CSC 595 – Final Project Report
November 15, 2017
Caroline Cao, Brandon Markwalder, David Scroggins

1. Overview of Dataset

The National Longitudinal Study of Adolescent Health tracked American adolescents grades 7-12 from 1994 to 2008. The dataset was developed from the Quality Education Database, which contains 26,666 U.S. High Schools. The complete dataset can be accessed here: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600 along with relevant user guides for study design, variable description etc.

A stratified sample of 80 schools with "probability of selection proportional to school size" were selected. Schools were stratified by region, urbanicity, school type, ethnic mix and size. One feeder school for each high school was recruited producing one school pair in 80 different communities. School size varied from 100 to 3,000+ students. Most the dataset was produced via a sequence of surveys. First, an In-School Survey in 1994 completed by over 90,000 students. The Wave 1 In-Home Survey (1995), in which adolescents were selected with unequal probability of selection from those who complete in-school survey was drawn from the In-School survey via stratifications by grade and sex and then random selection of ~17 students from each stratum to yield a sample of 200 adolescents from each school pair. The In-Home survey was repeated three additional times: Wave 2 In-Home Survey (1996); Wave 3 In-Home Survey (2001) and Wave 4 In-Home Survey (2008) (sampling criteria for each wave is detailed in the study's downloadable User Guide). In addition to survey results, Wave 4 includes several biomarker datasets including inflammation and immune function, measures of glucose homeostasis as well as lipids.

The composition of the dataset varies by Wave, however, the results from the "Wave 1: In-Home Questionnaire" give an adequate sense of scale. The dataset contains 2,794 columns/variables and 6,281 rows/instances for a total of 17,549,114 unique data points. A thorough summary of the variables is beyond the scope of this proposal. The variables range from basic demographic questions: date of birth, racial and ethnic background, intimate partner status, school status, daily activities, general health, access to health services, parental relationships, pregnancy history, relations with siblings etc. They provide a comprehensive self-reported description of the student's current living situation and personality. Complete details can be obtained from the user guides available at the dataset's main site.

2. Challenges and Approach

There are three primary challenges faced when working with this dataset:

- The size of the dataset(s) provides a rich field for exploration and analysis, but can also be overwhelming, especially in the context of initial exploration. Each wave contains over 17 million unique data points and this does not include the additional contextual datasets provided for each wave. Merely determining *what* one wants to focus on can

be a week-long project. Our project is, in part, to provide an interface to mitigate this challenge.

- The size of the dataset is complicated by its abstract encoding. Features are (understandably) encoded with compact string codes, such as "H1DA1." H1 indicates that this question belongs to the in-home survey for wave 1, DA1 indicates that this is the first question in the Daily Activities section for the relevant survey. The actual question for H1DA1 is "During the past week, how many times did you do work around the house, such as cleaning, cooking, laundry, yardwork, or caring for a pet?" Furthermore, responses to questions are encoded abstractly using integers (e.g. a response of 1 to H1DA1 indicates a real-world answer of "1 or 2 times"). No direct key/value mapping for codes and questions is available, and details for each feature must be obtained from the code books for each wave, which typically exceed 800 pages in length. A searchable database is provided on the website (Variables) which will display the question string (includes only features for In-Home Questionnaires). The interface is valuable for quick look ups of variables, however does not transition well to use in a visualization as it requires additional navigation to access to the response codes for the specific question.
- The distribution of the question responses can vary significantly. Consider H1DA1's distribution (not at all: 255; 1 or 2 times: 1759; 3 or 4 times: 2016; refused: 2; don't know: 6) compared to H1TO4 ("How old were you when you first started smoking cigarettes regularly (at least 1 cigarette every day for 30 days)?") where 5219 of respondents were encoded as "97", i.e. "legitimate skip." This heavy imbalance toward skips is especially prevalent in the more specific regions of the dataset that target non-typical behaviors for adolescents and can render classification tasks tricky. While an encoding for "skip" might be undesirable for classification, it still represents a meaningful response in that 1) it is a positive response to the question and not merely a missing data point and 2) often directly affects where a respondent reenters the questionnaire (a legitimate skip might result in a respondent skipping and entire subsection of the questionnaire). We wanted to design an interface that could capture not just a high volume of skips, but also where the individuals who skipped a question reentered the dataset.

Our intent was to build a prototype for visual exploration of the dataset to expedite the process of understanding its structure as well as identification of its unique encoding schemes. Consequently, we left the original shape of the data largely intact. We wanted to design a visual explorer that adheres to the Shneiderman mantra seriously, composed of an interface that eases the user into the dataset gradually enabling an understanding of the high-level structure as well as the interconnectedness of the question responses and laying the groundwork for a transition to classification tasks. The user should not only be able to visually appreciate the scale and structure of the dataset, but also be able to see how individual respondents progress from question to question. Additionally, since Wave 4 provides a rich set of additional contexts, we wanted to provide an initial interface for considering how certain responses map to the biomarker data contained in the Wave 4 files. Since our focus was on building a relatively quick

prototype, we focused on the health themes within all four waves, rather than attempting to build out a visualization for the entire dataset.

3. Initial Design

The general type of visualization required for an initial overview of the dataset is directly recommended by the structure of the In-Home Questionnaires. The dataset is essentially laid out as a tree: with each wave representing a node, question sections representing each wave's children, and specific responses (and associated counts) representing leafs. A quick consensus emerged that hierarchical visualizations would be the best approach to overview the dataset, and three primary styles were considered:

- Collapsible Indented Tree (of the style represented by this implementation https://bl.ocks.org/mbostock/1093025): This representation was a strong contender for overviewing each wave. It provides an intuitive interface for navigation and the size of the data can be managed via collapsing submenus that are not relevant to the user's current focus. We still think that this might prove useful if we were to expand the visualization to include the entire dataset, however, we eventually opted for another visualization as the focus of this visualization is less on a total overview of structure and more on the specific content of nodes. As mentioned our data codes are abstract and in the absence of provided key/value mappings, we felt that emphasizing the encodings rather than the structure of the data would limit the usefulness of this approach as an interface. We still believe this might serve when integrated with the broader system (perhaps one could generate an on demand indented tree for a specific node).
- Treemap: Treemaps provide a cellular overview of the data's structure, however, are spatially limited given a large number of nodes. For instance, consider the Tobacco, Alcohol, Drugs section in Wave 2, which contains over 80 questions (each a child node) and each of those questions 4+ responses. To avoid either an excessively large tree map or an excessively crowded tree map, we realized we would need to generate an initial overview Treemap per wave that featured just the sections that would re-render a Tree Map for a node when clicked on. While this certainly represents a zoom feature, we felt it failed as an overview feature, since one would struggle to get a clear overview of the *entire* dataset.
- Sunburst: Sunburst visualizations recommended themselves as a viable approach for two reasons: 1) They allow a *complete* overview of the entire structure of the dataset (all nodes are easily rendered visible in a single top level visualization) and 2) with the addition of a zoom feature, they allow for a user to drill into the dataset immediately to view each level of the hierarchy in more detail. With a zoom feature the addition of simple tooltips on hover to reveal a node's specific identity, Sunbursts allowed for a complete implementation of the Schneiderman mantra. Consequently, they were chosen as our primary interface to the dataset.

For a more detailed view of each wave's data, a consensus for two additional coordinated visualizations was more quickly reached. We wanted a visualization that would allow for

comparison of in-wave trends as individual respondents moved from question to question as well as a comparison of cross-wave trends (i.e. if a respondent is active in Wave 1 is that active lifestyle likely to persist across Waves 2, 3, and 4). The multivariate, high-dimensional, time oriented nature of our data reads as an advertisement for the benefits of parallel coordinates. We felt such a visualization would allow for easy comparison of how responses to specific questions progress as well as rendering meaningful individual decisions to skip certain questions, since skips could still be analyzed within the context of the entire questionnaire and not just in terms of a single question's distribution. The inclusion of D3 brushing would add filtering features allowing for more targeted analysis of trends based on specific responses to questions.

Given our end focus was on a visualization structure that pointed toward classification questions we gravitated toward a scatter plot matrix using the Wave 4 biomarkers as labels. This visualization would be coordinated with the parallel coordinates plot to allow for a targeted analysis of response clusters based on biomarkers (since we were focusing on a health theme, the use of biomarkers as labels seemed a natural progression, though we believe this generalizes well to the rest of the dataset). The addition of brushing would allow additional focus on certain regions of the dataset and enable quick cross-comparison of feature clusters based on the selected questions. Some issues with how our dataset mapped to this visualization would emerge. This will be discussed in more detail below. We also decided to keep the design static for the sake of speed and easy hosting directly from the repository. This would eventually present some design limitations which will be discussed below.

Our final design decisions can be outlined as follows:[1]

- Top level Sunburst-based interface: this is mostly a dummy visualization intended to serve as a general interface, however it does encode both the count of the waves as well as the size of the waves. From this top-level interface, two primary exploratory paths are available
- Path 1: Wave-Specific
    - Clicking on a specific wave will render a Sunburst visualization displaying the entire structure of the Wave. Each click on the nodes will drill down into the visualization by one level. A click in the center will move up a level in the wave. Hovering over a specific node will produce a tooltip indicating that specific node's relevant encoding (either section id, question id or response id).
    - Clicking to analyze a wave will render a drop-down menu that allows the user to select a question section and a specific question for up to five columns that produces a coordinated parallel coordinates plot and a scatter plot matrix.[2]
- Path 2: Cross-Wave

[1] A diagram of this design can be viewed in Appendix D

[2] In the absence of an available key/value mapping from code to question content, we had to build our own key/value mapping using a simple JavaScript object. This limited our scope to selecting only certain questions. Since this is a prototype, we felt this decision was reasonable, however, in the future we need to either create or obtain a complete key/value mapping to allow for full exploration (one option is to contact the curators of the dataset to see if they can make this available).

- Clicking in the center of the interface will render a drop-down menu that allows the user to selection a wave, question section and a specific question for up to five columns that produces a coordinated parallel coordinates plot and a scatter plot matrix.

4. Challenges and Iteration of Design

The implementation of the Sunburst visualizations proceeded mostly as expected with the primary challenges being coaxing the data into an appropriate JSON format and adapting available online Sunburst code into a structure that allowed for D3 functions to render labels, tooltips, etc.

Implementation of the parallel coordinates and scatter plot matrix components was mostly straightforward, however, we encountered some limits of the visualizations with regard to the structure of our data that we probably should have anticipated in our initial design phase. What follows is a brief discussion of these problems as well as an overview of steps taken to address them.

A significant portion of the dataset represents categorical responses to questions rather than continuous ranges. This mixed feature set limited to effectiveness of our chosen visualizations which are continuous data-centric. Both the parallel coordinates plot and scatter plot matrix suffered from over-plotting (the effects were more pronounced with the scatter plots). Categorical features (keeping in mind there are ~6,000 responses to each question) had the effect of hiding individual responses plotted at a certain x,y coordinate in the scatter plots. While trends were still visible in the parallel coordinates plot, the frequency of responses was also occluded in cases where the features were discrete rather than continuous.

Our initial solution to the over-plotted scatter plots was to abandon the visualization entirely in favor of grouped bar charts. Each row would represent a selected question, each group a response, and each bar a specific class label for a biomarker. We still believe this is a good solution and one that should be perhaps included in the future as an option for how the data is displayed (giving the user the ability to select not just questions of interest but specific types of coordinated visualizations to return), however, it proved difficult for us to quickly implement given the shape of our .csv files. The best approach both in terms of speed and efficiency was to dynamically render counts from a database, however, we were limited by our decision to keep the design static (and, frankly, by the limited time scale before the presentation). At this point we had a fairly large functioning static implementation, and restructuring to a dynamic design seemed cost-prohibitive. We reluctantly decided to push this design change to the future.

Our second solution to the over-plotted scatter plots (and the one reflected in this submission) was to jitter the data to render labels more visible, alter the opacity of plotted coordinates to render stacked instances more visible as well as a small stroke border to make them visually more distinguishable. We also added a zoom feature to the scatter plots to allow the user to more easily distinguish grouping. The solution was mildly successful in that it does allow one to

more easily discern grouping by label, however, the sheer quantity of instances plotted still point toward an alteration of the visualization. As suggested by Professor Brown in class, a better option here might be a contingency table. We are eager to implement this; however, we were limited by design decisions made regarding static vs dynamic, and have again reluctantly made the decision to push this overhaul of the structure.[3]

The parallel coordinates plot was likewise limited by over-plotting. While trends were still observable (particularly when the data was continuous) it was not always possible to tell the quantity of individuals moving from response to response. Our initial plan to resolve this (and the one represented in the submission) was to use shading, with the darkness of a line indicating a higher number of instances. This was somewhat successful but was limited by the fact that 1) color is itself not often valuable for encoding numerical data and 2) it was not always easy to discern the density of the shading given the thinness of the parallel coordinate lines. The best long term solution is to transition to an adaptation of Sankey diagram, however, given the time scale, we were not able to make this transition in time for the submission.

A walkthrough of our prototype data explorer can be located in Appendix E and the hosted visualization can be viewed directly (along with source code) at "nlsaahvisproject.github.io".

5. Final Discussion and Proposals for the Future

Primary lessons learned by the group can be summarized as follows:

- Fine grained details between visualizations can make a significant difference in directing final choices. This is especially evident in selecting the proper hierarchical visualization for our dataset. Identifying that we needed a hierarchical visualization was only half of the process as we had three viable options for showing the hierarchy, with each possessing their own list of unique pros and cons. All three probably would have worked in this context (and perhaps in a more complex system, perhaps more than one of them should be used in a coordinate fashion), but the proper decision was selecting the one that best met our criteria for the interface: a *complete*, single view of the data with the ability to zoom and present details on demand.
- Early design decisions regarding the structure of the technology stack are crucial in determining flexibility for alteration later in the build. This is largely self-evident, and I believe we all already knew this, but in the compressed time frame of building and presenting a project for a graduate class, we learned the lesson in a more intense way. We do not regret the decision to keep our design static: it allowed us to quickly build out the prototype and focus on altering the visualizations. It did, however, limit our ability to alter our visualizations when we needed to ability to dynamically alter our data as well as quickly transition to new visualizations when they were needed.
- Careful attention to feature types should drive the choice of visualization. Parallel coordinates was the correct choice conceptually. The scatter plot matrix was a viable

[3] This should not be interpreted as a performative statement. We all fully intend to keep developing this repository as a quality location for applying our visualization and development skills.

choice that ultimately proved to not be the best choice. In the former case, paying more attention to the mixed types of data and not just the overall structure and distribution of the data would have allowed us to anticipate the necessary alterations to the visualization (i.e. encoding quantity to account for over-plotting). In the latter case, this would have perhaps pushed us away from the scatter plot matrix in the early design phase, altering how we approach our data (perhaps we would have decided to pursue a dynamic approach early) and almost certainly alleviating some stress and frustration in the late stages of the quarter.

Despite some limitations, however, we believe our visualization represents a quality step toward developing a visual explorer for the National Longitudinal Study of Adolescent to Adult Health, a rich dataset with a great deal of classification potential but also with an overwhelming scale and abstract encoding strategy. Our top-level Sunburst visualizations provide the user the ability to immediately ascertain the scale and structure of the individual waves, as well as explore the contents of the waves and, using the zooming feature even make preliminary decisions as to the distribution of responses to a given question. We believe this will scale well to the entire dataset, though it may require an additional layer of overview. Parallel coordinates, as hoped, allows a clear overview of how respondents move between individual questions (as well as between waves) and also allows clear visual identification of the ways in which 'skips' and 'refusal to respond' shape the structure of individual questions. The scatter plot matrix provides a serviceable entry point toward classification tasks by mapping question responses to class labels, and despite needing additional alteration (or more likely replacement) provides an example of how coordinating visualizations adds depth to exploratory analysis in a single view.

By way of conclusion, we would like to outline our plans for next steps:

- Alteration/replacement of afore mentioned scatter plot with a contingency matrix
- Replacement of parallel coordinates with Sankey diagram
- Transition current static build to Flask-based dynamic build
- Add "data cleaning" capacity to parallel coordinates view (allow temporary removal of outliers etc.)
- Create machine learning backend for classification tasks (for instance, one idea we did not implement was using KNN to display a table or graph view of the n most similar instances to a selected data point).
- Create or access key/value mappings to make overview features more 'human readable' as well as build out question options for coordinated views section

Appendix A: Caroline Cao Individual Report

Personal Role: At the first meeting, each of us contributed our thoughts and agreed three main graphs (sunburst, parallel coordinates, scatter plot matrix) would be good visualizations to our data. Based on our group discussion from brain storming, I created an initial diagram as shown in Appendix D. Additionally, we naturally divided our tasks based on three main different visualizations.

In the second step, I focused on the question list with the dropdown and parallel coordinates. Since David took care the project structure and created GitHub repository for us, I can be more concentrated on my specific tasks.

The question list creation is time-consuming, but it was a good experience for me to get a better understanding of the questions and the data related. The questions, based on different waves and sections, are documented in the pdf files. Instead of working on coding how to parse the text from pdf, I worked on putting the questions into an object that defines which wave and section each question belongs to. Meanwhile, since Brandon and I share the same dropdown list to explore the data with both parallel coordinates and scatter plot matrix, I worked on creating the dropdown list for both separate wave and cross waves.

For the visualization of parallel coordinates, I referred to the example from Jason Davies (https://bl.ocks.org/jasondavies/1341281). The main problem I got was the integration. Most of part of integration is very smooth, because Brandon and I discussed and agreed to use the array as parameter for questions code and his code is very clean and reusable. However, since the code of scatter plot was based on version 4 of D3 but the parallel coordinates was built on version 3, the upgrade part became a heavy task of integration. Even more, as the integration took place at the last phase of the project, the time was very limited. For most of the upgrade work, I refer to the official guideline (https://github.com/d3/d3/blob/master/CHANGES.md). However, the brush function of version 4 is changed significantly, but this was a good exercise for me to know how the brush works. Luckily, we have a very good team. When I was facing the problem, David and Brandon are always there to help.

After the presentation, I adjusted the parallel coordinates based on the feedback. By changing the opacity of each line and accumulating them, the darkness of the line could indicate an overview of how large amount of people answer each pair questions.

Personal reflection: From this project, I learned that whether data visualization is good or not depends on how well the graph presents your data and idea. Choosing a proper data visualization might be the task throughout the process of the project. Even when we refer to someone else's code, we need to have our own thought of presenting our idea better. For example, I adjusted the darkness of the lines to present the trend better. Furthermore, when we have multiple versions of the library, we might need to decide which library will be used for the whole project. Otherwise, we might be in trouble at the last phase. Overall, it was one of

my best experience of teamwork. Besides learning from data visualization, I learned how to work earlier on each task from them.

Appendix B: Brandon Markwalder Individual Report

We as a team divided the project work into three visualizations. The bulk of my work was related to developing the scatter plot matrix, which as discussed in the main report, did not lend itself well to the dataset. Rather than repeat the lessons learned and future plans, I will walk through the design process and the attempt to coerce a less than ideal visualization for the given dataset, into one that is capable of providing a robust high level overview.

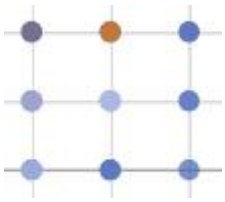
The scatter plot matrix was adapted from Mike Bostock's example and can be reviewed at: https://bl.ocks.org/mbostock/3213173. An iterative approach directed the design and the code was easily broken out into a standalone JavaScript file. I first reproduced the matrix in a more modularized manner using the original data. We identified the over-plotting problem immediately after feeding our data in, and we could have more seriously considered revisiting both our project design and choice of visualizations. With the compressed timeline for this project, we elected to move forward rather than make sweeping changes to our design. I did explore other visualizations including a grouped bar chart, but due to our static implementation, we were not able to easily feed our data in from the existing format. If we had chosen a dynamic implementation this would have been more realistic with database queries.

After the initial code adaptation was fleshed out and I had a working scatter plot matrix, it was handed over to Caroline for integration into her dropdown and parallel coordinates implementation. While I explored and attempted to implement more appropriate visualizations, I continued to refine the scatter plot matrix in the event that we were not able to implement a replacement visualization.

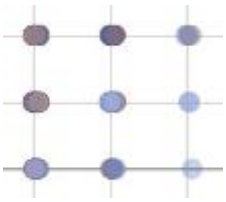
Refinements for the matrix included adding a zoom element, data jittering, and adjusting the opacity for each data point in the plot. We initially thought to jitter the data in separate csv files using the Python Pandas library, but a much more elegant solution presented itself quickly. The data was jittered on the x-axis with one line of code:

```
var xJitter = function(d) { return (x(d)) + Math.random()*2.5};
```

After jittering the data, the zoom, opacity, and brushing elements all proved to be redeeming efforts, modestly extending the usability of the visualization with the given dataset.



Pre-jittering and opacity adjustments



Post-jittering and opacity adjustments

Final touches to the scatter plot matrix included adding a legend, and modularizing the code by wrapping it into a callable function, allowing for easier code maintenance and reusability. The most important take away from this project is that all visualizations have their best uses, and for this dataset, a scatter plot matrix was not the best choice. Regardless of the enhancements, there are other options that better fit the data.

Appendix C: David Scroggins Individual Report

Personal Role: In the early stages I took responsibility for most of the semi-trivial organizational groundwork: setting up a GitHub repository, setting up a workflow board in Trello, setting up a Slack workspace (integrating with GitHub and Trello), and creating a template for the project. I also set up our initial datasets: cleaning the individual waves, merging them with the biomarkers, merging the individual wave datasets into a cross-wave dataset etc. This work was done entirely using the Pandas library (the scripts are not included, because they seem incidental to the focus of the project). I also tried to take the lead in organizing meetings and setting deadlines as well as writing the final report and handling submission. Caroline has a number of projects already on her hands and Brandon works fulltime (and has a family), so I saw this as an area to ease the stress on the group as a whole.

The initial design discussions really were truly collaborative (this was a great group to work with), and it is difficult for me to parse my exact contributions to this phase. We were largely on the same page from the start regarding how we wanted to approach the dataset and reached a quick consensus on the styles of visualizations we wanted to use and how we wanted to lay out the interface.

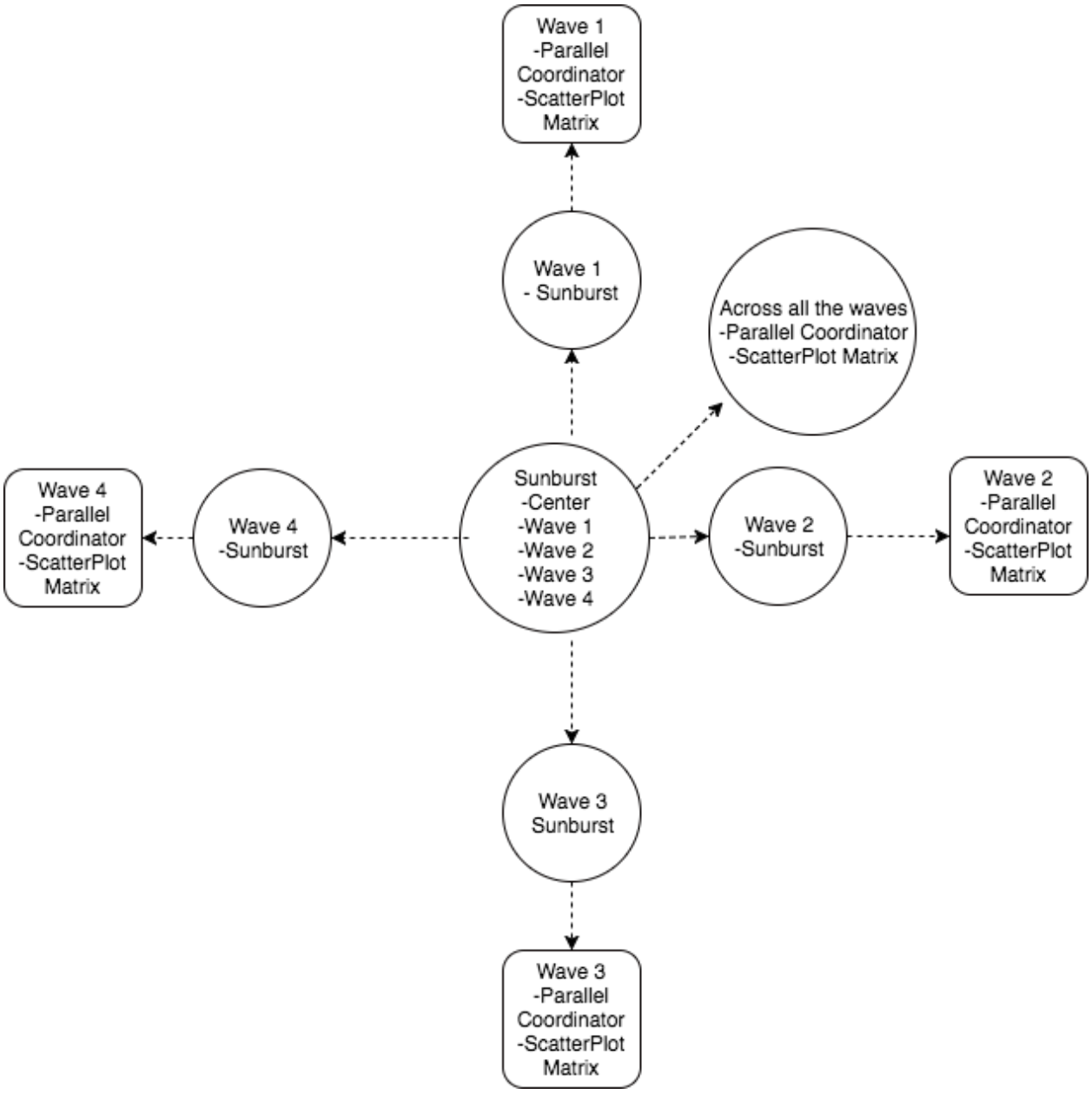
When it came to the actual building of the visualizations, I took on responsibility for the top-level interface as well as the wave specific sunbursts, which included preparing the hierarchical JSON files needed to this style of visualization. Getting the JSONs in working order ended up being the most challenging end of the project. Given the shape of our CSV, I never could quite get the parsing right, and in the interest of moving on to actually implementing the visualizations, I ended up editing some portions of them by hand. The final product renders properly, even if it wasn't produced in the most efficient manner. Implementing the Sunburst was a mostly straightforward process. I started with existing visualizations[4] and adapted them to the shape of the project. The adaptation had two primary facets: 1) refactoring to expose blocks of the code I needed access to customize and 2) adding D3 methods and functions to these exposed code blocks. I also included some straightforward D3 code to draw legends as well as a "button" (really just an svg element) to click through to the wave specific analyses.

My contribution to the other individual's code was standard and limited. We all worked pretty efficiently on our own regions of the project with well-defined connection points to keep everything standardized. Other than occasionally looking at Brandon and Caroline's code when asked, I feel we were able to efficiently work separately then converge when it was time to make sure it all integrated well. Organizing the presentation was, again, highly collaborative, and I feel that our contributions in this phase were again equally balanced. I also wrote the final report (it was reviewed by the team, naturally).

[4] Mike Bostock's Sunburst Partition: https://bl.ocks.org/mbostock/4063423 as well as a zoomable sunburst partition examples (there are multiple examples of this online and to be honest, I cannot remember which ones I used at this point, but the following is a decent example: http://bl.ocks.org/vgrocha/1580af34e56ee6224d33.

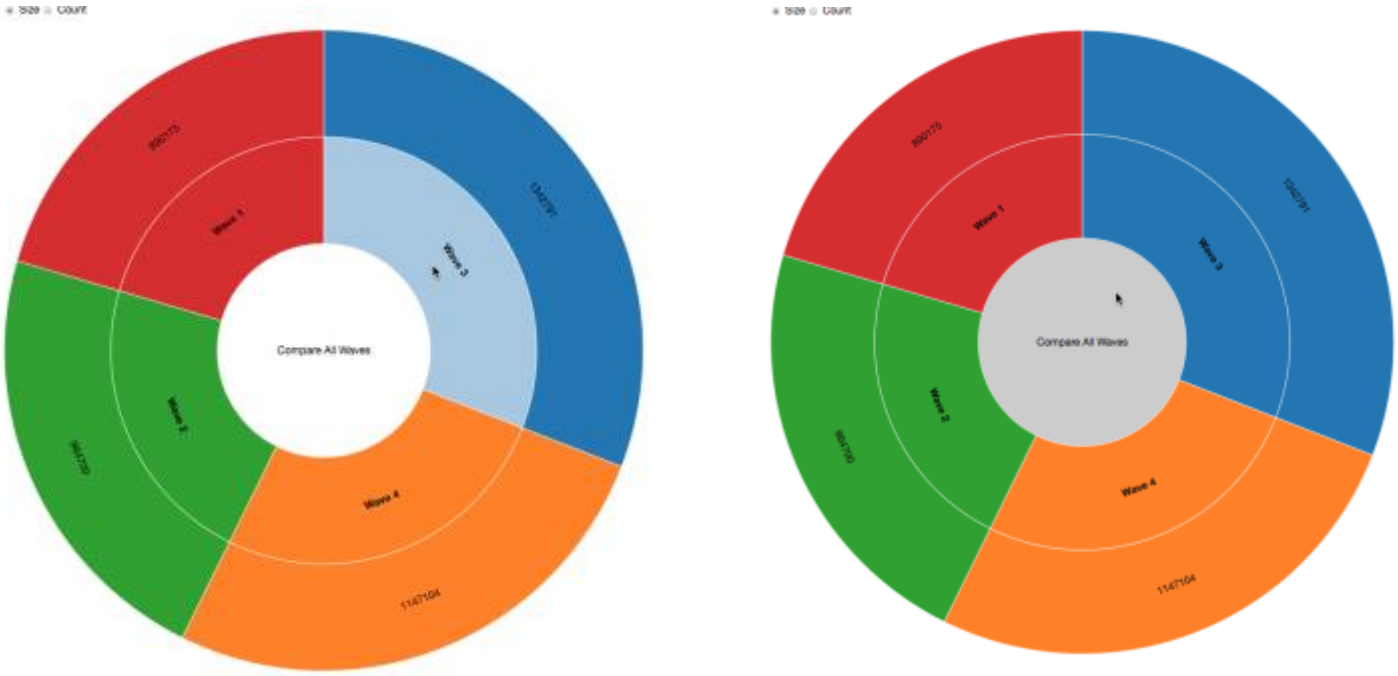
Personal reflection: My own personal observations largely reflect the body of the main report, in part because I wrote it, but also because when discussing it with the team, I realized that a lot of the personal observations I injected into the main report were reflective of how Caroline and Brandon felt. This project definitely drove home the importance of early implementation decisions (see our static vs dynamic decision) as well as emphasize my need to keep pushing to master a wider region of the stack. From a technical standpoint, this class pushed me heavily in terms of understanding how JavaScript works as well as digging into the D3 library of which I am beginning to realize I have only scratched the surface. I think the design process was also a valuable learning lesson in terms of paying closer attention to not just the overall structure of the data, but also the specific *types* of features we are working with. Some of the pain points discussed in the report could have been avoided up front. I think that we chose well in terms of the overall concept of our project. Sunburst visualizations allowed for both the complete overview of structure we were looking for as well as the ability to zoom into the dataset from this top level. While Parallel Coordinates and scatter plot matrices failed us a bit in terms of dealing with our categorical variables, conceptually they capture different facets of a thorough exploration of the dataset: space, time, clustering, concentration of labels etc. However, as we learned, specific visualizations may share certain conceptual components (consider Parallel Coordinates vs Sankey Diagrams), but work better given different types of features. If I had to identify a single most important learning experience throughout the project, this would be it. Overall, I think this project deepened my belief in visualization theory and techniques as absolutely essential in our much-hyped era of 'big data.' Our dataset doesn't really qualify as big data, but it *is* large, and its abstract encoding and thousands of pages of documentation make for a difficult up front learning curve. I believe our interface would reduce the steepness of this curve significantly, allowing a user to immediately ascertain the overall structure of the data as well as immediately begin generating general observations of patterns and trends. When this is scaled to the level of 'big data,' the benefits are magnified, but our current issues aren't merely that of the data we have collected, all of the data we confront and live in is, in a sense 'big'. Climate change, globalization, our increasingly networked lives all exceed our capacity to understand on a local level, and while quality visualizations aren't an antidote, they do represent an opportunity to reduce complex events to a comprehensible yet accurate visual space.

Appendix D: Initial Design Diagram

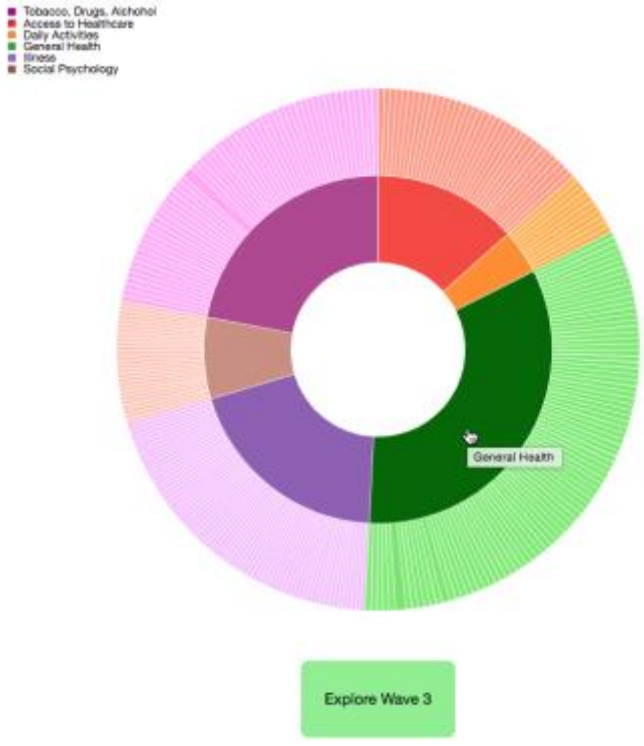


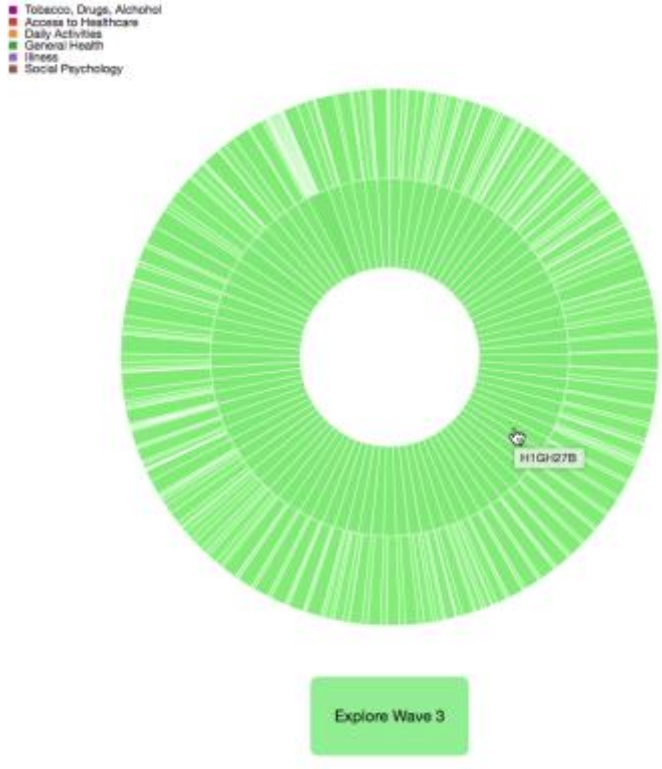
Appendix E: Screenshot Walk Through of Visualizations

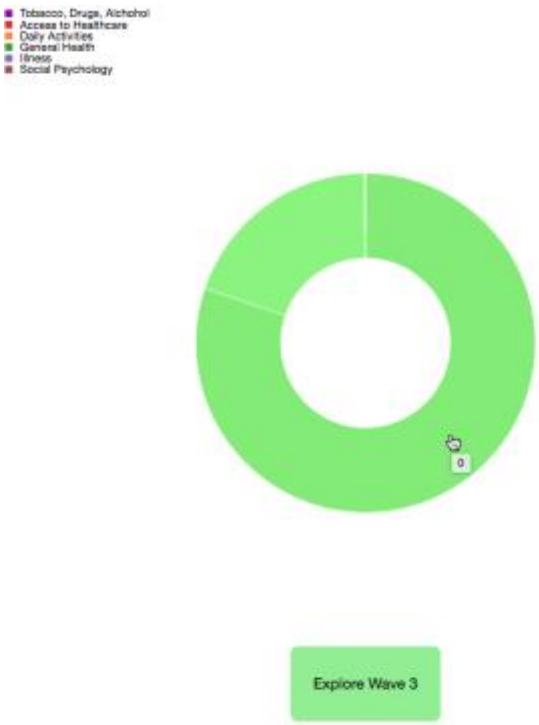
Top-level interface: Click for wave-specific analysis or to compare cross-wave



Wave-specific Sunburst: hover for node name, click node to descend one level, click center to go back one level

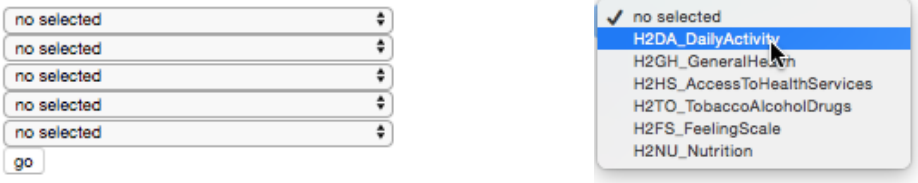


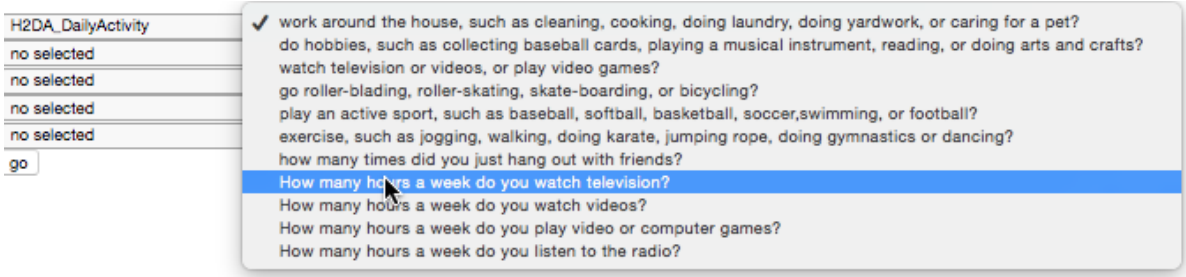




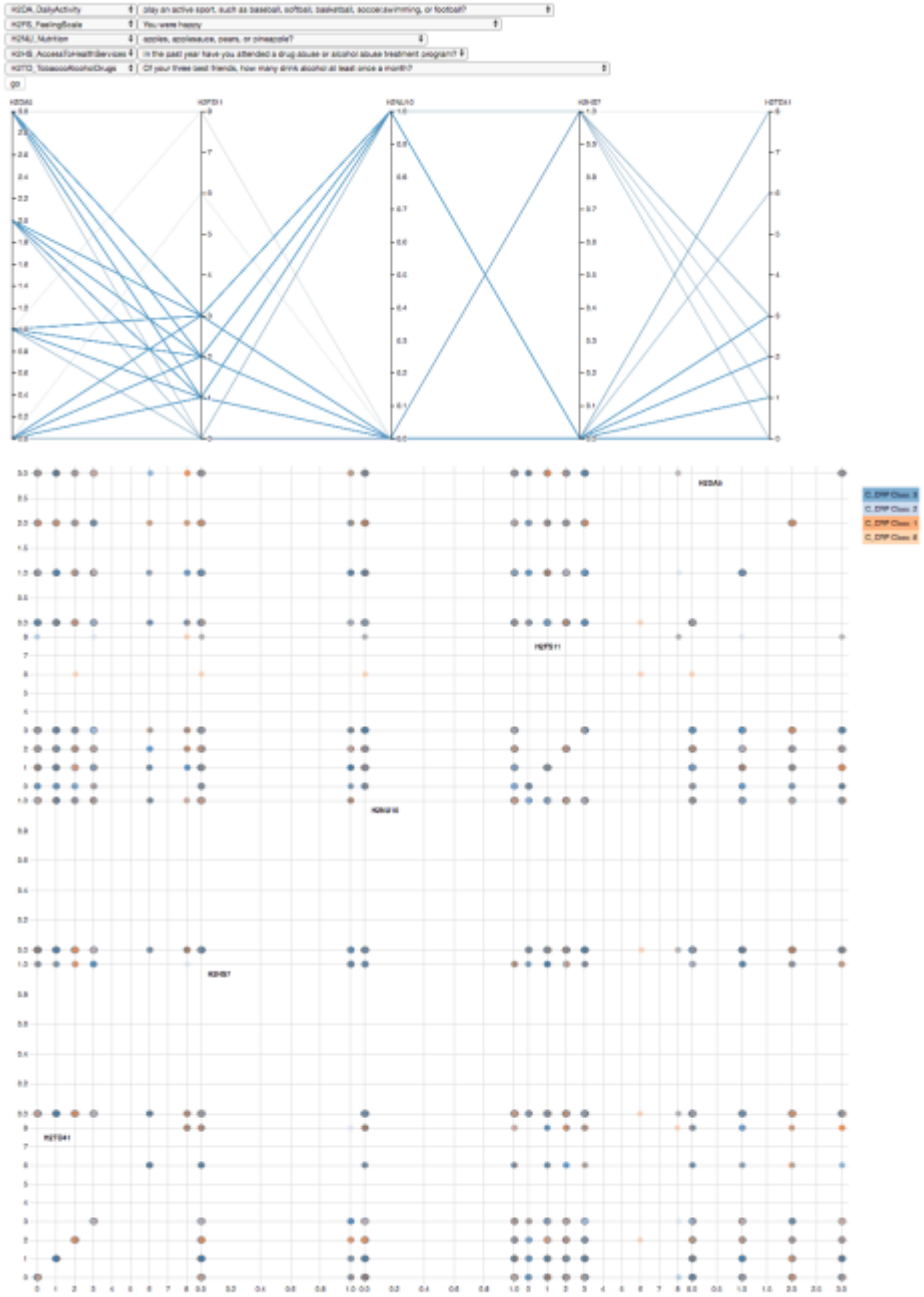
Wave-Specific Exploration

Drop down menu: select question section, specific question

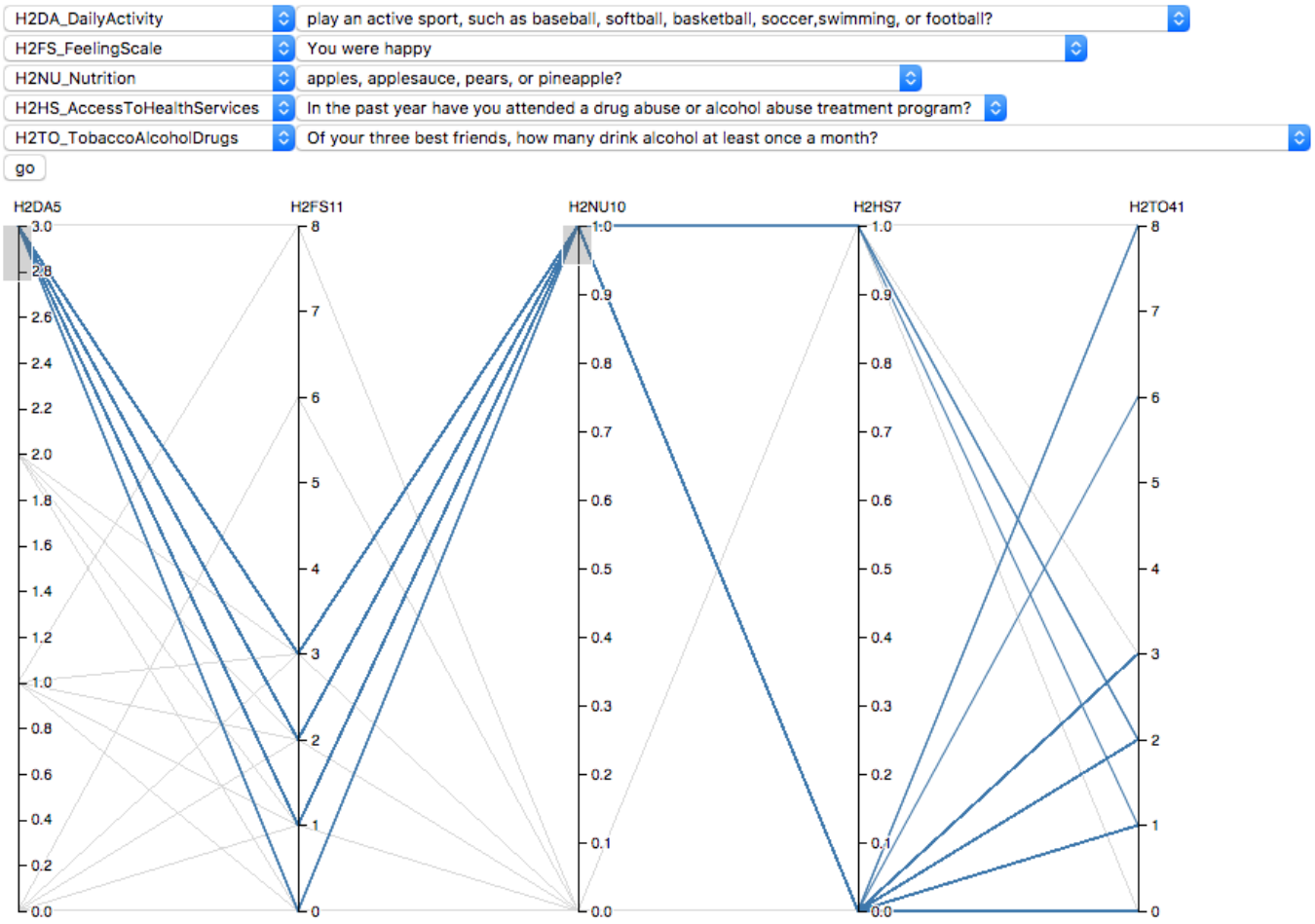




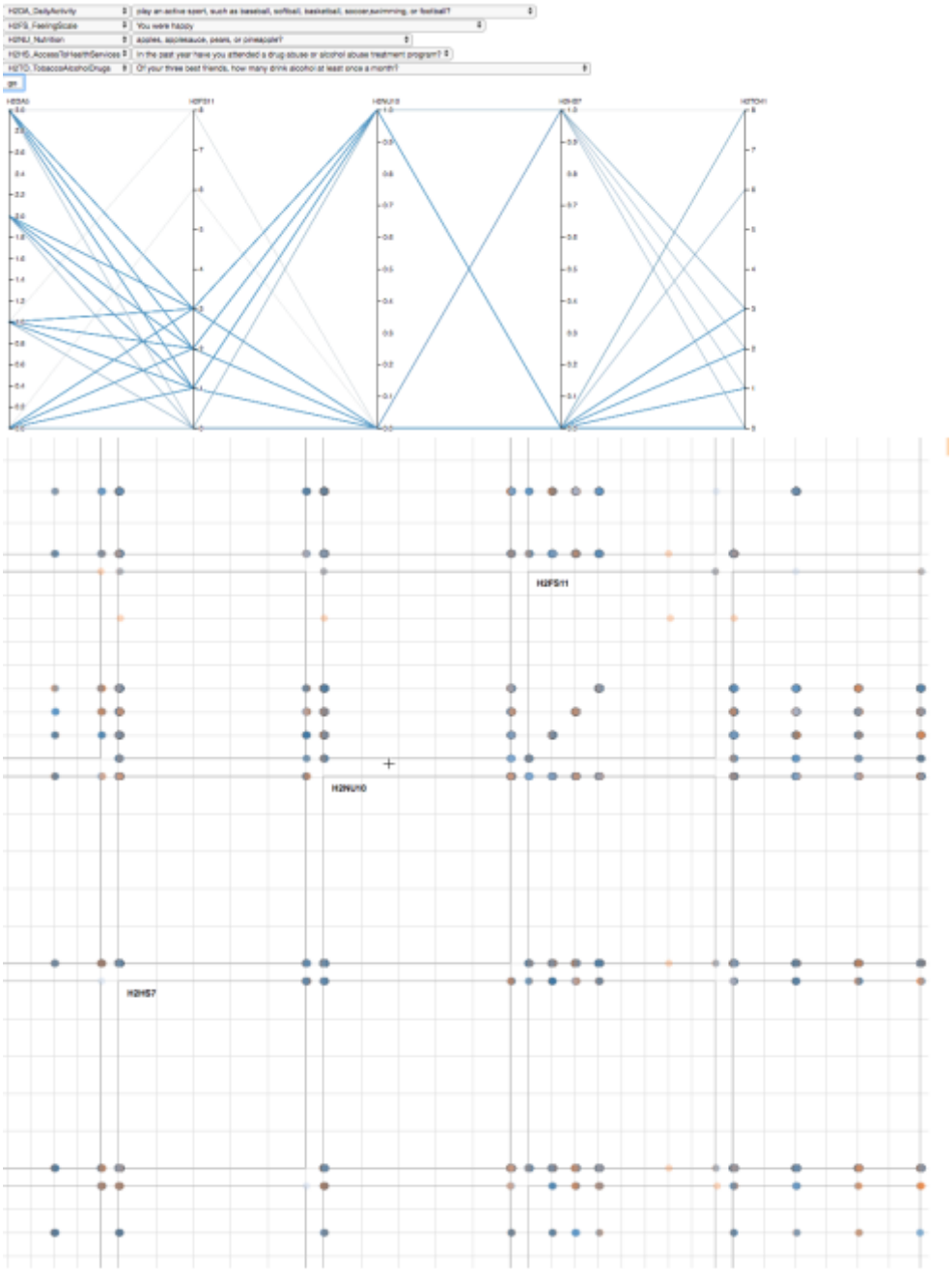
Results of question selections



Brush Parallel Coordinates to target patterns



Zoom scatter plot for closer observation



Cross-wave exploration: features are the same except you must choose a specific wave