Charity Vis

Visualizing the Charity Sector

http://groups.ischool.berkeley.edu/charityvis/
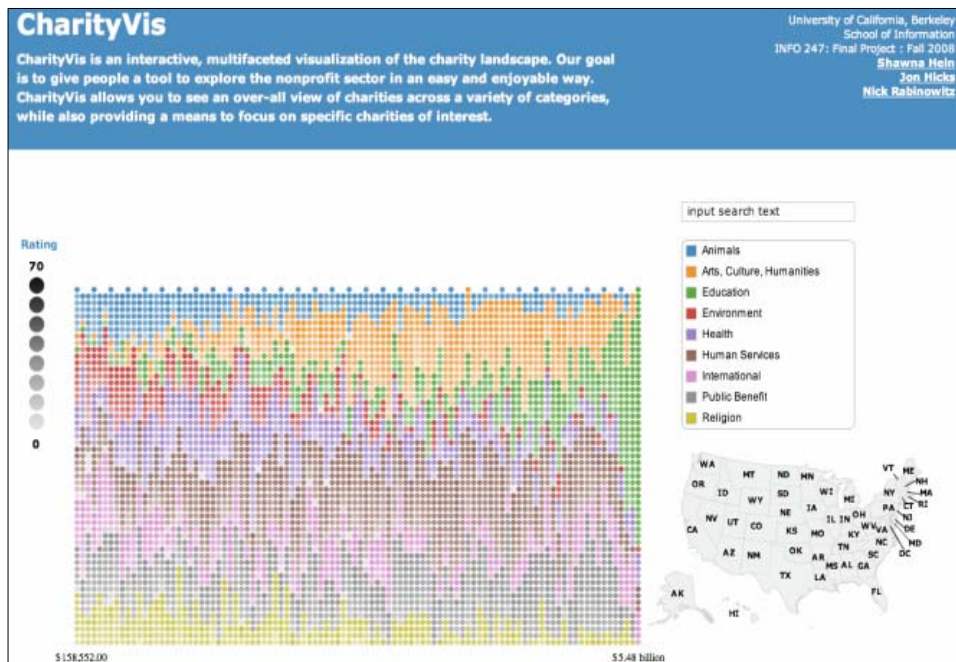
Shawna Hein, Jon Hicks, Nick Rabinowitz
Information Visualization
May 12, 2008

Abstract

Charity Vis is a data-driven visualization of information about charities in the United States. Charity Vis is designed to support a broad exploration of the non-profit sector while offering a meaningful context to specific charities within its dataset. We review the design process, tools utilized in its construction, and results and changes incorporated from our initial user evaluations. Our final model demonstrates the potential applications of such a visualization and presents possibilities for future research with the same data set.

Introduction

A variety of online resources are available to the interested public to help prospective donors and other users research and compare non-profit organizations. Despite the wealth of information available, none of these resources present visual information to facilitate exploration and analysis on a comparative, inter-organizational scale. In addition, the majority of these resources focus on presenting in-depth profiles of specific charities, rather than offering users an overall view of the non-profit sector.

Project Description

Charity Vis uses data scraped from several online charity research resources as the basis for an interactive visualization that allows users to better understand the non-profit sector as a whole while identifying specific agencies that meet those users' criteria. Ultimately, Charity Vis is intended to enable alternative ways to explore the non-profit space along a variety of data dimensions, including attributes such as size, efficiency, financial characteristics, geographic location, and external rating.

Target Audience

Charity Vis is designed with casual users in mind, though we hope it may also have applications for deeper research. Our initial use case pictured Charity Vis as a tool to filter through the at-times overwhelming corpus of sites like Charity Navigator and Guide Star, finding exceptional charities to which they might donate their time or money. However, as we explored a prototype of that corpus, we discovered nuances and unexpected patterns in the data that would not have been as readily apparent had we used traditional queries to parse the information. We imagine that researchers studying the non-profit sector might similarly be able to identify patterns or correlations that might provoke additional study.

Goals

We designed Charity Vis to support three central tasks:

1. Present the entire dataset across several dimensions and permit users to intuitively explore that landscape and access individual data points.
2. Allow users to filter data down along a variety of dimensions, arriving at a subset of charities that might meet that user's requirements or interests.
3. Once users have selected specific charities, offer them access to a large set of profile data available for each, including links to our primary data sources.

Related Work

A variety of prior work influenced the development of Charity Vis, either aesthetically, cognitively, or programmatically. Some of the most significant influences are discussed below.

NameVoyager|The Baby Name Wizard
(http://www.babynamewizard.com/voyager)

NameVoyager presents a chronological listing of the top thousand most popular names of babies born in the United States over the past hundred and twenty years, as reported by the Social Security Administration. Though our data sets are obviously different, NameVoyager dynamic filtering was perhaps the most important influence on the development of Charity Vis' visual style, with its color-separated bands of differing size indicating progression across the X axis. Our original design sketches (see "Design Process and Tools Used") reflect our desire to emulate an interface responsive, aesthetically striking, and quite useful to its intended audience.

Health Poverty Index Visualization Tool
(http://www.hpi.org.uk)

The Health Poverty Index (HPI) tool allows populations in England, differentiated by geography and socio-cultural identity, to be compared in terms of their "health poverty," a multivariate index defined by HPI creators. Their platform is built around providing a simple, scalable, factor-based visualization for comparing populations in a multivariate context stream, while providing the underlying data in as clear and accessible manner as possible for more in-depth analysis by tools of greater complexity. Charity Vis offers a similar service for comparing organizations and their resources, while providing an extra level of interactivity beyond HPI's ability to draw static graphs from user-selected data subsets.

80 Million Tiny Images
(http://people.csail.mit.edu/torralba/tinyimages/)

80 Million Tiny Images is a "a visualization of all the nouns in the English language arranged by semantic meaning." 80MTI was of interest to us not only for its staggering scope, but for its ability to "humanize" that scope by allowing a user to pluck from its structure any single point and from that selection determine immediately what that point represents. Echoes of its layout and selection method are visible in Charity Vis' info box.

Gap Minder
(http://www.gapminder.org)

Gap Minder is a non-profit foundation whose Trendalyzer software is used present animated time series graphs of complex, multivariate data. Their tools are chiefly used to explore the relationship between countries based on a wide range of statistical data. Our hope with Charity Vis is to present data of arguably similar complexity in an equally engaging, easy-to-navigate format. The visualizations presented by Gap Minder are eminently accessible, especially when compared to the complexity of other interactive multivariate visualizations, and are notable for their use of animation as an effective tool to provoke further comparison rather than as an aesthetic flourish.

Design Process and Tools Used

In order to create our final visualization, we followed an iterative design process, creating multiple rounds of mockups and prototypes and soliciting user feedback.

1. Data Sources and Dimensions

The first step in the process was to acquire the charity data to display. We identified a small number of charity data sources, each overlapping somewhat with the others. Most of the data on these sites comes directly from IRS public filings.

- **Charity Navigator** (http://www.charitynavigator.org/) has a clean, highly normalized data set, but only included a small (5000+) subset of all U.S. charities, and was lacking some data points that other services included. Given the high level of data consistency, however, and the custom quality ratings that we felt were important to our visualization, we decided to make the Charity Navigator records our core dataset.

- **GuideStar** *(*http://www.guidestar.org/*)* has a much larger dataset of charities than Charity Navigator, and in its complete records included several data points that Charity Navigator was lacking, such as founding year, total assets, and a breakdown of revenue sources. However, its records were inconsistently complete, with many only including a bare minimum of organizational data. GuideStar is also, in theory, a subscriber service; but a Google search turned up a free gateway to their data. We decided to use GuideStar's additional fields to supplement our data, though this was a tricky proposition as GuideStar and Charity Navigator didn't share a common unique ID for their databases.

- **The National Center for Charitable Statistics** *(*http://nccs.urban.org/*)* offers an extremely large dataset of charity data pulled from IRS filings, including a large range of historical data. While this seemed an appealing option, the data schema was significantly more complex than those of the other services, and was only accessible to researchers in either sub-optimal formats or through a pay service. We decided to forego this dataset.

- **The Better Business Bureau** *(*http://www.bbb.org/*)* offers a slightly different range of information on each charity, including another rating assessment for specific topics such as fraud. The dataset was sparse, however, with a relatively small number of charities and lots of missing data, and again there was no shared unique identifier. We decided to skip this dataset after our experience scraping GuideStar.

2. The Scraping Process

The data we wanted to acquire from Charity Navigator and GuideStar was only available on their respective sites in HTML format. We inquired with Charity Navigator whether it would be possible to access their data in a better format, and while it seemed like this might be an option, they weren't willing to release all of their data and the administrative process required seemed like it might take too long to work within our project timeframe. As the Charity Navigator site was consistently formatted and eminently scrapable, we decided to scrape the data.

- We used a homemade Ruby script to scrape all the records on the Charity Navigator site into a local MySQL database, using the Hpricot HTML parsing library to quickly identify and save page elements via XPath addresses. This took one scrape, though it required iterative debugging each time the script hit a new problem with an aberrant record or value. One issue we discovered in this process was that some of the financial figures we were dealing with were over the standard four-byte size (i.e. over 2.1 billion for signed integers), and required changing

the data type in our MySQL field. We ended up with 5,334 charity records.

- The next step was to scrape the corresponding records in the GuideStar database for additional data. This proved more challenging than we thought, and took a number of passes, including some manual sorting, due to the lack of a common unique identifier and other differences between the two databases. We started with a two-phase search and scrape, in which we ran a script to query GuideStar on the name, city, and state of each charity, then used the IDs of the top search results to run a second scrape for the individual records. This worked for approximately half the records.

- We then ran through three or four more passes to check the records we had scraped and find the records we had missed. We devised several heuristics to assess the accuracy of a match, including similarity of name, similarity of revenue (exact if the records were for the same tax year, but this was often not the case), matching CEO names, etc. We eventually ended up with some 400 missing records and another 80 or so that needed manual checks, which we divided up and ran through one by one.

By the end of this process, we had a much more complete dataset, but had we known the trouble it would take, it is unlikely that we would have tried to scrape GuideStar without a unique ID to match on. With 400 or more records still missing the extra fields, it is difficult to use the extra data in our visualization of the entire set, as the many null values cause visually disruptive patterns (e.g. the 400 records all cluster together incongruously in a way that's likely to confuse the user). We can, however, display the additional information in the "More info" window for individual charities, so the exercise was not entirely in vain.

3. Exploratory Data Analysis

Once we had a substantial local dataset in MySQL, our next step was to use existing tools for exploratory data analysis. The goal of this analysis was to get a better feel for the data, allowing us to understand some of the patterns and outliers and to try to visualize the data using traditional techniques. We used Tableau for the majority of this exploration, though we also tried several map-based views in ManyEyes.

Analysis Findings

- We were particularly interested in looking at the budgets of different organizations, as we felt like this was one of the key differentiators between charities that a user might be interested in. We found that the majority of charities were clumped below the $250M mark, as shown in Fig. 1, while a small number of very large organizations stretched out in a "long tail" that was in danger of dominating our visualization. Figure 1 also demonstrates some problems with the data – for example, the American Red Cross is shown with a zero budget, because its financial figures were too large, at $5+ billion, for Tableau to render.
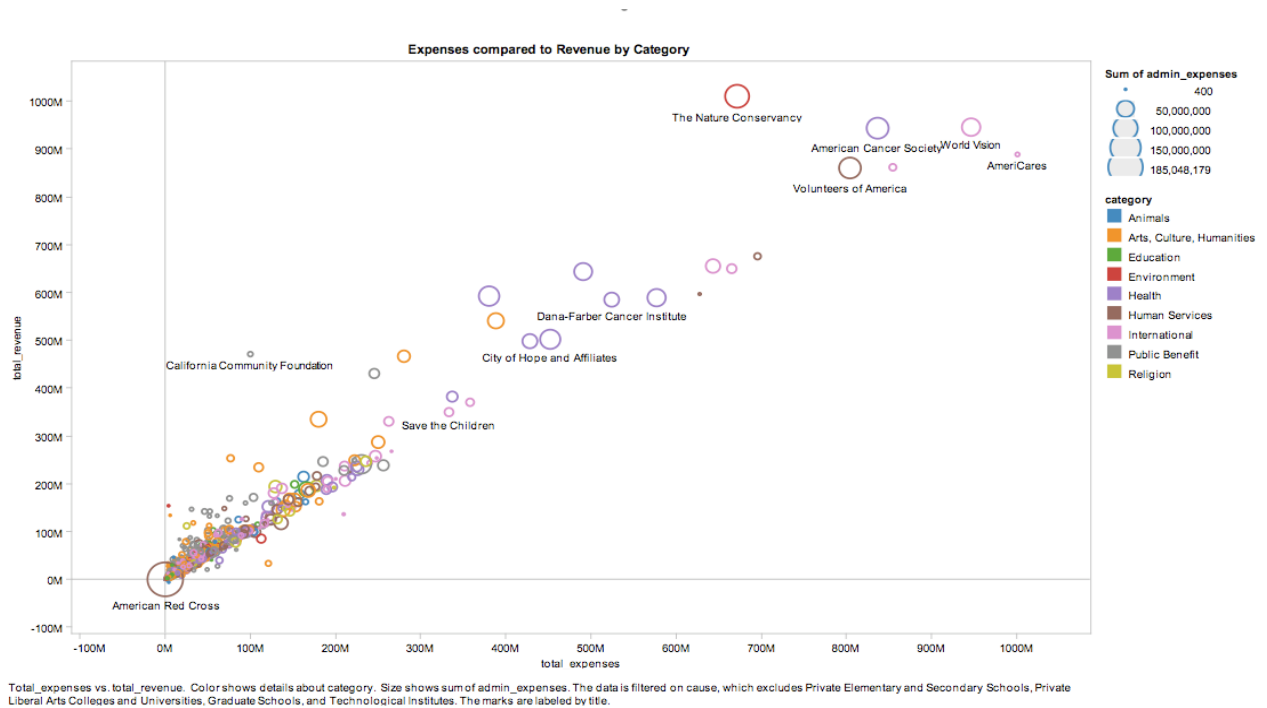
Figure 1: Scatterplot of expenses and revenue

- We found a similar pattern in the distribution of charities across states within our dataset. Two states, New York and California, accounted for a significant percentage of the charity records, with almost 700 each, while many of the smaller states had only a handful (see Fig. 2).
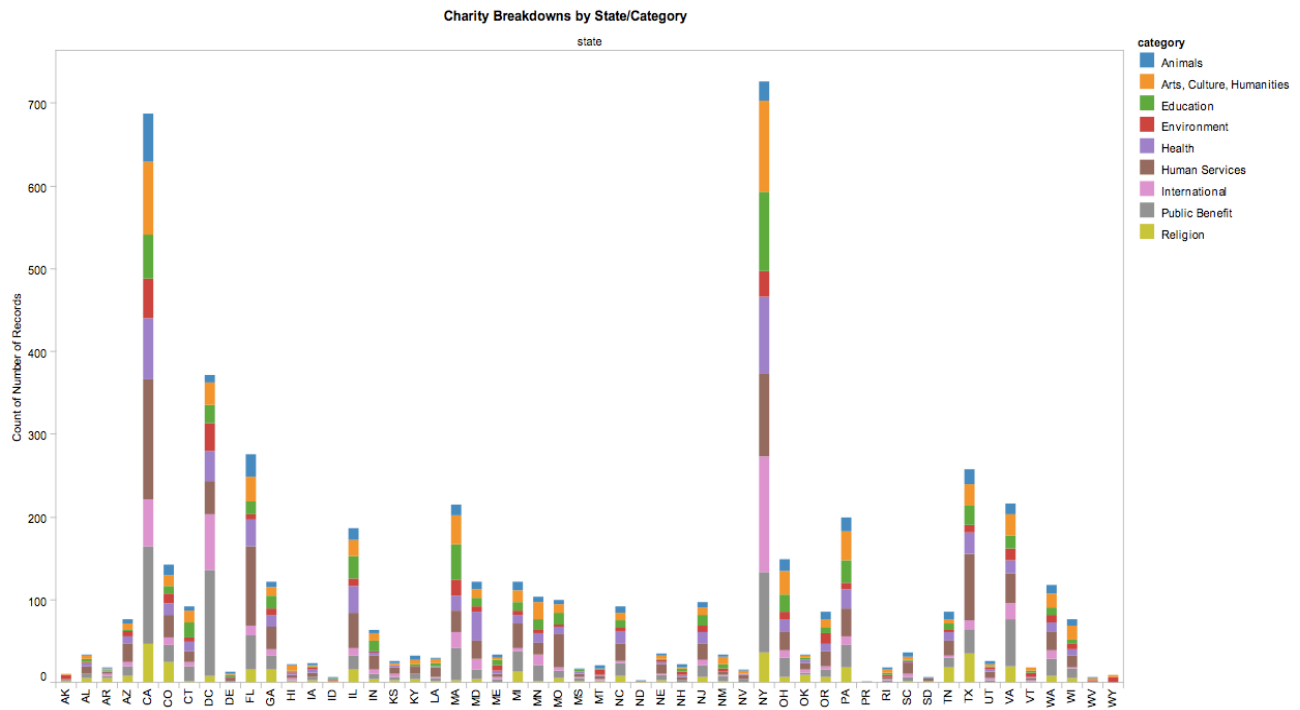
Figure 2: Charities by state and category

- The Charity Navigator ratings were another point of interest. The majority of the charities fell into the middle range of ratings, what Charity Navigator designates as "three-star" charities, while a few categories of service did particularly well (universities, community organizations) or particularly poorly (public broadcasting) – see Figure 3. When graphed with expenses, it became apparent that despite the popular conception of large non-profits as poorly run and inefficient, there seemed to be a clear correlation between large budgets and high ratings, especially in the rarified sector above $250M – see Figure 4.
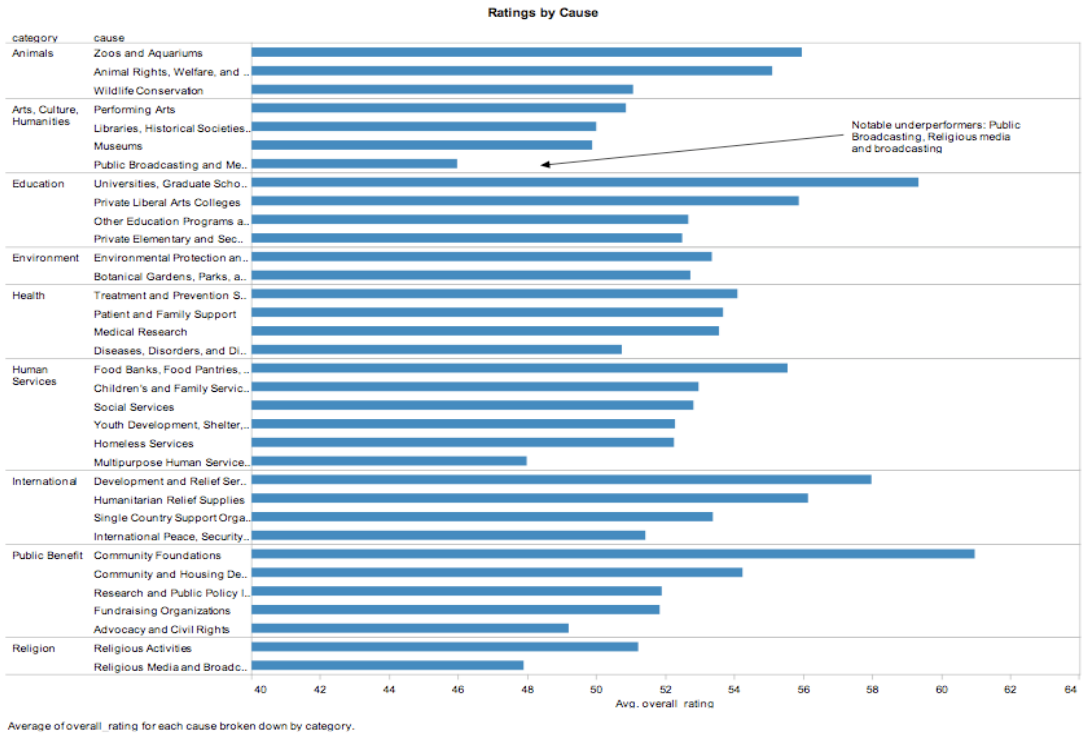
Figure 3: Average rating by category of service
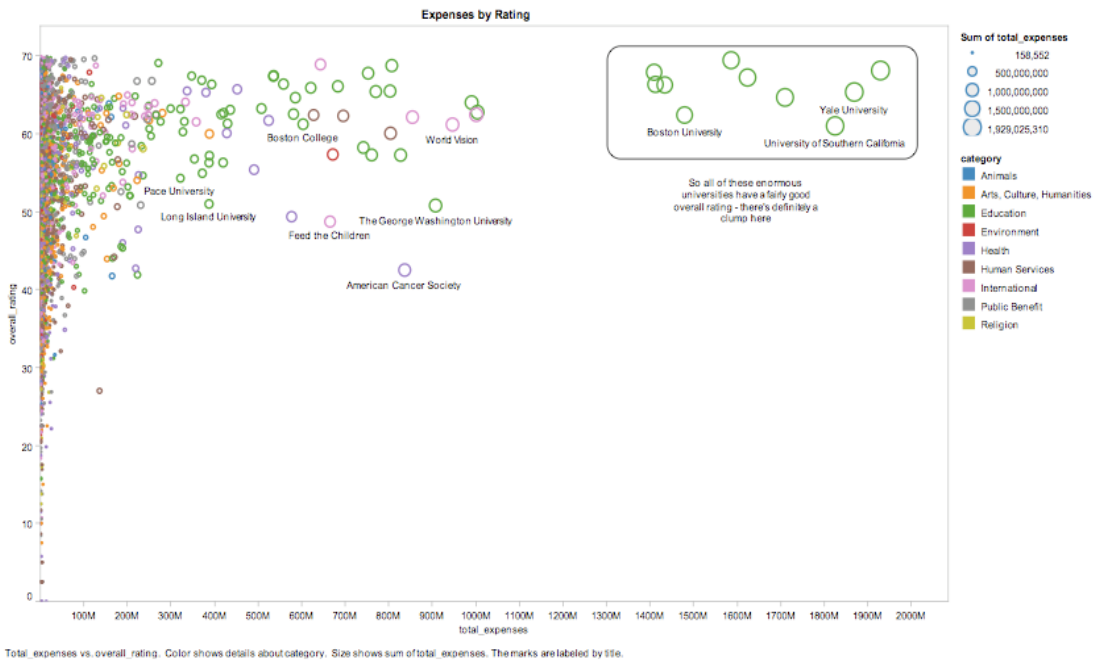
Figure 4: Rating vs. budget

4. Rough Mockup

Based on our findings in exploratory data analysis, we began to brainstorm a rough concept for an improved visualization. We had identified the dimensions we were most interested in displaying: budget, rating, category of service, and geographical location. We had also identified some issues we would need to address in our visualization, such as the tendency of large organizations to visually dominate the space, and the difficulty of showing values such as budget on a linear axis when there was such disparity between the smaller and larger organizations.

We also focused on the main feature that was missing from Tableau, the ability to select and learn more about a specific charity in the context of an aggregate view. While Tableau offered this affordance in a scatterplot showing all values, in practice it was very difficult to select any but the largest organizations for any of the layouts we tried.

With these goals and restrictions in mind, we put together sketches of an initial layout, two of which are shown in Figures 5 and 6, using Adobe PhotoShop.

- We would map budget to the X-axis using (we thought at the time) a logarithmic scale to stretch out the distance between the smaller values and compress it between the larger ones.
- Category would be mapped to color; we had nine categories, which seemed reasonably below the limit of distinguishable color groups.
- Rating would be mapped to the lightness value of the color, so that higher-rated charities would stand out as darker and fully saturated, while lower-ranked charities would fade to almost white. In addition, we would map rating to the Y-axis within each category, so that higher-rated charities would be higher up within their category.
- While we decided not to add another visual encoding for geographic position, as it might seem overwhelming and we felt this was the least important dimension to show in this context, we imagined a map on the screen that could be used to filter the data or display the location of the current selection.
- Each charity would be individually selectable, with additional information shown on selection.
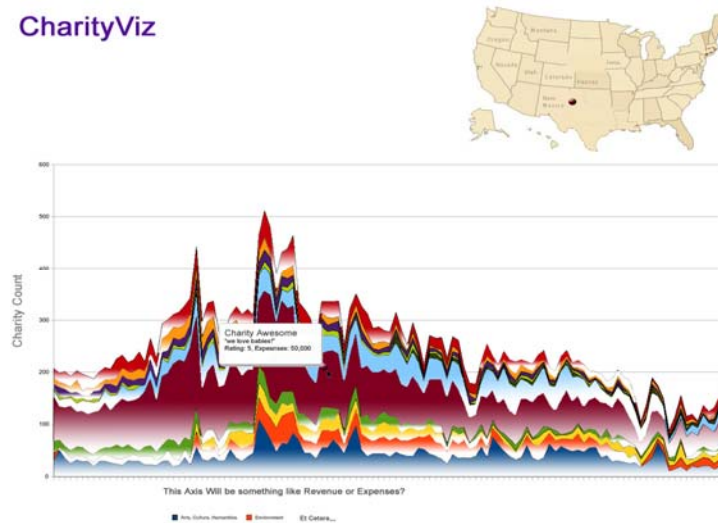
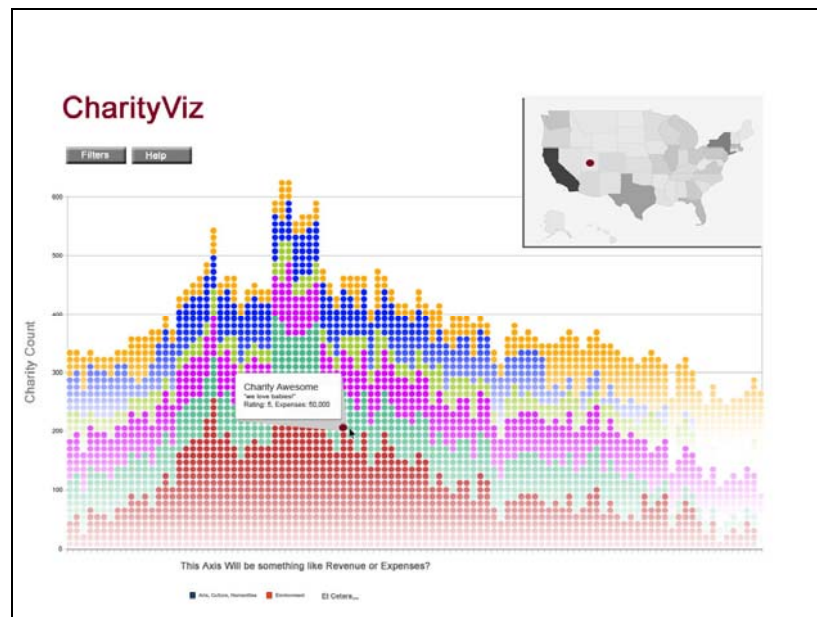*Figure 5: Initial Sketch #1*



Figure 6: Initial Sketch #2

5. Initial User Feedback

Soon after we finished our initial design sketches, we presented them in class and received valuable comments and critiques on our initial model. Based on the feedback from that discussion, we developed a more refined mock-up of our visualization. Using PhotoShop again, we created a second iteration of sketches with significantly more detail.

- We included a category legend following a suggestion made during our presentation, and, per another suggestion, designed it to be a filter selection control as well, allowing the user to select

a category in the legend and see only charities within that category.

- We tried several styles of visual representation for the individual charities in our dataset. The individual charity nodes in our initial sketch, shown in our class presentation, had drawn several complaints that we felt were an artifact of the sketch, and we knew that if we were going to user-test our second mock-up we would need to be careful about how each node was drawn. We wrote a short Javascript function for PhotoShop to place the nodes in a stack with varying levels of opacity so that the layout would be more realistic and involve less manual labor, and this allowed us to try out several different styles of charity node, as shown in Figure 7.
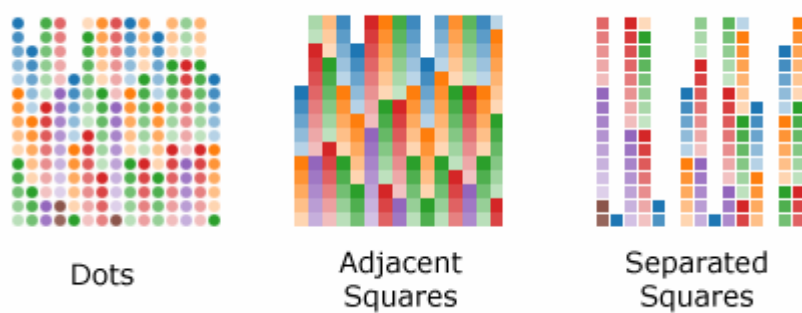
Figure 7: Charity Node Styles

- We also wanted to use visual elements, not just tabular information, to show detailed data for each charity. To this end, we designed small bullet graphs in the style of Stephen Few (see Figure 8) to indicate at a glance how highly rated the selected charity was and where its budget fell on the span of all the values in the dataset.
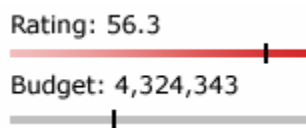
Figure 8: Charity detail bullet graphs

- Finally, we experimented with a view that broke categories down into subcategories when a category was clicked in the legend. Instead of showing one small layout of the category, we would show several layouts, one for each subcategory, as shown in Figure 9.
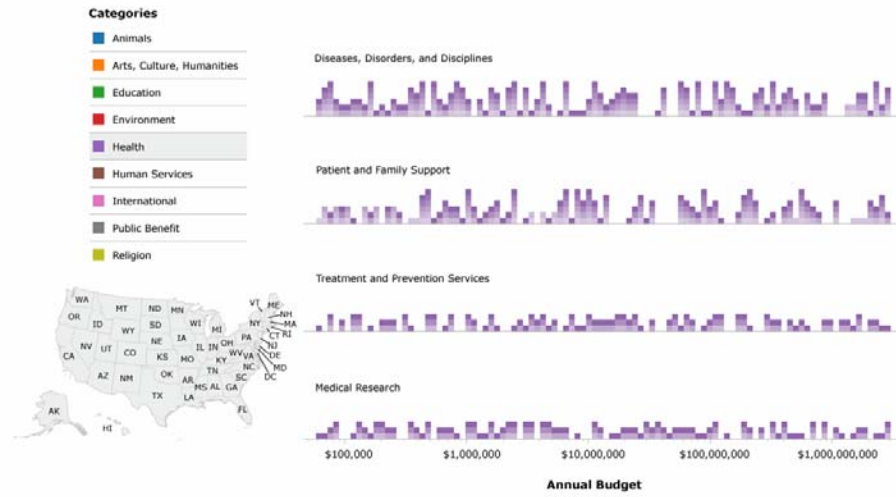
Figure 9: Subcategory breakdown view

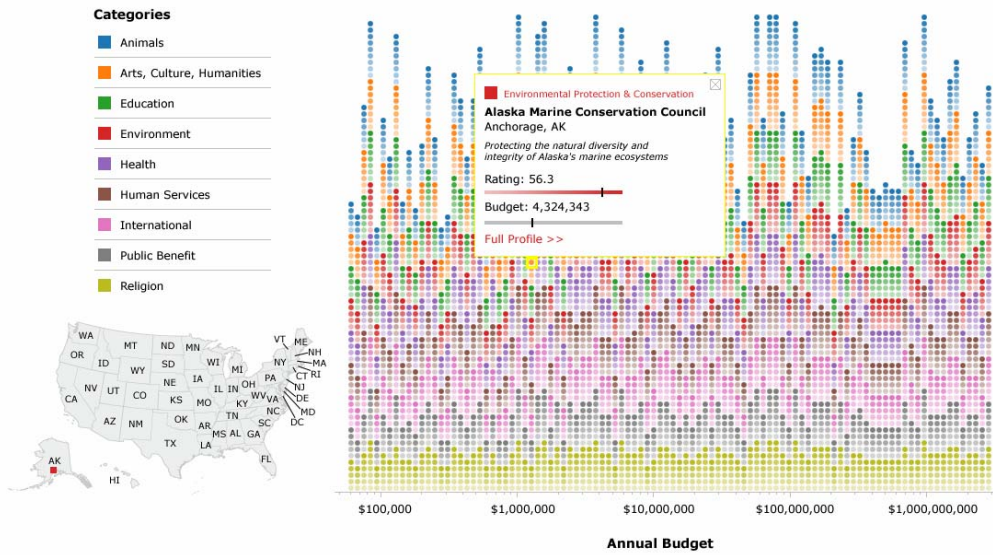The final sketch for the main view can be seen in Figure 10:

Figure 10: Final Mock-up

6. Initial User Testing

Our next step was to conduct informal user testing on our refined mockups. We showed the mockups to three potential users, asking them about their impressions of the visualization and how they thought it might work. We asked them to describe what how they would be likely to interact with the visualization, what they might click on, and how they would expect the controls to behave. We also asked for specific feedback on the node styles, among the options outlined above, and the bullet graphs, which users seemed to grasp without any problem. Overall the feedback was positive, in that users were interested by the visualization and indicated they would spend time exploring it. We also got specific feedback on improvements and user needs that we could pursue in our implementation, as outlined in Table 1.

User	Technical Experience	Feedback
P1	Novice	• preferred circles over squares to represent points, seemed more tactile • preferred color scheme on white background
P2	Expert	• thought points need to be bigger • wanted to have a clearer sense of ratings and their scale • wanted a clearer legend
P3	Expert	• wanted a full graph of each category instead of breakdowns • wanted slider controls on rating to control thresholds • wanted a clearer legend

Table 1: User Testing Feedback, Round One

7. Initial Flex Implementation

Based on our refined sketch and the feedback from our first round of user tests, we began the coding work to implement the project. Working with ActionScript 3 in a Flex environment, with the support of the Flare libraries, we implemented a simple interactive space in which we could test our visualization on users directly and from them further refine Charity Vis.

Our initial prototype implemented the basic layout of the nodes and the category legend. We quickly ran into an important issue with the scale used for the X-axis. We were creating a layout in which it was necessary to group the nodes into "buckets," but Flare doesn't provide a logarithmic bucket scale, and it turned out to be quite difficult to produce one on our own. When we finally succeeded, with the help of superior mathematicians among our peers, it turned to be the wrong

option for data that didn't follow a pure Zipf distribution. In the end, we decided to use the Flare-provided quantile scale, which simply divides the data evenly into a specified number of buckets, offering, with 100 buckets, a "percentile" distribution of data (i.e. the top bucket represents the 99th percentile, etc).

8. Additional User Testing

Using our prototype in conjunction with sketches, we informally tested Charity Vis on a second series of users (Table 2). For each user, we asked them to imagine that they were investigating a subset of American charities and observed how they interacted with the prototype and what as-yet unimplemented elements of the sketch they found useful or interesting.

User	Technical Experience	Feedback
P4	Novice	- had a specific charity in mind from the outset and wanted to search for it - dismayed that no text search was available - preferred darker background to lighter one
P5	Novice	- got distracted from charity search by playing with filters and transitions - like idea of full context to each point, but wanted more depth to info box - wanted a way to see similar charities, like "amazon recommends"

Table 2: User Testing Feedback, Round 2

Their feedback was incorporated into the final model when possible, as explained in more depth below.

Final Implementation

We ended up with an interactive visualization that we were quite happy with. Our visualization ultimately allows users to explore data about charities in ways they might not have been able to with the tools previously available.
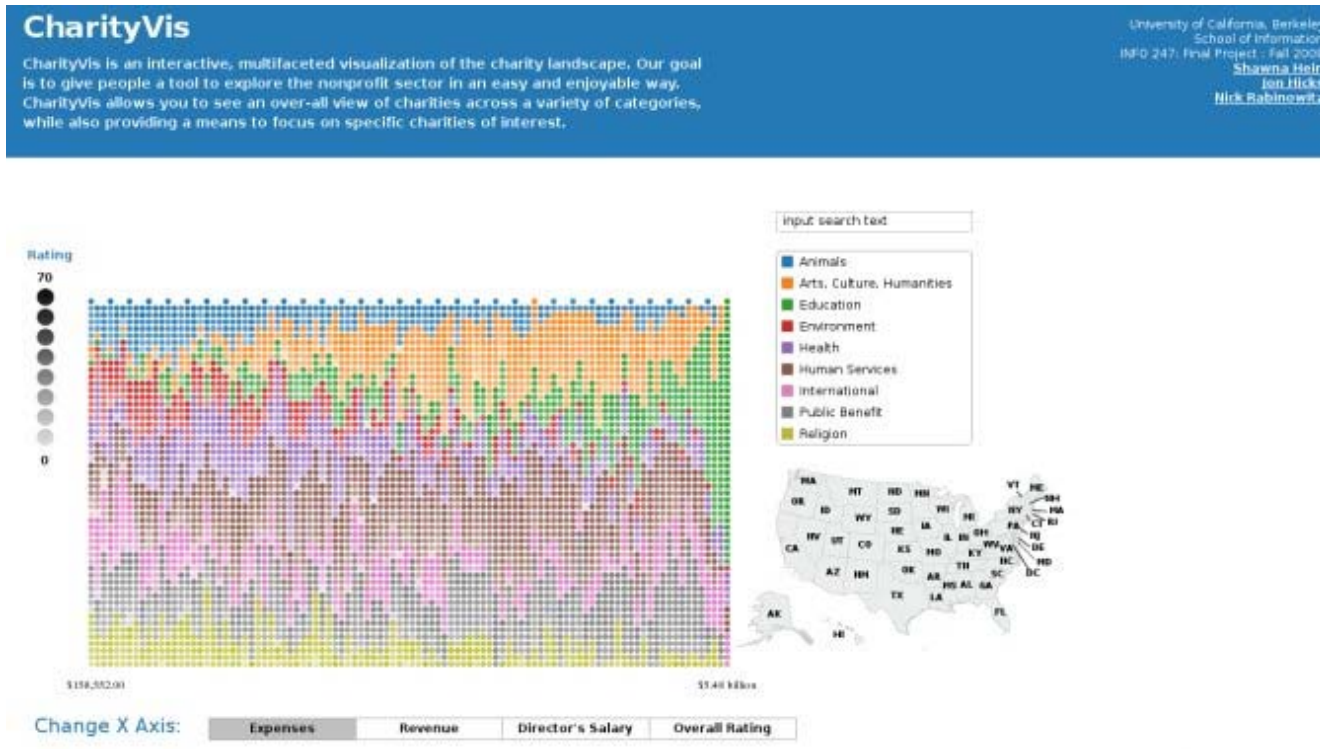

Figure 11 presents the main view of our visualization's interactive components: the main graph, the user-specified X-axis, the interactive legend and map filters, and a search box.

The Graph

The main graph in the center shows our charity dataset, with each charity visualized as a small colored circle of varying opacity. The color represents category (referenced via the legend) and the saturation of the circle represents the Charity Navigator rating (as the legend to the left of the graph implies). The more saturated the color, the higher the rating. Within each category, the higher up on the Y-axis a charity is, the more highly that charity has been rated. The categories are stacked on top of each other so as to facilitate cross-category comparisons. As elaborated below, the X-axis is user-determined, though it defaults to charity expenses, as we felt that was the easiest scale from which a novice user might determine a charity's total size. The X-axis is arranged in buckets, so each vertical

column of charities represents one range of values. Additionally, each charity on the graph is interactive. To see more about one of the charities, the user clicks on one of the circles, which highlights the node and opens a small info box (see Figure 14). The user can then click on "More Info" to see extended information about the charity (see Figure 15).
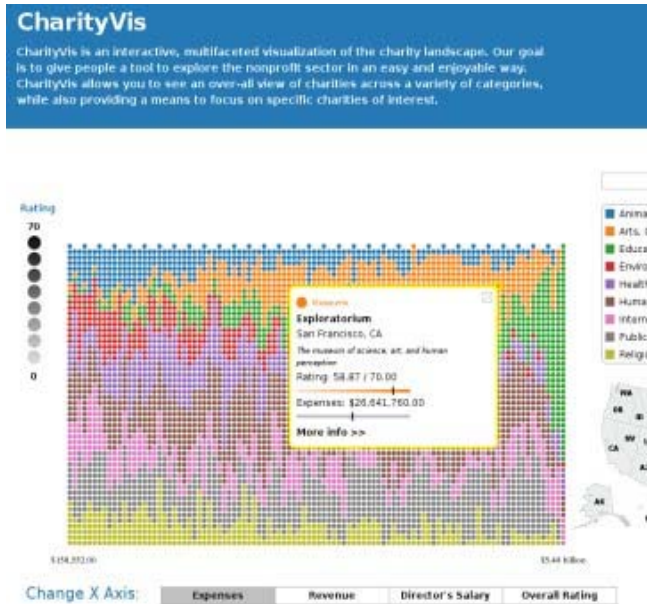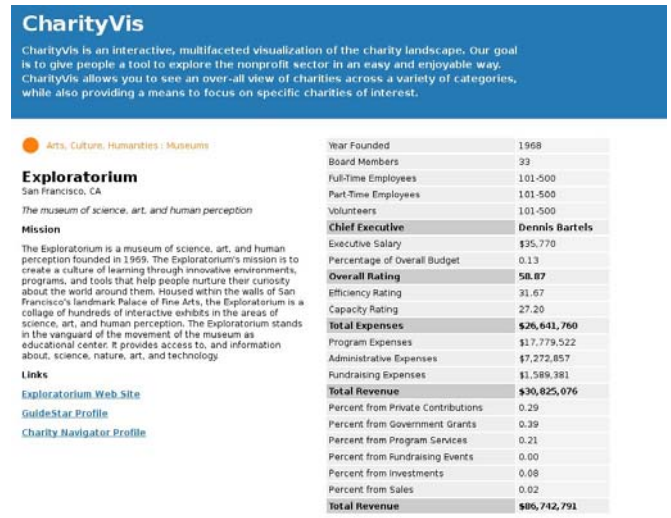

Figure 12: Charity Selected

Figure 13: Charity Detail

The X-Axis

Our X-axis is dynamic, and can be changed depending on what dimension of the data the user is interested in exploring. The user simply clicks on one of the buttons below the axis to change it. In the main screen, the Expenses button was selected. In Figure 16, the user has selected "Director's Salary" as the X-Axis, and the charities have been rearranged according to this value.
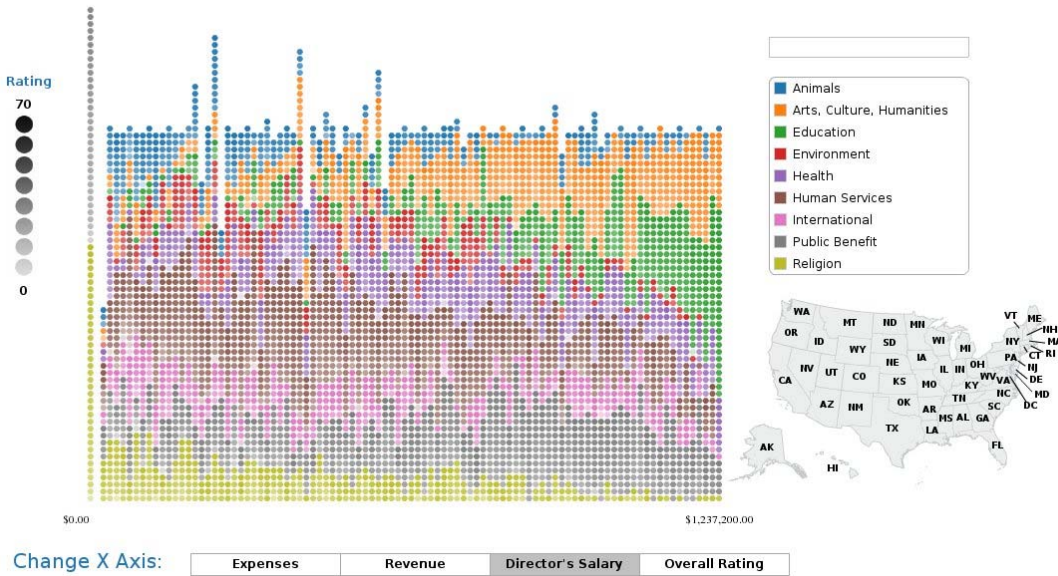
Figure 14: Dynamic X Axis

Filtering

Charity Vis allows for filtering based on category and location. From our design, we expect that a majority of users will want to drill down on only one (or a small group) of categories or states.   A few potential use cases, and then means by which a user could realize them, are described below.

1. By Category via the Legend

The interactive legend allows the user to filter by category. Clicking on a single legend item selects that category, and clicking while holding down the Shift key selects multiple categories. In Figure 15, the user filters for charities categorized under "Art, Culture, Humanities."
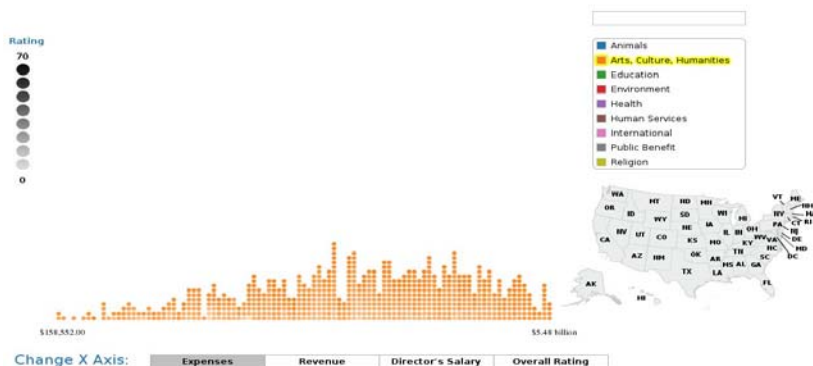
If a user clicks on one of the X-axis buttons at this stage, the filters will stay on and the charities in the particularly filtered categories will rearrange themselves.

2. By State via the Map

Charity Vis also allows for filtering by state. We created an interactive map for this purpose. Like the legend, a user can click on one state (via its label) or several states by holding down the shift key. This will automatically filter the charities by state, and only charities based in the selected states will show up on the graph. Note that this filter records where charities are located, not necessarily the communities that these charities serve. In Figure 16, a user has decided to view only the charities located in California and New York.
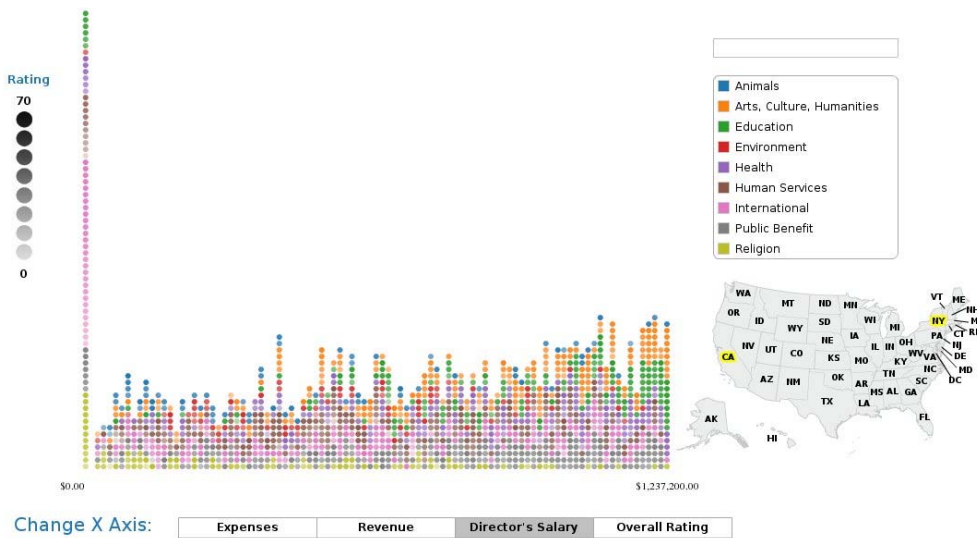

Figure 16: Filter by State

Users can also filter by State and Category at the same time, as seen in Figure 17. The dataset is generally large enough to support at least one instance of every category in each state, though some of the least populous states (e.g. Wyoming) will have empty categories.
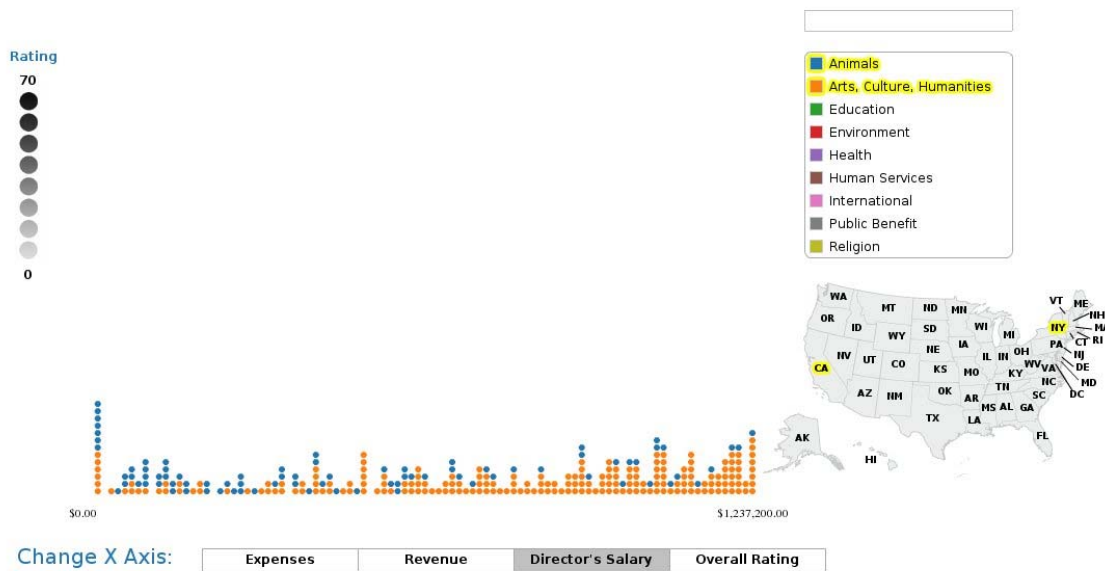
Figure 17: Filtering by Category and State

The Search Box

Several users in our tests noted how useful a text-based search function would be. To that end, we include a modest search box which does a simple text search of charity names. In Figure 18, a user has searched for "exploratorium," and has found the charity whose name matches that term. The search performs a simple substring match on the charity name and category.
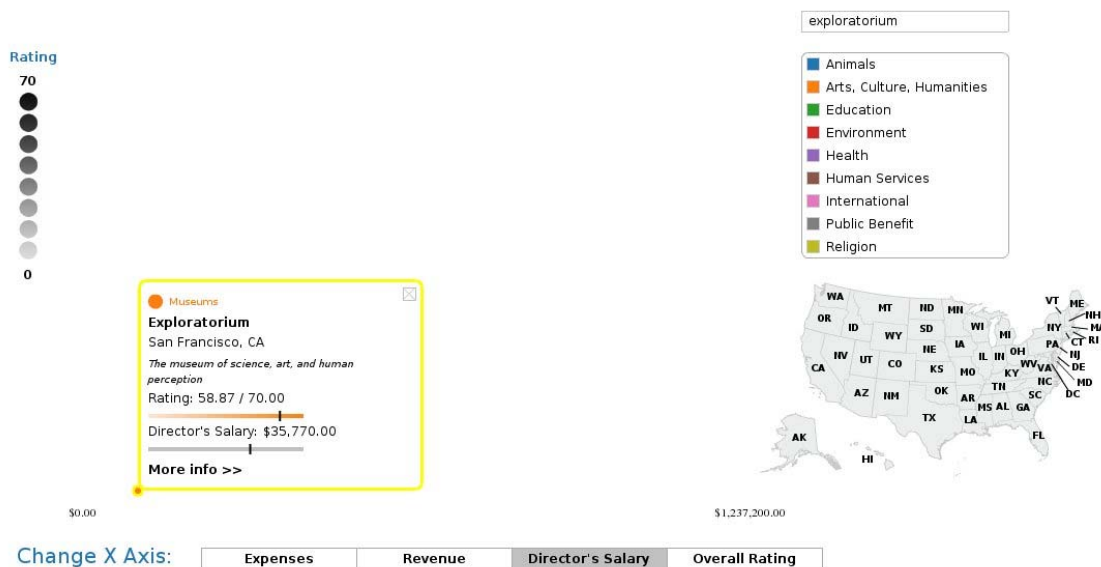

Figure 18: Search

Future Work

Although Charity Vis allows its users to complete the three major tasks we initially envisioned, there remains space for improvement. Some of those possibilities are outline here.

- *Mini info boxes with just the charity name that appear on rollover, like tooltips.*

  Right now it's impossible to know what charities are represented without selecting them invidually. Displaying the charity name on rollover might make it easier for the user to explore the space, though user testing would be required to ensure that the effect was not visually overwhelming.

- *Allowing users to drag charities to a separate space and then explore those particular charities in more detail.*
  Charity Vis currently offers an "all-or-nothing" scenario, in which users must lose focus on the main visualization in order to examine a charity more closely. If we could provide a space to store individual points for later perusal, users could browse more casually, and then compare the charities in their "shopping cart" with one another in more detail.

- *Direct donation linked up to the nodes, perhaps at the more info page.*
  Charity Navigator and GuideStar both offer browsers ways to donate money to charities through their sites. Besides being an easy way to help resources arrive at the charities that need them, direct donation would perhaps encourage charities to maintain more accurate and complete profiles.

- *Selecting multiple charities at once.*

  Though the functionality itself is not particularly dramatic, it could enable a variety of comparative evaluations, much like the separate space scenario described above.

- *Brushing and linking from selected charities to the legend and map. Dynamically filtering those ranges as the user expands or contracts them.*

  As it stands, the patterns produced by the visualization provoke a lot of questions. A good way to filter along user-supplied ranges or patterns, such as via brushing and linking, would go a long way toward potentially answering some of those questions and would enable the kind of fine control an ideal Charity Vis might offer.