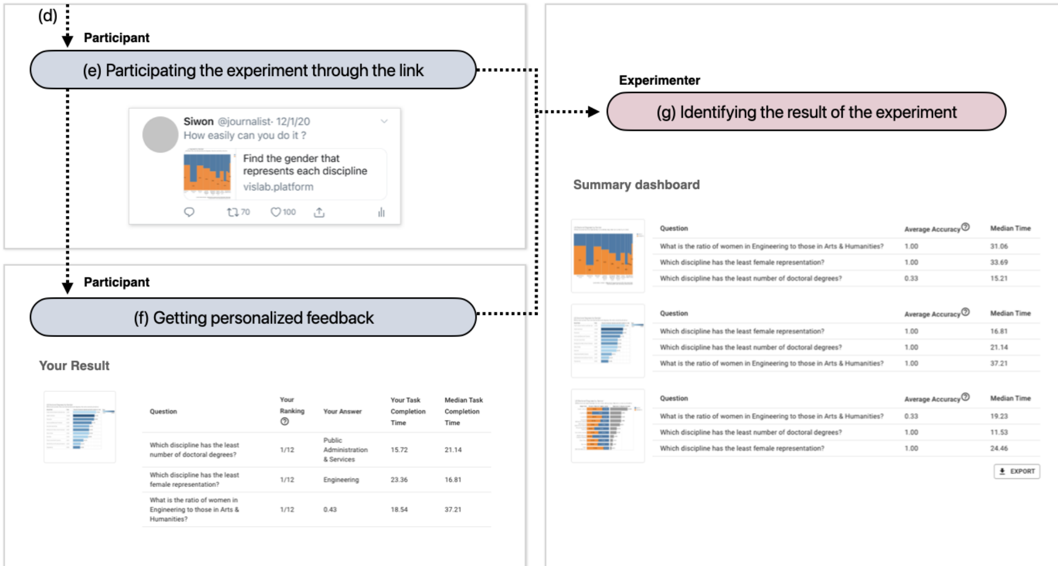

Fig. 3. A workflow for a graphical perception experiment: (e) participating in the experiment through the shared link, (f) a participant's interface confirming performance outcome at the end, and (g) an experimenter's interface for inspecting the result.

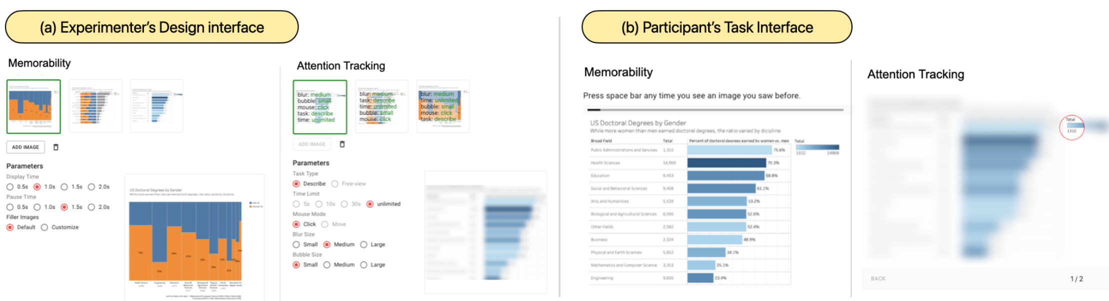

Fig. 4. (a) Experimenter's design interface and (b) Participant's task interface for memorability and attention tracking experiments.

3.3 Deploying Experiments and Analyzing Results

Once completing an experiment design, a user can preview the outcome in advance before deploying it (Figure 2c). After verifying the experiment, the user can deploy by sharing the generated link that will direct participants to the experiment (Figure 2d). If the link is shared on social media, auto-generated meta tags using the title and stimuli of the experiment will appear. After the experiment is finished, the experimenter can use the analytic dashboard to monitor the experiment results. They can inspect outcome measures including accuracy and time for perception (Figure 3g), hit rate for memorability, and click maps for attention tracking. They can see aggregate statistics as well as individual results to see if any malicious attempts exist and filter them as necessary to assure the quality of the interpretation of the results. If the experiment has pre- or post-surveys, they can also analyze the statistics of survey responses.

3.4 Motivating Online Participation through Personalized Feedback

A potential participant can easily click the link shared by the experimenter to participate in the experiment (Figure 3e). We intentionally designed the link not to require sign-up for participants to lower the barrier for participation [31]. The implication of this is that the identity of participants will not be tracked although it limits the ability to run longitudinal studies at the same time. To maintain the integrity of each experiment, we logged IP addresses to prevent duplicate participation and use similar precautionary measures used in crowdsourcing such as attention questions and vigilant images [8]. To motivate voluntary participation, we provided personalized feedback at the end of each experiment as a learning benefit (Figure 3f). A participant can inspect individual performance as well as comparisons to the aggregate result across all participants.

4 USER SCENARIO

To explain the usage context of VisLab, we illustrate how a visual data journalist can use the platform to evaluate multiple alternative designs of a visualization in Figure 6. We also expect VisLab to support two additional scenarios: education and research. For instance, an instructor for a visualization course can use VisLab to teach science behind visualization by asking students to replicate seminal experiments such as one by Cleveland & McGill [11] (Figure 5). On the other hand, researchers browser the publicly released results of experiments in VisLab and may find an experiment that provides an initial research hypothesis and merits further investigating through a rigorous controlled experiment.

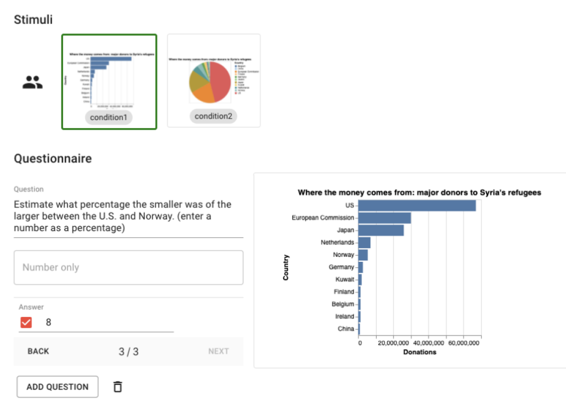

Fig. 5. Using in-the-wild pie and bar charts to replicate the experiment by Cleveland & McGill [11].

5 DISCUSSION

Practitioner-sourcing for visualization research. Our platform aims to not only support practitioners but also enable researchers to explore the vast visualization design space through practitioner-driven experiments. This is now possible because of millions of visualization users nowadays. The collective intelligence of practitioners can provide an enormous knowledge base and resources to advance visualization research. In particular, for the types of research that requires tedious human efforts such as generating and annotating training examples, practitioner-sourcing would be a viable solution. For instance, Draco [28] allows us to encode the results of empirical studies into a set of constraints for the automatic generation of visualizations. We could structure the outcome of experiments in our platform to meet the definition of constraints to enable crowdsourcing the constraints.

Toward developing practical visualization guidelines. One motivation for developing VisLab was to address the lack of guidelines available for practitioners [14]. Although there are guidelines developed by practitioners themselves, such as DatatoVis[4] and DataVizProject[5], there is currently no visualization design knowledge based on empirical studies. The problem is that people often blindly follow the design principles such as "don't ever use pie charts!" or "always have

[4]https://www.data-to-viz.com/
[5]http://datavizproject.com/

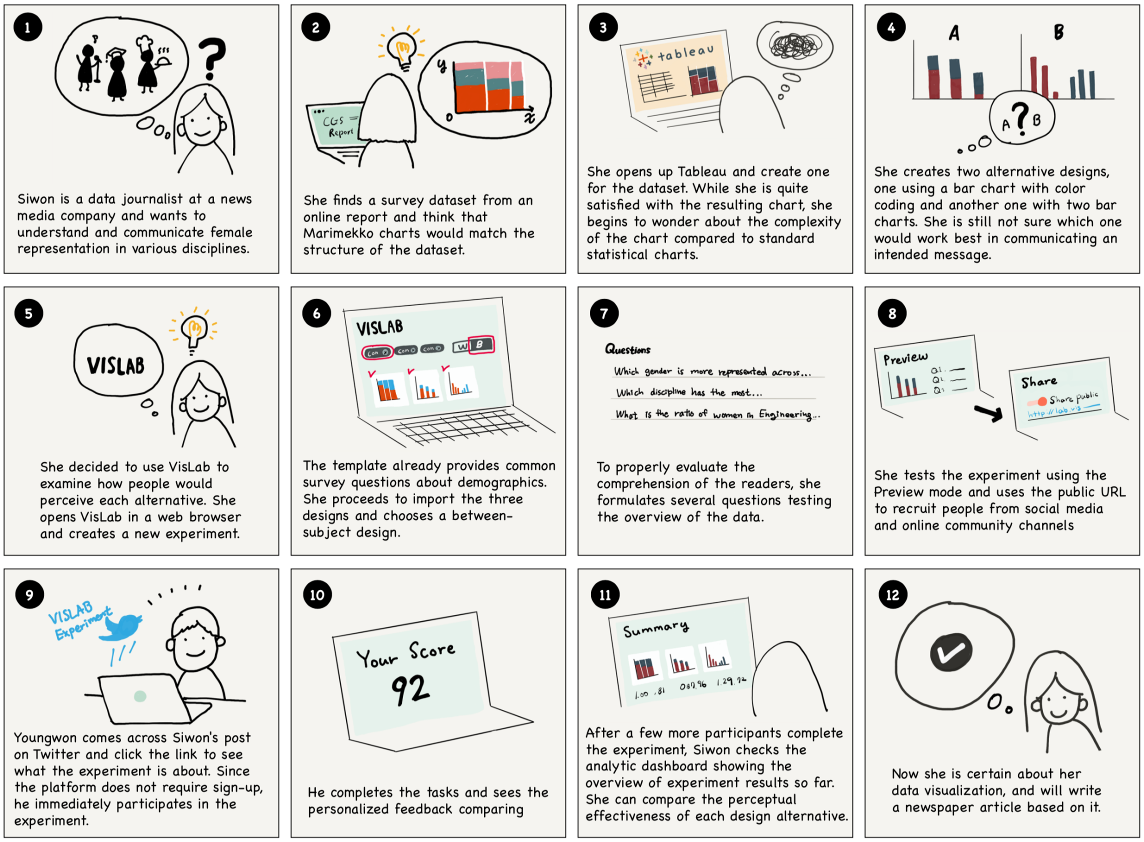

Fig. 6. A user scenario showing how a visual data journalist would use VisLab to answer questions about the trade-off of her design alternatives and derive design feedback on how actual participants would react to her visualizations.

a zero-baseline". Our platform is an initial step toward addressing this issue, providing opportunities for practitioners to test some of the existing design principles using their own visualizations. A next step would be generating actionable guidelines from the experiment results and made than publicly available on the platform to benefit other practitioners.

6 CONCLUSION AND FUTURE WORK

In this work, we demonstrated VisLab that aims to address the practical need of visualization designers in the wild by allowing them to evaluate visualizations through easy-to-deploy experiments rather than hunches and heuristics. In future work, we plan to deploy the platform to engage with actual users including various types of users such as practitioners, as well as researchers and educators. We also plan to incorporate statistical testings into the analytic dashboards to succinctly show which design variation is more likely to be effective. Also, currently, sharing an experiment with potential participants is possible but publicizing experiment designs is not possible. We hope to support forking existing experiments of others for reproduction and extension.

REFERENCES

[1] [n.d.]. Marimekko Chart. https://en.wikipedia.org/wiki/Marimekko#Marimekko_chart. Accessed: 2020-12-31.

[2] [n.d.]. Muller plot. https://en.wikipedia.org/wiki/Muller_plot. Accessed: 2020-12-31.

[3] [n.d.]. Tornado Plots chart. http://www.bewitched.com/demo/tornado/tornado.html. Accessed: 2020-12-31.

[4] [n.d.]. VisPerception. http://visperception.com/. Accessed:2020-12-31.

[5] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W Grant, Irene Reppa, Luciano Floridi, and Min Chen. 2012. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2759–2768.

[6] Rita Borgo, Luana Micallef, Benjamin Bach, Fintan McGee, and Bongshin Lee. 2018. Information visualization evaluation using crowdsourcing. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 573–595.

[7] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond memorability: Visualization recognition and recall. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 519–528.

[8] Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315.

[9] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385.

[10] Sheelagh Carpendale. 2008. *Evaluating Information Visualizations*. Springer Berlin Heidelberg, Berlin, Heidelberg, 19–45. https://doi.org/10.1007/978-3-540-70956-5_2

[11] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.

[12] William S Cleveland and Robert McGill. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229, 4716 (1985), 828–833.

[13] Vickie Curtis. 2015. Motivation to participate in an online citizen science game: A study of Foldit. *Science Communication* 37, 6 (2015), 723–746.

[14] Alexandra Diehl, Alfie Abdul-Rahman, Mennatallah El-Assady, Benjamin Bach, Daniel A Keim, and Min Chen. 2018. VisGuides: A Forum for Discussing Visualization Guidelines.. In *EuroVis (Short Papers)*. 61–65.

[15] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2016. The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 471–480.

[16] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. 2004. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*. Ieee, 17–24.

[17] Steve Haroz, Robert Kosara, and Steven L Franconeri. 2015. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1191–1200.

[18] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1943–1952.

[19] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 203–212.

[20] Jeffrey Heer, Jock Mackinlay, Chris Stolte, and Maneesh Agrawala. 2008. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE transactions on visualization and computer graphics* 14, 6 (2008), 1189–1196.

[21] Kevin Hu, Neil Gaikwad, Michiel Bakker, Madelon Hulsebos, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. VizNet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*. ACM.

[22] Ya-Hsin Hung and Paul Parsons. 2017. Assessing user engagement in information visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1708–1717.

[23] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 36.

[24] Younghoon Kim and Jeffrey Heer. 2018. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 157–167.

[25] Robert Kosara and Jock Mackinlay. 2013. Storytelling: The next step for visualization. *Computer* 46, 5 (2013), 44–50.

[26] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2011. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics* 18, 9 (2011), 1520–1536.

[27] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[28] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 438–448.

[29] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut instinct: Creating scientific theories with online learners. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 6825–6836.

[30] Jason Radford, Andy Pilny, Ashley Reichelmann, Brian Keegan, Brooke Foucault Welles, Jefferson Hoye, Katherine Ognyanova, Waleed Meleis, and David Lazer. 2016. Volunteer science: An online laboratory for experiments in social psychology. *Social Psychology Quarterly* 79, 4 (2016), 376–396.

[31] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM, 1364–1378.

[32] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. 2018. *Data-driven storytelling*. CRC Press.

[33] Warren Sack. 2011. Aesthetics of information visualization. *Context providers: Conditions of meaning in media arts* (2011), 123–50.

[34] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 341–350.