HCDE 411
Winter 2015

Mechanical Turk and Information Visualization

Using Mechanical Turk for Validating Information Visualization Principles

Group 5:
Keting Cen, Derek Lee, Hunter Mask, and Daniel Sandoval

Table of Contents

Abstract

Data visualization is, in essence, the practice of telling a story. Using a set of data, an author creates a compelling representation of the information contained in order to impart new, previously undiscovered insights to the target audience.

As visual storytellers, data visualizers need to be able to effectively communicate the domain and nature of their information to their audience. This report proposes a method of determining the most favorable method of exhibiting a data set to a target audience. As discussed in the report, the method utilizes the crowdsourcing of input on A/B visualization tests in order to create a final, "crowdsourced" visualization composed of the most favorable visual elements, as identified by the target user base.

Though limited by a narrow time frame, our results did indicate the feasibility of such a system for determining ideal data presentation techniques. An initial test of this system was used to create a composite visualization of grant data in the San Francisco Bay area. Further studies might investigate ways to improve this process, including, but not limited to, the impact of further process automation and alternate A/B testing methodologies.

Introduction

Visualization of data is a key way to communicate critical pieces of information that might otherwise be lost on a target audience. After studying the various ways to visually encode such information, our group sought out to determine how we might best communicate a particular set of data. After a series of design iterations, our focus turned to determining how best to communicate any data set in general as well as how one might discover this information. This led to our initial hypothesis that the use of crowdsourcing could be helpful in creating an effective visualization.

To perform a crowdsourced data collection, we decided to utilize Amazon's Mechanical Turk ("MTurk") technology. Mechanical Turk is a service developed by Amazon initially created to eliminate duplicate product listing pages from the retail company's own website. The premise of the service is to hire humans to perform tasks that, while generally minor, would be computationally taxing or impossible if performed by a digital system (NPR). It is now used across a variety of academic disciplines, being utilized to perform a range of tasks from building speech and language data (Callison-Burch et al.) to human behavioral analysis within the realm of theoretical biology (Rand).

Previous Work

Several similar studies have been conducted previously that consider the role of Mechanical Turk in usability testing. Results are mixed, but concur that the Mechanical Turk service can be extremely effective if properly utilized under the correct conditions.

In the study, "What Makes a Visualization Memorable?", authors Borkin et al. utilize the Mechanical Turk to determine the aspects of data visualizations that create a sense of "memorability" with the viewer -- that is to say, aspects which allow a visualization to persist for the longest duration in the viewer's memory. The study utilized pre-existing visualizations, imported from print and digital publications, which were then sorted into categories based on their type (for example, bar, line or scatter plots). Users were tested by being presented with a sequence of visualizations, responding by pressing a key when they recognized a repeated visualization. Based on the results from this testing, Borkin et al. were able to determine that more unique visualizations, such as matrix/grid representations, were significantly more memorable than more traditional visualization types such as bar graphs. They also concluded that minimal visualizations with a high "data-to-ink" ratio were less memorable than their more noisy counterparts (Borkin et al.). The wealth of insight gained by Mechanical Turk during this study establishes a precedent for the feasibility of our project's intended goal.

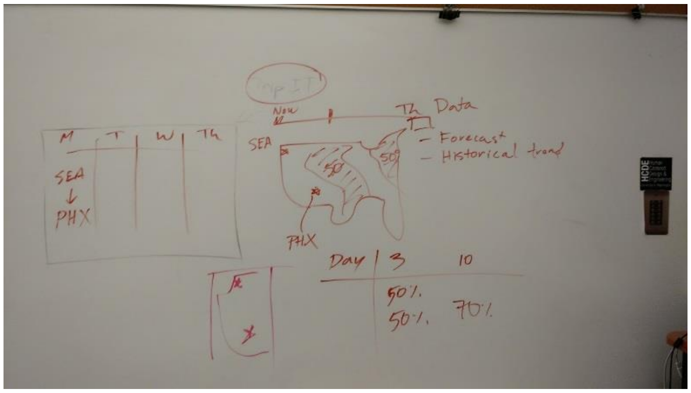
Our study differs somewhat from that conducted by Borkin et al. in that it focuses less on the memorability of a visualization and more on its ability to effectively communicate a given concept to the end user. Whereas Borkin et al.'s study is concerned primarily with ways to better instill a given visualization into the viewer's memory, our study seeks to establish a method with which to make the most widely understandable visualization possible through the utilization of crowdsourcing.

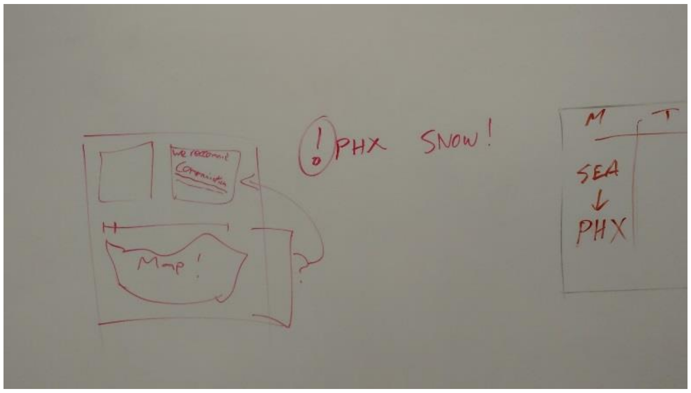
Our study's core foundation is based on the findings of the study, "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design", by Heer & Bostock. This study similarly uses mechanical turk to determine the best methodology for visualization. As noted by the study's findings, Mechanical Turk is well-suited for "research on the development of cultural markets". Our study seeks to modify Heer & Bostock's methodology by also noting the contextual data of participants, including cultural background and usage patterns.

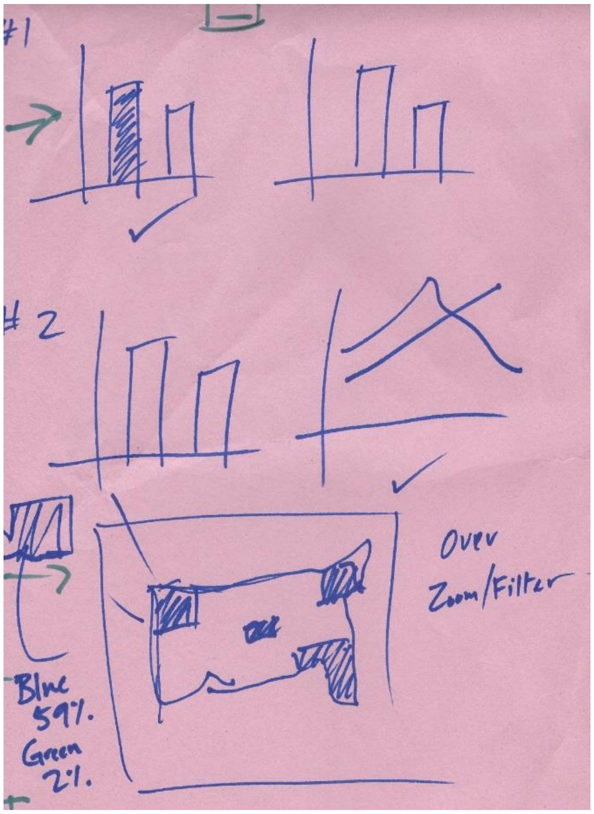
Design Process

Initial Ideation

In the initial phase of the ideation process, we decided to focus on the domain of transportation, and the problems faced by commuters and travelers when moving from one location to another. This design aimed to visualize publically available transportation data in a way that assisted commuters or regular users of public transportation in making decisions such as when to leave, what method of transit to take, or how long their journey would be. We sketched out a series of ideas (below) and even interviewed various commuters and travelers to better understand how to disseminate transportation information in a visual form.







We quickly found that, although these goals were valid, they had largely already been explored and solved by a number of other studies. We then set out to determine how else we might better portray transportation data. At this point, our focus widened to the larger question of how we might better portray any given data set. This led us to arrive at our final topic of exploration - evaluating and refining visualization techniques in order to empower users to create the best possible visualization, tailored specifically to maximize a given data set's impact on a given target audience.

Second Iteration

At this point in our design process, we set out to explore ways in which we could improve the visualization creation process for data storytellers everywhere. We formulated an initial user story for feature discovery:

> *"As a data storyteller, I need to be able to determine how best to visually encode information so that I can communicate my message effectively."*

As data visualizers ourselves, we knew that any given data set has a wealth of information best accessible through visual encoding. However, we also know that the accessibility of that information, even when visually encoded, is dependent on the understanding of the target audience.

Several standards on data visualization exist that argue certain techniques and methodologies in order to maximize this understanding. Articles such as Shneiderman's "The Eyes Have It" have been widely regarded as the basis for which information visualizations should be constructed by. Our new design sought out a methodology to verify these kinds of rules on a per-instance basis. As noted by others in the field of information visualization, such standards are often prone to old data, generalizations or biases such as those based on gender or culture (Marcus). With this in mind, we decided to use the power of crowdsourcing to obtain live, contextually relevant information on how best to portray a given set of information visually.

In designing the A/B testing component of our tool, we chose to utilize Amazon's Mechanical Turk service in order to recruit respondents. Mechanical Turk has a strong history of usage in research and testing, with scopes ranging from psychological studies to those in the domain of theoretical biology (Callison-Burch et al., Rand). Mechanical Turk has also been shown to attract a wide demographic of participants, varied fairly evenly in culture, gender, income, education level and age (Ipeirotis). As such, we felt confident in the validity of results gained from this service and their subsequent ability to achieve the needs of our tool.

Data Analysis

The underlying concept of our revised design was to provide a means for data visualizers to create effective, targeted visualizations that maximized understanding within the context of the given audience and data set. With this in mind, we envisioned that the data collected by our tool would allow visualizers to customize a visualization to this end. For this iteration of the design, much of this process involved manual manipulation of both testing data and visualization characteristics. The A/B testing segment, created in PHP, fed results into a MySQL database, which could then be queried by visualization creators to gain insights on how best to structure their final work. A minimally-designed set of web pages were also created to automatically deliver readable results to the end user, albeit in a less manipulable form. In future iterations, as noted in the discussion section below, we hope to further automate this process, including automatic test generation and final visualization analysis, as a means of further aiding in the visualization creation process.

For our pilot test of this tool, we used a dataset provided by the City of San Francisco on grant funding provided to various non-profit organizations within the San Francisco Bay Area. The data provided a good way to test our hypothesis using geographically-encoded information. Furthermore, most dimensions are easy to understand and close to people's life, so the complexity or abstractness of the data are unlikely to be a burden on survey participants.

Upon accepting our Human Intelligence Task (HIT) in Mechanical Turk, respondents would be asked to follow a link to our survey. As noted in the discussion section below, we later discovered, based on further discussion with other Mechanical Turk researchers, that a bias does exist by "Turkers" against this kind of link. At this point, respondents were to be asked a "screener" question, designed to eliminate invalid responses from participants who were not paying attention. Upon failure to pass the screener, respondents were presented with a message stating that they had failed to pay attention and would not be given a confirmation code, thus denying payment for the HIT. On the contrary, if the respondent passed the screener, they would be taken to the page with the relevant segment of the survey.

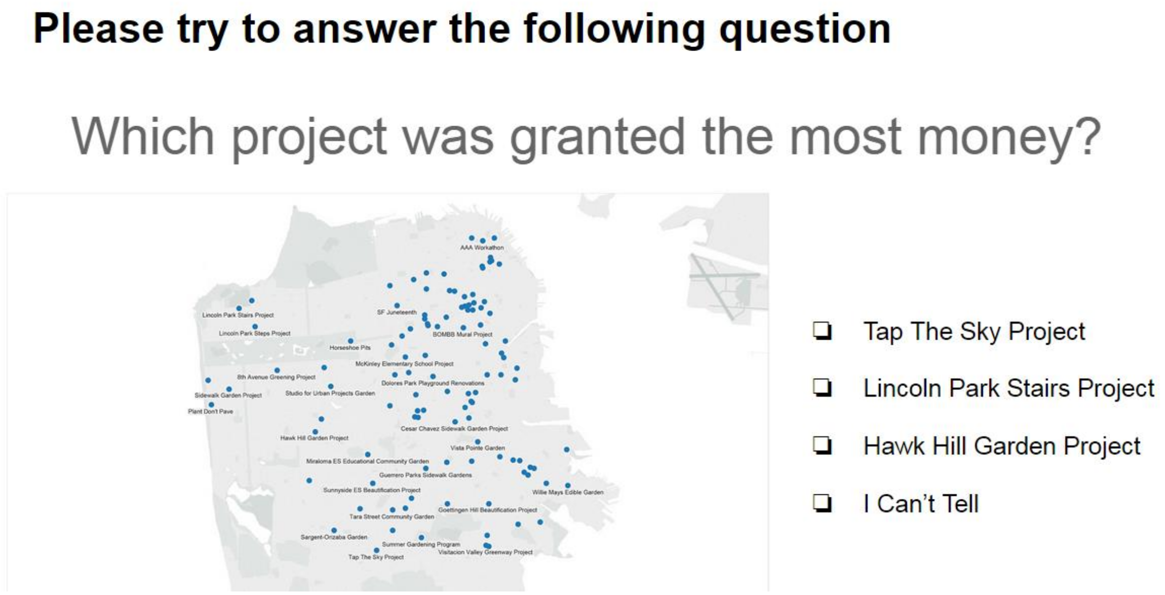
This section initially presented users with a side-by-side comparison of two visualization tests, each with a modified characteristic, and asked them to select which visualization better communicated the intent of the visualization. In our tests, we asked respondents to make a decision based on the displayed distribution of grant funding in the Bay Area. Upon submission, choices were recorded in the database and the user was given a "validation code" of which they were to enter on the Mechanical Turk website. This code was then used to approve payment to those participants who had successfully completed the test.

Data Collection Obstacles

Screener

The literature behind having a screener is strong. The logic behind the formulation of our screener largely comes from "Are Your Participants Gaming the System? Screening Mechanical Turk Workers" by Downs, Holbrook, Sheng, and Cranor (2010). Screeners are key to collecting valid data. The authors state that "studies aiming to correlate performance or responses on different tasks rely on meaningful data from each participant. Rather than aggregating noisy data, an alter[n]ative strategy would be to develop a reliable method of screening participants to remove the subset of those gaming the system" (Downs et al. 1). The paper states that the most effective screeners are ones that are "designed to appear as a formality, following the logic of the study task" (Downs et al. 2). It further states that basic screeners, such as those that simply ask the participants to answer if they are reading the question, "only catch the most egregious of participants, violate Gricean norms, ... and set a tone of distrust for the remainder of the task" (Downs et al. 2). Screening questions are important for data validity and to ensure that respondents are actually performing the task at hand.

From the beginning, we understood that payment could be an issue. Our understanding of Mechanical Turk and crowdsourcing human intelligence tasks (HITs) was that payment could be contingent on a 'valid response'. Discussions of the payment were included in the earliest phases of our research. Knowing that we would be paying for these HITs out-of-pocket, we wanted to ensure that our money was not being wasted on data that we would ultimately have to exclude.



Please try to answer the following question

Which project was granted the most money?

❏ Tap The Sky Project

❏ Lincoln Park Stairs Project

❏ Hawk Hill Garden Project

❏ I Can't Tell

AAA Workathon
Lincoln Park Stairs Project
SF Juneteenth
Lincoln Park Steps Project
BOMBB Mural Project
Horseshoe Pits
McKinley Elementary School Project
8th Avenue Greening Project
Dolores Park Playground Renovations
Sidewalk Garden Project
Studio for Urban Projects Garden
Plant Don't Pave
Cesar Chavez Sidewalk Garden Project
Hawk Hill Garden Project
Vista Pointe Garden
Miraloma ES Educational Community Garden
Guerrero Parks Sidewalk Gardens
Sunnyside ES Beautification Project
Willie Mays Edible Garden
Goettingen Hill Beautification Project
Tara Street Community Garden
Sargent-Orizaba Garden
Summer Gardening Program
Visitacion Valley Greenway Project
Tap The Sky Project

Unfortunately for us, our solution had major issues that simply we did not foresee. Our idea of a screener was a page that would prevent illegitimate users from proceeding with the HIT. We created a screener that presented a simple task. The map showed different data points on the San Francisco city map representing different government project grants, but without any meaningful variation between them. The question was posed, "Which project was granted the most money?" and four options were presented: three different projects and "I can't tell". If users chose anything other than "I can't tell", we redirected them to a page that informed them that they were not eligible for the HIT and did not pay them.

After launching the Mechanical Turk HIT, we noticed low data collection rates (three responses in 48 hours). One of our initial concerns was with the screener. We decided to investigate by performing rapid usability testing. We pulled the step one screener page up to nine different participants and recorded whether they chose "I can't tell" and a heuristic judgement on how quickly they performed the task (fast, medium, or deliberated). In the studies, two out of nine chose the wrong answer. Additionally, four out of nine deliberated or took a medium amount of time (approximately between 30 and 90 seconds).

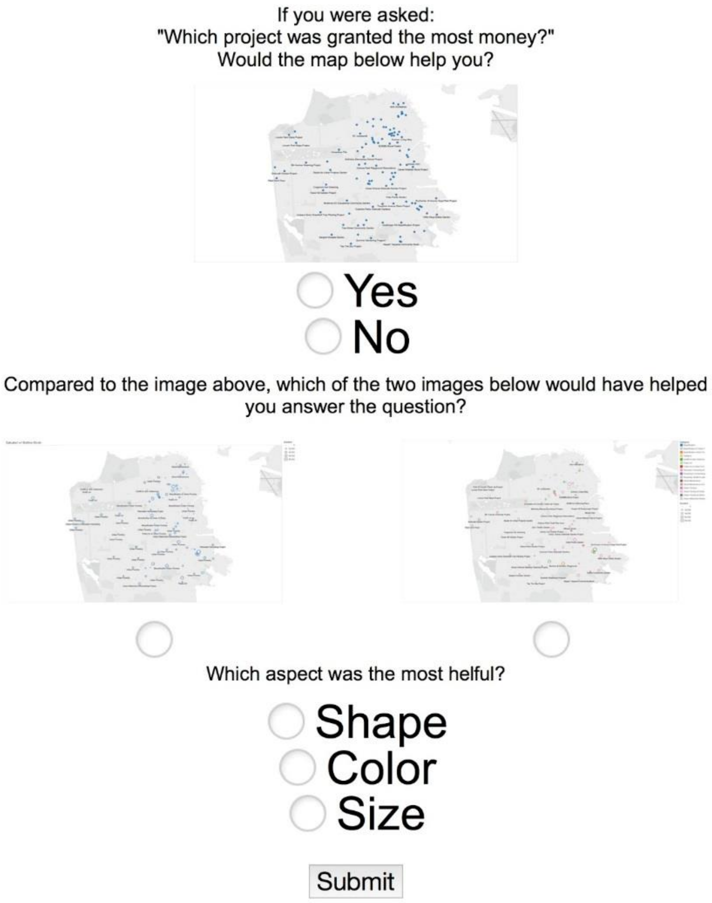
The rapid usability testing revealed the major issues that the screener introduced. As the intent was to eliminate only the people who were 'button-mashing for money', it should not have screened out legitimate participants. Similarly, it should not have taken more than 30 seconds. The screener was intended as a low-friction page, meant to move participants on quickly. Adding even 30 seconds, let alone 90 seconds, was bad for our completion times, and consequentially our hourly HIT pay-rate per hour. As it turned out, we should have also paid the participants that failed the screener, if at a lower rate than if they successfully completed the HIT.

Ultimately, our suspicions were largely confirmed by talking to Gary Hsieh, Associate Professor in Human Centered Design and Engineering (HCDE) at the University of Washington and one of our participants, Lucas Colusso, an HCDE PhD candidate. There are a lot of nuances with Mechanical Turk data collection, and having an unpaid screener did not bode well for our study's participation rates. According to them, Turkers have a lively community where they trade information about HITs, including which to avoid for what they perceive to be low paying. They often collude to avoid HITs like ours, due to the rigidity of our screener.

"Triad" Design

The concept of using a triad of visualizations for comparison comes from the 2014 Demiralp, Bernstein, and Heer paper "Learning Perceptual Kernels for Visualization Design". Demiralp et. al measured different so-called perceptual kernels and used crowdsourcing to determine the visual similarities different encodings, such as square, triangles, and circles. They tested five different judgement types: pairwise ratings and Likert scales of five and nine, triplet ranking with matching and discrimination, and spatial arrangement. Through this study, they found that triplets "exhibit the least inter-subject variance, are less sensitive to subject count, and enable the most accurate prediction of bivariate kernels from univariate inputs" (Demiralp et al. 9).

Using this, our HIT involved a triad, where users chose which visualization was the most helpful, using a reference map.



If you were asked:
"Which project was granted the most money?"
Would the map below help you?

○ Yes
○ No

Compared to the image above, which of the two images below would have helped you answer the question?

○ ○

Which aspect was the most helful?

○ Shape
○ Color
○ Size

Submit

As seen in screenshot, users picked between two of the randomly paired map encodings, compared against a neutral representation of the data. This neutral representation was the map used during the screener, which we eliminated (see "Screener" section).

Unfortunately, using triplets may have been confusing for the participants. After adding the social media collection method, we were opened up to comments and questions that were not possible with Mechanical Turk. We received several questions and comments regarding that it didn't make much sense why it was formed as a triplet rather than a pairwise comparison. By the time we switched to the social media collection, it was too late to investigate or iterate the issue further. As such, the utility of the triplets as outlined in Demiralp et al. may have been compromised.

Revised Design

Due to the relative lack of responses received from Mechanical Turk within our timeframe (approximately two weeks), possible causes of which are noted in future sections, we decided to expose our survey to a wider audience through social media. In doing this, we added a "referral code" to the survey's metadata. This code, attached to the link clicked by respondents, represented a source through which respondents were recruited to the survey. Responses could then be separated based on source, thus adding a mechanic to perform A/B testing on multiple, focused demographics.

In our pilot test, we exposed our A/B testing element to several demographics. Referral code 1 was assigned to participants from Mechanical Turk, the original source. Referral code 2 was assigned to participants from UX-focused social media groups such as HCDE student body groups. Referral code 3 was assigned to participants from general social media sources, recruited through means such as Facebook posts or direct messages. Finally, referral code 4 was assigned as the "catch-all" for participants recruited through any other source, including direct interaction.

Results

Though mired by a few setbacks, our pilot test of the tool we created yielded a great deal of information on how our hypothesized design methodology could and could not work. This experience provided insight into the validity of the general visual design standards in a general context.

Our pilot went through several iterations during the testing process. The end result first asked users if the "stock" visualization would help them determine the flow of grant money, then asked which of the A/B test visualizations, compared to the original, would have helped them. Finally, it asked users which aspect of the visualization chosen was most helpful.

At the conclusion of our pilot study, 16 responses were received from Mechanical Turk respondents. 76 responses were received from UX-centered social media groups, 117 were received from general social media, and 2 were received from untagged sources.

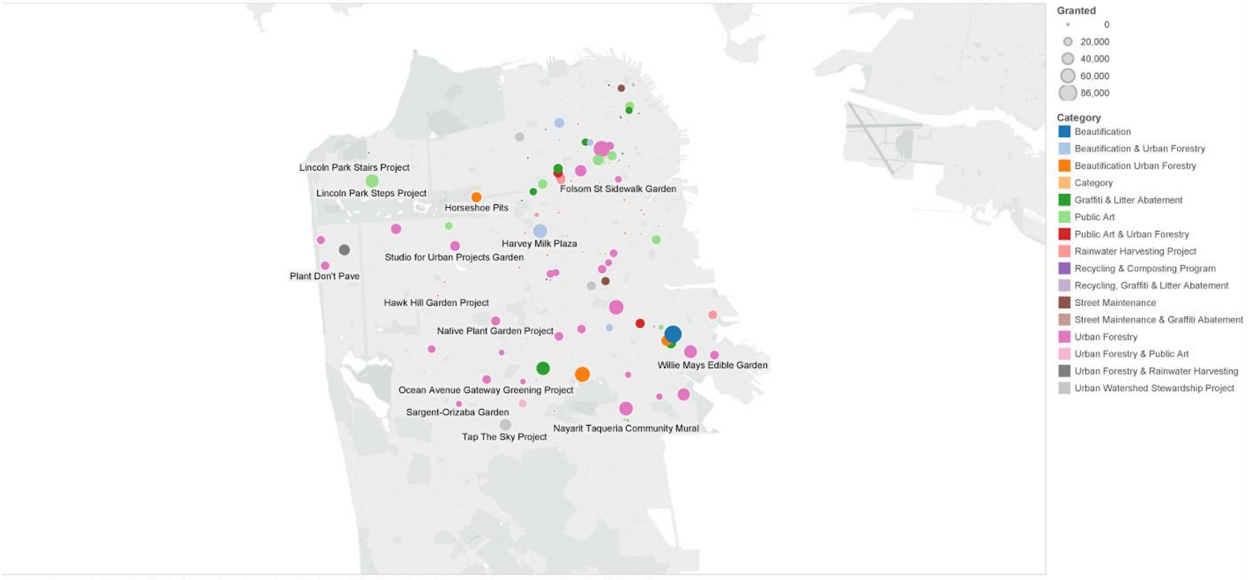
Respondents appeared to show high preference toward certain visual characteristics in our pilot test. For example, an extremely high preference was shown toward filled shapes versus unfilled ones to represent data points -- 37 respondents tested chose the filled option while only 5 chose unfilled. In addition, 32 participants preferred a labeled circular shape versus 10 who preferred a labeled square one (22 also preferred the unlabeled circle versus 11 for the unlabeled square), and 27 participants preferred size to be used as a dimension to encode funding amount versus 6 who did not.

On some characteristics, however, participants were more indecisive. Regarding the use of labels, participants were roughly split, with 19 preferring them and 17 against them. Along with the various choices and their reasons, we also recorded the time taken to complete the survey.

Visualization

After collecting enough data from both Mechanical Turk and social media, we finalized our dataset with a total of 211 responses total. From this dataset, we evaluated the amount of responses for each presented comparison in order to create a single "crowdsourced" visualization:

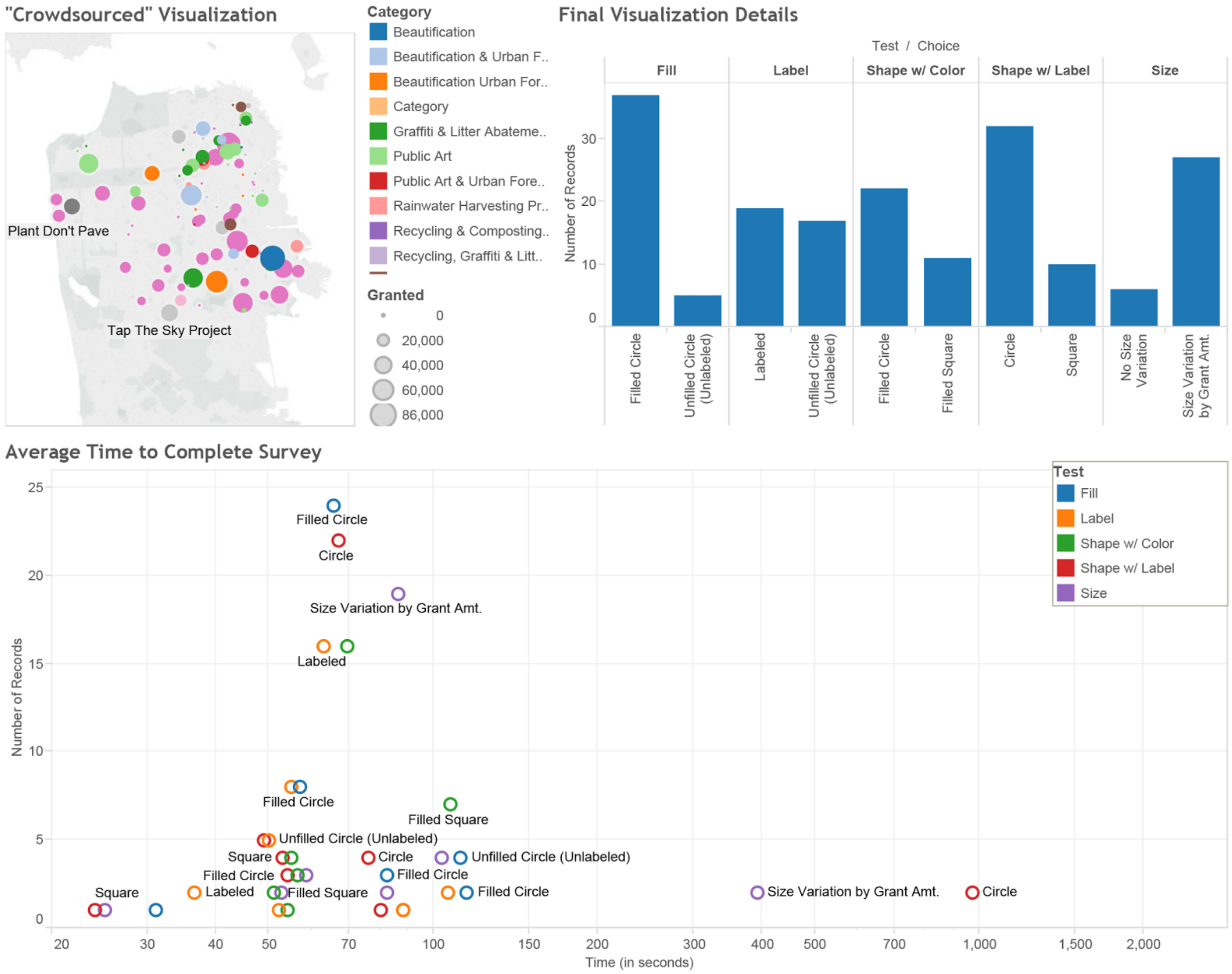
'Crowdsourced" Visualization


Lincoln Park Stairs Project
Lincoln Park Steps Project
Folsom St Sidewalk Garden
Horseshoe Pits
Harvey Milk Plaza
Studio for Urban Projects Garden
Plant Don't Pave
Hawk Hill Garden Project
Native Plant Garden Project
Willie Mays Edible Garden
Ocean Avenue Gateway Greening Project
Sargent-Orizaba Garden
Tap The Sky Project
Nayarit Taqueria Community Mural
Granted
0
20,000
40,000
60,000
86,000
Category
Beautification
Beautification & Urban Forestry
Beautification Urban Forestry
Category
Graffiti & Litter Abatement
Public Art
Public Art & Urban Forestry
Rainwater Harvesting Project
Recycling & Composting Program
Recycling, Graffiti & Litter Abatement
Street Maintenance
Street Maintenance & Graffiti Abatement
Urban Forestry
Urban Forestry & Public Art
Urban Forestry & Rainwater Harvesting
Urban Watershed Stewardship Project

Map based on Longitude and Latitude. Color shows details about Category. Size shows sum of Granted. The marks are labeled by Project Title.

We can see from this visualization that the collected data has built a crowdsourced map which seems to match the generally accepted visualization principles. Categories are mapped to color and text labels allow for viewers to gather more detailed information. The size is mapped to the grant amount, which is a common best-practice due to the human visual system being able to compare size relatively quickly (Scott, Slide 10). This verification of accepted best-practices makes sense due to the fact that we collect data from a wide audience with different backgrounds. In future studies, we would like to use our same methodology to challenge general visualization principles by collecting data from within certain cultural and societal contexts.

To explain this final visualization to an interested party - for instance, an information visualization researcher using our survey tool - we created an interactive dashboard to explore the reasons why this visualization proved to be the one chosen by the group of respondents. Our dashboard implements brushing-and-linking to support this exploration. The following image shows the default state of the dashboard.



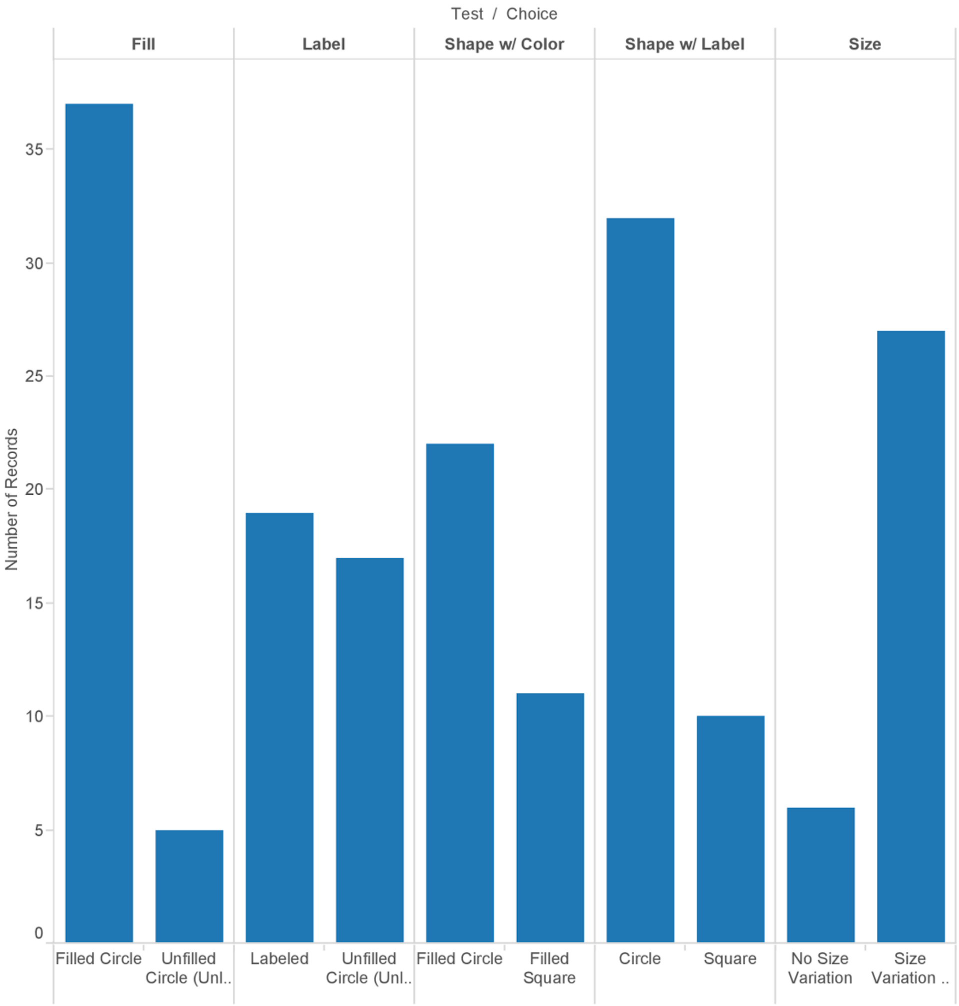
Live and interactive versions of the dashboard are available at the links below:

With mTurk Data:
https://public.tableau.com/views/MetaViz-FinalVisualizationLogScalewithmTurkData/Dashboard2?:embed=y&:showTabs=y&:display_count=yes

Without mTurk Data:
https://public.tableausoftware.com/views/MetaViz-FinalVisualizationLogScalewomTurkData/Dashboard2?:embed=y&:showTabs=y&:display_count=yes

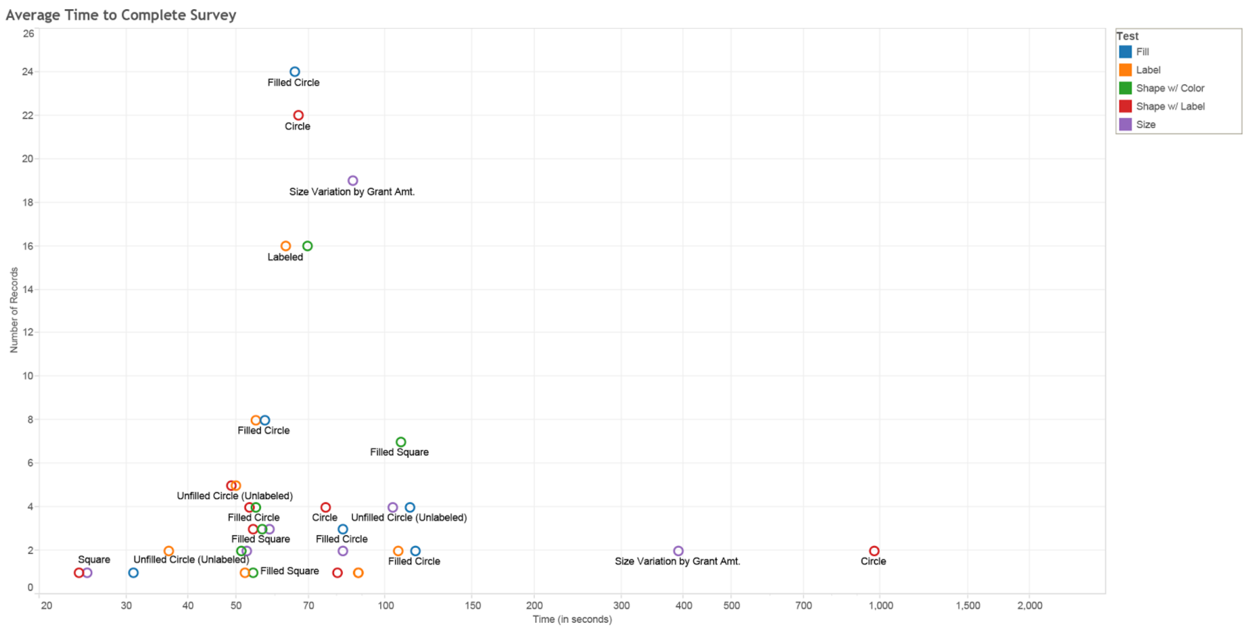
In order to explore why particular measures were chosen to be mapped to specific visualization elements, an interested researcher could first use the “Final Visualization Details” chart to understand how significant the chosen mappings were in the context of the survey’s comparisons.

Final Visualization Details



Sum of Number of Records for each Choice broken down by Test.

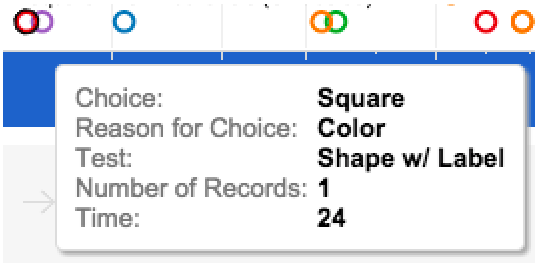
In our survey, we sought to compare multiple ways which dimensions could be displayed on a map. The results of these comparisons are directly represented on the adjacent map. In order to further understand these comparisons, a researcher could then look to the "Average Time to Complete Survey" to see if these choices were difficult or easy for the respondent to make.



Average of Time vs. sum of Number of Records. Color shows details about Test. The marks are labeled by Choice. Details are shown for Reason for Choice. The data is filtered on Referral, which keeps 1, 2, 3 and 4.

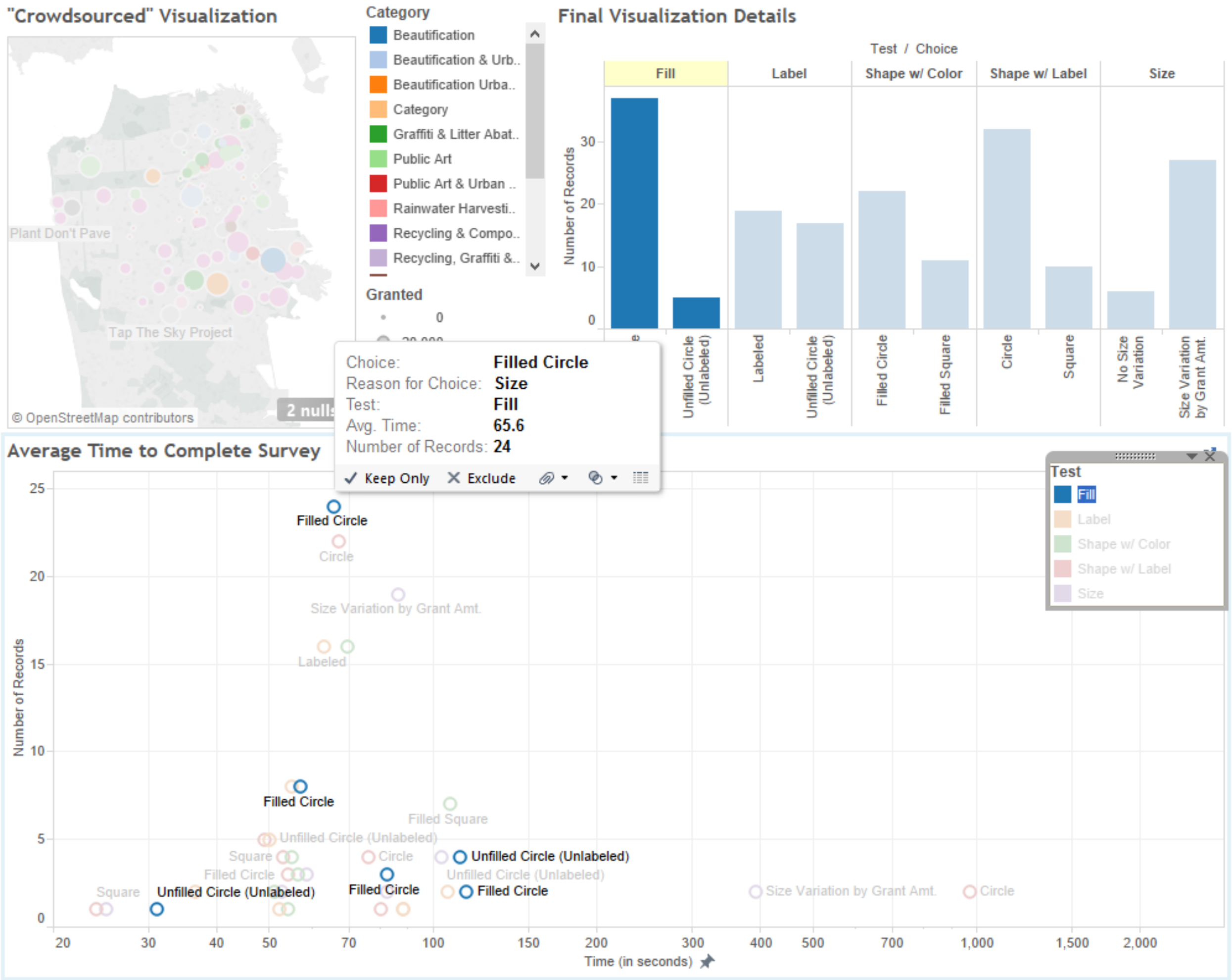
Our data shows that it took 24 seconds for our quickest response, a choice of a “Square” in the comparison between shapes with labels (square versus circle). Hovering over this data point (figure below) shows that the user thought the “color” was the most helpful.

Our data shows that it took 24 seconds for our quickest response, a choice of a “Square” in the comparison between shapes with labels (square versus circle). Hovering over this data point (figure below) shows that the user thought the “color” was the most helpful.



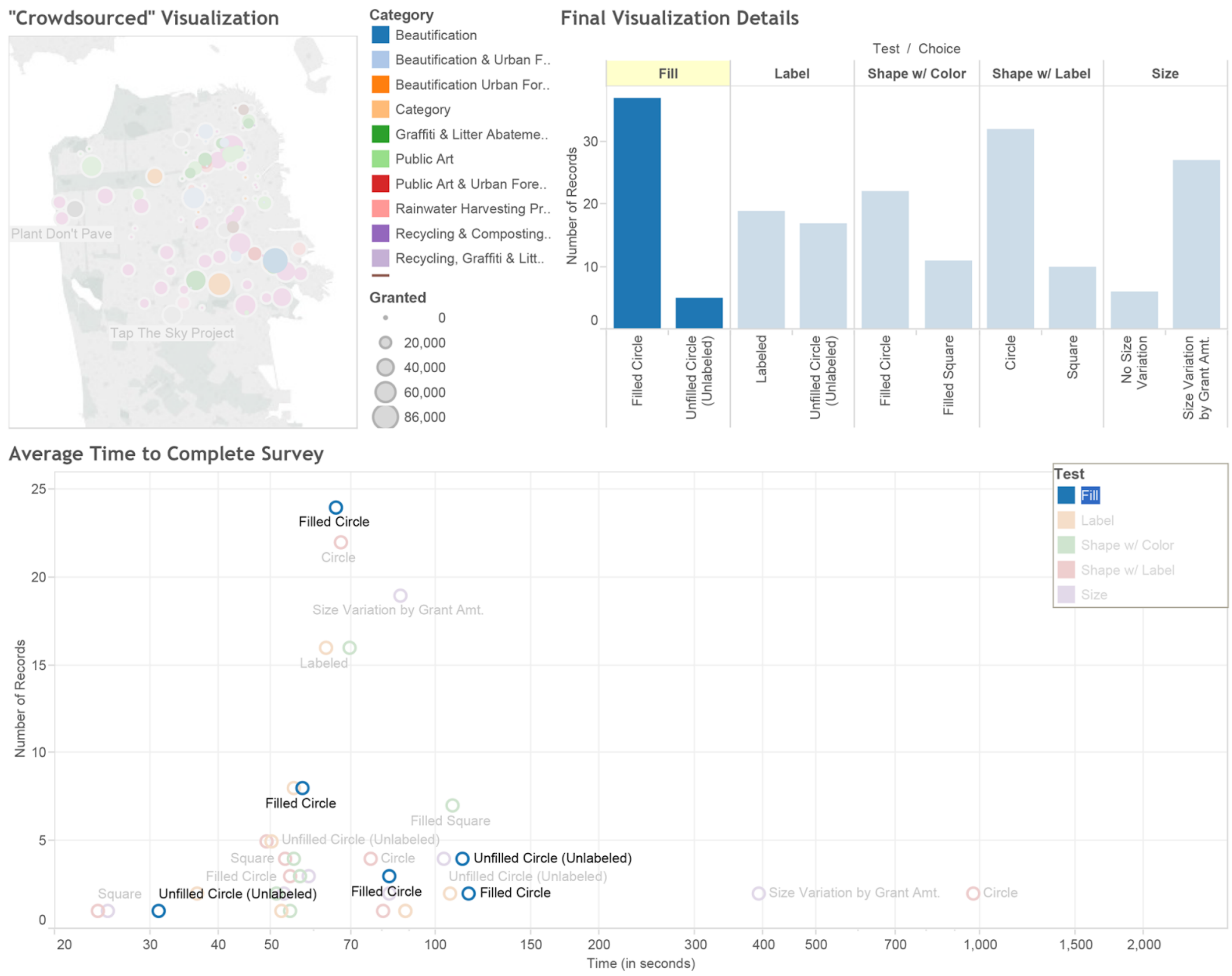
Choice: **Square**
Reason for Choice: **Color**
Test: **Shape w/ Label**
Number of Records: **1**
Time: **24**

This detail, along with additional information about why the respondent decided to choose one image over another (displayed in a detail-on-hover pop-up) allows for the researcher to verify the effectiveness of the chosen measure display from a qualitative standpoint. In the example given above, where “square” was chosen in 24 seconds, there is only 1 record in our dataset to display. It would be more useful, then, for a researcher to investigate data points with multiple records supporting the given choice. Selecting a particular comparison, for instance “fill”, and investigating the point with the highest number of records, displays the average time it took to make the given comparison along with additional information to better understand the reasons and difficulty of the decision (figure below).



Choice: Filled Circle
Reason for Choice: Size
Test: Fill
Avg. Time: 65.6
Number of Records: 24

Unfortunately, due to the fact that we did not standardize the collection to require the same amount of responses for each test case, researchers cannot compare the Final Visualization Details across tests. However, the provided details on demand between comparisons for an individual test - along with the brushed-and-linked comparisons for any given highlighted choice - display enough data to provide a basis for both a quantitative and qualitative analysis of the aggregate choices of survey respondents. By selecting a particular "Test", the researcher minimizes the scope of their analysis to better support comparisons within a particular test of a displayed measure.



Usability Testing

In order to better understand the usefulness of our final product, we contacted four researchers who used information visualizations in their academic endeavors. For each test, we described our process and methods as well as showed the survey as an example of what respondents were asked to do. We then displayed the dashboard and asked them to think-aloud as they described how it would be useful or not in any particular context. Every tester described the visualizations in clockwise order - discussing the map, then the details, and lastly the average time display.

Jeff Heer

Firstly, we talked to Jeff Heer, an Associate Professor in the Department of Computer Science and Engineering and director of the Interactive Data Lab. Notably, Professor Heer was a contributing author on several of the papers that guided this research. Professor Heer first pointed out that this tool would be useful for an overview of responses. He liked that he could quickly remove outliers from the data and compare aggregate choices within a category.

He did, however, mention an important aspect of our data that seemed to be missing, a qualitative understanding of the choices that the respondents made. He suggested that we add a free response question as well the ability to encode these responses into navigable data. After further prompting, he stated that this new information would allow researchers to understand more about who exactly is within this target population and why, in particular, they made the decisions displayed on the visualization. “Looking at this data,” he said, “it’s hard to tell if people are choosing the ‘best visualization’ or just the map they prefer.”

Along with this critique, he added a few suggestions that would make the current view more useful for a researcher:

1. Multiple views to walk users through a two-step workflow - displaying the “crowdsourced” visualization first, then displaying the meaningful data in a secondary screen.
2. Statistical representations of the data (i.e. mean, standard deviation, and other statistics) to allow researchers garner a basic understanding of the breadth of information available in the dataset.
3. A comparison of the collected data with the current design principles to show how the data-collection population’s visualization differs from one generated with the general design guidelines in mind.

Lucas Colusso

Following our interview with Jeff, we interviewed PhD candidate, Lucas Colusso, of Human Centered Design and Engineering at the University of Washington. One of his first comments supported Jeff’s as he stated the tool “would be useful to create a hypothesis from the dataset.” He even expressed that our dashboard would have been helpful in his current research as he studies responses from people with various demographics. On top of this, he mentioned that the visualization could only be used to “get a broad sense of what is happening,” saying that he would have to depend on other tools and visualizations to further explore the meaning behind the choices.

When asked why this was the case, he questioned the validity of the data and suggested we implement a way to filter out responses based on the amount of time respondents spent taking the survey. “Under ideal conditions,” he said, “time could be a screener” and it would be useful if our dashboard were able to determine an “ideal” time window. In this suggested workflow, as Lucas proposed, researchers could specify a custom response time cut-off, perhaps with a suggestion of such a time to filter out responses they might want to exclude.

Finally, Lucas suggested that we color-code the comparisons to make them easier to analyze.

Marilyn Ostergren

We also interviewed Marilyn Ostergren, instructor of Informatics 424, Information Visualization and Aesthetics at the University of Washington. Her feedback agreed with both Jeff and Lucas as she would have liked to see “trusted data” only. She made various suggestions as to how we could increase the validity of the responses through a larger sample and a more coherent survey design.

Like Jeff, she worried that our respondents were choosing the answers they preferred versus what was really most effective. A survey design with a stronger qualitative aspect, she stated, would have resolved this issue. She also supported Jeff's suggestion to compare our crowdsourced visualization with current design standards, which we were unable to implement into our final design in the given timeframe.

She had a few notes concerning our naming conventions and advised us to be careful when explaining that each of the five tests are separate, despite being in the same histogram. In our Final Visualization, the color-coding suggested by Lucas addresses this problem as it separates the various comparisons.

Daniel Perry

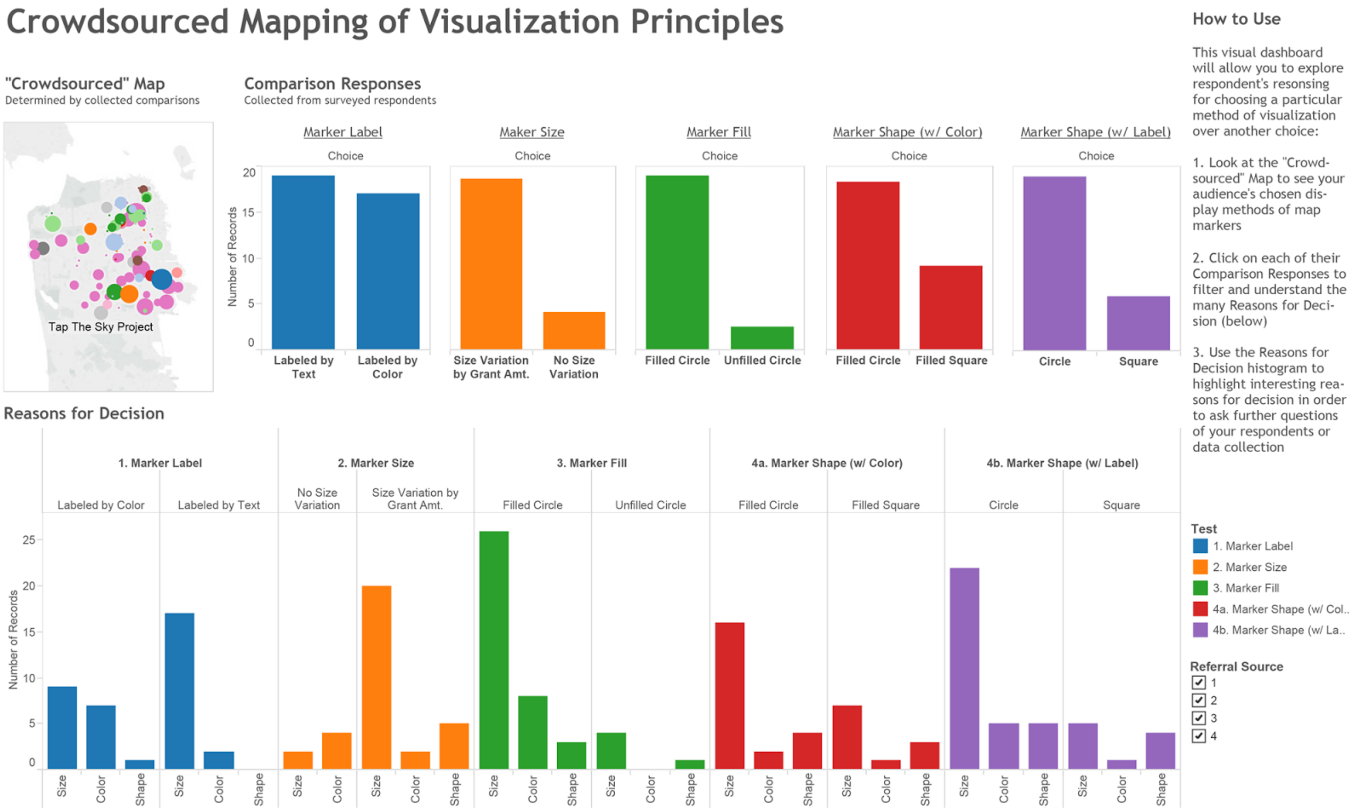
For our final usability test, we interviewed Daniel Perry, another PhD candidate in HCDE. Daniel was able to spend a significant amount of time with our product and supported many of the statements gathered from previous usability testers. He agreed that colors to separate the choices would make their segmentation more clear as well as allow for a quicker analysis. Unlike the other testers, however, he had a hard time wrapping his head around the meaning of the final visualization. After further explanation of the survey methods and goals of the experiment, he agreed that the combination of our data collection and visualization was a useful tool in understanding how a target population responds to the way geospatial data is displayed.

Daniel suggested that we support our experiment with further research. He would have liked to have seen more information about respondents including geolocation, reference source, and an ability to control our collected response time. In particular, he focused on how time is useful in this analysis at all, since we do not know the setting or cognitive load of the surveyed individual. In response to his critique, we decided to remove time from the final visualization. "If [we] wanted to use [time] in the future," he suggested, "[we] could collect only the amount of time the respondent spent comparing the two images."

Since he did not believe that time would be helpful for a more detailed view, Daniel and our group worked together to find a better way to represent the qualitative data related to our respondent's reasons for their decisions. He suggested adding a way to breakdown respondents' choices with a filter, rather than hiding the information behind a tooltip. We agreed and responded to this suggestion by adding a brushing-and-linking interaction to a set of two histograms - one containing the respondent's decisions, the other containing their reasoning. This final implementation agrees with Daniel's "ideal visualization" where "there are sections of colors and relevant sections can be zoomed and seen [up close]."

Final Visualization

Our final visualization combines our responses to the suggestions mentioned above as well as further implementations that we feel may be helpful to data visualizers.



Crowdsourced Mapping of Visualization Principles

How to Use

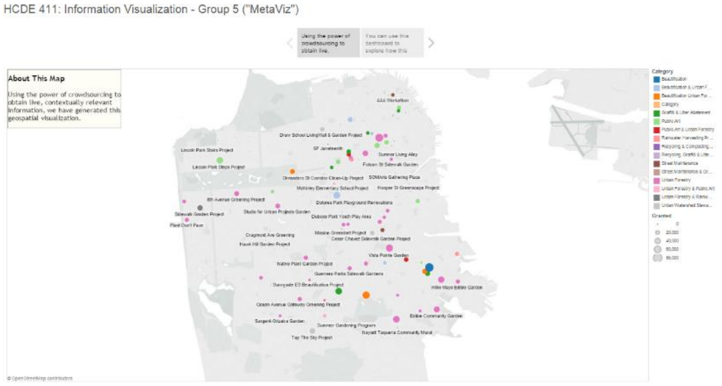
This visual dashboard will allow you to explore respondent's resoning for choosing a particular method of visualization over another choice:

1. Look at the "Crowd-sourced" Map to see your audience's chosen display methods of map markers

2. Click on each of their Comparison Responses to filter and understand the many Reasons for Decision (below)

3. Use the Reasons for Decision histogram to highlight interesting reasons for decision in order to ask further questions of your respondents or data collection

We have created two versions of the visualization, one for first-time users and another for returning users. Our first-time user visualization uses Tableau's "Story" feature to introduce the visualization. If they understand the visualization and would like a more responsive version of the dashboard (one that fits in the screen), our users can navigate to the version without Tableau's "Story" feature - which tends to add unnecessary scrollbars to horizontal displays (see screenshot below).

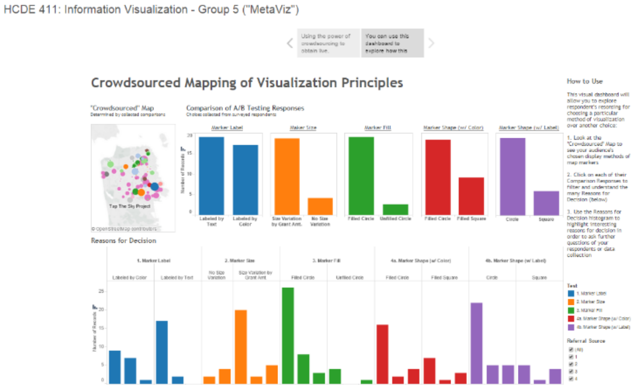


HCDE 411: Information Visualization - Group 5 ("MetaViz")

About This Map

Using the power of crowdsourcing to obtain live, contextually relevant information, we have generated this geospatial visualization.

Using the power of crowdsourcing to obtain live.
You can use this dashboard to explore how this
Category
Granted



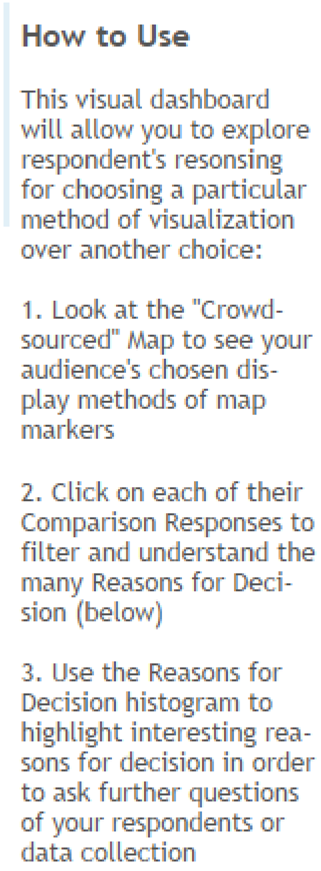
HCDE 411: Information Visualization - Group 5 ("MetaViz")

Using the power of crowdsourcing to obtain live.
You can use this dashboard to explore how this
Crowdsourced Mapping of Visualization Principles
"Crowdsourced" Map
Comparison of A/B Testing Responses
Reasons for Decision
How to Use
Test
Referral Source

The story version of the visualization allows our users to understand the context of the "Crowdsourced" Visualization by displaying first, before letting them see the data that formed it.

Final Viz w/ Story
https://public.tableau.com/views/CrowdsourcedMappingofVisualizationPrinciplesHCDE411-Story/Story1?:embed=y&:showTabs=y&:display_count=yes

Final Viz w/o Story
https://public.tableau.com/views/CrowdsourcedMappingofVisualizationPrinciplesHCDE411-FinalVisualization/Dashboard?:embed=y&:showTabs=y&:display_count=yes



How to Use

This visual dashboard will allow you to explore respondent's resonsing for choosing a particular method of visualization over another choice:

1. Look at the "Crowdsourced" Map to see your audience's chosen display methods of map markers

2. Click on each of their Comparison Responses to filter and understand the many Reasons for Decision (below)

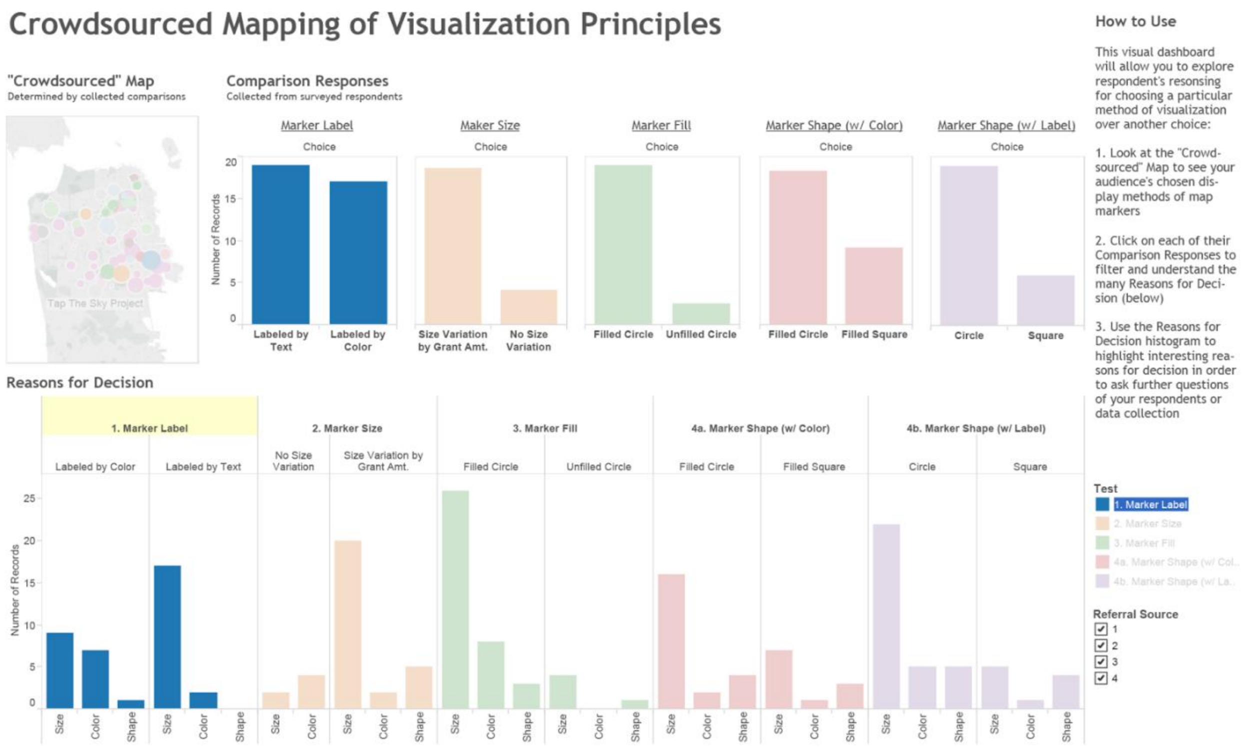
3. Use the Reasons for Decision histogram to highlight interesting reasons for decision in order to ask further questions of your respondents or data collection

The final visualization now has a title, “Crowdsourced Mapping of Visualization Principles” which gives viewers a context of understanding within the dashboard itself. Along with this contextual reminder, we have also provided more detailed titles for each visualization within the dashboard. As in the story, the map is explained as a combination of “selected comparisons” while the details of the responses themselves are described as “Collected from surveyed respondents” in the subtitle of the “Comparison Responses” visualization.

Along with more detailed titles, our dashboard also provides a help menu entitled, “How to Use” which walks through the three visualizations in a workflow we observed during our usability tests. We hope that this new format will help first-time users understand the dashboard without the need of any additional context.

Our new dashboard extends the functionality of our first as it implements brushing-and-linking to give our users an in-depth understanding of the comparisons which generated the “Crowdsourced”

Map. Clicking on any particular test (on the right), highlights that test in both the "Comparison Responses" and "Reasons for Decision" visualizations (screenshot below). This allows for a researcher to focus on the highlighted decision without the interference of the other colors on the dashboard (Heer & Shneiderman, 12).



How to Use

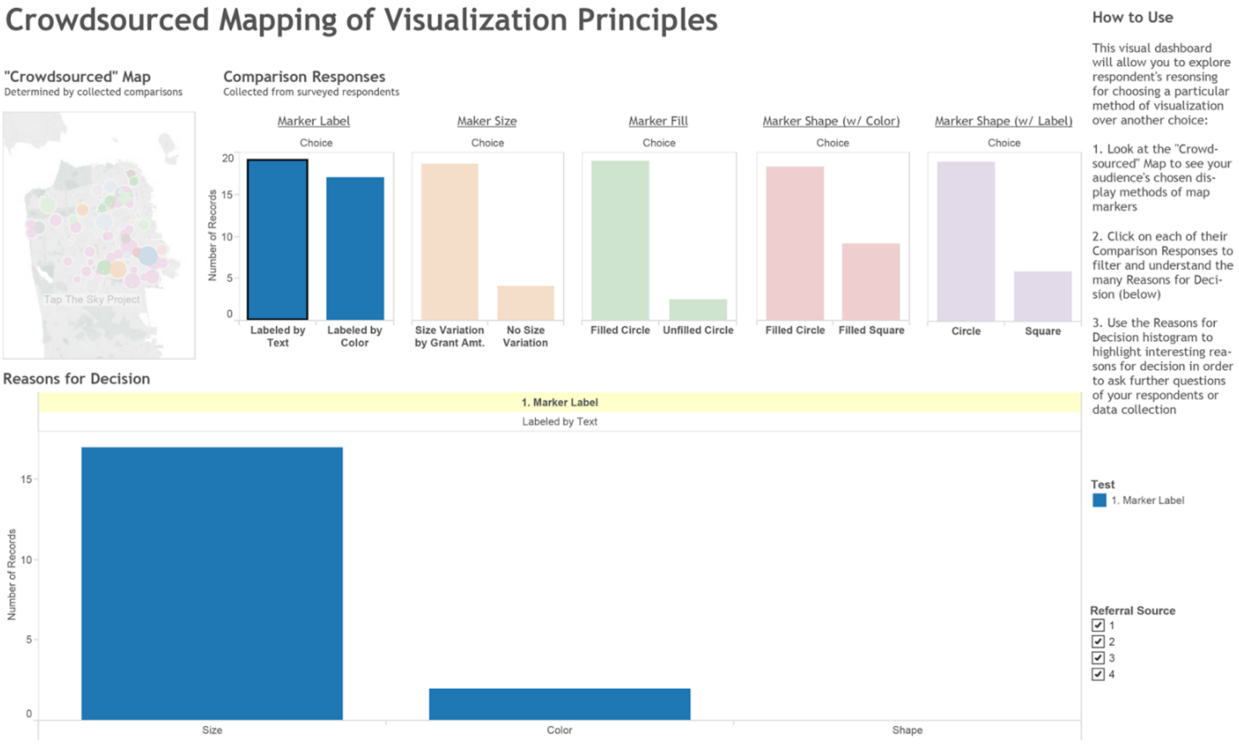
This visual dashboard will allow you to explore respondent's resonsing for choosing a particular method of visualization over another choice:

1. Look at the "Crowd-sourced" Map to see your audience's chosen display methods of map markers

2. Click on each of their Comparison Responses to filter and understand the many Reasons for Decision (below)

3. Use the Reasons for Decision histogram to highlight interesting reasons for decision in order to ask further questions of your respondents or data collection

Clicking on a particular choice filters the bottom visualization to show the particular reasons for the decision. This dynamic filter makes the reasons themselves significantly more visible to our user and is supported by our usability tests with both Jeff and Daniel (Heer & Schneiderman, 13).



Crowdsourced Mapping of Visualization Principles

"Crowdsourced" Map
Determined by collected comparisons

Comparison Responses
Collected from surveyed respondents

Reasons for Decision

How to Use

This visual dashboard will allow you to explore respondent's resonsing for choosing a particular method of visualization over another choice:

1. Look at the "Crowd-sourced" Map to see your audience's chosen display methods of map markers

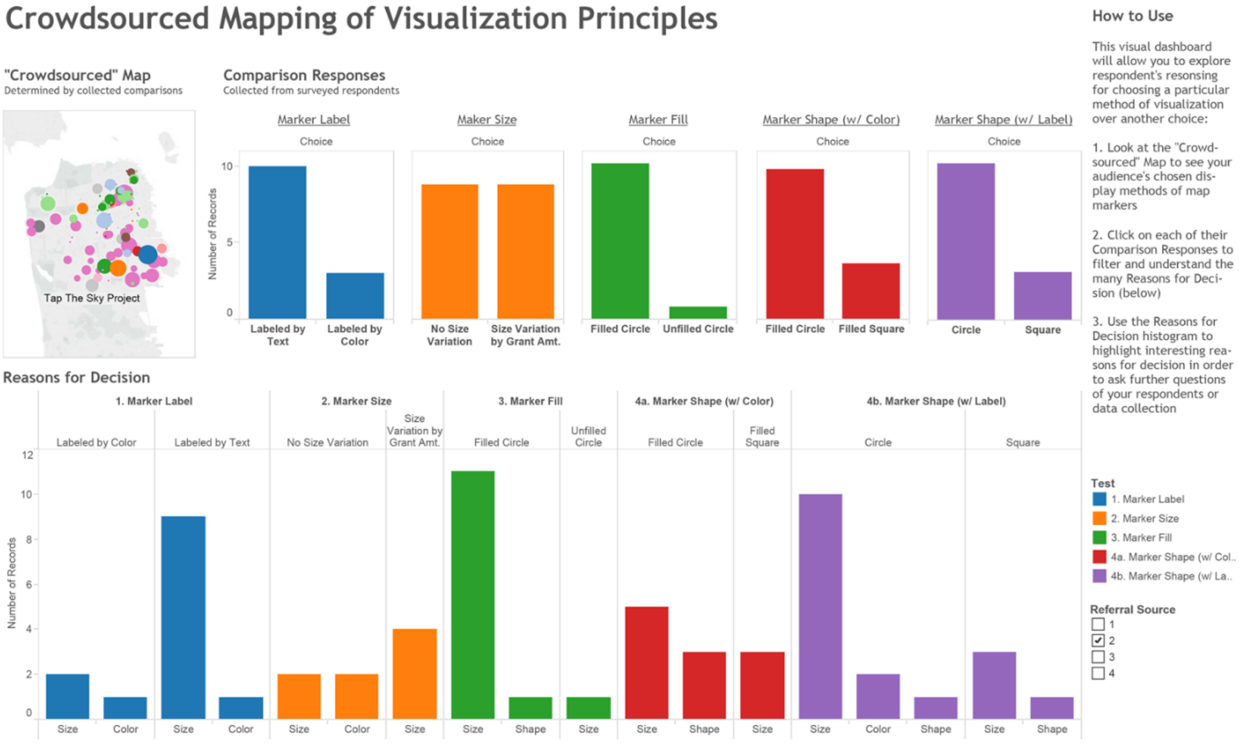
2. Click on each of their Comparison Responses to filter and understand the many Reasons for Decision (below)

3. Use the Reasons for Decision histogram to highlight interesting reasons for decision in order to ask further questions of your respondents or data collection

Test
🟦 1. Marker Label

Referral Source
☑ 1
☑ 2
☑ 3
☑ 4

We were also able to implement a filter based on the referral link we sent out. Each link has a particular source associated with it but for the sake of abstraction and protecting our respondents, we decided to display the numbers 1-4 to demonstrate the proof-of-concept.



Crowdsourced Mapping of Visualization Principles

"Crowdsourced" Map
Determined by collected comparisons

Comparison Responses
Collected from surveyed respondents

Reasons for Decision

How to Use

This visual dashboard will allow you to explore respondent's resonsing for choosing a particular method of visualization over another choice:

1. Look at the "Crowd-sourced" Map to see your audience's chosen display methods of map markers

2. Click on each of their Comparison Responses to filter and understand the many Reasons for Decision (below)

3. Use the Reasons for Decision histogram to highlight interesting reasons for decision in order to ask further questions of your respondents or data collection

Test
🟦 1. Marker Label
🟧 2. Marker Size
🟩 3. Marker Fill
🟥 4a. Marker Shape (w/ Col..
🟪 4b. Marker Shape (w/ La..

Referral Source
☐ 1
☑ 2
☐ 3
☐ 4

Future Goals

We were able to complete this study within the 10-week constraint of the quarter. Given more time, we would have liked to implement further automation of both survey set generation as well as the final visualization. As collection times increase and larger datasets are used with our dashboard, we would also like to add the ability to compare multiple survey collection results in one comprehensive dashboard. In response to our usability testing, we would have also liked to implement the ability for statistical analysis, either in a separate view or as an additional detail in the current dashboard. In response to Jeff's suggestion for a comparison against current visualization standards, the first view of our story could be used as a way for researchers to make side-by-side comparisons with a similar map which represents the current best-practices. This information, along with a computer-generated list of what differs from or validates a particular best-practice, would provide a way for researchers to find value in this method of data collection outside of their current ability to use this tool as a catalyst for further scrutiny.

Finally, we would like to spend more time developing a comprehensive survey that can be successfully used with Mechanical Turk (or other similar products) as a collection source. The addition of a free response question combined with an automated or semi-automated method of qualitative coding would allow for researchers to better understand the reasons behind the collected choices of their survey respondents.

References

Borkin, Michelle A., Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. "What Makes a Visualization Memorable?" *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013): 2306-315. Print.

Callison-Burch, Chris, and Mark Dredze. "Creating speech and language data with Amazon's Mechanical Turk." *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010.

Downs, Julie S., Mandy B. Holbrook, Steve Sheng, and Lorrie F. Cranor. "Are Your Participants Gaming the System? Screening Mechanical Turk Workers." *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010): 2399-402. Web. 14 Feb. 2015. <http://lorrie.cranor.org/pubs/note1552-downs.pdf>.

Demiralp, Cagatay, Michael S. Bernstein, Jeffrey Heer, "Learning Perceptual Kernels for Visualization Design," *Visualization and Computer Graphics, IEEE Transactions on* , vol.20, no.12, pp.1933,1942, 31 Dec. 2014

Heer, Jeffrey, and Michael Bostock. "Crowdsourcing graphical perception: using mechanical turk to assess visualization design." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010.

Heer, Jeffery. Scheiderman, Ben. “Interactive Dynamics for Visual Analysis.” *Stanford University. University of Maryland, College Park*. (2012). Print.

Ipeirotis, Panagiotis G. "Demographics of mechanical turk." (2010).

Marcus, Aaron. "Cross-cultural user-experience design." *Diagrammatic Representation and Inference*. Springer Berlin Heidelberg, 2006. 16-24.

"Post A Survey On Mechanical Turk And Watch The Results Roll In." *NPR*. NPR. Web. 08 Mar. 2015. <http://www.npr.org/blogs/alltechconsidered/2014/03/05/279669610/post-a-survey-on-mechanical-turk-and-watch-the-results-roll-in>.

Rand, David G. "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments." *Journal of theoretical biology* 299 (2012): 172-179.

Scott, Taylor. "Encodings - Graphic Excellence and Integrity" HCDE 411 - Information Visualization. HCDE Design Lab, Seattle, WA. January 13, 2015. Lecture.