

# CS 4804 Data Visualization Final - Process Book

*Written by Matthew McAlarney, Randy Huang, Joe Dobbelaar, and Priyanka Narisimhan*

## Overview and Motivation

Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

For our final project, we decided to expand our a3 experiment and conduct one that is more comprehensive and follows a new difficulty progression. Thus, we essentially began our final project by building off of our a3 assignment code. In our final project we decided to conduct a longer and more comprehensive experiment using icon arrays and control text to represent data. Our motivation for conducting a longer and more comprehensive experiment stems from our curiosity to see if icon arrays are more or less effective at communicating proportional data than textual descriptions of those proportions given a range of sample sizes. In addition, icon arrays are an important data visualization used in the medical community to communicate the risk of developing certain conditions; as a group, we have a shared interest in determining if icon arrays are a better tool for this task than textual descriptions of data.

Our project has the following goals broken into a few categories:

Intermediate Goals:

1. Determine whether users can accurately interpret small proportions in an icon array.
2. Determine whether users can accurately interpret and apply proportions in an icon array to different sample sizes.
3. Determine whether users can accurately interpret and apply proportions from a textual description to different sample sizes.

Overarching Goals:

1. Determine whether icons arrays are less effective at communicating proportional medical data than textual descriptions of data.

## Related Work

Anything that inspired you, such as a paper, a website, visualizations we discussed in class, etc.

There were a few research papers that our group referenced that closely influenced the outline and design of our experiment:

1. How to evaluate data visualizations across different levels of understanding:
  - a. <https://arxiv.org/pdf/2009.01747.pdf>
  - b. This paper emphasizes a framework of visualization tasks that informs the types of questions that we ask users in our Icon Array 6-Question Quiz. Professor Harrison recommended that we reference this research when formulating our questions, and we ultimately decided to follow this

framework of visualization tasks as it allowed us to pinpoint what exactly we wanted to learn from each question.

2. Improving Bayesian Reasoning:

- a. <https://www.cs.tufts.edu/~remco/publications/2015/InfoVis2015-Bayes.pdf>
- b. This is another paper that Professor Harrison recommended that we use to influence the design of our experiment. Improving Bayesian Reasoning is an academic experiment that also evaluates the effectiveness of icon arrays in communicating certain types of data against textual description of data. After discussing this particular research with Professor Harrison, we confirmed that we needed to use control text in our experiment; when comparing against icon arrays, we reasoned that text would provide a relatively simple means of communicating data that most users would interpret correctly in baseline questions. We initially thought that communicating proportional data through icon arrays on the other hand would not be relatively simple depending on the sample size(s) and size of the proportion.

## Questions

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

The questions our group is trying to answer closely overlap with our goals listed above, and as a result, these questions did not evolve drastically over the course of the project. Also keep in mind that since our final project is a direct extension of our a3 assignment, our hypothesis remained the same, and we improved our experiment to be more robust in addressing our hypothesis. When analyzing our resulting user data, our group certainly considered new questions about the effectiveness of communicating proportional data through icon arrays compared to control text.

The major questions that our group is trying to answer are the following; Are icons arrays less effective at communicating proportional data to the user than textual descriptions of proportional data? Is the user able to predict proportions for future sample sizes through icon arrays less effectively than they can through textual descriptions of proportional data?

## Data

Source, scraping method, cleanup, etc.

In terms of the data driving our experiment, our group decided not to obtain data from any official source for our final project. When discussing an improved icon array experiment with Professor Harrison, there was an emphasis on implementing controlled data randomization tailored to each question in the experiment. As a group, we collectively agreed that data randomization tailored to the needs of each question was a key direction we needed to follow; we knew that we needed to make our icon array experiment follow a tight-knit difficulty

progression and ensure that a certain number of users would answer the easier baseline questions correctly and the more challenging questions incorrectly. As a result of planning and implementing data randomization, we also determined that it was best to not impose the randomization on official data originating from a credible source in the medical community. Early on in our project, we discussed with Professor Harrison about the drawback of presenting misleading medical data to users, and thus, the content of our data is not based on real medical research. Instead, we followed an experiment practice conducted in the Bayesian Reasoning research where the diseases that icon arrays represent are fictional. The data in our experiment represents the proportion of people out of a given sample size who will get a certain fictional disease at some point in their life. We chose the following made-up disease names for each of the size questions in our experiment:

- 1) Lake Disease - Question 1
- 2) Swamp Disease - Question 2
- 3) Cavern Disease - Question 3
- 4) Jungle Disease - Question 4
- 5) Mountain Disease - Question 5
- 6) Ocean Disease - Question 6

## Exploratory Data Analysis

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

To initially look at functional medical data and how it could be represented, we put together an icon array randomization demo where we sketched different icon arrays and prototyped questions:

Icon Array Randomization Demo:

Our group created a [demo](#) to illustrate how we planned to display randomized icon arrays, and ran it past Professor Harrison for general feedback on the direction of our icon array designs and questions. The questions we proposed to ask users in our final experiment are structured similarly to the following two questions:

- 1) How likely is it to contract disease X?
- 2) So in a population size of 40000, how many people are likely to contract disease X?

We gained valuable insights about our icon arrays in the above demo from Professor Harrison's feedback on both the visualizations and corresponding questions. Below, we have summarized the feedback that he gave us on the Icon Array Randomization Demo:

Icon Array Randomization Demo Feedback Notes:

- Need to have about two questions that form a positive baseline in the experiment (a baseline question is one that users are expected to answer correctly the majority of trials)
  - Iteration 4 Version 1 in the Icon Array Randomization Demo is a good example of this; one icon out of ten total icons is highlighted

in red. We ask the user to identify this proportion for the sample size of ten, and then ask them to apply this proportion to a marginally larger sample size (1:10 -> 2:20 for example).

- We should consistently use rows of 10 in our icon arrays; **using simple, round numbers is important.**
- We should stick to asking users questions about relatively small sample sizes (between 10 and 1000 should be a reasonable range with the simpler questions spanning sample sizes of 10 - 20 total icons).
- As for an overall layout for our experiment, it would be a good idea to have multiple trials that are relatively simple (with measurable differences between the relatively simple trials of course), and 1-2 trials that are objectively more challenging.

We should use data randomization with lower and upper bounds for both the number of colored icons and number of dimensions in an icon array tailored to each question in the experiment. This means that we will have specific lower and upper bounds for the randomization tailored to each question.

We will make up disease names and reference the Bayesian Reasoning paper that also makes up disease names (also approved by Professor Harrison). The made-up disease names will allow us to avoid the scenario of presenting false or misleading data.

## Design Evolution

What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

As mentioned before, since our final project is a direct extension of our a3 assignment, we knew that the only types of visualizations that we would consider including in our application are icon arrays and control text (textual descriptions of proportional data). As a result, our experiment did not deviate from the core ideas communicated in our proposal. To understand the design decisions that we made with the icon arrays, it is important to first understand the encompassing design of our improved experiment. Our group used Professor Harrison's feedback on our Icon Array Randomization Demo to design our improved experiment and the icon arrays that are currently used in our application. Below is a detailed outline of our experiment design:

### **Final Project Experiment Design (6 questions, 1 trial per question):**

- **Note: We think that having exactly 6 questions (double the number of questions we included in a3) is a reasonable question number for our final project experiment as it enables us to expand our a3 experiment while not requiring the user to bear through a particularly long survey.**
- Hypothesis: Icon arrays are not as effective at communicating proportional data to the average user as textual descriptions of proportional data. In other words, the average user cannot deduce proportions of a sample size and predict

proportions for future sample sizes through icon arrays as effectively as they can through textual descriptions of data.

- Resources and research used to influence our visualizations and questions:
  - How to evaluate data visualizations across different levels of understanding: <https://arxiv.org/pdf/2009.01747.pdf>
  - Improving Bayesian Reasoning: <https://www.cs.tufts.edu/~remco/publications/2015/InfoVis2015-Bayes.pdf>
    - For our Final project, we have chosen to create our own fictional names for diseases presented in our problem interfaces. The use of these fictional names ensures that we avoid the potential situation of presenting misleading data that is connected to real research from the medical community.
  - Levels and corresponding tasks of the visualization taxonomy we are testing:
    - 1. Application: Use a percentage and total population to calculate a number.
    - 2. Synthesis: Predict a future value.
- Important Notes:
  - Each icon array has exactly 10 icons per row to ensure we are using round numbers.
  - Icons colored in red represent the number of people who are going to get a certain disease. Icons colored in gray represent the number of people who are not going to get the disease.
- Problem Difficulty Progression:
  - Difficulty Level 1 - Baseline (Problems 1 and 2):
    - We want to know: Can the user accurately interpret and apply small proportions pertaining to the disease-positive individuals in small icon arrays?
  - Difficulty Level 2 - Intermediate (Problems 3 and 4):
    - We want to know: Can the user accurately interpret and apply larger proportions pertaining to the disease-positive individuals in larger icon arrays?
  - Difficulty Level 3 - Advanced (Problems 5 and 6):
    - We want to know: Can the user distinguish between the disease-positive and disease-negative individuals in an icon array? Can the user accurately interpret and apply proportions pertaining to the disease-negative individuals in larger icon arrays?
- Problems (1 visualization shown per problem):
  - 1 (Taxonomy Mapping: Application, Use a percentage and total population to calculate a number):
    - Visualization Construction:
      - (1 or 2) X 10 Icon Array, (1-2) person icon(s) highlighted in red.

- Data Randomization Bounds:
    - Number of red colored icons: **1 or 2**
    - Number of dimensions (rows of 10): **1 or 2**
  - Question: About how many out of the **(total number of icons)** people in the sample will get Lake Disease sometime during their lives?
    - What we want to know from this question: Can the user accurately interpret a small proportion shown in an icon array with few dimensions?
- 2 (Taxonomy Mapping: Synthesis, Predict a future value):
  - Control Text: For every **(10)** people, **(1)** person will get Swamp Disease at some point during their lives.
    - Data Randomization Bounds:
      - Number of red colored icons: **1**
      - Number of dimensions (rows of 10): **1**
    - Question: About how many out of a sample of 20 people will get Swamp Disease at some point during their lives?
      - What we want to know from this question: Can the user accurately apply a proportion from a textual description to a different sample size that is also relatively small?
    - Answer: 2
  - 3 (Taxonomy Mapping: Application, Use a percentage and total population to calculate a number):
    - Visualization Construction:
      - **(3-5) X 10** Icon Array, **(10, 15, 20, or 25)** person icon(s) highlighted in red.
      - Data Randomization Bounds:
        - Number of red colored icons: **10, 15, 20, or 25**
        - Number of dimensions (rows of 10): **3-5**
    - Question: About how many out of the **(total number of icons)** people in the sample will get Cavern Disease sometime during their lives?
      - What we want to know from this question: Can the user accurately interpret a larger proportion shown in an icon array with a few more dimensions?
  - 4 (Taxonomy Mapping: Synthesis, Predict a future value):
    - Control Text: For every **(30)** people, **(10)** individuals will get Jungle Disease at some point during their lives.
      - Data Randomization Bounds:
        - Number of red colored icons: **10**
        - Number of dimensions (rows of 10): **3**
    - Question: About how many out of a sample of 60 people will get Jungle Disease at some point during their lives?

- What we want to know from this question: Can the user accurately apply a larger proportion from a textual description to a different sample size that is also larger?
  - Answer: 20
- 5 (Taxonomy Mapping: Application, Use a percentage and total population to calculate a number):
  - Visualization Construction:
    - **(3-5)** X 10 Icon Array, **(10, 15, 20, or 25)** person icon(s) highlighted in red.
    - Data Randomization Bounds:
      - Number of red colored icons: **10, 15, 20, or 25**
      - Number of dimensions (rows of 10): **3-5**
  - Question: About how many out of the **(total number of icons)** people in the sample will **never** get Mountain Disease sometime during their lives?
    - What we want to know from this question: Can the user accurately interpret a proportion shown in a larger icon array corresponding to the icons that are colored in gray (those individuals who will not get the disease)?
- 6 (Taxonomy Mapping: Synthesis, Predict a future value):
  - Control Text: For every **(30)** people, **(5)** individuals will get Ocean Disease at some point during their lives.
    - Data Randomization Bounds:
      - Number of red colored icons: **5**
      - Number of dimensions (rows of 10): **3**
  - Question: About how many out of a sample of 60 people will **never** get Ocean Disease at some point during their lives?
    - What we want to know from this question: Can the user accurately apply a proportion corresponding to the gray icons (those individuals who will not get the disease) in a textual description of a larger icon array to a different sample size that is also larger?
  - Answer: 10

Clarifications about our experiment and icon array designs:

1. We decided to include three pieces of control text and three icon arrays in our experiment. To improve on our a3 experiment and gain a more accurate insight as to the effectiveness of icon arrays compared to control text, we wanted to conduct an even comparison between control text and icon arrays. Each level of difficulty in our experiment design includes both an icon array and control text that are geared to answer the sub-question listed that we want to answer. Essentially, we knew that we needed to structure this experiment so that an icon array is pinned against control text for evaluating effectiveness towards the same visualization task.

2. Our difficulty progression is designed so that two questions have baseline difficulty, two have intermediate difficulty, and two have advanced difficulty. We knew that the two baseline questions had to evaluate small icon array proportion interpretation and application for getting a certain disease with small sample sizes. This way, we could be fairly certain that most users would get the first two baseline questions correct. The two intermediate questions also evaluated icon array proportion interpretation and application for getting a certain disease, but this time with larger proportions and sample sizes. Our assumption here is that it is more difficult for users to interpret and apply larger proportions for larger sample sizes than it is with smaller proportions and sample sizes. The two advanced questions evaluated icon array proportion interpretation and application for *never* getting a certain disease and used larger numbers within the same randomization bounds as the two previous intermediate questions. For the advanced question that employed an icon array, we ultimately wanted to see if the user could distinguish between the meaning of the red and black icons; the red icons denoting the number of people who will contract a certain disease at some point in their lives and the black icons denoting those who will never contract a certain disease at some point in their lives. Similarly with the advanced question that uses control text, we wanted to see if the user could deduce the proportion of people who will never get a certain disease from textual description.
3. Each question has its own set of data randomization bounds. For the two baseline questions, we chose to have the generated icon array keep between 1-2 dimensions of 10 icons and between 1-2 icons highlighted in red. For the two intermediate questions and two advanced questions, we chose to have the generated icon array keep between 3-5 dimensions of 10 icons and either 10, 15, 20, or 25 icons highlighted in red. We wanted to make sure that the generated proportions were either small or substantial, but also did not take up the vast majority of the sample sizes. Each piece of control text uses a set of constant numbers (shown in the randomization bound notation) to ensure that the text displays the same data across all users, and thus, functions as a traditional experiment control.

## Implementation

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

The implementation of our final project application directly stems from our a3 assignment application. As a result, our final project application uses a similar folder structure, the same React component structure, and a very similar experiment UI. We began changing our a3 assignment code through removing the static images corresponding to our previously employed icon arrays and instead implementing data randomization functionality for our newly designed icon arrays (as described in the experiment design outline above):




### Icon Array Randomization Functionality:

The code we wrote for the icon array data randomization originally utilized D3 to randomize the total number of icons in an icon array and also randomize the number of icons colored red in an array. The logic behind our icon array randomization follows the data randomization bounds that we set in our experiment design outline for each question. We set the default number of icons in a row to 10 for all questions as we knew that icon arrays that use even, rounded row numbers are easiest for users to read and removes unnecessary barriers of entry to the visualization tasks. We also made it so that the total number of people in an icon array is always a multiple of 10; we needed to employ even, rounded numbers for the total sample sizes that would be understandable to users when defining the dimensions. We also had created a lower bound on the total number of people in an icon array to make sure that the resulting image would always have a minimum of 10 total icons. In terms of the number of red icons selected in each of our icon arrays, we also made sure that we were employing even, rounded numbers that incremented by 1s for the baseline questions and by 5s for the intermediate questions.

In addition to the icon array randomization functionality that we implemented and fine-tuned throughout the course of developing our final application, we also improved the way that we implemented text in our experiment. Since our experiment design now calls for evaluating to what extent users can also effectively deduce a proportion from a sample size based on the number of black icons present, we knew that we needed to include additional informational text that gives the user context about distinguishing between the red and black icons. This need for informational text led us to create a new React component for our interface; a basic paragraph that displays the message **“The red people icons represent those who will get the disease and the black people icons represent those who will not get the disease”** for every question that has an icon array:

### Lake Disease Icon Array



The red people icons represent those who will get the disease and the black people icons represent those who will not get the disease. About how many out of the total 10 people in the sample will get Lake Disease sometime during their lives?

Submit

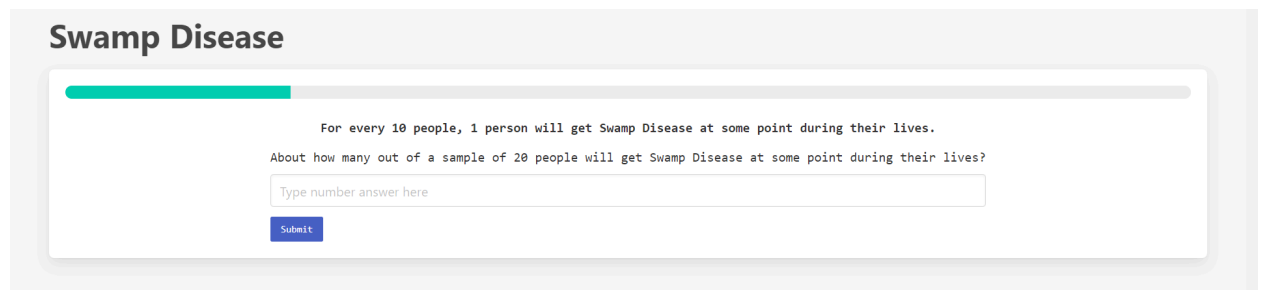
Icon Array 6-Question Quiz - Joe Dobbelaar, Priyanka Narasimhan, Randy Huang, Matthew McAlarney

Signed in as: Matthew McAlarney

Sign Out

In terms of the design of the icon arrays themselves in our experiment, the most important aspect here was making sure that every icon was visually identifiable in regards to the semantic context of our disease scenarios. To accomplish this, we knew that the icons representing those who would get a certain disease at some point in their life needed to have a distinct color, and the icons representing those who would not get a certain disease at some point in their life also needed to have a distinct color. The contrast between red (for the disease-positive icons) and black (for the disease-negative icons) colors for the icons allowed us to ensure that the user could also clearly distinguish between people who will get a certain disease and people who will not get a certain disease. Additionally, we were careful about choosing the shape of the icons themselves to hold in our application; we needed to make sure that the icons clearly represented people from a user's perspective. Although this may seem like a minor design choice, it is still quite important as the combination of our icons resembling people and having distinct colors is what allowed us to keep our overarching experiment design simple and understandable for all users.

In addition to the simple and clear design of our icon arrays, we also put some thought into the design of our control text. For each question in our experiment that corresponds to a piece of control text, we noticed that the resulting interface always held two lines of text with the same font and font size (control text + question). In an effort to distinguish the control text from the question text, we styled each piece of control text to have bold font weight and have a small bottom margin from the question:



**Swamp Disease**

For every 10 people, 1 person will get Swamp Disease at some point during their lives.

About how many out of a sample of 20 people will get Swamp Disease at some point during their lives?

Submit

## Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

We had 10 participants take our Icon Array 6-Question Quiz, and each participant completed exactly one trial of each question. After compiling the answer results of all 10 participants, we computed the average log2error per question using the individual log2errors calculated in our database file. Below is the raw ranking from the best performing visualizations to the worst performing visualizations in our experiment using the average log2error as the

metric of comparison, along with the means (average log2errors), lower bounds, and upper bounds:

```
{
  "visualizationNames": [
    "Visualization 2",
    "Visualization 4",
    "Visualization 3",
    "Visualization 5",
    "Visualization 6",
    "Visualization 1"
  ],
  "means": [
    0.21432290951073488,
    0.3646570232202958,
    0.7102070546024226,
    1.228342118067947,
    2.6069583231790885,
    4.230315508522389
  ],
  "lowerBounds": [
    0.21432290951073488,
    0.3646570232202958,
    0.7102070546024226,
    1.1819989358564074,
    1.4526625873969923,
    2.750778857237328
  ],
  "upperBounds": [
    0.4286458190214698,
    0.7293140464405917,
    1.0827554257806522,
    1.2503532503616492,
    1.4526625873969916,
    4.257971845894892
  ]
}
```

To synthesize the data taken from our output.json file above and make it more understandable, below is a polished ranking of our visualizations from best to worst performance with the corresponding question number, name and visualization type:

Visualization 2 - Question 2 Swamp Disease - Control Text  
Visualization 4 - Question 4 Jungle Disease - Control Text  
Visualization 3 - Question 3 Cavern Disease Icon Array - Icon Array  
Visualization 5 - Question 5 Mountain Disease Icon Array - Icon Array  
Visualization 6 - Question 6 - Ocean Disease - Control Text  
Visualization 1 - Question 1 - Lake Disease Icon Array - Icon Array

In an effort to understand which visualization type performed better overall in our experiment, it is crucial to compare pairs of visualizations within their respective difficulty groupings as previously shown in our experiment design outline. As a reminder, here are problem difficulty groupings used in our experiment:

Problem Difficulty Progression:

- Difficulty Level 1 - Baseline (Problems 1 and 2):
  - We want to know: Can the user accurately interpret and apply small proportions pertaining to the disease-positive individuals in small icon arrays?
- Difficulty Level 2 - Intermediate (Problems 3 and 4):
  - We want to know: Can the user accurately interpret and apply larger proportions pertaining to the disease-positive individuals in larger icon arrays?
- Difficulty Level 3 - Advanced (Problems 5 and 6):
  - We want to know: Can the user distinguish between the disease-positive and disease-negative individuals in an icon array? Can the user accurately interpret and apply proportions pertaining to the disease-negative individuals in larger icon arrays?

Here is the comparison of how visualizations performed within their difficulty pairs with each pair containing a ranking from best to worst performance:

Difficulty Level 1 - Baseline

- Best Performance: Visualization 2 - Question 2 Swamp Disease - Control Text
- Worst Performance: Visualization 1 - Question 1 - Lake Disease Icon Array - Icon Array

Difficulty Level 2 - Intermediate

- Best Performance: Visualization 4 - Question 4 Jungle Disease - Control Text
- Worst Performance: Visualization 3 - Question 3 Cavern Disease Icon Array - Icon Array

Difficulty Level 3 - Advanced

- Best Performance: Visualization 5 - Question 5 Mountain Disease Icon Array - Icon Array
- Worst Performance: Visualization 6 - Question 6 - Ocean Disease - Control Text

To analyze the rankings of visualization performances when placed into their respective difficulty groupings above, control text performed better than icon arrays in both the baseline and intermediate difficulty groupings. Interestingly though, icons arrays actually performed better than control text in the advanced difficulty grouping. Overall, these comparisons suggest that textual descriptions are more effective at communicating proportional data to the average user than icon arrays. Furthermore, our findings generally support our original hypothesis stating that icon arrays are less effective at communicating proportional data to the average user than textual data descriptions. However, there is a key question that our group is now wondering about our experiment; why did users perform better with icon arrays than with textual description in the advanced difficulty grouping but not in the baseline and intermediate groupings? One possibility is that there is some degree of on-the-quiz learning that may have happened with most users that participated in our experiment; for instance, if a user was not completely sure about how to answer question 3 given the notably larger sample size and proportion of red icons presented, then it could be that the user was able to readjust the accuracy of their interpretation when they encountered question 5. This question about our design of the questions in the advanced difficulty grouping leads our group to consider ways that this experiment could be improved even further. A few steps that we could take to improve our experiment further would encompass surveying a greater pool of participants from a wider variety of backgrounds, and reconsidering the ordering of questions within difficulty groupings (make it so that control text does not necessarily always appear after icon arrays). We think that these two steps would provide the foundation for our group to collect more meaningful results, and allow us to more accurately pinpoint weaknesses in our experiment design for future iteration.